# 이질적인 연합 학습 기반 미세조정을 위한 LoRA 기법

하시현, 전요셉
포항공과대학교

{ha6884311, yoseb.jeon}@postech.ac.kr

# LoRA Method for Heterogeneous Federated Fine-Tuning

Sihyeon Ha, Yo-Seb Jeon*
Pohang Univ. of Science and Technology (POSTECH)

## 요 약

본 논문은 federated fine-tuning 환경에서 rank-heterogeneity 문제가 존재하는 LoRA 모듈의 집계를 보다 효과적으로 수행하기 위한 새로운 프레임워크를 제안한다. LoRA 모듈을 rank-wise 로 분해한 후, rank 별로 가중치를 적용하는 집계 문제를 일반화된 형태로 수식화한다. 또한, 기존의 zero-padding 및 replication 기반 방법들이 이 일반화 문제의 특수한 해로 귀결됨을 보인다. 본 연구에서는 각 클라이언트의 데이터셋 크기와 LoRA rank 정보를 기반으로 가중치를 설정하는 rank-aware 집계 방식을 제안하였다. 제안 기법은 다양한 rank 설정에서도 효과적인 모델 통합이 가능하며, 실험을 통해 기존 방식 대비 높은 성능과 빠른 수렴 속도를 확인할 수 있었다.

## Ⅰ. 서 론

최근 대규모 모델을 미세 조정(fine-tuning)하여 제한된 환경에서도 효율적으로 활용하려는 연구가 활발히 진행되고 있으며, 대표적인 방법으로 LoRA(Low-Rank Adaptation)가 주목받고 있다 [1]. LoRA 는 전체 모델 파라미터를 고정한 채, 소수의 low-rank 행렬만을 학습함으로써 연산 비용을 절감하고 다양한 장치 환경에서의 미세 조정을 가능하게 한다. 이러한 특성은 연합 학습(federated learning) 환경에서 특히 유용하며, 클라이언트가 로컬 데이터에 맞춰 LoRA 모듈만을 학습하고 이를 서버에서 통합하는 방식의 federated fine-tuning 시나리오에 적합하다.

하지만 클라이언트 간 자원의 이질성으로 인해 서로 다른 LoRA rank 설정이 적용되는 경우가 많다. 이로 인해 발생하는 rank heterogeneity 문제는 LoRA 모듈의 차원 불일치를 초래하며, 단순한 평균이나 합산 기반의 집계가 불가능하다. 기존 연구에서는 서로 다른 rank 를 맞추기 위해 zero-padding (ZP) [2]이나 replication [3]과 같은 방식이 제안되었다. 하지만, 이러한 방법들은 rank 간 정보 활용이 비효율적이며, 보다 효과적이고 일반화 가능한 집계 방식으로 확장할 여지가 존재한다.

본 연구에서는 각 rank 에 대해 가중치를 적용하는 집계 문제를 일반화된 형태로 정의한다. 이를 기반으로, 기존의 ZP 및 replication 방식이 모두 해당 형태의 특수한 해임을 보이며, 이를 통해 보다 유연하고 해석 가능한 관점을 제시한다. 또한, 각 클라이언트의 데이터셋 크기와 LoRA rank 정보를 기반으로 가중치를 결정하는 heuristic 한 기법을 도입하여, 다양한 rank 설정 하에서도 효과적으로 모델을 통합하고 높은 미세조정 성능을 달성할 수 있도록 한다.

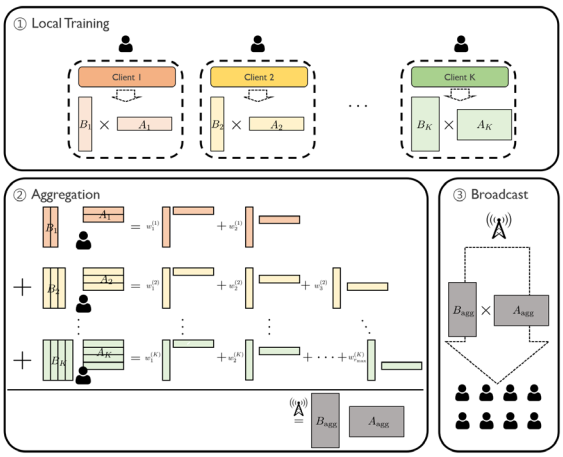


그림 1. Federated fine-tuning 파이프라인

## Ⅱ. 본론

본 연구는 federated fine-tuning 시나리오에서 클라이언트마다 서로 다른 LoRA rank 를 사용하는 rank heterogeneity 문제를 해결하기 위한 효과적인 집계 프레임워크를 제안한다. 클라이언트가 각자의 환경, 데이터 규모, 연산 자원에 따라 LoRA rank 를 다르게 설정하게 되면, 서버는 서로 다른 구조의 LoRA 모듈을 수신하게 된다. 특히, rank 가 다른 클라이언트로부터의 LoRA 업데이트는 차원이 일치하지 않아 직접적인 평균이 불가능하다는 어려움이 있다.

