https://doi.org/10.7840/kics.2023.48.11.1387

모델 포이즈닝 공격에 강건한 개인화 연합학습을 위한 부분 공유 알고리즘

박 희 원◆, 김 미 르*, 권 민 혜°

Robust Partial Share Federated Learning Algorithm against Model Poisoning Attack

Heewon Park◆, Miru Kim*, Minhae Kwon°

요 약

엣지 디바이스 기술의 발전과 상용화로 인해 방대한 양의 분산된 데이터가 증가하고 있으며, 이와 함께 연합학습은 분산된 데이터 환경에 적합한 인공지능 학습 기술로 활발히 연구되고 있다. 연합학습은 여러 디바이스에 위치한 데이터의 노출 없이 인공지능 모델을 학습할 수 있는 기술이다. 하지만 기존 연합학습 방식은 데이터 분포의 특성이 상이한 디바이스가 학습에 참여할 시 개별 데이터에 최적화된 모델을 만들 수 없다는 점과 비잔틴 공격에 취약하다는 한계가 있다. 이러한 한계점들을 극복하기 위하여 본 논문에서는 새로운 부분공유 알고리즘을 제안한다. 부분공유 알고리즘은 각 디바이스의 로컬 모델을 개인화 부분과 공유 부분으로 나눈 후 학습을 진행한다. 이는 각 디바이스가 개별 데이터 특성에 최적화된 모델을 만들 수 있게 하며, 잠재적인 공격으로부터 공유 부분만을 노출함으로써 공격에 강건한 모델을 생성할 수 있다. 본 논문에서 실험을 통하여 제안하는 알고리즘의 개인화 측면과 공격에 대한 강건성 측면에 대한 성능이 다른 기존 연합학습 알고리즘보다 우수함을 확인하였다.

Key Words : federated learning, personalization, model poisoning attack, data privacy

ABSTRACT

The exponential growth of decentralized data sources has propelled Federated Learning to the forefront of research. This approach facilitates the training of models across multiple devices without the need for direct data exchange. Nevertheless, the conventional federated learning method encounters inherent challenges when confronted with heterogeneous data distributions among clients. Furthermore, it remains susceptible to Byzantine attacks. To address these challenges, we propose a novel partial share algorithm. This algorithm trains local models by partitioning them into personalized and shared components, enabling clients to create personalized models that are tailored to their local data. Concurrently, it preserves robustness against potential attacks by exposing only the shared portion of the local model. Through an extensive series of experiments, we comprehensively evaluate the performance of the proposed algorithm in terms of personalization and robustness against attacks.

※ 본 연구는 정보(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(IITP-2021-0-00739) 및 대학 ICT 연구센터 지원사업(IITP-2022-2020-0-01602)의 연구결과로 수행되었음
◆ First Author : Soongsil University School of Electronic Engineering, heewon012@soongsil.ac.kr, 학생회원
° Corresponding Author : Soongsil university School of Electronic Engineering and Department of Intelligent Semiconductors, minhae@ssu.ac.kr, 종신회원
* Soongsil University Department of Intelligent Semiconductors, mirukim00@soongsil.ac.kr, 학생회원
논문번호 : 202306-133-B-RN, Received June 27, 2023; Revised August 25, 2023; Accepted August 28, 2023

Ⅰ. 서 론

기술의 발전과 함께 독립적으로 데이터를 수집하고 생성하는 디바이스가 증가함에 따라, 분산된 데이터 환경이 조성되고 있다[1]. 이러한 분산된 데이터 환경에서 일반적인 중앙집중식 학습으로 인공지능 모델을 학습하기 위해 각 디바이스의 데이터를 중앙 서버로 공유해야하며, 이는 디바이스에 위치한 개인의 사생활 데이터 노출의 위험으로 이어진다. 따라서 각 디바이스의 데이터를 활용하여 인공지능 모델을 훈련하면서도 데이터의 보안을 지킬 수 있는 기술로 연합학습이 제안되었다[2].

연합학습은 분산된 디바이스들이 데이터의 직접적인 공유 없이 각 디바이스에 위치한 모델 정보의 공유만으로 다수의 디바이스들과 연합하여 학습하는 기술로, 데이터 노출의 위험이 없어 활발히 연구되고 있는 기술이다[2-12]. 연합학습에서 매 통신라운드 $r(1 \le r \le R)$마다 전체 디바이스 집합 D에서 무작위로 통신에 참여할 C개의 디바이스를 선별하고, 통신참여집합 $D^{(r)}$을 생성한다. 통신참여집합 $D^{(r)}$에 포함된 d번째 디바이스는 개별 로컬 데이터 $\boldsymbol{X}_d$를 이용하여 로컬 모델 $\boldsymbol{W}_d$를 학습한다. 학습된 로컬 모델 $\boldsymbol{W}_d$은 중앙 서버와 공유되며, 중앙 서버는 통신 참여 집합 $D^{(r)}$에 포함된 디바이스로부터 수신한 다수의 로컬 모델을 취합하여 글로벌 모델 $\boldsymbol{W}$을 생성한다. 연합학습은 이렇게 생성된 글로벌 모델 $\boldsymbol{W}$을 평균 손실함수 $f(\boldsymbol{W}) = \frac{1}{|D|}\sum_{d=1}^{|D|} f_d(\boldsymbol{W})$를 최소화하는 값 $\hat{\boldsymbol{W}}$으로 수식 (1)과 같이 최적화하는 것을 목표로 한다.

$$\hat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W}} f(\boldsymbol{W})$$
$$f_d(\boldsymbol{W}) := \mathbb{E}_{\boldsymbol{X}_d \sim \boldsymbol{q}_d}[L_d(\boldsymbol{W}; (\boldsymbol{X}_d, \boldsymbol{y}_d))] \quad (1)$$

수식 (1)의 $f(\boldsymbol{W})$는 중앙 서버에서의 평균 손실함수로, 취합한 전체 로컬 모델의 예측 결과와 실제 값의 차이를 측정한 손실함수 $f_d(\boldsymbol{W})$의 평균으로 구해진다. $\boldsymbol{q}_d$는 d 번째 디바이스의 데이터 분포를 나타내며, $L_d(\boldsymbol{W}; (\boldsymbol{X}_d, \boldsymbol{y}_d))$는 글로벌 모델정보 $\boldsymbol{W}$에 대하여 개별 데이터 $\boldsymbol{X}_d$에 대한 추론 결과와 실제 값인 $\boldsymbol{y}_d$ 사이의 손실 값을 의미한다. 또한 $|D|$는 전체 디바이스 집합 D에 포함된 전체 디바이스의 수를 나타낸다.

하지만 이러한 연합학습은 각 디바이스에 위치한 개별 데이터의 특성을 반영한 개인화된 모델을 생성할 수 없다는 한계를 가지고 있다[2]. 기존 연합학습 시스템은 각 디바이스의 로컬 모델의 평균값으로 구한 하나의 글로벌 모델을 생성하고 모든 디바이스에 공유하여 사용하기 때문에, 개별 데이터의 특성에 최적화된 모델을 생성할 수 없다. 이러한 문제점을 해결하기 위해 개인화 연합학습 관련 많은 연구가 진행되고 있다[3-6].

연합학습은 또한 기본적으로 학습에 참여하는 디바이스들의 연합에 의존하며, 모든 디바이스가 신뢰할 만하다는 전제 하에 중앙 서버에서 로컬 모델을 취합하여 글로벌 모델을 생성한다. 이러한 연합학습에는 전체 시스템의 성능 저하를 목표로 하는 공격자가 연합학습에 참여하여 비정상적인 로컬 모델을 공유하는 모델 포이즈닝 공격의 위험이 있다[7-9]. 따라서 공격에 강건한 연합학습을 통해 글로벌 모델의 신뢰성을 높이기 위한 연구가 필요하다.

본 논문에서는 연합학습의 두 가지 주요 과제인 개인화의 필요성과 강건한 모델의 필요성을 해결하기 위하여 부분 공유 알고리즘인 pFedFrz를 제안한다. pFedFrz는 로컬 모델을 개인화 부분과 공유 부분으로 나누고, 모델 학습을 두 단계에 나누어 진행하며, 학습된 로컬 모델의 공유 부분만을 중앙 서버와 공유하는 학습 알고리즘이다. 로컬 모델의 개인화 부분을 활용하여 각 디바이스는 개별 데이터 특성에 최적화된 개인화 모델을 만들 수 있으며, 연합학습 시스템 내 공격자가 공격을 적용할 수 있는 로컬 모델의 범위를 제한함으로써 공격에 강건한 모델을 만들 수 있다.

