Deep Information Theory Based Robust Kernelized Autoencoders for Anomaly Detection in Latent Space

Anonymous submission

Abstract

Unsupervised outlier detection in subspaces is of great importance when discriminating the anomalies from the normal ones is difficult in high-dimensional space. Both density estimation and distance-based methods to detect anomalies in latent space have been explored in the past. These methods prove that retaining useful properties of the input data in the latent space helps in better reconstructing the original data. Kernel machines achieved great success when operated in a well-defined Reproducing Kernel Hilbert Space can capture important properties of original input data in the latent space. In this work, we try to enable kernelized auto-encoders with a robust distance measure using the Mahalanobis distance in the latent space and a differentiable measure of dissimilarity between positive semi-definite matrices using the Cauchy-Schwarz divergence and Renyi's entropy. The multi-objective function has two objectives - it tries to identify anomalies in the latent space using a robust hybrid reconstruction error, side by side, and it also tries to maximize the entropy of the original data while simultaneously preserving useful information between latent and prior space. This joint optimization aims to detect anomalies in the subspace through a strictly data-driven process of learning the similarity between prior and latent space.

Introduction

Autoencoders are a kind of neural network that learn data representations by optimizing a reconstruction loss such that input can be exactly reconstructed at the output. They have been used in the context of unsupervised learning for effectively learning latent representations in a low dimensional space (Zhou and Paffenroth 2017; Zong et al. 2018) when it is difficult to do density estimation in high dimensions (Chandola, Banerjee, and Kumar 2009). Training autoencoders using a reconstruction error corresponds to maximizing the lower bound of the mutual information between input and the learned representation (Vincent et al. 2010). Therefore, regularization methods are of great significance for robustness to noise, efficiently detecting anomalies in latent space, and also preserving useful information from the input space. In kernelized autoencoders, the pairwise similarities in the data are encoded in the form of a kernel matrix that we call a prior kernel and the auto-encoder tries to reconstruct it from the learned latent dimension representations (Cho and Saul 2009; Bengio, Courville, and Vincent 2013). This helps in learning data representations with predefined pairwise relationships encoded in the prior kernel matrix. Based on this idea, the Deep Kernelized Autoencoders (Kampffmeyer et al. 2017) optimize the reconstruction accuracy of input samples and a misalignment error in the form of a kernel matrix between a prior and inner product of latent space. The training loss is a multi-objective function that tries to optimize a reconstruction loss and a kernel alignment loss between the prior space and latent space in order to detect anomalies in the latent space.

An autoencoder has an encoding network that provides a mapping from the input domain to a latent dimension and the decoder tries to reconstruct the original data from the latent dimension. It is assumed that the data outside the inlier class cannot be effectively compressed and reconstructed (Pimentel et al. 2014). Often in a non-regularized latent space, the latent codes are unable to preserve the similarity of the input space and hence it may be difficult to reconstruct effectively (Lee et al. 2018). Empirical results suggest that reconstruction-based approaches alone fail to capture particular anomalies that lie near the latent dimension manifold (Denouden et al. 2018). Therefore, incorporating the Mahalanobis distance to better capture outlier samples in the latent space in addition to the reconstruction loss is explored in this paper. Since the outliers follow a more skewed distribution than the normal data, we introduce a robust version of the Mahalanobis distance estimation in this paper that employs the median absolute deviation as the parameter of estimation of scale and the median as the parameter of estimation of location in presence of outliers (Huber 1992; Hampel 2001). Moreover, information potential (IP) estimators for non-parametric density estimation have been explored previously in the context of clustering (Principe 2010) and optimizing neural network architecture (Tishby and Zaslavsky 2015). However, the usefulness of these information theory principles to identify anomalies in the latent space demands further investigation. In our method, we use a robust hybrid reconstruction loss which is accompanied by a kernel alignment loss between prior and latent space using mutual information in the form of the Cauchy-Schwarz divergence. This method when experimented with standard communication and cyber-security datasets shows better convergence and improvement in terms of precision, recall, and F1 and AUC scores(3% to 6 % approx.) over baseline models that rely

on reconstruction loss and simple matrix-based distances as misalignment error.

Related Work

Unsupervised anomaly detection in the latent space has been an interesting domain of research for a long time. Often, reconstruction-based approaches assume that anomalies cannot be effectively compressed and reconstructed from latent subspaces. These methods include Principal Component Analysis (PCA) (Jolliffe 2002), the Kernel PCA (Schölkopf, Smola, and Müller 1998), Robust deep AE (Zhou and Paffenroth 2017), (Yang et al. 2017), (Zhai et al. 2016). But there is a fundamental limitation is considering just reconstruction loss since sometimes anomalies lie near the latent dimension manifold defined by the model (Denouden et al. 2018). In such cases, anomalies with high reconstruction error and residing far away from the latent dimension manifold can be easily detected but those having low reconstruction error and lying close to the manifold with normal samples are highly unlikely to be detected. Therefore a robust method to detect such anomalies needs to be investigated which can detect anomalies in low-density areas in the subspaces.

