creative commons

C O M M O N S D E E D

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

Master's Thesis

Self-supervised auto-encoder for anomaly detection

Wonbin Kim (김 원 빈)

Department of Computer Science and Engineering

Pohang University of Science and Technology

2019

dCollection @ postech

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

이상 탐지를 위한 자율-지도 오토인코더

Self-supervised auto-encoder for anomaly detection

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

Self-supervised auto-encoder for anomaly detection

by

Wonbin Kim
Department of Computer Science and Engineering
Pohang University of Science and Technology

A thesis submitted to the faculty of the Pohang University of Science and Technology in partial fulfillment of the requirements for the degree of Master of Science in the Computer Science and Engineering

Pohang, Korea
12. 21. 2018
Approved by
Seung Jin Choi
Academic advisor

Self-supervised auto-encoder for anomaly detection

Wonbin Kim

The undersigned have examined this thesis and hereby certify that it is worthy of acceptance for a master's degree from POSTECH

01. 02. 2019

Committee Chair	Seung Jin Choi
Member	Hwan Jo Yu
Member	Su Ha Kwak

MCSE 김 원 빈. Wonbin Kim
20172498 Self-supervised auto-encoder for anomaly detection,
이상 탐지를 위한 자율-지도 오토인코더
Department of Computer Science and Engineering , 2019,
33p, Advisor : Seung Jin Choi. Text in English.

ABSTRACT

Anomaly detection is to find some unique patterns of a subset of the data that is pre-defined as "normal" by human or is the majority, which allows to detect "anomaly" that is the complement of the set by similarity measures. Recent deep anomaly detection methods use a score measure based on reconstruction loss to measure anomality. However if the dataset is heterogeneous and some of the classes share some properties with the possible anomalous, then the model can be confused to distinguish the normal and the anomalous by only learning single task. To alleviate this, we proposed Self-Supervised Autoencoder that learns reconstruction and classification by self-orgarnized supervision, jointly. Our approach has following advantages: 1) By learning not related directly two tasks, the model can explore different two data spaces of the data, so both tasks cooperate to detect anomalies. 2) Joint architecture improves performance from simply ensembling two models. By the experiment results on three benchmark image dataset, we demonstrate the advantage and the effectiveness of our proposed model.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

Contents

I. Introduction **1**

II. Related Work **5**

III. Proposed Method **6**

- 3.1 Problem Formulation 6
- 3.2 Image Rotation Prediction 7
  - 3.2.1 Dataset Generation 7
  - 3.2.2 Prediction 8
- 3.3 Pipeline 10
- 3.4 Anomaly Score 11

IV. Experiment **14**

- 4.1 Dataset 14
- 4.2 Configuration 14
- 4.3 Baseline 15
- 4.4 Implementation Detail 15

V. Result **18**

- 5.1 MNIST 18
- 5.2 Fashion MNIST 21
- 5.3 CIFAR10 23
- 5.4 Both tasks affect to each other? 26

VI. Conclusion **28**

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

Summary (in Korean) 29

References 30

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

I. Introduction

Anomaly Detection is to solve the task that the model learns the common features of the data regarded as *NORMAL* with the constraint that the model can obtain only the normal, then identifying rare events or human-determined exception defined as *ANOMALY* by some similarity meaurement. Since the dataset is so skewed towards the normal that the agent cannot learns an accurate decision boundary between the normal and the anomaly, the methods for anomaly detection are expected to behave abnormally only if the input is anomalous.

The problem has following challenges [1]. 1) Anomalies infrequently occur, i.e., compared to normal data, the anomalies are difficult or impossible to observe, the dataset is inevitably imbalanced. 2) Determining definition of the normal region is difficult and ambiguous. i.e., an object may have plenty features that describe it, but some key features that divide data into the normal and the anomalous are partial and may have complex relations. No reference exists by which to distinguish normal data from anomalous data except the supervised one, so selecting and extracting key features are difficult tasks. 3) The definition of the anomaly is not stationary. Although the dataset is the same, the definition of anomaly can be changed when the domain is changed. 4) Data with noise can be regarded as anomalous. 5) Additionally, distinct data can have the same semantic information on image dataset.

So far, anomaly detection problem have been studied on the assumption that the normal data consists of homogeneous dataset[2, 3, 4]. But anomaly detection on heterogeneous normal dataset is more practical, because a large number of image samples are required to train a deep model, and there are likely to be some hidden classes in the samples even if human expert cannot notice. Heterogeneous

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

normal dataset is more complex problem than homogeneous normal dataset. The expressive power of a neural network is powerful, but there is no theoretical guideline to determine network architecture for the appropriate expressive power so that the network have overflowing the expressive power. As a result, the network returns an output that corresponds to a sample of a class that is similar to the anomalous sample.

Recent research on the heterogeneous problem leverages this property. Reconstruction based model learn a kind of identity function $f_R(x) = x$, then it detects anomalous sample by reconstruction error $\mathcal{L}_R(x) = ||f_R(x) - x||_2^2$. In reconstruction pipeline, anomalous sample is reconstructed to the sample learned by the model or to the weird sample. An *et al.* [5] uses Variational Auto-encoder to estimate density by approximating expected negative log-likelihood (or reconstruction probability) of each data point. Zenati *et al.* [6] trains BIGAN[7, 8], then estimates the anomaly score by combining the reconstruction error and the discriminator loss. Akcay *et al.* [9] use a discriminator to trains an auto-encoder adversarially, and use another encoder to regress bottleneck features from reconstructed images, then estimates anomaly score by measuring distance between both bottleneck features. Aytekin *et al.* [10] clusters bottleneck feature after training an auto-encoder with $l2$ normalizing, then detects anomaly. But they use extra information which is the number of classes, we regarded this as a kind of supervision and thought that [10] is not relevant to us.

We believe the reason why the above methods show low performance is that data a single property of data can be shared on anomalies and one of the underlying classes, so these methods that are concentrated on learning single particular task can lead to failures on multi-class normal dataset. For example, suppose that one digit is anomalous class and other digits are normal class on MNIST [11], we can make ten dataset as others [5, 6, 9]. Thus the reconstruction

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

based method can detect anomalies on some configuration, but not on other configuration (Fig. 1.1).

We propose a new framework that learns the reconstruction task and unrelated another task for capturing features from different perspectives jointly. To introduce other tasks, We were inspired by Self-supervised [12] methods and created self-labeled dataset by specific task from the target dataset in unsupervised fashion. Thus our framework can model the given dataset with supervision. We showed that 1) our framework is that the model learns different tasks that are not related to each other, so can exploit more plentiful features to learn about given data, 2) coupling and jointly learning improves the ability to detect anomalies from simple ensemble model of two tasks.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

