확률적 연합 언러닝

류지현, 양희철
충남대학교 컴퓨터공학과
ryujihyunmail@gmail.com, hcyang@cnu.ac.kr

Probabilistic Federated Unlearning

Jihyun Ryu, Heecheol Yang
Department of Computer Science & Engineering, Chungnam National University

요 약

본 논문은 연합학습(Federated Learning) 환경에서 악성 클라이언트의 모델 공격으로 인한 보안 위협을 완화하기 위해, 탐지 불확실성을 확률적으로 모델링하여 반영하는 확률적 연합 언러닝(Probabilistic Federated Unlearning) 기법을 제안한다. 기존의 연합 언러닝 기법들은 탐지 결과를 정상 혹은 악성으로 단정하는 결정론적(Deterministic) 방식을 채택하여 오탐 및 미탐에 따른 성능 저하 문제를 겪었으나, 본 연구는 탐지 결과를 연속적인 확률값으로 해석하고 쌍곡탄젠트 기반의 비선형 가중치 함수를 사용하였다. CIFAR-10 데이터셋을 이용한 백도어 공격 시나리오 실험 결과, 제안 기법은 다양한 탐지 불확실성 상황에서도 모델의 정확도를 안정적으로 유지하면서 백도어 공격 성공률을 낮출 수 있음을 보였다.

Ⅰ. 서 론

최근 인공지능 기술의 발전과 함께 데이터 프라이버시 보호가 중요해지면서, 데이터를 중앙 서버로 수집하지 않고 로컬에서 학습하는 연합학습(Federated Learning)이 주목받고 있다[1]. 그러나 연합학습은 서버가 개별 클라이언트의 원본 데이터를 직접 확인할 수 없다는 구조적 특성으로 인해 악의적 클라이언트가 전역 모델을 오염시키는 모델 중독 공격이나 백도어 공격에 취약하다. 특히 백도어 공격은 특정 입력 패턴에 대해서만 오작동을 유도하기 때문에 탐지가 매우 어렵다.

이러한 위협에 대응하여 학습된 모델에서 특정 데이터의 영향을 제거하는 머신 언러닝(Machine Unlearning) 기술이 연합 언러닝(Federated Unlearning)으로 확장되고 있다[2]. 기존의 대표적인 알고리즘인 FedRecovery[3]는 그래디언트 잔차를 사용하여 재학습 없이 성능을 유지하며 , FedEraser[4]는 클라이언트의 업데이트 정보를 주기적으로 저장한 후 언러닝 시 이를 보정하여 모델을 재구성한다. 그러나 기존의 연합 언러닝 기법들은 언러닝 수행 전 악의적 클라이언트를 정확히 탐지해야 한다는 전제를 가지며, 탐지 결과를 정상 혹은 악성 두 가지로만 구분하는 결정론적 구조를 취한다. 이러한 방식은 탐지 과정의 불완전성으로 인해 오탐(False Positive) 발생 시 정상 클라이언트의 기여가 불필요하게 제거되어 성능이 저하되고, 미탐(False Negative) 발생 시 악성 업데이트의 영향이 남아 있는 한계를 지닌다. 본 연구에서는 탐지 결과를 연속적인 확률값으로 해석하고 이를 비선형 가중치로 언러닝 과정에 반영함으로써, 탐지 불확실성 상황에서도 안정적이고 강건한 학습을 가능하게 하는 확률적 연합 언러닝 기법을 제안한다.

Ⅱ. 본론

본 연구에서 제안하는 시스템 모델은 중앙 서버와 N개의 클라이언트로 구성되며, 전체 과정은 학습과 언러닝 두 단계로 구분된다.

먼저 학습 과정에서 서버는 전역 모델 W_G를 학습 참여 클라이언트에 배포한다. 클라이언트는 서버로부터 받은 모델을 클라이언트의 로컬 데이터셋 D_i로 학습하며, 이 과정에서 악성 클라이언트로 지정된 클라이언트는 백도어 트리거가 삽입된 데이터를 사용하여 백도어 공격을 수행한다. 악성 클라이언트가 아닌 클라이언트는 정상적으로 로컬 데이터로 학습을 수행한다. 클라이언트의 로컬 업데이트 파라미터 $W_1, W_2, \dots, W_N$를 서버로 전송하여 서버는 이를 사용하여 전역 모델을 업데이트한다.

이후 언러닝 과정에서 클라이언트가 악성일 가능성을 나타내는 악성 확률 벡터 $p = [p_1, p_2, \dots, p_N]$을 가중치로 사용하여 서버는 언러닝을 수행한다. 각 p_i는 0과 1 사이의 확률값으로 본 연구에서는 확률값이 주어졌다고 가정하고 사용하였다.

