연합학습을 활용한 사물인터넷 기기 침입 탐지

TEAM 7 강현재, 이승화, 최현석

발표 목차

01

서론

- IoT 보안 문제의식
- 프로젝트 목표
- 베이스라인 논문 소개
- 베이스라인 논문 분석 및 결론

02

본론

- 베이스라인 재현 (Non-IID)
- Dev 1: 최신 데이터셋 적용 (CICIoT2023)
- Dev 2: 개인화된 연합 학습 (PFL)

03

결론

- 최종 성능 평가
- 결론

서론: IoT 보안의 위기

초연결 사회의 그림자

수십억 개의 IoT 디바이스가 연결된 현대 사회는 전례없는 보안 위협에 직면해 있습니다.

- **제한된 자원:** IoT 기기는 고성능 백신 탑재 불가
- **Mirai 봇넷:** 취약한 기기를 좀비화하여 공격에 악용
- **데이터 폭증:** 모든 데이터를 중앙 서버로 보내기엔 대역폭 부족

핵심 딜레마

기존의 중앙 집중식 보안 모델은 증가하는 데이터양을 감당할 수 없으며, 개인 프라이버시 침해 위험이 크다.

프로젝트 목표

Step 1

베이스라인 재현

- 논문: Rahman et al. (2020)
- 데이터: NSL-KDD
- 목표: 논문 모델 검증

→

Step 2

최신 데이터셋 적용

- 데이터: CICIoT2023
- 목표: 현실적인 최신 위협 반영

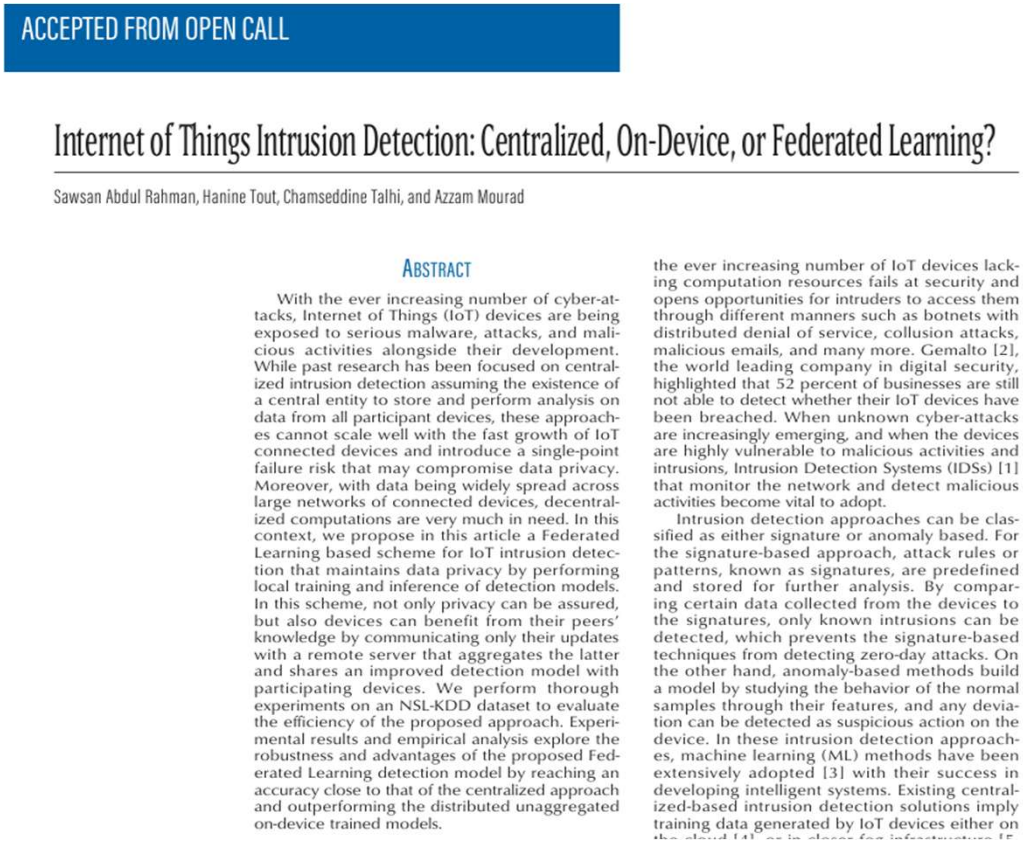
→

Step 3

학습방법 고도화

- 학습방법: PFL
- 목표: 성능 향상

베이스라인 논문 소개



ACCEPTED FROM OPEN CALL

Internet of Things Intrusion Detection: Centralized, On-Device, or Federated Learning?

Sawsan Abdul Rahman, Hanine Tout, Chamseddine Talhi, and Azzam Mourad

ABSTRACT

With the ever increasing number of cyber-attacks, Internet of Things (IoT) devices are being exposed to serious malware, attacks, and malicious activities alongside their development. While past research has been focused on centralized intrusion detection assuming the existence of a central entity to store and perform analysis on data from all participant devices, these approaches cannot scale well with the fast growth of IoT connected devices and introduce a single-point failure risk that may compromise data privacy. Moreover, with data being widely spread across large networks of connected devices, decentralized computations are very much in need. In this context, we propose in this article a Federated Learning based scheme for IoT intrusion detection that maintains data privacy by performing local training and inference of detection models. In this scheme, not only privacy can be assured, but also devices can benefit from their peers' knowledge by communicating only their updates with a remote server that aggregates the latter and shares an improved detection model with participating devices. We perform thorough experiments on an NSL-KDD dataset to evaluate the efficiency of the proposed approach. Experimental results and empirical analysis explore the robustness and advantages of the proposed Federated Learning detection model by reaching an accuracy close to that of the centralized approach and outperforming the distributed unaggregated on-device trained models.

the ever increasing number of IoT devices lacking computation resources fails at security and opens opportunities for intruders to access them through different manners such as botnets with distributed denial of service, collusion attacks, malicious emails, and many more. Gemalto [2], the world leading company in digital security, highlighted that 52 percent of businesses are still not able to detect whether their IoT devices have been breached. When unknown cyber-attacks are increasingly emerging, and when the devices are highly vulnerable to malicious activities and intrusions, Intrusion Detection Systems (IDSs) [1] that monitor the network and detect malicious activities become vital to adopt.

Intrusion detection approaches can be classified as either signature or anomaly based. For the signature-based approach, attack rules or patterns, known as signatures, are predefined and stored for further analysis. By comparing certain data collected from the devices to the signatures, only known intrusions can be detected, which prevents the signature-based techniques from detecting zero-day attacks. On the other hand, anomaly-based methods build a model by studying the behavior of the normal samples through their features, and any deviation can be detected as suspicious action on the device. In these intrusion detection approaches, machine learning (ML) methods have been extensively adopted [3] with their success in developing intelligent systems. Existing centralized-based intrusion detection solutions imply training data generated by IoT devices either on

논문 제목

Internet of Things Intrusion Detection: Centralized, On-Device, or Federated Learning?

