# 지식 증류 기반 연합학습의 강건성 평가

조윤기[1], 한우림[1], 유미선[1], 윤수빈[1], 백윤흥[1]
[1]서울대학교 전기정보공학부, 반도체공동연구소
ygcho@sor.snu.ac.kr, wrhan@sor.snu.ac.kr, msyu@sor.snu.ac.kr, sbyun@sor.snu.ac.kr, ypaek@snu.ac.kr

# A Evaluation on Robustness of Knowledge Distillation-based Federated Learning

Yun-Gi Cho[1], Woo-Rim Han[1], Mi-Seon Yu[1], Su-bin Yun[1], Yun-Heung Paek[1]
[1]Department of ECE and ISRC, SNU

요 약

연합학습은 원본 데이터를 공유하지 않고 모델을 학습할 수 있는 각광받는 프라이버시를 위한 학습 방법론이다. 이를 위해 참여자의 데이터를 수집하는 대신, 데이터를 인공지능 모델 학습의 요소들(가중치, 기울기 등)로 변환한 뒤, 이를 공유한다. 이러한 강점에 더해 기존 연합학습을 개선하는 방법론들이 추가적으로 연구되고 있다. 기존 연합학습은 모델 가중치를 평균내는 것으로 참여자 간에 동일한 모델 구조를 강요하기 때문에, 참여자 별로 자신의 환경에 알맞은 모델 구조를 사용하기 어렵다. 이를 해결하기 위해 지식 증류 기반의 연합학습 방법(Knowledge Distillation-based Federated Learning)으로 서로 다른 모델 구조를 가질 수 있도록(Model Heterogenousity) 하는 방법이 제시되고 있다.

연합학습은 여러 참여자가 연합하기 때문에 일부 악의적인 참여자로 인한 모델 포이즈닝 공격에 취약하다. 수많은 연구들이 기존 가중치를 기반으로한 연합학습에서의 위협을 연구하였지만, 지식 증류 기반의 연합학습에서는 이러한 위협에 대한 조사가 부족하다. 본 연구에서는 최초로 지식 증류 기반의 연합학습에서의 모델 성능 하락 공격에 대한 위협을 실체화하고자 한다. 이를 위해 우리는 GMA(Gaussian-based Model Poisoning Attack)과 SMA(Sign-Flip based Model Poisoning Attack)을 제안한다. 결과적으로 우리가 제안한 공격 방법은 실험에서 최신 학습 기법에 대해 평균적으로 모델 정확도를 83.43%에서 무작위 추론에 가깝게 떨어뜨리는 것으로 공격 성능을 입증하였다. 우리는 지식 증류 기반의 연합학습의 강건성을 평가하기 위해, 새로운 공격 방법을 제안하였고, 이를 통해 현재 지식 증류 기반의 연합학습이 악의적인 공격자에 의한 모델 성능 하락 공격에 취약한 것을 보였다. 우리는 방대한 실험을 통해 제안하는 방법의 성능을 입증하고, 결과적으로 강건성을 높이기 위한 많은 방어 연구가 필요함을 시사한다.

## 1. 서론

최근 AI는 급속도로 발전하여 나날이 혁신을 일으키고 있다. 이러한 AI의 성능 발전은 데이터에 기반을 두고 있는데, 즉 학습 데이터의 크기가 클수록 더 좋은 AI성능을 보인다. 그러나 최근 GDPR과 같은 강화된 개인정보보호법들은 데이터 수집을 어렵게하여, 결과적으로 AI 성능을 향상시키는 데 걸림돌이 된다. 이러한 어려움을 해결하기 위해 데이터를 직접 수집하지 않고 AI 모델을 학습시킴으로 프라이버시를 보존하고자 하는 새로운 학습 패러다임인 연합학습이 등장하였다.

