J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

연합학습에서의 손실함수의 적응적 선택을 통한 효과적인 적대적 학습

Effective Adversarial Training by Adaptive Selection of Loss Function in Federated Learning

이 수 철[1*]
Suchul Lee

요 약

연합학습은 보안 및 프라이버시 측면에서 중앙 집중식 방법보다 안전하도록 설계되었음에도 불구하고 여전히 많은 취약점을 내재한다. 적대적 공격(adversarial attack)을 수행하는 공격자는 신중하게 제작된 입력 데이터, 즉 적대적 예제(adversarial examples)를 클라이언트의 학습 데이터에 주입하여 딥러닝 모델을 의도적으로 조작하여 오분류를 유도한다. 이에 대한 보편적인 방어 전략은 이른바 적대적 학습(adversarial training)으로 적대적 예제들의 특성을 선제적으로 모델에 학습시키는 것이다. 기존의 연구에서는 모든 클라이언트가 적대적 공격 하에 있는 상황을 가정하는데 연합학습의 클라이언트 수가 매우 많음을 고려하면 실제와는 거리가 있다. 본 논문에서는 클라이언트의 일부가 공격 하에 있는 시나리오에서 적대적 학습의 양상을 실험적으로 살핀다. 우리는 실험을 통해 적대적 예제에 대한 분류 정확도가 증가하면 정상 샘플에 대한 분류 정확도의 감소하는 트레이드오프 관계를 가짐을 밝혔다. 이러한 트레이드오프 관계를 효과적으로 활용하기 위해 클라이언트가 자신이 공격받는지 여부에 따라 손실함수를 적응적으로 선택하여 적대적 학습을 수행하는 방법을 제시한다.

☞ 주제어 : 연합 학습, 적대적 공격, 적대적 학습, 트레이드오프

ABSTRACT

Although federated learning is designed to be safer than centralized methods in terms of security and privacy, it still has many vulnerabilities. An attacker performing an adversarial attack intentionally manipulates the deep learning model by injecting carefully crafted input data, that is, adversarial examples, into the client's training data to induce misclassification. A common defense strategy against this is so-called adversarial training, which involves preemptively learning the characteristics of adversarial examples into the model. Existing research assumes a scenario where all clients are under adversarial attack, but considering the number of clients in federated learning is very large, this is far from reality. In this paper, we experimentally examine aspects of adversarial training in a scenario where some of the clients are under attack. Through experiments, we found that there is a trade-off relationship in which the classification accuracy for normal samples decreases as the classification accuracy for adversarial examples increases. In order to effectively utilize this trade-off relationship, we present a method to perform adversarial training by adaptively selecting a loss function depending on whether the client is attacked.

☞ keyword : Federated learning, adversarial attacks, adversarial training, trade-off relationship

1. 서 론

딥러닝은 대량의 데이터로부터 인공 신경망을 활용해 자동으로 특징을 추출하여 학습함으로써 다양한 범주의 CS(computer science)문제에 대한 효과적인 해결 도구로 계속해서 주목을 받아왔다. 통상적으로 딥러닝 모델의 학습은 중앙 집중식 클라우드에 클라이언트의 데이터를 일원화하여 고성능의 서버 컴퓨팅 자원을 활용하여 효율성이 극대화 될 수 있도록 구현된다[1].

그러나 raw데이터의 중앙 집중화는 보안 및 개인정보의 유출 문제와 데이터의 전송을 위한 통신 대역폭 확보 등의 선결과제를 필연적으로 도출한다[2]. 예를 들어 셀룰러 망의 경우 스마트폰 사용자의 앱 설치 및 이용현황, 웹사이트 전송기록, 연락처, 멀티미디어 데이터가 클라우

1 Dept. of Computer Science and Information Engineering, Korea National University of Transportation, Uiwang, Kyunggi, 16106, Korea.
* Corresponding author (sclee@ut.ac.kr)
[Received 31 December 2023, Reviewed 9 January 2024(R2 22 February 2024, R3 27 March 2024), Accepted 28 March 2024]

드로 전송될 수 있다. 이는 대역폭을 소모할 뿐만 아니라 심각한 개인정보 유출의 원인이 되며, 개인정보 유출에 대한 우려로 스마트폰 사용자가 데이터 수집과정으로부터 이탈할 수 있다. 심지어 공격자는 raw데이터의 수집과정에서 의도적으로 데이터의 무결성을 손상시킴으로써 모델 학습을 방해할 수 있다[3][4].

연합학습(federated learning)에서는 각 학습 라운드마다 클라이언트가 딥러닝 모델 학습을 위해 raw데이터를 클라우드로 전송하는 대신 분산 방식으로 각각의 로컬 모델을 학습한다. 학습된 로컬 모델은 클라우드로 전송되며 모델 집계(model aggregation)과정을 거쳐 글로벌 모델에 통합된다. 글로벌 모델은 각 클라이언트로 배포되어 다음 라운드의 로컬 학습의 시점이 된다. 이 과정을 반복함으로써 클라이언트는 분산적으로 글로벌 모델 학습에 참여한다. 연합학습에서는 클라이언트의 raw데이터가 외부와 공유되지 않고 고정 크기의 모델의 매개변수만 전송하면 되기 때문에 개인정보 유출문제로부터 어느 정도 자유로울 수 있을 뿐만 아니라 통신에 필요한 대역폭도 감소한다.

