단일 픽셀 공격을 완화하기 위한 이미지 처리 기법

이연지 [1], 이일구 [2]
[1] 성신여자대학교 융합보안공학과 박사과정
[2] 성신여자대학교 융합보안공학과, 미래융합기술공학과공학과 교수

cselab.lyj@gmail.com, iglee19@gmail.com

Image Processing Technique to Mitigate One-Pixel Attack

Yeon-Ji Lee[1], Il-Gu Lee[1,2]
[1]Dept. of Convergence Security Engineering, Sungshin Women's University
[2]Dept. of Future Convergence Technology Engineering, Sungshin Women's University

요 약

최근 이미지 분류, 자율 주행 등 다양한 분야에 인공지능 기술이 접목됨에 따라 인공지능 기술을 이용한 새로운 위협이 등장하고 있다. 적대적 공격 중 단일 픽셀 공격은 이미지의 픽셀 하나를 왜곡하여 인공지능의 올바른 분류를 방해하는 공격 기법이다. 본 논문은 단일 픽셀 공격을 완화하는 이미지 처리 기법을 제안한다. 실험 결과에 따르면 제안한 방법을 적용하면 이미지의 사이즈를 27x27 로 조절하였을 때 100 개의 단일 픽셀 공격 이미지 중 94 개를 복구하였으며, 이미지의 신뢰도를 68.89% 개선하였다.

1. 서론

최근 이미지 분류, 객체 감지, 음성 제어, 자율 주행 등 다양한 분야에서 인공지능 기술이 빠르게 발전하고 있다[1,2]. 이와 같은 발전은 우리의 삶과 사회에 전반적인 영향을 미치고 있으며, 향후에는 더욱 많은 분야에 접목되어 더 큰 혁신과 발전을 이끌어낼 것이라 예측되고 있다. 그러나 이러한 인공지능 기반의 이미지 처리 기술이 발전함과 동시에 인공지능 기술을 악용하려는 시도 또한 증가하고 있으며, 적대적 샘플(adversarial examples)을 활용한 적대적 공격(adversarial attacks)과 같은 새로운 위협을 야기하고 있다.

적대적 공격은 이미지를 왜곡하여 적대적 샘플을 생성하는 공격을 의미한다. 이와 같은 적대적 공격은 자율 주행, 음성 인식 등의 분야에서 심각한 영향을 미치고 있다[2]. 그러므로 최근 이러한 지능적인 공격에 대응하기 위한 기술이 중요해지고 있다.

본 연구는 적대적 공격 중에서도 단일 픽셀에만 왜곡하는 적대적 단일 픽셀 공격(one-pixel attack)을 구현한 뒤, 단일 픽셀 공격을 완화하는 이미지 처리 기술을 제안한다. 본 연구의 주요 기여점은 다음과 같다.

- 단일 픽셀 공격을 완화하는 이미지 처리 기법을 제안한다.
- 단일 픽셀 공격과 공격에 대응하는 기술들의 성능을 평가하는 평가 프레임워크를 제안한다.

본 논문의 구성은 다음과 같다. 2 장에서 단일 픽셀 공격에 대응하기 위한 종래의 연구들을 분석하고 한계점을 도출한다. 3 장에서 제안하는 방법에 대해 설명하고 4 장에서 실험 및 성능 평가 결과에 대해 서술한 뒤 5 장에서 결론을 맺는다.

2. 관련 연구

최근 적대적 공격을 탐지하고 방어하기 위한 다양한 기법들이 연구되고 있다.

Muhammad A. H[3]는 단일 픽셀 공격 탐지 및 완화를 위하여 단일 픽셀 공격 이미지에 Accelerated Proximal Gradient 접근 방식을 적용한 강력한 주성분 분석(Robust Principal Component Analysis, RPCA) 기반의 이미지 복원 모델을 제안하였다. 제안 모델은 정규화 매개변수인 lambda 값의 하이퍼 파라미터 튜닝을 통해 적절한 값을 찾고, RPCA 모델이 이미지를 올바르게 복원할 수 있음을 입증하였다. 하지만 제안하는 모델은 복구를 위한 모델을 가지고 있어야 하며, 이미지마다 적용되는 파라미터 값이 다르다는 한계가 존재한다.

