NISQ 환경에서의 LLM 파인튜닝을 위한 양자 알고리즘 프레임워크

노지민, 이호영*, 박수현**, 김중헌
고려대학교, *단국대학교, **숙명대학교
(emilyjroh, joongheon)@korea.ac.kr, *hyeonglee@dankook.ac.kr, **soohyun.park@sookmyung.ac.kr

A Quantum Algorithmic Framework for Fine-Tuning Large Language Models in the NISQ Era

Emily Jimin Roh, Hoyeong Lee*, Soohyun Park**, Joongheon Kim
Korea Univ., *Dankook Univ., **Sookmyung Women's Univ.

요 약

본 논문은 NISQ(noisy intermediate-scale quantum) 환경에서 대규모 언어 모델(large language models, LLM)의 파인튜닝을 가능하게 하는 양자 알고리즘 프레임워크를 제안한다. 제안된 프레임워크는 양자 - 고전 하이브리드 학습 구조를 기반으로 하여, 제한된 큐비트 수와 잡음이 존재하는 현실적 양자 하드웨어 제약을 고려한다. 특히, 파라미터화된 양자 회로를 활용하여 LLM 파인튜닝 과정의 핵심 연산을 효율적으로 근사하는 방법을 제시한다. 이론적 분석을 통해 알고리즘의 수렴 특성과 계산 복잡도를 논의하며, NISQ 환경에서의 실현 가능성을 평가한다. 실험 결과는 제안된 접근법이 기존 고전적 파인튜닝 기법 대비 특정 조건에서 경쟁력 있는 성능과 잠재적 이점을 가짐을 보여준다.

Ⅰ. 서 론

대규모 언어 모델(Large language models, LLMs)은 자연어 처리 전반에서 뛰어난 성능을 보이며 다양한 응용 분야의 핵심 기술로 자리 잡았다. 그러나 이러한 모델의 파인튜닝은 막대한 계산 자원과 에너지를 요구하며, 모델 규모가 커질수록 학습 비용과 환경적 부담이 급격히 증가하는 한계를 지닌다. 구체적으로, 그림 1은 스마트 헬스케어, 스마트 빌딩, 재난 대응 등 도메인 특화 응용에서 대규모 언어 모델을 그대로 활용하기에는 계산 자원과 비용 제약이 크다는 점을 보여준다. 이러한 한계를 극복하기 위해 전체 모델을 재학습하는 대신, 파라미터 효율적 파인튜닝(parameter-efficient fine-tuning, PEFT)을 통해 핵심 파라미터만 조정하는 접근이 필요함을 강조한다. 즉, 경량화된 LLM 파인튜닝은 제한된 자원 환경에서도 실용적인 도메인 적응을 가능하게 한다. 이에 따라 기존 고전적 계산 패러다임을 보완하거나 대체할 수 있는 새로운 계산 프레임워크에 대한 연구 필요성이 대두되고 있다.

양자 컴퓨팅은 특정 계산 문제에서 고전적 방법 대비 잠재적인 계산 우위를 제공할 수 있는 기술로 주목받아 왔다. 특히 최근에는 완전한 오류 보정 양자 컴퓨터 이전 단계인 NISQ(noisy intermediate-scale quantum) 환경에서 실현 가능한 양자 알고리즘과 응용에 대한 연구가 활발히 진행되고 있다. NISQ 환경은 제한된 큐비트 수와 잡음, 짧은 코히런스 시간이라는 제약을 가지지만, 하이브리드 양자 - 고전 알고리즘을 통해 실질적인 이점을 탐색할 수 있는 현실적인 연구 무대이기도 하다.

이러한 배경 속에서, 양자 머신러닝과 변분 양자 알고리즘(variational quantum algorithms)은 NISQ 환경에 적합한 접근법으로 제안되어 왔다. 특히 파라미터화된 양자 회로는 고전적 최적화 기법과 결합되어 복잡한 함수 근사 및 표현 학습에 활용될 수 있다 [1]. 그러나 기존 연구들은 주로 소규모 모델이나 특정 머신러닝 태스크에 국한되어 있으며, LLM 파인튜닝과 같은 대규모 학습 문제에 대한 체계적인 양자 알고리즘 프레임워크는 아직 충분히 탐구되지 않았다.