저자

Sawsan Abdul Rahman, Hanine Tout, Chaneseddine Talhi, and Azzam Mourad

출처

IEEE Network(November/December 2020)

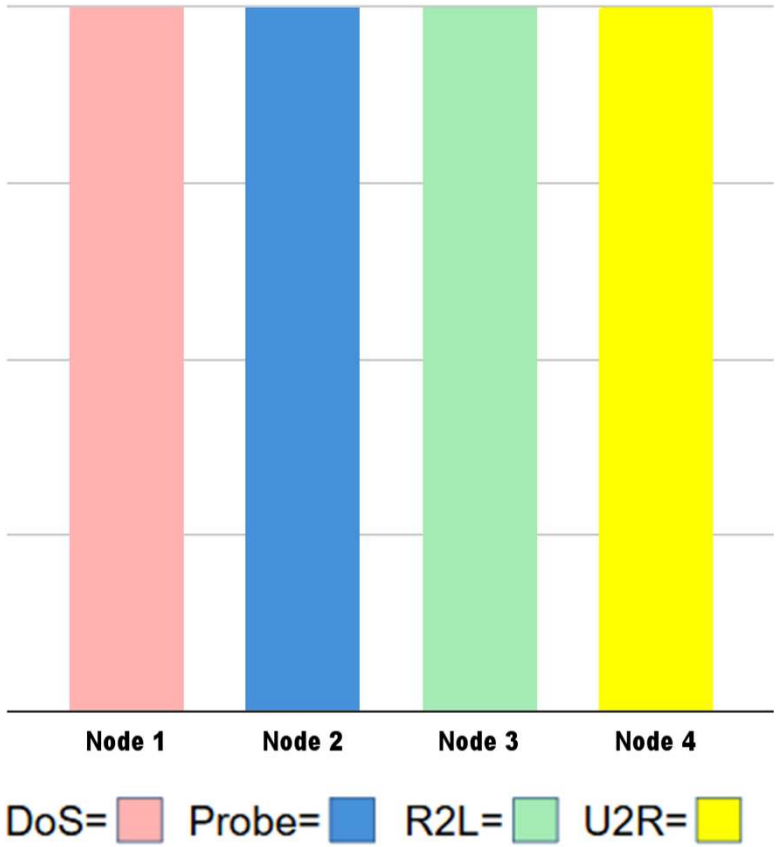
베이스라인 논문 분석

왜 기존 방식이 아닌 '연합' 학습인가?

구분	중앙 집중식 (Centralized)	온디바이스 (On-Device)	연합 학습 (Federated)
학습 위치	중앙 클라우드 서버	개별 로컬 기기	로컬 학습 + 서버 집계
데이터 이동	모든 원본 데이터 전송	이동 없음	가중치(Weight)만 전송
정확도	최상(글로벌 데이터 학습)	낮음(로컬 데이터 한정)	높음(중앙 집중식에 근접)
프라이버시	취약(데이터 유출 위험)	우수(데이터 외부 유출 없음)	우수(데이터 외부 유출 없음)
한계점	프라이버시 침해, 고비용	지식 고립(공격 취약)	클라이언트 관리 복잡성

IoT 침입 탐지를 위한 학습 패러다임 비교

베이스라인 논문 분석(Use Case 1)



비균일(Non-IID) 데이터 분포

- 4개 노드가 각각 다른 단일 공격 유형만 학습하는 극단적 상황

(Node 1: DoS, Node 2: Probe, Node 3: R2L, Node 4: U2R)

- 학습하지 않은 모든 공격 유형이 섞인 데이터로 진행
- Self-Learning(온디바이스)의 한계

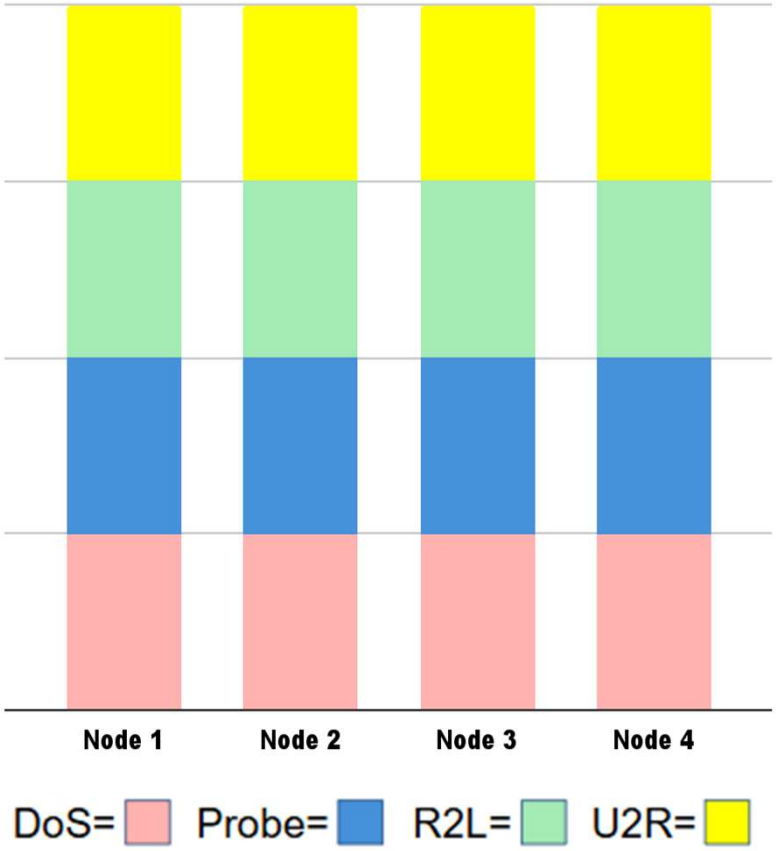
  학습하지 않은 공격 유형에 취약
- Federated-Learning의 성능

  모델 공유를 통해 공격 탐지 성공

  Node 3의 정확도 급상승 (49% -> 76.06%)

  중앙 집중식 모델(83.09%에 근접한 성능 달성)

베이스라인 논문 분석(Use Case 2)



균등한 데이터 분포 (Equal Data Distribution)

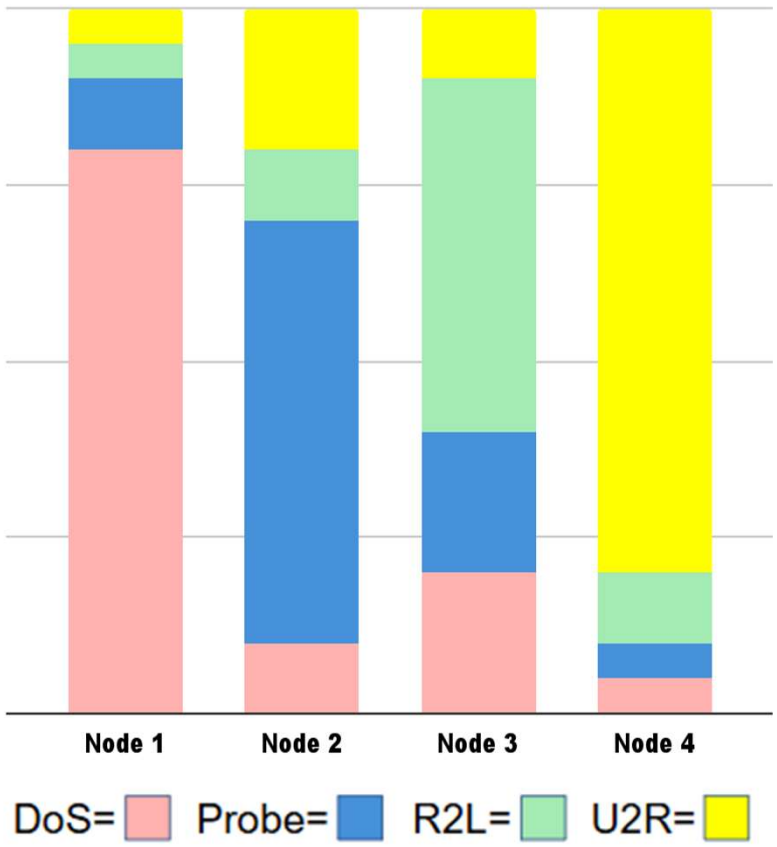
- 4개의 노드가 모든 공격 유형(DoS, Probe, R2L, U2R)을 골고루 학습하는 통계적으로 가장 이상적인 상황
- 데이터가 이상적으로 분포되었을 때, FL이 중앙 집중식 (Centralized) 모델만큼 똑똑해질 수 있는가?
- Self-Learning(온디바이스)의 결과

