The Deep Kernelized Autoencoder

Michael Kampffmeyer[a], Sigurd Løkse[a], Filippo M. Bianchi[a], Robert Jenssen[a,b], Lorenzo Livi[c]

[a]*Machine Learning Group, UiT–The Arctic University of Norway, http://site.uit.no/ml/*
[b]*Norwegian Computing Center, Oslo, Norway*
[c]*Department of Computer Science, University of Exeter, UK*

Abstract

Autoencoders learn data representations (codes) in such a way that the input is reproduced at the output of the network. However, it is not always clear what kind of properties of the input data need to be captured by the codes. Kernel machines have experienced great success by operating via inner-products in a theoretically well-defined reproducing kernel Hilbert space, hence capturing topological properties of input data. In this paper, we enhance the autoencoder's ability to learn effective data representations by aligning inner products between codes with respect to a kernel matrix. By doing so, the proposed *kernelized* autoencoder allows learning similarity-preserving embeddings of input data, where the notion of similarity is explicitly controlled by the user and encoded in a positive semi-definite kernel matrix. Experiments are performed for evaluating both reconstruction and kernel alignment performance in classification tasks and visualization of high-dimensional data. Additionally, we show that our method is capable to emulate kernel principal component analysis on a denoising task, obtaining competitive results at a much lower computational cost.

Keywords: Autoencoders; Kernel methods; Deep learning; Representation learning.

1. Introduction

Autoencoders (AEs) are a class of neural networks that gained increasing interest in recent years [28, 30, 44, 51, 53]. AEs are used for unsupervised learning of *effective* latent representations of data [4, 18]. However, what an *effective* representation consists of is highly dependent on the target task, such as clustering and classification [5]. In standard AEs, representations are derived by training the network to reconstruct inputs through either a bottleneck layer, thereby forcing the network to learn how to compress inputs, or through an over-complete representation. It can be shown that training autoencoders using a reconstruction error corresponds to maximizing the lower bound of the mutual information between input and the learned representation [53]. Regularization methods are commonly employed for enforcing sparseness, improving robustness to noise, avoiding trivial identity mappings, or penalizing sensitivity of the representation to small changes in inputs [5]. Nonetheless, regularization alone provides limited control over the nature of the hidden representation.

In this paper, we propose a method to learn representations that preserve desired similarities in input space with an AE. In our approach, similarities are encoded in form of a kernel matrix, which is used as a prior to be reproduced by inner products of the hidden representations learned by the AE. This allows us to learn data representations with specified pairwise relationships. The training loss minimizes a combination of reconstruction error and a term quantifying the misalignment of the prior and the inner products of the hidden representations; the misalignment is computed by means of the normalized Frobenius norm. We note that this process acts as a regularization for the hidden representations and resembles the well-known kernel alignment procedure [54]. Our contribution is in principle related to other well-established methods like those from the family of multidimensional scaling [7], where an explicit embedding of the data is computed by minimizing a measure of distortion based on inner products. Further, we will experimentally show that the proposed regularization method allows mitigating a problem often observed in non-regularized AEs, where codes for similar images are not similar themselves and the underlying manifold is disconnected [37].

1.1. Related Works

The proposed model, called *deep kernelized autoencoder*, is related to recent attempts to incorporate kernel and information theoretic learning methods within neural network architectures [9, 55]. Specifically, it is connected to works on interpreting neural networks from a kernel perspective [39] and the Information Theoretic-Learning Auto-Encoder [44], which imposes a prior distribution over the hidden representation in a variational autoencoder [30]. Achille and Soatto [1] proposed a regularization method exploiting information dropout, an information-theoretic generalization of dropout [48] for neural networks and show that an AE trained with such a regularization for a specific parameter setting simplifies to the variational autoencoder objective. Other information-theoretic learning concepts, such as the information bottleneck [49], have also recently emerged in the deep learning literature [47]. In [2] variational inference is used to optimize the lower bound on the information bottleneck to learn representations that maximize the mutual information between learned representation and output while minimizing the mutual information between input and hidden representation. Computing the information bottleneck is difficult, especially with high-dimensional data. Chalk et al. [8] proposed an efficient variational scheme for maximizing a lower bound of the original information bottleneck formulation, which also allows for non-linear mappings between input and compressed representation via kernel functions. Beside dimensionality reduction, neural networks utilizing kernel and information theoretic concepts have also been used to perform clustering [26].

In our work, we exploit kernel alignment to match the inner products of the learned representations with a similarity measure in the input space encoded as a kernel matrix. A recent related work in this direction by Horn and Müller [21] attempts to learn representations that preserve pairwise similarity by means of AEs. The authors specifically focus on dimensionality reduction, showing the possibility to approximate the pairwise data similarity in input space in linear fashion from the learned low-dimensional representation. In practice, given an input data point, the network is trained to recreate the related row of the similarity matrix. Recently, Chu and Cai [10] propose a similarity-preserving AE based on clustering data in input space. Hidden representations are learned in such a way that data points belonging to the same cluster are similar also in the hidden representation.

Another recent approach consists in integrating Wasserstein Generative Adversarial Neural Networks into the AE framework [28]. Similarly, Tolstikhin et al. [51] proposes the Wasserstein Autoencoder, which is based on a novel regularization technique minimizing the Wasserstein distance between the model distribution and a target distribution.

1.2. Contribution and paper organization

In addition to providing more control over hidden representations, our method also has several benefits that compensate for important drawbacks of traditional kernel methods. By means of an end-to-end training procedure, we learn an explicit approximate mapping function from the input to a kernel space, as well as the associated back-mapping to the input space. Once the mapping is learned, it can be applied to inputs and operations performed in kernel space can then be explicitly simulated by means of linear operations in code space, thus in practice allowing to perform non-linear operations in input space. Mini-batch training is used in the proposed method in order to lower the computational complexity inherent to traditional kernel methods and, especially, spectral methods [6, 24, 45]. Furthermore, our method can be used with arbitrary kernel functions, even those computed with an algorithmic procedure, i.e., where inner products in kernel space are not expressed by an analytic function. To stress this fact, in our experiments we consider the probabilistic cluster kernel (PCK), a kernel function that is the result of a feature generation procedure. PCK is robust with respect to hyperparameter choices and has been shown to often outperform counterparts such as the radial basis function (RBF) kernel [23].

A preliminary version of this method appeared in [25]. Here we extend our work by:

- providing a thorough literature background discussion, placing our work into a broader context;

- extending the experimental evaluation to additional datasets, namely (i) the image dataset CIFAR-10, (ii) the text dataset Reuters, and (iii) the remote sensing dataset Cloud;

- experimentally analyzing the effectiveness of the learned representations for classification tasks and visualizing high-dimensional data, and for generating new data samples beyond those seen during training.