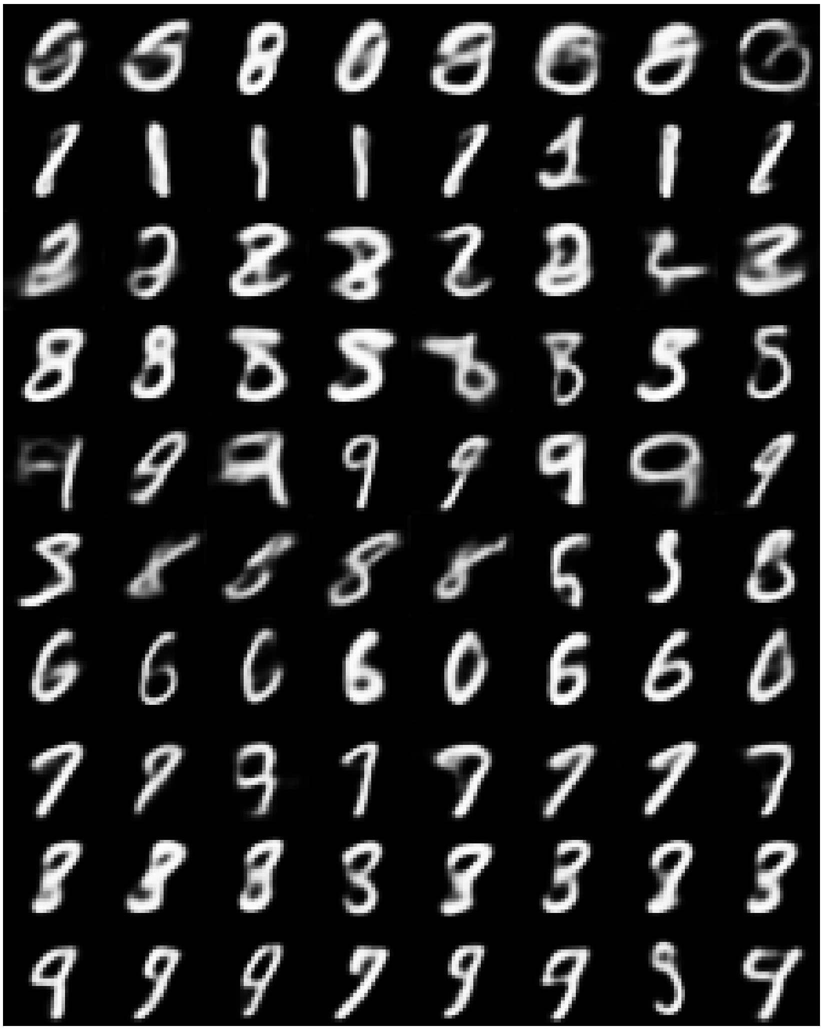


Figure 1.1: The reconstruction of anomalous images on our proposed model. Each row: Reconstruction of digit ’0’ to digit ’9’ respectively in the configuration that one digit is anomalous and other is normal. Almost images have weird shape or represent other digit but few are reconstructed the original, but all digit ’1’s are reconstructed as ’1’. It imply that learning particular reconstruction is inappropriate to distinguish anomalous instances on multi-class configuration.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

II. Related Work

Deep generative model based anomaly detection There are three types of recent deep learning methods for anomaly detection: 1) Reconstruction based model [6, 9], such as auto-encoder, learns an identify function $f_R(x) = x$ by using some symmetric architecture, e.g., an encoder and a decoder, with constraint, e.g., the dimensionality of output of intermediate layer is less than of the corresponding input, so that the model does not learn a trivial solution. Then the model detects anomalies by the difference between original sample and reconstructed sample. 2) Density estimation based [4, 13, 5] model finds a distribution that describes given data well. The main idea of the model is that a sample with low density or high energy is the anomaly in high probability. 3) GAN-base model [6] trains a discriminator with a generator in adversarial method until converge both to Nash Equilibrium. The model detects anomalies by measuring similarity with feature matching [14] score between the input and the corresponding sample that is reconstructed one or by discriminating with discriminator score.

Self-Supervised Learning [12] The framework inspired by ”context” of Natural Language Process (NLP) is one of unsupervised learning method for computer vision. By doing a specific task without no extra information on data and pairing processed data and the corresponding labels for data, a learner is given the self-labeled dataset to learn self-supervised tasks, such as finding a image patch relative position [12], inpainting removed region [15], solving jigsaw puzzle [16], predicting rotation of image [17] and giving noise [18]. These tasks learns generalized feature that can be useful to other application, e.g., classification, semantic segmentation, and object detection.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

III. Proposed Method

3.1 Problem Formulation

Problem Definition In this thesis, only images are considered for anomaly detection problem. Let $\mathcal{X}$ be the space of all ordinary images, and $X \subseteq \mathcal{X}$ be the set of images. We have a training set of image data $\mathcal{D}^{train}$ with

$$\mathcal{D}^{train} := \{(x_i, y_i) : x_i \in X^{(N)}\}_{i=1}^{N_{train}}, \tag{3.1}$$

and also have a test set of image data $\mathcal{D}^{test}$ with

$$\mathcal{D}^{test} := \{(x_i, y_i) : x_i \in X\}_{i=1}^{N_{test}}, \tag{3.2}$$

where N denotes the number of observation, superscript$^{(N)}$ denotes *Normal.* on the contrary, superscript$^{(A)}$ denote *Anomalous*, and $y_i \in \{1, ..., k\}$ and k are the corresponding all possible class of x_i and the number of all possible classes, respectively. But according to assumption of semi-supervised anomaly detection problem, we can only observe a training set of images $\mathcal{S}^{train}$ with

$$\mathcal{S}^{train} := \{x_i : x_i \in X^{(N)}\}_{i=1}^{N_{train}}, \tag{3.3}$$

and a test set of images $\mathcal{S}^{test}$ with

$$\mathcal{S}^{test} := \{x_i : x_i \in X\}_{i=1}^{N_{test}}, \tag{3.4}$$

Then our goal is to find detection function $f : \mathcal{X} \to \{0, 1\}$, where $f(x) = 0$ only if $x \in X^{(N)}$, otherwise $f(x) = 1$. But to find exact function f is impossible, because of the aforementioned challenge of anomaly detection. In practice, we use anomaly score function $\mathcal{A}_\lambda : \mathcal{X} \to \mathbb{R}$, where λ is anomaly threshold. If anomaly score of an sample is higher than λ, the sample is regarded as *anomaly.*

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

The anomaly threshold λ is a kind of parameter to detect as many anomalies as possible while avoiding Type I and Type II errors. But the best value for the threshold depends on a model and a dataset, and also to consider two type of errors, so Precision-Recall (PR) Curve are commonly used as the performance metric for anomaly detection.

3.2 Image Rotation Prediction

Out of the attributes of image data, the position, the size, and the shape of an object composed of elements are more significant than the position and the existence of each element. These semantic information of the object does not changed when the image is rotated. However to identify the object in the rotated image as the rotated rather than other objects, a learner has to be learned about the high level structure of the object, i.e., the learner should be able to answer following questions indirectly: where is the boarder line between the object and background? what is the top or bottom of the object? We produced self-labeled dataset from the target dataset by randomly rotating a image and giving the degree of rotation as label for rotated image to learn this property to our model. We considered rotation only multiples of 90°

3.2.1 Dataset Generation

Let $R_{90} : \mathcal{X} \times \{0, 1, 2, 3\} \to \mathcal{X}$ be image rotation operator and $R_{90}(x, d) = x^{90d}$, where x^{90d} is the image x rotated by $90d$ degrees counterclockwise. Then from given dataset S, generated self-labeled dataset $\hat{S}$ is defined by

Definition Self-labeled dataset

$$\hat{S} := \{(R_{90}(x_i, d), d) : x \in S, d \in \{0, 1, 2, 3\}\}. \tag{3.5}$$

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

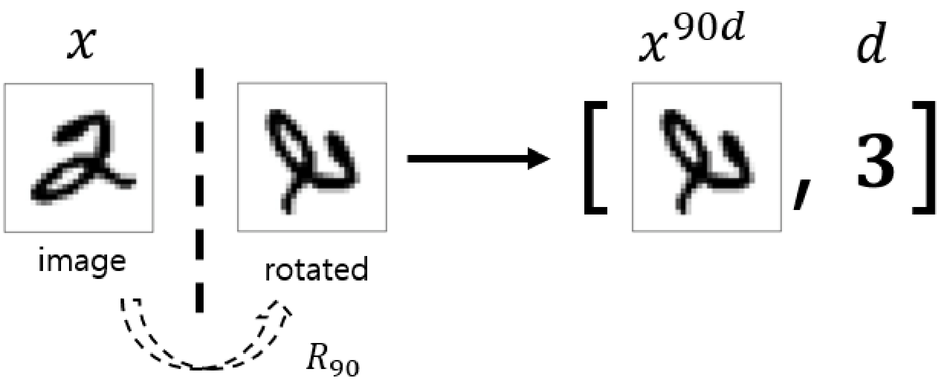


Figure 3.1: Example : Generation of self-labeled dataset.

