Feature-map Distillation 기반 모델 융합 연구 및 QKD 적용 가능성 논의

양다인, 이주형

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

A Study on Model Fusion Based on Feature-map Distillation and Discussion of Its Applicability to Quantum Key Distribution (QKD)

Dain Yang, Joohyung Lee
Gachon Univ.

요 약

엣지 AI (Artificial Intelligence) 환경에서는 데이터 프라이버시 보호와 통신 보안을 동시에 만족하는 학습 방식이 요구된다. 본 연구는 QKD (Quantum Key Distribution) 기반 보안 채널 위에서 탈중앙 연합학습에 Teacher 모델의 지식을 Student 모델에게 전달하는 지식 증류 기법 중 하나인 Feature-map Distillation 을 적용한 Def-FD (Decentralized Federated Learning via Feature-map Distillation) 프레임워크를 제안한다. Def-FD 는 탈중앙 연합학습 환경에서 모델을 융합할 때 모델의 출력값뿐만 아니라 중간 표현까지도 증류 대상에 포함시켜 지식 전달의 효과를 향상한다. 본 논문에서는 다양한 feature distillation 방식에 대한 실험과 가중치의 변화에 따른 성능 분석을 통해, 가장 효과적인 증류 방식과 설정을 규명하였다. 실험 결과는 distillation 구성에 따른 모델 정확도의 민감도를 확인하고, 효율적인 파라미터 설정의 가능성을 시사한다.

Ⅰ. 서론

엣지 AI (Artificial Intelligence) 는 사용자 단말 또는 네트워크 엣지에서 데이터를 실시간으로 처리하여 지연을 줄이고 프라이버시를 강화하는 장점이 있으나, 분산 구조 특성상 전송 중 보안 위협에 취약하다. 이러한 문제를 해소하기 위해 각 클라이언트가 로컬 데이터를 외부에 공개하지 않고 자체적으로 모델을 학습하는 연합학습(Federated Learning, FL) 방식이 제안되었으며, 이는 데이터 프라이버시를 효과적으로 보호할 수 있다. 양자암호통신 인프라 (Quantum Cryptography Network, QCN)와 같은 보안 채널을 통해 모델 관련 정보만을 교환함으로써 원본 데이터의 유출 없이 협력 학습이 가능하지만, 기존의 FL 은 서버 의존성, 확장성의 한계 등의 구조적 제약을 여전히 가지고 있다.

이를 해결하기 위해 탈중앙 연합학습(Decentralized Federated Learning, DFL)이 제안되었다. DFL 은 중앙 서버 없이 클라이언트 간의 Peer-to-Peer 방식으로 모델을 교환하며, 크게 두 단계로 이루어진다. 먼저 클라이언트는 로컬 데이터를 기반으로 모델을 학습한 뒤, 해당 모델을 이웃 클라이언트에게 전달한다. 모델을 전달받은 클라이언트는 이를 자신의 모델과 융합하여 글로벌 모델을 생성하게 된다. 기존 방식은 전달받은 모델의 파라미터를 평균 내는 방식으로 융합하지만, 이 경우 서로 다른 데이터 에서 학습된 파라미터를 단순히 합치는 데 따른 성능 저하 문제가 발생한다.

이러한 한계를 극복하기 위해 최근에는 단순 평균 방식 대신, 지식 증류 (Knowledge Distillation)를 기반으로 한 모델 융합이 주목받고 있다. 예를 들어, Def-KT (Decentralized Federated Learning via Mutual Knowledge Transfer) [1]는 상호 지식 전달을 활용하여 탈중앙 연합학습 환경에서도 더 나은 글로벌 모델을 형성할 수 있음을 보였다.

본 연구에서는 이러한 지식 증류 기반 융합의 효과를 더욱 극대화하기 위해, Feature Map Distillation 을 결합한 탈중앙 연합학습 프레임워크인 Def-FD (Decentralized Federated Learning via Feature-map based Knowledge Distillation)를 제안한다. 이 방식은 모델 간의 출력값 (logit)뿐만 아니라 중간 표현(feature map)까지도 증류 대상으로 포함시켜 보다 풍부한 정보 교환을 가능하게 한다. 이를 통해 모델 간 지식 전달의 효율을 높이고 글로벌 모델의 성능을 향상시킨다. 또한, QCN 의 강력한 보안 기능과 결합되어, 본 방식은 양자암호 네트워크에서도 실용적으로 활용이 가능하다.

Ⅱ. 본론

가. Feature-map distillation

지식 증류(Knowledge Distillation, KD)는 Student 모델이 Teacher 모델의 soft output 값을 참고하여 자신의 모델을 업데이트함으로써 Teacher 의 지식을 전달받는 방식이다. 여기서 soft output 은 softmax 함수를 적용한 logit 의 확률 분포를 의미한다. Feature-map distillation 은 모델의 출력값뿐만 아니라, convolutional layer 에서 추출된 중간 표현(feature map)까지도 증류 대상으로 포함시킨다. 이처럼 더 깊이

있는 정보를 공유함으로써, 탈중앙 연합학습 환경에서 모델 융합 시 글로벌 모델의 성능을 향상시킬 수 있다. 그림 1 에서 모델이 3 개의 Convolutional Layer 와 1 개의 Fully Connected (FC) Layer 로 구성되어 있다고 가정하면, 각 convolutional layer 의 출력값인 feature map 과 FC layer 의 출력값인 logit 모두 증류 대상으로 활용된다.

