적대적 사례 생성 기법 동향

오유진*, 김현지**, 임세진*, 서화정***
*한성대학교 IT융합공학부 (학부생)
**한성대학교 IT융합공학부 (대학원생)
***한성대학교 IT융합공학부 (교수)
oyj0922@gmail.com, khj1594012@gmail.com, dlatpwls834@gmail.com,
hwajeong84@gmail.com

A Study on generating adversarial examples

Yu-Jin Oh*, Hyun-Ji Kim**, Se-Jin Lim*, Hwa-Jeong Seo***
*Division of IT convergence Engineering, Hansung University
(Student)
**Division of IT convergence Engineering, Hansung University
(Graduate student)
**Division of IT convergence Engineering, Hansung University
(Professor)

요 약

인공지능이 발전함에 따라 그에 따른 보안의 중요성이 커지고 있다. 딥러닝을 공격하는 방법 중 하나인 적대적 공격은 적대적 사례를 활용한 공격이다. 이 적대적 사례를 생성하는 대표적인 4가지 기법들에는 기울기 손실함수을 활용하는 FGSM, 네트워크에 쿼리를 반복하여 공격하는 Deepfool, 입력과 결과에 대한 맵을 생성하는 JSMA, 잡음과 원본 데이터의 상관관계에 기반한 공격인 CW 기법이 있다. 이외에도 적대적 사례를 생성하는 다양한 연구들이 진행되고 있다. 그 중에서도 본 논문에서는 FGSM기반의 ABI-FGM, JSMA 기반의 TJSMA, 그 외에 과적합을 줄이는 CIM, DE 알고리즘에 기반한 One pixel 등 최신 적대적 사례 생성 연구에 대해 살펴본다.

1. 서론

현재 인공지능 시대라고 해도 과언이 아닐 만큼 얼굴 인식, 텍스트 분류 등 딥러닝을 이용한 서비스가 증가하고 점점 발전해 나가고 있다. 이렇게 딥러닝을 활용한 서비스가 늘어가면서 이를 공격하여 악용하려는 사례가 많아져 딥러닝에서의 높은 보안성이 요구되고 있다. 딥러닝을 공격하는 대표적인 공격으로 적대적 공격이 있고 이 적대적 공격은 적대적 사례를 활용한 공격이다. 적대적 사례를 생성하는 다양한 기법들이 연구되고 있고 본 논문에서는 적대적 사례 생성 기법에 관한 최신 연구에 대해 서술하고자 한다.

2. 관련 연구

2.1 인공신경망

인공신경망은 사람의 뇌 신경망(뉴런) 구조를 본떠 만든 구조이고 입력층, 은닉층, 출력층을 갖는다. 각 층은 노드들로 구성되어 있는데 입력노드들에 데이터를 입력받고 은닉노드에서의 계산을 통해 출력노드를 통해 결과를 도출해낸다. 이 중 2개 이상의 은닉층을 가지고 심층학습을 하는 것을 딥러닝이라고 한다.

인공신경망은 심층 신경망(Deep Neural Network), 합성곱 신경망(Convolutional Neural Network), 순환 신경망(Recurrent Neural Network) 등이 있다.

2.2 적대적 공격

적대적 공격은 적대적 사례를 이용한 공격으로 육안으로는 구분할 수 없는 작은 크기의 잡음을 데이터에 추가해 딥러닝 모델이 데이터를 오분류하게 만드는 공격이다. 적대적 공격은 공격 목표에 따라 표적공격 과 무표적 공격으로 나뉘고 공격자가 보유하고 있는 모델 정보의 양에 따라 화이트박스 공격

과 블랙박스 공격으로 구분할 수 있다.

2.2.1 표적공격

표적 공격은 공격자가 원하는 특정 클래스로 오인식하게 하여 원하는 결과를 도출해내는 공격이다. 클래스를 특정함으로써 무표적 공격보다 어렵고 클래스가 많을 시에는 공격이 더 어려워진다. 이 표적 공격이 안면인식 모델에 대해 성공한다면 특정 사람의 얼굴로 인식하여 안면 인식을 통한 본인 인증 시 권한이 도용될 수 있다.

