A TWO-CLASS HYPER-SPHERICAL AUTOENCODER FOR SUPERVISED ANOMALY DETECTION

Yuta Kawachi, Yuma Koizumi, Shin Murata, and Noboru Harada

NTT Media Intelligence Laboratories, Tokyo, Japan

ABSTRACT

Supervised anomaly detection has been a tough problem due to its necessity of special handling of unseen anomalies. In this paper, we present a heuristic implementation of variational auto-encoder with von-Mises Fisher prior applied to a supervised anomaly detector. The closed latent space like sphere is suitable for detecting unseen anomalies because we have a possibility to "fill" the space with seen training samples. If it ideally works, the reconstruction error will be high for all unseen anomalies. Experiments show that our model can separate normal and anomaly samples in the spherical latent space. It is also shown that he proposed model improves the performance for seen anomalies without degrading the performance for unseen anomalies.

***Index Terms*—** Anomaly detection, auto-encoder, von Mises-Fisher distribution

1. INTRODUCTION

Anomaly detection [1, 2] can be regarded as a binary classification problem [3, 4]. It is often the case with absence of anomalous data during data collection in real-world tasks, and hence, anomaly detection is grounded on outlier detection using only normal data [5, 6]. Widely used anomaly detection methods, for example, are one-class support vector machines [7] and support vector data descriptions (SVDD) [8]. SVDD with negative examples (SVDD-neg) [9], which is a supervised extension of SVDD with (small number of) negative examples, is also used when the anomaly labeled data is available. Auto-encoder based methods are also used for the tasks [10, 11].

In practical situations, small amount of anomalous data can be available during data collection. In this case, an anomaly detection method can be extended to a supervised binary classification problem. However, the supervised classification algorithm cannot be applied straightforwardly because of the following three problems: (i) imbalanced data, (ii) labeling cost, and (iii) presence of unseen anomalies [12]. In this paper, we tried to solve the last problem: we propose a solution for improving the detection performance for seen anomalies without degrading detection performance for unseen anomalies.

Simple supervised models have not been used to anomaly detection because the unseen (unknown) samples must be also classified as anomaly and this makes the problem difficult [13, 14]. Simple supervised models make classification surface irrelevant to such unseen samples which are far from the classes used to model training. In order to detect unseen samples as anomaly, the neural regression models with reconstruction error, such as auto-encoders (AE) or variational auto-encoders (VAE), are empirically known to detect them well and frequently used in real task [15–21]. If we combine the conventional reconstruction loss (RL-loss) based framework and another supervised classification framework, the detection performance for both seen and unseen anomalies should be high.

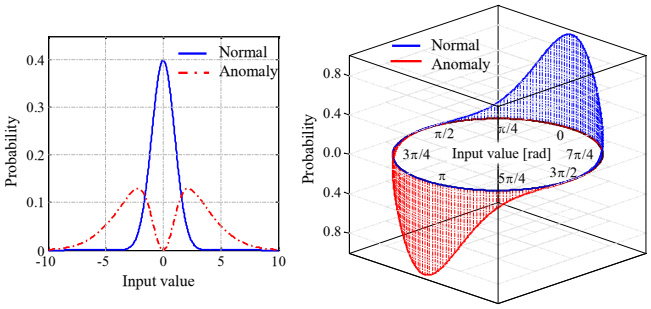

Fig. 1: Prior distribution for CS-VAE on the Euclidean space (left), and proposed 2C-vMF-VAE on hyper-spherical space (right).

In our previous work, we have proposed a supervised extension of VAE-based anomaly detector called complementary-set VAE (CS-VAE) [12]. It splits the latent space into two classes to detect normal and (seen) anomaly. In the CS-VAE, the prior distribution for the (seen) anomaly class is formulated as a complementary set of the that of normal class. However, this method has a flat latent space as same as the standard VAE. This results in the problem that we cannot define probability density function (PDF) on each class straightforwardly because the flat latent space has infinite area.

In this work, we present a VAE with (hyper-) spherical latent space. If we choose sphere as the closed latent space, we can split the surface into two finite areas. As a PDF on the latent space of VAE, we use the von Mises-Fisher (vMF) distribution to represent normal and anomaly then use the training method similar to [12], then we can construct the anomaly detector. Since the calculation of the KL divergence between two vMF distributions is not simple [22], the calculation of its gradient might also be complex and result in unstable training. For stability of the training of VAE, we formulate an approximated Kullback-Leibler (KL)-divergence between two vMF distributions. The prior distributions for the CS-VAE on a flat latent space [12] and for the two class vMF-VAE (2C-vMF-VAE) on a closed latent space (proposed) are shown in Fig.1. It should be noted that the concept of the complementary-set prior distribution represented on a flat latent space (left) can be translated into a simple distribution when represented on a closed latent space (right).