  데이터가 골고루 있어 혼자 학습해도 성능이 준수함

  (정확도 약 79.24% ~ 80.47% 도달)
- Federated-Learning의 성능

  중앙 집중식 모델의 정확도(83.09%)와 불과 1.6%p 차이로 대등한 성능을 입증

베이스라인 논문 분석(Use Case 3)



무작위 데이터 분포 (Random Data Distribution)

- 현실 세계처럼 데이터가 시간 흐름(Round)에 따라 노드에 무작위로 생성되고 누적되는 상황
- 예측 불가능한 현실적인 데이터 환경에서도 FL이 안정적인 성능을 유지하는가?
- Self-Learning(온디바이스)의 한계

  FL 모델에 비해 전반적으로 낮은 성능을 보임
- Federated-Learning의 성능

  모든 학습 라운드에서 Self-Learning보다 월등히 높은 정확도를 기록

  데이터 분포와 무관하게 중앙 집중식 모델에 근접하는 안정성을 보임

베이스라인 논문 결론

1. 대등한 성능

중앙 집중식 모델에 근접하는 높은 정확도(83.09%) 달성

2. 한계 극복

지식 공유를 통해, 모든 시나리오에서 개별 학습(Self-Learning)의 성능을 압도

3. 프라이버시 보호

데이터 외부 유출 없이 모델 파라미터만 공유하여 보안성 확보

베이스라인 논문 분석 (2): 실험 설계

Use Case #1: Non-IID 환경

4개의 노드가 **서로 다른 공격 유형만을 학습**하는 상황을 가정

Node 1

DoS

Node 2

Probe

Node 3

R2L

Node 4

U2R

핵심 검증 포인트

편향된 데이터만 학습한 노드가 연합학습을 통해 다른 공격유형도 탐지할 수 있을지 확인

베이스라인 데이터셋: NSL-KDD

NSL-KDD

- 1999년 KDD Cup 데이터셋을 개선한 버전.
- 침입 탐지 시스템 연구의 표준 벤치마크.
- 41개의 특성(Feature) 정보 포함.

NSL-KDD 침입 유형

1. DoS (Denail of Service, 서비스 거부)

- 시스템을 과부하 상태로 만들어 정상적인 사용자가 서비스를 이용하지 못하게 하는 공격유형
- NSD-KDD에서 가장 많은 비중을 차지하는 공격유형

2. Probe (Probing, 정보 수집)

- 추후 공격하기 위해 네트워크 구조, 열린 포트, 운영체제 버전 등을 수집하는 공격

3. R2L (Remote to Local, 원격 사용자 공격)

- 외부 사용자(Remote)가 내부(Local)계정을 탈취하는 공격
- 비밀번호 무작위 대입 공격 등이 포함

2. U2R (User to Root, 권한 상승)

- 일반 유저(User)가 관리자 권한(Root)을 탈취하는 공격
- 가장 위험도가 높은 공격 유형

재현 실험 환경 구축

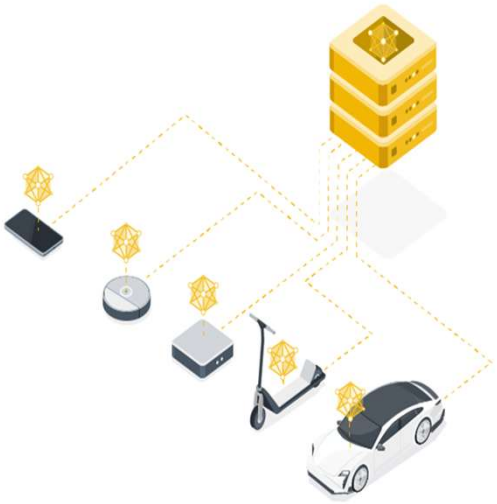
Flower 프레임워크 활용

실제 분산 학습 환경을 모사하기 위해

Flower(flwr) 라이브러리 사용.

- **Client:** 4개의 독립된 학습 프로세스
- **Server:** FedAvg 알고리즘으로 가중치 집계
- **Data Partition:** Use Case #1에 맞춰 공격 유형별 분할

