신뢰할 수 있는 수행 환경을 활용한 DNN 보안에 관한 연구 동향

김영주*, 강정환*, 권동현**
*부산대학교 정보융합공학과
**부산대학교 정보컴퓨터공학부
y0ungjupnu@pusan.ac.kr, jeonghwan@pusan.ac.kr, kwondh@pusan.ac.kr

A Study Trend on DNN security by using Trusted Execution Environment

Youngju Kim*, Jeong-Hwan Kang*, Dong-Hyun Kwon**
*Dept. of Information Convergence Engineering, Pusan National University
*Dept. of Computer Science and Engineering, Pusan National University

요 약

심층 신경망 기술은 실시간 예측 서비스를 위한 다양한 응용 분야에 적용되고 있다. 그뿐만 아니라 최근에는 민감한 개인 정보나 중요 정보들도 이러한 심층 신경망 기술을 통해 처리되면서 보안에 관한 관심이 높아지고 있다. 본 논문에서는 이러한 심층 신경망의 보안을 위해 하드웨어 기반의 안전한 수행환경에서 심층 신경망을 수행함으로써 연산 과정을 보호하는 연구들과 안전한 수행환경 내에서도 효율적인 심층 신경망 처리 기술들을 살펴볼 것이다. 그리고 이러한 연구 동향을 토대로 앞으로의 심층 신경망 연산 보호 기술의 연구 방향에 대해 논하도록 하겠다.

1. 서론

심층 신경망(deep neural network, DNN)은 실시간 예측 서비스를 제공하기 위해 스마트폰, 자율 주행 차량, 산업 자동화, IoT와 같은 다양한 응용 분야에 사용되고 있다. 빠른 예측을 위해 최종 사용자 장치에서 DNN을 수행하는 경우가 늘고 있다[1]. 특히, 최근에는 모바일 환경에서 인증 앱에 사용되는 생체 인식이나, 의료 앱에서 사용되는 의료 정보와 같은 민감한 정보를 포함한 데이터들도 DNN에 의해 처리되고 있다.

하지만 심층 신경망 연산도 소프트웨어로 동작하기 때문에 다른 소프트웨어들과 마찬가지로 보안 취약점이 존재한다. 연산 중 처리되는 정보들은 공격자에 의해 손상 혹은 유출될 수 있다. 이러한 취약점을 해결하기 위해 암호화 알고리즘을 적용하거나 신뢰할 수 있는 실행 환경(trusted execution environment, TEE)을 사용하는 방법 등이 도입되고 있다.

그중 본 논문에서는 TEE를 활용하여 DNN의 보안 취약점을 막기 위한 여러 방어 기법들을 살펴본다.

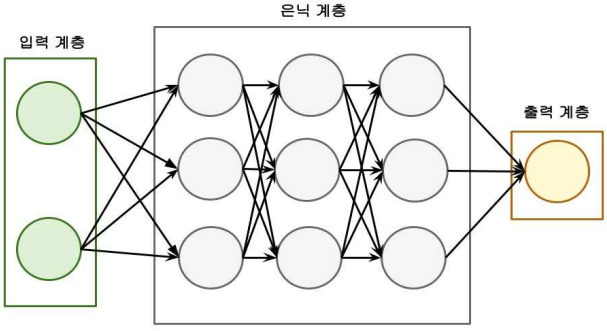
2. 배경 지식

2.1. 심층 신경망(DNN)

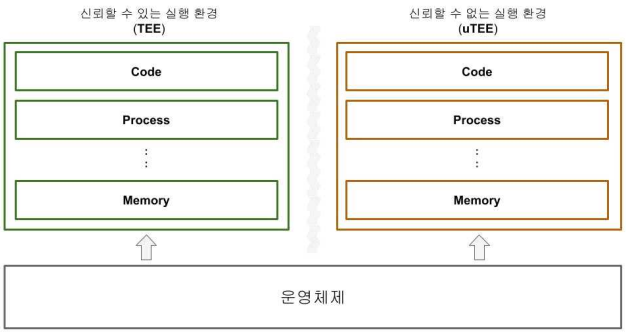
심층 신경망(deep neural network, DNN)은 기계학습 알고리즘 중 하나이다. 사람의 신경망(neural network, NN) 원리와 구조를 모방한 기계학습 알고리즘 중 DNN은 특히 여러 계층을 사용하여 점진적으로 더 높은 수준의 특징을 추출하여 결과를 예측한다(그림1). 계층을 여러 번 더 할수록 연산은 더 깊어진다. 그로 인해 한층 복잡한 학습을 수행해 정확한 예측을 만들어 낼 수 있다. DNN의 정확도 높은 결과를 위해서 수백 MB에서 최대 수 GB에 이르는 메모리 용량이 요구될 수 있다[2].

