딥 러닝 모델 추출 공격 기법 동향

백지훈[1], 문현곤[2]
[1]울산과학기술원 컴퓨터공학과 석사과정
[2]울산과학기술원 컴퓨터공학과 부교수

qorwlgns444@unist.ac.kr, hyungon@unist.ac.kr

A Survey on Deep Learing Model Extraction Attacks

Jihun Baek[1], Hyungon Moon[2]
[1]Dept. of Computer Science Engineering, UNIST
[2]Dept. of Computer Science Engineering, UNIST

요 약

딥 러닝 기술의 급속한 발전과 더불어, 이를 활용한 모델들에 대한 보안 위협도 증가하고 있다. 이들 중, 모델의 입출력 데이터를 이용해 내부 구조를 복제하려는 모델 추출 공격은 딥 러닝 모델 훈련에 높은 비용이 필요하다는 점에서 반드시 막아야 할 중요한 위협 중 하나라고 할 수 있다. 본 연구는 다양한 모델 추출 공격 기법과 이를 방어하기 위한 최신 연구 동향을 종합적으로 조사하고 분석하는 것을 목표로 하며, 또한 이를 통해 현재 존재하는 방어 메커니즘의 효과성을 평가하고, 향후 발전 가능성이 있는 새로운 방어 전략에 대한 통찰력을 제공하고자 한다.

1. 서론

현재 딥러닝 기술은 이제 우리 사회에서 빠질 수가 없는 요소 중 하나가 되었다. 이미지와 음성 인식, 자연어 처리, 스포츠와 의료 분야 등등 많은 곳에서 딥러닝 기술을 사용한 모델을 활용하고 있다.

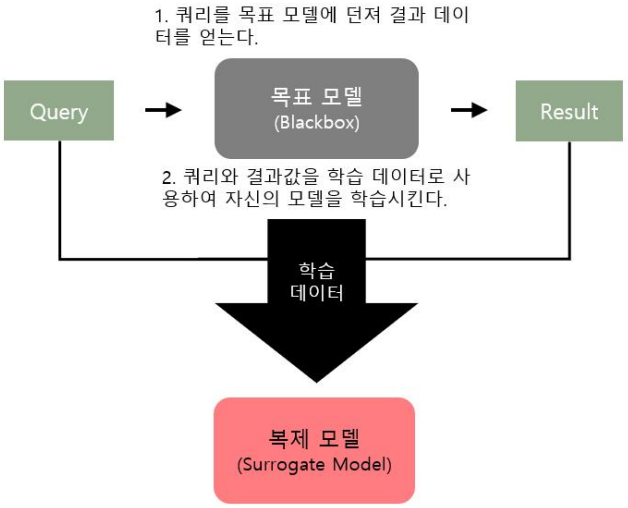
고성능 딥러닝 모델의 훈련에는 큰 비용과 노력이 필요하다. 딥러닝 모델의 성능은 학습 데이터의 질에 큰 영향을 받는데, 잘 정제되어 있고 바르게 라벨이 붙은 학습 데이터를 얻는 것은 매우 어렵거나 높은 비용을 필요로 한다. 혹은 고성능 모델의 훈련을 위해 복잡도가 높은 아키텍처(architecture)를 사용하는 경우, 높은 컴퓨팅 성능과 큰 메모리를 가진 연산장치를 장시간 큰 전력으로 구동시켜야 모델 훈련 과정을 완료할 수 있다. 이와 같은 여러 가지의 이유로 딥러닝 모델을 훈련하는 일은 높은 비용이 필요한, 도전적인 과정이다.

이와 같이 딥러닝 모델의 훈련에 높은 비용과 노력이 필요하다는 점은, 이를 악의적 공격으로부터 지킬 가치가 높은 대상으로 만든다. 이에 딥러닝 모델에 대한 다양한 보안 위협들이 논의되고 있는데 이들은 크게 딥러닝 모델의 기밀성에 대한 위협과 무결성에 대한 위협으로 분류될 수 있다. 모델이나 학습 데이터의 기밀성을 공격하는 방법에는 학습 데이터 추출 공격(Inversion Attack)과 모델 추출 공격(Extraction Attack)이 있고, 학습된 딥러닝 모델이나 그 활용(추론) 결과의 무결성을 공격하는 방법에는 오염 공격(Poisoning Attack)과 회피 공격(Evasion Attack)이 존재한다. 이 논문에서는 먼저 딥러닝 모델을 공격하는 네 가지 방법에 관해 설명한 후, 모델을 탈취하는 모델 추출 공격에 관한 여섯 가지 연구와 그 공격을 막는 방법에 관한 한 연구를 소개하려 한다.