기존 연합학습 관련 연구는 모델 학습의 최적화에 초점을 맞추어 수행되어 왔다. [5]에서는 클라이언트 로컬 데이터의 분포가 학습이 미치는 영향에 대해 실험적인 연구를 수행하였다. FedAVG는[6] 현재 가장 보편적으로 사용되는 모델 집계 방법이며 이종 데이터 하에서의 FedAVG를 개선한 FedPROX[7], Scaffold[8] 등이 제안되었다. 이 두 가지 모델 집계 방법은 글로벌 모델의 수렴에 소요되는 학습 라운드의 수를 줄여 모델을 전송하는데 필요한 통신량을 경감하였다.

연합학습은 본질적으로 보안 및 프라이버시 측면에서 중앙 집중식 학습방법보다 안전하도록 설계되어 있음에도 불구하고 여전히 많은 취약점을 가지고 있다. 예컨대 포이즈닝(poisoning) 공격을 들 수 있다. 포이즈닝은 학습 데이터를 왜곡하는 데이터 포이즈닝[9][10]과 학습 모델을 왜곡하는 모델 포이즈닝[11]으로 구분할 수 있다. 공격하의 클라이언트를 비잔틴(Byzantine) 노드로 간주 할 수 있으며 비잔틴 노드의 왜곡된 모델은 글로벌 모델로 집계되는 과정에서 차등 반영되거나 격리(quarantine)시키는 등의 다양한 방어 전략을 취할 수 있다.

강력한 주요 공격의 다른 유형은 로컬 데이터에 대한 GAN(Generative Adversarial Network)[3]기반의 적대적 공격(adversarial attack)이다. 적대적 공격의 공격 대상은 포이즈닝 공격과 유사할지 모르나 대응에 있어서는 전혀 다른 접근이 활용된다. 이는 선제적으로 적대적 예제(adversarial examples)의 특성을 학습하는 방법인데 이른바 적대적 학습(adversarial training)으로 불린다. 우리가 알고 있는 범위 내에서 문헌상 최초로 적대적 학습을 연합학습 환경에 적용한 연구는 FAT(Federated Adversarial Training)[12]이다. 적대적 학습에서는 상대적으로 유연한 min-max기반 손실함수(loss function)를 사용되는데 연합학습 환경에서 이러한 손실함수를 사용하면 클라이언트 간 데이터의 이질성을 악화시킬 수 있다. 학습 라운드가 어느 정도 진행되면 악화된 데이터의 이질성의 여파로 성능이 급감하는 현상이 관찰된다. [13]에서 이러한 현상을 완화하는 글로벌 모델 집계 방법인 SFAT(slacked-FAT)이 제안되었다.

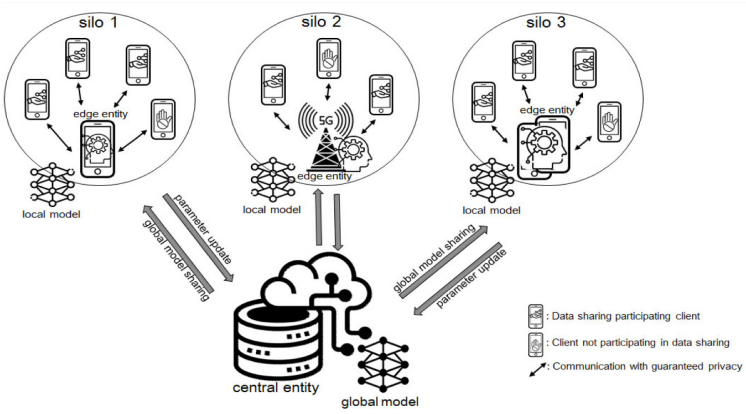
기존의 연합학습 환경에서의 적대적 학습에 관한 연구에서는 모든 클라이언트가 적대적 공격을 받는 시나리오를 가정한다. 그러나 연합환경에 수많은 클라이언트가 참여한다는 점을 감안하면 공격자가 모든 클라이언트를 특정할 수 있다는 가정은 자연스럽지 않다. 또한, 문제해결의 관점에서 일부 클라이언트가 공격 하에 있는 시나리오는 적대적 학습을 수행하기에 더욱 도전적인 상황이 된다.

본 논문에서는 보다 실제적인 상황 즉, 클라이언트의 일부가 공격을 받는 상황에서 적대적 공격이 딥러닝 학습의 성능에 대한 미치는 영향을 실험적으로 살펴본다. 실험결과에 따르면 '적대적 공격 하에서의 분류 정확도'와 '공격받지 않는 상황에서의 분류 정확도'간에는 상호 트레이드오프 관계에 있다. 우리는 각 클라이언트가 공격을 받고 있는지 여부에 따라 전략적으로 손실함수를 선택하여 이러한 트레이드오프 관계를 효과적으로 이용할 수 있는 적대적 학습 방법을 제안한다. CICMaldroid 2020 데이터셋을 활용한 초기 실험을 통해 우리는 연합학습의 클라이언트가 적대적 공격 하에서 제안하는 방법을 사용하여 효과적으로 적대적 훈련을 수행함을 보였다.

본 논문의 구성은 다음과 같다. 2장에서 관련 연구를 살펴본다. 3장에서 우리가 고려하는 시스템 및 공격 모델을 기술한다. 4장에서 실험환경을 서술하고 5장에서 결과를 제시한다. 끝으로 6장에서 본 논문을 맺는다.