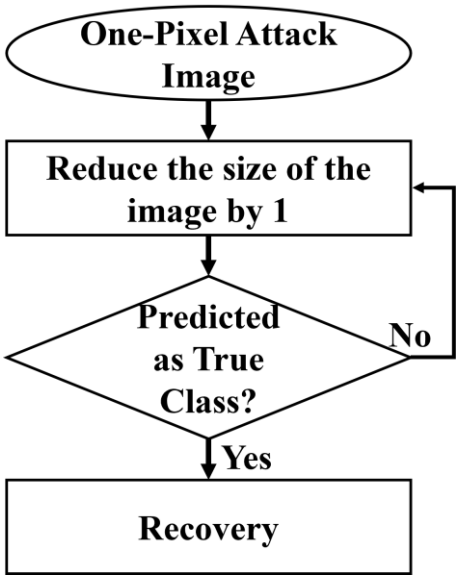
Rahul Paul[4]은 적대적 공격 기법들 중 FGSM(Fast Gradient Sign Method)과 단일 픽셀 공격을 완화하기 위하여 다중 초기화 기반 앙상블(multi-initialization

based ensemble) 모델을 제안하였다. 우선 생성한 단일 픽셀 공격 이미지를 테스트 데이터셋에 포함하여 학습을 진행한 결과에 따르면 정상 데이터셋 대비 정확도가 약 3% 감소한 결과를 보였다. 이후 생성된 공격 이미지를 완화하기 위하여 가중치 초기화 레이어를 7 개 포함시킨 CNN 모델 3 개를 결합한 앙상블 모델을 제안하였고, 결과적으로 단일 픽셀 공격 대비 8% 개선된 성능을 입증하였다. 하지만 제안하는 방법은 세 가지의 CNN 모델을 결합하여 복잡성이 크고, 복잡성에 비해 좋은 성능 개선을 이루지 못하고 있다는 점에서 한계가 있다.

종래 연구들은 단일 픽셀 공격 완화를 위해 복잡한 알고리즘을 제안하고 있다. 하지만 단일 픽셀 공격의 주요 공격 대상인 보안 카메라, 자율주행과 같은 사물인터넷 기기들은 복잡한 알고리즘이 적용되기에 적합하지 않으므로, 보다 단순한 완화 기법에 대한 연구가 필요하다.

3. 단일 픽셀 공격 완화를 위한 이미지 처리 기법

본 장에서는 제안하는 단일 픽셀 공격 완화 기법에 대해 설명하고 성능 평가 결과를 분석한다.



(그림 1) 단일 픽셀 공격 완화를 위한 이미지 처리 기법

그림 1 은 제안한 기법의 동작을 표현한 흐름도이다. 제안하는 모델은 단일 픽셀 공격을 받은 이미지의 해상도를 1 씩 낮춰가며 실제 클래스로 분류될 때까지 반복된 이미지 처리를 수행한다. 단일 픽셀 공격의 특성상 공격당한 픽셀이 주변의 픽셀들과 특성의 차이가 크다. 따라서 제안한 기법은 단일 픽셀 공격에 활용된 픽셀이 주변과 섞이도록 이미지의 해상도를 낮춤으로써 단일 픽셀 공격을 완화한다.

4. 성능 평가 및 분석

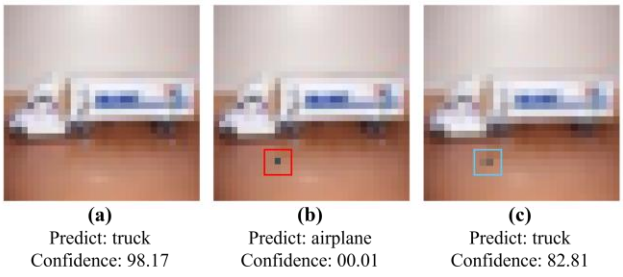
본 장에서는 제안한 모델을 실제로 구현하기 위한 환경과 성능 평가 결과를 분석한다.

