영차 최적화 알고리즘에 관한 연구 동향 조사

장혜지, 고현석*
한양대학교 전자공학과

j1608306@hanyang.ac.kr, *hyunsuk@hanyang.ac.kr

A Survey on research trends in Zeroth-Order Optimization Algorithms

Hyeji Jang, Hyunsuk Ko*
Hanyang Univ.

요 약

본 논문은 대규모 언어 모델(LLMs) 미세 조정(Fine-tuning) 시 발생하는 메모리 병목 문제를 완화하기 위한 영차(Zeroth-order, ZO) 최적화기법들의 최신 연구 동향을 조사한다. 대표적인 영차 최적화 기법인 MeZO를 시작으로, 그래디언트의 저랭크 구조를 활용한 LOZO, 헤시안(Hessian) 정보를 활용한 HiZOO를 중심으로 각 알고리즘의 핵심 원리와 구조적 차이를 비교 분석한다. 특히 SuperGLUE 벤치마크 실험 결과를 기반으로, 각 기법의 성능 특성과 적용상의 장단점을 정리하고, 영차 최적화 기법이 대규모 언어 모델 학습 환경에서 가지는 의미를 고찰한다.

Ⅰ. 서 론

확률적 경사하강법(SGD)이나 AdamW와 같은 1차 미분 기반(first-order, FO) 최적화 기법을 사용하여 대규모 언어 모델(LLMs)을 미세 조정하는 과정은, 주로 역전파(backward pass) 과정에서 발생하는 그래디언트 계산으로 인해 막대한 메모리 오버헤드를 요구한다. 이러한 한계를 완화하기 위해, 최근에는 영차(Zeroth-Order, ZO) 최적화 방법을 활용한 대규모 언어 모델 미세 조정에 대한 연구가 다시 주목받고 있다. 최근 제안된 영차 최적화 기법들, 예를 들어 MeZO [1]와 같은 방법은 모델의 순전파(forward pass) 결과만을 이용하여 그래디언트를 추정함으로써, 중간 활성값을 저장하거나 역전파를 수행할 필요를 제거한다. 이로 인해 그래디언트 계산에 필요한 메모리 사용량이 크게 감소하며, 결과적으로 영차 최적화 방법은 대규모 언어 모델 미세 조정에 유효한 대안으로 활용될 수 있다. 본 논문에서는 영차 최적화 기법들의 실용성을 입증한 MeZO와 이후 MeZO의 한계를 보완하기 위한 LOZO [2], HiZOO [3] 알고리즘 간 구조적 차이와 성능 특성을 비교 분석하고자 한다.

Ⅱ. 본론

고전적인 영차 최적화 기법인 ZO-SGD [4]는 손실 함수 값의 유한 차분만을 이용하여 그래디언트를 추정함으로써, 이론적으로는 순전파만으로 모델 파라미터를 갱신할 수 있다. 그러나 이러한 방식은 고차원 파라미터 공간에서 그래디언트 추정 분산이 급격히 증가하며, 모델 차원이 커질수록 수렴 속도가 저하되는 문제가 존재한다. 이로 인해 기존 영차 최적화 기법은 주로 적대적 예제 생성이나 입력 임베딩 조정과 같은 제한적인 문제에만 활용되어 왔으며, 수십억 파라미터 규모의 대규모 언어 모델을 직접 최적화하는 데에는 실질적인 한계가 있었다.

MeZO(Memory-efficient zeroth-order optimizer)는 기존 영차 최적화의 한계를 극복하기 위해, 고전적인 ZO-SGD에서 파라미터 섭동 및 복원을 in-place 방식으로 수행함으로써 역전파와 중간 활성값 저장을 제거하고, 추론과 유사한 수준의 메모리 사용량으로 모델 미세 조정을 가능하게 한다. 이를 통해 역전파 기반 미세 조정 대비 12배 이상 낮은 메모리 사용량으로도 대규모 모델 학습이 가능함을 보였다.

그러나 MeZO의 그래디언트 추정 방식은 구조적 한계를 가진다. 대규모 언어 모델의 1차 미분 그래디언트는 소수의 주요 방향에 정보가 집중된 저랭크 구조를 가지는 반면, MeZO의 그래디언트 추정치는 무작위 섭동에 의해 대부분의 파라미터 방향에 균등하게 분산되는 경향을 보인다. 이러한 불일치는 탐색 효율 저하로 이어질 수 있다.

LOZO(Low-rank ZO-SGD)는 저차원 부분공간에서의 섭동을 반복적으로 활용하는 영차 그래디언트 추정 방식을 제안하였다. 특히 동일한 저차원 부분공간을 유지하는 lazy sampling 전략을 통해, 대규모 언어 모델의 그래디언트의 구조를 보다 정밀하게 근사하고, MeZO 대비 향상된 성능과 안정적인 수렴 특성을 보였다.

MeZO는 손실 지형의 곡률을 고려하지 못해 특정 방향에서는 과도한 갱신이, 다른 방향에서는 정체가 발생할 수 있다. HiZOO(Hessian informed Zeroth-Order Optimizer)는 추가적인 순전파를 통해 대각 형태의 곡률 정보를 추정하고, 이를 이용해 영차 그래디언트를 사전 보정하는 방식을 제안하였다. 다만, HiZOO는 곡률 추정을 위한 추가 연산 및 메모리 비용이 요구된다는 한계가 존재한다.