기존 연구들은 이러한 문제를 해결하기 위해 몇 가지 heuristic 한 전략을 제안해왔다. 대표적으로 사용되는 방법은 ZP 와 replication 이다. ZP 는 낮은 rank 의 LoRA 모듈을 높은 rank 로 확장하기 위해 남은 부분을 0 으로 채우는 방식이다 [2]. 이 방식은 구현이 간단하고 통합 후의 모양이 일관되다는 장점이 있지만, padding 된 영역은 의미 있는 정보를 포함하지 않기 때문에 실제

집계 과정에서 뒷부분 rank 의 영향력이 감소하게 된다. 반면, replication 은 padding 된 영역을 기존 LoRA 행렬의 일부분으로 복제하여 채움으로써 정보 밀도를 유지하려는 시도이다 [3]. 하지만 이 방식은 후반부 rank 에 동일한 정보를 중복 삽입하는 셈이며, 이는 특정 rank 에 2 배의 가중치를 부여하는 것과 같은 효과를 낳는다. 게다가 이 방식은 두 개의 rank 수준 사이에서는 동작 가능하지만, 클라이언트가 세 개 이상 다양한 rank 를 가질 경우에는 어떻게 복제할지를 정의하기 어렵다.

이러한 한계를 극복하기 위해, 본 논문에서는 LoRA 모듈을 rank-wise 로 분해하여 표현하고, 각 rank 에 대해 별도의 가중치를 적용하는 방식으로 집계 기법을 일반화한다. LoRA 모듈은 다음과 같이 분해할 수 있다.

$$\mathbf{B}_{\text{agg}}\mathbf{A}_{\text{agg}} = \sum_{l=1}^{K}\sum_{r=1}^{r^{(l)}}\left(\boldsymbol{b}_r^{(l)} \oplus \cdots \oplus \boldsymbol{b}_r^{(K)}\right)\left(\boldsymbol{a}_r^{(l)^{\mathsf{T}}} \oplus \cdots \oplus \boldsymbol{a}_r^{(K)^{\mathsf{T}}}\right) \quad (1)$$

$\mathbf{B}_{\text{agg}}\mathbf{A}_{\text{agg}}$ 는 집계된 LoRA 모듈을 $\boldsymbol{b}_r^{(l)}$, $\boldsymbol{a}_r^{(l)}$ 은 $l$ 번째 클라이언트의 $r$ 번째 rank 의 열을 의미한다. $\oplus$ 는 집계하는 연산을 나타낸다. 이를 확장하면, rank 마다 서로 다른 가중치를 적용하여 다음과 같은 형태로 집계 기법을 표현할 수 있다:

$$\mathbf{B}_{\text{agg}}\mathbf{A}_{\text{agg}} = \sum_{r=1}^{r^{(K)}}\begin{bmatrix} | & \cdots & | \\ \boldsymbol{b}_r^{(1)} & \cdots & \boldsymbol{b}_r^{(K)} \\ | & \cdots & | \end{bmatrix} \cdot \begin{bmatrix} w_{a,r}^{(1)}w_{b,r}^{(1)} & w_{a,r}^{(1)}w_{b,r}^{(2)} & \cdots & w_{a,r}^{(1)}w_{b,r}^{(K)} \\ w_{a,r}^{(2)}w_{b,r}^{(1)} & w_{a,r}^{(2)}w_{b,r}^{(2)} & \cdots & w_{a,r}^{(2)}w_{b,r}^{(K)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{a,r}^{(K)}w_{b,r}^{(1)} & w_{a,r}^{(K)}w_{b,r}^{(2)} & \cdots & w_{a,r}^{(K)}w_{b,r}^{(K)} \end{bmatrix} \begin{bmatrix} | & \cdots & | \\ \boldsymbol{a}_r^{(1)} & \cdots & \boldsymbol{a}_r^{(K)} \\ | & \cdots & | \end{bmatrix}^{\mathsf{T}}$$

$$= \sum_{r=1}^{r^{(K)}} \mathbf{B}\mathbf{W}_r\mathbf{A}^{\mathsf{T}} \quad (2)$$

이때 각 $\mathbf{W}_r$은 rank $r$ 에 해당하는 LoRA 모듈에 부여되는 가중치를 나타내며, 결과적으로 이 문제는 전체 $r \times d \times d$ 크기를 갖는 가중치 $\mathbf{W}_r$을 설계하는 문제로 일반화된다. 이 표현은 다양한 집계 기법을 하나의 수식 틀 안에서 포괄적으로 설명할 수 있는 장점을 가진다.

