적대적 공격 감지와 GAN 을 이용한 복원

장준영 [1], 노민주 [1], 권준석 [2]
[1] 중앙대학교 소프트웨어학부 학부생
[2] 중앙대학교 소프트웨어학부 교수

junjang99@cau.ac.kr, romj98@cau.ac.kr, jskwon@cau.ac.kr

Adversarial Detection and Purification with GAN

Junyoung Jang[1], Minju Ro[1], Junseok Kwon[1]
[1]School of Computer Science and Engineering, Chung-Ang University

요 약

인위적인 공격뿐만 아니란 현실 세계에서도 이미지에 노이즈가 추가되는 경우가 있다. 이를 해결하기 위한 많은 연구가 이루어지고 있지만, 적대적 공격에 강건한 모델은 기존의 모델에 비해 원본 이미지에 대해 정확도가 떨어진다는 문제점이 있다. 따라서 본 논문은 생성 모델을 활용하여 적대적 예제에 강건한 모듈을 제안한다. 또한, 적대적 공격을 탐지하는 모듈을 활용하여 적대적 예제뿐만 아니라 원본 이미지에 대해서도 정확도를 높이는 방법을 제안한다.

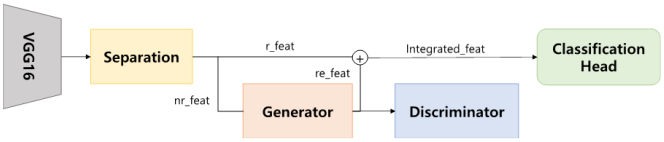
1. 서론

대표적인 공격인 FGSM(Fast Gradient Sign Method)과 PGD(Projected Gradient Descent)에 강건한 모델의 필요성을 느끼고 이를 보완한 모델을 제안한다. 또한, 적대적 공격에 의해 의도적으로 노이즈가 추가된 이미지뿐만 아니라 노이즈가 추가되지 않은 기존 이미지에 대해서도 정확도가 떨어지지 않는 모델을 만들고자 하였다. 따라서 Feature Separation and Recalibration(FSR)[1] 모듈을 활용하여 적대적 공격에도 강건하고 원본 이미지에 대해서도 강건한 모듈들을 제안한다. FSR_GAN 은 생성모델을 사용하여 노이즈에 강건하지 않은 특징에서 원래 특징을 복원할 수 있도록 하였다. 이를 통해 강건하지 않은 특징을 원래의 특징과 비슷하게 만들어 적대적 예제에도 강건한 모델을 생성할 수 있다. FSR_Attack 은 적대적 공격을 받은 특징인지 탐지하는 모듈을 추가하였다. 따라서 모든 입력이 아니라 공격을 받은 이미지에 대해서 선택적으로 적대적 방어를 하여 원본 이미지에 대한 성능을 향상시켰다.

2. 방법론

특징 복원

FSR_GAN 의 경우 강건한 특징과 강건하지 않은 특징으로 나눈 이후 강건하지 않은 특징을 GAN 의 생성자와 판별자를 적용하여 특징을 복원한다. 본 실험에서는 backbone 으로 VGG16[2]을 활용한다. FSR_GAN 은 쉽게 적용 가능한 프레임워크이므로 ResNet-18[3]과 같은 다양한 backbone 및 분류 모듈을 사용하는 것이 가능하다.



(그림 1) FSR_GAN 구조도

backbone 을 통과하여 얻은 특징 정보를 분리 모듈을 통해 강건한 것과 강건하지 않은 특징으로 분리한다. 분리 모듈의 출력 결과는 마스크의 형태가 되고 이때 각각은 0 과 1 사이의 값을 갖는 소프트 마스크를 활용한다. r_{feat}는 모델이 예측을 하는데 필요한 강건한 특징을 의미하고, nr_{feat}는 모델이 오분류를 유도하는 강건하지 않은 특징을 의미한다. $original_{feat}$의 경우 backbone 을 통과한 직후의 특징맵이다. r_{feat}와 nr_{feat}는 다음과 같은 식을 통해 구할 수 있다.

$$r_{feat} = mask \times original_{feat}$$
$$nr_{feat} = (1 - mask) \times original_{feat}$$

강건하지 않은 특징을 생성자를 통해 복원하는 과정을 거치는데 이때 WGAN 의 생성자를 사용한다. 생성자 모듈을 통과하여 얻은 특징은 판별자 모듈로 전달이 된다. 판별자의 경우 공격을 받지 않은 입력을 VGG16 에 통과시켜 얻은 특징 맵과 비교하여 판별을 진행한다. 복원하여 얻은 특징맵을 다시 음의 마스크 값을 곱하고 강건한 특징맵과 더하여 최종적인 특징맵을 계산한다. 해당 특징맵을 분류 모듈에 전달하여 분류 과제를 진행한다. 손실함수의 경우 아래의 식과 같다.

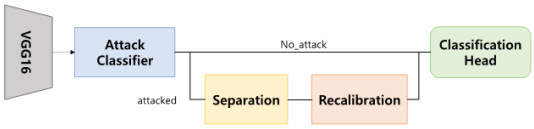
(수식 1)

$$L = L_{cls} + \lambda_{fake-validity} \cdot L_{fake-validity} + \frac{1}{|L|} \Sigma \left(\lambda_{sep} \cdot L_{sep}^{l} + \lambda_{rec} \cdot L_{rec}^{l} \right)$$

Defense GAN[4]의 fake validity score 를 계산하여 기존 FSR 손실함수에 더한다. 이를 통해 더 정확한 복원과정이 이루어지도록 유도한다. 판별자는 정화된 특징과 실제 특징을 비교하여 gradient penalty 를 통해 업데이트한다. [1]의 손실함수에 [4]의 fake validity score 를 추가로 고려하는 형태이다. λ는 하이퍼파라미터, L_{cls} 는 분류 손실함수이다.