클라이언트 C_i의 악성 탐지 확률 p_i가 주어질 때, 다음과 같이 쌍곡탄젠트함수를 사용한 가중치 함수를 정의하였다.

$$\phi_i(p_i;\alpha) = \frac{\tanh(\alpha(p_i - 0,5))}{\tanh(0.5\alpha)}$$

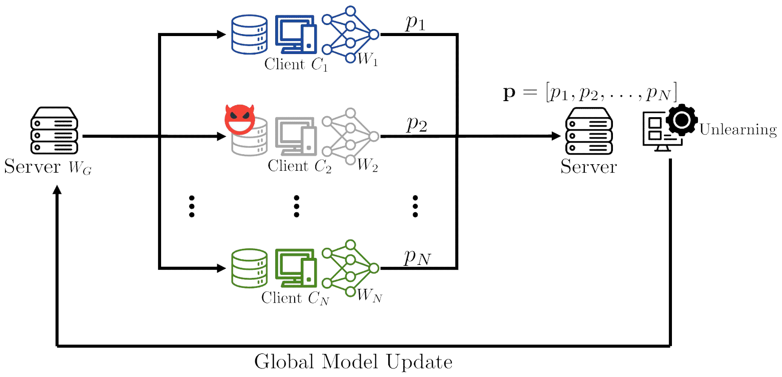
이를 사용하여 세 가지 언러닝 방식으로 사용하여 업데이트를 수행한다.
확률 가중치 경사 하강법(Gradient-Descent, GD)은 확률값이 작은 경우 즉, 정상 클라이언트로 탐지될 가능성이 있는 경우 큰 가중치를 부여하여 언러닝을 진행한다.

$$W^{(t+1)} = W^{(t)} - \eta \sum_{i=1}^{N} \overline{\omega_i} \nabla_W L_i(W^{(t)}), \text{ where } \overline{\omega_i} = \frac{\omega_i}{\sum_j \omega_j}, \quad \omega_i = \frac{1 - \phi_i}{2}$$

$W^{(t)}$는 라운드 t에서의 전역 모델, η는 학습률, $\overline{\omega_i}$는 가중치 평균 그리고 $L_i(W)$는 클라이언트 C_i의 로컬 손실함수이다.
확률 가중치 경사 상승법(Gradient-Ascent, GA)은 확률값이 큰 경우 즉, 악성으로 탐지될 가능성이 큰 경우로 큰 가중치를 부여하여 언러닝을 진행한다.

$$W^{(t+1)} = W^{(t)} + \eta \sum_{i=1}^{N} \overline{\gamma_i} \nabla_W L_i(W^{(t)}), \text{ where } \overline{\gamma_i} = \frac{\gamma_i}{\sum_j \gamma_j}, \quad \gamma_i = \frac{1 + \phi_i}{2}$$

$W^{(t)}$는 라운드 t에서의 전역 모델, η는 학습률, $\overline{\gamma_i}$는 가중치 평균 그리고 $L_i(W)$는 클라이언트 C_i의 로컬 손실함수이다.
마지막으로 확률 가중치 경사 하강/상승법(Gradient-Des/Asc, GDGA)의 경우 확률값 0.5를 기준으로 나누어서 정상으로 판단된 클라이언트는 경사 하강을 악성으로 판단된 클라이어트는 경사 상승으로 진행하여 업데이트를 한다.



〔그림1〕 시스템 모델

$$\sigma_i = \begin{cases} -1, & \phi_i < 0 \quad (\text{정상, Descent}) \\ +1, & \phi_i \geq 0 \quad (\text{악성, Ascent}) \end{cases}, \epsilon_i = \begin{cases} \frac{1-\phi_i}{2}, & \phi_i < 0 \\ \frac{1+\phi_i}{2}, & \phi_i \geq 0 \end{cases}$$

$$W^{(t+1)} = W^{(t)} + \eta \sum_{i=1}^{N} \sigma_i \bar{\epsilon}_i \nabla_W L_i(W^{(t)}), \text{where} \quad \bar{\epsilon}_i = \frac{\epsilon_i}{\sum_j \epsilon_j}$$

$W^{(t)}$는 언러닝 라운드 t에서의 전역 모델, η는 학습률, σ_i는 경사 하강 또는 상승을 결정하는 부호, $L_i(W)$는 클라이언트 C_i의 로컬 손실함수, ϵ_i는 확률 가중치 그리고 $\bar{\epsilon}_i$는 가중치 평균이다.

Ⅲ. 실험 및 결과 분석

1. 실험 설계 및 환경

실험을 위해 CIFAR-10 데이터셋의 60,000장 이미지를 활용하였으며, 데이터를 I.I.D.와 non-I.I.D. 방식으로 클라이언트에 분배하였다. 학습 모델은 ResNet-18을 사용하였으며, 악성 클라이언트는 이미지 우측 하단에 8×8 흰색 패치를 삽입하는 백도어 공격을 수행하도록 설정하였다. 전체 학습은 30라운드를 진행하고 언러닝을 30라운드 진행하였다. 탐지 불확실성을 가정한 확률 벡터 p를 [0.7, 0.6, 0.6, 0.22, 0.2, 0.2, 0.15, 0.13, 0.1, 0.1]로 설정하여 시나리오를 진행하였다.

