탈중앙 연합학습 환경에서 Feature-Map 기반 모델 융합 시 Layer 별 Distillation 효과 분석

양다인, 이주형*
가천대학교

dainyang@gachon.ac.kr, * j17.lee@gachon.ac.kr

Layer-wise Analysis of Feature-Map-Based Model Fusion in Decentralized Federated Learning

Dain Yang, Joohyung Lee*
Gachon Univ.

요 약

탈중앙 연합학습(Decentralized Federated Learning, DFL)은 서버 없이 클라이언트 간 peer-to-peer 방식으로 모델을 공유하며 학습하는 분산 학습 구조이다. DFL 환경에서는 각 클라이언트가 로컬 데이터로 학습한 모델을 교환·융합하는 과정에서, 클라이언트 간 데이터 분포 차이로 인해 융합된 모델의 성능이 저하되는 client-drift 문제가 발생한다. 기존 연구에서는 이러한 문제를 완화하기 위해 Knowledge Distillation 을 활용한 모델 융합 방식, 특히 feature-map 기반 distillation 기법이 제안되었으나, 증류가 적용되는 네트워크 layer 위치에 따른 성능 영향에 대한 분석은 충분히 이루어지지 않았다. 본 연구에서는 DFL 환경에서 feature-map 기반 distillation 을 적용할 때, 증류 대상 layer 에 따른 모델 성능 변화를 체계적으로 분석한다. 다양한 layer 에서의 feature-map distillation 실험을 통해, layer 위치에 따른 지식 전달 효과의 차이를 비교·분석하고, 가장 효과적인 feature-map distillation 전략을 규명한다.

Ⅰ. 서 론

연합학습(Federated Learning, FL)은 중앙 서버가 다수의 클라이언트로부터 로컬 데이터를 직접 수집하지 않고, 각 클라이언트가 보유한 데이터로 로컬 모델을 학습한 후 학습 결과만을 공유함으로써 글로벌 모델을 학습하는 분산 학습 패러다임이다 [1]. 이러한 구조는 데이터 프라이버시 보호와 통신 비용 절감 측면에서 장점을 가지며, 의료, 스마트 디바이스, 엣지 AI 환경 등 데이터 이동이 제한되는 다양한 응용 분야에서 주목받고 있다.

기존 FL 은 중앙 서버가 모델 집계를 담당하는 중앙집중형 구조를 기반으로 하지만, 서버 장애에 대한 취약성과 네트워크 병목 문제 등의 한계를 가진다. 이를 해결하기 위해 최근에는 중앙 서버 없이 클라이언트 간 직접 모델을 교환·융합하는 탈중앙 연합학습 (Decentralized Federated Learning, DFL) 구조가 제안되었으며, 이는 시스템의 확장성과 강건성을 향상시킨다.

그러나 DFL 환경에서는 클라이언트 간 Non-Independent and Identically Distributed (Non-IID) 데이터 분포로 인해 모델 차이가 누적되는 client drift 문제가 더욱 심각하게 발생한다. 단순한 파라미터 평균 기반의 모델 융합 방식은 이러한 분포 차이를 효과적으로 반영하지 못해 성능 저하를 초래한다.

이러한 client-drift 문제를 완화하기 위한 대안으로, 최근에는 Knowledge Distillation(KD)을 활용한 모델 융합 방식이 활발히 연구되고 있다 [2]. KD 기반 접근법은 하나의 모델이 다른 모델로부터 지식을 전달받는 형태로, 파라미터 공간이 아닌 출력 또는 표현 공간에서의 지식 전달을 가능하게 한다. DFL 환경에서도 KD 를 활용하여 클라이언트 간 모델을 융합함으로써, 비동질적 데이터 분포로 인한 모델 간 차이를 완화하고 보다 안정적인 학습을 달성하고자 하는 연구들이 제안되었다.