본 연구에서 사용한 CIFAR-10[5] 이미지 데이터셋은 10 개의 클래스가 각각 6,000 개의 이미지 데이터를 가지고 있어서 총 60,000 개의 이미지 데이터로 구성되었다. 각각의 클래스는 완전히 상호 배타적이므로 겹치는 이미지가 없고, 완전히 구분 가능하여 단일 픽셀 공격과 완화 기법을 테스트하기에 적합하다. 성능 평가를 위해 50,000 개의 데이터를 학습 데이터로 사용하였고, 9,900 개의 데이터는 테스트 데이터로 사용하였다. 이후 단일 픽셀 공격을 적용한 100 개의 평가 데이터를 생성하여 단일 픽셀 공격 완화 기법을 검증하였다. 성능 평가를 위해 학습과 분류에 사용된 모델은 전이학습 모델인 resnet(residual neural network) 모델로, 자세한 학습 모델 정보는 표 1 와 같다.

<표 1> 학습에 사용된 모델의 파라미터 값과 함수

Input Size	32, 32
Epochs	200
Batch Size	128
Channels	3
Training Activation	relu
Classification Activation	softmax
Loss Function	categorical_crossentropy

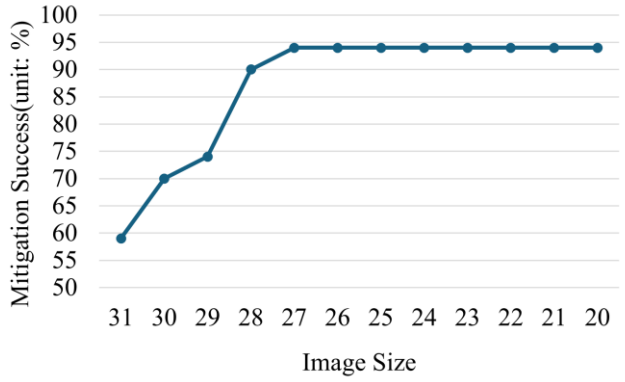
이러한 성능 평가 환경에서 학습 모델이 단일 픽셀 공격 이미지를 정상 클래스로 분류할 때까지 해상도를 1 씩 낮추며 신뢰도(confidence)를 평가하였다. 여기서 신뢰도는 모델이 정상 클래스로 판단한 예측 값(probability)을 의미하므로 공격 성공과 공격 완화를 입증하는 평가지표로 사용된다.



(그림 2) 단일 픽셀 공격 및 완화 기법 적용 예시

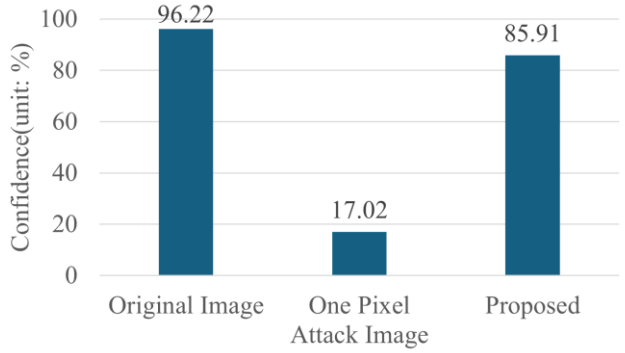
그림 2 는 제안하는 방법을 적용한 이미지의 예시이다. (a)는 공격을 적용하지 않은 정상 이미지이며 클래스를 올바르게 분류하고 있다. (b)는 단일 픽셀 공격

을 진행한 공격 이미지로 붉은 색으로 표시된 부분이 공격을 진행한 픽셀이다. 결과적으로 (b)를 다른 클래스로 오분류했고, 이에 따라 신뢰도가 낮게 측정되었다. (c)는 단일 픽셀 공격을 진행한 (b) 이미지에 이미지 처리 기법을 적용한 이미지로, 원본 사이즈인 32x32 를 31x31 로 1 만큼 축소한 결과이다. 이때 파란색으로 표시한 부분을 보면 공격 노이즈가 완화되었음을 알 수 있다. (c) 이미지는 다시 정상 클래스로 분류되었고, 그에 따라 신뢰도가 향상되었다. 이렇게 단일 픽셀 공격이 완화된 (c) 이미지의 신뢰도는 (a) 보다는 15.36% 낮게 측정되었지만, 단일 픽셀 공격 이미지에서 82.80% 개선되었다.



