KICSP 한국 융합신호처리학회

융합신호처리학회논문지 Vol. 23, No. 4 : 194~199 December. 2022
DOI : 10.23087/jkicsp.2022.23.4.002

이미지 분할 여부에 따른 VQ-VAE 모델의 적대적 예제 복원 성능 비교

김태욱[1], 현승민[2], 홍정희[1*]
[1]연세대학교 소프트웨어학부, [2]연세대학교 컴퓨터정보통신공학부

Comparison of Adversarial Example Restoration Performance of VQ-VAE Model with or without Image Segmentation

Tae-Wook Kim[1], Seung-Min Hyun[2], Ellen J. Hong[1*]
[1]Division of Software, Yonsei University
[2]Department of Computer & Telecommunications Engineering, Yonsei University

요 약 다양하고 복잡한 영상 데이터 기반의 산업에서 높은 정확도와 활용성을 위해 고품질의 데이터를 위한 전처리가 요구된다. 하지만 기존 이미지 또는 영상 데이터와 노이즈를 결합해 기업에 큰 위험을 초래할 수 있는 오염된 적대적 예제가 유입될 시 기업의 신뢰도 및 보안성, 완전한 결과물 확보를 위해 손상되기 이전으로의 복원이 필요하다. 이를 위한 대비책으로 기존에는 Defense-GAN을 사용하여 복원을 진행하였지만, 긴 학습 시간과 복원물의 낮은 품질 등의 단점이 존재하였다. 이를 개선하기 위해 본 논문에서는 VQ-VAE 모델을 사용함과 더불어 이미지 분할 여부에 따라 FGSM을 통해 만든 적대적 예제를 이용하는 방법을 제안한다. 먼저, 생성된 예제를 일반 분류기로 분류한다. 다음으로 분할 전의 데이터를 사전 학습된 VQ-VAE 모델에 전달하여 복원한 후 분류기로 분류한다. 마지막으로 4등분으로 분할된 데이터를 4-split-VQ-VAE 모델에 전달하여 복원한 조각을 합친 뒤 분류기에 넣는다. 최종적으로 복원된 결과와 정확도를 비교한 후 분할 여부에 따른 2가지 모델의 결합 순서에 따라 성능을 분석한다.

• 주제어 : VAE, VQ-VAE, FGSM, Adversarial Attack, Adversarial Example, Defense-VAE

Abstract Preprocessing for high-quality data is required for high accuracy and usability in various and complex image data-based industries. However, when a contaminated hostile example that combines noise with existing image or video data is introduced, which can pose a great risk to the company, it is necessary to restore the previous damage to ensure the company's reliability, security, and complete results. As a countermeasure for this, restoration was previously performed using Defense-GAN, but there were disadvantages such as long learning time and low quality of the restoration. In order to improve this, this paper proposes a method using adversarial examples created through FGSM according to image segmentation in addition to using the VQ-VAE model. First, the generated examples are classified as a general classifier. Next, the unsegmented data is put into the pre-trained VQ-VAE model, restored, and then classified with a classifier. Finally, the data divided into quadrants is put into the 4-split-VQ-VAE model, the reconstructed fragments are combined, and then put into the classifier. Finally, after comparing the restored results and accuracy, the performance is analyzed according to the order of combining the two models according to whether or not they are split.

• Key Words : VAE, VQ-VAE, FGSM, Adversarial Attack, Adversarial Example, Defense-VAE

Received 01 November 2022, Revised 08 December 2022, Accepted 10 December 2022

*** Corresponding Author** Ellen. J. Hong, Division of Software, Yonsei University, 1 Yonseidae-gil, Wonju, Gangwon-do, Korea.
E-mail: ellenhong@yonsei.ac.kr

Ⅰ. 서론

높은 정확도를 확보한 분류 모델일지라도, 입력데이터가 노이즈와 결합한 오염데이터일 경우 모델의 성능은 급격히 감소한다. 또한, 노이즈가 포함되어 오염된 이미지는 외적으로 식별이 불가능하도록 변조할 수 있으므로 기업은 결과물 완성도에 대한 위험에 노출되게 된다. 이를 방지하기 위해 기존에는 Defense-GAN[1]으로 노이즈를 제거하고자 하였으나 긴 학습 시간과 저품질의 복원 성능, 불완전한 노이즈 제거로 인해 보완이 불가피하였다.

