자원 제약적 IoT 환경을 위한 정보 병목 기반의 시맨틱 분할 연합 학습

김선민, 이주형*
가천대학교

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

A Semantic Split Federated Learning based on Information Bottleneck for Resource-Constrained IoT Environments

Sunmin Kim, Joohyung Lee*

Department of Computing at Gachon University

요 약

기존 분할 연합 학습(Split Federated Learning, SFL)은 학습 과정에서 대용량의 중간 데이터(Smashed Data)를 빈번하게 전송해야 하므로 통신 병목 현상이 발생하며, 무선 채널의 노이즈에 취약하다는 한계가 있다. 이를 해결하고자, 본 연구는 SFL 에 시맨틱 통신을 결합한 Semantic SFL (SSFL) 프레임워크를 제안한다. 제안하는 프레임워크는 Deep JSCC 과 Information Bottleneck(IB) 이론을 적용하여, 학습 태스크에 필수적인 시맨틱 정보만을 추출 및 압축하여 전송한다. 실험 결과, SSFL 은 기존 SFL 대비 통신량을 약 98% 절감하면서도 열악한 통신 환경에서 높은 정확도를 유지함을 확인하였다.

Ⅰ. 서 론

최근 IoT 및 모바일 디바이스의 폭발적인 보급으로 인해 엣지에서 생성되는 데이터의 양이 기하급수적으로 증가하고 있다. 이러한 방대한 데이터를 중앙 서버로 전송하여 학습하는 기존방식은 막대한 네트워크 비용을 유발할 뿐만 아니라, 민감한 개인정보 유출을 초래한다. 이에 대한 대안으로, 데이터를 로컬 디바이스에 유지한 채 모델의 가중치만을 공유하여 모델을 학습하는 연합학습(Federated Learning, FL)이 제안되었다[1]. FL 은 데이터 프라이버시를 보존할 수 있다는 장점이 있지만, 클라이언트가 모델 전체를 로컬에서 학습시켜야 하는 점에서 자원이 제한된 IoT 디바이스에 적용하기에는 한계가 있다[1].

이러한 FL 의 연산 부담을 완화하기 위해, 모델을 클라이언트와 서버가 나누어 학습하는 Split Learning (SL)이 있지만, 여러 클라이언트가 순차적으로 학습해야 하기 때문에 학습시간이 오래걸린다. SFL 은 SL 과 FL 을 결합한 프레임워크로써 클라이언트와 서버가 하나의 모델을 나눠 학습해 엣지 디바이스의 부하를 줄이면서 SL 의 단점인 순차적 학습을 FL 을 통해 병렬학습이 가능하게 하였다[2]. 하지만 SFL 은 라운드당 한 번만 가중치를 교환하는 FL 과 달리, SFL 은 학습 과정의 매 배치마다 대용량의 Smashed Data 와 그라디언트를 왕복으로 전송해야해 통신량이 많아질 수 있다. 특히, SFL 의 Smashed Data 는 태스크와 직접적으로 관련이 없는 배경 정보나 잡음과 같은 잉여 정보를 다수 포함하고 있다[3]. 또 한 학습 시 실제 무선 채널의 노이즈를 고려하지 않기 때문에 실제 통신 환경에서의 노이즈로 인한 데이터 왜곡 시 성능 저하를 겪게 된다.

본 연구는 이러한 문제를 해결하기 위해 시맨틱 통신을 SFL 에 통합한 SSFL 프레임워크를 제안한다. 제안하는 기법은 섀넌의 비트 중심 통신을 넘어, 학습 태스크 수행에 필수적인 시맨틱 정보만을 추출하여 전송하는 것을 목표로 한다. 이를 위해 우리는 SFL 의 Cut Layer 에 Deep JSCC 을 적용하고, Variational IB(VIB) 이론을 손실함수에 도입한다[4]. 이를 통해 Smashed Data 내의 불필요한 정보를 필터링하고 핵심 특징만을 압축하여 전송함으로써, 통신 효율을 극대화하고 채널 노이즈에 대한 강건성을 동시에 확보한다.