2. CONVENTIONAL VAE-BASED ANOMALY DETECTION

2.1. VAE for anomaly detection

In anomaly detection, the VAE is widely used to construct generative model of "normal data" $\boldsymbol{x} \in \mathbb{R}^D$ [23]. The probability of $\boldsymbol{x}$ can be

described with latent variable $\boldsymbol{z} \in \mathbb{R}^Q$ introduced as

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}.$$

To obtain more "realistic" model, we try to maximize evidence $\log p(\boldsymbol{x})$. However, the integration over the latent variable $\boldsymbol{z}$ is usually intractable. Hence, evidence lower bound (ELBO) is maximized instead of $\log p(\boldsymbol{x})$ itself. The ELBO is defined as

$$\text{ELBO}[\boldsymbol{x}] = \mathbb{E}_{q(\boldsymbol{z})}[\log p(\boldsymbol{x}|\boldsymbol{z})] - \text{KL}[q(\boldsymbol{z})||p(\boldsymbol{z})], \quad (1)$$

where $q(\boldsymbol{z})$ is an arbitrary distribution (it may depends on $\boldsymbol{x}$) over $\boldsymbol{z}$, $\mathbb{E}_{q(\boldsymbol{z})}$ denotes the expectation over $q(\boldsymbol{z})$, and $\text{KL}[q(\boldsymbol{z})|p(\boldsymbol{z})]$ denotes the Kullback-Leibler (KL) divergence between $q(\boldsymbol{z})$ and $p(\boldsymbol{z})$.

In the VAE setting, encoder and decoder are introduced to construct $q(\boldsymbol{z})$ and $p(\boldsymbol{x}|\boldsymbol{z})$, respectively. Parameters of the encoder and decoder, ψ_E and ψ_D, are trained to maximize the ELBO. When the encoder and decoder are trained properly, normal data is expected to increase the evidence. Lower value of the evidence implies that the data is not likely normal, i.e. anomal. Therefore, the ELBO, log likelihood, and/or KL divergence can be a candidate of anomaly score.

The choice of $p(\boldsymbol{x}|\boldsymbol{z})$, $p(\boldsymbol{z})$, and $q(\boldsymbol{z})$ is important to use the VAE as anomaly detector. In order to calculate the KL divergence in eq. (1) analytically and to sample $\boldsymbol{z}$ from $q(\boldsymbol{z})$ easily, the gaussian distributions are often chosen for $p(\boldsymbol{z}) \sim \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I})$ and $q(\boldsymbol{z}) \sim \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\sigma})$. When $\boldsymbol{x}$ is given, the encoder estimates the mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}$, that is,

$$\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{x}; \psi_E),\ \boldsymbol{\sigma} = \boldsymbol{\sigma}(\boldsymbol{x}; \psi_E). \quad (2)$$

Note that, the KL divergence in eq. (1) is now the KL divergence between two Gaussian distribution as follows:

$$\text{KL}[\mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\sigma})||\mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{1})] \quad (3)$$

When $p(\boldsymbol{x}|\boldsymbol{z})$ is also Gaussian distribution, the expectation of log likelihood of the distribution can be approximated by:

$$\mathbb{E}_{q(\boldsymbol{z})}[\log p(\boldsymbol{x}|\boldsymbol{z})] \approx -\frac{1}{2N}\sum_{i=1}^{N}[\boldsymbol{x} - \boldsymbol{x}'(\boldsymbol{z}_i; \psi_D)]^2, \quad (4)$$

where, $\boldsymbol{x}'(\boldsymbol{z}_i; \psi_D)$ denotes reconstruction by the decoder given the latent variable $\boldsymbol{z}_i$ sampled from $q(\boldsymbol{z})$. Note that, when $p(x|z)$ is Gaussian, the maximization of ELBO can be reduced to the minimization of $\text{RL}[x] + \text{KL}[q(z)||p(z)]$, where $\text{RL}[x]$ denotes the reconstruction (RL) loss.