따라서 본 연구에서는 Defense-VAE[2]에서 제안한 방식을 바탕으로 VQ-VAE 모델을 이용하여 이미지의 분할 여부, 임베딩 벡터 수, 모델 파이프라인 유무에 따른 복원 성능을 비교하고자 한다.

Ⅱ. 분할 이미지 복원 성능 검증

2.1 VQ-VAE 모델을 이용

보다 나은 복원 성능을 확보하기 위하여, 입력된 이미지 데이터가 존재하는 확률 분포를 찾아 Latent Vector를 이용해 이와 유사한 이미지를 생성할 수 있는 VAE 모델을 활용하고자 한다. 하지만 기존의 VAE 모델의 경우 디코더와 결합할 때 latent 들이 무시되는 Posterior Collapse 문제가 발생할 위험이 있으므로, 학습모델로 VQ-VAE[3]을 선정하였다.

2.2 VQ-VAE 모델 학습

학습 데이터는 MNIST[4] 데이터셋을 사용한다. Fig. 1의 (a)와 같이 이미지 분할을 적용하지 않은 MNIST 데이터셋을 학습한 VQ-VAE 모델을 No-Split-VAE라고 정의하고 No-Split-VAE 모델이 생성한 이미지를 확인한다. 마지막으로, Fig. 1의 (b)와 같이 이미지를 4등분으로 분할한 MNIST 데이터셋을 학습한 VQ-VAE 모델을 4-Split-VAE 모델이라고 정의하고 4-Split-VAE 모델이 생성한 이미지를 확인한다. No-Split-VAE 모델과 4-Split-VAE 모델 모두 Table 1과 같은 구성을 가지며, Input Shape에 따른 내부 노드 구성의 차이만을 가진다.

Fig. 1. Ground-truth Images and Reconstructed Images. (a) Non-Split; (b) Split

Table 1. Layers of VQ-VAE Model

Layer		Hyper Parameter
Encoder	Conv1	kernel=(4,4), filter=128, stride=2, padding=1
	Conv2	kernel=(4,4), filter=128, stride=2, padding=1
	Residual_Conv1	kernel=(3,3), filter=128, stride=2, padding=1
	Residual_Conv1	kernel=(1,1), filter=128, stride=2, padding=0
Embedding Layer		n_Embeddings=768
Decoder	Residual_Conv	kernel=(1,1), filter=128, stride=2, padding=0
	Residual_Conv	kernel=(3,3), filter=128, stride=2, padding=1
	ConvTranspose	kernel=(2,2), filter=128, stride=2, padding=0
	ConvTranspose	kernel=(3,3), filter=1, stride=2, padding=0

2.3 FGSM을 활용한 적대적 예제 생성

적대적 예제를 생성하기 위해, 신경망의 성능이 저하되는 방향의 Gradient를 Input Data 적용하여 적대적 예제를 생성하는 기법인 FGSM (Fast Gradient Sign Method) [5]을 사용한다. FGSM에서 적대적 예제의 훼손 정도인 Epslion 값은 0.3에서 육안으로도 확인할 수 있는 정도로 훼손 정도가 높았기에, 이를 바탕으로 0.05씩 증가하며 다양한 훼손 정도에 따른 MNIST 적대적 예제를 생성한다.

2.4 적대적 예제 복원 수행

첫째, MNIST 적대적 예제를 MNIST 분류 모델에 입력값으로 전달하여 분류 정확도를 측정한다. 둘째, No-Split-VAE 모델에 MNIST 적대적 예제를 입력하여 복원된 이미지를 얻는다. 이후, 복원된 이미지를 분류 모델에 입력값으로 전달하여 분류 정확도를 측정한다. 셋째, MNIST 적대적 예제를 4등분 후, 이를 4-Split-VAE 모델에 입력 값으로 전달하여 4개의 복원된 이미지를 얻는다. 이후, 4개의 이미지를 본래 이미지의 위치와 부합하도록 재구성하여 하나의 복원 이미지를 얻는다. 마찬가지로, 해당 이미지를 분류 모델에 입력값으로 전달한 뒤 분류 정확도를 측정한다. Fig 2의 (a), (b), (c)는 각각 적대적 예제, No-Split-VAE의 적대적 예제 복원 이미지, 4-Split-VAE 모델의 적대적 예제 복원 이미지를 보여준다.