본 논문은 다음과 같이 구성되어 있다. Ⅱ장에서는 선행적으로 연구된 기존 연합학습 기법의 장점과 단점에 대하여 서술한다. 이후 Ⅲ장에서는 제안하는 부분 공유 연합학습 기법에 대하여 서술하고, Ⅳ장에서 제안하는 방식의 우수성을 입증하기 위한 실험 결과에 대하여 서술한다. Ⅴ장에서는 제안하는 연합학습 기법에 대한 결론을 서술한다.

Ⅱ. 선행연구

연합학습은 분산된 데이터 환경에서 각 디바이스가 데이터를 직접 중앙 서버와 공유하는 것이 아닌 로컬 데이터를 사용하여 학습한 로컬 모델의 정보만을 중앙 서버와 공유하여 다수의 디바이스들이 협력하여 다수의 인공지능 모델을 학습하는 기술이다[2,3]. McMahan 연구팀은 연합학습 시스템에서 글로벌 모델 생성을 위한 알고리즘으로 FedAvg를 제안하였다[2]. 해당 방식에서 각 디바이스는 로컬 모델을 학습시킨 후, 통신참여집합에 포함되는 C개의 디바이스들은 학습된 모델의 모

든 정보 $\boldsymbol{W}_d$를 중앙 서버로 전송한다. 이후 중앙 서버는 수신한 로컬 모델을 평균 내어 다음 수식 (2)와 같이 글로벌 모델 $\boldsymbol{W}$을 생성한다.

$$\boldsymbol{W}=\frac{1}{C}\sum_{d=1}^{C}\boldsymbol{W}_d \tag{2}$$

이후 생성된 글로벌 모델은 다시 각 디바이스의 로컬 모델로 갱신된다.

이러한 FedAvg는 연합학습에서 가장 널리 사용되고 있으나, 각 디바이스에 존재하는 개별 데이터가 다른 특성을 가지는 경우, 글로벌 모델은 개별 디바이스의 특성을 반영하지 못하기 때문에 로컬 모델의 성능이 저하될 수 있다[3]. 이렇게 이질적인 데이터 환경에서의 연합학습의 개인화 성능을 위하여 Li 연구팀은 FedProx를 제안하였다[4]. FedProx는 proximal term을 이용하여 로컬 모델이 글로벌 모델로부터 멀어지는 것을 방지하며, 각 디바이스 마다 각 통신라운드 마다 별도의 에폭 수를 지정하여 학습을 진행한다. 이 외에도 개인화 연합학습을 위한 많은 연구가 진행되었다[5,6].

하지만 이러한 방식들은 모두 연합학습 시스템 내 악의적인 사용자가 존재하는 경우, 글로벌 모델의 신뢰성이 떨어질 수 있다. 특히 공격자가 모델 포이즈닝 공격을 통해 비정상적인 방법으로 생성된 로컬 모델 정보를 중앙 서버와 공유하게 되면, 중앙 서버에서 로컬 모델 집계 과정에서 오류가 생기며 이는 전체 시스템의 성능 저하를 유발한다[7,8,9]. Fang 연구팀은 모델 포이즈닝 공격에 강건한 알고리즘을 제안하였다[7]. 해당 연구에서 제안한 알고리즘인 LFR은 공격자의 조작을 탐지하기 위해 글로벌 모델 손실 값에 악영향을 주는 정도를 평가한다. 이후 손실값에 악영향을 적게 주는 신뢰할 수 있는 디바이스의 모델 정보를 우선 고려하는 방식으로 공격자의 조작을 식별하고 제거한다. 이를 통해 전체 모델의 성능을 보호할 수 있다. 하지만 해당 알고리즘은 중앙 서버의 많은 계산을 요구한다는 문제점을 가지고 있다.

위 두 한계점을 극복하면서 중앙 서버의 계산량을 줄일 수 있는 방법으로 부분공유 연합학습이 적용될 수 있다. 부분공유 연합학습은 로컬 모델을 개인화 부분과 공유 부분으로 나누어 중앙 서버에 공유 부분의 정보만을 공유하여 학습하는 기술이다[10-12]. 이 때 개인화 부분을 이용하여 개별데이터 특성에 맞춘 개인화 모델의 생성이 가능하다. 이를 위한 부분공유 알고리즘으로 FedSim과 FedAlt가 제안되었다[10]. FedSim은 공유 부분을 글로벌 모델 정보로 갱신한 뒤 개인화 부분과 공유 부분을 한번에 학습한다. 이때 연합학습 정보가 없는 개인화 부분과 연합학습 정보가 포함된 공유 부분 사이의 정보 격차로 인하여 개인화 부분과 공유 부분 사이의 호환성이 떨어진다. 따라서 FedSim은 로컬 모델 학습 과정에서 공유 부분의 연합학습 정보의 소실 우려가 존재한다. FedAlt는 호환성 문제를 해결하기 위하여 우선 개인화 부분만을 먼저 학습한다. 이후 개인화 부분을 고정하고 공유 부분만을 학습한다. FedAlt는 두 번째 단계에서 공유 부분이 개인화 부분에 과적합될 우려가 존재한다. 이에 본 논문에서 기존 알고리즘들의 문제점을 해결한 새로운 부분공유 알고리즘 pFedFrz을 제안한다.

Ⅲ. 본 론

본 장에서는 개인화 연합학습을 위한 부분공유 알고리즘을 제안한다. 우선, 연합학습 시스템에서의 각 디바이스에 대한 로컬 데이터셋과 로컬 모델을 정의한 후, 부분공유 연합학습에 대하여 설명한다. 이후, 제안하는 부분공유 알고리즘인 pFedFrz에 대하여 설명한다.

3.1 문제정의

부분공유 연합학습에서 학습에 참여하는 전체 디바이스 집합 D의 d번째 디바이스의 로컬 데이터셋 $\boldsymbol{X}_d\in\mathbb{R}^{M_d\times K}(1\le m\le M_d)$은 K개의 feature를 가진 M_d개의 데이터 샘플 x_d로 구성되며, 각 $\boldsymbol{X}_d$에 대한 label은 $\boldsymbol{y}_d$로 표시된다. 로컬 데이터셋 $\boldsymbol{X}_d$은 다음 수식 (3)과 같이 표시될 수 있다.

$$\boldsymbol{X}_d=[x_{d,1},x_{d,2},\cdots,x_{d,M_d}]^T,x_{d,m}\in\mathbb{R}^{K\times 1} \tag{3}$$

d 번째 디바이스에 내재된 로컬 모델 $\boldsymbol{W}_d$은 개인화 부분 $\boldsymbol{P}_d$와 공유 부분 $\boldsymbol{S}_d$의 집합 $\{\boldsymbol{P}_d,\boldsymbol{S}_d\}\in\boldsymbol{W}_d$으로 구성된다. 여기서 $\boldsymbol{P}_d$는 로컬 모델 $\boldsymbol{W}_d$의 개인화 부분으로 연합학습 진행 시 중앙 서버와 정보가 공유되지 않는다. $\boldsymbol{S}_d$는 로컬 모델 $\boldsymbol{W}_d$의 공유 부분으로, 연합학습 진행 시 중앙 서버와 공유되는 부분을 의미한다.

부분 공유 연합학습 시스템에서 중앙 서버는 r 번째 통신라운드에서의 통신 참여 집합 $D^{(r)}$의 d 번째 디바이스로부터 수신한 공유 부분 $\boldsymbol{S}_d$을 취합하여 글로벌 모델 $\boldsymbol{S}^{(r)}$을 생성한다. 이후 생성된 글로벌 모델 $\boldsymbol{S}^{(r)}$은 다시 각 디바이스의 로컬 모델로 공유되어 공유 부분

$\boldsymbol{S}_d$으로 갱신된다.