One of the recent works in employing kernelized autoencoders to detect anomalies in latent space is the DKAE (Kampffmeyer et al. 2017) framework which has a training loss that is a combination of reconstruction error and misalignment error between the prior and the inner products of the latent space codes. The kernel alignment loss is calculated by measuring the normalized Frobenius norm between code and prior space kernel alignment. Self-organizing information theoretic principles have played a significant role to design information-theoretic cost functions in unsupervised learning such as clustering (Principe 2010). The entropy and divergence-based cost functions with the concept of information potential and information forces can play a significant role in the context of non-parametric learning where the emphasis is put to learn from the interaction of samples without putting constraints on the data pdf (Principe 2010; Schweizer and Wolff 1981; Rényi 1959). On the basis of information-theoretic principles, a differentiable measure of dissimilarity between psd matrices using divergence and mutual information to measure kernel misalignment error between prior and latent space has been explored in this paper.

Joint learning of dimensionality reduction and density estimation using Gaussian Mixture Models has been explored in the past. In (Zong et al. 2018), the authors have proposed joint learning frameworks where parameters of the GMM can be directly estimated in various anomaly detection tasks. The joint optimization of a distance measure in the latent space and energy estimation in the latent space was found useful in detecting anomalies in latent space. Our method follows a similar approach using non-parametric entropy estimation using kernel methods to detect anomalies in the latent space. In addition, it uses a robust hybrid reconstruction error using robust Mahalanobis distance in the encoded space (Huber 1992; Hampel 2001). This is especially useful when the data follow a skewed distribution and cannot be considered under the conventional Gaussian distribution assumption. This is different from the parametric estimation methods using GMMs which assume the pdf of the data in advance. Also, mean-based estimations are unlikely to be robust as they have a low breakdown point in presence of anomalies(Rousseeuw and Croux 1993).

Deep Information-theoretic Robust Kernelized Autoencoder

The Information Theoretic Learning Robust kernelized Autoencoder(ITL-RKAE) has two main components, a latent dimension compression network and an entropy estimation network that maximizes the entropy of latent space and minimizes the divergence between latent space and prior space simultaneously.

Overview

Information theory measures have a solid background in balancing entropy and mutual information using principles of maximizing entropy and minimizing joint entropy which can be directly estimated from samples in a non-parametric way. The entropy is usually estimated by centering a Gaussian kernel at each sample drawn from a probability distribution and then summing them up to estimate the entire pdf.

Renyi's Entropy, Joint Entropy and Mutual Information as ITL descriptors

In our work, the focus was on estimating the entropy directly from samples and their interactions in a non-parametric way. It is often unwise to make assumptions about a parametric PDF model in advance. Instead of estimating the PDF and computing its entropy, we estimate quadratic Renyi's entropy directly from the samples.

Renyi's α order entropy can be considered as a generalization of Shannon entropy. Renyi's quadratic entropy is estimated as

$$\hat{H}_2(X) = -\log\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathcal{G}_{\sigma\sqrt{2}}(x_i - x_j)\right). \quad (1)$$

Now, the cross-entropy or the joint entropy can be directly estimated from the samples as

$$\hat{H}_2(X,Z) = -\log\frac{1}{N_X N_Z}\sum_{i=1}^{N_x}\sum_{j=1}^{N_Z}\mathcal{G}_{\sqrt{2}\sigma}\left(x_i - z_j\right), \quad (2)$$

where $\mathcal{G}_\sigma(x,z) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{(||x-z||^2}{2\sigma^2}\right)$ is the Gaussian kernel and σ is the kernel bandwidth. The argument in equation 1 is called the information potential in this self-organizing principle (Principe 2010).

Again, the argument of equation 2 is termed as the cross information potential that measures the interaction between samples. The goal is to maximize the variance of the latent space samples so that the spread is maximum.

Matrix based Renyi's Entropy and Joint Entropy:

Parametric models rely on estimating the underlying distribution based on an assumption of the data which is difficult

to presume in high dimensional space. Therefore, we rely on kernel density estimations to calculate the entropy in a non-parametric way.

Let $\mathcal{G}$ be the gram matrix obtained from evaluating a positive definite kernel on all pairs of samples in the original input space. Thus, (1) can be updated with the matrix based analogue of Renyi's α entropy of order 2 for a normalized positive semi-definite matrix X of size $N \times N$, defined as: (Yu et al. 2021)

$$\hat{H_2}(X) = -\log_2(tr(X^2)) = -\log_2\left(\sum_{i=1}^{N}\lambda_i(X)^2\right), \quad (3)$$

where $X_{i,j} = \frac{1}{N}\frac{\mathcal{G}i,j}{\sqrt{\mathcal{G}_{ii}\mathcal{G}_{jj}}}$ and $\lambda_i(X)$ denotes the i^{th} eigenvalue of X. We used this measure to estimate the entropy of the original data space.

Similarly, (2) can be replaced by the matrix-based analog of α order joint entropy between latent space Z and prior space X, defined as

$$\hat{H_2}(X, Z) = H_2\left(\frac{X \circ Z}{tr(X \circ Z)}\right), \quad (4)$$

where $\circ$ denotes the Hadamard product of the matrices.