2.2.2 무표적 공격

무표적 공격은 특정 표적 클래스 없이 임의의 클래스로 오인식하게 하는 공격이다. 무표적 공격은 목표 클래스가 없어 표적 공격 보다는 좀 더 수월하게 공격될 수 있다.

2.2.3 화이트박스 공격

화이트박스 공격은 공격자가 사전에 모델의 구조, 학습데이터, 알고리즘 등 모델에 대한 정보를 보유하고 있을 때 행하는 공격이다. 화이트박스는 모델에 대한 대부분의 정보를 갖고 있기 때문에 공격 성공률이 100%에 가깝지만 모든 정보를 알고 있다는 것 자체에서 비현실적이다. 그로 인해 실제로는 화이트박스 공격보다는 블랙박스 공격이 많이 시도되고 있다.

2.2.4 블랙박스 공격

블랙박스 공격은 공격자가 모델에 정보를 보유하고 있지 않을 때 행하는 공격이다. 공격자는 단지 입력값에 대한 결과값만을 알 수 있다. 블랙박스 공격은 화이트박스의 공격에 비해 낮은 성공률을 갖고 있지만 현실적이므로 많은 연구들이 진행되고 있다. 블랙박스 공격은 transfer attack 과 query-based attack으로 나뉜다[1].

Transfer attack은 적대적 사례를 생성하기 위해 화이트박스 모델에서의 적대적 사례를 블랙박스 모델에 전송하여 공격한다. 이 공격은 한 모델에 대한 적대적 사례가 다른 모델을 공격 시에도 효과적으로 작용한다는 전이 가능성에 기반을 둔다. 초기에는 대리 모델과 공격하는 대상모델의 구조가 다를 수 있기 때문에 낮은 성공률을 보였다가 최근 연구들에서 화이트 박스 모델의 기울기를 추정하거나 decision boundary에 접근하는 방식으로 블랙박스 공격을 시도하여 성공률을 높였다[2].

Query-based attack은 많은 쿼리를 통해 해당 모델의 기울기를 추정하고 이 기울기로 적대적 사례를 생성한다. Query-based attack은 transfer-attack에 비해 성공률은 높지만 이 성공률을 높이기 위해서는 매우 많은 쿼리가 필요하다[2].

2.3 적대적 사례[3]

적대적 사례는 데이터에 잡음을 추가하여 모델이 오인식하게 만드는 데이터 샘플이다. 이 잡음은 매우 작은 크기의 잡음으로 육안으로는 식별할 수 없지만 딥러닝 모델에 의해서는 식별되어 다른 클래스로 오인식하게 되는 것이다. 다음은 적대적 사례를 생성하는 대표적인 기법이다.

2.3.1 FGSM(Fast Gradient Sign Method)[4]

적대적 사례를 생성하는 기본적인 알고리즘 기법으로 목표 딥러닝 모델의 기울기 손실 함수에 대한 기울기를 계산하고 잡음을 추가하여 학습을 방해하는 알고리즘이다. 공격 대상 모델 학습 시에 사용되는 경사하강법 (gradient descent)과 반대되게 gradient absent 방식을 이용하여 기울기의 부호대로 노이즈를 추가하는 방식으로 모델이 클래스를 오인식 하게 한다.

2.3.2 Deepfool[5]

Deepfool 알고리즘은 분류 레이블을 변경하는 적대적 사례를 최소한으로 생성하기 위해 딥러닝 네트워크에 쿼리를 반복하는 무표적 공격이다. 딥러닝 모델 네트워크는 비선형 구조임에 비해 Deepfool은 선형구조라고 가정한 공격이기 때문에 여러 번의 쿼리를 반복해준다. 입력 벡터를 decision boundary에 투영한 결과로 최소한의 섭동을 만들 수 있고 이를 통해 섭동 벡터의 최소 크기를 다소 정확하게 추정할 수 있다.

2.3.3 JSMA (Jacobian-based Saliency Map Attack) [6]

JSMA 기법은 특정 입력에 따른 특정 결과를 도출해내어 모델이 오분류하게 하는 기법이다. 야코비 행렬을 계산하고 최소한의 잡음과 기울기를 계산하여 입력과 결과에 대한 맵을 생성한다. 이 기법은 특정 클래스로 오분류 할 수도 있다는 장점이 있으나 높은 계산복잡도가 요구 된다는 단점이 있으나 최소한의 잡음으로 모델이 오분류 하게끔 만드는 장

점이 있다.