부분공유 연합학습은 이렇게 생성된 글로벌 모델 $\boldsymbol{S}$를 평균 손실함수 $f(\boldsymbol{P},\boldsymbol{S})=\frac{1}{|D|}\sum_{d=1}^{|D|}f_d(\boldsymbol{P}_d,\boldsymbol{S}_d)$를 최소화하는 값 $(\hat{\boldsymbol{P}},\hat{\boldsymbol{S}})$으로 수식 (4)와 같이 최적화하는 것을 목표로 한다.

$$(\hat{\boldsymbol{P}},\hat{\boldsymbol{S}})=\arg\min_{(\boldsymbol{P},\boldsymbol{S})} f(\boldsymbol{P},\boldsymbol{S})$$
$$f_d(\boldsymbol{P}_d,\boldsymbol{S}_d):=\mathbb{E}_{\boldsymbol{X}_d\sim\boldsymbol{q}_d}[L_d((\boldsymbol{P}_d,\boldsymbol{S}_d);(\boldsymbol{X}_d,\boldsymbol{y}_d))] \quad (4)$$

수식 (4)의 $f(\boldsymbol{P},\boldsymbol{S})$는 중앙 서버에서의 평균 손실함수로, 각 디바이스의 손실함수 $f_d(\boldsymbol{P}_d,\boldsymbol{S}_d)$의 평균으로 구해진다. 이를 위하여 $f(\boldsymbol{P},\boldsymbol{S})$를 최소화 하는 글로벌 모델 정보 $\boldsymbol{S}$와 개인화 부분 정보의 집합 $\boldsymbol{P}=[\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_{|D|}]$을 찾는 것을 목표로 한다.

3.2 부분공유 연합학습 기법

부분공유 연합학습은 최초 통신라운드에 디바이스의 모델을 초기화한 후 중앙 서버에서 로컬 모델의 공유 부분을 취합하여 글로벌 모델을 생성하는 과정과 개별 디바이스에서 로컬 모델을 학습하는 과정으로 구성된다. 글로벌 모델을 생성하는 과정과 개별 디바이스에서 로컬 모델을 학습하는 과정은 각각 Algorithm 1과 Algorithm 2에 명시하였다.

초기화 과정으로 첫 번째 통신라운드에서 중앙 서버는 모든 디바이스에게 초기화된 글로벌 모델의 정보 $\boldsymbol{S}^{(1)}$를 전송한다. 이후 d 번째 클라이언트는 수신한 글로벌 모델 정보 $\boldsymbol{S}^{(1)}$을 로컬 모델의 공유 부분의 정보 $\boldsymbol{S}_d$로 갱신하고 개인화 부분의 정보 $\boldsymbol{P}_d$를 무작위로 초기화 하여 초기 로컬 모델의 정보 $(\boldsymbol{P}_d,\boldsymbol{S}_d)\in\boldsymbol{W}_d$를 생성한다. 이후 로컬 데이터를 이용하여 로컬 모델을 학습한다.

이후 글로벌 모델 생성과정으로 통신 라운드 $r(1\le r\le R)$ 에서 통신 참여 집합 D^r에 포함된 d 번째 디바이스는 중앙 서버로 $\{\boldsymbol{P}_d,\boldsymbol{S}_d\}\in\boldsymbol{W}_d$ 중 $\boldsymbol{S}_d$만을 중앙 서버로 공유한다. 중앙 서버에서는 로컬 모델 정보를 취합하여 글로벌 모델 $\boldsymbol{S}^{(r)}$을 다음 수식 (5)와 같이 생성한다.

$$\boldsymbol{S}^{(r)}=\frac{1}{C}\sum_{d\in D^{(r)}}\boldsymbol{S}_d \quad (5)$$

Algorithm 1 Partial Share Federated Learning

Input: Initial state of $\{\boldsymbol{P}_d\}_{d=1}^{|D|}$ and $\boldsymbol{S}^{(1)}$, number of communication rounds R, number of participants C, number of local epoch E, and learning rate η
for communication round r: 1 to R **do**
  $D^{(r)}\leftarrow$server samples C devices
  for each device $d\in D^{(r)}$ in parallel **do**
    $\{\boldsymbol{P}_d,\boldsymbol{S}_d\}\leftarrow$pFedFrz$(\boldsymbol{S}^{(r)},E,\eta)$
    Device sends $\boldsymbol{S}_d$ to server
  end for
  Server updates $\boldsymbol{S}^{(r)}=\frac{1}{C}\sum_{d\in D^{(r)}}\boldsymbol{S}_d$
  $\boldsymbol{S}^{(r+1)}\leftarrow\boldsymbol{S}^{(r)}$
end for

Algorithm 2 pFedFrz for Local Model Update

Input: Global model $\boldsymbol{S}^{(r)}$, number of local epoch E, learning rate η
for local epoch e: 1 to E **do**
  $\boldsymbol{S}_d\leftarrow\boldsymbol{S}^{(r)}$
  Freeze $\boldsymbol{S}_d$
  $\boldsymbol{P}_d\leftarrow\boldsymbol{P}_d-\eta\nabla L((\boldsymbol{P}_d,\boldsymbol{S}_d);(\boldsymbol{X}_d,\boldsymbol{y}_d))$
end for
for local step e: 1 to E **do**
  Unfreeze $\boldsymbol{S}_d$
  $\{\boldsymbol{P}_d,\boldsymbol{S}_d\}\leftarrow\{\boldsymbol{P}_d,\boldsymbol{S}_d\}-\eta\nabla L(\{\boldsymbol{P}_d,\boldsymbol{S}_d\};(\boldsymbol{X}_d,\boldsymbol{y}_d))$
end for
Return $\{\boldsymbol{P}_d,\boldsymbol{S}_d\}$

생성된 글로벌 모델은 다시 각 디바이스의 로컬 모델로 전송되고, 각 클라이언트는 수신한 글로벌 모델의 정보 $\boldsymbol{S}^{(r)}$로 로컬 모델의 공유 부분 정보 $\boldsymbol{S}_d$를 갱신한다.

다음 로컬 모델 학습은 제안하는 pFedFrz를 적용하여 진행된다. 통신 라운드 $r(2\le r\le R)$ 부터는 로컬 모델의 학습 과정을 두 단계로 나누어 진행한다. 우선 각 디바이스는 글로벌 모델의 정보로 공유 부분의 정보를 갱신한 후 갱신된 공유 부분의 정보 $\boldsymbol{S}_d$를 고정한다. 이후 로컬 데이터를 이용하여 로컬 모델의 개인화 부분 $\boldsymbol{P}_d$에 대한 학습만을 수행한다. 첫 번째 단계를 통해

연합학습의 정보를 가지고 있는 공유 부분 $\boldsymbol{S}_d$과 연합학습의 정보가 없는 개인화 부분 $\boldsymbol{P}_d$ 사이의 정보의 격차를 해소하여 개인화 부분과 공유 부분 사이의 호환성을 확보할 수 있다. 이후 로컬 모델 학습의 두 번째 단계로 공유 부분의 고정을 푼 후 로컬 모델의 개인화 부분과 공유 부분 $\{\boldsymbol{P}_d, \boldsymbol{S}_d\} \in \boldsymbol{W}_d$을 동시에 학습하는 단계로 구성된다. 두 번째 단계에서 로컬 모델 의 전체 부분에 대한 학습을 진행하는 과정은 d 번째 개별 디바이스가 보유하고 있는 로컬 데이터 $\boldsymbol{X}_d$에 대하여 개인화된 모델을 생성하는 단계이다. 이 때, 첫 번째 단계에서 개인화 부분과 공유 부분 사이의 호환성을 확보한 뒤 로컬 모델의 학습이 수행되기 때문에 학습 과정에서 공유 부분에 포함된 연합학습 정보의 소실을 최소화 할 수 있다.

본 논문에서 제안하는 pFedFrz를 이용한 부분공유 연합학습 시 개별 디바이스는 개별 데이터 특성에 최적화된 개인화 모델을 생성할 수 있다. 또한 연합학습에 참여하는 디바이스 중 모델 포이즈닝 공격자가 존재할 때, 공격 범위를 모델의 일부로 제한할 수 있기 때문에 기존 연합학습 대비 공격에 강건한 모델을 생성할 수 있다.

Ⅳ. 실험 결과

본 장에서는 기존 연합학습 방법과 제안하는 방법의 비교 실험을 통하여 제안하는 pFedFrz 방법이 기존 방법들 대비 높은 개인화 성능을 가짐과 동시에 모델 포이즈닝 공격에 대해 강건함을 입증한다.