2.2. 신뢰할 수 있는 실행 환경(TEE)

신뢰할 수 있는 실행 환경 기술은 일부보안상 중요한 애플리케이션을 위한 전용 수행환경이다. 이는



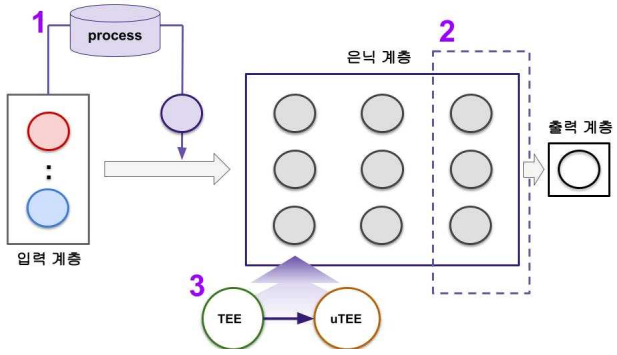
(그림 1) 심층 신경망 구조



(그림 2) TEE의 개요

일반 애플리케이션을 위한 신뢰할 수 없는 실행 환경(untrusted execution environment, uTEE)과 구분된다(그림2). 즉 TEE는 이렇게 중요 애플리케이션에 격리된 실행 환경을 제공하여 신뢰할 수 없는 수행환경이 공격을 받더라도 중요 애플리케이션의 실행 코드의 무결성, 프로세스의 정보보호, 보안 메모리에 대한 데이터 기밀성을 보장한다. 결과적으로 TEE는 중요 애플리케이션에 대한 공격 가능한 부분을 크게 줄여 보안성을 향상한다. 대표적으로 Intel과 ARM은 각각 SGX[3]와 TrustZone[4]이라는 TEE를 제공한다.

이러한 TEE는 일반적으로 TEE에서 사용할 수 있는 메모리의 크기가 제한되어 있다. 예를 들어, Intel SGX의 경우 EPC (enclave page cache)라고 하는 정해진 메모리 영역만을 사용해야 한다[3]. 한편 ARM TrustZone의 신뢰할 수 있는 수행환경에서 동작하는 OP-TEE[5]라고 하는 소프트웨어의 경우에는 단지 7MB의 메모리만을 사용한다.



(그림 3) 여러 TEE 활용한 DNN 보안 취약점 방어 기술 요약. (기본적으로 왼쪽에서 오른쪽으로 DNN의 과정이 이루어진다. 보라색의 1, 2, 3은 각 기술마다 취약점을 방어하기 위한 주요 기술들을 간략하게 보여준다.)

3. TEE 활용 DNN 보안 취약점 방어 기술

DNN의 데이터 보호를 보장하면서 동시에 DNN의 사용성을 해치지 않기 위해서는 정확성과 빠른 연산속도를 유지하는 것이 중요하다. 이로 인해 동형 암호화(homomorphic encryption, HE)와 같은 암호화 알고리즘을 적용하는 것[6]은 데이터 보호에는 효과적이지만 연산속도 면에서 많은 성능 부하가 발생한다. 그 때문에 이보다 TEE 상에서 DNN을 실행하는 접근 방식이 적은 성능 부하와 높은 기밀성을 동시 보장할 수 있는 기술로 여겨져 다양한 연구들이 있어왔다.

이에 이번 장에서는 TEE를 활용하여 DNN의 보안 취약점을 방어한 여러 사례를 살펴본다.

TensorSCONE[7]의 경우 입력 계층의 크기를 줄여 TEE에 실행하는 방식을 제안하였다.(그림3의 1) 즉 이를 통해 TEE에서 처리되는 입력의 크기를 줄여 TEE의 적은 메모리로도 DNN을 동작할 수 있게 하였다. 하지만 입력의 크기를 줄임에 있어 양자화 기술이 도입되는데[8], 이는 추론의 정확도를 떨어뜨린다. **SGX-BigMatrix**[9]도 유사하게 TEE에 입력 데이터에 대한 별도의 데이터 분석 방식을 도입(그림 3의 1)하여 효율적인 연산을 수행한다. 하지만 정렬과 같은 특정 분석 작업에 대해서만 효율적으로 작동한다는 한계를 지닌다.

