컨볼루션 오토인코더를 활용한 이미지 적대적 공격 방어 기법 연구

박태정* , 손배훈**, 서성관**, 윤주범***

*세종대학교 (학부생), **세종대학교 (대학원생), ***세종대학교 (교수)

Research of Defense Method on Image Adversarial Attack Using Convolution AutoEncoder

Tae-Jeong Park*,Seong-Kwan Seo,Bae-Hun Son**,Joo Beom Yun***

*Sejong University(Undergraduate student), **Sejong University(Graduate student), *Sejong University(Professor),

요 약

심층신경망을 이용한 인공지능 기술은 이미지나 음성 분류 문제에 높은 성능을 보여주지만, 공격자가 입력데이터에 Perturbation을 첨가하여 원본을 조작 해 오작동을 일으키는 적대적공격 (Adversarial Attack) 에 취약하다는 것이 여러 연구들을 통해 확인되고 있다. 본 논문에서는 적대적 공격을 이미지를 낮은 차원으로 압추갛고 복원하는 과정에서 노이즈가 제거되는 AutoENcoder 의 특징과 이미지 특징 추출에 높은 성능을 보여 분류와 영상처리에서 사용되는 CNN (Convolutional AutoEncoder)를 결합하여 CAE (Convolutional AutoEncoder)를 구성하였고 이를 활용하여 방어하는 기술을 실험하고 그 구조를 제안한다. 공격에 사용 된 모델은 CNN 기반 분류모델 Inception3 이고 노이즈가 제거 된 적대적 예제는 약 90% 정확도로 제 성능을 다 하지 못하여 CAE를 활용해 노이즈를 제거하는 구조는 적대적 공격을 방어하는 기술로 활용 될 수 있음을 알 수 있다.

I. 서론

심층신경망을 이용한 이미지 분류 / 인식 및 오디오 분류 / 인식의 성능은 기계학습 알고리즘들과 비교하여 성능이 뛰어남을 알 수 있다. 그러나 신경망의 입력에 사용되는 값에 perturbation을 첨가한 적대적 예제(Adversarial Attack) 는 심층신경망의 오작동을 일으킬 수 있음이 여러 연구들을 통해서 확인되고 있다.

적대적 공격의 대부분은 위에서 언급한 것처럼, 실제 Input 데이터에 물리적으로 혹은 입력값 자체에 Perturbation을 첨가한다. 이를 오분류하게 만드는 Perturbation는 최적화 기법들을 통해 찾을 수 있는데 대표적으로 FGSM , PGD 가 있다. 현실에서 사용되기 위해서는 물리적 공격이나 디지털공격을 가해야 한다. 이를 모두 막는 방법은 대표적으로 적대적 예제를 학습시키거나 입력 값을 검증하는 방법이다. 적대적 예제를 학습시키는 것은 적대적 예제가 생성되는 것 이상으로 학습을 시켜야 하기 때문에 어려움을 겪고 있다. 하지만 입력값을 검증하는 방법을 사용한다면, 학습 데이터를 많이 갖추어야 할 수고를 덜 수 있다.

이에 본 논문에서는 이미지를 낮은 차원으로 압축하고 복원하는 과정에서 노이즈가 제거되는 AutoEncoder 특징과 이미지 분류와 특징 추출에 높은 성능을 보이는 CNN (Convolution Neural

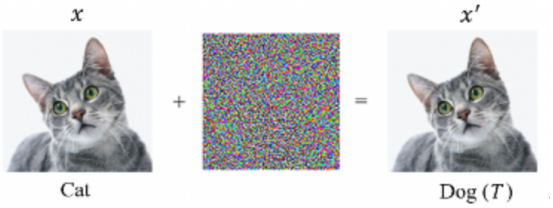
Netowrk)를 결합하여 CAE (Convolution AutoEncoder)를 구성한다. CAE 는 Perturbation 이 섞여 있을 수도 있는 이미지가 신경망의 Input 으로 사용되기 전 AutoEncoder 로 이미지의 노이즈를 제거한다. 이를 CNN 기반 이미지 분류 모델 Inception3에서 적대적 공격과 본 논문의 방어 기법을 실험하여 적대적 공격을 방어하는 기법을 제시한다.

