언어모델의 멀티태스크 능력을 위한 Router-Free 병합 방식에 대한 연구

이찬빈
울산과학기술원

chblee@unist.ac.kr

Router-Free Merging Methods for the Multitask Ability of Language Models

Chanbin Lee
Ulsan National Institute of Science and Technology

요 약

언어 모델이 다양한 태스크를 수행하도록 하기 위해 개별 태스크에 학습시킨 LoRA(Low-Rank Adaptation) 가중치를 라우터를 통해 연결하는 Mixture-of-LoRAs(MoA)와 같은 방법이 널리 사용된다. 그러나 이러한 접근은 추가적인 라우터 학습과 추론 복잡도를 요구한다는 단점이 있다. 본 연구에서는 태스크별로 구별되는 입력 형식을 사용할 경우, 명시적 라우팅 없이 단순한 수학적 가중치 병합만으로 멀티태스크 성능을 확보할 수 있는지 분석한다. 이를 위해 여러 태스크에 대해 독립적으로 학습된 LoRA 모듈을 연결(Concatenation), 가중합(Weight Summing), 특이값 분해(SVD) 기반 방식으로 병합하고 멀티태스크 성능을 MoA 및 oracle 설정과 비교하였다. 실험 결과, 1.5B 규모 모델에서 MoA 성능의 최대 95% 수준에 도달하였고, 태스크 간 LoRA 가중치의 직교성이 이러한 현상과 밀접한 관련이 있음을 확인하였다. 결론적으로 본 연구는 멀티태스크 능력을 위한 라우팅의 필요성을 재검토하고, 단순하면서도 효율적인 대안 가능성을 제시한다.

Ⅰ. 서론

언어 모델의 활용이 확대됨에 따라, 하나의 모델이 다양한 태스크를 수행할 수 있도록 하는 멀티태스크 학습 기법에 대한 관심이 증가하고 있다. LoRA(Low-Rank Adaptation)는 효율적 미세조정 기법으로서, 태스크별 성능을 효과적으로 확보할 수 있다는 점에서 널리 사용되고 있다[1]. 멀티태스크 능력을 위해서는 각각의 LoRA 모듈을 병합하여 사용하는 방식이 일반적이다.

기존 연구에서는 Mixture-of-LoRAs(MoA)와 같이 입력에 따라 적절한 LoRA 모듈이 선택되도록 하는 라우터 기반의 방식이 주로 사용되었다[2]. 이러한 방식은 높은 성능을 보이나, 추가 학습이 필요한 라우터를 포함함으로써 구조적 복잡성과 계산 비용을 증가시킨다. 이는 실제 시스템 적용 관점에서 중요한 제약이 될 수 있다.

이에 본 연구는 태스크별 입력 형식이 명확히 구별되는 경우, 라우터 구조가 반드시 필요한지에 대한 질문을 던진다. 입력 프롬프트 자체가 태스크 정보를 충분히 포함한다면, 추가 학습이 필요하지 않은 LoRA 가중치 병합만으로 멀티태스크 처리가 가능할 수 있다는 가설을 제시하고, 이를 실험적으로 검증한다.

Ⅱ. 실험 과정

2-1. Router-Free LoRA 모듈 병합 방식

본 연구에서는 각 태스크에 대해 독립적으로 학습된 LoRA 가중치를 라우터 없이 하나의 모델로 통합하는 세 가지 병합 전략을 설계한다.

(1) Concatenation 방식은 태스크별 LoRA 모듈을 입력에 순차적으로 적용하고 alpha 와 rank 값으로 스케일링하여 합하는 가장 단순한 방식이다.

