OAAE: Adversarial Autoencoders for Novelty Detection in Multi-modal Normality Case via Orthogonalized Latent Space

Sungkwon An[1*], Jeonghoon Kim[3*], Myungjoo Kang[2], Shahbaz Razaei[4], Xin Liu[4†]

[1]Computational Science and Technology, Seoul National University, Korea
[2]Department of Mathematics, Seoul National University, Korea
{sk_an, mkang} @snu.ac.kr
[3]Department of Mathematics, University of California, Davis, CA
[4]Department of Computer Science, University of California, Davis, CA
{jhxkim, srezaei, xinliu}@ucdavis.edu

Abstract

Novelty detection using deep generative models such as autoencoder, generative adversarial networks mostly takes image reconstruction error as novelty score function. However, image data, high dimensional as it is, contains a lot of different features other than class information which makes models hard to detect novelty data. The problem gets harder in multi-modal normality case. To address this challenge, we propose a new way of measuring novelty score in multi-modal normality cases using orthogonalized latent space. Specifically, we employ orthogonal low-rank embedding in the latent space to disentangle the features in the latent space using mutual class information. With the orthogonalized latent space, novelty score is defined by the change of each latent vector. Proposed algorithm was compared to state-of-the-art novelty detection algorithms using GAN such as RaPP and OCGAN, and experimental results show that ours outperforms those algorithms.

Introduction

Novelty detection, also called anomaly detection in broader perspective, is regarded to be a task of recognising the test data that differs in some respect from the data that are previously seen. Novelty detection has been actively researched since the demand has been increasing due to its significance and broad applications in security, AI safety, healthcare industry.

Deep learning has recently shown tremendous performances in learning distribution and representations of various complicated data such as high-dimensional data, time series data. Deep learning for novelty detection aims to learn feature representations and output novelty scores through the neural network to detect data, which has different feature representations from the previously observed data. Many deep learning algorithms for novelty detection has been proposed recently, showing significantly better performances than traditional novelty detection methods. Deep generative models such as autoencoder (AE), generative adversarial networks (GANs) and their variational models are recognized as one of the biggest breakthrough in deep learning. Since they show great performances in pattern recognition in general, they are adopted for novelty detection in deep learning framework frequently. Deep generative models-based novelty detection algorithms such as OCGAN (Perera, Nallapati, and Xiang 2019), RaPP (Kim et al. 2019), AnoGAN (Schlegl et al. 2017), (An and Cho 2015), and (Sakurada and Yairi 2014) usually takes image reconstruction error or extension of it as a novelty score function. The key in novelty detection is to differentiate whether the input data is normal or novelty. However, as image data itself has a lot of inherent traits, e.g. rotations and thickness of the digit in images in MNIST dataset, image reconstruction error can be magnified by those factors, which eventually increases the wrong novelty detection cases potentially as shown in Figure 1. This gets worse in multi-modal normality case, which we aim to tackle. To the best of our knowledge, there has not been any precedent deep generative approaches to tackle novelty detection in multi-modal normality cases.

In this paper, we propose a new framework of novelty score function using orthogonalized latent space. Detection of novelty class in latent space has several benefits. Latent space is lower dimensional space with the feature information than the original high dimensional data, which is easier to be handled. Furthermore, features in latent space can be disentangled and highlight the class information to detect novelty class well. Low dimensional trait of latent space enable us to handle the features in the data easier. In this regard, we propose a novelty function using the change of angle in latent vectors by embedding input data in latent space orthogonal to each class using mutual class information.

Related work

One-class Novelty Detection. In recent years, one-class novelty detection has received tremendous attentions as a traditional representation learning research problem. There have been many classical approaches to tackle this problem such as Principal Component Analysis (PCA). Deep learning, which has shown great performances in a variety of fields such as computer vision, cybersecurity, medical assistance, and etc., finds a way to learn representation and detect

*Authors contributed equally
†Corresponding author

Figure 1: Limitation of novelty detection using image reconstruction error. Top: Input images. Middle: Output images of adversarial autoencoder (AAE). Bottom: mean squared error (MSE) of all images. We set the images of digits of 0-8 and 9 as normal and novelty, respectively. Since mean of novelty scores among the image of digit 9 (novelty class) is 7.4, MSE values of normal image bigger than 7.4 lead to wrong novelty detection.

the based on previously seen representation. AE-based novelty detection mostly put reconstruction error such as mean squared error as a novelty detection function after learning the representation of the data. GAN-based novelty detection usually takes discriminator's prediction in the image space as a tool of measuring reconstruction error. One-Class novelty detection using GAN (OCGAN) shows a great performance in novelty detection in uni-modal normality data.