The paper is structured as follows. Section 2 provides the reader with a discussion of the relevant background, such as AEs and kernel methods; notably in Section 2.3 we introduce PCK, adopted here for obtaining kernel matrices to be used in our method. Section 3 describes the proposed methodology. Experimental results are discussed in Section 4 and Section 5. Finally, Section 6 draws conclusions and points to future research directions.

2. Background

2.1. Autoencoders and stacked autoencoders

AEs simultaneously learn two functions. The first one, the *encoder*, provides a mapping from an input domain, $\mathcal{X}$, to a code domain, $\mathcal{C}$, i.e., the hidden representation. The second function, the *decoder*, maps from $\mathcal{C}$ back to $\mathcal{X}$. For a single hidden layer AE, the encoding function $E(\cdot)$ and the decoding function $D(\cdot)$ are defined as

$$\begin{aligned}
\mathbf{h} &= E(\mathbf{x}) = \sigma(\mathbf{W}_E \mathbf{x} + \mathbf{b}_E) \\
\tilde{\mathbf{x}} &= D(\mathbf{h}) = \sigma(\mathbf{W}_D \mathbf{h} + \mathbf{b}_D),
\end{aligned} \tag{1}$$

where $\sigma(\cdot)$ denotes a suitable transfer function (e.g., a sigmoid applied component-wise), $\mathbf{x}$, $\mathbf{h}$, and $\tilde{\mathbf{x}}$ denote, respectively, a sample from the input space, its hidden representation also called *code*, and its reconstruction; finally, $\mathbf{W}_E$ and $\mathbf{W}_D$ are the weights, and $\mathbf{b}_E$ and $\mathbf{b}_D$ the bias of encoder and decoder, respectively.

In order to minimize the discrepancy between the original data and its reconstruction, model parameters in Equation 1 are learned by minimizing, usually through stochastic gradient descent (SGD), a reconstruction loss of the form

$$L_r(\mathbf{x}, \tilde{\mathbf{x}}) = \|\mathbf{x} - \tilde{\mathbf{x}}\|_2^2 . \tag{2}$$

Differently from Equation 1, a stacked autoencoder (sAE) consists of several hidden layers [18]. Deep architectures are capable of learning complex representations by transforming input data through multiple layers of nonlinear processing [5]. The optimization of the weights is harder in this case and pretraining is beneficial, as it is often easier to learn intermediate representations, instead of training the whole architecture end-to-end [4]. A common application of pre-trained sAE is the initialization of layers in deep neural networks [53]. Pretraining is performed in different phases, each of which consists of training a single AE layer. After the first AE has been trained, its encoding function $E(\cdot)$ is kept fixed and is applied to the input and the resulting representation is used to train the next AE in the stacked architecture. Each layer, being trained independently, aims at capturing more abstract features by trying to reconstruct the representation in the previous layer. Once all individual AEs are trained, their hidden layers (encoding and decoding functions) are extracted and stacked on each other, yielding a pre-trained sAE.

2.2. A brief introduction to kernel methods

Kernel methods process data in a reproducing kernel Hilbert space (RKHS) $\mathcal{K}$ associated with an input space $\mathcal{X}$ through an implicit (non-linear) mapping $\phi : \mathcal{X} \rightarrow \mathcal{K}$. There, data are more likely to become separable by linear methods [11], which produces results that are otherwise only obtainable by nonlinear operations in the input space. Explicit computation of the mapping $\phi(\cdot)$ and its inverse $\phi^{-1}(\cdot)$ is, in practice, not required. In fact, operations in the kernel space are expressed through inner products (kernel trick), which are computed as Mercer kernel functions in input space: $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$.

As a major drawback, kernel methods scale poorly with the number of samples n: traditionally, memory requirements of these methods scale with $\mathcal{O}(n^2)$ and computation with $\mathcal{O}(n^2 \times d)$, where d is the input dimension [13]. For example, kernel principal component analysis (kPCA) [45], a common dimensionality reduction technique that projects data into the subspace that preserves the maximal amount of variance

3

in kernel space, requires to compute the eigendecomposition of a kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$, with $K_{ij} = \kappa(x_i, x_j), x_i, x_j \in \mathcal{X}$, with computational and memory costs scaling as $\mathcal{O}(n^3)$ and $\mathcal{O}(n^2)$, respectively. For this reason, kPCA is not applicable to large-scale problems. The availability of efficient (approximate) mapping functions, however, would reduce the complexity, thereby enabling these methods to be applicable to larger datasets [9]. In this direction, Rahimi and Recht [41] and Vedaldi and Zisserman [52] proposed approximate mappings preserving the dot product structure by using low-dimensional randomized features, hence allowing the use of fast linear methods in an explicit way. Furthermore, finding an explicit inverse mapping from $\mathcal{K}$ to the input domain is a central problem in several applications, such as image denoising performed with kPCA, also known as the pre-image problem [3, 20].

Our proposed method instead, attempts to approximate the operations in the kernel space using an AE architecture that scales to large datasets, provides an implicit inverse mapping, and, once trained, can process new samples efficiently.

2.3. Probabilistic cluster kernel

The Probabilistic Cluster Kernel (PCK) [23] is a robust kernel function, which automatically adapts to the inherent structures in the data. Its robustness comes from the fact that it does not depend on any critical user–specified hyperparameters, like the width in Gaussian kernels. The PCK is trained by fitting multiple Gaussian Mixture Models (GMMs) to the input data using the EM algorithm and combining these models to generate a single kernel. In particular, GMMs are trained using different number of mixture components $g = 2, 3, \ldots, G$, each with different randomized initial conditions $q = 1, 2, \ldots, Q$. Let $\boldsymbol{\pi}_i(q, g)$ denote the *posterior distribution* for data point $\mathbf{x}_i$ under a GMM with g mixture components and initial condition q. The PCK is then defined as

$$\kappa_{\mathrm{PCK}}(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{Z} \sum_{q=1}^{Q} \sum_{g=2}^{G} \boldsymbol{\pi}_i^T(q, g) \boldsymbol{\pi}_j(q, g), \tag{3}$$

where Z is a normalizing constant.

Intuitively, the posterior distribution under a mixture model contains probabilities that a given data point belongs to a certain mixture component in the model. Thus, the inner products in Equation 3 are large if data pairs often belong to the same mixture component. By averaging these inner products over a range of g values, the kernel function has a large value if these data points are similar on both global scale (small $g \to$ large mixture components) and local scale (large $g \to$ small mixture components).

The PCK has previously been used for semi-supervised learning [22] and spectral clustering [23]. Additionally, variations of the method for handling missing data have been proposed for both time series [38] and vectorial data [35].

