UiT The Arctic University of Norway

Faculty of Science and Technology
Department of Physics and Technology

Advancing Deep Learning with Emphasis on Data-Driven Healthcare

Kristoffer Knutsen Wickstrøm

A dissertation for the degree of Philosophiae Doctor - July 2022

This thesis document was typeset using the *UiT Thesis LaTeX Template*.
© 2022 – `http://github.com/egraff/uit-thesis`

Abstract

Vast amounts of data are being collected at hospitals across the world on a continuous basis, and exploiting this data will play a vital role in ushering a new generation of healthcare. A promising direction for exploiting the aforementioned data is through data-driven methods, which are methods that learn to perform tasks based on patterns in collected data. Great advances have been made in data-driven healthcare over the last couple of years, with algorithms reaching clinician-level performance on some tasks, or solving challenges that seemed impossible for automatic systems only a few years ago.

A key driving force behind these advances in contemporary data-driven healthcare is deep learning. The success of deep learning is often attributed to its ability to automatically extract relevant features from data without the need for hand-crafted features. However, this ability also has limitations, since the complex feature extraction process introduces some fundamental challenges for deep learning-based systems in data-driven healthcare. Deep learning algorithms lack explainability and do not provide a notion of uncertainty. If these challenges are not tackled, data-driven healthcare systems based on deep learning algorithms will lack trustworthiness and reliability. Moreover, deep learning-based systems struggle when tasked with learning from unlabeled data. As most healthcare data is unlabeled, this is a fundamental limitation that needs to be addressed to exploit healthcare data in an efficient manner. Towards tackling these challenges, we present four lines of work where we develop new explainability, uncertainty, and unsupervised learning methodology.

In the intersection between explainability and uncertainty, we propose two new methods for capturing uncertainty in explanations. We develop the first approach to capture uncertainty in explainability, which is accomplished through a Bayesian method. The usability of the new methodology is illustrated in the context of semantic segmentation of colorectal polyps. We also develop a new ensemble approach for modeling uncertainty in the explanations for clinical time series, which is motivated by common characteristics associated with clinical time series. Further, we show how the uncertainty estimates can be used to create uncertainty-filtered explanations, which are demonstrated to have higher quality and less ambiguity.

In the intersection between explainability and unsupervised learning, we propose the first framework for explaining representations, as opposed to predictions. Using the new framework, we show how it allows for new insights into self-supervised learning, multi-view clustering, and traditional feature extraction techniques. Finally, we present a new self-supervised approach to learning from unlabeled data that exploits domain knowledge to extract clinically relevant features. The new approach is also coupled with the representation learning explainability framework to provide a novel analysis of explainability in representation learning.

While the emphasis in this thesis has been on data-driven healthcare, we believe that the advances introduced in this thesis can play an important role in other domains for designing more reliable and trustworthy deep learning systems that can effectively exploit data with little label information.

Acknowledgements

A number of people deserves praises and thanks for supporting and guiding me through this thesis.

First and foremost, I would like to thank my supervisor Professor Robert Jenssen for his everlasting support. As the supervisor on both my master and doctoral thesis, it is hard to overstate your impact on my journey in academia. You have a knack for knowing when to be supportive and when to demand more, which I think is the sign of a great supervisor. Thank you for always believing in me and all the opportunities you have presented me with.

To my co-supervisors, Associate Professor Michael Kampffmeyer and Associate Professor Karl Øyvind Mikalsen, thank you for your constant support during this thesis. I consider myself very lucky to have had the guidance of such a highly qualified team of researchers and I have learned a great deal from you both. There is no doubt that the quality of my works have been greatly improved by your contributions, to which I am very grateful.

I would also like to thank Professor Gustau Camps-Valls for having me as a guest researcher in the Image Processing Laboratory at the University of Valencia. Despite the stay being cut short due to the outbreak of the coronavirus, I had a great experience and learned a great deal. A special thanks to J. Emmanuel Johnson for letting me stay with him during my visit and showing me the city. One day I will return to experience the Falles!

Another big thanks goes to Associate Professor Marina M.-C. Höhne for hosting me in the Understandable Machine Intelligence Lab at the Technical University of Berlin. Both in terms of scientific and social experiences I enjoyed myself immensely, and hope to keep up the good collaborations we started. Also thanks to Kirill Bykov, Philine Bommer, and Anna Hedström for sharing their time and office with me. I hope our paths cross again.

To the entire machine learning group at the University of Tromsø. Being part of a large research group is undoubtedly beneficial when it comes to doing good and interesting research. However, the benefit of having colleagues to share

lunch and coffee breaks with, highs and lows, and social interactions cannot be emphasized enough. A great thanks to everyone in the group for making my PhD journey a highly enjoyable experience.

To my friends and family, thank you for supporting and encouraging me throughout these four years. I am looking forward to spending more time with all of you after the delivery of the thesis.

Lastly, to my dear Sigrid. If somebody told me earlier this year that the delivery of my thesis would not be the highlight of the year I would have never believed them. However, the delivery of the thesis has been eclipsed by the expected birth of our twins later this year. I cannot wait to experience the joys and struggles of raising them together with you.

Contents

Abstract i

Acknowledgements iii

List of Figures vii

List of abbreviations ix

1 Introduction 1

1.1 Key challenges 2
1.2 Key objectives 4
1.3 Key solutions 4
1.4 Brief summary of included papers 5
1.5 Other papers 7
1.6 Reading guide 9
1.7 Open science 11

I Methodology and context 13

2 Deep learning 15

2.1 Multilayer perceptrons 18
2.2 Convolutional neural networks 21

3 Explainability in deep learning 23

3.1 An XAI taxonomy 27
3.2 Propagation-based explainability 27
3.3 Perturbation-based explainability 28
3.4 Example-based explainability 29
3.5 Surrogate-based explainability 30

4 Uncertainty in deep learning 31

4.1 Bayesian methods 33
4.2 Ensemble methods 34

4.3 Test-time augmentation 35

5 Self-supervised deep learning **37**
5.1 Contrastive self-supervised learning 39
5.2 Clustering-based self-supervised learning 40
5.3 Siamese self-supervised learning 40

6 Data-driven healthcare **41**
6.1 Deep learning in data-driven healthcare 42

II Summary of research and concluding remarks **47**

7 Paper I **49**

8 Paper II **51**

9 Paper III **55**

10 Paper IV **57**

11 Concluding remarks **59**
11.1 Limitations and future work 60

III Included papers **63**

12 Paper I **65**

13 Paper II **85**

14 Paper III **97**

15 Paper IV **125**

Bibliography **139**

List of Figures

1.1 Overview of papers. 6
1.2 Included paper hierarchy. 10

2.1 Change in representation by neural network. 16
2.2 Illustration of simple MLP 18
2.3 Illustration of regularization in deep learning. 20
2.4 Illustration of simple CNN architecture. 21
2.5 Illustration of convolution operation. 21
2.6 Illustration of max pooling operation. 22

3.1 Example of explainability in the classification setting 24
3.2 A CNN classifier exploits spurious artifacts. 26
3.3 A taxonomy of XAI methods. 27

4.1 Uncertainty in simple MLP. 32
4.2 Overview of uncertainty modeling methods. 32

5.1 Illustration of a self-supervised framework 38

6.1 Overview of data-driven healthcare 43

7.1 Illustration of methodology in Paper I 50

8.1 Illustration of methodology in Paper II 52

9.1 Illustration of methodology in Paper III 56

10.1 Illustration of methodology in Paper IV 58

List of abbreviations

AI artificial intelligence

BYOL boostrap your own latent

CAM class activation mapping

CBIR content-based image retrieval

CNN convolutional neural network

CT computed tomography

EHR electronic health records

FCN fully convolutional network

Grad-CAM gradient-class activation mapping

LIME local interpretable model-agnostic explanations

LRP layerwise relevance propagation

MLP multilayer perceptron

MRI magnetic resonance images

NLP natural language processing

PET positron emission tomography

ProtoPNet prototypical part network

ReLU rectified linear unit

RISE randomized input sampling explanation

SGD stochastic gradient descent

SHAP Shapley additive explanations

SVM support vector machines

TCAV testing with concept activation vector

XAI explainable artificial intelligence

1

Introduction

Vast amounts of data are being collected at hospitals across the world on a continuous basis, and using this data will play a vital role in ushering a new generation of healthcare [1, 2]. A promising set of algorithms for filling this role are data-driven algorithms, which learn to perform a task by recognizing patterns in data. These algorithms improve their performance when given more data [3, 4], and can therefore keep improving with more examples.

Recent research has shown how data-driven algorithms can significantly improve automatic support systems for healthcare applications. Esteva et al. [5] demonstrated how a system for automatic classification of skin lesions could achieve performance comparable with domain-experts. Such a system could have a major impact as a low-cost healthcare solution on a global scale if implemented in mobile devices. Campanella et al. [6] developed a system for computational pathology that could provide comparable performance with domain-experts, and in some cases even surpass their performance. One promising use case for this system is to exclude slides in whole slide images and reduce the workload of pathologist, which are domain experts in high demand [7]. Kuttner et al. [8] proposed a framework for automatically extracting an arterial input function directly from positron emission tomography (PET) images. This could have a major impact in dynamic PET since it would alleviate the need for blood sampling in diagnosis, a process that is both time-consuming and painful for the patient. These are just some examples of how data-driven algorithms could aid in providing better and more efficient treatment.

A major driving force behind these advances in contemporary data-driven healthcare is deep learning [9], particularly in critical healthcare domains such as computer vision [10, 11] and natural language processing (NLP) [12, 13]. The success of deep learning is often attributed to its ability to automatically extract relevant features from data without the need for hand-crafted features [14]. However, this ability is a double-edged sword, since the complex automatic feature extraction process introduces some fundamental challenges for deep learning-based systems in data-driven healthcare. Deep learning algorithms lack explainability [15], do not provide a notion of uncertainty [16], and struggle when tasked with learning from unlabeled data [17]. If these challenges are not tackled, data-driven healthcare systems based on deep learning could lack trustworthiness, and might not be able to exploit healthcare data efficiently.

The goal of this thesis is to tackle these challenges by developing new methodology in the field of deep learning. These challenges are presented in the following section, and addressed in the included papers of this thesis.

1.1 Key challenges

This thesis will focus on three key challenges for data-driven healthcare: (1) the lack of explainability, (2) how to model uncertainty, and (3) learning from unlabeled data. These challenges and related work on addressing these challenges in the context of data-driven healthcare will be discussed in more detail in Chapter 6.

Lack of explainability

A fundamental problem in deep learning is the lack of explainability. This lack has been highlighted as one of the major factors that impedes data-driven healthcare based on deep learning from being implemented in clinical practice [18]. Explainability refers to the ability to explain why a particular prediction was made, typically by indicating what input features are most important to the prediction. Without this ability, healthcare providers could be reluctant to fully trust the system, since it is well known that deep learning-based systems can exploit artifacts and confounding factors to make their decision [19, 20].

Recently, major advances have been made within the within the field of explainable artificial intelligence (XAI) [21, 22], which aims at tackling the lack of explainability. However, there are still major issues in the field of XAI that have been left unattended. First, explanations are often presented without any notion of uncertainty. This can give an unwarranted trust in the reliability of an explanation, which could deteriorate trust instead of enforcing it. For instance, uncertainty quantification of importance maps plays an important role in field of computational neuroimaging [23], but such considerations are not present in current XAI methodology. While some initial studies have looked into uncertainty in explainability [24, 25], very little work have been done in this direction. Second, XAI methods have mainly been focused on explaining scores in the form of predictions or decisions. But this excludes many important deep learning models that do not produce such a score. For instance, representation learning through self-supervision has gained a lot of recent attention [26]. In such frameworks, the output is typically a vector representation that can be used for other tasks. Some work have made strides towards unsupervised explainability [27], but explaining representations is something that current XAI methods are not capable of.

How to model uncertainty

When working with real-world medical problems, uncertainty is unavoidable. Knowing the confidence of a prediction is crucial information when faced with life-or-death decisions. In a recent survey of clinicians, uncertainty modeling was highlighted as a key component for any system intended to be used in clinical practice [16]. Nevertheless, deep learning algorithms does not have the capability to provide uncertainty estimates with its prediction. This significantly limits the potential of deep learning in data-driven healthcare.

Development of methods for capturing uncertainty in deep learning systems have progressed significantly over the last couple of years [28, 29]. Nevertheless, this progression has mainly been focused on modeling uncertainty in predictions. Capturing uncertainty in e.g. explanations has received very little focus, which limits the impact of uncertainty analysis in deep learning.

Learning from unlabeled data

The success of deep learning has been mostly confined to scenarios where data have been accompanied by labels provided by human annotators [17]. However, medical data is typically obtained without label-information, which is why learning from unlabeled data has been highlighted as one of the main obstacles in data-driven healthcare [17]. Labeling medical data can be both costly and time consuming, as it requires efforts from numerous domain experts. Moreover, for challenging and noisy data, domain experts might disagree on the correct annotation, which makes labeling data even more problematic. If data-driven healthcare systems are not able to learn from unlabeled data, large amounts of collected information is rendered irrelevant. Development of deep learning methodology that can learn without labels is therefore of paramount importance if data-driven healthcare is going to fulfil its potential.

The field of unsupervised learning is a fundamental research area in machine learning, which is focused on learning without supervision. Recently, self-supervised representation learning has emerged as a major research direction within unsupervised learning [26], with impressive results on extraction of information without human supervision [30, 31]. However, contemporary self-supervised frameworks for images are developed with natural images in mind, and not customized for the characteristics found in medical images. This hinders the application of self-supervised approaches within data-driven healthcare.

1.2 Key objectives

The key objectives in this thesis is to address the aforementioned key challenges. In particular, our focus will be in the intersection of these challenges. That is, modeling uncertainty in explanations, and how to explain systems that do not rely on labels to learn. These objectives can be summarized as:

1 Develop methodology for modeling uncertainty in explanations.

2 Develop methodology for explaining systems that learn without labels.

3 Develop methodology for learning from unlabeled data.

As a secondary objective, we emphasize on evaluating the proposed methodology on healthcare data.

1.3 Key solutions

We propose two new approaches to undertake the problem of modeling uncertainty in explanations. First, we propose a Bayesian approach coupled with a propagation-based explainability method to capture uncertainty in explanations (Paper I). The proposed method is demonstrated in the context of semantic segmentation of colorectal polyps. Second, we introduce an ensemble approach that captures uncertainty in explanations by measuring the agreement in explanations across ensemble members (Paper II). This approach is applied in the context of classification of clinical time series.

To address the challenges with explaining models that learn from unlabeled data, we introduce the first framework for explaining representations of data (Paper III), as opposed to predictions. Moreover, Paper III also connects with the problem of modeling uncertainty in explanations, as we show how uncertainty can be captured in the proposed framework.

In order to tackle the issue of learning from unlabeled data we propose a new self-supervised approach that incorporates clinical knowledge into the training process (Paper IV), and enables deep content-based image retrieval (CBIR) systems to focus on particular organs in the feature extraction process. This work also connects with the challenge of explaining systems that learn from unlabeled data, since the proposed self-supervised approach cannot be explained without the framework proposed in Paper III.

1.4 Brief summary of included papers

This section presents a list of papers included in this thesis, along with a brief summary of each paper. Additionally, a list of other articles published during this PhD project is included in the next section. Figure 1.1 gives an overview of which challenges and parts of the machine learning field that the included and others papers are associated with. Figure 1.2 displays a hierarchy of the included papers and how they contribute to different parts of different fields.

I Kristoffer K. Wickstrøm, Michael C. Kampffmeyer, Robert Jenssen. "Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps". In Medical Image Analysis, 2020.

II Kristoffer K. Wickstrøm, Karl Øyvind Mikalsen, Michael C. Kampffmeyer, Arthur Revhaug, Robert Jenssen. "Uncertainty-aware deep ensembles for reliable and explainable predictions of clinical time series". In IEEE Journal of Biomedical and Health Informatics, 2020.

III Kristoffer K. Wickstrøm, Daniel J. Trosten, Sigurd Løkse, Ahcène Boubekki, Karl Øyvind Mikalsen, Michael C. Kampffmeyer, Robert Jenssen. "RELAX: representation learning explainability". Submitted to International Journal of Computer Vision.

IV Kristoffer K. Wickstrøm, Eirik A. Østmo, Keyur Radiya, Karl Øyvind Mikalsen, Michael C. Kampffmeyer, Robert Jenssen. "A clinically motivated self-supervised approach for content-based image retrieval of CT liver images". Submitted to Computerized Medical Imaging and Graphics.

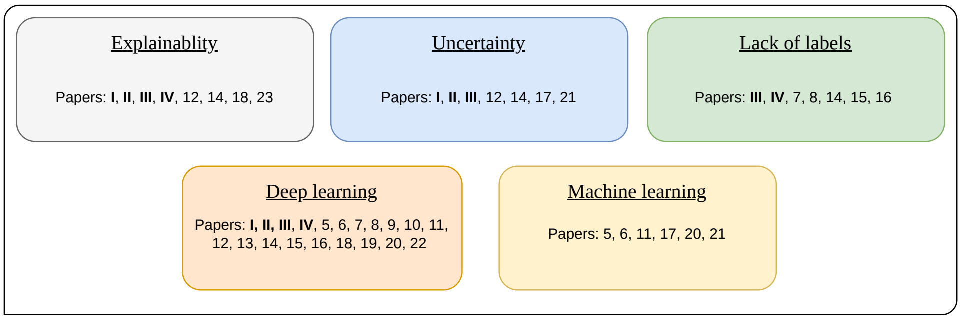

Explainablity
Papers: I, II, III, IV, 12, 14, 18, 23
Uncertainty
Papers: I, II, III, 12, 14, 17, 21
Lack of labels
Papers: III, IV, 7, 8, 14, 15, 16
Deep learning
Papers: I, II, III, IV, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 22
Machine learning
Papers: 5, 6, 11, 17, 20, 21

Figure 1.1: Overview of papers.

Paper I Introduces a new method for modeling uncertainty in the explanations, illustrated in the setting of semantic segmentation of colorectal polyps. We propose to sample explanations from a trained network through a noise injection procedure and compute uncertainty estimates by taking the standard deviation across all samples. We show how incorporating uncertainty is crucial to produce trustworthy explanations and that some parts of an explanations can have higher uncertainty than others.

Paper II Presents a new ensemble-based approach to uncertainty in explanations of clinical time series. The motivation for such an approach is to exploit common characteristics in clinical time series, namely small models and few data samples. These characteristics makes ensembles suitable for clinical time series, since they can be computationally demanding in e.g. computer vision where big models and large datasets are more common. Furthermore, we propose to filter the explanations using the uncertainty estimates, and we show how this provides more concise explanations with higher quality.

Paper III Proposes RELAX, the first method for explaining representations as opposed to predictions. The core idea of RELAX is to measure similarities between the representation of an image and perturbed versions of the same image. We provide theoretical guarantees and analysis of RELAX, and show that seemingly similar deep feature extraction models can utilize very different input features. Also, we show how RELAX can be used to explain hand-crafted feature extractors, and the explanation illustrate why such approaches can provide inferior performance compared to deep learning.

Paper IV This paper introduces a new clinically motivated self-supervised learning framework for CBIR. We propose to exploit know properties of the liver in CT images to train a feature extractor without labels using self-supervised learning. Experiments illustrate how the proposed approach achieves superior performance across several metrics. Moreover, this paper leverages the

RELAX framework to explain the representations produced by the proposed self-supervised framework. These explanations reveal insights into the feature extraction process that would not be obtainable without RELAX.

1.5 Other papers

5 Kristoffer K. Wickstrøm, Sigurd Løkse, Michael C. Kampffmeyer, Shujian Yu, Jose Principe and Robert Jenssen. ”Analysis of Deep Neural Networks using Tensor Kernels and Matrix–Based Renyi's Entropy”. Workshop on Information Theory and Machine Learning, Neural Information Processing Systems 2019.

6 Kristoffer K. Wickstrøm, Sigurd Løkse, Michael C. Kampffmeyer, and Robert Jenssen. ”Modelling the information plane of recurrent neural networks”. Extended abstract and oral presentation at the Northern Lights Deep Learning Conference, 2019.

7 Van Nhan Nguyen, Sigurd Løkse, Kristoffer K. Wickstrøm, Michael C. Kampffmeyer, Davide Roverso and Robert Jenssen. ”SEN: a novel dissimilarity measure for prototypical few–shot learning networks". Workshop on Visual Learning with Limited Labels, Conference on Computer Vision and Pattern Recognition, 2020.

8 Van Nhan Nguyen, Sigurd Løkse, Kristoffer K. Wickstrøm, Michael C. Kampffmeyer, Davide Roverso and Robert Jenssen. ”SEN: a novel dissimilarity measure for prototypical few–shot learning networks”. European Conference on Computer Vision, 2020.

9 Samuel Kuttner, Kristoffer K. Wickstrøm, Gustav Kalda, S Esmaeil Dorraji, Montserrat Martin-Armas, Ana Oteiza, Robert Jenssen, Kristin Fenton, Rune Sundset, Jan Axelsson. "Machine learning derived input-function in a dynamic 18F-FDG PET study of mice". Biomedical Physics and Engineering Express, 2020.

10 Shujian Yu, Kristoffer K. Wickstrøm, Robert Jenssen, Jose C Principe. "Understanding convolutional neural networks with information theory: An initial exploration". IEEE Transactions on Neural Networks and Learning Systems, 2020.

11 Andreas Kvammen, Kristoffer K. Wickstrøm, Derek McKay, Noora Partamies. "Auroral image classification with deep neural networks". Journal of Geophysical Research: Space Physics, 2020.

12 Kristoffer K. Wickstrøm, Karl Øyvind Mikalsen, Michael C. Kampffmeyer, Arthur Revhaug, Robert Jenssen. "Uncertainty-aware deep ensembles for reliable and explainable predictions of clinical time series". Extended abstract and oral presentation at the Northern Lights Deep Learning Conference, 2021.

13 Samuel Kuttner, Kristoffer K. Wickstrøm, Mark Lubberink, Andreas Tolf, Joachim Burman, Rune Sundset, Robert Jenssen, Lieuwe Appel, Jan Axelsson. "Cerebral blood flow measurements with 15O-water PET using a non-invasive machine-learning-derived arterial input function". Journal of Cerebral Blood Flow and Metabolism, 2021.

14 Kristoffer K. Wickstrøm, Daniel J. Trosten, Sigurd Løkse, Ahcène Boubekki, Karl Øyvind Mikalsen, Michael C. Kampffmeyer, Robert Jenssen. "RELAX: representation learning explainability". Abstract and oral presentation at the NOBIM conference, 2021.

15 Daniel J. Trosten, Kristoffer K. Wickstrøm, Shujian Yu, Sigurd Løkse, Robert Jenssen and Michael C. Kampffmeyer. "Deep clustering with the Cauchy-Schwarz divergence". Workshop on Information Theory for Deep Learning, Conference on Artificial Intelligence 2022.

16 Kristoffer K. Wickstrøm, Michael C. Kampffmeyer, Karl Øyvind Mikalsen, Robert Jenssen. "Mixing up contrastive learning: self-supervised representation learning for time series". Pattern Recognition Letters, 2022.

17 Kristoffer K. Wickstrøm, J. Emmanuel Johnson, Sigurd Løkse, Gustau Camps-Valls, Karl Øyvind Mikalsen, Michael C. Kampffmeyer, Robert Jenssen. "The kernelized Taylor diagram". Norwegian Artificial Intelligence Symposium, 2022.

18 Kristoffer K. Wickstrøm. "Hva gjør vi når kunstig intelligens gir oss kunnskap vi ikke forstår?". Forskersonen.no, 2022. `https://forskersonen.no/kunstig-intelligens-meninger-populaervitenskap/hva-gjor-vi-nar-kunstig-intelligens-gir-oss-kunnskap-vi-ikke-forstar/1957326`

19 Samuel Kuttner, Luigi T. Luppino, Kristoffer K. Wickstrøm, Nils T. D. Midtbø, S. Esmaeil Dorraji, Ana Oteiza, Montserrat Martin-Armas, Kristin Fenton, Laurence Convert, Otman Sarrhini, Roger Lecomte, Rune Sundset, Michael C. Kampffmeyer, Robert Jenssen. "Deep learning derived input-function in dynamic 18F-FDG PET imaging of mice". Extended abstract and top-rated oral presentation at the Annual Congress of the European Association of Nuclear Medicine, 2022.

20 Kristoffer K. Wickstrøm, Sigurd Løkse, Michael C. Kampffmeyer, Shujian Yu, Jose Principe and Robert Jenssen. "Information plane analysis of deep neural networks via matrix-based Renyi's entropy and tensor kernels". Submitted to Pattern Recognition.

21 Ane Blázquez-García, Kristoffer K. Wickstrøm, Shujian Yu, Karl Øyvind Mikalsen, Ahcene Boubekki, Angel Conde, Usue Mori, Robert Jenssen, Jose A. Lozano. "Selective imputation for multivariate time series datasets with missing values". Submitted to Transactions on Knowledge and Data Engineering.

22 Andreas Kvammen, Kristoffer K. Wickstrøm, Samuel Kociscak, Jakub Vaverka, Libor Nouzak, Arnaud Zaslavsky, Kristina Rackovic, Audun Theodorsen, Amalie Gjelsvik, David Pisa, Jan Soucek, and Ingrid Mann. "Machine learning classification of dust impact signals observed by the solar orbiter". Submitted to Annales Geophyicae.

23 Anna Hedström, Kristoffer K. Wickstrøm, Dilyara Bareeva, Wojciech Samek, Sebastian Lapuschkin, Marina M-C Höhne. "Can I count on you?: scrutinising the evaluation of AI explainers". Submitted to IEEE Transactions on Artificial Intelligence.

1.6 Reading guide

The remainder of this thesis is organized into three parts; methodology, summary of research, and included papers. The "methodology"-part consists of five chapters that introduce the relevant background material for all papers. Chapter 2 presents an overview of the essential components of deep learning and introduces convolutional neural networks. Chapter 3 introduces the field of XAI and gives a short overview of different explainability methods. Chapter 4 outlines uncertainty modeling in deep learning and presents Bayesian, ensemble, and test-time augmentation methods for uncertainty estimation. Chapter 5 presents the main components of self-supervised deep learning and briefly introduces contrastive, clustering, and siamese-based self-supervised learning. Chapter 6 gives an overview of data-driven healthcare, with a particular focus on the role of deep learning. The "summary of research and concluding remarks"-part consist 5 chapters, where Chapter 7-10 provides a brief overview of the scientific contributions of each paper in this thesis. Additionally, Chapter 11 includes some concluding remarks and discusses the limitations and potential future works based on our research. Lastly, the "included papers"-part contains the research papers included in this thesis.

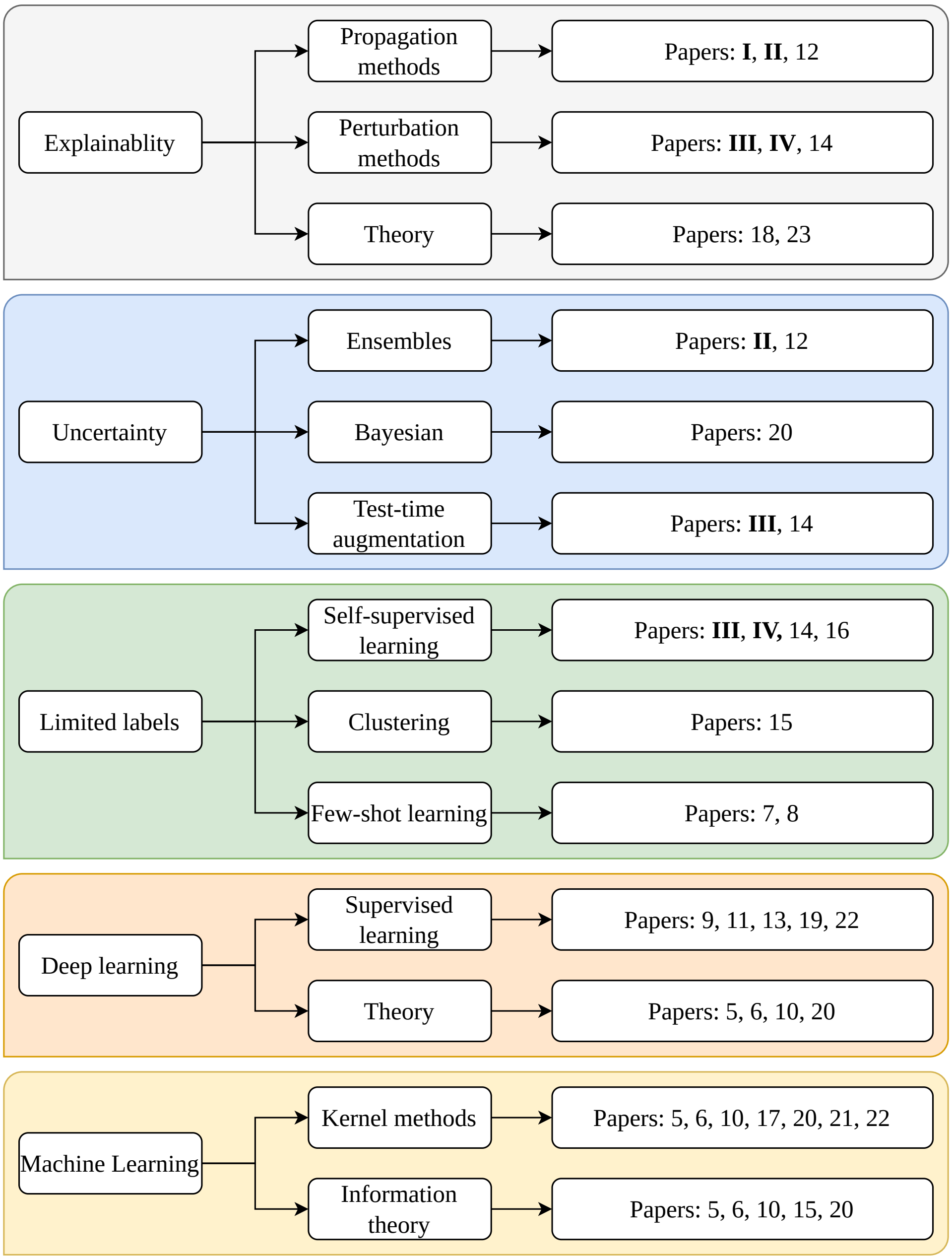


Figure 1.2: Included paper hierarchy.

1.7 Open science

Reproducibility is becoming increasingly important in all areas of science [32]. In deep learning, making research open could be achieved by sharing resources such as code and data, or making sure that all necessary details to reproduce an experiment is openly available. Towards making the research conducted in this thesis as open as possible, we have made code and other resources publicly available. These resources are further described in relation to each research paper in Chapter 7-10.

Part I

Methodology and context

2

Deep learning

Deep learning is part of the representation learning field, where the goal is to learn a data representation that is beneficial for performing some task [33, 34]. In deep learning, the new representation is created through a stack of successive transformations, where the transformations are parametrized by a neural network. Neural networks are trained to find patterns in large sets of data by adjusting its internal parameters to minimize some desired mathematical objective. The fact that neural networks can learn to automatically extract useful features directly from raw data is one of the major advantages compared to competing machine learning methods, which requires hand-crafted features to achieve good performance.

Figure 2.1 illustrates how altering the representation of data can be beneficial to perform a desired task. This example is concerned with binary classification of 2-dimensional data that are not linearly separable. The input data is shown in the leftmost plot in Figure 2.1. A simple neural network is trained to classify samples into two classes. The middle plot shows how the input data is first transformed into a 3-dimensional representation by the trained network. In this representation, the two classes are now linearly separable and much easier to distinguish. The rightmost plot shows the final transformation from the 3-dimensional representation to a 1-dimensional representation. Compared to the original representation, the data is now easily separable with a linear classifier. While the example shown in Figure 2.1 is simple, the core procedure it depicts is the same as in larger and more complex neural networks used to solve real-world tasks.

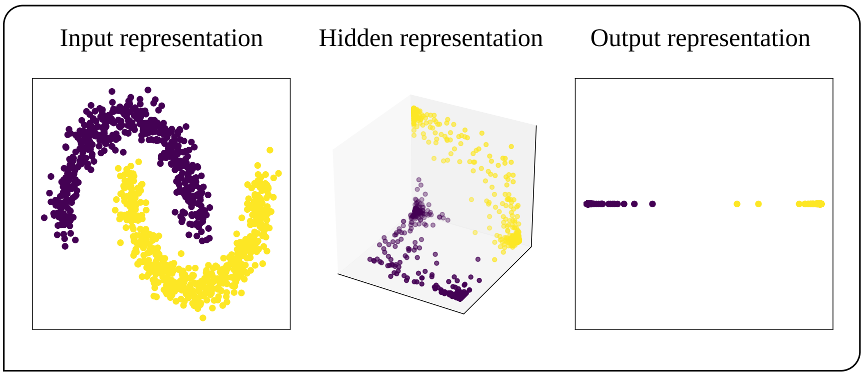


Figure 2.1: Illustration of how a simple neural network changes the data representation to solve a binary classification task.

The origin of deep learning is often traced back to the McCulloch-Pitts neuron [35], the perceptron algorithm [36], and the first learning-based neurons by Widrow and Hoff [37]. Some also argue that one could go back even further [38], to works from the early 1800s. Early research on artificial intelligence was inspired by biological systems, and neurons were designed to mimic neurons and synapses in the human brain. It is based on such early research that the name artificial neural networks first occurred, or simply neural networks. Early artificial intelligence (AI) research showed some promise, but went into a long period of little attention after a seminal paper by Minsky and Papert [39]. Minsky and Papert showed that the perceptron was unable to solve the XOR problem, a simple binary classification task. While this could potentially be addressed by stacking multiple layers of perceptrons in succession, the sentiment at the time was that the computational resources and algorithms for training such models were too limited for neural networks to be successful. The period following the work of Minsky and Papert is often referred to as the first winter of AI.

In the late 1980s and early 1990s, the development of the backpropagation algorithm [40, 41, 42] and improvements in computers brought new interest in neural network-based AI research, a period called the first neural network renaissance [43]. This era saw some initial papers on unsupervised training of neural networks [44, 45, 46] and impressive performance on real-world tasks such as in digit [47] and fingerprint recognition [48]. Despite such improvements, deep neural networks remained hard to train, which hampered their usability. Moreover, a milestone work by Hochreiter [49] identified the vanishing and exploding gradient problem, a fundamental challenge related to the training of neural networks. Following Hochreiter's work, neural network research entered the second winter of AI with support vector machines (SVMs) [50, 51] and random forest algorithms [52, 53] taking the center stage.

In the mid 2000s, unsupervised pretraining through restricted Boltzmann machines [54] and autoencoders [55] improved the stability and reliability in training of deep neural networks. These improvements, combined with better computational resources, more data, and theoretical advances in training algorithms led to neural network-based AI research entering the second neural network renaissance [43], a period we are still experiencing today. The convincing victory by Alexnet [14] in the 2012 edition of the ImageNet large scale visual recognition challenge [56] is often recognized at the start of the deep learning revolution. Alexnet was much deeper (had more layers) and contained many more trainable parameters than previous networks, and this increase in complexity has continued to the present day.

Today, deep learning is the de facto standard in important domains such as computer vision [10, 57] and NLP [58, 59]. The Alexnet architecture that ushered the current period was considered enormously complex at the time (in terms of the number of parameters). But recent deep learning architectures have increased the amount of parameters from millions to billions [58], and it has been shown that this increase can be crucial for performance [60, 58]. Given this over-parametrization, deep learning ignores the dangers of overfitting (memorizing the data) described in traditional learning theory [61], which should lead to worse performance. This is not the case, since deep learning regularly have more parameters than data points to train on and still show excellent performance on unseen examples. At this time, there are no definite explanations for the success of over-parametrized models, but an interesting research direction is through the concept of an inductive bias in neural networks. Belkin et al. [62] recently provided evidence for the double-descent phenomenon, in which heavily over-parametrized models that are capable of perfectly fitting the training data defied the traditional bias-variance trade-off and increased performance as the number of parameters increased. Belkin et al. hypothesized that the specific optimization used to train neural networks can lead to an inductive bias towards low-norm configurations that generalize well to unseen data. Theoretical analysis of simple neural networks have shown the presence of such an inductive bias [63, 64, 65, 66], but it remains to be shown for more complex architectures used in real-world applications.

This chapter provides a brief review of the core components in deep learning, which form the bedrock of this thesis.

2.1 Multilayer perceptrons

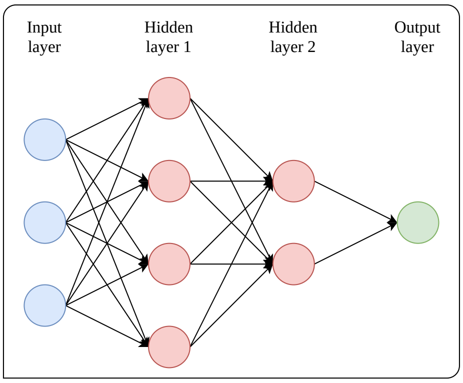


Figure 2.2: MLP with three hidden layers consisting of four and two units and one output unit.

Multilayer perceptrons (MLPs) form the basis for deep learning. They are constructed by stacking layers of transformations in succession. Each layer contains a number of units, commonly referred to as neurons. For clarity purposes, we will focus our discussion on MLPs in the setting of supervised classification. The goal of supervised learning is to learn a function f that transforms data from an input space X to an output space Y, i.e. $f : X \mapsto Y$. This goal is achieved by minimizing a loss function $L(f(\mathrm{x}), y)$ using a finite dataset of N sample pairs $D = \{(\mathrm{x}_i, y_i),\ i = 1, \cdots, N\}$, where x is a sample from the input space and y is a sample from the output space that indicates the desired output. In addition to minimizing the loss, f should also generalize and perform well on unseen samples from the input space. Figure 2.2 shows an example of a simple MLP.

The fundamental computation in a MLP takes place inside each unit. For a single sample, unit j in layer l computes a weighted sum of the output from the previous layer:

$$a_j^l = \sigma\Big(\sum_{j=1}^{k_l}\sum_{k=1}^{k_{l-1}} w_{jk}^l a_k^{l-1} + b_j^l\Big), \tag{2.1}$$

where b_j^l is the bias of the jth unit in the lth layer, w_{jk}^l is the weight connecting the kth unit in the $(l-1)$th layer with the jth unit in the lth layer, k_l is the number of units in the lth layer, k_{l-1} is the number of units in the $(l-1)$th layer, a_k^{l-1} is the output of the kth unit in the $(l-1)$th layer, and σ is the activation function. The computation in Equation 2.1 is carried out for each neuron in layer l, such that all neurons in layer l is connected with all neurons in layer $l-1$. Therefore, a single layer in a MLP is often referred to as a fully connected layer.

Activation function The activation function acts as an approximate unit step function that indicates if the neuron is "firing", and enables the MLP to learn non-linear transformations. Traditionally, the sigmoid activation function was often used (Equation 2.2). However, modern networks mostly use the Rectified Linear Unit (ReLU) activation function (Equation 2.3), as it improves the gradient flow and allows for training of deeper networks [67]

$$f_{\text{sigmoid}}(x) = \frac{1}{1 + \exp(-x)} \tag{2.2}$$

$$f_{\text{ReLU}}(x) = \max(0, x). \tag{2.3}$$

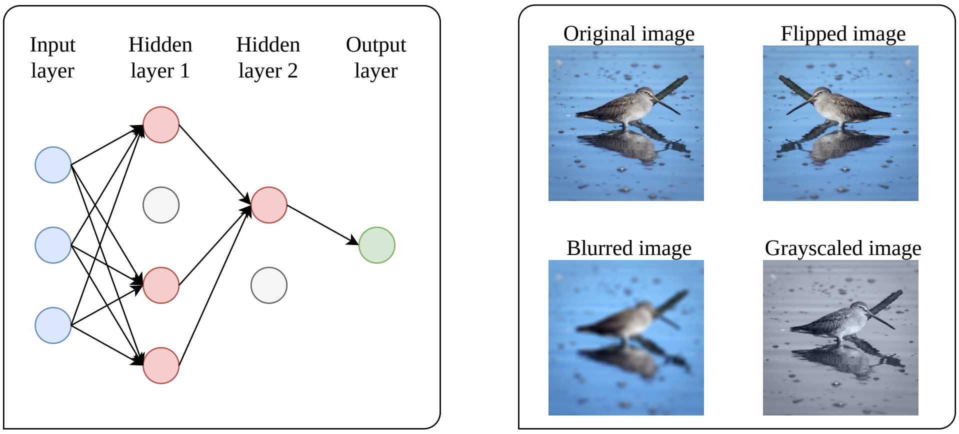
Additionally, an activation functions suited for the specific task is typically employed in the output layer. For the task of classification, the softmax activation function, $f_{\text{softmax}}(x)_c = e^{x_c} / \sum_{i=1}^{C} e^{x_i}$, is mostly used, where C is the number of classes. The softmax function maps the input values into the range $(0, 1)$ and guaranties that they sum to 1, such that the output of the softmax function can be interpreted as pseudo-probabilities for each class.

Loss function and optimization The loss function assesses how well the MLP is performing the desired task. In the setting of classification, the cross-entropy loss functions is a popular option, and is defined as:

$$L(f(\mathrm{x}), y)_{CE} = \frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} f(\mathrm{x})_c(i) \log(y_c(i)). \tag{2.4}$$

The loss function provides a signal that guides the training procedure, which is typically conducted through some variant of gradient descent. A variant of gradient descent that is used regularly is stochastic gradient descent (SGD), but more sophisticated alternatives like the ADAM [68] or LARS optimizer [69] are also popular choices. The core idea in any gradient descent optimizer is to adjust the parameters in such a way that the loss function is minimized. In deep learning, the gradient of the loss function with respect to all the learnable weights and biases in the network is computed through the backpropagation algorithm [40, 41, 42].

Regularization Deep learning algorithms can have a large number of learnable parameters that allow them to learn complex relationships. However, this also enables them to fit the data they are trained on perfectly, which can result in algorithms that do not generalize well to unseen data. The problem of memorizing the training data is referred to as overfitting and is a fundamental problem in deep learning. Many techniques exists to combat overfitting, and these techniques are referred to as regularization techniques. Below we present two widely used regularization techniques that are also important in the context of uncertainty modeling and learning from unlabeled data.



(a) (b)

Figure 2.3: Illustration of (a) dropout applied in a MLP and (b) common augmentations for natural images. Image is taken from the Imagenet dataset [70].

- **Dropout** is a stochastic regularization technique that randomly drops units in the network during training [71]. The motivation for the dropout procedure is to avoid co-adaptation in the units, but it can also considered as an ensemble approach of thinned networks [71]. A simple MLP with dropout applied is illustrated in Figure 2.3. Dropout was originally intended to be used during training. However, Gal and Ghahramani [72] proposed that Dropout could be used after training to model uncertainty, by sampling predictions from thinned versions of the trained network.

- **Data augmentation** is a technique to tackle overfitting by exploiting known invariances in the data to increase the amount of training data. For instance, if an object is invariant to rotation, i.e. it does not change characteristics after being rotated, randomly rotating images can be incorporated in training to create more training samples. This artificially increases the size of the training data which makes it harder to fit completely. A set of common augmentations are illustrated in Figure 2.3. Furthermore, data augmentation also plays a key part in recent research on self-supervised learning [30, 73], where useful representation are learned by exploiting known invariances to common data augmentations.

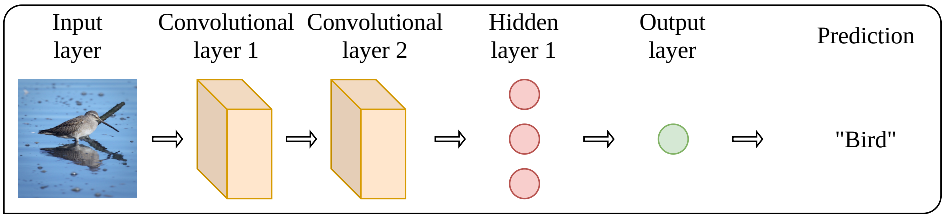


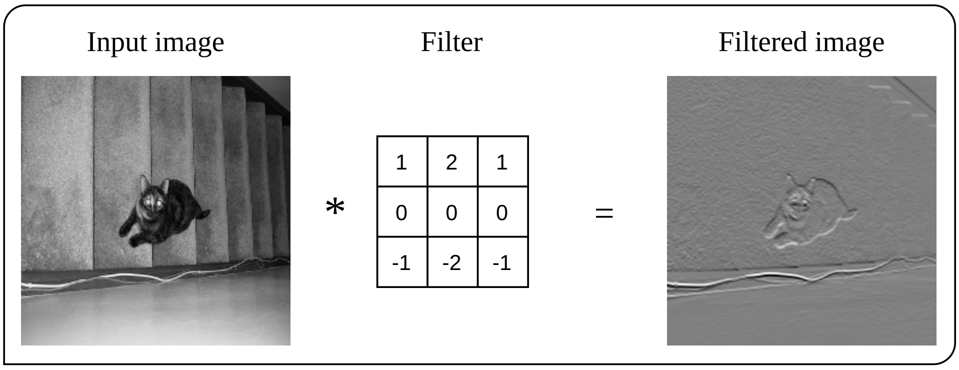
Figure 2.4: Illustration of simple CNN architecture.

2.2 Convolutional neural networks

A convolutional neural network (CNN) is a neural network where one or more layers are convolutional layers, as illustrated in Figure 2.4. Convolutional layers are layers that process the output from the previous layer through the convolution operation. Mathematically, the convolution operation measures the overlap of of two functions

$$s^*(t) = (s * k)(t) = \int s(a)k(t-a)da, \tag{2.5}$$

where s is the input signal, k is a filter, and s^* is the filtered version of s. Figure 2.5 displays an example where an image is convolved with a simple edge-detecting filter, resulting in a filtered version of the original image. Different types of filter will activate on different parts of the input, and can therefore be used to extract information. Prior to deep learning, filters for extracting information from e.g. image data were hand-crafted by researchers [74]. In CNNs, the filters in in each layers are learnt as a part of the optimization procedure, which results in improved performance compared to hand-crafted feature extractors [74].



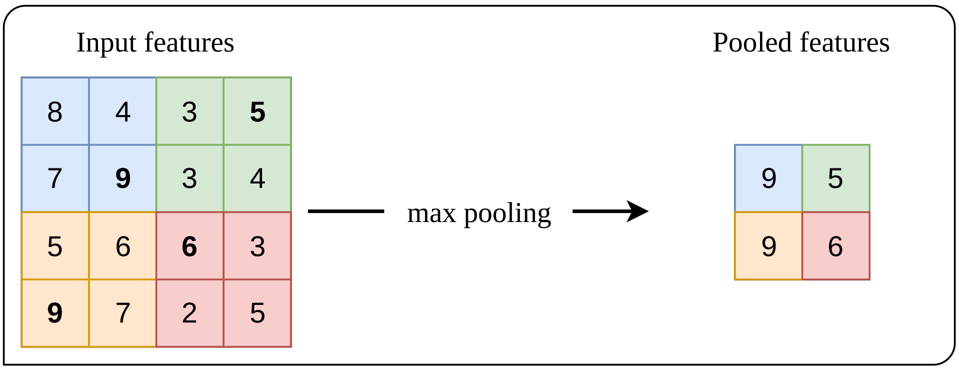
1	2	1
0	0	0
-1	-2	-1

Figure 2.5: Image taken from PASCAL VOC [75] convolved with horizontal edge detector.

LeCun et al. [33] argue that there are some key motivations for convolutional layers in deep learning architectures. First, in grid-like data such as time series, images, and video, there is often high correlation between values in local areas which form distinct motifs. These motifs can be detected by localized filters that are slid across the input through the convolution operation, a process which has several beneficial effects. It greatly reduces the number of learnable parameters compared to a fully connected layer, since you avoid the need to connect each neuron in one layer with the next. Also, the motifs can typically appear in any region of the grid-like input data, which allows the filter to be reused and parameters shared across the entire input. Lastly, the convolution operation introduces spatial invariances into the network, since the filter detects an object even if it has moved to another location in the input.

The second motivation is that many natural signals are composed in a hierarchical structure, where low-level features like edges and shapes are combined into high-level features like motifs or object parts. The layerwise structure in CNNs allows the signal to be decomposed in a similar fashion. Filters in the lower layers of CNNs have been shown to resemble Gabor filters [14], and detect basic components such as shapes and edges. Filters in the higher layers combine these basic components into high-level features, which are then used to identify objects in e.g. images or video.

