계층별 모델 역추론 공격

권현호 [1], 김한준 [2]
[1] 연세대학교 전기전자공학과 통합과정
[2] 연세대학교 전기전자공학과 교수

hyunho@yonsei.ac.kr, hanjun@yonsei.ac.kr

Layer-wise Model Inversion Attack

Hyun-Ho Kwon[1], Han-Jun Kim[1]
[1]Dept. of Electrical and Electronic Engineering, Yonsei University

요 약

모델 역추론 공격은 공격 대상 네트워크를 훈련하기 위해 사용되는 훈련 데이터셋 중 개인 데이터셋을 공개 데이터셋을 사용하여 개인 훈련 데이터셋을 복원하는 것이다. 모델 역추론 방법 중 적대적 생성 신경망을 사용하여 모델 역추론 공격을 하는 과거의 논문들은 딥러닝 모델 전체의 역추론에만 초점을 맞추기 때문에, 이를 통해 얻은 원본 이미지의 개인 데이터 정보는 제한적이다. 따라서, 본 연구는 대상 모델의 중간 출력을 사용하여 개인 데이터에 대한 더 품질 높은 정보를 얻는 데 초점을 맞춘다.

본 논문에서는 적대적 생성 신경망 모델이 원본 이미지를 생성하기 위해 사용되는 계층별 역추론 공격 방법을 소개한다. MNIST 데이터셋으로 훈련된 적대적 생성 신경망 모델을 사용하여, 원본 이미지가 대상 모델의 계층을 통과하면서 얻은 중간 계층의 출력 데이터를 기반으로 원본 이미지를 재구성하고자 한다. GMI 의 공격 방식을 참고하여 공격 모델의 손실 함수를 구성한다. 손실 함수는 사전 손실 및 정체성 손실항을 포함하며, 역전파를 통해서 원본 이미지와 가장 유사하게 복원할 수 있는 표현 벡터 Z 를 찾는다. 원본 이미지와 공격 이미지 사이의 유사성을 분류 라벨의 정확도, SSIM, PSNR 값이라는 세 가지 지표를 사용하여 평가한다. 공격이 이루어지는 계층에서 복원한 이미지와 원본 이미지를 세 가지 지표를 가지고 평가한다. 실험 결과, 공격 이미지가 원본 이미지의 대상 분류 라벨을 정확하게 가지며 원본 이미지의 필체를 유사하게 복원하였음을 보여준다. 평가 지표 또한 원본 이미지와 유사하다는 것을 나타낸다.

1. 서론

머신 러닝의 발전과 광범위한 사용으로, 개인의 프라이버시를 해치는 적대적 공격이 새로운 주요 도전 과제로 떠오르고 있다. 이러한 공격 중에서도 딥러닝 모델에서의 모델 역추론 공격은 얼굴 인식을 통한 잠금 해제와 같이 보안이 중요한 작업과 직접적으로 관련이 있어 특히 주목받고 있다. 역추론 공격에서는 악의적 사용자가 학습 모델을 훈련하는 데 사용된 개인 데이터셋을 복구하는 것을 목표로 한다. 모델에 대한 성공적인 역추론은 훈련된 데이터의 현실적이고 다양한 샘플을 생성하여 개인의 프라이버시를 심각하게 해친다.

딥러닝 모델을 역추론 공격하는 다양한 방법이 제안되었으며, 현재 최신 역추론 공격 방법으로는 DeepInversion [1], GMI [2], AMI [3] 등이 있다. 이 논문들은 출력 분류 라벨에 의해 입력 이미지를 생성할 수 있는 모델 역추론을 만들려 시도한다. 특히, GMI 모델은 모델의 파라미터가 접근 가능할 때 적대적 생성 신경망을 사용하여 공격한다.

GMI 모델은 모델 전체의 역추론에 초점을 맞추기 때문에, 중간 계층에서 원본 이미지를 재구성하는 정도를 판단하기 어렵다. 본 연구는 대상 모델의 계층을 통과하는 동안 얻은 중간 계층 출력 값을 사용하여 원본 이미지를 재구성하려 시도한다. 그러나 각 계층의 직접적인 역연산은 수학적으로 불가능하므로, 적대적 생성 신경망 모델을 활용하여 원본 이미지와 매우 유사한 공격 이미지를 생성하는 계층별 모델 역추론 공격을 소개한다. MNIST 숫자를 구분하는 합성곱 신경망 모델을 공격 대상 모델로 고려하며, MNIST 데이터셋으로 훈련된 적대적 생성 신경망 모델을 사

용하여, 원본 이미지가 대상 모델의 계층을 통과하면서 얻은 중간 출력 데이터를 기반으로 원본 이미지를 재구성한다. 복원된 공격 이미지는 원본 이미지의 분류 라벨 뿐만 아니라 필체도 복구한다.

