TrapMI: 분할 학습에서 모델 전도 공격을 회피할 수 있는 훈련 데이터 보호 방법[1)]

나현식[1], 최대선[2]
[1]숭실대학교 소프트웨어학과 석박사통합과정
[2]숭실대학교 소프트웨어학과 교수
rnrud7932@soongsil.ac.kr, suchoi@ssu.ac.kr

TrapMI: Protecting Training Data to Evade Model Inversion Attack on Split Learning

Hyun-Sik Na[1], Dae-Seon Choi[1]
[1]Dept. of Software, Soongsil University

요 약

Edge AI 환경에서의 DNNs 학습 방법 중 하나인 분할 학습은 모델 전도 공격으로 인해 입력 데이터의 프라이버시가 노출될 수 있다. 본 논문에서는 분할 학습 환경에서의 모델 전도 공격에 대한 기존 방어 기술들의 한계점을 회피할 수 있는 TrapMI 기술을 제안하고, 이를 통해 입력 이미지를 원본 데이터 세트의 도메인에서 특정 타겟 이미지 도메인으로 이동시킴으로써 이미지 복원의 가능성을 최소화시킨다. 추가적으로, 테스트 과정에서 타겟 이미지의 정보를 알 수 없는 제약을 회피하기 위해 AutoGenerator를 구축한 후 실험을 통해 원본 데이터 보호 성능을 검증한다.

1. 서론

분할 학습(Split Learning, SL)은 Edge AI 환경에서 Deep Neural Networks(DNNs) 모델을 학습하기 위한 대표적인 접근법 중 하나로 주목을 받고 있다[1]. 이것은 사용자의 local device에 있는 모델 앞단(Client-side model)에서 입력 데이터의 중간 특징값을 연산한 후 서버에 해당 값을 전송하여 서버 측에 있는 모델 뒷단(Server-side model)에서 나머지 연산을 수행함으로써 최종 출력값을 획득할 수 있다. 이러한 환경에서 사용자는 개인의 데이터를 직접 전송하지 않고 중간 특징값(Intermediate feature)만을 서버에게 제공하기 때문에 데이터 유출 문제를 방지할 수 있다는 장점이 있다.

하지만, 해당 서버가 악의적으로 모델 전도 공격(Model inversion attack)[2]을 통해 사용자의 입력 데이터를 중간 특징값을 통해 복원할 수 있으며, 이것은 SL 환경이 여전히 데이터 유출 문제를 완전히 방지하는 데 한계가 있음을 의미한다.

모델 전도 공격을 방지하기 위해, 이전 연구들은 중간 특징값에 Laplacian noise를 주입하거나, 입력 데이터와 중간 특징값 사이의 거리 상관 관계를 최소화하는 방식(NoPeekNN)으로 복원 성능을 제한하였다[3]. 하지만 이러한 방어 방법들은 노이즈 크기 및 방어 강도에 따라 모델 훈련 성능에 큰 악영향을 미칠 수 있다. 또한, 해당 방어 방법들이 적용된 중간 특징값을 모델 전도 공격을 통해 복원했을 때, 복원 이미지는 훼손된 형태로 보이게 되고, 공격자는 방어 기술이 적용되었음을 유추할 수 있다. 즉, 공격자가 시각화된 복원 이미지를 통해 방어 기술을 의심한다면 적응형 공격을 시도할 수 있으며, 결국 방어를 회피할 수 있는 상황이 발생할 가능성이 있다. 추가적으로, 이러한 방어 방법들은 원본 중간 특징값 분포의 도메인으로부터 크게 벗어나지 않기 때문에 복원 데이터들이 원본 입력 이미지의 형태를 크게 벗어날 수 없다는 한계가 존재한다.

본 논문에서는 이러한 문제를 해결하기 위해 TrapMI 방어 접근법을 제안한다. 사용자는 Client-side model에 데이터를 입력하기 전에 다른 타겟 이미지로 변환을 수행하여 해당 이미지를 입력한다. 즉, 서버는 변환된 이미지와 연결되는 중간 특징값을 전송받게 되면서 모델 전도 공격 시 원본 이미지가 아닌 변환 이미지를 복원하게 된다. 또한, 타겟 이미지는 원본 이미지와 전혀 다른 도메인에 위치하게 되면서 복원 이미지를 통해 원본의 도메인을 유추할 수 없으며, 복원된 이미지는 [3]과 같이 깨지거나 지저분하지 않아 공격자는 방어 기술의 존재를 의심하기 어려워진다.

