Performance of Data Enhancements and Training Optimization for Neural Network: A Polyp Detection Case Study

Fredrik Lund Henriksen*, Rune Jensen*, Håkon Kvale Stensland*, Dag Johansen†, Michael Riegler§, Pål Halvorsen‡§
*Simula Research Laboratory, Norway
†UiT - The Arctic University of Norway, Norway
‡SimulaMet, Norway
§Oslo Metropolitan University, Norway

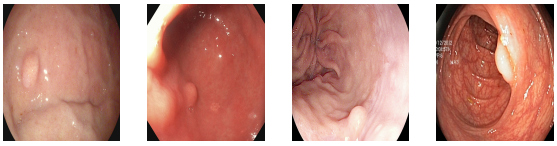
***Abstract*—Deep learning using neural networks is becoming more and more popular. It is frequently used in areas like video analysis, image retrieval, traffic forecast and speech recognition. In this respect, the learning and training process usually requires a lot of data. However, in many areas, data is scarce which is definitely the case in our medical application scenario, i.e., polyp detection in the gastrointestinal tract. Here, colorectal cancer is on the list of most common cancer types, and often, the cancer arises from benign, adenomatous polyps containing dysplastic cells. Detection and removal of polyps can therefore prevent the development of cancer. Due to high cost, time consumption, patient discomfort and in-accuracy of existing procedures, researchers have started to explore systems for automatic polyp detection to assist and automate current examination procedures. Following the current gained traction for neural networks, and the typical lack of medical data, we explore how data enhancements affect the training and evaluation of the networks in terms of polyp detection accuracy and particularly if it can be used to increase the detection rate. We also experiment with how various training techniques can be used to increase performance. Our experimental results show how data enhancement and training optimization can be used to increase different aspects of the performance, but we also point out mechanisms that have no and even a negative effect.**

***Keywords*-data enhancements; neural networks; disease detection;**

I. Introduction

Lately, the interest in deep learning using neural networks has exploded. Experiments in various areas analyzing and retrieving data are performed using this approach, including various area of multimedia data analysis. Here, neural networks mimic how a biological brain step-wise learns, where each layer (step) can learn different abstraction levels of the data using the input of previous layers until a final layer combines all abstractions in a stack-wise manner and provides the final result of the analysis. Current research proves that this is a promising approach. However, neural networks also come with several challenges where one is that training or learning is complicated. Usually, a vast amount of data is required to achieve an accurate result in the analysis, and in many scenarios, data availability is a challenge.

In this paper, we analyze the effect of enhancing data and optimizing training for neural networks in scenarios where data availability is a limitation for training. As a multimedia case study, we have selected a societal relevant scenario, where we build a system to automatically assist doctors in detecting polyps in videos and images from parts of the human gastrointestinal tract, i.e., the colon. Polyps are precursor lesions for colorectal cancer where the cancer often arises from benign, adenomatous polyps containing dysplastic cells. Colorectal cancer is the third most commonly diagnosed cancer in both men and women. In 2017, there will be an estimated 95.000 new cases of colon cancer in the US alone where about 27.000 men and 23.000 women will die due to the disease [1]. Thus, detection and removal of polyps therefore prevents the development of cancer, and our detection challenge is shown in figure 1, where the polyps must be detected and identified among images and video frames also containing normal mucosa and various other anatomical landmarks and abnormalities.



(a) Polyp (b) Polyp (c) Polyp (d) Polyp

Figure 1. Example of polyp images from the colon [2].

With respect to using neural networks for this type of abnormality detection, research have already proven that neural networks can be suitable, e.g., for similar problems detecting breast cancer [3], lung cancer [4], and for polyp detection in particular [5], [6], [7], [8], [9]. In this paper, we enhance our neural network-based EIR system (named after a Norse goddess associated with medical skills) with data enhancement methods and assess their effect on detection performance. In particular, we use the Tensorflow [10], [11] open source neural network library, and we "artificially" increase the dataset size making copies by changing rotation, brightness, contrast and reflection masking.

Using a real-world medical dataset [12], we have used various splits of the data in k-fold cross-validation experiments to make non-overlapping training and test sets and tested various combinations of the methods above. In summary,

for the GI polyp detection scenario, our experimental results show that rotation seems to be a good method method to extend the dataset size. Attempting to improve quality by reflection masking and higher contrasts had both positive and negative effects depending on the data splits. Overall, rotation is an interesting approach due to an increased F1-score, and the combination of rotation and contrast increases the positive recall.