2. 배경 지식

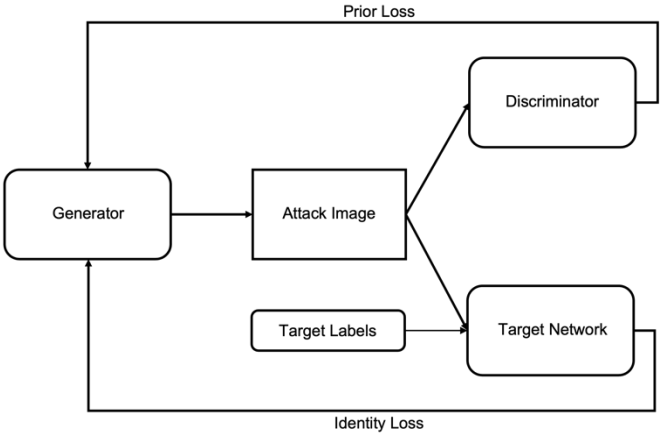
적대적 생성 신경망은 생성기(generator)와 판별기(discriminator)라는 두 개의 신경망으로 구성된 머신러닝 프레임워크의 한 유형이다. 생성기는 무작위 잠재 z 벡터를 입력으로 받아 실제 이미지와 유사한 가짜 이미지를 생성한다. 판별기는 실제 이미지와 가짜 이미지를 정확하게 구분하는 것을 목적으로 훈련된다. 이 두 신경망은 서로 적대적인 관계에 있으며, 생성기는 판별기가 실제라고 잘못 인식하는 가짜 이미지를 생산하려고 노력하고, 판별기는 생성기가 생성한 가짜 이미지를 가짜로 정확하게 분류하려 한다. 이러한 적대적 관계를 통해 생성기와 판별기는 각각의 성능을 향상시켜 생성기가 실제와 매우 유사한 가짜 이미지를 생성할 수 있게 된다.

2014 년 이안 굿펠로우(Ian Goodfellow)에 의해 소개된 적대적 생성 신경망 모델[4]은 훈련 중 불안정함을 겪었다. 이를 해결하기 위해 여러 가지 적대적 생성 신경망 모델이 개발되었다. 그 중 하나인 Wasserstein 적대적 생성 신경망(WGAN) [5]은 전통적인 적대적 생성 신경망 모델에서 기울기 소멸 문제를 야기할 수 있는 KL(Kullback-Leibler) 발산 대신 Wasserstein 거리를 사용한다. Wasserstein 거리는 두 확률 분포 사이의 거리를 측정하는 지표다. WGAN 은 생성기의 출력과 실제 데이터 분포 사이의 거리를 최소화하는 것을 목표로 한다. 이는 안정적인 훈련과 고품질의 이미지 생성을 촉진한다. WGAN 은 안정적인 훈련을 보장하고 고품질의 이미지를 생성하기 때문에, 본 논문은 계층별 모델 역추론에 WGAN 을 사용한다.

모델 역추론은 악의적인 사용자가 모델 출력과의 상관관계를 사용하여 모델을 훈련하는 데 사용된 민감한 개인 데이터셋을 복구하려고 시도하는 공격이다. 모델의 특성에 따라 다양한 접근 방식이 각 제약 상황에 따라 적용된다.

이 중에서 GMI 방법은 모델 내부에 접근할 수 있는 화이트 박스 상황에서의 접근 방식이다. 이 방법은 학습 데이터의 일부 공개 데이터셋을 활용하여 적대적 생성 신경망을 통해 분포적 사전 지식을 학습하고 이를 역전파 과정에서 사용한다. 공개 데이터셋이 공격하는 대상에 대한 정보를 포함하고 있지 않을지라도, 공개 데이터셋에 대한 사전 훈련을 통해 얻은 정보는 원래 잘못 설정된 역추론 문제를 정규화하는 데 도움을 줄 수 있다. 이 방법은 공개 데이터셋이 적대자가 복구하려는 정체성을 포함하지 않고, 크기가 작고, 개인 데이터 또는 다른 분포에서 온 경우에도 매우 높은 성능을 발휘한다.