적대적 공격 탐지

FSR_Attack 은 기존의 적대적 방어 모델의 한계점인 원본 이미지에 대한 성능을 보완하기 위해 제안하는 모듈이다.



(그림 2) FSR_Attack 구조도

백본을 통과한 후에 얻은 특징맵에 대해 convolution block 과 분류 모듈을 적용하여 해당 입력이 적대적 공격을 받은 이미지인지, 원본 이미지인지 판단한다. 공격을 받지 않은 특징이라고 판단하면 분류와 복원 모듈을 거치는 것이 오히려 성능 저하를 유발하므로 바로 분류 모듈로 전달하여 분류 작업을 진행한다. 만약 적대적 공격을 받은 특징이라고 판단하면 분류와 복원 모듈에 통과시켜 얻은 특징맵을 분류 모듈에 전달한다. 손실함수 식의 경우 하단과 같다.

(수식 2)

$$L = L_{cls} + \lambda_{attack} \cdot L_{attack} + \frac{1}{|L|} \Sigma \left(\lambda_{sep} \cdot L_{sep}^{l} + \lambda_{rec} \cdot L_{rec}^{l} \right)$$

공격을 받았는지 판단하는 손실함수(L_{cls})와 [1]의 복원, 분리 손실함수의(L_{rec}, L_{sep}) 결합으로 구성한다. 각각을 λ로 표시한 하이퍼파라미터로 크기를 조정한다. 분리와 복원 손실함수의 경우 각 모듈이 정확한 분류를 하는지 확인하기 위한 보조 레이어를 통해 정확한 클래스를 예측할 때 높은 점수를 할당하는 방식으로 계산한다.

3. 실험 결과

실험은 본 논문에서 제시하는 모델인 FSR_GAN 과 FSR_Attack 모듈뿐만 아니라 이와 비교하기 위해 VGG16 [2], FSR[1] 모델을 동일한 환경에서 학습시켰다. GPU 로 RTX 3060 12GB 을 사용하였고 Window 10 운영체제 환경에서 실험을 진행하였다. 데이터 셋은 CIFAR-10 을 사용하였다. 각각 200 epoch, 256 batch size 로 학습시킨 결과이다. 해당 실험에서 시용한 attack 방법은 FSGM 과 PGD 이다. 결과 테이블에서 Original, FSGM, PGD-20, PGD-100 은 각각 원본 이미지, FSGM 공격, PGD 로 20 step, 100 step 만큼 공격을 받은 이미지에 대한 정확도를 나타낸다.

	Original	FSGM	PGD-20	PGD-100
VGG-16	81.46%	51.52%	41.41%	39.42%
FSR	80.67%	54.45%	46.22%	44.05%
FSR_GAN(ours)	77.71%	53.47%	**47.15%**	**46.20%**
FSR_Attack(ours)	**83.14%**	53.62%	45.64%	43.92%

기존의 FSR[1]의 경우 결과가 잘 나오는 값에 대해 점수를 높게 메기는 방식으로 진행되었다. FSR_GAN 에서는 판별자를 추가하여 적대적 학습 원리에 의해 노이즈를 없애는 과정을 진행하였으므로 좀 더 정확한 복원을 유도하다. 이러한 이유로 White Box model 에서도 더 높은 확률로 오분류를 예방한다.

표 1 의 결과를 확인해보면, FSR_Attack 모듈의 경우 VGG16 에 비해 공격받은 이미지에 대한 정확도가 높다. 또한, FSR 에 비해 원본 이미지에 대한 정확도가 높은 것을 확인할 수 있다. 이는 단순히 모든 이미지에 대해 노이즈를 제거하는 방식보다 공격을 받은 이미지를 구분하여 복원하는 과정을 거쳤기 때문에 정확도가 높아졌다. 따라서 적대적 방어에서의 한계점인 원본 이미지에 대한 결과를 보완하는 모델임을 증명할 수 있다.

4. 결론

본 논문에서는 적대적 공격 상황에서 기존의 FSR 에서 제안하는 복원 과정을 GAN 의 원리를 활용하여 고도화하는 FSR_GAN 을 제안한다. 또한, 적대적 방어 모델의 한계점인 원본 이미지에 대한 성능을 개선하기 위해 FSR_Attack 을 제안한다. 실험을 통해 결과를 비교하고 분류 과제에서의 성능이 우수함을 보임을 확인하였다.

참고문헌

[1] Kim, W. J., Cho, Y., Jung, J., & Yoon, S.-E. (2023). Feature separation and recalibration for adversarial robustness. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr52729.2023.00791

[2] Karen Simonyan and Andrew Zisserman. Very deep convo- lutional networks for large-scale image recognition. In ICLR, 2015.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.

[4] P. Samangouei, M. Kabkab, and R. Chellappa, “Defense-GAN: Protecting classifiers against adversarial attacks using generative models,” in Proc. Int. Conf. Learn. Representations(ICLR), 2018, pp. 1–17.