3.2.2 Prediction

Then we introduce a random variable $c \sim \mathcal{U}\{0,3\}$ to produce the rotated image among four rotation ($0°, 90°, 180°$, and $270°$) classes randomly. In training, we sampled rotated image by c (Fig. 3.1).

Given dataset $\hat{S}$, Let $g : \mathcal{X} \rightarrow [0,1]^4$ be our classifier model predicting a probability distribution over all possible rotaion with softmax activation σ, where g^j is the predicted probability for jth class of given input. Then the objective is defined as:

Definition

$$\min_{\sigma} -\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{c\sim\mathcal{U}\{0,3\}}\left[log(\sigma^{c}(R_{90}(x_i,c)))\right] \tag{3.6}$$

The advantage of this method is that the process can be applied on test dataset, so we can estimate the task with supervision at test time.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

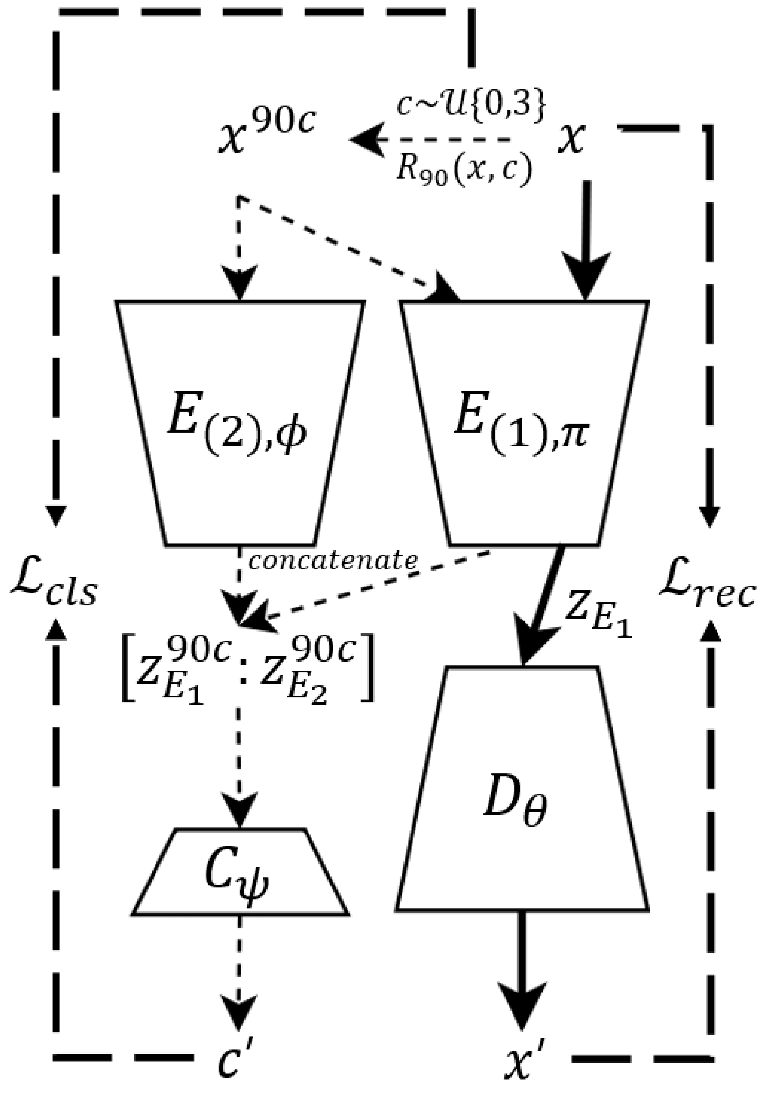


Figure 3.2: The pipeline of the proposed method. E, C and D: encoder, classifier and decoder respectively. Dotted: the pipeline for image rotation prediction, Solid: the pipeline for the auto-encoder. Dashed: the pipeline for each loss function. x^{90c}: counterclockwise $90d^\circ$ rotation of image x. z_{E_1} and z_{E_2}: output of first encoder($E_{(1)}$) and second encoder($E_{(2)}$) respectively. $[\cdot : \cdot]$: concatenation of both input.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

3.3 Pipeline

We now present the proposed model for semi-supervised anomaly detection problem. The model consists of three type sub-networks (Fig. 3.2): two encoders $E_{(1),\pi}, E_{(2),\phi}$, a decoder D_θ, and a classifier C_ψ, where π, ϕ, θ, and ψ are parameters of each sub-networks. We combine both task in the first encoder. All layers of the encoder compresses its inputs by only convolutional layers that is followed by Batch normalization [19] and leaky ReLU [20] activation in a row; All layer of the decoder upscales its inputs by only convolutional transpose layers that is followed by ReLU [21].

Auto-encoder The first encoder $E_{(1),\pi}$ and the decoder D_ϕ are coupled for general Auto-encoder pipeline (Dashed line, Fig. 3.1). In this pipeline $E_{(1),\pi}$ maps input to latent space, then D_θ generates bottleneck features. The pipeline is trained by the reconstruction loss $\mathcal{L}_{rec}$ defined as:

Definition

$$\mathcal{L}_{rec}(x, \pi, \theta) = ||x - D_\theta(E_{(1),\pi}(x))||_2^2, \tag{3.7}$$

By training the loss, the pipeline yields the reconstruction that is similar to original input image.

Classifier Both encoders $E_{(1),\pi}$ and $E_{(2),\phi}$ the classifier C_ψ coupled for image rotation prediction (Dotted line, Fig. 3.1). In this pipeline, C_ψ use both bottleneck features from $E_{(1),\pi}$ and $E_{(2),\phi}$.

By doing so, $E_{(1),\pi}$ simultaneously learns both the representation for reconstruction and the immediate features for prediction, allowing D_ϕ and C_ψ to influence each other. D_ϕ for reconstruction and C_ψ for prediction learn about different tasks, so that D_ϕ and C_ψ are mutually regularized through $E_{(1),\pi}$, i.e., the bottleneck features z obtained from $E_{(1),\pi}$ contain both comprehension of the structure of the image for reconstruction and comprehension of the structure of

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

the image for the prediction, so the model can learn each task with generalized features.

But comparing two given loss function, classification (Eq. 3.6) and reconstruction (Eq. 3.7), the auto-encoder yields images and C_ψ yields four-dimensional vector, so the reconstruction loss gives $E_{(1),\pi}$ larger signal to update than the classification loss. To mitigate this problem, we introduce second encoder $E_{(2),\phi}$ and ω that is hyper-parameter to determine the output channel size ratio of the two encoders for controlling back-prob signal from C_ψ to each encoder.