(a) (b) (c)

Fig. 2. Adversarial Examples and Restoration Images. (a) Adversarial Examples; (b) Non-Split (c) Split

2.5 적대적 예제 복원 성능 비교

Fig. 3 Original 그래프와 같이, 분류 모델은 노이즈가 포함된 적대적 예제 자체를 분류할 경우 Epsilon이 증가함에 따라 성능이 급격하게 낮아짐을 알 수 있다. Fig. 3 VAE의 경우, Epsilon의 증가에도 불구하고 적대적 예제를 분류할 경우보다 모델의 분류 및 복원 성능이 좋음을 알 수 있다.

Fig. 3의 4-Split-VAE의 경우, Epsilon 1.5 미만의 구간에서는 No-Split-VAE 모델보다 성능이 좋았으나, Epsilon 1.5 초과 구간에서는 No-Split-VAE 모델이 4-Split-VAE 모델보다 성능이 좋음을 확인할 수 있다.

Fig. 2의 (a)에서 볼 수 있듯, 이미지 분할 유무와 관계없이 Epsilon이 0.15보다 작은 경우에는 적대적 예제의 노이즈를 육안으로 구분하기 어렵다. 이러한 적대적 예제를 4-Split-VAE 모델이 No-Split-VAE 모델보다 정상적으로 복원했다는 점에서 IT 산업에서의 이미지 데이터에 가해질 수 있는 비가시적 적대적 공격에 대해 예방이 될 수 있다. 반면에 Epsilon이 0.15보다 큰 경우에는 육안으로도 적대적 예제임을 판별 가능하므로 본 연구의 취지에 부합하지 않음을 알 수 있다.

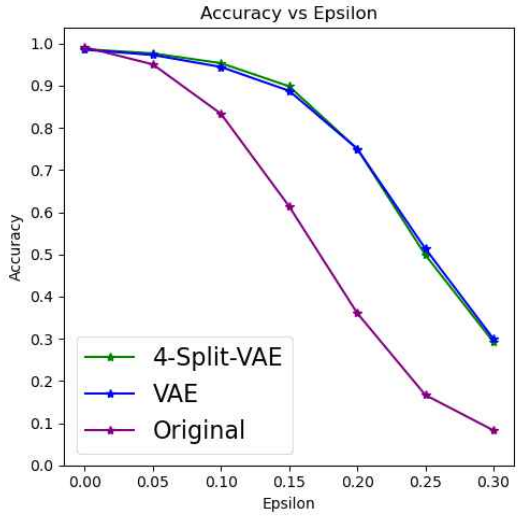


Fig. 3, Adversarial Example Restoration Result

Ⅲ. 임베딩 벡터 수에 따른 복원 성능 검증

3.1 Embedding Vector 수 조정

VQ-VAE 모델에서는 입력 데이터로부터 추출된 특징 벡터와 가장 유사한 이산공간 안의 임베딩 벡터를 선택한다. 선택된 임베딩 벡터는 디코더에 전달되고 이를 통해 모델을 학습시키게 된다. 이때 모델의 복원 성능을 극대화하기 위해 이미지 분할 여부 각각의 경우에 대해 Embedding Vector의 개수를 조정하여 Epsilon 증가에 따른 분류 정확도를 비교한다.

3.2 Embedding Vector 수 최적화

VQ-VAE 모델에서 Embedding Vector 수가 적을 경우, 모델의 학습은 데이터셋의 중요 특징들을 적은 수의 Embedding Vector만으로 표현하지 못할 수 있다. 반대로 너무 많을 경우 데이터셋에 과적합 되는 문제가 있으므로 적절한 Embedding Vector를 찾는 것이 중요하다. 본 연구는 비가시적인 적대적 예제의 복원 성능 개선을 목표로 하고 있으므로 Epsilon이 낮은 범위에서의 분류 정확도 수치를 기준으로 최적의 Embedding Vector의 수를 선택하였다. Table 2에서 알 수 있듯이, No-Split-VAE의 경우 Embedding Vector의 수가 900일 때 복원 성능이 가장 뛰어났다. 또한 Table 3에서 알 수 있듯이 4-Split-VAE 경우에는 Embedding Vector의 수가 300일 때 복원 성능이 가장 뛰어났다.