The following paper is structured as following. First, we give a short overview of the background and related work. This is followed by describing our data enhancement system. After that, we describe the conducted experiments where we present our findings. Finally, we give conclusions and an outlook for possible future work.

II. Background and Related Work

Machine learning is the concept where computers gain the ability to learn without being explicitly programmed. It has evolved from artificial intelligence research, and has been one of the hottest topics among researchers in recent years [13]. It learns by making data driven decisions or predictions instead of following static instructions, and it alters its own understanding in an iterative manner by evaluating its current interpretation of the data against past knowledge, i.e., creating a new and improved understanding where the improvements are kept, and changes for the worse are discarded. After many iterations, it will have gained a general understanding of the concept.

A machine learning approach that has been reborn and lately gained a lot of interest is *neural networks* [14] which is a type of machine learning which loosely mimics how a biological brain learns, i.e., being able to learn general concepts from concrete examples. Neural networks are also used in the medical domain, e.g., for micro-calcification detection in mammogram images [15], detecting breast cancer [3], colonic polyp detection [5], [6], [7], [8], [9] and lung cancer [4]. To perform the analysis of the data, deep neural networks (deep learning) contain multiple network layers where each layer can learn different abstraction levels of the data using the input of previous layers. The information travels through different routes in the network depending on each layers understanding in the same way as a brain works using neurons. The route ends up in a terminal, which is the output of the final layer giving the estimation made by the network, i.e., an estimated understanding of what the data is or contains.

Despite the promising initial research, neural networks come with several challenges, and in the context of our research, neural networks require a high amount of training data of sufficient quality and with ground truth. This is especially hard in the medical field since collecting such data requires the time of experts. In order to address the challenge of limited data for training, one can increase the data using various enhancement steps where copies of the image data are added by for example rotation, changing brightness and contrast, masking reflections, etc. A question, however, is how such enhancements affect the detection accuracy of such systems. Here, we have tested different enhancement techniques in order to research which methods really improve polyp detection in particular, and we compare various training optimizations.

III. Enhanced Neural Network Training

In our polyp detection system, as for computer vision systems in general, the input data greatly affects the results. The quality and quantity of the input data can be increased by using data enhancement. A higher quality could make the polyps easier to detect, while a higher quantity gives the system more samples to learn from, i.e., both potentially improving the situation where the input data to the neural network is too small or too narrow resulting in overfitting the network. To enhance our neural network training process, we have created a 5-stage training pipeline as shown in figure 2 where we use TensorBox. One can artificially increase the size of a dataset by transforming the existing dataset in various ways, such as rotation, translation, scaling, flipping, shearing and stretching [16].