A common component in CNNs is the pooling operation, which computes a summary statistic of a local region. Typical approaches to summarization is taking the average across the region or the maximum value within the region. The purpose of the pooling operation is twofold. First, summarizing local regions introduces invariance to slight shifts in the input. Second, by summarizing regions into points the spatial resolution of the input grid is reduced, which reduces the computational demand. An illustration of the max pooling operation is shown in Figure 2.6.



8	4	3	**5**
7	**9**	3	4
5	6	**6**	3
9	7	2	5

9	5
9	6

Figure 2.6: Illustration of max pooling operation.

3

Explainability in deep learning

Holzinger et al. [22] summarized the field of explainable artificial intelligence (XAI) as a tool to answer the question of "why?". Why does the algorithm think an image contains a certain object? Why did it make a mistake on this particular example? Why does it disagree with an experienced physicians for a particularly challenging illness to diagnose? Such questions will arise in almost all scenarios where automatic support systems are part of the decision process, and particularly in a safety critical domain such as healthcare. Answering the question of "why?" is crucial to create trustworthy, reliable, and informative automatic decision systems.

The topic of explainability has been investigated for a long time [76, 77], and explainability methods for neural networks were developed already in the mid 1990s [78]. But it was not until deep learning became a prominent force in machine learning that the need for XAI became critical. The lack of explainability was one of the most common criticisms during the start of the second neural network neural network renaissance [79]. Deep learning algorithms were regularly referred to as "black boxes" [15, 80], and the missing transparency was highlighted as a major obstacle for deep learning in healthcare applications [17]. However, major advances in XAI have been made over the last couple of years, and the field is now an important branch of deep learning research.

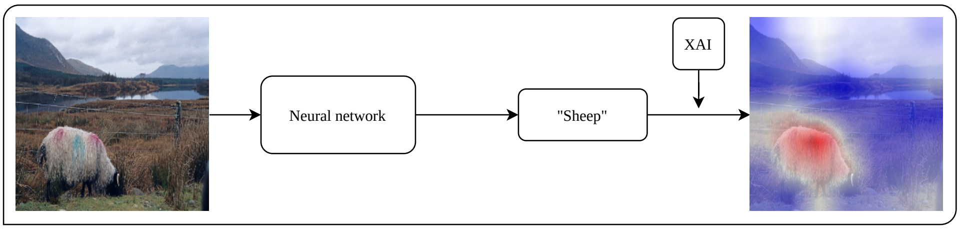


Figure 3.1: Example of explainability in the image classification setting. An image is passed through a neural network and classified as sheep. XAI allows for inspection of what pixels influence the prediction, which in this case are pixels associated with the sheep. Image is taken from PASCAL VOC [75].

Figure 3.1 illustrates a simple example of XAI in the classification setting, where an image is classified to the "sheep" class. Prior to the development of XAI, there were no way of investigating what features influenced the prediction of the model. Nobody could be sure if the network had learned features that actually correspond to the sheep or just some features associated with the sheep. For instance, Ribeiro et al. [81] showed how a CNN classifier trained to distinguish wolfs and huskies, learned to exploit the snowy background that was commonly present for the wolf class instead of features corresponding to the actual wolfs. But with the entrance of XAI, deep learning algorithms could now explain their predictions, and heatmaps such as the one shown in the rightmost part of Figure 3.1 could be used to investigate what in the input influenced the sheep prediction.

Why do we need explainability? The need for explainability in deep learning is regularly highlighted, and in particular in data-driven healthcare [18, 16]. But why do we need explainability? Lipton [82] motivates the need for explainability through the following desiderata:

- **Trust** There are several ways to think about trust in the context of deep learning. If the performance of a particular model is known to be very high, it might be trusted to make correct predictions. But can it be trusted in making predictions based on relevant input features? One well-known limitation of deep learning algorithms is that they can exploit confounding factors or artifacts in the data to make predictions, so called Clever Hans predictors [19]. Figure 3.2 shows an example from Gautam et al. [20] of a deep learning system that exploits artifacts in X-ray images to make predictions. Such a system could perform well in the design phase, but would fail when put into production. This example illustrates how XAI can be used to add another layer of trust to deep learning in addition to having a system with high precision, namely ensuring that predictions are based on relevant input features.

- **Causality** An enticing prospect for XAI is as a tool to establish causal connections between input features and a particular phenomenon. Machine learning algorithms are trained to identify patterns and not to establish causal relationships, but through XAI they can be used to guide domain experts towards identifying such relationships. In the context of healthcare, a precise predictive model combined with XAI could be used to produce hypotheses about the relationship between input features and a particular disease, which could later be investigated by domain experts. However, care must be taken when constructing such hypotheses, as several sets of input features might produce models with similar performance due to the Rashomon effect [83].

- **Transferability** Deep learning algorithms are often evaluated by splitting a dataset into a training and testing part. The model is trained using the training part and evaluated using the testing part. In this setting, the training and test data come from the same distribution, but this is often not the case when applying algorithms in real-world applications. For instance, a model trained on data from on hospital might vulnerable to distributional shift in the data when applied on data from a new hospital. XAI can be used to investigate if a model has learned to use patterns that can generalize to new settings or if they have picked up pattern that are domain-specific.

- **Informativeness** Predictive algorithms are often used as tools in an exploratory setting to guide domain experts when investigating new data. In such cases, providing more information than just a binary prediction is important to make informed and reliable decisions. If a patient is identified as having a particular disease, this would provide some information for the investigator. But if the algorithm also indicated that a particular input feature was important for identifying the disease, it could guide the investigator to look for similar patterns in other patients and aid in decision making.

- **Fair and ethical decision-making** Deep learning is being integrated into an increasing amount of domains where ethical concerns are of critical importance. For instance, in criminal justice, machine learning can be used to predict areas of potential criminal activity [84]. In this setting, explainability is important to ensure that the system takes ethical consideration into account, such as not making racially biased decisions. In some cases it can also be a legal requirement to provide an explanation for an automatic decision, such as the European Union's general data protection regulation ("right to an explanation") [85].

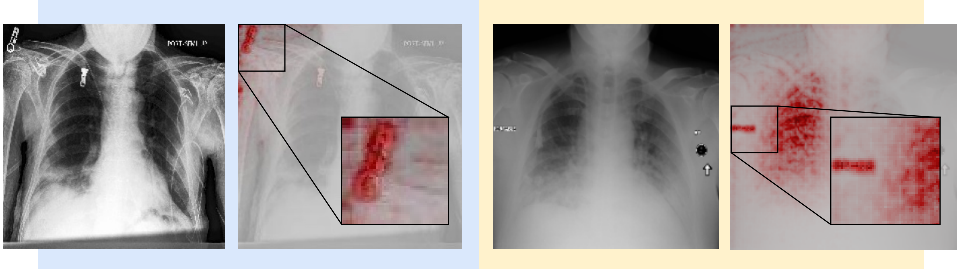


Figure 3.2: A CNN classifier exploits spurious artifacts in the data to classify images as pneumonia or non-pneumonia. Example from Gautam et al. [20] with permission of author.

Model-aware versus model-agnostic explainability An important distinction in explainability models is whether or not they require access to the inner workings of the model they want to explain. Methods that requires such information are referred to as model-aware or white-box methods, while methods that do not require information about the model are called model-agnostic or black-box methods. Randomized input sampling explanation (RISE) [86] and local interpretable model-agnostic explanations (LIME) [81] are two well-known model-agnostic methods, while gradient-class activation mapping (Grad-CAM) [87] and layerwise relevance propagation (LRP) [88] are widely-used model-aware methods. The great advantage of model-agnostic methods is that they are highly flexible and usually only require access to the predictions of the model that they want to explain. This allows them to be easily inserted into most deep learning-based systems. The disadvantage is that the information from the inner-workings of the model can be beneficial and might lead to explanations of higher quality [89].

Explainable versus non-explainable models Some machine learning models are inherently explainable. These models are often simple models where input features can be directly related to the prediction of a model. Linear models and decision trees are some typical examples of explainable models. The downside to using such models is that their simplistic nature often results in worse performance compared to deep learning algorithms, a phenomenon sometimes referred to as Occam's dilemma [83]. A recent direction in XAI is creating neural network architectures that have explainability built into them [90]. One well-known self-explainable architecture was proposed by Chen et al. [91]. Their prototypical part network (ProtoPNet) dissects an image into prototypical parts that are later combined to make the final classification. These prototypes allow the users to inspect what parts of the input influenced the prediction, such that explainability is included into the model. However, this comes at the cost of performance, since the ProtoPNet achieves worse performance compared to non-explainable baseline CNNs.

Local versus global explainability Both Figure 3.1 and 3.2 shows examples of local explanations. These are explanations that explain the prediction of a model for a single input sample. Global explanations investigate the general behaviour of the model and give an impression about the fundamental concepts and motifs that the model have learned. The deep dream framework [92] and partial dependence plots [93] are both global explanation methods that have been used to analyse deep learning algorithms.

3.1 An XAI taxonomy

The field of XAI has developed at a rapid pace, and a vast amount of methods are now available. In a recent review by Samek et al. [21], 46 XAI methods were mentioned and discussed, and Molnar [94] also lists a large number of available approaches. In such a plethora of methods it can be useful with an overarching structure to categorize different methods. In Figure 3.3, we propose a taxonomy for XAI methods that encapsulates most of the major direction within XAI. Each branch of the taxonomy is discussed briefly below.

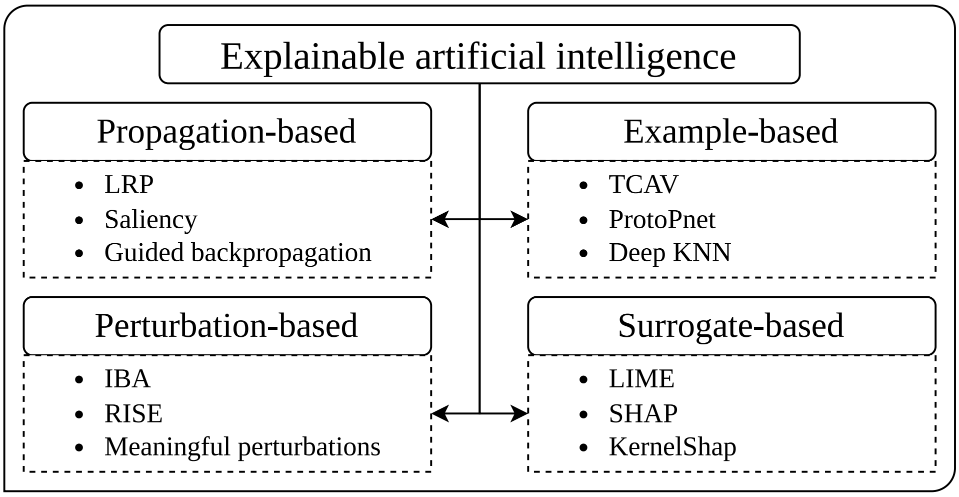


Figure 3.3: A taxonomy of XAI methods.

3.2 Propagation-based explainability

Propagation-based explainability are generally model-aware and local methods that propagate some score or prediction backwards through the neural network to the input. Using gradient information is a common propagation-based strategy that has been used already for several decades [78] and also for explaining other machine learning methods such as kernel methods [95]. The intuition for gradient-based methods is that we want to know how altering the input might influence the output, which is exactly what is computed through the gradient. In its most basic form, a gradient explanation can be computed as $\mathrm{e} = \frac{dy_c}{d\mathrm{x}}$, where y_c is the softmax output for class c and e is a vector indicated the importance of each input feature. A known limitation with gradient-based explanations is that they can be noisy [21], due to the shattered gradients problem [96]. Numerous works have attempted to tackle this problem, for instance by clipping gradients in the backward propagation process [97] or by adding noise in the parameter space and averaging across several explanations [98]. Another important propagation-based methods is LRP [88], which decomposed non-linear classifiers by propagating the output scores back to the input. LRP has been further developed in later works by Montavon et al. [99] and Kindermans et al. [100]. Lastly, class activation mapping (CAM) [101] is a popular propagation-based methods that exploits the global average pooling layer often found in CNNs. The CAM explanation indicates discriminative images regions that can later be used to identify regions associated with particular objects.

The strength of propagation methods is usually their simplicity and low computational demand, as they often require only a slight modification of the backpropagation procedure and a single forward and backward pass through the network. Their weakness is that they need access to the inner-working of a model which reduces their flexibility in certain applications, and can sometimes be noisy as described above.

3.3 Perturbation-based explainability

The common theme in perturbation-based explainability is to relate alteration of the input with changes in the output from the network. An early work by Zeiler and Fergus [102] showed how systematically occluding rectangular areas in images and monitoring the changes in the prediction score could provide coarse indications of input feature importance. Petsiuk et al. [86] introduced a more sophisticated and efficient occlusion scheme leading to RISE. While most perturbation-based methods are model-agnostic, there are some methods that insert noise into the layers of the network as opposed to the input, and thus require information about the model. A recent example is the information bottleneck approach of Schulz et al. [103], where the noise was injected into the layers of CNNs to provide explanations based on information theoretic quantities. There is also a large body of research on learning an optimal set of perturbations to identify the most relevant input features [104, 105, 106].

Perturbation-based methods are often simple and highly flexible, which makes them easy to use in analysis of many different models. On the other hand, they often require extensive sampling of occlusion or noise injection which can be computationally demanding. Also, perturbing inputs in a manner that preserves its characteristics can sometimes be a challenging task, and care must be taken when designing the perturbation scheme in these methods.

3.4 Example-based explainability

Example-based methods explain a model by presenting examples that are similar to a particular instance. The most basic form of example-based explainability is a nearest neighbour approach, where an instance is explained by presenting the most similar examples in the training data. Papernot and McDaniel [107] designed a deep learning framework for nearest-neighbour learning with explainability as a key motivation. A more advanced form of example-based explainability relates instances to concepts or prototypes, such as the aformentioned ProtoPNet [91]. Kim et al. [108] introduced testing with concept activation vectors (TCAVs), where neural networks could be explained through human-friendly concepts. A different example-based approach is counterfactual explanations, where an instance is explained by creating contradicting examples [109, 110]. For instance, if a patient is predicted to obtain an infection after surgery, a counterfactual explanation might generate a similar patient but with a lower measurement for some input feature. This would indicate that this particular feature was important for the prediction and that the generated patient would not be predicted to obtain an infection.

The great benefit of example-based methods is that they provide explanations that are easy to comprehend for humans. For complex input data it might be difficult to understand why a given set of input features are important for a prediction and it might be easier to ascertain if two instances looks similar or not. A downside of example-based methods is that they can be computationally demanding for large datasets. Also, for some counterfactual explanations it can be challenging to train a generator that provides high quality examples. Another challenge is that the stochastic generation of counterfactual explanations might produce contradicting examples, which can cause confusion in the human user.

3.5 Surrogate-based explainability

Surrogate methods exploit interpretable machine learning models to explain black-box models. Perhaps the most well-known and widely used surrogate explanation method is LIME [81]. The core idea is to train an interpretable model that locally approximates the black-box predictor. A dataset is generated by perturbing the input and collecting the prediction of the black-box model, which in turn is used to train the interpretable model. Another popular approach is Shapley additive explanations (SHAP) [111], which is based on Shapely values from game theory. SHAP computes the contribution of each feature to the prediction through a linear model.

Surrogate methods are simple to apply in most use-cases due to their highly flexible model-agnostic nature. Also, since the surrogate model is typically a simple machine learning model, it allows for a stronger theoretical analysis with more guarantees on convergence and optimality of explanations, compared to other approaches. One limitation of surrogate methods is that they require a model to be trained for each instance, which can be computationally demanding for complex instances or if we want to explain numerous instances. Another limitation is that the stochasticity in model training can lead to less robust explanations, as described by Alvarez-Melis and Jaakkola [112].

4

Uncertainty in deep learning

Uncertainty modeling is a fundamental research area in machine learning. In any real-world application, there will always be elements of uncertainty that must be taken into account to provide safe and reliable automated systems. This becomes especially apparent in healthcare applications, where decisions can have fatal consequences. Very few medical practitioners will trust an automated system without a notion of the systems confidence for a given case, which was highlighted in a recent survey of clinicians [16].

Deep learning algorithms do not capture model uncertainty. In classification, the softmax output is sometimes interpreted as model confidence, but this is not advisable [72]. As illustrated in Figure 4.1, softmax probabilities can give high confidence even if a sample lies far outside the data distribution, which has also been shown in quantitative studies [113]. Therefore, deep learning algorithms need to be modified in order to capture uncertainty.

Uncertainty is often categorized into two groups, aleatoric and epistemic uncertainty [114]. Aleatoric uncertainty, the word alea meaning rolling of dice in Latin, is the intrinsic randomness in a process that can not be reduced even if more data is collected. Examples of aleatoric uncertainty can be sensor or measurement noise. Epistemic uncertainty, the word episteme meaning knowledge in Greek, is the uncertainty that stems from lack of knowledge. In classifica-

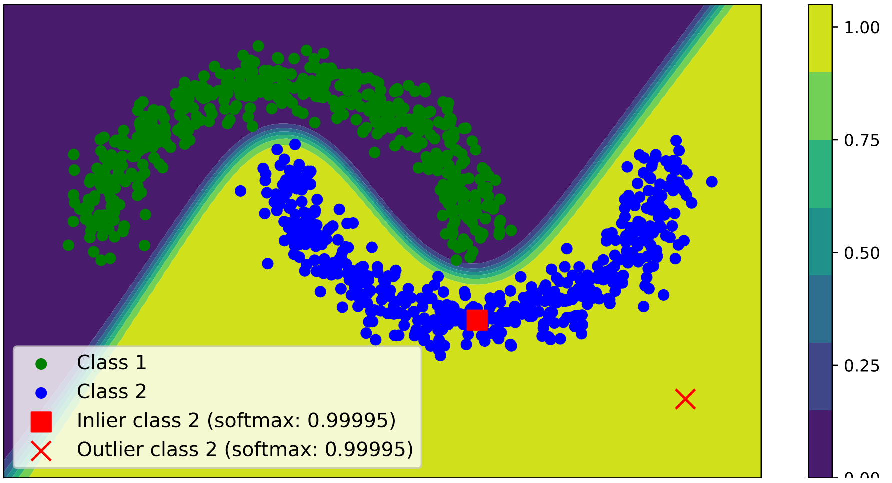


Figure 4.1: Simple MLP for binary classification gives same softmax probability for samples within and outside the data distribution.

tion, epistemic uncertainty can arise if one class is not well-represented in the training data, but could be removed if more examples of said class was collected. Aleatoric uncertainty can be divided further into homoscedastic and heteroscedastic uncertainty [115]. Homoscedastic uncertainty is uncertainty that is constant for different input, while heteroscedastic uncertainty is dependent on the degree of noise being different for some inputs.

Here, we focus on three of the most widely used methods for modeling uncertainty in deep learning algorithms; Bayesian, ensemble, and test-time augmentation methods, all of which are described below. Additionaly, a high-level overview of the different methods is presented in Figure 4.2.

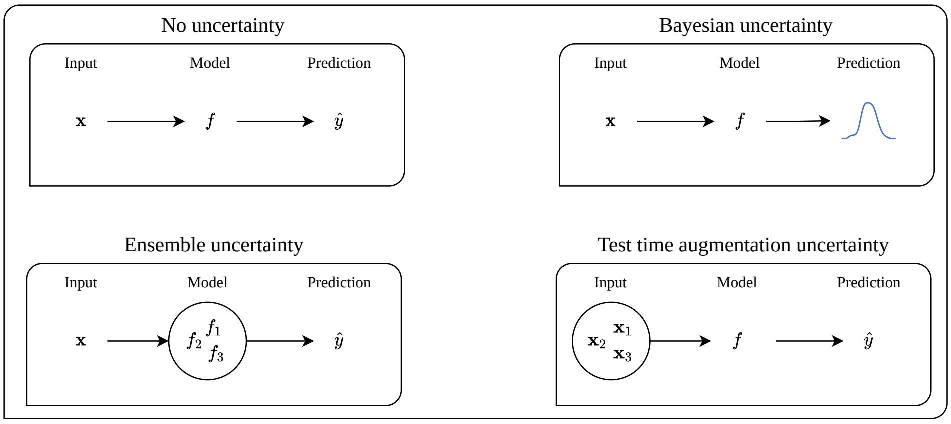


Figure 4.2: High-level overview of methods to model uncertainty in deep learning.

4.1 Bayesian methods

One of the most common approaches to uncertainty modeling in deep learning is through Bayesian methods. Reasoning about uncertainties can be naturally accomplished in a Bayesian setting by treating predictions or weights as distributions instead of point estimates. An early work by Blundell et al. [116] proposed to capture uncertainty by modelling the posterior distribution of the weights of a neural network. Due to the number of weights in most neural networks used for real-world applications it is computationally intractable to calculate the posterior distribution of the weights. Instead, they approximated the posterior through variational inference, which is a family of techniques for calculating intractable integrals. A widely used approach proposed by Gal and Ghahramani [72] is Monte Carlo dropout, which performs Bayesian inference to compute the posterior distribution of the prediction of a neural network. The key idea is to use dropout [71] to sample sets of weights close to an optimal weight configuration and average across predictions based on the sampled sets of weights. However, this requires dropout to be present in the architecture, which is not always the case. An alternative approach is Monte Carlo batch norm [117], which leverages the batch normalization technique [118] to sample sets of weights. This does require batch normalization to be present in the architecture, but batch normalization is a much more common inclusion in recent CNNs compared to dropout and could therefore be more applicable.

An alternative approach to variational inference is using Laplace's method [119] to approximate the intractable posterior. The main idea of Laplace's method is to replace the intractable integral with an integral that can be computed analytically. Ritter et al. [120] proposed a Laplace-based approach to uncertainty modeling which utilized a Kronecker factored Laplace approximation to model the posterior distribution of the weights of a neural network.

Modeling uncertainty through Bayesian methods has the benefit of a strong theoretical foundation. Also, averaging across distributions of models instead of a single final model gives a natural framework for reasoning about uncertainties in a model. The downside of a Bayesian approach is that modeling posterior distributions in deep learning is generally not computationally feasible. Therefore, it is necessary to estimate the posterior, which leads to a compromise between obtaining a good approximation and computational demand. For instance, assuming a particular distribution for the posterior can lead to efficient computation and analytic expressions for desired quantities. However, the assumption might not hold, which can lead to imprecise estimates. On the other hand, approximating the posterior might require extensive sampling to obtain good estimates, which limits its practicality. Determining how to model the posterior is therefore a crucial component in Bayesian uncertainty modeling that requires careful analysis of the data and model that is considered.

4.2 Ensemble methods

The original goal of ensemble methods was not to capture uncertainty, but rather improve performance by combining predictions from several statistical models [121], typically referred to as ensemble members. However, ensembles provide an intuitive approach to capturing uncertainty by aggregating summary statistics across the predictions of the ensemble. Essentially, uncertainty estimates can be thought of as agreement between ensemble members. If all members give a similar prediction, uncertainty will be low. On the other hand, if there is much variation in the predictions of the members, uncertainty will be high. Ensemble methods were among the first to model uncertainty in deep learning [122], and showed good performance compared to competing methods at the time [122]. An important aspect of ensemble methods is to have variety among the ensemble members. Gawlikowski et al. [29] lists 4 approaches for introducing variety in ensembles for deep learning. First, variability can be introduced through the random initialization and optimization of neural networks, since parameter configurations can converge to numerous local optimas. Second, bagging and boosting are two well-known strategies for ensuring variety in ensembles [121]. Bagging refers to uniformly resampling the training data with replacement, and boosting is the process of training models sequentially and optimizing based on the performance of prior models. Third, data augmentation can be used both during training and inference to introduce variation in the dataset. Lastly, variation can be introduced by having different deep learning architectures as each ensemble member.

The strength of ensemble methods for uncertainty modeling lies in their simplicity and their reliability. The only requirement for capturing uncertainty with ensembles is simply to train several models. But despite their simplicity, ensemble methods have been shown to provide state-of-the-art performance when compared with more complex methods [123], even with only some ensemble members. On the other hand, the limitation of ensemble methods is the computational demand. For large and complex datasets that requires big models to solve a task in a satisfactory manner, training a single model may require significant amounts of compute. Introducing numerous such big models will inevitably slow down both training and inference, and could in some cases even not be accomplished due to memory constraints. Therefore, the usability of ensemble methods for uncertainty quantification is highly dependent on the data and task at hand.

4.3 Test-time augmentation

Test time augmentation resembles the ensemble approach to uncertainty quantification, but instead of aggregating across several models, the aggregation is across numerous versions of the same input. Test-time augmentation augmentation has been particularly well-used in the medical setting, with several works demonstrating the benefits of obtaining uncertainties from augmentations [124, 125]. At its core, test-time augmentation amounts to generating multiple versions of an input using a suitable data augmentation scheme. Each version of the input is the passed into the deep learning system, and summary statistics such as the mean and standard deviation can be extracted from the prediction of each version.

The strength of test-time augmentation is again similar to that of the strength in ensembles, namely its simplicity. The procedure does not require modification to the algorithms and can be applied in a black-box manner, i.e. without access to the inner workings of the network. For the limitations of test-time augmentation methods, they suffer from similar computational restrictions that ensemble methods suffer from. However, their restriction is limited to the inference phase, but does not require extra resources during training. This is due to need for creating numerous versions of the input and conducting several forward passes through the network. However, they also have a limitation that is unique to test-time augmentation methods, namely that correct predictions might be altered from correct to incorrect due to the augmentation, as described by Shanmugam et al. [123]. This could have severe effects in classification tasks, since a correct decision might be altered by a component in the decision support system as opposed to characteristics in the data. Therefore, it is important to thoroughly analyze the augmentation used to obtain the uncertainty estimates from the test time augmentation procedure.

5

Self-supervised deep learning

Learning from unlabeled data through self-supervision has gained tremendous attention recently and has achieved impressive results in field such as computer vision [126], NLP [12], and time series analysis [127], in some cases even rivaling the performance of supervised models [30, 73]. Self-supervised learning is part of the representation learning field, and the goal is to learn a function that can transform data into a "useful" representation, where the function is typically modeled by a neural network. This is usually accomplished by exploiting known invariances in the data, such as invariance to data augmentation in vision or temporal invariance in time series. For brevity, we will limit this overview to self-supervised learning in the field of computer vision.

The core idea in self-supervised learning, namely that of learning a representation by maximizing agreement between different views of the same data, was introduced already in the early 1990s by Becker and Hinton [45]. Several works followed in the coming years [46, 128, 129] that contained many of the components seen in modern self-supervised frameworks, but the research direction did not receive much attention. Despite some works in the beginning of the second neural network renaissance such as by Dosovitskiy et al. [130], it was not until the late 2010s that the field really started to receive interest again [131, 132, 126]. Today, self-supervised learning with neural networks constitute a major research direction in learning from unlabeled data. Recent studies have

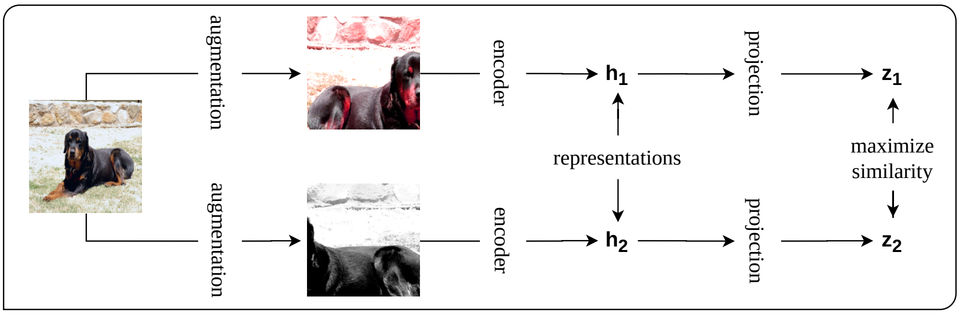


Figure 5.1: Illustration of a self-supervised framework. Image is taken from PASCAL VOC [75].

shown that self-supervised learning frameworks benefit from bigger models [60] and more data [133]. This can explain why self-supervised learning with neural networks did not catch on until recently, since the data and resources required for successful training were not available previously.

Numerous frameworks exist within contemporary self-supervised learning, but some components are shared across most approaches. These components are described below. Figure 5.1 displays an illustration of a self-supervised framework where the components are depicted.

- **Encoder** The core of self-supervised learning is learning a function that maps data into a new representation. In deep self-supervised learning, this function is modelled by a neural network and often referred to as an encoder. For computer vision, the encoder is often a large CNN [126, 30], but recent works have explored the potential of vision transformers [134] as the encoding unit [135]. After the encoder has been trained, it can later be used to transform data into a new representation where a desired task can be performed, or it can be used in transfer learning by initializing the encoder for a new downstream task where limited amounts of labeled data is available.

- **Projection head** The vast majority of current self-supervised framework employ a projection head that maps the output of the encoder into a new space where the loss is applied [30, 73, 31]. This projection head is usually a small MLP consisting of only a couple of hidden layers. While it has been shown in several works that the projection head is crucial for high performance, it is only recently that the purpose of the projection head has become clear. The mapping from the output of the encoder to the projection head where the loss function is applied avoids dimensional collapse in the representations of the encoder [136, 137], which facilitates a more informative representation for downstream tasks.

- **Loss function** The loss function is typically applied on the output of the projection head, and provides the signal for the gradient-based optimization of the learning framework. The two most common forms of loss functions in current deep self-supervised learning are classification losses like a cross-entropy loss with pseduo-labels [126, 31], or contrastive losses where similarity between positive pairs are maximized and similarity between negative pairs are minimized [30, 73].

- **Data augmentation** The purpose of data augmentation in self-supervised learning is for creating positive pairs of samples by exploiting known invariances in the data. Through one or several stochastic augmentations, two different views are created from a single sample and which constitute a positive pair. The two views must be sufficiently different such that the encoder must learn a high quality representation in order to identify the positive pairs, but also not too different such that the characteristics in the original image are distorted. It is therefore of crucial importance to carefully design the set of augmentations to obtain representations that encodes useful information from the input data.

Numerous self-supervised frameworks have been proposed over the last couple of years, but three variants are most commonly seen in the literature. These are contrastive, clustering, and siamese approaches, which are described below. Additionally, task-specific self-supervised framework have also been investigated in the literature. For instance, a useful representation can be learned by predicting the rotation of an image [138] or solving jigsaw puzzles [139], but these approaches will not be discussed in this thesis.

5.1 Contrastive self-supervised learning

The goal in contrastive self-supervised learning is identifying positive pairs among negative samples. This is achieved by attracting positive pairs and repulsing negative pairs in the representation space. The positive pairs are created through the data augmentation procedure. The negative pairs can be constructed in different ways and prevents all samples from being mapped to one point in the representation space (which would be the trivial solution), a phenomenon known as representation collapse. Examples of some widely used contrastive learning frameworks are SimCLR [30] and MoCo [73], which mainly differ in how they create the negative samples.

5.2 Clustering-based self-supervised learning

In clustering-based self-supervised learning, a clustering algorithm is used to create pseudo-labels that the encoder learns to predict. Most frameworks alternate between clustering and prediction, which enforces and enhances the structure of the data in the new representation. The trivial solution in clustering approaches would be to assign all samples to one cluster, which is typically avoided by treating other cluster centroids as negative prototypes that prevents collapse to one cluster. The DeepCluster framework is one of the most well-known clustering-based self-supervised frameworks [126], where pseudo-labels are created through k-means-clustering. The more recent SwAV framework combines both contrastive and clustering-based learning [31], which they demonstrate can be beneficial to performance.

5.3 Siamese self-supervised learning

Siamese approaches consist of two encoders that align different views of the same sample, without the need for negative samples or pseudo-labels. For siamese approaches, the trivial solution is to map all samples to the same point in the representation space which would maximize alignment. This solution is handled in different ways. Two of the most recent and successful siamese approaches are the boostrap your own latent (BYOL) [140] and SimSiam [141] frameworks. The BYOL framework employs a teacher-student setup while the SimSiam framework uses a stop-gradient operation that avoid the need for training two encoders. Both approaches have been shown to avoid the problem of representation collapse [137].

6

Data-driven healthcare

The right to health is a fundamental human right [1], but numerous challenges face those who wish to comply. A recent paper by Figueroa et al. [2] list a profusion of obstacles in contemporary healthcare such as shortage of trained health personnel, increases in costs and workload, aging population, and challenging diagnosis, to name a few. Tackling these and other problems is crucial to provide high quality and reliable healthcare to people around the world.

Many researchers and healthcare professionals believe that data-driven healthcare has the potential to solve many of of these problems [142, 17]. Data-driven methods are based on algorithms that learn to perform tasks by identifying patterns in data, and often improve in line with the amount of data. Recent advances in data-driven approaches based on deep learning have shown remarkable performance in healthcare applications [5, 6], with some suggesting that current AI methods will bring about the 4th healthcare revolution [1].

In the healthcare domain, ample amounts of data are collected each day, which is a major reason for the optimistic view on data-driven healthcare. A diverse set of measurements are collected during the treatment or assessment of patients. Data can come in many forms, such as time series of blood samples, CT images of organs, or composition of data types in electronic health records (EHRs). In some cases, the data is accompanied by annotations from domain experts. This could for instance be the diagnosis associated with a particular time series, or

[1] `https://www.ohchr.org/en/health`

a segmentation mask delineating a tumor in a PET image. But a much more common scenario is that the data is not partnered with such information and it must be processed without information from physicians.

There are numerous ways that data-driven healthcare can aid in solving pressing issues in healthcare. For instance, screenings programs examine many patients without symptoms, with the goal of detecting cancer before it spreads. Early detection is crucial to save lives, and also allows for a less exhaustive and costly treatment procedure. However, processing the large amount of patients is both time and resource demanding. A system based on data-driven algorithms could automatically process the data and indicate which samples require extra attention from physicians, thus enabling better and more time efficient treatment. Such screening systems have been investigated and shown encouraging results in e.g. screening for breast [143, 144] and colorectal cancer [145].

Another promising research area is data-driven decision support systems that aid physicians with diagnosis of challenging diseases. This could be rare conditions that require specialist knowledge to detect reliably, or it could be a disease that is problematic to identify due to its complexity. Precise predictive systems could guide practitioners in such demanding scenarios. There are many examples of such systems with promising developments. One example is predicting complications that might arise during surgery [146]. Another is identifying cancer patients trajectories based on free text data [147].

Data-driven algorithms can also be used in the development of new medicaments that could be used to provide better, safer, and more efficient treatment. A recent survey by Kim et al. [148] showed that data-driven approaches have great potential in aiding with selecting and designing potential drugs. Data-driven algorithms can also play an important role in the development of vaccines, as demonstrated in recent paper by Raeven et al. [149].

6.1 Deep learning in data-driven healthcare

Another important reason for the positive outlook for data-driven healthcare is the recent advances in deep learning [9, 1]. Deep learning has the benefit of being able to process raw data without the need for complex pre-processing. Also, the high precision that deep learning exhibit is often attributed to this ability [74]. Furthermore, deep learning algorithms usually increase their performance when presented with more data [4]. Since data is gathered continuously in the healthcare sector, this means that systems can keep improving as time goes on. An overview of data-driven healthcare is shown in Figure 6.1.

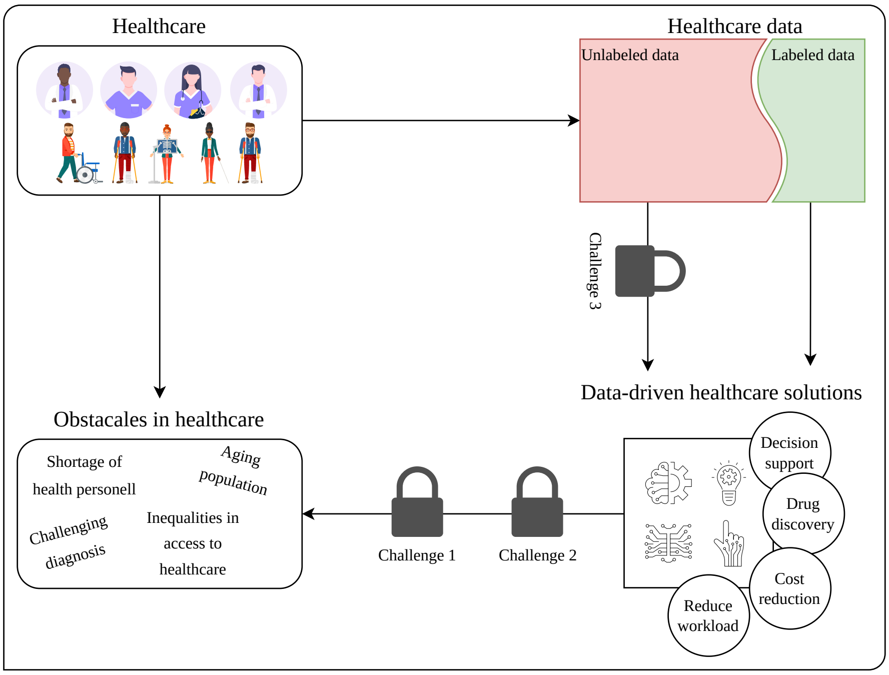


Figure 6.1: Overview of data-driven healthcare. Images are taken from `www.mostphotos.com`.

The field where deep learning has had the biggest impact is in medical computer vision. Across several imaging types such CT, PET, and X-ray images and in numerous medical image related tasks, deep learning has brought great improvements [150]. Campanella et al. [6] developed a deep learning algorithm that could make predictions with a similar level of precision as domain experts in pathology on whole slide images. Whole-slide images can be very large and processing them is time consuming and challenging. A precise data-driven system could automatically identify patients that require further inspection by pathologist, thus reducing their workload and increasing the time spent analysing patients that require more detailed care. Dong et al. [151] also showed how deep learning can play a vital role processing whole-slide images, e.g. to conduct fast and accurate breast cancer segmentation. Zhou et al. [152] introduced a deep learning systems that could identify abnormalities in CT scans of COVID-19 patients with high precision. Such a system could play a vital role in fast and accurate treatment of patients infected by the coronavirus. Kuttner et al. [8] proposed a deep learning solution automatically models the arterial input function solely based on image data. Prior solutions relied on blood sampling to correct the prediction of the arterial input functions. But sampling blood is a challenging and painful procedure. Therefore, great benefit in terms of cost and treatment efficiency can be gained from such a solution.

But deep learning has also lead to great improvements in other medical fields. In medical NLP, Li et al. [153] introduced a transformer-based deep learning system for processing text data in EHRs. When trained and evaluated on a corpus of millions of patients, the system showed great improvements over prior solutions when it comes to predicting the likelihood of a patient getting a new medical condition in the future. In medical graph data, Choi et al. [154] exploited the graph-like structure of diagnosis codes to train a deep learning system that could conduct disease phenotyping with high accuracy. And in medical time series, Harutyunyan et al. [155] proposed a multitask deep learning system that demonstrated high performance across four clinical time series prediction tasks.

Despite all of these promising developments, there are still some major obstacles that needs to be overcome in order to achieve the full potential of deep learning in data-driven healthcare. The lack of explainability is regularly listed as one of the top challenges that needs to be tackled [18, 16]. Without explainability, physicians will be reluctant to trust the data-driven system. For instance, it has been shown in several studies that deep learning algorithms can exploit artifacts and confounding factors instead of generalizable patters [19, 20]. This can have detrimental affects, as systems might fail unexpectedly or report erroneous evidence for a disease. Several recent works investigate XAI in the context of healthcare applications, as illustrated by a recent review by van der Velden et al. [156]. For instance, Weina Jin [157] investigated how well explanation fulfill clinical requirements in multi-modal medical imaging. Another work by Thomas et al. [158] introduced an interpretable deep learning systems for classification of non-melanoma skin cancer. Despite all of these advances, there are some gaps in the XAI literature. First, very little work have been done on capturing uncertainty in explanations, apart from some preliminary works [24, 25]. Without uncertainty, explanations might provide an unwarranted trust in an automated system. Second, apart from some notable exceptions [27], current XAI methods mostly operate in the supervised setting on a score or a prediction. But none of the current methods are capable of explaining representation in the unsupervised setting, which can occur regularly in e.g. self-supervised learning for medical data.

Another component that is missing in deep learning is uncertainty quantification. Such a component is regularly highlighted as highly desired component in any automatic support system for healthcare tasks [16, 28]. Being able to reason about uncertainties is of utmost importance when operating in a setting where decision can have fatal consequences. For a physician receiving support from an automatic system it is important to know if the system is providing a suggestion that is highly certain or not, since intrusive and exhaustive treatments might be recommend based on the suggestion. Several works have looked into uncertainty quantification in deep learning-based systems for healthcare applications. Carneiro et al. [159] showed how Bayesian methods could be used to capture uncertainty in classification of polyps from colonoscopy images. Herzog et al. [160] proposed to use Monte Carlo dropout to asses the uncertainty in a deep learning system for stroke analysis based on magnetic resonance images (MRIs). Leibig et al. [161] demonstrated how leveraging uncertainty estiamtes could be used to improve diagnostic performance. Nevertheless, very little attention have been given on how to capture uncertainty in explanations.

An additional challenge for deep learning-based data-driven healthcare is learning without supervision. As mentioned above, most of the data that is gathered in the healthcare sector is unlabeled. This constitutes a problem for deep learning algorithms, since they struggle to obtain optimal performance without label information [17]. Overcoming this problem is therefore of great significance, as it would allow deep learning-based systems to exploit the full amount of data available in the healthcare sector. This could potentially increase performance and lead to more robust and reliable deep learning-based support systems. Self-supervised learning have received an increasing amount of attention in the healthcare domain. Hansen et al. [162] proposed a new method for few-shot medical image segmentation that was based on self-supervised learning. Bozorgtabar et al. [163] used self-supervised learning to improve the performance in anomaly detection for X-ray images. Yang et al. [164] showed how self-supervised learning could be used to effectively exploit unlabeled histopathological images by learning. Dong and Voiculescu [165] showed how the combination of contrastive and federated learning could be used to effectively makes use of decentralized unlabeled medical data. But no works have looked into how to incorporate domain-expertise into self-supervised learning for CBIR, nor have they been able to explain the representations produced by a self-supervised framework.

Part II

Summary of research and concluding remarks

/7

Paper I

Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps

Kristoffer K. Wickstrøm, Michael C. Kampffmeyer, Robert Jenssen

Medical Image Analysis

To the best of our knowledge, this work is the first to estimate uncertainties in explanations in the XAI field. We present a Bayesian approach that relies on a variational approximation of the posterior distribution of explanations through the dropout technique [71]. We leverage the guided backpropagation technique [97] to obtain explanations. The usability of the proposed methodology is demonstrated for semantic segmentation of colorectal polyps.

Experiments were conducted on two real-world datasets, which indicate that deep models are utilizing the shape and edge information of polyps to make their prediction. Moreover, inaccurate predictions show a higher degree of uncertainty compared to precise predictions. Lastly, we investigate how uncertainties in explanations of polyp prediction behave, and show that an explanation

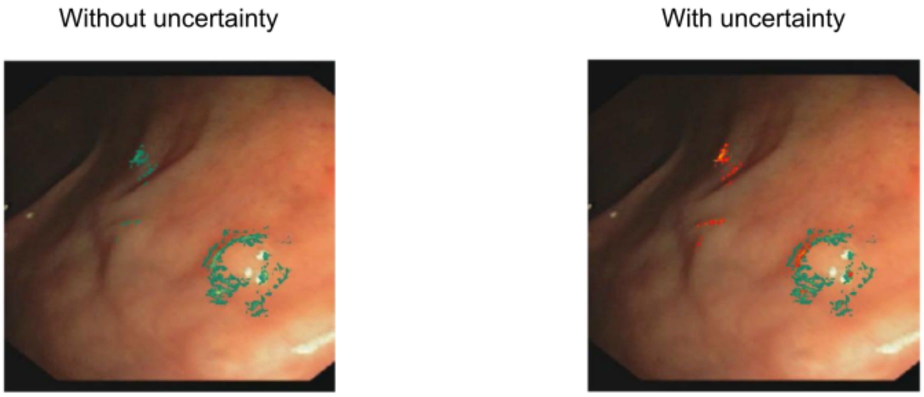


Figure 7.1: Illustration of explanation without (leftmost figure) and with (rightmost figure) uncertainty estimates included. Example taken from Paper I.

can have varying degrees of uncertainty associated with different parts of an explanation. Figure 7.1 illustrates one such example taken from Paper I, where the uncertainty analysis reveals that parts of the explanation associated with non-polyp pixels have a higher degree of uncertainty.

This work was presented as a part of an invited talk at the Big Insight center for research-based innovation in Oslo. It was also featured during a spotlight presentation at the Visual Intelligence center for research-based innovations. Video recordings of both presentations are listed below, together with code used in the paper.

- Invited talk at the Big Insight center for research-based innovation: `https://www.youtube.com/watch?v=STInTtflcyU`.

- Spotlight presentation at the Visual Intelligence center for research-based innovation: `https://www.youtube.com/watch?v=CluEu7lp3RM`.

- Code: `https://github.com/Wickstrom/uc-in-xai.git`.

Contributions by the author

1. The idea was conceived by me and further developed with all co-authors.

2. The implementation and experiments were conducted by me.

3. I wrote the main draft of the paper.

/8

Paper II

Uncertainty-aware deep ensembles for reliable and explainable predictions of clinical time series

Kristoffer K. Wickstrøm, Karl Øyvind Mikalsen, Michael C. Kampffmeyer, Arthur Revhaug, Robert Jenssen

IEEE Journal of Biomedical and Health Informatics

This paper presents a deep ensemble approach to capture uncertainty in explanations of predictions for time series. Our core idea is to train an ensemble of neural networks that produce a set of predictions and explanations, and capture uncertainty by taking the standard deviation across the explanations. The uncertainty estimates can be considered as a measure of disagreement between the ensemble members on what features are important to performed the desired task. Moreover, we propose to use the uncertainty estimates to filter out uncertain parts of an explanations, a method we refer to as uncertainty-filtered explanations.

Using ensembles to capture uncertainty is motivated by two common characteristics in clinical time series. First, clinical time series can be successfully

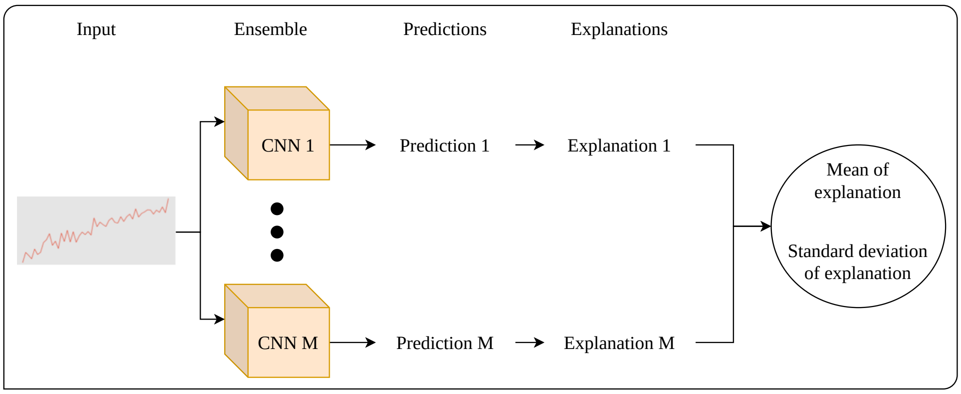


Figure 8.1: Illustration of deep ensemble approach proposed in Paper II.

processed by deep learning architectures with low computational demand, which allows for more members in the ensemble. Second, many clinical time series datasets contain a limited amount of examples. This eases the computational demand and reduces the time needed to train the deep ensemble. We chose to process the clinical time series using the fully convolutional network (FCN) proposed by Wang et al. [166], due to its encouraging performance on time series classification benchmarks [166]. Each member of the ensemble was explained using the CAM explanation method, as the structure of the FCN architecture is particularly suitable for CAM method. An illustration of the proposed deep ensemble is shown in Figure 8.1.

Experiments were conducted on both synthetic and real-world datasets. Results demonstrate that the proposed ensemble is more accurate in locating relevant time steps and is more consistent across random initializations. Furthermore, we also show how the proposed uncertainty-filtering can provide clearer and more understandable explanations. The explanations from the deep ensemble were also analysed with the aid of a domain expert. This analysis revealed that the deep learning system had learned to identify patterns that closely agreed with clinical knowledge, which illustrates how XAI can be used to establish trust.

