클라우드 기반 머신러닝 서비스 보안 프레임워크

서한결*, 강동윤*
*성균관대학교 컴퓨터공학과
shy7338@naver.com, ehddbs1231@naver.com

A Security Framework for ML service based on Cloud

Han-Gyeol Seo*, Dong-Yoon Kang*
*Dept. of Computer Engineering, Sungkyunkwan University

요 약

AI 모델 서비스 제공에 강제되는 높은 메모리 사용량을 해결하기 위해 일반적으로 클라우드 컴퓨팅 기술을 이용한다. 클라우드 기반 서비스는 개발자로 하여금 메모리 사용량에 대한 걱정을 덜어주고 서비스 이용자에게는 편리하게 양질의 서비스를 제공받을 수 있게 한다. 하지만 보안 대책이 미흡한 클라우드 서비스는 서비스를 제공받아 얻는 이익만을 생각하기에는 보안사고로 인한 피해가 막대할 수 있다. AI 기술이 인간의 삶에 깊이 파고든 현 상황에서 우리가 대부분 이용하는 클라우드에 기반 서비스의 보안 문제는 그 중요도가 굉장히 높다고 할 수 있다. 이를 위해 본 논문에서는 클라우드 기반 머신러닝 서비스를 분석하여 어떤 공격이 이루어질 수 있는지 분석하고 그에 대한 연구된 방어법들의 효과를 확인하여 효과적인 것들을 선별하고 접목시키는 시도를 한다.

1. 서론

인공지능(AI)에 대한 활발한 연구를 통해 비약적인 기술 발전이 이루어졌다. 여러 기업들은 모델을 스스로 훈련(train)시킬 능력이 부족한 사용자에게 기업의 모델을 서비스로 제공하는데, 모델을 서비스하는 과정에서 많은 메모리 사용은 불가피하기 때문에 기업에서는 클라우드 컴퓨팅 기술로 이를 해결하려고 한다. 하지만 클라우드 서비스를 이용함으로써 여러 보안 문제가 발생할 수 있다 [9]. 모델을 복사(copy)하여 비슷한 성능의 모델을 만들어 기업의 경제 활동에 피해를 입힐 수 있고, 모델의 파라미터(parameter)를 손상시키거나 원래의 모델에 악의적인 입력 값을 넣음으로써 모델의 성능을 떨어뜨릴 수도 있다.
클라우드 기반 머신러닝 서비스는 AI 모델 자체가 보유한 보안 취약점과 클라우드를 이용함으로써 발생할 수 있는 보안 문제 두 가지 모두에 대응할 수 있어야 한다. 본 논문은 현재까지 연구된 방어법들 중 몇 가지를 선별하고 접목시켜 AI 모델이 갖는 취약점과 클라우드의 보안 문제점 일부를 동시에 방어할 수 있는 하나의 보안 프레임워크를 제시하고자 한다.
본 논문은 구성은 다음과 같다. 우선 2장에서 본 논문이 중점을 둔 클라우드 기반 머신러닝 서비스가 가질 수 있는 취약점에 대해 설명하고 3장에서 그에 대해 연구된 방어법들을 소개한다. 4장에서는 본 논문에서 제시하는 여러 방어법들을 접목시킨 프레임워크를 소개하며 5장에서 그 효과를 실험을 통해 확인하고 결과를 분석하며 마무리한다.