데이터를 공유하지 않는 대신, 연합학습의 각 참여자는 학습의 요소들(e.g., 가중치, 모델 파라미터, 모델 예측 등)을 교환하여 학습한다. 기존에 많이 사용되던 방식은 단순히 각 참여자가 로컬 모델을 가지고 자신의 데이터에 학습한 뒤, 각 로컬 모델 파라미터들을 합산하여, **합의된 모델 파라미터**를 다시 나누어 가지는 방식이다. 그러나 이러한 방식은 참여자 간에 동일한 모델 구조를 강제한다. 각 참여자는 자신의 상황에 알맞은 모델 구조가 있더라도 합의된 모델을 따라야 하기 때문에 작은 규모의 모델 사용이 필요하더라도 큰 모델을 사용해야하는 등 여러 문제가 발생한다.

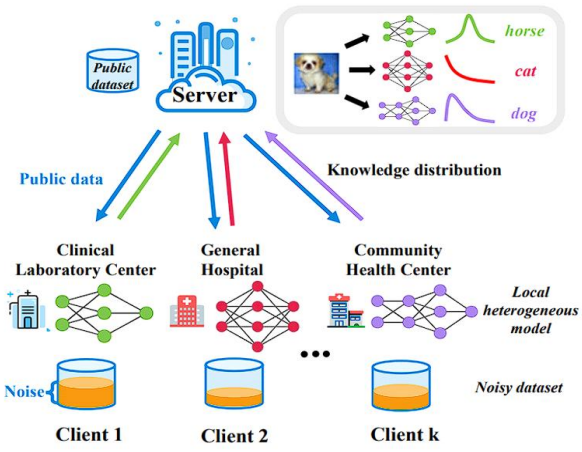
이에 다양한 모델 구조로 연합학습을 진행할 수 있도록 지식 증류 기반의 연합학습(KD-FL)이 등장하였다. 각 참여자의 모델이 공공데이터 셋(public set)에 대하여 동일한 결과값을 내도록 한다. 구체적으로, 먼저 로컬 모델을 각자의 프라이빗 데이터에 학습한다. 이후 모델 가중치가 아닌 공공데이터 셋에

대한 서로 다른 로컬 모델의 출력값을 공유하고 평균낸 뒤, 서로 다른 로컬 모델이 **합의된 출력값**(평균값)을 출력하도록 다시 학습한다. 수렴할 때까지 이를 반복한다.

한편, 연합학습과 같은 분산 컴퓨팅은 연산 참여자들 중 악의적인 참여자로 인한 보안 위협이 존재한다[1]. 연합학습 또한 이러한 위협이 존재하는데, 악의적인 참여자가 학습을 방해하여 최종적으로 모델 성능을 떨어뜨릴 수 있다[2,3,4]. 이는 안전한 연합학습을 위해 고려해야하는 심각한 위협으로 여겨진다. 그러나 기존 가중치를 더하는 방식의 연합학습에서는 많은 연구가 이루어졌지만, 지식 증류 기반 연합학습(이하 KD-FL)에서는 아직 탐구되지 않았다. *본 논문에서는 최초로 KD-FL에서 악의적인 참여자에 대한 위협을 탐구하고자 한다.*

기존 연합학습에 대한 공격은 업로드하는 모델 파라미터를 오염시켰는데[2], 모델 파라미터는 공유되지 않기 때문에 새로운 공격 방식이 고려되어야 한다. 모델 파라미터 대신 우리는 공유되는 퍼블릭 데이터에 대한 출력값을 오염시킨다. 이를 위해 우리는 이전 연구들[3,8,9]에서 영감받아 GMA(Gaussian-based Model Poisoning Attack)과 SMA(Sign-Flip based Model Poisoning Attack)를 제안한다. 각 공격은 perturbation을 넣는 방식에 차이가 존재한다. 가우시안 노이즈는 예기치 못한 변화를 야기하여 모델 학습을 방해할 수 있다[9]. 이를 활용하여 GMA는 가우시안 노이즈를 합의된 출력값에 더하는 방식으로 다른 모델 학습을 방해한다. 또 다른 방법으로는 가중치의 부호를 뒤집는 방식으로, 학습의 방향성을 뒤집기 때문에 완전히 반대 방향으로 학습을 유도하여 모델 성능을 감소시키는 방식이 있다[3]. 이를 기반으로 SMA는 공유되는 출력값의 부호를 뒤집어 다른 모델의 학습을 방해한다. 여기에 더해 우리는 업로드하는 출력값을 N배 크기를 키워 업로드 한다. 이처럼 큰 값을 업로드 할 경우, 상대적으로 작은 정직한 다른 참여자의 업로드한 값의 영향력을 줄일 수 있다[8].

