활성화 기반 저랭크 분해를 활용한 대형언어모델 경량화

하시현, 전요셉*
포항공과대학교

{sihyeon.ha, yoseb.jeon}@postech.ac.kr

Activation-based Low-Rank Decomposition for LLM Compression

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

요 약

대형언어모델(Large Language Model, LLM)은 수십억 개 이상의 파라미터를 갖고 있으며, 이러한 거대 규모는 성능 향상의 핵심이자 동시에 배포 효율의 가장 큰 걸림돌로 작용한다. 이에 따라 다양한 모델 경량화 연구가 진행되어 왔으나, 기존의 양자화와 프루닝 방식은 하드웨어 의존성과 성능 저하 문제를 완전히 해소하지 못하였다. 최근에는 입력 활성의 통계적 특성을 반영하여 정보 손실을 최소화하는 저랭크 분해 기법이 주목받고 있다. 이러한 접근은 단순히 가중치 행렬의 크기를 줄이는 것이 아니라, 모델이 실제로 처리하는 입력의 통계적 구조를 고려함으로써 표현 능력을 최대한 유지할 수 있다는 점에서 효과적이다. 본 논문에서는 활성화 기반 저랭크 분해의 수학적 원리와 기존 SVD 기반 경량화 방식의 차이를 분석하고, 별도의 미세 조정 없이도 성능을 유지할 수 있는 효율적 경량화 전략으로서의 가능성을 살펴본다.

Ⅰ. 서 론

대형언어모델은 사전학습을 통해 방대한 데이터로부터 언어적 표현과 추론 능력을 학습하며, 최근의 GPT-4, Claude, LLaMA 등은 이미 수백억 개 이상의 파라미터를 포함하고 있다. 그러나 이러한 거대 모델은 일반적인 GPU 환경에서 추론조차 어려우며, 실제 응용 시스템에 배포하기 위해서는 효율적인 파라미터 축소와 연산 최적화가 필수적이다.

대표적인 모델 경량화 방법으로는 양자화와 프루닝이 있다. 양자화는 가중치의 비트 정밀도를 줄여 메모리 사용량을 감소시키는 방식이지만, 실제 추론 과정에서는 비양자화(dequantization)가 필요하므로 연산 효율의 이점을 온전히 누리기 어렵다. 프루닝은 중요도가 낮은 가중치를 제거하여 모델을 단순화하지만, 성능을 유지하기 위해 복잡한 탐색 알고리즘이 필요하며, 단순한 규칙 기반 프루닝을 적용할 경우 모델 정확도가 크게 저하되는 문제가 있다.

이러한 한계를 보완하기 위한 대안으로 저랭크 분해 기반 경량화 방식이 주목받고 있다. 저랭크 분해는 대규모 가중치 행렬을 두 개의 저차원 행렬로 분해하여 파라미터 수를 줄이는 기법으로, 구조적 변형 없이 모델의 계산 복잡도와 메모리 사용량을 동시에 감소시킨다. 특히 파라미터 수가 줄어드는 비율에 비례하여 저장 공간과 연산량이 모두 줄어들기 때문에, 실제 추론 속도 향상으로도 이어지는 것이 장점이다.

본 논문에서는 이러한 저랭크 분해 기반 경량화 접근을 소개하고자 한다.

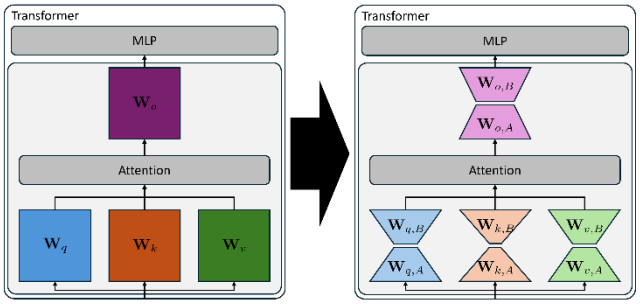

그림 1. 저랭크 분해 기반 LLM 경량화

Ⅱ. 본론

1. 전통적 저랭크 분해

저랭크 근사는 대규모 가중치 행렬 $W \in R_{d_{out} \times d_{in}}$ 을 두 개의 작은 행렬 $B \in R_{d_{out} \times r}$, $A \in R_{r \times d_{in}}$ 로 분해하여 $W \approx BA$ 로 근사하는 방식이다. 랭크 r 이 $\min(d_{out}, d_{in})$ 보다 충분히 작을 때, 전체 파라미터 수는 $d_{out} d_{in}$ 에서 $r(d_{out} + d_{in})$ 으로 감소하며 연산 복잡도도 이에 비례해 줄어든다.

가장 대표적인 방법은 특이값 분해(SVD decomposition)로, 다음 최적화 문제를 해결한다.

$$\min_{rank(BA) \le r} \|W - BA\|_F^2 \quad (1)$$

Eckart–Young–Mirsky 정리에 따르면 (1)의 해는 W의 상위 r 개 특이값과 대응하는 특이벡터를 이용한 분해이며, 최적해를 보장한다 [1]. 즉, 특이값 분해는 추가적인 학습 없이도 행렬 자체를 가장 효율적으로

압축할 수 있는 폐형식 해를 제공한다. 이러한 특성으로 특이값 분해는 모델 경량화에 고려되어 왔다. 가중치 행렬을 두 개의 저차원 행렬로 치환하면, 행렬 곱셈 연산의 비용이 대폭 감소하고 저장 메모리도 선형적으로 줄어든다. 또한 SVD 는 재학습이나 파라미터 탐색이 필요 없기 때문에, 간단하고 안정적인 학습 후(post-training) 경량화 기법으로 적합하다.

