모델 경량화를 통한 통신 및 연산 효율적 연합학습 기법에 관한 연구 동향

송윤재, 홍준표
홍익대학교

yjsong1003@mail.hongik.ac.kr, jp_hong@hongik.ac.kr

A Survey on Communication- and Computation-Efficient Federated Learning via Model Compression

Yoon Jae Song, Jun-Pyo Hong
Hongik Univ.

요 약

본 논문은 연합학습(FL: Federated learning)의 주요 도전 과제인 통신 및 연산 자원 효율성 개선을 위한 효과적인 방안으로 모델 경량화 기법에 대해 살펴본다. 즉, 연합학습 중 학습 및 자원 상황에 따라 적응적으로 일부 모델 파라미터를 제거 혹은 비활성화해 모델을 경량화함으로써, 큰 성능 저하 없이 지역학습 및 지역모델 전송에 필요한 연산·통신 자원을 크게 낮출 수 있다. 모델 경량화 기반의 기존 연합학습 기법들의 주요 특징과 장단점을 소개하고, 향후 연구방향으로 빈도주의적(frequentist) 접근이 갖는 한계를 보완할 수 있는 베이지안(Bayesian) 접근 기반의 불확실성 정량화와 이를 활용한 모델 경량화 방법을 제안한다.

Ⅰ. 서 론

연합학습(FL: Federated learning)은 다수의 클라이언트들이 각자 본인이 소유한 지역 데이터를 기반으로 분산 학습을 수행하고, 지역 학습된 모델을 전송해 서버에서 이들을 합성함으로써 데이터의 직접적인 전송 없이 학습을 수행할 수 있는 기법이다 [1]. 전역 모델의 수렴을 위해서는 이와 같은 지역 학습과 전송을 여러 차례 반복해야 하고, 이는 상당한 연산 및 통신 부하를 발생시켜 연합학습의 주요 병목으로 고려되고 있다. 이와 같은 연학학습의 병목현상을 완화하기 위한 효과적인 방안으로 모델의 차원을 줄이는 모델 경량화 연구가 활발히 진행되고 있다. 본 논문에서는 Lottery Ticket 가설을 바탕으로 pruning, dropout 을 설명하고 FL 에서의 적용방법에 대해서 살펴본다. 나아가 기존 대부분의 FL 기법이 채택하고 있는 빈도주의적(frequentist) 접근법들의 한계점 극복을 위한 베이지안(Bayesian) 접근 기반의 모델 경량화를 설명하고 이를 FL 환경에 적용하는 후속 연구 방향을 제안한다.

Ⅱ. 본 론

Lottery Ticket 가설과 관련 실험들을 통해 거대한 신경망 안에 실제로는 훨씬 작은 규모의 효율적인 서브넷(subnetwork)가 존재하며, 이 부분만으로도 원래 전체 모델과 거의 동일한 성능을 낼 수 있음이 잘 알려져 있다 [2]. 해당 가설을 기반으로 한 모델 경량화를 통해 FL 환경에서 연산 및 통신 효율성을 향상을 동시에 기대할 수 있다 [3]. 본 장에서는 (i) 로컬에서 서브넷을 도출하는 기준과 절차 (ii) 클라이언트별로 상이한 서브넷을 글로벌 모델로 결합하는 방법을 기준으로 다양한 경량화 기법들을 설명한다.

2.1 Adaptive Pruning

Pruning 은 신경망 일부 파라미터를 0 으로 만들어 모델을 경량화하는 방법이다. FL 에서는 로컬 학습 과정에서 중요도 지표(누적 squared gradient 등)를 기반으로 파라미터를 제거하여 클라이언트별 경량화된 신경망을 형성하고, 재구성 라운드에서 서버가 여러 클라이언트의 중요도 지표를 바탕으로 목표 밀도/마스크를 주기적으로 재설계함으로써 학습 진행상황에 따라 적응적으로 모델이 변화한다(PruneFL) [4]. 서버에서의 지역 모델 결합은 (i) 표준 가중 평균(FedAvg) 또는 (ii) 겹치는 파라미터만 집계(overlapped-only aggregation) 같은 규칙을 사용한다. 전자는 단순하고 보편적이며, 후자는 개인화 정보 보존에 유리하다. 한편 pruned global model 이 항상 모든 클라이언트에 대해 winning ticket 임을 보장하지는 않으나, FL 환경에서 초기값으로 되돌려 재학습해도 pruning 되지 않은 원래의 모델로 학습했을 때와 비슷한 성능을 보임을 통해 글로벌 모델 기준 winning ticket 성질을 확인할 수 있다. 하지만 PruneFL 특성상 재구성 라운드에서 full-model 과 중요도 지표를 전송함에 따라 통신량이 증가하고, 학습과정에서 일반적인 FL 에 비해 더 많은 하이퍼파라미터에 의존하는 한계가 존재한다.