Flower
A Friendly Federated Learning Framework



NSL-KDD 데이터셋

데이터 전처리

- 정답(정상, 공격)데이터를 이진 클래스로 레이블링
- 카테고리별 원-핫 인코딩 적용
- 수치형 데이터 Min-Max 스케일링 적용

Node (On-Device):
각 노드는 한가지 공격유형만 전담해 학습을 진행

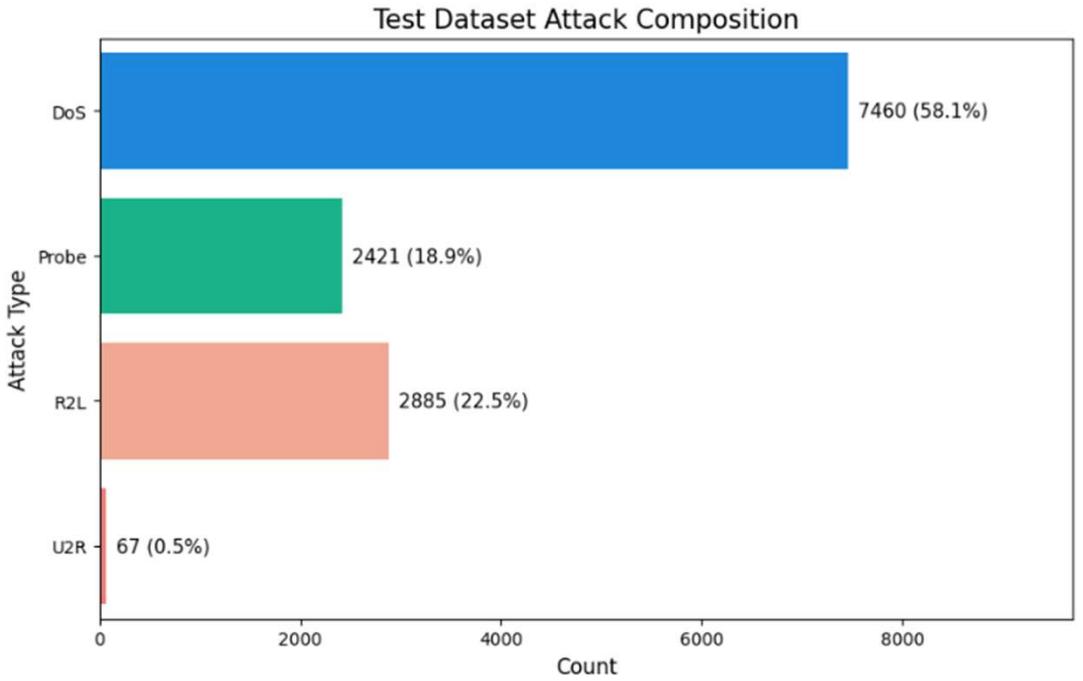
Test 테스트 상황

테스트 상황

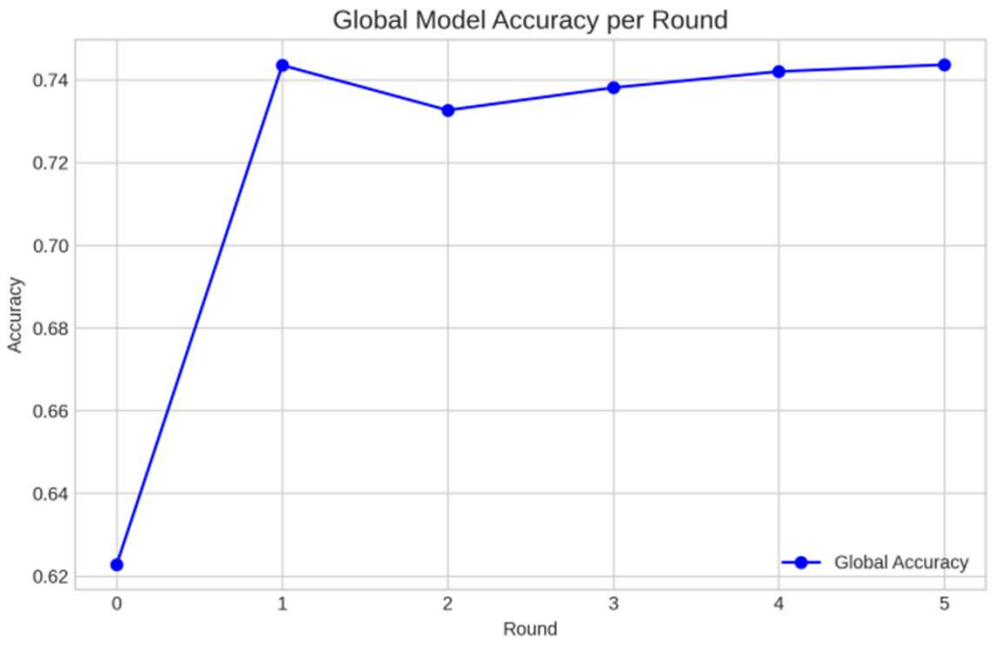
학습(Train) 데이터는 노드별로 공격 유형이 나뉘어 있지만,

성능 평가(Test) 데이터는 모든 공격 유형이 혼합되어 있음

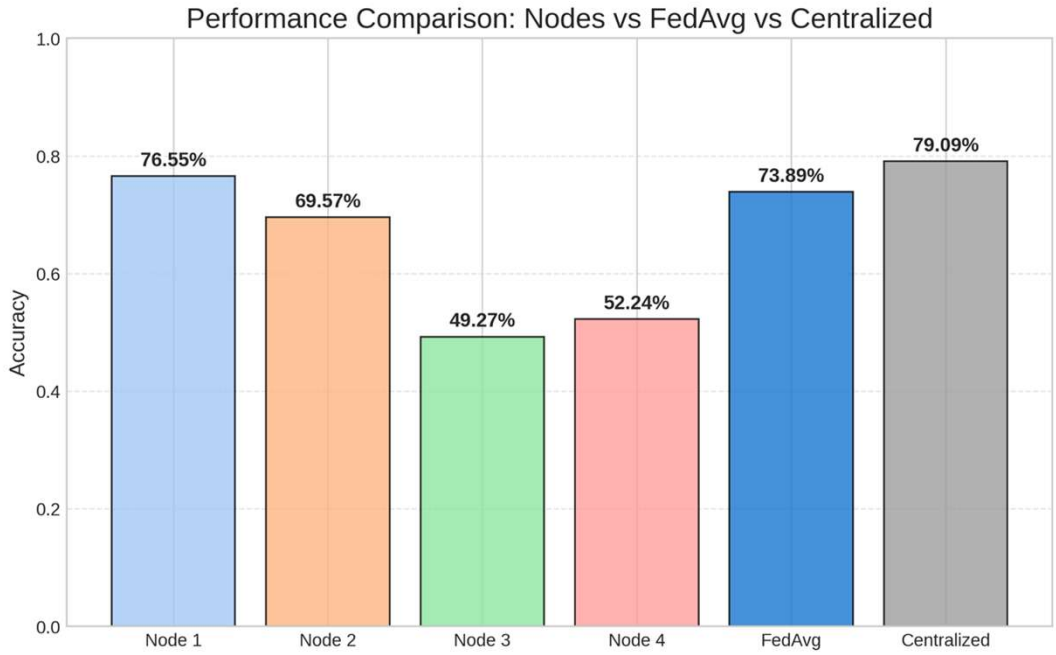
> **Node (On-Device)의 상황:**
> 학습은 DoS만 했지만, Test상황에서는
> DoS, Probe, R2L, U2R 을 모두 탐지해야해
> 정확도가 떨어짐



학습 결과

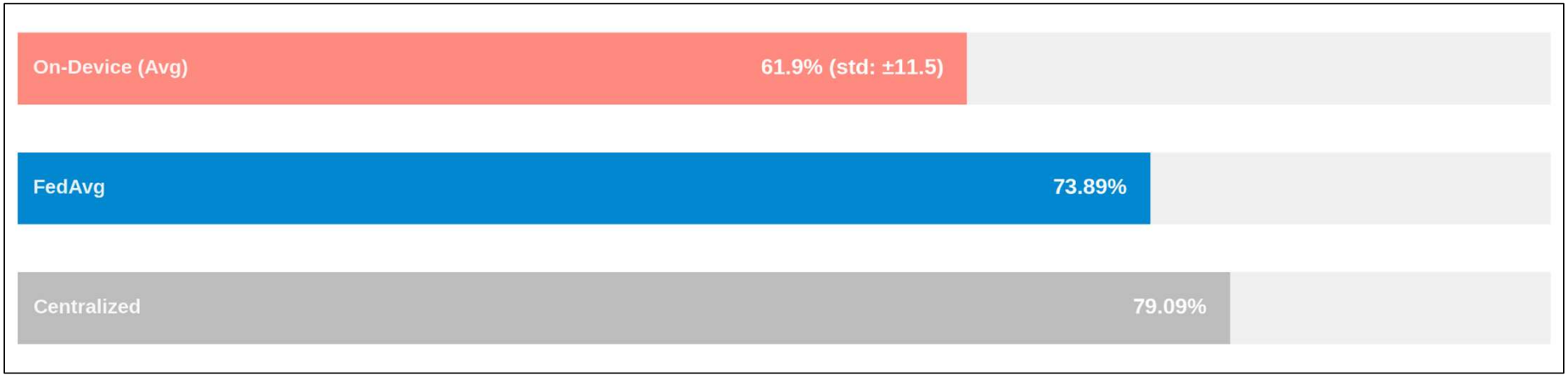


글로벌 모델(FedAvg) 정확도



정확도 비교

학습 결과



On-Device의 특징

자신이 학습한 공격 유형에 대해서는 탐지 성능이 높음

FL의 성과

파라미터(가중치) 공유를 통해 **모든 공격 유형에 대해 고른 탐지 성능**을 확보. 지식의 고립 문제 해결.

