

Exploiting Training Stage Reconstruction Instability for Image Anomaly Detection

Yihua Wang

School of Mechanical, Electrical & Information Engineering, Shandong University, Weihai,
Shandong, 264200, China

202100800728@mail.sdu.edu.cn

Abstract. Image anomaly detection is an important and practically significant research area. Methods such as reconstruction and feature extraction are mainly utilized. For reconstruction methods, purely normal input images should be provided, which is close to supervised learning. For feature extraction methods, unsupervised learning ones without a designated feature learning orientation perform much worse than supervised learning ones. In this paper, the goal is to achieve anomaly detection based on unsupervised learning. Generative Adversarial Network (GAN) is used as the approach, similar to unsupervised learning feature extraction methods, but it focuses more on the feature variation rate rather than using the feature outputs for further classification. The results show an innovative phenomenon in the GAN training process and achieve clear anomaly detection in unsupervised learning. The feature variations reflect the learning state of the model. The feature variations of similar images will converge in the certain learning state, while anomalies will appear as outliers. This phenomenon reveals how the model is learning, particularly about the details. The learning process of the model is intermittent and tries to learn the main features before the details. A more difficult task makes the fluctuations more significant.

Keywords: Anomaly detection, generative adversarial network, image reconstruction.

1. Introduction

Image anomaly detection needs a mass of normal images or labeled images for normal ones and anomalies. Although using only normal images as input is considered unsupervised learning, it still requires that anomalies be picked out beforehand. These approaches require tremendous labor cost, and an incorrectly provided image will have a significantly negative effect on the performance, especially if the input data consists of only normal images. Additionally, datasets like tire X-ray images, lung CT images, and retina images are originally in the same format with minor anomalies. In these cases, a completely unsupervised learning method shows its superiority.

Unsupervised Anomaly Detection with Generative Adversarial Networks (AnoGAN) is a classical and traditional method [1]. The reimplementation of AnoGAN illustrates the perspective above. If the input also includes a minority of anomalies, using the anomaly score—which combines residual loss and discrimination loss—may fail to properly reconstruct images without anomalies after sufficient training epochs. One innovative approach is to focus on the latent space, as AnoGAN helps us obtain the latent vector z for each corresponding image through z -space searching.

Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and recent diffusion reconstruction models [2, 3], despite employing different architectures, all focus on reconstruction to establish a mutual mapping between the latent space and the input data. VAEs show clearer boundaries in latent space compared to GANs, which makes them more competent in interpolation tasks [4]. But in the latent space, every feature can become a meaningful image. There is no clustering phenomenon happened here, so that classification without labels to find the outliers based on conventional latent space cannot work. Apart from that, the boundaries between images within the same category contribute to the model's adaptability, whereas the boundaries between different categories may lack practical significance, which enhances the probability of detecting wrongly. A good approach to address all the issues mentioned above is Cluster Generative Adversarial Network (ClusterGAN). This article utilizes a mixture of discrete and continuous latent variables in order to create a non-smooth geometry in the latent space [5]. The discrete latent variables,

such as one-hot encodings, are used to predict the cluster to which the image belongs, while the continuous latent variables aim to provide diversity in the generated images. The results of the article display an effective clustering outcome, allowing a completely unsupervised Anomaly Detection can be designed using clusterGAN [5].

Regarding unsupervised feature extraction method, previous research is an insightful method [6]. Unlike conventional methods, this article uses K-means after a neural network to provide gradient backpropagation to the previous layers of the neural network, which gives a clear learning orientation for identifying cohesion within clusters and estrangement between clusters [6]. This method is also feasible for the task.

In this work, the reimplementation of AnoGAN reveals an interesting phenomenon: the model tends to learn the main features at an early learning stage. At this stage, the anomaly details seem not to be noticed, and the reconstruction ability is also not stable. Subsequently, once the model is perfectly trained, the situation reverses. The relationship between these stages shows an ability to find the anomalies.

2. Method

2.1. Dataset

Tire X-ray images are provided as dataset, which is a private dataset. It is respectively named as A, B, and C. Datasets A, B, and C all consist of 64 images. In dataset A, only the eighteenth image is an anomaly image with iron filings in it. Dataset B is similar but contains two anomaly images: the first and the nineteenth. Dataset C is a different category from Datasets A and B, and it is harder to distinguish visually than Datasets A and B. Dataset C has only one anomaly image, the nineteenth.