2. 공격법(Attacks)

2.1 모델 하이퍼파라미터 탈취 공격(Model Hyperparameter Stealing Attack)

모델 하이퍼파라미터 탈취 공격 [1]이란 공격 대상 AI 모델 정보와 그 하이퍼파라미터를 탈취하는 공격으로, 모델의 정보를 도용하게 되면 학습자의 지적 재산권과 알고리즘의 기밀성을 저하시킬 수 있고, 이후 공격자가 회피 공격(evasion attack), 반전 공격(inversion attack) 등 다른 공격을 이어서 진행 할 수도 있다. 그 중에서 특히 본 논문은 유사 데이터(surrogate data)를 이용하여 모델을 복사(copy)하는 공격[2]에 집중한다. 해당 공격이 입력에 대한 결과 값만으로도 이루어 질 수 있다는 점이 클라우드 기반 머신 러닝 서비스의 특징을 잘 활용한 공격으로 볼 수 있기 때문이다. 강아지 이미지를 입력으로 주었을 때 해당 강아지의 품종을 분류하는 모델을 예로 들어보자. 만약 강아지가 아닌 고양이 사진을 입력으로 넣는다면 모델은 낮은 확신을 갖고 특정 품종으로 분류할 것이다. 이때, 고양이 이미지를 유사 데이터

(surrogate data)라고 한다. 대량의 고양이 이미지를 입력으로 넣는다면 모델이 분류한 결과값으로 품종이 정해진 고양이 데이터가 새로이 만들어질 것이다. 해당 데이터를 훈련 데이터로 하여 새로운 모델을 훈련시킨다면 원래의 강아지 품종 분류 모델과 유사한 성능을 갖는 모델을 만들어 낼 수 있다는 것이 알려져 있다.[2]

2.2 개인 정보 탈취 공격(Private Data Stealing Attack)

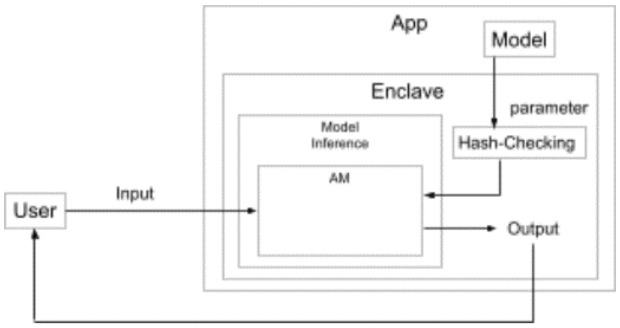
서비스를 이용하는 사람들의 개인 정보를 탈취하는 공격이다. 서비스 이용자는 해당 서비스를 제공하는 기업을 신뢰하고 개인 정보를 입력하지만, 해당 정보는 완벽히 안전하다고는 할 수 없다. 외부 공격자에 의해 도난 당할 가능성도 있지만, 가장 경계해야할 것은 서버 관리자가 악의를 품는 것이다. 일반적으로 서버 관리자는 서비스 이용자들의 개인 정보에 쉽게 접근이 가능할 것이다. 따라서 서버 관리자를 신뢰하지 못한다면 해당 서비스를 이용하는 것은 힘들 것이다[5].

2.3 모델 하이퍼파라미터 오염 공격(Model Hyperparameter Poisoning Attack)

모델의 하이퍼파라미터 값을 개발자의 의도와는 다르게 변경시켜 모델의 성능을 저하시키거나 공격자들의 의도대로 결과 값이 나오게 만드는 공격이다. 2.2 개인 정보 탈취 공격과 마찬가지로 외부 공격자와 서버 관리자에 의해 이루어 질 수 있다.[5]

3. 프레임워크(Framework)

본 논문이 제시하는 프레임워크는 앞서 3장에서 언급한 방어법들을 접목시킨 것이다. 전체 과정은 (그림 2)와 같다.



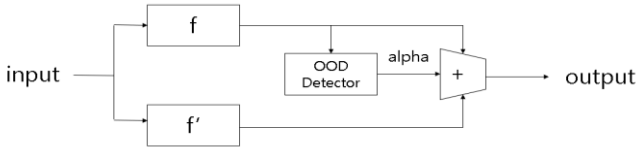
(그림 2) 프레임워크

(그림 2)를 보면, 이용자가 입력 값을 넣으면 바로 Enclave로 들어가서 개인 정보를 보호한다. 곧바로 외부에 있는 파라미터를 Enclave 내부로 load 하는데, 이 때 파라미터의 무결성을 확인하기 위해 해시 체크를 수행한다. 그 다음, Adaptive Misinformation 알고리즘을 이용하여 입력 데이터가 유사 데이터인지 여부를 판단하고 만약 입력 데이터가 분포 내 데이터라면 원래 모델 f에 의한 결과를 출력 값으로, 입력 데이터가 분포 외 데이터라면 허위 정보를 만드는 모델 f' 에 의한 결과를 출력 값으로 내보내 유사 데이터를 통한 모델 하이퍼파라미터 탈취 공격을 방어한다.