2. 관련 연구

연합학습은 2017년 McMahan *et al.*의 연구[6]에서 모델 집계의 근간이 되는 알고리즘인 FedAVG가 소개된 이래 꾸준히 연구되어 왔다. 연합학습은 Non-IID(independent and identically distributed) 로컬 데이터 하에서 데이터 이질성으로 인한 성능 저하 문제를 겪는다. 이를 해결하기



(그림 1) FL 시스템 모델 (중앙 클라우드 서버 및 3개의 엣지 엔터티)
(Figure 1) FL system model(central cloud server and three edge entities)

위해 FedPROX[7], Scaffold[8] 등 FedAVG를 최적화하려는 시도가 있었다. FedPROX는 데이터 이질성으로 인해 글로벌 모델의 수렴 지연을 완화하기 위해 FedAVG에 proximal term을 도입하였으며, Scaffold는 control variate를 활용하여 로컬 업데이트의 기울기 변화를 줄임으로써 글로벌 모델의 수렴을 가속하는 방법을 제안했다.

연합학습에서 개인정보 유출 및 보안 기술에 대한 다각도의 연구가 진행되어 왔다. 연합학습에 대한 주요 공격 유형으로 포이즈닝[9-11]과 적대적 공격[12-15]을 들 수 있다. 데이터 포이즈닝[9][10]은 로컬 데이터를 의도적으로 오염시켜 편향된 글로벌 모델 학습을 유도하는 것이다. 이러한 공격은 탐지로부터 회피하기 위해 여러 라운드에 걸쳐 반복적으로 글로벌 모델에 작은 오염을 주입하고 축적한다. 따라서 성능이 크게 저하되는 경우가 많다.

적대적 공격[15][16]은 공격을 위해 신중하게 제작된 입력 데이터, 즉 적대적 예제를 학습 데이터에 포함시킴으로써 딥러닝 모델을 의도적으로 조작한다. 이러한 공격은 딥러닝 모델의 의사 결정 프로세스의 취약점을 악용하여 잘못된 분류 혹은 잘못된 출력을 유발한다. 이를 실제로 구현하는 방법으로 PGD(Projected Gradient Descent)나 FGSM(Fast Gradient Signed Method) 등이 알려져 있다. 특히, PGD에 견고성(robustness)을 가지도록 학습된 딥러닝 모델은 다른 적대적 공격에도 견고성을 지니는 특성이 있는데 이를 1st order adversary라 칭한다[16]. 중앙 집중화된 환경에서 적대적 학습은 min-max기반 손실함수를 사용하여 최악의 경우 - 자연 데이터(공격받지 않은 데이터)와 가장 유사한 적대적 예제를 분류하는 경우 - 를 최적화 하도록 구현된다[3]. Zhang *et al.* [17]의 연구는 적대적 예제에 대한 예측 오류를 자연 오류와 경계 오류로 구분하고 이들 간 상호 균형을 유지할 수 있는 손실함수인 TRADES를 제안했으며, [18]에서는 예측 오류를 개선하는 MART손실함수를 제안했다.

문헌상 연합학습 환경에 적대적 학습을 최초로 적용한 연구로 Zizzo *et al.*[12]을 들 수 있다. [12]에서는 IID 및 Non-IID 데이터 하에서 연합학습과 적대적 학습을 화학적으로 결합한 FAT을 제안했다. 그러나 FAT은 Non-IID 데이터 하에서 학습 라운드가 어느 정도 진행되면 성능이 급격히 감소한다. [13]에서 이러한 현상을 '모델 드리프트'라고 명명하고 이를 해결하기 위한 모델 집계 알고리즘인 SFAT(slacked-FAT)을 제안했다. '모델 드리프트'는 적대적 예제가 학습 데이터에 추가됨으로써 학습 데이터의 이질성이 심화되기 때문에 발생하는 현상으로 SFAT은 데이터 이질성이 글로벌 모델의 집계에 미치는 영향을 완화하기 위해 가중치를 기반으로 로컬 모델을 집계하도록 한 것이다.

본 논문에서는 기존 연합학습 환경에서 적대적 학습에 대한 연구와 달리 클라이언트 중 일부가 공격받고 있는 시나리오를 가정한다. 연합학습 클라이언트가 자신이 공격받는지 여부에 따라 전략적으로 손실함수를 선택하여 적대적 학습을 수행하는 방법을 제안한다.

3. 시스템 모델

3.1 FL모델

본 논문에서는 그림 1과 같이 셀룰러 망에서 악성 안드로이드 앱을 탐지하기 위한 목적의 크로스 사일로(cross-silo) 연합학습 모델을 가정한다[19]. 제안하는 연합학습 모델에서는 클라우드 서버가 중앙 엔터티가 된다. 중앙 엔터티에서는 엣지 엔터티와 협력하여 글로벌 모델을 학습하고 로컬 모델 매개변수 값을 업데이트한다. 각 사일로에서 엣지 엔터티는 사일로에 속한 클라이언트를 대표하는 몇몇의 클라이언트 또는 기지국이 된다.