2. 배경지식

딥러닝 모델에 대한 대표적인 보안 위협은 크게 다음과 같은 네 가지가 있다. 첫 번째는 학습 데이터 추출 공격(Inversion Attack)으로 모델의 학습을 위해 사용한 이미지 데이터를 복원하여 가져오는 방식이다. 데이터 분류를 위한 모델은 결과물로 결과의 신뢰도를 함께 출력하는데, 이 값을 사용하여 학습에 사용한 데이터를 추론하는 방법이다.

두 번째는 오염 공격(Poisoning Attack)으로 학습 데이터에 오염된 데이터를 추가하여 모델을 망가뜨



(그림 1) Knockoff Nets의 공격 흐름.

리는 공격 방법이다.

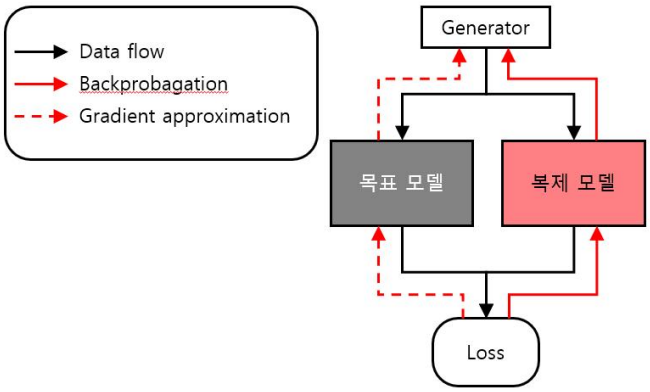
세 번째는 회피 공격(Evasion Attack)으로 입력 데이터에 노이즈를 추가하여 모델을 속이는 공격이다. 이미지 데이터에 노이즈를 추가하게 되면 인간의 눈으로는 큰 차이가 없지만, 모델은 다른 데이터로 인식할 수 있는데, 이 방식을 활용해 모델을 공격할 수 있다.

마지막은 이번 논문에서 주로 다루게 될 방식으로, 추출 공격(Extraction Attack)이다. 이 공격은 모델에 쿼리(Query)를 계속 던지면서 결괏값을 분석하여 모델을 복제하는 방식으로 진행한다. 이때 추출 공격의 타겟은 Graybox 모델과 Blackbox 모델이 있는데, 공격 방법도 두 가지로 나뉘게 된다. Graybox 모델은 목표 모델(Victim Model)에 대해 모델의 구조, 학습 데이터와 같은 사전 지식을 알고 있는 모델을 칭한다. 이 경에는 모델에게 미세하게 달라지는 데이터를 보내어 결괏값의 변화를 관찰하는 방식이나, 결정 경계선(Decision boundary) 근처의 데이터가 많은 정보를 가지고 있음을 활용하여 공격할 수 있다. Blackbox 모델은 목표 모델에 대해 알고 있는 사전 지식이 하나도 없을 때의 경우이다. 이럴 때는 학습 데이터에 대해 정보가 없기에 제한된 API만을 이용해 모델을 복제한다.

2. 대표적인 모델 추출 공격 기법들

1) Knockoff Nets

Knockoff Nets [1]는 이미 학습이 끝난 모델을 blackbox로 이용해 복제 모델을 학습하는 방식이다. Knockoff Net의 특징은 목표 모델을 학습시킨 학습 데이터를 모르기 때문에 그냥 다른 학습 데이터를 가져와 쿼리로 활용한다는 것이다. 또한 목표 모델의 구조도 모르기 때문에 복제 모델의 구조 또한 임의로 지정한다. 목표 모델의 임의의 데이터를 쿼리로 던진 후, 나오는 결과물을 학습 데이터로 활용하여 복제 모델을 학습한다.



(그림 2) Data-Free Model Extraction의 공격 다이어그램.