$$\Delta W = \sum_{i=1}^{N} \frac{\alpha}{r_i} A_i B_i \quad (1)$$

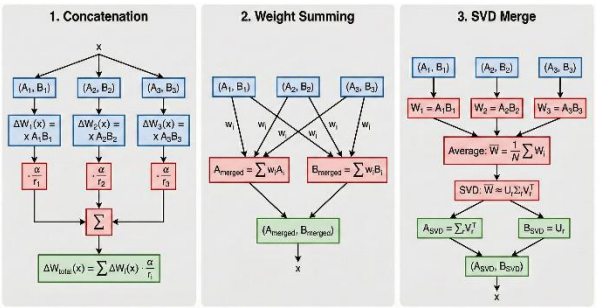
(2) Weight Summing 방식은 각각의 LoRA 모듈의 A 행렬과 B 행렬의 가중합을 먼저 계산하여 병합한다.

$$A_{merged} = \sum_{i=1}^{N} w_i A_i \,, B_{merged} = \sum_{i=1}^{N} w_i B_i \quad (2)$$

(3) SVD Merge 방식은 특이값 분해(SVD)를 통해 저차원 근사를 수행함으로써 하나의 압축된 LoRA 모듈을 생성한다.

$$\bar{W} = \frac{1}{N}\sum_{i=1}^{N} A_i B_i = U\,\Sigma V^{\top},$$

$$A_{merged} = \Sigma_{1:r} V^{T}_{1:r}, B_{merged} = U_{:,1:r} \quad (3)$$



[그림 1] 3종류의 Router-Free 병합 방식.

2-2. 실험 세부사항

실험은 5개의 자연어처리 태스크(요약, 번역, 질의응답, 감정분석, 자연어추론)를 이용해 진행하였다. 먼저 각각의 태스크에 대해 LoRA 가중치를 학습하고, 이후

모든 태스크가 섞인 평가용 데이터셋을 구성해 성능을 측정하였다. 이때 각각의 태스크는 고유의 구조화된 프롬프트를 사용하며, 이를 통해 입력 단계에서 저마다 서로 다른 형식을 사용하도록 설계하였다. 그림 2는 이러한 입력 형식의 예시이다.

Prompt Format – Question Answering

```
### Context:
{context}
### Question:
{question}
### Answer:
```

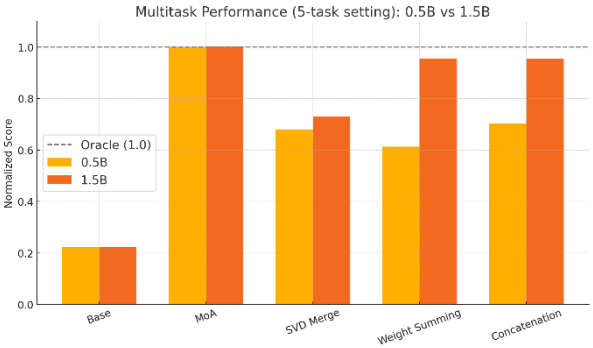
[그림 2] 질의응답 태스크에서 사용된 프롬프트 형식.

평가 방식에 있어서는, 태스크별로 서로 다른 성능 지표를 사용하므로, 각 태스크의 성능을 해당 태스크의 oracle 성능(태스크별 LoRA 를 단독 적용한 경우)으로 정규화한 뒤 평균을 계산하여 멀티태스크 성능 점수(multitask score)로 산출했다. 이는 특정 태스크에 편향되지 않고 병합 방식에 따른 전반적인 성능을 비교하기 위함이다. 또한, 실험에 사용된 언어모델은 모두 Qwen2 모델로 통일하였다.

Ⅲ. 결과 및 분석

3-1. 모델 크기에 따른 성능 분석

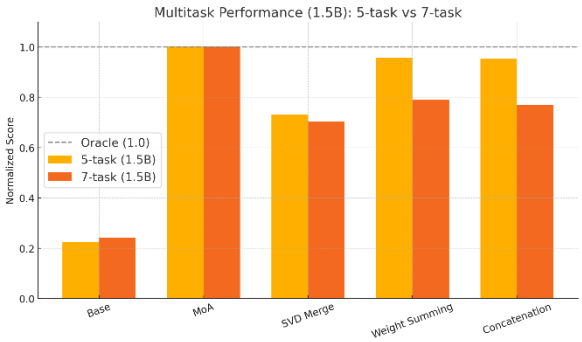
병합 방식별 멀티태스크 성능은 그림 3에 제시하였다. 0.5B 모델에서는 MoA 와 비교하였을 때 최대 70%에 불과한 성능을 보이는 것을 확인할 수 있다. 그러나 1.5B 모델에서는 Weight Summing 및 Concatentation 기반 병합 방식에서 MoA 의 최대 95% 수준의 성능에 도달하였다. 이는 모델의 크기가 일정 수준 이상 확보된다면 모델이 라우터 없이도 자체적으로 입력 형식에 드러난 태스크 정보를 활용해 알맞은 모듈을 사용할 수 있다는 것으로 풀이된다.



