

# 연합 학습에서 엣지 디바이스를 위한 계층적 전문가 혼합

김재현<sup>1</sup>, 최봉준<sup>2</sup>

<sup>1</sup>승실대학교 컴퓨터학과 석사과정

<sup>2</sup>승실대학교 컴퓨터학과 교수

teeraiser@soongsil.ac.kr, davidchoi@soongsil.ac.kr

## Hierarchical Mixture-of-Experts for Edge Device in Federated Learning

Jae-Heon Kim<sup>1</sup>, Bong-Jun Choi<sup>2</sup>

<sup>1</sup>Dept. of Computer Science and Engineering, Soongsil University

<sup>2</sup>Dept. of Computer Science and Engineering, Soongsil University

### 요 약

연합 학습은 최근 기계 학습 분야에서 발생할 수 있는 다양한 문제들을 해결했지만, 학습의 주체가 서버에서 클라이언트로 이동함에 따라 클라이언트 장치의 컴퓨팅 자원의 한계가 새로운 문제로 부각되었다. 클라이언트의 장치는 중앙 집중적 서버와 비교하여 상대적으로 적은 컴퓨팅 자원을 보유하고 있으며, 특히 엣지 디바이스와 같은 클라이언트의 장치는 현저히 적은 컴퓨팅 자원으로 인해 일반적인 연합 학습 절차에 효과적으로 참여하기 어렵다. 본 연구는 클래스 계층 구조와 계층적 전문가 혼합을 통해 엣지 디바이스가 연합 학습에 효과적으로 참여할 수 있도록 하였다. 이 기법은 CIFAR-100 과 Tiny ImageNet 데이터셋을 통해 효용성을 입증하였으며, 기존 기법과 비교해 라운드당 학습 시간과 메모리 사용량이 낮음을 보인다.

### 1. 서론

연합 학습(Federated Learning)[1]은 개인정보 보호 기계 학습의 한 방식으로, 중앙 집중형 기계 학습 환경에서 발생할 수 있는 여러 문제를 해결하기 위해 제안되었다. 기존 기계 학습 환경에서는 사용자의 데이터를 서버로 전송함으로써 개인정보가 유출될 수 있는 문제가 존재하며, 사용자가 생성하는 데이터의 품질이 증가함으로써 높은 통신 비용이 발생한다. 연합 학습은 사용자의 데이터를 서버로 전송하는 것이 아닌, 클라이언트의 장치에서 직접 학습을 수행하고, 중앙 서버는 학습된 모델을 수집, 집계 및 전송하는 역할을 수행하는 방식으로 학습의 주체를 중앙 서버에서 클라이언트로 옮겨 데이터 유출을 방지하는 동시에 통신 비용을 줄였다.

연합 학습은 기존에 있던 여러 문제를 해결했지만, 새로운 문제점들을 야기하였다. 각 클라이언트가 자신의 데이터와 기기를 활용하여 학습하기에 각 클라이언트의 환경이 이질적일 수 있으며, 이는 모델 학습에 부정적인 영향을 끼친다. 클라이언트의 데이터

분포가 비독립 동일 분포(Non-Independent and Identically Distributed; Non-I.I.D)일 경우, 모델이 적절하게 수렴되기 어려우며 학습된 글로벌 모델이 클라이언트의 환경에 적합하지 않을 수 있다. 또한 클라이언트가 보유한 기기 간의 컴퓨팅 자원이 이질적일 경우, 낙오자(Straggler)로 인해 연합 학습 수행 속도가 느려지거나 글로벌 모델의 성능이 저하될 수 있으며, 연합 학습에 참여할 수 있는 클라이언트가 제한될 수 있다[2].

한편, 기존 기계 학습 모델의 아키텍처는 분리할 수 없는 단일 구조를 가지고 있다. 이러한 구조를 가진 경우, 해당 모델은 새로운 클래스의 추가 등으로 인한 목표 작업이 변경되는 경우 학습한 기존 모델을 사용할 수 없으며, 새로운 모델을 학습해야 한다. 연합 학습에서 이러한 문제는 클라이언트의 많은 자원을 소요하여 학습한 지식을 잃는다는 것을 의미하며, 클라이언트에게 새로운 작업에 대한 학습을 다시 요청해야 한다[3].