Task	SST-2	RTE	WSC	WiC
13B Zero-shot	58.8	59.6	38.5	55.0
13B MeZO	91.4	66.1	63.5	59.4
13B LOZO	91.7	**70.4**	63.5	**60.8**
13B HiZOO	**92.1**	68.2	**65.4**	59.4
13B FT	91.8	70.9	84.1	76.9
30B Zero-shot	56.7	52.0	38.5	50.2
30B MeZO	90.6	66.4	63.5	56.3
30B LOZO	**92.8**	65.3	**64.4**	**57.2**
30B HiZOO	90.3	**69.3**	63.5	53.4

표 1. SuperGLUE 데이터셋의 4개 벤치마크 과제에서 OPT-13B및 OPT-30B 모델의 성능 비교 (모든 수치는 5회 반복 실험에 대한 평균 정확도). FT는 Adam 기반 전체 파라미터 미세 조정을 의미하며, FT를 제외한 방법 중 최고 성능을 굵은 글씨로 표시하였다.

표 1은 SuperGLUE 데이터셋의 네 가지 벤치마크 과제(SST-2, RTE, WSC, WiC)에 대해 OPT-13B 및 OPT-30 모델을 대상으로 영차 최적화 기법들의 성능을 비교한 결과를 나타낸다. Zero-shot 결과와 비교할 때, MeZO, LOZO, HiZOO는 두 모델 규모 모두에서 전반적으로 큰 성능 향상을 보이며, 영차 최적화가 대규모 언어 모델 미세 조정에 실질적으로 효과적임을 확인할 수 있다.

OPT-13B 모델 기준으로, MeZO는 모든 과제에서 zero-shot 대비 유의미한 성능 개선을 달성한다. LOZO는 대부분의 과제에서 MeZO 대비 추가적인 성능 향상을 보이는데, 이는 대규모 언어 모델의 그래디언트가 본질적으로 저차원 구조를 가진다는 가정이 실제 성능 향상으로 이어질 수 있음을 시사한다. HiZOO 역시 OPT-13B 모델에서 WSC 과제에서 65.4%로 가장 높은 성능을 기록하고, 거의 모든 과제에서 MeZO 대비 추가적인 성능 향상을 달성하여, 곡률 정보를 활용한 사전 보정이 효과적임을 확인할 수 있다.

OPT-30B 모델에서도 유사한 경향이 관찰되며, 모델 규모가 증가하더라도 영차 최적화 기법이 안정적으로 동작함을 보인다. LOZO는 RTE를 제외한 대부분의 과제에서 가장 높은 정확도를 달성하여, 곡률 정보를 활용한 사전 보정이 특히 해당 벤치마크 과제에서 효과적임을 확인할 수 있다. OPT-30B에 대한 FT 결과는 계산 비용 문제로 인해 본 실험에 포함되지 않았다.

종합적으로 보면, MeZO는 실용적인 기준선 역할을 수행하며, LOZO는 저랭크 구조를 활용함으로써 성능과 안정성 측면에서 가장 균형 잡힌 개선을 보인다. HiZOO는 곡률 정보를 반영하여 일부 과제에서 가장 우수한 성능을 보이지만, 추가 연산 비용을 수반하며, OPT-30B 모델과 실험한 벤치마크 과제에서 MeZO보다 낮은 성능을 보인다는 점에서 적용 환경에 따른 고려가 필요하다.

Ⅲ. 결론

본 논문에서는 대규모 언어 모델 미세 조정 과정에서의 메모리 병목 문제를 완화하기 위한 영차 최적화 기법들의 연구 동향을 조사하고, MeZO, LOZO, HiZOO를 중심으로 구조적 특성과 성능을 비교 분석하였다. SuperGLUE 벤치마크 실험 결과, 영차 최적화 기법들이 순전파만을 활용함에도 불구하고 zero-shot 대비 유의미한 성능 향상을 달성함을 확인하였다. 또한 저랭크 구조 활용 여부와 곡률 정보 반영 방식에 따라 각 기법이 서로 다른 성능 특성과 연산 비용을 보인다는 점을 정리하였다. 본 논문의 분석은 영차 최적화 기법들의 특성과 적용 가능성을 이해하기 위한 기초 자료로 활용될 수 있을 것으로 기대된다.

참 고 문 헌

[1] Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D. Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. In Neural Information Processing Systems, 2023.

[2] Yiming Chen, yuan zhang, Liyuan Cao, Kun Yuan, and Zaiwen Wen. Enhancing zeroth-order fine-tuning for language models with low-rank structures. In International Conference on Learning Representations, 2025.

[3] Yanjun Zhao, Sizhe Dang, Haishan Ye, Guang Dai, Yi Qian, and Ivor Tsang. Second-order fine-tuning without pain for LLMs: A hessian informed zeroth-order optimizer. In International Conference on Learning Representations, 2025.

[4] J.C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. IEEE Transactions on Automatic Control, 37(3):332- 341, 1992.