Approaches on Novelty Score Function. There has been other approaches to determine novelty scores other than reconstruction error or discriminator's prediction. Generative Probabilistic Novelty Detection with adversarial autoencoders (GPND) (Pidhorskyi, Almohsen, and Doretto 2018) identifies novelty data by considering it to be an inlier or an outlier. GPND has done this by utilizing a probabilistic approach and computing how likely it is that a new data was generated by the normal distribution effectively. RaPP: Novelty Detection with Reconstruction along Projection Pathway (RaPP)(Kim et al. 2019) introduces a new way to quantify novelty scores using values in hidden space activation obtained from a deep autoencoder. RaPP compares input and its autoencoder reconstruction both of in the input space and in all of the hidden spaces. However, in order to enforce their metrics, RaPP network is required to be symmetric, which makes designing network architecture and training network a very expensive work. As the data becomes more complicated, it becomes more expensive due to fully-connected layers in encoder and decoder caused by its structural problem. RaPP also showed a great performance in multi-modal normality case.

Proposed Method: OAAE

In this section, We propose a new AAE novelty detection algorithm using orthogonalized latent space (OAAE) for multi-modal normality case. The key idea is to disentangle latent space using mutual class information by employing orthogonal low rank embedding (OLE) loss(Lezama et al. 2018), which enables us to achieve minimizing the variance of latent vectors in intra-class as well as maximizing margins of inter-class latent vectors (in terms of angle; equivalently orthogonalize inter-class latent vectors). With such an orthogonalized latent space, we estimate a novelty score by quantifying the change of angle in each latent vector.

Orthogonal Latent Embedding

OLE is carried out using rank function (Lezama et al. 2018). Mathematical formulations of OLE begins with the following equation:

$$\arg\min_{\mathbf{T}} \sum_{c=1}^{C} rank(\mathbf{T}\mathbf{X}_c) - rank(\mathbf{T}\mathbf{X}), \text{ s.t. } ||\mathbf{T}||_2 = 1, \tag{1}$$

where $\mathbf{X}$ denotes input dataset, $\mathbf{X}_c$ denotes the set of data points with class c in a subspace of $\mathbf{R}^d$, $\mathbf{T}$ is a linear transformation on the data (i.e., feed forward network for deep learning framework), $||\cdot||_2$ is the matrix Euclidean norm. We interpret this formulation term by term intuitively (Qiu and Sapiro 2015). Minimizing the first term $\sum_{c=1}^{C} rank(\mathbf{T}\mathbf{X}_c)$ keeps the transformed data from the same subspace a consistent representation, and maximizing the second term $rank(\mathbf{T}\mathbf{X})$ encourages the transformed data from different subspace to represent a diverse representation. Additionally, the normalization constraint $||\mathbf{T}||_2 = 1$ avoids the trivial solution, i.e., $\mathbf{T} = 0$. Since it is known that the nuclear norm ($||\mathbf{A}||_\star$; the sum of the singular values of the matrix $\mathbf{A}$) is the convex envelop of $rank(\mathbf{A})$ over the unit ball of matrices (Fazel 2003), and due to efficiency of optimization (Candès et al. 2011; Recht, Fazel, and Parrilo 2010), we reformulate the equation using the nuclear norm as follows:

$$\arg\min_{\mathbf{T}} \sum_{c=1}^{C} ||\mathbf{T}\mathbf{X}_c||_\star - ||\mathbf{T}\mathbf{X}||_\star, \text{ s.t. } ||\mathbf{T}||_2 = 1. \tag{2}$$

