엣지 기반 정확도 검증을 통한 계층적 연합학습의 악의적 클라이언트 방어 기법

김건호, 정현수, 길준민*

제주대학교 컴퓨터공학과

kgh9941@stu.jejunu.ac.kr, jhs990909@stu.jejunu.ac.kr, jmgil@jejunu.ac.kr

Defense Scheme Against Malicious Clients in Hierarchical Federated Learning by Edge-Based Accuracy Validation

Geonho Kim, Hyunsu Jeong, Joon-Min Gil*

Dept. of Computer Engineering, Jeju National University

요 약

대규모 네트워크 환경에서 계층적 연합학습(Hierarchical Federated Learning, HierFL)은 서버 - 엣지 - 클라이언트 구조 하에서 확장성과 통신 효율성을 동시에 제공한다. 그러나 라벨 플립(label-flipping)과 같은 악의적 클라이언트는 전역 모델 성능을 심각하게 저하시키며, 더욱이 기존의 단순 집계 방식만으로는 이를 효과적으로 방어하기 어렵다. 본 연구에서는 이러한 문제를 해결하기 위해 엣지에서 정확도 검증을 수행하는 필터링 방식을 제시한다. 제안 기법은 각 클라이언트로부터 전달된 로컬 모델을 엣지가 독립적인 검증 데이터셋으로 평가하고, 정확도가 특정 기준에 부합하는 업데이트만을 서버로 전달하도록 한다. 이러한 방식을 통해 악의적 클라이언트의 비율이 증가하는 상황에서도 제안 기법은 기존 단순 평균 기반 집계 방식 대비 전역 모델의 정확도를 안정적으로 유지할 수 있음을 보여주었다. 이는 본 논문에서 제안하는 기법이 계층적 연합학습 환경에서 보안성과 신뢰성을 향상시킬 수 있음을 증명한다.

Ⅰ. 서 론

현대의 분산 학습 환경에서는 개인정보 보호와 통신 효율성을 동시에 확보할 수 있는 기법이 필수적이다. 연합학습(Federated Learning, FL)은 데이터를 중앙 서버로 직접 전송하지 않고, 클라이언트가 로컬 데이터로 학습한 모델 파라미터만을 공유하여 전역 모델을 구축하는 방식으로 이러한 요구를 충족한다. 그러나 클라이언트 수가 증가하고 네트워크 규모가 확장될수록, 단일 서버 구조에서는 통신 지연과 부하 집중 문제가 발생하여 확장성과 안정성에 한계를 드러낸다. 이를 보완하기 위해 현재 활용되고 있는 계층적 연합학습(Hierarchical Federated Learning, 이하 HierFL)은 기존 연합학습 구조에 서버 - 엣지 - 클라이언트의 계층적 구조를 도입하여 통신 효율성을 높이고 대규모 학습 환경에서도 안정적으로 모델을 학습할 수 있도록 한다. 이 과정에서 엣지(edge)는 복수의 클라이언트 업데이트를 먼저 집계한 뒤 서버로 전달하는 중간 계층으로서, 네트워크 확장성과 학습 속도를 보장하는 중요한 역할을 한다. 그러나 기존 HierFL 연구들은 엣지를 단순한 집계 노드로만 활용하는 경향이 강하다[1]. 이러한 구조에서는 악의적 클라이언트가 조작한 업데이트가 있을 경우 엣지를 통해 그대로 서버로 전달됨으로 모델 성능 저하가 발생하며, 더욱이 라벨 플립(label-flipping)과 같은 모델 오염 공격은 전역 모델의 정확도를 크게 저하시킨다. 즉, HierFL은 확장성과 통신 효율성 측면에서 장점을 가지지만, 보안성과 강건성 측면에서는 여전히 취약하다.

따라서, 본 연구에서는 이러한 문제를 해결하기 위해 엣지에서의 정확도 검증 기반 필터링 기법을 제안한다. 제안 기법에서 엣지는 단순히 클라이언트 모델을 집계하는 것이 아니라, 독립적인 검증 데이터셋을 활용하여 각 클라이언트 업데이트의 정확도를 평가하고, 기준 미달인 업데이트는 서버 전송 단계에서 배제한다. 이를 통해 악의적 클라이언트가 전역 모델에 미치는 영향을 최소화하고, 학습 안정성과 신뢰성을 강화시킨다.