2.2. Network Architecture and Implementation Details

The classical Deep Convolutional Generative Adversarial Network (DCGAN) is utilized as a prototype. Use Rectified Linear Unit (ReLU) activation in generator for all layers except for the output, which uses Tanh. Use LeakyReLU activation in the discriminator for all layers. Use the output of the second-to-last layer of the discriminator as a feature [7, 8]. The feature is usually used for feature matching [9, 10], feature extraction, as inception score to assess image quality [11], or as part of the loss function [1, 8]. However, the variation of the features has not received sufficient attention.

In this research, an iteration has 100 epochs as main part for training both of generator and discriminator, then 50 epochs as side part only for training discriminator. The discriminator after the side part training will be saved at each iteration without disrupting the main training process. This method helps stabilizing the discriminator for each generator in a certain state, making the convergence effect more noticeable. These parameters can be changed to achieve a better performance. When the model is aware of the flaw, it is uncontrollable. Therefore, comparing the generated images at certain epochs is not effective. The best solution to this issue is to compare the feature variation rate over iterations for each image. The algorithm tries to draw a line chart to make it more visual.

The input for the discriminators is always the real input data. Discriminators at different stages will have varying abilities to extract features from them. The absolute value of the subtraction of feature outputs between two adjacent iteration is feature variation rate. The feature has 20 dimensions. For convenience, the feature will be reduced to a single dimension by summing the values.

2.3. Evaluation Indexes

Methods to evaluate the differences between the generated images at early and later stages have been tested. Firstly, If the pixel difference between the images is used as the score directly, similar images with a slightly shift will have a higher score rather than images with minor flaws. A neural network model based on the feature evaluation is more feasible. Eventually, the feature output of the GAN's discriminator tends to be the best evaluation. At early stage, the discriminator does not have

the ability to extract the feature of a detail as well as flaw, whereas at the later stage, the discriminator will have the ability. As a result, the absolute value of the difference in feature outputs for normal images will be comparatively lower than for anomaly images.

3. Result

Fig. 1 and Fig. 2 provide a clear detection of the anomaly images. Fig. 3 is more complex but still clear enough to detect the anomaly image. The results are generally consistent with the previous assumption. Line chart A is the clearest display of the assumption. After learning the features of main body, the model is going to learn the details information. This is the time to find the anomaly images.

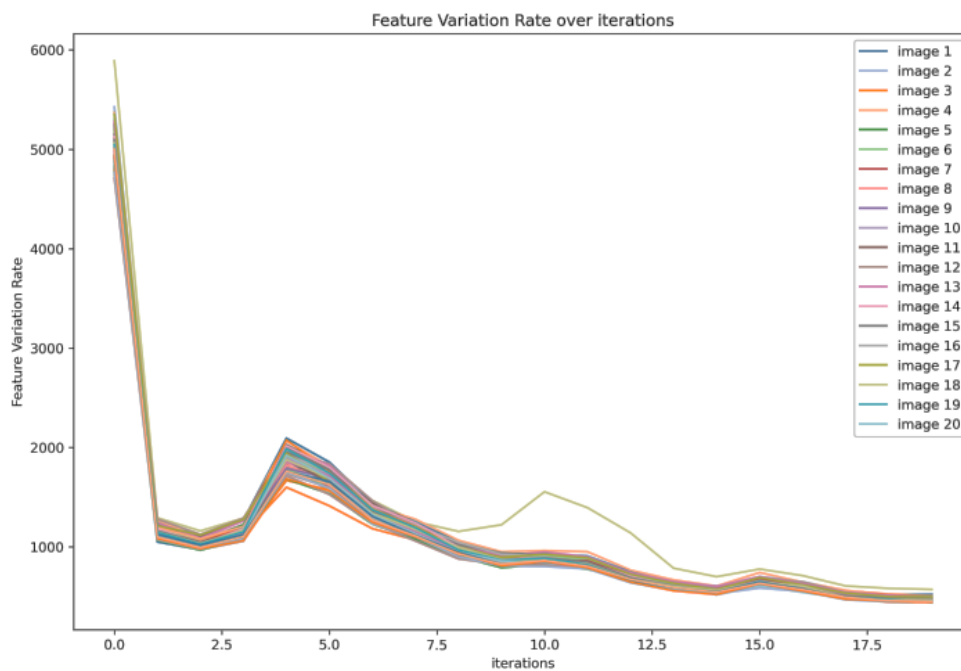


Figure 1. Result on dataset A, where image 18 is anomaly (Figure Credits: Original)

