SVD 기반 LLM 경량화에서 Adaptive LoRA 적용을 통한 성능 분석 연구

마예은, 김명준, 박서윤, 안진현*
명지대학교

k41031614@mju.ac.kr, tjmjtjmj@mju.ac.kr, dhn04100@mju.ac.kr

*wlsgus3396@mju.ac.kr

A Study on Performance Analysis of Adaptive LoRA in SVD-based Large Language Model Compression

Yeeun Ma, Myoungjun Kim, Seoyun Park, Jin-Hyun Ahn*
Myoungji Univ.

요 약

본 논문은 대규모 언어모델(LLM)의 효율적 경량화를 위해 제안된 SVD-LLM(Singular Value Decomposition for Large Language Models) 프레임 워크의 LoRA(Low-Rank Adaptation) 모듈을 파라미터 효율적 미세조정(PEFT: Parameter-Efficient Fine-Tuning) 기법 중 하나인 적응형 저랭크 적응(AdaLoRA)으로 대체하여 성능 및 자원 효율성을 분석한다. 기존 SVD-LLM은 모든 레이어에 동일한 고정 랭크를 적용하기 때문에 레이어별 중요도를 반영하지 못하여 과소 혹은 과잉 근사 문제가 발생할 수 있다. 이를 개선하기 위해, 본 연구에서는 SVD 기반 저차원 공간 위에서 AdaLoRA를 적용하여 레이어별 중요도에 따라 동적으로 랭크를 재분배하도록 하였다. WikiText-2 데이터셋을 활용한 실험 결과, 제안 방법은 기존 LoRA적용 대비 원본 모델의 지식과 분포를 잘 보존함을 확인하였다. 이러한 결과는 단순한 압축과 미세조정 접근의 결합을 넘어, 구조적 저랭크 근사와 적응형 저랭크 최적화의 융합이 LLM 경량화 연구에서 새로운 가능성을 제시함을 보여준다.

Ⅰ. 서 론

최근 대규모 언어모델은 다양한 자연어 처리(NLP) 분야에서 번역, 질의응답, 요약 등 다양한 과제에 걸쳐 탁월한 성능을 보이고 있다. 그러나 수십억 개 이상의 파라미터를 포함한 모델은 학습 및 추론 시 막대한 메모리와 연산 자원을 요구한다는 한계가 있다. 이를 해결하기 위해 최근에는 다양한 모델 압축 및 파라미터 효율적 미세조정(PEFT) 기법이 활발히 연구되고 있다.

대표적인 모델 압축 기법으로는 지식 증류(Knowledge Distillation), 가지치기 (Pruning), 양자화(Quantization), 저랭크 근사(Low-Rank Approximation) 등이 있다. 이 중 저랭크 근사 방법은 모델의 가중치 행렬을 저차원 공간으로 투영하여 메모리 사용량과 연산 복잡도를 줄일 수 있는 장점이 있다.

최근 제안된 SVD-LLM[1]은 모델의 가중치 행렬을 특이값 분해(SVD)하여 저차원 근사로 변환함으로써 메모리 사용량과 연산량을 크게 줄일 수 있는 대표적인 경량화 기법이다. 기존 SVD-LLM은 SVD 분해 후 낮아진 성능을 복구하기 위해서 LoRA(Low-Rank Adaptation)[2]를 통해 미세조정을 진행하게 된다. 이때 사용되는 LoRA는 대표적인 PEFT(PEFT: Parameter-Efficient Fine-Tuning) 기법으로, 사전 학습된 언어모델에 소수의 저랭크 행렬을 삽입하여 효율적으로 파라미터를 학습하는 방법이다. 따라서 전체 모델을 재학습하지 않고도 효율적인 미세조정을 가능하게 한다. 하지만 모든 레이어에 동일한 고정 랭크를 사용하는 방식으로 인해 표현력이 부족해지는 문제가 발생할 수 있다.

AdaLoRA(Adaptive LoRA)[3]는 LoRA를 개선하기 위해 등장하였으며, 학습 중 파라미터 중요도를 동적으로 평가하고, 그에 따라 랭크를 재분배하는 방식으로 자원 활용을 최적화한다. 본 연구에서는 SVD 기반 저차원 근사 모델인 SVDLLM의 LoRA 모듈을 AdaLoRA를 대체하여, 구조적 저차원 근사와 적응형 저랭크 최적화를 결합한 새로운 프레임워크를 제시하고 그 효과를 실험적으로 분석한다.