셀룰러 망에서는 대부분의 단말이 스마트폰이고, 사용자 수가 매우 많기 때문에 크로스 디바이스(cross-device) 모델을 적용하는 것이 합리적으로 보일 수 있다. 그러나 연합학습의 주요 목적이 글로벌 모델 학습임을 고려하면 사용자의 개인 데이터는 글로벌 모델의 거름이 될 로컬 모델을 학습하기에 충분하지 않은 경우가 많다. 더 중요한 것은 크로스 디바이스 모델에서는 모델 업데이트를 위해 중앙 서버와 클라이언트 간의 과도한 통신 대역폭이 필요하다는 것이다. 이는 개별 스마트폰 사용자에게는 반갑지 않은 일이다. 따라서 제안하는 연합학습 모델에서는 클라이언트에 비해 상대적으로 컴퓨팅 및 통신 성능이 더 좋은 기지국이 엣지 엔터티가 되거나(그림 1의 사일로 2), 클라이언트가 연합을 형성하고 클라이언트 중 일부가 연합을 대신해 엣지 엔터티 역할을 한다고 가정한다(그림 1의 사일로 1 또는 3).

3.2 적대적 공격 모델

적대적 공격과 관련한 기존 연구의 주요 내용은 '데이터에 아주 작은 왜곡(perturbation)을 통해 딥러닝 모델의 예측오류를 발생시킬 수 있는가?' 즉, '노말 데이터에 아주 작은 왜곡을 줌으로써 강력한 적대적 예제를 제작할 수 있는가?'에 대한 것이다. 위 질문에 답하기 위해 [3]에서 FGSM을 통한 적대적 예제를 제작할 수 있는 방법이 제안되었다. 이는 수식 (1)로 설명할 수 있다.

$$x + \epsilon sgn(\nabla_x L(\theta, x, y)) \quad (1)$$

수식 (1)에서 θ는 모델의 파라미터이며 x는 원래의 데이터, y는 타겟, 그리고 $L(\theta, x, y)$는 모델을 학습하는데 필요한 코스트를 의미한다. 작은 ϵ 값으로 제한된 두 번째 항이 이러한 작은 왜곡을 의미한다. 예컨대 이미지 데이터라면 유관으로는 구분하기 매우 어려운 노이즈를 추가하는 것과 유사하다.

수식 (1)에서 알 수 있듯 FGSM에서는 작은 왜곡을 한 번만 주입한다. PGD는 FGSM을 통한 왜곡의 주입을 반복적으로 수행할 수 있도록 일반화한 형태로 반복적 왜곡의 주입으로 훨씬 강력한 적대적 공격을 수행할 수 있다. [16]에서 실험적으로 밝혀낸 바에 따르면 PGD공격에 대해서 견고성을 지니도록 학습된 모델은 여타의 적대적 공격에도 견고성을 지니게 된다. 이러한 특성을 [16]에서는 1st order adversary로 정의하고 있다. 따라서 PGD는 많은 적대적 공격 관련 연구에서 공격 모델로 채택되었다. 본 논문에서는 제안하는 연합학습 모델에서 일부 엣지 엔터티에 대하여 PGD-20공격(20회 FGSM이 반복)을 수행한다.

4. 실험 환경 설정

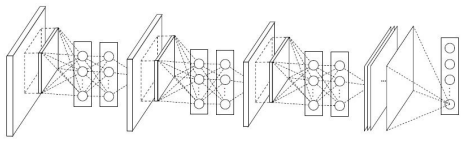
4.1 데이터셋 및 실험 환경

본 논문의 실험에서는 CICMalDroid 2020 데이터셋을 사용하였다[20]. CICMalDroid 2020 데이터셋은 VirusTotal, Contagio 블로그 등 써드 파티로부터 2017년 12월부터 2018년 12월까지 안드로이드 앱 설치 파일인 APK 형태로 수집되었다. 악성코드를 포함해 총 17,341개의 샘플로 구성되어 있으며 모든 샘플은 정상 및 악성 클래스 4종(애드웨어, 뱅킹, 리스크웨어 및 SMS) 중 하나에 속한다. APK 파일을 스트림 오더(stream order)[5]를 사용하여 이미지 파일로 변환하였으며 본 논문에서 학습된 딥러닝 모델은 이미지로 변환된 APK파일의 클래스 분류를 수행하는 것이다. 이미지로 변환된 총 17,341개의 안드로이드 앱 샘플 중 14,000개의 샘플을 무작위로 추출하여 학습 데이터로 사용하고, 나머지 3341개의 샘플은 평가 데이터로 사용한다.

4.2 FL 및 적대적 학습 매개변수 설정

4.2.1 FL 네트워크 모델

적대적 학습은 min-max손실함수의 안쪽 함수에서 데이터 왜곡의 양상을 학습하는 방식으로 구현된다[16]. 이를 위해서는 학습에 사용되는 신경망의 구조가 왜곡의 양상을 담을 수 있도록 충분히 복잡한 구조를 가지고 있어야 한다. 지면관계상 본 논문에서 실험결과를 제시하지는 않으나 적대적 학습을 간단한 합성곱 신경망 - 2~3개



(그림 2) 전면에 여러 MLP 컨볼루션 레이어와 전역 평균 풀링 레이어를 사용하는 NIN

(Figure 2) NIN using multiple MLP convolutional layers on the front and a global mean pooling layer

의 합성곱층과 완전연결층으로 구성된 신경망 - 을 통해 수행하였을 때 성능이 크게 저하됨을 관찰하였다. 이에 착안하여 기존의 잘 알려진 신경망 중에서 상당한 복잡도를 가지고 있는 NIN(Network in Network)을 채택하였다. 그림 2에서 확인할 수 있듯 NIN신경망은 여러 개의 MLP합성곱층(multi-layer perception convolutional)을 연결함으로써 구현된다. 우리는 NIN의 이중 구조가 적대적 예제의 왜곡을 충분히 학습하기 위해 충분하다고 판단하였다.