먼저, 우리는 수학적인 방법을 통해서 모델 역추론을 시도한다. 입력에 대한 정보를 추출하기 위해 컨볼루션 연산을 행렬 곱셈 연산으로 바꾸는 방법을 사용한다. 입력 X 와 커널 K 사이의 컨볼루션 연산을 행렬과 벡터 곱셈으로 풀어서 작성하면 행이 열보다 많은 커널 행렬 U 와 입력 X 를 벡터로 풀어서 작성할 수 있다. 우리는 커널 K 와 출력 Y 에 대한 정보를 알고 있는 상황을 가정하기에, X 를 구하기 위해서는 U 행렬과 U 의 전치행렬의 곱의 역행렬을 구해야 한다. 하지만, 역행렬을 구할 수 없어 수학적인 방법으로 모델 역추론을 하는 것은 불가능하다는 결론에 도달하였다.



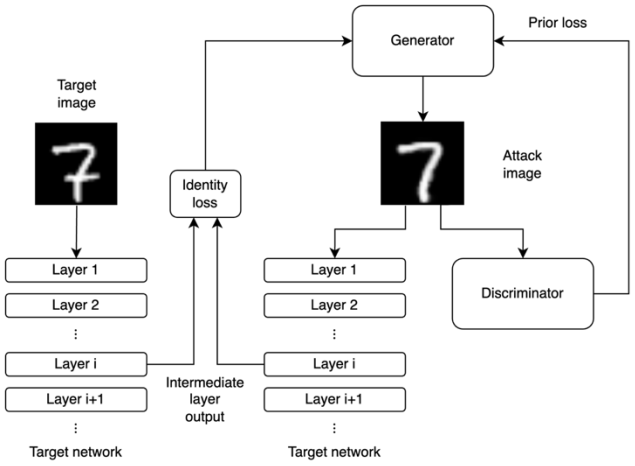
(그림 1) 적대적 생성 신경망을 사용한 모델 역추론 공격의 개략도.

3. 방법론

수학적인 방법으로 모델 역추론 공격을 하는 것이 불가능하기 때문에, 본 논문은 적대적 생성 신경망을 활용한 방식으로 계층별 모델 역추론 공격에 접근한다. 그림 1 은 적대적 생성 신경망을 사용한 모델 역추론 공격인 GMI 의 개략도를 보여준다.[2] 생성기에 의해 생성된 공격 이미지는 판별기와 공격 대상 네트워크 모두에서 손실을 계산하는 데 사용된다. 대상 네트워크에서는 공격 이미지가 입력으로 제공될 때 출력 분류 라벨이 얻어진다. 정체성 손실(identity loss)은 이 출력 분류 라벨과 대상 분류 라벨 간의 교차 엔트로피 손실로부터 얻어지는 반면 사전 손실(prior loss)은 판별기로부터 진짜 이미지와 유사도를 기반으로 얻어진다. 이 손실들은 생성기의 잠재 벡터를 업데이트하여 공격 이미지를 향상시키는 데 사용된다.

본 논문은 GMI 모델을 기반으로, 중간 계층에서 공격하는 계층별 모델 역추론 공격을 구현한다. 우리는 표현 벡터 Z 를 그라디언트 사이의 오차를 최소화하

는 방식으로 업데이트하는 이전의 공격 방식과 달리 본 모델에서는 그라디언트 대신 원본 이미지와 공격 이미지 사이의 중간 계층 출력에서 발생하는 손실을 줄이는 방향으로 생성기의 잠재 벡터가 업데이트된다. 자세한 개략도는 그림 2 에서 보여준다.



(그림 2) 계층별 모델 역추론 공격의 자세한 개략도.

공격 과정은 다음 다섯 단계로 구성된다. 첫째, 공격하려는 원본 대상의 중간 계층 출력에 대한 정보를 얻는다. 둘째, 훈련된 생성기가 임의의 잠재 벡터에 대한 공격 이미지를 생성하고, 이 이미지는 공격 대상 네트워크와 판별기에 입력된다. 셋째, 공격 대상 네트워크에서 공격 이미지에 대한 순전파가 진행된다. 순전파의 결과로 중간 계층 출력이 얻어진다. 넷째, 공격 이미지의 중간 계층 출력과 원본 이미지에 대한 중간 계층 출력에 대한 정체성 손실을 구하고, 판별기에서 생성된 공격 이미지에 대한 이전 손실을 구한다. 마지막으로, 생성기의 잠재 벡터가 이 손실들에 의해 업데이트된다. 이 과정을 반복하여 가장 원본 이미지와의 정체성 손실이 작은 잠재 벡터를 찾아 원본 이미지와 가장 유사하게 복원한 공격 이미지를 얻는다.