4.1 실험 설정

데이터셋: 본 논문에서는 3개의 데이터셋 MNIST, EMNIST, 그리고 CIFAR-10에 대하여 제안하는 부분 공유 알고리즘 pFedFrz의 성능을 평가하였다. MNIST 데이터셋은 7만개의 손글씨 흑백의 숫자 이미지로 구성된 28×28 픽셀의 데이터셋으로, 10개의 label을 가지고 있다[13]. EMNIST 데이터셋은 MNIST 데이터셋의 확장 버전으로, 손글씨 숫자 및 알파벳의 이미지로 구성된 28×28 픽셀의 데이터셋이다[14]. CIFAR-10 데이터셋은 6만개의 컬러 이미지로 32×32 픽셀의 데이터셋으로, 10개의 label을 가지고 있다[15].

성능비교: 본 논문에서 제안하는 pFedFrz의 우수성을입증하기 위하여 다음 세 개의 연합학습 알고리즘과의 성능 비교를 수행하였다.

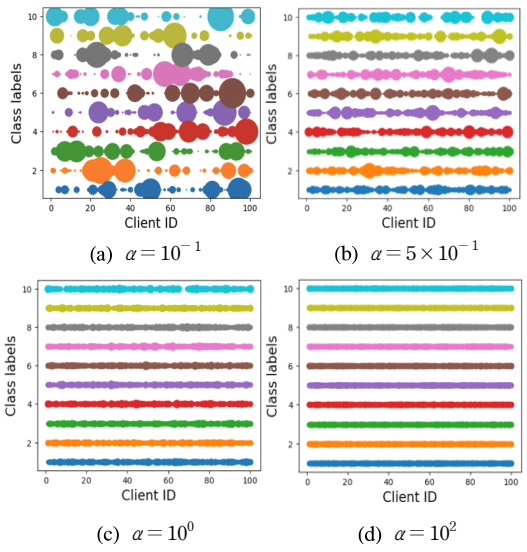
- FedAvg[2]: 대표적인 연합학습 알고리즘으로, 각 디바이스는 학습한 로컬 모델의 모든 정보를 중앙 서버와 공유한다. 해당 알고리즘과의 비교를 통하여 부분 공유의 이점을 확인할 수 있다.
- FedSim[10]: 부분 공유 알고리즘의 한 종류로, 글로벌 모델의 정보로 공유 부분의 정보를 갱신한 후, 개인화 부분과 공유 부분 사이의 호환성 확보 없이 로컬 모델학습을 진행한다. 해당 알고리즘과의 비교를 통하여 개인화 부분과 공유 부분 사이의 호환성 확보의 이점을 확인할 수 있다.
- FedAlt[10]: 부분 공유 알고리즘의 한 종류로, 두 단계에 걸쳐 로컬 모델을 학습한다. 첫 번째 개인화 부분에 대한 학습만을 진행한다. 두 번째 단계로 학습된 개인화 부분의 정보를 고정하고 공유 부분에 대한 학습을 진행한다. 두 번째 단계에서 공유 부분의 학습이 고정된 개인화 부분에 대해 과대적합 될 수 있다는 우려가 존재한다. 해당 알고리즘과의 비교를 통하여 로컬 모델 학습 시 개인화 부분의 고정 유무의 차이를 확인할 수 있다.
- LFR[7]: 모델 포이즈닝 공격에 강건한 연합학습 알고리즘 중 한 종류로 중앙 서버에서 로컬 모델을 집계할 때 손실함수에 악영향을 주는 디바이스를 식별하여 제거하는 알고리즘이다. 해당 알고리즘과의 비교를 통하여 제안하는 알고리즘이 얼마나 모델 포이즈닝 공격에 강건한지 확인할 수 있다.

통신비용: 제안하는 pFedFrz는 로컬 모델의 모든 정보를 중앙 서버와 공유하지 않기 때문에 통신비용을 효과적으로 감소시킬 수 있다. Table 1에서 볼 수 있듯이 세 개의 데이터셋에 사용한 모델 모두 한 번의 통신 라운드 당 통신비용이 FedAvg보다 pFedFrz가 적게 요구된다. 특히 MNIST와 EMNIST의 경우, pFedFrz가 FedAvg의 약 $\frac{1}{10}$에 해당하는 크기의 통신비용을 가지도록 설정하였다.

데이터 분포 환경: 디바이스 간 데이터 분포는 디리클레 분포(Dirichlet distribution)를 사용하여

표 1. 한 통신 라운드 당 통신비용
Table 1. Communication cost per communication round

	MNIST	EMNIST	CIFAR-10
FedAvg	98.5KB	5.18MB	657KB
pFedFrz	9.25KB	0.49MB	377KB



(a) $\alpha = 10^{-1}$ (b) $\alpha = 5 \times 10^{-1}$

(c) $\alpha = 10^{0}$ (d) $\alpha = 10^{2}$

그림 1. α 크기에 따른 디리클레 분포 시각화
Fig. 1. Visualization of Dirichlet distribution according to the size of α.

I.I.D.(Inde-pendently and Identically Distributed)한 분포부터 Non-I.I.D.한 분포까지 고려할 수 있도록 데이터를 분배하였다. 디리클레 분포는 분포 계수인 α를 조절하여 디바이스 간 데이터 분포의 이질성을 조절할 수 있다. α의 크기가 작을수록 데이터 분포의 이질성이 커지고, α의 크기가 클수록 디바이스 간 데이터 분포가 균등해지며, 이는 Figure 1에서 확인 가능하다.

개인화 성능(Performance of Personalization; PoP): 일반적으로 모델의 성능을 나타내는 정확도는 데이터의 모든 클래스에 대한 정확도를 평균 내어 측정한다. 하지만 이러한 방식은 Non-I.I.D.한 데이터 분포 환경에서 각 디바이스의 데이터 특성에 대한 개인화 성능을 나타내기에는 한계가 있다. 따라서 개인화 성능을 측정하기 위하여 본 논문에서는 PoP 방식의 측정을 제안하고 적용한다. PoP 성능을 측정할 때, t 번째 label 데이터가 디리클레 분포를 $\boldsymbol{Q}_t \sim Dir(\alpha), \boldsymbol{Q}_t = [\boldsymbol{q}_{t,1}, \boldsymbol{q}_{t,2}, \cdots, \boldsymbol{q}_{t,|D|}]$와 같이 따른다고 가정한다. 위 수식에서 $\boldsymbol{Q}_t$는 t 번째 label 데이터의 확률 집합이며, $\boldsymbol{q}_{t,d}$는 k 번째 label의 데이터가 d 번째 디바이스에 분배되는 비율을 나타낸다. PoP는 다음 수식 (6)과 같이 표현된다.

$$PoP = \frac{1}{|D|} \frac{1}{N_{\text{total}}} \sum_{d=1}^{|D|} \sum_{t=1}^{T} \mathrm{N}_{\mathrm{t,d}} \mathbf{q}_{\mathrm{t,d}} \qquad (6)$$

위 수식에서 T는 데이터의 총 label 수를 나타내며, N_{total}은 테스트 데이터의 총 개수를 나타낸다. 또한 $N_{t,d}$는 t 번째 label 데이터에 대해 d 번째 디바이스가 정확히 예측한 횟수를 나타낸다.