3. Deep kernelized autoencoders

In this section, we describe our contribution, which is a method combining deep AEs with kernel methods: the deep kernelized AE (dkAE). A dkAE is trained by minimizing the following loss function

$$L = (1 - \lambda) L_r(\mathbf{x}, \tilde{\mathbf{x}}) + \lambda L_c(\mathbf{C}, \mathbf{P}), \tag{4}$$

where $L_r(\cdot, \cdot)$ is the reconstruction loss in Equation 2. $L_c(\cdot, \cdot)$ is the code loss, a distance measure between two matrices, $\mathbf{P} \in \mathbb{R}^{n \times n}$, the kernel matrix given as prior, and $\mathbf{C} \in \mathbb{R}^{n \times n}$, the inner product matrix of codes associated to the input data. The objective of $L_c(\cdot, \cdot)$ is to enforce the similarity between $\mathbf{C}$ and the kernel matrix $\mathbf{P}$. λ is a hyperparameter ranging in $[0, 1]$, which weights the importance of the two objectives in Equation 4; for $\lambda = 0$, the loss function simplifies to the traditional AE loss in Equation 2. A depiction of the training procedure is reported in Figure 1.

We implement $L_c(\cdot, \cdot)$ as the normalized Frobenius distance between $\mathbf{C}$ and $\mathbf{P}$. Each matrix element C_{ij} in $\mathbf{C}$ is given by $C_{ij} = E(\mathbf{x}_i) \cdot E(\mathbf{x}_j)$ and the code loss is computed as

$$L_c(\mathbf{C}, \mathbf{P}) = \left\| \frac{\mathbf{C}}{\|\mathbf{C}\|_F} - \frac{\mathbf{P}}{\|\mathbf{P}\|_F} \right\|_F. \tag{5}$$
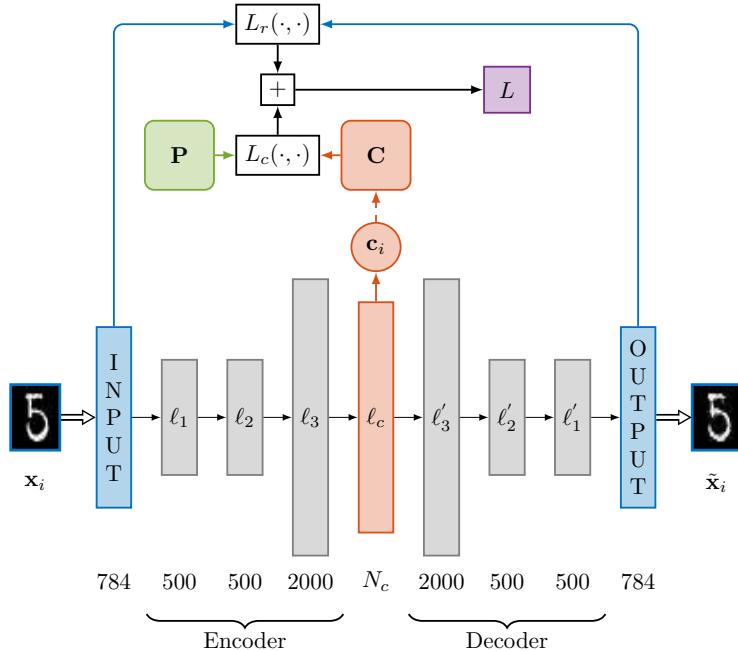
4

Figure 1: Schematic illustration of dkAE architecture. Loss function L depends on two terms. First, $L_r(\cdot, \cdot)$, is the reconstruction error between the true input $\mathbf{x}_i$ and the output of the dkAE, $\tilde{\mathbf{x}}_i$. The second term, $L_c(\cdot, \cdot)$, is the distance measure between matrices $\mathbf{C}$ (computed as inner products of codes $\{\mathbf{c}_i\}_{i=1}^n$) and the target prior kernel matrix $\mathbf{P}$. For mini-batch training the matrix $\mathbf{C}$ is computed over the codes of the data in the mini-batch and that distance is compared to the submatrix of $\mathbf{P}$ related to the current mini-batch.

It is worth noting that minimizing the normalized Frobenius distance between the kernel matrices is equivalent to maximizing the traditional kernel alignment cost, since

$$\left\| \frac{\mathbf{C}}{\|\mathbf{C}\|_F} - \frac{\mathbf{P}}{\|\mathbf{P}\|_F} \right\|_F = \sqrt{2 - 2A(\mathbf{C}, \mathbf{P})}, \tag{6}$$

where $A(\mathbf{C}, \mathbf{P}) = \frac{\langle \mathbf{C}, \mathbf{P} \rangle_F}{\|\mathbf{C}\|_F \|\mathbf{P}\|_F}$ is exactly the kernel alignment cost function [12, 54]. Note that the distance in Equation 6 can be implemented also with more advanced differentiable measures of (dis)similarity between positive-definite matrices, such as divergence and mutual information [14, 32]. However, these options are not explored in this paper and are left for future research.

In this paper, the prior kernel matrix $\mathbf{P}$ is computed by means of the PCK algorithm introduced in Section 2.3, such that $\mathbf{P} = \mathbf{K}_{\text{PCK}}$. However, our approach is general and *any* kernel matrix can be used as prior in Equation 5.

Note, that the kernel alignment also acts as a regularization, discouraging the learning of trivial mappings. Furthermore, we also employ tied weights in the encoder and decoder as additional regularization following [27].

3.1. Mini-batch training

We use mini batches of k samples to train the dkAE, thereby avoiding the computational restrictions of kernel and especially spectral methods outlined in Section 2.2. In particular, the memory complexity of the algorithm can be reduced to $\mathcal{O}(k^2)$, where $k \ll n$. Finally, we note that the computational complexity scales linearly with regards to the network parameters. Given a mini batch of k samples, the dkAE loss function
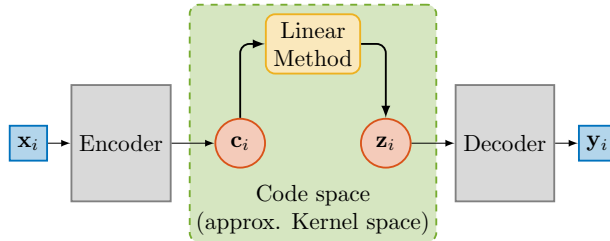
Figure 2: The encoder maps input $\mathbf{x}_i$ to $\mathbf{c}_i$, which lies in code space. In dkAEs, the code domain approximates the space associated to the prior kernel $\mathbf{P}$. A linear method receives input $\mathbf{c}_i$ and produces output $\mathbf{z}_i$. The decoder maps $\mathbf{z}_i$ back to input space. The result $\mathbf{y}_i$ can be seen as the output of a non-linear operation on $\mathbf{x}_i$ in input space.

is defined by taking the average of the per-sample reconstruction cost

$$L_{\text{batch}} = \frac{1-\lambda}{kd} \sum_{i=1}^{k} L_r(\mathbf{x}_i, \tilde{\mathbf{x}}_i) + \lambda \left\| \frac{\mathbf{C}_k}{\|\mathbf{C}_k\|_F} - \frac{\mathbf{P}_k}{\|\mathbf{P}_k\|_F} \right\|_F, \tag{7}$$

where d is the dimensionality of the input space, $\mathbf{P}_k$ is a subset of $\mathbf{P}$ that contains only the k rows and columns related to the current mini-batch, and $\mathbf{C}_k$ contains the inner products of the codes related to the mini-batch. Note that $\mathbf{C}_k$ is re-computed for each mini batch ($\mathcal{O}(k^2)$), while $\mathbf{P}_k$ is obtained by means of indexing operations with cost $\mathcal{O}(k)$.