II. 관련 연구

이 장에서는 적대적 교란 신호 (Perturbation) 과 CAE (Convolutional AutoEncoder) 에 대해 서술한다.

2.1 적대적 공격 (Adversarial Attack)

이 절에서는 적대적 공격과 적대적 공격 기법을 서술한다. 적대적 공격이란, 인공지능에 사용되는 입력을 조작하여 딥러닝 모델에 Input 으로 사용되었을 때 조작되기 전과 다른 클래스로 분류되도록 하는 방법이다. 그림1 은 적대적 공격의 예시로 이미지에 Perturbation (Noise) 를 추가하여 실제 데이터와는 완전히 다른 클래스로 분류되도록 한다.



[그림1] 적대적 공격 예시

적대적 공격을 위해 input 에 perturbation 을 추가하는 다양한 기법이 적용 되었다. Goodfellow[1] 가 제시한 FGSM (Fast Gradient Sign Method) 방법은 손실함수의 미분 값이 target class 로 향하도록 perturbation을 생성하고 bias를 만들어 올바른 클래스로 분류하지 못하도록 한다.

$$minimized(x, x^{'})$$

$$s.t. F(x^{'}) = T, x^{'} \ is \ valid$$

식1. perturbation 최소화

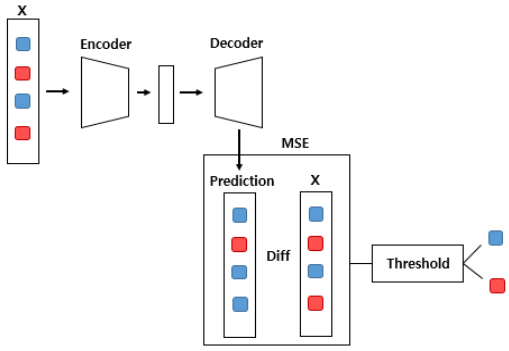
그림에서 보이는 것처럼, 완전히 다른 클래스로 인지 되도록 하지만, 사람의 눈에 명확하게 noise 가 보인다면 실제로 사용되기 어려운 적대적 예제이기 때문에 식에서 보이는 것처럼 이 차이는 최소화 되어야 한다. 이처럼 적대적 공격은 Input 값을 조작하여 원하는 class 로 오분류 되도록 하는 것이 목적이다.

2.2 적대적 공격 방어 기법 연구

적대적 공격을 알아차리거나 (감지) 적대적 예제 임에도 불구하고 올바르게 분류해내는 것을 적대적 방어 (Adversarial Defense) 라고 정의한다. 이 적대적 방어를 만드는 기법 중 적대적 훈련 (Adversarial Training) 은 가장 고전적이고 기본적인 기법으로. 적대적 예제를 생성해내는 보델에서 생성 된 적대적 예제와 원래 학습 데이터셋으로 함께 모델을 학습시켜 적대적 공격에 강건해지는 모델을 만드는 방법이다. 적대적 훈련은 Szegedy[6] 가 제안 한 L-BFGS에서 소개 된 방어 기법이다. 본 논문에서는 적대적 훈련 (학습) 하는 것이 아니라 노이즈를 제거하고 특징을 추출하여 학습 기법보다 조금 더 효율적인 방법을 제시한다.

2.2 CAE (Convolutional AutoEncoder)

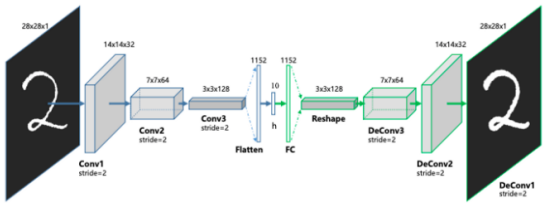
이 절에서는 본 논문에서 사용 된 딥러닝 신경망 구조인 AE 와 CNN 구조를 AutoEncoder 에 결합시킨 CAE (Convolutional AutoEncoder) 에 대해 서술한다.



