LETTER

Few-Shot Anomaly Detection Using Deep Generative Models for Grouped Data

Kazuki SATO[†a)], **Satoshi NAKATA**[††b)], ***Nonmembers***, **Takashi MATSUBARA**[†††c)],
and **Kuniaki UEHARA**[††††d)], ***Members***

SUMMARY There exists a great demand for automatic anomaly detection in industrial world. The anomaly has been defined as a group of samples that rarely or never appears. Given a type of products, one has to collect numerous samples and train an anomaly detector. When one diverts a model trained with old types of products with sufficient inventory to the new type, one can detect anomalies of the new type before a production line is established. However, because of the definition of the anomaly, a typical anomaly detector considers the new type of products anomalous even if it is consistent with the standard. Given the above practical demand, this study propose a novel problem setting, few-shot anomaly detection, where an anomaly detector trained in source domains is adapted to a small set of target samples without full retraining. Then, we tackle this problem using a hierarchical probabilistic model based on deep learning. Our empirical results on toy and real-world datasets demonstrate that the proposed model detects anomalies in a small set of target samples successfully.

key words: *anomaly detection, deep generative model, disentangled representation learning*

1. Introduction

Anomaly detection, a problem of identifying samples with patterns substantially differing from normal ones, has been attracting much attention, such as the discovery of defective parts for manufactured products in factories [1] and diagnosis of lesions in medical image analysis [2], [3]. Because the anomaly has a wide variety of patterns produced non-orderly, the unsupervised anomaly detection is preferable to the supervised anomaly classification [4], [5].

When focusing on the industrial use, an anomaly detector needs to learn typical patterns from numerous samples of a type of products. Given a new type (i.e., target type) of products, one has to collect numerous samples of the product from a production line and to train another anomaly detector from scratch. When one diverts a model trained with old types (i.e., source types) of products with sufficient inventory to the new type, one can detect anomalies of the new type before a production line is established.

In this study, we consider a novel problem setting of detecting anomalies within a small set of samples of a target type by diverting a model that is already trained with sources types, which we refer to as *few-shot anomaly detection*. Under this problem setting, we propose to use a deep learning-based probabilistic model, which separates domain-level common features in grouped samples from sample-level features. We demonstrate the effectiveness of the proposed method in experiments using toy datasets and a real-world dataset containing aerial images of chemical factories.

2. Few-Shot Learning Anomaly Detection

The domain adaptation includes various problem settings [6]. In general, the purpose is to recognize samples (called target samples) obtained from a domain (called target domain), but there exists a reason not to train a model efficiently. Hence, additional samples (called source samples) are obtained from different domains (called source domains) and used for training. For classification, target samples are often assumed to be unlabeled, whereas source samples are labeled. For unsupervised anomaly detection, the number of target samples is assumed to be limited [7]–[9]. In any cases, domain adaptation methods train a model using both source and target samples.

The few-shot learning is a special case of the domain adaptation, where the number of available target samples is extremely limited (typically, 1–10 samples) and most domain adaptation methods are inapplicable [10]. Especially, few-shot learning methods train a model only using source samples and, *after training*, adjust the model every time a set of target samples is given [11]–[13]. The few-shot learning setting is free from the following bottlenecks. First, domain adaptation methods adapt a model to a target domain that is used at the training phase. Therefore, to adapt another target domain, they have to train another model from scratch. Second, even when domain adaptation methods train many models for many target domains, one has to determine which model is best for a given set of samples by using a domain classifier. Given an additional target domain, the domain classifier is retrained using all samples obtained from all target domains. In both cases, a huge amount of computational resources is required. Conversely,

Manuscript received July 17, 2021.
Manuscript publicized October 25, 2021.

