

연합학습 기반 동적 LoRA 랭크 조절 연구 및 양자 암호 네트워크 응용 논의

최현준, 이주형
가천대학교hjc405@naver.com, j17.lee@gachon.ac.kr

A Study of Federated Dynamic LoRA Rank Adaptation and Its Potential in QKD Networks

Choi hyun jun, Joohyung Lee
Gachon Univ.

요 약

본 연구에서는 이질적 연산·메모리 제약을 지닌 엣지 클라이언트 환경에서 대형 사전학습 모델을 효율적으로 파인튜닝하기 위한 Federated MagLoRA(Max-flops and GPU Low-rank Adaption)를 제안한다. 각 클라이언트는 클라이언트가 초당 처리 가능한 연산량과 사용할 수 있는 메모리 크기를 바탕으로 LoRA(Low-rank Adaption) 어댑터의 목표 랭크를 동적으로 산정·프루닝하고, 로컬 업데이트된 파라미터를 FedAvg로 통합해 이기종 분산 환경에서 중앙 집중형 학습과 비슷한 정확도를 유지한다. GLUE-SST2 실험에서 MagLoRA는 고정 랭크 대비 FLOPs/sample을 최대 30% 절감하고 파라미터 수를 85% 줄이면서도 정확도 손실을 0.6%p 이내로 억제했으며, QKD(Quantum Key Distribution) 네트워크 채널 예측·침입 탐지 등 리소스 제약형 엣지 ML(Machine Learning) 태스크에도 직접 적용 가능성을 시사한다.

I. 서 론

최근 자연어처리(NLP) 및 컴퓨터 비전 분야에서는 수십억 개 파라미터의 사전학습 모델이 다양한 다운스트림 작업에서 최첨단 성능을 보인다. 하지만 다운 스트림 환경에서는 고품질의 데이터를 통한 미세조정(Fine-tuning)이 필수적인데 실제 환경에서의 클라이언트는 충분한 양의 방대한 고품질 데이터를 보유하기 어렵다.

이로 인해 데이터를 한 곳에 집중하지 않고도 효과적으로 모델을 학습할 수 있는 연합 학습(Federated Learning, FL)[1]이 주목받고 있다. FL은 각 클라이언트가 로컬에서 학습한 파라미터만 중앙 서버에 전송하여 집계함으로써, 데이터 프라이버시를 보호하면서도 높은 성능을 달성할 수 있는 분산 학습 방법이다.

그러나 FL 환경에서는 각 클라이언트의 연산 능력과 메모리 용량은 매우 이질적이다. 예를 들어, 성능이 낮은 엣지 디바이스에서는 모델의 모든 파라미터를 학습하거나 업데이트하는 과정에서 과도한 메모리 및 연산 비용으로 인해 심각한 성능 저하나 학습 지연이 발생할 수 있다. 이를 완화하기 위해 LoRA(Low-Rank Adaptation)[2]와 같은 파라미터 효율적 튜닝(PEFT) 기법을 활용한다. LoRA는 가중치 업데이트를 낮은 차원의 행렬 곱으로 대체하여 계산 및 메모리 부담을 효과적으로 줄인다. 하지만 이때 모든 클라이언트에 동일한 저차원 행렬의 크기(랭크, rank)를 적용하면 클라이언트 간의 자원 활용 비효율이 발생한다. 예를 들면, 랭크가 클 경우 저사양 클라이언트에서는 메모리나 연산이 과부하되고, 반대로 랭크가 작을 경우 고사양 클라이언트는 표현력 부족으로 충분한 성능을 달성하지 못하게 된다.

본 논문에서는 FL 환경에서 각 클라이언트의 자원 제약을 고려하여, 각 클라이언트에 맞는 랭크를 동적으로 결정하고 주기적으로 중요도가 낮은 랭크 성분을 제거하는 Federated MagLoRA(Max-flops and GPU LoRA)를 제안한다. 이를 통해 이기종 분산 환경에서도 통신·메모리 비용을 최소화하면서 학습 성능 저하를 억제한다. 또한 QKD(Quantum Key Distribution) 네트워크의 채널 상태 예측·침입 탐지 같은 엣지 ML(Machine Learning) 태스크에도 직접 적용할 수 있음을 논의한다.