An extended abstract (paper 12) of preliminary work that lead to this paper was presented at the 3rd Northern Lights Deep Learning Conference, Tromsø, Norway in 2021. This work was also presented as a part of an invited talk at the Big Insight center for research-based innovation in Oslo. Video recordings of both presentations are listed below, together with code used in the paper.

- Invited talk at the Big Insight center for research-based innovation: `https://www.youtube.com/watch?v=STInTtflcyU`.

- Oral presentation from the Norhern Lights Deep Learning Conference 2021: https://www.youtube.com/watch?v=Odv3YD1OFxE.

- Code: https://github.com/Wickstrom/TimeSeriesXAI

Contributions by the author

1. The idea was conceived by me and further developed with all co-authors.

2. The implementation and experiments were conducted by me.

3. The discussion and analysis was conducted in collaboration with domain-expert Arthur Revhaug.

4. I wrote the main draft of the paper.

9

Paper III

RELAX: Representation Learning Explainability

Kristoffer K. Wickstrøm, Daniel J. Trosten, Sigurd Løkse, Ahcène Boubekki, Karl Øyvind Mikalsen, Michael C. Kampffmeyer, Robert Jenssen

Submitted to International Journal of Computer Vision

The vast majority of XAI research have been focused on explaining scores and predictions, but no methods are designed for explaining representations of data. In this paper we present the first representation learning framework, entitled RELAX. RELAX is a perturbation-based method that works by measuring similarities between representation of an input and a masked version of itself. An illustration of the RELAX framework is presented in Figure 9.1. To provide a better understanding of RELAX, we provide a theoretical analysis that links the explanations of representations to linear scoring functions between data points and a class-conditional mean. Also, we derived a bound on the number of masks required to obtain precise and reliable explanations. Since no methods have been developed to explain representations, we develop gradient-based methods as baseline methods to compare RELAX with.

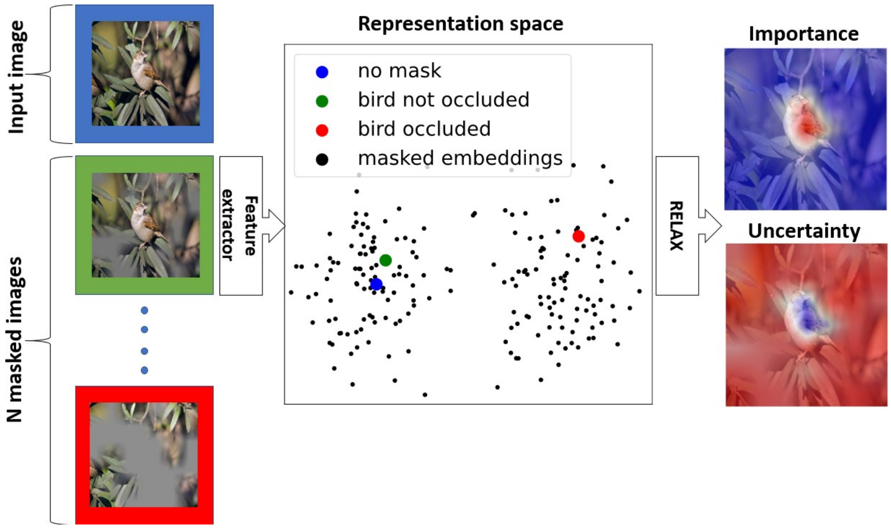


Figure 9.1: Conceptual illustration of RELAX taken from Paper III.

The results on public datasets demonstrate the benefit of RELAX compared to baseline methods across several metrics. Moreover, we show how RELAX can be used in the context of multi-view clustering and to explain traditional feature extraction methods commonly used prior to deep learning. Lastly, we also conduct a user study with human examiners, with results indicating that the RELAX explanations agreed the most with explanations from humans.

An abstract (paper 14) of preliminary work was presented at the NOBIM conference, Oslo, Norway in 2021. Code used in the paper is listed below.

Code: `https://github.com/Wickstrom/RELAX`

Contributions by the author

1. The idea was conceived by me and further developed with all co-authors.
2. The implementation and experiments were conducted by me.
3. I wrote the main draft of the paper.

10

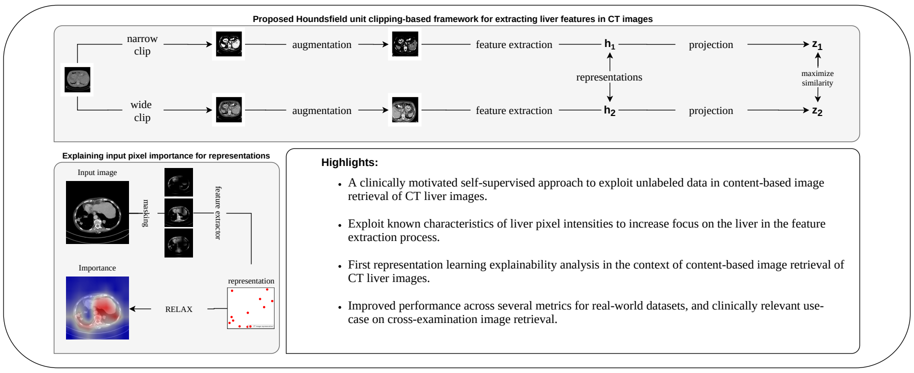
Paper IV

A clinically motivated self-supervised approach for content-based image retrieval of CT liver images

Kristoffer K. Wickstrøm, Eirik A. Østmo, Keyur Radya, Karl Øyvind Mikalsen, Michael C. Kampffmeyer, Robert Jenssen

Submitted to Computerized Medical Imaging and Graphics

In this paper, we propose a self-supervised framework for CBIR of CT liver images. Current self-supervised frameworks often rely on data augmentations that are designed for natural images and does not take into account the characteristics of medical images. We propose a clinically motivated self-supervised framework that exploits known invariances in CT liver images to train feature extractors that focus on clinically relevant features. The main idea is to have a narrow and wide clipping of the pixel intensities for different views of the same image. This clipping should preserve the liver features but to a varying degree remove non-liver pixels. This will encourage the feature extractor to learn that pixels related to the liver are important and should receive attention. Furthermore, we leverage the RELAX framework from Paper III to conduct a



Highlights:

- A clinically motivated self-supervised approach to exploit unlabeled data in content-based image retrieval of CT liver images.
- Exploit known characteristics of liver pixel intensities to increase focus on the liver in the feature extraction process.
- First representation learning explainability analysis in the context of content-based image retrieval of CT liver images.
- Improved performance across several metrics for real-world datasets, and clinically relevant use-case on cross-examination image retrieval.

Figure 10.1: Graphical abstract with figures taken from Paper IV.

novel representation learning explainability analysis in the context of CBIR. A graphical abstract of the paper can be viewed in Figure 10.1.

Experiments show that the proposed framework improve the retrieval performance and leads to feature extractors that focus more on liver-related features. Moreover, we present the first representation learning explainability analysis in the context of CBIR of CT liver images. Our analysis reveals that feature extractors with seemingly similar performance can focus very different types of features, and that feature extractors trained on non-medical datasets (Imagenet-pretrained feature extractors) focus on edge-information and not organ-information. Lastly, we present a case-study where results of a cross-examination CBIR analysis is compared with the analysis of a domain expert. Our results show that the CBIR system can achieve high agreement with domain experts.

Code used in the paper is listed below.

Code: `https://github.com/Wickstrom/clinical-self-supervised-CBIR-ct-liver.git`

Contributions by the author

1. The idea was conceived by me and further developed with all co-authors.
2. The implementation and experiments were conducted by me.
3. The discussion and analysis was conducted in collaboration with domain-expert Keyur Radia.
4. I wrote the main draft of the paper.

11

Concluding remarks

In this thesis, we advanced deep learning with emphasis on data-driven healthcare. We focused on addressing key challenges that limit the usability of deep learning in data-driven healthcare; (1) the lack of explainability, (2) how to model uncertainty, and (3) learning from limited labels.

A Baeysian approach was proposed to capture uncertainties in explanations, thus working in the intersection of Challenge 1 and 2. The usability of the new methodology was illustrated in the context of semantic segmentation of colorectal polyps. Modeling uncertainty in explanation was also investigated in the context of clinical time series, a common data modality encountered in healthcare applications. Motivated by particular characteristics in clinical time series, we proposed a deep ensemble approach that could capture uncertainties in explanations. The uncertainty estimates were used to to create uncertainty-filtered explanations, which were shown to have higher quality and less ambiguity.

In the intersection between Challenge 1 and 3, the first framework for explaining representations, as opposed to predictions. Using the new framework, we showed how it allowed for new insights into self-supervised learning, multi-view clustering, and traditional feature extraction techniques. Furthermore, we proposed a new self-supervised framework that exploits domain-knowledge in the data augmentation procedure to train deep learning architectures without label information. The framework was used in CBIR of CT liver images, and results demonstrated how it was beneficial to performance and in extracting clinically relevant features. Lastly, we showed how explainability could give insights into limitation and strengths of different feature extraction models in the CBIR setting.

The thesis has focused on healthcare applications, and we have demonstrated how our contributions can be successfully applied across several tasks and data modalities. But the proposed methodology in the included research paper is not limited to healthcare applications, and could see use in other domains. We believe that the advances introduced in this thesis can play an important role in designing more reliable and trustworthy deep learning systems that can effectively exploit data with little label information.

11.1 Limitations and future work

Any research paper will have both strengths and limitations. In this section, we discuss the limitations of the papers included in this thesis. Furthermore, we examine promising direction for future research related to the methodology presented in this thesis.

Paper I The methodology proposed in Paper I relies on dropout being part of the neural network architecture. This is not always the case, and simply adding dropout to a network that has not been trained with such a regularization can significantly alter the performance of the network. The purpose of dropout in our methodology is to sample a set of weights from the trained network. This could be achieved through other means. For instance, the Monte Carlo batch normalization [117] procedure leverages batch normalization [118] to sample predictions that could be used to capture uncertainty. Since batch normalization is a more common inclusion in deep learning architecture this could increase the flexibility of the methodology. Another approach could be to inject noise into the parameters of the network as done by Bykov et al. [98], in order to sample weights from the trained network. Such an approach would be applicable for any neural network architecture, but also requires the selection of the noise level. Lastly, quantitative evaluation of explanations was less evolved during the development of Paper I, and therefore this work relied mainly on qualitative analysis. A thorough quantitative analysis such as in Paper II and III would provide further insights into the strengths and weaknesses of the proposed methodology.

Paper II Using ensembles to capture uncertainty, as proposed in Paper II, can be challenging in cases where it is necessary to use models with high computational demand, for instance when processing 3D medical images. In such cases, the computational demand could be decreased in several ways. One promising direction is quantization [167], where 32-bit floating-point activations and weights are mapped to 8-bit integers to improve computational efficiency. A more traditional approach is network pruning, where redundant neurons are removed from the network. A recent paper by Shomron et al. [167] showed how explanations could be used to prune CNNs.

Paper III The RELAX framework introduced in Paper III provides an explanation for the representation of an image. This explanation is computed by numerous forward passes through the network. This can be computationally demanding, particularly if a quantitative analysis across many images is to be carried out. An interesting approach to build the representation explainability into the network, thus only requiring a single-forward pass, is self-masking. Such approaches have been developed in the context of classification [168, 169], but not in a completely unsupervised setting. Another potential avenue for future research is extending RELAX to new data modalities. This is not a trivial task, since the masking must be tailored to the specific data type. For instance, masking graph data could be carried out in numerous ways, and can be performed on both the edges and nodes of a graph.

Paper IV Paper IV illustrated how feature extractors trained without labels could learn to retrieve similar examples as domain experts. However, it also showed that the feature extractors lacked spatial awareness due to being trained only on single slice images. This could be improved by incorporating neighbouring slices into the self-supervised training procedure, or by training features extracts that process the full 3D CT volume. Another interesting question to investigate is how the proposed wide and narrow clipping strategy could be used to train feature extractors that focus on particular organs. Our focus was on the liver, and the clipping was tailored to this purpose. But other clipping strategies are possible, which could be used to improve the performance of CBIR for other organs.

Future directions The number of XAI methods have gone from only a handful to several dozens during the last decade. The large selection of available methods raises an interesting question; how do we determine what XAI method to use? Or stated differently, how do we determine what makes an explanation good? In general, there exists no "correct" explanation, and therefore it is not possible no directly asses which explanation method is superior. As more XAI methods emerge and deep learning becomes integrated into new domains with requirements on explainability, the question of what make a good explanation needs to be answered.

At the moment, there are two main directions that stand out in answering this question. First, quantitative analysis is becoming an increasingly important part of XAI. Recently, Hedström et al. [170] introduced the Quantus toolbox, which have collected and unified numerous quantitative measures that indirectly evaluate how good an explanation is. Second, self-explainable model can eliminate the question all together, since model can explain itself without the need for external XAI methods. A notable work in this direction is the ProtoPNet described in Chapter 3, but more recent works have also shown that self-explainable deep learning can be a promising direction [171].

Several uncertainty methods have been investigated in the research papers, but they have all been considered in isolation. In Paper I and II, the uncertainty stemming from the model was considered, while Paper III investigated uncertainty in the data. An interesting avenue of research is modeling uncertainty in both data and model simultaneously. This would require carefully coupling of methods to ensure that the multiplicative effect of e.g. adding noise in both the data and the model is kept under control.

The works presented in this thesis have mainly been focused on XAI as a tool that can be used after a model is trained. But XAI can also be used to improve the performance of a deep learning model. For instance Sun et al. [172] showed how XAI could be used to fine-tune a model for image captioning. Also, Silva et al. [173] demonstrated how explanations could be used to guide a model for CBIR, leading to improved performance. Such approaches providing promising avenues for future research. For instance, quantitative XAI measures like those collected in Quantus [170] could be used as part of the training objective, which could lead to improved explanations and performance.

Part III

Included papers

/12

Paper I

Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps

Kristoffer Wickstrøm, Michael Kampffmeyer, Robert Jenssen

Medical Image Analysis

Medical Image Analysis 60 (2020) 101619

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps

Kristoffer Wickstrøm[1,*], Michael Kampffmeyer[1], Robert Jenssen[1]

Department of Physics and Technology, UiT The Arctic University of Norway, Tromsø NO-9037, Norway

ARTICLE INFO

Article history:
Received 10 May 2019
Revised 14 November 2019
Accepted 14 November 2019
Available online 20 November 2019

Keywords:
Polyp segmentation
Decision support systems
Fully convolutional networks
Monte carlo dropout
Guided backpropagation
Monte carlo guided backpropagation

ABSTRACT

Colorectal polyps are known to be potential precursors to colorectal cancer, which is one of the leading causes of cancer-related deaths on a global scale. Early detection and prevention of colorectal cancer is primarily enabled through manual screenings, where the intestines of a patient is visually examined. Such a procedure can be challenging and exhausting for the person performing the screening. This has resulted in numerous studies on designing automatic systems aimed at supporting physicians during the examination. Recently, such automatic systems have seen a significant improvement as a result of an increasing amount of publicly available colorectal imagery and advances in deep learning research for object image recognition. Specifically, decision support systems based on Convolutional Neural Networks (CNNs) have demonstrated state-of-the-art performance on both detection and segmentation of colorectal polyps. However, CNN-based models need to not only be precise in order to be helpful in a medical context. In addition, interpretability and uncertainty in predictions must be well understood. In this paper, we develop and evaluate recent advances in uncertainty estimation and model interpretability in the context of semantic segmentation of polyps from colonoscopy images. Furthermore, we propose a novel method for estimating the uncertainty associated with important features in the input and demonstrate how interpretability and uncertainty can be modeled in DSSs for semantic segmentation of colorectal polyps. Results indicate that deep models are utilizing the shape and edge information of polyps to make their prediction. Moreover, inaccurate predictions show a higher degree of uncertainty compared to precise predictions.

1. Introduction

Colorectal Cancer (CRC) is one of the leading causes of cancer-related deaths worldwide (Siegel et al., 2017; Chen et al., 2016; Larsen, 2016), with an estimated five-year survival rate for an advanced stage CRC diagnosis of 14%. The estimated survival rate for early diagnosis is 90% (Larsen, 2016). Currently, the gold standard for CRC prevention is through regular colonoscopy screenings. One of the main tasks during a screening is to locate small abnormal growths called polyps, which are known to be possible precursors to CRC. Hence, increasing the detection rate of polyps is an important component for reducing mortality rates. However, such screenings are manual procedures performed by physicians and are therefore affected by human factors such as fatigue and experience. One study has estimated the polyp miss rate during a screening to be between 8–37%, depending on the size and type of the polyps (Van Rijn et al., 2006). A possible method for increasing polyp detection rate is to design Decision Support Systems (DSSs), which could aid physicians during or after the procedure. A dependable and robust DSS would have the advantage of not being influenced by human factors and could also provide a second opinion for inexperienced practitioners.

One popular approach for developing DSSs has been through machine learning, with promising results on a range of different tasks like brain tumor segmentation (Havaei et al., 2017), retinal vessel segmentation (Guo et al., 2019), melanoma lesion segmentation (Nida et al., 2019), and colorectal polyp detection (Bernal et al., 2015; 2014; Liu, 2017; Ribeiro et al., 2016). In the context of CRC prevention, there have been a number of studies on the detection of polyps with encouraging results (Tajbakhsh et al., 2016; Hwang et al., 2007; Alexandre et al., 2007; Wimmer et al., 2016; Häfner et al., 2015), but polyp segmentation has proven to be a challenging task and the necessary precision has been difficult to obtain (Bernal et al., 2015; 2014; Condessa and Bioucas-Dias, 2012).

[*] Corresponding author.
E-mail address: kristoffer.k.wickstrom@uit.no (K. Wickstrøm).
[1] UiT Machine Learning Group (http://machine-learning.uit.no).

https://doi.org/10.1016/j.media.2019.101619

However, as a consequence of increasing amounts of publicly available colon imagery combined with advances in deep learning research for image analysis, recent studies based on deep learning for colorectal polyp segmentation have shown promising results and a significant increase in precision (Vázquez et al., 2016; Brandao et al., 2017; Urban et al., 2018).

High precision is a crucial component of any reliable DSS, but other constituents are also vital in order to engineer dependable DSSs. Physicians are tasked with making decisions that can have fatal consequences and they go to great lengths in order to ensure that the decision they make is likely to have a favorable outcome. Therefore, a trustworthy DSS should provide a measure of uncertainty to accompany its prediction such that physicians can make well-informed decisions. Another integral part of a dependable DSS is to communicate to the user what factors influences a prediction. Without such information, the user can not determine if the model is detecting features that are actually associated with the disease in question or if it is exploiting artifacts in the data. For instance, a study by Zech et al. (2018) uncovered that a deep learning model tasked with diagnosing disease from x-ray images had learned to exploit information in metal tokens included in the x-ray images for inference instead of detecting disease-specifics features. When the model is then presented with an image without these artifacts the precision drops considerably.

Despite the obvious benefit of increased performance, systems based on deep learning have no inherent way of representing the uncertainty associated with a model's prediction nor do they provide any indication as to what features in the input influences a particular prediction. This lack of theoretical understanding for the underlying mechanics of deep models have resulted in deep learning based models often being referred to as ”black boxes” (Alain and Bengio, 2017; Shwartz-Ziv and Tishby, 2017; Yu and Príncipe, 2018). Multiple recent studies have proposed methods that, to some extent, address the lack of transparency (Gal and Ghahramani, 2016; Kendall and Gal, 2017; Springenberg et al., 2015; Zeiler and Fergus, 2014; Bach et al., 2015; Simonyan et al., 2013), and they have seen some use in analysis of medical images (Dubost et al., 2019; Zech et al., 2018) However, these methods have yet to be utilized in DSSs for colorectal polyp segmentation based on deep learning.

Our contributions are the following:[2]

- We incorporate and develop recent advances in the field of deep learning for semantic segmentation of colorectal polyps in order to create deep models that provide uncertainty measures along with their prediction. Results indicate that erroneous predictions show a significantly higher degree of uncertainty compared to correct predictions. Furthermore, we model input feature importance to create interpretable deep models. Results show that our models are considering shape and edge information in order to segment polyps.
- We propose a novel method for estimating uncertainty in the importance of input features, which we refer to as Monte Carlo Guided Backpropagation, and demonstrate how this method can be used in the context of colorectal polyp segmentation.

To the authors' knowledge, none of the above points have been previously explored in the context of semantic segmentation of colorectal polyps.

2. Models and methods

In this section we introduce Fully Convolutional Networks (FCNs) and describe the three architectures utilized in this study. Next, we explain how we incorporate uncertainty and interpretability in deep learning based DSSs (Sections 2.2 and 2.3). Finally, we present our method for estimating the uncertainty associated with the importance of input features (Section 2.4).

2.1. Fully convolutional networks

FCNs are CNNs particularly suited to tackle per pixel prediction problems like semantic segmentation, i.e. providing a probability score for what class each pixel belongs to. For instance, in the case of semantic segmentation of colorectal polyps, each pixel is labeled as a polyp or as part of the colon (background class). Segmentation is considered a more challenging task than detecting or localizing an object in an image, but provides more information. The shape information provided by a meaningful segmentation map can for example be used to study anatomical structures or inspect other regions of interest (Sharma et al., 2010).

We investigate three architectures for the task of polyp segmentation, namely the Fully Convolutional Network 8 (FCN-8) (Shelhamer et al., 2017), U-Net (Ronneberger et al., 2015) and SegNet (Badrinarayanan et al., 2017) for the following reasons. These networks have been applied in a number of different domains and are chosen to form a well-understood foundation for our studies. This enables uncertainty and interpretability experiments to be the main focus. Previous use of the FCN-8 for polyp segmentation has shown promising results (Vázquez et al., 2016; Brandao et al., 2017). SegNet has been shown to achieve comparable results to the FCN-8 in some applications but is a less memory intensive approach with fewer parameters to optimize. U-Net has previously demonstrated encouraging results on medical tasks and does also contain fewer parameters than the FCN-8, thus providing a lightweight alternative. We include these different networks in this study in order to compare what features are considered important by different models and how uncertainty estimates differ among networks. The interested reader can find a detailed description along with figures of the three models in Appendix A.

2.2. Uncertainty in fully convolutional networks

Despite their success on a number of different tasks, CNNs are not without flaws. One of these flaws, which becomes especially apparent for medical applications, is their inability to provide any notion of uncertainty in their prediction. When a physician is considering the symptoms of a patient and contemplates what medication to prescribe there might be several viable options, and the final decision might spell the difference between a fatal or favorable outcome. Since the stakes are so high, physicians will have to weight the different options and reflect on which choice is most likely to have a favorable outcome. If a physician decides to consult a DSS based on a CNN, she or he would be presented with a recommendation that has no indication as to how likely a desirable outcome is, thus making it difficult for the physician to trust the system. Although the softmax output regularly found at the end of a CNN is sometimes interpreted as model confidence, this is generally ill-advised (Gal and Ghahramani, 2016) and other approaches must be considered.

In contrast, Bayesian models provide a framework which naturally includes uncertainty by modeling posterior distribution for the quantities in question. Given a dataset $\mathcal{D} \equiv \left\{ \mathbf{x}_n \in \mathbb{R}^D, \mathbf{y}_n \in \mathbb{R}^C \right\}_{n=1}^{N}$, where $\mathbf{x}_n$ denotes an input vector and $\mathbf{y}_n$ denotes its corresponding one-hot encoded label vector, the predictive distribution of a Bayesian neural network for a new pair of

[2] This work significantly extends our preliminary study (Wickstrøm et al., 2018) by: (1) Including U-Net in our analysis; (2) significantly extending our experimental section by including new experiments on the 2015 MICCAI polyp detection challenge (Bernal et al., 2017) and the Endoscene dataset (Vázquez et al., 2016) (3) proposing a novel method for estimating uncertainty in the importance of input features and evaluating our proposed method on two polyp segmentation datasets; (4) providing a more thorough literature background discussion and placing our work into a broader context.

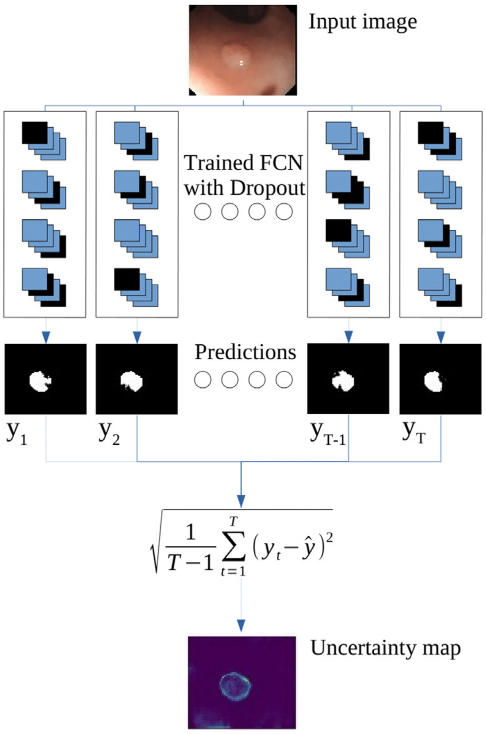


Fig. 1. Illustration of the Monte Carlo Dropout procedure. The same input image is passed through a trained FCN with Dropout applied T times, resulting in T different predictions. The standard deviation of each pixel is then estimated based on these T predictions.

samples $\{\mathbf{x}_*, \mathbf{y}_*\}$ can be modeled as:

$$p(\mathbf{y}_*|\mathbf{x}_*, \mathcal{D}) = \int p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{W})p(\mathbf{W}|\mathbf{x}_*, \mathcal{D})d\mathbf{W} \quad (1)$$

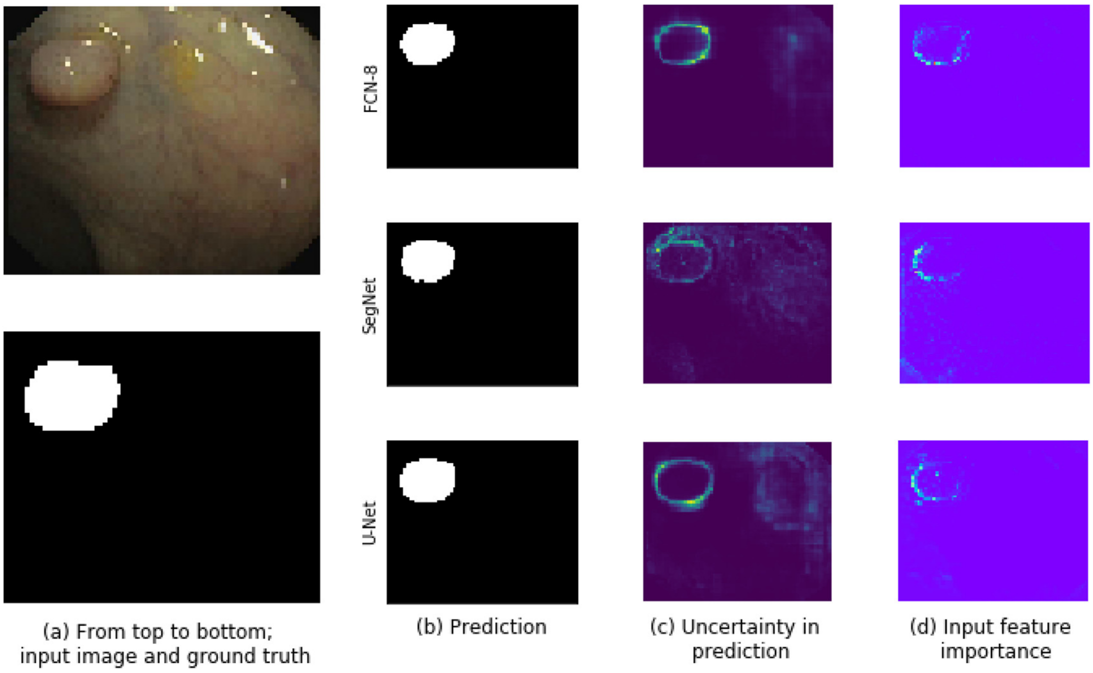
In Eq. (1), $\mathbf{W}$ refers to the weights of the model, $p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{W})$ is the softmax function applied to the output of the model, denoted by $f_{\mathbf{W}}(\mathbf{x}_*)$, and $p(\mathbf{W}|\mathbf{x}_*, \mathcal{D})$ is the posterior over the weights which capture the set of plausible model parameters for the given data. Obtaining $p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{W})$ only requires a forward pass of the network, but the inability to evaluate the posterior of the weights analytically makes Bayesian neural networks computationally infeasible. To sidestep the problematic posterior of the weights, (Gal and Ghahramani, 2016) proposed to incorporate Dropout as a method for sampling sets of weights from the trained network to approximate the posterior of the weights. The predictive distribution from Eq. (1) can then be approximated using Monte Carlo integration as follows:

$$p(\mathbf{y}_*|\mathbf{x}_*, \mathcal{D}) \approx \frac{1}{T}\sum_{t=1}^{T}\text{Softmax}(f_{\mathbf{W}_t^*}(\mathbf{x}_*)) \quad (2)$$

where T is the number of sampled sets of weights and $\mathbf{W}^*{}_t$is a set of sampled weights. In practice, the predictive distribution from Eq. (2) can be estimated by running T forward passes of a model with Dropout applied to produce T predictions and then computing the standard deviation over the softmax outputs of the T samples. We will refer to these uncertainty estimates as uncertainty maps. This method of utilizing Dropout for sampling from the posterior of the predictive distribution is referred to as Monte Carlo Dropout, and the method is illustrated in Fig. 1.

2.3. Interpretability in fully convolutional networks

Another desirable property which CNNs lack is interpretability, i.e. being able to determine what features induce the network to produce a particular prediction. For instance, a physician might be interested in discerning what information the prediction of a given DSS is based on, and if it concurs with medical knowledge. A CNN-based DSS has no inherent way of providing such an explanation. However, several recent works have proposed different methods to increase network interpretability (Zeiler and Fergus, 2014; Bach et al., 2015). In this paper, we evaluate and develop the Guided Backpropagation (Springenberg et al., 2015) technique for FCNs on the task of semantic segmentation of colorectal polyps in order to



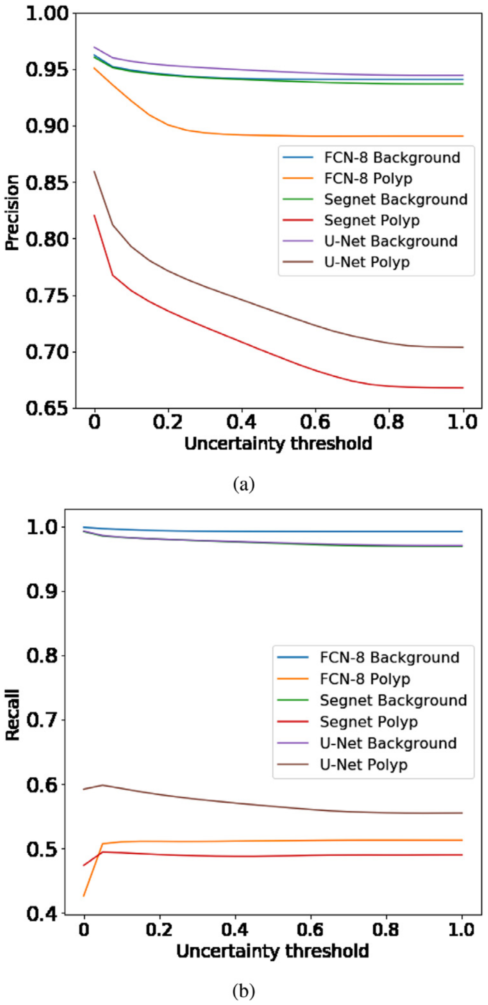
(a) From top to bottom; input image and ground truth

(b) Prediction

(c) Uncertainty in prediction

(d) Input feature importance

Fig. 2. Figure displays the prediction, uncertainty map, and interpretability map for the FCN-8, SegNet and U-Net, for the input image shown in the leftmost column. Best viewed in color.



(a)

(b)

Fig. 3. Precision and recall vs uncertainty plot for background and polyp class on the Endoscene test set.

assess which pixels in the input image the network deems important for identifying polyps. We choose Guided Backpropagation as it is known to produce clearer visualizations of salient input pixels compared to other methods (Zeiler and Fergus, 2014; Simonyan et al., 2013). We refer to these visualizations of salient pixels as interpretability maps.

The central idea of Guided Backpropagation is the interpretation of the gradients of the network with respect to an input image. Simonyan et al. (2013) exploited that, for a given image, the magnitude of the gradients indicate which pixels in the input image need to be changed the least to affect the prediction the most. By utilizing backpropagation (Rumelhart et al., 1988; Werbos, 1974), they obtained the gradients corresponding to each pixel in the input such that they could visualize what features the network considers essential. Springenberg et al. (2015) argued that positive gradients with a large magnitude indicate pixels of high importance while negative gradients with a large magnitude indicate pixels which the networks want to suppress. If these negative gradients are included in the visualization of important pixels it might result in noisy visualization of descriptive features. In order to avoid noisy visualizations the Guided Backpropagation procedure alters the backward pass of a neural network such that negative gradients are set to zero in each layer, thus allowing only positive gradients to flow backward through the network and highlighting pixels that the system finds important.

2.4. Monte carlo guided backpropagation: Uncertainty in input feature importance

To determine the uncertainty associated with an input feature's importance for the prediction, we propose a novel approach inspired by Monte Carlo Dropout combined with Guided Backpropagation. In Section 2.2 we discussed CNNs inability to produce any notion of uncertainty and described Monte Carlo Dropout, which provides a method to obtain approximate measures of uncertainty for CNNs by utilizing Dropout during inference. Accompanying a model's prediction with an uncertainty estimate adds the option to assess if a particular prediction is highly certain or a case that could require further analysis from a human expert. In Section 2.3 we described Guided Backpropagation, a technique developed to visualize the relative importance of input features for CNNs by considering the positive gradients from a backward pass through the network. But, determining the importance of the input features based on gradients from a single backward pass encounters the same issue we discussed regarding decisions based on predictions from a single forward pass. How confident are we that these features are important for the decision of the network?

Given a new sample $\mathbf{x}_*$, we want to find the gradients that correspond to the input features, denoted by $\boldsymbol{\delta}^0$. Taking a similar approach as in Section 2.2, the approximate predictive distribution for the gradients of the input features is given by

$$q(\boldsymbol{\delta}^0|\mathbf{x}_*) = \int p(\boldsymbol{\delta}^0|\mathbf{x}_*, \boldsymbol{\theta})q(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{3}$$

Calculating $p(\boldsymbol{\delta}^0|\mathbf{x}_*, \boldsymbol{\theta})$ is done through the backpropagation algorithm, i.e. computing the gradients with respect to the output of the network and then using the chain rule to work backward toward the input gradients. Also, we modify the backward pass such that negative gradients are canceled, following the Guided Backpropagation procedure. For clear notation, we denote this procedure as $\nabla_{\boldsymbol{\theta}} f^{gb}(\mathbf{x}_*; \boldsymbol{\theta})$, where $\nabla_{\boldsymbol{\theta}}$ indicate finding the gradients of each layer with respect to the parameters of the network and $f^{gb}(\mathbf{x}_*; \boldsymbol{\theta})$ is the prediction of the model with the modified backward pass. The predictive distribution in Eq. (1) can then be approximated using Monte Carlo integration as follows:

$$q(\boldsymbol{\delta}^0|\mathbf{x}_*) = \frac{1}{T}\sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} f^{gb}(\mathbf{x}_*; \mathbf{W}_t^*). \tag{4}$$

In practice, this amounts to performing T forward and backward passes with Dropout applied and computing the standard deviation over the gradients of each input pixel over all T samples. We refer to this method of estimating gradient uncertainty as Monte Carlo Guided Backpropagation.

3. Experiments

3.1. Experimental setup

We evaluate our methods on a recent benchmark dataset for polyp segmentation, namely the EndoScene dataset (Vázquez et al., 2016), which consists of 912 RGB images obtained from colonoscopies of 36 patients. Each input image has a corresponding annotated (labeled) image provided by physicians, where pixels belonging to a polyp are marked in white and pixels belonging to the colon are marked in black. We consider the binary task of classifying each pixel as polyp or part of the colon (background class). Following the approach of Vázquez et al. (2016) we separate the dataset into a training, validation, and test set. The training set consists of 20 patients and 547 images, the validation set consists of 8 patients and 183 images, and the test set consists of 8 patients and 182 images. All RGB input images are normalized to the range [0,1]. All models were trained using ADAM (Kingma and Ba, 2014)

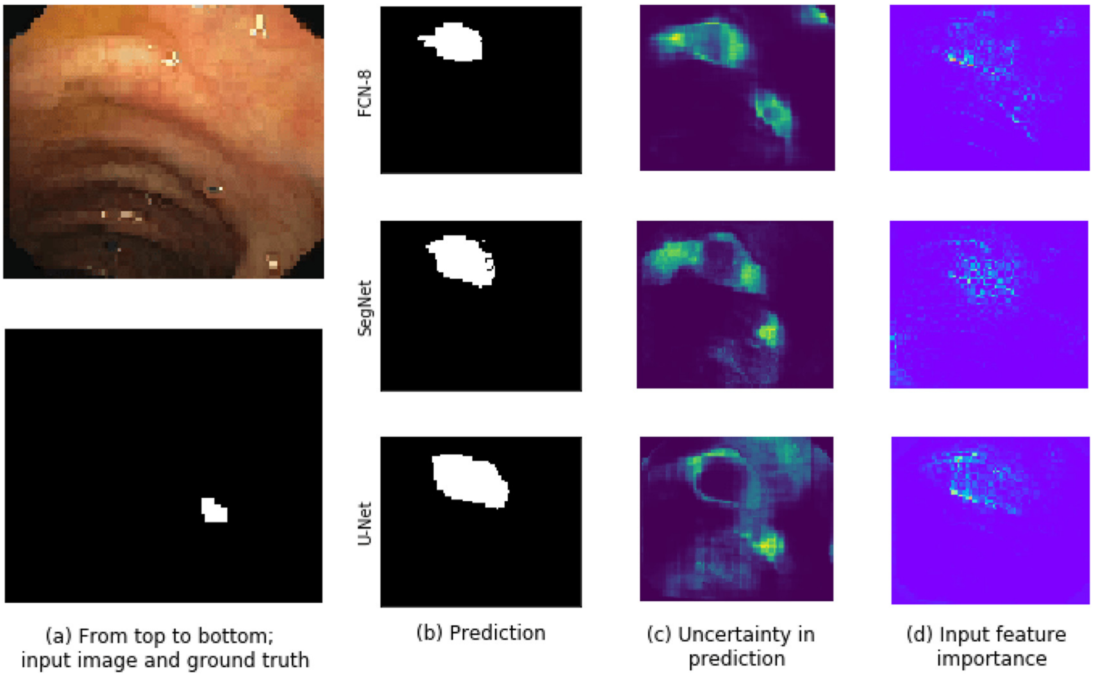

FCN-8
SegNet
U-Net
(a) From top to bottom; input image and ground truth
(b) Prediction
(c) Uncertainty in prediction
(d) Input feature importance

Fig. 4. Figure displays the prediction, uncertainty map, and interpretability map for the FCN-8, SegNet and U-Net, for the input image shown in the leftmost column. Best viewed in color.

Table 1
Results on the EndoScene test dataset.

Model	# Parameters(M)	IoU background	IoU polyp	Mean IoU	Global Accuracy
SDEM (Bernal et al., 2014)	-	0.799	0.221	0.412	0.756
U-Net	27.5	0.945	0.516	0.723	0.945
SegNet	29.5	0.933	0.522	0.727	0.935
FCN-8 (Vázquez et al., 2016)	134.5	**0.946**	0.509	0.727	**0.949**
FCN-8	134.5	**0.946**	**0.587**	**0.767**	**0.949**

with a batch size of 10 and a cross-entropy loss. We use the validation set to apply early stopping by monitoring the polyp IoU score with a patience of 30. For performance evaluation, we calculate the Intersection over Union (IoU) metric and global accuracy (per-pixel accuracy) on the test set. For a given class c, prediction $\hat{y}_i$ and ground truth y_i, the IoU is defined as

$$\text{IoU}(c) = \frac{\sum_i (\hat{y}_i == c \land y_i == c)}{\sum_i (\hat{y}_i == c \lor y_i == c)} \tag{5}$$

where $\land$ is the logical *and* operation and $\lor$ is the logical *or* operation.

Additionally, we evaluated our proposed method for estimating uncertainty in input feature importance on the 2015 MICCAI polyp detection challenge (Bernal et al., 2017). As the test images of this dataset are of high quality and our proposed approach is mostly a visual technique, assessing our method on this data will provide further validation of our method.

3.2. Quantitative and qualitative results

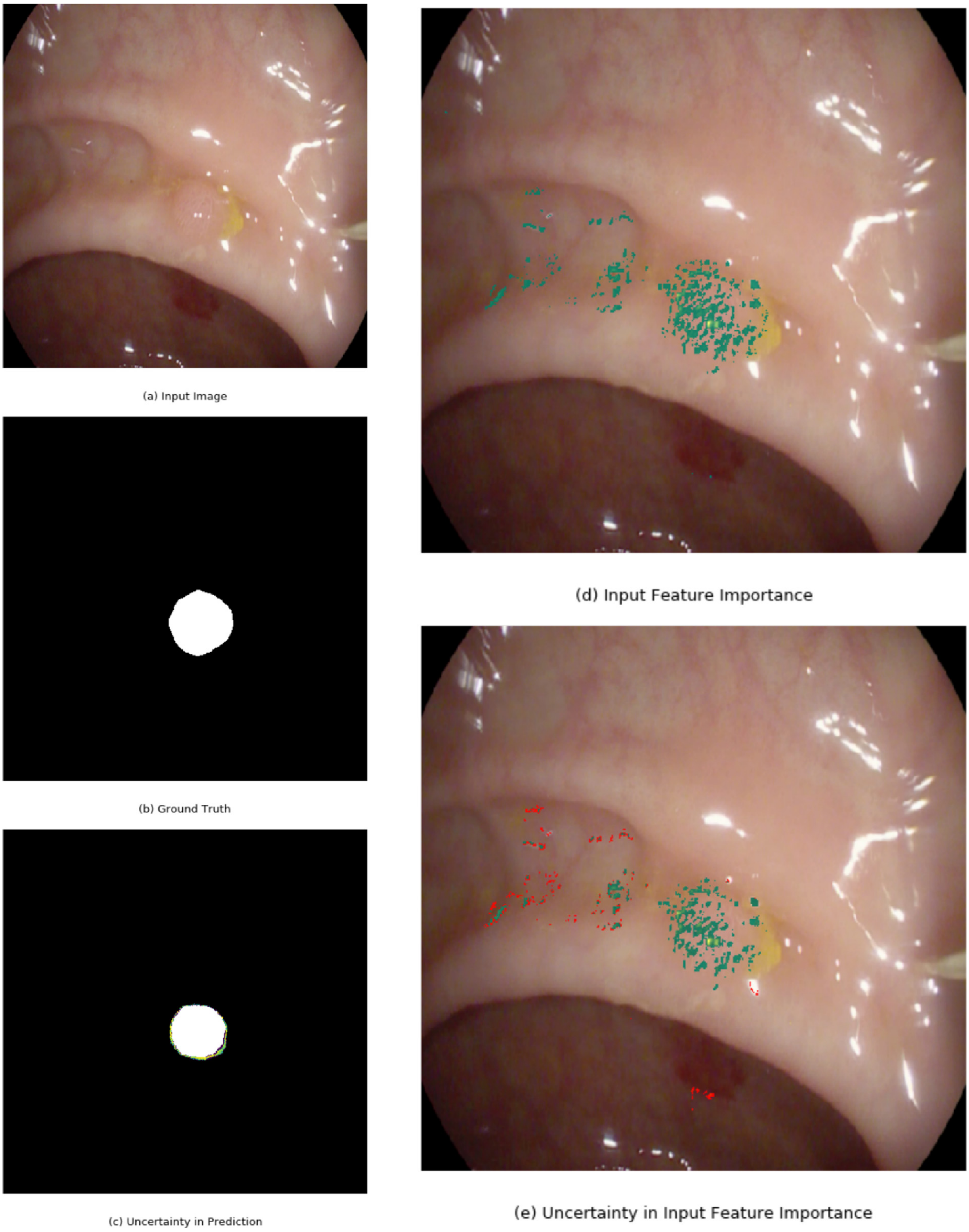
*Quantitative results*In Table 1 we report our results for the FCN-8, SegNet and U-Net along with the results of previous works on polyp segmentation from both traditional machine learning and deep learning based approaches. The traditional machine learning method computes a histogram based on the pixel values and uses peaks and valleys information from the histogram to perform segmentation. It is referred to as the Segmentation from Energy Maps (SDEM) algorithm (Bernal et al., 2014). For the deep learning approach, segmentation is performed using the FCN-8, but without Batch Normalization or transfer learning. This approach is referred to as FCN-8 in Table 1. The results show that all deep learning approaches significantly outperform the more traditional machine learning approach, and the difference in performance between our implementation of the FCN-8 and that of Vázquez et al. (2016) demonstrates that including recent advances in deep learning methodology can improve performance.

Qualitative results Fig. 2(b) and 4(b) displays some qualitative results on the test data for the FCN-8, SegNet and U-Net. Fig. 2 shows a typical example where a large, elliptical polyp is located with high precision by all three models. In Fig. 4 we present a more challenging example where all models fail to locate the small polyp present in the image. Interested readers can find additional results in Appendixs B and C.

3.3. Modeling uncertainty in prediction

Figs. 2(c) and 4(c) present examples of uncertainty estimation for the FCN-8, SegNet and U-Net, respectively, using Monte Carlo Dropout. These uncertainty maps are obtained by sampling 10 predictions from each model with a dropout rate of 0.5 and estimating the standard deviation for each pixel. Pixels displayed in bright green are associated with high uncertainty while pixels displayed in dark blue are associated with low uncertainty.

