이상 탐지를 위한 합성 데이터 생성 및 성능 분석

황주효 · 진교홍*

창원대학교

Synthetic Data Generation and Performance Analysis for Anomaly Detection

Ju-hyo Hwang · Kyo-hong Jin*

Changwon National University

E-mail : ajsj2200@gmail.com, khjin@changwon.ac.kr

요 약

자기 지도 학습을 이용한 이상 탐지는 일반적으로 합성 데이터를 생성해 정상과 이상을 학습하고, 실제 이상 데이터를 테스트 데이터로 사용하여 이상 탐지 성능을 측정한다. 정상 데이터와 유사한 합성 데이터를 생성하기 위해 기존 연구에서는 원본 이미지에서 특정 패치를 자르고 붙이는 식으로 합성 데이터를 생성한다. 이런 방식에서 정상 데이터와 유사한 정도는 패치 개수와 크기에 따라 달라지므로 이상 탐지 성능에 영향을 미칠 수 있다. 본 연구에서는 패치 크기 및 개수를 다르게 하여 합성 데이터를 생성한 뒤 사전 학습된 모델을 사용하여 정상 데이터와의 유사성 측정 및 분석을 진행하였고 모델을 학습시켜 이상 탐지 성능을 측정하여 보았다.

ABSTRACT

Anomaly detection using self-supervised learning typically generates synthetic data to learn to classify normal and ab normal, and uses real abnormal data as test data to measure anomaly detection performance. In a study using this meth od to generate synthetic data similar to normal data, anomaly detection was carried out by generating synthetic data by cutting and pasting a specific patch from the original image. In this way, the degree of similarity to normal data depen ds on the number and size of patches, which affects anomaly detection performance. In this paper, synthetic data were generated by varying patch sizes and numbers, and then similarity and analysis with normal data were conducted using a pre-trained model, and anomaly detection performance was measured by learning the model.

키워드

Self-Supervised Learning, Anomaly Detection, Synthetic Data, Image Embedding

Ⅰ. 서 론

이상 탐지란 정상 데이터에서 벗어난 이상 데이터를 탐지하는 작업을 의미한다. 그러나 실제 현장에서는 이상 데이터가 없거나 그 수가 매우 적어 충분한 이상 데이터 확보가 필요한 지도 학습을 적용하기 어렵다. 따라서 이상 데이터가 필요하지 않은 자기지도 학습 방식이 연구되고 있다.

자기지도 학습 방식을 사용한 이상 탐지[1, 2]는 주로 정상 데이터를 이용해 정상 데이터와는 다른 합성 데이터를 생성한 후, 정상 데이터와 합성 데이터를 학습에 활용한다. 기존 연구에서는 원본 이미지에서 특정 패치를 자르고 붙이는 식으로 합성 데이터를 생성하였다. 이런 방식은 패치 개수와 크기에 따라 정상 데이터와 유사한 정도가 달라진다. 정상 데이터와 유사한 합성 데이터일수록 정상 데이터와 비슷한 특징 벡터를 가지게 된다. 모델은 이를 분류하기 위해 더 복잡한 결정 경계를 가지게 되므로, 패치 개수와 크기는 이상 탐지 성능에 영향을 미치는 파라미터가 될 수 있다.

본 논문에서는 패치 크기 및 개수를 다르게 하여 합성 데이터를 생성한 뒤 모델을 학습시켜 정상 데이터와 임베딩 거리 측정 및 분석을 진행하였다. 그 후 각 모델을 학습시켜 이상 탐지 성능을 측정하였다.

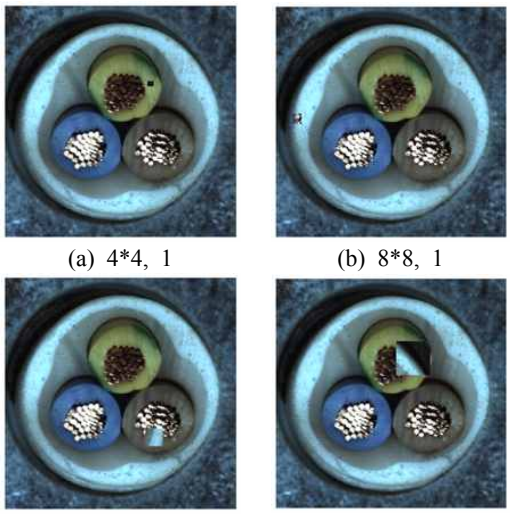
* corresponding author

II. 파라미터 및 학습 모델

본 논문에서는 MVTec AD 데이터셋[3]을 이용해 합성할 패치 크기와 개수(패치 가로 크기 * 패치 세로 크기, 개수)를 (4*4, 1), (4*4, 2), (8*8, 1) (8*8, 2), (16*16, 1), (32*32, 1)로 변경해가며 실험을 진행하였다.

아래 그림 1과 그림 2는 합성 패치 크기 및 개수에 따른 합성 데이터 생성 결과를 나타낸 것이다. 합성한 패치 크기가 커질수록 합성한 패치의 주변 픽셀과 더 불연속적인 특징을 가지는 것을 확인할 수 있다. 이러한 특징은 정상 데이터와의 유사성에 영향을 주는 요소가 될 수 있다[2].

