신뢰실행환경 기반 온디바이스 심층 신경망 보호에 관한 연구

유준승[1], 강기봉[1],백윤흥[2]
[1]서울대학교 전기정보공학과 석박통합과정, 반도체공동연구소
[2]서울대학교 전기정보공학과 교수, 반도체공동연구소
jsyou@sor.snu.ac.kr, kbkang@sor.snu.ac.kr, ypaek@snu.ac.kr

A Study on Protecting On-Device Deep Nerual Networks with Trusted Execution Environments

Junseung You[1], Kibong Kang[1], Yunheung Paek[1]
[1]Dept. of Electrical and Computer Engineering and Inter-Univeristy Semiconductor Research Center (ISRC), Seoul National University

요 약

온디바이스(On-device) 딥러닝은 사용자의 프라이버시를 보호할 수 있다는 장점으로 인해 점차 주목받고 있다. 그러나 이러한 접근은 딥러닝 모델이 사용자 디바이스에 전달됨에 따라, 오히려 모델 자체의 프라이버시가 취약해지는 문제를 야기한다. 이에 때라 최근 연구들은 신뢰실행환경(Trusted Execution Environment, TEE)을 활용하여 딥러닝 모델의 프라이버시를 보호하고자 시도하고 있다. 그러나 기존 TEE 기반 딥러닝 모델 보호 기법들이 전제하는 TEE의 위협 모델은 모델 프라이버시를 충분히 보장하지 못하는 한계를 가진다. 본 논문에서는 기존의 TEE 위협 모델을 분석하고, 이를 바탕으로 기존 보호 기법들이 방어하지 못하는 공격 시나리오들을 제시한다. 또한, 이러한 공격에 대응하기 위한 기본적인 방안과 그 한계점에 대해 논의한다.

1. 서론

사용자 프라이버시 강화를 위해 온디바이스 딥러닝은 최근 그 도입이 점점 확대되고 있다. 이는 로컬 디바이스 상에서 심층신경망 추론을 수행함으로써, 사용자의 얼굴이나 지문과 같은 민감한 정보가 원격 서버에 직접 노출되는 것을 방지한다. 그러나 이러한 접근은 모델 프라이버시에 대한 보안 이슈를 야기하고 있다. 특히, 딥러닝 모델이 디바이스로 전달되어야 하기 때문에 경제적으로 가치가 큰 고유 자산인 모델이 사용자에게 직접적으로 노출될 수 있으며, 이를 활용해 모델 학습에 사용된 다른 사용자의 사적 데이터까지 유출할 수 있다. 이에 따라, 다양한 연구들은 디바이스 내 악의적일 수 있는 프로그램 및 시스템 구성요소로부터 안전하게 격리된 환경을 제공하는 신뢰실행환경 기술을 활용하여 그 내부에서 모델을 실행하는 방어 기법들을 제시한다. 특히, 온디바이스 딥러닝이 대부분 Arm 기반 시스템 온 칩(모바일, 테블릿 등)에서 실행되기 때문에 기존 보호 기법들은 Arm 프로세서에 초점을 둔다.

이러한 TEE 기반 온디바이스 딥러닝 보호 기법들은 TEE가 하나의 (완전무결한) 블랙박스라는 가정을 전제로 한다. 하지만 안타깝게도 이는 사실이 아니다. TEE에 대한 취약점 및 공격은 지속적으로 발견되고 있으며, 이를 활용하여 TEE 내부에서 실행되고 있는 모델 정보를 유출할 수 있다. 예를 들어, 하드웨어 기반 TEE인 Arm TrustZone의 경우 메모리에 저장하는 데이터를 암호화하지 않기 때문에 메모리 스누핑이나 인터포징 또는 콜드 부트 공격과 같은 물리적인 공격들로부터 취약하며, 가상화 기반 TEE의 경우 가상화를 관리하는 하이퍼바이저가 공격되면 TEE가 제공하는 격리가 깨질 수 있다.

