# 계층적 연합 학습 구조에서 엣지 서버의 동적 필터링과 가중 집계를 이용한 악의적 클라이언트 방어 기법

김건호[1], 정현수[2], 길준민[3*]
[1,2,3]제주대학교 컴퓨터공학과
[1]kgh9941@stu.jejunu.ac.kr, [2]jhs990909@stu.jejunu.ac.kr, [3]jmgil@jejunu.ac.kr

# Malicious Client Defense Scheme Using Dynamic Filtering and Weighted Aggregation on Edge Servers in a Hierarchical Federated Learning Architecture

Geonho Kim[1], Hyunsu Jeong[2], Joon-Min Gil[3*]
[1,2,3]Dept. of Computer Engineering, Jeju National University

요 약

계층적 연합 학습(Hierarchical Federated Learning, HierFL)은 높은 확장성과 통신 효율성으로 인해 대규모 분산 환경에 적합한 학습 구조로 주목받고 있다. 그러나 악의적인 클라이언트가 의도적으로 라벨을 고의로 변조하는 라벨 플립(Label-Flipping) 공격을 수행할 경우, 글로벌 모델의 성능이 심각하게 저하되는 취약점이 존재한다. 본 논문에서는 이러한 문제를 해결하기 위해, 엣지 서버 단에서 수행되는 2단계 방어 기법을 제안한다. 첫 번째 단계는 정확도 기반 동적 필터링으로, 엣지 서버는 독립적인 검증 데이터셋을 활용해 각 클라이언트 모델의 정확도를 평가하고, 상위 그룹의 평균 정확도를 기준으로 동적으로 설정된 임계값 이하의 모델을 집계에서 제외한다. 두 번째 단계는 정확도 기반 가중 집계로, 필터링을 통과한 모델들에 대해 최근 K 라운드의 평균 정확도를 가중치로 적용하여 안정적인 집계를 수행한다. CIFAR-10 데이터셋을 이용한 실험 결과, 악의적 클라이언트 비율이 높은 공격 환경에서도 제안 기법은 안정적인 학습을 유지하며 베이스라인 대비 뚜렷한 성능 향상을 보였다. 이러한 결과는 제안한 악의적 클라이언트 방어 기법이 계층적 연합 학습 시스템의 신뢰성을 효과적으로 향상시킴을 보여준다.

## Ⅰ. 서 론

최근 IoT 및 엣지 컴퓨팅 기술의 확산과 더불어, 대규모 분산 환경에서 개인정보를 보호하며 인공지능 모델을 학습하는 기술의 중요성이 그 어느 때보다 높아지고 있다. 연합학습은 데이터를 중앙 서버로 이동시키지 않고, 각 클라이언트 기기에서 모델을 로컬로 학습한 뒤 업데이트 결과만을 공유하는 방식으로 이러한 요구를 효과적으로 충족시킨다[1]. 그러나 클라이언트 수가 수천, 수만 단위로 확장됨에 따라 중앙 서버에 직접 연결되는 전통적인 연합학습 구조는 심각한 통신 병목과 서버 부하 문제를 초래한다. 이를 해결하기 위해 서버 - 엣지 - 클라이언트의 3계층 구조를 갖는 계층적 연합학습(Hierarchical Federated Learning, HierFL)이 제안되었으며, 엣지 서버가 인접한 클라이언트 그룹의 중간 집계자로서 통신 효율성과 학습 속도를 개선하는 대안으로 주목받고 있다[2]. 하지만 HierFL 구조는 여전히 보안 위협에 취약하다. 특히 일부 악의적 클라이언트가 데이터의 라벨링을 변조하여 학습을 방해하는 라벨 플립 공격은 글로벌 모델의 성능을 심각하게 훼손한다[3]. 따라서 기존 HierFL에서는 엣지 서버가 악의적 업데이트를 별도의 검증 없이 상위 서버로 전달하기 때문에, 소수의 공격만으로도 전체 모델이 쉽게 오염된다. 이는 엣지 서버가 단순한 중계 노드의 역할로 머무는 한 구조적으로 해결하기 어려운 문제이다.

본 논문에서는 이러한 취약점을 해결하기 위해, 엣지 서버를 수동적인 집계자에서 능동적 방어 노드(active defense node)로 전환하는 2단계 방어 기법을 제안한다. 제안 기법의 핵심은 엣지 서버가 각 클라이언트가 생성한 지역 모델의 정확도를 독립적으로 평가하고, 이를 바탕으로 정확도 기반의 집계를 수행하는 것이다. 첫 번째 단계에서는 정확도 기반 동적 필터링을 통해 정확도가 낮은 클라이언트를 집계에서 제외하고, 두 번째 단계에서는 가중 집계를 통해 정확도에 따라 클라이언트의 기여도를 다르게 책정한다. 게다가, 다양한 공격 시나리오에서의 실험 결과, 제안 기법은 악의적 클라이언트에 대한 방어가 없는 기존 기법 대비 현저히 높은 신뢰성을 보였으며, 극한의 공격 환경에서도 안정적인 학습을 유지함을 확인하였다.