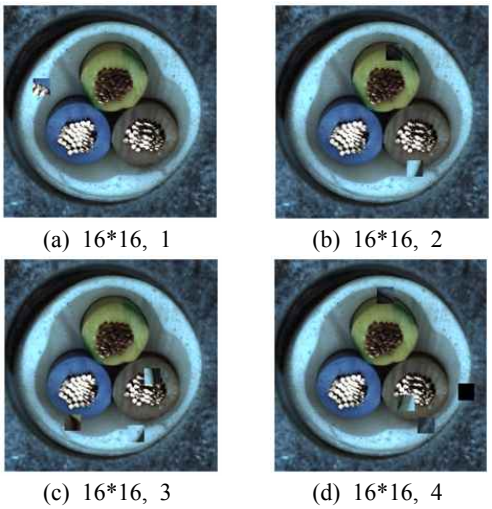
아래 그림 3과 같이 학습 모델은 ImageNet-21K에서 사전 학습되어 가중치가 고정된 MLP-Mixer(L/16)[5]를 사용하였다. 그 뒤 완전 연결층(Fully Connected Layer)을 추가하여 분류 학습을 수행하였다.



(a) 4*4, 1 (b) 8*8, 1

(c) 16*16, 1 (d) 32*32, 1

그림 1. 합성 패치 크기에 따른 합성 데이터 생성 결과



(a) 16*16, 1 (b) 16*16, 2

(c) 16*16, 3 (d) 16*16, 4

그림 2. 합성 패치 개수에 따른 합성 데이터 생성 결과

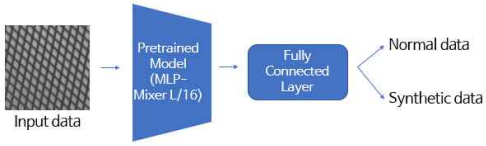


그림 3. 학습 모델

III. 실험 결과

모델 학습을 진행한 후 이상 데이터와 합성 데이터를 입력하여 특징 벡터를 추출하였다. 그 뒤 Umap[4]을 통해 특징 벡터를 2차원으로 차원 축소 후 임베딩 거리를 측정하였다. 임베딩 거리는 훈련 정상 데이터의 특징 벡터로부터 마할라노비스 거리(Mahalanobis Distance)를 사용하였다. 이상 탐지 성능은 테스트 데이터로 정상 데이터와 이상 데이터를 입력하여 Binary Cross-entropy Error(BCE) 및 AUC 점수로 측정하였다.

아래 표 1은 각 클래스에 따른 정상 데이터의 특징 벡터와 이상 데이터의 특징 벡터간의 마할라노비스 거리(A), 정상 데이터의 특징 벡터와 합성 데이터의 특징 벡터간의 마할라노비스 거리(B)를 측정하여 둘의 차를 아래 식 (1)을 통해 정규화한 값(Z)와 테스트 에러(BCE) 및 AUC 점수를 나타내었다.

$$Z = \frac{A - B}{A} \quad (1)$$

Z 값이 클수록 정상 데이터와 이상 데이터의 임베딩 거리보다 정상 데이터와 합성 데이터의 임베딩 거리가 더 가깝다는 것을 의미한다. 즉 Z 값이 클수록 정상 데이터와 유사한 합성 데이터가 생성되었다는 것을 의미한다. 표 1에서 합성 패치의 크기가 커질수록, 합성 패치의 개수가 많아질수록 Z 값이 감소하는 것을 확인할 수 있다. 그리고 Z 값이 클수록 평균적으로 더 낮은 테스트 에러를 가진 것을 확인할 수 있다.

표 1. 클래스별 실험 결과

class	case	Z	BCE	AUC
grid	4*4, 1	0.32	0.28	96.6
	4*4, 2	0.06	0.27	96.6
	8*8, 1	0.29	0.35	94.6
	8*8, 2	-0.12	1.05	88.6
	16*16, 1	-0.25	0.41	94.3
	32*32, 1	-1.60	3.89	73.7
tile	4*4, 1	-0.07	0.52	86.3
	4*4, 2	-0.10	0.78	74.0
	8*8, 1	0.23	0.58	79.8
	8*8, 2	-0.63	0.56	79.0
	16*16, 1	-0.05	0.54	74.9
	32*32, 1	-0.13	1.38	69.0
cable	4*4, 1	-0.04	0.81	67.1
	4*4, 2	-0.76	0.84	67.6
	8*8, 1	-0.42	0.76	76.3
	8*8, 2	-1.25	1.05	68.0
	16*16, 1	-0.98	1.05	74.7
	32*32, 1	-1.29	1.20	75.6

Ⅳ. 결론

본 논문에서는 합성 데이터 생성 시 합성 패치 크기와 개수에 따른 이상 탐지 성능을 분석하였다. 그 결과 합성 패치의 크기가 커질수록, 합성 패치의 개수가 많아질수록 정상 데이터와 임베딩 거리가 멀어지는 것을 확인할 수 있었다. 뿐만 아니라 정상 데이터와 유사한 합성 데이터를 생성하여 학습할수록 평균적으로 이상 탐지 성능이 증가하는 것을 확인할 수 있었다.

Acknowledgement

“이 연구는 산업통상자원부에서 시행하는 공정혁신 시뮬레이션 센터 구축사업의 지원을 받아 수행되었음“

References

[1] Li CL, Sohn K, Yoon J, and Pfister T. CutPaste: Self-Supervised Learning for Anomaly Detection and Localization. 2021. *arXiv: 2104. 04015* [cs.CV].

[2] Hannah MS, Jeremy T, Benjamin H, Bernhard K. Natural Synthetic Anomalies for Self-Supervised Anomaly Detection and Localization. 2022. *arXiv: 2109.15222* [cs.CV].

[3] Bergmann P, Batzner K, Fauser M, Sattlegger D, and Steger C. “The MVTec Anomaly Detection Dataset: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection,” eng. In *International journal of computer vision* 129.4. 2021, pp. 1038-1059.

[4] McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426.*

[5] Ilya T, Neil H, Alexander K, Lucas B, Xiaohua Z, Thomas U, Jessica Y, Daniel K, Jakob U, Mario L, et al. Mlp-mixer: An all-mlp architecture for vision. *arXiv:2105.01601*, 2021. [cs.CV].