2.2. Complementary-set VAE (CS-VAE) for anomaly detection

The previous subsection, the case without anomalous sample during training period is mainly concerned. When some anomalous samples are obtained, i.e. seen anomalies, it would be better to train ψ_E and ψ_D with those samples. The anomaly can be defined as a complement of the normal set [6]. The probability for anomaly is shown in Fig. 1 (left) and is formulated as follows:

$$\mathcal{C}(\boldsymbol{z}; s) = \prod_{q=1}^{Q}\mathcal{N}(z_q; 0, s^2)\left[\frac{1}{\sqrt{2\pi}} - \mathcal{N}(z_q; 0, 1)\right]. \quad (5)$$

While KL divergence for normal samples are still eq. (3), KL divergence for anomalous samples are now defined as

$$\text{KL}_+[\mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\sigma})||\mathcal{C}(\boldsymbol{z}; s)]. \quad (6)$$

This KL-divergence can be analytically calculated [12]. By training ψ_E and ψ_D using both (3) and (6), it would be expected that the KL divergence decreases for "seen" anomalies, and the reconstruction loss increases for "unseen" anomalies.

3. PROPOSED METHOD

3.1. Two class von Mises-Fisher VAE for anomaly detection

In this paper, we consider the prior PDFs on hyper-spherical space. There are several DNNs that handle spherical space, such as spherical CNN [24], and those DNNs are applied to language modeling applications [25]. In this paper, we adopt the von Mises-Fisher distribution as a prior distribution. Note that, there is a prior work of a VAE with a spherical uniform prior [26]. The major differences between this prior work and the proposed method are shown below:

Prior selection: we selected vMF distribution as prior PDFs because we need to concentrate the representations of normal and anomaly in the latent spherical space to classify seen anomaly in this space.

Reparametrization trick: our approach emits standard Eular angle under a Gaussian-distributed assumption. We also use a rotation matrix and a fixed vector to produce the vectors concentrated on a point on the hypersphere. Our approach does not require additional sampling procedure like [26].

In the proposed method, as show in Fig. 1 (right), vMF distributions with mean direction $\boldsymbol{\mu}_0^-$ and $\boldsymbol{\mu}_0^+$ are used for the normal and anomaly prior PDFs, respectively. The same concentration parameter κ_0 is used for each distribution, and $\left\|\boldsymbol{\mu}_0^-\right\|_2 = \left\|\boldsymbol{\mu}_0^+\right\|_2 = 1$. In following discussion, we describe the calculation procedures of RL-loss and KL-loss on vMF prior PDFs. In this paper, for simplifying the discussion, we explain the 3-D latent space (spherical surface) case.

Reconstruction loss: In the proposed method, the reconstruction loss is defined as the squared-error in the same manner as the Gaussian-VAE. In contrast to the Gaussian-VAE, the encoder emits the Eular angle variables $\boldsymbol{\theta} = (\alpha, \beta, \gamma)^\top$ and a variance parameter σ as

$$\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{x}; \psi_E), \sigma = \sigma(\boldsymbol{x}; \psi_E). \quad (7)$$

Next, a noise $\boldsymbol{\epsilon}$ is added in the way same as the Gaussian-VAE as

$$\boldsymbol{\theta}' = \boldsymbol{\theta} + \exp(\sigma)\boldsymbol{\epsilon}, \quad (8)$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; \boldsymbol{0}, \boldsymbol{1})$. Then, a rotation matrix $\mathbf{R}(\boldsymbol{\theta})$ is composed from the noised angle variables $\boldsymbol{\theta}'$, and a fixed vector $\boldsymbol{v}$ is rotated using this matrix as

$$\boldsymbol{z}' = \mathbf{R}(\boldsymbol{\theta}')\boldsymbol{v}, \quad (9)$$

where $\|\boldsymbol{v}\| = 1$ is an arbitrary vector. Finally, the rotated vector is inputted to the decoder and the reconstruction error is calculated as

$$\text{RL}[\boldsymbol{x}] = \|\boldsymbol{x} - \boldsymbol{x}'\big(\boldsymbol{z}'; \psi_D)\big]\|_2^2. \quad (10)$$

This reparametrization process is clearly differentiable because the rotation matrix is composed of trigonometric functions therefore back-propagation is possible.