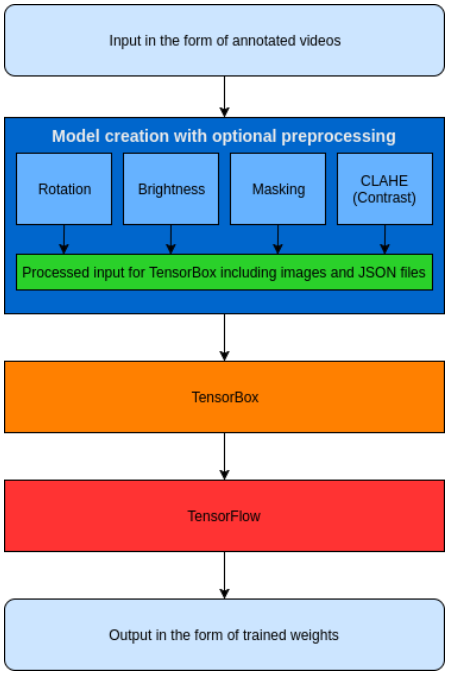
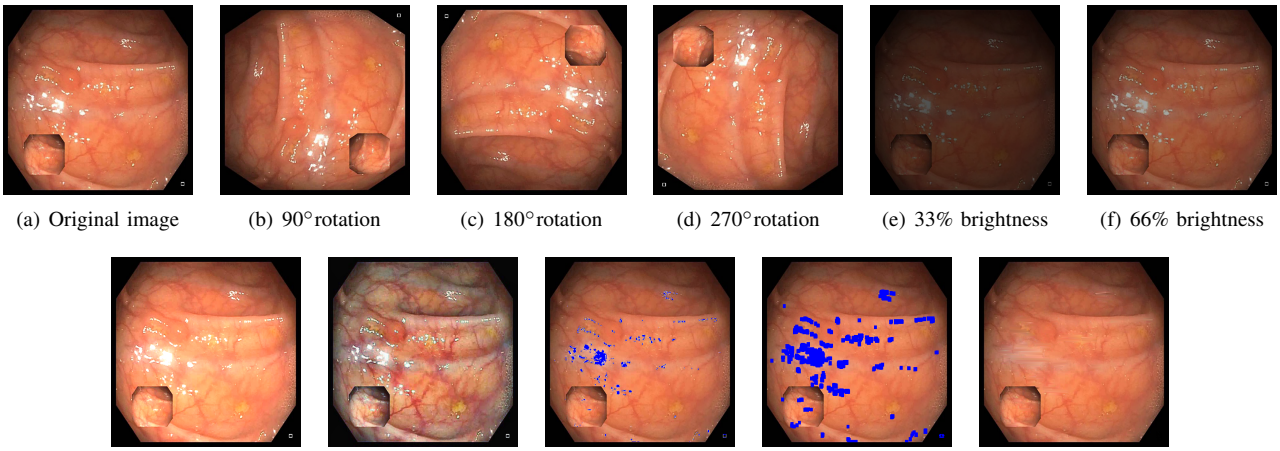


Figure 2. Training system

Polyps have no logical up or down as they can be found anywhere inside the colon. Thus, **rotation** of images may potentially improve the detection performance. In this respect, Amaral et al. [17] has performed experiments using rotated images and were able to increase the detection rate by between 8% and 42% using a very a low number of images as input, i.e., arguing that the main benefit of rotation is to increase the amount of input data. In our experiments,



(a) Original image (b) 90°rotation (c) 180°rotation (d) 270°rotation (e) 33% brightness (f) 66% brightness

(g) 133% brightness (h) Enhanced contrast (i) Marked bright spots (j) Padded bright spots (k) Filled bright spots

Figure 3. An example of the original image compared to the various additional enhancements: rotation, brightness, contrast and reflection maskings.

rotation of images are performed on each image that contains a polyp. As shown in figures 3(b)-3(d), the original image (figure 3(a)) is duplicated three times, with 90°, 180° and 270° rotated variants, in addition to the original.

Using different endoscopes, one can experience different types of lighting, i.e., **brightness** levels in different parts of the image. As with rotation, only images that contain a polyp and are part of the training data are brightness adjusted. The original image in figure 3(a) is again duplicated three times, with 33%, 66% and 133% brightness level variants, shown in figures 3(e)-3(g), respectively.

As can be observed in figure 1, it can be a challenge to distinguish the polyps from the surrounding areas in the GI tract. A possible way to improve polyp detection is to **enhance the contrast** in the images. Yadav et al. [18] were able to increase the number of detectable edges in images with heavy fog by enhancing the contrast. While we have no images with fog, contrast enhancement could be beneficial by increasing the detail level in low-contrast areas. In this respect, adaptive histogram equalization [19] is a technique to perform contrast enhancement. In contrast to ordinary histogram equalizations, the adaptive variant uses the neighboring regions to derive a transformation function. The benefit of this is that dark and light regions within the image are also sufficiently enhanced, since it adapts the function to local areas in the image. Furthermore, to deal with a possible over-amplification of noise in adaptive histogram equalization, contrast limited adaptive histogram equalization (CLAHE) [20] is an optimization limiting the amplification. It clips the histogram at a predefined limit, and distributes the clipped part among surrounding areas, preserving the clipped part while limiting the amplification. Using OpenCV, we implemented a CLAHE tool for each RGB channel in the image. At the end, the original image is replaced with the enhanced version, making updating file names unnecessary and avoids having duplicate images. An example of this can be seen in figure 3(h).