[그림2] 임계값 설정으로 이상탐지에 사용되는 AE 구조

AE(AutoEncoder) 신경망은 대표적인 비지도

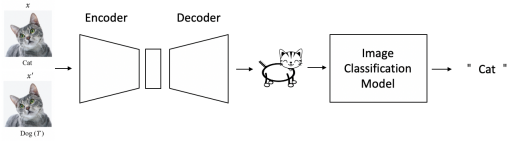
학습 모델로 별도의 레이블 없이도 인코더 계층과 디코더 계층으로 구성되어 입력과 동일 한 값이 출력되도록 학습하는 모델이다. 정상 데이터가 신경망에 입력되면 인코더 층에서 압축되고 디코더 층에서 입력과 똑같이 복원되는 것을 목표로 복원 데이터의 손실을 최소화하는 과정을 통해 정상 데이터의 특징을 학습한다. 그림 1은 테스트 데이터 X가 모델의 학습된 신경망에 입력되는 과정을 나타낸 것이다. X는 정상과 비정상 데이터로 구성되어 있다. X가 신경망을 거쳐 복원값(Prediction)이 반환되면 원본 X와 복원값을 비교하여 그 차이가 임계값(Threshold) 보다 클 경우 비정상 데이터로 분류하고, 작은 경우 정상 데이터로 분류하며 이상탐지를 수행한다. 이처럼 AutoEncoder 는 모델자체의 특성을 이용하여 특징추출, 이상탐지, 노이즈 제거 등 여러 가지 분야에 적용 될 수 있다. 이를 이미지나 영상분야에 적용시키기 위하여 이미지/영상 분류에 특히 높은 성능을 보이는 CNN (Convolutional Neural Network) 구조를 오토인코더의 네트워크 구조에 결합시킨 것이 CAE (Convolutional AutoEncoder) 이다. AE 의 인코와 디코더 층을 Convolutional Layer를 사용한다면 더욱 정확한 차원축소가 가능하여 특징을 더욱 잘 추출하고, 복원 해 낼 수 있다. 실제로 CAE 에 관한 연구도 활발하게 진행되고 있다.

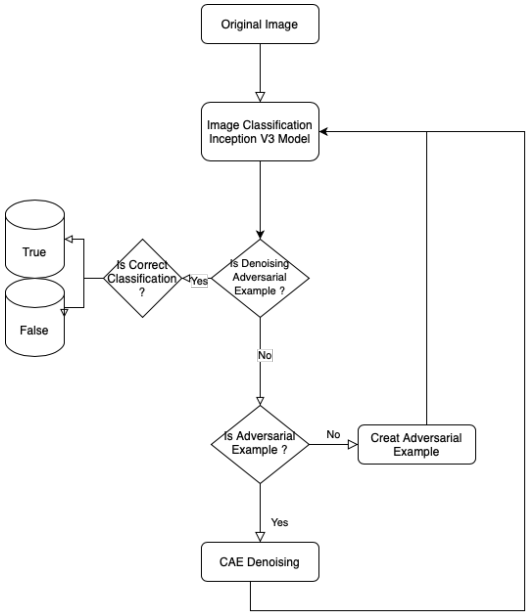


[그림3] CAE 구조

III. 구조제안과 실험결과

이 장에서는 해당 논문의 핵심주제인 CAE를 활용하여 Perturbation을 제거하고 이미지 분류 모델에서도 정상 작동하는지 확인하기 위한 실험을 진행하고, 그 구조를 제안한다.





[그림4] CAE를 활용한 Perturbation 제거와 실험과정

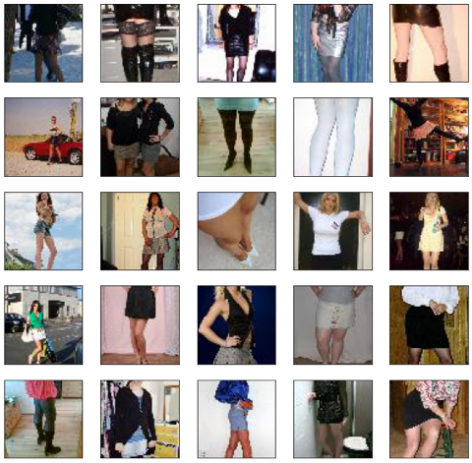
기존 적대적 공격은 고양이처럼 보이는 사진을 Perturbation을 첨가하여 개의 클래스로 인식 시키게 만들었다. 이를 본 논문에서는 Perturbation 이 첨가되어 실제로 고양이처럼 보이지만 개의 클래스로 분류되는 이미지를 CAE 에 넣고, 다시 한 번 이미지 분류모델에 넣는 실험구조를 설계했다.