본 연구에서는 방대한 실험을 통해 제안하는 GMA와 SMA의 효용성을 입증하였다. 널리 사용되는 프라이버시가 중요한 개인 금융 및 정보 데이터(Adult, Census)[10]를 통한 실험에서는 모델 성능을 무작위 추론에 가깝게 떨어뜨리는 것을 보였다. 이를 통해 *우리는 현재 KD-FL의 취약성을 제안하는 공격을 통해 입증함으로 앞으로의 연구에 대한 방향성을 제시하였다.*




(그림 1) 지식 증류 기반 연합학습 개요도 [7].

## 2. 지식 증류 기반 연합학습(KD-FL)

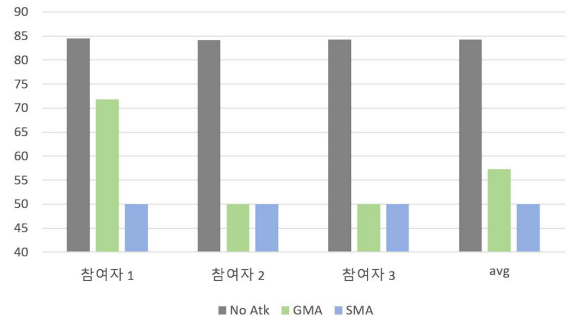
FedMD[5]는 KD-FL 알고리즘들 중 가장 기저로 사용되는 학습 알고리즘이다[6,7]. 그림 1에서 보이듯, 참여자들과 학습을 도울 하나의 서버가 존재한다. 각 참여자는 각자의 프라이빗 데이터셋을 가지고 있고, 학습을 위해 퍼블릭 데이터셋이 존재한다. 가장 먼저 각 참여자는 자신의 로컬 모델을 프라이빗 데이터셋으로 학습한다. 그 뒤 퍼블릭 데이터셋에 대한 로컬 모델의 출력값을 서버에 업로드한다. 서버는 업로드된 참여자들의 출력값을 평균낸 뒤, 다시 배포한다. 평균 출력값은 합의된 출력값이라 칭한다. 이후 참여자들은 자신의 로컬 모델이 퍼블릭 데이터에 대응되는 합의된 출력값을 출력하도록 학습시킨다. 이를 반복한다.

## 3. 위협 모델

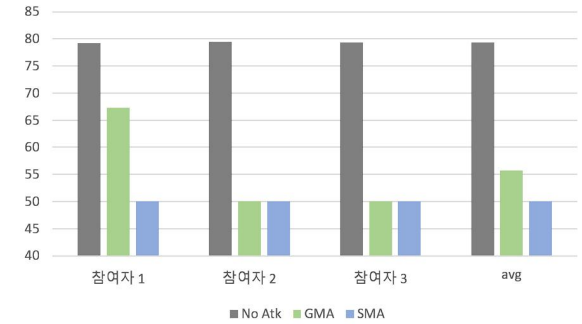
**공격자의 목표.** 공격자는 참여자 중 한 명으로, 공격자의 목표는 다른 참여자의 로컬 모델이 최종적으로 올바르게 기능하지 못하도록 하는 것이다.

**공격자의 능력.** 공격자는 자신의 모든 로컬 학습 과정을 조작할 수 있다. 하지만 연합하여 진행해야하는 모든 프로토콜은 완벽히 따라야하며, 서버에 접근할 수 없고, 다른 참여자들의 데이터 또한 모델에 대한 접근도 불가능하다.