2.2 Dropout

Dropout 은 원래 과적합 방지 기법이지만, FL 에서는 통신·연산 비용을 줄이기 위한 방안으로 활용될 수 있다 [5]. 로컬 서브넷 도출은 서버가 전역 모델에서 클라이언트별 제약(연산/대역/지연)에 맞춘 크기의 서브넷을 무작위로 생성·배정하거나, 클라이언트가 자신의 dropout 비율에 따라 부분 모델만 활성화해 학습하는 방식으로

이뤄진다. 글로벌 결합은 각 클라이언트가 서브넷 파라미터만 업로드하면, 서버가 비활성 파라미터는 직전 전역값으로 복원한 뒤 평균 집계로 전역 모델을 갱신한다. 이 방식은 pruning 대비 구현이 단순하고 클라이언트별 통신·연산 부담을 크게 낮출 수 있는 장점이 있으나, 서브넷 매칭·복원 규칙에 따라 성능이 크게 변화하는 단점을 갖는다.

2.3 Bayesian Neural Network

기존 대부분의 FL 기법들은 빈도주의 접근에 기반하고 있어 모델 및 추론 결과에 대한 신뢰도 혹은 불확실도를 정량화하지 못한다. 즉 학습된 모델에서 어떤 파라미터가 더 중요한 역할을 하는지, 추론 결과는 얼마나 신뢰할 만한지 확인하기 어렵다. 특히, 이와 같은 특징은 데이터가 적거나 Non-IID 한 환경에서 학습 성능을 크게 저하시키는 문제를 낳을 수 있다 [6]. 따라서, 모델 파라미터 및 추론 결과의 불확실도를 정량화하고, 이를 학습 및 전송에 활용해 적응적인 FL 을 가능하게 하는 베이지안 FL 이 최근 연구되고 있다 [7]. 로컬 서브넷/분포 획득은 각 클라이언트가 로컬 데이터로 posterior 분포(Gaussian Approximation; 평균·분산)를 학습하거나, 분산(불확실성)을 고려한 Bayesian pruning 으로 신뢰도 기반 희소 마스크를 설계할 수 있다. Bayesian pruning 에서 파라미터의 신뢰도 정량화를 위한 지표로는 대표적으로 SNR, SPR, BMR 등이 있다. 글로벌 결합은 기존 FL 의 점 추정 평균과 달리, 클라이언트에서 학습한 local posterior 간의 곱으로 global posterior 를 업데이트를 수행함으로써 신뢰도가 고려된 보다 안정적인 모델 합계를 가능하게 한다. 이는 불확실성 표현과 데이터가 적은 클라이언트에 대한 정규화 효과 측면의 이점을 줄 수 있으며, 파라미터(평균·분산)의 효율적 전송을 통해 통신량을 줄일 수 있다.

III. 결 론

본 논문에서는 연산 및 통신 효율성 향상을 위한 모델 경량화 기반의 연합학습 기법들에 대해 살펴보았다. 이들 대부분의 연구들은 frequentist 접근에 기반하고 있어 학습 데이터가 적고, 분포가 상이한 환경에서 학습성능이 저하되는 단점을 갖는다. 이와 같은 단점 극복을 위해 최근 Bayesian FL 이 연구되어 있으나 아직 초기 단계에 머물러 있는 상황이다. 이에 (i) 경량화된 BNN 의 FL global aggregation 규칙 (ii) BNN 경량화 기법 (iii) Bayesian 경량화, FL 집계의 공동 설계를 통해 정확도–효율–불확실성 트레이드오프의 정량 검증에 대한 후속연구를 수행할 계획이다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (RS-2025-16065671)

참 고 문 헌

[1] H. Brendan McMahan et al., “Communication-Efficient Learning of Deep Networks from Decentralized Data,” AISTATS, 2017.

[2] J. Frankle and M. Carbin, “The lottery ticket hypothesis: Finding sparse, trainable neural networks,” in *Proc. 7th Int. Conf. Learn. Represent. (ICLR)* New Orleans, LA, USA, May 2019, pp. 6–9.

[3] S. Seo, S.-W. Ko, J. Park, S.-L. Kim, and M. Bennis, “Communication-Efficient and Personalized Federated Lottery Ticket Learning,” in *Proc. IEEE 22nd Int. Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2021, pp. 581–585

[4] Y. Jiang, S. Wang, V. Valls, B. J. Ko, W.-H. Lee, K. K. Leung, and L. Tassiulas, “Model Pruning Enables Efficient Federated Learning on Edge Devices,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 10374–10389, Dec. 2023

[5] D. Wen, K.-J. Jeon, and K. Huang, “Federated Dropout—A Simple Approach for Enabling Federated Learning on Resource Constrained Devices,” *IEEE Wireless Communications Letters*, vol. 11, no. 5, pp. 923–927, May 2022

[6] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra,“Weight Uncertainty in Neural Networks,” in *Proc. 32nd International Conference on Machine Learning (ICML)*, Lille, France, Jul. 2015, pp. 1613–1622.

[7] J.-P. Hong, H. Seo, and K. Lee, “Distribution-level AirComp for wireless federated learning under data scarcity and heterogeneity,” arXiv:2506.06090, 2025.