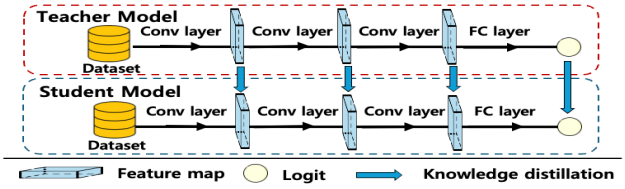

그림 1. Feature-map distillation

나. Def-FD

제안된 Def-FD 는 그림 2 와 같이 학습 과정을 진행한다. 먼저, 한 클라이언트는 자신의 로컬 데이터셋을 사용하여 모델을 로컬 업데이트한 후, 업데이트된 모델을 이웃 클라이언트에게 무작위로 전송한다. 모델을 전달받은 클라이언트는 자신의 모델과 전달받은 모델을 융합하여 새로운 글로벌 모델을 생성하는데, 이 과정에서 feature-distillation 을 활용하여 양측 모델의 정보를 최대한 효과적으로 통합한다. 이를 통해 데이터 프라이버시를 강화하면서도 서로 다른 모델간의 지식 증류 효과를 극대화할 수 있는 효과적인 협력 학습이 가능하다.

한편, 양자 컴퓨팅의 발전으로 기존 암호 알고리즘이 무력화될 가능성이 커지면서, 절대적 보안을 제공하는 양자 키 분배 (Quantum Key Distribution, QKD) 기반의 QCN 인프라의 필요성이 부각되고 있다. QKD 는 양자역학 원리를 통해 도청을 원천 차단하고 무결성을 보장한다. 이러한 QKD 인프라 위에서 Def-FD 기반의 AI 모델 학습을 수행하면 엣지 환경에서의 민감 데이터 보호와 모델 파라미터 전송의 안전성은 물론, 성능 향상까지 동시에 확보할 수 있다.

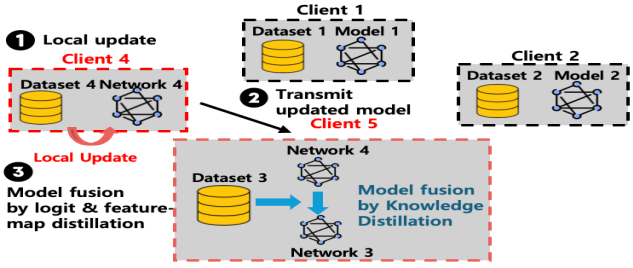

그림 2. Def-FD 의 프레임 워크

다. 실험 결과

제안한 FD-KT 기법의 성능을 분석하기 위해, Feature distillation 손실 방식(L2, COS, MSE)과 해당 loss 의 가중치(0.01- 0.03) 변화에 따른 실험을 수행하였다. Cross-entropy loss 및 logit distillation loss 는 각각 1 로 고정하였고, 실험은 Fashion-MNIST 데이터셋 기반의 Non-IID (Non-Independent and Identically Distributed) 환경에서 진행되었다. 각 클라이언트는 10 개 클래스 중 8 개만 보유하며, Global accuracy 는 전체 테스트셋으로, Local accuracy 는 로컬 데이터셋의 20%로 측정하였다.

그림 3 은 손실 방식 및 weight 변화에 따른 정확도를 시각화한 결과이며, 그림 4 는 각 방식에서의 최적 weight 기준 Global/Local 정확도를 비교한 것이다. 결과적으로, MSE 는 Global accuracy 에서, L2 는 Local accuracy 에서 가장 우수한 성능을 나타냈다.

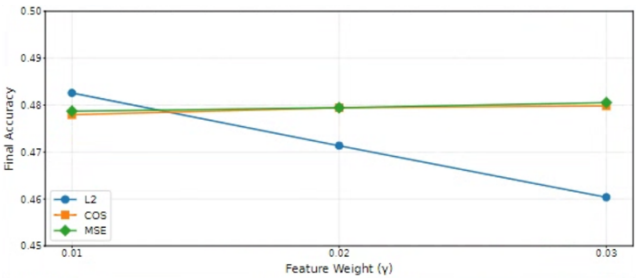

그림 3. Accuracy by Feature-map Distillation Loss Type and Weight

Method	Weight (γ)	Global Acc (%)	Local Acc (%)
L2	0.01	40.95	81.95
COS	0.03	41.21	81.73
MSE	0.03	42.32	81.84

그림 4. Global and Local Accuracy Comparison Table

Ⅲ. 결론

본 연구는 DFL 환경에서 feature-map distillation 을 활용한 Def-FD 프레임워크를 제안하고, 다양한 손실 함수 유형과 가중치 설정에 따른 성능을 분석하였다. 실험 결과, distillation 기법과 해당 loss 의 가중치 조합에 따라 모델 성능이 달라지며, 적절한 설정을 통해 정확도 향상이 가능함을 확인하였다. 제안한 방식은 QKD 기반 보안 채널에서도 적용 가능하여, 보안이 요구되는 엣지 AI 환경에 효과적으로 활용될 것으로 기대된다.

ACKNOWLEDGMENT

본 연구는 한국과학기술정보연구원(KISTI)의

위탁연구개발과제로 수행한 것입니다.

(과제번호 K25L5M2C2/P25030)

참 고 문 헌

[1] 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.