심층신경망에 대한 적대적 공격과 탐지 방법 조사

박태우, 최석환, 최윤호

부산대학교

A Survey of Adversarial Attack and Detection Methods for DNN

Taeu Bahk, Seok-Hwan Choi, Yoon-Ho Choi

Pusan National University

요 약

심층신경망(DNN) 기술은 최근 다양한 작업에서 놀라운 발전을 하고 있다. 하지만 심층신경망은 노이즈나 적대적 섭동(Adversarial Perturbation)에 취약하다. 심층신경망을 기만하기 위하여 데이터에 적대적 섭동을 추가한 산출물이 적대적 예제이다. 이 과정을 적대적 공격(Adversarial Attack)이라고 한다. 적대적 공격을 통해 생성되는 적대적 예제는 심층신경망이 잘못된 예측을 출력하도록 야기한다. 이는 실제로 심층신경망 기반 시스템을 도입 및 적용하는데 제한 요인이 된다. 따라서 적대적 공격과 탐지 기술은 중요하다. 본 논문에서는 적대적 공격과 탐지 방법을 조사하고자 한다.

I. 서론

심층신경망(DNN, Deep Neural Network)은 최근 몇 년 동안 다양한 작업에서 매우 발전하여 널리 사용되고 있다. 하지만 내재적으로 불확실성을 가진 심층신경망은 노이즈나 적대적 섭동(Adversarial Perturbation)에 취약하다. 이로 인해 데이터에 적대적 섭동을 명시적으로 생성하여 심층신경망을 기만 및 회피하는 적대적 공격(Adversarial Attack)이 등장하였다. 이 과정에서 생성되는 적대적 공격의 산출물이 적대적 예제(Adversarial Example)이다. 적대적 예제는 심층신경망이 높은 신뢰도(Confidence)로 잘못된 예측을 출력하게 만들 수 있다. 즉, 적대적 예제는 심층신경망의 오작동을 야기한다. 적대적 예제는 현실에서도 문제를 초래할 수 있다. 예컨대, 딥 페이크(Deep Fake)[1]와 같이 개인의 신원을 나타내는 영상을 조작하거나, 정지를 의미하는 표지판을 최저속도로 제한하는 표지판[2]으로 기만하여 자율주행 자동차가 위험한 행동을 수행하도록 만드는 범죄 등의 악용 사례가 발생할 수 있다. 적대적 공격과 탐지 방법에 대해 많은 연구가 있다. 하지만, 특히 적대적 탐지에 대해 조사하거나 연구한 국내 논문은 드물다. 따라서 본 논문에서는 대표적인 적대적 공격과 적대적 탐지 방법에 대해 조사하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서 적대적 공격 방법에 대해 소개한다, 3장에서는 적대적 탐지 방법에 대해 기술한다. 마지막으로, 4장에서는 본 논문의 결론을 맺는다.

II. 적대적 공격 방법

본 장에서는 적대적 공격 방법들에 대해 소개한다. 적대적 공격 방법은 대표적으로 Fast Gradient Sign Method (FGSM)[3]와 Basic Iterative Method (BIM)[4], Projected Gradient Descent (PGD)[5], C&W[6]가 있다.

2.1 Fast Gradient Sign Method (FGSM)

Ian Goodfellow 등[3]은 적대적 예제를 생성하기 위해 이미지에 추가할 적대적 섭동을 찾는 방법을 2014년에 최초로 제안하였다. FGSM은 인공신경망의 학습 방법인 경사 하강법(Gradient Descent)을 역방향으로 수행하는 방법이다. 즉, 심층신경망의 손실함수의 경사를 부호(sign) 방향이 가장 가파른 방향으로 증가시킨다.

2.2 Basic Iterative Method (BIM)

Kurakin 등[4]은 FGSM[3] 방법을 확장하여, 경사를 반복적으로 갱신하는 방법을 제안하였다. 각 반복마다 손실함수의 값을 계산하여 경사를 이동시킨다. 또한, 각 반복에서 변경 가능한 적대적 예제의 픽셀 값을 제한한다.

2.3 Projected Gradient Descent (PGD)

Madry 등[5]은 FGSM 알고리즘을 사용하여 $\| \delta \|_\infty \leq \epsilon$을 만족하는 제약조건상에서 손실함수 경사의 방향으로 적대적 섭동인 δ을 반복적으로 갱신하는 방법을 제안하였다. 다시 말해 δ의 값 범위를 ϵ(적대적 섭동의 크기)값 안으로 제한시킨 채로, 손실함수의 경사를 반복적으로 갱신하여 적대적 예제를 생성한다.

2.4 C&W

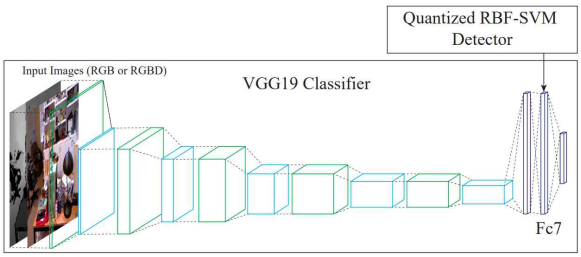
Carlini와 Wagner 등[6]은 적대적 섭동의 크기를 최소로 만들기 위해 세 개(l_∞, l_0, l_2)의 거리 메트릭(Distance metric)을 이용하는 적대적 공격을 제안하였다. 적대적 예제를 생성하기 위해, 적대적 섭동 계산을 반복하고 적대적 크기가 최소인 값 하나를 최종적으로 선택한다.