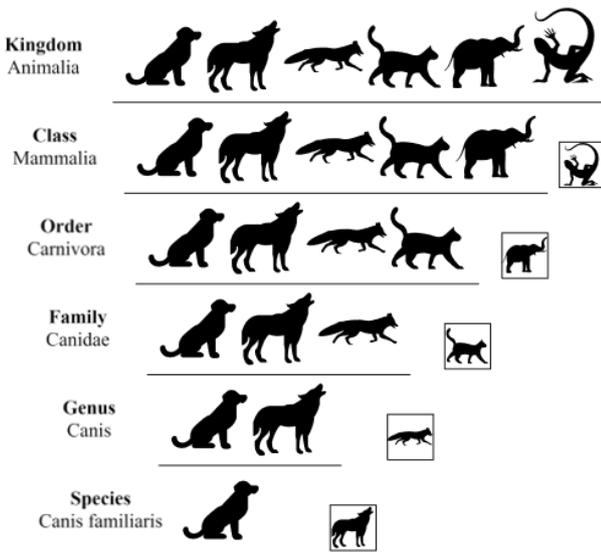
본 연구에서는 엣지 디바이스를 대상으로 하는 연

합 학습 환경에서 계층 분류(Hierarchical Classification) 문제를 계층적 전문가 혼합(Hierarchical Mixture-of-Experts)을 통해 해결한다. 전문가들의 독립성과 계층 모델의 활용을 통해 기존의 하나의 작업을 여러 단순한 작업으로 분할하여 독립적으로 수행함으로써 (1) 각 라운드 당 클라이언트에게 요구되는 학습 비용과 통신 비용을 줄임으로써 다양한 종류의 클라이언트가 학습에 참여할 수 있도록 하는 동시에 낙오자를 줄이며, (2) 각 작업에 대한 지식을 계층 구조를 기반으로 독립적으로 유지함으로써 작업의 확장성을 보장하여 새로운 작업을 적은 비용으로 학습할 수 있으며, 클라이언트에게 작업에 대한 유연성을 제공하여 클라이언트의 모델을 추가적인 학습 없이 환경에 적합하게 개인화 할 수 있다.

2. Preliminaries

2.1. 계층 분류

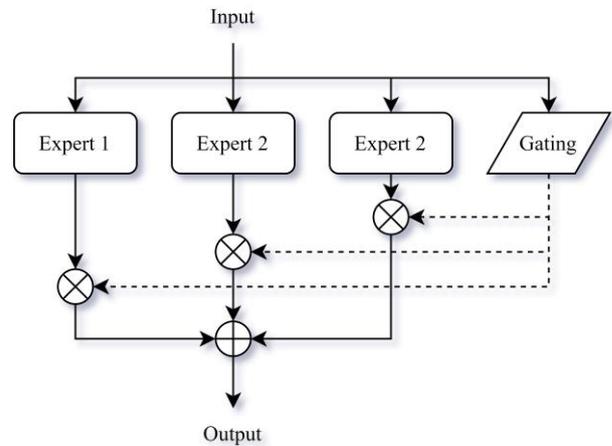
계층 분류(Hierarchical Classification)는 다중 레이블 분류(Multi-label Classification) 문제 중 하나로, 여러 수준의 클래스가 트리 또는 DAG(Directed Acyclic Graph)로 구성되어 각 수준에 해당하는 클래스로 분류하는 문제이다. 루트 클래스를 제외한 모든 클래스에는 상위 클래스가 존재하며, 트리 또는 DAG의 상위 수준에 속해 있을수록 더 포괄적인 개념을 가진다. 계층 분류 문제는 요구되는 분류 수준에 따라 반드시 리프 노드 중 하나로 분류되어야 하는 필수 리프 노드 예측(Mandatory Leaf Node Prediction)과 상위 노드로도 예측이 가능한 비필수 리프 노드 예측(Non-Mandatory Leaf Node Prediction)으로 분류된다. 그림 1은 계층 분류의 대표적인 예시인 생물 분류 문제를 보여준다.



(그림 1) 계층 구조에 따른 생물 분류 문제. 상위 분류는 모든 하위 분류가 보유한 공통 특성에 따라 분류됨. 테두리가 표시된 생물은 상위 분류에는 속하였으나, 하위의 표기된 분류와 다른 범주로 분류되는 생물.