The example shown in Fig. 2 shows that all models have high confidence for most pixels in their prediction, with the exception of pixels around the border of the polyp itself. This is reasonable, as it is difficult to assess exactly where the polyp starts and the colon ends. In the example shown in Fig. 4, where all models make



(a) Input Image

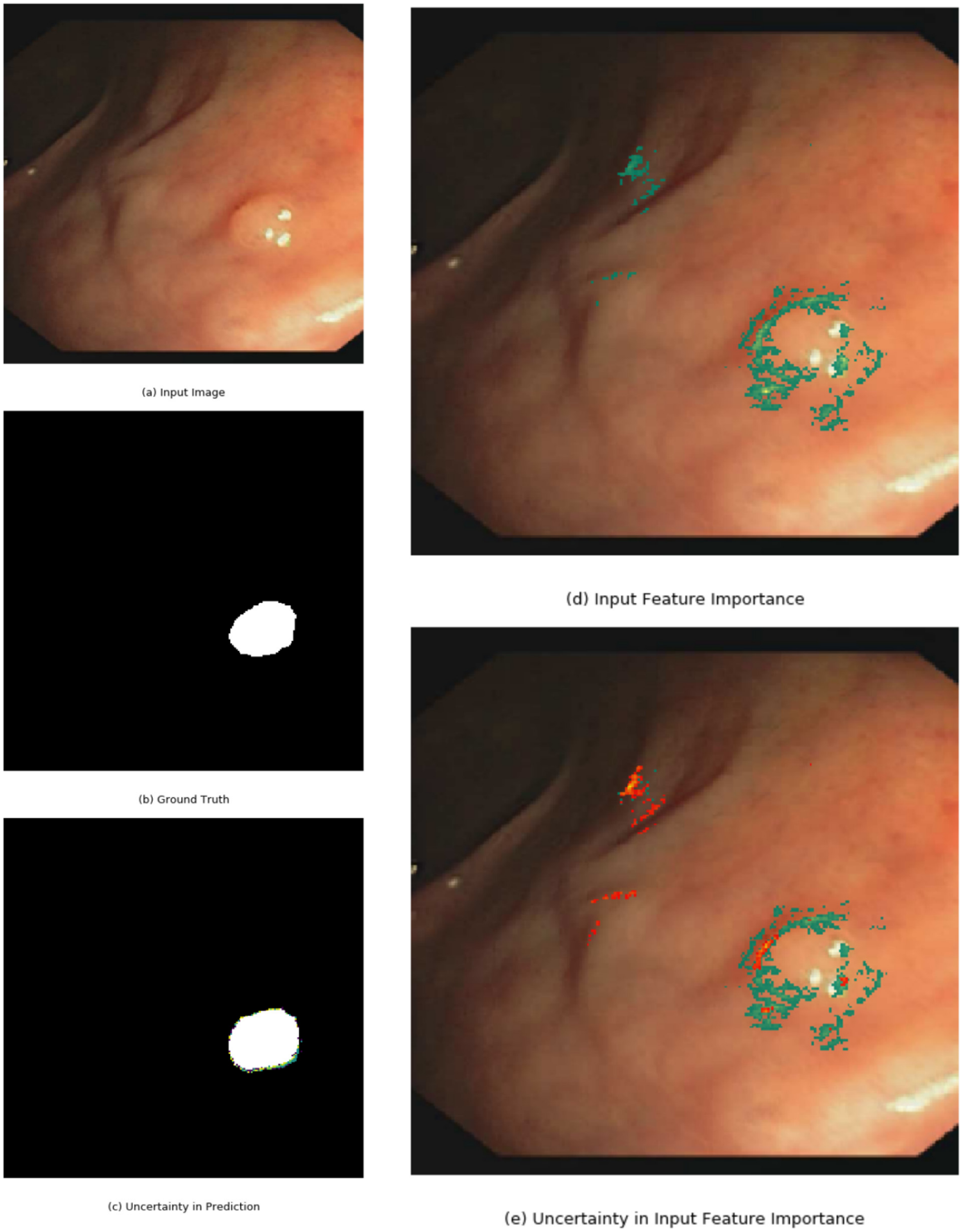
(b) Ground Truth

(c) Uncertainty in Prediction

(d) Input Feature Importance

(e) Uncertainty in Input Feature Importance

Fig. 5. Figure displays input image (a), ground truth (b), prediction with uncertainty overlaid (c), input feature importance (d), and uncertainty in input feature importance (e). For the uncertainty in input feature importance results, pixels colored green indicate that the features are important for the prediction of polyps and that the model is certain of its importance. Pixels colored red indicate features that might be important for the prediction of polyps but the model is uncertain of its importance. Best viewed in color. Input image originated from the MICCAI dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(a) Input Image

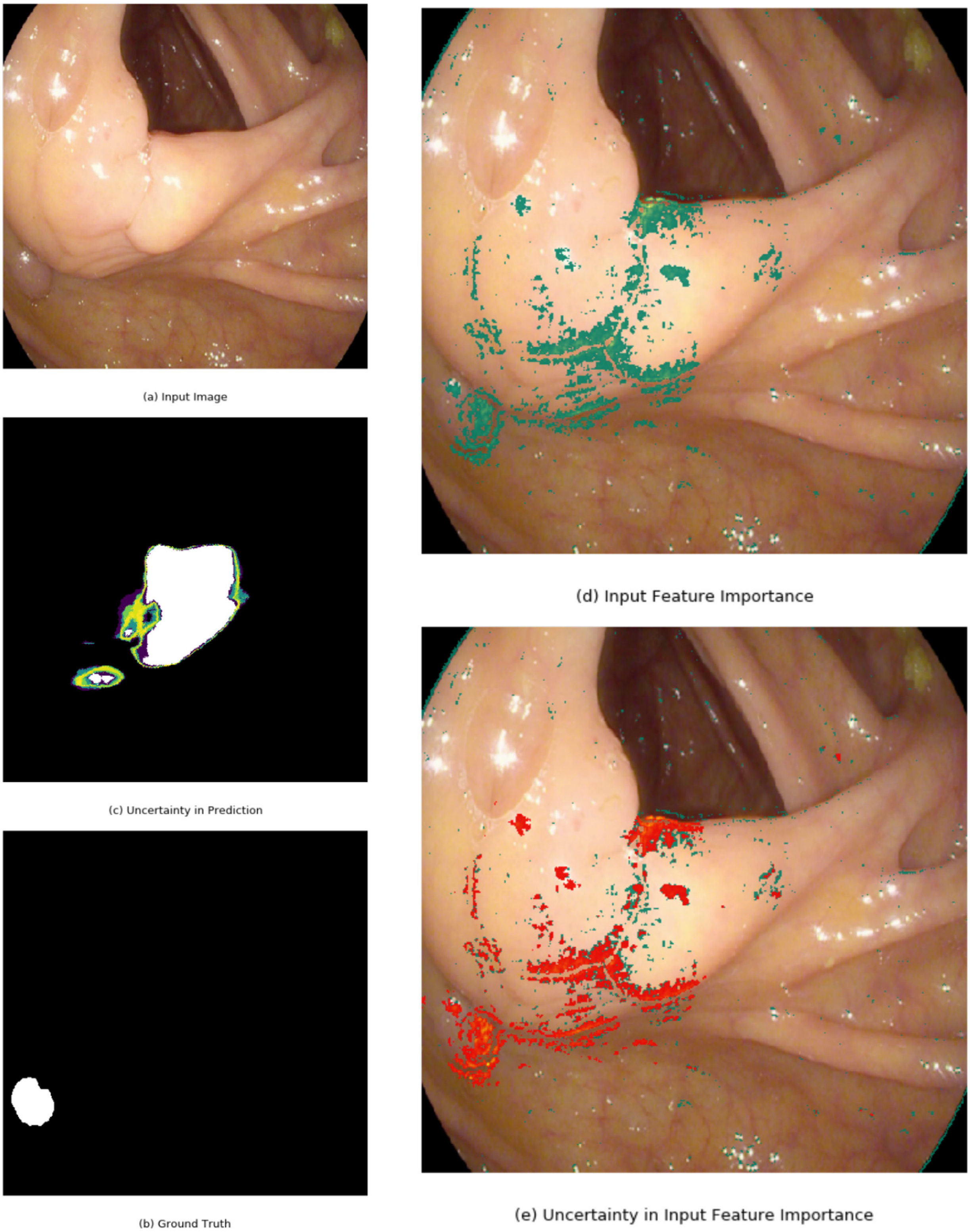
(b) Ground Truth

(c) Uncertainty in Prediction

(d) Input Feature Importance

(e) Uncertainty in Input Feature Importance

Fig. 6. Figure displays input image (a), ground truth (b), prediction with uncertainty overlaid (c), input feature importance (d), and uncertainty in input feature importance (e). For the uncertainty in input feature importance results, pixels colored green indicate that the features are important for the prediction of polyps and that the model is certain of its importance. Pixels colored red indicate features that might be important for the prediction of polyps but the model is uncertain of its importance. Best viewed in color. Input image originated from the Endoscene dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(a) Input Image

(d) Input Feature Importance

(c) Uncertainty in Prediction

(b) Ground Truth

(e) Uncertainty in Input Feature Importance

Fig. 7. Figure displays input image (a), ground truth (b), prediction with uncertainty overlaid (c), input feature importance (d), and uncertainty in input feature importance (e). For the uncertainty in input feature importance results, pixels colored green indicate that the features are important for the prediction of polyps and that the model is certain of its importance. Pixels colored red indicate features that might be important for the prediction of polyps but the model is uncertain of its importance. Best viewed in color. Input image originated from the MICCAI dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

inaccurate predictions, the uncertainty estimates look notably different, with large regions of uncertainty for all three models. The examples shown in Figs. 2 and 4 demonstrate how seemingly similar predictions can have different uncertainty estimates for the different types of networks investigated in this work, and that erroneous predictions show distinctively different uncertainty estimates than correct predictions.

Fig. 3 displays how precision and recall is related to uncertainty in predictions. It shows the overall precision and recall for each class on the Endoscene test dataset when pixel with a mean-class uncertainty above a certain threshold are excluded. The estimated uncertainty for each class have been normalized into values between 0 and 1. Results in Fig. 3 (a) display how precision decreases as more pixel predictions with high uncertainty are included. This connection between precision and uncertainty agrees with the qualitative examples in Figs. 2 and 4 discussed above. Results in Fig. 3 (b) show how recall slightly increases for the polyp class at a low uncertainty threshold, but then remains unchanged for both classes. The interested reader can find a similar experiment on the MICCAI dataset in Appendix C.

3.4. Modeling input feature importance

Figs. 2 (d) and 4 (d) show examples where Guided Backpropagation has been used to analyze the FCN-8, SegNet and U-Net, respectively. Pixels displayed in bright green are associated with pixels that are important to the prediction of the model while pixels displayed in blue are associated with pixels that are less important to the final prediction.

Fig. 2 indicates that all models are considering the edges of the polyp to make their prediction, where particularly the leftmost and bottom edge of the polyp is highlighted as important by all models. Fig. 4, where all models fail to locate the polyp, displays more disagreement between the models as to what pixels are important.

3.5. Modeling uncertainty in input feature importance

In order to focus on the new methodology we only use one model to evaluate our proposed method. The overall best performing segmentation model, FCN-8, was chosen to evaluate the proposed methodology for estimating uncertainty in input feature importance and demonstrate its merit. Figs. 5–7 presents examples of uncertainty estimation for input feature importance for the FCN-8 using Monte Carlo Guided Backpropagation. These results are obtained by sampling 10 gradient estimates from each model with a dropout rate of 0.5. The figures display: (a) the input image; (b) the ground truth; (c) prediction with uncertainty overlaid; (d) input feature importance; and (e) uncertainty in input feature importance. For the uncertainty in input feature importance results, pixels colored green indicate that the features are important for the prediction of polyps and that the model is certain of its importance. Pixels colored red indicate features that might be important for the prediction of polyps but the model is uncertain of its importance. Examples shown in Figs. 5 and 7 are from the test set of the MICCAI dataset while the example shown in Fig. 6 is from the test set of the Endoscene dataset. Interested readers can find additional examples of uncertainty estimation for input feature importance in Appendix B.

Fig. 5 displays an example where the FCN-8 makes a successful segmentation. The interpretability map in Fig. 5 (d) indicates that there are two regions of importance in the input image, one corresponding to the polyp and one region towards the leftmost part of the image. However, the uncertainty in the input feature importance map in Fig. 5 (e) shows that the model is uncertain of the leftmost feature's importance, while the features corresponding to the polyp itself have a high degree of certainty.

Fig. 6 shows another example where the FCN-8 makes a successful segmentation, but also highlight important input features towards the leftmost part of the image, in addition to the polyp itself. Fig. 6 (e) displays that the FCN-8 is highly confident in the importance of the features corresponding to the polyp itself, but indicate a high degree of uncertainty for the highlighted regions towards the leftmost part of the image.

Fig. 7 exhibits an example from the MICCAI dataset where the FCN-8 fails to locate the polyp present in the image, but instead segments a large portion of the colon as polyp. While the interpretability maps in Fig. 7 (d) show large regions of important pixels, it is evident from Fig. 7 (e) that none of the regions have a high degree of importance. As the prediction with uncertainty overlayed in Fig. 7 (e) also indicates regions of uncertainty, practitioners would be wary to trust the model's prediction in this case.

4. Conclusion

In this work we have demonstrated how DSSs based on deep learning can be interpretable and provide uncertainty estimates with their predictions. Moreover, we presented a novel method for estimating uncertainty in input feature importance and demonstrated how this technique can be used to model uncertainty in input pixel importance. Our results demonstrate that the models considered in these experiments exploit edge and shape information of polyps in order to make their predictions and that uncertainty differs significantly between false and correct predictions.

Declaration of Competing Interest

All authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPU used for this research.

Appendix A. Network details

In order to perform per pixel predictions, FCNs employ an encoder-decoder architecture and are capable of end-to-end learning. The encoder network extracts useful features from an image and maps it to a low-resolution representation. The decoder network is tasked with mapping the low-resolution representation back into the same resolution as the input image. Upsampling in FCNs is performed using a fixed upsampling approach, like bilinear or nearest neighbor interpolation, or by learning the upsampling procedure as part of the model optimization via transposed convolutions. Learned upsampling filters add additional parameters to the network architecture, but tend to provide better overall results (Shelhamer et al., 2017). Upsampling can further be improved by including skip connections, which combine coarse level semantic information with higher resolution segmentation from previous network layers. Due to the lack of fully connected layers, inference can be performed on images of arbitrary size.

A1. FCN-8

The FCN-8 was introduced by Shelhamer et al. (2017) and consists of an encoder network and a decoder network, where the encoder network is based on the VGG-16 architecture (Simonyan and Zisserman, 2015) and consists of five encoders. The decoder network consists of three decoders. Dropout (Srivastava et al., 2014), a regularization technique that randomly set units in a layer to zero, is included between all layers of the first decoder. Upsampling is

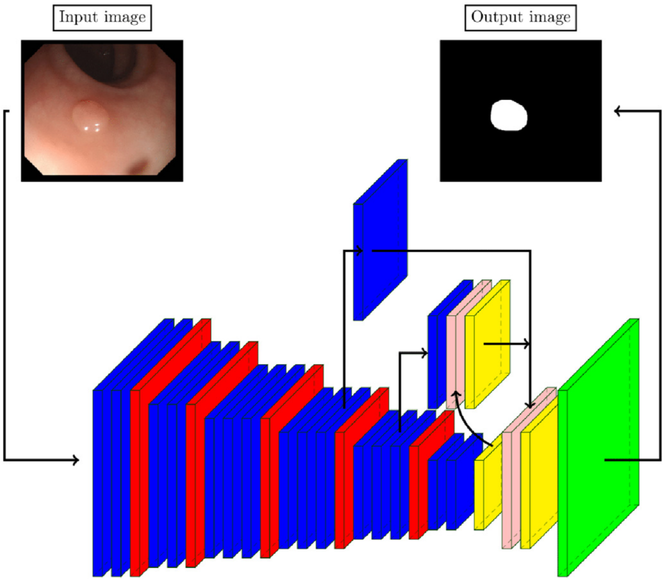


Fig. A.8. An illustration of the FCN-8. Color codes description: Blue - Convolution (3x3), Batch Normalization and ReLU; Yellow - Upsampling; Pink - Summing; Red - Pooling (2x2); Green - Soft-max. Dropout was included as proposed by Simonyan and Zisserman (2015) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

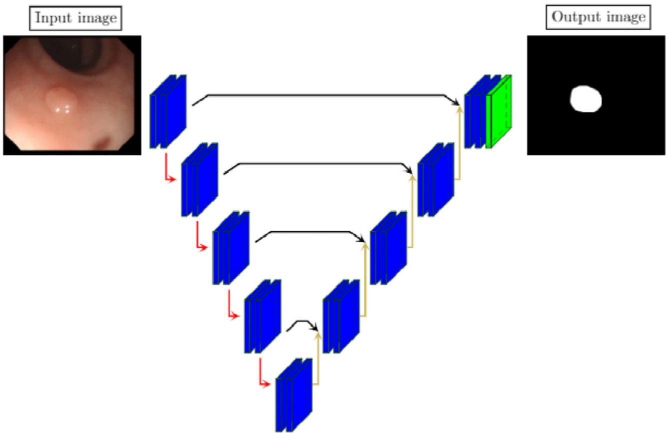


Fig. A.9. An illustration of the U-Net. Color codes description: Blue - Convolution (3x3), Batch Normalization and ReLU; Green - Soft-max; Yellow arrow - Upsampling; Black arrow - Concatenate; Red arrow - Pooling (2x2) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

performed using transposed convolutions at the end of each encoder and skip connections are included between the three central encoders and the decoders. Note that we have added Batch Normalization (Ioffe and Szegedy, 2015) in our implementation and that the encoder weights are initialized with pretrained weights from a VGG16 model (Simonyan and Zisserman, 2015) that was previously trained on the ImageNet dataset (Deng et al., 2009).

A2. U-Net

One of the first networks to build upon FCNs was the U-Net (Ronneberger et al., 2015), which is comprised of an encoder network consisting of five encoders and a decoder network consisting of four decoders. U-Net introduced an alternative method to recover the resolution of the data where the feature maps produced in the fifth encoder is upsampled by a factor of two using transposed convolution and concatenated with the feature maps produced by the fourth encoder. These combined feature maps are passed into the first decoder, which in turn is upsampled and concatenated with the feature maps of the third encoder. This process is repeated until the resolution of the input feature map is recovered. The final decoder is followed by a 1 × 1 convolutions that maps the feature vector into the desired number of classes and a softmax function. Dropout is applied after each layer of the final encoder. We included Batch Normalization after each layer, except for layers preceding a transposed convolution and the final layer.

A3. SegNet

Both the FCN-8 and the U-Net rely on transposed convolutions to recover feature maps with the same resolution as the input features. SegNet (Badrinarayanan et al., 2017), instead, presents another option and is made up of a symmetrically structured encoder decoder network, where the encoder network consists of five encoders based on the VGG-16 (Simonyan and Zisserman, 2015) and the decoder consists of five decoders. The decoder network is identical to the encoder network but with the max-pooling operation replaced by a max-unpooling operation. When a feature map is downsampled the max-pooling indices are stored and used at a later stage to perform non-linear upsampling, a procedure with several advantages. Firstly, it produces sparse feature maps that are computationally attractive and implicit feature selectors. Secondly, it removes the need to learn additional filter for upsampling, thus reducing the number of parameters in the model. Dropout was included after the three central encoders and decoders inspired by Kendall et al. (2015).

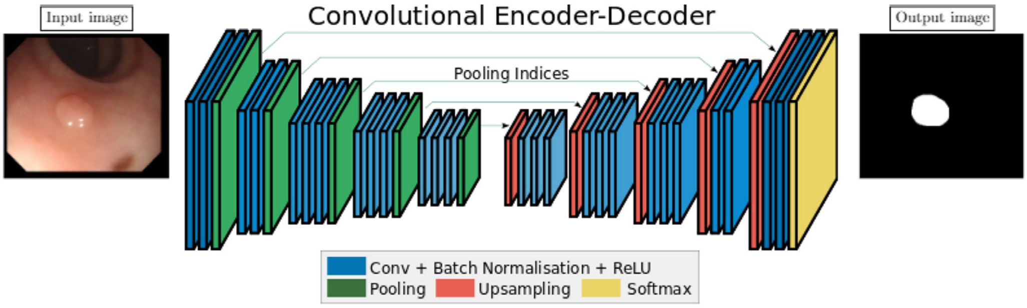

Input image
Convolutional Encoder-Decoder
Pooling Indices
Output image
Conv + Batch Normalisation + ReLU
Pooling
Upsampling
Softmax

Fig. A.10. An illustration of SegNet, originally obtained from Badrinarayanan et al. (2017). Color codes description: Blue - Convolution (3x3), Batch Normalization and ReLU; Green - Soft-max; Yellow arrow - Upsampling; Black arrow - Concatenate; Red arrow - Pooling (2x2) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

Appendix B. additional qualitative results

Figs. B.11–B.13 display additional results on test images from the Endoscene dataset for the FCN-8, SegNet and U-Net, respectively. Each row represents, from top to bottom, input image, ground truth, prediction, uncertainty map, and interpretability map. Results were obtained using the same procedure as described in the main paper.

Figs. B.14–B.16 display additional results of estimating uncertainty in input feature importance for the FCN-8. These results are also obtained following the same procedure described in the main paper.

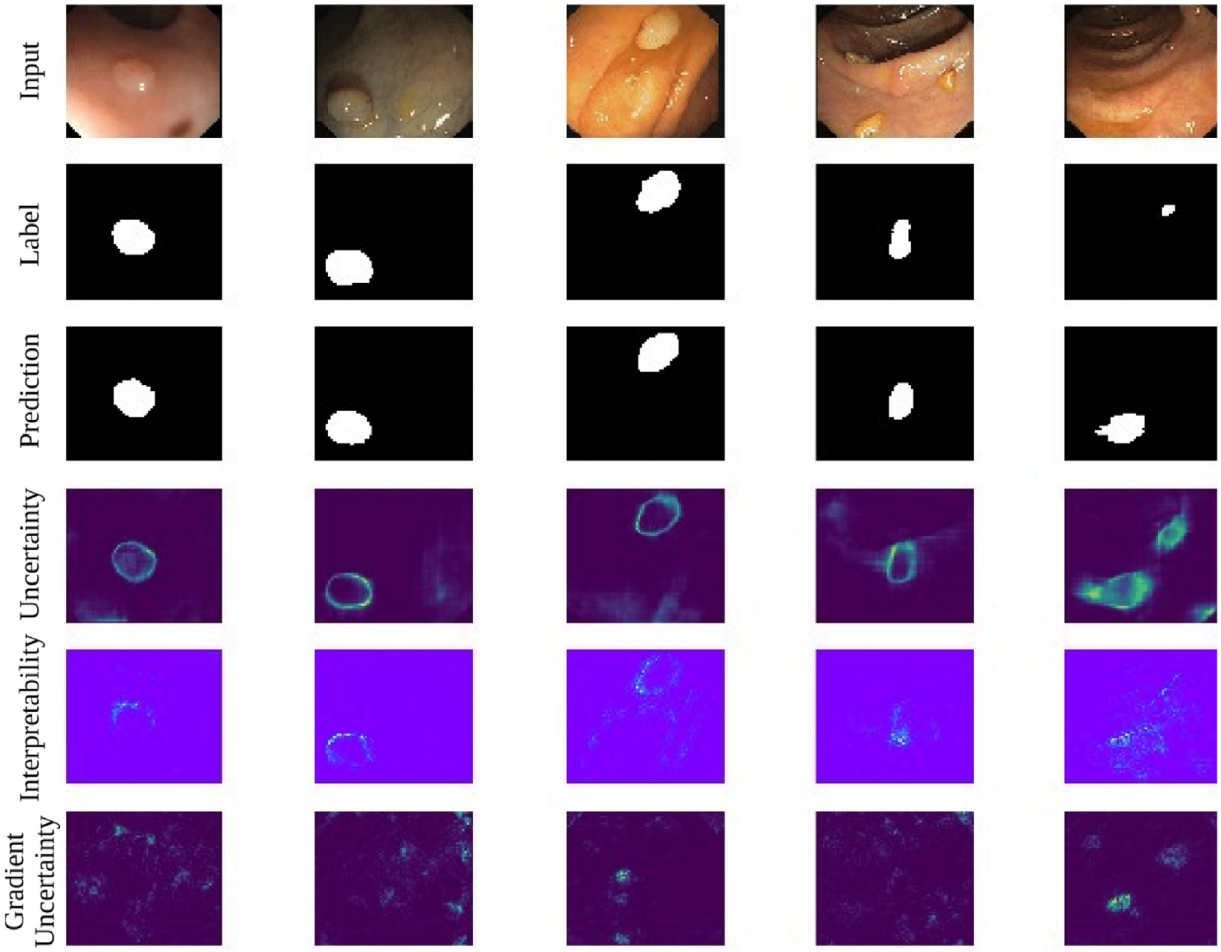


Fig. B.11. Figure displays FCN-8's predictions, the uncertainty map associated with the predictions, and the input features the network deems important. Each row represents, from top to bottom, input image, ground truth, prediction, uncertainty map, and interpretability map. White pixels are classified as polyps and black pixels are classified as background class. For the uncertainty maps, dark blue pixels are associated with low uncertainty and bright green pixels are associated with high uncertainty. For the interpretability maps, bright green pixels are considered important to the prediction of the network. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

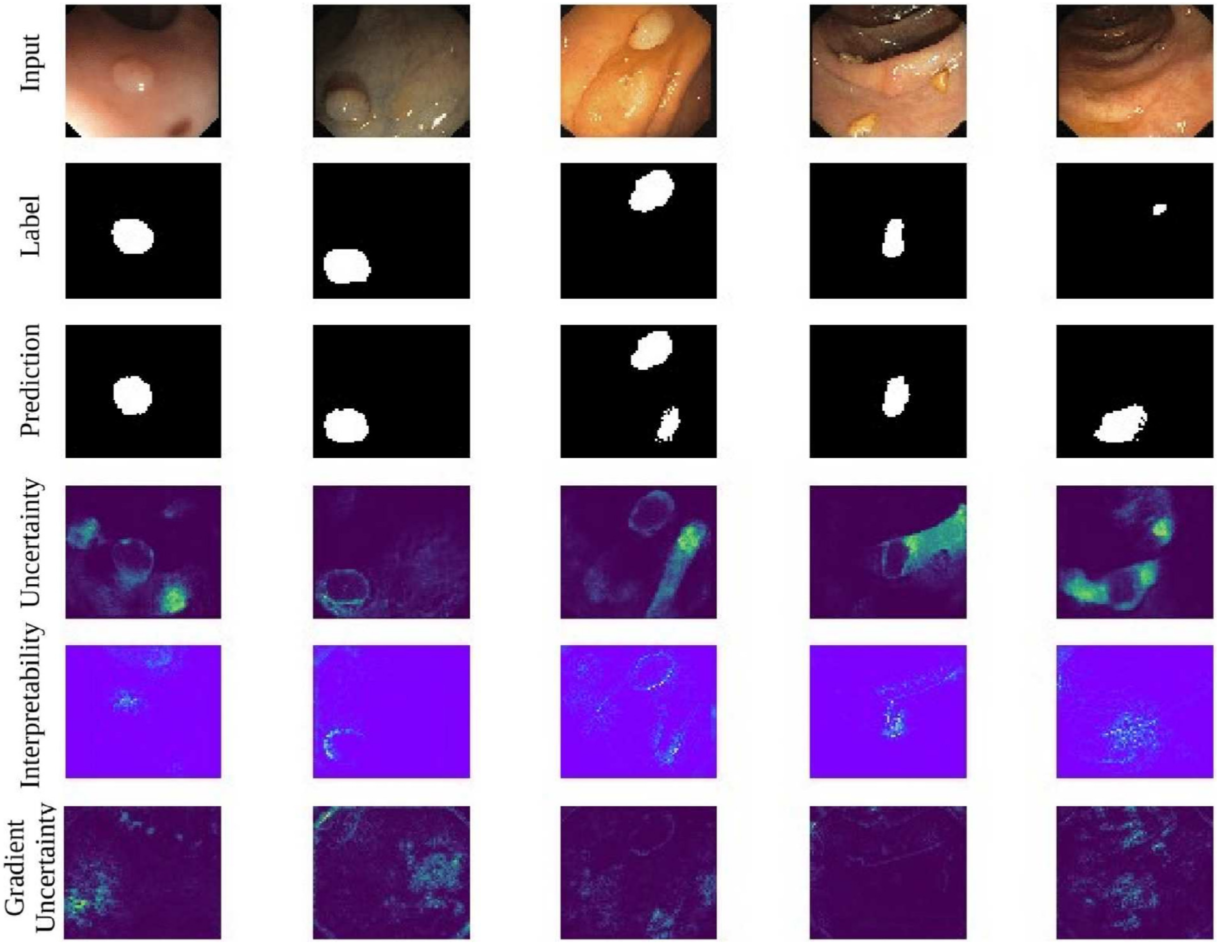


Fig. B.12. Figure displays SegNet's predictions, the uncertainty map associated with the predictions, and the input features the network deems important. Each row represents, from top to bottom, input image, ground truth, prediction, uncertainty map, and interpretability map. White pixels are classified as polyps and black pixels are classified as background class. For the uncertainty maps, dark blue pixels are associated with low uncertainty and bright green pixels are associated with high uncertainty. For the interpretability maps, bright green pixels are considered important to the prediction of the network. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

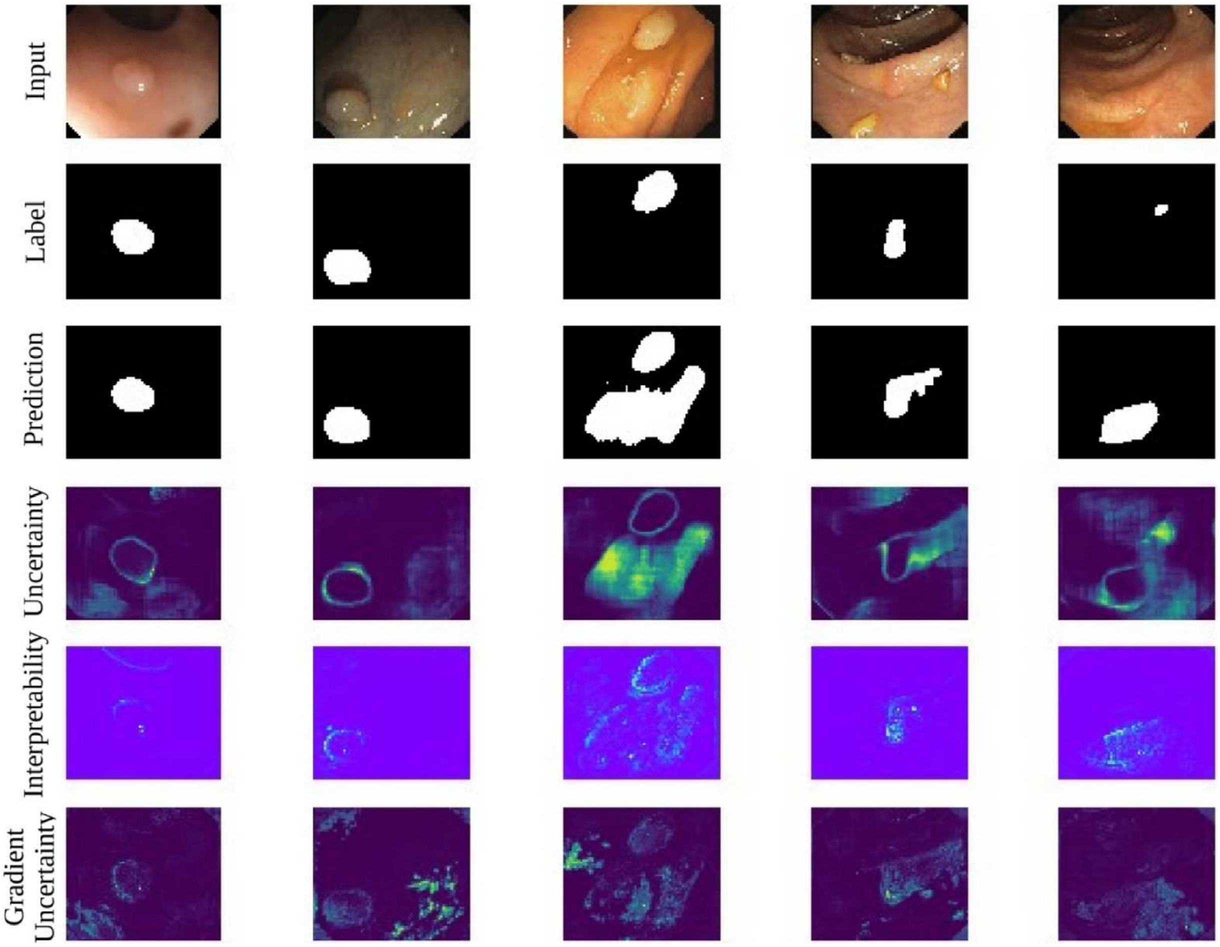


Fig. B.13. Figure displays U-Net's predictions, the uncertainty map associated with the predictions, and the input features the network deems important. Each row represents, from top to bottom, input image, ground truth, prediction, uncertainty map, and interpretability map. White pixels are classified as polyps and black pixels are classified as background class. For the uncertainty maps, dark blue pixels are associated with low uncertainty and bright green pixels are associated with high uncertainty. For the interpretability maps, bright green pixels are considered important to the prediction of the network. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

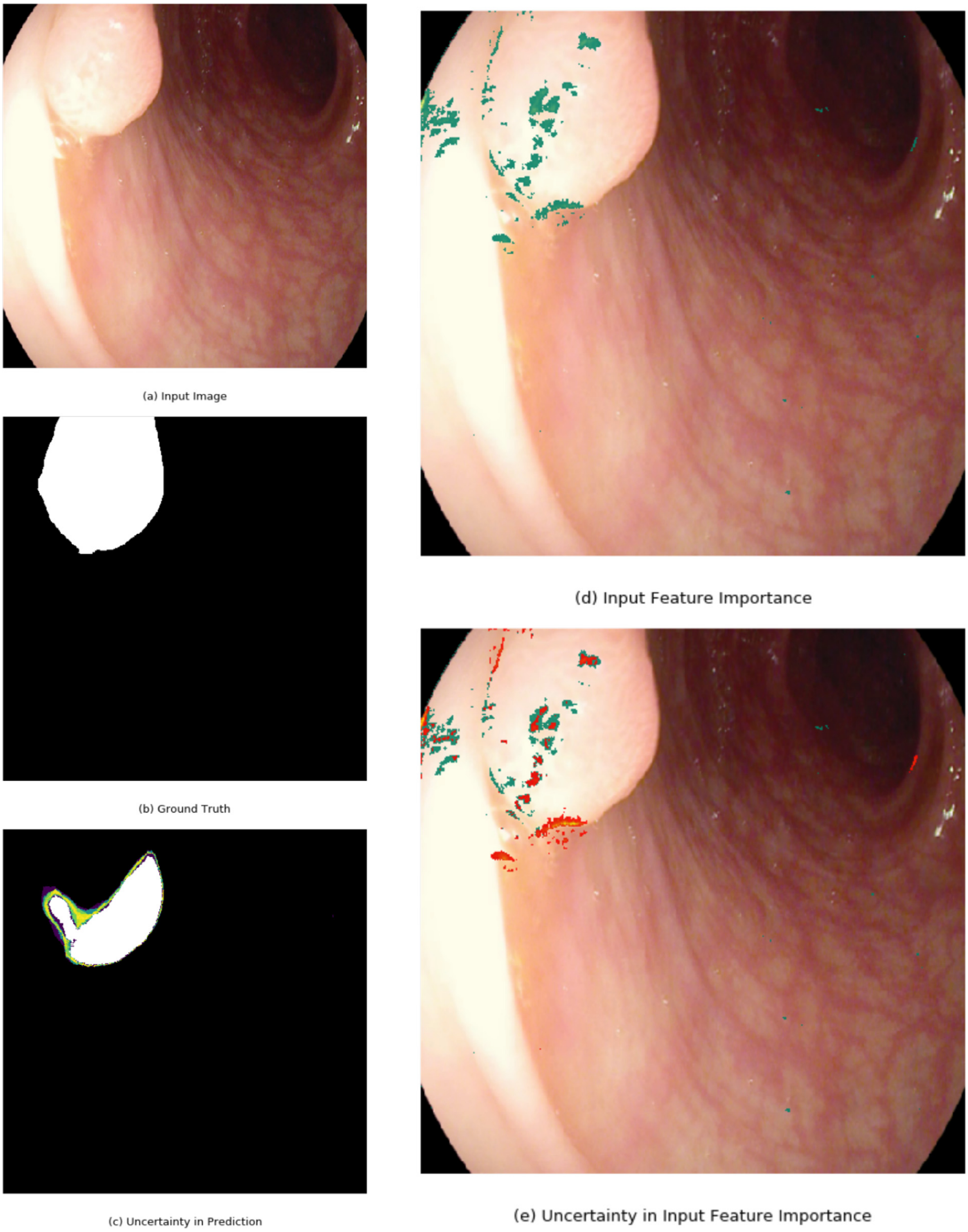
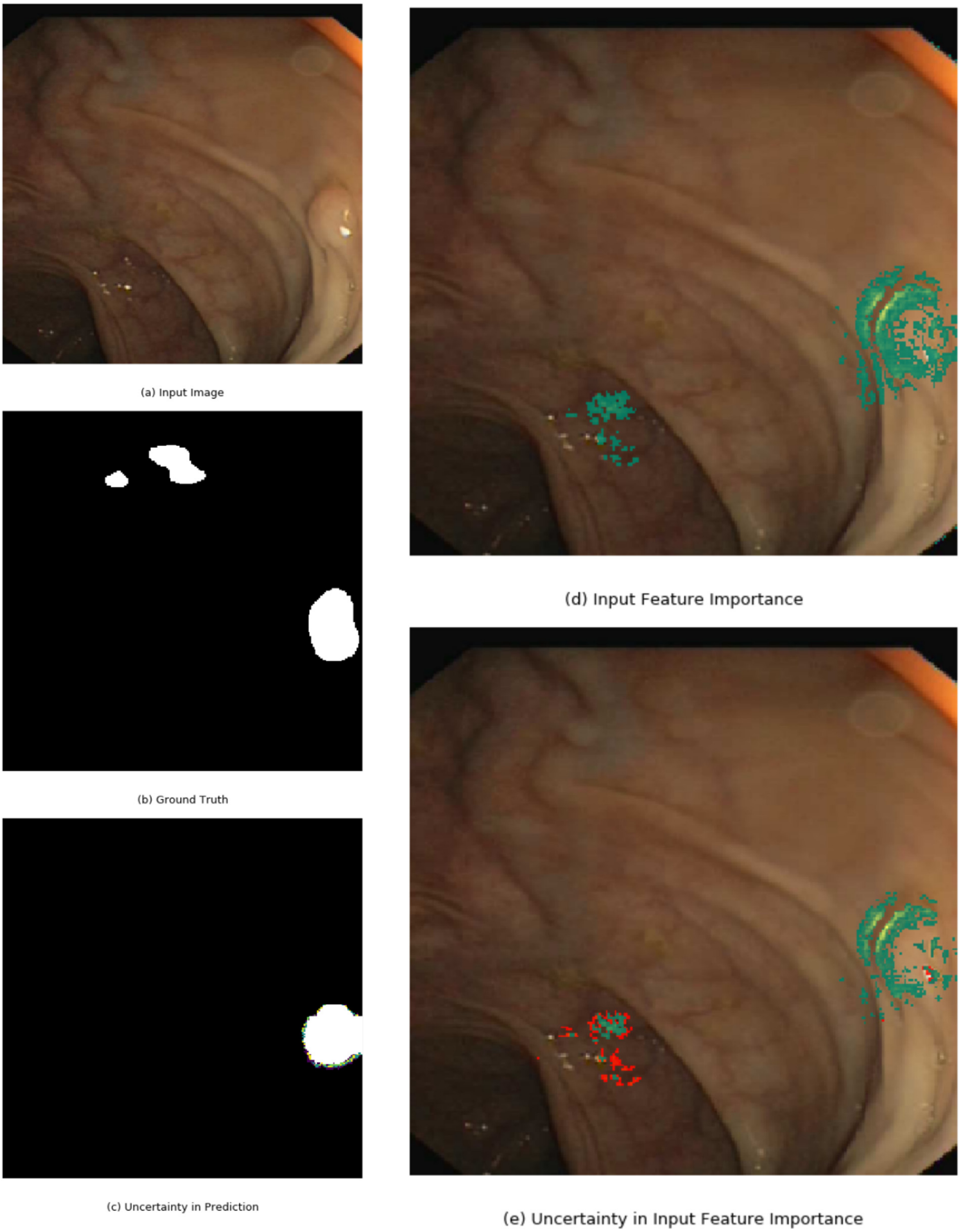

(a) Input Image
(b) Ground Truth
(c) Uncertainty in Prediction
(d) Input Feature Importance
(e) Uncertainty in Input Feature Importance

Fig. B.14. Figure displays input image (a), ground truth (b), prediction with uncertainty overlaid (c), input feature importance (d), and uncertainty in input feature importance (e). For the uncertainty in input feature importance results, pixels colored green indicate that the features are important for the prediction of polyps and that the model is certain of its importance. Pixels colored red indicate features that might be important for the prediction of polyps but the model is uncertain of its importance. Best viewed in color. Input image originated from the MICCAI dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(a) Input Image

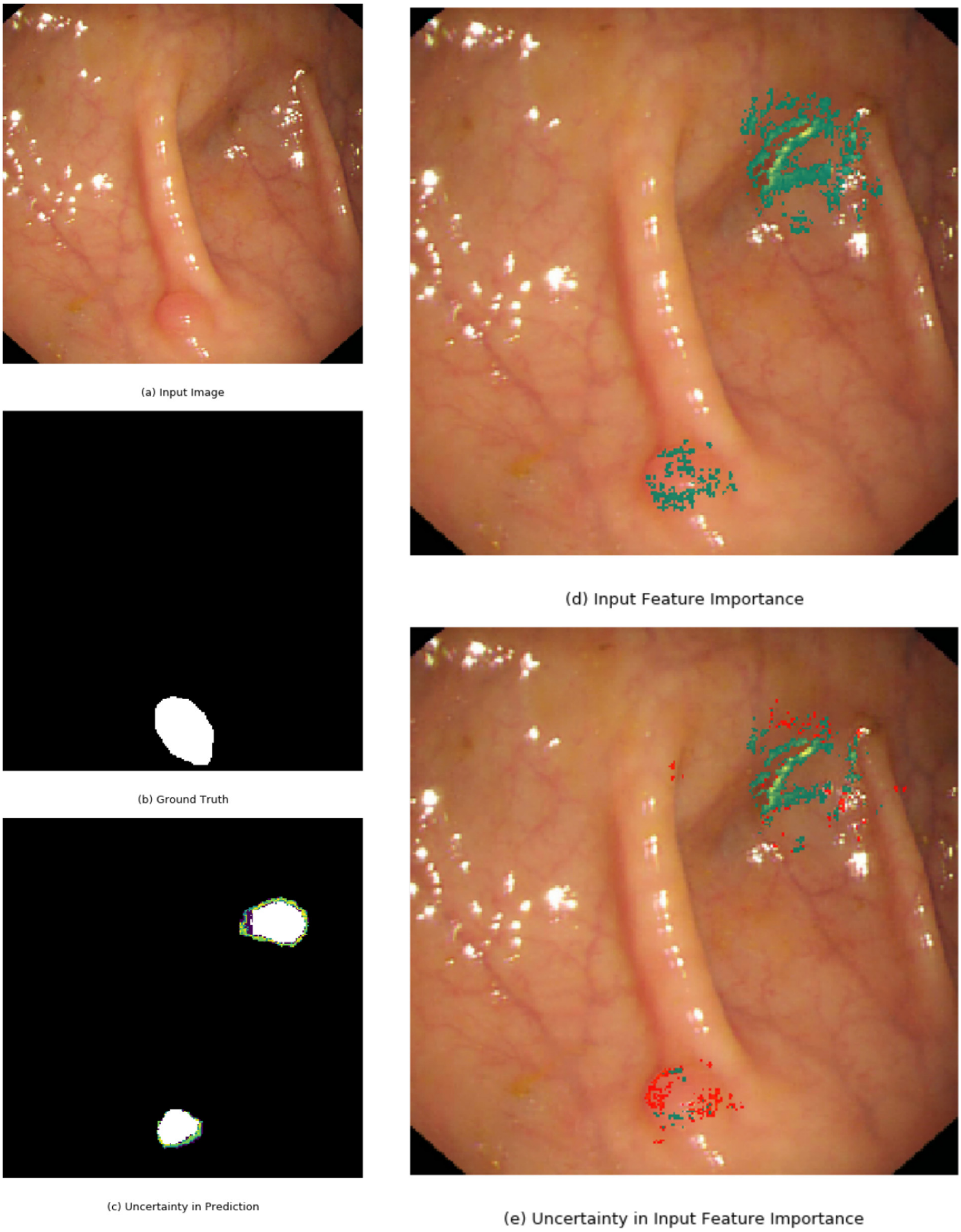
(b) Ground Truth

(c) Uncertainty in Prediction

(d) Input Feature Importance

(e) Uncertainty in Input Feature Importance

Fig. B.15. Figure displays input image (a), ground truth (b), prediction with uncertainty overlaid (c), input feature importance (d), and uncertainty in input feature importance (e). For the uncertainty in input feature importance results, pixels colored green indicate that the features are important for the prediction of polyps and that the model is certain of its importance. Pixels colored red indicate features that might be important for the prediction of polyps but the model is uncertain of its importance. Best viewed in color. Input image originated from the Endoscene dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(a) Input Image

(b) Ground Truth

(c) Uncertainty in Prediction

(d) Input Feature Importance

(e) Uncertainty in Input Feature Importance

Fig. B.16. Figure displays input image (a), ground truth (b), prediction with uncertainty overlaid (c), input feature importance (d), and uncertainty in input feature importance (e). For the uncertainty in input feature importance results, pixels colored green indicate that the features are important for the prediction of polyps and that the model is certain of its importance. Pixels colored red indicate features that might be important for the prediction of polyps but the model is uncertain of its importance. Best viewed in color. Input image originated from the Endoscene dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Appendix C. Additional Qualitative Results on MICCAI dataset

Fig. C.17 and C.18 display additional results on test images from the MICCAI dataset for the FCN-8, SegNet and U-Net, respectively. Results were obtained using the same procedure as described in the main paper. Fig. C.19 displays how precision and recall is related to uncertainty in predictions on the MICCAI test data, similar to the experiment described in Section 3.3.

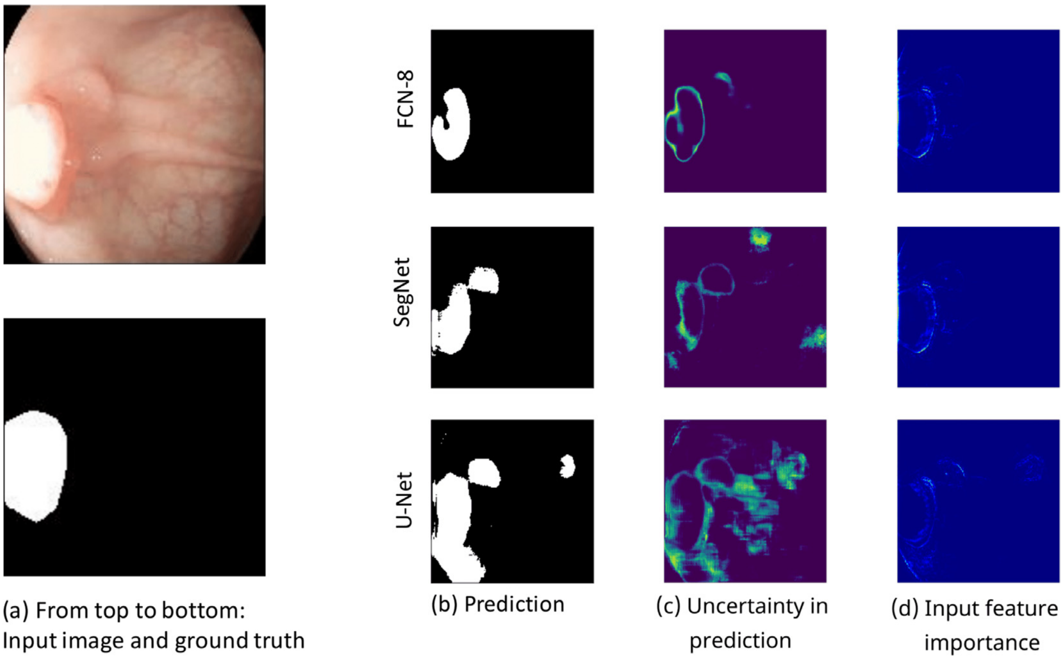

FCN-8
SegNet
U-Net
(a) From top to bottom:
Input image and ground truth
(b) Prediction
(c) Uncertainty in prediction
(d) Input feature importance

Fig. C.17. Figure displays the prediction, uncertainty map, and interpretability map for the FCN-8, SegNet and U-Net, for the input image from the MICCAI dataset shown in the leftmost column. Best viewed in color.

