https://doi.org/10.7840/kics.2024.49.2.308

다중 안테나 무선통신 기반 연합학습을 위한 차등 정보보호 기법

박 정 욱•, 윤 상 석°

Differential Privacy for Multiple Antenna Wireless Communication Assisted Federated Learning

Junguk Park•, Sangseok Yun°

요 약

본 논문에서는 무선통신 기반 연합 학습의 보안 취약점을 극복하는 차등 정보보호 기법에 관한 연구를 수행한다. 특히, 차등 정보보호 기법을 다중 안테나 무선통신의 관점에서 바라보고, 차등 정보보호 기법을 사용했을 때 발생하는 추론 정확도 감소를 최소화하는 빔 형성 기법을 제시한다. 또한, 제안하는 빔 형성 기법을 활용했을 때, 차등 정보보호를 보장하기 위한 전력 할당 기법을 제시하였다. 모의 실험을 통해 제안하는 기법이 연합 학습의 추론 성능 감소를 효과적으로 극복할 수 있음을 확인하였다.

Key Words : Dirferential privacy, federated learning, multiple antenna, wireless communication

ABSTRACT

In this work, we study differential privacy (DP) for federated learning over multiple antenna wireless communications. Specifically, we revisit DP from the perspective of wireless communications and propose an efficient beamforming scheme that can reduce inevitable degradation of inference accuracy due to intentional noises from DP. We also propose a power allocation scheme guaranteeing DP for the proposed beamforming scheme. Simulation results demonstrate that the proposed scheme can effectively prevent inference accuracy loss caused by DP.

Ⅰ. 서 론

연합 학습 (federated learning, FL)은 개인 데이터를 공개하지 않으면서도 다수 사용자의 개인 데이터를 활용해 협력적으로 신경망을 훈련시킬 수 있는 기법으로, 개인 데이터가 민감 정보를 포함하고 있어 훈련 데이터의 확보가 어려운 의료, 금융 등과 같은 분야에서 활발히 연구 및 활용되고 있다[1]. 하지만 최근 FL 과정에서 사용자가 중앙 서버로 전달하는 로컬 모델 및 기울기 (gradient) 정보로부터 각 사용자의 훈련 데이터를 복원할 수 있는 모델 전도 공격이 보고되었다[2].

이러한 공격을 방지하기 위한 기술로 차등 정보보호 (differential privacy, DP) 기법이 연구되고 있으나[3,4] DP 기법 적용시 추론의 정확도가 감소하는 치명적인 문제가 발생한다. 최근 다수의 FL 사용자들이 공간적으로 상관된 잡음을 전송하여 정확도 손실 없이 보안 위협을 극복하는 기법이 제안되었으나[5-6], 이를 위해서는 모든 FL 사용자가 서로 완벽한 동기 상태에 있다는 비현실적인 가정이 필수적이다. 이에 본 연구에서는 다중 안테나를 활용해 빔 형성을 수행함으로써 DP 기법에서 사용된 의도적 잡음으로 인한 추론 정확도 감소를 저감하는 기법을 제시하고, 모의 실험을 통해 제안 기법의 성능을 검증하였다.

Ⅱ. 시스템 모델

2.1 연합 학습 및 모델 전도 공격

본 논문에서는 M명의 사용자가 FL을 수행하는 환경을 고려한다. 즉, i번째 FL 라운드에서 중앙 서버는 이전 라운드에 사용자로부터 수신한 기울기를 활용해 수식 (1)과 같이 새로운 글로벌 모델 $\mathbf{w}^i$를 생성하고 다시

※ 이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No. 2022-0-00088, MEC (Mobile Edge Computing) 에지 클라우드 기반 초저지연 블록체인 서비스 플랫폼 핵심 기술개발)과 한국연구재단의 지원 (No. NRF-2021R1G1A1094982)을 받아 수행된 연구임

• First Author : Korea Advanced Institute of Science and Technology, Information & Electronics Research Institute, pjuk@kaist.ac.kr, 연수연구원, 정회원