2.2. 계층적 전문가 혼합

전문가 혼합(Mixture-of-Experts)은 복잡한 입력 공간을 여러 개의 전문가 네트워크가 분할하여 처리하는 기계 학습 기법이다. 일반적인 전문가 혼합은 그림 2와 같이 하나의 게이팅 네트워크와 둘 이상의 전문가 네트워크로 구성된다. 전문가 네트워크는 전체 작업 중 특정 부분에 특화되어 있으며, 게이팅 네트워크는 여러 전문가 네트워크의 처리 결과에 대한 가중치를 결정한다. 이를 통해 하나의 복잡한 작업을 여러 개의 단순한 작업으로 나눈 후, 각 전문가 네트워크가 서로 다른 작업에 특화됨으로써 기존의 복잡한 작업을 해결한다. 계층적 전문가 혼합은 이를 계층적으로 구성하여 더 복잡한 작업을 효과적으로 처리할 수 있다.

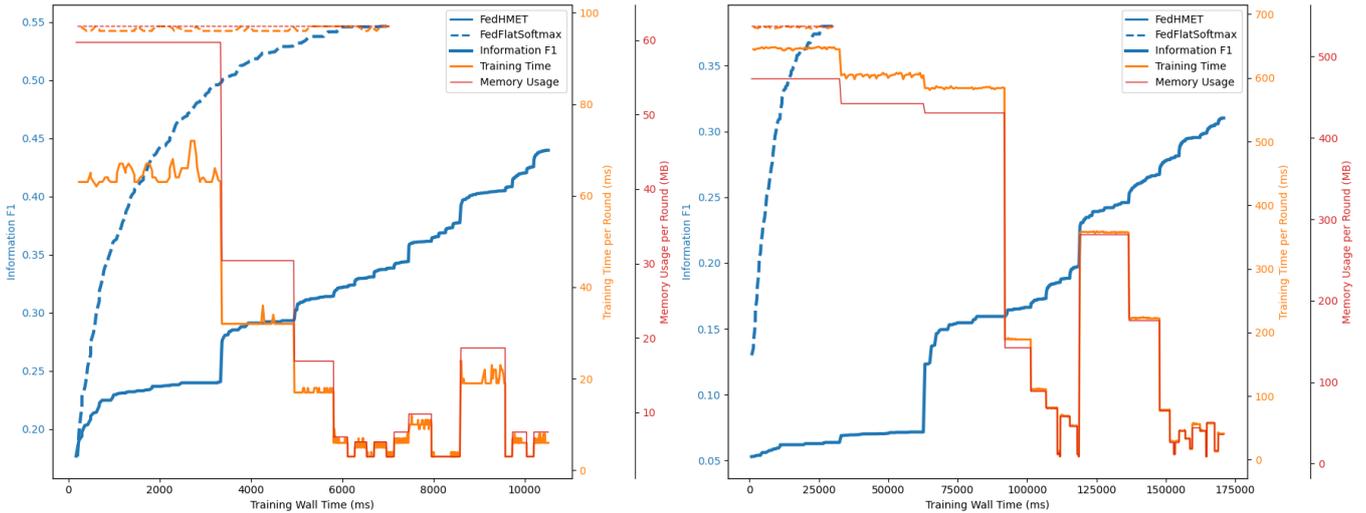


(그림 2) 3 개의 전문가 네트워크와 게이팅 네트워크로 구성된 전문가 혼합 모델. 각 전문가는 입력의 특정 특성에 특화되어 있으며, 게이팅 네트워크는 입력을 기반으로 전문가 네트워크 결과의 가중치를 결정함.

3. 연구 제안

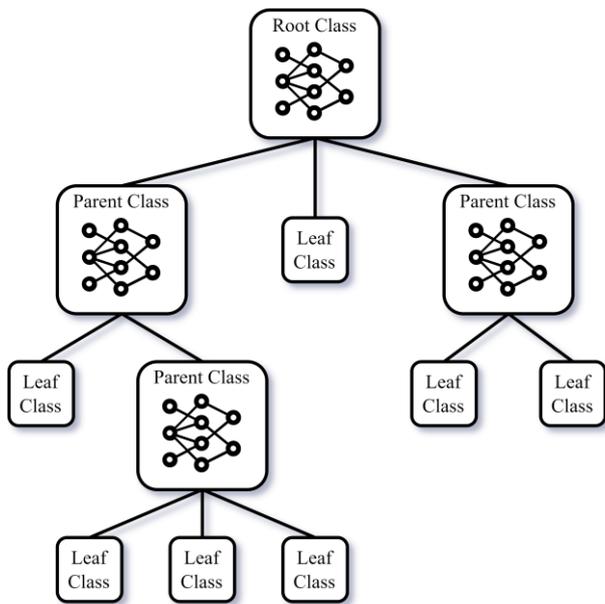
본 연구에서는 연합 학습 환경에서 엣지 디바이스가 계층 분류 문제를 효과적으로 해결하기 위한 계층적 전문가 혼합 모델을 제안한다.

