Spatially-weighted Anomaly Detection

Minori Narita\textsuperscript{1}, Daiki Kimura\textsuperscript{2}, Ryuki Tachibana\textsuperscript{2}
\textsuperscript{1}The University of Tokyo, \textsuperscript{2}IBM Research AI
Email: narita@g.ecc.u-tokyo.ac.jp, \{daiki, ryuki\}@jp.ibm.com

Abstract—Many types of anomaly detection methods have been proposed recently, and applied to a wide variety of fields including medical screening and production quality checking. Some methods have utilized images, and, in some cases, a part of the anomaly images is known beforehand. However, this kind of information is dismissed by previous methods, because the methods can only utilize a normal pattern. Moreover, the previous methods suffer a decrease in accuracy due to negative effects from surrounding noises. In this study, we propose a spatially-weighted anomaly detection method (SPADE) that utilizes all of the known patterns and lessens the vulnerability to ambient noises by applying Grad-CAM, which is the visualization method of a CNN. We evaluated our method quantitatively using two datasets, the MNIST dataset with noise and a dataset based on a brief screening test for dementia.

I. INTRODUCTION

A. Anomaly detection by images

These days, many types of anomaly detection methods have been proposed and applied to fields such as production quality checking and medical screening \cite{1, 2, 3}. In real situations, a part of the anomaly patterns is often known beforehand. For example, various typical anomaly patterns from participants of a screening test for dementia called the Yamaguchi Fox-Pigeon Imitation Test (YFPIT) \cite{4} have been reported. With this test, participants made specific gestures to imitate a fox or a pigeon, and they were then screened on the basis of the result. Fig. 1 gives the overview of a method where anomaly detection is applied to YFPIT.

A straightforward approach is to train a model for only normal patterns, and then to classify by deviation. When using images for detection, it is common to train an auto-encoder such that reconstruction errors toward the normal images are minimized, and then use the error for detection. An advantage of this method is that users can check the results of the detection with their own eyes. However, they cannot utilize the information of anomaly patterns, even if they already know it. Moreover, the results might fluctuate due to error from noises, the model will make the noise values average.

Another approach is to solve the problem by training a classification model for both the normal and anomaly patterns. If the input is images, it is natural to perform by using a convolutional neural network (CNN). The advantage of this approach is that all of the known patterns can be utilized. On the other hand, there is a serious vulnerability in that the output of the anomaly patterns is intrinsically unpredictable. It is important to maintain robustness against unknown patterns, especially when we aim for application to medical fields. For these reasons, a new method is required that can mitigate ambient noises and maintain robustness against unknown patterns, while being based on anomaly detection.

B. Spatial information in images

To maintain robustness against unknown patterns, only a normal pattern should be used during the training process for anomaly detection. On the other hand, a method such as weighting a region that is important for detection by using all of the training data can be implemented to buffer the effect of surrounding noises and utilize the known patterns. Generally speaking, object recognition methods such as Faster R-CNN \cite{5} and SSD \cite{6} seem to be suitable in this context, but these methods need region information for training. This is problematic because labeling all of the objects by hand is daunting work, and also, it is difficult to check quantitatively if the best labeling has been done.

Recently, Class Activation Mapping (CAM), which visualizes the region of interest (ROI) from the gradients of a CNN \cite{10}, and Grad-CAM \cite{11}, which is a generalization of CAM, have been proposed. Grad-CAM obtains the ROI from the gradients flowing into the last convolutional layer of a CNN. The significant advantage of Grad-CAM is that the regional information can be obtained without having to train the data that has the regional information.

C. Proposed method

In this paper, we propose a new method called SPAtially-weighted Anomaly DEtection (SPADE), which conducts anomaly detection by using Grad-CAM to weight the important region in an image. We also modified the Grad-CAM algorithm under this work. The overview of SPADE is presented in Fig. 2.