Ⅱ. 제안 방법

A. 연구 개요

언어 모델의 가중치 행렬 $W \in \mathbb{R}^{d_1 \times d_2}$은 특이값 분해를 통해 $W = U\Sigma V^T$로 표현될 수 있다. SVD-LLM에서는 상위 r개의 특이값만 유지하여 근사하여 메모리 사용량을 줄인다. 그러나 단순 근사만으로는 최적화 과정에서 수치 불안정성이 발생할 수 있어, Cholesky 분해를 추가적으로 활용한다. 즉, 공분산 행렬을 $C = SS^T$ 꼴로 분해하고 WS로 계산하여 whitening을 수행한다. 이후 근사된 가중치는 분해된 두 블록으로 재구성된다.

$$W_u' = U \cdot [Trunc(\Sigma)]^{\frac{1}{2}},\ W_v' = [Trunc(\Sigma)]^{\frac{1}{2}} \cdot V^T \cdot S^{-1}$$

SVD-LLM은 단순 SVD 근사만으로는 성능 저하가 발생할 수 있기 때문에, 이를 보완하기 위해 LoRA(Low-Rank Adaptation) 모듈을 추가적으로 삽입한다. LoRA는 가중치 업데이트를 $\Delta W = BA$ 형태로 두어, $W' = W + \Delta W$로 표현한다. 이를 W_u', W_v' 각각에 적용하면 $\Delta W_u' = B_u A_u$, $\Delta W_v' = B_v A_v$이고, 보정된 근사 가중치는 다음과 같다.

$$W \approx (W_u' + \Delta W_u)(W_v' + \Delta W_v)$$

그러나 기존 LoRA는 모든 레이어와 모든 U,V에 대해 고정된 랭크 r을 사용하기 때문에, 레이어별 중요도 차이를 반영하지 못한다.

B. AdaLoRA의 적용

AdaLoRA는 각 보정항을 세 개의 파라미터 집합으로 나타낸다:

$$\Delta W_u' = P_u \Lambda_u Q_u, \qquad \Delta W_v' = P_v \Lambda_v Q_v$$

중요도 스코어는 $s(w_{ij}) = |w_{ij} \nabla_{w_{ij}} \mathcal{L}|$로 정의되며, 이는 파라미터의 크기와 학습 민감도를 동시에 반영한다. 따라서 최종 SVD-LLM + AdaLoRA 모델은 다음과 같이 표현된다.

$$W \approx (W_u' + P_u \Lambda_u Q_u)(W_v' + P_v \Lambda_v Q_v)$$

C. AdaLoRA 적용의 이론적 정당성

모델의 손실함수 $\mathcal{L}(W)$를 가중치 W에 대해 2차 테일러 전개하면, 특정 파라미터 성분 ΔW_i를 제거했을 때 손실 변화는 다음과 같이 근사된다.

$$\Delta \mathcal{L}_i \approx -\nabla_{W_i} \mathcal{L} \cdot \Delta W_i + \frac{1}{2} \Delta W_i^T H_i \Delta W_i$$

1차 근사만 고려하면, 손실 증가량의 크기는 $S_i \approx |\Delta W_i| \cdot |\nabla_{W_i} \mathcal{L}|$로 근사된다. 이는 파라미터의 크기와 손실 민감도를 동시에 반영하며, AdaLoRA가 학습 과정에서 사용하는 핵심 지표이다.

제한된 랭크/파라미터 예산 B가 주어졌을 때, 목표는 전체 손실 증가량을 최소화하는 것이다.

$$\min_{\{r_i\}} \sum_i f(S_i, r_i) \ s.t. \sum_i r_i \le B,$$

여기서 r_i는 성분 i에 할당된 랭크 수, $f(S_i, r_i)$는 성분 i개를 r_i개 만큼 유지했을 때의 손실 기여를 나타낸다. 라그랑주 승수법을 적용하면, 최적 해에서는 모든 활성 성분에 대해 한계 효용이 동일해야 한다:

$$\frac{\partial f(S_i, r_i)}{\partial r_i} = \lambda, \forall i \in \mathcal{S}.$$

이는 곧 중요도 스코어 S_i가 큰 성분에 더 많은 랭크를 배분하는 정책과 일치한다. 즉, AdaLoRA가 동적으로 랭크를 재분배하는 것은 이론적으로 전역 최적화 조건을 만족하는 전략이다.

중요도 스코어는 Signal-to-Noise Ratio(SNR)과도 연결된다. 만약 $\nabla_{W_i} \mathcal{L}$이 기댓값 μ_i, 분산 σ_i^2를 갖는 확률변수라면, $S_i \propto \frac{|\mu_i|}{\sigma_i}$, 즉, 파라미터가 가진 신호(gradient 기댓값)가 잡음(gradient 분산) 대비 얼마나 강한지를 나타낸다. 이는 AdaLoRA의 중요도 기반 선택이 단순한 휴리스틱이 아니라 통계적 근거를 가진 방법임을 의미한다. 따라서 AdaLoRA는 SVD-LLM 환경에서 고정 랭크 LoRA보다 손실 최소화 측면에서 이론적으로 손실 최소화에 유리하다.