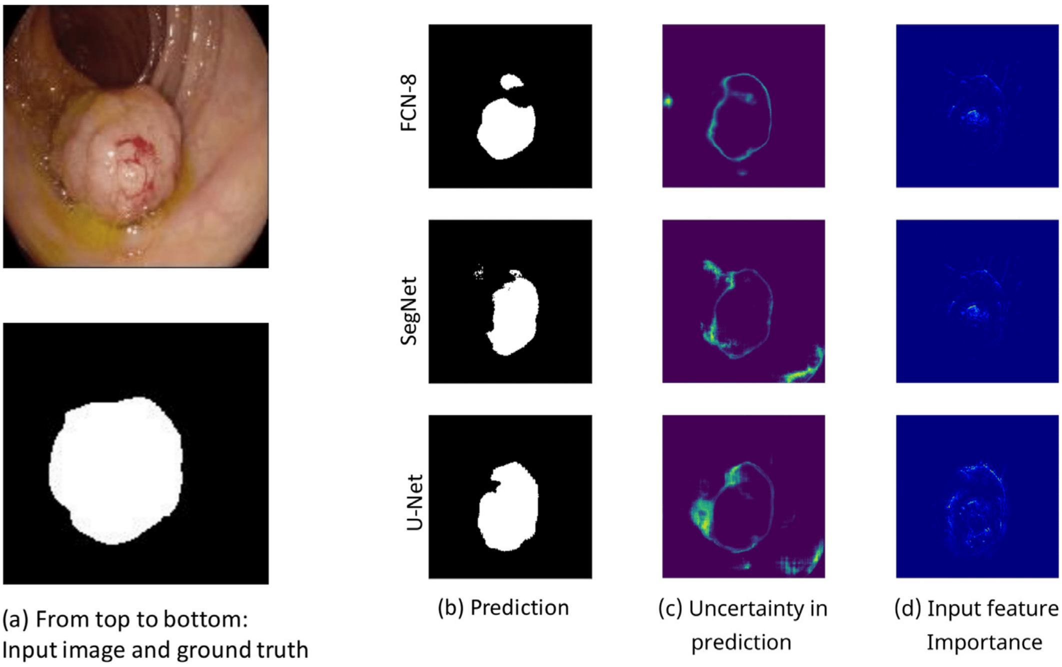
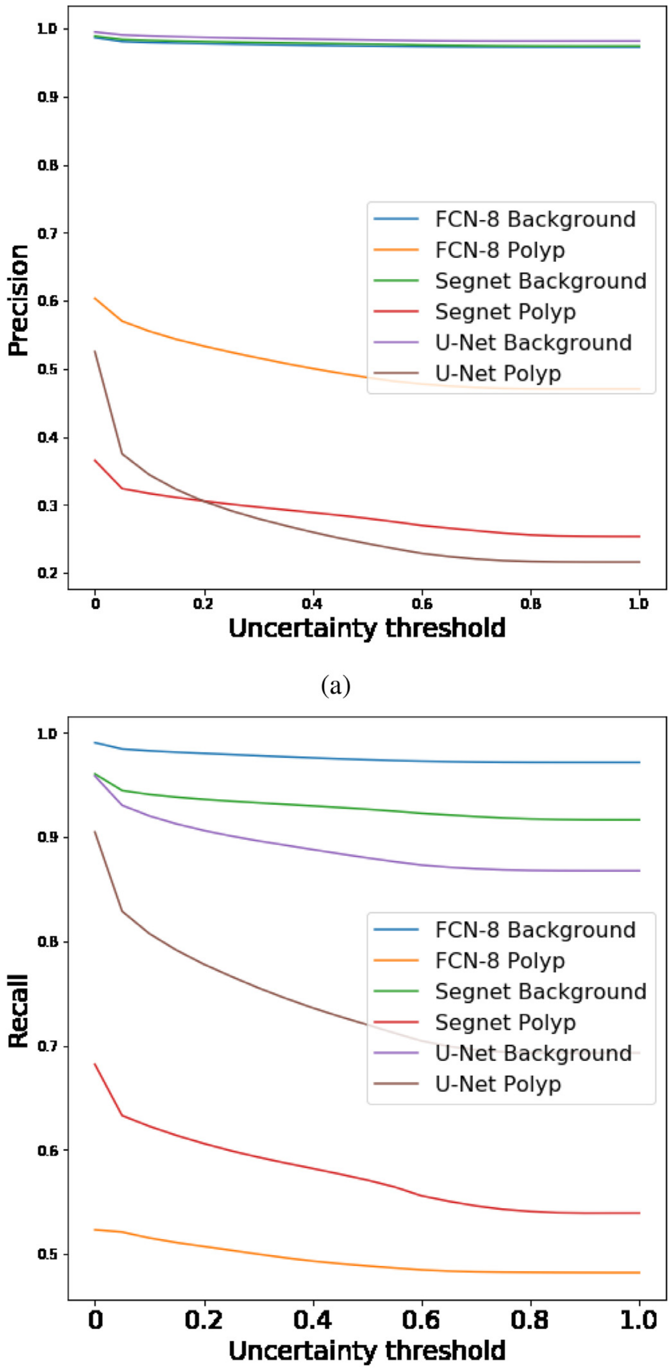

FCN-8
SegNet
U-Net
(a) From top to bottom:
Input image and ground truth
(b) Prediction
(c) Uncertainty in prediction
(d) Input feature Importance

Fig. C.18. Figure displays the prediction, uncertainty map, and interpretability map for the FCN-8, SegNet and U-Net, for the input image from the MICCAI dataset shown in the leftmost column. Best viewed in color.



(a)

(b)

Fig. C.19. Precision and recall vs uncertainty plot for background and polyp class on the MICCAI test set.

References

Alain, G., Bengio, Y., 2017. Understanding intermediate layers using linear classifier probes. ArXiv: 1610.01644.

Alexandre, L.A., Casteleiro, J., Nobreinst, N., 2007. Polyp detection in endoscopic video using svms. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (Eds.), Knowledge Discovery in Databases: PKDD 2007. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 358–365.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS One 10 (7), e0130140.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE TPAMI 2481–2495.

Bernal, J., Núñez, J.M., Sánchez, F.J., Vilariño, F., 2014. Polyp segmentation method in colonoscopy videos by means of Msa-Dova energy maps calculation. In: Workshop on Clinical Image-Based Procedures. Springer, pp. 41–49.

Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F., 2015. Wm-Dova maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. Comput. Med. Imaging Graph. 43, 99–111.

Bernal, J., Tajkbaksh, N., Sánchez, F.J., Matuszewski, B.J., Chen, H., Yu, L., Angermann, Q., Romain, O., Rustad, B., Balasingham, I., Pogorelov, K., Choi, S., Debard, Q., Maier-Hein, L., Speidel, S., Stoyanov, D., Brandao, P., Córdova, H., Sánchez-Montes, C., Gurudu, S.R., Fernández-Esparrach, G., Dray, X., Liang, J., Histace, A., 2017. Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. IEEE Trans. Med. Imaging 36 (6), 1231–1249. doi:10.1109/TMI.2017.2664042.

Chen, W., Zheng, R., Baade, P.D., Zhang, S., Zeng, H., Bray, F., Jemal, A., Yu, X.Q., He, J., 2016. Cancer statistics in china, 2015. CA: A Cancer J. Clinic. 66 (2), 115–132. doi:10.3322/caac.21338.

Condessa, F., Bioucas-Dias, J., 2012. Segmentation and detection of colorectal polyps using local polynomial approximation. In: Campilho, A., Kamel, M. (Eds.), Image Analysis and Recognition. Springer Berlin Heidelberg, pp. 188–197.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A Large-Scale Hierarchical Image Database. In: Proceedings of the CVPR09, pp. 1097–1105.

Dubost, F., Adams, H., Bortsova, G., Ikram, M.A., Niessen, W., Vernooij, M., de Bruijne, M., 2019. 3D regression neural network for the quantification of enlarged perivascular spaces in brain mri. Med. Image Anal. 51, 89–100. doi:10.1016/j.media.2018.10.008.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of the ICML. JMLR.org, pp. 1050–1059.

Guo, S., Wang, K., Kang, H., Zhang, Y., Wang, K., Li, T., 2019. Bts-dsn: deeply supervised neural network with short connections for retinal vessel segmentation. Int. J. Med. Inf. doi:10.1016/j.ijmedinf.2019.03.015.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., Larochelle, H., 2017. Brain tumor segmentation with deep neural networks. Med. Image Anal. 35, 18–31. doi:10.1016/j.media.2016.05.004.

Hwang, S., Oh, J., Tavanapong, W., Wong, J., de Groen, P.C., 2007. Polyp detection in colonoscopy video using elliptical shape feature. In: Proceedings of the IEEE International Conference on Image Processing, 2. II–465–II–468. doi: 10.1109/ICIP.2007.4379193.

Häfner, M., Tamaki, T., Tanaka, S., Uhl, A., Wimmer, G., Yoshida, S., 2015. Local fractal dimension based approaches for colonic polyp classification. Med. Image Anal. 26 (1), 92–107. doi:10.1016/j.media.2015.08.007.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the ICML, pp. 448–456.

Kendall, A., Badrinarayanan, V., Cipolla, R., 2015. Bayesian segnet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv: 1511.02680.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.) Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 5574–5584.

Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv: 1412.6980.

Larsen, I., 2016. Cancer in norway 2015 - cancer incidence, mortality, survival and prevalence in norway. oslo: Cancer registry of norway; 2016.

Liu, Q., 2017. Deep learning applied to automatic polyp detection in colonoscopy images : master thesis in system engineering with embedded systems.

Nida, N., Irtaza, A., Javed, A., Yousaf, M.H., Mahmood, M.T., 2019. Melanoma lesion detection and segmentation using deep region based convolutional neural network and fuzzy c-means clustering. Int. J. Med. Inf. 124, 37–48. doi:10.1016/j.ijmedinf.2019.01.005.

Brandao, P., Mazomenos, P., Ciuti, G., Caliò, R., Bianchi, F., Menciassi, A., Dario, P., Koulaouzidis, A., Arezzo, A., Stoyanov, D., 2017. Fully convolutional neural networks for polyp segmentation in colonoscopy. Proc. SPIE 10134. 10134–10134–7. doi: 10.1117/12.2254361.

Ribeiro, E., Uhl, A., Häfner, M., 2016. Colonic polyp classification with convolutional neural networks. In: Proceedings of the IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS), pp. 253–258. doi:10.1109/CBMS.2016.39.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), MICCAI. Springer International Publishing, Cham, pp. 234–241.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1988. Neurocomputing: foundations of research. Nature 696–699.

Sharma, N., Ray, A., Shukla, K., Sharma, S., Pradhan, S., Srivastva, A., Aggarwal, L., 2010. Automated medical image segmentation techniques. J. Med. Phys. 35 (1), 3.

Shelhamer, E., Long, J., Darrell, T., 2017. Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (4), 640–651.

Shwartz-Ziv, R., Tishby, N., 2017. Opening the black box of deep neural networks via information. arXiv: 1703.00810.

Siegel, R.L., Miller, K.D., Jemal, A., 2017. Cancer statistics, 2017. CA: A Cancer J. Clinic. 67 (1), 7–30. doi:10.3322/caac.21387.

Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv: 1312.6034.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. ICLR.

Springenberg, J., Dosovitskiy, A., Brox, T., Riedmiller, M., 2015. Striving for simplicity: The all convolutional net. In: Proceedings of the ICLR (Workshop track), p. N/A.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.

Tajbakhsh, N., Gurudu, S.R., Liang, J., 2016. Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans. Med. Imaging 35 (2), 630–644. doi:10.1109/TMI.2015.2487997.

Urban, G., Tripathi, P., Alkayali, T., Mittal, M., Jalali, F., Karnes, W., Baldi, P., 2018. Deep learning localizes and identifies polyps in real time with 96 percent accuracy in screening colonoscopy. Gastroenterology doi:10.1053/j.gastro.2018.06.037.

Van Rijn, J.C., Reitsma, J.B., Stoker, J., Bossuyt, P.M., Van Deventer, S.J., Dekker, E., 2006. Polyp miss rate determined by tandem colonoscopy: a systematic review. Am. J. Gastroenterol. 101 (2), 343.

Vázquez, D., Bernal, J., Javier Sánchez, F., Fernández-Esparrach, G., López, A., Romero, A., Drozdzal, M., Courville, A., 2016. A benchmark for endoluminal scene segmentation of colonoscopy images. J. Healthcare Eng. 2017.

Werbos, P., 1974. Beyond regression: New tools for predicting and analysis in the behavioral sciences. Ph.D. thesis. Harvard University.

Wickstrøm, K., Kampffmeyer, M., Jenssen, R., 2018. Uncertainty modeling and interpretability in convolutional neural networks for polyp segmentation. In: Proceedings of the IEEE (MLSP), pp. 1–6. doi:10.1109/MLSP.2018.8516998.

Wimmer, G., Tamaki, T., Tischendorf, J., Häfner, M., Yoshida, S., Tanaka, S., Uhl, A., 2016. Directional wavelet based features for colonic polyp classification. Med. Image Anal. 31, 16–36. doi:10.1016/j.media.2016.02.001.

Yu, S., Príncipe, J.C., 2018. Understanding autoencoders with information theoretic concepts. arXiv: 1804.00057.

Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K., 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLOS Med. 15 (11), 1–17. doi:10.1371/journal.pmed.1002683.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), Proceedings o the ECCV. Springer International Publishing, Cham, pp. 818–833.

/13

Paper II

Uncertainty-aware deep ensembles for reliable and explainable predictions of clinical time series

Kristoffer Wickstrøm, Karl Øyvind Mikalsen, Michael Kampffmeyer, Arthur Revhaug, Robert Jenssen

IEEE Journal of Biomedical and Health Informatics

/14

Paper III

RELAX: Representation Learning Explainability

Kristoffer Wickstrøm, Daniel Trosten, Sigurd Løkse, Ahcène Boubekki, Karl Øyvind Mikalsen, Michael Kampffmeyer, Robert Jenssen

Submitted to International Journal of Computer Vision

RELAX: Representation Learning Explainability

Kristoffer K. Wickstrøm[1*], Daniel J. Trosten[1], Sigurd Løkse[1], Ahcène Boubekki[1], Karl Øyvind Mikalsen[1], Michael C. Kampffmeyer[1] and Robert Jenssen[1]

[1*]Department of Physics and Technology, UiT The Arctic University of Norway, Hansine Hansens veg 18, Tromsø, 9019, Troms, Norway.

*Corresponding author(s). E-mail(s): kristoffer.k.wickstrom@uit.no;
Contributing authors: daniel.j.trosten@uit.no; sigurd.lokse@uit.no; ahcene.boubekki@uit.no; karl.o.mikalsen@uit.no; michael.c.kampffmeyer@uit.no; robert.jenssen@uit.no;

Abstract

Despite the significant improvements that self-supervised representation learning has led to when learning from unlabeled data, no methods have been developed that explain what influences the learned representation. We address this need through our proposed approach, RELAX, which is the first approach for attribution-based explanations of representations. Our approach can also model the uncertainty in its explanations, which is essential to produce trustworthy explanations. RELAX explains representations by measuring similarities in the representation space between an input and masked out versions of itself, providing intuitive explanations and significantly outperforming the gradient-based baselines. We provide theoretical interpretations of RELAX and conduct a novel analysis of feature extractors trained using supervised and unsupervised learning, providing insights into different learning strategies. Moreover, we conduct a user study to assess how well the proposed approach aligns with human intuition and show that the proposed method outperforms the baselines in both the quantitative and human evaluation studies. Finally, we illustrate the usability of RELAX in several use cases and highlight that incorporating uncertainty can be essential for providing faithful explanations, taking a crucial step towards explaining representations.

Keywords: representation learning, explainability, uncertainty, self-supervised learning

1 Introduction

Interpretability is of vital importance for designing trustworthy and transparent deep learning-based systems (Pedreschi et al, 2019; Tonekaboni et al, 2019), and the field of explainable artificial intelligence (XAI) has made great improvements over the last couple of years (Antoran et al, 2021; Schulz et al, 2020). However, there exists no methods for attribution-based explanations of *representations*, despite the tremendous advances in representation learning using e.g self-supervised learning (Chen et al, 2020; Caron et al, 2020; He et al, 2020). This lack of explainability makes representation learning less trustworthy and dependable, and there is therefore a need for representation learning explainability. To be able to explain learned representations would provide crucial information in several use-cases. For instance, a typical clustering approach is applying K-means to the representation produced by a feature extractor trained on unlabeled data (Lin

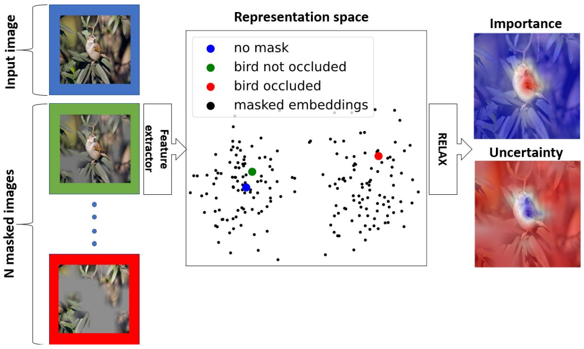


Fig. 1 Conceptual illustration of RELAX. An image is passed through an encoder that produces a new vector representation of the image. Similarly, masked images are embedded in the same latent space. Input feature importance is estimated by measuring the similarity between the representation of the unmasked input with the representations of numerous masked inputs.

et al, 2021; Wen et al, 2020; Yang et al, 2017), but there is no method for investigating which features are characteristic for the members of a cluster.

Representation learning explainability would also allow for a new approach for evaluating representation learning frameworks. Representation learning frameworks are typically evaluated by training simple classifiers on the representation produced by the feature extractor or through a downstream task (Chen et al, 2020; He et al, 2020; Caron et al, 2020). However, such approaches provide only limited information about the features used by the models, and might ignore important distinctions between them. For instance, a similar accuracy on some downstream task does not necessarily equate to the representations being based on the same features. This highlights the need for an explanatory framework for representations, as many of the current evaluation methods are not sufficient for illuminating differences in the what features are used by different feature extractors.

However, any explanatory framework can make over or under-confident explanations. Hence, uncertainty is a key component for designing trustworthy models, since trusting an explanation without knowing the uncertainty of the explanation might lead to an unjustified trust in the model. A recent survey where clinicians were asked what was necessary for making trustworthy models, found that explainability alone was not enough and that uncertainty was also of high importance (Tonekaboni et al, 2019). Our experiments show that uncertainty can be used to increase the faithfulness of explanations, by removing uncertain parts of an explanations. Nevertheless, little work has been done on uncertainty in explanations of representations.

In this work, we present the first framework for explaining representations, entitled REpresentation LeArning eXplainability (RELAX), which is also equipped with uncertainty quantification with respect to its own explanations. The framework is illustrated in Figure 1. RELAX measures the change in the representation of an image when compared with masked versions of itself. The core idea is that when informative parts of the input are masked out, the representation should change significantly. When averaging over numerous masks, RELAX reveals the important regions of the input. RELAX is an intuitive and highly versatile framework that can explain any representation, given a suitable similarity function and masking strategy. To provide insight into the geometrical properties of RELAX, we show that the importance of a pixel can be seen as the result of a scoring function based on an inner product between the input and the mean of the masked representations in the representation space. Figure 2 shows an example where RELAX is used to investigate the explanations and the corresponding uncertainties for a selection of widely used feature extraction models, which demonstrate that RELAX is a versatile framework for highlighting the emphasis that feature extractors put on pixels and regions in the input (top row).

Our contributions are:

- RELAX, a novel framework for explaining representations that also quantifies its uncertainty.
- A threshold approach called U-RELAX that removes uncertain parts of an explanation and increases the faithfulness of the explanations.
- A theoretical analysis of the framework and an estimation of the number of masks needed to obtain a given level of confidence.
- A comprehensive experimental section that compares widely used supervised and self-supervised feature extraction models and evaluates a number of hyperparameters.
- A user study that examines how well the explanations align with human evaluation.

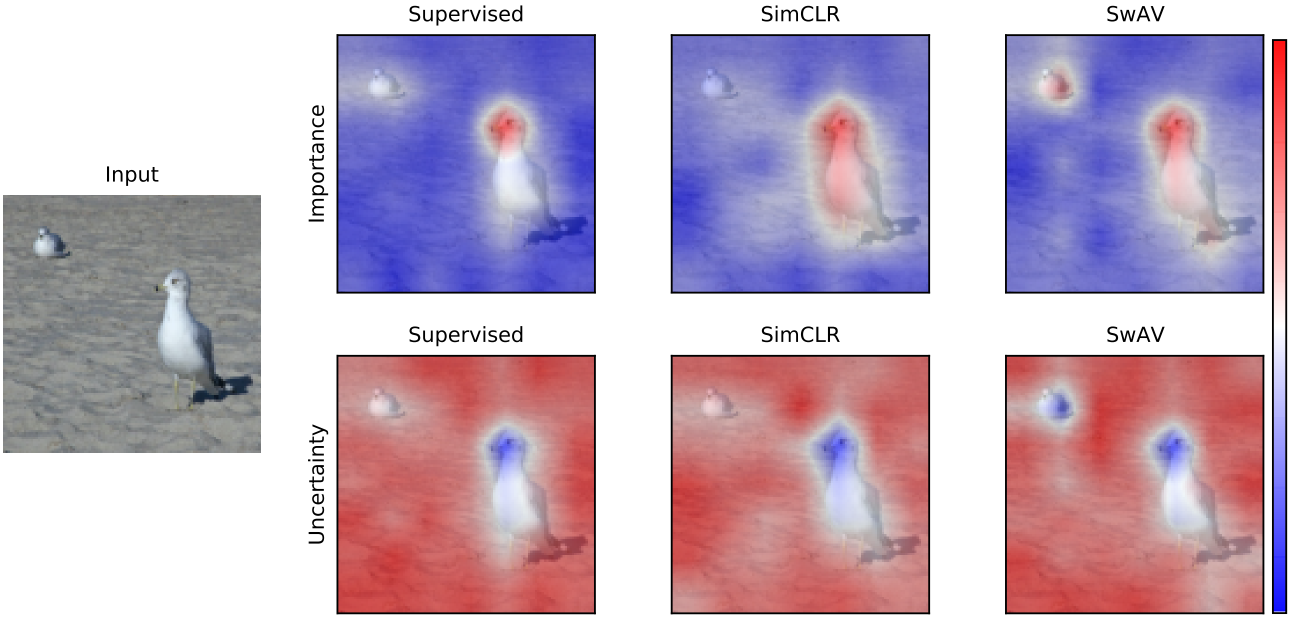


Fig. 2 The figure shows the RELAX explanation and its uncertainty for the representation of the leftmost image for three widely used feature extractors. The first row displays the explanations for the representation and the second row shows the uncertainty associated with the different explanations. Red indicates high values and blue indicates low values. In this example, two objects are present in the image, one bird prominently displayed in the foreground, and another more inconspicuous bird in the background. The plots show that all models emphasize the bird in the foreground with low uncertainty. On the other hand, there is more disagreement on how much emphasis to put on bird in the background, also with a differing degree of uncertainty. The example illustrates that different feature extractors utilize different features in the representation of the image, and with different amounts of uncertainty. The image is taken from VOC (Everingham et al, 2009).

- Two use cases for RELAX. First, RELAX enables explainability in state-of-the-art incomplete multi-view clustering. This illustrates the usability of RELAX in recent cutting-edge research. Second, RELAX allows for explanation of classic computer vision techniques such as Histogram of Oriented Gradients (HOG). This demonstrates that RELAX is a flexible framework, which is capable of explaining representations produced by any method, not just those produced by deep neural networks.

Code for RELAX is available at https://github.com/Wickstrom/RELAX.

2 Related Work

In this section, we present the previous works that are most closely related to our work. The focus will be on attribution-based explanations where each input feature is assigned an importance. Therefore, we will not consider other explainability methods such as example-based explanations (Koh and Liang, 2017; Karimi et al, 2020) or global explanations (Mordvintsev et al, 2015).

Occlusion-based explainability. There exist a number of occlusion-based explainability methods. Systematically occluding an image with a gray rectangle and then measuring the change in activations could be used to provide coarse explanations for CNNs (Zeiler and Fergus, 2014). A more sophisticated occlusion approach can improve explanations, in which smooth masks are generated and accumulated to produce explanations for the prediction of a model (Petsiuk et al, 2018). A slightly different approach is meaningful perturbations, where a spatial perturbation mask that maximally affects the model's output is optimized (Fong and Vedaldi, 2017). A follow up work proposed extremal perturbations, where a perturbation can be considered extremal if it has maximal effect on the network's output among all perturbation of a given, fixed area (Fong et al, 2019). On a different note, an information theoretic approach to XAI has been proposed, where noise is injected in order to measure the information in different regions of the input (Schulz et al, 2020). Similarly, Kolek et al (2021) introduced a rate-distortion perspective to explainability. Note that none of these methods are capable of providing explanations for representations.

Explaining representations. Attribution-based explainability methods are extensively used to explain specific sample predictions (Bach et al, 2015; Petsiuk et al, 2018; Schulz et al, 2020). However, to the best of our knowledge, no attribution-based explainability method exists for explaining representations. While initial attempts have been made to explain representations such as the Concept Activation Vectors (Kim et al, 2018), which uses directional derivatives to quantify the model prediction's sensitivity, these explanations only relate the representations to high-level concepts and require label information. Similarly, network dissection has been proposed to interpret representations (Bau et al, 2017), but requires predefined concepts and label information without indicating the importance of individual pixels. A different direction is designing models that have the capability to explain their own decisions built into the system (Chen et al, 2019; Alvarez-Melis and Jaakkola, 2018). Two drawbacks of such an approach is that it might lead to models with weaker performance and does not explain representations. Another approach maps semantic concepts to vectorial embedding (Fong and Vedaldi, 2018). However, this requires segmentation masks that are not available in the unsupervised setting. Representations have also been investigated from learnability and describability perspectives (Laina et al, 2020), but this was achieved through human-annotators that are typically not available. Lastly, the inspectability of deep representations have been investigated through an information bottleneck approach (Losch et al, 2021), but with a focus on segmentation and predefined concepts.

Uncertainty in explainability. Modeling uncertainty in explainability is a rapidly evolving research topic that is receiving an increasing amount of attention. One of the earliest works proposed to use Monte Carlo Dropout (Gal and Ghahramani, 2016) in order to estimate the uncertainty in gradient-based explanations (Wickstrøm et al, 2018, 2020), which was later followed by a similar approach that was based on Layer-wise Relevance Propagation (Bykov et al, 2020). Uncertainties that are inherent in the widely used LIME method (Ribeiro et al, 2016) have been explored (Zhang et al, 2019). Also, ensemble-based approaches, where uncertainty estimates are obtained by taking the standard deviation across the ensemble, have also been proposed (Wickstrøm et al, 2021). Recently, Counterfactual Latent Uncertainty Explanations (CLUE) was presented (Antoran et al, 2021), where uncertainty estimates from probabilistic models can be interpreted. Nevertheless, none of these approaches were designed for quantifying the uncertainty in explanations of representations, as they either require label information or are computationally impractical.

3 Representation Learning Explainability

We present RELAX, our proposed method for explaining representations, equipped with uncertainty quantification. Furthermore, we leverage RELAX's ability to quantify uncertainty and introduce as a new concept a method for filtering out uncertain parts of the explanations, which we entitle U-RELAX. This is important, as uncertain explanations might give an unwarranted trust in the model. Our framework is inspired by RISE (Petsiuk et al, 2018). However, RISE was designed for explaining predictions and is not transferable for explaining representations or quantifying uncertainty. Note that the proofs of the theorems in this section are given in Appendix C.

3.1 RELAX

The central idea of RELAX is that when informative parts are masked out, the representation should change significantly. Let $\mathbf{X} \in \mathbb{R}^{H \times W}$ represent an image[1] consisting of $H \times W$ pixels, and f denote a feature extractor that transforms an image into a representation $\mathbf{h} = f(\mathbf{X}) \in \mathbb{R}^D$. To mask out regions of the input, we apply a stochastic mask $\mathbf{M} \in [0, 1]^{H \times W}$, where each element M_{ij} is drawn from some distribution.

[1] To enhance readability, we do not include image channels, but this can be easily included by letting the masks span the channel dimension.

The stochastic variable $\bar{\mathbf{h}} = f(\mathbf{X} \odot \mathbf{M})$, where $\odot$ denotes element-wise multiplication, is a representation of a masked version of $\mathbf{X}$. Moreover, we let $s(\mathbf{h}, \bar{\mathbf{h}})$ represent a similarity measure between the unmasked and the masked representation. Intuitively, $\mathbf{h}$ and $\bar{\mathbf{h}}$ should be similar if $\mathbf{M}$ masks *non-informative* parts of $\mathbf{X}$. Conversely, if *informative* parts are masked out, the similarity between the two representations should be low.

Motivated by this intuition, we define the importance R_{ij} of pixel (i, j) as:

$$R_{ij} = \mathrm{E}_{\mathbf{M}}\big[s(\mathbf{h}, \bar{\mathbf{h}})M_{ij}\big]. \tag{1}$$

Equation (1) is core to our framework as it computes the importance of a pixel (i, j) as a weighted similarity score for masked versions of a given image. However, integrating over the entire support of $\mathbf{M}$ is not computationally feasible. Therefore, we approximate the expectation in Equation (1) by sampling N masks and computing the sample mean:

$$\bar{R}_{ij} = \frac{1}{N}\sum_{n=1}^{N} s(\mathbf{h}, \bar{\mathbf{h}}_n)M_{ij}(n). \tag{2}$$

Here, $\bar{\mathbf{h}}_n$ is the representation of the image masked with mask n, and $M_{ij}(n)$ the value of element (i, j) for mask n. The explanations of RELAX are computed through Equation (2), and an illustration of RELAX is given in Figure 1. As a similarity measure we use the cosine similarity

$$s(\mathbf{h}, \bar{\mathbf{h}}) = \frac{\langle \mathbf{h}, \bar{\mathbf{h}} \rangle}{\|\mathbf{h}\| \|\bar{\mathbf{h}}\|}, \tag{3}$$

where $\|\cdot\|$ denotes the Euclidean norm of a vector. There are several motivations for this choice. First, Liu et al (2021) argued that angular information preserves the essential semantics in neural networks, in contrast to magnitude information. Since the cosine kernel normalizes the representation, essentially discarding magnitude information, such a similarity measure would be suited to capture key information encoded in the representations. Second, the cosine kernel does not rely on hyperparameters that must be selected, which may be beneficial in an unsupervised setting where we cannot do cross validation. Third, a large portion of feature extractors trained using self-supervised learning use the cosine kernel in their loss function (Chen et al, 2020; Chen and He, 2021). Therefore, it is the natural choice for measuring similarities in their latent space. However, based on the two first points, the cosine kernel is still suitable for models trained without the cosine kernel. Lastly, other alternatives for the kernel functions, such as the radial basis function or polynomial kernel, requires careful tuning of hyperparameters. We consider an investigation of such alternatives and their hyperparameters as a direction for future research.

Note that we recognize that the masking strategy can introduce a shift in the distribution of pixel intensities. However, in our experiments, we observed that this potential shift did not impact the explanations. An experiment where the distribution is approximately preserved is included in Appendix A.

Masking distribution. There are several ways to sample the masks in Equation (2), for instance by letting each $M_{ij}(n)$ be iid. Bernoulli. However, sampling masks with the same size as the input results in a massive sample space, and simultaneously makes it challenging to create smooth masks that cover different portions of the image [2].

To avoid these problems, we generate masks as suggested by Petsiuk et al (2018). Binary masks of smaller size than the input image are generated, where each element of these smaller masks is sampled from a Bernoulli distribution with probability p. These masks are then upsampled using bilinear interpolation to the same size as the image. The distribution for M_{ij} is then a continuous distribution between 0 and 1. Specifically: we sample N binary masks, each with size $h \times w$, where $h < H$ and $w < W$. We upsample these masks to size $(h+1)C_H \times (w+1)C_W$, where $C_H \times C_W = \lfloor H/h \rfloor \times \lfloor W/w \rfloor$ is the size of the cell in the upsampled masks. Lastly, we crop the final masks of size $H \times W$ randomly from the $(h+1)C_H \times (w+1)C_W$ masks.

Number of masks required. In order to minimize the computational cost of RELAX, we derive the following lower bound on the number of masks required for a certain estimation error.

Theorem 1. *Suppose $s(\cdot, \cdot)$ is bounded in $(0, 1)$.*[3] *Then, for any $\delta \in (0, 1)$ and $t > 0$, if N in*

[2] See Appendix B for evaluation of masking strategies.

[3] This holds for the cosine similarity, since the representations considered are assumed to be ReLU outputs (non-negative).

Equation (2)*, satisfies:*

$$N \geq -\frac{\ln(\delta/2)}{2t^2}, \tag{4}$$

we have $\mathrm{P}(|\bar{R}_{ij} - R_{ij}| \geq t) \leq \delta$*.*

Theorem 1 states that if N satisfies Equation (4), we are able to estimate R_{ij} to an absolute error of less than t with probability at least $1 - \delta$. See Appendix C for proof and verification of bound. In all of our experiments, we generate 3000 masks, which ensures an estimation error below 0.01 with a probability of 0.99.

RELAX from a kernel perspective. To provide insights into the geometrical properties of RELAX, we present a kernel viewpoint of Equation (2).

Theorem 2. *Suppose the similarity function* $s(\cdot,\cdot)$ *is a valid Mercer kernel (Mercer, 1909). The importance* $\bar{R}_{ij}$ *then acts as a linear scoring function between* $\mathbf{h}$*, and the weighted mean of* $\bar{\mathbf{h}}_1, \ldots, \bar{\mathbf{h}}_N$*, in the reproducing kernel Hilbert space (RKHS) induced by* $s(\cdot,\cdot)$*. That is:*

$$\bar{R}_{ij} = \langle \phi(\mathbf{h}), \frac{1}{N}\sum_{n=1}^{N} \phi(\bar{\mathbf{h}}_n) M_{ij}(n) \rangle_{\mathcal{H}}, \tag{5}$$

where $\phi : \mathbb{R}^d \to \mathcal{H}$ *is the mapping to the RKHS,* $\mathcal{H}$*, induced by the kernel* $s(\cdot,\cdot)$*, and* $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ *is the inner product on* $\mathcal{H}$*.*

Theorem 2 provides interesting insight, as many scoring functions are based on inner-products, e.g. between points of interest and class-conditional means (e.g., Fisher discriminant analysis, Bayes classifier under Gaussian distributions with equal covariance structure). This means that even though RELAX is a novel approach, it is founded in well-known statistical concepts (McCullagh and Nelder, 1989).

Additionally, RELAX has the following interpretation from non-parametric statistics

Theorem 3. *Suppose* $s(\cdot,\cdot)$ *is a valid Parzen window (Theodoridis and Koutroumbas, 2009). Then:*

$$\bar{R}_{ij} \propto p_{ij}(\mathbf{h}), \tag{6}$$

where $p_{ij}(\cdot)$ *is a weighted Parzen density estimate (Parzen, 1962) of the density of the masked embeddings:*

$$p_{ij}(\cdot) = \frac{1}{\sum_{n'=1}^{N} M_{ij}(n')} \sum_{n=1}^{N} s(\cdot, \bar{\mathbf{h}}_n) M_{ij}(n). \tag{7}$$

A high RELAX score is obtained when the unmasked representation $\mathbf{h}$ is close to mean of masked representations, which aligns well with out intuition for RELAX.

3.2 Uncertainty in Explanations

Trusting an explanation without a notion of uncertainty can lead to an unjustified faith in the model. Therefore, we introduce an approach that allows uncertainty quantification to be incorporated into the RELAX framework. Our intuition for this approach stems from what happens when informative and uninformative parts are masked out. If informative parts are masked out, the similarity score will not only drop, but drop with varying degree. If there is a big variation in the similarity scores for a given pixel, it indicates that the explanation for said pixel is uncertain. Based on this intuition, we propose to estimate the uncertainty in input feature importance as:

$$U_{ij} = \mathrm{Var}_{\mathbf{M}}[s(\mathbf{h}, \bar{\mathbf{h}})M_{ij}]. \tag{8}$$

Again, it is not feasible to integrate over all of $\mathbf{M}$ and U_{ij} is therefore approximated by the sample variance:

$$\bar{U}_{ij} = \frac{1}{N}\sum_{n=1}^{N} (s(\mathbf{h}, \bar{\mathbf{h}}_n) - \bar{R}_{ij})^2 M_{ij}(n). \tag{9}$$

Equation (9) estimates the uncertainty of the RELAX-score for pixel (i,j) by measuring the difference between the similarity score and the explanations. To estimate Equation (9), we must first estimate the importance of a pixel. The uncertainty estimates provided in Equation (9) can be thought of as measuring the spread of pixel importance values in relation to importance estimated using Equation (2). There are several benefits of our method. First, it requires no labels, which is sometimes used in other uncertainty estimation methods (Antoran et al, 2021). Secondly, it avoids computationally intense sampling methods, for instance through Monte Carlo sampling (Teye et al, 2018; Gal and Ghahramani, 2016). Lastly, the uncertainty estimation can be combined with the computation of Equation (2), as explained in Section 3.4.

3.3 U-RELAX: Uncertainty Filtered Explanations

All parts of an explanation do not have the same level of uncertainty associated with it. In such cases, it could be beneficial to remove input features that are indicated as important but also have high uncertainty, while only keeping important input features with low uncertainty. This could increase the faithfulness of an explanation and provide clearer explanations. Therefore, we propose a thresholding approach where explanations with high uncertainty are removed from the explanation. We define our U-RELAX importance score as:

$$\bar{R}'_{ij} = \begin{cases} \bar{R}_{ij}, & \text{if } \bar{U}_{ij} < \epsilon \\ 0, & \text{otherwise} \end{cases}, \tag{10}$$

where ϵ is a threshold chosen by the user. Essentially, Equation (10) provides the possibility to only consider explanations of a particular certainty level, depending on ϵ. We propose two ways of choosing epsilon. First as:

$$\epsilon = \frac{\gamma}{HW} \sum_{i}^{H} \sum_{j}^{W} \bar{U}_{ij}, \tag{11}$$

that is, the average uncertainty for a particular image, weighted by hyperparameter γ. This provides a simple and intuitive way of selecting the threshold, which is motivated by only wanting to consider pixels that have high importance and low uncertainty. Alternatively, ϵ can be computed by taking the median uncertainty for a particular image.

We refer to this uncertainty-filtered version of RELAX as U-RELAX. Figure 3 shows an example of the U-RELAX explanation compared with the RELAX explanation. In this case, the emphasis on the bird in the background is removed as the uncertainty was too high for this part of the explanation.

3.4 One-Pass Version of RELAX

Computing Equation (9) requires first computing Equation (2), since the uncertainty estimation requires an estimate of the importance in order to be computed. This introduces additional computational overhead. We refer to computing Equation (2) followed by Equation (9) as the *two-pass* version of RELAX. To improve computational efficiency, we propose an online version of RELAX where importance and uncertainty is computed simultaneously, which we refer to as the *one-pass* version of RELAX. One-pass RELAX is based on well-known estimators of running mean and variance (West, 1979). Importance is computed as:

$$\begin{aligned} \bar{R}_{ij}^{(n)} =& \bar{R}_{ij}^{(n-1)} + \\ & M_{ij}(n) \frac{s(\mathbf{h}, \bar{\mathbf{h}}_n)(n) - \bar{R}_{ij}^{(n-1)}}{W_{ij}(n)}, \end{aligned} \tag{12}$$

where $\bar{R}_{ij}^{(n)}$ is the importance of pixel (i, j) at mask n, and $W_{ij}(n) = \sum_{n'=0}^{n} M_{ij}(n')$ is the sum of the mask elements (i, j) for the first n masks. Uncertainty is computed as:

$$\begin{aligned} \bar{U}_{ij}^{(n)} =& \bar{U}_{ij}^{(n-1)} + M_{ij}(n)(s(\mathbf{h}, \bar{\mathbf{h}}_n) - \\ & \bar{R}_{ij}^{(n)})(s(\mathbf{h}, \bar{\mathbf{h}}_n) - \bar{R}_{ij}^{(n-1)}), \end{aligned} \tag{13}$$

where $\bar{U}_{ij}^{(n)}$ is the uncertainty in the importance of pixel (i, j) after the nth mask. Pseudo-code is shown in Algorithm 1. All experiments are carried out using the one-pass version of RELAX. See Appendix D for a comparison of the one-pass versus two-pass version.

4 Evaluation and Baseline

4.1 Evaluation of Explanations

Evaluation is a developing subfield of XAI, and a unifying score is not agreed upon Doshi-Velez and Kim (2017), even more so for explanations of representations. To evaluate the explanations, we use two of the most widely used explainability evaluation scores, namely localisation and faithfulness (Samek et al, 2017; Petsiuk et al, 2018; Fong et al,

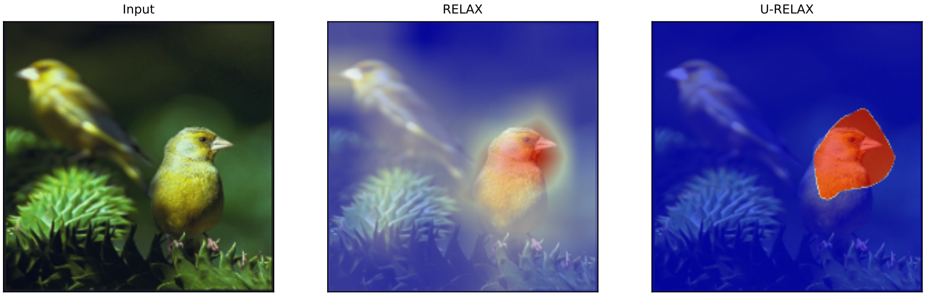


Fig. 3 Comparison of RELAX and U-RELAX on an image taken from PASCAL VOC, where red indicates high importance and blue indicates low importance. In this case, the emphasis on the bird in the background is removed as the uncertainty was to high for this part of the explanation.

Algorithm 1 Pytorch-like pseudocode for RELAX.

```
# f           - feature extractor
# X[1,C,H,W]- input image
# R[H,W]      - importance (init as zeros)
# U[H,W]      - uncertainty (init as zeros)
# W[H,W]      - sum of masks (init with
#               small positive number)<
for mask in mask_generator: # [1,1,H,W]
    W += mask
    h, h_mask = f(x), f(x*mask)
    s = cosine_similarity(h, h_mask)
    R_prev = R
    R += m*(s-R)/W
    U += (s-R)*(s-R_prev)*m
return R, U/(W-1)
```

2019; Schulz et al, 2020). All scores are computed using the Quantus toolbox [4].

Localisation. The explanations should put emphasis on input regions corresponding to the objects present in an image. Localisation measures to which degree the explanation agrees with the ground truth location of an object. High performance in localisation indicates that the explanations often align with the bounding boxes or segmentation masks provided by human annotators. We consider three localisation scores, the *pointing game* (Zhang et al, 2017), *top-k intersection*, and *relevance rank accuracy* (Arras et al, 2022). The pointing game measures whether the pixel with the highest importance is located within the object location. Top-k intersection considers the binarized version of the top-k most important pixels and measures the intersection with the ground truth mask. Relevance rank accuracy is measured by taking the ratio of high intensity relevances within the ground truth mask. Since RELAX operates in the unsupervised setting we do not have explanations for individual classes. Therefore, the bounding boxes/segmentation masks are collected into one unified bounding box/segmentation mask. This results in unsupervised version of localisation that is suitable for explaining representations.

Faithfulness. Pixels assigned with high importance should be indicative of "true" importance. Faithfulness is typically measures by monitoring the classification accuracy of a classifier as input features are iteratively removed. High faithfulness indicates that the explanation is capable of identifying features that are important for classifying an image correctly. We measure faithfulness using the *monotonicity* score. Nguyen and Martinez (2020) proposed to measure monotonicity by computing the correlation of the absolute values of the attributions and the uncertainty in the probability estimation. This will indicate if an explanation is correctly highlighting important features in the input.

4.2 Representation Explainability Baseline

While there are are no existing methods that provide attribution-based explanations for representations, it is possible to adopt certain methods to provide such explanations. One of the most common baselines in the field of explainability is saliency explanations (Springenberg et al, 2015;

[4] https://github.com/understandable-machine-intelligence-lab/Quantus

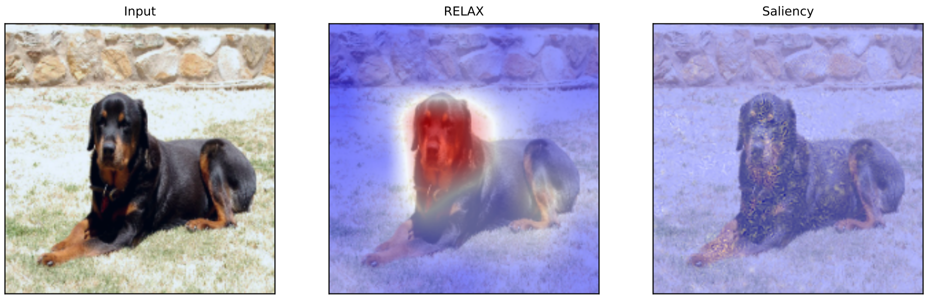


Fig. 4 Comparison of RELAX and saliency explanation for an image from PASCAL VOC. The example shows how both explanations focus on the dog, but the saliency explantion is much more erratic and unfocused than the RELAX explanations.

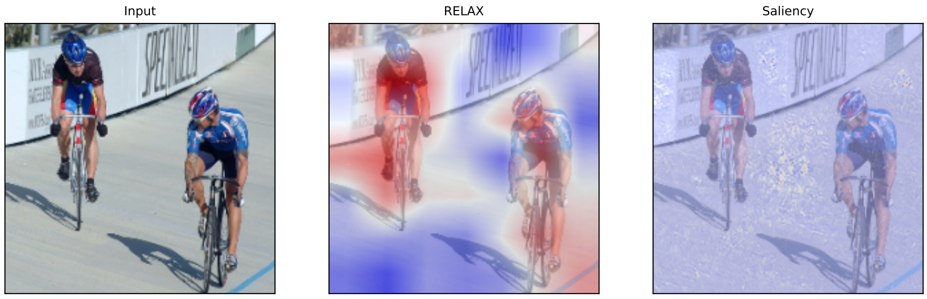


Fig. 5 Comparison of RELAX and Saliency explanation for an image from PASCAL VOC. The example shows how RELAX captures information about both objects, while the saliency explanation is focused on the gap in between the two objects.

Adebayo et al, 2018), which utilize gradient information to attribute importance. An explanation is obtained by computing the gradient for a prediction with respect to the input. However, it is not trivial to extend such methods for explaining representations. We propose the following for a saliency approach:

$$\mathbf{S} = \frac{1}{D}\sum_{d=1}^{D} \nabla f(\mathbf{X})_d, \tag{14}$$

where D is the dimensionality of the representation and S_{ij} is the importance of pixel (i, j) for the given representation. The gradient for each dimension of the representation will give an explanation, and Equation (14) takes the mean across all explanations. This is the most straight-forward and intuitive approach for explaining representations with gradients. It also illustrates the challenges that arise when adopting gradient-based explanations for representation, as some form of agglomeration of the explanations is required. Figure 4 and Figure 5 shows a qualitative comparison between the RELAX and saliency explanation for a representation of an image. Both Figures illustrate how RELAX provides more intuitive and clear explanations that are able to capture information related to the objects in the image, when compared with the saliency explanation.

Once the saliency approach from Equation (14) have been established, it is also possible to adopt improvements of the standard saliency explanations. For instance, Guided Backpropagation is a widely used explainability technique that uses gradient information (Springenberg et al, 2015). Guided Backpropagation differs from Equation (14) by zeroing out negative gradients in the backward pass of the backpropagation scheme. We define the Guided Backpropagation procedure for representations as:

$$\mathbf{S}_{\text{GB}} = \frac{1}{D}\sum_{d=1}^{D} \nabla_{\text{GB}} f(\mathbf{X})_d. \tag{15}$$

Second, SmoothGrad is another gradient-based explainability method that can be adopted from Equation 14 (Smilkov et al, 2017). SmoothGrad injects noise into the input and produces an explanation by averaging over multiple explanations. We define SmoothGrad for representation as:

$$\mathbf{S}_{\text{SG}} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{D}\sum_{d=1}^{D}\nabla f(\mathbf{X}_m)_d, \tag{16}$$

where M is the number of explanations computed based on the noisy input.

5 Experiments

To evaluate RELAX, we conduct numerous experiments and report both quantitative and qualitative results. We evaluate several features extraction models, both deep and non-deep, and trained with and without supervision. Our experiments show the advantageous of RELAX compared to the baselines, and illustrates how RELAX enables new approaches for analysing and understanding representation learning.

Implementation details. For the supervised model, we use the pretrained model from Pytorch (Paszke et al, 2019). For the models trained without labels but with self-supervision, we use the SimCLR (Chen et al, 2020) and SwAV (Caron et al, 2020) frameworks, both of which have seen recent widespread use. These methods are chosen to represent two major types of self-supervised learning frameworks, namely contrastive instance learning (SimCLR) and clustering-based learning (SwAV). For SimCLR and SwAV, we use the pretrained models from Pytorch Lightning Bolts (Falcon and Cho, 2020). We use a ResNet50 (He et al, 2016) as the backbone for the feature extractors, and all models are trained on ImageNet (Deng et al, 2009).

Similarly as in previous works (Fong et al, 2019; Schulz et al, 2020), we use the test split of the PASCAL VOC07 (VOC) (Everingham et al, 2009) and the validation split of MSCOCO2014 (COCO) (Lin et al, 2014) for evaluating the localisation scores, since they contain information about the location of the objects in the images. For the faithfulness score, we use the validation set of ImageNet (Deng et al, 2009). For all datasets, we randomly sample 1000 images for evaluation and repeat all experiments 3 times. Since we are interested in investigating how RELAX and U-RELAX varies due to the stochastic masking process, we use the same 1000 images across the repeated experiments. We generate 3000 masks to ensure a low estimator error. We set $h = w = 7$ and resize all images to $H = W = 224$, as suggested by Zhang et al (2017). For the monotonicity score, we use Alexnet (Krizhevsky et al, 2012) as the classifier, as suggested by Samek et al (2017). We also experiment with the VGG13 (Simonyan and Zisserman, 2015) as the classifier for monotonicity score. These results are reported in Appendix F. The threshold for U-RELAX is determined with median aggregation and $\gamma = 1.0$, based on the empirical evaluation conduced in Section 5.4.