II. SPADE

We assume that a part of the anomaly patterns is already known. Therefore, there are three classes in this problem setting: a normal class \( \mathcal{X} \), known anomaly classes \( \mathcal{A} \), and test data including unknown anomaly classes \( \mathcal{U} \). During the training process, \( \mathcal{X} \) and \( \mathcal{A} \) are used. \( \mathcal{U} \) is used for evaluation.
The outline of the overall process is as follows. First, in the training process, a CNN is trained as the binary classifier and obtains the parameters related to the interest region information, and a Variational Auto-Encoder (VAE) \([9]\) is also trained using only the normal class. After that, the error weighted from the interest region in the image is obtained by multiplying the ROI given by the CNN using Grad-CAM and the reconstruction error given by VAE. Anomaly detection is performed depending on the value of this weighted error.

### A. Training

The normal class \(X\) is given to VAE as input, and parameters \(\theta, \phi\) of VAE are optimized to minimize the cost function

\[
\mathcal{L}(\theta, \phi, x_i) = -D_{KL}(q_{\phi}(z|x_i)||p_{\theta}(z)) + \mathbb{E}_{q_{\phi}(z|x_i)}[\log p_{\theta}(x_i|z)],
\]

(1)

where \(x_i\) is a sample data, \(z\) is a hidden space, \(D_{KL}\) is KL divergence, and \(\mathbb{E}_{q_{\phi}(z|x_i)}[\log p_{\theta}(x_i|z)]\) is the expected value of a logarithmic likelihood \(\log p_{\theta}(x_i|z)\) of input \(x_i\) for an encoder \(q\). The reconstruction error from input \(e_u\) is computed by these trained parameters \(\theta\) and \(\phi\).

At the same time, all of the training data \((X + A)\), which are known patterns, are input to the CNN and trained to classify a correct class and incorrect classes by optimizing the parameter \(\omega\) to minimize the cost function below.

\[
\mathcal{L}(\omega, x_i, t_i) = ||t_i - f(x_i; \omega)||,
\]

(2)

where \(t_i\) is the lable of a sample \(x_i\), which has the value of 1 if \(x_i\) is the correct pattern, and 0 if not.

### B. Detection

The detection algorithm is performed as follows. First, we calculate a reconstruction error \(e_u (= |u - \hat{u}|)\) from a reconstructed image \(\hat{u}\) by inputting a test data \(u\) to the trained VAE network. At the same time, this test data \(u\) is also input to the CNN network and the ROI is estimated by using the visualization method, Grad-CAM. Grad-CAM is executed as follows. We first compute the gradient of the score for class \(c\) (the normal pattern), or the partial derivative of the feature maps \(A^k\) of a convolutional layer concerning \(y_c\) (before the softmax layer) (i.e., \(\frac{\partial y_c}{\partial A^k}\)). The gradient flowing back is global-average-pooled to obtain the importance weights of the \(k_{th}\) feature map for the class \(c\), \(\alpha^c_k\).

\[
\alpha^c_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c_i}{\partial A^k_{ij}}
\]

(3)

Next, we perform a weighted combination of forward activation maps, and obtain the ROI, \(L^{\text{CAM}}\).

\[
L^{\text{CAM}} = f \left( \sum_k \alpha^c_k A^k \right),
\]

(4)

where \(f\) is the activation function. \(f\) is the absolute function (Abs) for input images, and the ReLU function for reconstruction images (described later). The eventual ROI \(L^{\text{CAM}}\)

\[
L^{\text{CAM}} = \text{Abs} \left( \sum_k \alpha^c_k A^k \right) + \text{ReLU} \left( \sum_k \alpha^c_k A^k \right)
\]

(5)

We obtain a weighted reconstruction error \(e'_u\) focusing on the important region by multiplying \(e_u\) by \(L^{\text{CAM}}\). The input such that \(\text{sum}(e'_u)\) is less than a threshold is classified as the correct pattern, and larger than the threshold is classified as an incorrect pattern. The outline of the algorithm above is shown in Algorithm 1. We describe the reasons and the details of the modification from a simple Grad-CAM below.