The light source, which is needed to capture video inside the colon, can potentially create sharp reflections (e.g., figure 3(a)) since the colon surface can be uneven and contain fluids. Zhou et al. [21] use these reflections for detection and polyp measurements. In our case, images both with and without polyps contain similar types of reflections. As such, it could be beneficial for the polyp detection to **remove the reflections**, letting the neural network focus on other features of the polyp. As shown in figures 3(i)-3(k), our tool to the masking of reflections consists of three steps; marking bright areas, padding marked areas and filling marked areas with surrounding colors. For each marked area, we find the color to the left and right, and color the pixels as a gradient color between the left and right color. If no valid color is found in either direction, we try the pixel above instead. The RGB limit is set to (240, 150, 150), and the padding radius is 5 pixels.

IV. Improving Dataset Size and Quality

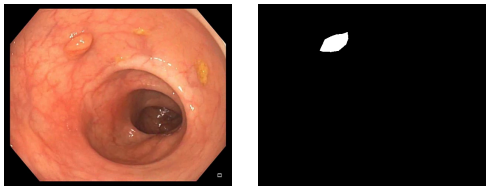
To evaluate our pre-processing tools, we have performed a large number of experiments using each method and their combinations, various numbers of training iterations and confidences. We used the ASU Mayo Clinic polyp dataset [22], which is an annotated dataset of polyp videos as training and evaluation data. It consists of 20 videos, 10 of which contain polyps and 10 that do not. An example can seen in figure 4 where a mask is used to show the location of a polyp. Furthermore, we have used a 5-fold cross validation [23] where we divided the 10 videos into 5 separate partitions, so in each cross validation, we use 8 videos for training and 2 for evaluation. The result from each cross validation is then averaged to produce a single estimation. This ensures that the results are not based on a single partitioning, which could be a deviation, but rather an average over all partitions. Each experiment is

Combination	Positive precision	Positive recall	Positive F1	Negative precision	Negative recall	Negative F1	**Weighted F1**	**Weighted Precision**	**Weighted Recall**
100k	28.50	29.08	27.52	96.19	96.76	96.46	**92.95**	**92.77**	**93.27**
200k	31.27	26.60	28.03	96.12	98.01	97.05	**93.51**	**92.80**	**94.36**
300k	31.50	26.16	27.73	96.10	98.15	97.11	**93.56**	**92.81**	**94.47**
400k	31.64	26.31	27.85	96.11	98.15	97.11	**93.57**	**92.83**	**94.47**
500k	31.60	26.29	27.85	96.11	98.14	97.10	**93.56**	**92.82**	**94.46**

Table I
RESULTS WITHOUT USING ANY ENHANCEMENTS (NP).

Combination	Positive precision	Positive recall	Positive F1	Negative precision	Negative recall	Negative F1	**Weighted F1**	**Weighted Precision**	**Weighted Recall**
NP	31.60	26.29	27.85	96.11	98.14	97.10	**93.56**	**92.82**	**94.46**
R	36.51	31.80	33.67	96.35	97.41	96.88	**93.62**	**93.22**	**94.06**
B	20.24	20.70	19.87	95.62	96.01	95.80	**91.85**	**91.73**	**92.05**
RB	25.09	27.66	25.79	96.17	96.60	96.38	**92.78**	**92.52**	**93.13**
M	24.74	22.26	22.85	96.00	97.76	96.86	**93.20**	**92.52**	**93.98**
C	27.26	24.62	25.14	96.15	97.60	96.86	**93.28**	**92.69**	**93.99**
MC	27.84	24.82	25.92	96.12	97.88	96.98	**93.45**	**92.74**	**94.22**
RBM	24.86	30.24	27.06	96.23	95.29	95.75	**92.25**	**92.56**	**92.01**
RBC	20.86	31.12	24.40	96.28	93.80	95.01	**91.43**	**92.40**	**90.70**
RBMC	19.42	25.00	21.05	95.98	95.20	95.57	**91.77**	**92.03**	**91.69**
RM	30.84	31.61	30.67	96.37	96.85	96.60	**93.24**	**93.01**	**93.56**
RC	30.80	37.79	33.44	96.63	95.93	96.27	**93.08**	**93.28**	**92.97**
RMC	24.97	29.99	25.95	96.24	95.73	95.97	**92.41**	**92.61**	**92.40**

Table II
SUMMARY OF ALL THE RESULTS FROM THE DIFFERENT DATA ENHANCEMENT METHODS, GIVEN 90% CONFIDENCE AND 500K TRAINING ITERATIONS



(a) Frame containing polyp. (b) Ground truth mask.