From the objective (Eq. 3.6) of image rotation predction, the classification loss $\mathcal{L}_{cls}$ for the proposed model is re-written as:

Definition

$$\mathcal{L}_{cls}(x, \psi, \phi, \pi) = \mathbb{E}_{c \sim \mathcal{U}\{0,3\}} \left[-\log \sigma^c (C_\psi(z_{E_1}^{90c}, z_{E_2}^{90c})) \right], \tag{3.8}$$

where σ is softmax function, superscript^{90c} represents the degree of rotation of the corresponding input, and

$$z_{E_1} = E_{(1),\pi}(x), \qquad\qquad z_{E_2} = E_{(2),\phi}(x). \tag{3.9}$$

As noted above, $E_{(1),\pi}$ has two objective, So it minimizes the total loss function defined as:

Definition

$$\mathcal{L} = w_{cls}\mathcal{L}_{cls} + w_{rec}\mathcal{L}_{rec}, \tag{3.10}$$

where w_{cls} and w_{rec} are the weighting parameters to control impact of each loss.

3.4 Anomaly Score

We use the reconstruction loss (Eq. 3.7) to measure anomaly scores, and propose three other extra methods.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

1. **Reconstruction loss** $\mathcal{A}_{rec}$: As noted previous, Almost reconstruction based method use this measure (Eq. 3.7) for anomaly score. Last two methods consider the two tasks.

2. **Negative Rotation Prediction Score** $\mathcal{A}_{cls}$: We believe that the model cannot predict the rotation of the image well if the image does not belongs to model's knowledge, so the image with higher prediction score than anomaly threshold λ is regarded as the normal, otherwise the anomaly. The negative average prediction score $\mathcal{A}_{cls}$ is defined as:

   Definition

   $$\mathcal{A}_{cls}(x) = -0.25 \sum_{d=0}^{3} C_{\psi}^{d}(z_{E_1}^{90d}, z_{E_2}^{90d}) \tag{3.11}$$

3. **Weighted sum** $\mathcal{A}_{ws}$: It is a convex combination of the reconsctruction loss and the prediction score. The weighted sum score of the two score $\mathcal{A}_{ws}$ is defined as:

   Definition

   $$\mathcal{A}_{ws}(x) = \alpha \mathcal{A}_{rec}(x) + (1-\alpha)\mathcal{A}_{cls}(x), \tag{3.12}$$

   where α the trade-off parameter between two anomaly score method.

4. **Maximum** $\mathcal{A}_{max}$: Maximum value out of the two score is only used to measure anomalies. The maximum $\mathcal{A}_{max}$ is defined as:

   Definition

   $$\mathcal{A}_{max}(x) = \max(\alpha \mathcal{A}_{rec}(x), (1-\alpha)\mathcal{A}_{cls}(x)). \tag{3.13}$$

We normalize the two primary scores $\mathcal{A}_{rec}$ and $\mathcal{A}_{cls}$ with training data $\mathcal{S}^{train}$ to have a equal effect in determining the overall scores $\mathcal{A}_{max}$ and $\mathcal{A}_{ws}$ as far as possible:

$$\mathcal{A}'_{\cdot}(x) = \frac{\mathcal{A}_{\cdot}(x) - \min \mathcal{A}_{\cdot}(x')}{\max \mathcal{A}_{\cdot}(x') - \min \mathcal{A}_{\cdot}(x')}, \tag{3.14}$$

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

where $x \in \mathcal{S}^{test}$ and $x' \in \mathcal{S}^{train}$.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

IV. Experiment

To evaluate the proposed method, we benchmark with three image dataset, MNIST [11], Fashion-MNIST [22], and CIFAR10 [23] image dataset.

4.1 Dataset

MNIST It consists of 28 × 28, grayscale, and handwritten digit from 0 to 9 with 60000 images for training and 10000 images for testing. It has data imbalance between classes, so we reduced the data in each class to the number in the class that has the smallest number.

Fashion-MNIST It consists of 32 × 32, grayscale and ten types of fashion products with 60000 images for training and 10000 images for testing, is similar to MNIST which is widely used in the field of machine learning, and has been introduced to provide more difficult tasks.

CIFAR10 It consists of 32 × 32 and colour, and ten types of mutually exclusive objects with 60000 images for training and 10000 images for testing.

4.2 Configuration

To replicate the configuration of [6, 9, 5], a class is regarded as *anomaly*, and other classes are regarded as *normal*. By doing so, we get 10 different dataset from target dataset in which each class is treated as anomalous once. Then we train the model on $\mathcal{D}^{train}$ until completing 25 epochs.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

4.3 Baseline

We compare with methods that are state-of-the-art in the configuration, Efficient-GAN Based Anomaly Detection (EGBAD) [6], and GANomaly [9] implemented them as described in the original paper.

EGBAD uses BIGAN [7, 8] architecture that introduces an encoder, trains jointly with a decoder, and a discriminator judges a sample x as real or fake by x and the corresponding latent representation z. To estimate anomaly score, it uses reconstruction loss, and discriminator loss or feature matching loss.

GANomaly uses an adversarial autoencoder, and introduces extra encoder to regress the latent representation of a sample x from the reconstructed sample x' with different parameter. To estimate anomaly score, it uses the distance between the latent representation z of x from original encoder and the latent representation z' of x' from the extra encoder.

4.4 Implementation Detail

We use the same structure for both encoder $E_{(1)}$ and $E_{(2)}$. With the guideline of Radford *et al.* [24], we use the convolutions followed by Batch Normalization and Leaky ReLU with $\alpha = 0.2$ as the activation for the encoders, and the transposed convolutions followed by ReLU as the activation for the decoders; we do not use Batch Normalization at the decoder. The kernel size for all convolutions is 4×4. The number of kernels for the encoder is 32, 64, 128, and a, and the number of kernels for the decoder is 256, 128, 64, 32, b, where a is a hyper-parameter for latent dimension and b is the channel of images; Both a and b depend on target dataset. We set $a = 16$ for MNIST and Fashion MNIST, and set $a = 200$ for CIFAR10.

To prevent the autoencoder learning a trivial solution, instead of using fully

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

connected layer at the last layer of the encoder and the first layer of the decoder, we use the convolution with 1 stride and no padding, and the transposed convolution with 1 stride and no padding symmetrically; other layers are 2 strides and use padding. As a result, the last layer of the encoder compresses 4×4 feature map to 1×1 feature map, and the first layer of the decoder upsamples the bottleneck features to 4×4 feature map. Thus the model learns bottleneck features that have a global perspective rather than a regional perspective.

For controlling the impact of each encoder for the prediction and the reconstruction, with the weighting parameters w_{cls} and w_{rec} (Eq. 3.10) we introduce a hyper-parameter λ_E which controls the depth of the last layer of the second encoder $E_{(2)}$. Let $a_{(1)}$ and $a_{(2)}$ denote the depth of the last layer of the first and second encoder, then $a_{(2)} = \lambda_E a_{(1)}$.

To predict the rotation, we only use two fully connected layer with Batch Normalization as the classifier.