4.2.2 FL 매개변수 설정

본 논문에 제시된 실험결과는 NIN모델, CICMaldroid 2020 데이터셋, 모델 집계 알고리즘인 FedAVG에 대하여 경험적 최적화를 충분히 진행한 후 얻은 결과이다. 주지할만한 매개변수의 설정은 다음과 같다. 전체 엣지 엔터티의 수는 100개이고 매 라운드마다 10개(10%)의 엣지 엔터티가 랜덤하게 선택되어 연합학습에 참여한다. 로컬 학습에서의 라운드 수와 로컬 배치 크기는 10으로 설정되었다. 학습율(learning rate)은 모델의 수렴속도에 영향을 미치며 0.01로 세팅되었다. 이러한 연합학습의 하이퍼 파라미터 값들은 학습 성능에 지대한 영향을 미치나 기존의 연구에서 영향력이 충분히 논의되었기에 본 논문에서는 기존 연구[5-6]에서의 기본값을 변경 없이 사용한다.

4.2.3 적대적 학습 설정

데이터셋은 100개의 엣지 엔터티에 IID형태로 분산되어 있다고 가정한다. 매 학습 라운드에서 무작위로 선택된 10개의 엣지 엔터티가 해당 라운드의 로컬 학습을 수행한다. 무작위 선택은 매 라운드마다 이루어지므로 10개의 엣지 엔터티는 매 라운드마다 다르다고 볼 수 있다.

100개의 엣지 엔터티 중 AR(attack rate)의 비율만큼의 엣지 엔터티가 적대적 공격 하에 있다고 가정한다. 즉, AR이 1이면 모든 엣지 엔터티가 적대적 공격 하에 있는 상황으로 기존의 연합학습 환경에서의 적대적 학습에 대한 연구에서 고려하는 시나리오와 동일하다[12][13]. 본 논문에서 AR의 값은 0.1, 0.3, 1로 설정한다. 연합학습에 참여하는 엣지 엔터티가 매 라운드마다 바뀌는 것과 달리 공격 받는 대상은 고정되어 있다고 가정하는데 이것이 실제 상황과 더 부합하기 때문이다.

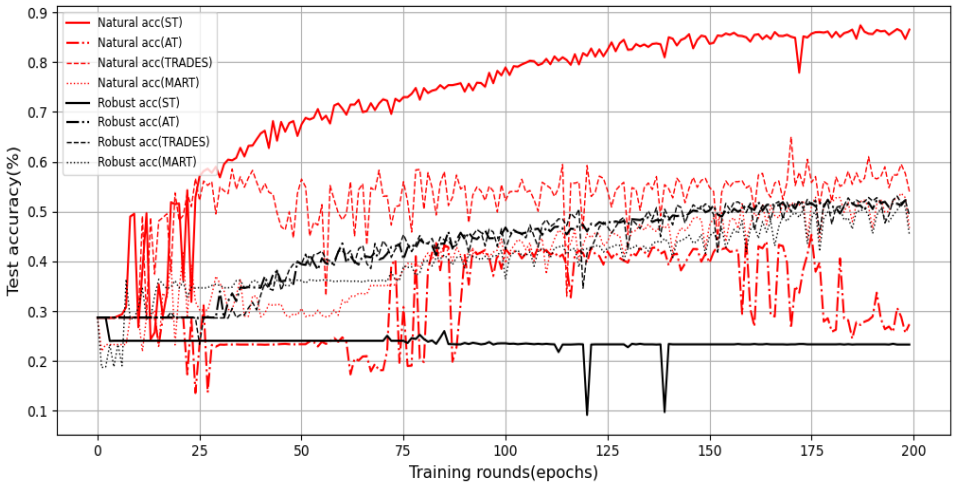
5. 실험 결과

5.1 모든 클라이언트가 적대적 공격을 받는 경우

그림 3에서 모든 엣지 엔터티가 적대적 공격을 받는 시나리오에서의 3종(AT, TRADES, MART)의 딥러닝 모델의 악성앱 탐지 성능을 도식하였다. ST(standard training)는 적대적 공격 하에 있지 않으며 적대적 학습을 수행하지 않고 학습된 모델이다. AT, TRADES, 그리고 MART는 모든 엣지 엔터티가 적대적 공격 하에 있을 때 적대적 학습을 수행하여 학습된 모델이다. TRADES와 MART는 적대적 학습에서 일반적으로 사용되는 min-max 기반 손실함수 대신 각각 [17]과 [18]에서 제안한 손실 함수가 사용되었다. Natural acc와 robust acc는 각각 PGD공격을 받지 않은 테스트 셋에서의 분류정확도, PGD공격 하에서의 테스트 셋에서의 분류 정확도를 의미한다.