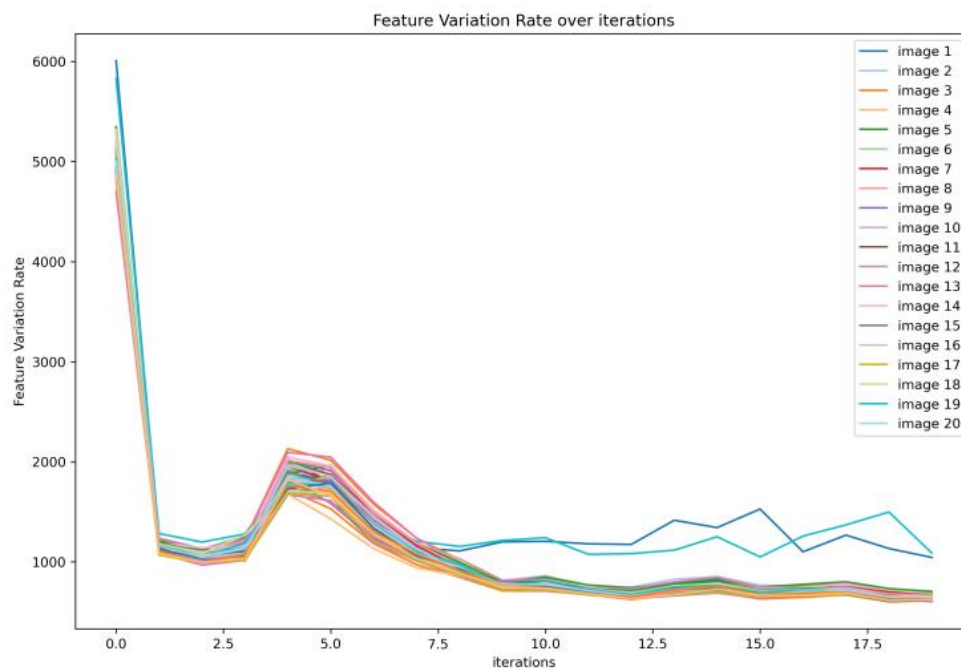


Figure 2. Result on dataset B, where image 1 and 19 are anomaly (Figure Credits: Original)

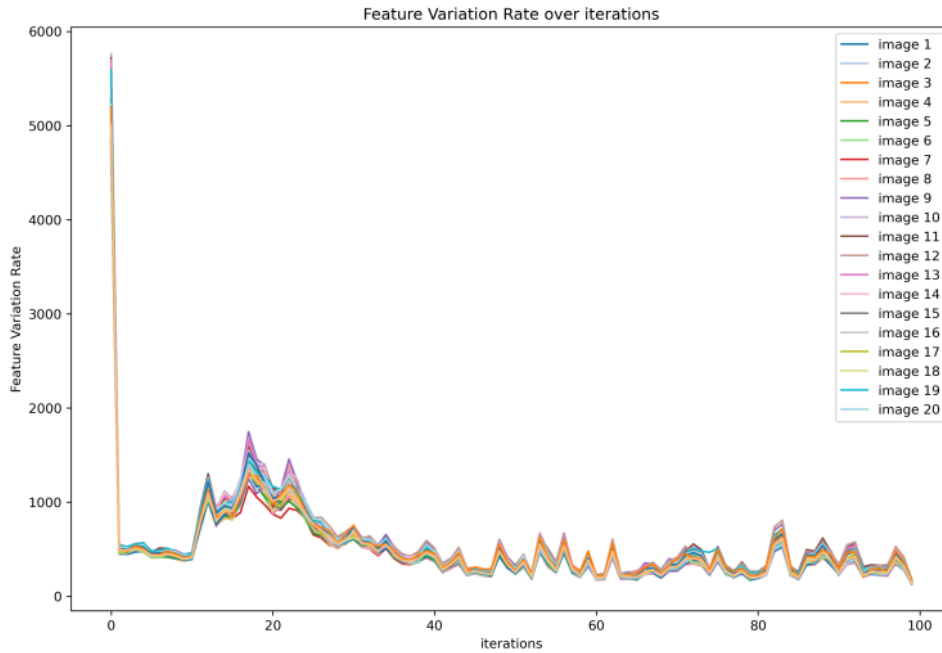


Figure 3. Result on dataset C, where image 19 is anomaly (Figure Credits: Original)

Another interesting observation is that anomaly images often change rapidly during the very first iteration, but their values are unstable and not useful for classifying the anomalies.