베이스라인의 한계

"1999년 데이터로 2025년의 보안을 논할 수 없다"

오래된 프로토콜, 단순한 공격 시나리오, 합성 데이터의 한계

Dev 1: 최신 데이터셋 적용으로 이동

Dev 1: CICIoT2023 데이터셋 도입

데이터셋 특징

캐나다 사이버 보안 연구소(CIC)에서

2023년 공개한 최신 벤치마크

- 105개의 실제 IoT 디바이스 기반 트래픽
- 4,600만 건 이상의 대규모 데이터
- 33종의 다양한 공격 시나리오

CICIoT2023 데이터셋

DDoS / DoS

서비스 거부

74.6% / 17.7%

Mirai

IoT 특화 봇넷, 좀비감염

5.8% (309,768개)

Spoofing

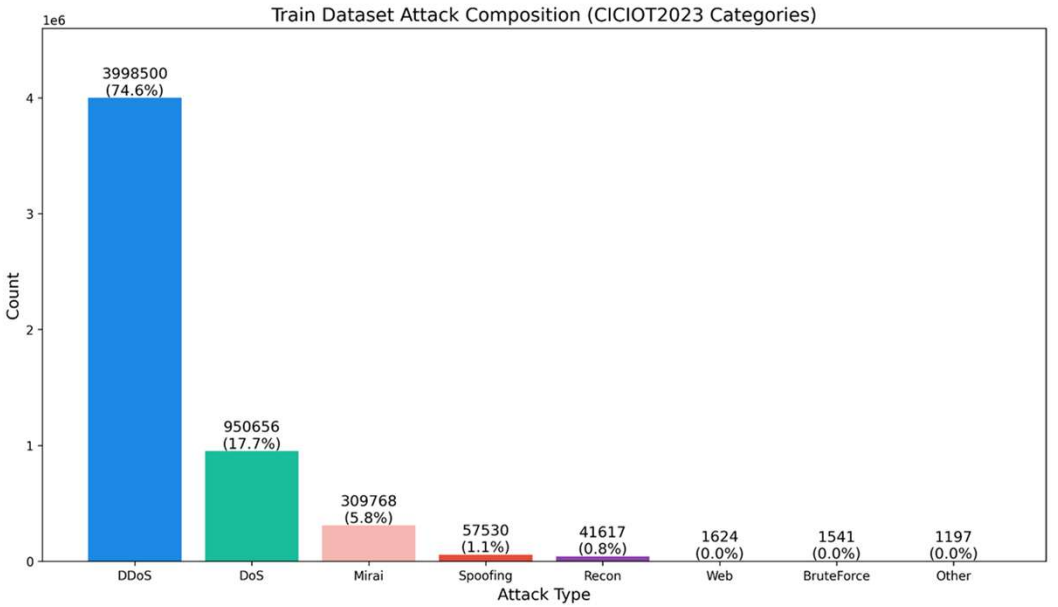
ARP/DNS 변조

1.1% (57,530개)

상위 4개 공격 유형

가장 많은 비중의

공격유형 4종 활용



데이터 전처리 (Preprocessing)

CICIoT2023

최신 데이터셋으로, NSL-KDD대비

전처리 과정이 간단함

- **One-hot Encoding:**

데이터셋에 이미 적용되어 있음

- **feature Dropping:**

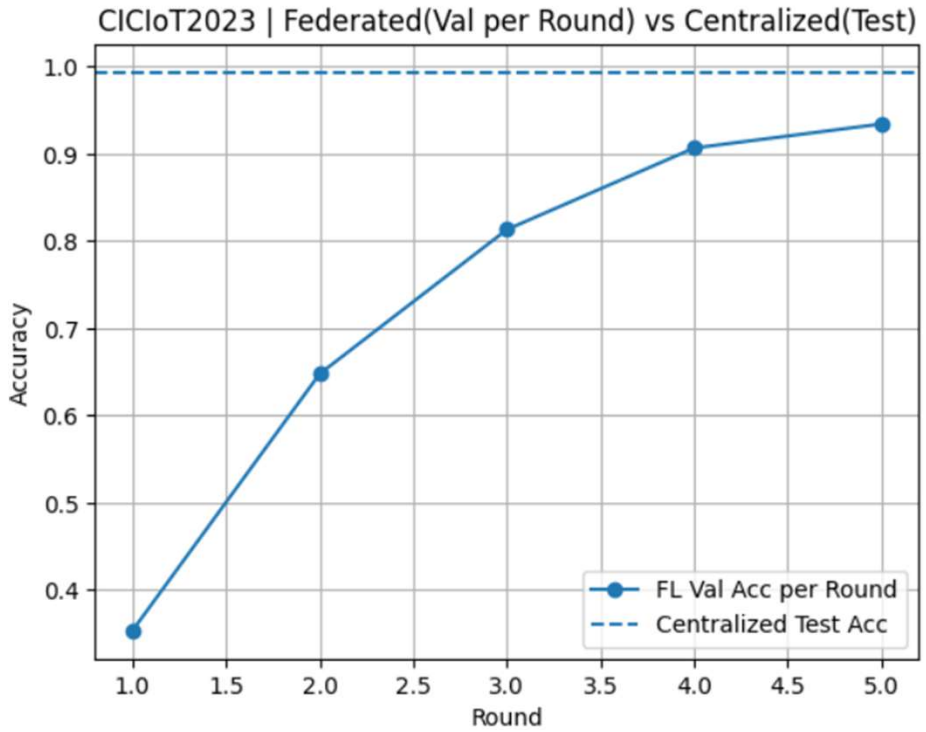
학습에 방해되는 열들을 사전에 제거

train.csv (1.62 GB)

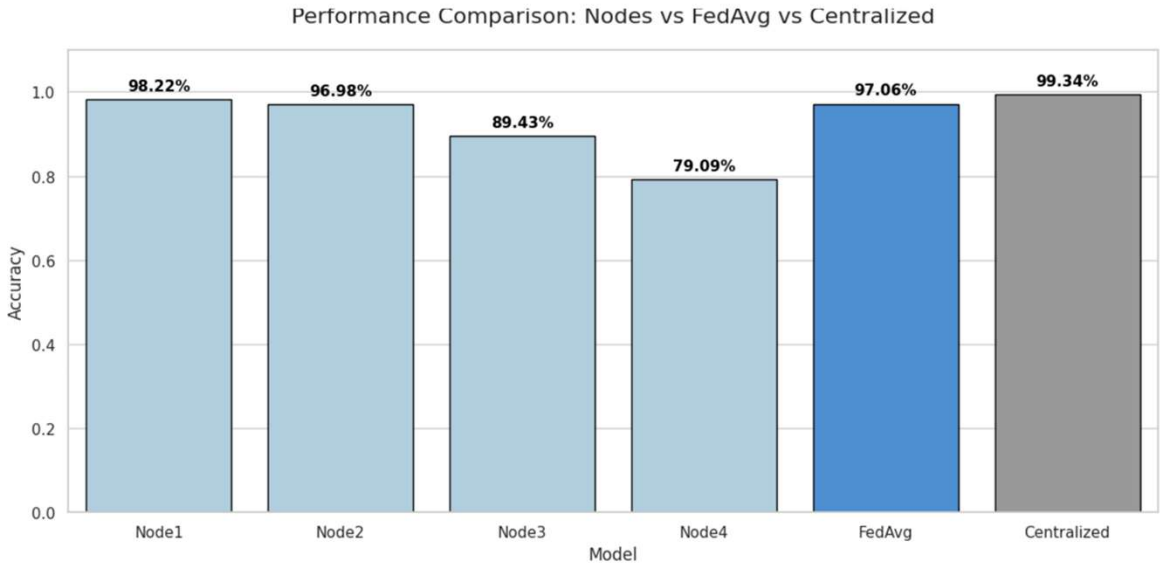
Detail Compact Column 47 of 47 columns

# Magnitue	# Radius	# Covariance	# Variance	# Weight	∆ label
41.84554580588544	761.4567603516936	305219.32230083307	0.95	141.55	DDoS-ACK_Fragmentation
10.392304845413264	0.0	0.0	0.0	141.55	DDoS-SYN_Flood
10.462812588467663	2.4452861688730807	16.85311781532524	0.19	141.55	DDoS-PSHACK_Flood
34.409301068170514	0.0	0.0	0.0	141.55	Mirai-greeth_flood