기존 KD 기반 모델 융합 연구의 대부분은 logit 기반 distillation, 즉 모델의 최종 출력 값을 증류 대상으로 활용한다. 이러한 방식은 구현이 간단하다는 장점이 있으나, 학습 결과만을 반영할 뿐 모델 내부의 표현(representation)에 대한 정보는 충분히 전달하지 못한다는 한계를 가진다. 특히, 클라이언트 간 데이터 분포 차이가 큰 DFL 환경에서는 출력 수준의 지식만으로는 모델 간 표현 격차를 효과적으로 줄이기 어렵다. 이에 따라, 중간 layer 의 feature-map 을 증류 대상으로 활용하는 feature-map 기반 distillation 기법이 보다 풍부한 표현 정보를 전달할 수 있는 대안으로 주목받고 있다.

Ⅱ. 본론

2.1 Feature-Map 기반 탈중앙 연합학습

Feature-map-based knowledge distillation 은 서로 다른 데이터 분포에서 학습된 모델 간의 중간 표현을 직접 정렬함으로써, 출력 로짓(logit) 수준의 지식 전달만으로는 포착하기 어려운 의미적 특징 정보(semantic information)를 효과적으로 공유하는 기법이다. 기존의 지식 증류 기법은 주로 최종 출력 분포 간의 차이를 최소화하는 방식에 의존하였으나, 이러한 접근은 Non-IID 환경에서 각 클라이언트가 학습한 특징 공간의 차이를 충분히 반영하지 못하는 한계를 가진다.

탈중앙 연합학습 환경에서는 각 클라이언트가 서로 다른 데이터 분포를 보유하고 있기 때문에, 동일한 입력에 대해서도 모델 내부의 feature representation 이

크게 상이하게 형성될 수 있다. 이로 인해 단순한 로짓 기반 지식 전달은 출력 공간의 일치만을 강제할 뿐, 모델이 데이터를 해석하는 내부 표현 구조까지 일관되게 정렬하지 못한다. Feature-map-based distillation 은 이러한 문제를 해결하기 위해, 모델의 중간 layer 에서 추출된 feature map 을 직접 증류 대상으로 활용한다.

[그림 1] 처럼, 먼저 로컬 업데이트 클라이언트는 자신의 데이터로 로컬 모델을 업데이트한다. 그 후 업데이트 한 모델을 서버에게 전송한다. 서버는 받은 모델과 자신의 모델을 로컬 모델로 학습하며 knowledge distillation 을 사용하여 서로의 지식을 전달 하도록 한다. 제안된 FM-kt 는 Feature-map 과 logit 을 모두 증류하여 더 풍부한 지식을 공유하도록 하여 융합된 모델의 generalization 성능을 높인다. 지식증류 과정에서 Loss function 은 교차 엔트로피 손실, Kullback- Leibler divergence 를 이용한 로짓 기반 지식 증류 손실, 그리고 두 모델 간 feature representation 의 L2 거리를 측정하는 feature-map distillation 손실로 구성된다.