Following (Lezama et al. 2018), (2) becomes the following loss using minibatch as below to be applied to the deep learning framework:

$$\mathbf{L}_{OLE}(\mathbf{Y}) := \sum_{c=1}^{C} \max(\Delta, ||\mathbf{Y}_c||_\star) - ||\mathbf{Y}||_\star \tag{3}$$

$$= \sum_{c=1}^{C} \max(\Delta, ||\Phi(\mathbf{X}_c;\theta)||_\star) - ||\Phi(\mathbf{X};\theta)||_\star. \tag{4}$$

To optimize (3) using backpropagation, the projected subgradient for the nuclear norm and the descent direction for (3) are obtained in by using SVD decomposition on matrix $\mathbf{A}$, i.e., $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, and zero filling matrix $\mathbf{Z}_c$ as follows:

$$g_{||A||_\star}(A) = \mathbf{U}_1\mathbf{V}_1^T, \tag{5}$$

$$g_{\mathbf{L}_{OLE}}(\mathbf{Y}) = \sum_{c=1}^{C} \left[\mathbf{Z}_c^{(l)} | \mathbf{U}_{c1}\mathbf{U}_{c1}^T | \mathbf{Z}_c^{(r)}\right] - \mathbf{U}_1\mathbf{V}_1^T. \tag{6}$$

where $\mathbf{U}_1$ and $\mathbf{V}_1$ be the first s columns of $\mathbf{U}$ and $\mathbf{V}$, respectively, corresponding to eigenvalues larger than a small threshold value δ. Similarly, $\mathbf{U}_{c1}$ and $\mathbf{V}_{c1}$ be left and right singular vectors of $\mathbf{Y}_c$ where their corresponding singular values are greater than the threshold δ. Using $\mathbf{L}_{OLE}$ loss, we

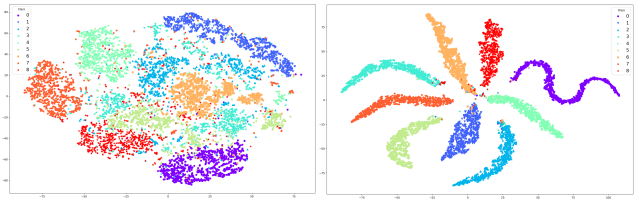

Figure 2: Embedding of trained latent space using t-SNE. Left: AAE without OLE loss. Right: AAE with OLE loss. Reduced variance of intra-class clusters of latent vectors was observed.

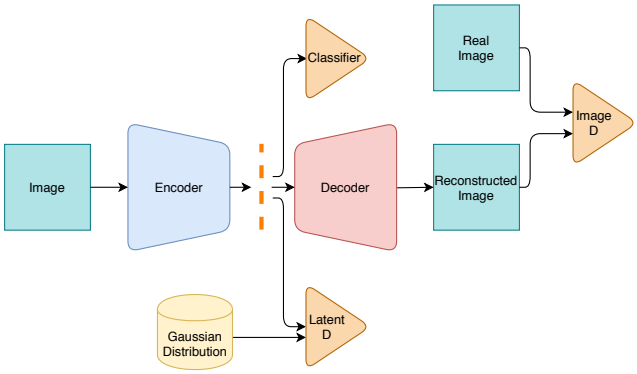

Figure 3: OAAE architecture

embed our high dimension dataset in orthgonoalized latent space with the two main benefits: reduced variance of intra-class, maximized angle margins of clusters of inter-class as shown in Figure 2.

Architecture

The architecture of the proposed network is based on AAE (Makhzani et al. 2015) and classifier was added to use mutual class information in OLE loss shown in Figure 3. Each of encoder and decoder in our model has five layers with three convolutional layers and two fully-connected layers at the end. Details of training of our algorithm is described in Algorithm1. Main key in our algorithm is to adopt OLE loss to use mutual information and disentangle features in latent space and returns novelty score using the change of angles in latent vectors.

Experiment

Datasets

MNIST. The MNIST database, which stands for Modified National Institute of Standards and Technology database, consists of a large number of 28×28 gray scale images of handwritten digits (10 classes; 0∼9). The MNIST dataset is commonly and widely used for various computer vision, image processing researches due to its simplicity. In our experiments, we choose images of one handwritten digit and every other images of remaining nine different handwritten digits as a novelty class, normal class data, respectively.