5.1 Qualitative Results

Figure 2 and 6 displays the explanation and the uncertainty in the explanations provided by RELAX for an image from the PASCAL VOC and MS COCO dataset, respectively. See Appendix G for additional qualitative results. The input to the feature extractors is shown on the left, the first row shows the explanations, and the second row shows the uncertainties.

Are all instances of the same object equally important?

Figure 2 shows an example with two objects, one bird prominently displayed in the foreground, and another more inconspicuous bird in the background. An interesting question that RELAX allows us to answer is: are both of these birds important for the representation of this image? And, are both of them equally important? First, all models indicate that the bird in the foreground is important, and that the explanations for this bird have low uncertainty. Second, SimCLR puts little emphasis on the bird in the background. In contrast, both the supervised feature extractor and SwAV are highlighting the second bird as having an influence on the representation. However, the uncertainty estimates for the second bird is slightly higher than those of the first bird, but still low compared to the remaining parts of the image.

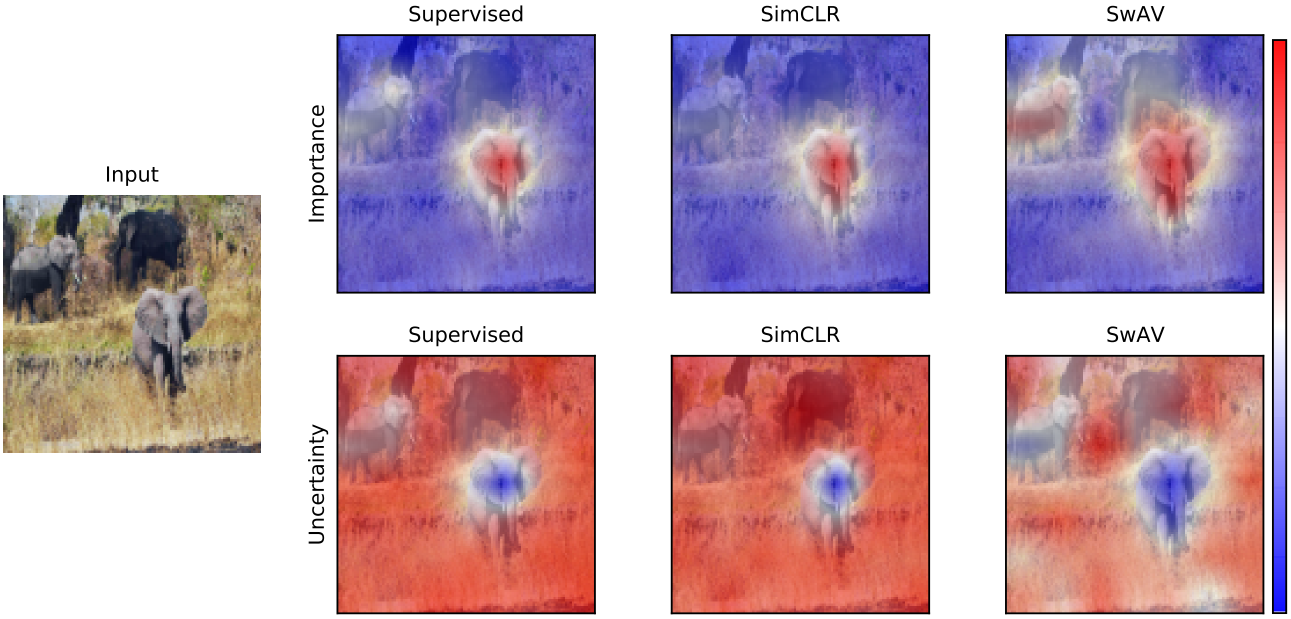


Fig. 6 The figure shows the RELAX explanation and its uncertainty for the representation of the leftmost image for a number of widely used feature extractors. The first row displays the explanations for the representation and the second row shows the uncertainty associated with the different explanations. Red indicates high values and blue indicates low values. In this example, three elephants are visible in the image. The results show that all models highlight the elephant in the foreground as important for the representation, but there is more disagreement about the elephants in the background. Moreover, the uncertainty of the explanation for the elephant in the foreground is very low compared to the remaining regions of the image. Image is taken from MS COCO.

What features are important in complex images with numerous objects?

Figure 6 shows an image with 3 elephants, one in the foreground and two in the background. Additionally, the background is more diverse and the objects have different lighting and perspective. Again, RELAX enables investigation of interesting aspects of the representations, such as: are the models capable of recognizing all elephants and utilizing the information? Does the models focus on background information instead of the objects? All models highlight the elephant in the foreground as important with high certainty. However, there is little emphasis on the shaded elephant, and the associated region of the image also has a high degree of uncertainty. Both the supervised model and SwAV put some importance on the third elephant with some degree of certainty, while SimCLR uses little or no information about the third elephant.

In both Figure 2 and 6, the SwAV feature extractor is focusing on several regions in the input, but with some regions of high uncertainty. While it is difficult to say exactly why, we hypothesize that it can be related to its self-supervised training procedure. SwAV relies on matching image views to a set of prototypes. Therefore, different parts of the input can be related to different prototypes, which we conjecture can lead to SwAV considering several regions of the input.

5.2 Quantitative Results

Table 1 and 2 displays the quantitative evaluation of our proposed methodology compared with the gradient-based baselines described in Section 4.2. The results show how the proposed method outperforms the baselines across all scores. The low standard deviation for RELAX show that the proposed methodology is robust to the stochasticity in the masks. Furthermore, the feature extractor trained using supervised learning achieves the highest performance compared to the feature extractors trained using self-supervised learning, which illustrates that label information does provide additional useful information for these scores.

For the localisation scores, RELAX provides the highest performance. The segmentation masks

Scores	Methods	Supervised		SimCLR		SwAV	
		COCO	VOC	COCO	VOC	COCO	VOC
pointing game	Saliency	67.1±0.0	82.8±0.0	59.9±0.0	75.9±0.0	60.0±0.0	76.3±0.0
	Smooth Saliency	62.8±0.0	79.5±0.0	60.1±0.0	75.9±0.0	59.8±0.0	76.4±0.0
	Guided Saliency	66.6±0.0	82.9±0.0	58.4±0.0	73.3±0.0	59.5±0.0	75.8±0.0
	RELAX	**72.6±0.1**	**86.6±0.2**	**68.7±0.3**	**85.2±0.3**	**67.8±0.2**	**84.7±0.2**
	U-RELAX	72.1±0.3	86.4±0.4	68.6±0.2	85.0±0.5	66.7±0.7	84.1±0.4
top k	Saliency	62.2±0.0	80.1±0.0	56.5±0.0	71.3±0.0	56.5±0.0	71.4±0.0
	Smooth Saliency	59.2±0.0	74.1±0.0	56.4±0.0	71.1±0.0	56.4±0.0	71.3±0.0
	Guided Saliency	62.2±0.0	80.2±0.0	55.1±0.0	69.0±0.0	56.3±0.0	71.1±0.0
	RELAX	**72.8±0.4**	**86.9±0.1**	**69.0±0.3**	**85.6±0.2**	**68.1±0.4**	**85.1±0.2**
	U-RELAX	72.2±0.4	86.5±0.2	68.8±0.4	85.3±0.1	66.6±0.4	84.2±0.3
relevance rank	Saliency	46.8±0.0	59.5±0.0	41.2±0.0	53.6±0.0	40.9±0.0	53.4±0.0
	Smooth Saliency	42.6±0.0	54.6±0.0	41.1±0.0	53.4±0.0	40.9±0.0	53.3±0.0
	Guided Saliency	46.8±0.0	59.8±0.0	40.6±0.0	53.0±0.0	40.9±0.0	53.3±0.0
	RELAX	**56.4±0.0**	**70.2±0.1**	**54.2±0.2**	**69.8±0.1**	**52.4±0.1**	**69.1±0.0**
	U-RELAX	52.4±0.0	64.7±0.1	50.7±0.1	63.3±0.1	46.2±0.1	59.5±0.0

Table 1 Pointing game, top k, and relevance rank scores in percentages and averaged over 3 runs. Higher is better and bold numbers highlight the top performance. Results show that our method improves on the baseline across all scores.

Scores	Methods	Supervised	SimCLR	SwAV
monotonicity	Saliency	12.8±0.2	14.8±0.5	14.6±0.3
	Smooth Saliency	15.4±0.1	14.3±0.3	14.0±0.3
	Guided Saliency	15.3±0.3	15.3±0.2	14.2±0.6
	RELAX	18.3±0.5	20.2±0.4	**21.3±0.4**
	U-RELAX	**23.6±0.4**	**22.9±0.1**	18.3±0.6

Table 2 Monotonicity scores averaged over 3 runs. Higher is better and bold numbers highlight the top performance. Results show that our method improves on the baseline.

or bounding boxes can in many cases be large, and U-RELAX might remove uncertain points close to the boundaries of the segmentation masks. This might be desirable from a human perspective, as it provides clearer explanations with less uncertainty, but it will decrease the localisation scores. For the faithfulness score, U-RELAX provides a significant boost in performance for two encoders. The removal of uncertain explanations allows the classifier to focus on a smaller subset of highly relevant features. This can lead to the classifier having a more stable decrease in accuracy and a higher faithfulness score.

5.3 Human Evaluation

The localisation and faithfulness scores are both proxies for human evaluation that allow for quantitative analysis. However, the ultimate goal of XAI is to provide explanations that are understandable for people and align well with human intuition. Therefore, we conduct a user study with human evaluation of explanations. In this user study, 13 people were asked to select their preferred explanation from a selection of explanations across 10 different images. See Appendix E for a detailed description of the user study.

Table 3 reports the results of the human evaluation. The results clearly indicate that RELAX and U-RELAX were the methods that aligned

	RELAX	U-RELAX	Saliency	Smooth Saliency	Guided Saliency	Random
Counts	79	29	9	4	8	1

Table 3 Human evaluation of representation explainability methods across 10 images from the PASCAL VOC dataset. Results show that the majority of the votes were cast for RELAX and U-RELAX.

most closely with human intuition. Some participants highlighted that when both RELAX and the gradient-based methods indicated an object as important, they often preferred the more object focused explanation of RELAX, as opposed to the more edge focused explanations of the baselines. It was also noted that for some images the participants disagreed with most explanations, and would have provided a different explanation if possible. We believe that these are valuable insights that will be useful for improving explainability methods and also for designing future user studies.

5.4 U-RELAX Hyperparameter Evaluation

Table 4 and 5 reports localisation and faithfulness scores for different values of the hyperparameters in U-RELAX. Mean versus median aggregation is considered, and a selection of values for γ. The results indicate that setting γ to less than 1, typical degrades performance. This can be understood by the thresholding being to strict and removing to many pixel indicated as important. Also, the differences between mean and median aggregation of the uncertainties is mostly low, but median aggregation gives a slight improvement, particularly for the relevance rank score and the monotonicity score.

5.5 Use Case I: Multi-View Clustering

To further illustrate the ability of RELAX to obtain insights into new tasks, we conduct an experiment on multi-view clustering. We learn a feature extractor using the Completer framework (Lin et al, 2021), which uses an information theoretic approach to fuse several views into a new representation. Completer uses individual encoders for each view, and concatenates the representation from each encoder to produce a unified representation. Clustering is performed by applying K-means to the learned representations. To adopt RELAX for such a setting, we generate individual masks for each view and monitor the change in the representation in the unified representation space. While there is no way to investigate which parts of the different views that influence the unified representation in the Completer framework, using RELAX allows us to answer this question. Figure 7 shows an example on Noisy MNIST (Wang et al, 2015), where one view is a digit and the other view is a noisy version of the same digit. The result shows that the Completer framework is exploiting information from both views to produce a new representation, even if one view contains more noise. Such insights would not be obtainable without RELAX.

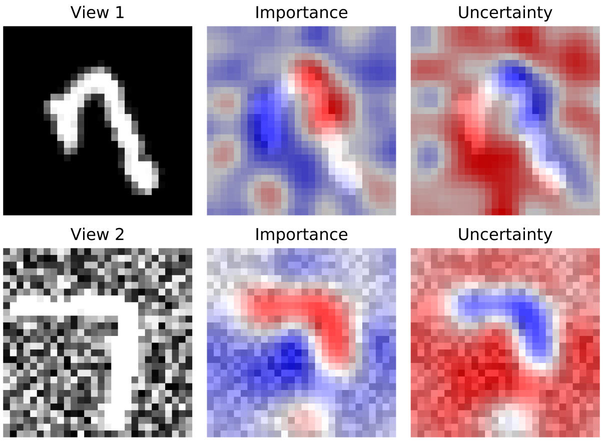


Fig. 7 RELAX explanation and uncertainty for the representation of an example from Noisy MNIST image for a number of widely used feature extractors. The first row displays input, explanation, and uncertainty for view 1, and the second row for view 2. Red indicates high values and blue indicates low values. The Figure shows that Completer is extracting complementary information from the two views for creating its unified representation.

5.6 Use Case II: Explaining HOG Features

RELAX is not limited to representations produced by deep neural networks. It can be used to explain the representation produced by any function that transform an image into a vector representation. To illustrate the versatility of

Scores	(aggregation, γ)	Supervised		SimCLR		SwAV	
		COCO	VOC	COCO	VOC	COCO	VOC
pointing game	(mean, 0.95)	71.1±0.4	86.5±0.2	67.6±0.1	83.9±0.3	63.3±0.7	81.1±0.5
	(mean, 0.99)	71.8±0.4	86.4±0.5	68.6±0.4	**85.0±0.4**	66.4±0.6	**84.2±0.4**
	(mean, 1.0)	71.7±0.1	86.5±0.2	68.6±0.1	**85.0±0.3**	**66.7±0.7**	84.1±0.2
	(median, 0.95)	71.2±0.2	**86.6±0.1**	67.6±0.4	84.2±0.2	63.6±0.2	80.9±0.1
	(median, 0.99)	71.8±0.3	86.5±0.4	**68.8±0.3**	**85.0±0.2**	66.3±0.6	84.0±0.3
	(median, 1.0)	**72.1±0.3**	86.4±0.4	68.6±0.2	**85.0±0.5**	**66.7±0.7**	84.1±0.4
top k	(mean, 0.95)	71.3±0.4	86.2±0.2	67.1±0.1	83.2±0.3	62.8±0.2	79.5±0.4
	(mean, 0.99)	**72.2±0.4**	**86.6±0.2**	**68.8±0.3**	85.2±0.2	66.4±0.2	84.0±0.3
	(mean, 1.0)	**72.2±0.4**	86.5±0.2	**68.8±0.4**	**85.3±0.1**	**66.7±0.4**	**84.3±0.2**
	(median, 0.95)	71.2±0.4	86.1±0.2	67.1±0.2	83.2±0.4	62.7±0.2	79.1±0.4
	(median, 0.99)	**72.2±0.4**	86.5±0.2	68.7±0.3	85.2±0.2	66.4±0.2	83.9±0.3
	(median, 1.0)	**72.2±0.4**	86.5±0.2	**68.8±0.4**	**85.3±0.1**	66.6±0.4	84.2±0.3
relevance rank	(mean, 0.95)	45.9±0.0	55.7±0.0	41.6±0.1	52.3±0.1	39.6±0.1	51.0±0.0
	(mean, 0.99)	50.3±0.0	61.2±0.1	48.6±0.1	59.8±0.1	44.0±0.1	56.0±0.1
	(mean, 1.0)	51.4±0.1	63.0±0.1	50.3±0.1	62.2±0.1	45.6±0.1	58.2±0.1
	(median, 0.95)	46.8±0.0	57.2±0.1	42.4±0.1	53.3±0.1	40.4±0.1	52.1±0.0
	(median, 0.99)	51.2±0.0	63.0±0.1	49.1±0.1	60.8±0.1	44.6±0.1	57.3±0.1
	(median, 1.0)	**52.4±0.0**	**64.7±0.1**	**50.7±0.1**	**63.3±0.1**	**46.2±0.1**	**59.5±0.0**

Table 4 Evaluation of U-RELAX hyperparameters in terms of pointing game, top k, and relevance rank scores in percentages and averaged over 3 runs. Higher is better and bold numbers highlight the top performance

Scores	(aggregation, γ)	Supervised	SimCLR	SwAV
monotonicity	(mean, 0.95)	16.3±0.5	11.8±0.3	12.4±0.3
	(mean, 0.99)	22.2±0.2	20.4±0.5	16.2±0.3
	(mean, 1.0)	23.2±0.1	21.8±0.3	18.0±0.0
	(median, 0.95)	17.9±0.7	12.8±0.2	13.5±0.2
	(median, 0.99)	23.0±0.7	21.1±0.1	17.1±0.4
	(median, 1.0)	**23.6±0.4**	**22.9±0.1**	**18.3±0.6**

Table 5 Evaluation of U-RELAX hyperparameters in terms of monotonicity score in percentages and averaged over 3 runs. Higher is better and bold numbers highlight the top performance

RELAX, we explain representation produced by the Histogram of Oriented Gradients (HOG) feature extraction method (Dalal and Triggs, 2005), which have been used extensively in the computer vision literature. Figure 8 and 9 shows two examples where the explanation for the HOG representation is compared with the SimCLR and SwAV representations. We consider the representations from these two methods since they are also unsupervised like the HOG features.

Features produced by deep neural networks are typically allow for higher performance than those from algorithms such as HOG and other handcrafted feature extraction methods. RELAX provides insights into why this is. In Figure 8, both the SimCLR and the SwAV feature extractors focus on the cat in the center of the images. The HOG algorithm has a more widespread focus. Also, much of the emphasis is put on the cord going along the staircase. Since the HOG algorithm is utilizing gradient information, these sharp lines will have a big influence on the representation, and it is therefore not surprising that the

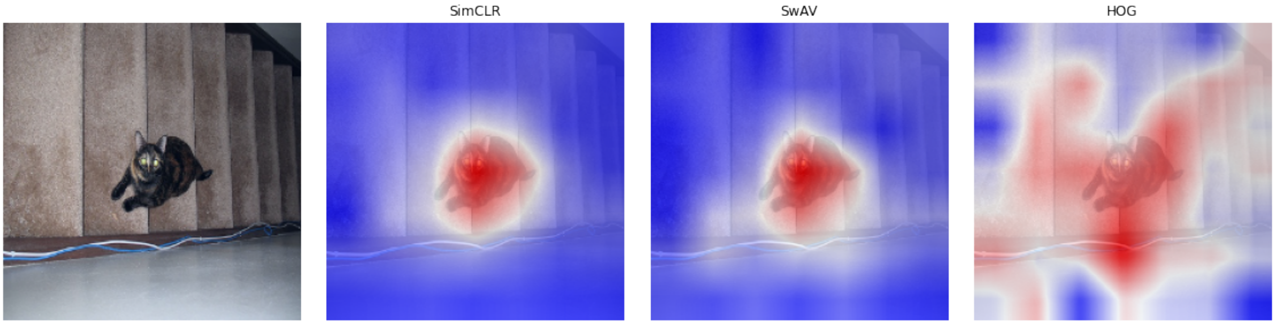


Fig. 8 The figure shows the RELAX explanation for two deep learning-based feature extractors compared with the traditional HOG algorithm. Figure shows how HOG features focus on more indistinct regions in the input, while deep learning methods focus mainly on the cat. Image is taken from PASCAL VOC.

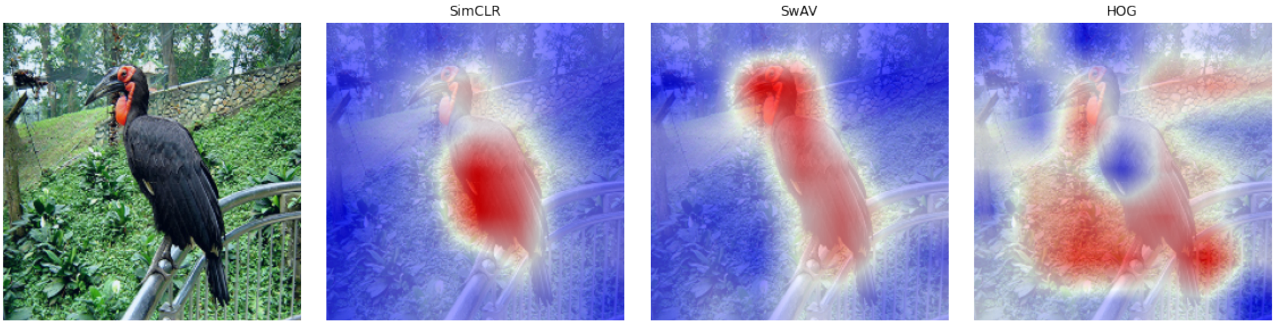


Fig. 9 The figure shows the RELAX explanation for two deep learning-based feature extractors compared with the traditional HOG algorithm. Figure shows how HOG features puts little attention on the bird and mostly focus on the background. Image is taken from PASCAL VOC.

cat receives less attention. In Figure 8, both SimCLR and SwAV focus on the bird, while the HOG features are more focused on other regions in the image. For instance, the iron rod and a tree in the background and are indicated as being important for the representation of this image. Both examples provide insights into why HOG features lead to inferior performance, when compared with features produced by deep neural networks. This information would not be available without the proposed RELAX framework.

6 Conclusion

In this work, we presented RELAX, a framework for explaining representations produced by any feature extractors. RELAX is based on masking out parts of an image and measuring the similarity with an unmasked version in the representation space. We introduced a principled approach to quantifying uncertainty in explanations. RELAX was evaluated by comparing several widely used feature extractors. Results indicate that there can be a big difference in the quality of the explanations. It was shown that filtering out parts of an explanation based on its uncertainty can improve the faithfulness, and that RELAX can have a facilitating role, providing explainability for several downstream applications such as multi-view clustering. We believe that RELAX can be an important addition in the intersection between XAI and representation learning.

Acknowledgments. The authors would like to thank Nils Midtbø, Caroline Granås, Theodor Ross, Suaiba Salahuddin, Sigrid Vold Jensen, Thomas Johansen, Julianne Nyvold, Kristoffer Furøy, Jostein Henriksen, Erland Grimstad, Jonas Mørch-Lampe, Inger Solheim, and Andreas Kvammen for participating in the user study through human evaluation.

Statements and Declarations

6.1 Competing Interests

The authors have no competing interests to declare that are relevant to the content of this article.

6.2 Funding

This work was financially supported by the Research Council of Norway (RCN), through its Centre for Research-based Innovation funding scheme (Visual Intelligence, grant no. 309439), and Consortium Partners. The work was further partially funded by RCN FRIPRO grant no. 315029, RCN IKTPLUSS grant no. 303514, and the UiT Thematic Initiative “Data-Driven Health Technology”.

Appendix A

An alternative approach for creating the random variable $\bar{\mathbf{h}}$ is the following:

$$\bar{\mathbf{h}} = f(\mathbf{X} \odot \mathbf{M} - \mathbf{D}(1 - \mathbf{M})), \tag{A1}$$

where each element of $\mathbf{D}$ follows $N(\mu_{x_{ij}}, \sigma_{x_{ij}})$. The mean $\mu_{x_{ij}}$ and standard deviation $\sigma_{x_{ij}}$ is estimated by averaging across all samples in the data. Such a strategy could avoid potential distribution shifts that might occur when zeroing out large parts of the image, but also required determining the mean and variance of the data distribution.

Table A1 displays localisation scores scores the two masking strategies outlines in Section 3.1, namely zero masking or insertion of normally distributed noise. While there is some variation in the results, masking out with zeros provide the highest performance overall.

Appendix B

Figure B1 shows alternative strategies for masking out part of the input. One alternative is to apply Bernoulli noise to the input, which is equivalent to using Dropout (Srivastava et al, 2014) on the input. However, However, this does not introduce noise with spatial awareness, and therefore results in failing to explain the representation of the image. Another option is to drop regions of the input, such that objects could be fully or partially removed from the input. This could be achieved using the DropBlock algorithm (Ghiasi et al, 2018). However, this requires tuning the size of the mask on the input, which will be highly dependent on the objects present in the image. Such a per-image tuning would be impractical in most scenarios.

Appendix C

In this section we present the proofs for all theorems in the main paper.

C.1 Proof of Theorem 1

Proof First, let the Bounded difference assumption be defined as follows:

Definition 3.1 (Bounded difference assumption)**.** *Let a be some set and $f : A^N$ $f : A^N \to \mathbb{R}$. The function f satisfies the bounded differences assumption if if there exists real numbers $c_1, \dots, c_N \geq 0$ so that for all $i = 1, \dots, N$,*

$$\sup_{x_1,\dots,x_N,x_i \in A} |f(x_1, \dots, x_N, x_i') - f(x_1, \dots, x_N, x_i')| \tag{C2}$$

We then have the following lemmas:

Lemma 3.1 (McDiarmid’s inequality)**.** *Let $X_1, \dots, X_N$ be arbitrary independent random variables on set A and $f : A^N \to \mathbb{R}$ satisfies the bounded difference assumption. Then, for all $t > 0$*

$$P(|f(X_1, \dots, X_N) - \mathrm{E}[f(X_1, \dots, X_N)|] \geq t) \leq 2e^{\frac{-2t^2}{\sum_{n=1}^{N} c_n^2}} \tag{C3}$$

Proof See McDiarmid (1989). □

Lemma 3.2. *Let $X_1, \dots, X_N$ and f be defined as in Lemma 3.1, then if each X_n satisfies $X_n \in (a_n, b_n)$ and $f(X_1, \dots, X_N) = \sum_{n=1}^{N} X_n$, then $c_n = b_n - a_n$.*

Proof See McDiarmid (1989). □

We are now ready to prove the theorem. First, let

$$X_n = \frac{s(\mathbf{h}, \bar{\mathbf{h}}_n) M_{ij}(n)}{N}, \tag{C4}$$

and

$$f(X_1, \dots, X_n) = \sum_{n=1}^{N} X_n. \tag{C5}$$

Scores	Methods	Supervised		SimCLR		SwAV	
		COCO	VOC	COCO	VOC	COCO	VOC
pointing game	RELAX (zeros)	**72.6±0.1**	**86.6±0.2**	**68.7±0.3**	**85.2±0.3**	**67.8±0.2**	84.7±0.2
	RELAX (noise)	72.0±0.5	86.0±0.3	66.6±0.1	84.3±0.7	67.7±0.5	**85.1±0.3**
top k	RELAX (zeros)	**72.8±0.4**	**86.9±0.1**	**69.0±0.3**	**85.6±0.2**	68.1±0.4	85.1±0.2
	RELAX (noise)	72.4±0.4	86.5±0.1	66.0±0.3	84.2±0.2	**68.2±0.3**	**85.3±0.2**
relevance rank	RELAX (zeros)	56.4±0.0	**70.2±0.1**	**54.2±0.2**	**69.8±0.1**	52.4±0.1	69.1±0.0
	RELAX (noise)	**56.7±0.0**	70.1±0.1	53.5±0.1	68.5±0.0	**52.8±0.1**	**69.2±0.0**

Table A1 Evaluation of zero versus noise masking strategy in terms of pointing game, top k, and relevance rank scores in percentages and averaged over 3 runs. Higher is better and bold numbers highlight the top performance. Results indicate that zero masking provides the best performance.

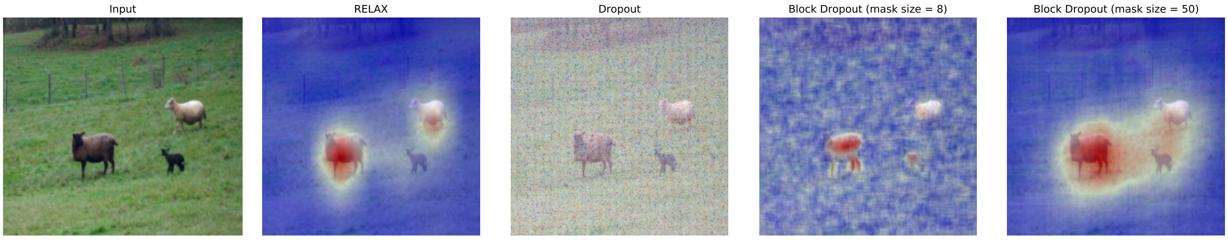


Fig. B1 Comparison of different masking strategies. Leftmost image shows input, and second to left is the RELAX explanations with the masking presented in the main paper. The center image is with Bernoulli-noise (Dropout) directly on the input, and the remaining two images are with Block Dropout with different block size. The example illustrates that other masking strategies either fail completely, or require per-image parameter tuning, which is impractical in most scenarios.

Since $s(\cdot,\cdot)$ is bounded in $(0, 1)$ (we use the cosine similarity between vectors with non-negative elements (ReLU outputs)), we have $a_n = 0$ and $b_n = 1/N$, which gives $c_n = 1/N$ by Lemma 3.2.

Now, observe that

$$f(X_1, \ldots, X_n) = \frac{1}{N}\sum_{n=1}^{N} s(\mathbf{h}, \bar{\mathbf{h}}_n) M_{ij}(n) = \bar{R}_{ij}. \tag{C6}$$

Combining Lemmas 3.1 and 3.2 then gives

$$P(|\bar{R}_{ij} - R_{ij}|] \geq t) \leq 2e^{\frac{-2t^2}{\Sigma_{n=1}^{N}(1/N)^2}} \tag{C7}$$

for all $t > 0$. Inserting $N = -\ln(\delta/2)/2t^2$ gives

$$P(|\bar{R}_{ij} - R_{ij}|] \geq t) \leq 2e^{\frac{-2t^2}{\Sigma_{n=1}^{N}(1/N)^2}} \tag{C8}$$

$$= 2e^{-2t^2\left(-\frac{\ln(\delta/2)}{2t^2}\right)} \tag{C9}$$

$$= 2e^{\ln(\delta/2)} \tag{C10}$$

$$= \delta, \tag{C11}$$

which concludes our proof. □

In Figure C2 we show an empirical validation the bound. We calculate the absolute error as the number of masks increase, averaged over 10 randomly sampled images from the PASCAL VOC dataset. To obtain a value for R_{ij}, we use 10000 masks and average over 10 runs for a single sample. The results indicate that the true error is much lower than the proposed bound, which we attribute to setting $a_n = 0$. While it is possible to obtain a similarity of 0, it is highly unlikely since our masking strategy never removes all information in an image.

C.2 Proof of Theorem 2

Proof Since $s(\cdot,\cdot)$ is a valid Mercer kernel, we can write $s(\mathbf{h}, \bar{\mathbf{h}}_n) = \langle \phi(\mathbf{h}), \phi(\bar{\mathbf{h}}_n)\rangle_{\mathcal{H}}$. This gives

$$\bar{R}_{ij} = \frac{1}{N}\sum_{n=1}^{N} \langle \phi(\mathbf{h}), \phi(\bar{\mathbf{h}}_n)\rangle_{\mathcal{H}} M_{ij}(n) \tag{C12}$$

$$= \langle \phi(\mathbf{h}), \frac{1}{N}\sum_{n=1}^{N} \phi(\bar{\mathbf{h}}_n) M_{ij}(n)\rangle_{\mathcal{H}} \tag{C13}$$

by the bilinearity of the inner product on $\mathcal{H}$. □

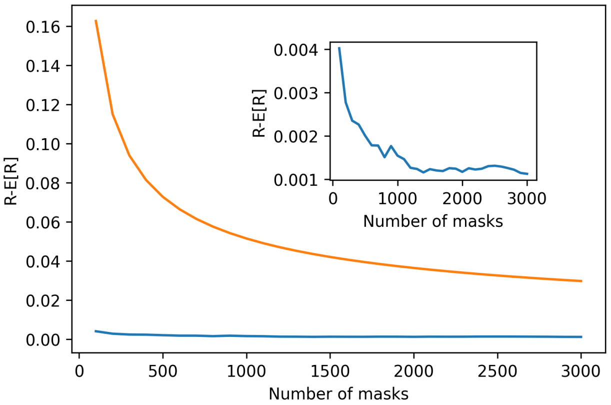


Fig. C2 Empirical evaluation of the derived bound for the number of masks necessary for low estimation error. We calculate the absolute error as the number of masks increase, average over 10 randomly samples images from the PASCAL VOC dataset. To obtain a value for R_{ij}, we use 10000 masks and average over 10 runs for a single sample. Results indicate that the estimation error is much lower than the predicted bound.

C.3 Proof of Theorem 3

Proof Observe that

$$\bar{R}_{ij} \cdot \frac{N}{\sum_{n'=1}^{N} M_{ij}(n')} \tag{C14}$$

$$= \frac{N}{\sum_{n'=1}^{N} M_{ij}(n')} \cdot \frac{1}{N} \sum_{n=1}^{N} s(\cdot, \bar{\mathbf{h}}_n) M_{ij}(n) \tag{C15}$$

$$= \frac{1}{\sum_{n'=1}^{N} M_{ij}(n')} \sum_{n=1}^{N} s(\cdot, \bar{\mathbf{h}}_n) M_{ij}(n) \tag{C16}$$

$$= p_{ij}(\mathbf{h}) \tag{C17}$$

$\bar{R}_{ij}$ is therefore proportional to $p_{ij}(\mathbf{h})$. □

Appendix D

We investigate the potential differences between the one-pass and two-pass version of RELAX. For a given image, we calculate the absolute error between the one-pass and two-pass estimates for different number of masks. The results are shown in Figure D3 and illustrate that the difference between the two methods is very small, particularly as the number of masks increases. However, since the one-pass version computes both the importance and uncertainty in one pass through the data, it requires only half the number of masks compared to the two pass version, thus increasing the computational efficiency of RELAX.

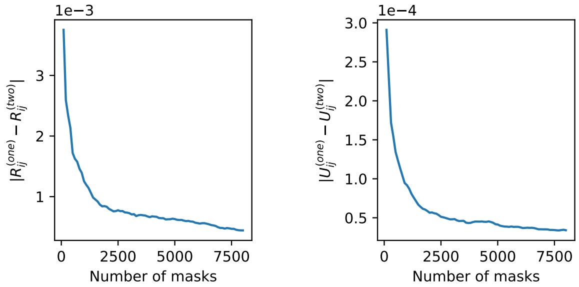


Fig. D3 Absolute error of one-pass versus two-pass version of RELAX for importance (leftmost figure) and uncertainty (rightmost figure), averaged over 50 images from the VOC dataset. The figure shows how the difference between the versions is small for both the importance and uncertainty estimates.

Appendix E

The user study in the main manuscript was conducted by having a group of participants select among competing explanations for a random selection of images from the PASCAL VOC dataset. The group of participants consisted of men and women, where some had knowledge of machine learning and other were uneducated. None of the participants have been involved in the development of this work. Figure E4 displays an example from the study. The participants were shown an image with 6 competing explanations, and asked to chose which one they preferred. To determine which explanation each participant judged to be the "best", they were told to ask themselves the following questions:

> "Which of these explanations agree the most with how you would explain the important content in the given image?

For each image, the explanations were shuffled randomly. The participants were shown 10 images, and asked to only pick on explanation. Overall, 13 people participated in the study.

There are several limitations. Both the number of images and the number of participants could have been greater. The participants had to chose one explanation, when in some cases they might have wanted to select none or more explanations. Also, the images could have been selected from other datasets. There are also potential biases with the study. Most participants are from one country and from a limited age segment. Lastly,

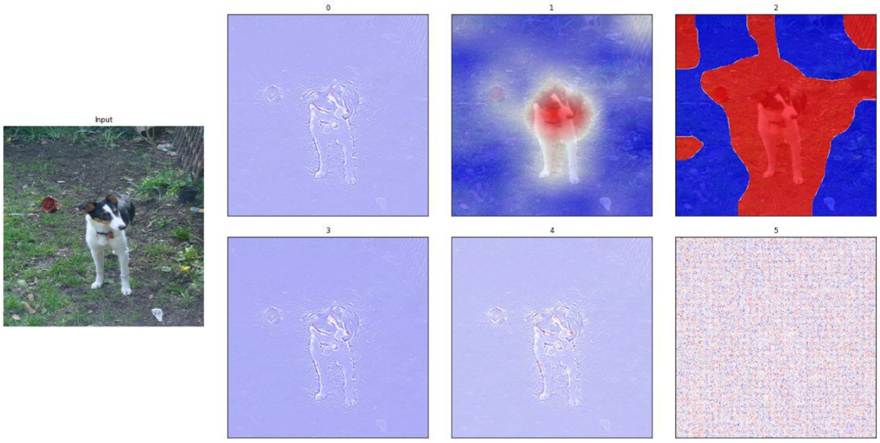


Fig. E4 Example from the human evaluation experiment. Participants were asked to select which explanation they preferred out of the 6 alternatives. For each of the images, the explanations were shuffled in a random order. One of the explanations for each image was randomly sampled from random noise, in order to assess if any participants would select a nonsensical explanation.

we did not control the type of screen that participants performed their evaluation on, which could also have an undesirable affect.

Appendix F

Appendix G

This section presents additional qualitative results. Figure G5 to G14 displays examples of explanations and their associated uncertainty, provided by RELAX, for images from the VOC and COCO dataset. Figure G5 displays an example where all feature extractors agree in terms of importance, but the degree of uncertainty varies. Figure G6 shows an example where only SwAV highlight both objects as important for the representation. Similarly, Figure G7 displays an example where only SwAV is considering both the person and the car as important for the representation. Figure G7 to G14 shows similar examples where RELAX provides insights into the different feature extractors.

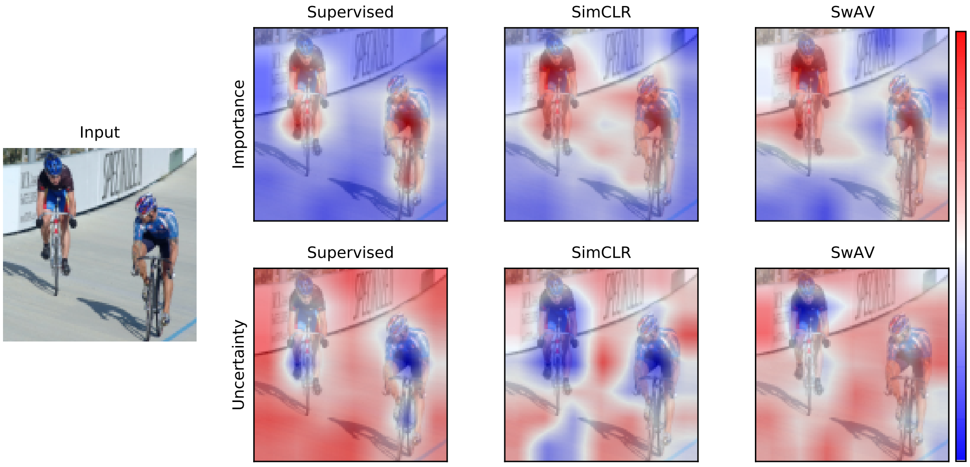


Fig. G5 Example from the VOC dataset.

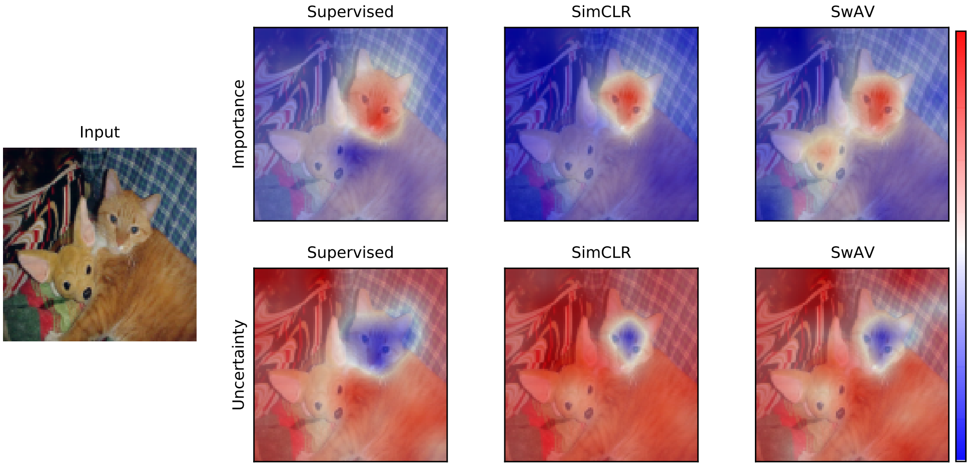


Fig. G6 Example from the COCO dataset.

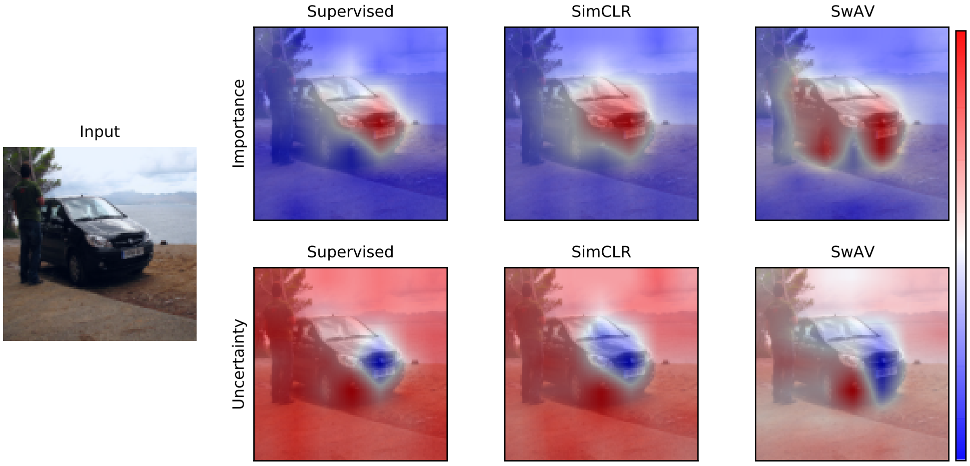


Fig. G7 Example from the VOC dataset.

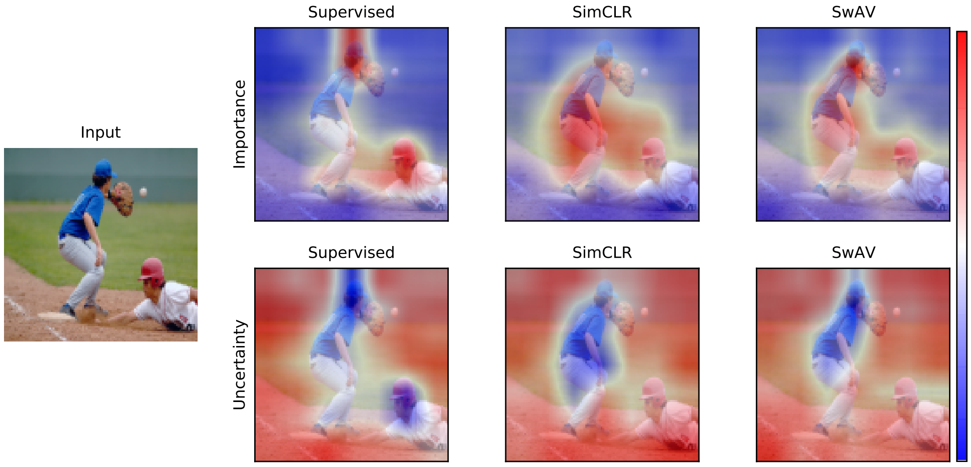


Fig. G8 Example from the COCO dataset.

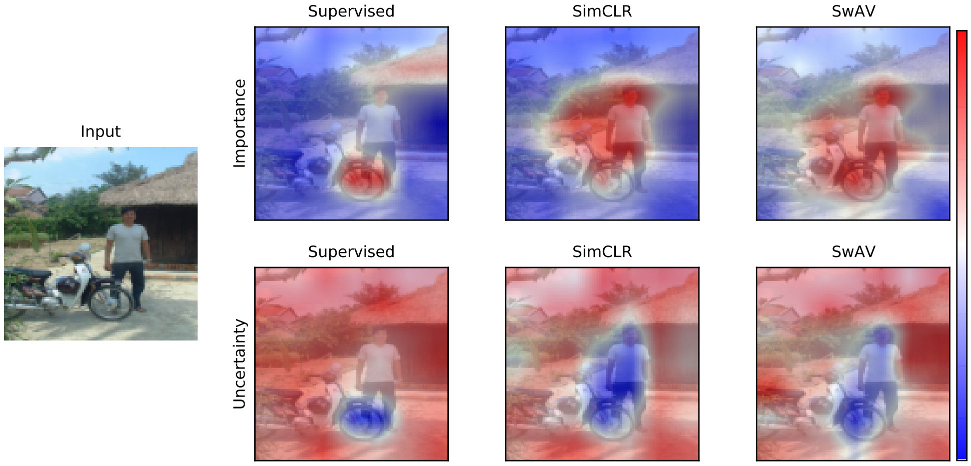


Fig. G9 Example from the VOC dataset.

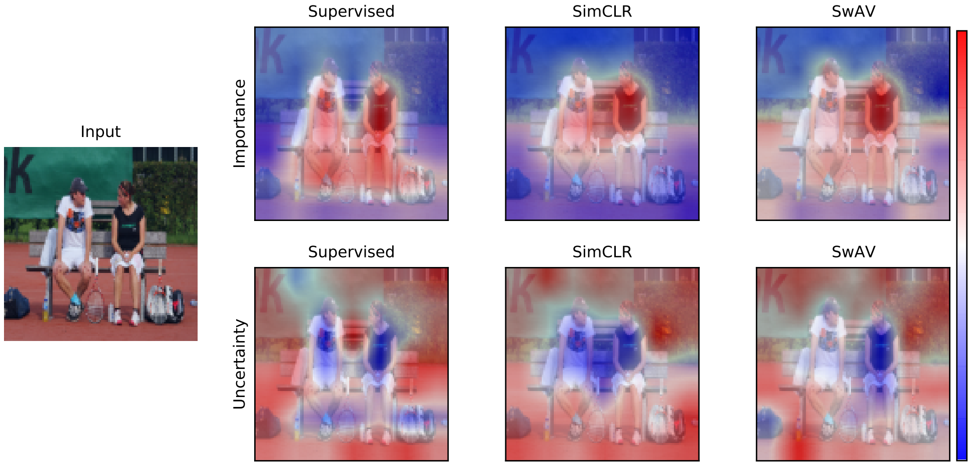


Fig. G10 Example from the COCO dataset.

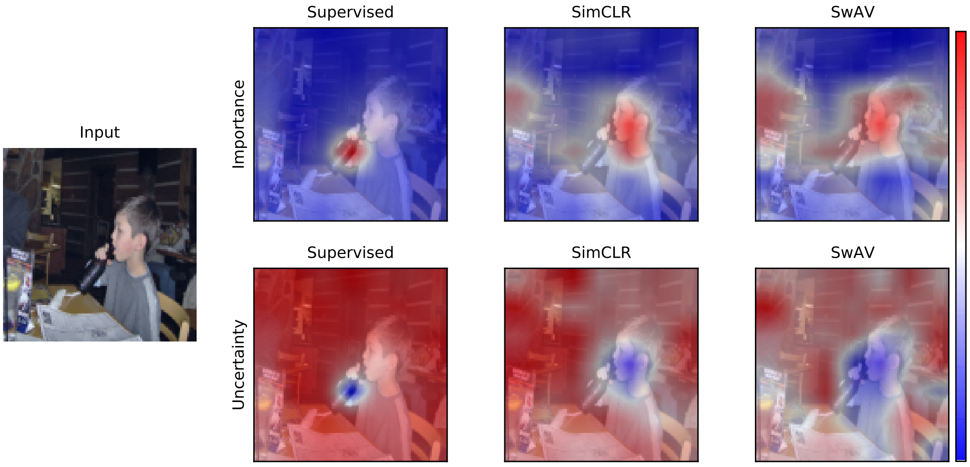


Fig. G11 Example from the VOC dataset.