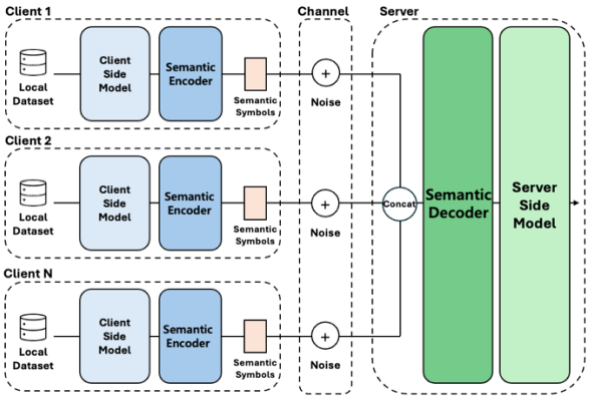

그림 1. SSFL 의 프레임워크

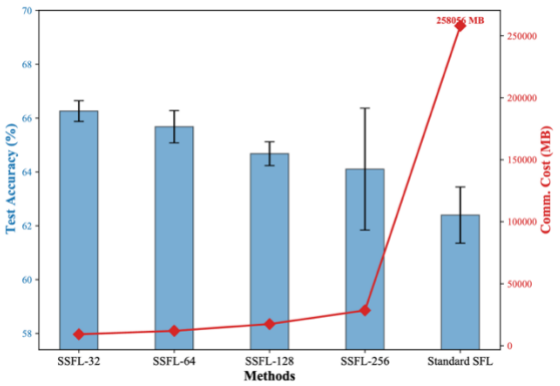

그림 2. 통신 데이터양의 따른 정확도 비교

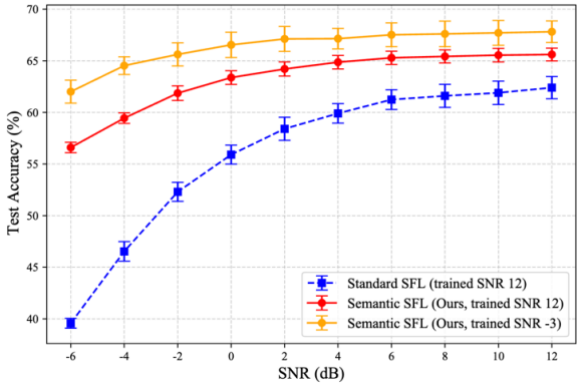

그림 3. 다양한 SNR 환경에서의 성능비교

Ⅱ. 본론

가. SSFL 프레임워크

1) Deep JSCC 기반 시맨틱 전송: 제안하는 SSFL 의 전체 구조는 [그림 1]과 같다. 클라이언트 모델과 서버 모델 사이의 Cut Layer 에 시맨틱 인코더/디코더를 배치한다. 클라이언트는 입력 데이터 x로부터 특징 h를 추출한 후, 시맨틱 인코더를 통해 이를 채널 전송에 적합한 압축된 심볼 z로 변환한다. 기존의 디지털 통신이 소스 코딩과 채널 코딩을 별도로 수행하는 것과 달리, 제안하는 Deep-JSCC 구조는 이 과정을 하나의 신경망으로 통합하여 학습한다.

2) 정보 병목(IB) 기반 최적화: 통신 효율성과 모델의 일반화 성능을 동시에 확보하기 위해, 본 연구는 IB 이론을 적용한다. IB 의 목표는 입력 x에 대한 정보량은 최소화하고, 타겟 y에 대한 정보량은 최대화하는 최적의 표현 z를 학습하는 것이다. 하지만 상호정보량을 직접 계산하는 것은 불가능하므로, 우리는 VIB 이론을 활용하여 이를 근사한다. 인코더는 결정된 값이 아닌 확률 분포(평균 μ, 분산 σ)를 출력하며, 재매개변수화 트릭을 통해 잠재 변수 z를 샘플링한다. 이후 서버는 수신된 z를 기반으로 예측값 $\hat{y}$를 산출하며, 최종적으로 SSFL 은 다음과 같은 손실 함수를 통해 학습된다.