**공격자의 지식.** 공격자는 자신의 로컬 모델과 프라이빗 데이터를 가지고 있다. 공격자는 모델 학습의 목적과 연합학습 프로토콜을 알고 있다.




(그림 2) Adult 데이터셋에 대한 실험 결과 (정확도, 낮을수록 효과적인 공격)



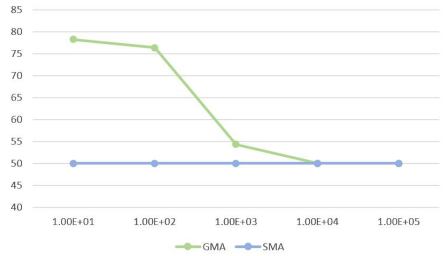

(그림 3) Census에서의 실험 결과 (정확도, 낮을수록 효과적인 공격)

## 4. 방법론

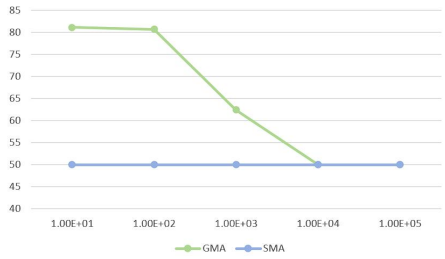
우리는 이 장에서 GMA와 SMA의 방법을 자세하게 서술한다. 앞서 설명한 것과 같이 GMA와 SMA는 perturbation을 넣는 방식에 차이가 존재한다.

공격자는 먼저 자신의 로컬 모델을 자신의 프라이빗 데이터로 학습시킨다. 그 후 KD-FL의 프로토콜에 따라 퍼블릭 데이터에 대한 출력값을 업로드한다. 이때 GMA공격은 공격자 모델의 출력값에 가우시안 노이즈를 더하게 된다. 더해진 노이즈는 최종적으로 합의된 출력값에 영향을 준다. 다른 참여자들의 로컬 모델이 노이즈가 낀 합의된 출력값을 뱉도록 학습하게된다. 결과적으로 각 로컬 모델은 입력값에 대응되는 값이 아닌 단순 노이즈를 출력하도록 하기 때문에 모델 성능을 감소시킬 수 있다.

SMA 공격은 공격자 모델에서 퍼블릭 데이터에 대한 출력값의 부호를 바꾼 뒤 서버에 업로드한다. 마찬가지로 이러한 perturbation을 최종적으로 다른 참여자들의 로컬 모델이 오염된 출력값을 출력하도록 학습하게 된다. 이때 오염된 출력값은 올바른 출력의 반대이기 때문에 오히려 다른 참여자들의 모델은 올바르지 않은 출력을 하게되어 제대로 기능하지 못하게 된다.




(그림 4) Adult에서 $\mu$의 변화에 따른 실험 (정확도, 낮을수록 효과적인 공격)




(그림 5) Census에서 $\mu$의 변화에 따른 실험 (정확도, 낮을수록 효과적인 공격)

단순히 perturbation을 주는 것은 다른 참여자들의 학습 과정을 제대로 방해하지 못할 수 있다. 이를 해결하기 위해 우리는 logit-replacement approach를 제안한다. 이는 [9]에서 영감을 받아, 벡터의 평균을 내는 과정에서 특정 벡터의 크기를 비정상적으로 키움으로, 다른 작은 벡터의 영향력을 효과적으로 줄이고, 최종적으로는 평균 벡터를 특정 벡터로 교체하는 접근법이다. 결과적으로 수식은 다음과 같다.

$$\hat{v} = \mu \cdot p + v \qquad (1)$$

$v$는 공격자의 원본 출력값이고, $\mu$는 logit-replacement를 위한 하이퍼파라미터이다. $p$는 공격 방식(GMA, SMA)에 따른 perturbation 벡터를 의미한다. 이러한 방식을 통해 다른 로컬 모델에 대해 효과적으로 학습을 방해할 수 있다.