Table 2. No-Split-VAE Accuracy

Embedding	Accuracy by Epsilon						
	0	0.2	0.4	...	2.6	2.8	3.0
100	0.976	0.971	0.963		0.522	0.433	0.346
200	0.977	0.972	0.966		0.531	0.423	0.325
300	0.978	0.974	0.968		0.484	0.372	0.282
400	0.979	0.974	0.968		0.556	0.462	0.381
500	0.977	0.974	0.968		0.635	0.553	0.465
600	0.976	0.972	0.967	...	0.566	0.466	0.371
700	0.980	0.975	0.969		0.607	0.530	0.434
800	0.983	0.979	0.974		0.478	0.367	0.285
900	0.983	0.981	0.975		0.488	0.385	0.298
1000	0.981	0.978	0.971		0.573	0.476	0.385

Table 3. 4-Split-VAE Accuracy

Embedding	Accuracy by Epsilon						
	0	0.2	0.4	..	2.6	2.8	3.0
100	0.9836	0.9787	0.974		0.401	0.330	0.278
200	0.9816	0.9781	0.973		0.403	0.322	0.259
300	0.9874	0.9838	0.978		0.495	0.397	0.316
400	0.9847	0.9820	0.977		0.471	0.388	0.312
500	0.9859	0.9822	0.978		0.368	0.286	0.220
600	0.9874	0.9833	0.980	...	0.403	0.329	0.265
700	0.9873	0.9836	0.979		0.444	0.352	0.277
800	0.9868	0.9842	0.979		0.393	0.312	0.239
900	0.9872	0.9840	0.979		0.413	0.323	0.256
1000	0.9873	0.9845	0.980		0.412	0.323	0.250

Ⅳ. 모델 파이프라인 복원 성능 검증

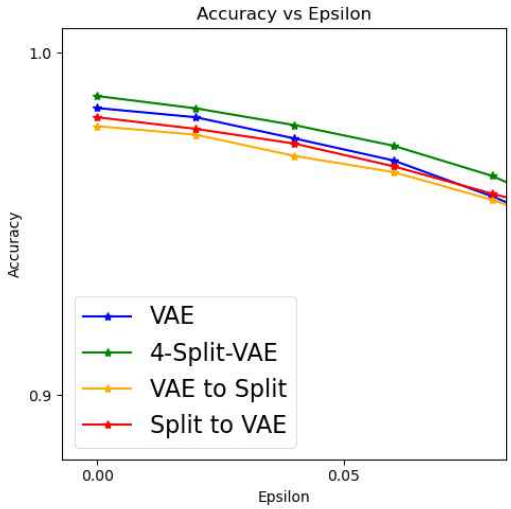
4.1 VQ-VAE 모델 파이프라인 설계

실험 3.2에 따르면 Epsilon이 낮은 경우에는 이미지를 분할하였을 때 복원 성능이 전체적으로 우수했으며, 반대로 Epsilon이 높은 경우에는 이미지를 분할하지 않고 복원하였을 때의 성능이 더 우수했다. 따라서 이러한 Trade-Off를 최소화하고 복원 성능의 보편성을 높이고자 VAE 모델과 4-Split-VAE 모델을 결합하여 결합 순서에 따른 두 가지 파이프라인 모델을 형성하였다. VAE to Split 파이프라인 모델의 경우 적대적 예제를 4등분하여 4-Split-VAE 모델을 통과시킨 후, 생성된 1차 복원 이미지 4장을 결합한 뒤 VAE 모델에 통과시켜 최종 복원 이미지를 생성한다. Split to VAE 파이프라인 모델의 경우 적대적 예제를 VAE 모델에 통과시킨 후, 생성된 1차 복원 이미지를 4등분한 뒤, 4-Split-VAE 모델에 통과시켜 최종 복원 이미지를 생성한다. 최종적으로 VAE, 4-Split-VAE, VAE to Split, Split to VAE의 총 4가지 모델을 Epsilon에 따른 분류 정확도 그래프를 통해 적대적 예제 복원 성능을 비교한다.