1) 이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2021-0-00511, 엣지 AI 보안을 위한 Robust AI 및 분산 공격탐지기술 개발)

2. 위협 모델

본 논문은 SL 환경에서의 DNNs를 활용한 분류 작업의 입력 데이터 보호를 목표로 하며, 공격자(서버)는 다음과 같은 보유 지식에 따라 다른 공격을 시도할 수 있다:

- Client-side model f_{θ_1}의 아키텍처
- Client-side model f_{θ_1}의 파라미터(가중치 및 편향)
- 훈련에 직접 사용되는 훈련 데이터 세트 X_{train}
- 훈련 데이터 세트 X_{train}의 분포
- Client-side model f_{θ_1}에 대한 질의 가능 여부

본 논문에서는 현실적인 제약을 고려하여 공격자는 f_{θ_1}의 아키텍처 및 파라미터를 알 수 없고, X_{train}에 포함된 데이터들은 알 수 없지만 X_{train}의 특징 및 분포를 알 수 있다는 가정을 하였다. 또한, 공격자가 f_{θ_1}에 임의의 데이터를 입력하여 중간 특징값을 획득할 수 있다는 설정을 하였다.

이에 따라 공격자는 임의의 역 네트워크(Inverse network) $\hat{f}_{\theta_1}$를 구축한 후, X_{train}와 유사한 분포를 가진 공격 데이터 세트 X_{attack}을 f_{θ_1}에 반복적으로 질의하여 출력된 중간 특징값을 $\hat{f}_{\theta_1}$에 입력해 재복원한다. 즉, 다음 수식을 통해 $\hat{f}_{\theta_1}$를 최적화시킨다:

$$\hat{f}_{\theta_1} = argmin_{\hat{f}_{\theta_1}} \frac{1}{m}\sum_{i=1}^{m} \| \hat{f}_{\theta_1}(f_{\theta_1}(\hat{x}_i)) - \hat{x}_i \|_2^2$$

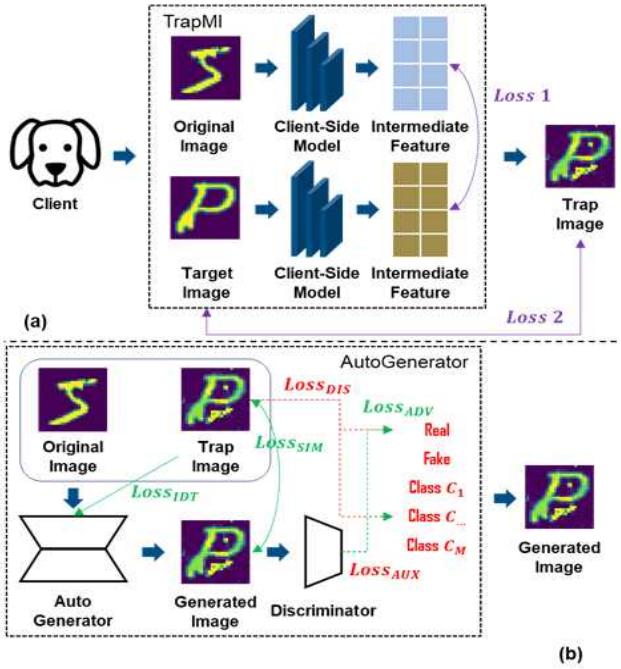
여기서 $\hat{x}$는 X_{attack} 내 각 데이터이며, 공격자는 $\hat{f}_{\theta_1}$를 통해 복원된 이미지 $\hat{f}_{\theta_1}(f_{\theta_1}(\hat{x}_i))$와 원본 공격 이미지 $\hat{x}_i$ 간 L_2 거리를 최소화하면서 최적화된 복원 모델을 구축할 수 있다.