For the all experiments, we use Adam [25] with $\beta_1 = 0.5$. We choose a batch size of 64, a learning rate of 10^{-4} for training, and Normal distribution with 0.01 standard deviation as the initializer of all models.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

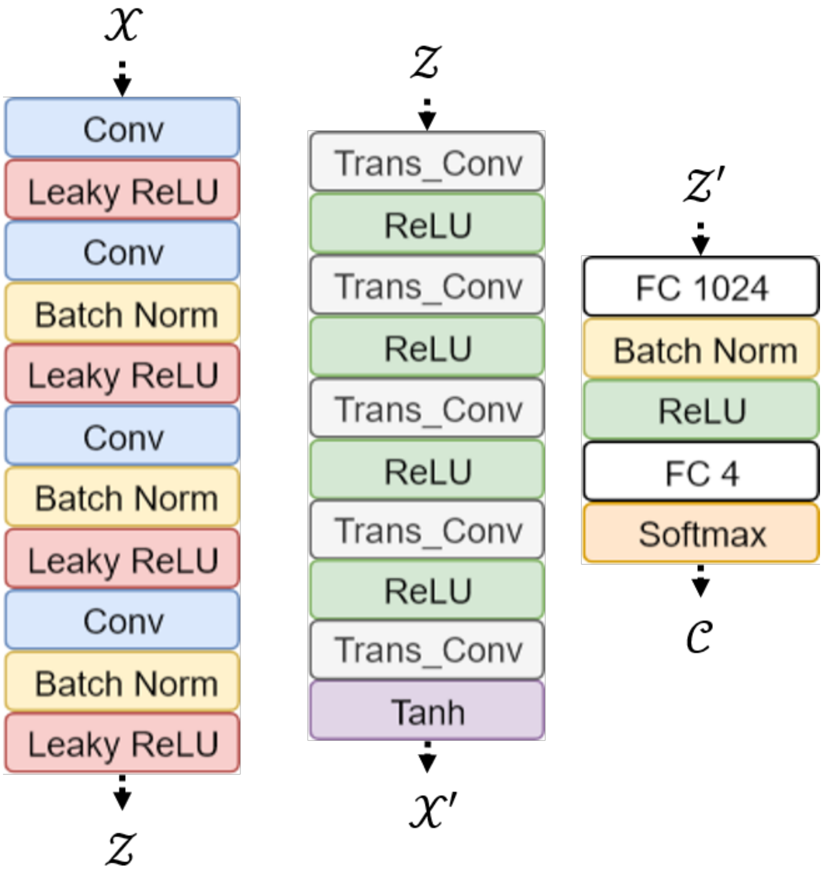


Figure 4.1: Network architecture of the encoder E (left), the classifier C (middle), and the decoder D (right). The kernel size for the convolutions (Conv) and the transposed convoloutions (Trans_Conv) are 4×4. The strides are 1 for the last layer of E and the first layer of D; others are 2. The $\alpha = 0.2$ for Leaky ReLU.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

V. Result

According to Davis and Goadrich [26], if the number of positive (anomalous) sample and the number of negative (normal) are skewed, Receiver operating characteristic (ROC) curves tends to bias performance to an overly good point of view for the algorithm. As noted above, anomaly occur infrequently, so to avoid that imbalance is impossible. To measure performance equitably, it is recommended to use Precision Recall (PR) curves. Anomaly detection problems suffer skewed observation between *normal* and *anomaly*, and in that case, the PR curve shows clearer and less biased algorithm performance than the ROC curve, so we measured the algorithm performance with PR Area Under Curve (AUC).

We tested each configuration with three different seeds and then took an average (Fig. 5.1a, 5.2a, 5.3a), and averaged all result from each dataset (Table. 5.1). Vertical lines (Fig. 5.1a, 5.2a, 5.3a) denote the standard deviation of three run. We use $\mathcal{A}_{ws}$ and $\mathcal{A}_{max}$, and compare the performances to the baselines. Additionally we report our two base models: *AE* that only train and use the autoencoder from our model; *Rotation* that only learns classification from our model, and $\mathcal{A}^*_{max}$ and $\mathcal{A}^*_{ws}$ denote the two overall score methods obtained by using *AE* and *Rotation*. For all experiments, We choose $\alpha_{\mathcal{A}} = 0.5$.

5.1 MNIST

For the experiments on MNIST, we choose $\lambda_E = 0.1$, $w_{cls} = 1$, and $w_{rec} = 100$, empricially. Our model surpasses the baselines on the all configuration (Fig. 5.1a).

However as the result of reconstruction of *AE* (Fig. 1.1) shows the recon-

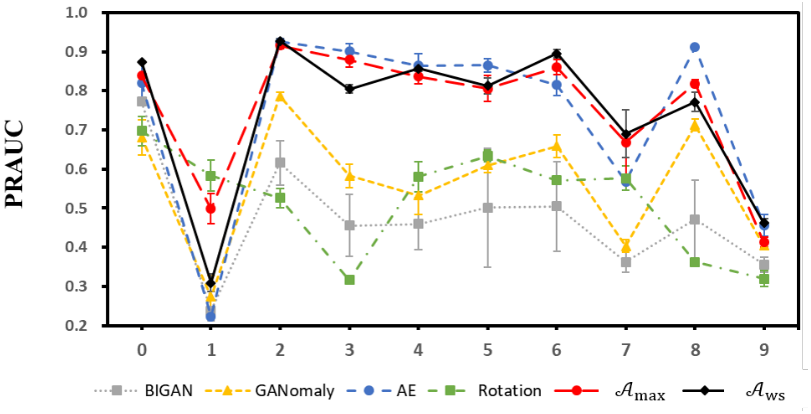
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

struction of anomalous sample on each configuration, the reconstuction methods (AE, $\mathcal{A}_{rec}$) shows very poor performance at the configuration with 1. Our model with $\mathcal{A}_{ws}$ and $\mathcal{A}^{*}_{ws}$ also shows very poor performances at the configuration 1.

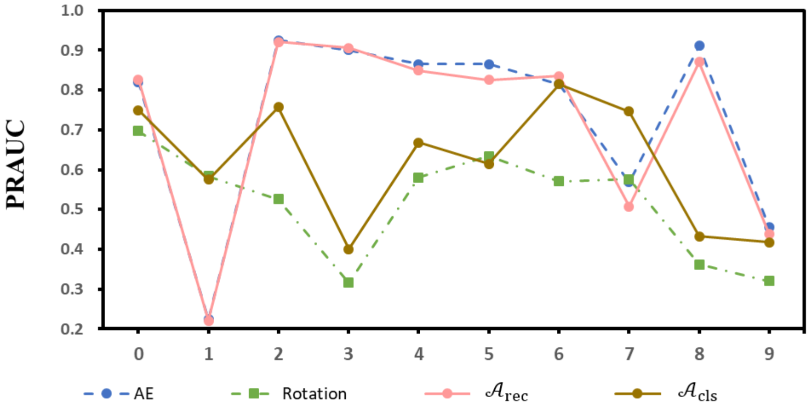
Basically the ability to detect anomalous by the two overall scores which uses the primary scores depend on the ability of the primary scores, so the reconstruction score has a decisive effect on the ability and the classification scores sometimes decrease the ability, vice versa on the configuration with 1. In other word, the overall scores tend to be in the middle of the primary scores.

However in some case, e.g. on the configuration with 0, 6, and 7, the ability of the overall score methods improve the primaries, but the improvement barely appears without joint learning (Fig. 5.1c). The average PRAUC (Table 5.1) of our model with $\mathcal{A}_{max}$ and $\mathcal{A}_{ws}$ outperforms others.