Figure 4. Polyp annotation example in the ASU Mayo dataset [22]

trained between 100k to 500k training iterations, where a training iteration is defined as one round of a feedback and adjustment loop. For the presented evaluation, a 90% confidence is used, which means that the system has to have a 90% or higher certainty that the image contains a polyp, before it is classified as positive. The different results are labeled according to method used using NP (Non-preprocessed), R (Rotation), B (Brightness variations), M (Masking reflections) and C (Contrast enhancement), and where combinations like RB means that rotation and brightness have been used. For further details about the setup, please see [24].

We started by using non-preprocessed data to see how the system performed with no data enhancement method applied, giving us a basis for comparison with the data enhancement methods. The results can be seen in table I. The system without any enhancements (NP) achieves a weighted recall of 94.46, a weighted precision of 92.82 and a weighted F1-score of 93.56. From the NP results, we can see that we get an increase in scores from 100k to 300k training iterations, but after 300k there are small variations, but very little to no gain. However, due to space restrictions, we therefore only present 500k for the remainder of this section. More details can again be found in [24] where we also show that iterations up to 1000k have negligible impacts.

Next, table II shows the results from the experiments, using the enhancements both individually and combined. The most promising combinations have been selected. We have tried rotation (R) and brightness (B) variations to increase the dataset, and we used masked reflections (M) and contrast enhancement (C) to improve the quality of the dataset. Additionally, we have experimented using combinations. The results show that rotation has an ability to improve positive recall, but also produces a slight decrease in negative recall. However, the improvement in positive recall is enough to offset the decrease in negative recall, resulting in rotation being the only data enhancement method able to beat the F1-score of NP. Brightness variations shows an inability to improve any aspect of the results, both when used independently and in combinations. It lowers both positive and negative recall, and thus affects weighted F1-score

Version	Positive precision	Positive recall	Positive F1	Negative precision	Negative recall	Negative F1	**Weighted F1**	**Weighted Precision**	**Weighted Recall**
Different types of networks									
Inception	14.52	86.67	24.87	98.04	56.62	71.78	**68.10**	**91.49**	**58.98**
Resnet	15.44	84.68	26.12	97.89	60.66	74.82	**71.00**	**91.43**	**62.44**
Default RNN	70.66	66.32	68.42	97.15	97.66	97.40	**95.13**	**95.07**	**95.20**
Comparing training techniques									
None	13.01	8.23	10.08	95.97	97.55	96.75	**93.05**	**92.43**	**93.74**
LSTM	18.61	22.42	20.34	96.51	95.63	96.07	**92.84**	**93.19**	**92.51**
Rezoom	16.24	12.26	13.97	96.13	97.18	96.65	**93.12**	**92.72**	**93.56**
Both	17.52	17.10	17.31	96.31	96.41	96.36	**92.99**	**92.95**	**93.02**
Comparing optimizers									
Rezoom + SGD	54.70	29.28	36.58	96.11	99.27	97.66	**94.41**	**94.02**	**95.49**
LSTM + SDG	00.00	00.00	00.00	94.74	100.0	97.29	**92.20**	**95.04**	**94.74**
Rezoom + ADAM	15.78	35.94	20.65	96.15	86.81	91.18	**87.65**	**92.12**	**84.17**

Table III
RESULTS OF USING REZOOM AND LSTM AND THEIR COMBINATIONS WITH 90% AS CONFIDENCE

negatively. Masking reflections has a varying effect on the results, depending on the videos. The best experienced result was an increase in positive recall of 5% and negative recall of 1.5%, while the worst experienced result was a decrease in positive recall of 19% and negative recall of 1.3%. Due to this, masking reflections had an overall negative effect on the F1-score. A more advanced implementation of masking reflections that are able to handle reflections of many shapes and colors, may be able to increase the results in additional videos, making masking reflections able to improve overall performance. Contrast enhancement has a varying effect on the results, depending on the polyps in the videos. If there are polyps with defined outlines, it is able to enhance the polyp, making it more detectable. If there are no outlines, for instance if the polyp is part of other structures in the colon, the polyp will not be more detectable, but instead other structures that resemble that of a polyp may be mistaken as such, producing additional false positives.