기울기 반전 공격(Gradient Sign Flip Attack): 각 연합학습 방법의 모델 포이즈닝 공격에 대한 강건성을 확인하기 위하여, 대표적인 기울기 반전 공격을 수행하는 공격자의 비율을 0%부터 90%까지 순차적으로 높이며 성능 변화를 관찰하였다. 기울기 반전 공격의 "기울기"는 손실함수의 변화율 $\nabla L(\boldsymbol{W}_d; (\boldsymbol{X}_d, \boldsymbol{y}_d))$을 나타낸다. 공격자인 디바이스 a는 글로벌 모델 정보 $\boldsymbol{W}$를 이용하여 로컬 모델 $\boldsymbol{W}_a$을 갱신하고, 기울기 $\nabla L(\boldsymbol{W}_a; (\boldsymbol{X}_a, \boldsymbol{y}_a))$를 계산한다. 이후, 반전 기울기 $-\nabla L(\boldsymbol{W}_a; (\boldsymbol{X}_a, \boldsymbol{y}_a))$를 생성한다. 생성된 반전 기울기는 로컬 모델 $\boldsymbol{W}_a$를 갱신할 때 다음 수식 (7)과 같이 사용된다.

$$\boldsymbol{W}_a \leftarrow \boldsymbol{W}_a - \eta(-\nabla L(\boldsymbol{W}_a; (\boldsymbol{X}_a, \boldsymbol{y}_a))) \qquad (7)$$

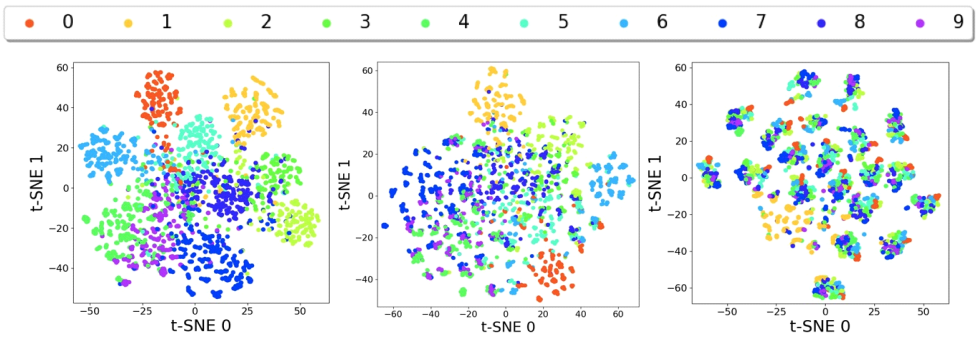
위 수식의 η는 학습 속도(learning rate)를 의미한다. 공격자는 위와 같이 생성된 비정상 모델 정보 $\boldsymbol{W}_a$를 중앙 서버로 공유하고, 이러한 정보는 글로벌 모델 생성에 악영향을 주어 성능 저하를 유발한다.

4.2 실험 결과

4.2.1 호환성 확보

본 논문에서는 기존 제안된 부분공유 알고리즘인 FedSim이 개인화 부분과 공유 부분 사이의 호환성이 부재한 상태로 로컬 학습이 진행된다는 사실을 발견하였다. 이 문제를 해결하기 위하여 로컬 모델 학습 과정에서 공유 부분을 고정하고 개인화 부분만을 학습하는 방식을 도입하여 해결하였으며, 이는 Figure 2의 실험 결과를 통해 확인할 수 있다.

Figure 2는 개인화 부분과 공유 부분 사이의 호환성을 확인하기 위하여 학습이 완료된 시점에서 각 부분공유 알고리즘 별로 개인화 부분의 출력값을 t-SNE (t-Distributed Stochastic Neighbor Embedding)를 적용시켜 시각화한 결과이다. t-SNE는 고차원 데이터의 시각화를 위한 알고리즘으로, 데이터의 패턴을 시각화하여 이해하기 위하여 사용된다. 이를 통해 로컬 모델 학습 방식에 따라 로컬 모델의 개인화 부분이 얼마나 데이터의 특징을 잘 추출하였는지 볼 수 있다. Figure



(a) pFedFrz (b) FedAlt[10] (c) FedSim[10]

그림 2. 부분 공유 방안에 따른 개인화 부분의 출력 t-SNE 시각화
Fig. 2. Visualization of t-SNE according to partial share algorithms

2의 각 그래프에서 같은 색의 점은 같은 label을 가진 데이터를 나타내며, 같은 색 데이터 별로 잘 모이고 다른 색 데이터와는 잘 분리될수록 높은 군집화 성능을 나타낸다. Figure 2.(c)에서 FedSim으로 학습된 로컬 모델의 개인화 부분은 데이터의 특징을 잘 추출하지 못하여 개인화 부분의 출력값을 t-SNE에 적용 시 데이터의 클래스 label 별로 낮은 군집화 성능을 확인할 수 있다. 이는 FedSim으로 학습 시 개인화 부분과 공유 부분 사이의 호환성이 확보되지 않은 상태로 학습이 진행되었기 때문이다. 반면 Figure 2.(a), Figure 2.(b)에서 볼 수 있듯이 로컬 학습 과정에서 개인화 부분과 공유 부분사이의 호환성을 우선 확보한 FedAlt와 pFedFrz의 경우 FedSim에 비해 개인화 부분의 출력값이 데이터의 특성을 잘 반영하여 t-SNE 확인 결과 데이터가 label 별로 잘 구분됨을 볼 수 있다.

4.2.2 과적합 해결

본 논문에서는 기존에 제안된 부분공유 알고리즘인 FedAlt가 로컬학습의 두 번째 단계에서 개인화 부분을 고정하고 공유부분만을 학습하는 과정에서 공유 부분이 개인화부분에 과적합되는 방향으로 학습되는 문제가 있음을 발견하였다. 이 문제를 해결하기 위하여 로컬 학습 두 번째 단계에서 개인화 부분의 고정을 하지 않고 로컬 모델의 모든 부분을 동시에 학습하도록 하여 해당 문제를 해결하였음을 실험을 통하여 확인하였다.

Figure 3은 알고리즘 별 서버에서 생성된 글로벌 모델로 로컬 모델의 공유 부분을 갱신하기 전의 손실값과 공유 부분 갱신 후의 손실값의 차이를 나타낸 그래프이다. 5번의 실험에 대한 평균을 실선으로, 표준편차는 음영으로 나타내었다. x축은 통신라운드 수를 나타내며, y축은 로컬 모델의 공유 부분을 글로벌 모델로 갱신하기 전과 후의 손실값의 차이를 나타낸다. FedAlt의 경우 공유 부분이 개인화 부분에 과적합되어 각 디바이스별로 중앙 서버에 송신하는 공유 부분의 정보가 서로 상이한 값을 가지게 된다. 따라서 중앙 서버에서 취합하고 평균내어 생성된 글로벌 모델은 취합 전 공유 부분과의 차이가 크게 난다. 이는 Figure 3에서 볼 수 있듯이 FedAlt의 손실값 차이가 pFedFrz에 비해 학습 전반에 걸쳐 높게 나타남을 통해 확인할 수 있다. FedSim의 차이가 가장 크게 나타나는 이유는 앞선 단계에서 호환성 확보 없이 학습을 진행하기 때문에 글로벌 모델 정보의 소실이 일어나 공유 부분 갱신 전과 후의 손실값 차이가 큰 것이다.

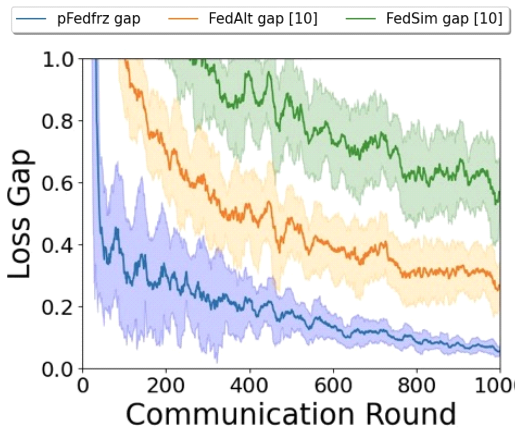
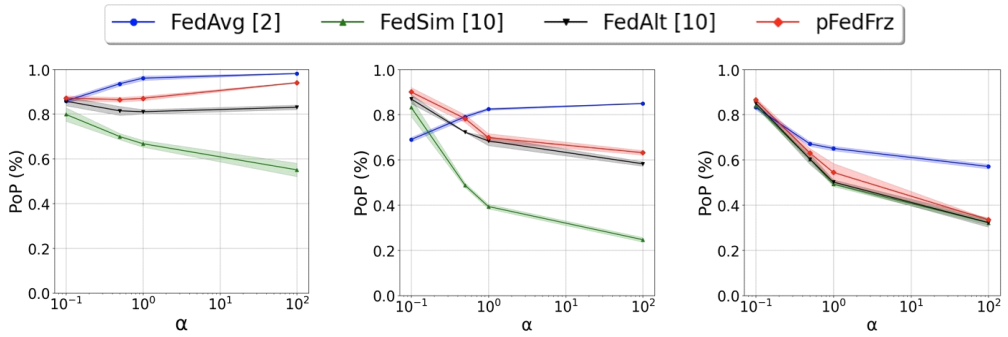


그림 3. 공유 부분 갱신 전 후의 손실 값 차이
Fig. 3. Gap in loss value of before and after updating the shared part with global model

4.2.3 개인화 성능 비교

Figure 4는 세 개의 데이터셋에 대하여 각 알고리즘



(a) MNIST (b) EMNIST (c) CIFAR-10

그림 4. α 크기에 따른 알고리즘의 개인화 성능(PoP) 비교
Fig. 4. Performance of Perosnalization (PoP) comparison according to the size of α

의 PoP를 측정한 10번의 실험의 평균을 실선으로, 표준편차는 음영으로 나타낸 그래프이다. Figure 4의 각 그래프의 x축은 α의 크기를 나타내며, α의 크기가 작을수록 디바이스 간 데이터 분포가 Non-I.I.D 하다. y축은 알고리즘의 PoP를 나타낸다.