3.2. Operations in code space

Linear operations in code space can be performed as shown in Figure 2. The encoding scheme of the proposed dkAE implicitly approximates $\phi(\cdot)$, mapping an input $\mathbf{x}_i$ onto the kernel space. In particular, in dkAEs, the feature vector $\phi(\mathbf{x}_i)$ is approximated by the code $\mathbf{c}_i$. Our non-linear encoder maps the inputs into a space where they are more likely to be linearly separable, as there the code vectors preserve a non-linear similarity computed in the input space. A linear operation on $\mathbf{c}_i$ produces a result in the code space, $\mathbf{z}_i$, relative to the input $\mathbf{x}_i$. Unlike other kernel methods where the explicit mapping back to the input space is not defined, we can map codes back by means of a decoder, which in our case approximates the inverse mapping $\phi(\cdot)^{-1}$ from the kernel space back to the input domain. This enables dkAEs to provide visualization and interpretation of the results in the original space; we further explore these perspectives in the experiments.

4. Analysis of dkAE

In this section, we perform an analysis of the proposed method by considering three experiments. Section 4.1 delineates the experimental setting. In Section 4.2, we evaluate the sensitivity of the two terms in the objective function (Equation 7) when varying the λ hyperparameter (in Equation 4) and the size of the code layer (i.e., number of neurons in the innermost hidden layer). Successively, in Section 4.3 we evaluate the reconstruction accuracy and kernel alignment performance implemented by dkAEs. Further, in Section 4.4 we compare dkAEs approximation accuracy of the prior kernel matrix with kPCA as the number of retained principal components increases.

4.1. Experimental setting

The analysis is performed on the MNIST dataset, which consists of 60000 handwritten digit images [33]. We use a subset of 20000 samples due to the computational restrictions imposed by the PCK, which we use to illustrate dkAEs ability to learn arbitrary kernels, even if they originate from an ensemble procedure.

We train PCK by fitting GMMs on a subset of 200 training samples using parameters $Q = G = 30$. These parameters are sufficiently large to ensure robust results [35]. Once trained, the GMM models are
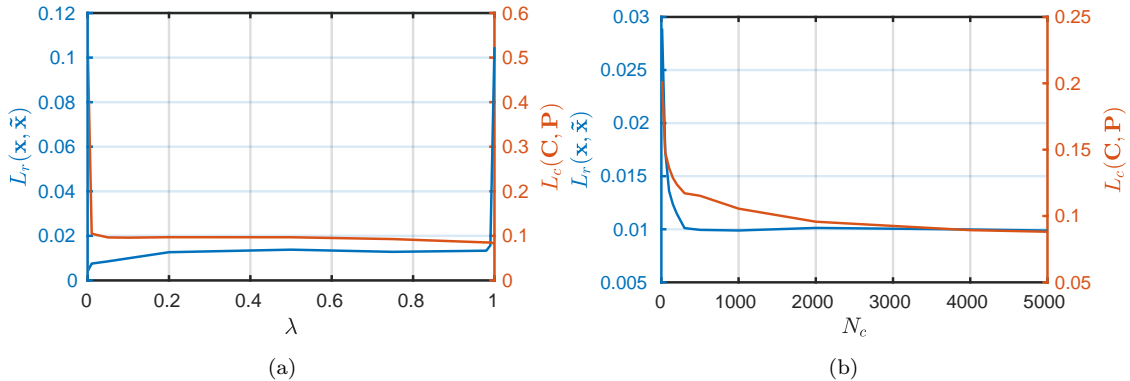
6

Figure 3: (a): Tradeoff when choosing λ. High λ values result in low L_c, but high reconstruction cost, and vice-versa. (b): Both L_c and reconstruction costs decrease when code dimensionality N_c increases.

applied to the remaining data to calculate the whole kernel matrix. We use 70%, 15% and 15% of the data for training, validation, and testing, respectively.

The network architecture used in the experiments is $d - 500 - 500 - 2000 - N_c$ (see Figure 1), which has been demonstrated to perform well on several datasets, including MNIST, for both supervised and unsupervised tasks [19, 36]. Here, N_c refers to the dimensionality of the code layer. Training was performed using the sAE pretraining approach outlined in Section 2.1. To avoid learning the identity mapping on each individual layer, we applied a common [27] regularization technique where the encoder and decoder weights are tied, i.e., $W_E = W_D^T$. This is done during pretraining and fine-tuning. Unlike in traditional sAEs, to account for the kernel alignment objective, the code layer is optimized according to Equation 4 *also* during pretraining.

Size of mini-batches for training was chosen to be $k = 200$ randomly, independently sampled data points; in our experiments, an epoch consists of processing $(n/k)^2$ batches. Pretraining is performed for 30 epochs per layer and the final architecture is fine-tuned for 100 epochs using gradient descent based on Adam [29]. The dkAE weights are randomly initialized according to Glorot et al. [15].

4.2. Sensitivity analysis of hyperparameter λ and size N_c of code layer

Here, we evaluate the influence of the two main hyperparameters influencing the resulting model. Note that the experiments shown in this section are performed by training the dkAE on the training set and evaluating the performance on the validation set. We evaluate both the out-of-sample reconstruction L_r and L_c. This is done in order to select the optimal parameters for evaluating the test set in the successive experiments. Figure 3(a) illustrates the effect of λ for a fixed value $N_c = 2000$ of neurons in the code layer. It can be observed that the reconstruction loss L_r increases as more and more focus is put on minimizing L_c (obtained by increasing λ). This quantifies empirically the trade-off in optimizing the reconstruction performance and the kernel alignment at the same time. By inspecting the results, specifically the near constant losses for λ in range $[0.1, 0.9]$ the method appears robust to changes in hyperparameter λ.

Analyzing the effect of varying N_c given a fixed $\lambda = 0.1$ (Figure 3(b)), we observe that both losses decrease as N_c increases. This could suggest that an even larger architecture, characterized by more layers and more neurons w.r.t. the architecture adopted here might work well, as the dkAE does not seem to overfit; due also to the regularization effect provided by the kernel alignment.