III. 적대적 탐지 방법

본 장에서는 적대적 공격에 대응하기 위해, 입력 이미지를 정상 이미지와 적대적 예제 중 하나로 분류하는 적대적 탐지(Adversarial Detection) 방법을 소개한다. 적대적 탐지 방법은 대표적으로 Safetynet[7], [8]이 있다.

3.1 Safetynet

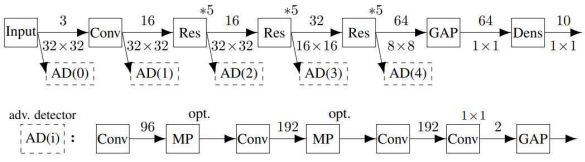
Lu 등[7]은 주어진 입력 이미지를 적대적 예제와 합법적(정상) 예제 중 하나로 이진 분류하기 위해, 기존 심층신경망에 적대적 예제 탐지 네트워크를 추가로 붙인 아키텍처를 제안하였다. 이 논문의 저자는 소프트맥스(Softmax) 함수의 직전에 위치한 활성화 함수가 적대적 예제와 합법적 예제에 대하여 서로 다른 결과(패턴)를 나타낼 것이라고 가정하였다. 여기서 VGG19 분류 네트워크를 통해 최종 활성화 함수의 출력을 양자화한다. 그 다음, RBF-SVM 기반 적대적 예제 탐지 네트워크는 양자화를 통해 요약된 표현의 특징을 입력받아서 적대적 예제와 합법적 예제의 분포 차이를 학습한다. 다음 그림 1은 Safetynet의 아키텍처이다.



[그림 1] Safetynet 아키텍처

3.2 On detecting adversarial perturbations

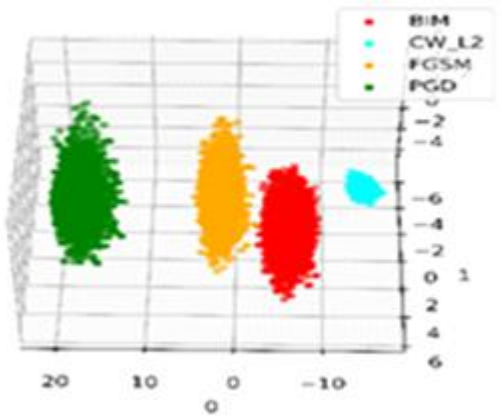
Metzen 등[8]은 주어진 입력 이미지가 적대적 예제일 확률을 계산하여, 입력 이미지를 적대적 예제와 합법적 예제 중 하나로 이진 분류하는 적대적 예제 탐지 네트워크를 제안하였다. 이 논문의 아키텍처도 Safetynet[7]와 마찬가지로, 기존 심층신경망에 적대적 예제 탐지 네트워크를 추가하였다. Safetynet과의 가장 큰 차이점은 심층신경망의 최종 활성화 함수의 특징이 아닌, 중간 활성화 함수의 특징을 사용한다. 적대적 예제 탐지 네트워크는 심층신경망 중간 계층 특징을 사용하여 입력 이미지가 적대적 예제일 확률을 학습한다. 다음 그림 2는 이 논문에서 제안한 모델의 아키텍처이다.



[그림 2] On detecting adversarial perturbations 논문에서 제안한 아키텍처:
(위) 심층신경망(ResNet);
(아래) 적대적 예제 탐지 네트워크(adv. detector)

IV. 결론

본 논문에서는 적대적 공격과 탐지 방법들을 조사하였다. 적대적 공격을 통해 생성된 적대적 예제는 심층신경망 기반 시스템을 기만하고 무력화 시킬 수 있다. 따라서 많은 연구자들의 관심이 필요하다. 본 논문을 확장하여, CIFAR-10 데이터셋으로 생성된 적대적 예제가 4가지 유형의 적대적 공격 방법(FGSM[3], BIM[4], PGD[5], C&W[6]) 중 어떠한 공격에 의해 생성되었는지 탐지할 수 있는 연구를 다음 그림 3 과 같이 수행하고 있다.



[그림 3] 적대적 공격 탐지 실험 결과

ACKNOWLEDGMENT

본 논문은 한국연구재단 논문연구과제 (NRF-2018R1D1A3B07043392) 및 4단계 BK21, 동남권4차산업혁명리더양성사업단에 의하여 지원되었습니다.

[참고문헌]

[1]ROSSLER, Andreas, et al. Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE International Conference on Computer Vision. 2019. p. 1-11.

[2]LU, Jiajun, et al. No need to worry about adversarial examples in object detection in autonomous vehicles. arXiv preprint arXiv:1707.03501, 2017.

[3]GOODFELLOW, Ian J.; SHLENS, Jonathon; SZEGEDY, Christian. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

[4]A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” arXiv preprint arXiv:1607.02533, 2016.

[5]MADRY, Aleksander, et al. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.

[6]CARLINI, Nicholas; WAGNER, David. Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). IEEE, 2017. p. 39-57.

[7]LU, Jiajun; ISSARANON, Theerasit; FORSYTH, David. Safetynet: Detecting and rejecting adversarial examples robustly. In: Proceedings of the IEEE International Conference on Computer Vision. 2017. p. 446-454.

[8]METZEN, Jan Hendrik, et al. On detecting adversarial perturbations. arXiv preprint arXiv:1702.04267, 2017.