학습 결과



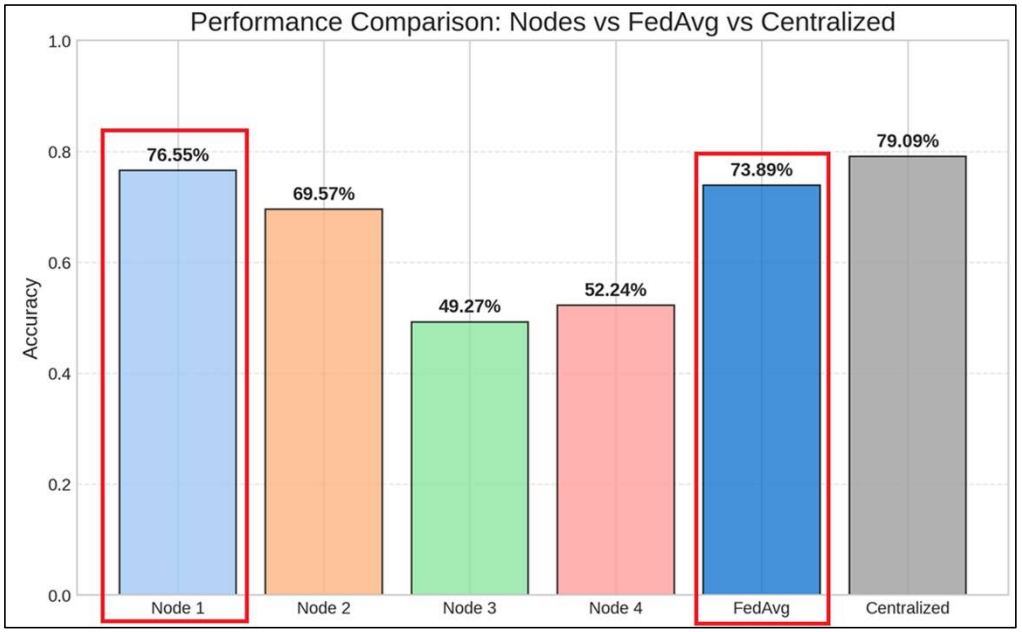
글로벌 모델(FedAvg) 정확도



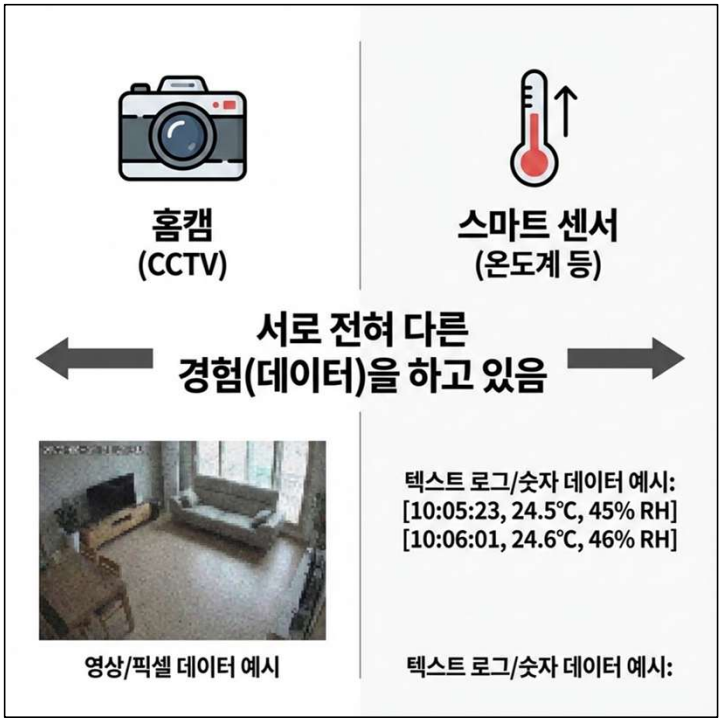
정확도 비교

Dev2:
Personalized Federated Learning
개인화된 연합학습

Dev 2: 개인화(Personalization)의 필요성

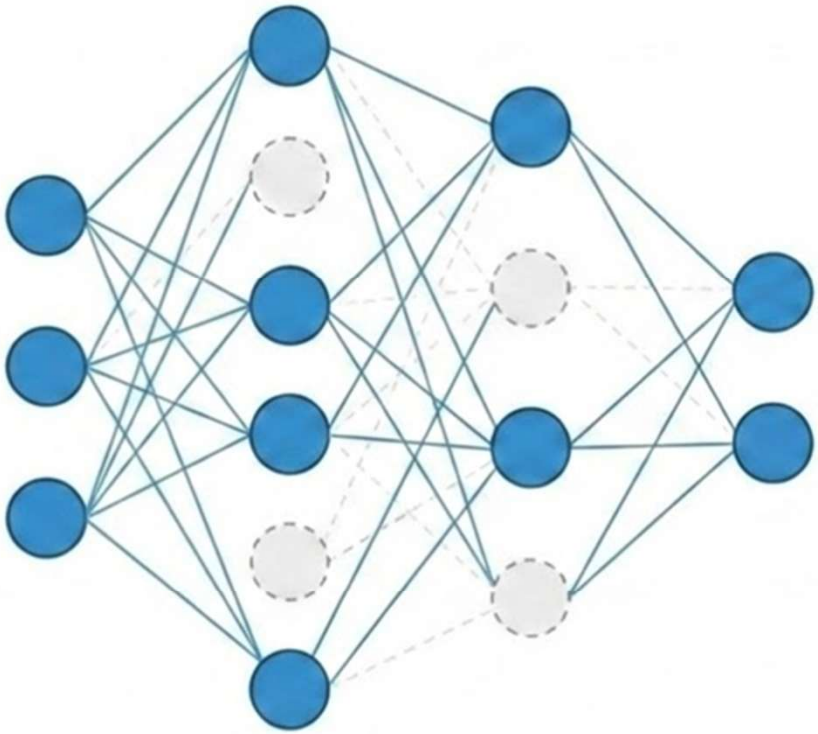


<baseline UseCase #1 결과>



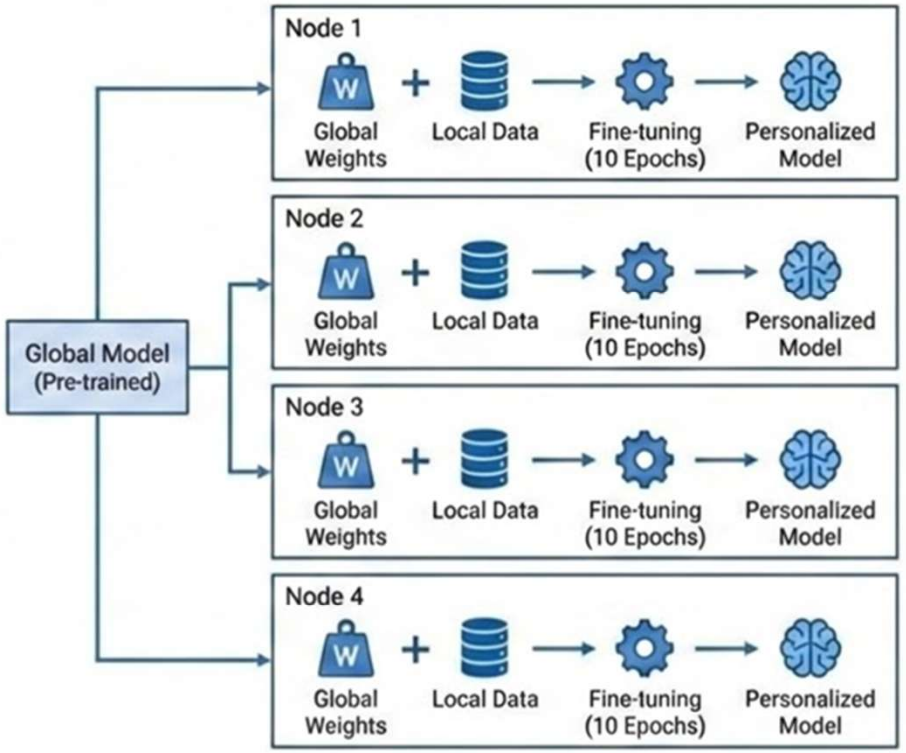
'평균의 함정'에 빠짐

Dev 2: PFL 구현 및 설계



모델 구조 개선

Dropout 비율: 0.2 -> 0.5 (과적합 방지 및 일반화 성능 향상)



PFL 적용 방법

Global Model (Pre-trained) 기반, 각 Node의 Local Data로 Fine-tuning (10 Epochs) 수행, 개인화된 모델 생성

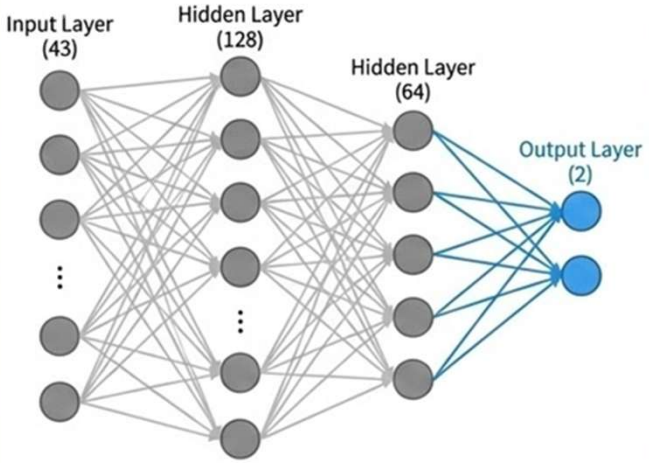
Dev 2: PFL 적용방법(1) - Head만 파인튜닝