Ⅲ. 실험 환경 및 설정

A. 데이터셋

SVD-LLM에서 LoRA 모듈을 AdaLoRA로 대체하여 다양한 압축 비율에 대하여 Wikitext-2[4] 데이터셋으로 성능을 확인한다.

B. 학습 환경

실험은 NVIDIA A100 GPU 40GB 환경에서 수행하였다. 베이스 모델로는 Llama 7B를 사용하였다. SVD-LLM에서 압축율은 20%, 40%로 설정하였으며, 학습 에폭(epoch)은 3, LoRA에서 랭크(r)는 8로 설정하였다. AdaLoRA에서 학습 에폭은 LoRA와 동일하게 3으로 설정하였으며 초기 랭크(r_{init})는 12, 최종 랭크(r_{final})는 8로 두어 랭크 축소 과정을 거치도록 하였다.

Ⅳ. 실험 결과 및 분석

Table 1: Compression ratio

Method	PPL (20%)	PPL (40%)	ARR(%) (20%)	ARR(%) (40%)	Teacher-KL (20%)	Teacher-KL (40%)	Top-10 Recall (20%)	Top-10 Recall (40%)
Llama 7B(Original)	5.68	5.68	-	-	-	-	-	-
SVD-LLM(W)	7.88	13.76	88.61	76.23	0.4903	1.0322	0.6821	0.5648
SVD-LLM	7.56	**9.40**	90.16	**84.41**	0.4150	0.6680	0.7005	0.6291
SVD-LLM(AdaLoRA)	**7.42**	9.49	**90.65**	83.52	**0.3744**	**0.6692**	**0.7145**	**0.6320**

Table 1은 각각 압축 비율 20%와 40%에서의 실험 결과를 나타낸다. 압축 비율을 20%로 설정하였을 때, AdaLoRA를 적용한 SVD-LLM은 Perplexity(PPL), ARR(Agreement Rate of Ranking), Teacher-KL(Llama 7B와의 토큰 순위 일관성 유지), Top-10 Recall 등 모든 지표에서 SVD-LLM 대비 성능 저하를 효과적으로 완화하는 것을 확인하였다. 이는 AdaLoRA의 동적 랭크 재분배가 모델 표현력을 효과적으로 보존하고 있음을 의미한다. 한편, 압축 비율을 40%로 설정하였을 때에는 PPL과 ARR 측면에서 SVD-LLM이 소폭 우위를 보였으나, AdaLoRA는 Teacher-KL 및 Top-10 Recall 지표에서 여전히 경쟁력을 유지하였다. 이는 압축률이 높아질수록 절대적인 성능 격차는 줄어들지만, AdaLoRA가 Teacher 모델의 분포와 순위를 더 잘 근사하며 언어적 지식의 정합성을 더 잘 반영한다는 점에서 의미가 있다. 즉, 단순 PPL 개선에 그치지 않고 Teacher 모델의 언어적 지식을 더 충실히 보존한다는 점에서 의의가 크다.

Ⅴ. 결론

본 연구에서는 SVD-LLM의 LoRA 모듈을 AdaLoRA로 대체하여, 구조적 저랭크 근사와 적응적 저랭크 학습을 결합하는 새로운 경량화 방법을 제안하였다. 실험 결과, AdaLoRA 적용은 기존 LoRA 대비 극단적인 압축 상황에서도 원본 모델과의 분포 정합성을 유지하는 강점을 확인하였다. 이는 AdaLoRA의 동적 랭크 조정이 레이어별 중요도를 반영하면서도, 압축으로 인한 성능 저하를 완화하는 데 기여함을 보여준다. 따라서 본 연구는 LLM 경량화 연구에서 단순 압축과 미세조정의 결합을 넘어, 높은 압축에서도 안정적인 성능을 유지할 수 있는 레이어별 적응성을 반영한 효율적 파라미터 학습 전략을 제시하는 데에 기여할 수 있을 것으로 기대한다.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00212836)

참 고 문 헌

[1] Wang, Xin, Yu Zheng, Zhongwei Wan, and Mi Zhang. "SVD-LLM : Truncation-aware singular value decomposition for large language model compression" *ICLR* , 2025.

[2] Hu, Edward J., et al "Lora: Low-rank adaptation of large language models." *ICLR* , 2022.

[3] Zhang, Qingru, et al. "Adalora: Adaptive budget allocation for parameter-efficient fine-tuning." *ICLR,* 2023.

[4] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. "Pointer sentinel mixture models." ICLR, 2017.