3. 제안 방법

3.1. TrapMI(Trap against Model Inversion)

[그림 1(a)]는 본 논문에서 제안하는 TrapMI의 전체 구조도이다. 사용자(Client)는 원본 이미지(Original Image) x와 특정 타겟 이미지(Target Image) x_{tar}를 입력하여 Client-side model f_{θ_1}를 통해 각 이미지에 해당하는 중간 특징값(Intermediate Feature) I_{org}과 I_{tar}를 추출할 수 있다. 또한 사용자는 TrapMI를 통해 x_{tar}와 유사하면서 I_{org}과 I_{tar}의 차이를 최소화한 Trap Image x_{trap}를 생성할 수 있다. 즉, 다음 수식을 통해 최적의 x_{trap}를 생성한다:

$$x_{trap} := minimize_{\bar{x}} (Cos(I_{org}, I_{\bar{x}}) + \alpha \times (1 - SSIM(\bar{x}, x_{tar}))$$



(그림 1) (a) TrapMI 구조도 (b) AutoGenerator 구조도

여기서 x_{trap}의 초기값은 $\bar{x} = x_{tar}$이며, Cos, $SSIM$, α는 각각 Cosine Distance Loss, Structural Similarity Loss, Loss의 가중치를 의미한다. TrapMI는 두 손실 함수 최소화를 통해 $\bar{x}$를 반복 업데이트하여 최적의 x_{trap}를 도출할 수 있다.

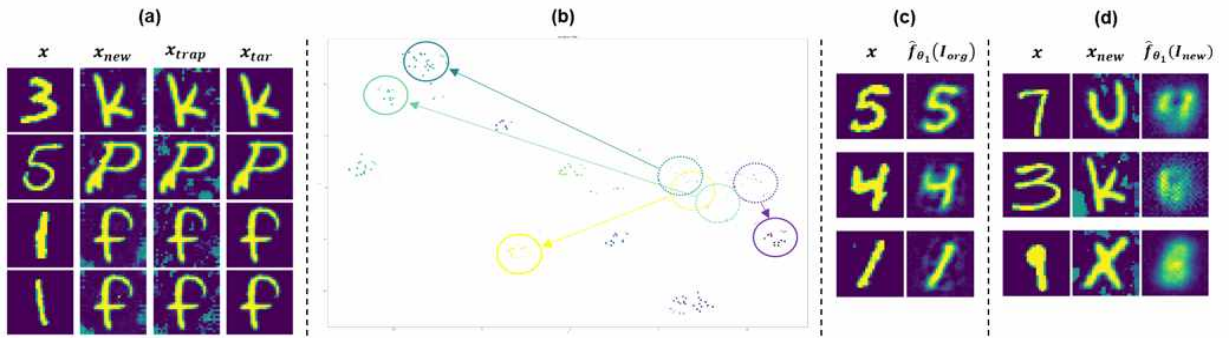
3.2. AutoGenerator

사용자는 훈련 단계에서 x_{tar}에 대한 접근이 가능하지만, 입력 데이터의 레이블을 예측하는 분류 작업의 테스트 단계에서는 접근을 할 수 없다. [그림 1(b)]는 이러한 환경에서 각 x에 대해 자동으로 x_{tar}에 가까운 이미지 x_{new}를 생성할 수 있는 AutoGenerator의 학습 구조도이다. 이것은 전통적인 Conditional Generative Adversarial Networks (CGANs)[4]의 구조를 기반으로 생성기 G와 판별기 D로 구성되어 있으며, 각 모델은 다음 수식을 통해 최적화한다:

$$G = minimize\,(\alpha_{ADV} Loss_{ADV}(D(G(x)), y_{Fake}) + \alpha_{SIM} Loss_{SIM}(x_{trap}, G(x)) + \alpha_{IDT} Loss_{IDT}(x_{trap}, G(x_{trap})))$$

$$D = minimize\,(\alpha_{DIS} Loss_{DIS}((D(G(x)), y_{Fake}) + (D(x_{trap}), y_{Real})) + \alpha_{AUX} Loss_{AUX}(D(G(x)), y_{Class}))$$

여기서 $Loss_{ADV}$은 D가 x_{new}를 가짜 이미지라고 판별한 손실값, $Loss_{SIM}$은 x_{trap}과 x_{new}의 L_2 거리, $Loss_{IDT}$는 G의 TrapMI 도메인 분포 학습을 강화하



(그림 2) (a) AutoGenerator 및 TrapMI를 통해 생성된 이미지 (b) 원본, 생성, Trap, 타겟 이미지 분포 시각화 (c) 방어 기술이 적용되지 않은 SL 환경에서의 모델 전도 공격 (d) 제안 방법이 적용된 모델 전도 공격

는 용도로 사용된다. 또한, $LOSS_{DIS}$와 $Loss_{AUX}$는 일반적인 CGANs에서 사용되는 손실 함수[4]이다.