모델 구조

- Input Layer
- Dense (66)
- Batch Norm
- Dropout
- Dense (67)
- Batch Norm
- Dropout
- Dense (68, Output)

적용 방법

출력 부분만 파인튜닝




```
for layer in p_model.layers[:-1]:
    layer.trainable = False
```

<결과>

FINAL PERFORMANCE COMPARISON				
Model Strategy	Node 1 (DDoS)	Node 2 (DoS)	Node 3 (Mirai)	Node 4 (Spoofing)
Self-Learning	0.9813	0.9521	0.5503	0.8572
FedAvg (Global)	0.9816	0.9816	0.9816	0.9816
Personalized FL	0.9810	0.9644	0.5740	0.9724

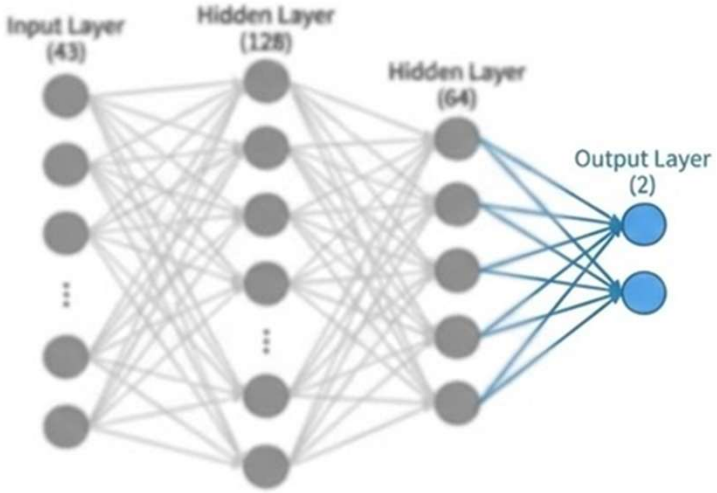
실험 결과 분석
대부분 on-device 보단 성능이 소폭 상승하지만, 오히려 글로벌 모델 대비 대폭 하락함