본 논문은 이러한 공백을 메우기 위해 NISQ 환경에서 LLM 파인튜닝을

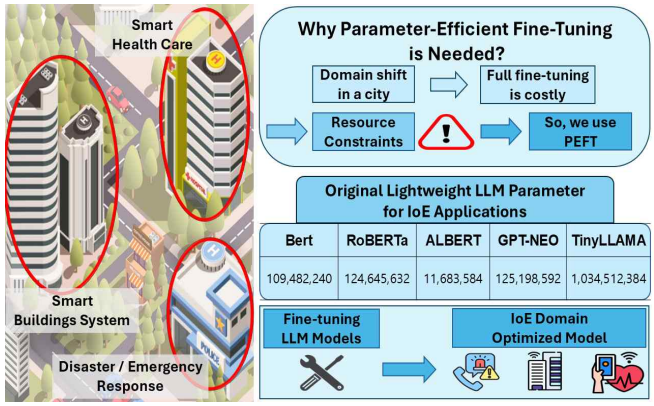

그림 1. LLM 파인튜닝의 배경 및 동기

수행하기 위한 양자 알고리즘 프레임워크를 제안한다. 제안하는 접근법은 LLM 파인튜닝 과정의 핵심 연산을 양자 회로로 부분적으로 대체하거나 근사하는 하이브리드 구조를 채택함으로써, 현실적인 양자 하드웨어 제약을 고려한다 [2]. 또한 알고리즘의 이론적 특성과 계산 복잡도를 분석하고, 실험적 평가를 통해 NISQ 환경에서의 적용 가능성과 잠재적 장점을 검증한다. 본 연구는 양자 컴퓨팅과 대규모 언어 모델 연구를 연결하는 초기 단계의 시도로서, 향후 양자 강화 인공지능 연구를 위한 기초적 프레임워크를 제공하는 데 그 의의가 있다.

Ⅱ. Quantum Algorithm for Parameter-Efficient Fine-Tuning

본 절에서는 사전학습된 대규모 언어 모델의 파라미터 효율적 파인튜닝을 위해 제안하는 양자 알고리즘 구조를 설명한다 [3]. 제안하는 접근법은 기존 Transformer 기반 LLM의 구조를 유지하면서, attention 모듈 내 선형 변환 과정에 양자 서브레이어인 quantum sublayer(QSL)를 삽입하는 방식으로 설계된다. 이를 통해 전체 모델 파라미터를 재학습하지 않고도 제한된 양자 자원을 활용한 효율적인 도메인 적응을 가능하게 한다.

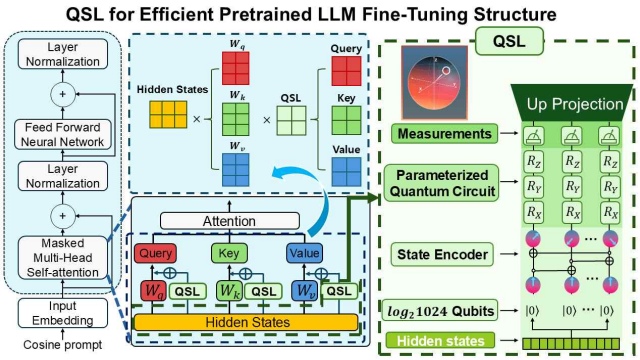

그림 2. PEFT를 위한 QSL 구조도

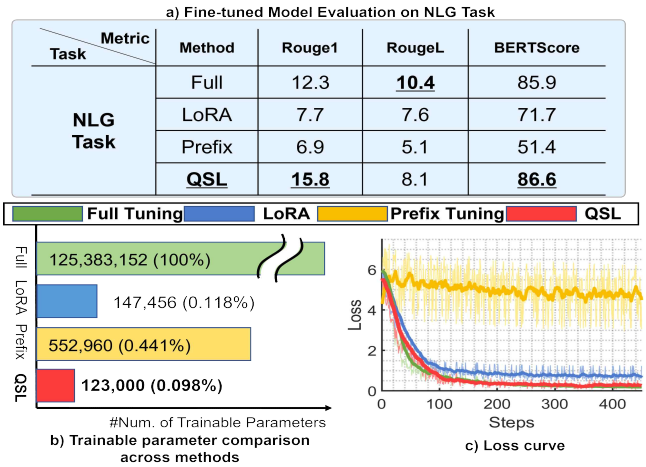

그림 3. QSL 기반 NLG Task에서의 성능평가 결과

QSL의 삽입 구조는 그림 2에서와 같이, QSL은 Transformer 블록의 masked multi-head self-attention 내부에서 query, key, value 생성을 위한 선형 변환 행렬과 결합된다. 기존의 고전적 선형 변환 대신, 은닉 상태(hidden states)의 일부 또는 저차원 투영된 표현이 QSL을 통해 처리되며, 해당 결과가 각 attention head의 계산에 반영된다. 이 과정에서 원래의 attention 연산 흐름은 유지되므로, 모델의 안정성과 호환성이 보장된다. QSL 내부에서는 입력 은닉 상태가 상태 인코더(state encoder)를 통해 양자 상태로 변환된다. 차원 축소된 은닉 표현은 $\log_2 d$ 개의 큐비트에 인코딩된다. 이후 파라미터화된 양자 회로가 적용되며, 회로는 단일 큐비트 회전 게이트(RY)와 얽힘 게이트(CNOT)로 구성된다. 이러한 구조는 NISQ 환경에서 구현 가능하도록 회로 깊이를 제한하면서도 충분한 표현력을 확보하도록 설계된다.