이러한 기존 연구들의 TEE의 보안 수준에 대한 가정과 TEE 공격 가능성 사이의 불일치에 따른 온디바이스 딥러닝 프라이버시 유출을 막기 위해 본 논문에서는 기존 보호 기법들의 TEE에 대한 위협 모델을 분석하고, 이를 바탕으로 각 기법들이 대응하지 못하는 공격을 제시한다. 또한, 공격들에 대응하기 위해 최근 공개된 Arm의 하드웨어 기반 TEE인 Confidential Compute Architecture (CCA)를 활용한 기본적인 해결책을 논의함과 동시에 그 활용에서 새롭게 파생되는 보안 문제에 대해 설명한다.

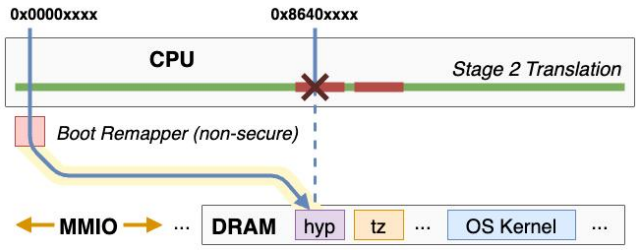
2. 기존 TEE 기반 온디바이스 딥러닝 보호 기법

<표 1> 기존 TEE 기반 온디바이스 딥러닝 보호 기법들이 차용한 TEE의 종류 및 신뢰 모델.

	신뢰실행환경	Trust Model
DarkneTZ[1]	TrustZone	blackbox
ShadowNet[2]	TrustZone	blackbox
ASGARD[3]	Virtualization	trusted hyper.

기존 TEE 기반 온디바이스 딥러닝 보호 기법들은 크게 두 가지 종류의 TEE를 사용한다(표1). 첫 번째 종류는 하드웨어 기반 TEE이다. 이는 TEE 내부에서 실행되는 프로그램 및 데이터를 보호하기 위한 격리-다른 시스템 컴포넌트가 보호하고자 하는 코드 및 데이터 메모리를 접근하지 못하게 하는 것-를 하드웨어가 보장해준다. 구체적으로, Arm에서 제공하는 TrustZone의 경우, 프로그램을 두 가지 중 하나의 환경에서 실행되한다. 하나는 normal world로, 이는 일반적인 프로그램이 실행되는 환경이다. 또 다른 하나는 secure world로, 보호하는 프로그램이 실행되는 환경이다. 이 때 시스템의 메모리는 페이지 별로 어느 환경에서 사용되는 메모리인지로 구분되며, TrustZone Address Space Controller라는 하드웨어 컴포넌트를 통해 normal world에서 실행되는 프로그램은 secure world에서 사용하는 메모리를 접근하지 못하게 한다. 두 번째 종류는 가상화 기반 TEE이다. 이는 TEE를 위한 격리를 가상화를 통해 제공한다. 이 때 가상화란 각 프로세스/프로그램을 서로 다른 가상머신 단위로 관리하는 것을 뜻한다. 가상화는 Arm의 경우 2단계 주소 변환을 통해 이루어지는데, 1단계 주소 변환이 1단계 페이지 테이블을 이용하여 가상 주소를 간이(intermediate) 물리 주소로 변환 후, 2단계 주소 변환이 2단계 페이지 테이블을 이용하여 가상 물리 주소를 실제 물리 주소로 변환한다. 이 때 2단계 주소 변환을 담당하는 2단계 페이지 테이블에 대한 관리 권한을 프로그램이나 가상 머신 커널보다 더 높은 권한을 가지는 시스템 소프트웨어인 하이퍼바이저에게 부여한다. Arm의 경우 이러한 권한을 Exception Level (EL)을 통해 차등적으로 부여하며, 일반 프로그램은 EL0, 커널은 EL1, 하이퍼바이저는 EL2에서 실행된다. 이를 통해 하이퍼바이저는 2단계 페이지 테이블 엔트리(entry)에 저장되는 간이 주소-물리 주소 간의 매핑을 조정하여 허용되지 않은 프로그램이 TEE의 메모리를 접근하려고 할 시 주소 변환 오류가 나게끔 하여 격리를 실현할 수 있다.

