부채널 분석을 이용한 딥러닝 네트워크 공격 동향

김덕영[1], 김현지[2], 김현준[2], 서화정[3]
[1]한성대학교 융합보안학과 석사과정
[2]한성대학교 정보컴퓨터공학과 박사과정
[3]한성대학교 융합보안학과 교수
dudejrdl123@gmail.com, khj1594012@gmail.com, khj930704@gmail.com,
hwajeong84@gmail.com

Deep learning network attack trends using side channel analysis

Duk-Young Kim[1], Hyun-Ji Kim[2], Hyun-Jun Kim[2], Hwa-Jeong Seo[3]
[1]Dept. of IT Convergence Security, Han-sung University
[2]Dept. of Infomation Computer Engineering, Han-sung University
[3]Dept. of IT Convergence Security, Han-sung University

요 약

최근 빠른 속도로 개발되고 있는 인공지능 기술은 여러 산업 분야에서 활용 되고 있다. 그러나 최근 딥러닝 네트워크에 대한 부채널 공격 기법들이 등장하고 있으며, 이는 해당 모델을 재구현하여 자율 주행 자동차에 대한 해킹 등과 같이 치명적인 보안 위협이 될 수 있으므로 이에 대한 이해와 대응책이 필요하다. 본 논문에서는 딥러닝 네트워크에 대한 부채널 공격 기법 동향에 대해 살펴보고, 이에 대한 대응 기술 또한 함께 알아본다.

1. 서론

최근 4차 산업혁명을 선도하는 기술 중 하나는 인공 지능이며, 특히 자율 주행 자동차, 이미지 생성, 가상 음성 생성 등에 활용되는 딥러닝의 연구가 활발하게 이루어지고 있다[1]. 공격자가 딥러닝 가속기에 접근하거나 이를 탈취하는 경우, 부채널 분석[2]을 통해 가속기의 내부 비밀 정보인 가중치나 편향 값을 복구할 수 있다. 이와 같이 부채널 분석을 통해 딥러닝 모델의 내부 구조가 알려지게 될 경우, 공격자는 해당 모델을 재구현할 수 있으므로 자율 주행 자동차에 대한 해킹 등과 같은 치명적인 보안 위협이 될 수 있다. 따라서 이에 대한 대응책이 필요하다. 본 논문에서는 딥러닝 네트워크에 대한 부채널 공격 기법[]과 그에 대한 대응책을 살펴본다.

2. 관련연구

2.1 딥러닝(Deep Learning)

딥러닝은 인공 지능의 한 분야로, 인공 신경망을 기반으로 한 머신러닝 알고리즘이다. 이 알고리즘은 다층 인공신경망을 사용하며 입력 데이터로부터 복잡한 패턴을 학습하고, 이를 통해 데이터를 분석하고 예측한다. 딥러닝 알고리즘에는 다양한 구조가 있으며, 대표적으로 DNN, MLP, CNN이 있다. 그림 1과 같은 MLP는 입력층, 은닉층, 출력층으로 구성된 가장 기본적인 딥러닝 네트워크이다. 이러한 구조는 처리 속도가 빠르고 1차원 데이터 처리에 용이하다. 그림 2에서 나타나는 CNN은 2차원 이상의 데이터를 처리하기 위한 모델이다. 컨볼루션 레이어(Convolution layer)와 풀링 레이어(Pooling layer)를 추가하여 입력 데이터의 특징을 추출하고 데이터 압축을 수행하여 대표 값을 추출하며, 특히 이미지 관련 작업에 굉장히 효과적이다. 이러한 딥러닝 모델은 역전파 알고리즘을 사용하여 실제 값과 계산된 값의 오차를 계산하고 가중치를 조정하면서 데이터를 학습한다. 이 과정을 반복하여 오차를 최소화하고 입력과 출력 간의 관계를 학습한다.

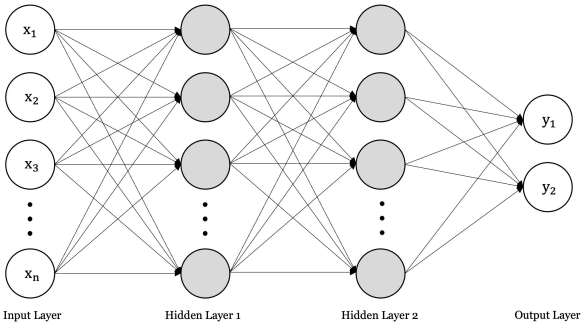

그림 1. DNN architecture

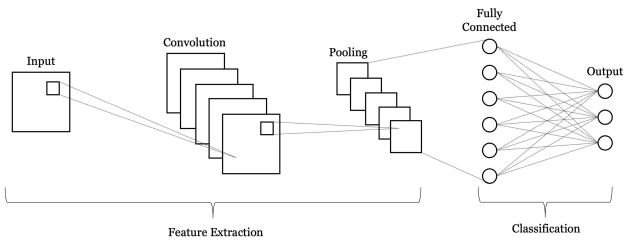

그림 2. CNN architecture

2.2 부채널 분석(Side Channel Analysis)