2.3.4 CW(Carlini Wagner)[7]

잡음이 증가하면 공격 성공률이 높아지는 대신 원본 데이터와의 차이도 증가할 것이고, 그와 반대로 최소한의 잡음으로 원본 데이터와의 차이를 최소화 하면 공격 성공률이 낮아진다. CW는 이 차이를 최소화 하면서 성공률을 높이는 최적의 적대적 사례를 찾는 기법이다. 이 기법은 화이트 공격 시에 사용되며 성공률이 100%에 달한다. 이는 공격 모델에 대해 알기 때문에 이러한 성공률이 나타나는 것이다.

2.4 Adabelief[8]

Adabelief는 적응형 학습 속도 최적화 알고리즘이다. 이것은 현재 기울기 방향에 대한 belief, 즉 믿음에 따라 스텝 크기를 조정한다. 예를 들어, 관측된 기울기가 예측한 값과 가까운 경우, 그것을 기반으로 기울기가 더 작은 기울기에서 손실 함수의 감소를 가속화하고 손실 함수가 더 잘 수렴하도록 큰 조치를 취한다.

3. 적대적 사례 생성 기법

본 장에서는 최근 연구된 적대적 사례 생성 기법들에 대해 살펴본다. FGSM 기반의 ABI-FGM, JSMA 기반의 TJSMA과 이 외에 One pixel 기법 및 CIM 기법이 있다.

3.1 ABI-FGM (Adabelief Iterative Fast Gradient Method)[9]

ABI-FGM은 대표적인 공격기법인 FGSM을 단일 단계에서 다중 단계로 반복한 I-FGSM에 adabelief 최적화를 도입하여 개선한 것이다.

이 기법에서는 반복과정에서 손실 함수의 기울기 방향을 따라 속도 벡터를 축적하고 예측 기울기와 weight 기울기 사이 차의 제곱 값도 축적한다. 그 후 두 벡터로 파라미터 방향을 얻은 후 실제 기울기와 예측 기울기 사이의 편차에 따라 단계를 조정하기 위해 파라미터를 조정하여 수렴(convergence) 속도와 효과를 보장한다.

이러한 개선은 수렴 프로세스를 최적화하여 전이 가능성이 더 많은 적대적 사례를 생성할 수 있다.

이 기법은 단일로 사용할 시 화이트 공격 시에 공격 성공률 100%지만 블랙박스 공격 시에는 약 50%의 공격 성공률 결과가 도출되었다. 다만 이 기법을 다른 알고리즘과 결합하여 사용하면 블랙박스 공격에서도 더 높은 성공률이 나타날 수 있다.

3.2 CIM(Crop-Invariant Attack Method)[9]

CIM은 이미지의 정확한 분류에 영향을 거의 미치지 않는 에지를 범위 내에서 잘라 과적합을 줄이는 것을 기반으로 하는 기법이다.

대부분의 이미지들은 중앙에 가장 중요한 부분이 있고 에지에 가까울수록 덜 중요하다. 이미지의 에지를 잘라내면 덜 중요한 부분을 제거했으므로 이미지 손실 보존 변환을 실현할 수 있다. 이를 바탕으로 이미지의 크롭 복사본에 대한 적대적 사례를 최적화하여 전이 가능한 적대적 사례를 생성한다.

3.4.TJSMA(Taylor JSMA)[10]

TJSMA는 JSMA 기반으로 JSMA보다 더 균형 잡힌 맵을 얻을 수 있는 알고리즘으로 맵에 간단한 가중치를 주고 입력에 추가로 패널티를 부과한다. 높은 계산 복잡도로 인해 생성 속도가 높은 JSMA의 단점을 보완하였다.

표적 공격에서의 실험 결과, MNIST 데이터에서 JSMA에 비해 성공률은 훈련단계에서 10.98% 포인트, 테스트단계에서 11% 포인트 정도 증가했고, CIFAR-10 데이터에서도 훈련단계에서 11.23% 포인트, 테스트 단계에서 12% 포인트가 증가했다.

또한 MNIST 이미지를 성공적으로 만드는데 필요한 시간을 측정한 결과 JSMA보다 약 1.41배 빠른 것으로 확인되었다.