Based on the above definitions, we calculate the matrix-based normalized Renyi's entropy of the latent space encoded data, joint entropy between the prior space and latent space, and finally the mutual information between latent and prior space.

The Cauchy-Schwarz Divergence

The measure of closeness between two pdfs $p_1(x)$ and $p_2(x)$ is provided by information-theoretic measures like the Kullback-Leibler divergence (Goldberger et al. 2003). In this paper, we focus on the Cauchy-Schwarz divergence since it is differentiable and can be used as a loss function. If two pdfs $p_1(x)$ and $p_2(x)$ are non-negative and integrate to one, then the Cauchy-Schwarz divergence can be defined as

$$D_{cs}(p_1, p_2) = -\log\frac{\int p_1(x)p_2(z)\,dx\,dz}{\sqrt{\int p_1^2(x)\,dx \int {p_2}^2(z)\,dz}}. \quad (5)$$

The CS divergence is a symmetric measure, i.e, $0 \leq D_{cs} \leq \infty$ and it's minimum when $p_1(x) = p_2(x)$. If we consider $x_i,\ i = 1, \ldots, N$ be data points drawn from the density $p_1(x)$ and $x_j,\ j = 1, \ldots, N$ be the data points drawn from $p_2(x)$. The estimators of these distributions using Gaussian kernels are given as

$$\hat{p_1}(x) = \frac{1}{N_1}\sum_{i}^{N}\mathcal{G}_\sigma\left(x - x_i\right),$$

and

$$\hat{p_2}(z) = \frac{1}{N_2}\sum_{i}^{N}\mathcal{G}_\sigma\left(z - z_j\right).$$

Now, the actual densities can be replaced by their estimators as

$$\int p_1(x)p_2(z)\,dx\,dz = \int \hat{p_1}(x)\hat{p_2}(z)\,dx\,dz, \quad (6)$$
$$= \frac{1}{N_1 N_2}\sum_{i,j=1}^{N_1,N_2}\int \mathcal{G}_\sigma\left(x - x_i\right) \mathcal{G}_\sigma\left(z - z_j\right)\,dx\,dz.$$

Based on the above expressions and using the convolution property of Gaussian functions, the non-parametric sample-based estimator using the CS pdf divergence is given by

$$\hat{D}_{cs}(p1, p2) = -\log\frac{\sum_{i,j=1}^{N_1,N_2}\mathcal{G}_{\sqrt{2}\sigma}\left(x_i - z_j\right)}{\sqrt{\left(\sum_{i=1}^{N}\mathcal{G}_\sigma\left(x - x_i\right)\right)^2\left(\sum_{i=1}^{N}\mathcal{G}_\sigma\left(z - z_j\right)\right)^2}}. \quad (7)$$

Mutual Information as CS divergence between Prior and Latent Space

Information theoretic learning has a number of divergences that are connected with Renyi's entropy. Mutual information characterizes the fundamental compromise of maximum information retained in latent space from the prior space as well as the minimal information that can distinguish anomalies from the original data. The principle of relevant information (PRI) proposed by Jose Principe (Principe 2010) has been explored in the context of clustering. In this paper, we explored the mutual information in terms of the Cauchy-Schwarz divergence between input and the latent space.

The mutual information between input and latent space when expressed in terms of Cauchy-Schwarz divergence becomes

$$\hat{D}_{CS}(p_x||p_z) = \log_2\frac{\hat{H_2}(X)\hat{H_2}(Z)}{\hat{H_2}^2(X, Z)},$$

where $H_2(X, Z)$ is the second order joint entropy between latent and prior space.

The Cauchy-Schwarz divergence obeys the Cauchy-Schwarz inequality which guarantees the divergence to be only equal to zero when the pdfs are equal. The CS divergence is a symmetric measure. Minimizing the divergence over the prior or original data space p_x, i.e, $\min_{p_x} D_{CS}\left(p_x||p_z\right)$ is a trade-off between maximizing the entropy of the prior space i.e, p_x and minimizing the cross-entropy of the prior space with respect to the latent space. Maximizing the entropy causes the samples from p_x to spread out while minimizing the cross-information potential results in the samples from the latent space aligning toward samples from prior space.

The ITL Autoencoder and the Info-Theory cost function

The basic information theoretic autoencoder is trained by optimizing the following loss function

$$\mathcal{L} = (1 - \lambda)\,\mathcal{L}_r\left(X, \tilde{X}\right) + \min_x\left(H_2\left(X\right)\right) + \lambda\left(D_{CS}\left(X||Z\right)\right), \quad (8)$$

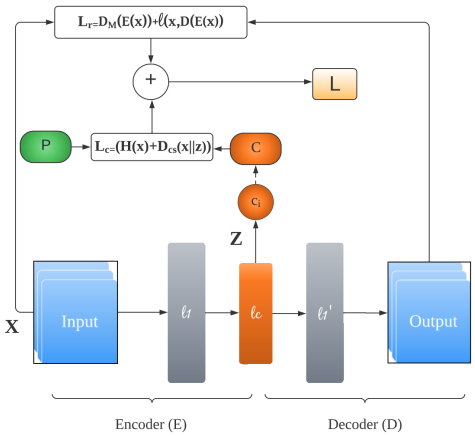


Figure 1: Deep ITL Robust KAE

where $L_r(\cdot,\cdot)$ is the reconstruction loss and λ is a hyperparameter which ranges in $[0,1]$. The regularization parameter tries to weigh the importance of the two objectives. The second part of the objective function calculates the misalignment error between the prior and the latent space expressed as psd matrices. We use the mutual information expressed in terms of the Cauchy-Schwarz divergence as a measure of dissimilarity between input and the latent space.