$$\mathcal{L}_{Total} = \mathcal{L}_{CE}(y, \hat{y}) + \beta \cdot D_{KL}(p(z|x)||r(z)$$

여기서 $\mathcal{L}_{CE}$는 서버에서 계산되는 태스크 손실로, y와 $\hat{y}$ 간의 오차를 줄여 모델의 예측 정확도를 높이고 D_{KL}은 클라이언트에서 계산되는 압축 손실로, 인코딩된 분포 $p(z|x)$와 정규분포 $r(z)$ 간의 차이를 줄여 정보량을 제한한다. β 는 압축률과 정확도의 균형을 조절하는 하이퍼파라미터 이다. 클라이언트는 로컬에서 계산된 압축 손실의 그라디언트와 서버로부터 전송받은 태스크 손실의 그라디언트를 결합하여 모델을 업데이트 한다. 따라서 인코더는 태스크와 무관한 배경이나 잉여 정보를 필터링하고 핵심적인 시맨틱 정보만을 추출하도록 학습되며, 이는 결과적으로 통신 데이터양을 줄이도록 유도한다.

나. 실험 결과

본 연구의 성능을 검증하기 위해 CIFAR-10 데이터셋과 ResNet-18 모델을 사용하여 실험을 수행하였다. 각 클라이언트는 데이터의 70%를 하나의 클래스로, 나머지 30%를 무작위 클래스로 구성하였다. 채널은 AWGN 을 적용하여 무선 환경을 구성하였다. 제안하는 방법은 $SSFL-d$로 표기하며, 여기서 d는 기존의 4096 차원을 d차원으로 압축한 심볼의 차원을 의미한다.

1) 통신 효율성 및 정확도: [그림 2] 는 통신 데이터양에 따른 정확도를 보여준다. 기존 SFL 은 약 258,000 MB 의 데이터를 전송해야 하지만, 제안하는 SSFL-64 는 약 4,000 MB 라는 기존 SFL 대비 약 98% 의 통신량 절감효과를 보이면서 더 높은 성능을 달성했다. 이는 IB 기반의 압축이 불필요한 정보를 효과적으로 제거했음을 의미한다. 특히, 압축률이 가장 높은 SSFL-32 가 가장 우수한 성능을 보였는데, 이는 IB 기반의 압축이 과적합을 방지하는 정규화 효과를 가져왔기 때문이다.

2) 노이즈 강건성: [그림 3]은 다양한 SNR 에서의 성능비교이다. 기존 SFL 은 SNR 이 낮아질수록 성능이 급격히 저하되는 반면, SSFL 은 모든 구간에서 기존 SFL 보다 안정적인 성능을 유지하였다. 이는 Deep JSCC 가 채널 노이즈에 대해 강건한 특징을 학습했음을 시사한다.

Ⅲ. 결론

본 연구는 IoT 환경 내 SFL 의 통신 병목과 노이즈 민감성 해결을 위해, IB 이론과 Deep JSCC 를 결합한 SSFL 프레임워크를 제안하였다. SSFL 은 태스크에 필수적인 시맨틱 정보만을 전송하여 통신효율을 극대화한다. 실험결과, 통신량을 기존 대비 98% 이상 절감하였으며 낮은 SNR 환경에서도 우수한 성능을 유지한다. 이는 SSFL 이 효율성과 강건성을 동시에 만족시킴을 시사한다.

ACKNOWLEDGMENT

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

참 고 문 헌

[1] W. Y. B. Lim et al., "Federated learning in mobile edge networks: A comprehensive survey," IEEE Communications Surveys & Tutorials, vol. 22, no. 3, pp. 2031-2063, 2020.

[2] C. Thapa et al., "SplitFed: When federated learning meets split learning," AAAI Conference on Artificial Intelligence, vol. 36, no. 8, pp. 8485-8493, Jun. 2022.

[3] Y. Matsubara et al., "Supervised compression for resource-constrained edge computing systems," IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 2685-2695, 2022.

[4] J. Shao, Y. Mao, and J. Zhang, "Task-oriented communication for multi-device cooperative edge inference," IEEE Transactions on Wireless Communications, vol. 22, no. 11, pp. 7374-7389, Nov. 2023.