4.3. Reconstruction error and kernel alignment

By considering the previous results, in the following experiments we set $\lambda = 0.1$ and $N_c = 2000$. Figure 4 illustrates the results in Section 4.2 qualitatively by displaying a set of original images from our test set and their reconstruction error for the chosen λ value and a non-optimal one. Similarly, the prior kernel
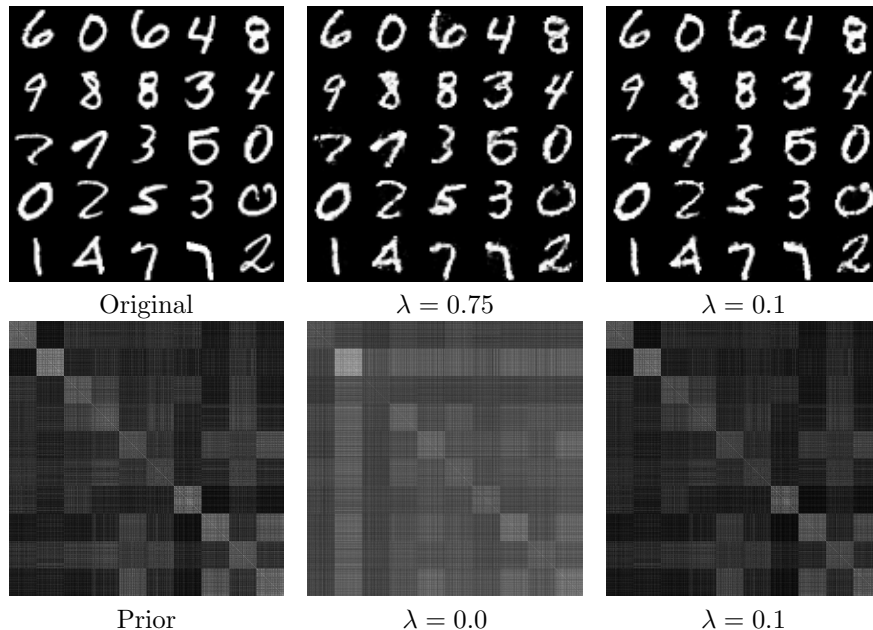
| | Original | $\lambda = 0.75$ | $\lambda = 0.1$ |

Figure 4: Illustrating the reconstruction error and kernel alignment trade-off in for different λ values. We note that the reconstruction for a small λ is generally better (see also Figure 3(a)), but that small λ yields high L_c.

| Kernel | Improvement [%] vs. | | | $L_c(\cdot, \mathbf{K}_I)$ |
	$\mathbf{P}$	$\mathbf{K}_{AE}$	$\mathbf{C}$	
$\mathbf{P}$	0	12.7	-0.2	1.0132
$\mathbf{K}_{AE}$	-11.3	0	-11.4	1.1417
$\mathbf{C}$	0.2	12.9	0	1.0115

Table 1: We compute L_c with respect to the ideal kernel matrix $\mathbf{K}_I$ for our test dataset (10 classes) and compare the relative improvement for the three kernels in Figure 4. It can be seen that the kernel matrix produced by dkAE ($\mathbf{C}$) is quantitatively comparable to the prior kernel ($\mathbf{P}$) with regards to its distance from the ideal kernel matrix and outperforms the traditional sAE ($\mathbf{K}_{AE}$).

(rows/columns sorted by class in the figure, to ease the visualization) and the dkAEs approximated kernel matrices, relative to test data, are displayed for two different λ values. Note that, to illustrate the difference to a traditional sAE, one of the two λ values is set to zero. It can be clearly seen that, for $\lambda = 0.1$, both the reconstruction error and kernel matrix closely resemble the original, which agrees with the plots in Figure 3(a).

Inspecting the kernels obtained in Figure 4, we compare the distance between the kernel matrices, $\mathbf{C}$ and $\mathbf{P}$, and the ideal kernel matrix, obtained by considering supervised information. We build the ideal kernel matrix $\mathbf{K}_I$, where $K_I(i, j) = 1$ if elements i and j belong to same class, otherwise $K_I(i, j) = 0$. Table 1 illustrates that the kernel approximation produced by dkAE outperforms a traditional sAE with regards to kernel alignment with the ideal kernel. Additionally, it can be seen that the kernel approximation $\mathbf{C}$ actually is more similar to the ideal kernel than the kernel prior, which we hypothesize is due to the reconstruction objective, which allows the codes to capture additional information (w.r.t. to PCK) about the structure of the input space.

4.4. Approximation of kernel matrix given as prior

In order to quantify the kernel alignment performance, we compare dkAE to the approximation provided by kPCA when varying the number of retained principal components. For this test, we take the kernel matrix $\mathbf{P}$ of the training set and compute its eigendecomposition. We then select an increasing number of components m (with $m \geq 1$ components related to the largest eigenvalues) to project the input data as
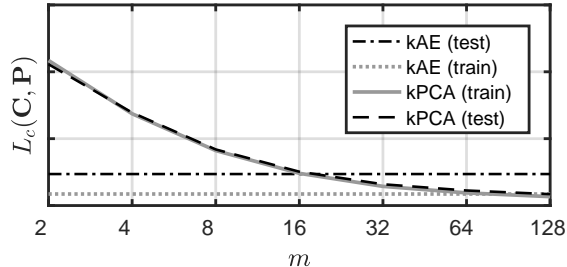
Figure 5: Comparing dkAEs approximation of the kernel matrix to kPCA for an increasing number of components. The plot shows that dkAE reconstruction is more accurate for low number (i.e., $m < 16$) of components.

follows: $\mathbf{Z}_m = \mathbf{E}_m \mathbf{\Lambda}_m^{1/2}, d = 2, ..., N$. The approximation of the original kernel matrix (prior) is then given by $\mathbf{K}_m = \mathbf{Z}_m \mathbf{Z}_m^T$. We compute the distance between $\mathbf{K}_m$ and $\mathbf{P}$ following Equation 6 and compare it to the dissimilarity between $\mathbf{P}$ and $\mathbf{C}$. For evaluating the out-of-sample performance, we use the Nyström approximation for kPCA [45] and compare it to the dkAE kernel approximation on the test set.

Figure 5 shows that the approximation obtained by means of dkAEs achieves a more accurate reconstruction then kPCA when using a small number of components, i.e., $m < 16$. Note that it is common in spectral methods to chose a number of components equal to the number of classes in the dataset [40], in which case, for the 10 classes in MNIST, dkAE would outperform kPCA. As expected, when the number of selected components increases, the approximation provided by kPCA is better. However, as shown in the previous experiment (Section 4.3), this does not mean that the approximation performs better with regards to the ideal kernel. In fact, in that experiment the kernel approximation of dkAE actually performed at least as well as the prior kernel (kPCA with all components taken into account).

5. Applications of dKAEs in classification, denoising, and visualization of high-dimensional data

In this section, we evaluate the effectiveness of dkAEs learned representations on multiple tasks. In Section 5.1, we compare classification performance on different benchmarks and illustrate how dkAEs can be used also for visualization of high-dimensional data. In Section 5.2, we present an application of our method for image denoising, where we apply PCA in dkAE code space $\mathcal{C}$ to remove noise.