[†] The author is with Graduate School of System Informatics, Kobe University, Kobe-shi, 657–8501 Japan.
[††] The author is with Mitsubishi Chemical Systems, Inc., Tokyo, 131–0045 Japan.
[†††] The author is with Graduate School of Engineering Science, Osaka University, Toyonaka-shi, 560–8531 Japan.
[††††] The author is with Faculty of Business Administration, Osaka Gakuin University, Toyonaka-shi, 564–8511 Japan.
a) E-mail: ksato@ai.cs.kobe-u.ac.jp
b) E-mail: nakata.satoshi.mx@mitsubishichem-sys.co.jp
c) E-mail: matsubara@sys.es.osaka-u.ac.jp
d) E-mail: kuniaki.uehara@ogu.ac.jp

DOI: 10.1587/transinf.2021EDL8063

the few-shot learning methods are assumed to quickly adjust a trained model to fit a given set of target samples without full retraining.

Given the above, we propose the *few-shot anomaly detection*. A model is trained using samples obtained only from source domains. After training, a small set of samples is obtained from a target domain. Then, the model is expected to detect anomalies in the small set. In typical unsupervised anomaly detection, the anomaly is defined as a group of samples that rarely or never appears in the training phase; the anomaly is sometimes referred to as "unseen samples", and any target samples are detected as anomalies. However, the purpose of the few-shot anomaly detection is to detect anomalies (in some sense depending on datasets) among target samples. The anomaly is defined as a sample with patterns substantially differing from the remaining of the given set of target samples.

The few-shot anomaly detection reflects a practical situation in a factory in which we aim to detect defective parts from a new type (i.e., target type) of product for which a production line has not yet been established. In this situation, it may become necessary to divert a model trained with older products (i.e., source samples) with sufficient inventory to detect anomalies in newer products. To our best knowledge, this is the first time to tackle the few-shot anomaly detection.

3. Methods

Variational Autoencoder: Autoencoder (AE) is a kind of DNNs generally consisting of two networks, called encoder and decoder [14]. The encoder first maps a sample x to a low-dimensional variable z as a compressed feature, and then the decoder maps it to a reconstruction $\tilde{x}$ so that the original sample x is recovered.

Variational autoencoder (VAE) is an AE extended from the viewpoint of generative model [15]. For a sample x in a data space $\mathcal{X} \subset \mathbb{R}^{N_x}$, consider a probabilistic model with latent variable z in a low-dimensional latent space $\mathcal{Z} \subset \mathbb{R}^{N_z}$ such that $p_\theta(x) = \int_z p_\theta(x|z)p(z)$. By introducing a variational distribution $q_\phi(z|x)$, the marginal log-likelihood of a sample x is lower-bounded by the evidence lower bound (ELBO) $-\mathcal{L}(x)$ as

$$\begin{aligned} &\log p_\theta(x) \geq -\mathcal{L}(x) \\ &\quad := -D_{KL}(q_\phi(z|x)\|p(z)) + \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right]. \end{aligned} \tag{1}$$

Then, the parameters θ and ϕ are updated to maximize the ELBO. $q_\phi(z|x)$ and $p_\phi(x|z)$ are implemented as an encoder and a decoder of the AE. Instead of the point estimates, they output the parameters of distributions. The second term of the ELBO corresponds the reconstruction error.

Because the encoder compresses an input x into a low-dimensional variable z, they retain only salient information and discard others. The AEs produce a large error for a group of samples that rarely or never appears at the training phase. Hence, the AEs are trained using normal samples, and their error is used as an anomaly score [16]. However, in few-shot anomaly detection, all target samples are detected as anomalies because they never appear at the training phase.

Multi-level VAE for Few-Shot Anomaly Detection: Suppose that a group $G = \{x_1, \ldots, x_{N_G}\}$ is composed of two or more observations that belong to the same class and is given as input. Multi-Level VAE (MLVAE) aims to generalize to unseen classes by extracting the common feature c_G (*content*) and the varying feature s_i (*style*) within a group [17], [18]. The ELBO for a group G is given by

$$\begin{aligned} -\mathcal{L}(G) := &-D_{KL}(q_{\phi_c}(c_G|G)\|p(c_G)) \\ &- \textstyle\sum_{x_i \in G} D_{KL}(q_{\phi_s}(s_i|x_i)\|p(s_i)) \\ &+ \textstyle\sum_{x_i \in G} \mathbb{E}_{q_{\phi_c}(c_G|G)} \mathbb{E}_{q_{\phi_s}(s_i|x_i)}\left[\log p_\theta(x_i|c_G, s_i)\right]. \end{aligned} \tag{2}$$