First, we took the sum of the ROI of a reconstructed image from a VAE as well as an original image. If we estimate the regional information only from the original image, we may fail to recognize the target object and underestimate the reconstruction loss for an anomaly pattern. Since a reconstruction image is close to a correct pattern, this modification enables us to weight the region that should be focused on.

Second, we changed the activation function of Grad-CAM from the ReLU function to the absolute function. Unlike Grad-CAM, in SPADE, we should focus on the region that has dissimilarity to the target as well. By activating by the absolute function, we can obtain the regional information that is significant for the detection more accurately. An example that fails to focus on the hand region when using the ReLU function is given in Fig. [3]. Meanwhile, the activation function is ReLU when a reconstruction image is input. This is because the feature of a reconstructed image should always be close to the correct pattern.

Third, we normalized the weighting size. When the value of regional information (weighting) is large, the error will be large as well. By normalizing each ROI, we can get rid of the bias that comes from the region size. Examples of the difference between activation functions are shown in Fig. [1]-[3].

We define SPADE as a method that weights the reconstruction error of VAE with the above modifications. The substantial difference from previous methods is that we utilize the regional information in images automatically by applying the visualization method of a CNN. This enables us to create a model robust against ambient noise unrelated to detection.
III. Experiments

A. Comparative approach

We evaluated the proposed method (SPADE) and compared it with anomaly detection by the reconstruction error of a VAE (VAE-based) [8], detection depending on the likelihood of a CNN (CNN-based), and Naïve SPADE, which simply multiplies the reconstruction error of a VAE and the output of Grad-CAM. For Naïve SPADE, we computed regional information only from an input image, which is not normalized, and activated by the ReLU function. In the experiment, we used the known normal and anomaly patterns for training data, and calculated the AUC of the ROC curve using all of the evaluation images including unknown anomaly patterns.

B. MNIST with noise dataset

In its original form, the MNIST dataset does not have noises, and a figure is written using the entire image space. Therefore, it is not suitable for this problem (noises are included, and a target object exists in a specific position). Hence, we added normalized noise to images and tripled the image size, and also changed the size and position of a figure depending on each image to create an MNIST dataset that includes noises. Examples of images are shown in Fig. 5.

The value range of the MNIST dataset is chosen from inside this range.

As for the network structure, a VAE has 4 convolutional layers and a 128-dimensional hidden space, and a CNN consists of 3 convolutional layers. We assume that 0 is the correct pattern, and 1–9 are the incorrect patterns, and one of them is the known pattern while the others are the unknown patterns. Specifically, we computed the result in which we assume that one of 1, 3, 5, 7, 9 is the known incorrect pattern.

The results are given in Table I. Our method has the comparative result than CNN-based baseline method.

C. Hand gesture

We made a pigeon gesture dataset in reference to the Yamaguchi Fox-Pigeon Imitation Test [4]. The hand gesture dataset was created by taking 189,000 images as a total, comprising seven types of gestures reported by [4] from 18 participants who had different body shapes and gender. We used depth images from Kinect, as the shape and depth are more important for detection than the color. An example of images is shown in Fig. 6.

We took pictures following the process below. A participant sits on a chair and makes a gesture of each pattern while facing toward Kinect, and Kinect takes pictures continuously at the frequency of around 30 fps. We had participants make the gesture while changing their position, angle of hands, and angle of posture. In addition, we asked participants to change their seated position on the chair, left, middle, and