For our classification experiments, apart from MNIST, we consider also the following datasets:

- CIFAR-10, which consists of 60000 32×32 color images belonging to 10 classes (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck) [31]. Similar to the MNIST dataset, we consider a subset of 20000 samples.

- Cloud, a dataset containing three multispectral satellite images captured over Spain and France. Each pixel in the images is represented by 19 dimensions, where 13 dimensions represent spectral bands from the MEdium Resolution Imaging Spectrometer (MERIS) instrument on board the Environmental Satellite (ENVISAT) [42], while the remaining six dimensions are related to physical features [16]. Each pixel is labeled according to the presence of a cloud in that particular area. This is a binary classification task, where the goal is to identify areas in the image which are obscured by clouds. The dataset is identical to the one previously used in [17].

  Similar to the MNIST dataset, we consider a subset of 20000 samples, where the training set consists of pixels sampled from one image, the validation set is sampled from a different image and the test set is sampled from the remaining image.

- Reuters, which consists of 800000 news stories that have been manually categorized into a category tree [34]. Similar to [56] we choose the four root categories as labels and remove stories that are labeled

9

Method	MNIST	CLOUD	CIFAR-10	REUTERS
SVM	90.60	99.50	36.60	91.40
kSVM	93.80	99.60	36.93	93.23
cSVM	94.80	99.63	38.17	93.77
scSVM	96.23	99.70	42.73	94.17

Table 2: Quantitative analysis of the learned feature representation of dkAE for classification tasks. A linear SVM operating in code space (cSVM) is compared with a linear SVM and a kernel SVM (kSVM) operating directly in input space. We also considered a linear SVM operating in code space where the prior $\mathbf{P}$ for the alignment is given by the outer product of class labels (scSVM).

with multiple root categories. To represent each news story we compute feature vectors consisting of the Term Frequency-Inverse Document Frequency (TF-IDF) of the 2000 most frequently occurring word stems and then use a Singular Value Decomposition (SVD) to produce 20 dimensional vectors prior to training. The SVD is performed on the training set, with out-of-sample transformations for validation and test sets.

5.1. Visualization and classification in code space

In order to evaluate the learned representation and illustrate the use of our method on an independent test set, we evaluate the classification performance of the learned representation. Here we make use of a linear support vector machine (SVM) operating in the code space and compare it to a linear and a non-linear kernel SVM (kSVM) operating directly in input space. The dKAE is trained on the training dataset, the SVMs model parameters are optimized on the validation set and the final accuracy is shown on the test dataset. Table 1 shows that linear SVM trained in the code space (cSVM) outperforms the SVM models operating in input space on all datasets.

As a consequence of the fact that our code representation is controlled by an arbitrary kernel matrix, we can also extend our work to learn representations in a supervised manner by aligning the code matrix $\mathbf{C}$ with the ideal kernel matrix. Similar to the experiments for the unsupervised representation, we train a linear SVM in the code space representation that has been learned (by exploiting supervised information) and provide the achieved accuracy (scSVM) in Table 2. As expected, when exploiting supervised information to learn representations, improvements are observed for all datasets. To illustrate the robustness of our approach with respect to architectural choices, we make use of the same architecture for all datasets, namely the one described in Section 4. Note, however, to avoid overfitting to the training data when training the supervised representation for the CLOUD dataset, the architecture for this particular dataset was reduced to $d - 50 - 50 - 200 - 200$ for all experiments.

Now we assess the capability to visualize high-dimensional data. Figure 6 shows the visualization of a low-dimensional representation learned by dkAE for the MNIST dataset; here, we consider 2000-dimensional codes. We utilize PCA to map the learned codes to two-dimensional vectors. We take into account also the low-dimensional representation learned by four alternative methods, namely an autoencoder without the use of kernel alignment, a denoising autoencoder (DAE) [53] with 20% masking noise, and kernel entropy component analysis (KECA) [24] as well as ISOMAP [50], two popular non-linear dimensionality reduction methods. Note, that the visualization here is presented for the test set and not the training data. For KECA we utilize an RBF kernel with σ being set to 15 percent of the median pairwise euclidean distances between datapoints, following a rule of thumb from [24]. We use KECA to reduce the dimensionality to 10 dimensions, the number classes in the dataset, before using PCA to reduce it further down to 2. In order to provide a quantitative evaluation of the visualizations, we consider the generalization error on a 1-Nearest Neighbor classification task following the example of [43]. Results are shown in Table 3, which demonstrate the superior performance obtained by means of dkAE.

5.2. Denoising and visualizing code space traversal in input space

Here, we highlight the potential of performing explicit operations in code space as initially described in Section 3.2. We try to emulate kPCA by performing PCA in our learned code space and evaluate the
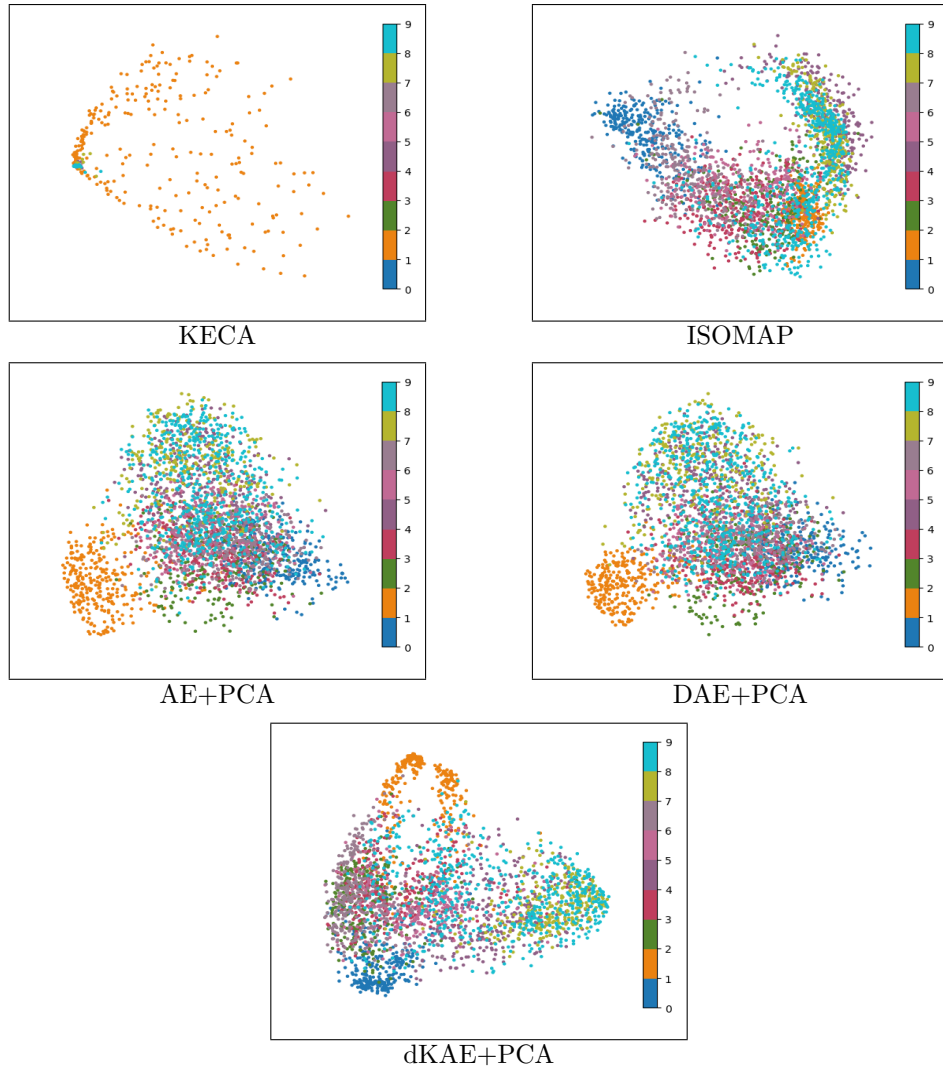
Figure 6: MNIST data. Two dimensional embedding of the code space obtained using standard AEs, DAEs and our dKAE. The codes are projected to two dimensions using PCA. We compare their preformance to non-linear dimensionality reduction techniques KECA and ISOMAP.