KL-divergence loss: Since the calculation of the KL divergence between two vMF distributions is not simple, the calculation of its gradient might also be complex and result in unstable training. Thus, we defined an alternative of the true KL divergence based on the cosine similarity. First, the direction vector $\mathbf{R}(\boldsymbol{\theta})\boldsymbol{v}$ is gained by rotating the fixed vector without noise $\boldsymbol{\epsilon}$. Then, the cosine similarities between the direction vector and the mean directions of normal and anomalous prior are calculated as

$$\phi^- = \boldsymbol{\mu}_0^- \cdot \mathbf{R}(\boldsymbol{\theta})\boldsymbol{v}, \quad (11)$$

$$\phi^+ = \boldsymbol{\mu}_0^+ \cdot \mathbf{R}(\boldsymbol{\theta})\boldsymbol{v}, \quad (12)$$

Algorithm 1 vMF-VAE loss calculation

Input: $\boldsymbol{x}$, $\boldsymbol{\mu}_0$ ($= \boldsymbol{\mu}_0^-$ or $\boldsymbol{\mu}_0^+$), κ_0,$\boldsymbol{v}$, λ
Output: $\mathrm{loss}_{\mathrm{RL}} + \lambda \mathrm{loss}_{\mathrm{KL}}$
  $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{x}; \psi_E), \sigma = \sigma(\boldsymbol{x}; \psi_E)$
  $\mathrm{loss}_{\mathrm{KL}} \leftarrow \exp(2\sigma) - \kappa_0 \boldsymbol{\mu}_0^\top \mathbf{R}(\boldsymbol{\theta})\boldsymbol{v}$
  $\mathrm{loss}_{\mathrm{RL}} \leftarrow 0$
  for $k \in [1, \cdots, K]$ **do**
    $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, I)$
    $\boldsymbol{\theta}'_k = \boldsymbol{\theta} + \exp(\sigma)\boldsymbol{\epsilon}_k$
    $\mathrm{loss}_{\mathrm{RL}} \leftarrow \mathrm{loss}_{\mathrm{RL}} + \frac{1}{K}\|\boldsymbol{x} - \boldsymbol{x}'(\mathbf{R}(\boldsymbol{\theta}'_k)\boldsymbol{v}; \psi_D)\|_2^2$
  end for

respectively. Then, the KL-loss for normal is defined as

$$\mathrm{KL}[\boldsymbol{x}] = \kappa_1 - \kappa_0 \phi^-, \tag{13}$$

where $\kappa_1 = \exp(2\sigma)$. In the same manner, that for anomaly is defined as

$$\mathrm{KL}_+[\boldsymbol{x}] = \kappa_1 - \kappa_0 \phi^+. \tag{14}$$

This definition of KL-divergences are simplified version of the upper bound of the KL-divergence between two vMF distiribution [22]. When $d = 3$ in Eq. (13) of [22], the upper bound of the KL-divergence can be written as Eq. (13).

3.2. Comparison with CS-VAE

As we can see in Fig. 1, the definition of complemental sets causes overlap between normal and anomaly. This may leads to false detections near the intersections. (This may be important property in case of real-world applications and needs to be studied more in detail.) In contrast, the proposed prior tends to assign one "pole" to the normal and the other to anomaly. Therefore, the intersection rarely occurs.

Other advantage of the vMF prior arises when the complements is concerned. The complements of vMF distribution can be defined straightforwardly because of its closure property. However, it is difficult to calculate analytically the KL divergence of those complemental distribution. This is an important future work.

3.3. Implementation

We describe the detail of the training procedure of the proposed method. Training method which optimizes the generative objectives for the two classes at the same time is not obvious. Our prior work [12] tried to optimize the two classes alternately by the epoch. In this procedure, we need to choose which epoch (normal or anomaly) is better to stop the optimization. Also there is another problem that the convergence is not so clear. To overcome this problem, we created a normal and anomaly paired batch. The paired batch consists of a pair of randomly selected normal and anomalous samples. Then, the loss defined by **Algorithm 1** is calculate for both normal and anomaly in each batch. Finally, the parameters of the encoder and decoder are updated to minimize the loss.

4. EXPERIMENTS

4.1. Comparison methods

We implemented these models for this experiment.

- AE (RL): Unsupervised two units bottleneck layer autoencoder with reconstruction loss anomaly score. The training data only contain the normal data.

Table 1: Experiment conditions.