The posterior $q_{\phi_c}(c_G|G)$ is defined as the product of the distributions of c_i obtained from each member of the group:

$$q_{\phi_c}(c_G|G) \propto \textstyle\prod_{x_i \in G} q_{\phi_c}(c_i|x_i). \tag{3}$$

Assuming a multivariate normal distribution for $q_{\phi_c}(c_i|x_i)$, the accumulated posterior distribution $q_{\phi_c}(c_G|G)$ also becomes a multivariate normal distribution and thus the content variable c_G can be obtained using the reparameterization trick. Methods of disentangling style and content such as MLVAE have been used for various downstream tasks such as image-to-image translation task [19] and few-shot classification task [20] as they perform well even for classes that were not observed during training.

We propose to use MLVAE for the few-shot anomaly detection, where a domain is used instead of a class. During the training phase, a small set of samples obtained from one of the source domains is fed to MLVAE as an input group G. After training, a small set of samples obtained from a target domain is given to MLVAE. Then, the sample-wise reconstruction error $\mathbb{E}_{q_{\phi_c}(c_G|G)} \mathbb{E}_{q_{\phi_s}(s_i|x_i)}\left[\log p_\theta(x_i|c_G, s_i)\right]$ is obtained and used as the anomaly score of the sample x_i. Intuitively, MLVAE learns the permissible range of a given domain at the training phase. Then, MLVAE detects a sample as an anomaly if it is out of the permissible range estimated using the given set of target samples. We emphasize that, while MLVAE is never retrained using the target domain, it is expected to be generalized to any target domains, reducing false positives.

4. Experiments

Toy Datasets: We evaluated the proposed method with two toy datasets: the Street View House Numbers (SVHN) dataset [21] and the CIFAR-10 dataset [22]. The SVHN dataset consists of images of house numbers collected from Google Street View and contains color images of size 32 × 32, 73,257 for training and 26,032 for testing. The CIFAR-10 dataset contains color images of size 32×32 for 10 different classes, 5,000 per class for training and 1,000 per class for testing. We used one class as the target domain, and the remaining classes as the source domains. We kept all training samples unmodified as normal samples. We duplicated each test sample, converted one of two to an anomaly by rotating it 90 degrees clockwise, and kept the other as a normal sample; hence, from CIFAR-10, we obtained 1,000 normal samples and 1,000 anomalous samples per class for testing.

Factory Roof Dataset: We further evaluated the proposed method using a factory roof dataset consisting of aerial images of different chemical factories. The dataset contains color images of size 3000 × 4000, 759 for training and 25 for testing. All of the test images are of factories with rust on their roofs, which we aimed to detect. Different roofs have different appearances, as shown in Fig. 1. Preliminary experiments found that an AE trained using some roofs mistakenly detected the whole of an unseen (i.e., target) roof due to the difference in the appearances. Hence, we treated a roof as a domain, and examined the dataset in the few-shot anomaly detection. We obtained patches with a size of 100 × 100 from each roof image.

Models: Because domain adaptation methods train a model using both source and target samples [7]–[9], they are inapplicable to the few-shot learning. We compared MLVAE with the basic methods, AE and VAE. The encoder of the AE or VAE outputted an N_z-dimensional latent variable z. The encoder of MLVAE outputted a pair of an N_c-dimensional content c_G and an N_c-dimensional style s_i.