Algorithm 1 Novelty Detection algorithm

Input : Image x with class c, N Epochs, K Iteration
Training phase
for epochs 0 to N **do**
  for iteration 0 to K **do**
    $n \leftarrow \mathcal{N}(0, I)$
    $z \leftarrow \mathcal{N}(0, I)$
    Discriminator training phase
    $\mathcal{L}_{latent} \leftarrow \mathcal{D}_{latent}(z, 1) + \mathcal{D}_{latent}(Enc(x + n), 0)$
    $\mathcal{L}_{image} \leftarrow \mathcal{D}_{image}(x, 1) + \mathcal{D}_{image}(Dec(z), 0)$
    Back-propagate and update
    Encoder, Decoder and Classifier training phase
    if K%5 == 0 **then**
      $\mathcal{L}_{recon} \leftarrow ||x - Dec(Enc(x + n))||_2^2$
      $\mathcal{L}_{Enc} \leftarrow \mathcal{D}_{latent}(Enc(x + n), 1)$
      $\mathcal{L}_{Dec} \leftarrow \mathcal{D}_{image}(Dec(z), 1)$
      $\mathcal{L}_{ole} \leftarrow OLE(Enc(x + n), c)$
      $\mathcal{L}_{cls} \leftarrow CrossEntropy(C(Enc(x+n)), c)$
      Back-propagate and update
    end if
  end for
end for
Test phase
Test image x
$z_0 \leftarrow Enc(x)$
$z_1 \leftarrow Enc(Dec(Enc(x)))$
$Novelty\_Score \leftarrow angle(z_0, z_1)$

Fasion MNIST (f-MNIST). The fashion-MNIST is a dataset of 28×28 grayscale images 10 different classes (T-shirt, Trouser, Pullover, Dress, Coat, Sandals, Shirt, Sneaker, Bag, Ankle boots). It shares the same image size with the original MNIST dataset but f-MNIST is regarded as a harder data to learn in general because of the complexity that semantic images have. Similar to the previous experiments on MNIST dataset, we choose images with one class (e.g., T-shirt) and every other images of remaining nine different class as a novelty class, normal class data, respectively.

CIFAR10. The CIFAR10 dataset consists of 60000 32×32 coloured images with evenly distributed 10 classes (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck). This dataset was selected due to its complexity. CIFAR10 dataset is usually treated as harder data to train than MNIST or f-MNIST in general due to its multi-channel trait.

Architectures of Baseline Algorithms

We compare performance of our models to that of two state-of-the-art GANs-based novelty detection algorithms: OCGAN, RaPP. We briefly explain how those two algorithms work in the following sections.

OCGAN. OCGAN solves classical one-class novelty detection problem and aims to determine whether the new input is from the same class or not. The key idea of OCGAN is to learn latent representations of normal class data using a denoising autoencoder network and to directly force the latent space to entirely represent the given class. OCGAN is

Table 1: AUROC of OAAE and the baselines

	MNIST										
	0	1	2	3	4	5	6	7	8	9	Mean
OCGAN	0.91	0.08	0.76	0.81	0.77	0.72	0.87	0.37	0.923	0.46	0.67
RaPP	**0.99**	0.89	**0.98**	**0.95**	0.92	**0.97**	**0.98**	0.97	0.96	0.89	0.95
OAAE	0.98	**0.97**	0.97	**0.95**	**0.95**	**0.97**	0.975	**0.972**	**0.982**	**0.968**	**0.970**

	f-MNIST										
	T-shirt	Trouser	Pullover	Dress	Coat	Sandals	Shirt	Sneaker	Bag	Boots	Mean
OCGAN	0.577	0.750	0.596	0.723	0.557	0.801	0.546	0.769	0.877	0.726	0.692
RaPP	0.70	0.78	0.65	0.82	0.57	**0.85**	0.58	0.61	**0.98**	**0.82**	0.736
OAAE	**0.915**	**0.88**	**0.816**	**0.847**	**0.853**	0.716	**0.791**	**0.789**	0.966	0.799	**0.837**

	CIFAR10										
	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck	Mean
OCGAN	0.54	0.71	0.40	0.52	0.31	0.58	0.40	0.61	0.44	0.69	0.52
RAPP	0.469	0.654	0.416	0.578	0.357	0.604	0.382	0.579	0.553	0.681	0.527
OAAE	**0.706**	**0.777**	**0.579**	**0.713**	**0.660**	**0.742**	**0.620**	**0.683**	**0.652**	**0.786**	**0.692**

particularly focused on learning uni-modal normality data.