공격의 초기 단계에서 임의의 잠재 벡터를 10 개를 무작위로 생성하여 해당 이미지를 생성한다. 초기 잠재 벡터가 지나치게 지배적이어서 이미지가 지역 최소값에 갇힐 가능성 때문에, 다양한 이미지 변화를 생성하기 위해 여러 잠재 벡터를 사용하는 접근법이 도입된다. 그 후에 생성된 벡터들 중에서 가장 낮은 손실을 가진 이미지가 원본 이미지와 가장 유사한 이미지로 판별되어 해당 이미지를 최종적인 공격 이미지로 선정한다.

4. 실험

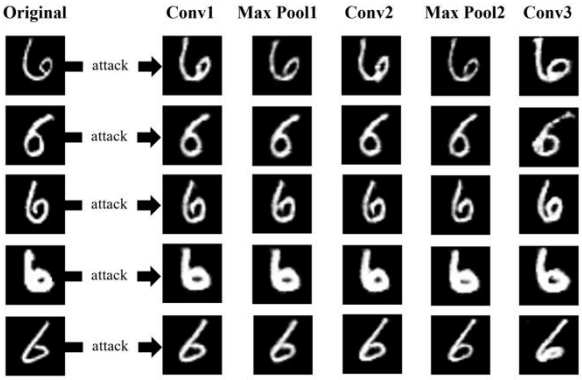
MNIST 테스트 데이터셋에서 무작위로 선택한 1000 개의 이미지에 대해 계층별 모델 역추론 공격을 수행하였다. 이는 모델의 어떤 계층에서 공격한 이미지가 원본 이미지와 가장 유사한 이미지를 생성하는지 확인하기 위함이다. 생성된 공격 이미지의 분류 라벨을 공격 대상의 모델보다 더 좋은 성능을 가진 모델을 통해서 확인하고 원본 이미지의 분류 라벨과 비교하여 분류 라벨 정확도를 측정한다. 또한, SSIM(Structural Similarity Index) 및 PSNR(Peak Signal-to-Noise Ratio)과 같은 평가 지표를 사용하여 원본 이미지와 공격 이미지 사이의 유사성을 측정한다.

<표 1> 계층별 모델 역추론 공격 실험 결과

계층	파라미터	라벨 유사도	SSIM	PSNR
Conv 1	64x28x28	0.963	0.909	21.685
Max Pool 1	64x14x14	0.959	0.908	21.826
Conv 2	128x10x10	0.941	0.887	20.840
Max Pool 2	128x5x5	0.91	0.849	19.919
Conv 3	256x1x1	0.896	0.648	14.106
Output	10x1x1	0.999	0.354	10.552

표 1 는 대상 모델의 Conv 1, Max Pool 1, Conv 2, Max Pool 2, Conv 3 및 출력 계층에서 수행된 모델 역추론 공격의 파라미터 수, 분류 라벨 정확도, 이미지 유사도를 나타내는 SSIM 및 PSNR 지표를 비교한다. 출력 계층에 대한 공격은 필체를 고려하지 않고 원본 이미지의 분류 라벨 정보만을 사용하여 무작위 숫자 이미지를 생성한다. 결과적으로 생성된 이미지가 분류 라벨과 완벽하게 일치하여 정확도는 0.999 이지만, SSIM 및 PSNR 은 각각 약 0.4 와 11 로 낮아, 생성된 이미지가 원본 이미지와 다름을 나타낸다. 입력에 더 가까운 계층에서 수행된 공격은 일반적으로 더 높은 유사성을 보인다. 입력 이미지가 더 나은 분류를 위해 이미지 특징에 대한 정보를 추출하는 ReLU 및 Max Pooling 연산을 거치면서 원본 이미지의 세부적인 정보는 압축된다. 따라서, 입력에 더 가까운 계층에서 수행된 공격은 원본 이미지의 세부 데이터를 더 많이 보유하고 있기 때문에 원본 이미지를 더 잘 재구성할 수 있다. Conv 1 과 Max pooling 1 을 통과한 후의 공격 이미지 간의 SSIM 과 PSNR 의 차이는 상대적으로 작다. 그러나 Max pooling 1 을 통과한 후와 Conv 2 를 통과한 후의 공격 이미지 간의 SSIM 과 PSNR 의 차이는 이전 경우에 비해 크다. 이는 Conv 계층을 통과할 때 이미지에 대한 정보가 Max pooling 계층을 통과할 때보다 더 압축되어 두 계층 사이의 유사성 차이가 더 크게 발생하기 때문이다.