두 종류의 TEE는 온디바이스 딥러닝 보호 기법에 활용된다. 이 때, 각 보호 기법들은 TEE를 활용함에 있어서 TEE의 종류에 따라 다른 신뢰 모델을 차용한다. 하드웨어 기반 TEE를 활용하는 기법들의 경우 TEE를 하나의 블랙박스로 전제한다. 즉, 해당 연구들에 있어서 TEE는 하나의 도구로써 그 내부에서 딥러닝 모델을 실행 시키는 것만으로 사용자(디바이스 내 다른 프로그램)가 모델에 접근할 수 없다고 가정한다. 가상화 기반 TEE의 경우 EL2에서 실행되는 하이퍼바이저를 신뢰한다. 즉, 하이퍼바이저 코드는 악의적이지 않으며, EL0에서 실행되는 프로그램들로부터 안전하다고 가정한다.



(그림 1) 낮은 권한(EL0) 프로그램이 하이퍼바이저 메모리를 접근하는 CVE-2022-22063[4] 취약점.

3. TEE의 신뢰 모델과 공격 기법

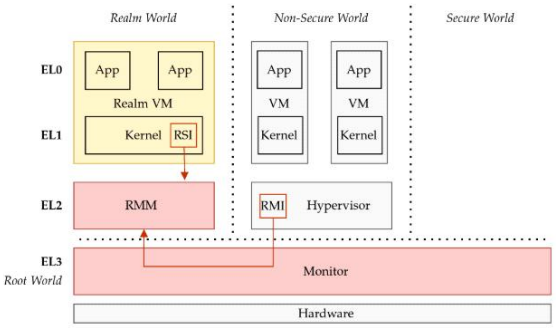
안타깝게도, 1장에서 언급하였듯이, TEE는 완전무결하지 않고, 기존 TEE 기반 딥러닝 보호 기법들이 가정한 TEE 신뢰 모델이 적합하지 않게끔 하는 공격들이 존재한다. 우선, Arm에서는 두 종류의 TEE 모두 TEE의 코드와 데이터를 모두 메모리에 평문(plaintext)로 저장한다. 이는 물리적인 공격자로부터 취약하다. 즉, 메모리에 물리적으로 접근하고 그 값을 읽을 수 있는 공격자들로부터는 TEE 데이터가 안전하지 않다. 이러한 물리 공격은 메모리 버스 스누핑, 메모리 인터포징, 콜드 부트 등의 다양한 방법으로 가능하다. 물리 공격이 TEE의 신뢰 모델에 포함되지 않은 이유는 일반적으로 TEE의 활용 시나리오가 클라우드 환경를 전제로 하기 때문이다. 즉, 데이터센터나 클라우드 서버의 경우 관찰해야 하는 메모리가 방대할 뿐만 아니라 관리 회사 차원에서 악의적인 관찰을 꼼꼼하게 관리할 수 있기 때문에 물리적인 공격이 실질적으로 어렵다. 하지만, 이와 반대로 온디바이스 환경에서는 소수의 인원이 개인적으로 디바이스를 사용하기 때문에 이러한 메

모리 관찰을 방지하기 어려우며, 이에 따라 물리 공격에 대한 위험성이 훨씬 커지게 된다.