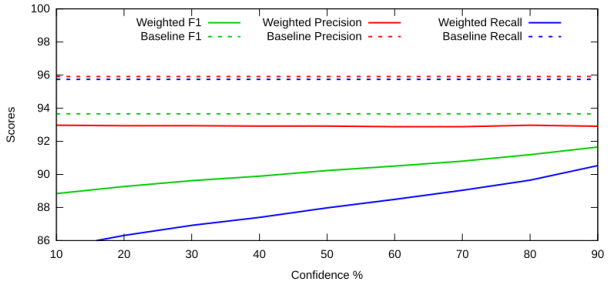
Rotation can be combined with any data enhancement method, where it is able to increases the positive recall while only slightly lowering the negative recall, making the addition of rotation an overall improvement. When masking reflections and contrast enhancement are combined, they improve each others results by providing mutual gains, improving the performance from their individual results. By introducing rotation to MC, the mutual gains are not present, making RMC produce worse results than RM and RC. When introducing rotation to either masking reflections or contrast enhancement, we see the highest increase in positive recall, where RMs 31.62% is up from 22.26% and RCs 37.79% is up from 24.62%, which makes RCs positive recall an increase of 11.50% compared to NP. The combinations that include contrast enhancement tend to have a higher positive recall, while those that include masking reflections tend to have a higher negative recall. Because of the imbalance in the dataset, masking reflections achieves a higher F1-score than contrast enhancement. We still view RC as the most interesting combination because of its high positive recall and relatively high negative recall, and it is thus a suitable combination for various scenarios.

V. OTHER TRAINING OPTIMIZATION

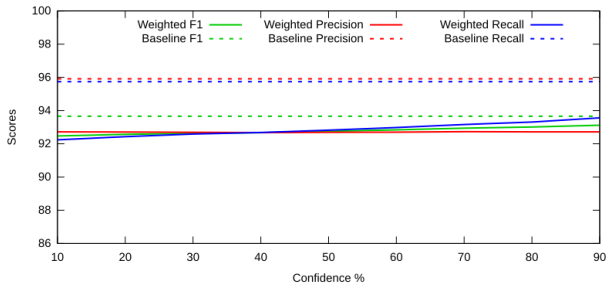
We have also performed experiments with various training optimizations where an excerpt of our results are shown in table III. First, we performed a limited experiment with different types of neural networks where Inception and Resnet were tested to see the effects of the different types of networks. They both produced very high positive recall, but at the same time very low negative recall, making them non-optimal for most common polyp detection scenarios. Due to this, we did not perform in-depth experiments with Inception and Resnet, and decided to keep the default neural network.

We then experimented with different training techniques, consisting of LSTM and Rezoom. First a limited experiment was performed with the different combinations since LSTM and Rezoom can both be enabled and disabled, resulting in four different combinations. LSTM produced the highest positive recall while Rezoom produced the highest weighted F1-score, which is then used for the next experiment.

Our next experiment examine examined how different optimizers affect the results. Root Mean Square Propagation (RMS), Stochastic gradient descent (SGD) and Adaptive Moment Estimation (Adam) were compared, where RMS has been used as the optimizer for all experiments until now. SGD was run with both LSTM and Rezoom from the previous step, where Rezoom and SGD produced the better results, therefore only Rezoom was tested with Adam. The results of SGD saw the highest increase in positive recall without a large sacrifice in negative recall, almost doubling its positive recall from 32.38% to 60.04%. Adam produced both lower positive and negative recall compared to SGD with no other benefits, and as such SGD was the optimizer used in the last experiment.



(a) After 100k iterations



(b) After 500k iterations.

Figure 5. Confidence variations using 100k and 500k iterations.

For the last experiment, we combined the optimized training with the most optimal data enhancements. From the experiments with data enhancements, R produced the highest weighted F1-score and RC the highest positive recall, which is why R and RC were used in this experiment. In combination with SGD, both produced similar results, with R slightly higher than RC in every aspect for every confidence. By applying R, improvements can be had, especially in positive recall, with little no no decrease in negative recall. This shows that data enhancement and training optimization should be combined to achieve optimal results.