The information maximization principle and the maximum entropy optimization: According to the maximum entropy principle (MaxEnt) (Huber 1992; Jaynes 2003), one should ideally choose a distribution that maximizes the entropy of the data given the constraints. Maximizing the entropy adds the least of our prior knowledge of the problem and therefore lets the data speak for itself. Formally, we have

$$\max_p H_2(X) = -\sum_{i=1}^{N} p_{x_i} \ln p_{x_i},$$

$$\text{subject to} \sum_{i=1}^{N} p_{x_i} g_k(x_i) = a_k; \sum_{i=1}^{N} p_{x_i} = 1; \quad k = 1,\ldots,m.$$

The constraints are put on the expected values of the functions $g_i(X)$, $\sum_{i=1}^{N} p_i g_k(x_i) = a_k$, $k = 1,2,..m$, where the random variable X takes values $x_1,...x_N$ with corresponding probabilities $p_1,p_2...,p_N$

Minimum Cross-Entropy optimization principle The divergence like the entropy can be used to create a different optimization principle. When applied to the prior data space, the maximum entropy principle tries to optimize the entropy of the prior information and find a latent distribution Z that is closest to the prior space X. The latter can be achieved by minimizing the Cauchy-Schwarz divergence between the input and the prior space. Minimizing the divergence can be formulated as follows:

$$\min_p D_{CS}(X||Z), \text{ subject to} \sum_{i=1}^{N} p_{x_i} g_k(x_i) = a_k;$$

$$\sum_{i=1}^{N} p_{x_i} = 1; \quad k = 1,\ldots,m.$$

This optimization principle based on Cauchy-Schwarz divergence is called the minimization of cross-entropy between input and latent space. The information maximization principle as stated by Ralph Linsker (Linsker 1992) when applied in this case, should transfer as much information as possible from the input to the latent space. This results in maximizing the mutual information between the input and the latent space. The optimization is expressed by local interaction rules learning the complexity of the data. This cost function simultaneously attempts to maximize the entropy of the prior data space and minimize the Cauchy-Schwarz divergence between input and the latent space. This gives rise to opposing forces that make the data samples settle in different stationary configurations that correspond to a self-organizing clustering in the data.

Applying Kernel Smoothing in the latent space: Bandwidth selection is crucial while doing a non-parametric kernel density estimation. Optimum bandwidth selection is directly linked to the smoothness of the learning. A small bandwidth can lead to under-smoothing whereas a large bandwidth may result in over-smoothing. Following the idea of the paper (Álvarez-Meza, Cárdenas-Pena, and Castellanos-Dominguez 2014), we employed an adaptive strategy to tune the optimum kernel bandwidth in the latent space that aims to identify an RKHS that maximizes the variance of the information force among the samples in the latent space by maximizing the information potential variability of the Gaussian Kernel-based pdf estimation. The kernel bandwidth must be tuned precisely to estimate an RKHS that holds the most important data relationships. The non-parametric estimation of the Renyi's entropy directly from a set of N samples in the latent space, $z = \{z_j : j \in [1,N]\}$, can be given by

$$f(z) = p_z(z|\theta) = \mathbf{E}\{\mathcal{G}(z - z_j, \theta) : j \in [1,N]\},$$

where $\mathcal{G}\{\cdot,\theta\} \in R^+$ is a symmetric Gaussian kernel function with parameter set θ and $\mathbf{E}$ is the expectation operator. The estimator for Renyi's quadratic entropy can be given by

$$H_2(Z) = -\log \sum_{z_i \in Z} p_z^2(z_i|\theta) = -\log V(Z).$$

The argument of the logarithm in the above equation is termed the information potential (IP) of the latent space. The IP for the Gaussian kernel is defined as:

$$V(Z) = \mathbf{E}\{\mathcal{G}(z_i - z_j, \sigma^2) : i,j \in [1,N]\}.$$

It is inferred that the IP results in an entropy estimate that is based on the summation of pairwise sample interactions through the Gaussian kernel function (Singh and Principe 2010). Now, the information force (IF), $F_i \in \mathbf{R}^p$ is defined

as the force acting on a sample z_i due to all other samples in the latent space Z. It is given by the derivative of the IP with respect to z_i in the latent space. Formally, we have

$$\mathbf{F_i} = \frac{\partial V(Z)}{\partial z_i} = (N\sigma)^{-2} \sum_{z_j \in Z} \mathcal{G}\left(z_i - z_j, \sigma^2\right)(z_i - z_j), \\ = \mathbf{E}\{F(z_i|z_j) : j \in [1, N]\}.$$