° Corresponding Author : Pukyong National University, Department of Information and Communications Engineering, ssyun@pknu.ac.kr, 조교수, 정회원

논문번호 : 202310-101-C-LU, Received October 12, 2023; Revised October 30, 2023; Accepted October 30, 2023

사용자에게 배포한다.

$$\mathbf{w}^i = \mathbf{w}^{i-1} + \eta|D|^{-1}\sum_{m=1}^{M}\Delta\mathbf{w}_m^{i-1}|D_m|. \quad (1)$$

여기서 $|D_m|$과 $\Delta\mathbf{w}_m^i$는 사용자 m의 훈련 데이터 샘플의 수 및 i번째 라운드에 전송한 기울기이며, η는 학습률, $|D| = \sum_{m=1}^{M}|D_m|$이다. 한편, 사용자 m은 수신한 모델 $\mathbf{w}^i$ 및 자신의 훈련 데이터를 활용해 i번째 라운드의 기울기 $\Delta\mathbf{w}_m^i$를 계산하고 이를 중앙 서버에 전송한다. FL에서는 전술한 과정을 모델이 수렴할 때까지 반복 수행한다.

모델 전도 공격은 중앙 서버에서 사용자로 전송되는 글로벌 모델 및 각 사용자가 서버에게 전송하는 기울기 정보로부터 특정 사용자의 훈련 데이터를 복원하는 공격이다. FL이 I라운드 수행되는 경우, 공격자는 각 라운드 $i \in \{1, \cdots, I\}$에 대한 $\mathbf{w}^i$ 및 $\Delta\mathbf{w}_m^i$들을 관측해 사용자 m의 훈련 데이터 복원을 시도한다[2].

2.2 차등 정보보호

DP 기법은 모델 전도 공격을 방지하기 위한 기술로 기울기 클리핑을 통해 기울기 정보에 대한 특정 사용자 데이터 샘플의 영향을 제한하고, 또한 기울기 정보에 의도적으로 잡음을 더해 민감 정보 마스킹을 수행한다[3]. 전술한 FL 환경에서 $d = \min_i|D_i|$라 할 때, 각 사용자의 최대 기울기가 C가 되도록 클리핑을 수행한 후, 평균이 0이고 표준편차 σ가 아래와 같은 가우시안 잡음을 더해 서버로 전송하면 (ϵ, δ)-DP를 달성할 수 있음이 알려져 있다[4].

$$\sigma = \frac{2\sqrt{2\ln(1.25/\delta)}IC}{d\epsilon} \quad (2)$$

이때, (ϵ, δ)-DP는 특정 데이터 샘플의 유무에 대한 로그 가능도 비 (log-likelihood ratio)가 ϵ보다 작지 않을 확률이 δ보다 작다는 의미로, DP의 관점에서 낮은 ϵ과 δ 값에 대해 (ϵ, δ)-DP를 만족하는 경우 더욱 높은 보안을 달성하는 것을 의미한다.

하지만 이를 위해서는 기울기 정보에 더 큰 잡음을 더해야 하므로 모델의 수렴이 느려지고 추론 정확도가 감소하게 된다. 따라서 이를 극복하기 위한 DP 기법을 다음 장에서 제안한다.