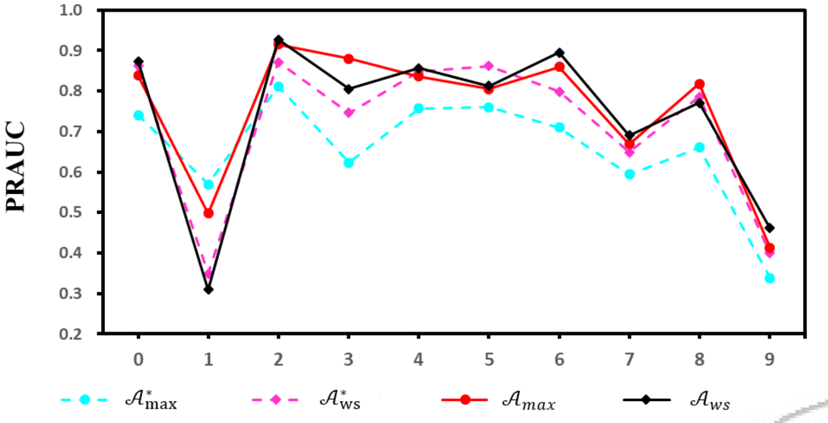
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교



(a) Baseline vs Base model vs OURS with standard deviation.



(b) Base model vs OURS components.



(c) Without joint learning vs OURS.

Figure 5.1: The configuration that anomaly (the indicated class) vs normal (others) on MNIST. *AE*: only auto-encoder; *Rotation*: only classifier; $(\circ)^{*}$: use the overall scores by *AE* and *Rotation*

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

5.2 Fashion MNIST

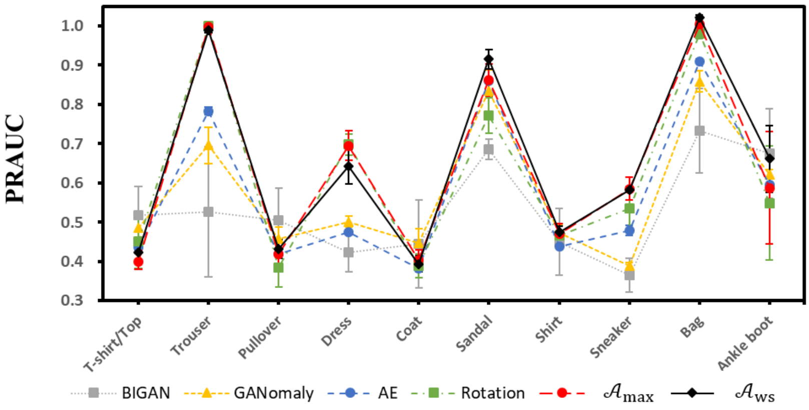
For the experiments on Fashion MNIST, we choose $\lambda_E = 2$, $w_{cls} = 1$, and $w_{rec} = 0.01$, empricially. Our model outperforms the baseline on the almost configuration (Fig. 5.2a).

However our model showed low performance in T-shirt, Pullover and Coat, because these three fashion items are similar to each other and other anomalous items may show similar pattern if some of these fashion items are learned.

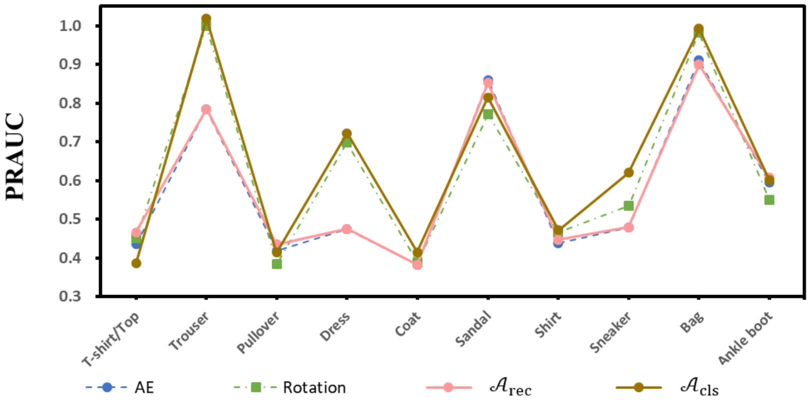
Classification based methods (OURS and *Rotation*) show better performance than reconstruction based methods (OURS and *AE*) at almost configurations, so two overall score methods tend to depend on the classification.

However as the same as MNIST, on the configuration with Sandal, Shirt, Bag, Ankel, and Sneaker, the ability of the overall score methods, especially $\mathcal{A}_{ws}$, improve the primaries. Even though the baseline gets to the top on some configuration, the average PRAUC (Table 5.1) of our model with $\mathcal{A}_{max}$ and $\mathcal{A}_{ws}$ outperforms others.

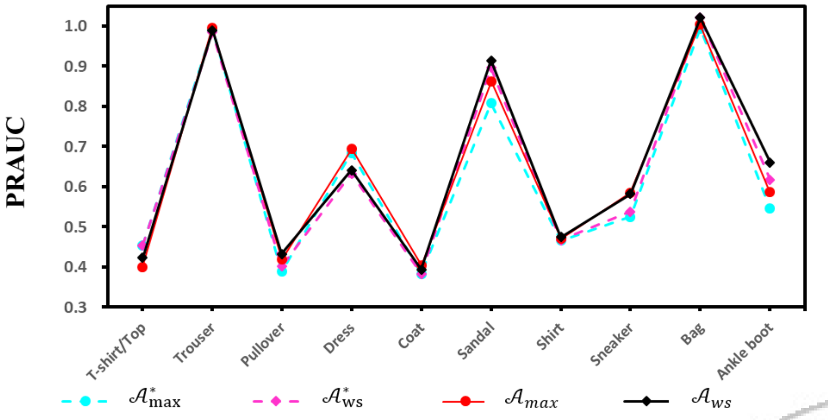
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교



(a) Baseline vs Base model vs OURS with standard deviation.



(b) Base model vs OURS components.



(c) Without joint learning vs OURS.

Figure 5.2: The configuration that anomaly (the indicated class) vs normal (others) on Fashion MNIST. *AE*: only auto-encoder; *Rotation*: only classifier; $(\circ)^{*}$: use the overall scores by AE and Rotation

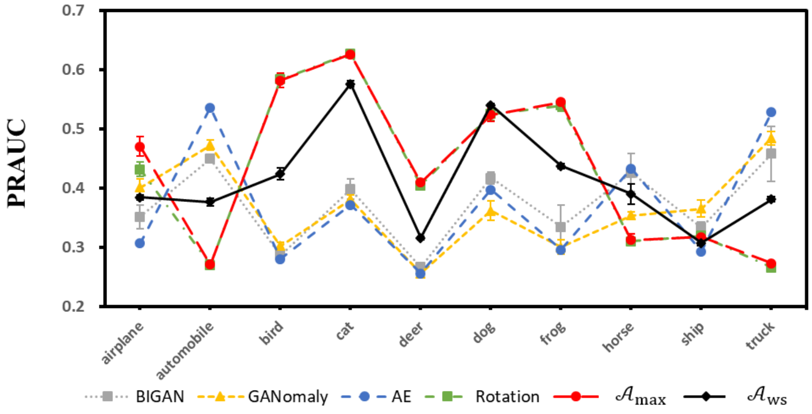
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

5.3 CIFAR10

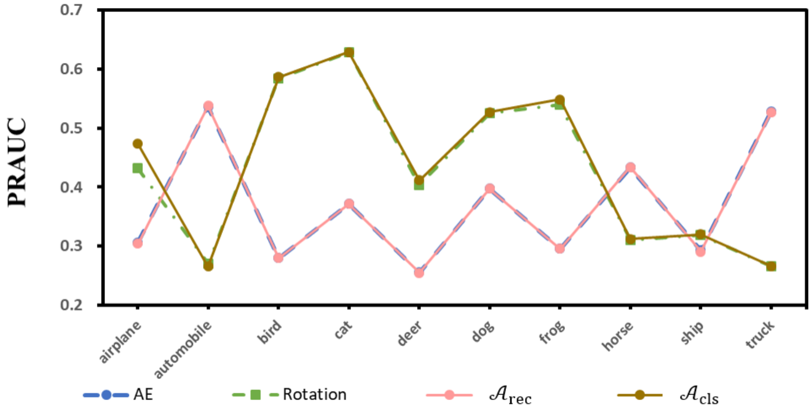
Reconstruction based models and classification based models show conflicting behaviors: classification based method obtains better results than reconstruction based method on the most configuration. This behaviors cause the two overall score methods to behave different: $\mathcal{A}_{max}$ follows the classification; $\mathcal{A}_{ws}$ is in the middle of the two primaries.

