엔트로피를 활용한 분할 연합 학습에서의 효율적인 오프라인 추론 활성화에 관한 연구

김선민, 이주형

가천대학교
kimsm0566@gachon.ac.kr, j17.lee@gachon.ac.kr

A study on enabling efficient offline inference in entropy based splifet federated learning

Sunmin Kim, Joohyung Lee
School of AI Software at Gachon University

요 약

기존 Split Learning 은 추론 시 서버 의존성으로 인한 통신 비효율성 문제가 존재한다. 이를 해결하고자, 본 연구는 SL 을 기반으로 클라이언트 오프라인 추론을 가능하게한 Federated Split learning framework via Mutual Knowledge Distillation (FSMKD)[1]에 엔트로피 기반 적응형 추론을 통합해 예측 엔트로피를 기준으로 오프라인 또는 온라인 추론을 선택적으로 수행하는 Federated Split learning framework via Mutual Knowledge Distillation with Generalization Personalization(FSMKD_GP)를 제안한다.

Ⅰ. 서론

엣지 Artificial Intelligence (AI)환경은 분산 장치 간의 협력으로 지능형 서비스를 제공하지만, 데이터 유출 및 통신 보안 위협이 심각하다. 특히 양자 컴퓨팅 위협에 대비해 최고수준의 보안이 요구되는 분야에서는 양자 암호 통신 네트워크(Quantum Cryptography Network, QCN) 와 같은 차세대 보안 기술로 통신보안을 강화하는 것이 필수적이다.

이러한 보안 통신 인프라 위에서, 각각의 클라이언트가 로컬 데이터는 유지하면서 모델학습을 수행하는 Federated Learning(FL)과 Split Learning (SL)은 데이터 프라이버시를 보호한다. 이들은 QCN 등으로 보호된 채널에서 모델 관련 정보만을 교환하여 원본 데이터 노출 없이 학습을 가능하게 하지만, 각각 개인화 제약, 서버 의존성 등의 문제를 가지며, FL 과 SL 을 결합한 초기 방식 역시 같은 문제를 가지고 있다.

이러한 배경에서, FL 과 SL 을 결합한 FSMKD (Federated Split learning framework via Mutual Knowledge Distillation)[1]가 제안되어 학습 성능을 개선하였으나, 모델의 일부분이 서버 측에 존재하는 SL 의 구조로 인해 추론 시에 서버와 연결되어야 하는 온라인 추론 제약이 존재한다. 이는 실시간성이 중요한 엣지 AI 환경에서 통신 비용과 지연 문제를 발생시킨다.

본 연구는 온라인 추론 제약을 극복하고자, FSMKD [1] 프레임워크에 엔트로피 기반 적응형 추론 메커니즘을 통합한 FSMKD_GP 를 제안한다. 이는 SplitGP [2]의 분포 외(Out Of Distribution, OOD) 데이터 처리 아이디어와 불확실성 측정을 활용해, 추론 시 엔트로피를 기준으로 오프라인 추론 또는 온라인 추론을 선택한다. 이를 통해 통신 비용과 지연을 줄이고, 서버 일반화 능력을 선택적으로 활용하여 효율성, 개인화, 일반화의 균형을 맞춘 실용 적인 엣지 AI 프레임워크를 제공한다. 더 나아가, 해당 프레임워크는 QCN 의 양자 키 분배 운용 자동화에서도 높은 보안성을 유지하며, 활용될 수 있다.

