클라우드 기반 머신러닝 서비스 보안 프레임워크

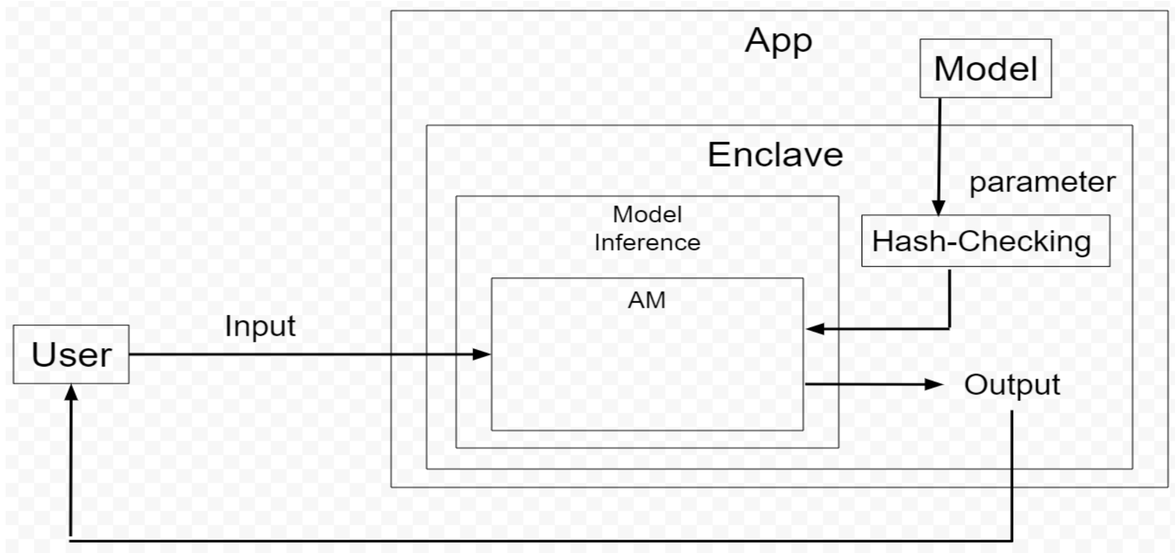
Security Framework for ML Service based on Cloud

개 요

AI 기술이 인간의 삶에 깊이 파고든 현 상황에서 우리가 사용하는 AI 기술이 적용된 애플리케이션 그 중에서 많은 부분을 차지하고 있는 클라우드 기반 애플리케이션의 보안은 그 중요도가 높다고 할 수 있습니다. 보안 대책이 미흡한 클라우드 서비스는 서비스를 제공받아 얻는 이익만을 생각하기에는 보안사고로 인한 피해가 막대할 수 있습니다.

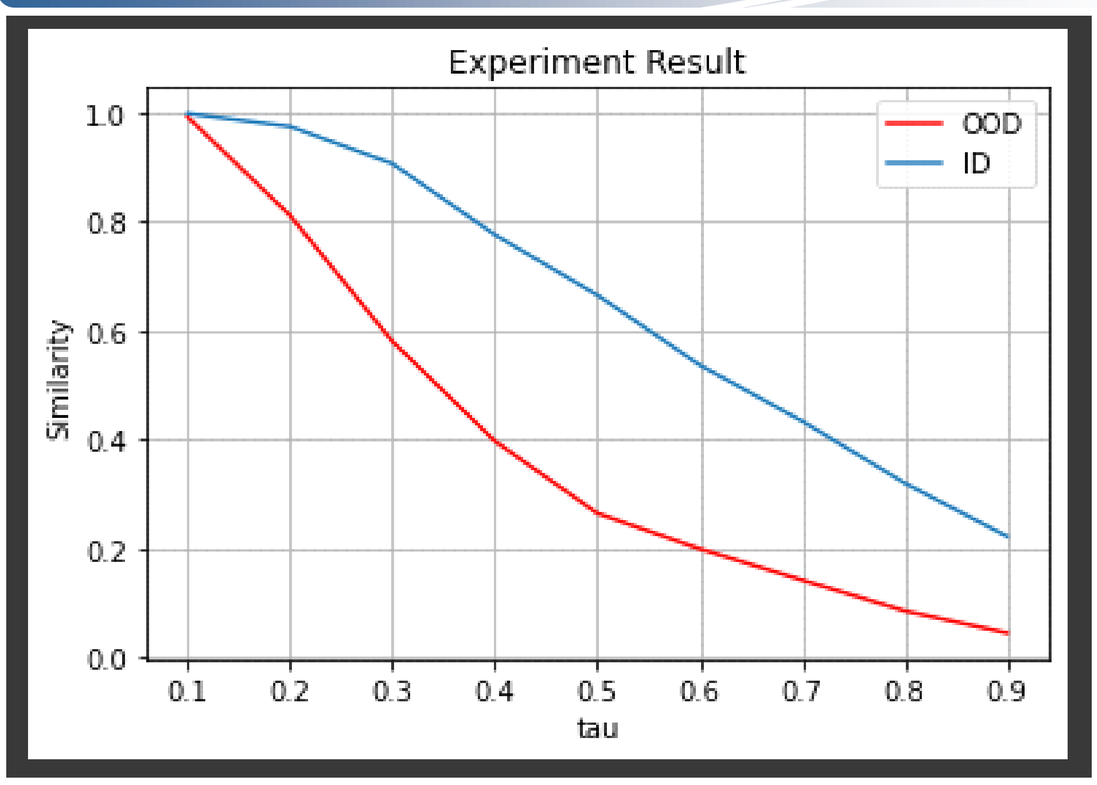
이를 위해 본 논문에서는 클라우드 기반 AI 서비스를 분석하여 어떤 공격이 이루어질 수 있는지 분석하고 그에 대한 연구된 방어법들의 효과를 확인하여 효과적인 것들을 선별하고 접목시키는 시도를 합니다.