그러나 전통적인 방식은 기본적으로 가중치 행렬 자체의 재구성 오차만을 최소화하므로, 모델이 실제 입력 데이터를 처리하는 과정에서의 통계적 특성이나 출력 변화를 직접적으로 반영하지는 못한다.

2. 활성화 기반(activation-aware) 관점

이러한 한계를 해결하기 위해 데이터 분포를 고려한 근사, 즉 활성화 인식 접근이 제안된다. 전통적인 저랭크 분해가 W 자체의 오차를 최소화한다면, 활성 인식 근사는 다음과 같은 문제를 푼다.

$$\min_{rank(BA)\le r} \mathbb{E}_x \|Wx - BAx\|_F^2 \quad (2)$$

(2)는 입력 x 의 분포를 고려한 출력 활성화 공간에서의 근사 오차를 최소화한다. 이는 단순히 가중치 크기가 아닌 입력 방향별 활성화된 정보량을 기준으로 차원을 축소한다.

이 접근은 입력 통계에 따라 중요한 주성분 방향을 우선적으로 보존하며, 별도의 미세 조정 과정이 필요 없다는 측면에서 학습 후 (post-training) 압축으로 분류할 수 있다. 입력 통계 집계에 필요한 순전파 연산만으로 충분하므로, 추가 연산 비용이 적고 실제 적용이 간단하다는 장점이 있다.

Ⅲ. 실험적 관찰 및 분석

활성화 기반 저랭크 분해 접근은 실제 대형언어모델에 적용하여 그 효율성과 성능 유지 능력을 검증할 수 있다. 본 절에서는 대표적인 실험 사례를 통해 이러한 접근의 경량화 효과와 성능 특성을 살펴본다. 우선 실험은 LLaMA-2-7B 모델의 다중 헤드 어텐션(Multi-head Attention) 구조에 포함된 선형 투영 계층(Q, K, V, O)에 저랭크 분해를 적용하는 형태로 수행되었다. 입력 활성의 공분산 행렬은 WikiText-2 데이터셋에서 추출된 일부 문장(약 512 개, 길이 1024 토큰)으로부터 추정하였으며, 별도의 미세 조정(fine-tuning) 과정은 수행하지 않았다. 즉, 순전파를 통해 계산된 입력 활성 통계만을 이용하여 각 계층별 분해가 이루어졌다.

평가에는 두 가지 지표가 사용되었다. 첫째, 언어 모델링 성능을 평가하기 위해 WikiText-2 와 Penn Treebank 데이터셋에서의 perplexity 를 측정하였다. 둘째, 추론 및 일반화 능력을 검증하기 위해 ARC-Easy, HellaSwag, PIQA 등 여러 zero-shot reasoning 벤치마크에서의 정확도를 비교하였다. 표 1 에 나타난 결과에서 볼 수 있듯이, 제안된 방식은 전체 파라미터의 약 15%만을 유지한 상태에서도 dense 모델 대비 언어 모델링 성능의 저하가 매우 제한적이었으며, zero-shot 평가에서도 원본 모델과 유사한 정확도를 유지하였다. 특히 기존의 단순 SVD 기반 방식은 동일한 압축 비율에서 성능이 급격히 하락한 반면, 활성화 기반 접근은 소폭의 성능 감소만을 보였다. 또한, 모델의 경량화가 실제 추론 속도 향상으로도 이어지는지를 검증하기 위해 FLOPs 와 추론 시간의 비례 관계를 측정한 결과(표 2), 파라미터 감소율에 비례하여 연산량이 선형적으로 줄어들었으며, 실제 GPU 환경에서의 추론 속도 또한 약 1.21 배 향상되었다. 이는 단순한 저장 공간 절감에 그치지 않고, 연산 효율 측면에서도 실질적인 개선 효과가 있음을 의미한다.

표 1. PPL 및 Zero-shot 정확도

기법	PPL (WikiText-2)	PPL (PTB)	Zero-shot (Average)
Dense	10.46	128.01	0.54
SVD	851.48	2512.70	0.31
FWSVD [2]	940.68	1896.48	0.41
Ours	**12.67**	**147.55**	**0.48**

표 2. FLOPs 와 Speedup

압축 비율	FLOPs in MHA	Speedup
Dense	1.00×	1.00×
0.05	0.84×	1.09×
0.10	0.69×	1.15×
0.15	0.53×	1.21×

Ⅳ. 결론

본 논문에서는 대형언어모델의 효율적 경량화를 위한 활성화 기반 저랭크 분해 방법을 다루었다. 기존의 양자화와 프루닝은 하드웨어 의존성 및 성능 저하 문제로 인해 실질적인 추론 효율을 확보하기 어렵지만, 저랭크 분해는 모델 구조를 유지하면서 파라미터 수와 연산량을 동시에 줄일 수 있는 효과적인 대안으로 평가된다. 특히 입력 활성의 통계적 특성을 반영하는 방식은 출력 왜곡을 최소화하며, 별도의 미세 조정 없이도 성능을 유지할 수 있음을 실험적으로 확인하였다. 이러한 결과는 활성화 기반 저랭크 분해가 성능과 효율을 균형 있게 달성할 수 있는 대형언어모델 경량화의 유망한 방향임을 시사한다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. RS-2025-24683888).

참 고 문 헌

[1] Eckart, Carl, and Gale Young. "The approximation of one matrix by another of lower rank." Psychometrika 1.3 (1936): 211-218.

[2] Hsu, Yen-Chang, et al. "Language model compression with weighted low-rank factorization." arXiv preprint arXiv:2207.00112 (2022).