The encoder was composed of five convolution layers for the toy datasets and six convolution layers for the factory roof dataset. Each intermediate layer had a kernel size of 4, a stride of 2, and a padding of 1, and was followed by the batch normalization and the ReLU function. The last layer had a stride of 1 and no padding, and its kernel size was 2 for the toy datasets and 3 for the factory roof dataset. The decoder was composed of transposed convolution layers with feature maps of the same sizes as those of the encoder. As hyperparameters of each model, we searched the number of dimensions of latent variables in $N_z, N_c, N_s \in \{2, 4, 8, 16, 32, 64, 128, 256\}$ and the number of channels of the convolutional layer in $N_{conv} \in \{16, 32\}$. For MLVAE, we set the group size to $N_G = 10$, which is recommended in the original work [17]. At the training phase, we randomly took ten normal source samples and fed it to MLVAE as an input group G. For evaluating a target sample in SVHN and CIFAR-10 datasets, we took nine samples additionally from the same target domain (i.e., the same class) and built an input group G. Even through most samples in the target domain are expected to be normal because of the definition of the anomaly, the assumption is unnatural that all additional samples as normal. We varied the number of anomalous samples from zero to two and denote it in a bracket; for example, MLVAE (1/9) denotes MLVAE using a group G composed of a sample taken from the anomalous subset and eight samples taken from the normal subset in addition to a sample to be tested. For the factory roof dataset, we took ten patches randomly from the same domain (i.e., the same roof), built an input group G, and evaluated each pixel. We did not adjust the fraction of anomalous pixels in each patch.

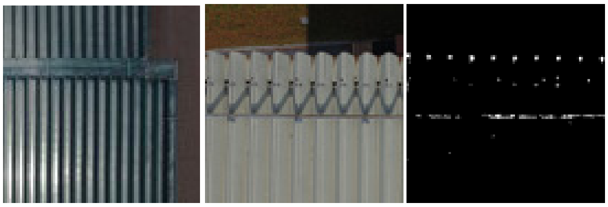


Fig. 1 Samples and an annotation from the factory roof dataset.

5. Results and Discussion

To evaluate the performance, we calculated the area under the receiver operating characteristic curve (ROC-AUC) for each model, as shown in Tables 1 and 2. For the factory roof dataset, we also provide the intersection over union (IoU) when the threshold was determined so that the true positive rate (TPR) was 50, 90, or 95%. MLVAE achieved the best performance for all datasets and evaluation metrics. While MLVAE got a lower performance with more anomalous samples for SVHN and CIFAR-10 datasets, it was always superior to AE and VAE. This result implies that MLVAE is robust to anomalies in the target domain. Figure 2 shows the histogram of anomaly scores of VAE and MLVAE. For these histograms, the Wasserstein distances between the distributions of normal and anomalous samples are shown in Table 3. For both source and target domains, the MVLAE achieved the largest Wasserstein distance, indicating that it achieved better separation between normal and anomalous samples.

Figure 3 shows a sample and the corresponding annotation of anomalous area from the factory roof dataset and heatmaps of anomaly scores. While anomaly scores of AE and VAE were high even in non-anomalous areas, MLVAE showed low anomaly scores in non-anomalous areas, confirming that MLVAE contributes to reduction of the false positive rate as described in Sect. 3.

When a single threshold value is determined, samples

Table 1 Resultant ROC-AUCs for SVHN and CIFAR-10 datasets.

	SVHN		CIFAR-10	
Model	target	source	target	source
AE	0.644	0.656	0.601	0.606
VAE	0.700	0.717	0.605	0.616
MLVAE (0/9)	**0.723**	**0.745**	**0.640**	**0.642**
MLVAE (1/9)	**0.719**	**0.741**	**0.637**	**0.635**
MLVAE (2/9)	**0.718**	**0.739**	**0.634**	**0.633**

Table 2 Resultant performances for factory roof dataset.

		IoU		
Model	ROC-AUC	@TPR95	@TPR90	@TPR50
AE	0.901	0.031	0.102	0.268
VAE	0.948	0.051	0.104	0.235
MLVAE	**0.972**	**0.130**	**0.197**	**0.294**