모든 모델 학습 방법에서 공통적으로 발견되는 실험결과는 robust acc가 증가하면 정도의 차이는 있으나 natural acc가 감소한다는 것이다. 이를 통해 딥러닝 모델에서 natural acc와 robust acc가 트레이드오프 관계에 있음을 알 수 있다. 이러한 트레이드오프 관계를 보이는 이유는 다음과 같이 설명될 수 있다. 적대적 공격은 데이터에 왜곡을 주입하여 딥러닝 모델에서 학습해야할 '클래스 간 구분 경계'의 위치를 이동시키거나 모호하게 한다. 이동되어진 경계는 일반적인 딥러닝 학습 방법을 통해서 학습하기 용이하나 모호해진 경계는 상대적으로 유연한 min-max 기반 손실함수를 도입하여 학습함이 효과적이다.

ST의 경우 85%의 natural acc를 보이며, 23.3%의 robust acc를 보인다. ST는 PGD공격이 주입한 왜곡을 전혀 학습하지 못함을 보여주는 결과이다. 데이터셋이 5종의 클래스로 구성됨을 고려하면 ST는 PGD기반 적대적 예제에 대한 분류기능이 거의 전무하다고 볼 수 있다. 충분한 학습이 이루어진 3종의 적대적 학습의 robust acc는 50%정도를 보인다. MART의 경우 robust acc커브에 다소 진동이



(그림 3) 4종(ST, AT, TRADES, 그리고 MART)의 딥러닝 모델의 Natural 정확도 및 robust 정확도
(Figure 3) Natural accuracy and robust accuracy of four types of deep learning models (ST, AT, TRADES, and MART)

관찰되나 수렴된 모델의 분류 성능은 큰 차이를 보이지 않았다. 오히려 다른 손실함수를 사용함으로써 발생하는 차이는 natural acc에서 두드러진다. TRADES는 natural acc가 robust acc보다 높은 55%를 보이며, MART는 natural acc와 robust acc가 비슷한 수준이다. 이 결과로 TRADES와 MART의 손실함수가 상대적으로 적은 natural acc의 감소를 보여 robust acc와 natural acc의 트레이드오프 관계를 적절히 활용한다고 볼 수 있다. 충분한 학습이 진행되면 AT의 robust acc는 여타의 적대적 학습과 동등한 성능을 내나 natural acc커브에서는 160라운드 이후 하방 드리프트가 관찰되며 성능이 크게 저하됨을 볼 수 있다.

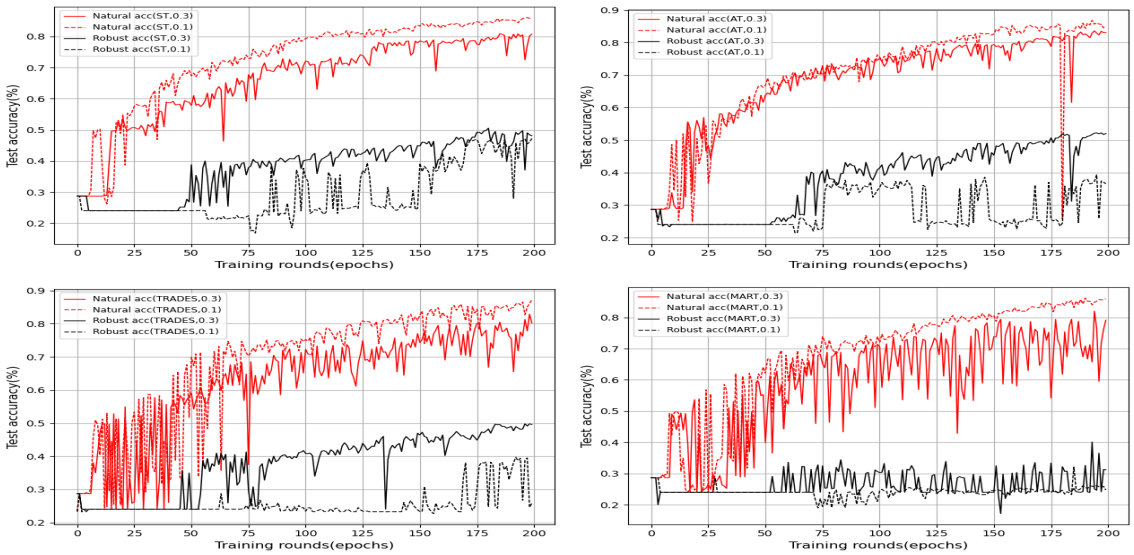
5.2 클라이언트의 일부가 적대적 공격을 받는 경우

그림 4에서 엣지 엔터티 중 일부가 적대적 공격을 받는 시나리오에서 딥러닝 모델의 악성앱 탐지 성능을 도식하였다. 매 학습 라운드마다 10개의 엣지 엔터티가 연합학습에 참여하므로 AR값이 0.1과 0.3이면 평균적으로 1, 3개의 엣지 엔터티가 공격을 받는 상황이다.

ST의 경우 AR값이 0.1일 때 적대적 공격으로 인한 natural acc의 감소가 거의 없다. 그러나 robust acc의 경우 45%정도를 보여준다. 이는 다음과 같은 이유로 설명할 수 있다. 적대적 예제의 특성에 대한 학습을 전혀 진행할 수 없었던 상황과 비교하여 robust acc가 크게 향상되었으나 학습 기회가 제한적이므로 robust acc커브의 수렴이 느리며 수렴과정에서 상당한 진동이 관찰된다. AR값이 0.3일 때는 natural acc의 수렴지점이 80%정도로 낮아지며 robust acc는 48%정도에 수렴한다. 이는 AR값이 0.1일 때와 비교하여 적대적 예제의 특성에 대한 학습이 비교적 잘 수행되고 있음을 의미한다.