# of hidden units / MLPs	2, 10, 100, 300, 500, 700, 1000, 2000
batch size	100
# of epochs	200
normal prior vector $\boldsymbol{\mu}_0^-$	$(0, 0, +1)^\top$
anomaly prior vector $\boldsymbol{\mu}_0^+$	$(0, 0, -1)^\top$
κ prior κ_0	1
CS-VAE KL coff. C (train)	10 (only for anomaly)
vMF-VAE KL coff. C (train)	10 (normal, anomaly)
# of Monte-Carlo sampling (VAE, CS-VAE)	1 (train), 3 (test)
dataset definitions	
normal (train)	MNIST 1, 2, 3 [t123]
seen anomaly (train)	MNIST 4, 5, 6 [t456]
normal (eval.)	MNIST 1, 2, 3 [e123]
seen anomaly (eval.)	MNIST 4, 5, 6 [e456]
unseen anomaly (eval.)	MNIST 7, 8, 9 [e789]
unseen anomaly (eval.)	1000 samples from Omniglot [eOmn]

- VAE (RL): Unsupervised two latent variables variational autoencoder with reconstruction loss anomaly score [15]. The training data only contain the normal data.
- CS-VAE(RL/KL/ELBO): Two-class supervised flat (two latent variables) latent space variational autoencoder with three kinds of anomaly score [12]. The training data contain both normal and anomaly data.
- 2C-vMF-VAE(RL/KL/ELBO): Two-class supervised spherical latent space variational autoencoder with three kinds of anomaly score. The training data contain both normal and anomaly data (proposed).

All models contain encoder and decoder which are conventional single hidden layer perceptrons. The CS-VAE uses the standard Gaussian-VAE KL divergence term for normal data. For anomaly data, the model uses the cost function derived from the KL divergence from the anomaly distribution. We compared the anomaly detection stability and seen/unseen anomaly detection performance of these models. To train and test these methods, we used MNIST [27] and Omniglot [28] dataset. The all data are converted into 784 dimensions, as same as the MNIST dataset, in advance. Other conditions are listed in Table 1.

4.2. Spherical latent space aquisition

To investigate whether the 2C-vMF-VAE is trained so as to separate normal and anomaly in the latent space, we decoded and visualized the latent space of the 2C-vMF-VAE in Fig. 2. Normal digits were place at around the north area, and anomalous digits were placed at around south area. These patterns shows us that the two-class spherical representation in the latent space is properly acquired.

4.3. Objective evaluation

We conducted objective evaluations of these models. The results are evaluated in the area-under-the-receiver-operating-characteristic-curve (AUC) scores. Evaluation data sets (denoted as [e123], [e456], [e789] and [eOmn]) different from training data sets (denoted as [t123] and [t456]) are used. In Figs. 3 to 5, the AUC scores corresponding to eight conditions for the number of hidden units listed on Table 1 are calculated and displayed in box plots.

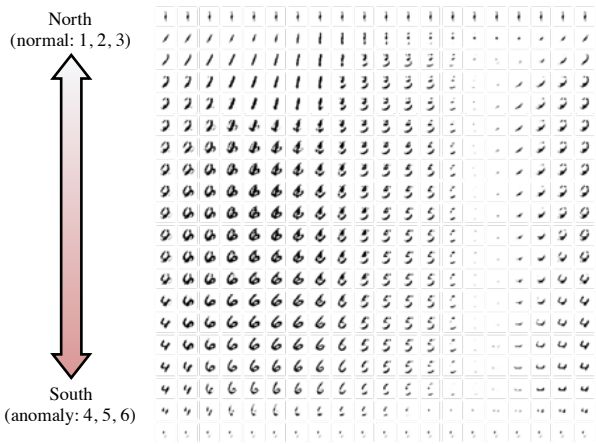

Fig. 2: Latent space visualization using 2C-vMF-VAE decoder (Mercator projection). Normal digits are 1, 2, and 3 [t123], and seen anomalous digits are 4, 5, and 6 [t456]. Mean direction of normal is north, and that of anomaly is south.