Here $\mathbf{F}(z_i|z_j) = \left(N\sigma^2\right)^{-1} \mathcal{G}\{(z_i - z_j), \sigma^2\}(z_i - z_j)$ is known as the conditional IF acting on z_i due to z_j. It aims to seek an RKHS that maximizes the overall information potential variability with respect to the kernel bandwidth parameter so that all the IF magnitudes spread most widely on $\mathcal{G}$. In the latent space, this variability is maximized in terms of the kernel bandwidth parameter, given by

$$\sigma^* = \arg\max_{\sigma} var\{p_z(z|\sigma)\}.$$

Taking the derivative of the variance with respect to the bandwidth and equating it to zero, the bandwidth update can be given by

$$\sigma^2(k+1) = \frac{V_k(Z)\,\mathbf{E}\{(\mathbf{F}_k(z_i|z_j))^T(z_i - z_j) : i, j \in [1, N]\}}{\mathbf{E}\{\mathbf{F}_k^2(z_i|z_j) : i, j \in [1, N]\}},$$

where $V_k(Z)$ and $\mathbf{F}_k(z_i|z_j)$ are the IP and conditional IF obtained when $\sigma = \sigma(k)$, respectively.

The kernel smoothing operation in the latent space on the encoded data leads to smooth convergence of the entropy loss function.

Robust hybrid reconstruction error with the Mahalanobis distance in latent space

The reconstruction error is considered a novelty measure while identifying anomalies data from the normal data in the latent space. But reconstruction error is not a sufficient distance metric in the latent space (Pimentel et al. 2014). Those samples which lie extremely close to the latent space manifold will have low reconstruction error and will be difficult to get detected. We combined the Mahalanobis distance (D_M) between the encoded data and the mean vector of the encoded data in the latent space with the reconstruction error (l) of the training samples. The multi-objective function of (8) now becomes

$$\mathcal{L} = \alpha \cdot D_M(Z) + \beta \cdot \mathcal{L}(X, D(E(X)) + \min_x (H_2(X) + D_{CS}(Z||X). \tag{9}$$

During training, the regularization parameter α is set to the reciprocal of the mean absolute deviation of the Mahalanobis distance between the encoded validation data and the mean of the encoded train data. Likewise, β is set to the reciprocal of the mean absolute deviation of the reconstruction error on the validation data.

An alternative: The MAD and the Median: The Mahalanobis distance (M_d) is a distance measure based on how many standard deviations an encoded sample z_i is from the mean vector of the encoded data in the latent space. In the encoded space, it can be written as (Mahalanobis 1936):

$$\hat{D}_M(z) = \sqrt{(z - \hat{\mu})\,\hat{\Sigma}^{-1}\,(z - \hat{\mu})},$$

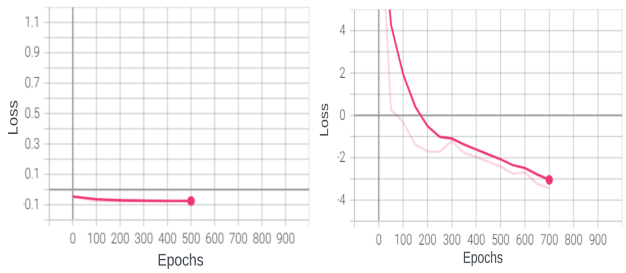
where $\hat{\mu}$ and $\hat{\Sigma}$ are the estimated mean vector and the estimated covariance matrix of the encoded data, respectively. This distance metric accounts for the correlation between the dimensions in the encoded space while estimating their distribution. The histogram of the data we experimented with shows that the data is not strictly following a Gaussian distribution and it is well known in the literature (Huber 2011; Rousseeuw and Croux 1993) that the mean and the variance are not the ideal parameters of estimation when data follow a skewed distribution like ours. We found our reconstruction error stagnating after some epochs. This led us to use more robust estimators of location and scale parameters such as the Median and the Median Absolute Deviation (MAD). The absolute deviation from the median(MAD) was popularized by Hampel (Hampel 2001). The median, like the mean, is a measure of central tendency but is found to be insensitive to the presence of outliers. It has a high breakdown point and becomes absurd only when 50 % of the observations are infinite. The MAD is also totally immune to the sample size.

Training and Experimental Results

For our experiments, we used a deep autoencoder model with 30 hidden layers and constrain our encoder(E) and decoder(D) to have the same architecture, i.e, $W_E = W_D^T$. Unlike, traditional AEs, we employ the kernel alignment objective in the code space as defined in (8) during training and fine-tuning. During training, we adjust the batch size from 200 to 600 based on the size of the training dataset. It is well-known that if the dimensionality is high, ITL divergence measures require a larger batch size to reliably estimate the entropy and the Cauchy-Schwarz divergence in the latent and prior space. The hyper-parameters of the model that needs to be tuned are the kernel bandwidth and the regularization parameter λ. These two hyper-parameters determine the behavior of our model. It is observed that the reconstruction loss increases when more weight is put on the kernel alignment loss. This demands a trade-off between the reconstruction performance and the kernel alignment when optimized simultaneously. The regularization parameter λ is kept as 0.5 during the training process. We also observe that the reconstruction error stagnates after a few epochs but the robust Mahalanobis distance (MD) in the latent space based on the MAD and median converges smoothly. Therefore, during our training and validation, we put more weight on the robust MD than the reconstruction loss. Further details of the training can be found in the supplementary materials.