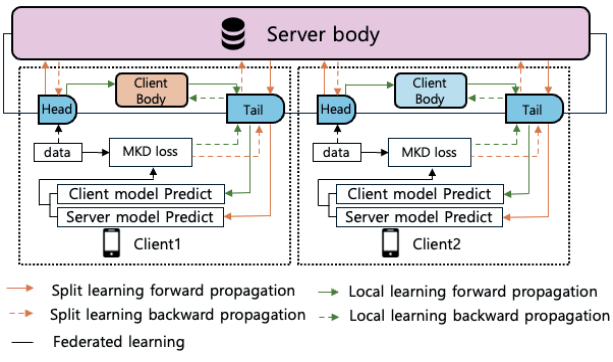

그림 1. FSMKD_GP 의 프레임워크

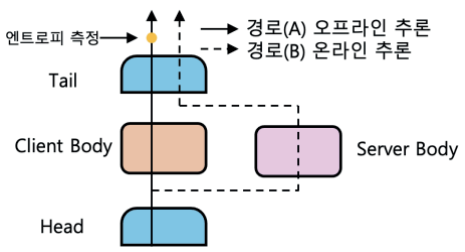

그림 2. FSMKD_GP 추론 경로

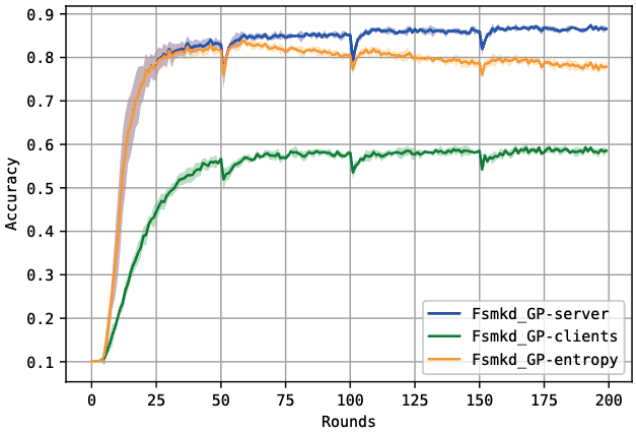

그림 3. 최적임곗값(Eth=0.1)일때의 수렴 그래프

Ⅱ. 본론

가. FSMKD 와 SplitGP

FSMKD[1]는 기존 FL 과 SL 의 결합 방식의 한계를 극복하고자, 상호 지식 증류 (MKD) 를 통해 FL 과 SL 성능을 양방향으로 향상시키는 프레임워크다. 핵심적으로 head:client-body:tail(개인화) 및 head:server-body:tail (공유 글로벌)의 이중-body 구조를 사용하여, 클라이언트별 개인화 모델 생성과 강력한 공유 글로벌 모델 학습을 동시에 지원한다.

SplitGP[2]는 개인화와 일반화 동시 달성 및 효율적추론을 목표로 하는 SL 기법이다. 핵심은 모델을 분할하고 클라이언트와 서버 측에 명확히 다른 역할을 부여하는 것이다. 클라이언트 측은 개인화에, 서버 측은 일반화 (특히 분포 외 OOD 데이터처리)에 집중하도록 학습하여, 효율적인 추론과 함께 일반화와 개인화라는 두 목표를 동시에 달성한다.

나. FSMKD_GP

제안 된 FSMKD_GP 의 학습 과정은 FSMKD[1]와 동일하다([그림 1]). 클라이언트는 head 처리 후 중간값 (h_act)을 서버로 전송하고, 서버는 server-body 처리 후 결과(B_act)를 반환한다. 클라이언트는 B_act 를 tail 에 통과시켜 글로벌 경로로 예측하면서, h_act 를 client-body:tail 에 통과해 로컬 경로 예측도 수행한다. 학습을 위해, 클라이언트는 각 로컬, 글로벌 경로마다 실제 레이블을 이용한 Cross Entropy 손실과 MKD 손실을 결합한 손실 함수를 계산한다. 클라이언트는 로컬 경로 손실로 client-body 를 업데이트하며, 글로벌 경로 손실로 head, tail 을 업데이트한다. 서버도 글로벌 경로 손실을 바탕으로 글로벌 server-body 를 업데이트하고, 주기적으로 FedAvg 를 통해 head/tail 가중치를 취합 및 갱신하여 클라이언트에게 재배포한다.

제안 된 FSMKD_GP 의 핵심 아이디어는 엔트로피 기반의 적응형 추론경로 선택이다. 추론 시, 데이터를 클라이언트 로컬 모델 전체를 통과시켜 예측 엔트로피를 계 산해 불확실성을 측정한다. 계산된 엔트로피가 사전에 정의된 임곗값 미만이면, 추가 통신 없이 해당 예측을 최종 결과로 사용하는 오프라인 추론을 수행한다 ([그림 2]의 경로(A)). 반대로 엔트로피가 임곗값 이상이면 (높은 불확실성, OOD 데이터를 의미), 클라이언트는 서버와 통신하여 FSMKD_GP 의 글로벌 경로를 통해 예측을 수행하는 온라인 추론을 진행한다 ([그림 2]의 경로(B)).