본 논문에서 제시하는 그림4 의 구조를 반영 할 경우 CNN을 적용한 AutoEncoder 구조 (CAE)를 통과하면서 노이즈(Perturbation)가 제거된다. 노이즈가 제거 된 이 이미지를 적대적 공격에 성공했던 이미지 분류 모델 inveption v3 에 다시 input 으로 사용하면 노이즈가 섞여 'dog'로 분류되던 이미지가 정상적으로 'cat ' 으로 분류되게 된다.

실험에 사용 된 이미지 분류모델은 Google 의 Inception v3 이다. inception v3 모델에 적대적 공격을 수행하고, 해당 output을 본 논문에서 제시한 구조에 input 으로 사용하면서 CAE 가 노이즈(Perturbation)를 제거하여 해당 Input 이 적대적 예제로 기능을 상실하여 적대적 공격을 방어하는지 실험 했다.

Denosing 역할을 하는 CAE (Convolutional AutoEncoder) 는 ImageNet Dataset을 활용했다. 해당 Image Data 에 Gausian Noise를 추가하여 Noise 가 없는 원본

이미지로 targeting 되도록 학습시켜 노이즈를 제거하는 CAE 구조를 설계했다.



그림[5]. 데이터셋. ImageNet

원본이미지

toaster 로 인식되는 조작된 이미지

[그림6] 실험에 사용 된 원본이미지와 적대적 예제

입력사진 Label	공격label	CAE solution 적용 후	정확도
persian cat	toaster	persian cat	86.14%
Border collie	snail	border collie	96.25%
baseball	basketball	baseball	89.55%
car	snake	car	92.35%

표1. 실험결과

그림5처럼 눈에 보이지 않는 노이즈가 섞였을 때 오탐을 일으키는 이미지 적대적 공격은 모두 성공했다. 이 이미지들을 CAE 에 넣은 뒤 다시 Inception 모델에 넣었을 때 높은 정확도로 원본 이미지의 Label을 나타냈다. 정확도는 사진마다 100번 테스트 한 뒤 모델 confidence 의 평균을 나타낸 것이다.

IV.결론

그림6와 표1에서 보이는 것처럼, 사람 눈에 보이지 않는 정도의 노이즈도 고양이에서 토스트기로 전혀다른 클래스로 분류했다. 이처럼 적대적 공격은 인공지능의 분류 성능에 큰 문제를 일으킬 수 있으며, 적대적 공격 방어 기법에 관한 연구는 지속적으로 진행되어야 할 것이다.

적대적 공격을 방어하기 위한 본 논문에서 적용 한 CAE를 활용하여 노이즈를 제거하는 방법은 90% 의 효과가 있음을 보였으며, 이미지 분류를 사용하는 산업에서 더 강건한 인공지능을 사용하는 것에 한 방법으로 사용 될 수 있을 것이다. 더 강건한 모델을 만들기 위해 이후 AutoEncoder 의 특징을 살려 MSE (Mean Squared Error) 에 기반 한 임계값에 따라 적대적 예제 데이터를 사후처리하는 연구를 진행할 예정이다.

[참고문헌]

[1] C. Xie, Y. Wu, L. Maaten, A. L. Yuille and K. He, “Feature Denoising forImproving Adversarial Robustness,” Proceedings of the IEEE Conference onComputer Vision and Pattern Recognition (CVPR), 2019.

[2] I. J. Goodfellow, J. Shlens and C. Szegedy, “Explaining and HarnessingAdversarial Examples,” Proceedings of the 3rd International Conference onLearning Representations (ICLR), 2015.

[3] A. Madry, A. Makelov, L. Schmidt, D. Tsipras and A. Vladu, “Towards DeepLearning Models Resistant to Adversarial Attacks,” Proceedings of the 6thInternational Conference on Learning Representations (ICLR), 2018.