Ⅲ. 제안하는 기법

3.1 다중 안테나 무선통신 기반 연합학습

연합 학습의 성능 향상을 위해서는 다수 사용자의 참여가 필수적이다. 따라서 본 논문에서는 다수의 사용자 참여를 유도하기 용이한 4G/5G 등의 무선통신 단말을 활용한 연합 학습 환경을 고려한다. 특히, 다중 안테나를 보유한 무선통신 단말 기반 FL을 고려하며, FL 사용자는 T개의 안테나, 중앙 서버와 공격자는 단일 안테나를 가지고 있다고 가정한다. 사용자 m은 i번째 FL 라운드의 송신 신호 $\mathbf{x}_m^i = \mathbf{f}_m^i s_m^i + \mathbf{Z}_m^i\mathbf{w}_m^i$를 중앙 서버로 전송한다. 이 때, s_m^i는 기울기 $\Delta\mathbf{w}_m^i$의 아날로그-진폭 변조 송신 심볼, $\mathbf{t}_m$은 DP를 위한 $(T\text{-}1)\times 1$ 잡음 벡터이며 $\mathbf{t}_m$의 각 성분 $t_m \sim \mathcal{N}(0, (T-1)^{-1})$이다. 또한, $\mathbf{f}_m^i$와 $\mathbf{N}_m^i$는 각각 기울기와 잡음 신호를 위한 $T \times 1$ 빔 형성 벡터 및 $T\times(T\text{-}1)$ 빔 형성 행렬이다. 수식의 간결성을 위해 지금부터 명확한 경우 FL 라운드 인덱스 i는 생략한다.

사용자 m과 중앙 서버, 그리고 공격자 사이의 채널을 각각 $\mathbf{h}_m^T$, $\mathbf{g}_m^T$라 하면, 중앙 서버와 공격자가 사용자 m으로부터 수신한 신호 y_m과 z_m은 수식 (3)과 같이 나타낼 수 있다.

$$\begin{aligned} y_m &= \mathbf{h}_m^T(\mathbf{f}_m s_m + \mathbf{N}_m\mathbf{t}_m) + u_m, \\ z_m &= \mathbf{g}_m^T(\mathbf{f}_m s_m + \mathbf{N}_m\mathbf{t}_m) + v_m. \end{aligned} \quad (3)$$

이 때, 열 잡음 $u_m \sim \mathcal{N}(0, \sigma_u^2)$, $v_m \sim \mathcal{N}(0, \sigma_v^2)$이다.

본 논문에서는 내부 공격자를 고려하며, 따라서 중앙 서버와 공격자는 채널 추정을 통해 $\mathbf{h}_m^T$와 $\mathbf{g}_m^T$를 모두 알고 있다. 중앙 서버와 공격자는 송신 심볼의 검출을 위해 zero-forcing equalization을 수행하며, 중앙 서버와 공격자의 추정 송신 심볼 $\hat{s}_m$, $\hat{s}_{m,e}$는 수식 (4)와 같이 나타내어진다.

$$\begin{aligned} \hat{s}_m &= s_m + (\mathbf{h}_m^T\mathbf{f}_m)^{-1}(\mathbf{h}_m^T\mathbf{N}_m\mathbf{t}_m + u_m), \\ \hat{s}_{m,e} &= s_m + (\mathbf{g}_m^T\mathbf{f}_m)^{-1}(\mathbf{g}_m^T\mathbf{N}_m\mathbf{t}_m + v_m). \end{aligned} \quad (4)$$

중앙 서버는 추정 심볼 $\hat{s}_m^{i-1}$을 활용해 수식 (1)의 $\Delta\mathbf{w}_m^{i-1}$을 대체함으로써 글로벌 모델을 갱신한다.

3.2 제안하는 빔 형성 기법

본 논문에서 제안하는 기울기 빔 형성 벡터 $\bar{\mathbf{f}}_m$는 채널 $\mathbf{h}_m^T$에 대한 maximum ratio transmission 벡터이고, 잡음 빔 형성 벡터 $\bar{\mathbf{N}}_m$는 채널 $\mathbf{h}_m^T$의 영 공간 (null space)을 생성하는 기저 벡터로 구성된 행렬이다. 이처럼 빔 형성을 수행하는 경우, 중앙 서버의 실효 잡음 신호 $\mathbf{h}_m^T\bar{\mathbf{N}}_m\mathbf{t}_m = 0$이 되어 중앙 서버는 기울기 정보를

정확하게 수신할 수 있는 반면 공격자는 잡음 성분 $\mathbf{g}_m^T\bar{\mathbf{N}}_m\mathbf{t}_m \neq 0$을 수신하므로 공격자의 채널만을 선택적으로 열화 시킬 수 있다. 여기서 송신 전력 제약을 위해 기울기 빔 형성 벡터 $\mathbf{f}_m = \sqrt{\phi}\bar{\mathbf{f}}_m$, 잡음 빔 형성 행렬 $\mathbf{N}_m = \sqrt{1-\phi}\bar{\mathbf{N}}_m$로 신호를 전송하며, ϕ는 신호와 잡음 신호의 전력 할당 계수이다.