다만 JSMA를 개선했음에도 불구하고 입력이 커질 때 맵의 높은 계산 비용 때문에 대규모 데이터 셋이 확장 불가능하고 작은 데이터 세트를 위한 것을 고려해야 한다.

3.3 One pixel[11]

One pixel 기법은 픽셀 하나만을 수정해서 적대적 사례를 생성해 모델이 오인식할 수 있게 하는 기법이다. 이것은 최적화 문제를 해결하는 EA 알고리즘인 DE 알고리즘[12]을 기반으로 한 픽셀에 적대적 교란을 생성한다.

One pixel 기법을 CIFAR-10 데이터 셋을 활용하여 합성곱 신경망에서 공격을 해본 결과 표적공격에서는 19.82%, 비표적 공격에서는 68.71% 의 성공률을 보였고 신뢰도는 79.40%를 보였다.

다른 적대적 사례 생성 알고리즘을 사용한 공격에 비해 one pixel 생성 기법을 사용한 공격은 성공률은 낮지만 적대적 정보가 덜 필요하여 블랙박스 공격에서 사용할 수 있다.

4. 결론

본 논문에서는 최근 연구된 적대적 사례 생성 기법인 ABI-FGM, CIM, TSMA, one pixel 에 대해 알아보았다. 이 뿐만 아니라 현재 다양한 생성 기법들이 연구되고 있다. 적대적 사례 생성 기법 연구가 증가하는만큼 우리는 이 적대적 사례를 이용한 적대적 공격을 막기 위한 방어책들을 마련하여 딥러닝 모델의 안전성에 대한 위협을 감소시켜 나가야 할 것이다.

5. Acknowledgment

이 논문은 부분적으로 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(No.2018-0-00264, IoT 융합형 블록체인 플랫폼 보안 원천 기술 연구, 50%) 그리고 부분적으로 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2020R1F1A1048478, 25%) 그리고 부분적으로 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2021-0-00540, GPU/ASIC 기반 암호알고리즘 고속화 설계 및 구현 기술개발, 25%).

참고문헌

[1] Suyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu, “Improbing blck-box adversarial attacks with a transfer-based prior.”, In Advances in Neural Information Processing Systems (NeurIPS), 2019.

[2] Jie Li, Rongrong Ji, Hong Liu, Jianzhuang Liu, Bineng Zhong, Cheng Deng, Qi Tian, “Projection & Probability-Driven Black-Box Attack.” In Proceedings of the IEEE Conference on Co puter Vision and Pattern Recognition(CVPR),2020, pp.362-371

[3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” Int’l Conf. Learning

[4] Ian J. Goofellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples.” arXiv prepring arXiv:1412.6572, 2014

[5] S.M Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “Deepfool: a Simple and Accurate Method to Fool Deep Neural Networks.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp. 2574-2582, 2016.

[6] Papernot, Nicolas, et al. “The limitations of deep learning in adversarial settings.” 2016 IEEE European symposium on security and privacy, 2016

[7] N. Carlini, D. Wagner, “Towards Evaluating the Robustness of Neural Networks.” 2017 IEEE Symposium on Security and Privacy, IEEE, pp.39-57,2017

[8] Juntang Zhuang, Tommy Tang, Yifan Ding, Sekhar Tatikonda, Nicha Dvornek, Xenophon Papademetris, and James Duncan.“Adabelief optimizer: Adapting stepsizes by the belief in observed gradients.” NeurIPS, 2020.

[9] B. Yang, H. Zhang, Y. Zhang, K. Xu, and J. Wang. “Adversarial example generation with adabelief optimizer and crop invariance.” arXiv preprint arXiv:2102.03726, 2021.

[10] Combey, T., Loison, A., Faucher, M., & Hajri, H. (2020). Probabilistic Jacobian-based saliency maps attacks. Machine Learning and Knowledge Extraction, 2(4), 558-578.

[11] Su, Jiawei, Danilo Vasconcellos Vargas, and Kouichi Sakurai. "One pixel attack for fooling deep neural networks." IEEE Transactions on Evolutionary Computation 23.5 (2019): 828-841.

[12] Das, S., & Suganthan, P. N. (2010). Differential evolution: A survey of the state-of-the-art. IEEE transactions on evolutionary computation, 15(1), 4-31.