Only on the configuration with dog class, $\mathcal{A}_{ws}$ improves two primary score methods. The ability of reconstruction based method is lower than classification based model, so the methods related to classification obtain better PRAUC. Therefore $\mathcal{A}_{cls}$ and $\mathcal{A}_{max}$ outperforms others (Table 5.1).

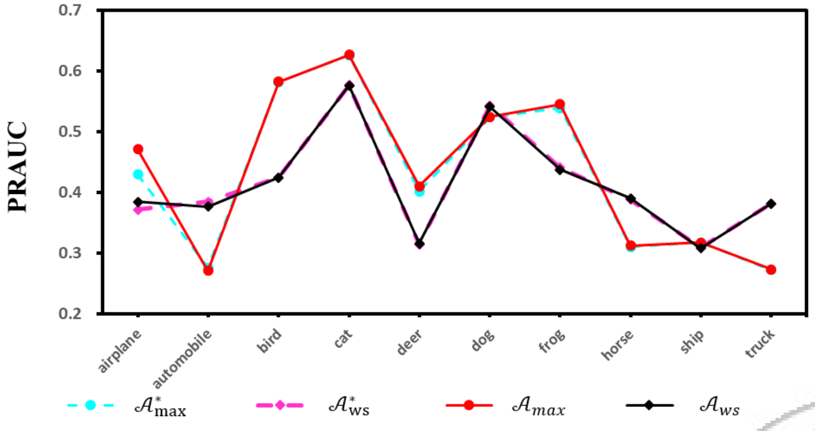
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교



(a) Baseline vs Base model vs OURS with standard deviation.



(b) Base model vs OURS components.



(c) Without joint learning vs OURS.

Figure 5.3: The configuration that anomaly (the indicated class) vs normal (others) on CIFAR10. *AE*: only auto-encoder; *Rotation*: only classifier; $(\circ)^*$: use the overall scores by *AE* and *Rotation*

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

Table 5.1: Average PRAUC on the configurations that one (anomaly) versus others (normal) of each dataset. Bold: Best performance.

Method	MNIST	Fashion	CIFAR10
BIGAN	0.4738	0.4827	0.3722
GANomaly	0.5645	0.5264	0.3678
AE	0.7350	0.5278	0.3696
Rotation	0.5165	0.5723	0.4279
$\mathcal{A}^*_{max}$	0.6567	0.5740	0.4277
$\mathcal{A}^*_{ws}$	0.7173	0.5896	0.4134
OURS $\mathcal{A}_{rec}$	0.7201	0.5327	0.3692
OURS $\mathcal{A}_{cls}$	0.6175	0.5953	**0.4340**
OURS $\mathcal{A}_{max}$	0.**7534**	0.5925	0.4332
OURS $\mathcal{A}_{ws}$	0.7403	**0.6034**	0.4133

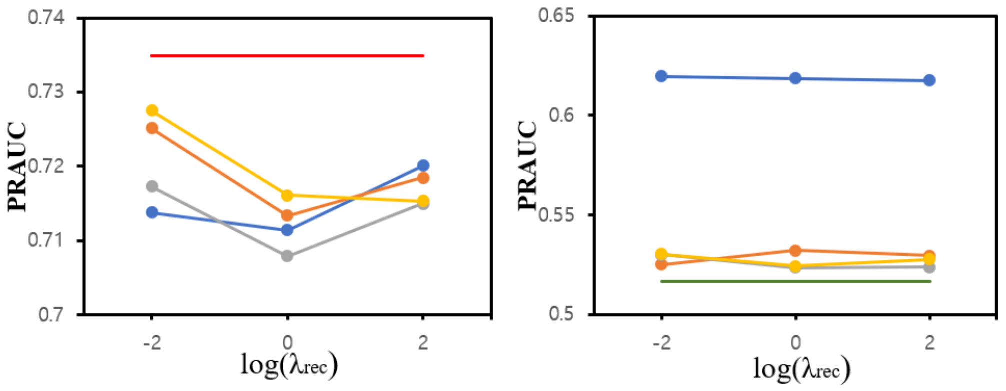


Figure 5.4: Comparison average PRAUC on all the configuration of MNIST by the reconstructed loss (left) and by the negative classification score (right) of OURS, and AE and Rotation. X axis: log scale of λ_{rec} with fixed $\lambda_{cls} = 1$. λ_E: the ratio of the depth of last layer of second encoder.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

5.4 Both tasks affect to each other?

Fig. 5.4 show the average PRAUC on all the configuration of MNIST by each primary loss when two hyper-parameters are changed, and the comparison between the primary loss of our model, and *AE* and *Rotation*. Both λ_E and λ_{rec} do not have a great effect on the PRAUC by $\mathcal{A}_{rec}$; λ_E has a great effect on the PRAUC by $\mathcal{A}_{cls}$. The PRAUC increases around 0.1 when the classification is used and $a_{(2)} = 1$, i.e., the classifier improves the performance when the autoencoder and the classifier are coupled more tightly. Although the autoencoder degenerates the performance, the improvement causes the overall score method to get higher performance.

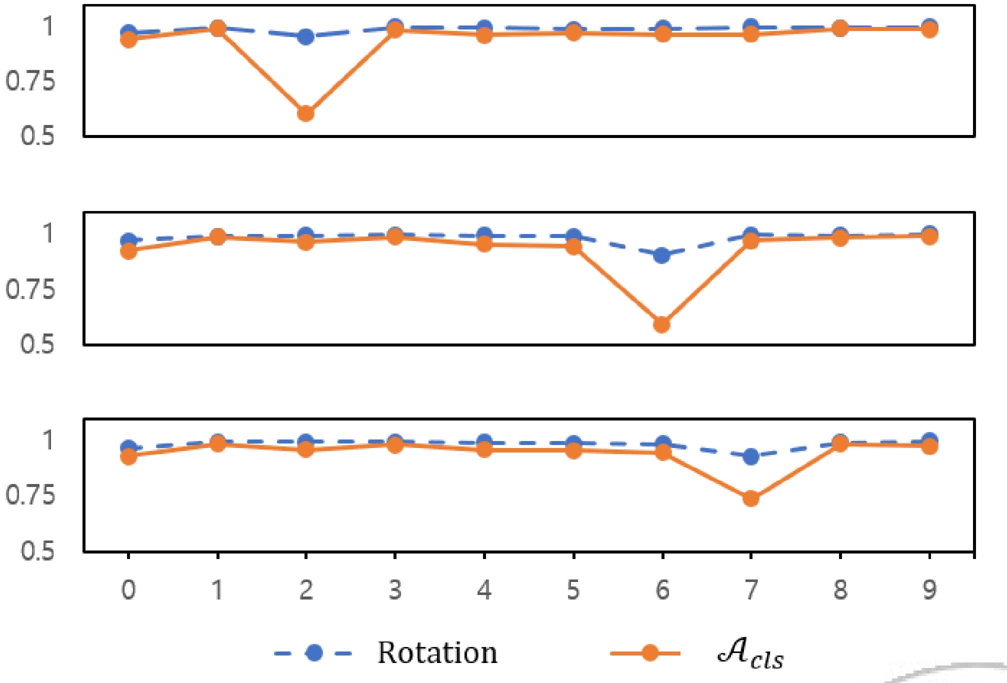


Figure 5.5: Average accuracy over all rotation of *Rotation* vs Our model on the configuration with 2 (top), 6 (middle), and 7 (bottom) of MNIST.

