엣지 디바이스 환경에서의 경량 연합학습 시스템 설계

김주령§, 정택준§, 임동건§, 한하영§, 방인규■†, 김태훈■§

§국립한밭대학교 컴퓨터공학과, †국립한밭대학교 지능미디어공학과

{20222021, 20211929, 20211893, 20222517}@edu.hanbat.ac.kr, {ikbang, thkim}@hanbat.ac.kr

Design of a Lightweight Federated Learning System for Edge Device Environments

Juryeong Kim§, Taekjun Jeong§, Donggeon Im§, Inkyu Bang■†, Taehoon Kim■§

§Department of Computer Engineering, Hanbat National University

†Department of Intelligence Media Engineering, Hanbat National University

요 약

저궤도 위성 환경과 같이 통신 자원이 제한되고 연산 성능이 제약되는 환경에서, 데이터 프라이버시와 통신 효율을 동시에 확보하기 위한 대안으로 연합학습의 중요성이 부각되고 있다. 본 연구에서는 Jetson Nano 두 대를 클라이언트로 활용하여, FP16 양자화와 Top-K 희소화를 결합한 경량 연합학습 시스템을 제안한다. 실험을 통한 성능 평가 결과, 제한된 연산 자원과 대역폭 환경에서도 학습이 안정적으로 수행됨을 확인할 수 있었고, 제안한 알고리즘은 기존 FedAvg 대비 약 90~95%의 통신량을 절감하면서도 유사한 수준의 전역 모델 정확도를 유지하였다. 이를 통해 제안한 시스템이 저사양 엣지 및 위성 통신 환경에서도 효율적이고 안정적인 분산 학습이 가능함을 입증하였다.

Ⅰ 서론

저궤도 위성(Low Earth Orbit; LEO) 네트워크의 확산은 전 세계 어디서나 초저지연(ultra-low latency) 통신이 가능한 지능형 엣지 인프라 환경의 기반을 마련하고 있다. 이러한 환경에서는 수많은 IoT 및 엣지 디바이스가 위성 네트워크를 통해 연결되어 데이터를 로컬에서 처리·학습함으로써, 중앙 서버 의존도를 낮추는 분산 지능 구조가 요구된다 [1]. 그러나 LEO 위성 기반 통신은 위성의 빠른 궤도 이동과 제한된 대역폭으로 인해 간헐적 연결, 세션 단절, 통신 지연이 빈번하게 발생하며, 이는 모델 학습의 안정성과 효율성을 저하시킨다 [2].

이러한 한계를 극복하기 위한 대안으로 연합학습(Federated Learning; FL) 기술이 주목받고 있다. 연합학습은 각 참여 노드가 로컬 데이터를 유지한 채 모델 파라미터만 교환함으로써, 데이터 프라이버시를 보호하면서도 분산 학습을 수행할 수 있는 구조를 제공한다. 특히, 데이터 전송이 어려운 LEO 네트워크 환경에서는 연합학습을 통해 중앙 서버로의 데이터 이동을 최소화함으로써 통신 효율을 향상시킬 수 있다 [3].

그러나 실제 위성 기반 또는 저사양 엣지 환경에서의 연합학습 구현은 제한된 연산 자원, 불안정한 네트워크, 발열 및 메모리 제약 등의 물리적 제약으로 인해 여전히 어려운 과제이다. 이에 본 논문에서는 연산 자원의 제약된 상황을 모사하기 위해 Jetson Nano 두 대를 클라이언트로, macOS 환경을 중앙 서버로 구성한 경량 연합학습 시스템을 설계 및 구현하고, FP16 양자화와 Top-K 희소화를 결합한 경량 연합학습 알고리즘을 제안하여 성능의 우수성을 검증하고자 한다.