Choosing an appropriate kernel matrix as prior

We used an RBF kernel that uses a multivariate Gaussian density function to do the kernel density estimation in a non-parametric way in order to estimate the probability density of the data. The smoothing is adjusted with the kernel bandwidth parameter. We used the Scikit-learn kernel density estimator that is capable of performing density estimation in



(a) (b)

Figure 2: (a) DKAE kernel loss (Frobenius distance); (b) DITL-RAE entropy loss (Mutual Information). Using the mutual information as kernel alignment loss results in better convergence.

any number of dimensions of the data. The kernel bandwidth and the regularization parameter were selected via a grid search algorithm. A Gaussian kernel with $\sigma = 0.16$ is found to be the best fit for our input data. The model is fine-tuned for 1000 epochs using gradient descent based on the Adam optimizer (Kingma and Ba 2014).

Algorithm 1: Algorithm for Estimating Mutual Entropy from Renyi's quadratic entropy using CS Divergence

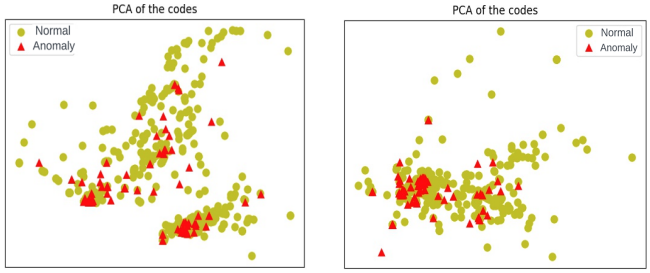
1: Input: $\mathbf{X} \in \mathbf{R}^d$ observation vector,batch size m, latent matrix $\mathbf{Z} = \mathbf{E}(\mathbf{X})$
2: Output: Estimated MI between prior and latent space.
3: Begin:
- Initialize autoencoder encoder and decoder weights: $\mathbf{W}_0, \mathbf{W}_1, ..., \mathbf{W}_n$
- Loop
  - Sample mini-batch $\{X_i\}_1^m \in \mathbf{R}^d$
  - Compute normal Gram matrices with Gaussian kernels of size $m \times m$ for each batch
  - Update kernel bandwidth σ
  - Calculate normalized second order Renyi's entropy, joint entropy and mutual entropy $\mathbf{MI}(\mathbf{X}, \mathbf{Z})$.
  - Evaluate code error $\mathbf{L_c} = \min_x(\mathbf{H_2}(\mathbf{X})) + \lambda(\mathbf{D_{CS}}(\mathbf{X}||\mathbf{Z}))$
  - Update $\mathbf{W} \leftarrow$ Optimize MI(X, Z)
- Until $||\mathbf{W}^{(k+1)} - \mathbf{W}^{(k)}|| < \epsilon$

4: End
5: $\mathbf{MI} = \mathbf{MI}^{(k+1)}$.

Datasets

We used three benchmark datasets: CSE-CIC-IDS2018, NSL-KDD, and Deepsig for our experiments.

- **CSE-CIC-IDS2018:** This is the publicly available cybersecurity dataset from joint collaboration between Communications Security Establishment (CSE) and the Canadian Cybersecurity Institute. The data has seven different attack scenarios like Brute-force, Heartbleed, Botnet, DoS, DDoS, Web attacks, and infiltration of the network. For our experiments, we performed a binary classification problem by considering all the anomalies under one anomaly class. It has a total of 79 traffic features. We did a correlation analysis among the features with the anomaly flag and selected the 29 most correlated features to create the model.
- **NSL-KDD:** This is also a publicly available dataset from the CIC. It has a total of 43 features for traffic data. We performed a similar correlation analysis as above and selected the 20 most important features related to the label.
- **Deepsig:** This is a communication dataset collected from different sensors at Deepsig. It has a total of 19 features and we selected the most correlated ones like the bandwidth, frequencies, sampling rate, etc.



(a) (b)

Figure 3: (a) Latent dimension projection of trained DKAE; (b) Latent dimension projection of trained DITL-RAE.

Table 1: Dataset details:

Datasets	Dimensions	Correlated feats	Instances	Anomaly ratio
CSE-CIC-IDS	79	29	5000	0.1
NSL-KDD	43	20	5000	0.1
Deepsig	19	6	5000	0.1

All the datasets are normalized with a mean of 0 and a standard deviation of 1. We have taken four subsamples from each of the datasets randomly, each containing 5k samples wherein anomaly and normal samples are distributed in the ratios as mentioned in table 1. We plotted the histograms to ensure that all the subsets follow the identical data-generating distribution.