[그림 3] 0.5B, 1.5B 파라미터 모델에서 LoRA 모듈 병합 방식에 따른 멀티태스크 점수.

3-2. 태스크 개수에 따른 성능 분석

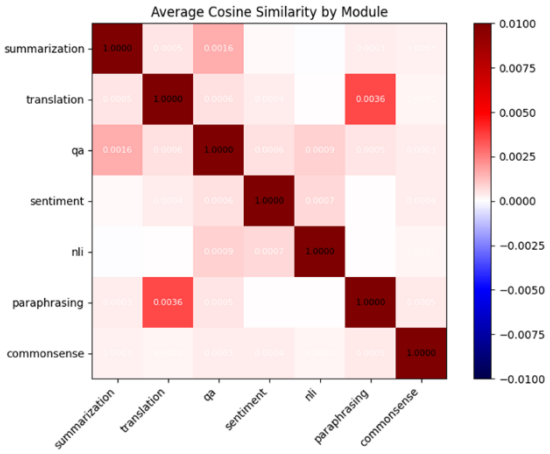
이어서 본 연구에서는 태스크 개수가 증가함에 따라 제시한 방식들의 멀티태스크 성능이 어떻게 변화하는지 관찰하였다. 이를 위해 2개의 추가 태스크(재서술, 상식추론)를 도입하였고, 1.5B 모델에서의 실험 결과를 그림 4에 나타내었다. 태스크가 늘어남에 따라 모든 Router-Free 병합 방식에서 성능 감소가 나타났지만, SVD Merge 방식의 경우에는 태스크 개수 증가에 따른 안정성이 높은 것으로 나타났다.



[그림 4] 각각 5, 7개 태스크에서 LoRA 모듈 병합 방식에 따른 멀티태스크 점수.

3-3. LoRA 모듈의 직교성 분석

병합 방식이 효과적으로 작동하는 이유를 분석하기 위해, LoRA 가중치 간 코사인 유사도를 측정하였다. 그림 5에 나타난 것처럼 측정 결과 서로 다른 태스크의 경우에는 코사인 유사도가 0에 매우 가깝게 관측되는데, 이러한 직교성은 가중치를 선형적으로 결합하는 경우에 최소한의 간섭으로 병합을 수행할 수 있다는 것을 의미한다. 이것이 일부 병합 방식이 라우터 없이도 좋은 멀티태스크 성능을 달성할 수 있는 이유로 해석된다.



[그림 5] Task 별 LoRA 가중치의 cosine similarity.

Ⅳ. 결론

본 연구는 LoRA 기반 멀티태스크 병합에 있어, 라우팅 구조의 필요성을 실험적으로 재검토하였다. 태스크별 입력 형식이 명확하게 구별되는 환경에서 언어모델이 라우터 없이도 자체적으로 알맞은 모듈을 활용할 수 있는 능력이 있음을 보였으며, 이러한 발견을 바탕으로 더욱 효율적인 병합 방식의 설계가 가능할 것으로 기대한다.

참 고 문 헌

[1] Hu, E. J. et al. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the 10th International Conference on Learning Representations (ICLR), 2022.

[2] Feng, W. et al. Mixture-of-LoRAs: An Efficient Multitask Tuning Method for Large Language Models. Proceedings of the 2024 Joint International Conference on Computational Linguistics and Language Resources Evaluation (LREC-COLING 2024), 11371-11380, 2024.