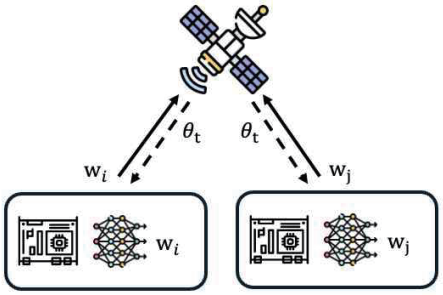

그림 2. LEO 위성 기반 연합학습 시스템

Ⅱ 제안 알고리즘

본 논문에서는 저사양 엣지 환경에서 통신·메모리 제약을 완화하면서도 FedAvg에 준하는 수렴 특성을 유지하기 위한 경량 연합학습 알고리즘을 제안한다. 제안하는 알고리즘은 로컬 모델 파라미터(w_i), 전역 모델 파라미터(θ_t), 로컬 학습 후 파라미터 변화량($\Delta w_i^{(t)}$)에 대해, K%의 상위 중요도 파라미터만 선택(Top-K)하고 이를 FP16 정밀도로 양자화(Quantization)하여 통신량을 줄인다.

$$\widetilde{S}_i^{(t)} = \mathrm{Q16}(\mathrm{TopK}(\Delta w_i^{(t)}, K)) \quad (1)$$

여기서, $\widetilde{S}_i^{(t)}$는 t번째 라운드에서 전역 서버로 보낼 경량화된 로컬 파라미터이다. 서버는 수신된 파라미터를 FP32로 복원한 후, FedAvg 기반의 가중 평균 방식으로 전역 모델을 갱신한다.

■ Corresponding Authors: Inkyu Bang (ikbang@hanbat.ac.kr), Taehoon Kim (thkim@hanbat.ac.kr)

$$\theta_{t+1} = \theta_t \cdot \frac{\sum_{i=1}^{m} n_i \cdot \mathrm{FP32}(\widetilde{S}_i^{(t)})}{\sum_{i=1}^{m} n_i} \quad (2)$$

여기서, θ_t는 라운드 t에서의 전역 모델 파라미터이며 n_i는 클라이언트 i의 데이터 샘플 수, m은 참여 클라이언트 수이다.

Ⅲ. 시스템 구현 및 성능 평가

본 장에서는 제안한 알고리즘의 학습 정확도와 통신 효율성을 평가하기 위한 실험 환경을 구성하였다. 중앙 서버는 Flower 서버를 실행하여, 두 개의 클라이언트들이 학습 세션에 참여할 수 있도록 gRPC 통신 채널을 개방한다.

서버는 초기 모델 파라미터를 각 클라이언트로 전송하고, 클라이언트는 이를 기반으로 자신에게 불균등하게 할당된 MNIST 데이터 샤드(Shard)로 로컬 분류 학습을 수행한다. 학습이 완료되면 클라이언트는 업데이트된 weight와 로컬 메트릭(loss, accuracy)을 서버로 전송하며, 서버는 이를 FedAvg 방식으로 통합하여 전역 모델을 갱신한다. 라운드 종료 시 서버는 새로운 전역 weight를 클라이언트로 배포하며, 지정된 라운드 수만큼 이 과정을 반복한다. 하드웨어 구축에 사용된 서버와 클라이언트에 대한 상세 사양은 표 1에 요약되어 있다.

표 1 실험 환경 구성 요약

구분	항목	내용
Hardware	Server	8-core CPU, 16GB RAM
	Client	NVIDIA Jetson Nano (2GB) × 2
Software	Server OS	macOS Sequoia 15.6.1
	Client OS	Ubuntu 18.04 LTS
	Framework	Flower 0.18.0
	Deep Learning Library	PyTorch 1.11
	Dataset	MNIST
Training	Strategy	FedAvg
	Communication	gRPC (insecure mode)

그림 2는 제안한 알고리즘의 파라미터 변화량(Δw_i)을 기반으로, 상위 중요 파라미터만을 사용하는 Top-K 비율(K%)에 따른 라운드별 전역 모델 검증 정확도를 나타낸다. Top-K 비율이 증가할수록 FedAvg와 유사한 수준의 정확도를 유지하였으며, 특히 Top-K=20 이상에서는 수렴 속도와 최종 정확도 모두에서 FedAvg와 거의 동일한 성능을 보이는 것을 확인할 수 있었다.

그림 3은 라운드별 누적 통신량 변화를 비교한 결과이다. Top-K=10의 경우 FedAvg 대비 약 95% 이상의 통신량 감소를 달성하였으며, Top-K=20에서도 약 90% 수준의 효율 향상을 보였다. 이러한 결과는 제안한 알고리즘이 통신 자원이 제한된 환경에서도 전송 효율을 극대화하면서, 학습 안정성과 수렴 특성을 유지할 수 있음을 보여준다.