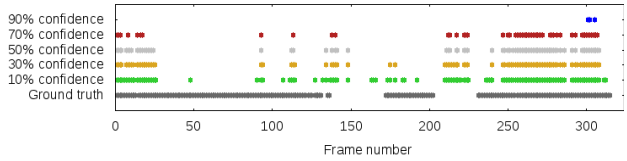
VI. Training iterations and confidences

As seen above, the network is very unstable before 300k iterations have been reached, as this is the phase where the network makes major adjustments to its weights. From 300k until 500k iterations, the network stabilizes to a larger extent, and while it is still adjusting its weights, the changes in the results become minor. In [24], we demonstrated that iterations beyond 500k iterations (up to 1000k) have no real impact. The reason a network will often fluctuate initially, is that it is less sure about its classifications until additional iterations have been executed. The confidence of the network for its classifications increases gradually, and will thus result in less fluctuations.

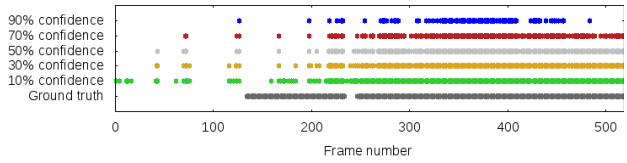
Every classification a network performs is accompanied by a confidence, a number between 0 and 1 indicating how sure the network is of the existence of a polyp within the image. When evaluating, a certain confidence number needs to be used as to what will be defined as a polyp, for instance, using 0.9 would mean that the classification needs to have a 90% certainty before we classify it as an image containing a polyp. A comparison of confidence variations between 100k and 500k iterations is shown in figure 5. With many confidence numbers towards the middle of the spectrum, the results become more randomized. When training the network further, the confidences were to a higher degree either close to 0 or 1.

When requiring a confidence of 0.9 in evaluation, the number of false positives are kept relatively low, but at the same time some true positives may be missed. By changing the required confidence, it is possible to trade false positives for true positives. A general balance does not exist, as different scenarios may focus on different aspects.

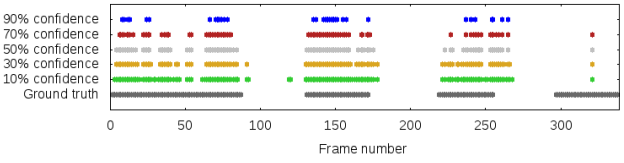
All detections for every video using a confidence of 10%, 30%, 50%, 70% and 90% are displayed in figure 6 together with the ground truth that indicates which frames contain a polyp. The figures show how the detection is spread throughout the videos, where all true and false findings can be read. They clearly show how different confidences affect detection and how lowering the confidence increases detection, but also produces additional false positives. Thus, choosing the appropriate confidence is important, and to avoid false positives, we have used a confidence of 90% in most of your experiments.



(a) Split 1.



(b) Split 2



(c) Split 3.

Figure 6. Effect of various confidence scores for selected polyp-videos.

VII. Conclusion

It has been shown many times that neural networks can be used for anomaly detection in medical videos. A challenge, however, is often that data is scarce and has artifacts. The question we have addressed in this paper is whether data preprocessing methods improve the detection rate. Using GI anomaly detection (polyps) as a case study,

we showed that rotation increases the overall performance, and a combination of rotation and contrast enhancement results in the highest number of detected polyps. Additionally, both masking reflections and contrast enhancement show potential depending on the video. Brightness variations, on the other hand, seems to be unable to produce positive effects. Thus, some data enhancements improve detection accuracy, where using Rezoom as a training technique and SGD as a training optimizer, seems to produce the best results where the detection is increased while keeping the number of false positives relatively stable.

REFERENCES

[1] American Cancer Society, Inc. (2017) Colorectal Cancer Statistics, 2017. [Online]. Available: http://pressroom.cancer.org/CRCstats2017

[2] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nuyen, M. Lux, P. T. Schmidt, M. Riegler, and P. Halvorsen, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proc. of MMSYS*, june 2017, pp. 164–169.

[3] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," *arXiv preprint arXiv:1606.05718*, 2016.

[4] F. Ciompi, K. Chung, S. J. van Riel, A. A. A. Setio, P. K. Gerke, C. Jacobs, E. T. Scholten, C. Schaefer-Prokop, M. M. Wille, A. Marchiano *et al.*, "Towards automatic pulmonary nodule management in lung cancer screening with deep learning," *arXiv preprint arXiv:1610.09157*, 2016.