Baseline Methods

- **DKAE:** The Deep Kernelized autoencoder (Kampffmeyer et al. 2017) has a code loss that enforces similarity between a kernel prior matrix represented as the inner product matrix of input data representation. The kernel alignment loss is calculated as the normalized Frobenius distance between the latent dimension code matrix and the prior kernel matrix.
- **DAGMM:** The Deep Autoencoding Gaussian Mixture Model (Zong et al. 2018) is an unsupervised anomaly detection model that utilizes deep autoencoders to generate low dimensional representation and reconstruction error

Table 2: Experiment Results on CSE-CIC-IDS data.

Subsets	DAGMM				DKAE			
	Precision	Recall	F1	AUC	Precision	Recall	F1	AUC
1	0.676	0.681	0.679	0.511	0.626	0.609	0.617	0.516
2	0.651	0.660	0.656	0.483	0.674	0.674	0.680	0.653
3	0.666	0.671	0.668	0.495	0.725	0.676	0.698	0.519
4	0.550	0.578	0.537	0.526	0.566	0.547	0.514	0.536
	Random Forest				DITL-RKAE			
1	0.800	0.890	0.840	0.500	**0.728**	**0.706**	**0.716**	**0.523**
2	0.830	0.910	0.870	0.500	**0.730**	**0.684**	**0.704**	**0.528**
3	0.800	0.890	0.840	0.500	**0.733**	**0.694**	**0.711**	**0.534**
4	0.790	0.890	0.830	0.500	**0.570**	**0.573**	0.572	0.522

Table 3: Experiment Results on NSL-KDD data.

Subsets	DAGMM				DKAE			
	Precision	Recall	F1	AUC	Precision	Recall	F1	AUC
1	0.948	0.958	0.953	0.953	0.994	0.994	0.994	0.986
2	0.967	0.953	0.960	0.974	0.996	0.996	0.996	0.976
3	0.940	0.868	0.891	0.888	0.986	0.986	0.986	0.970
4	0.954	0.865	0.893	0.919	0.996	0.996	0.996	0.990
	Random Forest				DITL-RKAE			
1	1.000	1.000	1.000	0.995	0.956	0.956	0.956	0.918
2	0.990	0.990	0.990	0.990	0.962	0.963	0.962	0.938
3	0.990	0.990	0.990	0.991	0.954	0.950	0.952	0.878
4	0.990	0.990	0.990	0.991	0.990	0.990	0.990	0.952

for each input data point which is then fed to a Gaussian Mixture Model (GMM). The DGAMM model jointly optimizes the parameters of the deep autoencoder and the mixture model simultaneously leveraging an estimation network to facilitate the learning of the mixture model.

- **Random Forests:** Random Forest or random decision trees is an ensemble supervised learning method used in classification and regression. It tries to do the estimation by fitting a number of decision tree classifiers on various sub-samples of a dataset and averages them to improve the accuracy and reduce over-fitting. We used the Scikit-learn built-in Ensemble Random Forest Classifier to generate the results.

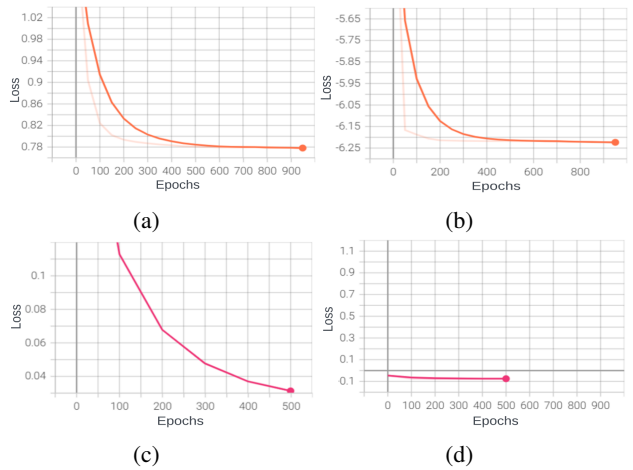


Figure 4: (a) DITL-RKAE reconstruction loss; (b) DITL-RKAE entropy loss; (c) DKAE reconstruction loss; (d) DKAE kernel loss. Robust reconstruction loss and entropy loss of DITL-RKAE show better convergence during training.

Metrics and Observation

We considered the standard metrics like precision, recall, AUC, and f1 scores to analyze our anomaly detection performance. In each of our experiments, we sample 5k samples of training data where anomaly distribution is 10% of the subset. In each case, we keep 500 samples as validation data that

Table 4: Experiment Results on Deepsig data.

Subsets	DAGMM				DKAE			
	Precision	Recall	F1	AUC	Precision	Recall	F1	AUC
1	0.951	0.886	0.905	0.938	0.933	0.965	0.948	0.952
2	0.951	0.886	0.905	0.938	0.950	0.954	0.952	0.878
3	0.950	0.954	0.952	0.878	0.992	0.992	0.992	0.964
4	0.986	0.986	0.986	0.986	0.986	0.986	0.986	0.950
	Random Forest				DITL-RKAE			
1	0.998	0.998	0.998	0.973	0.996	0.996	0.996	0.990
2	0.998	0.998	0.998	0.973	0.950	0.954	0.952	0.878
3	0.998	0.998	0.998	0.973	0.976	0.976	0.976	0.953
4	0.998	0.998	0.998	0.973	0.955	0.956	0.955	0.909

lies outside our training subset. We observe that increasing the number of samples and batch size for the entropy and divergence estimation for DITL-RAE show improvements over the baselines like GMM whose performance degrades with an increase in sample size. This reaffirms the fact that ITL divergence measures require larger batches for reliable estimation in the latent space. In some cases, stacking the auto-encoder also improves the results, however, we keep it for future scope of research on improving the architecture with more hidden and stacked layers.