부채널 분석은 전자기기가 동작할 때 발생하는 소비 전력, 전자파, 시간 등의 부채널 정보를 이용하여 내부 비밀 정보를 복원하는 기술이다. 복원하는 방법에는 단순 전력 분석(Simple Power Analysis, SPA) [3], 상관 전력 분석(Correlation Power Analysis, CPA) [4], 차분 전력 분석(Differential Power Analysis, DPA) [5] 등이 있으며, 그중에서도 상관 전력 분석이 가장 대표적이다.

2.2.1 단순 전력 분석

단순 전력 분석은 단일 파형으로 분석하는 방법이며, 공격자는 공격 지점의 연산 과정과 구현 방법을 정확히 알아야하는 기법이다. 예를 들어 RSA에서의 square and multiply 연산을 수행할 경우, data가 1이면 square and multiply연산, data가 0이면 square 연산만을 수행하는 것이다. 이처럼 어떤 경우에 어떠한 명령어가 수행되는지 알아야 가능한 부채널 분석 방법이다.

2.2.2 상관 전력 분석

상관 전력 분석은 다수의 파형을 이용하는 통계적 분석을 통해 내부 비밀 정보를 복원하는 방법으로 전자기기에서 수집한 부채널 정보와 공격자가 추측한 중간값과의 상관계수를 계산하여 가장 유의미한 결과를 도출한 중간값을 내부 비밀 정보로 결정한다.

2.2.3 차분 전력 분석

차분 전력 분석은 해밍웨이트 모델으로써, 수행하는 연산, 사용되는 데이터, 노이즈, 기본 소비전력에 의해 총 전력이 결정된다. 전력 파형을 얻을 경우 특정 연산 지점에 대한 세부적인 값 차이는 데이터에 의한 차이이고, 그래프의 위상 차이는 연산의 종류에 따라 결정된다. 즉, 전력 소비가 높은 연산이라면 해밍웨이트가 높을 것이다. 이처럼 전체 그래프 패턴을 보고 데이터의 값 차이인지, 명령어의 차이인지 등을 알 수 있으며, 알고리즘 내부의 함수 패턴을 파악할 수 있다.

3. 딥러닝 네트워크 부채널 공격 소개 및 대응책

3.1 딥러닝 네트워크에 대한 부채널 공격 기법

딥러닝 네트워크에 대한 부채널 공격은 주로 내부 파라미터 및 내부 구조를 복원하는 과정으로 나뉜다. 내부 파라미터 복구의 경우, 네트워크를 형성하는 가중치, 입력값, 활성 함수 등과 같은 다양한 요소가 대상이며 내부 구조의 경우 네트워크 층의 수나 뉴런의 수가 주요 대상이다. MLP (Multi Layer Perceptron)과 CNN (Convolutional Neural Network) 모델을 대상으로 MNIST 데이터 셋을 사용하여 모델의 가중치, 활성 함수, 레이어와 뉴런의 수 등을 높은 정확도로 복원하는 등 내부 비밀 정보 복구에 대한 연구가 활발히 이루어지고 있다. 초기 연구에서는 가중치의 범위나 활성 함수에 제한을 두는 등 비현실적인 가정이었지만, 최근 연구에서는 최소한의 부채널 정보를 이용하여 네트워크 내부 가중치와 활성 함수를 복원하는데 성공하는 등 다양한 시도가 계속 되고 있다.

최근 연구 동향으로, Yoshida et al.[6]는 네트워크의 내부 구조를 알고 있는 상황에서 Chain CPA(Chain Differential Cryptanalysis with chosen plaintext)라는 선택 평문 공격을 이용하여 블록 암호의 내부 구조를 파악하고 암호화 키를 찾아내는 공격 기법을 활용하여 가중치 정보를 복구하는 방안을 제시했다. 그러나 동일 저자의 후속 연구에서는 단순전력분석의 정확도가 낮다는 한계를 언급하였다.[7] 또한, Maji et al.[8]은 그레이 박스 환경에서 CNN과 BNN 모델의 가중치, 편향, 입력값을 복구하는 기술을 제안하며, 측정된 파형의 SNR(signal to noise ratio) 및 모델의 복잡성을 최소화할 수 있음을 입증했다. 그러나 위의 기술도 그레이 박스 환경은 공격 대상 네트워크의 구조를 사전에 알고 있다는 한계가 있다.

이러한 연구 결과들은 전력 및 전자파 분석과 같은 부채널 분석이며, 이에 대응하기 위한 연구의 중요성이 강조되고 있다. 또한, 최근 활발히 사용되고 있는 딥러닝 가속기에서 인공지능 모델을 연산하는 경우가 많으므로 딥러닝 가속기의 안전성에 대한 연

구 필요성도 증대되고 있다.