4. 방어법(Defenses)

4.1 조정적 허위 정보 주입(Adaptively Injecting Misinformation)

첫 번째 취약점인 유사 데이터를 이용한 모델 하이퍼파라미터 탈취 공격에 대한 방어법으로 참고논문 [2]에서 제시한 조정적 허위 정보 주입을 자세히 소개한다.



(그림 1) Adaptive Misinformation 알고리즘

AM(Adaptive Misinformation)은 (그림 1)과 같은 구조를 띄고 있다. 가장 처음 입력 x에 대해서 분포 외 데이터 탐지기가 해당 입력이 분포 내 데이터(In Distribution)인지 분포 외 데이터(Out Of Distribution)인지 원래 모델 f의 결과값 f(x)를 보고 판단한다. 만약 분포 내 데이터라면 원래 모델 f에 의한 예측을 결과값(output)으로, 분포 외 데이터라면 허위 정보를 만들어 내는 모델 f' 에 의한 예측을 결과값으로 준다.

4.1.1 분포 외 데이터 탐지기(OOD Detector)

분포 외 데이터 탐지기는 조정적 허위 정보 주입 알고리즘 핵심 아이디어이다. f(x) 값을 소프트맥스 확률(Softmax Probability)로 나타냈을 때, 그 중 예측값이 될 최댓값이 특정값(임계치) 이상인지 아닌지를 확인함으로써 입력 x가 분포 외 데이터인지 아닌지 판단한다. 소프트맥스 확률의 최댓값은 해당 값의 색인(index)이 예측값이 될 확률로 볼 수도 있는데, 그 확률이 낮을수록 입력값이 분포 외 데이터일 확률

은 높아진다는 사실로부터 나온 아이디어이다. 판단 결과를 값 alpha로 반환 하는데, 이 alpha값은 3.1.3. 장에서 설명할 것이다.

4.1.2 허위 정보 생성 모델 f'

f' 은 분포 외 데이터 입력에 대해 허위 정보를 생성시키기 위한 모델로써 보통 정답을 맞출 확률을 높이는 훈련에서 사용하는 크로스 엔트로피 손실 함수(cross entropy loss function) [식 1] 과 달리 [식 2]를 사용하여 정답을 맞출 확률을 낮추는 훈련을 시킨다.

$$loss = -\sum_{i} t_i log(s_i)$$

[식 1]

$$loss = -\sum_{i} t_i log(1 - s_i)$$

[식 2]

* ti는 i번째 정답, si는 예측값을 소프트맥스 확률로 나타낸 것의 i번째 요소

4.1.3. 결과값(output)

마지막 결과값은 분포 외 데이터 탐지기에서 판단한 결과를 바탕으로 분포 내 데이터라면 f(x), 분포 외 데이터라면 f' (x)로 결정되어야 한다. 이를 반영한 식은 [식 3]과 같다.

$$output = (1 - \alpha) * f(x) + \alpha * f'(x)$$

$$\alpha = S(ymax - \tau)$$

$$S(z) = \frac{1}{1 + e^{\nu z}}$$

[식 3]

S(z)는 역 시그모이드 함수(reverse sigmoid function)로 z 값이 0보다 크면 0.5보다 작은 값을 반환하고 0보다 작으면 0.5보다 큰 값을 반환한다. 따라서 분포 외 데이터 탐지기에서 ymax - τ(임계치) 값을 확인하여 결정된 α 값으로 인해 ID일 경우 (ymax - τ > 0 일 때, α -> 0 이면 output -> f(x)) f(x) 값이, OOD일 경우 (ymax - τ < 0 일때, α -> 1 이면 output -> f' (x)) f' (x)값이 output이 된다.