Conclusion

In this paper, we propose the Deep Info-theory Learning Robust Kernelized Autoencoder for robust anomaly detection in latent subspaces. The DITL-RAE consists of two major components: a hybrid robust reconstruction loss using the median and MAD-based Mahalanobis distance and a divergence measure between prior and latent space using mutual information between psd matrices. Selecting a good regularization value while jointly optimizing the reconstruction loss and the kernel alignment loss helps improve anomaly detection in the latent space and in better convergence, especially when the data follow a skewed distribution. We follow non-parametric entropy estimation techniques using kernel density estimation without any prior assumption of the data. In future work, we will address the scalability aspect of the model with massive datasets using some decomposition techniques, which is often a requirement in the industrial setup. Although some initial experiments show minor improvement after stacking autoencoders and increasing

the number of hidden layers, we would analyze the performance of these large models with big datasets in the context of anomaly detection in the latent space. We also will explore the automatic hyper-parameter tuning of the regularization parameter and kernel bandwidth for different sizes of data.

References

Álvarez-Meza, A. M.; Cárdenas-Pena, D.; and Castellanos-Dominguez, G. 2014. Unsupervised kernel function building using maximization of information potential variability. In *Iberoamerican Congress on Pattern Recognition*, 335–342. Springer.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828.

Chandola, V.; Banerjee, A.; and Kumar, V. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3): 1–58.

Cho, Y.; and Saul, L. 2009. Kernel methods for deep learning. *Advances in neural information processing systems*, 22.

Denouden, T.; Salay, R.; Czarnecki, K.; Abdelzad, V.; Phan, B.; and Vernekar, S. 2018. Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance. *arXiv preprint arXiv:1812.02765*.

Goldberger, J.; Gordon, S.; Greenspan, H.; et al. 2003. An Efficient Image Similarity Measure Based on Approximations of KL-Divergence Between Two Gaussian Mixtures. In *ICCV*, volume 3, 487–493.

Hampel, F. R. 2001. Robust statistics: A brief introduction and overview. In *Research report/Seminar für Statistik, Eidgenössische Technische Hochschule (ETH)*, volume 94. Seminar für Statistik, Eidgenössische Technische Hochschule.

Huber, P. J. 1992. Robust estimation of a location parameter. In *Breakthroughs in statistics*, 492–518. Springer.

Huber, P. J. 2011. Robust statistics. In *International encyclopedia of statistical science*, 1248–1251. Springer.

Jaynes, E. T. 2003. *Probability theory: The logic of science*. Cambridge university press.

Jolliffe, I. T. 2002. *Principal component analysis for special types of data*. Springer.

Kampffmeyer, M.; Løkse, S.; Bianchi, F. M.; Jenssen, R.; and Livi, L. 2017. Deep kernelized autoencoders. In *Scandinavian conference on image analysis*, 419–430. Springer.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31.

Linsker, R. 1992. Local synaptic learning rules suffice to maximize mutual information in a linear network. *Neural computation*, 4(5): 691–702.

Mahalanobis, P. C. 1936. On the generalized distance in statistics. National Institute of Science of India.

Pimentel, M. A.; Clifton, D. A.; Clifton, L.; and Tarassenko, L. 2014. A review of novelty detection. *Signal processing*, 99: 215–249.

Principe, J. C. 2010. *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media.

Rényi, A. 1959. On measures of dependence. *Acta mathematica hungarica*, 10(3-4): 441–451.

Rousseeuw, P. J.; and Croux, C. 1993. Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424): 1273–1283.

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

Schweizer, B.; and Wolff, E. F. 1981. On nonparametric measures of dependence for random variables. *The annals of statistics*, 9(4): 879–885.

Singh, A.; and Principe, J. C. 2010. Kernel width adaptation in information theoretic cost functions. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2062–2065. IEEE.

Tishby, N.; and Zaslavsky, N. 2015. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, 1–5. IEEE.

Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.-A.; and Bottou, L. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12).

Yang, X.; Huang, K.; Goulermas, J. Y.; and Zhang, R. 2017. Joint learning of unsupervised dimensionality reduction and gaussian mixture model. *Neural Processing Letters*, 45(3): 791–806.

Yu, S.; Alesiani, F.; Yu, X.; Jenssen, R.; and Principe, J. 2021. Measuring dependence with matrix-based entropy functional. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 10781–10789.

Zhai, S.; Cheng, Y.; Lu, W.; and Zhang, Z. 2016. Deep structured energy based models for anomaly detection. In *International conference on machine learning*, 1100–1109. PMLR.

Zhou, C.; and Paffenroth, R. C. 2017. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 665–674.

Zong, B.; Song, Q.; Min, M. R.; Cheng, W.; Lumezanu, C.; Cho, D.; and Chen, H. 2018. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*.