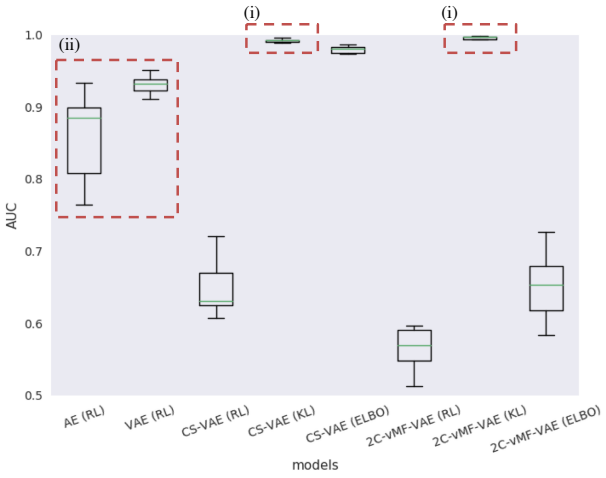

Fig. 3: Detection performance of seen MNIST anomalies (i.e. normal: [e123] vs. anomaly: [e456]).

Seen anomaly detection: Fig. 3 shows the result of seen anomaly detection ([e123] vs. [e456]). Marked as (i) in this figure, we can see use of KL-loss as anomaly score achieved highest performance on both supervised anomaly detection methods. In addition, marked as (ii), both supervised methods significantly improved the performance of AUC score for seen anomalies than the unsupervised methods. Thus, our supervised methods have efficiently worked for detection of seen anomalies.

Unseen anomaly detection: Fig. 4 and 5 show the results for unseen anomaly detection for similar patterns (i.e. digits, but unseen anomalies) and significantly different patterns (i.e. non-digits), respectively. In the similar patterns case shown in Fig. 4, we could not obtain expected results. In 2C-vMF-VAE, the AUC score of KL-loss was still higher than that of RL-loss (i), even though KL-loss score was almost same to that of the unsupervised method, i.e. VAE (ii). This is because that the proposed method used closed latent space and the similar anomalies might be projected closer to the normal area. Thus, unseen anomaly also has small reconstructing error. This result indicates that additional learning for overlooked anomalies is required. On the other hand, in the significantly-different-

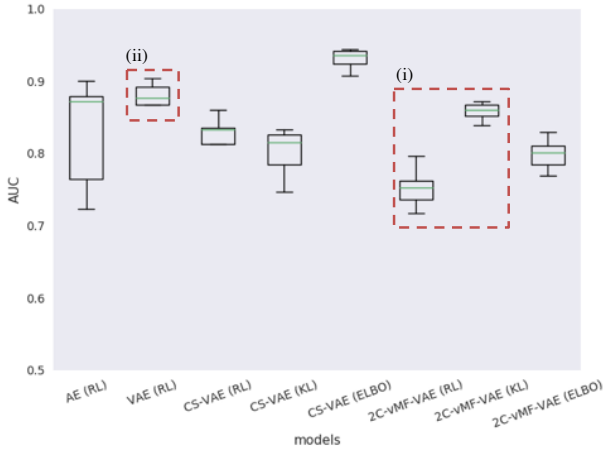

Fig. 4: Detection performance of unseen MNIST anomalies (i.e. normal: [e123] vs. anomaly: [e789]).

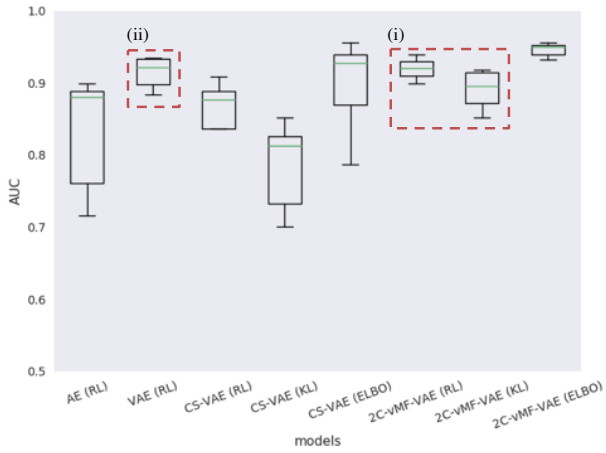

Fig. 5: Detection performance of unseen Omniglot character anomalies (i.e. normal: [e123] vs. anomaly: [eOmn]). The models were trained with MNIST dataset, i.e. [t123] and [t456].

anomalous-pattern case shown in Fig. 5, we obtained the expected results. Namely, the AUC score of RL-loss were higher than that of KL-loss (i), and RL-loss score was almost same to that of the VAE (ii). Thus, in the significantly-different-anomalous-pattern case, the proposed method achieved the purpose of supervised anomaly detector and it improved the performance for seen anomalies without degrading performance for unseen anomalies.

