연합학습에서의 모델 탈취 공격 방어를 위한 노이즈 기반 글로벌 모델 보호

권대호[1], 최봉준[2]
[1]숭실대학교 수학과
[2]숭실대학교 컴퓨터학과
{kwndh01, davidchoi}@soongsil.ac.kr

Noise-based Global Model Protection for Defending Against Model Theft Attacks in Federated Learning

Dae Ho Kwon[1], Bong Jun Choi[2]
[1]Dept. of Mathematics, Soongsil University
[2]School of Computer Science and Engineering, Soongsil University

요 약

크라우드소싱 기반 연합학습에서 글로벌 모델은 모델 제작 의뢰자의 중요한 자산이다. 하지만, 연합학습 플랫폼 운영 주체는 모델 학습을 위해 글로벌 모델의 파라미터를 클라이언트에게 공유해야 한다. 이 과정에서 악의적인 클라이언트로부터 모델 탈취 공격이 발생할 수 있다. 본 연구는 모델의 파라미터에 노이즈를 추가함으로써, 공유되는 모델의 정확도를 낮은 수준으로 유지하는 동시에 취합된 로컬 모델로부터 만들어진 글로벌 모델의 높은 정확도는 보장하고자 하는 데 목적이 있다.

1. 서론

크라우드소싱 기반 연합학습(Crowdsourcing-based Federated Learning)[1]은 모델 제작 의뢰자가 연합학습 플랫폼 운영 주체에 모델 제작을 요청하고, 플랫폼 운영 주체는 클라이언트를 모집해 글로벌 모델을 학습하는 새로운 연합학습 체계다.

크라우드소싱 기반 연합학습에서 글로벌 모델은 모델 제작 의뢰자의 중요한 자산이다. 하지만, 학습 라운드마다 모든 클라이언트에게 글로벌 모델이 공유되는 과정에서 악의적인 클라이언트가 모델을 탈취할 위험이 상존한다.

본 연구에서는 이러한 문제를 해결하기 위해, 글로벌 모델의 파라미터에 노이즈를 추가하여 학습 과정에서 클라이언트에게 배포되는 글로벌 모델의 정확도를 의도적으로 낮춤으로써 모델 탈취 공격을 효과적으로 방어하는 방안을 제시한다. 이러한 과정에서 최종적으로 모델 제작 의뢰자에게 전달되는 글로벌 모델의 정확도는 최대한 보존하여, 모델의 유용성을 저해하지 않고자 한다.

2. 노이즈 기반 글로벌 모델 보호

노이즈 기반 글로벌 모델 보호는 딥러닝 모델의 파라미터에 노이즈를 가해 클라이언트에게 배포되는 모델의 정확도를 떨어뜨리지만, 글로벌 모델의 정확도는 높이는 데 목적이 있다. 본 연구에선 세 개의 글로벌 모델 GM_s, GM_a, GM_b을 활용해 학습을 진행한다. GM_s는 모델 제작 의뢰자에게 인도될 글로벌 모델, GM_a과 GM_b는 학습 과정에서 클라이언트에게 공유되는 노이즈가 가해진 글로벌 모델이다. 정규분포 $N(0,\sigma^2)$를 따르는 랜덤한 노이즈를 생성한 후, 아래와 같이 GM_a, GM_b의 각각 첫 번째 컨볼루션층의 파라미터에 더하거나 뺀다.

$$w_{a,t} \leftarrow w_{a,t} + N(0,\sigma^2),\ w_{b,t} \leftarrow w_{b,t} - N(0,\sigma^2).$$

노이즈의 강도(σ)는 매 라운드 $GM_{a,t}$, $GM_{b,t}$의 정확도가 상한선을 넘기지 않도록 동적으로 정해진다.

이렇게 변형된 파라미터를 클라이언트에게 공유해 로컬에서 학습시킨 뒤, 학습된 로컬 모델과 서버사이드(server-side) 클라이언트 $k \in S_{s,t}$를 FedAvg[2]과정에 포함해 $w_{s,t}$를 업데이트한다. 서버사이드 클라이언트는 플랫폼 운영 주체가 $GM_{s,t-1}$의 파라미터를 가지고만 있도록 임의로 생성한 클라이언트다. 각 서버사이드 클라이언트의 샘플 수 n_k는 K클라이언트 중에서 한 명씩 뽑아 샘플 수만 전달받는다.

$w_{a,t}$와 $w_{b,t}$는 서버사이드 클라이언트를 포함하지 않고 로컬 모델로만 파라미터를 업데이트한다. 이후 다음 라운드에서 새로 생성되는 서버사이드 클라이언트는 $w_{s,t}$를 가지고, 학습에 참여하는 클라이언트는 $w_{a,t}$ 또는 $w_{b,t}$를 배포 받는다.

Algorithm 1: Noise-based Federated Averaging. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, η is the learning rate, C is the fraction of clients, n_k is the number of samples on client k and P_k is the set of indexes of data points on client k.
60 is a hyperparameter that serves as the upper bound of the accuracy for $GM_{a,t}$ and $GM_{b,t}$ in each round.