2.1 Federated MagLoRA algorithm

LoRA[2]는 사전학습된 가중치인 $W = W^0 + BA$, $B \in R^{d_{out} \times r}$, $A \in R^{r \times d_{in}}$ 에서 여기서 r (랭크)은 추가할 행렬의 차원으로 $r \ll \min(d_{out}, d_{in})$ 을 택해 파라미터 효율을 달성한다. FedAvg는 각 클라이언트 k 가 로컬 업데이트한 모델 파라미터 θ_k 를 로 평균화하여 글로벌 모델을 동기화한다.

$$\theta_{glob} \leftarrow \sum_k \frac{n_k}{N} \theta_k \quad (1)$$

이때 θ_k 는 클라이언트 k 가 로컬에서 업데이트한 모델 파라미터, n_k 는 클라이언트 k 의 데이터 수, N 은 전체 데이터 수이다. 클라이언트는 자신의 연산 자원인 (max_flops, max_memory)를 이용해 목표 랭크인 r_{tgt} 를 결정한다. r_{mem} 은 클라이언트가 최대 메모리를 사용할 때 가능한 랭크, r_{flop} 은 최대 연산량을 사용할 때 가능한 랭크로 최종 목표 랭크는 이 값 중 작은 값으로 결정된다. 이를 식으로 표현하면 (2)식과 같이 표현 가능하다. d_{in} 은 입력 차원으로 입력 특성 벡터의 길이를 의미하며 d_{out} 은 출력 차원으로 LoRA 적용 대상의 선형 계층의 출력 차원이다.

$$r_{mem} = \left\lceil \frac{max_memory}{d_{in} + d_{out}} \right\rceil, r_{flop} = \left\lceil \frac{max_flops}{d_{in} + d_{out}} \right\rceil$$

$$r_{tgt} = \max(1, \min(r_{max}, r_{mem}, r_{flops})) \quad (2)$$

이후 정해진 주기인 ΔT 마다 현재 랭크 r_{cur} 에서 중요도가 낮은 성분을 제거하여 목표 랭크로 맞춘다.

$$S_i = \|B_{:,i}\|_2 * \|A_{i,:}\|_2 \quad (3)$$

여기에서[식 3] S_i 는 랭크 i 의 중요도를 나타내고 중요도는 LoRA 저차원 행렬 내 특정 성분의 기여도를 의미한다. B 는 LoRA 차원 축소 행렬로 원래 Weight에 추가되는 보정값을 생성하며 A 는 차원 복원 행렬로 입력 임베딩을 저차원에서 다시 복원하는 역할을 한다. 해당 수식에 의해 기여도가 낮은 성분부터 순차적으로 제거되며, LoRA 저차원 행렬의 해당 성분의 $L_2 - norm$ 곱으로 계산된다. 이는 해당 성분이 전체 업데이트에서 차지하는 기여도를 간접적으로 측정할 수 있으며, 이후 값이 작은 성분부터 순차적으로 제거(pruning)된다. 제안된 알고리즘은 다음 pseudocode로 요약된다

Algorithm Federated MagLoRA

II. 본 론

Input: 클라이언트 제약 $\{(max_flops, max_mem)\}$, rounds R ,
prune interval ΔT

Output: 글로벌 모델 파라미터 θ

Initialize global θ^0

for $t=1$ to R do

for each client k in parallel do

$\theta_k \leftarrow \theta^{t-1}$

LocalTrain(θ_k)

if $t \bmod \Delta T == 0$ then

DynamicPrune(θ_k, B, θ_k, A)

end if

end for

$\theta^t \leftarrow \text{FedAvg}(\{\theta_k\}, \{n_k\})$

End for

2.2 QKD 네트워크 후처리 응용

제안된 Federated MagLoRA 는 QKD(양자 키 분배) 네트워크의 채널 상태 예측 및 침입 탐지 같은 후처리 ML 태스크에도 적용 가능하다. 각 QKD 노드는 키 생성 속도(key rate)와 메모리 버퍼 한계에 대응해 연산 능력을 지정하고, MagLoRA 가 이를 기반으로 LoRA 어댑터의 랭크를 동적으로 조정한다. 노드별로 경량화된 모델이 빠르게 수렴하면서도 FedAvg 를 통해 공유된 글로벌 모델이 이질적 채널 환경 전반을 안정적으로 학습하는 효과를 기대할 수 있다. 예컨대, QKD 노드가 주변 채널 상태에 따라 노이즈 패턴을 학습하거나, 이상 신호를 감지하는 작업은 제한된 연산 환경에서도 적응형 경량 모델이 요구된다.

2.3 실험결과

본 절에서는 SST-2 데이터셋을 이용해 3 개의 이질적 클라이언트(각각 서로 다른 FLOPs·메모리 제약)를 가정한 연합학습 시나리오에서 제안하는 MagLoRA 와 고정 랭크 LoRA($r = 8, 4$)를 비교 평가한 결과를 서술한다. 모델 성능을 단순 정확도 외에도 연산 자원 대비 효율적으로 평가하기 위해, 본 논문에서는 기존 연구[3]를 참고하여 다음과 같은 효율성 지표(Efficiency Score) 를 활용한다.

$$\text{Efficiency Score} = \frac{\text{Accuracy}}{\frac{\text{FLOPs}}{\text{sample}} * 10^{-6}}$$

해당 지표는 모델이 소비한 단위 연산량(FLOPs) 대비 얻은 예측 정확도를 나타내며, 값이 클수록 정확도를 덜 희생하고도 연산 효율이 높다는 것을 의미한다. 따라서, 자원 제약 환경에서의 실제 적용 가능성을 평가하는 데 효과적이다.