뿐만 아니라 가상화 기반 TEE의 경우, 하이퍼바이저에 대한 공격이 이를 차용한 보호 기법의 신뢰 모델을 부적합하게 만든다. 일반적으로 더 낮은 권한 레벨에서 더 높은 권한 레벨을 얻는 것(privilege escalation)은 어려운 것으로 여겨지지만, 여러 연구들에서 그러한 공격이 실제로 가능하다는 것을 보여주고 있다. 예를 들어, 그림 1은 EL0에서 하이퍼바이저의 메모리에 접근을 가능토록 하는 CVE-2022-22063[4] 취약점을 보여준다. 이 취약점은 부트 리매퍼(boot remapper)라는 첫 64KiB 또는 128KiB 가상 주소를 설정 가능한 메모리 구역으로 매핑할 수 있게끔 하는 하드웨어 컴포넌트를 사용한다. 이 기능의 원래 목적은 첫 CPU가 부팅시 ROM 메모리를 읽다가 다른 CPU들이 부팅될 때 특정 메모리 구역(펌웨어 등)을 바로 읽어서 실행할수 있도록 만드는 것인데. 이를 부팅 이후에 활용하면 EL0에서 실행되는 일반 프로그램이 하이퍼바이저 메모리 영역을 접근 가능하다. 특히, 부트 리매퍼를 통한 메모리 접근은 2단계 주소 변환을 통하지 않게 때문에 하이퍼바이저 자체적으로 이에 대해 방어하지 못한다. 특히, 접근하는 하이퍼바이저 메모리 영역이 하이퍼바이저가 관리하는 2단계 페이지 테이블일 경우, 이를 조작하여 가상화 기반 TEE의 격리를 깰 수 있다. 온디바이스 딥러닝 보호를 위해 가상화 기반 TEE를 활용하는 ASGARD[3]의 경우, 격리를 제공하는 2단계 페이지 테이블을 관리하는 하이퍼바이저(TEEVisor)를 normal world의 EL2에서 실행하기 때문에 이러한 공격에 더욱 취약하다.

4. (잠재적) 해결 방안

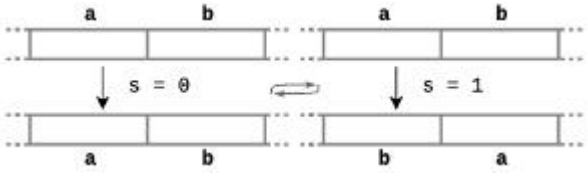
3장에서 설명한 공격 방향은 크게 물리 공격과 normal world에서 실행되는 하이퍼바이저에 대한 공격으로 분류된다. 이 두 공격으로부터 TEE를 보호할 수 있는 방안 중 하나는 Arm에서 최근 공개한 하드웨어 기반 TEE인 Confidential Computer Architecture (CCA)를 사용하는 것이다. CCA는 normal, secure 외에 새로운 실행 환경인 Realm을 제공하며, 가상화를 기반으로 TEE를 가상 머신 단위로 제공한다(그림 2). 이 때, 3장에서 설명했던 가상화 기반 TEE와는 다르게 CCA는 Realm world의 EL2에서 실행되는 하이퍼바이저인 Realm Management Monitor (RMM)이 2단계 페이지 테이블을 관리한다. 이와 더불어 CCA는 TEE VM 메모리를 암호화해서 저장하는 Memory Encryption Context(MEC)를 제공하며, 각 VM은 서로 다른 키를 사용하여 메모리를 암호화한다. 이러한 CCA의 특징은 앞서 설명한 TEE들이 지니고 있던 취약점을 해결한다. TEE 데이터가 암호화해서 메모리에 저장되기 때문에, 공격자가 메모리 트래픽을 모니터링하더라도 데이터에를 유출 시킬 수 없다. 또한, TZASC를 통해 secure world 메모리가 normal world에서 접근 불가하였듯이, normal world에서 Realm world가 사용하는 메모리를 하드웨어가 접근할 수 없게끔 만들기 때문에 Realm world 하이퍼바이저가 관리하는 2단계 페이지 테이블도 normal world 공격자로부터 안전하다.