Algorithm 1 Spatially-weighted Anomaly Detection (SPADE)

```
procedure TRAINING
φ, θ, ω ← Initialize parameters
repeat
    ∀x ∈ X^M, g ← ∇θ_L(θ, x)
    φ, θ ← Update parameters using gradients g
until convergence of parameters (φ, θ)
repeat
    ∀{x_i, t_i} ∈ (X + A)^M, g ← ∇ω_L(ω, x_i, t_i)
    ω ← Update parameters using gradients g
until convergence of parameter ω
end procedure

procedure TEST
for all x ∈ U do
    ̂u ← R(u; θ, φ) (obtain the reconstructed image)
    e_u ← |̂u − u|
    L_CAM ← Abs (∑_k α_u^k A^k) + ReLU (∑_k α_u^k A^k)
    e_u' ← e_u ∗ L_CAM / |L_CAM|
    if sum(e_u') > threshold then
        Detect as incorrect pattern
    else
        Detect as correct pattern
    end if
end for
```

|                  | 1  | 3  | 5  | 7  | 9  | Avg. |
|------------------|----|----|----|----|----|------|
| VAE-based [8]    | .63| .63| .63| .63| .63| .63  |
| CNN-based        | .73| .88| .96| .88| .96| .88  |
| Naïve SPADE      | .67| .59| .65| .76| .53| .64  |
| SPADE            | .85| .87| .92| .86| .90| .88  |

Table 1
AUROC for MNIST dataset with noises
right. By collecting data comprising different types of hands and postures, we aimed to create a strongly robust classifier.

As for the network structure, the VAE has 4 convolutional layers and a 256-dimensional hidden space. We used the structure of ResNet [12] for the CNN. We assume that $b$ (pigeon gesture) is the correct pattern, and $c$–$h$ are the incorrect patterns, with one of them known and the others unknown. Specifically, we computed the result in which we assume that one of $c$, $d$, $e$ is the known incorrect pattern.

The results are given in Table II. Our method outperformed the comparative methods for all of the given patterns.

|                  | $c$ | $d$ | $e$ | Avg. |
|------------------|-----|-----|-----|------|
| VAE-based [8]    | .95 | .95 | .95 | .95  |
| CNN [12]-based   | .86 | .97 | .90 | .91  |
| Naïve SPADE      | .80 | .71 | .78 | .76  |
| SPADE            | .98 | .97 | .96 | .97  |

IV. CONCLUSION

In this paper, we proposed a new method that utilizes the information of known anomaly patterns for an anomaly detection problem in which a part of the anomaly patterns is already known. We verified that the proposed method which combines anomaly detection based on a VAE with Grad-CAM, which is the visualization method of a CNN, outperformed previous methods. The three points below are the main contributions of this paper.

- We proposed an new anomaly detection method that utilizes information of the known anomaly patterns for a problem in which a part of the anomaly patterns is already known
- The proposed method outperformed previous methods including an anomaly detection method using a VAE and the method using the likelihood of a CNN
- We demonstrated through the results of real environment dataset that this method can be applied to a wide variety of real situations.

From these points, we conclude that our method has potential applications to many fields including production quality checking and fraud detection.

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly Detection: A Survey”, ACM Computing Surveys, 2009.
[2] A. Taboada-Crispi, H. Sahli, M. Orozco Monteagudo, D. Hernández-Pacheco, and A. Falcón Ruiz, “Anomaly Detection in Medical Image Analysis”, Research on Advanced Techniques in Diagnostic Imaging and Biomedical Applications, 2009.
[3] N.S. Halvaei and M.K. Akbari, “A novel model for credit card fraud detection using artificial immune systems”, Appl. Soft Comput., 24 (2011) 4049.
[4] H. Yamaguchi, Y. Maki, and T. Yamagami, “Yamaguchi Fox-Pigeon Imitation Test: A Rapid Test for Dementia”, Dementia and Geriatric Cognitive Disorders, 2010.
[5] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: towards real-time object detection with region proposal networks”, NIPS, 2015
[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, “Ssd: Single shot multibox detector”, ECCV, 2016.
[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks”, NIPS, 2012.
[8] J. An and S. Cho, “Variational autoencoder based anomaly detection using reconstruction probability”, SNU Data Mining Center Tech. Rep., 2015.
[9] D. Kingma and M. Welling, “Auto-Encoding Variational Bayes”, ICLR, 2013.
[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning Deep Features for Discriminative Localization”, CVPR, 2016.
[11] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization”, ICCV, 2017.
[12] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition”, CVPR, 2016.