Figure 4.(a)는 MNIST 데이터셋에 대한 결과이다. 그래프를 보면 $\alpha = 10^2$ 일 때, 즉 디바이스 간 데이터 분포가 I.I.D한 환경에서는 로컬 모델의 모든 부분을 중앙 서버와 공유하여 학습하는 FedAvg의 성능이 가장 우수 나타난다. 이렇게 I.I.D.한 환경에서 FedAvg가 부분공유 알고리즘보다 좋은 성능을 내는 이유는 데이터 분포가 I.I.D.한 경우 모든 디바이스가 학습하고자 목표로 하는 모델이 동일하여 개별 디바이스 별 개인화가 필요하지 않기 때문이다. 따라서 I.I.D.한 분포에서는 개인화 부분 생성보다 모델의 모든 정보를 공유하는 것이 학습에 더 유리하기 때문에 개인화 부분 없이 전체 모델 정보를 공유하는 FedAvg가 개인화를 진행하기 위해 모델을 일부만 공유하는 부분공유 알고리즘보다 좋은 성능을 보이는 것이다. 하지만 $\alpha = 5 \times 10^{-1}$일 때부터 FedAvg의 성능이 감소하기 시작하며, $\alpha = 10^{-1}$ 일 때는 pFedFrz의 성능보다 낮게 측정된다. 이는 Non-I.I.D.한 환경에서 FedAvg는 개별 디바이스의 데이터 특성에 맞춘 개인화 모델을 생성할 수 없어 성능이 I.I.D.한 환경에 비해 떨어진 반면, pFedFrz를 포함한 부분공유 알고리즘은 개인화 모델의 생성을 통하여 Non-I.I.D.한 환경에서 개별 디바이스의 데이터 특성에 최적화된 모델을 만들어 보다 좋은 성능을 낼 수 있게 되었기 때문이다. MNIST 데이터셋은 다른 데이터셋에 비해 간단한 데이터셋이기 때문에 pFedFrz는 모델의 모든 정보를 공유하지 않고도 I.I.D.한 데이터 분포에서 좋은 성능을 나타낸다. Non-I.I.D.한 분포에서도 pFedFrz는 우수한 개인화 모델 생성을 통하여 FedAvg에 비해 좋은 성능을 나타낸다.

Figure 4.(b)는 EMNIST 데이터셋에 대한 결과로 FedAvg의 성능은 $\alpha = 10^2$에서 모든 알고리즘 중에서 가장 높게 나타나며, $\alpha = 10^0$일 때부터 감소하기 시작한다. 반면 부분공유 알고리즘은 α의 크기가 작아질수록 성능이 높아진다. $\alpha = 5 \times 10^{-1}$일 때는 pFedFrz가 FedAvg보다 성능이 높으며 $\alpha = 10^{-1}$ 일 때는 모든 부분공유 알고리즘이 FedAvg보다 성능이 높게 측정된다. 특히 제안하는 pFedFrz는 $\alpha = 5 \times 10^{-1}$ 이하일 때, 우수한 개인화 모델 생성으로 성능이 모든 알고리즘 중 가장 높게 측정된다.

Figure 4.(c)는 CIFAR-10 데이터셋에 대한 실험 결과를 나타낸다. Figure 4.(c)에서는 이전 결과들과 다른 양상을 확인할 수 있다. 이전 결과에서는 α 크기가 작아질수록 FedAvg의 성능이 떨어짐을 확인할 수 있었는데, CIFAR-10의 결과는 α 크기가 작아질수록 FedAvg를 비롯한 모든 알고리즘의 성능이 증가하고 있다. pFedFrz의 경우 $\alpha = 10^{-1}$일 때 FedAvg보다 더 좋은 성능을 나타냄을 확인할 수 있다. CIFAR-10 에서만 FedAvg의 성능이Non-I.I.D.한 데이터 분포환경에서 I.I.D. 환경 보다 더 좋은 성능을 나타내는 것은 이전 여러 연구에서도 비슷한 경향성을 나타내는 것으로 보아 CIFAR-10 데이터셋의 특성이라고 볼 수 있다[16,17].

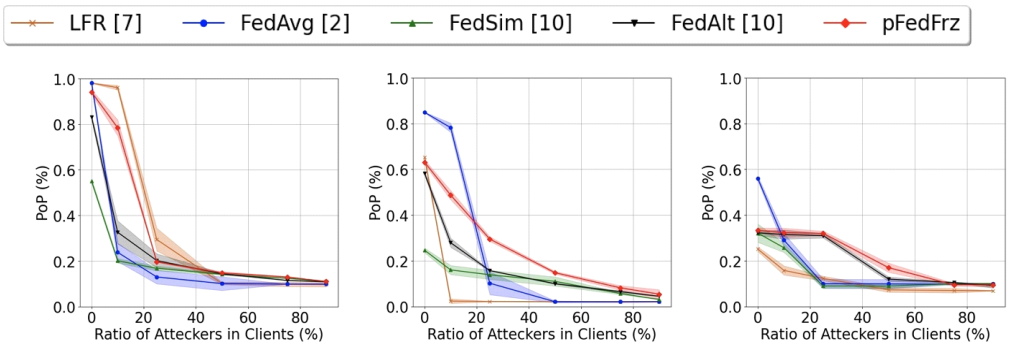
위 Figure 4의 결과를 통하여 제안하는 pFedFrz가 개별 디바이스의 데이터 특성에 최적화된 개인화 모델을 생성하여 Non-I.I.D.한 데이터 분포 환경에서 FedAvg와 다른 부분공유 알고리즘에 비하여 좋은 성능을 보이는 것을 확인하였다.

4.2.4 모델 포이즈닝 공격에 대한 강건성 비교

Figure 5와 Figure 6은 각각 $\alpha = 10^2$일 때의 I.I.D.한 분포환경과 $\alpha = 10^{-1}$일 때의 Non-I.I.D.한 데이터 분포에서 연합학습 시스템 내 공격자의 비율에 따른 각 알고리즘의 성능을 측정한 10번의 실험의 평균을 실선으로, 표준편차는 음영으로 나타낸 그래프이다. 그래프의 x축은 전체 디바이스 중 공격자의 비율을 나타내며, y축은 각 알고리즘이 개인화 성능(PoP)을 나타낸다.