이 계층적 전문가 혼합 모델은 계층 분류 문제를 분류 체계(Taxonomy)를 따르는 클래스 계층 구조(Class Hierarchy) 기반의 부모 노드별 지역 분류기(Local Classifier per Parent Node; LCPN) 접근법으로 해결한다. 클래스 계층은 트리 구조로 되어 있으며, 각 부모 노드에 전문가 네트워크를 지역 분류기로 지정함으로써 각 전문가 네트워크는 이미지를 해당 노드에서 자식 클래스 중 하나로 분류하는 작업에 특화된다. 또한 부모 노드에 위치한 전문가는 예측 결과에 따라 하나의 자식 노드의 전문가를 활성화시키는 게이팅 네트워크의 역할을 수행한다. 그림 3은 추상적



(그림 4) CIFAR-100 과 Tiny ImageNet 에서 제안한 방법과 Flat Softmax 를 Information F1, 학습 시간, 메모리 사용량을 비교한 그래프. 왼쪽 그래프는 CIFAR-100 의 실험 결과이고, 오른쪽 그래프는 Tiny ImageNet 의 실험 결과.

인 계층 전문가 혼합 모델을 보인다.



(그림 3) 계층 전문가 혼합 모델 예시. 각 전문가는 계층 구조의 부모 클래스에 위치해 자식 클래스 중 하나로 분류하는 작업에 특화됨. 이때 각 전문가와 작업은 다른 전문가와 작업에 독립적.

각 전문가들은 서로 독립적인 기계 학습 모델이며 서로 다른 작업을 명시적으로 할당하였기 때문에, 한 전문가의 학습은 독립적으로 수행할 수 있다. 연합 학습에서는 각 라운드마다 한 개의 전문가를 학습함으로써 기존 단일 모델 기반의 연합 학습과 비교하여 한 라운드에 요구되는 학습 시간과 메모리 요구량을 줄일 수 있다. 한편, 기존의 데이터 라벨은 원본 작업, 즉 기존의 단일 수준의 분류 작업을 위해 지정되어 있지만, 전문가 혼합 모델에서는 각 전문가가 원

본 작업이 아닌 분할된 소형 작업을 수행한다. 이때, 분류 체계 기반 계층 모델에서는 여러 하위 클래스를 포함하는 유일한 상위 클래스가 존재하므로, 데이터가 속한 범주의 라벨로 재지정함으로써 기존 데이터를 계층 모델 학습에 사용할 수 있다.

모델의 추론은 효과적인 컴퓨팅 자원 활용을 위해 하향식 탐욕적 예측을 사용한다. 먼저, 루트 전문가가 입력을 예측하여 자식 클래스 중 하나로 분류한다. 분류된 클래스가 리프 노드가 아니라면 해당 노드에 위치한 전문가에서 예측을 다시 수행하여 그 자식 클래스 중 하나로 분류한다. 이를 반복하여 리프 노드에 도착할 때까지 반복한다. 이를 통해 한 번에 활성화되는 모델을 제한하는 동시에 입력을 최종 예측까지 하향식으로 분류할 수 있다.

#### 4. 성능 평가

본 연구에서는 제안한 방법을 평가하기 위해 단일 모델 기반으로 계층 분류를 수행하는 Flat Softmax[4]와 비교하였으며, 불균형 클래스 계층을 다루는 계층 분류 문제에서 모델 성능을 적절하게 평가할 수 있는 Information Recall 과 Information Precision[4]을 기반으로 한 Information F1 을 통해 평가하였다. 사용한 데이터셋은 CIFAR-100 데이터셋[5]과 Tiny ImageNet 데이터셋[6]으로, CIFAR-100 은 단일 수준의 분류를 위한 fine label 과 다섯 개의 fine label 로 구성된 상위 분류인 coarse label 로 구성되어 있으며, 이를 기반으로 클래스 계층을 구성하였다. Tiny ImageNet 은 ImageNet[7] 데이터셋의 일부분을 축소된 저해상도 이미지로 구성한 데이터셋으로, ImageNet 의 기반인 WordNet[8]을 사용해 클래스 계층을 구성하였다. 그

< 표 1 > CIFAR-100 과 Tiny ImageNet 에서 제안한 방법과 Flat Softmax 를 Information F1, 학습 시간, 메모리 사용량을 비교한 표. 학습 시간과 메모리 사용량의 백분율은 Flat Softmax 대비 제안한 방법의 사용량을 나타냄.

