Non-IID 데이터 기반 연합학습에서 채널별 대칭 양자화에 따른 통신 효율성 및 정확도 분석

이동욱, 이웅희*

한성대학교

{1971455, whlee}@hansung.ac.kr

Communication Efficiency and Accuracy Analysis of Federated Learning on Non-IID Data with Channel-Wise Symmetric Quantization

Dongwook Lee, Woonghee Lee*

Hansung University

요 약

본 논문은 non-IID 데이터 기반 연합학습 환경에서 클라이언트들이 학습 후 모델을 업로드 시 원본 모델과 채널별 양자화 모델 간에 발생하는 통신량 차이와 글로벌 정확도 변화 양상을 실험적으로 분석한다. 이를 통해 채널별 양자화가 non-IID 데이터 기반 연합학습 환경에서 전송 비트를 크게 줄이면서도 원본 모델 대비 동일 수준 이상의 정확도가 도출됨을 확인한다. 이러한 분석을 바탕으로, 본 논문은 통신 제약 환경에서의 연합학습을 위한 모델 전송 정책 수립에 필요한 실증적 정보를 제공한다.

Ⅰ. 서 론

최근 연합학습은 개인정보를 중앙에 수집하지 않고도 모델을 학습할 수 있다는 장점 때문에 엣지·모바일 환경에서 널리 활용되고 있다. 그러나 라운드마다 클라이언트가 서버로 모델을 전송해야 하므로, 통신 자원이 제한된 환경에서는 전송 비용이 학습 성능을 좌우할 수 있다. 본 논문에서는 경량 CNN 모델인 SqueezeNet[1]을 사용하고 연합학습 방식으로 널리 쓰이는 FedAVG[2]를 적용한 환경에서 모델 업로드 시 원본 모델 대비 채널별 양자화[3] 적용에 따른 통신량 절감과 전역 정확도 변화를 실험적으로 분석한다.

Ⅱ. 본론

본 논문에서는 채널별 대칭 양자화 기법을 사용하였다. 본 기법은 각 채널의 가중치 분포를 0을 중심으로 한 균일 정수 격자에 사상하는 방식이다. 채널마다 하나의 스케일 s_c를 두어 실수 가중치 $w_{c,i}$를 정수 범위 $[q_{\min}, q_{\max}]$의 최근접 격자점으로 표시한 뒤, s_c로 복원을 수행한다. 구제척 정의는 식(1)-(3)에 제시한다.

식 (1)에서 채널 c의 스케일 s_c는 해당 채널의 실수 가중치 $w_{c,i}$ 중 최대 절댓값을 양자화 범위의 최댓값 $q_{\max}$로 나누어 정의한다.

$$s_c = \frac{\max_i |w_{c,i}|}{q_{\max}} \quad \text{.....(1)}$$

식 (1)에서 도출된 채널별 스케일 s_c를 바탕으로 식 (2)에서는 채널별 가중치 $w_{c,i}$를 채널 c의 스케일 s_c로 나누어 반올림 후 $[q_{\min}, q_{\max}]$로 클리핑하여 정수 $q_{c,i}$로 변환하며, 이를 양자화 가중치로 사용한다.

$$q_{c,i} = \mathrm{clip}\left(\mathrm{round}\left(\frac{w_{c,i}}{s_c}\right), q_{\min}, q_{\max}\right) \quad \text{.....(2)}$$

식 (2)에서 도출된 양자화 가중치를 바탕으로 식 (3)은 채널 스케일 s_c를 곱해 양자화 가중치를 실수 가중치로 복원한다.

$$\hat{w} = s_c \cdot q_{c,i} \quad \text{.....(3)}$$

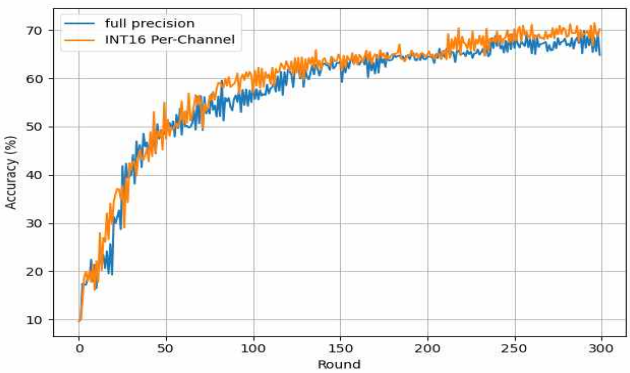

그림 1. 업로드 시 원본 모델과 양자화 모델간의 정확도 비교

그림 1은 CIFAR-10 데이터셋에 대해 Drichlet 분포(α=0.5)로 non-IID 환경을 구성하고, 클라이언트 기기 10개, 로컬 학습 5회, 글로벌 라운드 300, INT16을 양자화 비트로 설정하여 수행한 실험 결과를 나타낸다. 실험 결과, 원본 모델 대비 모든 라운드 구간에서 유사하거나 소폭 상승한 정확도를 보였고, 특히 초기 라운드에서 수렴 속도가 더 빠른 경향을 보였다. 이는 양자화가 정규화 효과를 제공하여 각 클라이언트의 로컬 데이터셋에 대한 과적합을 억제하면서 전역 일반화 성능을 개선할 수 있음을 시사한다.

Model	Traffic (MB)
SqueezeNet(FP32)	28.41
SqueezeNet(INT16)	14.66

표 1. 모델 업로드 시 통신량 크기 비교

표 1은 매 라운드마다 업로드 되는 통신량을 비교한 결과이다. 원본 모델은 28.41MB, INT16 양자화 모델은 14.66MB로 약 48.4%의 전송량 감소가 확인되었으며, 이는 통신 제약적인 환경에서도 효율적인 연합학습 운용이 가능함을 시사한다.

Ⅲ. 결론

본 논문에서는 non-IID 데이터 기반 연합학습 환경에서 모델 업로드 시 원본 모델 대비 채널별 INT16 양자화의 효과를 실험적으로 분석하였으며, 실험 결과 약 48%의 전송량 감소와 소폭 향상된 전역 정확도 및 초기 수렴 가속을 보여주었다. 이는 통신 제약 환경에서도 효율적인 연합학습 운용과 모델 전송 정책 수립이 가능함을 시사한다. 후속 연구에서는 클라이언트별 통신 상태를 인지하여 클라이언트별 모델 업로드 시 양자화를 동적으로 조정하는 적응형 양자화 연합학습 기법을 고안할 계획이다.

ACKNOWLEDGMENT

이 논문은 2025년 정부(방위사업청)의 재원으로 국방기술진흥연구소의 지원을 받아 수행된 연구임(KRIT-CT-23-041).

참 고 문 헌

[1] Iandola, Forrest N. “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and<0.5MB model size.” arXiv preprint arXiv:1602.07360 (2016).

[2] McMahan, Brendan, et al. “Communication-efficient learning of depp networks from decentralized data.” Artificial intelligence and statistics. PMLR, 2017.

[3] Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. arXiv preprint arXiv:1806.08342, 2018.