표 1. 집계 기법별 $\mathrm{W}_r$ 결정

| | $\mathbf{W}_r,\ w_r^{(k)} = w_{a,r}^{(k)} = w_{b,r}^{(k)}$ |
|---|---|
| FedAvg | $w_r^{(k)} = \frac{\lvert D^k \rvert}{\sum_{k'}\lvert D^{(k')} \rvert}$, $\lvert D^k \rvert$ ≔클라이언트 $k$가 보유한 데이터셋의 샘플 수 |
| ER | $w_r^{(k)} = \frac{C_1}{C_r}$, $C_r$ ≔rank $r$을 가진 클라이언트의 수 |
| Proposed | $w_r^{(k)} = \frac{\lvert D^{(k)} \rvert}{\sum_{k'}\lvert D^{(k')} \rvert} \cdot \frac{\lvert D_1 \rvert}{\lvert D_r \rvert}$, $\lvert D_r \rvert$ ≔ rank $r$을 가진 클라이언트들이 보유한 데이터셋 크기의 총합 |

표 1 은 기존 방식들과 본 논문에서 제안하는 방식이 위 수식 내에서 $\mathbf{W}_r$을 어떻게 설정하는지를 요약한 것이다. FedAvg 방식은 각 클라이언트가 보유한 데이터셋의 크기에 비례하여 $\mathbf{W}_r$ 을 선형 결합하는 방식으로, 데이터 양을 중심으로 한 평균 기반 집계를 수행한다. Extended replication (ER) 방식은 replication 기법을 클라이언트가 3 명 이상일 때에도 적용 가능하도록 일반화한 방식이다. 본 논문에서 제안하는 기법은 위의 두 방식을 결합한 형태로, rank 간 정보량의 불균형을 완화하고, 클라이언트 수 및 데이터 크기를 동시에 고려하여 각 $\mathbf{W}_r$ 에 대한 집합적인 중요도를 정량적으로 반영하는 구조를 갖는다.

이러한 설계를 검증하기 위해, 서로 다른 데이터셋과 LoRA rank 를 사용하는 federated fine-tuning 환경을 구성하고, Food-101 데이터셋으로 실험을 수행하였다. 실험에는 vision transformer (ViT) 모델의 16b variant 를 기반으로 한 사전학습된 모델을 사용하였으며, 총 4 개의 클라이언트 그룹이 참여하였다. 각 그룹은 LoRA rank 가 각각 2, 4, 8, 16 으로 설정된 2 개의 클라이언트로 구성되었다. 성능 평가는 round 에 따른 미세조정 정확도를 기준으로 수행되었다, 그림 2 는 제안 기법이 기존 방식들보다 더 빠르게 수렴하고, 더 높은 성능을 달성함을 보여준다.

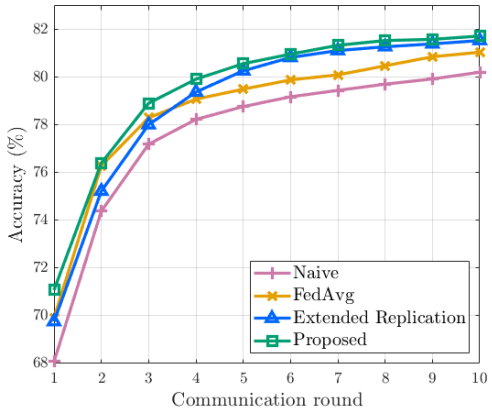


그림 2. 집계 기법별 Accuracy 측정

## Ⅲ. 결론

본 논문은 federated fine-tuning 환경에서의 rank heterogeneity 문제를 해결하기 위해, LoRA 모듈을 rank-wise 로 분해하고 각 rank 에 가중치를 적용하는 집계 방식을 제안하였다. 제시된 수식 구조는 기존의 ZP 와 replication 방식이 포함되는 일반화된 틀을 제공하며, 클라이언트 수와 데이터셋 크기를 고려한 가중치 설계를 통해 더 유연하고 정보 반영적인 집계를 가능하게 한다. 실험을 통해, 제안 기법이 기존 방식 대비 높은 미세조정 성능과 빠른 수렴 속도를 달성함을 확인하였다. 향후에는 데이터 품질 등 추가 정보를 반영한 가중치 설계로의 확장을 고려할 수 있다.

## ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. RS-2024-00453301).

## 참 고 문 헌

[1] E. Hu et al. “Lora: Low-rank adaptation of large language models,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Jan. 2022, pp.

[2] Y. J. Cho, et al. "Heterogeneous lora for federated fine-tuning of on-device foundation models," arXiv preprint arXiv:2401.06432 (2024).

[3] Y. Byun and J. Lee, “Towards federated low-rank adaptation of language models with rank heterogeneity,” arXiv preprint arXiv:2406.17477 (2024).