KECA	ISOMAP	AE+PCA	DAE+PCA	dkAE+PCA
29.5	36.8	30.5	31.2	**39.6**

Table 3: 1-nearest neighbor classification accuracy on representations shown in Figure 6. The overall best result is highlighted in bold.
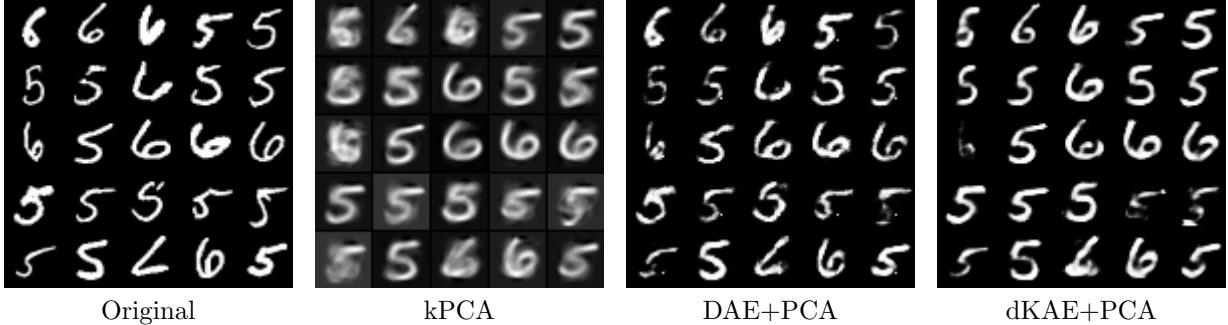
| Original | kPCA | DAE+PCA | dKAE+PCA |

Figure 7: Denoising with kPCA in input space and PCA in code space.

Noise std.	kPCA	DAE+PCA	dkAE+PCA
0.25	0.0427	0.0173	0.0358

Table 4: MSE of reconstruction.

performance on a denoising task. Denoising is a task that requires both a mapping to the kernel space, as well as a back-projection to the input space. Traditional kernel methods cannot perform back-projection explicitly; approximate solutions have been proposed in the literature [3, 20]. We choose the method proposed by Bakir et al. [3], where they use kernel ridge regression, such that a different kernel (in our case an RBF) can be used for back-mapping. Due to the challenge of finding a good σ for the RBF kernel that works on all MNIST numbers, we performed this test on the 5 and 6 class only. The regularization parameter and the σ required for the back-projection where found via grid search, where the best regularization parameter according to mean squared error (MSE) reconstruction was found to be 0.5 and σ as the median of the Euclidean distances between the projected feature vectors.

Both models are fitted on the training set and additive Gaussian noise is added to the test set. For both methods, 32 principal components are used. Table 4 shows that dkAE+PCA outperforms kPCAs reconstruction in terms of MSE. However, as MSE is not necessarily a good measure for denoising [3], we also visualize the results in Figure 7. It can be seen that dkAE yields sharper images in the denoising task. We further compare the results to a denoising autoencoder (DAE+PCA). We observe that the denoising autoencoder is able to outperform the dkAE with regards to the MSE measure as it is explicitly trained for the denoising task. Qualitatively, however, we observe in Figure 7 that the qualitative difference between these two is small, with DAE outperforming the dkAE on some images while producing more washed out images on others. For example, the reconstruction of the first image in the first row is better reconstructed using the DAE, while the second and fifth image in the first row are better reconstructed by the dKAE.

dkAE allows to explicitly explore the code space beyond the image of the dataset at hand and accordingly generate new instances with related representations in input space. To this end, we visualize the effect of movements in code space, illustrated in Figure 8. For this experiment, we perform k-means clustering in code space and chose the number of clusters to be equal to the number of classes in MNIST. We select pairs of cluster centroids at random and interpolate between two centroids following a straight path in code space; in future works, we will consider also non-linear methods to obtain a smoother interpolation between the centroids [46]. The first and last image in Figure 8 correspond to the cluster centers. The intermediate images are generated by mapping points along the aforementioned path in code space back to the input space by means of the trained decoder. In the first two panels, we observe a smooth transition of an 8 and a 7 to a 0. The third panel, instead, illustrates that k-means found two clusters in the 1s class, one for the far leaning ones and one for the straight ones. Interpolating between these two allows us to generate numbers with a varying degree of leaning to the right.

Figure 8: First and the last image of each panel show two k-means centroids in code space obtained on the MNIST dataset. Additional numbers are generated by "walking" on interpolated points between the two centroids.

6. Conclusions

We proposed a novel model for autoencoders, dubbed deep kernelized autoencoders, that exploits information provided by a user-defined kernel matrix to learn similarity-preserving data representations. The proposed model is trained end-to-end in an unsupervised way. By means of a parameter-free kernel alignment procedure based on inner products between codes, we are able to approximate arbitrary kernel functions defined in input space. This allows us to learn an explicit mapping from the input space to the code space, as well as the backward mapping. We evaluated the learned data representations on classification tasks and illustrated how the learned backmapping can be used to visualize operations performed directly in code space. In addition, the proposed autoencoder enables us to emulate well-known kernel methods for unsupervised learning, such as kernel PCA; however, our approach scales well with the number of data points as it is not based on eigendecomposition procedures.

In future work, we will continue to investigate this line of research by exploring alternative loss functions for kernel alignment, beyond those based on Frobenius norm. In particular, we will investigate the use of information-theoretic divergence measures and formulations based on mutual information between positive semi-definite matrices.

Acknowledgment

References

[1] A. Achille and S. Soatto. Information dropout: learning optimal representations through noise. *arXiv preprint arXiv:1611.01353*, 2017.

[2] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2017.