## Ⅱ. 시스템 모델

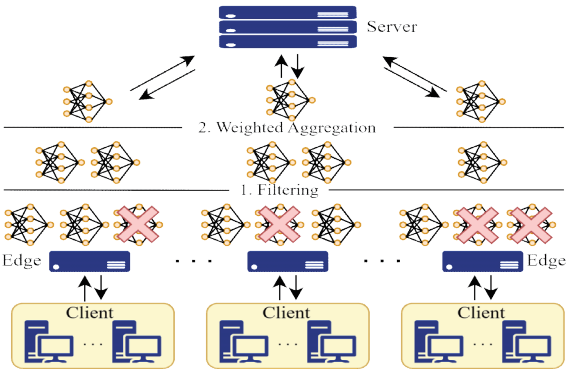



그림 1 동적 필터링과 가중 집계 기능을 갖는 계층적 연합학습 구조

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 모델

을 갱신한다. 갱신된 전역 모델은 다시 각 엣지 서버를 거쳐 클라이언트로 재배포되는 과정이 반복적으로 수행됨으로써 전체 시스템은 점차적으로 성능이 향상된 전역 모델로 완성된다.

## Ⅲ. 제안 기법

**Algorithm 1** Accuracy-based Filtering and Weighted Aggregation (at Edge Server)

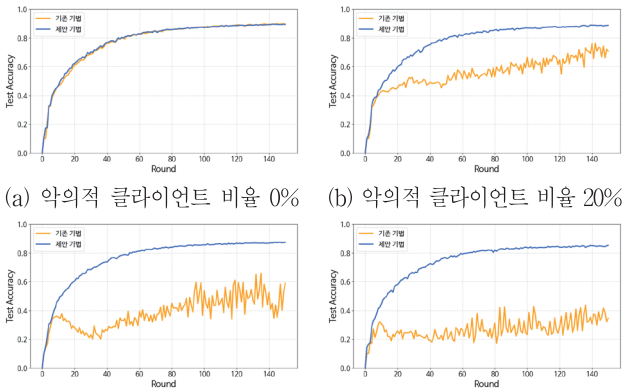
**Require:**
$r$: Current communication round
$W_r = \{w_c\}_{c=1}^{C}$: Set of client models at the edge
$D_{val}$: Public validation dataset at the edge
$H$: History of client accuracies $\{c : [acc_{r-K+1}, ..., acc_r]\}$
$N, M, K$: Hyperparameters
**Ensure:**
$\bar{w}_{edge}$: Aggregated edge model
$A \leftarrow []$
**for all** client model $w_c \in W_r$ **do**
  $acc_c \leftarrow \text{Evaluate}(w_c, D_{val})$
  Append $acc_c$ to $A$
  Update $H[c]$ with $acc_c$
**end for**
// Step 1: Accuracy-based Filtering
$A_{sorted} \leftarrow \text{Sort}(A, \text{descending})$
$acc_{avg\_topN} \leftarrow \text{Mean}(A_{sorted}[1 \text{ to } N])$
$T \leftarrow acc_{avg\_topN} \times (M/100.0)$
$C_{selected} \leftarrow \{\}$
**for** $c$ from 1 to $C$ **do**
  **if** $A[c] \geq T$ **then**
    Add $c$ to $C_{selected}$
  **end if**
**end for**
// Step 2: Weighted Aggregation
$\Omega \leftarrow []$
**for all** client $c \in C_{selected}$ **do**
  $\omega_c \leftarrow \text{Mean}(H[c])$
  Append $\omega_c$ to $\Omega$
**end for**
$\bar{w}_{edge} \leftarrow \frac{\sum_{c \in C_{selected}} \omega_c \cdot w_c}{\sum_{c \in C_{selected}} \omega_c}$
**return** $\bar{w}_{edge}$

알고리즘 1은 엣지 서버에서 2단계에 걸쳐 동작한다. 먼저, 엣지 서버는 독립적인 검증 데이터셋으로 모든 클라이언트의 정확도를 평가한다. 다음으로 첫 번째 단계인 동적 필터링 부분에서는 상위 N개의 모델의 평균 정확도를 기준으로 동적 임계값 M을 설정하여, 이에 미치지 못하는 정확도의 모델은 집계에서 제외한다(알고리즘 1에서 7~15번째 줄). 두 번째 단계인 가중 집계 부분에서는 필터링을 통과한 모델들을 대상으로 최근 K라운드의 평균 정확도를 각 모델의 가중치로 할당한다(알고리즘 1에서 21번째 줄).