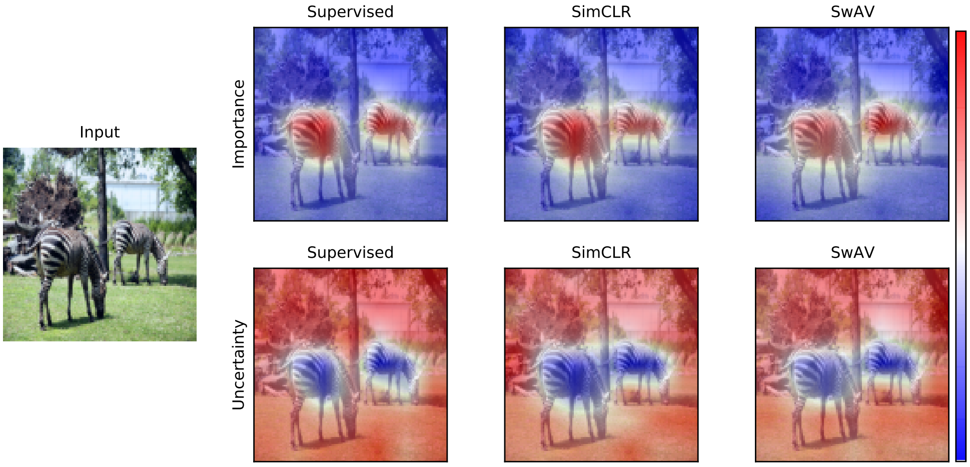


Fig. G12 Example from the COCO dataset.

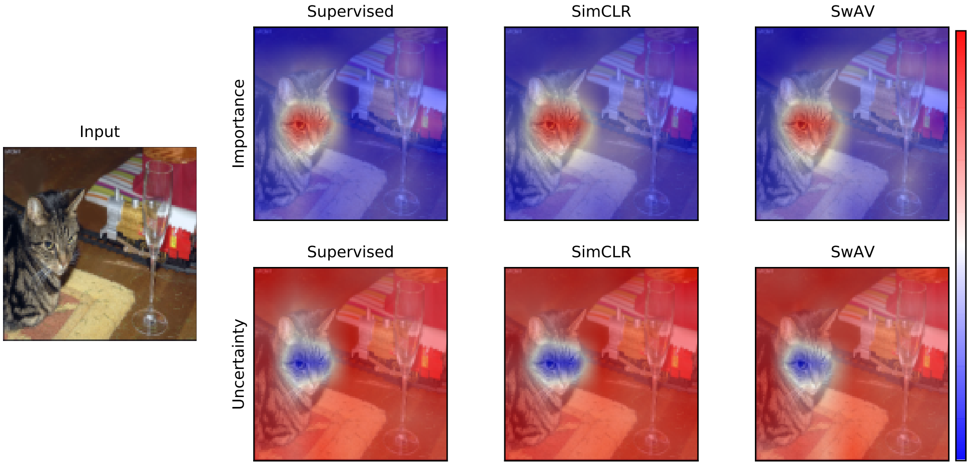


Fig. G13 Example from the VOC dataset.

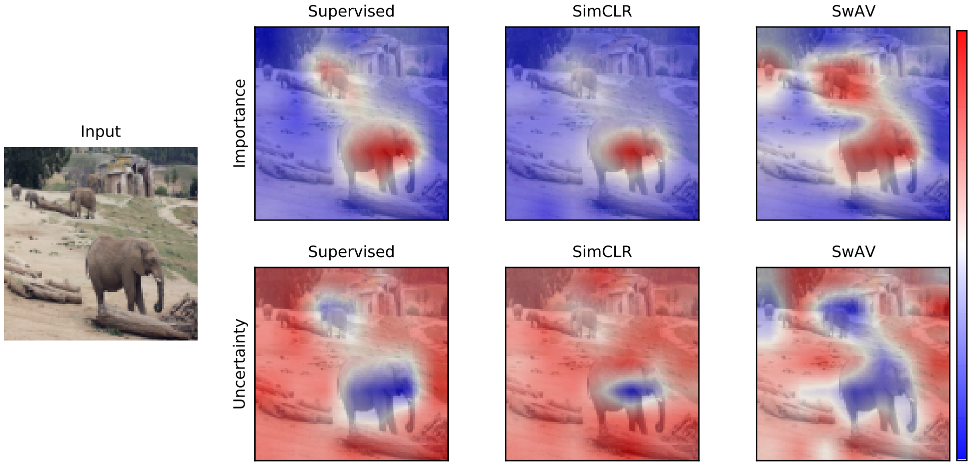


Fig. G14 Example from the COCO dataset.

References

Adebayo J, Gilmer J, Muelly M, et al (2018) Sanity checks for saliency maps. In: Advances in Neural Information Processing Systems. Curran Associates, Inc.

Alvarez-Melis D, Jaakkola TS (2018) Towards robust interpretability with self-explaining neural networks. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, NIPS'18, p 7786–7795

Antoran J, Bhatt U, Adel T, et al (2021) Getting a {clue}: A method for explaining uncertainty estimates. In: International Conference on Learning Representations

Arras L, Osman A, Samek W (2022) Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. Information Fusion 81:14–40. https://doi.org/https://doi.org/10.1016/j.inffus.2021.11.008, URL https://www.sciencedirect.com/science/article/pii/S1566253521002335

Bach S, Binder A, Montavon G, et al (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE 10(7):e0130,140. https://doi.org/10.1371/journal.pone.0130140

Bau D, Zhou B, Khosla A, et al (2017) Network dissection: Quantifying interpretability of deep visual representations. In: IEEE Computer Vision and Pattern Recognition

Bykov K, Höhne MM, Müller K, et al (2020) How much can I trust you? - quantifying uncertainties in explaining neural networks. CoRR abs/2006.09000. URL https://arxiv.org/abs/2006.09000, https://arxiv.org/abs/2006.09000

Caron M, Misra I, Mairal J, et al (2020) Unsupervised learning of visual features by contrasting cluster assignments. In: Advances in Neural Information Processing Systems, pp 9912–9924

Chen C, Li O, Tao C, et al (2019) This looks like that: Deep learning for interpretable image recognition. In: International Conference on Neural Information Processing Systems

Chen T, Kornblith S, Norouzi M, et al (2020) A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp 1597–1607

Chen X, He K (2021) Exploring simple siamese representation learning. In: IEEE Computer Vision and Pattern Recognition, pp 15,750–15,758

Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE Computer Vision and Pattern Recognition, pp 886–893 vol. 1, https://doi.org/10.1109/CVPR.2005.177

Deng J, Dong W, Socher R, et al (2009) Imagenet: A large-scale hierarchical image database. In: IEEE Computer Vision and Pattern Recognition, Ieee, pp 248–255

Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. 1702.08608

Everingham M, Gool LV, Williams CKI, et al (2009) The pascal visual object classes (VOC) challenge. International Journal of Computer Vision pp 303–338. https://doi.org/10.1007/s11263-009-0275-4, URL https://doi.org/10.1007/s11263-009-0275-4

Falcon W, Cho K (2020) A framework for contrastive self-supervised learning and designing a new approach. arXiv preprint arXiv:200900104

Fong R, Vedaldi A (2018) Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In: IEEE Computer Vision and Pattern Recognition, pp 8730–8738, https://doi.org/10.1109/CVPR.2018.00910

Fong R, Patrick M, Vedaldi A (2019) Understanding deep networks via extremal perturbations and smooth masks. In: IEEE International Conference on Computer Vision

Fong RC, Vedaldi A (2017) Interpretable explanations of black boxes by meaningful perturbation. In: IEEE International Conference on Computer Vision, pp 3449–3457, https://doi.org/10.1109/ICCV.2017.371

Gal Y, Ghahramani Z (2016) Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp 1050–1059

Ghiasi G, Lin TY, Le QV (2018) Dropblock: A regularization method for convolutional networks. In: International Conference on Neural Information Processing Systems, p 10750–10760

He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: 2016 CVPR, pp 770–778, https://doi.org/10.1109/CVPR.2016.90

He K, Fan H, Wu Y, et al (2020) Momentum contrast for unsupervised visual representation learning. In: IEEE Computer Vision and Pattern Recognition

Karimi AH, Barthe G, Balle B, et al (2020) Model-agnostic counterfactual explanations for consequential decisions. In: International Conference on Artificial Intelligence and Statistics, pp 895–905

Kim B, Wattenberg M, Gilmer J, et al (2018) Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: International Conference on Machine Learning, pp 2673–2682

Koh PW, Liang P (2017) Understanding black-box predictions via influence functions. In: International Conference on Machine Learning, p 1885–1894

Kolek S, Nguyen DA, Levie R, et al (2021) A rate-distortion framework for explaining black-box model decisions. 2110.08252

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges C, Bottou L, et al (eds) Advances in Neural Information Processing Systems, vol 25. Curran Associates, Inc., URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Laina I, Fong RC, Vedaldi A (2020) Quantifying learnability and describability of visual concepts emerging in representation learning. In: Advances in Neural Information Processing Systems

Lin TY, Maire M, Belongie S, et al (2014) Microsoft COCO: Common objects in context. In: Computer Vision – ECCV 2014. Springer International Publishing, p 740–755, https://doi.org/10.1007/978-3-319-10602-1_48, URL https://doi.org/10.1007/978-3-319-10602-1_48

Lin Y, Gou Y, Liu Z, et al (2021) Completer: Incomplete multi-view clustering via contrastive prediction. In: IEEE Computer Vision and Pattern Recognition, pp 11,174–11,183

Liu W, Lin R, Liu Z, et al (2021) Learning with hyperspherical uniformity. In: Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, vol 130. PMLR, pp 1180–1188, URL http://proceedings.mlr.press/v130/liu21d.html

Losch M, Fritz M, Schiele B (2021) Semantic bottlenecks: Quantifying and improving inspectability of deep representations. International Journal of Computer Vision 129(11):3136–3153. https://doi.org/10.1007/s11263-021-01498-0, URL https://doi.org/10.1007/s11263-021-01498-0

McCullagh P, Nelder J (1989) Generalized Linear Models, Second Edition. Chapman & Hall

McDiarmid C (1989) On the method of bounded differences, Cambridge University Press, p 148–188. https://doi.org/10.1017/CBO9781107359949.008

Mercer J (1909) Functions of positive and negative type, and their connection with the theory of integral equations. Philosophical Transactions of the Royal Society, London 209:415–446

Mordvintsev A, Olah C, Tyka M (2015) Inceptionism: Going deeper into neural networks. URL https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

Nguyen A, Martinez MR (2020) On quantitative aspects of model interpretability. ArXiv abs/2007.07584

Parzen E (1962) On estimation of a probability density function and mode. The Annals of Mathematical Statistics 33(3):1065–1076. https://doi.org/10.1214/aoms/1177704472, URL https://doi.org/10.1214/aoms/1177704472

Paszke A, Gross S, Massa F, et al (2019) Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, pp 8024–8035

Pedreschi D, Giannotti F, Guidotti R, et al (2019) Meaningful explanations of black box AI decision systems. Proceedings of the AAAI Conference on Artificial Intelligence 33:9780–9784. https://doi.org/10.1609/aaai.v33i01.33019780, URL https://doi.org/10.1609/aaai.v33i01.33019780

Petsiuk V, Das A, Saenko K (2018) Rise: Randomized input sampling for explanation of black-box models. In: Proceedings of the British Machine Vision Conference

Ribeiro MT, Singh S, Guestrin C (2016) "why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, pp 1135–1144

Samek W, Binder A, Montavon G, et al (2017) Evaluating the visualization of what a deep neural network has learned. IEEE TNNLS 28(11):2660–2673. https://doi.org/10.1109/TNNLS.2016.2599820

Schulz K, Sixt L, Tombari F, et al (2020) Restricting the flow: Information bottlenecks for attribution. In: International Conference on Learning Representations

Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations

Smilkov D, Thorat N, Kim B, et al (2017) Smoothgrad: removing noise by adding noise. In: International Conference on Machine Learning Visualization Workshop

Springenberg JT, Dosovitskiy A, Brox T, et al (2015) Striving for simplicity: The all convolutional net. In: ICLR Workshop

Srivastava N, Hinton G, Krizhevsky A, et al (2014) Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research pp 1929–1958

Teye M, Azizpour H, Smith K (2018) Bayesian uncertainty estimation for batch normalized deep networks. In: International Conference on Machine Learning, pp 4907–4916

Theodoridis S, Koutroumbas K (2009) Pattern Recognition, Fourth Edition. Academic Press

Tonekaboni S, Joshi S, McCradden MD, et al (2019) What clinicians want: Contextualizing explainable machine learning for clinical end use. In: Machine Learning for Healthcare Conference, pp 359–380

Wang W, Arora R, Livescu K, et al (2015) On deep multi-view representation learning. In: International Conference on Machine Learning, p 1083–1092

Wen J, Zhang Z, Xu Y, et al (2020) Cdimc-net: Cognitive deep incomplete multi-view clustering network. In: International Joint Conference on Artificial Intelligence

West DHD (1979) Updating mean and variance estimates: An improved method. Commun ACM 22(9):532–535. https://doi.org/10.1145/359146.359153, URL https://doi.org/10.1145/359146.359153

Wickstrøm K, Kampffmeyer M, Jenssen R (2018) Uncertainty modeling and interpretability in convolutional neural networks for polyp segmentation. In: IEEE International Workshop on Machine Learning for Signal Processing, pp 1–6

Wickstrøm K, Kampffmeyer M, Jenssen R (2020) Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. Medical Image Analysis 60:101,619. https://doi.org/https://doi.org/10.1016/j.media.2019.101619, URL https://www.sciencedirect.com/science/article/pii/S1361841519301574

Wickstrøm K, Mikalsen K, Kampffmeyer M, et al (2021) Uncertainty-aware deep ensembles for reliable and explainable predictions of clinical time series. IEEE Journal of Biomedical and Health Informatics 25(7):2435–2444. https://doi.org/10.1109/JBHI.2020.3042637

Yang B, Fu X, Sidiropoulos ND, et al (2017) Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In: International Conference on Machine Learning, p 3861–3870

Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B, et al (eds) European Conference on Computer Vision, pp 818–833

Zhang J, Bargal SA, Lin Z, et al (2017) Top-down neural attention by excitation backprop. International Journal of Computer Vision 126(10):1084–1102. https://doi.org/10.1007/s11263-017-1059-x, URL https://doi.org/10.1007/s11263-017-1059-x

Zhang Y, Song K, Sun Y, et al (2019) ”Why Should You Trust My Explanation?” Understanding Uncertainty in LIME Explanations. In: Workshop on AI for Social Good

15

Paper IV

A clinically motivated self-supervised approach for content-based image retrieval of CT liver images

Kristoffer Wickstrøm, Eirik Østmo, Keyur Radya, Karl Øyvind Mikalsen, Michael Kampffmeyer, Robert Jenssen

Submitted to Computerized Medical Imaging and Graphics

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

A clinically motivated self-supervised approach for content-based image retrieval of CT liver images

Kristoffer Knutsen Wickstrøm[a,*], Eirik Agnalt Østmo[a], Keyur Radiya[b], Karl Øyvind Mikalsen[a,b], Michael Christian Kampffmeyer[a,c], Robert Jenssen[a,c,d]

[a]*Machine Learning Group at the Department of Physics and Technology, UiT the Arctic University of Norway, Tromsø NO-9037, Norway*
[b]*Department of Gastrointestinal Surgery, University Hospital of North Norway (UNN), Tromsø, Norway*
[c]*Norwegian Computing Center, Department SAMBA, P.O. Box 114 Blindern, Oslo NO-0314, Norway*
[d]*Department of Computer Science, University of Copenhagen, Universitetsparken 1, 2100 København Ø, Denmark*

ARTICLE INFO

Article history:

Keywords:
Content-based image retrieval
Self-supervised learning
CT liver imaging
Explainability

ABSTRACT

Deep learning-based approaches for content-based image retrieval (CBIR) of CT liver images is an active field of research, but suffers from some critical limitations. First, they are heavily reliant on labeled data, which can be challenging and costly to acquire. Second, they lack transparency and explainability, which limits the trustworthiness of deep CBIR systems. We address these limitations by (1) proposing a self-supervised learning framework that incorporates domain-knowledge into the training procedure and (2) providing the first representation learning explainability analysis in the context of CBIR of CT liver images. Results demonstrate improved performance compared to the standard self-supervised approach across several metrics, as well as improved generalisation across datasets. Further, we conduct the first representation learning explainability analysis in the context of CBIR, which reveals new insights into the feature extraction process. Lastly, we perform a case study with cross-examination CBIR that demonstrates the usability of our proposed framework. We believe that our proposed framework could play a vital role in creating trustworthy deep CBIR systems that can successfully take advantage of unlabeled data.

1. Introduction

Content-based image retrieval (CBIR) is a core research area in medical image analysis, with numerous studies across many different image modalities (Barata and Santiago, 2021; Ramalhinho et al., 2021; Haq et al., 2021). CBIR supports clinicians in retrieving relevant images from a large database compared to a query image, which reduces labor-intensive manual search and aids in diagnosis. For instance, a physician might want to investigate how patients in a large database with a similar disease as a new patient, such as liver metastasis, were diagnosed. The information from the previous diagnoses can then be used to determine the proper treatment for the new patient. In analysis of CT image of the liver, CBIR have been an active and important area of medical image analysis for many years (Zhao et al., 2004; Chi et al., 2013; Yoshinobu et al., 2020). CBIR has the potential to make labour intensive tasks in the clinical workflow more time efficient, as illustrated in Section 7.3.

Currently, deep learning-based CBIR, or deep CBIR, constitute the state-of-the-art of CBIR (Silva et al., 2020; Yoshinobu et al., 2020; Haq et al., 2021), due to its high precision and efficiency. However, deep CBIR suffers from some critical limitations. First (1), current deep CBIR for CT liver images rely on labeled data for training (Yoshinobu et al., 2020). Obtaining labeled data can be costly and time-consuming, which therefore

*Corresponding author: kristoffer.k.wickstrom@uit.no;

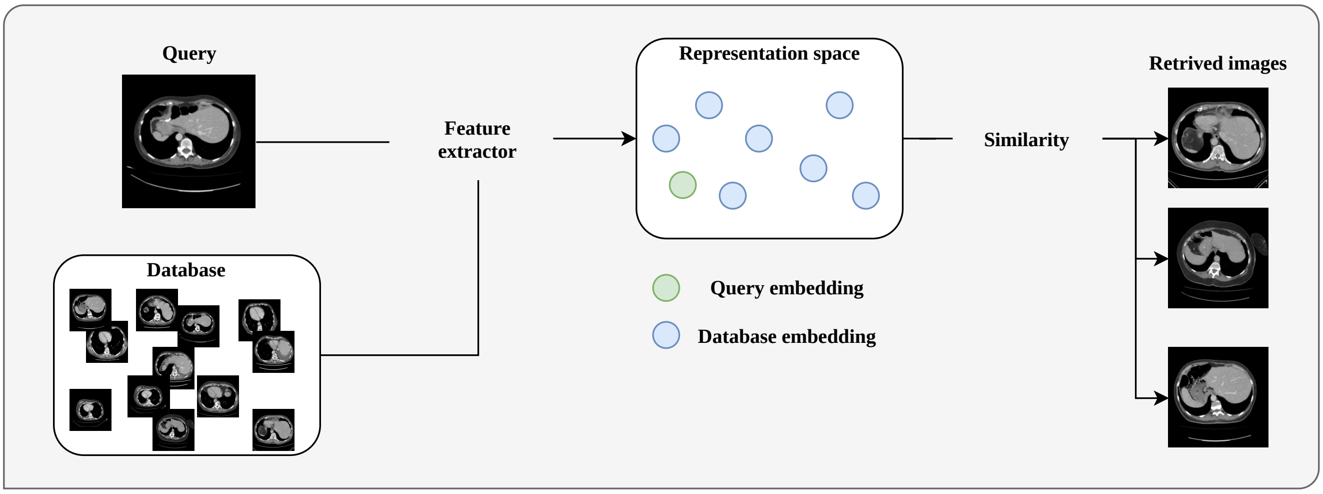


Fig. 1: Illustration of content-based image retrieval.

limits the usability of deep CBIR systems. However, recent works have shown how self-supervised learning can leverage unlabeled data to improved CBIR systems (Siradjuddin et al., 2019; Monowar et al., 2022), but such approaches have not been explored in the context of CBIR of CT liver images. Second (2), deep CBIR suffer from a fundamental lack of explainability. This can have detrimental effects in a clinical setting, since deep learning-based systems are known to exploit confounding factors and artifacts to make their predictions. For instance, Gautam et al. (2022) showed that a deep-learning-based system learned to use tokens and artifacts in X-ray images to makes its predictions instead of clinically relevant features. These tokens and artifacts would not be present for new patients, and such a system would not work as intended if put into clinical practice. Therefore, it is not advisable to blindly trust the retrieved images from the deep CBIR system without investigating what input features influence the retrieval process through an explainability analysis.

A promising direction to address the first limitation is learning from unlabeled data through self-supervision. Recent self-supervised learning frameworks have shown remarkable results, in some cases even rivalling supervised learning (Chen et al., 2020; Caron et al., 2020; Chen and He, 2021). In a nutshell, contemporary self-supervised approaches train a feature extractor that extract informative representations by exploiting known invariances in the data. These representations can then be used for other tasks, such as CBIR by taking similarities between the new representation to retrieve similar images. These self-supervised approaches have been show to improve performance in the context of chest X-ray (Truong et al., 2021; Azizi et al., 2021) and dermatology classification (Azizi et al., 2021), organ andcardiac segmentation (Hansen et al., 2022), and whole heart segmentation (Dong et al., 2021), but have yet to be developed for CBIR of CT liver images.

In this paper, we propose a clinically motivated self-supervised framework for CBIR of CT liver images. Our proposed framework incorporates domain knowledge that exploits known properties of the liver, which leads to improved performance compared to well-known self-supervised baselines. Concretely, a novel Houndsfield unit clipping strategy that removes non-liver pixels from the input and allows the system to focus on the liver is incorporated into the self-supervised training. While the focus in this paper is on the liver in CT images, our proposed framework could also be used to focus on other organs by altering how the Houndsfield units are clipped.

For the second limitation, great improvements have been made in the field of explainable artificial intelligence (XAI) over the last couple of years, and numerous studies have shown how XAI can improve the reliability and trustworthiness of deep learning-based systems in healthcare (Silva et al., 2020; Gautam et al., 2022). However, the majority of these improvements have been in algorithms that can explain models which produce decisions, such as classification or similarity scores. When learning from unlabeled data through e.g. self-supervised learning, such a score or similarity measure might not be available and standard XAI techniques cannot be applied. But the recent field of representation learning explainability (Wickstrøm et al., 2021) aims at explaining vector representations, and can therefore tackle the lack of explainability in deep CBIR. But such a representation learning explainability analysis has not been performed in the context of CBIR of CT liver images.

Our contributions are:

- A clinically motivated self-supervised framework specifically designed to extract liver specific features.
- A novel explainability analysis that explains the representations produced in the feature extraction process.
- Thorough evaluation on real-world datasets.
- A case-study where images from the same patient are retrieved across different examinations.

2. Related work

2.1. *Content-based image retrieval*

The goal of content-based image retrieval (CBIR) is to find similar images from a large-scale database, given a query image. CBIR is an active area of research that span numerous medical imaging domains, such as X-ray (Haq et al., 2021; Silva et al., 2020), dermatology (Barata and Santiago, 2021; Ballerini et al., 2010), mammography (Jiang et al., 2014), and histopathology (Peng et al., 2019; Zheng et al., 2019). An illustration of a CBIR system in the context of CT liver images is shown in Figure 1.

2.2. *Content-based image retrieval of CT liver images*

CBIR of CT liver images have been extensively studied. Early studies relied on handcrafting features based on certain properties in the images. Gabor filters have been used to extract texture information (Zhao et al., 2004). Texture information have also been combined with density information in the context of focal liver lesion retrieval (Chi et al., 2013). Histogram-based features extraction have been explored to retrieve CT scans with similar liver lesions. Manifold learning have been utilized to facilitate CBIR of CT liver images (Mirasadi and Foruzan, 2019). Lastly, a Bayesian approach has been studied in connection with multi-labeled CBIR of CT liver images (Ramalhinho et al., 2021).

Recently, deep learning-based feature extraction have improved performance significantly in CBIR of CT liver images. The most straight forward approach for deep CBIR is to train a neural network for the task of CT liver image classification and use the intermediate features prior to the classification layer for calculating similarities. This has been demonstrated to produce good results when the network was trained for the task of focal liver lesions detection (Yoshinobu et al., 2020). However, all these approaches need labeled data to train the deep learning-based feature extractor.

2.3. *Self-supervised learning*

Learning from unlabeled data is a fundamental problem in machine learning. Recently, self-supervised learning have shown promising results in computer vision (Chen et al., 2020; Chen and He, 2021), natural language processing (Devlin et al., 2019; Brown et al., 2020), and time series analysis (Franceschi et al., 2019; Wickstrøm et al., 2022). Furthermore, recent studies have also demonstrated that self-supervised learning can improve performance across several imaging domains in medical image analysis (Azizi et al., 2021; Truong et al., 2021; Hansen et al., 2022; Dong et al., 2021).

For computer vision, there are three main approaches to self-supervised learning. First, contrastive self-supervised learning is performed by sampling positive pairs and negative samples and learning a representation where the positive pairs are mapped in close proximity and far from the negative samples. The SimCLR framework (Chen et al., 2020) is one of the most widely used approaches in this category. Second, clustering-based self-supervised learning utilizes clustering algorithms to produce pseudo-labels which in turn are used to learn a useful representation of the data. DeepCluster (Caron et al., 2018) and the SwAV framework (Caron et al., 2020) are two of the most widely used clustering-based self-supervised approaches in the literature. Lastly, siamese self-supervised approaches learns how to produce a useful representation by maximizing agreement between positive pairs of samples. The two main contemporary approaches in siamese self-supervised approaches is the SimSiam framework (Chen and He, 2021) and the BYOL framework (Grill et al., 2020).

2.4. *Explainability*

Explainability is of vital importance for machine learning systems in healthcare. Without it, clinicians cannot fully trust the algorithms decision and the system becomes less reliable. Many recent studies have shown how explainability can be incorporated into deep learning systems for medical image analysis, ranging from diabetic retinopathy (Quellec et al., 2021), dermatology (Barata and Santiago, 2021; Gu et al., 2021), X-ray (Khakzar et al., 2021), and endoscopic images (Wickstrøm et al., 2020; Vasilakakis et al., 2021).

Most of the widely used explainability techniques typically leverage the classification or similarity score to ascertain input feature importance (Springenberg et al., 2015; Schulz et al., 2020; Plummer et al., 2020), and such approaches have been explored in the context of deep CBIR. For models trained for classification tasks, explanations through gradient information have been shown to both provide new insights and improve performance for X-ray images (Silva et al., 2020). For models trained to output a similarity score, it has been shown how the similarity score can be used to provide explanations (Dong et al., 2019; Plummer et al., 2020). Similarity score explanations have been explored for X-ray images (Hu et al., 2022). Lastly, it has been shown that explanation by examples can be effective in histopathological images (Peng et al., 2019).

In the unlabeled setting where only the feature extraction model is available, these techniques are not applicable. In such cases, it is desirable to explain the vector representation of an image, since the decision is not available. Representation learning explainability is a very recent field of XAI, that has yet to be developed for CBIR. In this work, we leverage the RELAX framework (Wickstrøm et al., 2021) to explain the feature extractors trained using self-supervised learning. RELAX is the first method that allows for representation learning explainability and has been shown to provide superior performance to competing alternatives (Wickstrøm et al., 2021).

3. A clinically motivated self-supervised approach for CT liver images

In this section, we present our proposed clinically-motivated self-supervised approach and the SimSiam framework for self-supervised learning.

3.1. *A clinically motivated self-supervised approach for CT liver images*

We propose to incorporate clinical knowledge into our self-supervised framework to learn more clinically relevant features. In self-supervised learning, known invariances in the data are used to train a feature extractor that extracts relevant features from the input images. For instance, the liver can occur on both the left and right hand side of an image, depending on which direction the patient is inspected. Therefore, the feature extractor should be invariant to horizontal flips in the images, and this invariance can be learned by incorporating horizontal flipping into the self-supervised learning procedure. Identifying these invariances is crucial to make the self-supervised system work properly and focus on clinically relevant features in the input images. Our motivation is based on the knowledge that the pixel intensities of the liver lay within a certain range for CT images. A standard pre-processing step is to clip the pixel intensities of the CT images (Li et al., 2018a), such that unimportant pixels are removed prior to learning. The pixel intensities of CT images represent a physical quantity, namely the Houndfield unit. The same clipping is usually applied to all images. However, if this clipping was incorporated into the self-supervised learning procedure, the network could be guided to learn which feature are liver features and which ones are not. In a sense, we are exploiting the knowledge that the liver should be invariant to pixel intensity clipping for a certain range of clipping.

Based on this motivation, we propose a Houndsfield clipping strategy where the pixel values for the same image are clipped and scaled based on different ranges of Houndsfield units. Figure 2 shows how our proposed clipping scheme affects an image. The leftmost image has no clipping applied, and illustrates why it is important to remove some pixel intensities in order to highlight relevant structures in the images. The middle images show the narrow clipping strategy between 50 and 150 Houndsfield units. Notice how only the liver and some other organs are now visible in the image. The rightmost image shows the wide clipping strategy between -200 and 300 Houndsfield units. In this case, some redundant structures are removed, but more organs are left visible compared to the middle image. The images considered in this paper are intra venous contrast enhanced images taken in the portal venous phase. These two ranges were chosen based on the following. First, it is known that the liver typically has Houndsfield units in the range 50-60 (Tisch et al., 2019). Furthermore, we have collected all pixel intensities for the liver in the Decathlon dataset. These values are shown in Figure 3, and illustrates how the narrow clip will remove some of the liver pixels but keep the main proportion, while the wide clip will keep almost all liver pixels apart from some outliers. Our proposed framework for learning representations that focus on liver features is shown in Figure 4. Each image is clipped with the wide and narrow range, before the data augmentation is applied. Afterwards, we follow the SimSiam approach described below. During testing, we use the wide clipping to ensure that most liver pixels are kept in the images.

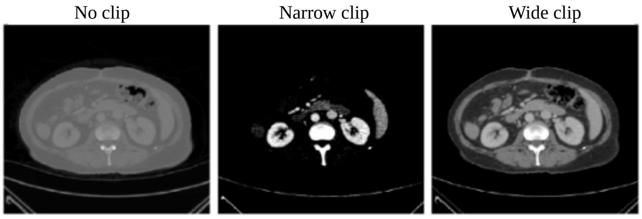


Fig. 2: Effect of Houndsfield unit clipping on CT liver images. From left to right, no clipping, narrow clip (50, 150), and wide clip (-200, 300).

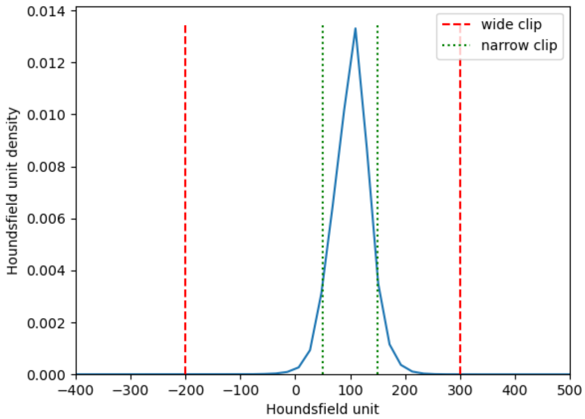


Fig. 3: Distribution of pixel intensity values for liver pixel from the Decathlon dataset and the two clipping strategies used in our proposed framework.

3.2. *SimSiam framework*

In this work, we build on the SimSiam framework. The main motivation for this choice is that both contrastive and clustering-based self-supervised approaches requires a large batch size during training to provide high quality representations (Chen et al., 2020; Caron et al., 2020). This can be computationally challenging, especially if the medical images in question are large. However, the siamese-approaches (Chen and He, 2021) are less sensitive to the batch size used during training. Furthermore, we opt for the SimSiam approach over BYOL to avoid training both a student and a teacher network used in BYOL, again to avoid additional computational overhead.

Let $\mathbf{X} \in \mathbb{R}^{H \times W}$ represent an input image with height H and width W and f a feature extractor that transforms $\mathbf{X}$ into a new d-dimensional representation $\mathbf{h} \in \mathbb{R}^d$, that is $f(\mathbf{X}) = \mathbf{h}$. Next, two views $\mathbf{X}_1$ and $\mathbf{X}_2$ are constructed by augmenting the original image. The task performed in SimSiam to learn a useful representation, is to maximize the similarity between the two views. The representation h is the new representation that can be used for downstream tasks, such as the CBIR. However, the loss is not computed directly on the output of the feature extractor f. Instead, a multilayer perceptron-based projection head g transforms $\mathbf{h}$ into a new representation $\mathbf{z}$, that is $g(\mathbf{h}) = \mathbf{z}$, where the loss is computed. This projector is a crucial component in most self-supervised frameworks (Chen et al., 2020; He et al., 2020), as it avoid dimensional collapse in the representation h (Jing et al., 2022), which is the one that will be used for downstream tasks such as CBIR. The learning is performed by minimizing the negative cosine similarity between the two views:

$$D(\mathbf{z}_1, \mathbf{z}_2) = -\frac{\mathbf{z}_1}{||\mathbf{z}_1||_2} \cdot \frac{\mathbf{h}_2}{||\mathbf{h}_2||_2}, \tag{1}$$

where $||\cdot||_2$ denotes the ℓ_2-norm.

$$L = D(\mathbf{z}_1, \mathbf{h}_2) + D(\mathbf{z}_2, \mathbf{h}_1) \tag{2}$$

An important component of the the SimSiam framework is a stop-gradient $(\mathrm{stopgrad})$ operation, which is incorporate in

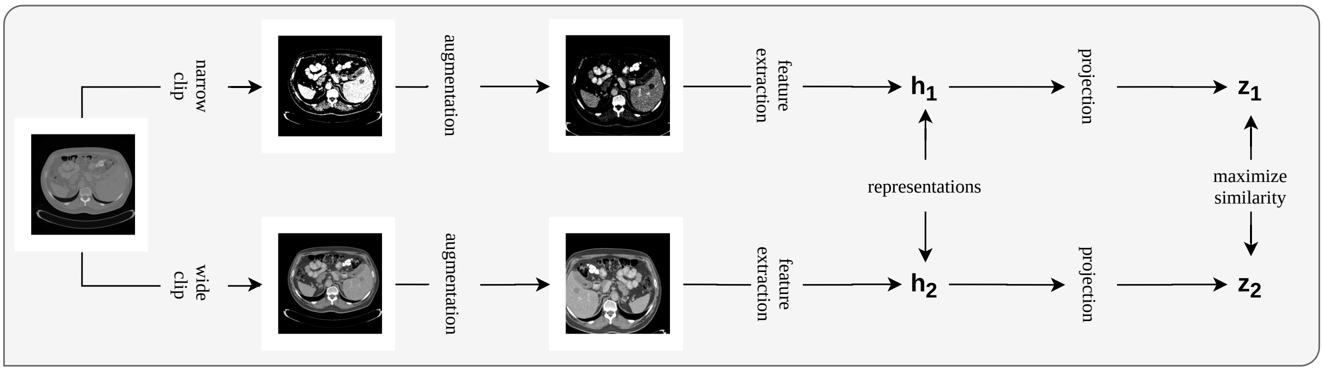


Fig. 4: Illustration of proposed self-supervised framework.

Equation 2 as follows:

$$L = \frac{1}{2}D(\mathbf{z}_1, \text{stopgrad}(\mathbf{h}_2)) + \frac{1}{2}D(\mathbf{z}_2, \text{stopgrad}(\mathbf{h}_1)) \quad (3)$$

The stop-gradient operation is applied to the projector network, such that the encoder on $\mathbf{X}_2$ no gradient from $\mathbf{h}_2$ in the first term, but it recieves gradients from $\mathbf{z}_2$ (and similarly for $\mathbf{X}_1$). The stop-grad operation allows SimSiam to mimic a teacher-student setup, but avoid the need to store two networks. Furthermore, it has been shown that the stop-grad operation is critical to avoid the problem of complete collapse in the representations (Tian et al., 2021).

Data augmentation. The prior knowledge inject through the data augmentation is of paramount importance to ensure that the models learns relevant features. The data augmentation used in SimSiam is similar to the standard approach in recent self-supervised learning (He et al., 2020; Chen et al., 2020):

1. Crop with a random proportion from [0.2, 1.0], and resize to a fixed size.
2. Flip horizontally with a probability of 0.5.
3. Color augmentation is performed by randomly adjusting the brightness, contrast, saturation, and hue of each image with a strength of [0.4, 0.4, 0.4, 0.1]
4. Randomly convert image to gray scale version with a probability 0.2.

Note that the input images are converted to pseudo RBG images by stacking the input image 3 times along the channel axis. Prior works have shown that the augmentation scheme listed above can lead to increased performance across several medical image related tasks (Azizi et al., 2021; Truong et al., 2021; Hansen et al., 2022; Dong et al., 2021), albeit not in the context of CBIR of CT liver images. However, these augmentations are selected with natural images in mind, and do not take into account the properties of CT liver images. Our proposed Hounsfield unit clipping scheme takes into account the particular characteristics of CT images of the liver, which we hypothesize can improve the self-supervised framework.

4. Explaining representations

Explainability is a critical component for creating trustworthy and reliable deep learning-based systems. For deep CBIR, we want to know what information the feature extractor is using to create the representation that the retrieval is based on. This requires explaining the vector representations produced by the feature extractor, which can not be accomplished with standard explainability techniques since they require a classification or similarity score to create the explanation. However, the recent field of representation learning explainability address the problem of explaining representations (Wickstrøm et al., 2021). In this work, we leverage the RELAX (Wickstrøm et al., 2021) framework to explain the representations used in the CBIR system.

4.1. RELAX

RELAX is an occlusion-based explainability framework that provides input feature importance in relation to a vector representation, as opposed to a classification or similarity score. The core idea of RELAX is to evaluate how the representation of an image changes as parts of the image are removed using a mask. Let $\mathbf{M} \in [0, 1]^{H \times W}$ represent a stochastic mask used for removing parts of the image. Next, $\bar{\mathbf{h}} = f(\mathbf{X} \odot \mathbf{M})$, where $\odot$ denotes element-wise multiplication, is the representation of a masked version of $\mathbf{X}$ and $s(\mathbf{h}, \bar{\mathbf{h}})$ is a similarity measure between the unmasked and the masked representation. The intuition behind RELAX is that when informative parts are masked out, the similarity between the two representations should be low, and vice versa for non-informative parts. Finally, the importance R_{ij} of pixel (i, j) is defined as:

$$\bar{R}_{ij} = \frac{1}{N} \sum_{n=1}^{N} s(\mathbf{h}, \bar{\mathbf{h}}_n) M_{ij}(n). \quad (4)$$

Here, $\bar{\mathbf{h}}_n$ is the representation of the image masked with mask n, and $M_{ij}(n)$ the value of element (i, j) for mask n. The similarity measure used in the cosine similarity, as proposed in prior works Wickstrøm et al. (2021). The RELAX framework is illustrated in Figure 5.

The mask generation is a crucial component in RELAX. In this work, we follow the strategy used in previous studies (Petsiuk et al., 2018; Wickstrøm et al., 2021). Binary masks of size

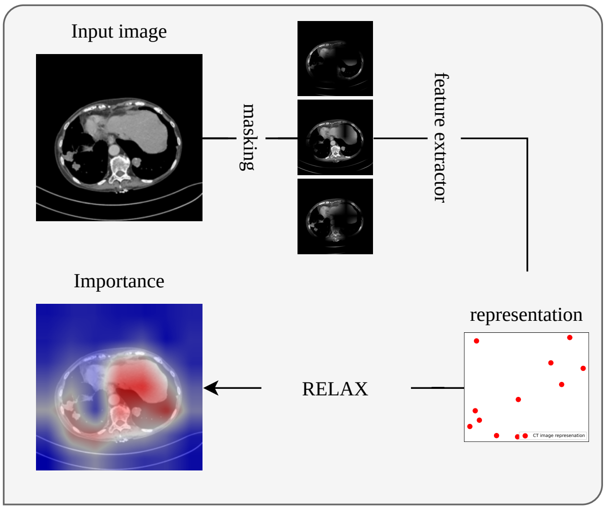


Fig. 5: Illustration of RELAX. A feature extractor produces a new representation of an input image, and RELAX determines what input features are important for the representation.

$h \times w$, where $h < H$ and $w < W$, are generated, where each element of the mask is sampled from a Bernoulli distribution with probability p. To produce smooth and spatially coherent masks, the small masks are upsampled using bilinear interpolation to the same size as the input image. Furthermore, the number of masks required to obtain reliable estimates of importance is an important hyperparameter. In this work, we generate 3000 masks to obtain an explanation for a single image, as suggested in a prior work (Wickstrøm et al., 2021).

5. Evaluation

We introduce the set of scores utilized to provided quantitative evaluation of our proposed framework.

5.1. Evaluating quality of CBIR

A standard approach to evaluate the quality of a CBIR system is to measure the class-consistency in the top retrieved images (Silva et al., 2020; Li et al., 2018b). One of the most common approaches to evaluate the class-consistency is through mean average precision (MAP):

$$\text{MAP} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{K}\sum_{k=1}^{K}\text{precision(k)}_n, \tag{5}$$

where N is the number of test samples (query images), K is the top-K retrieved images for each query image, and precision is defined as:

$$\text{precision(k)} = \frac{|\text{relevant images} \cap \text{k-retrieved images}|}{|\text{k-retrieved images}|}. \tag{6}$$

MAP evaluates the precision of the retrieved images across several values of K, which makes it robust towards fluctuations among the top retrieved images.

5.2. Evaluating quality of representations

The most widespread approach for evaluating the representation produced by a self-supervised learning framework is to train a simple classifier on the learned representations (Chen et al., 2020; Caron et al., 2020; He et al., 2020). The motivation for this, is that a simple classifier is highly dependent on the representation it is given in order to perform the desired task. In this work, we follow recent studies that use a k-nearest neighbors (KNN) classifier (Caron et al., 2021, 2020) to evaluate the quality of the representation. We opt for a KNN classifier over a linear classifier as it does not require any training, which can lead to ambiguities in the results (Kolesnikov et al., 2019), and has minimal hyperparameters to tune.

5.3. Evaluating the quality of explanations

Great improvements have been made in the field of XAI over the last couple of years. In contrast, the field of evaluation for explanations is still under active development (Doshi-Velez and Kim, 2017). However, recent advances have introduced new methods for providing quantitative evaluation of explanations. In this work, we use the relevance rank accuracy score (RR) (Arras et al., 2022). RR measures how many of the top-M relevant pixels lies within the ground truth segmentation mask. It can be considered a proxy for how well the explanation agrees with a human explanation for a given images. Let R_M denote the M most relevant pixels in an explanation, and S the segmentation mask for the liver. RR can then be defined as:

$$\text{RR} = \frac{1}{N}\sum_{n}^{N}\frac{|R_M(n) \cap S(n)|}{|S(n)|}. \tag{7}$$

The RR is computed using the Quantus toolbox (Hedström et al., 2022).

6. Data

In this section, we present the data used to evaluated our proposed framework.

6.1. Decathlon data

The medical segmentation decathlon is a biomedical image analysis challenge where several tasks and modalities are considered (Antonelli et al., 2021). One of the datasets in this challenge is a CT liver dataset acquired from the IRCAD Hopitaux Universitaires and consists of 201 contrast-ehanced CT liver images from patietns with mostly cancers and metastaic liver disease. However, we exclude 70 of these images as they do not include label information. Using every slide from each volume is computationally intractable. Therefore, we construct a slice-wise dataset as follows. From each volume, we sample 5 slices with no liver and 5 slices with liver. We construct the training set from the first 100 volumes and the test set from the remaining 31 volumes. This results in a balanced dataset with 1310 training images and 310 test images.

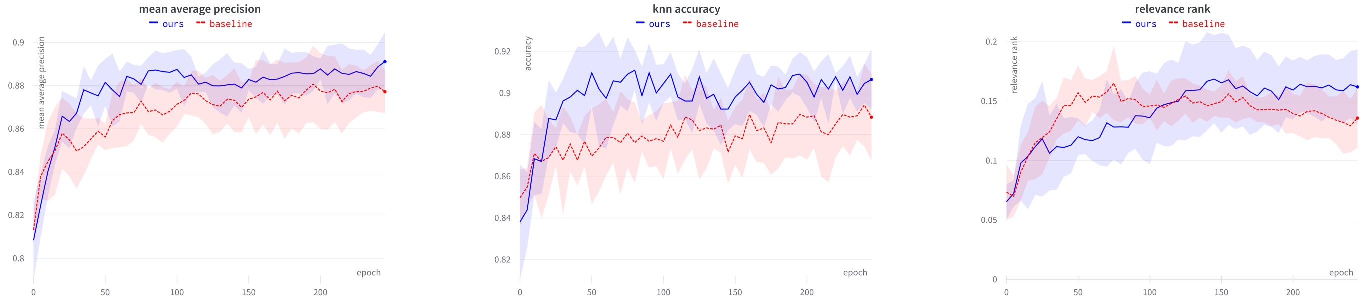


Fig. 6: From left to right, mean average precision, knn accuracy, and relevance rank scores versus epochs across 5 training runs on the test images from the Decathlon dataset The plot show how performance increase with training time, and that the proposed framework learns faster with better results.

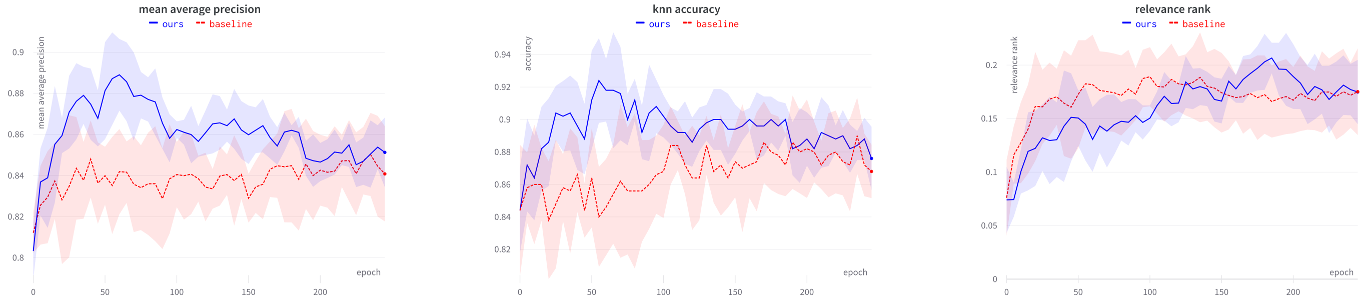


Fig. 7: From left to right, mean average precision, knn accuracy, and relevance rank scores versus epochs across 5 training runs on the test images from the UNN dataset. The plot show how performance increase with training time, and that the proposed framework learns faster with better results.

6.2. UNN data

The UNN dataset is from an extensive database of CT scans from The University Hospital of North Norway (UNN). It is under development through a close collaboration between UiT, The Arctic University of Norway, and UNN. The database contains CT volumes of 376 patients surgically treated for rectal cancer from 2006 to 2011 in North Norway. The examinations were conducted for diagnostic and routine follow-up purposes. The full dataset consists of CT with coronal, sagittal, and axial slices of mainly the thorax, abdomen, and pelvis. Examinations were conducted with different scanners and protocols at eight different hospitals in North Norway in the period 2005 to 2020.

From the full dataset a subset of 3347 axial volumes from 368 patients was selected based on descriptive keywords and DICOM metadata to limit it to contain mostly volumes of the liver and abdomen. This subset is similar to the CT liver partition of the medical segmentation decathlon dataset in terms of image resolutions and contents, but more diverse in terms of image quality, contrast enhancement levels, and artifacts because it is only curated using keywords and metadata, and not by manual assessment.

From the UNN subset 10 randomly selected volumes from 10 different and randomly selected patients with liver tumors were manually labeled with segmentation masks of the liver and metastatic regions by a clinician (co-author K.R.) to be used in our study. In addition, two volumes from a patient that had been treated with liver surgery to remove a metastatic liver segment were included. One volume was before the surgery, and one after the surgery. The study of these pre- and post-operative images is conducted as a use-case of cross-examination CBIR.

7. Experiments

We present the results of the experimental evaluation of our proposed framework. All models were trained with a batch size of 32 and for 250 epochs. Optimization was carried out using stochastic gradient descent with momentum=0.9, weight decay=0.0001, and learning rate=0.05 * batch size / 256, as used in the SimSiam framework (Chen and He, 2021). As in previous works (Chen et al., 2020; Chen and He, 2021), a ResNet50 (He et al., 2016) was used as the feature extractor, with the output of the average pooling layer as the final representation. For both the KNN classifier and the MAP we set K=5. Code is available at `https://github.com/Wickstrom/clinical-self-supervised-CBIR-ct-liver.git`.