(그림 3) 제안한 방식의 이미지 사이즈 조정에 따른 공격 완화 비율

그림 3 은 이미지 사이즈 조정에 따른 단일 픽셀공격 완화 비율 그래프이다. 이미지의 사이즈를 31x31 로 조절하였을 때 100 개의 단일 픽셀 공격 이미지 중 59 개가 완화되었고, 30x30 으로 조절하였을 때 70 개, 29x29 로 조절하였을 때 74 개, 28x28 로 조절하였을 때 90 개가 완화되었다. 공격 완화 성공률이 수렴하는 지점은 27x27 로서, 이때 100 개의 이미지 중 94 개의 이미지가 완화되었다. 100 개 중 6 개의 이미지는 이미지의 사이즈를 1x1 까지 조절하여도 오분류하여 결국 완화하지 못했다.



(그림 4) 단일 픽셀 공격 완화에 따른 신뢰도 평가

그림 4 는 원본 이미지와 단일 픽셀 공격 이미지, 그리고 이미지 처리 기법을 적용한 이미지의 정상 클래스에 대한 신뢰도를 측정한 값의 평균을 낸 결과이다. 완화되지 않은 6 개의 이미지를 제외한 결과이며, 이를 통해 단일 픽셀 공격이 성공적으로 적용되었음을 확인할 수 있었고, 이미지 처리 기법을 통한 완화 기법이 성공적으로 적용되어 공격 이미지의 신뢰도를 68.89% 개선하였음을 알 수 있다.

5. 결론

본 논문은 단일 픽셀 공격의 효율적인 완화를 위해 이미지 처리 기법을 적용하는 방법을 제안했다. CIFAR-10 데이터셋을 resnet 으로 학습하여 기본 모델을 생성하였으며, 100 개의 단일 픽셀 공격 이미지 데이터를 생성하여 완화 기법을 평가하였다. 실험 결과에 따르면 제안하는 모델은 단순한 동작으로 단일 픽셀 공격 이미지를 완화할 수 있었고, 이미지의 사이즈를 32x32 에서 27x27 로 줄였을 때 100 개 중 96 개의 이미지를 완화하였다. 또한, 단일 픽셀 공격을 적용한 이미지에서 68.89%의 신뢰도를 개선하여 제안 모델의 성능을 입증하였다.

후속 연구에서는 단일 픽셀 공격이 아닌 다중 픽셀 공격을 구현하여 이를 완화하기 위한 방법에 대해 연구할 계획이다. 또한 32x32 사이즈의 이미지가 아닌 고해상도 이미지에 단일, 다중 픽셀 공격을 적용하고 완화하기 위한 기법을 연구할 계획이다.

Acknowledgements

본 논문은 2024 년도 산업통상자원부 및 한국산업기술진흥원의 산업혁신인재성장지원사업 (RS-2024-00415520)과 과학기술정보통신부 및 정보통신기획평가원의 ICT 혁신인재 4.0 사업의 연구결과로 수행되었음 (No. IITP-2022-RS-2022-00156310)

참고문헌

[1] Shilin Qiu, Qihe Liu, Shijie Zhou and Chunjiang Wu, “Review of Artificial Intelligence Adversarial Attack and Defense Technologies,” applied sciences, volume 9, issue 5, 2019.

[2] Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang and Anil K. Jain, “Adversarial Attacks and Defenses in Images, Graphs and Text: A Review,” International Journal of Automation and Computing, volume 17, pp. 151-178, 2020.

[3] Muhammad Akbar Husnoo, Adnan Anwar, "Do not get fooled: Defense against the one-pixel attack to protect IoT-enabled Deep Learning systems," Ad Hoc Networks, volume 122, 2021.

[4] Rahul Paul , Matthew Schabath, Robert Gillies, Lawrence Hall, and Dmitry Goldgof, “Mitigating Adversarial

Attacks on Medical Image Understanding Systems," 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), USA, 2020.

[5] Tensorflow, "cifar10," last updated: 06 December, last accessed 22 April 2024, available: https://www.tensorflow.org/datasets/catalog/cifar10.