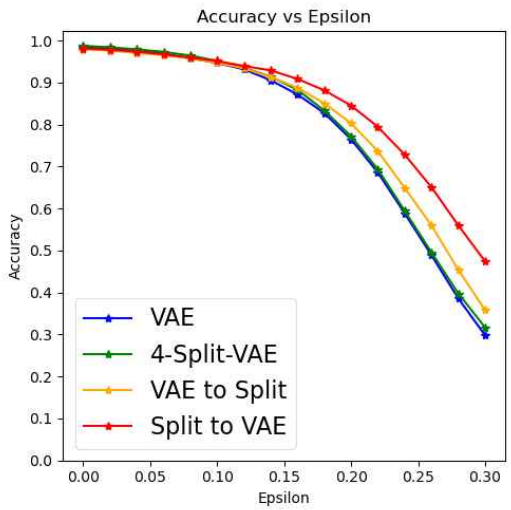
4.2 VQ-VAE 모델 파이프라인 복원 결과

Fig. 4 (a)에서 보듯이 Epsilon이 낮은 경우에는 4가지 모델 중 4-Split-VAE 모델이 가장 복원 성능이 우

수하며, 이와 반대로 (b)에서 보듯이 Epsilon이 높은 경우에는 Split-to-VAE 모델이 가장 우수하다.



(a)



(b)

Fig. 4 Adversarial Example Restoration Result
(a) Epsilon 0 to 0.08; (b) Epsilon 0 to 0.3

Ⅴ. 결론 및 향후 연구

본 연구에서는 이미지의 분할 여부에 따라 VQ-VAE 모델이 육안으로 노이즈를 판별할 수 없는 비가시적 적대적 예제로부터 어느 정도의 복원 성능을 보일 수 있는지를 실험하였다. 그 과정에서 이미지 분할 여부와 더불어 Embedding Vector의 수 조정 및 모델 파이프라인 형성에 따른 복원 성능을 비교하였다. 그 결과 Epsilon 크기가 작을 때와 클 때, 이미지 분할이 복원 성능향상에 미치는 영향의 Trade-off를 모델 파이프라인이 최소화하지는 못했다. 그러나 비가시적 구간에서의 복원 이미지에 대한 모델별 분류 정확도 격차가 미미하며 그 수치가 모두 1.0에 가까운 높은 정확도를 보인다. 이와 더불어 가시적 구간에서는 그 차이가 뚜렷하고 Split to VAE 파이프라인의 정확도가 가장 우수하다. 따라서 본 연구의 취지에 부합함과 더불어 모델 파이프라인을 통한 적대적 예제 복원 성능이 향상되었으므로 Defense-VAE로서의 의의가 있다. 이에 확장적으로 기존의 이미지 분할을 4등분에서 더욱 세분화하여 이미지의 크기 대비 어느 정도의 비율로서 분할을 진행하는 것이 Epsilon이 클 때 최적의 성능을 도출하는지에 관해 추가적인 연구가 필요하다.

ACKNOWLEDGMENTS

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No, 2022R1F1A1074273).

REFERENCES

[1] Pouya Samangouei, Maya Kabkab, Rama Chellappa. (2018). Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models. ICLR 2018, arXiv:1805.06605

[2] Xiang Li, Shihao Ji. (2019). Defense-VAE: A Fast and Accurate Defense against Adversarial Attacks. MLCS 2019, arXiv:1812.06570

[3] Aaron van den Oord, Oriol Vinyals, & Koray Kavukcuoglu. (2017). Neural Discrete Representation Learning. NIPS 2017, arXiv:1711.00937

[4] Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine, 29(6), 141-142.

[5] Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy. (2015). Explaining and Harnessing Adversarial Examples. ICLR 2015, arXiv:1412.6572

저자소개

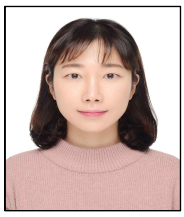
김 태 욱 (Tae-Wook Kim)

2021년 3월~현 재 : 연세대학교
소프트웨어학부(재학)
관심분야 : 딥러닝, 머신러닝,
컴퓨터비전

현 승 민 (Seung-Min Hyun)

2020년 3월~현 재 : 연세대학교
컴퓨터정보통신공학부(재학)
관심분야 : 딥러닝, 머신러닝,
컴퓨터비전

홍 정 희 (Ellen J. Hong)



2013년 2월 : KAIST
전기및전자공학과(공학박사)
2016년 10월~2018년 1월 :
동서대학교 컴퓨터공학부 조교수
2018년 2월~2019년 8월 : KT
융합기술원 선임연구원
2019년 9월~현재 : 연세대학교 소프트웨어학부 조교수
관심분야 : 인공지능, 시스템 모델링 시뮬레이션,
시뮬레이션 기반 최적화, 디지털 트윈, 지능형 시스템