[5] B. Li, Y. Fan, M. Q. H. Meng, and L. Qi, "Intestinal polyp recognition in capsule endoscopy images using color and shape features," in *Proc. of IEEE ROBIO*, Dec 2009, pp. 1490–1494.

[6] A. V. Mamonov, I. N. Figueiredo, P. N. Figueiredo, and Y. H. R. Tsai, "Automated polyp detection in colon capsule endoscopy," *IEEE Transactions on Medical Imaging*, vol. 33, no. 7, pp. 1488–1502, 2014.

[7] K. Pogorelov, S. Losada, C. Griwodz, T. de Lange, K. Ranheim Randel, D. T. Dang Nguyen, H. Kvale Stensland, F. De Natale, D. Johansen, M. Riegler, and P. Halvorsen, "A holistic multimedia system for gastrointestinal tract disease detection," in *Proc. of MMSys*, 2017.

[8] K. Pogorelov, M. Riegler, S. L. Eskeland, T. de Lange, D. Johansen, C. Griwodz, P. T. Schmidt, and P. Halvorsen, "Efficient disease detection in gastrointestinal videos – global features versus neural networks," *Multimedia Tools and Applications (MTAP)*, vol. 76, no. 21, 2017.

[9] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks," in *Proc. of IEEE ISBI*, 2015.

[10] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. J. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Józefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. G. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. A. Tucker, V. Vanhoucke, V. Vasudevan, F. B. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *CoRR*, vol. abs/1603.04467, 2016. [Online]. Available: http://arxiv.org/abs/1603.04467

[11] Google. (2017) Tensorflow. [Online]. Available: www.tensorflow.org/

[12] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE transactions on medical imaging*, vol. 35, no. 2, pp. 630–644, 2016.

[13] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015. [Online]. Available: http://science.sciencemag.org/content/349/6245/255

[14] J. Schmidhuber, "Deep learning in neural networks: An overview," *CoRR*, vol. abs/1404.7828, 2014. [Online]. Available: http://arxiv.org/abs/1404.7828

[15] I. Iseri, "An artifical intelligence based software application for microcalcification detection on mammogram images," in *Proc. of SIU*, May 2016, pp. 1973–1976.

[16] A. Zheng and M. Wang, "Convolutional neural networks-based plankton image classification system," 2015.

[17] T. Amaral, L. M. Silva, L. A. Alexandre, C. Kandaswamy, J. M. de Sá, and J. M. Santos, *Transfer Learning Using Rotated Image Data to Improve Deep Neural Network Performance*. Cham: Springer International Publishing, 2014, pp. 290–300. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-11758-4_32

[18] G. Yadav, S. Maheshwari, and A. Agarwal, "Contrast limited adaptive histogram equalization based enhancement for real time video system," in *Proc. of ICACCI*, Sept 2014, pp. 2392–2397.

[19] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. T. H. Romeny, and J. B. Zimmerman, "Adaptive histogram equalization and its variations," *Comput. Vision Graph. Image Process.*, vol. 39, no. 3, pp. 355–368, Sep. 1987. [Online]. Available: http://dx.doi.org/10.1016/S0734-189X(87)80186-X

[20] K. Zuiderveld, "Graphics gems iv," P. S. Heckbert, Ed. San Diego, CA, USA: Academic Press Professional, Inc., 1994, ch. Contrast Limited Adaptive Histogram Equalization, pp. 474–485. [Online]. Available: http://dl.acm.org/citation.cfm?id=180895.180940

[21] M. Zhou, G. Bao, Y. Geng, B. Alkandari, and X. Li, "Polyp detection and radius measurement in small intestine using video capsule endoscopy," in *Proc. of BMEI*, Oct 2014, pp. 237–241.

[22] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 630–644, 2016.

[23] J. Larsen and C. Goutte, "On optimal data split for generalization estimation and model selection," in *Proc. of Neural Networks for Signal Processing*, Aug 1999, pp. 225–234.

[24] F. L. Henriksen and R. Jensen, "Polyp detection using neural networks - data enhancement and training optimization," Master's thesis, University of Oslo, 2017.