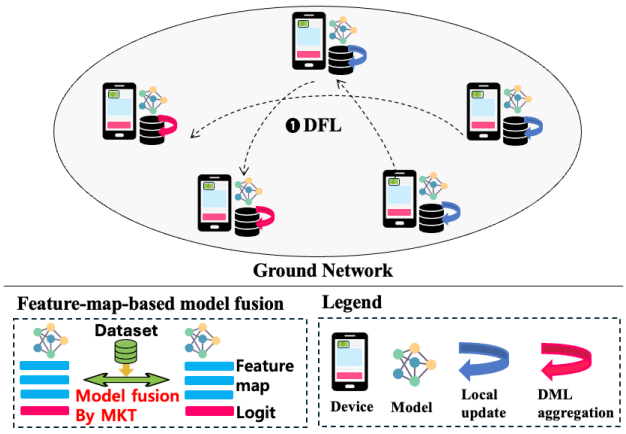

그림 1. FM-kt 프레임워크

2.2 실험결과

본 연구는 다양한 layer 에서 Feature-map distillation 을 적용한 실험을 수행하여, layer 위치에 따른 지식전달 효과의 차이를 분석하고, 최적의 feature-map distillation 방법을 찾는다. CIFAR10 데이터셋과 ResNet18 모델을 사용하여 실험을 진행하였다. 이 모델은 총 4 개의 layer 에서 feature-map 을 추출한다. 데이터 분포가 Non-IID 환경에서는 각 클라이언트가 전체 10 개의 클래스 중 8 개 (segment = 8)와 4 개 (segment = 4)만 가지도록 설정하였다. Segment 수가 8 인 경우에는 100 라운드에서의 정확도를 측정하였으며, Segment 수가 4 인 경우에는 50 라운드에서의 정확도를 측정하였다.

Layer	Segment = 8	Segment = 4
1	37.46	25.70
2	36.84	25.75
3	37.53	25.73
4	37.68	25.76
All	36.88	25.91

그림 2. Layer 별 정확도 차이

[그림 2]에서는 Segment8 인 경우, 1, 3, 4 번 째 layer 만 distillation 을 하였을 때 더 좋은 성능을 보이는 것을 알 수 있다. 모든 layer 에 대해 distillation 을 적용하는 경우 연산 비용이 크게 증가함에도 불구하고 성능 향상이 제한적이므로, 효율적인 feature-map distillation 전략이라고 보기 어렵다. Segment4 는 50 라운드 실행 결과인데, 50 라운드에서는 각 레이어의 정확도 차이가 크지 않은 것을 알 수 있다.

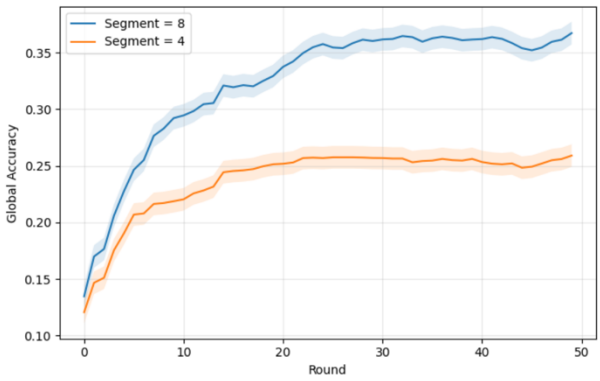

그림 3. Segment 별 Global accuracy 비교

[그림 3]은 feature-map distillation 이 클라이언트 간 서로 다른 표현 정보를 효과적으로 교환함으로써, 학습 과정에서 안정적인 학습 추이와 수렴 특성을 보임을 입증한다. 특히, 상대적으로 완화된 Non-IID 환경에서는 feature-map distillation 을 적용한 경우가 그렇지 않은 경우보다 더 빠른 수렴 속도와 높은 정확도를 달성하는 경향을 확인할 수 있다.

Ⅲ. 결론

본 논문에서는 탈중앙 연합학습 환경에서 feature-map 기반 knowledge distillation 의 layer 위치에 따른 성능 영향을 분석하였다. 실험 결과, 모든 layer 를 증류하는 방식보다 특정 중간 layer 에서의 feature-map distillation 이 더 효과적인 지식 전달과 성능 향상을 제공함을 확인하였다. 본 연구는 DFL 환경에서 효율적인 feature-map distillation 전략 설계에 대한 실험적 근거를 제시한다.

ACKNOWLEDGMENT

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 2026 년도 SW 중심대학사업의 결과로 수행되었음"(2021-0-01389)

참 고 문 헌

[1] McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." Artificial intelligence and statistics. PMLR, 2017.

[2] C. Li, G. Li, and P. K. Varshney, "Decentralized Federated Learning via Mutual Knowledge Transfer," IEEE Internet of Things Journal, vol. 9, no. 2, pp. 1136-1147, Jan. 2021.