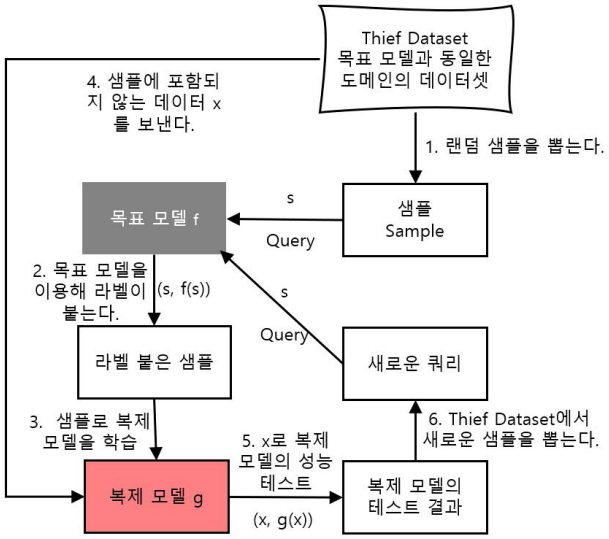
2) Data-Free Model Extraction

Data-Free Model Extraction [2] 역시 Knockoff Nets과 비슷하게 학습 데이터에 대한 정보 없이 공격하는 방식이다 (그림 2). 임의의 데이터를 목표 모델에 던져 결과로 얻을 수 있는 손실 함수를 활용하여 목표 모델과 비슷한 모델을 만드는 것이다.

모델의 학습 데이터의 정보가 없을 때, 특정 데이터가 아닌 임의의 데이터를 사용하여도 어느 정도 성능이 나오는 복제 모델을 만들 수도 있다는 사실을 Knockoff Nets와 Data-Free Model Extraction을 통해서 알 수 있다.

3) ActiveThief

Active Thief [3]는 목표 모델과 같은 도메인의 데이터셋(Thief Dataset)과 Subset selection strategies를 이용해 모델을 추출한다. 이때 Thief Dataset은 라벨이 없어도 상관이 없는데. Thief Dataset에서 샘플을 선택한 후, 목표 모델을 활용하여 라벨을 붙여주는 방식이다. 라벨이 붙은 샘플로 복제 모델을 학습시킨 후, 샘플로 뽑히지 않은 데이터로 학습된 복제 모델을 테스트한다. 이후 처음으



(그림 3) ActiveThief의 모델 추출 방식.

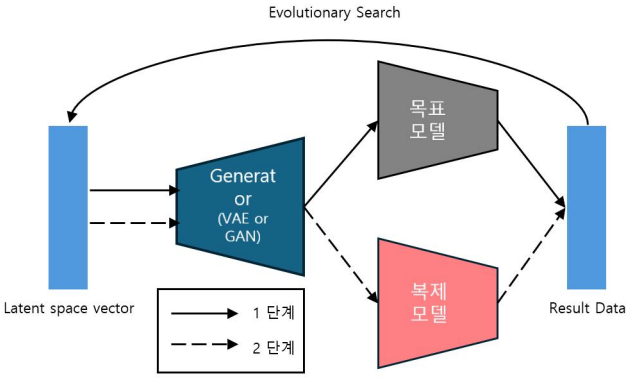
로 돌아가 아까의 샘플과는 다른 샘플을 뽑으며 과정을 반복한다.

이때 다른 샘플을 뽑는 방법을 Subset selection strategies라고 부른다. 임의로 뽑거나(Random), 학습이 덜 된 데이터를 선택하거나(Uncertainty), 학습된 데이터와 가장 다른 데이터를 선택하거나(k-center) 틀린 결과가 나온 데이터와 비슷한 데이터를 선택하는(DFAL) 등 다양한 방식이 있는데 연구 결과에는 k-center의 방식이 가장 accuracy가 높게 나왔다.

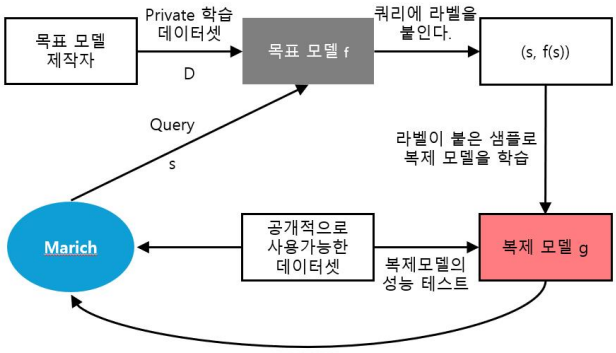
4)Black-Box Ripper

Black-Box Ripper [4]는 변분 오토인코더(VAE) 혹은 적대적 생성 신경망(GAN)을 제너레이터로 이용하여 Latent space vector로부터 학습 데이터를 만들어내는 방법이다.

Evolutionary Search 알고리즘으로 원하는 도메인의 이미지를 만들어 낼 Latent vector를 찾은 후, 제너레이터를 통해 라벨이 붙은 학습 데이터셋을 만들어 복제 모델에 학습시킨다. Black-Box Ripper는 ActiveThief와 다르게 직접 학습 데이터를 창조하여 모델을 학습하기에 원하는 도메인의 데이터가 없어도 공격이 가능하다는 장점이 있다.