(그림 2) Arm Confidential Compute Architecture

하지만 안타깝게도 최근 연구들은 CCA조차 온디바이스 딥러닝 보호에 일차원적으로 차용되었을 경우 모델 프라이버시 유출을 막기에 부족하는 것을 보여준다. 그 이유는 CCA의 메모리 암호화 알고리즘에 있다. CCA의 암호화는 다른 하드웨어 기반 가상머신 단위 TEE인 Intel Trusted Domain Extension이나 AMD Secure Encrypted Virtualization과 같은 암호화 알고리즘을 사용할 것으로 예상된다. 이 때 암호화 속도 및 메모리 최적화를 위하여 XEX 기법이 사용되는데, 이는 동일한 평문이 동일한

```
ct-swap(array a, array b, secret_bit s):
  1: mask = ~(s – 1);
  2: delta = mask & (a^b);
  3: a = a ^ delta;
  4: b = b ^ delta;
```



메모리 위치에 저장될 시 같은 암호문으로 저장된다는 특징을 가지고 있다. 이러한 특성을 이용하여 암호문 부채널 공격 (ciphertext side channel)이 가능하다. 예를 들어, 아래와 같이 *secret_bit s*에 따라 *a*와 *b*의 값이 바뀌는(swap)되는 함수가 있다. 이 때 공격자는 (암호화된) 메모리 값이 바뀌는지를 하여 *secret_bit*의 값이 0인지 1인지 확인할 수 있으며, 이를 활용하여 TEE 내부에서 실행되는 딥러닝 모델의 입력값을 추출하는 공격[5]까지 가능한 것으로 보고되고 있다.

5. 결론

본 논문에서는 신뢰실행환경 기반 온디바이스 딥러닝의 취약점들 및 그 해결 방안을 분석했다. 특히, 기존 TEE 기반 딥러닝 모델 보호 기법들의 하드웨어 기반과 가상화 기반의 TEE 신뢰 모델들을 분석하고, 이들이 모델 프라이버시를 보호하는데 불충분함을 공격 방법과 함께 설명했다. 이와 더불어 이를 막기 위한 잠재적 방안으로 Arm의 새로운 하드웨어 기반 TEE인 Confidential Compute Architecture가 기존 TEE 취약점을 어떻게 해결하는지 논의하고, 그 한계점인 암호문 부채널 공격에 대해 논의했다. 이에 따라, 온디바이스 딥러닝 보호의 특성을 고려하여 TEE를 추가적으로 가능한 공격들로부터 보호하는 연구가 계속되어야 할 것이다.

사사

이 논문은 2025년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의하여 지원되었으며, 2025년도 정부(과학기술정보통신부)의 재원으로, 정보통신기획평가원의 지원을 받아 수행된 연구(No.2021-0-00528, 하드웨어 중심 신뢰계산기반과 분산데이터보호박스를 위한 표준 프로토콜 개발, No.RS-2024-00406121, 자동차 보안 취약점 기반 위협 분석 시스템 개발(R&D), No.RS-2024-00438729, 익명화된 기밀실행을 이용한 전주기적 데이터 프라이버시 보호 기술 개발, IITP-2023-RS-2023-00256081)이자 한국연구재단의 지원을 받아 수행된 연구(RS-2023-00277326)이고, 반도체 공동연구소 지원의 결과물임을 밝힘.

참고문헌

[1] Fan Mo, et. al., “DarkneTZ: Towards Model Privacy at the Edge using Trusted Execution Environments”, Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services, 2020

[2] Zhichuang Sun, et. al., “ShadowNet: A Secure and Efficient On-device Model Inference System for Convolutional Nerual Networks”, 2023 IEEE Symposium on Security and Privacy (SP)

[3] Myungsuk Moon, et. al., “ASGARD: Protecting On-Device Deep Nerual Networks with Virtualization-Based Trusted Execution Environments”, Network and Distributed System Security (NDSS) Symposium 2025

[4] National Institute of Standards and Technology (NIST), “CVE-2022-22063”, https://nvd.nis.gov/vuln/detail/CVE-2022-22063“

[5] Yuanyuan Yuan, et. al., “CipherSteal: Stealing Input Data from TEE-Shielded Neural Networks with Ciphertext Side Channels”, 2025 IEEE Symposium on Security and Privacy (SP)