5. CONCLUSIONS

We proposed a von-Mises Fisher variational auto-encoder for constructing a two-class anomaly detector. For stability of the training of VAE, we formulate an approximated KL-divergence between two vMF distributions. In experiments, we confirmed that the proposed 2C-vMF-VAE model can properly acquire spherical two-class latent space and detect seen anomalies better than classical VAE. In addition, we confirmed reasonable results in significantly different anomalous pattern case. Thus, we conclude that the proposed method is effective for supervised anomaly detector. It improves the performance for seen anomalies without degrading the performance for unseen anomalies.

6. REFERENCES

[1] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial intelligence review*, vol. 22, no. 2, pp. 85–126, 2004.

[2] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, July 2009.

[3] I. Steinwart, D. Hush, and C. Scovel, "A classification framework for anomaly detection," *Journal of Machine Learning Research*, vol. 6, no. Feb, pp. 211–232, 2005.

[4] I. Steinwart, D. Hush, and C. Scovel, "Density level detection is classification," in *Proc. NIPS*, 2005, pp. 1337–1344.

[5] Y. Koizumi, S. Saito, H. Uematsu, and N. Harada, "Optimizing acoustic feature extractor for anomalous sound detection based on Neyman-Pearson lemma," in *Proc. EUSIPCO*. EURASIP, 2017.

[6] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, and N. Harada, "Unsupervised detection of anomalous sound based on deep learning and the Neyman-Pearson lemma," *IEEE Trans. ASLP*, 2018.

[7] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[8] D. M. J. Tax and R. P. W. Duin, "Data domain description using support vectors," in *Proc. ESANN*, 1999, vol. 99, pp. 251–256.

[9] D. M. J. Tax, *One-class classification*, Ph.D. thesis, Delft University of Technology, 2001.

[10] S. Zhai, Y. Cheng, W. Lu, and Z. Zhang, "Deep structured energy based models for anomaly detection," in *Proc. ICML*, 2016.

[11] A. Munawar, P. Vinayavekhin, and G. De Magistris, "Limiting the reconstruction capability of generative neural network using negative learning," in *Proc. MLSP*, 2017.

[12] Y. Kawachi, Y. Koizumi, and N. Harada, "Complementary set variational autoencoder for supervised anomaly detection," in *Proc. ICASSP*, 2018.

[13] N. Görnitz, M. Kloft, K. Rieck, and U. Brefeld, "Toward supervised anomaly detection," *Journal of Artificial Intelligence Research*, vol. 46, no. 1, pp. 235–262, Jan. 2013.

[14] B. Du and L. Zhang, "A discriminative metric learning based anomaly detection method," *IEEE Trans. Geoscience and Remote Sensing*, vol. 52, no. 11, pp. 6844–6857, 2014.

[15] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," 2015.

[16] Y. Ma, P. Zhang, Y. Cao, and L. Guo, "Parallel auto-encoder for efficient outlier detection," in *Proc. IEEE BigData*. IEEE, 2013, pp. 15–17.

[17] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in *Proc. MLSDA*, New York, NY, USA, 2014, MLSDA'14, pp. 4:4–4:11, ACM.

[18] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G. Shroff, "LSTM-based encoder-decoder for multi-sensor anomaly detection," in *Proc. ICML*, 2016.

[19] R. C. Aygun and A. G. Yavuz, "Network anomaly detection with stochastically improved autoencoder based models," in *Proc. IEEE CSCloud*, June 2017, pp. 193–198.

[20] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," in *Proc. SIGKDD*, New York, NY, USA, 2017, KDD '17, pp. 665–674, ACM.

[21] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," in *Proc. ICLR*, 2018.

[22] T. Diethe, "A note on the kullback-leibler divergence for the von mises-fisher distribution," *arXiv pre-print, arXiv:1502.07104*, 2015.

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

[24] T. S. Cohen, M. Geiger, J. Koehler, and M. Welling, "Spherical CNNs," in *Proc. ICLR*, 2018.

[25] K. Guu, T. B. Hashimoto, Y. Oren, and P. Liang, "Generating sentences by editing prototypes," *Trans. of the Association for Computational Linguistics*, 2018.

[26] T. R. Davidson, L. Falorsi, N. De Cao, T. Kipf, and J. M. Tomczak, "Hyperspherical variational auto-encoders," in *Proc. UAI*, 2018.

[27] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[28] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.