그림 3 은 분류 라벨 '6'이지만 다른 필체를 가진 다섯 개의 MNIST 이미지에 대해 계층별 모델 역추론 공격을 통해 생성된 이미지를 보여준다. 그림 3 은 중



(그림 3) 숫자 '6'의 다양한 원본 이미지에 대한 계층별 모델 역추론 공격 생성 이미지

간 출력을 사용하여 수행된 계층별 모델 역추론 공격들이 모두 원본 이미지의 필체를 복원하려고 하며, 공격을 통해서 생성된 이미지들은 계층별로 유사성 정도가 다르게 나타나고 있음을 보여준다. Conv 3 후에 추출된 중간 출력에서의 파라미터 수는 가장 적은 파라미터를 가진 중간 출력보다 12 배 이상 작다. 원본 이미지의 특징에 대한 정보가 분류 라벨 분류에 필요한 정보로 압축되어, 시각적으로 관찰될 때 다른 계층에서의 공격을 통해 생성된 이미지들에 비해 원본 이미지와의 유사성이 덜하다. 또한, Conv 계층 후 출력에서의 공격을 통해 생성된 이미지와 Max Pooling 계층 후의 공격을 통해 생성된 이미지 사이에는 시각적 차이가 거의 없는 것을 확인하였다.

5. 결론

본 논문에서는 대상 모델의 중간 계층 출력을 추출하고 사전 훈련된 적대적 생성 신경망을 사용하여 원본 이미지를 재구성하며 공격을 수행하는 계층별 모델 역추론 공격을 소개한다. 본 논문의 접근 방식의 아이디어는 원본 이미지의 특성에 대한 정보를 가진 중간 계층의 출력을 기반으로 원본 이미지를 추론하는 것이다. MNIST 데이터셋에 대해 수행한 실험에서 다음과 같은 결과를 얻었다. 본 논문의 공격은 MNIST 분류 라벨을 보존하며 필체도 포착하여 원본과 유사한 이미지를 성공적으로 재구성한다. 최종적으로, 계층이 입력에 더 가까울수록 복원의 유사도가 더 높다.

향후 연구로는, MNIST 분류 모델뿐만 아니라 다양한 합성곱 신경망 모델들에 대한 계층별 모델 역추론 공격을 수행할 것이다. 각 모델들에 대한 생성된 공격 이미지들과 원본 이미지 사이의 분류 라벨 정확도와 유사성을 측정한다. 이 분석을 통해 모델 역추론 공격에 대해 어떤 합성곱 신경망 모델이 더 취약하거나 강인한지 식별할 수 있을 것이라고 기대한다.

참고문헌

[1] H. Yin, P. Molchanov, J. M. Alvarez, Z. Li, A. Mallya, D. Hoiem, N. K. Jha, and J. Kautz, “Dreaming to distill: Data-free knowledge transfer via deepinversion,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8715–8724.

[2] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song, “The secret revealer: Generative model-inversion attacks against deep neural networks,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 253–261.

[3] Z. Yang, J. Zhang, E.-C. Chang, and Z. Liang, “Neural network inversion in adversarial setting via background knowledge alignment,” in Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS), 2019, pp. 225–240.

[4] Ian J. Goodfellow, Pouget-Abadie, Mirza, Bing Xu, D. Warde-Farly, S. Ozair, A. Courville, Y. Bengio, “Generative Adversarial Nets,” NIPS, 2014.

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou, “Wasserstein gan,” arXiv preprint arXiv:1701.07875, 2017.

* 이 논문은 2024 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2020-0-01361, 인공지능대학원지원(연세대학교); No. 2022-0-00050, 데이터 플로우 구조 기반 PIM 의 실행 및 프로그래밍 모델 개발; No. RS-2023-00277060, 개방형 엣지 AI 반도체 설계 및 SW 플랫폼 기술개발; No. RS-2024-00395134, 차세대 AI 반도체를 위한 DPU 중심의 데이터센터 아키텍처). 또한 이 논문은 삼성전자의 지원을 받아 수행된 연구임.