[3] G. H. Bakir, J. Weston, and B. Schölkopf. Learning to find pre-images. *Advances in Neural Information Processing Systems*, pages 449–456, 2004.

[4] Y. Bengio. Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[5] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, Aug 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.50.

[6] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[7] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching. *Proceedings of the National Academy of Sciences*, 103(5):1168–1172, 2006. doi: 10.1073/pnas.0508601103.

[8] M. Chalk, O. Marre, and G. Tkacik. Relevant sparse codes with variational information bottleneck. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 1957–1965. Currant Associates, Inc., 2016.

[9] Y. Cho and L. K. Saul. Kernel methods for deep learning. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, pages 342–350. Curran Associates, 2009.

[10] W. Chu and D. Cai. Stacked similarity-aware autoencoders. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1561–1567. AAAI Press, 2017.

[11] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[12] N. Cristianini, A. Elisseeff, J. Shawe-Taylor, and J. Kandola. On kernel-target alignment. *Advances in neural information processing systems*, 2001.

[13] B. Dai, B. Xie, N. He, Y. Liang, A. Raj, M.-F. F. Balcan, and L. Song. Scalable kernel methods via doubly stochastic gradients. *Advances in Neural Information Processing Systems*, pages 3041–3049, 2014.

[14] L. G. S. Giraldo, M. Rao, and J. C. Principe. Measures of entropy from data using infinitely divisible kernels. *IEEE Transactions on Information Theory*, 61(1):535–548, Nov. 2015. doi: 10.1109/TIT.2014.2370058.

[15] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. *In Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 249–256, May 2010.

[16] L. Gómez-Chova, G. Camps-Valls, J. Calpe-Maravilla, L. Guanter, and J. Moreno. Cloud-screening algorithm for envisat/meris multispectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 45(12):4105–4118, 2007.

[17] L. Gómez-Chova, R. Jenssen, and G. Camps-Valls. Kernel entropy component analysis for remote sensing image clustering. *IEEE Geoscience and Remote Sensing Letters*, 9(2):312–316, 2012.

[18] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786): 504–507, 2006. ISSN 0036-8075. doi: 10.1126/science.1127647.

[19] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7): 1527–1554, 2006.

[20] P. Honeine and C. Richard. A closed-form solution for the pre-image problem in kernel-based machines. *Journal of Signal Processing Systems*, 65(3):289–299, 2011.

[21] F. Horn and K.-R. Müller. Learning similarity preserving representations with neural similarity encoders. *arXiv preprint arXiv:1702.01824*, 2017.

[22] E. Izquierdo-Verdiguier, L. Gomez-Chova, L. Bruzzone, and G. Camps-Valls. Semisupervised kernel feature extraction for remote sensing image analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 52(9):5567–5578, 2014.

[23] E. Izquierdo-Verdiguier, R. Jenssen, L. Gómez-Chova, and G. Camps-Valls. Spectral clustering with the probabilistic cluster kernel. *Neurocomputing*, 149:1299–1304, 2015.

[24] R. Jenssen. Kernel entropy component analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32 (5):847–860, 2010.

[25] M. Kampffmeyer, S. Løkse, F. M. Bianchi, R. Jenssen, and L. Livi. Deep kernelized autoencoders. In P. Sharma and F. M. Bianchi, editors, *20th Scandinavian Conference on Image Analysis*, pages 419–430. Springer International Publishing, Tromsø, Norway, Jun. 2017. doi: 10.1007/978-3-319-59126-1_35.

[26] M. Kampffmeyer, S. Løkse, F. M. Bianchi, L. Livi, A.-B. Salberg, and R. Jenssen. Deep divergence-based clustering. In *IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–8, Sep. 2017. doi: 10.1109/MLSP. 2017.8168158.

[27] H. Kamyshanska and R. Memisevic. The potential energy of an autoencoder. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1261–1273, 2015. doi: 10.1109/TPAMI.2014.2362140.

[28] Y. Kim, K. Zhang, A. M. Rush, and Y. LeCun. Adversarially regularized autoencoders, 2018. URL `https://openreview.net/forum?id=BkM3ibZRW`.

[29] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[30] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[31] A. Krizhevsky. Learning multiple layers of features from tiny images. *Technical report, University of Toronto*, 2009.

[32] B. Kulis, M. A. Sustik, and I. S. Dhillon. Low-rank kernel learning with Bregman matrix divergences. *Journal of Machine Learning Research*, 10(Feb.):341–376, 2009.

[33] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[34] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5(Apr):361–397, 2004.

[35] S. Løkse, F. M. Bianchi, A.-B. Salberg, and R. Jenssen. Spectral clustering using pckid–a probabilistic cluster kernel for incomplete data. In *Scandinavian Conference on Image Analysis*, pages 431–442. Springer, 2017.

[36] L. Maaten. Learning a parametric embedding by preserving local structure. *International Conference on Artificial Intelligence and Statistics*, pages 384–391, 2009.

[37] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[38] K. Ø. Mikalsen, F. M. Bianchi, C. Soguero-Ruiz, and R. Jenssen. Time series cluster kernel for learning similarities between multivariate time series with missing data. *Pattern Recognition*, 76:569–581, 2018.

[39] G. Montavon, M. L. Braun, and K.-R. Müller. Kernel analysis of deep networks. *Journal Machine Learning Research*, 12: 2563–2581, Nov. 2011. ISSN 1532-4435.

[40] A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, pages 849–856, 2001.

[41] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems*, pages 1177–1184. Curran Associates, Inc., 2008.

[42] M. Rast, J. Bezy, and S. Bruzzi. The esa medium resolution imaging spectrometer meris a review of the instrument and its mission. *International Journal of Remote Sensing*, 20(9):1681–1702, 1999.

[43] G. Sanguinetti. Dimensionality reduction of clustered data sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):535–540, 2008.

[44] E. Santana, M. Emigh, and J. C. Principe. Information theoretic-learning auto-encoder. *arXiv preprint arXiv:1603.06653*, 2016.

[45] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[46] H. Shao, A. Kumar, and P. T. Fletcher. The riemannian geometry of deep generative models. *arXiv preprint arXiv:1711.08014*, 2017.

[47] R. Shwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

[48] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[49] S. Still. Information bottleneck approach to predictive inference. *Entropy*, 16(2):968–989, 2014. doi: 10.3390/e16020968.

[50] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. doi: 10.1126/science.290.5500.2319.

[51] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

[52] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):480–492, Mar. 2012. ISSN 0162-8828. doi: 10.1109/TPAMI.2011.153.

[53] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371–3408, 2010.

[54] T. Wang, D. Zhao, and S. Tian. An overview of kernel alignment and its applications. *Artificial Intelligence Review*, 43 (2):179–192, 2015.

[55] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing. Deep kernel learning. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 370–378, 2016.

[56] J. Xie, R. Girshick, and A. Farhadi. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pages 478–487. JMLR.org, 2016.