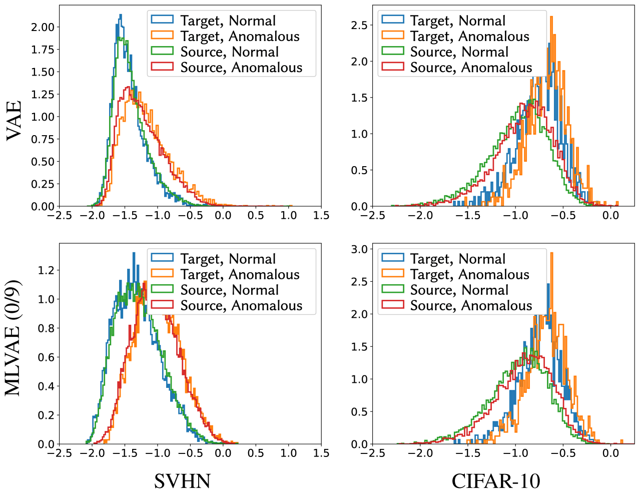
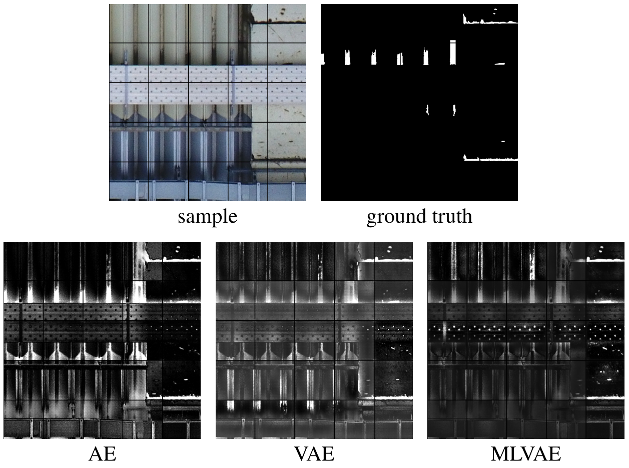


Fig. 2 Histogram for anomaly score when we set the class “1” as the target domain for the toy datasets.



sample

ground truth

AE

VAE

MLVAE

Fig. 3 A sample and the corresponding annotation of the rust from the factory roof dataset (top) and heatmaps of anomaly scores (bottom).

Table 3 Wasserstein distance between score distributions.

	SVHN		CIFAR-10	
source	target	source	target	source
VAE	0.050	0.052	0.053	0.067
MLVAE (0/9)	**0.172**	**0.195**	**0.060**	**0.073**
MLVAE (1/9)	**0.075**	**0.146**	**0.112**	**0.137**
MLVAE (2/9)	**0.072**	**0.183**	**0.108**	**0.130**

Table 4 Youden’s index for SVHN and CIFAR-10 datasets.

	SVHN			CIFAR-10		
Model	target	source	diff.	target	source	diff.
AE	0.210	0.195	0.015	0.163	0.121	0.042
VAE	0.294	0.279	0.015	0.181	0.150	0.031
MLVAE (0/9)	0.332	0.325	**0.007**	0.227	0.204	**0.023**
MLVAE (1/9)	0.326	0.314	**0.012**	0.223	0.198	**0.025**
MLVAE (2/9)	0.322	0.306	0.016	0.216	0.184	0.032

Table 5 Youden’s index for factory roof dataset.

Model	target	source	diff.
AE	0.688	0.687	**0.001**
VAE	0.731	0.720	0.011
MLVAE	0.754	0.753	**0.001**

or pixels with anomaly scores exceeding the threshold value are detected as anomalies. The threshold value is sometimes determined to maximize the difference between the true and false positive rates (called Youden’s index [23]). In practical few-shot anomaly detection, one has to divert the threshold determined only using source samples to target samples. Tables 4 and 5 summarize Youden’s index for the target samples when the threshold was determined using source or target samples. For the factory roof dataset, MLVAE achieved the best performance, and AE and MLVAE got the smallest performance gap. For both toy datasets, MLVAE achieved the best performance regardless of the fraction of anomalous samples, and MLVAE got the smallest performance gap with zero or one anomalous sample in the group G. Only when two anomalous samples are in the group G, the performance gap grows to a similar level to AE and VAE. This fact also implies that MLVAE is not just a good anomaly detector but also a good domain adapter to a small set of test samples even when some of them are anomalous, and hence it is suitable for few-shot anomaly detection.

6. Conclusion

We introduced the few-shot anomaly detection, a novel problem setting in which a group of samples belonging to the same domain is provided as input. To generalize to target domains, we proposed to use MLVAE, which separates features varying among a domain from sample-level features. The experimental results on the two toy datasets and the real-world dataset showed that our proposed method can robustly detect anomalies in target domains.