↓

성능 개선 실패

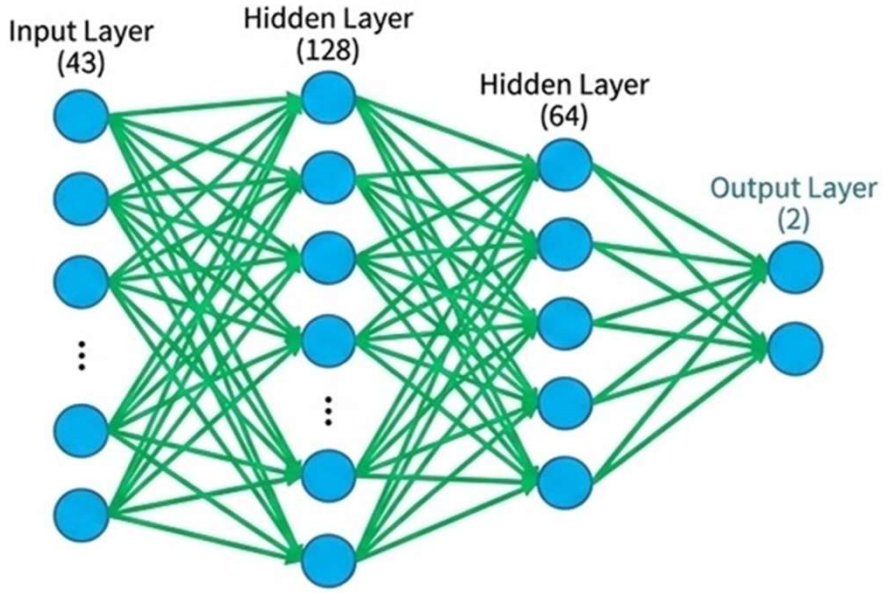
Dev 2: PFL 적용방법(2) - 모든 레이어 파인튜닝

부분 학습 (Partial Update) ❌



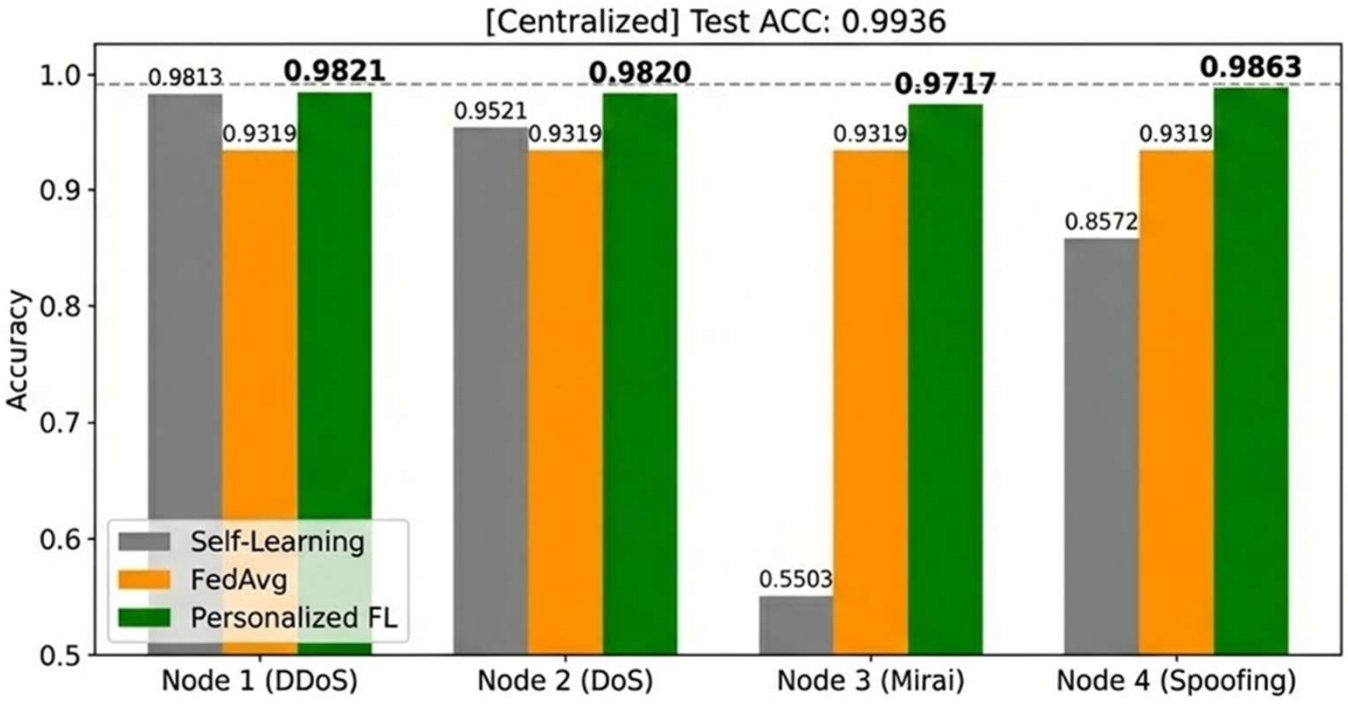
특징 추출 능력 고정으로 인한 한계

전체 미세 조정 (Full Fine-tuning) ✅




```
for layer in p_model.layers[:]:
    layer.trainable = True
```

Dev 2: PFL 적용방법(2) - 모든 레이어 파인튜닝 결과

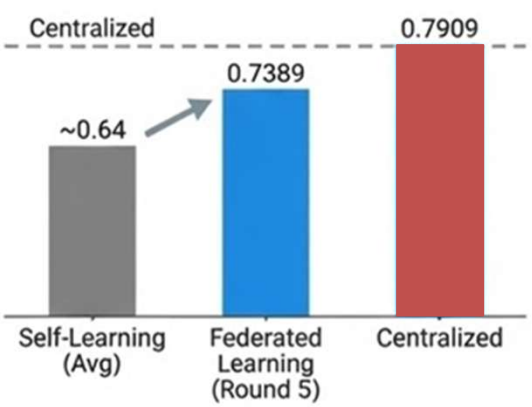



```
--- 1. 테스트셋 이진 레이블 분포 (y_test) ---
Normal (0)      27709
Attack (1)    1149142
Name: count, dtype: int64

=> 총 1176851개 중 Attack이 97.65% 차지

--- 2. 테스트셋 'Attack' 샘플의 카테고리별 분포 ---
AttackCategory
DDoS          855981
DoS           204245
Mirai          66749
Spoofing       12410
Recon           8812
Web              370
BruteForce       319
Other            256
Name: count, dtype: int64
```

< Test Dataset 분포 >

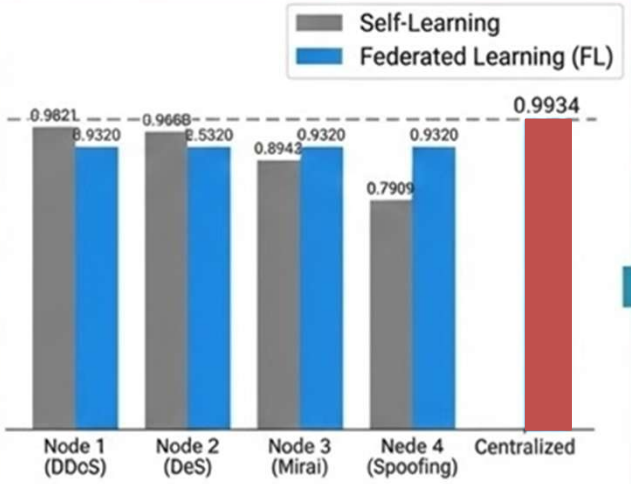
연합학습(FL) 모델 고도화 결과 정리

1. 베이스라인 (NSL-KDD)



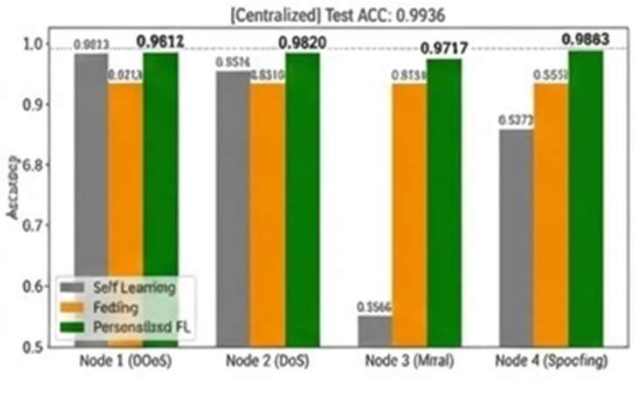
FL은 개별 학습된 노드들의 평균보다 높은 성능을 보이며, 중앙 집중식 모델과 큰 차이 없는 결과를 달성함.

2. 최신 데이터셋 (CICIoT2023)



특정 노드(Node 1, 2)에서는 FL이 Self-Learning보다 성능이 낮아지는 '평균의 함정' 발생.

3. 개인화 연합 학습 (PFL)



PFL을 통해 모든 노드에서 성능이 비약적으로 향상되었으며, 중앙 집중식 모델에 근접한 결과 달성.

결론

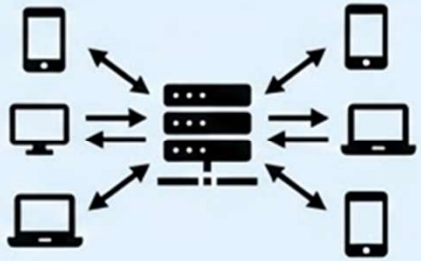
중앙집중형 학습

- 학습 용이성 높음
- 프라이버시 침해 위험
- 대역폭 문제

개별학습 (On-device)

- 프라이버시 보호 확실
- 성능 저하
- 데이터 고립

연합학습 (FL)



- 프라이버시 보호
- 준수한 성능
- 데이터 이질성 극복

연합학습이 현실적인 침입탐지, 보안의 주축이 될 것이다.

Q & A

감사합니다.

참고 자료

1 - Internet of Things Intrusion Detection: Centralized, On-Device, or Federated Learning? Sawsan Abdul Rahman, Hanine Tout, Chamseddine Talhi, and Azzam Mourad

2 - PERSONALIZED FEDERATED LEARNING VIA SEQUENTIAL LAYER EXPANSION IN REPRESENTATION LEARNING Jaewon Jang, Bonjun Choi Computer Science and Engineering