공격자의 추정 송신 심볼 $\hat{s}_{m,e}$을 활용해 모델 전도 공격을 수행한다. 추정 송신 심볼 $\hat{s}_{m,e}$은 송신 심볼 s_m에 실효 잡음 $\tilde{v}_m$이 더해진 형태이다. 공격자는 채널 $\mathbf{g}_m^T$와 $\mathbf{h}_m^T$ (즉, 빔 형성 벡터 $\mathbf{f}_m$와 $\mathbf{N}_m$)에 대한 정보를 가지고 있으므로, 실효 잡음의 첫 번째 항 $\tilde{v}_{m,1} = (\mathbf{g}_m^T\mathbf{f}_m)^{-1}\mathbf{g}_m^T\mathbf{N}_m\mathbf{t}_m$과 두 번째 항 $\tilde{v}_{m,2} = (\mathbf{g}_m^T\mathbf{f}_m)^{-1}v_m$의 조건부 분포 (conditional distribution)는 아래 수식 (5)와 같이 나타내어진다.

$$\tilde{v}_{m,1}\sim\mathcal{N}\left(0,(1-\phi)|\mathbf{r}_m|^2/\left(\phi\bar{f}_{g,m}^2(T-1)\right)\right)$$
$$\tilde{v}_{m,2}\sim\mathcal{N}\left(0,f_{g,m}^{-2}\sigma_v^2\right) \quad (5)$$

단, 여기서 $\mathbf{r}_m = \mathbf{g}_m^T\mathbf{N}_m$을 나타내며 $\bar{f}_{g,m} = \mathbf{g}_m^T\bar{\mathbf{f}}_m$을 의미한다. 결과적으로 도청자의 실효 잡음의 분산은 실효 잡음의 첫 번째 항 $\tilde{v}_{m,1}$과 두 번째 항 $\tilde{v}_{m,2}$의 분산의 합으로 결정된다.

2.2장에서 서술한 바와 같이 공격자가 수신하는 실효 잡음의 분산이 σ^2보다 큰 경우 제안 빔 형성 기법은 (ϵ,δ)-DP를 달성할 수 있다. 간단한 수식 변형을 통해 이러한 조건을 달성하기 위한 전력 할당 계수 ϕ의 조건을 다음과 같이 찾을 수 있다.

$$\phi \leq \min\left[1,\frac{\epsilon^2d^2\left(|\mathbf{r}_m|^2+\sigma_v^2(T-1)\right)}{\epsilon^2d^2|\mathbf{r}_m|^2+8\ln(1.25/\delta)I^2C^2\bar{f}_{g,m}^2}\right]$$

Ⅳ. 모의실험

제안 기법의 성능 검증을 위해 다중 안테나 무선 환경을 고려하지 않고 각 단말이 기울기 정보에 $\mathcal{N}(0,\sigma^2)$를 따르는 잡음을 삽입하는 기존 DP 기법의 성능을 함께 평가하였다. 실험을 위해 단말의 수 M= 10, 연합 학습의 학습률 η = 0.0001, 사용자의 안테나 수 T= 10, 중앙 서버 및 도청자의 열 잡음의 분산 $\sigma_u^2 = \sigma_v^2 = 1$, 기울기 클리핑 계수 C= 1.4, 목표 보안 계수 ϵ = 5, δ = 0.01인 환경을 고려하였다. 또한, 무선 채널 $\mathbf{h}_m^T$와 $\mathbf{g}_m^T$는 Rayleigh 분포에 따라 생성하였다.