This work was partially supported by JSPS KAKENHI (19H04172, 19K20344).

References

[1] D.-B. Perng, S.-H. Chen, and Y.-S. Chang, “A novel internal thread defect auto-inspection system,” International Journal of Advanced Manufacturing Technology, vol.47, no.5-8, pp.731–743, 2010.

[2] X. Chen, J. Wang, and H. Ge, “Training Generative Adversarial Networks via Primal-Dual Subgradient Methods: A Lagrangian Perspective on GAN,” International Conference on Learning Representations (ICLR), 2018.

[3] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville,Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, “Brain tumor segmentation with Deep Neural Networks,” Medical Image Analysis, vol.35, pp.18–31, 2017.

[4] K. Lee, K. Lee, H. Lee, and J. Shin, “A simple unified framework for detecting out-of-distribution samples and adversarial attacks,” Advances in Neural Information Processing Systems (NeurIPS), 2018.

[5] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, “Concrete problems in ai safety,” arXiv, 2016.

[6] G. Wilson and D.J. Cook, “A Survey of Unsupervised Deep Domain Adaptation,” ACM Transactions on Intelligent Systems and Technology, vol.11, no.5, pp.1–46, 2020.

[7] G. Pang, C. Shen, H. Jin, and A.v.d. Hengel, “Deep weakly-supervised anomaly detection,” arXiv, 2019.

[8] Y. Koizumi, S. Murata, N. Harada, S. Saito, and H. Uematsu, “Sniper: Few-shot learning for anomaly detection to minimize false-negative rate with ensured true-positive rate,” ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.915–919, 2019.

[9] Z. Yang, I.S. Bozchalooi, and E. Darve, “Anomaly Detection with Domain Adaptation,” arXiv, 2020.

[10] Y. Wang, Q. Yao, J.T. Kwok, and L.M. Ni, “Generalizing from a Few Examples: A Survey on Few-shot Learning,” ACM Computing Surveys, vol.53, no.3, pp.1–34, 2020.

[11] H. Edwards and A. Storkey, “Towards a Neural Statistician,” International Conference on Learning Representations (ICLR), 2016.

[12] C. Finn, P. Abbeel, and S. Levine, “Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks,” 2017.

[13] W.Y. Chen, Y.C.F. Wang, Y.C. Liu, Z. Kira, and J.B. Huang, “A closer look at few-shot classification,” International Conference on Learning Representations (ICLR), 2019.

[14] G.E. Hinton and R.R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” Science, vol.313, no.5786, pp.504–507, 2006.

[15] D.P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” International Conference on Learning Representations (ICLR), 2014.

[16] V. Škvára, T. Pevnỳ, and V. Šmídl, “Are generative deep models for novelty detection truly better?,” OOD, ACM SIGKDD 2018 Workshop, 2018.

[17] D. Bouchacourt, R. Tomioka, and S. Nowozin, “Multi-Level Variational Autoencoder: Learning Disentangled Representations from Grouped Observations,” AAAI Conference on Artificial Intelligence (AAAI), 2017.

[18] J. Nemeth, “Adversarial disentanglement with grouped observations,” Proc. AAAI Conference on Artificial Intelligence, vol.34, no.6, pp.10243–10250, 2020.

[19] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, “Multimodal unsupervised image-to-image translation,” Proc. European Conference on Computer Vision (ECCV), vol.11207, pp.179–196, 2018.

[20] M.F. Mathieu, J.J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun, “Disentangling factors of variation in deep representation using adversarial training,” Advances in Neural Information Processing Systems (NIPS), 2016.

[21] Y. Netzer and T. Wang, “Reading digits in natural images with unsupervised feature learning,” Advances in Neural Information Processing Systems (NIPS), 2011.

[22] Alex Krizhevsky, “CIFAR-10 and CIFAR-10 datasets.”

[23] W.J. Youden, “Index for rating diagnostic tests,” Cancer, vol.3, no.1, pp.32–35, 1950.