엔트로피 임계값	클라이언트측 모델 정확도	서버측 모델 정확도	엔트로피 활용 모델 정확도	서버측 모델 추론비율	클라이언트측 모델 추론비율	서버측 모델 대비 정확도
0.1	58.53%	86.60%	77.88%	46.14%	53.86%	89.93%
0.2	58.62%	86.15%	74.70%	37.68%	62.31%	86.71%
0.3	58.48%	86.27%	72.60%	31.52%	68.47%	84.15%
0.4	58.70%	86.56%	71.33%	27.62%	72.37%	82.41%
0.5	59.32%	86.81%	70.15%	22.84%	77.15%	80.81%
0.6	58.53%	86.10%	67.64%	18.96%	81.03%	78.56%
0.7	59.21%	86.18%	66.34%	14.48%	85.51%	76.98%
0.8	59.4%	86.65%	64.53%	9.96%	90.03%	74.47%
0.9	59.62%	86.6%	63.47%	7.34%	92.65%	73.29%

표 1. 임곗값별 실험 결과표

다. 실험 결과

제안된 FSMKD_GP 의 실험 결과, 핵심 파라미터인 엔트로피 임곗값(Eth)은 추론 경로 선택과 최종 성능에 중요한 영향을 미쳤다. Eth 를 0.1 부터 0.9 까지 변화시킨 실험에서, Eth=0.1 일 때 최적의 최종 정확도를 달성하였으며, 이는 불확실한 샘플을 효과적으로 서버로 보내 처리한 결과로 해석된다([표 1]). 이 설정에서 53.86%의 오프라인 추론 비율을 유지하며 통신 비용 절감 효과를 보였고, 서버 모델 정확도의 89.93%의 정확도를 달성하였다. 또한, FSMKD_GP 는 안정적인 학습 과정을 통해 높은 정확도에 수렴함을 확인하였다 ([그림 3]).

더 나아가, FSMKD_GP 의 입증된 통신 효율성과 안정적인 학습 성능은 QCN 환경 적용에 대한 긍정적인 가능성을 시사한다. 특히 QCN 내 노드 간 양자 키 분배운용 자동화 등에서, FSMKD_GP 는 QCN 의 높은 보안성을 기반으로 자원 제약적 엣지 노드의 효율적인 지능형 제어에 활용될 잠재력을 가진다.

Ⅲ. 결론

본 연구는 FSMKD[1]의 온라인 추론 비효율성 해결을 위해, 엔트로피 기반 적응형 추론 메커니즘을 통합한 FSMKD_GP 를 제안하였다. FSMKD_GP 는 예측 엔트로피를 기준으로 오프라인/온라인 추론을 선택하여 통신 효율성을 높인다. 실험 결과, 제안된 방식은 최적 임곗값 (Eth=0.1)에서 높은 정확도를 유지하며 통신 비용을 효과적으로 절감하였다. 이는 FSMKD_GP 가 QCN 등 보안 환경 하에서 효율성, 개인화, 일반화 간 실용적인 균형을 제공하는 효과적인 엣지 AI 프레임워크임을 입증하며, 향후 추가적인 성능 개선이 기대된다.

ACKNOWLEDGMENT

본 연구는 한국과학기술정보연구원(KISTI)의 위탁연구개발과제로 수행한 것입니다. (과제번호 K25L5M2C2/P25030)

참 고 문 헌

[1] L. Lyu, H. Yu, C. P. Lam, H. Li, Y. Shen, and Y. Liu, "Federated Split Learning via Mutual Knowledge Distillation," IEEE Transactions on Network Science and Engineering, vol. 9, no. 6, pp. 4385-4398, Nov.-Dec. 2022.

[2] H. Zhang, Y. Venkatesha, J. R. Zhang, H. Zhao, and R. K. Gupta, "SplitGP: Achieving Both Generalization and Personalization in Split Learning," Proc. 40thInternational Conference on Machine Learning (ICML), pp. 40148-40173, Jul. 2023