3.2 딥러닝 네트워크에 대한 부채널 공격 대응 기술

딥러닝 네트워크를 대상으로 한 부채널 공격은 딥러닝 모델의 보안을 위협하는 중요한 문제이다. 이러한 공격에 대비하여 다양한 방어 기술들이 연구되고 있으며, 대표적인 예로 셔플링(Shuffling)[9]과 마스킹(Masking)[10] 기법이 있다.

3.2.1 셔플링

셔플링은 부채널 공격으로부터 딥러닝 모델을 보호하기 위한 중요한 방법 중 하나로 이 방법은 DNN 모델의 내부 정보를 보호하는 데 사용된다. 셔플링은 보통 데이터나 가중치 등의 패턴을 무작위로 섞어서 공격자가 모델의 내부 정보를 추출하기 어렵게 만드는 기법으로 공격자의 공격 복잡성을 증가시키면서도 시스템의 오버헤드를 최소화할 수 있게 만든다. 최근 Liu et al.[9]의 연구에서 셔플링을 활용하여 대규모 DNN 모델의 내부 정보를 보호하고 확장 가능한 방법을 제안하였다.

3.2.2 마스킹

마스킹 기법은 신경망을 보호하는 또 다른 중요한 방어 기술로 신경망의 파라미터와 중간 계산 결과에 임의성을 추가하여 공격자가 내부 정보를 추출하기 어렵게 만든다. 최근 Althanasiou et al.[10]의 연구에서 마스킹은 보안성을 향상시키지만, 일부 성능 저하를 초래할 수 있다는 결과가 나왔지만 이러한 성능 저하는 정확도에 큰 영향을 미치지 않는 것으로 확인되었다.

4. 결 론

딥러닝 네트워크에 대한 부채널 공격은 현재 매우 중요한 보안 문제로 대두되고 있다. 딥러닝 네트워크에 대한 부채널 공격은 주로 내부 파라미터 및 내부 구조를 복원하므로 동일한 딥러닝 모델이 재구현될 수 있는 보안 취약점이 존재한다. 이러한 공격으로부터 딥러닝 모델을 보호하기 위해 셔플링과 마스킹과 같은 다양한 방어 기술들이 제안되고 있다. 최근 연구에서 셔플링 기법을 활용한 대규모 DNN 모델의 내부 정보를 보호하고 확장 가능한 방법과 마스킹 기법을 이용하여 내부 정보를 추출하기 어렵게 만들어 보안성을 향상시키는 등 다양한 기법들이 연구되고있다.

5. Acknowledgment

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.2018-0-00264, Research on Blockchain Security Technology for IoT Services, 50%) and this work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-00627, Development of Lightweight BIoT technology for Highly Constrained Devices, 50%).

참고문헌

[1] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436-444.

[2] Joye, Marc, and Francis Olivier. "Side-Channel Analysis." (2011): 1198-1204.

[3] Mangard, Stefan, Elisabeth Oswald, and Thomas Popp. Power analysis attacks: Revealing the secrets of smart cards. Vol. 31. Springer Science & Business Media, 2008.

[4] Kocher, Paul, Joshua Jaffe, and Benjamin Jun. "Differential power analysis." Advances in Cryptology—CRYPTO'99: 19th Annual International Cryptology Conference Santa Barbara, California, USA, August 15-19, 1999 Proceedings 19. Springer Berlin Heidelberg, 1999.

[5] Brier, Eric, Christophe Clavier, and Francis Olivier. "Correlation power analysis with a leakage model." Cryptographic Hardware and Embedded Systems-CHES 2004: 6th International Workshop Cambridge, MA, USA, August 11-13, 2004. Proceedings 6. Springer

Berlin Heidelberg, 2004.

[6] Yoshida, Kota, et al. "Model reverse-engineering attack using correlation power analysis against systolic array based neural network accelerator." 2020 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2020.

[7] Yoshida, Kota, et al. "Model reverse-engineering attack against systolic-array-based dnn accelerator using correlation power analysis." IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences 104.1 (2021): 152-161.

[8] Maji, Saurav, Utsav Banerjee, and Anantha P. Chandrakasan. "Leaky nets: Recovering embedded neural network models and inputs through simple power and timing side-channels—Attacks and defenses." IEEE Internet of Things Journal 8.15 (2021): 12079-12092.

[9] Liu, Yuntao, Dana Dachman-Soled, and Ankur Srivastava. "Mitigating reverse engineering attacks on deep neural networks." 2019 IEEE Computer Society Annual Symposium on VLSI (ISVLSI). IEEE, 2019.

[10] Athanasiou, Konstantinos, et al. "Masking feedforward neural networks against power analysis attacks." proceedings on privacy enhancing technologies (2022).