4.2 Intel SGX를 이용한 개인 정보 보호(Private Data Preserving By Using Intel SGX)

두 번째 취약점인 개인 정보 탈취 공격에 대한 방어법으로 참고 논문 [5]에서 제시한 Intel SGX를 이용한 개인 정보 보호를 소개하겠다. Intel SGX(Software Guard Extensions)는 Intel CPU에 적용된 코드 및 데이터를 메모리내에 격리하는 하드웨어 기반 보안 기술이다. Enclave라고 하는 private memory를 할당하여 해당 메모리 사용자를 제외한 다른 그 어떤 누구도 접근할 수 없도록 설계되어 있다. Intel SGX를 이용해 신뢰실행환경(TEE) 환경을 구축하고 공격으로부터 강하게 보호되는 Enclave 공간을 할당하여 그 안에서 모델 추론을 실행하는 방식이다. 하지만 Enclave를 만들 수 있는 공간의 크기는 크지 않기 때문에 모델의 파라미터를 보호하지 않기로 결정하고 이를 Enclave 외부에 두어 공간의 사용량을 줄였다. 그 후 필요한 파라미터만 그 때 즉시 on-demand 형식으로 Enclave에 load 하는 방법을 사용하였다. 하지만 모델의 파라미터를 보호하지 않기 때문에 모델이 오염될 수 있고, 이로 인해 오염된 파라미터가 load 되는 문제가 발생할 수 있다.

4.3 해시 체크를 통한 하이퍼파라미터 오염 감지(Detecting Hyperparameter Poisoning By Hash-Checking)

세 번째 취약점인 모델 하이퍼파라미터 오염 공격에 대한 방어법으로 해시 체크(Hash-checking)을 제안한다[5]. Enclave 내부에 기존에 계산된 Hash table을 두어 load된 파라미터를 지정된 해시 함수로 연산하여 비교한다면 파라미터의 무결성을 확인할 수 있을 것이다. 3.2장 미지막에서 설명한 두 번째 방어법의 취약점 또한 이 방법을 이용하여 대응할 수 있다.

5. 구현 및 결과 분석

5.1. 실험 모델 및 결과

AM을 실행해볼 모델로 사용하기 위해 kaggle dog-breed-identification (https://www.kaggle.com/c/dog-breed-identification/data)에서 데이터를 가져와 추려서 6898개의 강아지 이미지로 80개의 품종으로 분류하는 모델을 훈련시켰다. Pytorch를 이용한 코드를 통해 전체 6898개의 이미지 중 5%인 345개의 validation set을 대상으로 정확도(accuracy) 82.6%인 모델을 만들어낼 수 있었다.

원본 모델에서는 손실 함수를 torch.nn.CrossEntropy를 바로 사용하였지만, mislabel 모델을 구현할 때는 CrossEntropy가 LogSoftmax와 NLLLoss의 combine임을 고려하여 loss function을

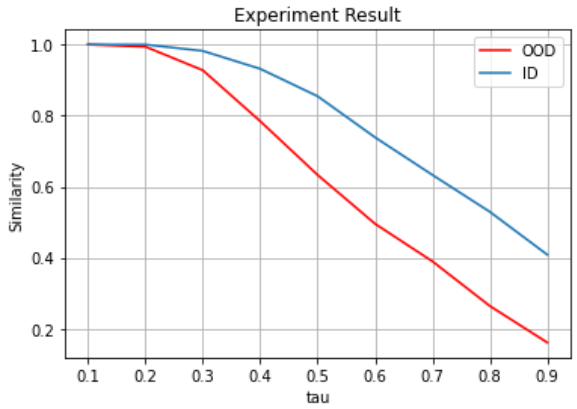
x=torch.nn.Softmax, torch.log(1-x), torch.nn.NLLLoss의 순차적방식으로 구현하였다.

```
train_loss: 0.0231, val_loss: 0.7242, val_acc: 0.8261
Finish Training.
```