4. Discussion

4.1. Discussion of Results

Fig. 1 is a simple and easy-to-understand example. After an initial change, the process always experiences an unexpected bottleneck before significant improvements occur. The significant improvement is always unstable, whereas they are the most important stage to learn the main features. After the main features have been learned, the feature discrimination in normal images no longer changes significantly, while the discriminator begins to learn the features of flaws in anomaly images. Once this stage is finished, there is nothing to learn anymore. The feature discrimination will no longer change. As for Fig. 2, an increase in the number of anomaly images in the dataset makes it more difficult for the discriminator to converge during the flaw learning phase.

In this research, Fig. 1 and Fig. 2 are presented for simple problems, while Fig. 3 is used for complex and difficult ones. For simple problems, the main feature variations will remain low stably and rapidly, with only one peak, while for difficult ones, the feature variations will have plenty of peaks, valleys, and bottlenecks. At peaks, the feature variations of images are more discrete, and more conspicuous during early iterations. In contrast, at bottlenecks or valleys, the variations are more aggregated. The outlier anomalies may appear in this time.

Additionally, the learning process seems to be intermittent. The feature variations may alternate between valleys and peaks until the main features are learned. After that, the improvement of discriminator's ability to detect details becomes more noticeable in anomaly images.

4.2. Other Attempts

Methods such as VAE and Wasserstein Generative Adversarial Network (WGAN) have also been attempted [12]. For the VAE, regardless of whether the features are from the encoder or the decoder, the results are not appropriate for classifying anomalies. For WGAN, the results of generated images using z-space searching from [1] show that these methods always ignore detailed information, which makes them unsuitable for anomaly detection. Using Wasserstein distance as a loss function seems to be insensitive to details, while other features are learned quite well. This phenomenon can be

directly applied to AnoGAN. Replace the GAN with WGAN, using all images as input data, anomalies included. The other approaches stay the same. However, the degree of details ignored by WGAN cannot be managed through its parameters.

5. Conclusion

The results show an innovative phenomenon in the GAN training process and achieve clear anomaly detection in unsupervised learning. The feature variations reflect the learning state of the model. The feature variations of similar images will converge in the certain learning state, while anomalies will appear as outliers. This phenomenon reveals how the model is learning, particularly about the details. The learning process of the model is intermittent and tries to learn the main features before the details. A more difficult task makes the fluctuations more significant.

This research is complete unsupervised anomaly detect, and fill a gap in the field. The more profits it provides are related to how the model is learning, particularly about the details, which will have a profound influence on deep learning training field.

However, the appearance of different learning states varies with each training session, and when the anomalies appear as outliers is uncontrollable. How to control the model to learn features of different extents is still an ongoing challenge. Another issue is that a large dataset should be tested in the future.

References

- [1] Schlegl, Thomas, et al. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. International conference on information processing in medical imaging, 2017: 146-157.
- [2] Ho Jonathan, Ajay Jain, Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 2020, 33: 6840-6851.
- [3] Wyatt Julian, Leach Adam, Schmon Sebastian M, et al. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 650-656.
- [4] Kingma Diederik P., Max Welling. Auto-encoding variational bayes. arXiv preprint, 2013: 1312.6114.
- [5] Mukherjee Sudipto, Asnani Himanshu, Lin Eugene, Kannan, Sreeram Clustergan: Latent space clustering in generative adversarial networks. Proceedings of the AAAI conference on artificial intelligence. 2019, 33 (1): 4610-4617.
- [6] Yang Bo, Fu Xiao, Sidiropoulos Nicholas D, et al. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. international conference on machine learning, 2017: 3861-3870.
- [7] Szandała Tomasz. Review and comparison of commonly used activation functions for deep neural networks. Bio-inspired neurocomputing, 2021: 203-224.
- [8] Radford Alec, Luke Metz, Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint, 2015: 1511.06434.
- [9] Salimans Tim, Goodfellow Ian, Zaremba Wojciech, et al. Improved techniques for training gans. Advances in neural information processing systems, 2016, (29):1-8.
- [10] Huang Xun, Serge Belongie. Theoretical Analysis of Feature Matching in GANs. arXiv preprint, 2017: 1705.04740.
- [11] Liu Shaohui, Wei Yi, Lu Jiwen, et al. An improved evaluation framework for generative adversarial networks. arXiv preprint, 2018: 1803.07474.
- [12] Martin Arjovsky, Soumith Chintala, Léon Bottou, Wasserstein GAN. arXiv preprint, 2017: 1701.07875.