양자 회로의 출력은 측정(measurement)을 통해 고전적 값으로 변환되며, 이 값들은 업-프로젝션(up-projection) 단계를 거쳐 원래의 hidden state 차원으로 복원된다. 복원된 출력은 기존 attention 연산에 결합되어, 최종적으로 self-attention 결과에 반영된다. 이때 학습 가능한 파라미터는 주로 양자 회로의 회전 각도에 국한되므로, 전체 학습 파라미터 수는 기존 LLM 파인튜닝 대비 현저히 감소한다.

Ⅲ. 성능 평가

그림 3은 Full Fine-Tuning, LoRA, Prefix Tuning, 그리고 제안된 QSL 기반 방법 간의 성능과 파라미터 효율성을 종합적으로 비교한 결과를 제시한다. QSL은 ROUGE, BERTScore 등 주요 자연어 생성 평가 지표에서 기존 PEFT 기법 대비 가장 우수한 성능을 기록하거나 Full Fine-Tuning에 근접한 결과를 달성하였다. 특히 BERTScore 기준으로는 QSL이 다른 모든 PEFT 방법을 상회하는 성능을 보이며 의미적 일관성 측면에서 강점을 나타낸다. 반면 LoRA와 Prefix Tuning은 파라미터 수를 크게 줄이는 장점이 있으나, 전반적인 성능 저하가 비교적 뚜렷하게 나타난다. 또한, ROUGE-1과 ROUGE-L 기준에서 QSL은 기존 LoRA 및 Prefix Tuning 대비 높은 점수를 기록하여, 단어 수준의 정보 보존뿐만 아니라 문장 수준의 구조적 일관성 측면에서도 유의미한 성능을 확인할 수 있다. 파라미터 비교 결과에서 QSL은 전체 모델 파라미터의 극히 일부만을 학습함에도 불구하고, Full Fine-Tuning 대비 효율적인 성능 - 파라미터 균형을 달성함을 확인할 수 있다. 학습 손실 곡선 분석에서도 QSL은 초기 수렴 속도가 빠르고 안정적인 감소 추세를 보여 학습 안정성이 우수함을 시사한다. 종합적으로, 실험 결과는 제안된 QSL 기반 양자 PEFT 기법이 제한된 계산 자원 환경에서도 실용적인 성능을 제공할 수 있는 유망한 대안임을 보여준다.

Ⅳ. 결론

본 논문에서는 NISQ 환경에서의 대규모 언어 모델 파인튜닝을 위해 양자 알고리즘 기반의 파라미터 효율적 학습 프레임워크를 제안하였다. 제안된 QSL 구조는 Transformer의 attention 모듈에 QSL을 통합함으로써, 전체 모델 구조를 유지하면서도 학습 파라미터 수를 효과적으로 감소시킨다. 이론적 분석과 실험 결과를 통하여, QSL 기반 방법이 기존 PEFT 기법 대비 우수하거나 경쟁력 있는 성능을 달성함을 확인하였다. 특히 제한된 학습 자원과 잡음이 존재하는 NISQ 환경에서도, 적은 큐비트 기반으로 안정적인 수렴 특성을 보였다.

본 연구는 양자 컴퓨팅을 LLM 파인튜닝이라는 대규모 학습 문제에 적용할 수 있음을 실증적으로 보여주는 초기 단계의 시도로서 의의가 있다. 향후 연구에서는 더 깊은 양자 회로 설계, 오류 완화 기법의 통합, 그리고 다양한 LLM 아키텍처로의 확장이 필요하다. 또한 실제 양자 하드웨어 상에서의 실험을 통해 실용적 성능을 검증하는 것이 중요한 과제로 남아 있다. 이러한 후속 연구를 통해 양자 강화 언어 모델 학습의 가능성이 더욱 구체화될 것으로 기대된다.

ACKNOWLEDGMENT

이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(RS-2024-00439803, SW컴퓨팅산업원천기술개발사업 (SW스타랩)); 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2025-00561377). 본 논문의 교신저자는 박수현, 김중헌임.

참 고 문 헌

[1] G. Brassard, I. Chuang, S. Lloyd, and C. Monroe, "Quantum Computing," *The National Academy of Sciences,* vol. 95, no. 19, pp. 11032-11033, September 1998.

[2] E. J. Roh, H. Baek, D. Kim, and J. Kim, "Fast Quantum Convolutional Neural Networks for Low-Complexity Object Detection in Autonomous Driving Applications", *IEEE Transactions on Mobile Computing,* vol.24, no.2, pp. 1031-1042, February 2025.

[3] E. J. Roh and J. Kim, "Quantum-amplitude embedded adaptation for parameter-efficient fine-tuning in large language models," in *Proc. ACM International Conference on Information and Knowledge Management (CIKM),* Seoul, Korea, November 2025.