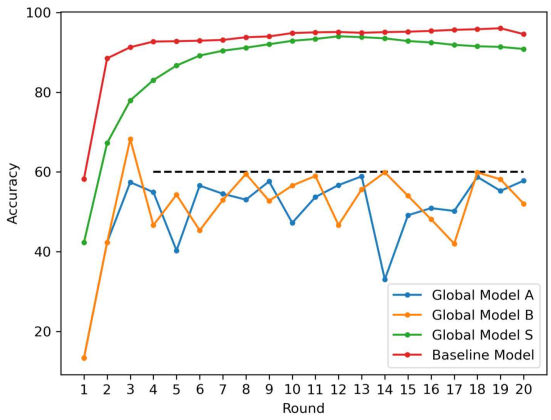
Server:
initialize w_0
for *each round* $t = 1,2,3,\ldots$ **do**
$m \leftarrow \max(C \cdot K, 1)$
$u \leftarrow 0$
$S_{a,t} \leftarrow$ (random set of $m/2$ clients)
$S_{b,t} \leftarrow$ (random set of $m/2$ clients)
$S_{s,t} \leftarrow$ (random set of $2m$ server-side clients)
while $Acc(GM_{a,t}) > 60$ **or** $Acc(GM_{b,t}) > 60$ **and** $t > 3$ **do**
$u \leftarrow u+1$
$w_{a,t}^{conv1} \leftarrow w_{a,t}^{conv1} + N(0,(u\sigma)^2)$//When re-entering the while
$w_{b,t}^{conv1} \leftarrow w_{b,t}^{conv1} - N(0,(u\sigma)^2)$ loop, the previous noises are removed.
for *each client* $k \in S_{a,t}$ **in parallel do**
$w_{a,t+1}^{k} \leftarrow \text{ClientUpdate}(k, w_{a,t})$
for *each client* $k \in S_{b,t}$ **in parallel do**
$w_{b,t+1}^{k} \leftarrow \text{ClientUpdate}(k, w_{b,t})$
if $t > 3$ **then**
$m_{s,t} \leftarrow \Sigma_{k \in S_{a,t} \sqcup S_{b,t} \sqcup S_{s,t}} n_k$
$w_{s,t+1} \leftarrow \Sigma_{k \in S_{a,t} \sqcup S_{b,t} \sqcup S_{s,t}} \frac{n_k}{m_{s,t}} w_{t+1}^{k}$
$m_t \leftarrow \Sigma_{k \in S_{a,t} \sqcup S_{b,t}} n_k$
$w_{t+1} \leftarrow \Sigma_{k \in S_{a,t} \sqcup S_{b,t}} \frac{n_k}{m_t} w_{t+1}^{k}$
$w_{a,t+1}$ and $w_{b,t+1} \leftarrow w_{t+1}$
if $t \leq 3$ **then**
$w_{s,t+1} \leftarrow w_{t+1}$
ClientUpdate(k, w):
$\beta \leftarrow$ (split P_k into batches of size B)
for *each local epoch* i *from* 1 *to* E **do**
for batch $b \in \beta$ **do**
$w \leftarrow w - \eta \nabla l(w;b)$
return w to server

악의적인 클라이언트가 학습 과정에서 획득한 $GM_{a,t}$, $GM_{b,t}$는 글로벌 모델$GM_{s,t}$의 정확도를 보장하지 않아 공격에 대한 수동적인 방어를 제공한다.

3. 성능 평가

MNIST 데이터를 활용하여 Convolutional Neural Network(CNN) 모델의 성능을 평가했다. 참가자는 100명이며 라운드마다 무작위로 10명을 추출했다. (그림 1)은 제안된 메커니즘을 통해 생성된 $GM_{s,t}$의 정확도와 매 라운드 클라이언트에게 공유되는 $GM_{a,t}$, $GM_{b,t}$의 정확도를 비교한 그래프이다.

실험 결과에 따르면, 모델 제작 의뢰자에게 전달될 $GM_{s,t}$는 Baseline Model과 유사한 수준의 높은 정확도를 보장하는 반면, 학습 과정에서 클라이언트에게 배포된 $GM_{a,t}$와 $GM_{b,t}$의 정확도는 60% 이하로 유지되어, 그 유용성이 제한적임을 확인할 수 있다.



(그림 1)

4. 결론

본 연구에서 제안한 글로벌 모델 파라미터에 랜덤 노이즈를 추가하는 방식은 로컬 모델의 성능 저하에 효과적이지만, 매 라운드 정확도 기반 노이즈 조절은 추가적인 연산 비용을 유발한다. 정확도 대신 다른 요소를 활용하여 현재의 정확도 상한선을 유지하면서 연산 비용을 절감할 수 있는 새로운 노이즈 조절 기법 개발이 필요하다. 따라서 후속 연구로에서는 제안된 알고리즘의 일반화를 통해 다양한 모델 구조와 데이터셋에 대한 적용 가능성을 검증해야 한다.

ACKNOWLEDGMENT

본 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2022 R1A2C4001270). 또한, 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음(IITP-2024-RS-2020-0-01602).

참고문헌

[1] S.R. Pandey, et al, “A Crowdsourcing Framework for On-Device Federated Learning”, IEEE Transactions on Wireless Communications, vol.19, no.5, pp.3241-3256, 2020
[2] McMahan, et al, “Communication-efficient Learning of Deep Networks from Decentralized Data”, Artificial Intelligence and Statistics, PMLR, pp. 1273-1282, 2017