Ⅱ. 제안 기법

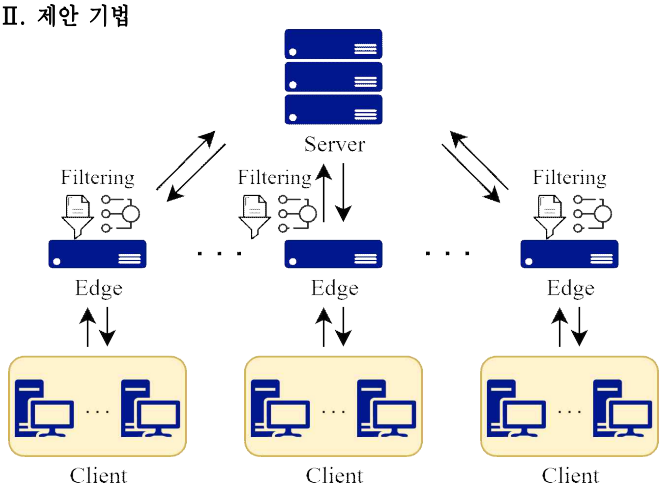


그림 1 클라이언트 필터링 기능을 갖는 계층적 연합학습 구조

2.1 시스템 모델

본 연구의 제안 기법은 그림 1과 같이 서버 - 엣지 - 클라이언트 구조를 갖는 계층적 연합학습 환경에 기반한다. 전역 서버는 기본 모델을 준비하고 배포하는 역할을 한다. 이후 클라이언트는 개별적으로 보유한 데이터를 기반으로 로컬 학습을 수행하며, 학습된 모델의 파라미터를 엣지 서버로 전송한다. 각 엣지 서버는 연결된 클라이언트들로부터 수집한 파라미터를 정확도 임계값을 기준으로 필터링 및 집계하여 엣지 모델을 형성한다. 이렇게 생성된 엣지 모델은 다시 상위의 전역 서버로 전달되며, 이때 전역 서버는 모든 엣지 서버로부터 모인 결과를 통합하여 최종 전역 모델을 갱신한다. 갱신된 전역 모델은 다시 각 엣지 서버를 거쳐 클라이언트로 재배포되는 과정이 반복적으로 수행됨으로써 전체 시스템은 점차적으로 성능이 향상된 전역 모델로 완성되어 간다.

2.2 정확도 검증 기반 필터링 알고리즘

Algorithm 1: Hierarchical FL with Edge Filtering

Input: Clients $\mathcal{C}$, Edges $\mathcal{E}$, Rounds R, Local steps τ, Validation set D_{val}
Output: Global model w^R

Initialize w^0 and broadcast to edges and clients;
for $r = 1$ **to** R **do**
  foreach *client* c **do**
    Update w_c^r from w^{r-1} by τ local steps;
    Send w_c^r to its edge;
  end
  foreach *edge* e **do**
    Evaluate $\{w_c^r\}$ on D_{val};
    Select $S_e = \{c \mid a_c \geq \text{median}(\{a_c\})\}$;
    Aggregate $w_e^r \leftarrow \frac{1}{|S_e|} \sum_{c \in S_e} w_c^r$;
    Send w_e^r to the server;
  end
  $w^r \leftarrow \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} w_e^r$;
  Broadcast w^r to all clients;
end
return w^R;

엣지 서버는 클라이언트들로부터 수집된 모델 업데이트를 집계하기 전에 알고리즘 1에서 제시한 정확도 기반 필터링 알고리즘을 적용한다. 엣지 서버는 전역 검증 데이터셋으로 각 클라이언트 모델을 평가하여 정확도 a_c를 산출하고, 필터링 정책으로 정해진 임계값에 따라 업데이트 수용 여부를 결정한다. 이는 기존 연구[2]에서 보고된 바와 같이, 악의적 클라이언트가 제출한 업데이트는 정상 클라이언트에 비해 전역 검증 데이터셋에서 성능이 현저히 낮게 나타난다는 결과에 근거한다. 본 연구에서는 필터링 정책으로 중앙값을 사용하며, 이는 각 엣지에서 평가된 클라이언트 업데이트 정확도들의 중앙값을 기준치로 삼아 $a_c \geq median(a_c)$를 만족하는 업데이트만 통과시킨다. 통과한 업데이트에 대해 집계함수인 FedAVG를 적용하여 엣지 모델을 산출하고 이를 전역 서버로 전송한다.