To analysis why our classifier performs better than the classifier without the

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

autoencoder, we demonstrate classification (Fig. 5.5) on the configuration that the larger improvements occur by our classification method: 2, 6, and 7. Only learning the rotation of images is generalized well so that *Rotation* predicts well even if the images are not learned; intermediate features of our model have the constraints that the features satisfy the two objective so that our classifier cannot predict the rotation when if the images are not learned.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

VI. Conclusion

We argue that detecting anomalies by single aspect has limitation, so we proposed self-supervised autoencoder by introducing self-supervised method to add extra unsupervised task and connecting both architectures at the encoder to learn both representation. We chose Image Rotation Prediction as self-supervised task. By learning the classification task that how much rotated is the image, the model learns high level structure of the image to recognize the rotated image as rotated rather than another image. Through unrelated two tasks, our approach can detect the anomaly with different aspects of the data, so our approach showed better results in heterogeneous normal dataset. Joint architecture led to degeneration of the generalization that reconstruction or classification of the image belonging to the anomaly, so the ability to detect anomalies improved.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

요약문

이상 탐지 (Anomaly detection)은 데이터로부터 정상이라 사전 정의된 부분집합 또는 데이터의 대다수를 구성하는 부분집합의 독특한 패턴들을 찾아서, 이를 통해 부분 집합의 여집합, 즉 ”비정상”을 찾는 것이다. 최근 딥러닝을 이용한 이상 탐지 방법은 재생성 손실을 통해 비정상을 측정한다. 하지만 데이터 집합이 여러 다른 종류들로 이뤄져있고 비정상 데이터와 정상 데이터가 어떠한 특성을 공유하는 경우 하나의 작업만을 학습한 모델로는 이들을 구분하는 것이 어려울 수 있다. 이 문제를 해결하기 위해 우리는 재생성 및 자율-지도 방식으로 생성된 데이터 집합을 통해 분류를 학습하는 자율-지도 오토인코더를 제안한다. 우리의 방식은 다음과 같은 장점이 있다. 1) 직접적으로 관련되지 않은 두 태스크를 학흡하는 것을 통해 데이터의 다른 두 공간을 활용하여 비정상을 탐지할 수 있다. 결합된 구조를 통해 단순히 앙상블하는 것보다 결과가 향상되었다. 이미지 데이터 집합의 실험을 통하여 우리가 제안한 방법의 장점과 효율성을 보였다.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

References

[1] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, July 2009.

[2] I. Golan and R. El-Yaniv. Deep anomaly detection using geometric transformations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9781–9791. Curran Associates, Inc., 2018.

[3] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft. Deep one-class classification. In Jennifer Dy and Krause andreas, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4393–4402, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[4] S. Zhai, Y. Cheng, W. Lu, and Z. Zhang. Deep structured energy based models for anomaly detection. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1100–1109, New York, New York, USA, 20–22 Jun 2016. PMLR.

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

[6] H. Zenati, C. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar. Efficient gan-based anomaly detection. *arXiv:1802.06222*, 2018.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

[7] J. Donahue, P. Krähenbühl, and T. Darrell. Adversarial feature learning. In *International Conference on Learning Representations*, 2016.

[8] V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro, and A. Courville. Adversarially learned inference. In *International Conference on Learning Representations*, 2016.

[9] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *14th Asian Conference on Computer Vision (ACCV)*. Springer, 2018.

[10] C. Aytekin, X. Ni, F. Cricri, and E. Aksu. Clustering and unsupervised anomaly detection with l2 normalized deep auto-encoder representations. *arXiv preprint arXiv:1802.00187*, 2018.

[11] Y. LeCun and C. Cortes. MNIST handwritten digit database. 2010.

[12] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *International Conference on Computer Vision (ICCV)*, 2015.

[13] 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 *International Conference on Learning Representations*, 2018.

[14] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2234–2242. Curran Associates, Inc., 2016.

[15] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE*

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

Conference on Computer Vision and Pattern Recognition (CVPR), pages 2536–2544, 2016.

[16] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision (ECCV)*, pages 69–84. Springer, 2016.

[17] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

[18] S. Jenni and P. Favaro. Self-supervised feature learning by learning to spot artifacts. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[19] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[20] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

[21] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[22] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[23] A. Krizhevsky, V. Nair, and G. Hinton. Cifar-10 (canadian institute for advanced research).

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

[24] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2016.

[25] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[26] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

Acknowledgements

가장 먼저 포항공대 기계학습 연구실에서 공부할 수 있는 기회를 주신 최승진 교수님께 감사 드립니다. 교수님의 가르침을 통해서 연구에 대해서 아무것도 모르던 제가 연구가 무엇인지, 어떤 연구를 지향해야하는 지에 대해서 배웠습니다. 그리고 언제나 열심히 일하시고 정진하시는 모습을 통해 연구자가 지향해야할 자세가 무엇인지 배울 수 있었습니다. 그리고 저에게 해주신 진심 어린 조언들을 새겨 듣고 열심히 살겠습니다.

그리고 바쁘신 와중에도 저의 문제로 학위 심사 일정을 변경해야했음에도 심사를 맡아주신 유환조 교수님, 곽수하 교수님께 감사 드립니다. 더 좋은 연구를 위해 부족한 부분과 연구 방향에 대해서 조언해 주신 덕분에 논문을 완성하는 데 도움이 되었고, 이후에 수행할 연구의 방향을 결정하는 데 많은 도움이 되었습니다.

또한 대학원에 입학하여 힘이 되어주고 도와준 연구실 사람들에게도 감사드립니다. 도움이 필요할 때 조언을 아끼지 않았던 정택 선배, 같은 분야를 연구하면서 많은 도움을 준 영남, 졸업 준비로 같이 수행해야할 업무를 묵묵하게 혼자서 수행해준 영석, 그리고 연구실 생활이 즐거운 추억이 되게 해준 주호 형, 세훈이 형, 본곤이 형, 지유, 민섭이형, 인혁이 형, 진휘 형, 윤호, 현호, 민교, 나영에게 감사를 전합니다.

마지막으로 항상 사랑과 지지를 보내주시는 부모님, 먼 지방에서 누구보다 열심히 살면서 오빠를 걱정해주는 동생에게 감사 드립니다. 그리고 제 석사 과정이 성공적으로 끝나기를 응원해준 친구들과 지인들에게 감사 드립니다.

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY 1986 포항공과대학교

Curriculum Vitae

Name : Wonbin Kim

Education

2010. 3. – 2017. 2.	Department of Computer Science and Engineering, Inha University (B.S.)
2017. 3. – 2019. 2.	Department of Computer Science and Engineering, Pohang University of Science and Technology (M.S.)

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교