시스템 구성



AM	**· surrogate data를 이용한 model stealing attack에 대한 방어** **· Enclave 내부에서 input이 Out-of-distribution인지 판단**
Enclave	**· Intel SGX를 이용한 강력하게 보호되는 메모리 공간** **· 크기가 작아 Model을 외부에 두고 on-demand로 load**
Hash-Checking	**· load된 parmeter의 무결성을 확인하기 위한 장치**
Progress	**1) input값을 주면 Enclave 내부에 저장** **2) Enclave 외부에서 원래 모델 f의 첫 번째 parameter block을 load** **3) load된 block hash-check** **4) 원래 모델 f의 마지막 block까지 2)~4)과정을 반복하여 f(x) 계산** **5) AM 알고리즘의 OOD Detector에 의해서 alpha값을 결정** **6) mislabel 시키는 모델 f'에 대하여 2)~4) 반복, f'(x)값을 결정** **7) output으로 (1-alpha)*f(x) + alpha*f'(x) 를 User에게 전달**

실험 및 분석



실험 과정

1) input x에 대해 원래 모델 f의 output f(x) 계산

2) input x에 대해 본 논문이 제시한 프레임워크를 거쳐 나온 output 계산

3) 1)와 2)의 output의 유사도 계산

4) 1) ~ 3) 과정을 OOD Detector가 OOD와 ID를 구별하는 기준값인 τ값을 [0.1,0.2,...,0.9] 로 다르게 설정하여 반복 실행하고 유사도를 그래프로 나타냄

분석

그래프를 보면 τ값이 작을 수록 OOD input을 판단하는 능력이 떨어지지만, τ값이 커질수록 ID input을 판단하는 능력도 떨어지게 되는 것을 확인. 따라서 적절한 τ값의 선택이 프레임워크의 성능을 높이는데 가장 중요한 요소.

해당 그래프에서는 τ가 0.3일 때 ID input에 대해서는 유사도 90.60%, OOD input에 대해서는 유사도 57.96%로, 원래 모델의 성능을 크게 저하시키지 않으면서 attacker의 공격 능력은 크게 저하시키는 가장 적절한 임계치라고 분석됨.

결론

여러 보안 알고리즘의 연구에 의해 제시된 보안 프레임워크는 그 성능이 점점 발전할 수 있고, 그 점에서 본 논문의 연구 성과를 찾을 수 있다고 생각합니다. 반면 프로세스 과정이 점점 늘어난다면 그만큼 속도 면에서 손해를 볼 수 있습니다. 본 논문에서는 프레임워크의 보안적인 측면에 초점을 맞춰 프로세스를 구성하고 그 효과에 대한 실험을 진행하였지만, 머신 러닝 서비스의 특성상 속도 또한 굉장히 중요한 요소이므로 이 부분에 대한 개선이 필요할 것입니다.