(그림 4) Black-Box Ripper의 단계



(그림 5) Marich의 공격 방식

5)Marich

Marich [5]는 예측 API를 사용하여 공개적으로 사용 가능한 데이터셋에서 목표 모델로 최소한의 쿼리를 보내어 공격할 수 있게 설계해주는 활성 샘플링 기반 쿼리 선택 알고리즘이다. 다른 연구와는 달리 공개적으로 사용 가능한 쿼리 데이터셋을 사용한다는 것이 차별점인데, 복제 모델의 학습한 결과를 Marich로 보내어 더 좋은 학습 데이터를 만들 수 있는 쿼리를 찾아내는 방식으로 공격을 진행한다.

6) D-DAE

앞에서 설명한 다양한 방식은, 목표 모델이 쿼리 결과를 반환하기 전 데이터에 방해하는 방식으로 공격을 어느 정도 막을 수 있다. 이때 D-DAE [6]은 이 방어를 무너뜨리는 모델 추출 공격 체제이다. D-DAE는 방해 감지와 방해 복구로 이루어진 두 가지 모듈의 설계로 일반 추출 공격 방식과 통합하여 사용할 수 있다. 방해 감지 모듈은 목표 모델로부터 얻은 쿼리 결과를 통해 방어 메커니즘을 추론하고,

그 결과를 활용하여 방해 복구 모델에서 깨끗한 쿼리 결과로 만들어 일반적인 모델 추출 공격이 가능하게 만든다.

7) MEA-Defender

딥러닝 알고리즘을 통해 만들어진 심층 신경망(DNN) 모델들은 소유자의 지적 재산(IP)을 보호하기 위해 워터마크를 사용했다. 그러나, 이런 워터마크의 존재는 기존 모델 추출 공격 방식을 막을 수 없었다. 그래서 MEA-Defender [7]라는 DNN 모델의 IP를 보호할 수 있는 새로운 워터마크가 제시되었다. 입력 도메인에서 두 개의 소스 클래스에서 나온 샘플들을 조합함으로써 워터마크를 획득하고, 워터마크 손실 함수를 통해 워터마크의 출력 도메인 위치를 샘플 도메인 위치로 변경한다. 만약 모델 추출 공격에 당한다면 입력 도메인과 출력 도메인도 함께 도난당하기 때문에, 워터마크 역시 함께 도난당한 모델로 추출되게 된다.

4. 결론

다양한 방식의 모델 추출 공격 방식과 해당 공격을 막는 방법을 정리하였다. 모델 추출 공격에는 목표 모델에 던질 좋은 쿼리를 찾는 것이 중요하고 이 쿼리를 찾는 다양한 방법이 연구되었다. 임의로 데이터셋을 쿼리로 던진 뒤, 결과를 분석하여 적당한 쿼리를 찾는 방법도 있었고, 직접 학습 데이터를 만들어내는 방법도 있었다. 또한 마지막으로 모델 추출 공격은 결국 입출력 데이터를 활용하여 공격하기에, 이러한 공격을 워터마크를 활용하여 막는 방법도 소개하여 다음 모델 추출 공격의 방향성을 제시할 수도 있었다.

Acknowledgement

본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2022R1F1A1076100). 또한 이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2021-0-00724, 임베디드 시스템 악성코드 탐지·복원을 위한 RISC-V 기반 보안 CPU 아키텍처 핵심 기술 개발)

참고문헌

[1] OREKONDY, Tribhuvanesh; SCHIELE, Bernt; FRITZ, Mario. Knockoff nets: Stealing functionality of black-box models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. p. 4954-4963.

[2] TRUONG, Jean-Baptiste, et al. Data-free model extraction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021. p. 4771-4780.

[3] PAL, Soham, et al. Activethief: Model extraction using active learning and unannotated public data. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2020. p. 865-872.

[4] BARBALAU, Antonio, et al. Black-box ripper: Copying black-box models using generative evolutionary algorithms. Advances in Neural Information Processing Systems, 2020, 33: 20120-20129.

[5] KARMAKAR, Pratik; BASU, Debabrota. Marich: A Query-efficient Distributionally Equivalent Model Extraction Attack. Advances in Neural Information Processing Systems, 2024, 36.

[6] CHEN, Yanjiao, et al. D-dae: Defense-penetrating model extraction attacks. In: 2023 IEEE Symposium on Security and Privacy (SP). IEEE, 2023. p. 382-399.

[7] LV, Peizhuo, et al. MEA-Defender: A Robust Watermark against Model Extraction Attack. arXiv preprint arXiv:2401.15239, 2024.