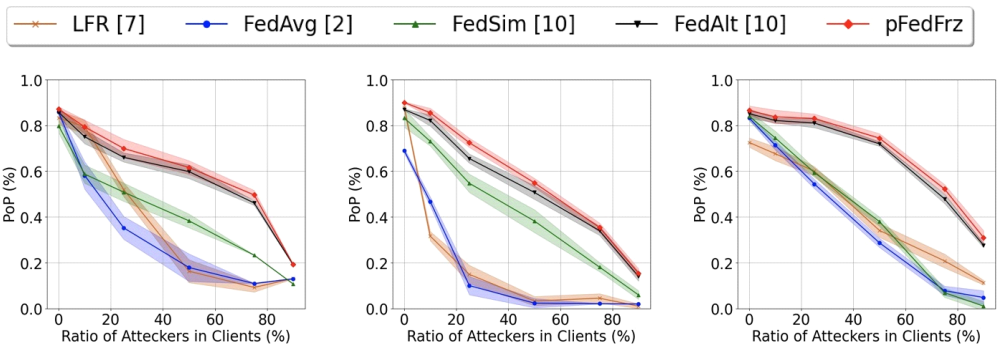
I.I.D.한 데이터 분포 환경에서 Figure 5에서 볼 수 있듯이 공격자가 0%일 때의 성능은 Figure 4에서 $\alpha = 10^2$일 때의 성능과 동일하다. 공격자가 0% 일 때 세가지 데이터셋 모두 FedAvg의 성능이 가장 높게 나타남을 볼 수 있다. 하지만 Figure 3.(a)의 경우 공격자가 10%일 때 FedAvg의 성능 감소율은 약 75%이며 pFedFrz은 약 17%의 감소율을 보였다. 기존 모델 포이즈닝 공격에 강건한 알고리즘으로 제안되었던 LFR은 약 0.04%의 가장 낮은 감소율을 보였다. LFR의 경우 MNIST를 제외한 다른 데이터셋에서의 결과를 보면 공격자가 10%만 참여하는 상황에서도 높은 성능 감소율을 보여 가장 낮은 성능을 보인다. 이를 통해 LFR은 공격자 비율이 10%보다 클 때와, MNIST보다 상대적으로 어려운 데이터셋에 대하여 제안하는 pFedFrz에 비해 모델 포이즈닝 공격에 대해 강건성이 낮다고 할 수 있다.

Figure 5.(b)를 보면 FedAvg의 성능은 공격자가 10%일 때까지 모든 알고리즘 중 가장 높지만 공격자가 25%일 때 급격한 성능 하락을 보였으며, 상대적으로 성능 하락이 완만하게 나타난 부분공유 알고리즘의 성능이 FedAvg보다 높게 나타남을 볼 수 있다. 특히 pFedFrz는 세가지 부분공유 알고리즘 중에서 공격자 증가에 따른 성능 감소율이 가장 낮게 나타나 공격자가 25% 이상일 때부터는 가장 높은 성능을 나타낸다. Figure 5.(c)에서도 pFedFrz의 성능 감소율이 가장 작아 공격자가 10% 이상일 때부터 가장 우수한 성능을



(a) MNIST (b) EMNIST (c) CIFAR-10

그림 5. I.I.D.한 데이터 분포에서의 공격자 증가에 따른 개인화 성능(PoP) 변화
Fig. 5. Performance of Personalization (PoP) according to the ratio of attackers in clients on I.I.D. setting



(a) MNIST (b) EMNIST (c) CIFAR-10

그림 6. Non-I.I.D.한 데이터 분포에서의 공격자 증가에 따른 개인화 성능(PoP) 변화
Fig. 6. Performance of Personalization (PoP) according to the ratio of attackers in clients on Non-I.I.D. setting

나타낸다.

Non-I.I.D.한 데이터 분포 환경에서 Figure 6에서 볼 수 있듯이 공격자가 0%일 때의 성능은 Figure 4에서 $\alpha = 10^{-1}$일 때의 성능과 동일하다. 공격자가 참여함에 따라 세 가지 데이터셋 모두 FedAvg의 성능은 급격하게 떨어짐을 확인할 수 있다. LFR의 성능은 MNIST 데이터셋에 대해서만 공격자 비율이 10%일 때 까지는 성능 감소율이 약 0.05%로 가장 낮지만 공격자 비율이 25% 이상일 때부터와 EMNIST, CIFAR-10 데이터셋에 대해서는 FedAvg와 비슷하게 급격한 성능 하락을 보인다. 반면 부분공유 알고리즘은 FedAvg보다 상대적으로 성능하락이 완만하게 나타나며, 공격자가 90%일 때 까지 천천히 성능이 감소함을 볼 수 있다. 특히 Figure 6.(a)와 Figure 6.(c)에서 pFedFrz와 FedAlt는 FedSim보다 성능 하락이 상대적으로 완만하게 나타남을 볼 수 있으며, 세 가지 데이터셋 모두 공격자가 0%일 때부터 90%일 때까지 pFedFrz는 강건한 개인화 모델을 만들어 가장 높은 성능을 나타냄을 볼 수 있다. 위 개인화 성능실험과 강건성 실험을 바탕으로 제안하는 pFedFrz가 연합학습 시스템 내 모델 포이즈닝 공격이 있을 경우 I.I.D.한 환경과 Non-I.I.D.한 환경 모두 강건한 개인화 모델 생성을 통하여 가장 좋은 성능을 냄을 실험적으로 입증하였다. 특히 I.I.D.한 데이터 분포 환경에서도 공격자가 없는 경우 FedAvg의 성능이 가장 좋았던 반면, 공격자 참여 시 pFedFrz는 모델 포이즈닝 공격에 대한 강건한 모델을 생성함으로써 I.I.D.한 환경임에도 FedAvg보다 좋은 성능을 나타냄을 확인하였다.

4.2.5 복잡도 분석

실험에 사용된 이미지 분류 모델인 합성곱 신경망 모델을 학습할 때, 한 합성곱 층 당 계산 복잡도는 다음과 같다[18].

$$K \times K \times C_{\text{in}} \times \text{H}_{\text{out}} \times \text{W}_{\text{out}} \times \text{C}_{\text{out}} = O(C^2K^2HW) \tag{8}$$

합성곱 신경망 모델의 합성곱 층에서는 입출력이 벡터가 아닌 $H \times W \times C$의 3차원 특징맵(feature map)으로 이루어진 데이터이다. 여기서 H, W, C는 각각 데이터의 높이, 너비, 채널 수를 의미한다. C_{in}과 C_{out}은 각각 입력과 출력 데이터의 채널 수를 의미하며, H_{out}은 출력 데이터의 높이, 그리고 W_{out}는 출력 데이터의 너비를 뜻한다. 합성곱 연산을 위한 커널의 크기가 $K \times K$라고 할 때 계산 복잡도는 위 수식 (8)과 같이 계산된다. 모델이 총 l개의 층으로 구성되며 한 학습 당 E 에폭만큼 반복한다고 할 때, 학습이 한 단계로 이루어지는 알고리즘의 경우 계산 복잡도는 수식 (9)와 같다.

$$E \times l \times K \times K \times C_{\text{in}} \times \text{H}_{\text{out}} \times \text{W}_{\text{out}} \times \text{C}_{\text{out}} = O(El C^2K^2HW) \tag{9}$$

제안하는 알고리즘 pFedFrz과 FedAlt의 경우 일반적인 알고리즘과는 달리 로컬 학습이 총 두 단계에 걸쳐 진행된다. 각 단계별로 E 에폭씩 반복한다고 하였을 때 계산 복잡도는 다음과 같다.

$$2 \times E \times l \times K \times K \times C_{\text{in}} \times \text{H}_{\text{out}} \times \text{W}_{\text{out}} \times \text{C}_{\text{out}} = O(2El C^2K^2HW) \tag{10}$$

따라서 제안하는 알고리즘과 학습이 한단계로 이루어진 일반적인 알고리즘과의 계산 복잡도를 비교해보면 $O(2lEC^2K^2HW) = O(lEC^2K^2HW)$로 동일함을 알 수 있다. 또한 단계 별 학습 에폭 수를 E에서 $\frac{1}{2}E$로 조절하면 일반적인 알고리즘과 계산 복잡도 및 수행 시간 또한 동일하도록 설정할 수 있다.