## Ⅳ. 성능 평가

제안하는 기법의 성능평가를 위해 CIFAR-10 데이터셋이 사용되었다. 전체 학습 데이터의 10%는 클라이언트 모델을 평가하고 필터링하기 위한 검증 데이터셋으로 사용하였다. 나머지 90%의 학습 데이터는 50개의 클라이언트에게 분배되었으며, Non-IID 환경을 구성하기 위해 10개의 클래스 중 클라이언트당 2개의 클래스만 가지도록 구성하였다. 악의적 클라이언트는 무작위로 선택되며, 자신의 로컬 데이터셋 라벨을 변조하는 라벨 플립(label-flipping) 공격으로 모사하였다. 공격 환경의 강도를 다양하게 평가하기 위해 악의적 클라이언트 비율은 20%, 40%, 60%로 설정하였으며, 공격 강도(라벨 변경 확률)는 100%로 설정하여 실험을 수행하였다. 학습은 총 125 라운드 동안 진행되었으며, 계층 구조는 50개의 클라이언트와 5개의 엣지 서버로 구성되었다. 각 클라이언트는 라운드마다 5회의 로컬 업데이트를 수행하였다. 학습 모델은 ResNet-18을 기반으로 SGD(Stochastic Gradient Descent) 옵티마이저를 적용하였다(learning rate=0.05, momentum=0.9, weight decay=5e-4). 한편, 제안하는 2단계 방어 기법의 하이퍼파라미터는 N은 3, 임계 비율 M은 80%, K는 5로 설정하였다. 이러한 환경에 기반하여, 제안하는 방어 기법과 방어 없이 단순 평균으로 집계하는 기존 기법을 정확도 관점에서 성능을 비교하였다.

그림 2는 악의적 클라이언트의 비율에 따른 라운드별 전역 모델 정확도의 변화를 보여준다. 그림 2의 결과를 살펴보면, 악의적 클라이언트 비율이 증가할수록 기존 기법은 성능이 불안정하게 학습이 진행되고 있는 모습을 보여준다. 반면, 제안 기법은 악의적 클라이언트 비율에 상관없이 안정적으로 학습이 이루어지고 있는 모습을 보여준다.




(a) 악의적 클라이언트 비율 0% (b) 악의적 클라이언트 비율 20%

(c) 악의적 클라이언트 비율 40% (d) 악의적 클라이언트 비율 60%

그림 2 악의적 클라이언트 비율에 따른 라운드별 테스트 정확도

## V. 결론

본 논문에서는 라벨 플립 공격에 취약한 계층적 연합학습의 보안 문제를 해결하기 위해, 엣지 서버를 수동적인 중계 노드에서 능동적인 방어 노드로 전환하는 2단계 방어 기법을 제안하였다. 제안 기법은 엣지 서버가 클라이언트가 학습한 지역 모델의 정확도를 평가하여 정확도 기반 동적 필터링으로 신뢰도가 낮은 모델을 배제하고, 가중 집계를 통해 과거 성능이 우수한 클라이언트의 기여도를 높이는 방식으로 동작한다.

다양한 공격 시나리오에 대한 실험 결과, 제안 기법은 악의적 클라이언트가 다수 존재하는 가혹한 환경에서도 기존 기법 대비 높은 신뢰성을 유지하며, 안정적이고 신뢰할 수 있는 모델 학습 성능을 달성함을 확인하였다. 향후 연구로 필터링 강도의 동적 조절을 통해 다양한 공격 시나리오에 대해서도 효과적으로 방어할 수 있도록 알고리즘을 개선할 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] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84 - 90, 2017.

[2] Q. Qiu, Z. Wu, H. Wang, Q. Yang, Y. Wang, and C. Su, "Hierarchical Aggregation for Federated Learning in Heterogeneous IoT Scenarios: Enhancing Privacy and Communication Efficiency," *Future Internet*, vol. 17, no. 1, p. 18, Jan. 2025.

[3] A. Alam, S. M. M. Rahman, A. K. Das, and M. S. Hossain, "Label flipping attacks in hierarchical federated learning for intrusion detection in IoT," *IEEE Internet of Things Journal*, vol. 11, no. 6, pp. 10592-10604, Mar. 2024.