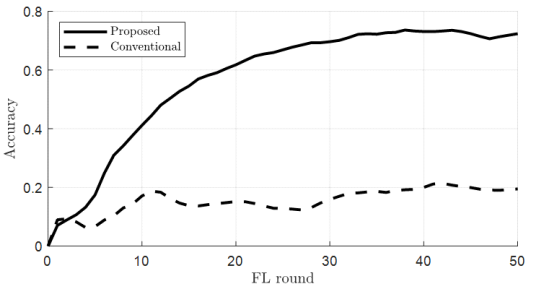

그림 1. 제안 기법 및 기존 기법의 성능 비교
Fig. 1. Performance comparison between the proposed scheme and the conventional scheme.

한편, 연합 학습의 추론 성능 평가를 위한 인공 신경망은 완전 연결 신경망 (Fully-connected network)을 사용했으며, 은닉 노드의 수는 128개, 활성화 함수로는 ReLU 활성화 함수를 사용하였다. 또한, 학습을 위한 훈련 데이터로는 i.i.d. (independent and identically distributed) MNIST 데이터를 사용했으며, 따라서 입력 노드의 수는 784개이다. 총 M= 10명의 사용자를 고려하므로 각 사용자의 훈련 데이터 수 $|D_1| = \cdots = |D_M| = 6000$이다.

제안 기법과 기존 DP 기법의 FL 라운드 변화에 따른 추론 정확도를 그림 1에 도시하였다. 기존 기법의 경우 (ϵ,δ)-DP 달성을 위해 인공적으로 삽입한 잡음에 의해 모델의 훈련 성능이 감소하나 제안 기법의 경우 (ϵ,δ)-DP 달성을 보장하면서도 다중 안테나 기반 빔 형성에 의해 DP를 위한 잡음이 상쇄되어 훈련이 잘 진행되는 것을 확인할 수 있다.

Ⅴ. 결 론

본 논문에서는 무선통신 기반 FL의 보안 취약점을 극복하기 위해 다중 안테나를 활용한 DP 기법을 제안하였다. 특히, DP 기법 사용시 불가피하게 감소하는 추론 정확도를 최소화하기 위한 효율적인 빔 형성 기법을 제시하였고, 제안한 빔 형성 기법 사용시 공격자가 받게 되는 유효 잡음 신호의 전력을 분석함으로써 보안 취약점을 극복할 수 있음을 보였다. 향후 연구 방향으로, 다중 안테나 외부 공격자 및 신뢰할 수 없는 중앙 서버가 존재하는 환경을 위한 무선 DP 기법에 관한 연구를 수행할 예정이다.

References

[1] H. B. McMahan, E. Moore, D. Ramage, S.

Hampson, and B. A. Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *Proc. AISTATS 2017*, pp. 1-11, Fort Lauderdale, FL, USA, Apr. 2017.
(https://doi.org/10.48550/arXiv.1602.05629)

[2] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, “Inverting Gradients - How easy is it to break privacy in federated learning?,” in *Proc. NeurIPS 2020*, pp. 1-11, Vancouver, Canada, Dec. 2020.
(https://doi.org/10.48550/arXiv.2003.14053)

[3] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, “Deep Learning with Differential Privacy,” in *Proc. CCS 2016*, pp. 308-318, New York, NY, USA, Oct. 2016.
(https://doi.org/10.1145/2976749.2978318)

[4] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. S. Quek, and H. Vincent Poor, “Federated learning with differential privacy: Algorithms and performance analysis,” *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 3454-3469, Apr. 2020.
(https://doi.org/10.1109/TIFS.2020.2988575)

[5] J. Liao, Z. Chen, and E. G. Larsson, “Over-the-air federated learning with privacy protection via correlated additive perturbations,” in *Proc. Allerton 2022*, pp. 1-8, Monticello, IL, USA, Sep. 2022.
(https://doi.org/10.1109/Allerton49937.2022.9929413)

[6] L. Maßny and A. W.-Zeh, “Secure over-the-air computation using zero-forced artificial noise,” in *Proc. ITW 2023*, pp. 1-6, Saint-Malo, France, Apr. 2023.
(https://doi.org/10.48550/arXiv.2212.04288)