Ⅳ. 결론

저궤도 위성 환경과 같이 통신 자원이 제한되고 연산 성능이 낮은 엣지 환경에서도 안정적인 학습이 가능하도록, FP16 양자화와 Top-K 희소화를 결합한 경량 연합학습 알고리즘을 제안하였다. 실험 결과, 기존 FedAvg 대비 약 90~95%의 통신량을 절감하면서도 유사한 수준의 전역 모델 정확도를 유지함을 확인하였다. 향후 연구에서는 본 알고리즘을 실제 LEO 위성 네트워크 및 다중 엣지 클러스터 환경으로 확장하여, 동적 연결성과 지연을 고려한 적응형 통신 스케줄링 및 모델 동기화 기법을 연구할 예정이다.

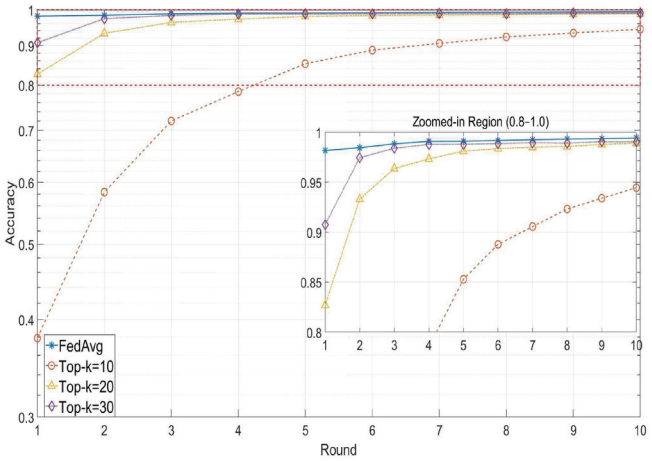

그림 3 Top-K 비율에 따른 라운드별 전역 모델 정확도 비교

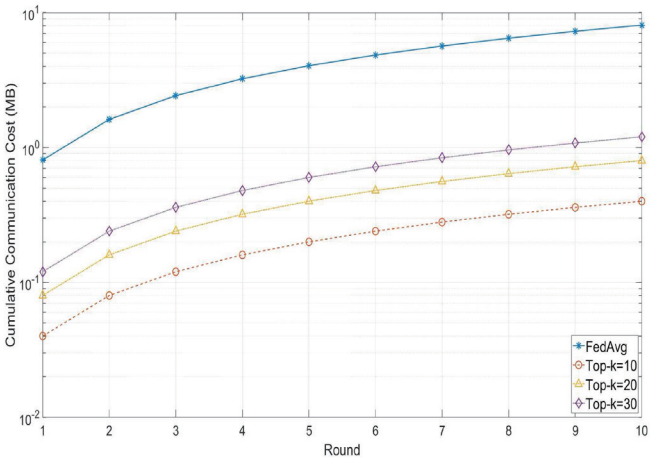

그림 3. Top-K 비율에 따른 누적 통신량 비교

ACKNOWLEDGMENT

본 연구는 2025년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과 (2022-0-01068) 및 2025년도 교육부 및 대전광역시의 재원으로 대전RISE센터의 지원을 받아 수행된 지역혁신중심 대학 지원체계(RISE)의 결과임 (2025-RISE-06-002)

참 고 문 헌

[1] J. Wen, J. Zhang, J. Wu, and C. Wu, “A survey on federated learning: challenges and applications,” *International Journal of Machine Learning and Cybernetics*, vol. 13, pp. 2523 - 2545, 2022.

[2] A. Kairouz *et al.*, “Advances and open problems in federated learning,” *Foundations and Trends in Machine Learning*, vol. 14, no. 1 - 2, pp. 1 - 210, 2021.

[3] E. Dritsas, M. Trigka, “Federated Learning for IoT: A Survey of Techniques, Challenges, and Applications,” *Journal of Sensor and Actuator Networks*, vol. 14, no. 9, 2025.