AT도 ST와 유사하게 AR값이 0.1일 때 적대적 공격으로 인한 natural acc의 감소가 거의 없다. 그러나 robust acc의 경우는 37%정도에서 수렴하며 상당한 진동이 관찰된다. AR값이 0.3일 때 적대적 공격으로 인한 natural acc는 83%정도로 그 감소량이 가장 적다. Robust acc는 52%정도에서 수렴한다. 이는 모든 엣지 엔터티가 적대적 공격하에 있는 상황에서 TRADES가 비교 스킴 중 가장 좋은 성능을 냈던 것과 대비된다.

TRADES의 경우 AR이 0.3일 때 robust acc가 50%정도이며 가장 진동이 적은 수렴 곡선을 보였다. Natural acc커브에서는 AT보다 진동이 많고 성능이 약간 내려간 80%정도였다. 이는 유연성이 좋은 손실함수를 사용하는 것이 공격을 받지 않은 데이터의 특성을 학습하는데 있어서는 오히려 역효과를 낼 수도 있다는 점을 확인하는 실험결과라고 할 수 있다. MART의 경우 AR이 0.3일 때 모든 비교 방법 중 natural acc커브에서 가장 심각한 진동을 보였으며 robust acc는 다른 적대적 학습방법과는 달리 30%정도의 성능을 보였다.



(그림 4) 엣지 엔터티 중 일부가 적대적 공격을 받는 시나리오에서 딥러닝 모델의 악성앱 탐지 성능
(Figure 4) Malicious app detection performance of deep learning models in scenarios where some of the edge entities are subject to adversarial attacks

정리하면 실험에 사용된 3종의 적대적 학습의 방법들은 엣지 엔터티의 일부가 공격받는 상황에서 특정한 학습방법이 다른 방법에 비해 우수한 성능을 내거나 하는 경우는 관찰하지 못했다. 반면, 모든 엣지 엔터티가 공격을 받는 상황에서는 각 학습방법의 장단을 실험적으로 확인할 수 있다. Robust acc와 natural acc사이의 트레이드오프 관계를 효과적으로 활용하는 적대적 훈련을 위해서는 TRADES 손실함수를 사용하는 것이 적절하다. TRADES 손실함수를 사용한 적대적 학습을 통해 얻은 딥러닝 모델의 robust acc는 여타의 적대적 학습방법과 동등하면서 natural acc에서 가장 우수하다. 반면, 기본적인 min-max 기반 손실함수를 사용하는 적대적 학습은 노말 샘플을 분류하는데 있어서 극심한 성능저하를 겪는다. MART손실함수를 사용하면 성능이 향상되나 전반적으로 TRADES를 사용한 적대적 학습보다는 성능이 낮다.

5.3 적대적 학습에서 손실함수의 적응적 선택

본 논문의 실험에서는 엣지 엔터티가 공격받고 있는지 여부와 관계없이 동일한 손실함수를 사용한다. 예컨대 AR이 0.1인 경우 연합학습의 매 학습 라운드에 참여하는 적대적 예제는 평균적으로 1개의 엣지 엔터티의 데이터에만 포함되어 있음에도 불구하고 모든 엣지 엔터티가 일률적으로 동일한 손실함수를 사용하여 딥러닝 모델의 학습을 수행한다. 우리의 실험결과에 의하면 유연성이 좋은 손실함수는 적대적 예제의 특성을 학습하는 데에는 강점이 있으나 오히려 정상적인 샘플을 학습하는데 있어 부정적인 영향을 줄 수 있음을 확인했다. 이를 우리는 적대적 학습의 트레이드오프 관계라 명명했다.

이러한 트레이드오프 관계를 효과적으로 활용하기 위해 우리는 엣지 엔터티가 공격 하에 있는지 여부에 따라 적응적으로 손실함수를 사용하는 방법론을 적용할 필요가 있다. 이를 위해서는 엣지 엔터티가 자신이 공격 하에 있음을 스스로 탐지할 수 있는 기법이 개발되어야 한다. 간단히 고려할 수 있는 방법으로 중앙 엔터티로부터 전송받은 글로벌 모델에 로컬 데이터를 테스트셋으로 한 모델 예측을 수행해보고 분류 정확도가 임계값보다 낮으면 공격 하에 있다고 추론하는 방법이다. 만약 엣지 엔터티가 스스로 자신이 공격 하에 있다고 판단한다면 TRADES나 MART등 유연한 손실함수를 사용하여 적대적 예제의 특성을 학습할 수 있도록 하고, 그렇지 않은 경우에는 일반적인 손실함수를 사용하여 정상 데이터에 대한 정확도를 향상시키도록 하는 것이다. 현재 본 연구팀은 이러한 손실함수의 적응적 선택방법을 실용적으로 구현하는 연구를 진행 중에 있다. 따라서 제안한 알고리즘에 대한 수학적인 분석이나 실험적 검증은 향후 과제로 남긴다.