Privado[10]의 경우 앞선 연구들과는 달리 입력을 줄이는 것이 아닌 DNN의 계층 중 일부만 TEE에서 실행함으로써 보안을 높인다. 하지만 일부의 연산만 수행(그림 3의 2) 할 경우 결과의 정확도가 떨어질 수 있다. 또한, 이렇게 일부 계층에 적용함에도 불구하고 충분히 작은 메모리에서 기술을 수행할 수 없어 한계점을 가진다.

Slalom[11]의 경우 DNN의 연산을 두 부분으로 나누고 신뢰할 수 있는 환경(TEE)에서 신뢰할 수 없는 환경(uTEE)으로 연산을 아웃소싱(그림3의 3)한다. 이러한 분할은 모든 연산을 수행하지만, 충분히 작은 메모리로도 작업을 수행할 수 있다. 하지만 이는 실행 환경 간에 잦은 교환을 유발하며 제어 탈취 공격(control hijacking attack)등으로 부터 취약할 수 있다.

4. 결론

DNN의 연산을 TEE에 실행하여 보호하는 방식은 암호화 연산과 같은 큰 성능 부하를 일으키지 않고

높은 보안을 보장할 수 있다. 하지만 앞선 연구들에서 볼 수 있듯이 TEE의 제한된 메모리로 인해 DNN 연산 전체를 TEE에서 실행하는 데 한계가 있다.

점점 더 커지는 DNN 처리량을 고려할 때, TEE에서 수행되는 DNN의 연산을 계층 단위로 DNN 나누는 것에서 나아가 한 계층을 안에서 세분화하여 나누는 등의 연산 분할에 관한 더 많은 연구가 필요할 것이다.

ACKNOWLEDGMENT

본 논문은 2021년 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임. (IITP-2019-0-01343)

참고문헌

[1] K. Ota, M. S. Dao, V. Mezaris, and F. G. B. De Natale, “Deep learning for mobile multimedia: A survey,” ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 13, no. 3s, pp. 34:1 – 34:22, 2017.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Conference on Computer Vision and Pattern Recognition(CVPR), 2016, pp. 779 – 788.

[3] McKeen, F., Alexandrovich, I., Berenzon, A., Rozas, C. V., Shafi, H., Shanbhogue, V., & Savagaonkar, U. R.(2013). Innovative instructions and software model for isolated execution. Hasp@isca, 10(1).

[4] Varanasi, P.,& Heiser, G.(2011, July). Hardware-supported virtualization on ARM. In Proceedings of the Second Asia-Pacific Workshop on Systems (pp. 1-5).

[5] OP-TEE Documentation, https://optee.readthedocs.io/en/latest/, 2021

[6] Oseni, Ayodeji, et al. "Security and Privacy for Artificial Intelligence: Opportunities and Challenges." arXiv preprint arXiv:2102.04661 (2021).

[7] Kunkel, R., Quoc, D. L., Gregor, F., Arnautov, S., Bhatotia, P., & Fetzer, C. (2019). TensorSCONE: a secure TensorFlow framework using Intel SGX. arXiv preprint arXiv:1902.04413.

[8] Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., ... & Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2704-2713).

[9] Shaon, F., Kantarcioglu, M., Lin, Z., & Khan, L. (2017, October). Sgx-bigmatrix: A practical encrypted data analytic framework with trusted processors. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 1211-1228).

[10] Tople, S., Grover, K., Shinde, S., Bhagwan, R., & Ramjee, R. (2018). Privado: Practical and secure DNN inference. arXiv preprint arXiv:1810.00602.

[11] Tramer, F., & Boneh, D. (2018). Slalom: Fast, verifiable and private execution of neural networks in trusted hardware. arXiv preprint arXiv:1806.03287.

[12] Gupta, N., Jati, A., & Chattopadhyay, A. (2020). MemEnc: A Lightweight, Low-Power and Transparent Memory Encryption Engine for IoT. IEEE Internet of Things Journal.

[13] Oseni, A., Moustafa, N., Janicke, H., Liu, P., Tari, Z., & Vasilakos, A. (2021). Security and Privacy for Artificial Intelligence: Opportunities and Challenges. arXiv preprint arXiv:2102.04661.

[14] Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018, April). Sok: Security and privacy in machine learning. In 2018 IEEE European Symposium on Security and Privacy (EuroS&P) (pp. 399-414). IEEE.