## 5. 실험

**실험 세팅.** 실험에 사용되는 데이터셋은 Adult와 Census이다. 이는 개인의 금융 정보, 직업, 나이, 학력 등 으로 이루어진 데이터이다. [7]을 따라 총 참여자는 4명으로 이 중 0번 참여자는 공격자이다. 각 로컬 모델은 5-layer의 fully-connected network를 사용한다. 각 로컬 모델은 로컬 프라이빗 데이터셋에 대해 1 epoch 학습한 뒤, 퍼블릭 데이터셋에 대

한 출력값을 합의하여 1 epoch 학습한다. 이를 반복하는 global epoch은 10회를 사용한다. 공격을 위해서 $\mu$는 기본적으로 100을 사용한다. 성능 비교를 위한 지표(metric)는 정확도를 사용한다. 이는 전체 테스트 샘플 중 올바르게 추론할 샘플의 확률을 나타낸다. 낮을수록 공격이 효과적인 것을 나타낸다.

**메인 실험.** 메인 실험의 결과는 그림 2, 3에서 나타난다. 대부분의 공격에서 다른 참여자들의 로컬 모델의 성능이 무작위 추론(50%)에 도달하는 것을 볼 수 있다. 하지만 GMA는 단순히 무작위성을 야기하기 때문에 완전히 틀린 출력만 내도록하는 SMA가 GMA보다 공격 성능이 더 좋은 것을 나타낸다.

**Ablation Study.** 제안하는 logit-replacement approach는 하이퍼파라미터로 $\mu$를 사용한다. 따라서 우리는 해당 하이퍼파라미터의 변화에 따른 공격 성능을 실험하여, logit-replacement attack의 효과를 입증한다. 그림 4, 5와 같이 $\mu$가 증가할수록 공격 성능이 더 올라가는 것을 보인다. 이는 해당 접근법이 공격을 유의미하게 강력하게 할 수 있다는 것을 의미한다.

## 6. 결론

본 논문에서는 KD-FL에서 악의적인 공격자에 대한 위협을 평가하기 위해 GMA와SMA를 제안한다. 이를 통해 현재 KD-FL의 취약함을 보이고, 방대한 실험을 통해 입증한다. 이를 통해 앞으로 안전한 KD-FL을 위해 연구해야 할 방향성을 시사한다.

## 6. ACKNOWLEMENT

이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (RS-2023-00277326). 이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (IITP-2023-RS-2023-00256081). 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2022-0-00516, 국가통계데이터에 적용 가능한 차등 정보보호 개념을 도출하고 통계분석의 유용성을 보장해야 하는 문제 해결). 이 논문은 2024년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의하여 지원되었음. 본 연구는 반도체 공동연구소 지원의 결과물임을 밝힙니다.

## 참고문헌

[1] Lamport, Leslie, Robert Shostak, and Marshall Pease. "The Byzantine generals problem." Concurrency: the works of leslie lamport. 2019. 203-226.

[2] Fang, Minghong, et al. "Local model poisoning attacks to {Byzantine-Robust} federated learning." 29th USENIX security symposium (USENIX Security 20). 2020.

[3] Shejwalkar, Virat, and Amir Houmansadr. "Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning." NDSS. 2021.

[4] Lee, Younghan, et al. "FLGuard: Byzantine-Robust Federated Learning via Ensemble of Contrastive Models." European Symposium on Research in Computer Security. Cham: Springer Nature Switzerland, 2023.

[5] Li, Daliang, and Junpu Wang. "Fedmd: Heterogenous federated learning via model distillation." arXiv preprint arXiv:1910.03581 (2019).

[6] Takahashi, Hideaki, Jingjing Liu, and Yang Liu. "Breaching FedMD: image recovery via paired-logits inversion attack." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[7] Fang, Xiuwen, and Mang Ye. "Robust federated learning with noisy and heterogeneous clients." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[8] Bagdasaryan, Eugene, et al. "How to backdoor federated learning." International conference on artificial intelligence and statistics. PMLR, 2020.

[9] Blanchard, Peva, et al. "Machine learning with adversaries: Byzantine tolerant gradient descent." Advances in neural information processing systems 30 (2017).

[10]Chaudhari, Harsh, et al. "SNAP: Efficient extraction of private properties with poisoning." 2023 IEEE Symposium on Security and Privacy (SP). IEEE, 2023.