6. 결 론

본 논문에서는 연합학습의 클라이언트의 일부가 적대적 공격 하에 있을 때, 적대적 학습을 효과적으로 수행하기 위한 제반사항을 실험적으로 살펴보았다. 실험결과에 대한 고찰을 통해 적대적 학습의 트레이드오프 관계를 효과적으로 활용하기 위한 방법의 설계방향에 대해 논했다. 제안하는 방법에 대한 수학적인 분석이나 실험적 검증은 향후 과제로 남긴다.

참고문헌(Reference)

[1] Drainakis, Georgios, et al. “Federated vs. centralized machine learning under privacy-elastic users: A comparative analysis,” 2020 IEEE 19th International Symposium on Network Computing and Applications (NCA), IEEE, 2020. http://dx.doi.org/10.1109/NCA51143.2020.9306745

[2] Preuveneers, Davy, et al. “Chained anomaly detection models for federated learning: An intrusion detection case study,” Applied Sciences, 8.(12), 2663, 2018. http://dx.doi.org/10.3390/app8122663

[3] Goodfellow, Ian, et al. “Generative adversarial nets,” Advances in neural information processing systems, 27, 2014. https://dl.acm.org/doi/10.5555/2969033.2969125

[4] Kang, Mingu, et al. “Resilience against Adversarial Examples: Data-Augmentation Exploiting Generative Adversarial Networks,” KSII Transactions on Internet & Information Systems, 15(11), 4105-4121, 2021. http://doi.org/10.3837/tiis.2021.11.013

[5] Lee, Suchul. “Distributed Detection of Malicious Android Apps While Preserving Privacy Using Federated Learning,” Sensors, 23(4), 2198, 2023. https://doi.org/10.3390/s23042198

[6] McMahan, Brendan, et al. “Communication-efficient learning of deep networks from decentralized data,” Artificial intelligence and statistics, PMLR, 2017. https://doi.org/10.48550/arXiv.1602.05629

[7] Li, Tian, et al. “Federated optimization in heterogeneous networks,” Proceedings of Machine learning and systems, 2, 429-450, 2020. https://doi.org/10.48550/arXiv.1812.06127

[8] Karimireddy, Sai Praneeth, et al. “Scaffold: Stochastic controlled averaging for federated learning,” International conference on machine learning, PMLR, 2020. https://doi.org/10.48550/arXiv.1910.06378

[9] Tolpegin, Vale, et al. “Data poisoning attacks against federated learning systems,” Computer Security–ESORICS 2020: 25th European Symposium on Research in Computer Security, ESORICS 2020, Guildford, UK, September 14–18, 2020, Proceedings, Part I 25. Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-58951-6_24

[10] Farhadkhani, Sadegh, Rachid Guerraoui, and Oscar Villemaud, “An equivalence between data poisoning and byzantine gradient attacks,” International Conference on Machine Learning, PMLR, 2022. https://doi.org/10.48550/arXiv.2202.08578

[11] Fang, Minghong, et al. “Local model poisoning attacks to Byzantine-Robust federated learning,” 29th USENIX security symposium (USENIX Security 20), 2020. https://doi.org/10.48550/arXiv.1911.11815

[12] Zizzo, Giulio, et al. “Fat: Federated adversarial training,” arXiv preprint arXiv:2012.01791, 2020. https://doi.org/10.48550/arXiv.2012.01791

[13] Zhu, Jianing, et al. “Combating Exacerbated Heterogeneity for Robust Models in Federated Learning,” arXiv preprint arXiv:2303.00250, 2023. https://doi.org/10.48550/arXiv.2303.00250

[14] Bai, Tao, et al. “Recent advances in adversarial training for adversarial robustness,” arXiv preprint arXiv:2102.01356, 2021. https://doi.org/10.48550/arXiv.2102.01356

[15] Hitaj, Briland, Giuseppe Ateniese, and Fernando Perez-Cruz, “Deep models under the GAN: information leakage from collaborative deep learning,” Proceedings of the 2017 ACM SIGSAC conference on computer and communications security, 2017. https://doi.org/10.1145/3133956.3134012

[16] Madry, Aleksander, et al. “Towards deep learning models resistant to adversarial attacks,” arXiv preprint arXiv:1706.06083, 2017. https://doi.org/10.48550/arXiv.1706.06083

[17] Zhang, Hongyang, et al. “Theoretically principled trade-off between robustness and accuracy,” International conference on machine learning, PMLR, 2019. https://doi.org/10.48550/arXiv.1901.08573

[18] Wang, Yisen, et al. “Improving adversarial robustness requires revisiting misclassified examples,” International conference on learning representations, 2019.

[19] Mothukuri, Viraaji, et al. “A survey on security and

privacy of federated learning," Future Generation Computer Systems, 115, 619-640, 2021. https://doi.org/10.1016/j.future.2020.10.007
[20] CICMalDroid 2020 datasets, 2020. https://www.unb.ca/cic/datasets/maldroid-2020.html

◐ 저 자 소 개 ◑

이 수 철(Suchul Lee)
2008년 서울대학교 컴퓨터공학부(공학사)
2014년 서울대학교 대학원 컴퓨터공학부(공학박사)
2014년~2016년 ETRI부설 국가보안기술연구소 선임연구원
2016년~현재 한국교통대학교 철도대학 데이터사이언스학과 부교수
관심분야 : 정보통신 및 보안, 인공지능
E-mail : sclee@ut.ac.kr