4. 실험 및 결과

4.1. 구현 및 실험 설정

제안 방법을 구현하기 위해 타겟 분류 모델 f_θ는 4개의 Convolutional Layer(CL)로 구성된 간단한 Convolutional Neural Networks 모델을 사용하였고, 앞단의 두 레이어를 f_{θ_1}로 설정하였으며, $\hat{f}_{\theta_1}$는 2개의 CL과 2개의 Linear Layer로 구성되었다. 그리고 G는 ResNet-18을, D는 8개의 CL을 기반으로 구축하였다. 또한, α는 0.8, α_{ADV}, α_{SIM}, α_{IDT}, α_{DIS}, α_{AUX}는 각각 1.0, 1.5, 1.0, 1.0, 1.5로 설정하였다.

한편, 실험을 위해 훈련 및 테스트 데이터 세트는 MNIST의 10개 클래스를 모두 사용하였고, 타겟 이미지는 EMNIST에서 MNIST의 각 클래스 당 다른 알파벳 이미지 한 장씩 선택하였다. 그리고 분류 모델, TrapMI, AutoGenerator의 반복 횟수는 각각 20, 50, 80회로 설정하였고, 학습률은 각각 0.001, 0.1, 0.001이며, Adam optimizer를 통해 최적화하였으며, 훈련 시 batch size는 64로 설정하였다. 실험 설정에 따라 방어 기술이 적용되지 않는 기본적인 f_θ의 분류 성능은 99.26%를 달성하였고, 제안 방법을 적용한 분류 작업 시 모든 훈련 및 테스트 데이터들은 TrapMI 기술과 AutoGenerator를 통과한 후 입력하였다.

4.2. 실험 결과

[그림 2(a)]의 각 열은 x, x_{new}, x_{trap}, x_{tar}이다. TrapMI로 생성된 x_{trap}은 x_{tar}과 유사하면서 x의 중간 특징값으로 유도되는 노이즈가 추가된 것을 확인할 수 있다. 또한, x_{new}는 x_{trap}과 유사하게 생성된 것을 확인할 수 있다. [그림 2(b)]는 클래스별 생성된 이미지들이 원본 도메인(점선 원)에서 타겟 도메인(실선 원)으로 모두 이동한 것을 보여주며, [그림 2(c)]와 달리 공격자는 모델 전도 공격을 통해 복원에 성공하여도 [그림 2(d)]와 같이 x를 유추할 수 없게 된다. 추가적으로, AutoGenerator를 통해 생성한 이미지의 분류 성능은 97.73%로, 기존 분류 성능에 가깝게 도달할 수 있었다.

4.2. 결론

본 논문에서는 SL 환경에서 훈련 및 테스트 과정 중 입력 데이터 보호를 위해 TrapMI를 제안하였으며 테스트 과정에서 타겟 이미지의 정보 없이 이미지를 변환하기 위해 AutoGenerator를 제안하여 실험 결과를 도출하였다. 서버의 관점에서 모델 전도 공격을 통한 입력 데이터 복원의 성공 가능성을 최소화하면서 기존 방어 기술의 단점을 회피할 수 있는 실험 결과를 보여주었다.

참고문헌

[1] Letaief K. B. et al. “Edge Artificial Intelligence for 6G: Vision, Enabling Technologies, and Applications” IEEE Journal on Selected Areas in Communications, 40, 1, 5-36, 2021.

[2] He Z. et al. “Model Inversion Attack Against Collaborative Inference” 35th Annual Computer Security Applications Conference, 2019, 148-162.

[3] Titcombe T. et al. “Practical Defences Against Model Inversion Attacks for Split Neural Networks” ArXiv preprint arXiv:2104.05743, 2021.

[4] Mirza M. and Simon O. “Conditional Generative Adversarial Nets” ArXiv preprint, arXiv:1411.1784, 2014.