표 2. 로컬 에폭 수에 따른 계산 복잡도, 수행 시간, 통신 비용 및 성능
Table 2. Computational complexity, performance time, communication cost and performance comparison

	Computation	Time/iter.	Communication/iter.	Accuracy
FedAvg(E)	$O(lEC^2K^2HW)$	1.01s	5.18MB	68.9(%)
FedSim(E)	$O(lEC^2K^2HW)$	1.01s	0.49MB	90.2(%)
FedAlt(E)	$O(lEC^2K^2HW)$	1.87s	0.49MB	95.3(%)
pFedFrz(E)	$O(lEC^2K^2HW)$	1.87s	0.49MB	**96.6(%)**
FedAlt($\frac{1}{2}E$)	$O(lEC^2K^2HW)$	1.01s	0.49MB	89.7(%)
pFedFrz($\frac{1}{2}E$)	$O(lEC^2K^2HW)$	1.01s	0.49MB	**92.9(%)**

표 2는 $\alpha = 0.1$인 Non-I.I.D. 한 분포에서 EMNIST 데이터셋에 대하여 학습 에폭 수를 E로 설정하여 진행한 FedAvg, FedSim, FedAlt, pFedFrz 의 실험 결과 및 학습 단계 별 에폭 수를 $\frac{1}{2}E$로 설정하여 진행한 FedAlt와 pFedFrz의 성능을 나타낸 표이다. 표 2에서 확인할 수 있듯이, 로컬 학습이 한 단계로 이루어지는 FedAvg와 FedSim의 경우 한 통신 라운드 당 수행 시간이 1.01s가 소요되며, 로컬 학습이 두 단계로 구성되는 FedAlt와 pFedFrz의 경우 1.87s 소요되는 것을 볼 수 있다. 학습 단계 별 에폭 수를 $\frac{1}{2}E$로 조절한 FedAlt와 pFedFrz는 수행 시간이 1.01s로 FedAvg와 FedSim과 동일하다. 가장 성능이 높은 알고리즘은 에폭 수가 E일 때의 pFedFrz의 성능이 96.6(%)로 가장 높게 나타난다. 또한 에폭 수를 $\frac{1}{2}E$로 조절하였을 때의 pFedFrz의 성능은 92.9(%)로 동일한 계산 복잡도와 수행 시간을 가진 알고리즘 중에서 가장 높게 나타남을 볼 수 있다. 이를 통하여 제안하는 알고리즘인 pFedFrz는 계산 및 수행 시간을 고려하여 에폭 수를 $\frac{1}{2}E$로 맞춘다면 동일 계산 및 수행 시간을 가진 다른 알고리즘 중에서 가장 높은 성능을 낼 수 있음을 확인할 수 있다.

V. 결 론

본 논문은 모델 포이즈닝 공격에 강건한 개인화 연합학습을 위한 부분공유 알고리즘을 제안하였다. 제안하는 pFedFrz는 로컬 모델의 공유 부분만을 중앙서버와 공유하도록 하여 공격자의 공격 범위를 제한하여 공격에 강건한 모델을 만들 수 있다. 또한, 로컬 모델의 학습을 공유 부분과 개인화 부분 사이의 호환성을 확보하는 단계와, 개별 데이터에 최적화된 모델로 학습하는 단계를 통해 우수한 개인화 모델을 생성하여 Non-I.I.D.한 데이터 분포 환경에서 좋은 개인화 성능을 낼 수 있다. 제안하는 pFedFrz의 성능을 확인하기 위하여 세 가지 데이터셋을 바탕으로 세 가지 기존 연합학습 알고리즘과의 성능 비교를 진행한 결과, pFedFrz는 Non-I.I.D. 한 데이터 분포 환경에서 개별 디바이스의 데이터 특성에 최적화된 개인화 모델을 생성한다. 이를 통하여 다른 알고리즘 들에 비해 개인화와 모델 포이즈닝 공격에 대한 강건성 측면에서 우수한 성능을 보임을 확인하였다.

References

[1] J. Howarth, *How many people own smartphones (2023-2028)*, Retrieved Aug. 11, 2020, from https://explodingtopics.com/blog/smartphone-stats

[2] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *International Conference on AISTATS,* pp. 1273-1282, Apr. 2017.

[3] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, and R. G. D’Oliveira, “Advances and open problems in federated learning,” in *Found. Trends Mach. Learn.,* vol. 14, no. 1-2, pp. 1-210, Jun. 2021. (http://dx.doi.org/10.1561/2200000083)

[4] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, “Federated optimization in heterogeneous networks,” in *Conference on MLSys,* vol. 2, pp. 429-450, Texas, USA, Mar. 2020.

[5] H. Zhu, J. Xu, S. Liu, and Y. Jin, “Federated learning on non-IID data: A survey,” *Neurocomputing*, vol. 465, pp. 371-390, 2021. (https://doi.org/10.1016/j.neucom.2021.07.098)

[6] T. Li, S. Hu, A. Beirami, and V. Smith, “Ditto: Fair and robust federated learning through personalization,” in *ICML,* pp. 6357-6368, Vienna, Austria, Jul. 2021.

[7] M. Fang, X. Cao, J. Jia, and N. Z. Gong, “Local model poisoning attacks to byzantine-robust federated learning,” in *USENIX Secur.*, pp. 1623-1640, Boston, USA, Aug. 2020.

[8] K. Pillutla, S. M. Kakade, and Z. Harchaoui, “Robust aggregation for federated learning,” in *IEEE TSP*, vol. 70, pp. 1142-1154, 2021. (https://doi.org/10.1109/TSP.2022.3153135)

[9] J. Zhang, J. Chen, D. Wu, B. Chen, and S. Yu, “Poisoning attack in federated learning using generative adversarial nets,” in *IEEE TrustCom,* pp. 374-380. Rotorua, New

Zealand, Aug. 2019. (https://doi.org/10.1109/TrustCom/BigDataSE.2019.00057)

[10] K. Pillutla, K. Malik, A. R. Mohamed, M. Rabbat, M. Sanjabi, and L. Xiao, "Federated learning with partial model personalization," in *ICML*, pp. 17716-17758, Baltimore, USA, Jun. 2022.

[11] M. Kim and M. Kwon, "Performance analysis of partial-share solution for personalized federated learning," in *KICS Winter Conference*, pp. 1333-1334, Pyeongchang, South Korea, Feb. 2023.

[12] H. Kye and M. Kwon, "Partial federated learning based network intrusion system for mobile devices," in *ACM MobiHoc*, pp. 283-284, Seoul, South Korea, Oct. 2022. (https://doi.org/10.1145/3492866.3561257)

[13] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541-551, 1989. (https://doi.org/10.1162/neco.1989.1.4.541)

[14] G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik, "EMNIST: Extending MNIST to handwritten letters," in *IEEE IJCNN*, pp. 2921-2926, Ancorage, Alaska, May 2017. (https://doi.org/10.1109/IJCNN.2017.7966217)

[15] A. Krizhevsky, V. Nair, and G. Hinton, "*CIFAR-10 (canadian institute for advanced research)*," 2009. Retrieved Jun. 27, 2023, from http://www.cs.toronto.edu/kriz/cifar.

[16] Q. Li, Y. Diao, Q. Chen, and B. He, "Federated learning on non-iid data silos: an experimental study," in *IEEE ICDE*, pp. 965-978, Kuala Lumpur, Malaysia, May 2022. (https://doi.org/10.1109/ICDE53745.2022.00077)

[17] Z. Qin, L. Yang, Q. Wang, Y. Han, and Q. Hu, "Reliable and interpretable personalized federated learning," in *IEEE CVF*, pp. 20422-20431, Vancouver, Canada, Jun. 2023.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, California, USA, Dec. 2017.

박 희 원 (Heewon Park)

2020년 3월~현재 : 숭실대학교 전자정보공학부 IT융합전공

<관심분야> 인공지능, 연합학습, 모바일 네트워크

[ORCID:0009-0006-0446-3151]

김 미 르 (Miru Kim)

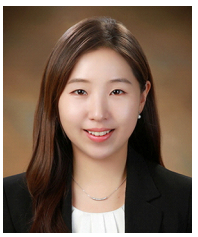
2019년 3월~2022년 8월 : 숭실대학교 전자정보공학부 IT융합전공

2022년 9월~현재 : 숭실대학교 지능형반도체학과 석사과정

<관심분야> 이상탐지기술, 인공지능, 연합학습

[ORCID:0000-0002-5394-4780]

권 민 혜 (Minhae Kwon)



2011년 8월 : 이화여자대학교 전자정보통신공학과 학사

2013년 8월 : 이화여자대학교 전자공학과 석사

2017년 8월 : 이화여자대학교 전자전기공학과 박사

2017년 9월~2018년 8월 : 이화여자대학교 전자전기공학과 박사 후 연구원

2018년 9월~2020년 2월 : 미국 Rice University, Electrical and Computer Engineering, Postdoctoral Researcher

2018년 9월~2020년 2월 : 미국 Baylor College of Medicine, Center for Neuroscience and Artificial Intelligence, Postdoctoral Researcher

2020년 3월~현재 : 숭실대학교 전자정보공학부 및 지능형반도체학과 조교수

<관심분야> 모바일네트워크, 이상탐지기술, 인공지능, 강화학습, 자율주행

[ORCID:0000-0002-8807-3719]