2. 평가 지표

· 정확도 (Accuracy, Acc): 모델의 전반적인 분류 성능을 측정하기 위한 지표이다. 테스트 데이터셋 D_{test}에 대해 모델이 올바르게 분류한 이미지의 비율로 정의되며, 언러닝 이후에도 모델이 원래의 성능을 얼마나 잘 유지하는지를 평가한다.

$$\text{Acc} = \frac{1}{|D_{test}|} \sum_{(x_i, y_i) \in D_{test}} 1 \cdot \{f(x_i) = y_i\}$$

· 공격 성공률 (Attack Success Rate, ASR): 백도어 공격의 성공 정도 및 언러닝을 통한 공격 완화 효과를 측정하는 지표이다. 입력 패턴(Trigger)이 삽입된 이미지가 공격자가 의도한 타깃 라벨로 분류되는 비율로 정의된다. 언러닝이 효과적으로 수행될수록 이 수치는 낮아지게 된다.

$$\text{ASR} = \frac{1}{D_{backdoor}} \sum_{(x_i, y_i) \in D_{backdoor}} 1 \cdot \{f(x_i) = y_{\text{target}}\}$$

3. 실험 결과

· 정확도 측면 : 전체 시나리오에서 확률적 방식은 결정론적 방식보다 일관되게 높은 정확도를 유지하였다. 특히 non-I.I.D. 환경에서 결정론적 방식의 급격한 정확도 하락을 효과적으로 완화함으로써 모델의 안정성을 보였다.

· 공격 성공률 측면 : 전체 시나리오에서 확률적 방식은 결정론적 방식보다 완만한 감소 경향을 보였다. 이는 확률 가중치에 의한 영향으로 인한 것으로 확률적 방식이 결정론적 방식의 결과에 수렴하는 결과를 보였다. 이를 탐지 오차에 의한 정상 기여도의 과도한 손실을 방지하면서도 공격 영향을 소거할 수 있음을 보였다.

구분	Probabilistic			Deterministic
	$\alpha = 1$	$\alpha = 10$	$\alpha = 100$	
GD	76.38(0.15)	76.10(0.32)	75.65(0.29)	75.65(0.29)
GA	73.90(0.48)	**72.54(1.69)**	71.43(2.88)	**71.43(2.88)**
GDGA	73.26(1.95)	**71.15(3.81)**	70.38(4.18)	**70.38(4.18)**

[표1] 시나리오 정확도 결과(%), I.I.D

구분	Probabilistic			Deterministic
	$\alpha = 1$	$\alpha = 10$	$\alpha = 100$	
GD	69.23(15.80)	51.23(22.28)	45.62(27.38)	45.62(27.38)
GA	22.58(36.45)	**19.37(35.03)**	19.50(36.12)	**19.50(36.12)**
GDGA	18.96(36.46)	**18.52(36.63)**	18.51(36.78)	**18.51(36.78)**

[표2] 시나리오 공격 성공률 결과(%), I.I.D

구분	Probabilistic			Deterministic
	$\alpha = 1$	$\alpha = 10$	$\alpha = 100$	
GD	74.84(0.54)	73.49(0.81)	72.97(1.18)	72.97(1.18)
GA	69.95(1.08)	**62.14(8.69)**	51.81(21.91)	**51.81(21.91)**
GDGA	66.51(3.60)	**61.55(8.14)**	58.00(12.46)	**58.00(12.46)**

[표3] 시나리오 정확도 결과(%), non-I.I.D

구분	Probabilistic			Deterministic
	$\alpha = 1$	$\alpha = 10$	$\alpha = 100$	
GD	89.26(8.14)	72.29(22.22)	63.96(33.05)	63.96(33.05)
GA	23.56(37.19)	**20.18(37.00)**	19.71(37.18)	**19.71(37.18)**
GDGA	18.37(33.33)	**17.69(34.48)**	17.03(33.47)	**17.03(33.47)**

[표4] 시나리오 공격 성공률 결과(%), non-I.I.D

Ⅳ. 결론

본 연구는 연합학습 환경에서 악성 클라이언트 탐지의 불확실성 문제를 해결하기 위해 확률 기반 연합 언러닝 기법을 제안하였다. 실험 결과, 제안 기법은 평가 지표인 정확도와 공격 성공률 측면에서 기존 방식보다 뛰어난 균형을 보여주었다. 이는 실제 연합학습 시스템의 보안성과 신뢰성을 높이는 데 실질적인 기여할 것으로 기대된다.

참고문헌

[1] B. McMahan, et al., "Communication-Efficient Learning of Deep Networks from Decentralized Data," Proc. AISTATS, pp. 1273-1282, 2017.

[2] L. Bourtoule, et al., "Machine Unlearning," Proc. IEEE SP, pp. 141-159, 2021.

[3] E. Bagdasaryan, et al., "How To Backdoor Federated Learning," Proc. AISTATS, pp. 2938-2948, 2020.

[4] L. Zhang, et al., "FedRecovery: Differentially Private Machine Unlearning for Federated Learning Frameworks," IEEE Trans. Inf. Forensics Security, vol. 18, pp. 4732-4745, 2023.

[5] G. Liu, et al., "FedEraser: Enabling Efficient Client-Level Data Removal from Federated Learning Models," Proc. IEEE/ACM IWQoS, 2021.