7.1. Quantitative results

Table 1 and 2 presents the MAP, accuracy of a 5NN classifier, and the RR on the test data from the Decathlon and UNN datasets. The results show that the proposed framework outperforms the standard self-supervised approach across most scores. Furthermore, self-supervised learning greatly improves upon simply using the feature extractor trained on the Imagenet dataset. Also, the improvements are transferable across datasets, as the feature extractors trained on the Decathlon data also leads to improved performance in the UNN data.

Figure 6 and 7 presents the evolution of MAP, accuracy of a 5NN classifier, and the RR on the test data from the Decathlon and UNN datasets across training. The plots highlight how the scores improve as training progresses and stabilizes. However, an interesting observation is that the MAP and KNN accuracy achieves its highest value earlier in training on the UNN dataset.

Table 1: Mean and std of mean average precision, knn accuracy and relevance rank score across 5 training runs on the test images from the Decathlon dataset. Results show that the proposed framework outperforms the baselines. Bold numbers indicates the highest performing model.

pretraining	MAP	ACC	RR
IN	79.4	80.3	5.00
IN + SS (baseline)	87.7 ± 1.0	88.8 ± 2.0	13.6 ± 2.5
IN + SS (ours)	**89.1** ± **1.3**	**90.6** ± **1.4**	**16.2** ± **3.1**

Table 2: Mean and std of mean average precision, knn accuracy and relevance rank score across 5 training runs on test images from the UNN dataset. Results show that the proposed framework outperforms the baselines. Bold numbers indicates the highest performing model.

pretraining	MAP	ACC	RR
IN	80.7	83.0	4.34
IN + SS (baseline)	84.1 ± 2.3	86.8 ± 1.6	**17.5** ± **4.0**
IN + SS (ours)	**85.1** ± **1.7**	**87.6** ± **1.9**	**17.5** ± **3.0**

7.2. *Explaining representations - qualitative results*

The relevance rank scores in Table 1 and 2 show that the proposed framework utilizes liver features in the image to a larger degree than the baseline approaches. However, the scores are far from perfect, which means that other parts of the image are also being used. Also, the feature extractor that is only trained on the Imagenet dataset has a very low relevance rank score, meaning that it is putting little attention on the liver. All of these observations can be investigated through XAI. In this section, we illuminate these observations through a new explainability analysis for CBIR by leveraging the RELAX framework that was described in Section 4.1. We show 4 qualitative examples, where the first example shows explanations for the feature extractor trained using Imagenet, and the remaining three examples shows explanations for the feature extractor trained using the proposed framework. In all examples, we show a query from the test set and the 5 retrieved images by CBIR system. Additionally, we show the explanation for the query and retrieved images. The explanation show which features in the input are the most important for the representation of the image, where important pixels are highlighted in red and non-important pixels in blue.

Example 1: the feature extractor pretrained on Imagenet focuses on hard edges such as the spine. Figure 8 displays an example where 2 of the 5 the retrieved images do not contain parts of the liver. When inspecting the explanations, it is clear that the feature extractor is not focusing on the liver, but rather on the tailbone. We hypothesize that since the feature extractor has never been presented with CT images, it utilizes prominent features with hard edges such as spine, as opposed to organs with softer boundaries. The behaviour discovered in this example is important, as it might also result in unexpected or poor retrievals for other queries.

Example 2: the feature extractor trained using the proposed framework focuses on liver features. Figure 9 shows an example where all the retrieved images contain liver. Additionally, it is evident that the feature extractor is putting more emphasis on the liver for all the images, which illustrates how the proposed self-supervised framework has enabled the feature extractor to focus on clinically relevant features.

Example 3: the feature extractor trained using the proposed framework uses features from organs that often cooccur with the liver. Figure 10 displays an example where CBIR system retrieves 5 images that contain the liver, but where the explainability analysis shows that it not focusing on part of the images where the liver is present. Instead, it puts attention on the kidneys, which are quite prominent in all images. The kidneys often occur together with the liver in many CT images, and it also has similar pixel intensities as the liver (in terms of Houndsfield units). Therefore, it is not surprising that the feature extractor has learned to utilize both liver and kidney features, which also explains the behaviour in this example. Such insights would not be obtainable without conducting the explainability analysis.

Example 4: the feature extractor trained using the proposed framework focuses on liver features, also for images from a different dataset. Lastly, Figure 11 shows and example from the UNN dataset. This example illustrates that also on this new and unseen dataset, the feature extractor is basing the representation of these images features associated with the liver.

7.3. *Case study: cross-examination CBIR*

A typical scenario in clinical practice is comparing a newly conducted examination with one ore more previous examinations. For instance, one might want to compare a particular slice from the new examination with a selection of slices from one or several previous examinations. Such a comparison can help physicians understand how a diseases has progress since the previous examination, such as the development of liver metastasis. But when conducting such a comparison, the physician must manually inspect the new examination, and potentially several previous examinations. The CT scans are often taken with different settings across examinations, and it is therefore not possible to simply select the same slice from different examinations, as this can image completely different parts of the patient. A precise and reliable CBIR system could make such a cross-examination more efficient, by automating the retrieval process for the physician.

A typical scenario in clinical practice is comparing a newly conducted examination with one or more previous examinations. Such as development, progress, or effect of treatment of liver metastasis, the status of liver cirrhosis, auto-immune diseases in the liver, or any morphological or anatomical changes in the liver over the course. One might want to compare a region of interest in the slice from the current examination with

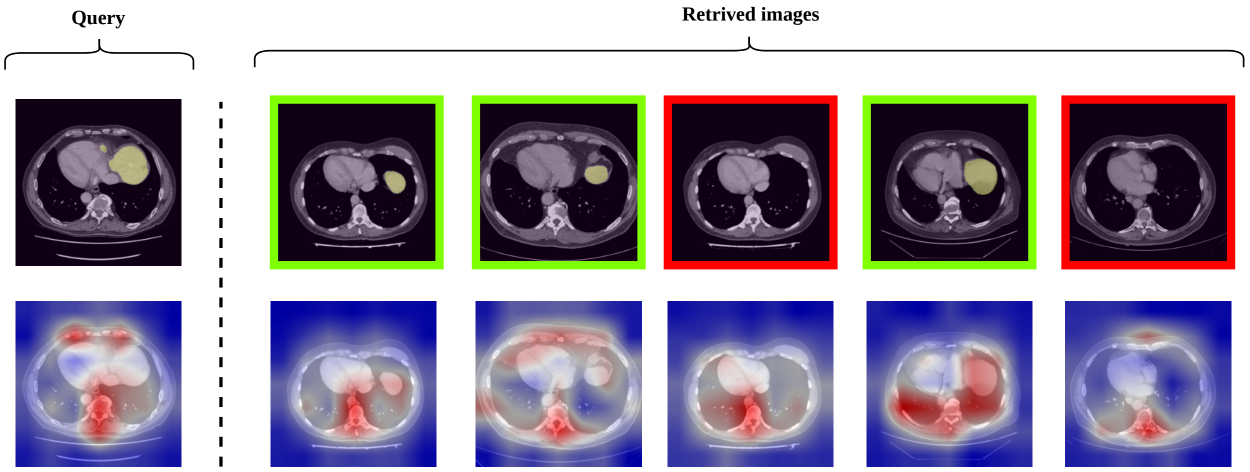


Fig. 8: **Example (1):** CBIR example from Decathlon dataset with feature extractor pretrained on Imagenet dataset. Top row shows, from left to right, the query and the top 5 retrieved images. Bottom row shows the important features for the representation of each image, with important features in red and less important in blue. Some of the retrieved images do not contain the liver, and the explainability analysis shows that the feature extractor is focusing on the spine and rib cage instead of the organs. This information is important to understand why non-relevant images are retrieved, and would not be available without the explainability analysis.

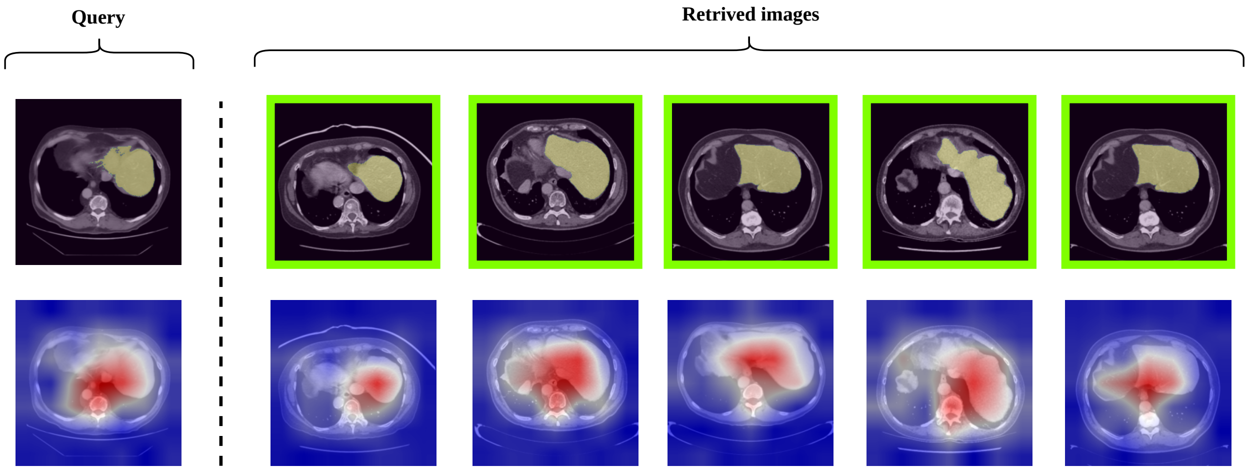


Fig. 9: **Example (2):** CBIR example from Decathlon dataset with feature extractor pretrained using the proposed self-supervised framework. Top row shows, from left to right, the query and the top 5 retrieved images. Bottom row shows the important features for the representation of each image, with important features in red and less important in blue. All retrieved images contain the liver, and the explainability analysis shows that the feature extractor is focusing on the liver.

previously conducted examinations. The comparison result will be applicable for evaluating the disease over the long course. In routine clinical practice, the selection of slices from the previous examinations is made manually to achieve the comparison, which is very time-consuming. CT scans taken over time are often conducted in an inconsistent sequence and contain unlike body regions in the same array of slices; therefore, selecting the identical arrays of slices from the different examinations is inaccurate. A precise and transparent CBIR system could make a cross-examination more efficient through the automatic retrieval process.

Figure 12 displays an example of such a cross-examination. The query is selected from a recent examination, and the retrieved images are from the previous examination of the same patient. This patient is from the UNN dataset and was selected since liver metastasis has been developed between the two examinations. The query was selected by an experienced physician (co-author K.R.), which also selected five images to examine from the previous examination. Ideally, the CBIR system should align well with the image selected by the physician. In this example, the CBIR system produces a successful retrieval, as it identifies the same images as the physician. However, an interesting observation is that the CBIR retrieved are not sorted in the same manner as the physician's retrievals. Probably this deviation is due to the CBIR system being trained on single slices without a sense of spatial coherence. Future works could address this by incorporating neighboring samples as positive pairs in the self-supervised training.

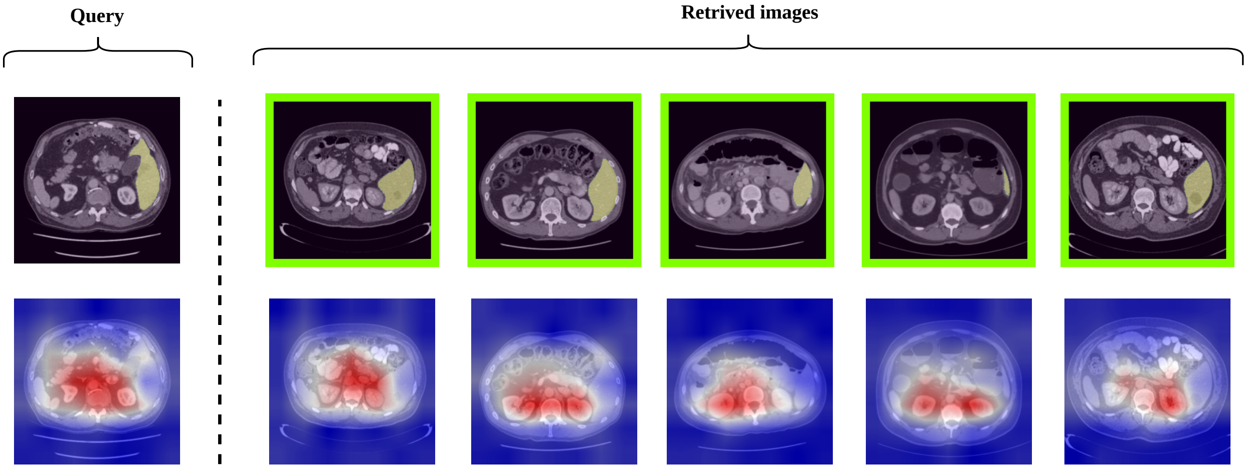


Fig. 10: **Example (3):** CBIR example from Decathlon dataset with feature extractor pretrained using the proposed self-supervised framework. Top row shows, from left to right, the query and the top 5 retrieved images. Bottom row shows the important features for the representation of each image, with important features in red and less important in blue. All retrieved images contain the liver, but the explainability analysis reveals that the focus is on the kidneys, not the liver.

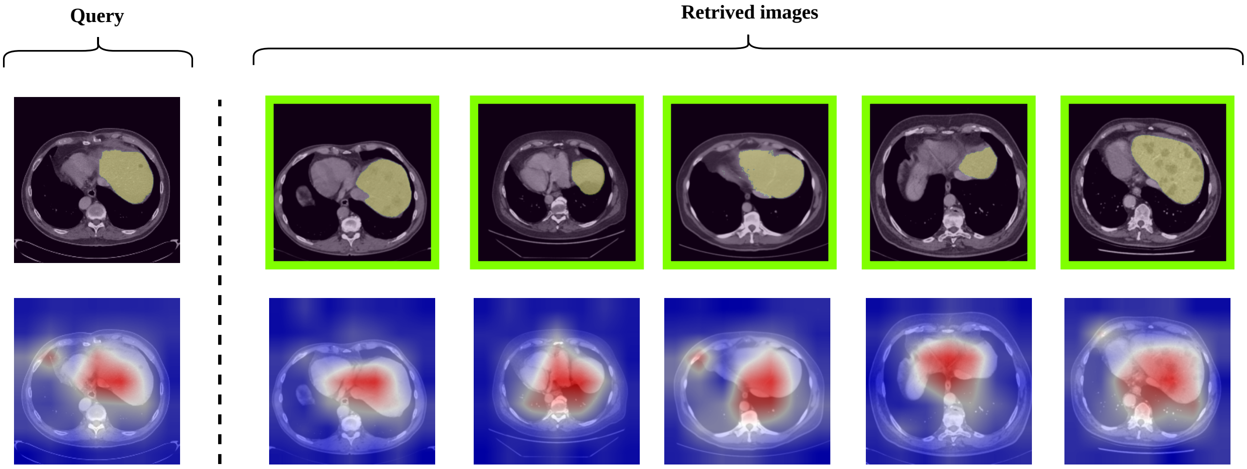


Fig. 11: **Example (4):** CBIR example from UNN dataset with feature extractor pretrained using the proposed self-supervised framework. Top row shows, from left to right, the query and the top 5 retrieved images. Bottom row shows the important features for the representation of each image, with important features in red and less important in blue. All retrieved images contain the liver and the feature extractor is focusing on the liver, which illustrates that the feature extractor trained on the Decathlon dataset transfers well to the UNN dataset.

8. Conclusion

We propose a clinically motivated self-supervised framework for CBIR of CT liver images. Our proposed framework exploits the properties of the liver to learn more clinically relevant features, which results in show leads to improved performance. Moreover, we leverage the RELAX framework to provide the first representation learning explainability analysis in the context of CBIR of CT liver images. Our analysis provides new insights into the feature extraction process and shows how self-supervised learning can provide feature extractors that extract more clinically relevant features compared to feature extractors trained on non-CT liver images. Our experimental evaluation also shows how the proposed framework generalizes to new datasets, and we present a clinically relevant user study. In future works, we intend to investigate how the proposed approach can be extended to extract features specific to other organs based on clipping strategies catered specifically to the desired organ. We believe that the proposed framework can play an essential role in constructing reliable CBIR that can effectively utilize unlabeled data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by The Research Council of Norway (RCN), through its Centre for Research-based Innovation

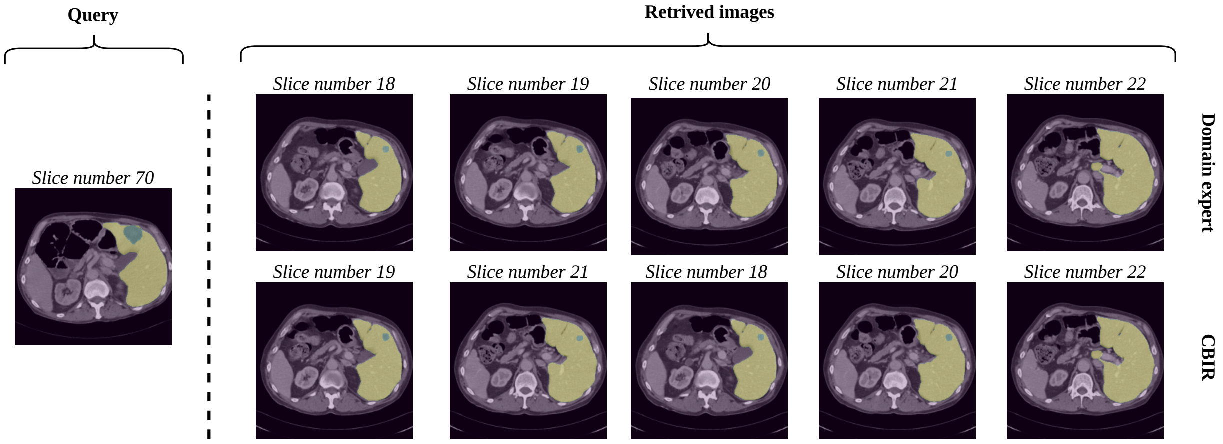


Fig. 12: An example of cross-examination CBIR. The query is from a recent examination, and the retrieved images are from a prior examination from the same patient. The goal of such a study is to investigate the development of liver metastasis. The query and retrieved images in the top row are selected by an experienced physician, and the bottom row are the retrieved images from the CBIR system. The CBIR successfully retrieves the same images as the physician, but lacks the spatial coherence to order the retrieved images.

funding scheme [grant number 309439] and Consortium Partners; RCN FRIPRO [grant number 315029]; RCN IKTPLUSS [grant number 303514]; and the UiT Thematic Initiative.

References

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., AnnetteKopp-Schneider, Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., van Ginneken, B., Bilello, M., Bilic, P., Christ, P.F., Do, R.K.G., Gollub, M.J., Heckers, S.H., Huisman, H., Jarnagin, W.R., McHugo, M.K., Napel, S., Pernicka, J.S.G., Rhode, K., Tobon-Gomez, C., Vorontsov, E., Huisman, H., Meakin, J.A., Ourselin, S., Wiesenfarth, M., Arbelaez, P., Bae, B., Chen, S., Daza, L., Feng, J., He, B., Isensee, F., Ji, Y., Jia, F., Kim, N., Kim, I., Merhof, D., Pai, A., Park, B., Perslev, M., Rezaiifar, R., Rippel, O., Sarasua, I., Shen, W., Son, J., Wachinger, C., Wang, L., Wang, Y., Xia, Y., Xu, D., Xu, Z., Zheng, Y., Simpson, A.L., Maier-Hein, L., Cardoso, M.J., 2021. The medical segmentation decathlon. `arXiv:2106.05735`.

Arras, L., Osman, A., Samek, W., 2022. Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. Information Fusion 81, 14–40. doi:`https://doi.org/10.1016/j.inffus.2021.11.008`.

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., Natarajan, V., Norouzi, M., 2021. Big self-supervised models advance medical image classification, in: International Conference on Computer Vision, pp. 3478–3488.

Ballerini, L., Li, X., Fisher, R.B., Rees, J., 2010. A query-by-example content-based image retrieval system of non-melanoma skin lesions, in: Medical Content-Based Retrieval for Clinical Decision Support. Springer Berlin Heidelberg, pp. 31–38. doi:`10.1007/978-3-642-11769-5_3`.

Barata, C., Santiago, C., 2021. Improving the explainability of skin cancer diagnosis using cbir, in: Medical Image Computing and Computer Assisted Intervention, Springer International Publishing. pp. 550–559.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language models are few-shot learners, in: Advances in Neural Information Processing Systems, pp. 1877–1901.

Caron, M., Bojanowski, P., Joulin, A., Douze, M., 2018. Deep clustering for unsupervised learning of visual features, in: European Conference on Computer Vision, pp. 132–149.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A., 2020. Unsupervised learning of visual features by contrasting cluster assignments, in: Neural Information Processing Systems, pp. 9912–9924.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers, in: International Conference on Computer Vision, pp. –.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, pp. 1597–1607.

Chen, X., He, K., 2021. Exploring simple siamese representation learning, in: Computer Vision and Pattern Recognition, pp. 15750–15758.

Chi, Y., Zhou, J., Venkatesh, S.K., Tian, Q., Liu, J., 2013. Content-based image retrieval of multiphase ct images for focal liver lesion characterization. Medical Physics 40, 103502.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186.

Dong, B., Collins, R., Hoogs, A., 2019. Explainability for content-based image retrieval, in: Computer Vision and Pattern Recognition Workshops, pp. 95–98.

Dong, N., Kampffmeyer, M., Voiculescu, I., 2021. Self-supervised multi-task representation learning for sequential medical images, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer. pp. 779–794.

Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning. `arXiv:1702.08608`.

Franceschi, J.Y., Dieuleveut, A., Jaggi, M., 2019. Unsupervised scalable representation learning for multivariate time series, in: Neural Information Processing Systems, p. 4650–4661.

Gautam, S., Höhne, M.M.C., Hansen, S., Robert Jenssen, M.K., 2022. Demonstrating the risk of imbalanced datasets in chest x-ray image-based diagnostics by prototypical relevance propagation, in: International Symposium on Biomedical Imaging, pp. –.

Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.Á., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M., 2020. Bootstrap your own latent - a new approach to self-supervised learning, in: Neural Information Processing Systems, pp. 21271–21284.

Gu, R., Wang, G., Song, T., Huang, R., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S., 2021. Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE Transactions on Medical Imaging 40, 699–711.

Hansen, S., Gautam, S., Jenssen, R., Kampffmeyer, M., 2022. Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels. Medical Image Analysis 78, 102385.

Haq, N.F., Moradi, M., Wang, Z.J., 2021. A deep community based approach for large scale content based x-ray image retrieval. Medical Image Analysis 68, 101847.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning, in: Computer Vision and Pattern Recognition, pp. 9726–9735.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: 2016 CVPR, pp. 770–778. doi:10.1109/CVPR.2016.90.

Hedström, A., Weber, L., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., Höhne, M.M.C., 2022. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations. arXiv:2202.06861.

Hu, B., Vasu, B., Hoogs, A., 2022. X-mir: Explainable medical image retrieval, in: Winter Conference on Applications of Computer Vision (WACV), pp. 1544–1554.

Jiang, M., Zhang, S., Metaxas, D.N., 2014. Detection of mammographic masses by content-based image retrieval, in: Machine Learning in Medical Imaging, pp. 33–41.

Jing, L., Vincent, P., LeCun, Y., Tian, Y., 2022. Understanding dimensional collapse in contrastive self-supervised learning, in: International Conference on Learning Representations, pp. –.

Khakzar, A., Zhang, Y., Mansour, W., Cai, Y., Li, Y., Zhang, Y., Kim, S.T., Navab, N., 2021. Explaining covid-19 and thoracic pathology model predictions by identifying informative input features, in: Medical Image Computing and Computer Assisted Intervention, pp. 391–401.

Kolesnikov, A., Zhai, X., Beyer, L., 2019. Revisiting self-supervised visual representation learning, in: IEEE Computer Vision and Pattern Recognition, pp. –.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A., 2018a. H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Transactions on Medical Imaging , 2663–2674.

Li, Z., Zhang, X., Müller, H., Zhang, S., 2018b. Large-scale retrieval for medical image analytics: A comprehensive review. Medical Image Analysis 43, 66–84.

Mirasadi, M.S., Foruzan, A.H., 2019. Content-based medical image retrieval of CT images of liver lesions using manifold learning. International Journal of Multimedia Information Retrieval 8, 233–240.

Monowar, M.M., Hamid, M.A., Ohi, A.Q., Alassafi, M.O., Mridha, M.F., 2022. Autoret: A self-supervised spatial recurrent network for content-based image retrieval. Sensors .

Peng, T., Boxberg, M., Weichert, W., Navab, N., Marr, C., 2019. Multi-task learning of a deep k-nearest neighbour network for histopathological image classification and retrieval, in: Lecture Notes in Computer Science, pp. 676–684.

Petsiuk, V., Das, A., Saenko, K., 2018. Rise: Randomized input sampling for explanation of black-box models, in: Proceedings of the British Machine Vision Conference, p. 151.

Plummer, B.A., Vasileva, M.I., Petsiuk, V., Saenko, K., Forsyth, D., 2020. Why do these match? explaining the behavior of image similarity models, in: European Conference on Computer Vision, pp. 652–669.

Quellec, G., Al Hajj, H., Lamard, M., Conze, P.H., Massin, P., Cochener, B., 2021. Explain: Explanatory artificial intelligence for diabetic retinopathy diagnosis. Medical Image Analysis 72, 102118.

Ramalhinho, J., Tregidgo, H.F.J., Gurusamy, K., Hawkes, D.J., Davidson, B., Clarkson, M.J., 2021. Registration of untracked 2d laparoscopic ultrasound to ct images of the liver using multi-labelled content-based image retrieval. IEEE Transactions on Medical Imaging 40, 1042–1054.

Schulz, K., Sixt, L., Tombari, F., Landgraf, T., 2020. Restricting the flow: Information bottlenecks for attribution, in: International Conference on Learning Representations, pp. –.

Silva, W., Poellinger, A., Cardoso, J.S., Reyes, M., 2020. Interpretability-guided content-based medical image retrieval, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Springer-Verlag, Berlin, Heidelberg, p. 305–314. doi:10.1007/978-3-030-59710-8_30.

Siradjuddin, I.A., Wardana, W.A., Sophan, M.K., 2019. Feature extraction using self-supervised convolutional autoencoder for content based image retrieval, in: International Conference on Informatics and Computational Sciences, pp. 1–5.

Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M., 2015. Striving for simplicity: The all convolutional net, in: International Conference on Learning Representations Workshop, pp. –.

Tian, Y., Chen, X., Ganguli, S., 2021. Understanding self-supervised learning dynamics without contrastive pairs, in: International Conference on Machine Learning, pp. 10268–10278.

Tisch, C., Brencicova, E., Schwendener, N., Lombardo, P., Jackowski, C., Zech, W.D., 2019. Hounsfield unit values of liver pathologies in unenhanced postmortem computed tomography. International Journal of Legal Medicine , 1861–1867.

Truong, T., Mohammadi, S., Lenga, M., 2021. How transferable are self-supervised features in medical image classification tasks?, in: Proceedings of Machine Learning for Health, pp. 54–74.

Vasilakakis, M., Sovatzidi, G., Iakovidis, D.K., 2021. Explainable classification of weakly annotated wireless capsule endoscopy images based on a fuzzy bag-of-colour features model and brain storm optimization, in: Medical Image Computing and Computer Assisted Intervention, pp. 488–498.

Wickstrøm, K., Kampffmeyer, M., Jenssen, R., 2020. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. Medical Image Analysis 60, 101619.

Wickstrøm, K., Kampffmeyer, M., Øyvind Mikalsen, K., Jenssen, R., 2022. Mixing up contrastive learning: Self-supervised representation learning for time series. Pattern Recognition Letters 155, 54–61. doi:https://doi.org/10.1016/j.patrec.2022.02.007.

Wickstrøm, K.K., Trosten, D.J., Løkse, S., Øyvind Mikalsen, K., Kampffmeyer, M.C., Jenssen, R., 2021. RELAX: Representation learning explainability. arXiv:2112.10161.

Yoshinobu, Y., Iwamoto, Y., Han, X., Lin, L., Hu, H., Zhang, Q., Chen, Y.W., 2020. Deep learning method for content-based retrieval of focal liver lesions using multiphase contrast-enhanced computer tomography images, in: International Conference on Consumer Electronics, pp. 1–4.

Zhao, C., Cheng, H., Huo, Y., Zhuang, T., 2004. Liver ct-image retrieval based on gabor texture, in: International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 1491–1494.

Zheng, Y., Jiang, B., Shi, J., Zhang, H., Xie, F., 2019. Encoding histopathological WSIs using GNN for scalable diagnostically relevant regions retrieval, in: Lecture Notes in Computer Science, pp. 550–558.

Bibliography

[1] Ian Weissman. Ai is the new reality: the 4th healthcare revolution in medicine. https://healthmanagement.org/c/hospital/issuearticle/ai-is-the-new-reality-the-4th-healthcare-revolution-in-medicine, 2019. Accessed: 16.06.22.

[2] Carah Alyssa Figueroa et al. Priorities and challenges for health leadership and workforce management globally: a rapid review. *BMC Health Services Research*, 2019.

[3] Sayan Mukherjee et al. Estimating dataset size requirements for classifying DNA microarray data. *Journal of Computational Biology*, pages 119–142, 2003.

[4] Junghwan Cho et al. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy. *ArXiv*, 2015.

[5] Andre Esteva et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, pages 115–118, 2017.

[6] Gabriele Campanella et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, pages 1301–1309, 2019.

[7] Philipp Jurmeister, Klaus-Robert Müller, and Frederick Klauschen. Artificial intelligence: a solution for the lack of pathologists? *Der Pathologe*, pages 218–221, 2022.

[8] Samuel Kuttner et al. Cerebral blood flow measurements with 15 o-water PET using a non-invasive machine-learning-derived arterial input function. *Journal of Cerebral Blood Flow &; Metabolism*, pages 2229–2241, 2021.

[9] Riccardo Miotto et al. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, pages 1236–1246, May 2017.

[10] Kaiming He et al. Deep residual learning for image recognition. *Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[11] Ehteshami Bejnordi et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, pages 2199–2210, 2017.

[12] Jacob Devlin et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[13] Yuqi Si et al. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, pages 1297–1304, 2019.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *International Conference on Neural Information Processing Systems*, page 1097–1105, 2012.

[15] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information, 2017.

[16] Sana Tonekaboni et al. What clinicians want: Contextualizing explainable machine learning for clinical end use. In *Machine Learning for Healthcare*, pages 359–380, 2019.

[17] T. Ching et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, page 20170387, 2018.

[18] Jianxing He et al. The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, pages 30–36, 2019.

[19] Christopher J. Anders et al. Finding and removing clever hans: Using explanation methods to debug and improve deep models. *Information Fusion*, pages 261–295, 2022.

[20] Srishti Gautam et al. Demonstrating the risk of imbalanced datasets in chest x-ray image-based diagnostics by prototypical relevance propagation. In *International Symposium on Biomedical Imaging*, pages –, 2022.

[21] Wojciech Samek et al. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, pages 247–278, 2021.

[22] Andreas Holzinger et al. Explainable AI methods - a brief overview. In *xxAI - Beyond Explainable AI*, pages 13–38. Springer, Cham, 2022.

[23] Lars Kai Hansen and Laura Rieger. Interpretability in intelligent systems – a new concept? In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 41–49. Springer International Publishing, 2019.

[24] Kirill Bykov et al. How much can I trust you? - quantifying uncertainties in explaining neural networks. *CoRR*, 2020. URL `https://arxiv.org/abs/2006.09000`.

[25] Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. "Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations. In *Workshop on AI for Social Good*, 2019.

[26] Linus Ericsson et al. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, pages 42–62, 2022.

[27] Grégoire Montavon et al. Explaining the predictions of unsupervised learning models. In *xxAI - Beyond Explainable AI*, pages 117–138. Springer International Publishing, 2022.

[28] Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. Second opinion needed: communicating uncertainty in medical machine learning. *npj Digital Medicine*, 2021.

[29] Jakob Gawlikowski et al. A survey of uncertainty in deep neural networks. *ArXiv*, 2022.

[30] Ting Chen et al. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607, 2020.

[31] Mathilde Caron et al. Unsupervised learning of visual features by contrasting cluster assignments. In *International Conference on Neural Information Processing Systems*, pages 9912–9924, 2020.

[32] Holly Else. A guide to plan s: the open-access initiative shaking up

science publishing. *Nature*, 2021.

[33] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, pages 436–444, 2015.

[34] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1798–1828, 2013.

[35] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, pages 115–133, 1943.

[36] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, page 386, 1958.

[37] Bernard Widrow and Marcian E. Hoff. Associative storage and retrieval of digital information in networks of adaptive "neurons". In *Biological Prototypes and Synthetic Systems*, pages 160–160. Springer US, 1962.

[38] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, pages 85–117, 2015.

[39] Marvin Minsky and Seymour Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, 1969.

[40] S. Linnainmaa. The representation of the cumulative rounding error of an algorithm as a taylor expansion of the local rounding errors. Master's thesis, University of Helsinki, 1970.

[41] P. J. Werbos. Applications of advances in nonlinear sensitivity analysis. In *Proceedings of the IFIP Conference*, pages 762–770, 1981.

[42] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, pages 533–536, 1986.

[43] Jürgen Schmidhuber. Very deep learning since 1991 - fast and deep / recurrent neural networks. `https://people.idsia.ch/~juergen/deeplearning.html`, 2022. Accessed: 06.06.22.

[44] Dana H. Ballard. Modular learning in neural networks. In *National Conference on Artificial Intelligence*, page 279–284, 1987.

[45] Suzanna Becker and Geoffrey E. Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, pages 161–163, 1992.

[46] Suzanna Becker and Geoffrey E Hinton. Learning to make coherent predictions in domains with discontinuities. In *International Conference on Neural Information Processing Systems*, pages 372–379, 1992.

[47] Y. LeCun et al. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, pages 541–551, 1989.

[48] Pierre Baldi and Yves Chauvin. Neural networks for fingerprint recognition. *Neural Computation*, pages 402–418, 1993.

[49] Sepp Hochreiter. Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, Institut für Informatik, Lehrstuhl Prof. Brauer, Technische Universität München, 1991.

[50] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Conference on Learning Theory Workshop on Computational Learning Theory*, pages 144–152, 1992.

[51] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, pages 273–297, 1995.

[52] Tin Kam Ho. Random decision forests. In *nternational Conference on Document Analysis and Recognition*, pages 278–282 vol.1, 1995.

[53] Leo Breiman. Random forests. *Machine Learning*, pages 5–32, 2001.

[54] Paul Smolensky. *Information Processing in Dynamical Systems: Foundations of Harmony Theory*, pages 194–281. MIT Press, 1987.

[55] Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, pages 233–243, 1991.

[56] Jia Deng et al. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[57] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

[58] Tom Brown et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901, 2020.

[59] Kevin Clark et al. Pre-training transformers as energy-based cloze models. In *EMNLP*, 2020. URL `https://www.aclweb.org/anthology/2020.emnlp-main.20.pdf`.

[60] Ting Chen and otheres. Big self-supervised models are strong semi-supervised learners. In *International Conference on Neural Information Processing Systems*, page 22243–22255, 2020.

[61] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer New York, 2000.

[62] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, pages 15849–15854, 2019.

[63] Suriya Gunasekar et al. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages –, 2017.

[64] Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pages 1772–1798, 2019.

[65] Suriya Gunasekar et al. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, volume 31, pages –, 2018.

[66] Daniel Soudry et al. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, pages 1–57, 2018.

[67] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2011.

[68] Jimmy Ba Diederik P. Kingma. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[69] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks, 2017. URL `http://arxiv.org/abs/1708.03888`.

[70] Jia Deng et al. Imagenet: A large-scale hierarchical image database. In

Conference on Computer Vision and Pattern Recognition, pages 248–255, 2009.

[71] Nitish Srivastava et al. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, pages 1929–1958, 2014.

[72] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016.

[73] Kaiming He et al. Momentum contrast for unsupervised visual representation learning. In *Computer Vision and Pattern Recognition*, pages 9726–9735, 2020.

[74] Loris Nanni, Stefano Ghidoni, and Sheryl Brahnam. Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recognition*, pages 158–172, 2017.

[75] Mark Everingham et al. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, pages 303–338, 2009.

[76] William R. Swartout. Explaining and justifying expert consulting programs. In *Computers and Medicine*, pages 254–271. Springer, 1985.

[77] A. Carlisle Scott et al. Explanation capabilities of production-based consultation systems. *American Journal of Computational Linguistics*, pages 1–50, 1977.

[78] N.J.S. Morch et al. Visualization of neural networks using saliency maps. In *International Conference on Neural Networks*, pages 2085–2090, 1995.

[79] Jatinder Singh et al. Responsibility and machine learning: Part of a process. *SSRN Electronic Journal*, 2016.

[80] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2017.

[81] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. “why should I trust you?”: Explaining the predictions of any classifier. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 97–101, 2016.

[82] Zachary C. Lipton. The mythos of model interpretability. In *International*

Conference on Machine Learning Workshop on Human Interpretability in Machine Learning, pages –, 2016.

[83] Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, pages 199 – 231, 2001.

[84] Kiana Alikhademi et al. A review of predictive policing from the perspective of fairness. *Artificial Intelligence and Law*, pages 1–17, 2021.

[85] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, pages 50–57, 2017.

[86] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference*, 2018.

[87] Ramprasaath R. Selvaraju et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *International Conference on Computer Vision*, pages 618–626, 2017.

[88] Sebastian Bach et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE,* page e0130140, 2015.

[89] Julius Adebayo et al. Debugging tests for model explanations. In *Advances in Neural Information Processing Systems,* pages 700–712, 2020.

[90] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, pages 206–215, 2019.

[91] Chaofan Chen et al. This looks like that: Deep learning for interpretable image recognition. In *International Conference on Neural Information Processing Systems*, 2019.

[92] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks, 2015. URL `https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html`.

[93] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, pages 1189–1232, 2001.

[94] Christoph Molnar. *Interpretable Machine Learning*. Bookdown, 2 edition,

2022. URL https://christophm.github.io/interpretable-ml-book.

[95] J. Emmanuel Johnson et al. Kernel methods and their derivatives: Concept and perspectives for the earth system sciences. *PLOS ONE*, page e0235885, 2020.

[96] David Balduzzi et al. The shattered gradients problem: If resnets are the answer, then what is the question? In *International Conference on Machine Learning*, page 342–350, 2017.

[97] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net, 2015.

[98] K. Bykov et al. Noisegrad: Enhancing explanations by introducing stochasticity to model weights. *Conference on Artificial Intelligence*, pages – , 2022.

[99] Grégoire Montavon et al. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, pages 211–222, 2017.

[100] Pieter-Jan Kindermans et al. Learning how to explain neural networks: Patternnet and patternattribution, 2018.

[101] B. Zhou, A. Khosla, Lapedriza. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. *Computer Vision and Pattern Recognition*, 2016.

[102] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833, 2014.

[103] Karl Schulz et al. Restricting the flow: Information bottlenecks for attribution. In *International Conference on Learning Representations*, 2020.

[104] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *International Conference on Computer Vision*, pages 2950–2958, 2019.

[105] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *International Conference on Computer Vision*, pages 3449–3457, 2017.

[106] Stefan Kolek et al. A rate-distortion framework for explaining black-box

model decisions. In *xxAI - Beyond Explainable AI*, pages 91–115. Springer, Cham, 2022.

[107] Nicolas Papernot and Patrick D. McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *CoRR*, 2018. URL `http://arxiv.org/abs/1803.04765`.

[108] Been Kim et al. Interpretability beyond feature attribution:quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pages 2668–2677, 2018.

[109] Sana Tonekaboni, Shalmali Joshi, Kieran Campbell, David K Duvenaud, and Anna Goldenberg. What went wrong and when? instance-wise feature importance for time-series black-box models. In *Advances in Neural Information Processing Systems*, pages 799–809, 2020.

[110] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *SSRN Electronic Journal*, pages –, 2017.

[111] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, page 4768–4777, 2017.

[112] David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. In *International Conference on Machine Learning Workshop on Human Interpretability in Machine Learning*, pages –, 2018.

[113] Yaniv Ovadia et al. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *International Conference on Neural Information Processing Systems*, page 14003–14014, 2019.

[114] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural Safety*, pages 105–112, 2009.

[115] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *International Conference on Neural Information Processing Systems*, page 5580–5590, 2017.

[116] Charles Blundell et al. Weight uncertainty in neural networks. In *International Conference on International Conference on Machine Learning*, page 1613–1622, 2015.

[117] Mattias Teye, Hossein Azizpour, and Kevin Smith. Bayesian uncertainty

estimation for batch normalized deep networks. In *International Conference on Machine Learning*, pages 4907–4916, 2018.

[118] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[119] Adriano Azevedo-Filho and Ross D. Shachter. Laplace's method approximations for probabilistic inference in belief networks with continuous variables. In *Uncertainty Proceedings*, pages 28–36. Elsevier, 1994.

[120] Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=Skdvd2xAZ`.

[121] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Ensemble learning. In *The Elements of Statistical Learning*, pages 605–624. Springer, 2008.

[122] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *International Conference on Neural Information Processing Systems*, page 6405–6416, 2017.

[123] Divya Shanmugam et al. Better aggregation in test-time augmentation. In *International Conference on Computer Vision*, pages 1194–1203, 2021.

[124] Nikita Moshkov et al. Test-time augmentation for deep learning-based cell segmentation on microscopy images. *Scientific Reports*, 10, 2020.

[125] Guotai Wang et al. Automatic brain tumor segmentation based on cascaded convolutional neural networks with uncertainty estimation. *Frontiers in Computational Neuroscience*, 13, 2019.

[126] Mathilde Caron et al. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision*, pages 132–149, September 2018.

[127] Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. Unsupervised scalable representation learning for multivariate time series. In *International Conference on Neural Information Processing Systems*, page 4650–4661, 2019.

[128] Jane Bromley et al. Signature verification using a "siamese" time delay neural network. In *International Conference on Neural Information Processing Systems*, page 737–744, 1993.

[129] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *Conference on Computer Vision and Pattern Recognition*, pages 1735–1742, 2006.

[130] Alexey Dosovitskiy et al. Discriminative unsupervised feature learning with convolutional neural networks. In *International Conference on Neural Information Processing Systems*, page 766–774, 2014.

[131] Xiaolong Wang, Kaiming He, and Abhinav Gupta. Transitive invariance for self-supervised visual representation learning. In *International Conference on Computer Vision*, pages 1338–1347, 2017.

[132] Zhirong Wu, Yuanjun Xiong, S. Yu, et al. Unsupervised feature learning via non-parametric instance discrimination. *Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[133] Mathilde Caron et al. Unsupervised pre-training of image features on non-curated data. In *International Conference on Computer Vision*, pages 2959–2968, 2019.

[134] Alexander Kolesnikov et al. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[135] Mathilde Caron et al. Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision*, pages 9630–9640, 2021.

[136] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, pages –, 2022.

[137] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, volume 139, pages 10268–10278, 2021.

[138] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=S1v4N2l0-`.

[139] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer International Publishing, 2016.

[140] Jean-Bastien Grill et al. Bootstrap your own latent - a new approach to self-supervised learning. In *International Conference on Neural Information Processing Systems*, volume 33, pages 21271–21284, 2020.

[141] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Computer Vision and Pattern Recognition*, pages 15750–15758, June 2021.

[142] Wullianallur Raghupathi and Viju Raghupathi. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2014.

[143] Andreas D. Lauritzen et al. An artificial intelligence–based mammography screening protocol for breast cancer: Outcome and radiologist workload. *Radiology*, 2022.

[144] Marthe Larsen et al. Possible strategies for use of artificial intelligence in screen-reading of mammograms, based on retrospective data from 122, 969 screening examinations. *European Radiology*, 2022.

[145] Cowan Ho et al. A promising deep learning-assistive algorithm for histopathological screening of colorectal cancer. *Scientific Reports*, 2022.

[146] Cristina Soguero-Ruiz et al. Predicting colorectal surgical complications using heterogeneous clinical data and kernel methods. *Journal of Biomedical Informatics*, pages 87–96, June 2016.

[147] Kasper Jensen et al. Analysis of free text in electronic health records for identification of cancer patient trajectories. *Scientific Reports*, 2017.

[148] Jintae Kim et al. Comprehensive survey of recent drug discovery using deep learning. *International Journal of Molecular Sciences*, page 9983, 2021.

[149] René H. M. Raeven et al. Systems vaccinology and big data in the vaccine development chain. *Immunology*, pages 33–46, 2018.

[150] Ravi Aggarwal et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *npj Digital Medicine*, 2021.

[151] Nanqing Dong et al. Reinforced auto-zoom net: Towards accurate and fast breast cancer segmentation in whole-slide images. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 317–325. Springer International Publishing, 2018.

[152] Longxi Zhou et al. An interpretable deep learning workflow for discovering subvisual abnormalities in CT scans of COVID-19 inpatients and survivors. *Nature Machine Intelligence*, pages 494–503, 2022.

[153] Yikuan Li et al. BEHRT: Transformer for electronic health records. *Scientific Reports*, April 2020.

[154] Edward Choi et al. Gram: Graph-based attention model for healthcare representation learning. In *International Conference on Knowledge Discovery and Data Mining*, page 787–795, 2017.

[155] Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 2019.

[156] Bas H.M. van der Velden et al. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470, 2022.

[157] Ghassan Hamarneh Weina Jin, Xiaoxiao Li. Evaluating explainable ai on a multi-modal medical imaging task: Can existing algorithms fulfill clinical requirements? *Conference on Artificial Intelligence*, pages –, 2022.

[158] Simon M. Thomas et al. Interpretable deep learning systems for multi-class segmentation and classification of non-melanoma skin cancer. *Medical Image Analysis*, page 101915, 2021.

[159] Gustavo Carneiro et al. Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy. *Medical Image Analysis*, page 101653, 2020.

[160] Lisa Herzog et al. Integrating uncertainty in deep neural networks for MRI based stroke analysis. *Medical Image Analysis*, page 101790, 2020.

[161] Christian Leibig et al. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, 2017.

[162] Stine Hansen et al. Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels. *Medical Image*

Analysis, page 102385, 2022.

[163] Behzad Bozorgtabar, Dwarikanath Mahapatra, Guillaume Vray, and Jean-Philippe Thiran. SALAD: Self-supervised aggregation learning for anomaly detection on x-rays. In *Medical Image Computing and Computer Assisted Intervention*, pages 468–478. Springer, 2020.

[164] Pengshuai Yang et al. Self-supervised visual representation learning for histopathological images. In *Medical Image Computing and Computer Assisted Intervention*, pages 47–57. Springer, 2021.

[165] Nanqing Dong and Irina Voiculescu. Federated contrastive learning for decentralized unlabeled medical images. In *Medical Image Computing and Computer Assisted Intervention*, pages 378–387. Springer International Publishing, 2021.

[166] Z. Wang, W. Yan, and T. Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *International Joint Conference on Neural Networks*, pages 1578–1585, 2017.

[167] Gil Shomron et al. Post-training sparsity-aware quantization. In *Advances in Neural Information Processing Systems*, pages 17737–17748, 2021.

[168] Saeid Asgari Taghanaki et al. Infomask: Masked variational latent representation to localize chest disease. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 739–747, 2019.

[169] Andrey Zhmoginov, Ian Fischer, and Mark Sandler. Information-bottleneck approach to salient region discovery. In *International Conference on Machine Learning Workshop in Self-Supervised Learning*, pages –, 2019.

[170] Anna Hedström et al. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations, 2022.

[171] Srishti Gautam et al. This looks more like that: Enhancing self-explaining models by prototypical relevance propagation, 2021.

[172] Jiamei Sun, Sebastian Lapuschkin, Wojciech Samek, and Alexander Binder. Explain and improve: LRP-inference fine-tuning for image captioning models. *Information Fusion*, pages 233–246, 2022.

[173] Wilson Silva, Alexander Poellinger, Jaime S. Cardoso, and Mauricio Reyes.

Interpretability-guided content-based medical image retrieval. In *Medical Image Computing and Computer Assisted Intervention*, page 305–314. Springer-Verlag, 2020.