데이터셋		CIFAR-100			Tiny ImageNet		
모델	지표	Information F1	라운드당 학습 시간 (ms)	라운드당 메모리 사용량 (MB)	Information F1	라운드당 학습 시간 (ms)	라운드당 메모리 사용량 (MB)
	Flat Softmax		0.54632	97 (100%)	61.9 (100%)	0.37979	681 (100%)
제안한 방법	최대	0.43977	72 (74.2%)	59.7 (96.4%)	0.31015	649 (95.3%)	472.7 (87.7%)
	최소		3 (3.1%)	4.1 (6.6%)		5 (0.7%)	8.6 (1.6%)

림 4 와 표 1 은 CIFAR-100 과 Tiny ImageNet 에서 Flat Softmax 와 제안한 방법의 Information F1, 각 라운드별 클라이언트의 평균 학습 시간과 메모리 사용량을 보여준다. 제안한 방법

은 두 데이터셋에서 Flat Softmax 에 비해 낮은 Information F1 을 보였으나, 클라이언트 기기에서 한 라운드 학습에 소요되는 시간이 적고, 메모리 사용량이 낮다.

### 5. 결론

본 연구는 엣지 디바이스를 위한 연합 학습에서, 계층적 전문가 혼합을 통해 계층 분류 문제를 자원 효율적으로 해결하는 방법을 제시하였다. 이 방법을 통해 한 라운드에 소요되는 학습 시간과 메모리 사용량을 줄임으로써 적은 컴퓨팅 자원을 가진 엣지 디바이스도 연합 학습에 참여할 수 있게 하여 클라이언트와 학습 데이터의 다양성을 확보하여 낙오자 발생을 억제할 수 있다. 또한 계층 모델 기반의 독립적인 전문가를 통해 모델이 확장성과 유연성을 가져 여러 상황에 적은 비용으로 대응할 수 있다. 그러나 제안한 방법은 하향식 탐욕적 예측 방식으로 인해 Flat Softmax 에 비해 상대적으로 낮은 Information F1 성능을 보인다. 또한 각 라운드에서 소요되는 컴퓨팅 자원은 줄었으나, 전체 모델의 크기는 커지면서 전체 학습 시간은 증가하였다. 향후 이러한 문제들을 해결하기 위해 다양한 접근법에 대한 검토가 필요하다.

### 사사문구

본 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2022R1A2C4001270). 또한, 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 융합보안핵심인재양성사업의 연구 결과로 수행되었음 (IITP-2024-RS-2024-00426853).

### 참고문헌

- [1] MCMAHAN, Brendan, et al. Communication-efficient learning of deep networks from decentralized data. *Artificial intelligence and statistics*. Fort Lauderdale, Florida, USA. PMLR, 2017. p. 1273-1282.
- [2] Li, Tian, et al. "Federated optimization in heterogeneous networks." *Proceedings of Machine learning and systems 2* (2020): 429-450.
- [3] Dong, Jiahua, et al. Federated class-incremental learning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. New Orleans, Louisiana, USA. 2022. p. 10164-10173.
- [4] Valmadre, Jack. "Hierarchical classification at multiple operating points." *Advances in Neural Information Processing Systems*. New Orleans, Louisiana, USA. 35 (2022): 18034-18045.
- [5] Krizhevsky, Alex, and Geoffrey Hinton. "Learning multiple layers of features from tiny images." (2009): 7.
- [6] Le, Ya, and Xuan Yang. "Tiny imagenet visual recognition challenge." *CS 231N 7.7* (2015): 3.
- [7] Deng, Jia, et al. Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*. Miami, Florida, USA Ieee, 2009. p. 248-255.
- [8] Fellbaum, Christiane. "WordNet: An electronic lexical database." *MIT Press google schola 2* (1998): 678-686..