(그림 3) 원본 모델의 훈련

```
train_loss: 0.0019, val_loss: 0.0000, val_acc: 0.0000
```

(그림 4) mislabel 모델의 훈련



(그림 5) 유사도 그래프

(그림 5) 임계치 τ의 값에 따라 OOD input과 ID input을 넣었을 때, AM을 적용했을 경우와 하지 않았을 경우 각각의 유사도 그래프. test image는 kaggle cat and dog (https://www.kaggle.com/tongpython/cat-and-dog)을 사용하였다.

5.2. 분석

(그림 5)의 그래프를 보면 τ값이 작을 수록 분포 외 데이터를 판단하는 능력이 떨어지지만, τ값이 커질수록 분포 내 데이터를 판단하는 능력도 떨어지게 되는 것을 확인할 수 있다. 따라서 적절한 τ값의 선택이 AM 알고리즘의 성능을 높이는데 가장 중요한 요소라고 할 수 있다. 해당 그래프에서는 τ=0.5가 분포 내 데이터에 대해서는 유사도 85.46%, 분포 외 데이터에 대해서는 유사도 63.30%로, 원래 모델의 성능을 크게 저하시키지 않으면서 공격자의 공격 능력은 크게 저하시키는 가장 적절한 임계치라고 분석된다.

6. 결론

본 논문에서는 클라우드 기반 머신러닝 서비스의 여러 보안 문제점들을 조사하고 여러 문제점들을 동시에 해결 가능한 보안 프레임워크를 제시했다. 5장에서 실험한 결과를 바탕으로 해당 프레임워크는 타겟으로 한 공격법을 효과적으로 방어할 수 있다는 사실을 검증하였다. 해당 프레임워크는 Enclave 내부에 단순히 프로세스를 추가함으로써 다른 방어법들을 추가 적용하는데 용이하다는 강점을 가지고 있다. 반면, Intel SGX가 GPU와 연동하지 못해 속도가 느리다는 결점이 있어 추후 개선이 필요하다.

참고문헌

[1] B. Wang and N. Z. Gong, “Stealing Hyperparameters in Machine Learning,” in IEEE 2018.

[2] S. Kariyappa and M. K. Qureshi, “Defending Against Model Stealing Attacks with Adaptive Misinformation,” in CVPR 2020.

[3] 이슬기, 김경한, 김병익, 박순태, “기계학습 모델 공격연구 동향: 심층신경망을 중심으로,” 정보보호학회지 2019.

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and Harnessing Adversarial Examples”, in ICLR 2015.

[5] T. Lee, Z. Lin, S. Pushp, C. Li, Y. Liu, Y. Lee, F. Xu, C. Xu, L. Zhang, J. Song, “Occlumency: Privacy-preserving Remote Deep-learning Inference Using SGX”, in ACM ISBN 2019.

[6] P. Samangouei, M. Kabkab, R. Chellappa, “Defense-GAN: protecting classifiers against adversarial attacks using generative models”, in ICLR 2018.

[7] 전상기, 최창준, 이종혁 (2017). Trusted Execution Environment의 구성 요소 및 응용 기술 조사. 한국통신학회 학술대회논문집, 65-66

[8] 류권상, 최대선, “인공지능 보안 공격 및 대응 방안 연구 동향”, 정보보호학회지 2020.

[9] C. M. R. d. Silva, J. L. C. d. Silva, R. B. Rodrigues, G. M. M. Campos, L. M. d. Nascrimento and V. C. Garcia, “Security Threats in Cloud Computing Models: Domains and Proposals,” in IEEE 2013.