Ⅲ. 성능 평가

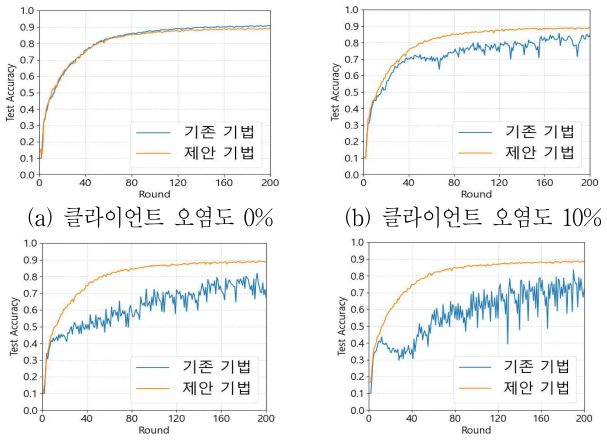
3.1 실험 환경 구성

전체 학습 데이터의 10%는 클래스별 균등 분할을 통해 전역 검증 데이터셋으로 분리하고, 각 엣지 서버에서 모델 업데이트의 신뢰도 평가에 활용하였다. 나머지 90%의 학습 데이터는 클라이언트 수에 맞추어 데이터를 분배하였다. 악의적 클라이언트는 무작위로 선택되며, 라벨 플립 공격으로 모사하였다. 공격 클라이언트 비율은 0 - 50% 범위에서 10% 간격으로 변화시켰다. 학습은 총 200 라운드 동안 진행되었으며, 각 클라이언트는 라운드마다 5회의 로컬 업데이트를 수행하였다. 학습 모델은 ResNet-18 모델을 기반으로, SGD 옵티마이저(learning rate=0.05, momentum=0.9, weight decay=$5\times e^{-4}$)를 적용하였다. 한편, 제안 기법의 성능평가를 위해 사용되는 데이터셋으로 CIFAR-10이 사용되었다. 이러한 데이터셋에 기반하여 클라이언트 오염의 검증 기능을 갖는 제안 기법과 클라이언트 오염 검증이 없는 기존 기법을 정확도 관점에서 성능을 비교하였다.

3.2 성능 결과

그림 2는 클라이언트 오염도에 따른 라운드별 전역 모델 정확도의 변화를 보여준다. 그림 2의 결과를 살펴보면, 오염 클라이언트 비율이 증가할수록 기존 기법은 모델 정확도의 성능 저하와 함께 라운드가 증가할수록 정확도의 변동성이 점점 증가한다. 반면, 제안 기법은 전반적으로 안정적인 수렴 과정을 보이며 비교적 일관된 성능을 유지한다.

그림 3은 baseline 대비 정확도 차이의 표준편차를 나타낸다(baseline으로 오염 클라이언트가 없는 기존 기법이 사용됨). 기존 기법은 클라이언트의 오염도가 높아질수록 표준편차가 급격히 증가하여 성능 변동성이 큰 반면, 제안 기법은 표준편차가 낮게 유지되며 안정적인 성능을 보인다.



(a) 클라이언트 오염도 0%　(b) 클라이언트 오염도 10%

(c) 클라이언트 오염도 20%　(d) 클라이언트 오염도 30%

그림 2 클라이언트 오염도에 따른 라운드별 테스트 정확도

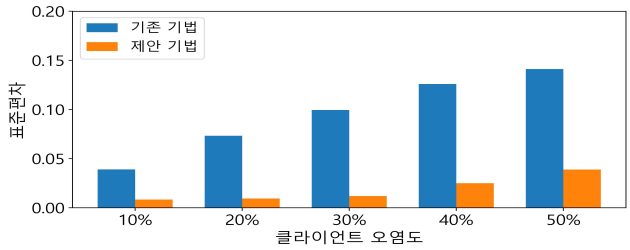


그림 3 Baseline과의 정확도 차이의 표준편차

Ⅳ. 결론

본 연구는 서버 - 엣지 - 클라이언트 구조의 계층적 연합학습에서 악의적 클라이언트 공격이 전역 모델에 미치는 영향을 분석하고, 이를 완화하기 위한 정확도 검증 기반 중앙값 필터링 기법을 제안하였다. 제안 기법은 각 엣지 서버가 전역 검증 데이터셋으로 클라이언트 업데이트를 평가하고, 중앙값 이상의 정확도를 보이는 모델만 수용하는 방식이다.

CIFAR-10 데이터셋과 ResNet-18 모델을 활용한 실험에서는 공격 클라이언트 비율이 증가할수록 기존 기법의 정확도가 급격히 하락한 반면, 제안 기법은 성능 저하를 효과적으로 억제하였다. 이러한 결과는 엣지에서의 필터링이 계층적 연합학습 환경에서 신뢰성 있는 학습을 보장하는 핵심 요소임을 시사한다. 향후 연구에서는 데이터 포이즈닝, 업데이트 조작 등 다양한 비정상 행위에 대한 확장과 복합 정책 적용을 통해 안정성을 더욱 강화할 예정이다.

ACKNOWLEDGMENT

본 과제(결과물)는 2025년도 교육부 및 제주도의 재원으로 제주RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다(2025-RISE-17-001).

참 고 문 헌

[1] T. Do, D. A. Tran, and A. Vo, “Edge assignment in edge federated learning,” SN Applied Sciences, vol. 5, no. 281, 2023.

[2] S. Venkateswaran, Q. Shaikh, S. Singh, J. Abraham, and A. Bochare, "Defense Mechanism to Thwart Model Poisoning on Non-IID Data based Federated Learning for Credit Fraud Detection System," 2025 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2025