RaPP. A new methodolgy for novelty detection is proposed in RaPP by adopting values in hidden space activation obtained from a deep AE. RAPP compares input and its AE or VAE reconstruction in the hidden spaces as well as in the input space. RaPP introduces two metrics combining those hidden activated values to measure novelty scores. In order to achieve this, RaPP requires the model to be symmetric to enforce its evaluation methodologies, which causes its structural limitation, and training model becomes a very expensive work as the data becomes more complicated due to fully-connected layers in encoder and decoder caused by their structural problem.

Training Details

All of our experiments were conducted by Python 3.6.9. Adam optimizer was adopted to train our model. For the stable adversarial learning, the encoder is trained with one iteration after every five iterations for the discriminator. Each experiment is carried out with 100 epochs with batch size as much as 64, and we set learning rate as 0.0004. Gaussian noises with standard deviation of 0.02 were added to the input image data at the training phase.

Experimental Results

We evaluate the performances of all experiments using Area Under the Receiver Operating Characteristic curve (AUROC) as shown in Table 1.

Discussion

Our methods showed a better performance than other previous GAN-based state-of-the-art novelty detection algorithms such as OCGAN, RaPP. Specifically, our approach provides a much higher AUROC values for experiments on more complicated data such as f-MNIST, CIFAR-10. It supports that as a tool of novelty score measurement, change of latent vector is more reasonable than image reconstruction errors since image reconstruction error can be more escalated in more complicated data. In training level, our approach leverages on class labels in normal dataset, which is sometimes a expensive work. Unsupervised learning framework without using normal class labels can be considered potentially.

Conclusion

We proposed a new novelty detection framework using deep generative models. Instead of evaluating novelty class using image reconstruction error, the change of angle in latent vector is regarded as a tool for novelty detection quantity. We adopt OLE loss using mutual class information to achieve disentanglement of latent vectors to maximize the effect of class information. Our new approach shows a greater performance in multi-modal normality scenarios than previously existing GAN based state-of-the-art novelty detection algorithms.

Acknowledgments

The work was partially supported by NSF through grants IIS-1838207, CNS 1901218, OIA-2040680, OIA-2134901 and USDA-020-67021-32855.

References

An, J., and Cho, S. 2015. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE* 2(1):1–18.

Candès, E. J.; Li, X.; Ma, Y.; and Wright, J. 2011. Robust principal component analysis. *Journal of the ACM (JACM)* 58(3):1–37.

Fazel, S. M. 2003. Matrix rank minimization with applications.

Kim, K. H.; Shim, S.; Lim, Y.; Jeon, J.; Choi, J.; Kim, B.; and Yoon, A. S. 2019. Rapp: Novelty detection with reconstruction along projection pathway. In *International Conference on Learning Representations*.

Lezama, J.; Qiu, Q.; Musé, P.; and Sapiro, G. 2018. Ole: Orthogonal low-rank embedding-a plug and play geometric loss for deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8109–8118.

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

Perera, P.; Nallapati, R.; and Xiang, B. 2019. Ocgan: One-class novelty detection using gans with constrained latent representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2898–2906.

Pidhorskyi, S.; Almohsen, R.; and Doretto, G. 2018. Generative probabilistic novelty detection with adversarial autoencoders. In *Advances in neural information processing systems*, 6822–6833.

Qiu, Q., and Sapiro, G. 2015. Learning transformations for clustering and classification. *The Journal of Machine Learning Research* 16(1):187–225.

Recht, B.; Fazel, M.; and Parrilo, P. A. 2010. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review* 52(3):471–501.

Sakurada, M., and Yairi, T. 2014. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, 4–11.

Schlegl, T.; Seeböck, P.; Waldstein, S. M.; Schmidt-Erfurth, U.; and Langs, G. 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, 146–157. Springer.