Method	Avg. Accuracy (%)	Final Loss	Avg. FLOPs /Sample	Trainable Params	Final Rank	Eff. Score	Resource
Fixed $r=8$	86.7	0.31	105,728	7,426	8	819.5	5.62
Fixed $r=4$	87.40	0.31	102,144	3,842	4	855.5	2.90
MagLoRA	87.13	0.32	99,456	1,154	1	875.6	0.87

표 1. 평균 성능 및 자원 효율 비교표

수렴 속도 및 정확도 측면에서 세 방법 모두 유사한 최종 성능을 보였다[그림 2]. Fixed $r = 8$ 모델은 10 라운드 기준 평균 정확도 86.70%, Fixed $r = 4$ 는 87.43%, 제안하는 MagLoRA 는 87.13%로[표 1], MagLoRA 가 고정 랭크 방식과 유사한 수준의 정확도를 유지함을 확인하였다. 특히 MagLoRA 는 초반 $r=8$ 로 시작한 후, 3 라운드부터 목표 랭크로 신속히 전환하여 4 라운드 이내 안정적인 수렴을 달성하였다.

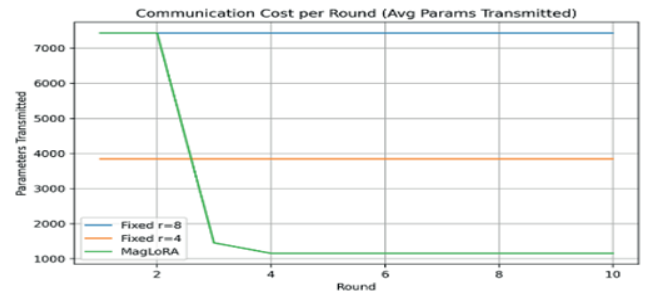


그림 2. 연합학습 라운드별 정확도 비교

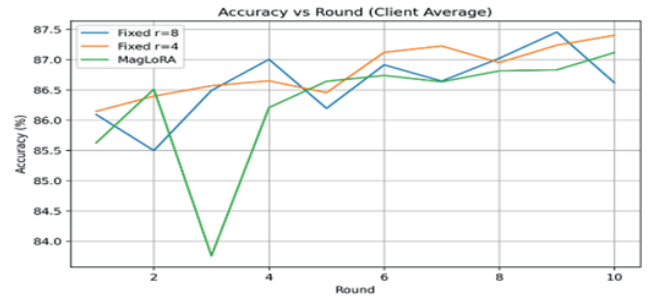


그림 3. 연합학습 라운드별 통신 비용 비교

하지만 연산 효율 측면에서 MagLoRA 는 샘플당 평균 FLOPs 를 기존 고정 $r = 8$ 대비 약 6.0%, $r = 4$ 대비 약 2.6% 절감하였으며[그림 3], 특히 파라미터 수와 통신량에서도 각각 최대 85% 및 70% 수준의 절감 효과를 보였다. 이를 통해 MagLoRA 는 성능 저하 없이 자원 효율성을 크게 향상시킴을 입증하였다.

마지막으로 효율성 지표[3]를 적용한 결과, Fixed $r = 8$ 은 평균 819.5, Fixed $r = 4$ 는 평균 855, MagLoRA 는 평균 875 을 기록하여 MagLoRA 가 가장 높은 효율성을 나타냈다. 즉, MagLoRA 는 연합학습 환경에서 계산·메모리·통신 비용을 현저히 줄이면서도 정확도 손실을 최소화하는 우수한 성능을 입증하였다.

III. 결론

본 논문에서는 이기종 연합 학습 환경에서 사전학습 대형 모델의 미세조정을 위한 federated MagLoRA 를 제안하였다. 각 클라이언트가 자신의 연산 능력에 기반해 목표 랭크를 동적으로 계산하며, 주기적 랭크 프루닝으로 불필요 파라미터를 제거하고 FedAvg 로 글로벌 모델 동기화를 통해, 통신량·메모리 비용은 대폭 절감(최대 85%/70%)되고, 정확도 손실은 0.6%p 이하로 미미함을 보였다. 또한, QKD 네트워크의 채널 상태 예측·침입 탐지와 같이 옛지 ML 태스크에도 MagLoRA 를 적용할 수 있어, 리소스 제약이 심한 다양한 옛지 기반 연합학습 및 QKD 네트워크 응용에서 실질적인 적용 가능성을 보여준다.

ACKNOWLEDGMENT

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

참고 문헌

- [1] H. Brendan McMahan, "Communication-Efficient Learning of Deep Networks from Decentralized Data," Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 1273–1282, 2017.
- [2] Edward J. Hu, "LoRA: Low-Rank Adaptation of Large Language Models," arXiv preprint arXiv:2106.09685, 2021.
- [3] Andrew G. Howard, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv preprint arXiv:1704.04861, 2017.