LLM 기반 의료진 보조 처방 시스템

김유나, 이성진, 정은성
홍익대학교 소프트웨어융합학과
e- mail : rladbsk2805@naver.com, sjlee7485@naver.com, ejung@hongik.ac.kr

A LLM- Based Clinical Decision Support System for Medical Prescription Assistance

YuNa Kim and SeongJin Lee and Eun-Sung Jung
Hongik University

Abstract

This study proposes an LLLM- based prescription support system that recommends appropriate drug lists based on patient diagnoses to assist clinicians amid the ongoing shortage of medical staff. By applying FA- LoRA, we reduced the number of trainable parameters by approximately 93.5% while maintaining performance, and enhanced efficiency through FlashAttention. Using Llama3.1- 8B- Instruct in a single- node setup, we evaluated the model via human assessment and GPT- 4, confirming its accuracy and appropriateness. Future work includes extending the system with federated learning and differential privacy to protect sensitive medical data.

I. 서론

최근 의료진 부족으로 응급 상황 대응과 필수 진료 서비스 제공에 어려움이 발생하고 있으며, 의료 시스템의 취약성과 인력 불균형 문제가 지속적으로 제기되고 있다. 또한 의료 취약 지역에는 공공의료 인력 부족이 장기간 지속되어 필수 의료 서비스 제공에 영향을 미치고 있다는 분석도 있으며, 코로나19 펜데믹 기간동안 의료진의 과중한 업무 부담으로 인해 의료 시스템의 취약성이 더욱 부각되고 있다. 이러한 상황은 의료 자원의 효율적인 분배와 병원 간 협력의 중요성을 다시 한번 일깨워 준다.

본 논문은 위와 같은 문제에 의료진 부담을 줄여줄 수 있는 자동화된 의료 시스템의 필요성을 인식했다. 이에 환자 진단에 알맞는 처방 약물 리스트를 예측하여 의료진 부담을 줄여주는 AI 시스템을 개발한다. 추가적으로 LLM의 최적화를 위해 FA- LoRA[1]와 FlashAttention2[2]을 적용하는 방법을 제안한다.

본 연구의 주요 기여사항은 다음과 같다.

1. 의료진을 보조하는 AI 시스템 구축 : 환자의 진단 정보를 입력 받아 적절한 처방 약물 리스트를 예측함으로써, 의료진의 진단 업무를 보조하는 AI 시스템을 제안한다.
2. FA- LoRA 및 FlashAttention2을 결합한 LLLM 최적화 구조 : 기존 모델 대비 학습 가능한 파라미터를 약 98.5% 감소시키는 FA- LoRA와, 연산 병목을 해결하고 GPU 활용 효율을 극대화하는 FlashAttention2를 결합하여 최적화된 LLM 구조를 구현하였다.

II. 본론

2.1 관련 연구

의료 AI 분야에서는 환자 데이터를 기반으로 다양한 예측을 수행하는 연구가 활발히 진행 중이다.

예를 들어, GIST 연구팀의 PANCDR 모델은 환자의 유전자 발현 정보와 약물 그래프를 기반으로 암 환자의 약물 반응을 예측하고[3], ImpriMed는 환자 개인에 적합한 항암제를 추천하는 정밀 의료 AI를 상용화한다[4]. 아주대 의료원은 감염질환 환자의 진단 및 약물 기록을 활용해 항생제 내성을 예측하며[5], KAISR DeepDDI는 약물 간 또는 약물-음식 간 상호작용을 예측하는 기술을 개발하였다[6].

이러한 연구들은 주로 특정 약물의 반응, 내성, 상호작용에 대한 예측을 중심으로 하고 있으며, 진단 정보를 기반으로 환자에게 적절한 약물 리스트를 예측하는 방식과는 다르다. 그러나 일부 병원 기반 시스템들은 통계 기반 임상 의사결정 지원(CDSS)를 통해 환자별 처방 패턴을 분석하기도하나, 이는 통계적으로 추적된 처방 빈도나 경향을 활용하는 것이지[7], LLM을 통해 진단과 처방 간을 학습하는 방식은 아니다.

본 연구는 기존 연구들과 달리, 의료진에게 보조적인 시스템으로 환자의 진단 정보만을 기반으로 LLM을 통해 처방 약물 리스트 전체를 생성하는 방식에 초점을 맞춘다. 이는 AI가 처방 제안을 먼저 제시하고 의료진이 검토하는 보조 구조를 구현하고자 하는 점에서 기존 접근 방식과 차별화 된다.

2.2 LoRA(Low-Rank Adaptation)

일반적으로 컴퓨팅 자원이 제한된 환경에서 LLM의 모든 파라미터를 미세 조정하는 데 한계가 있다. 이에 효율적인 파라미터 미세조정(PEFT, Parameter-Efficient Fine-Tuning)을 위해 제안된 기법 중 하나인 LoRA를 적용하여, 계산 비용을 크게 절감하면서도 모델의 효율적인 미세 조정이 가능하다.

LoRA[8]는 전체 파라미터 대신 일부 저차원 행렬만 학습함으로써 연산 효율을 높이고 메모리 사용량을 줄이는 파라미터 효율적 미세조정 기법이다.

본 연구는 Computing 자원이 제한된 환경에서도 대형 LLM 모델의 학습이 가능하도록 하는 방안을 모색한다. 하지만 더 효율적으로 파라미터 수를 줄이기 위해, 해당 연구에서는 두 저차원 행렬 중 A 행렬을 고정(동결)하고 B행렬만 학습 가능한 파라미터로 유지하는 방식인 FA-LoRA(Frozen-A Low Rank Adaptation)[1]를 제안한다. FA-LoRA는 더 작은 파라미터 수로 동일하거나 유사한 성능을 달성하며, 메모리 사용량 및 연산량을 줄여 더 가벼운 학습과 추론이 가능하다. 또한 특정 계층(Attention head, FFN 등)만 부분적으로 미세조정해 전체 모델의 안정성과 효율성을 보장한다는 장점이 존재한다.

2.3 FlashAttention

Transformer 모델에서 Self-Attention 연산은 기본적으로 입력 시퀀스의 모든 토큰 간의 관계를 계산하며, 시퀀스 길이가 N일 경우, N이 증가할 수록 계산량과 메모리 사용량이 $O(N^2)$ 으로 증가하는 문제가 발생한다. 이는 시퀀스 길이가 두배가 되면서, 연산량과 메모리 사용량이 네 배로 증가하며, 학습 과 추론 속도 저하 및 병목 현상을 초래한다. 이런 문제를 해결하고자 FlashAttention 기법을 도입한다.

FlashAttention1[9]은 Self-Attention 연산을 블록 단위로 수행하며, GPU의 글로벌 메모리 접근을 최소화하여 보다 빠른 공유 메모리를 적극 활용하여, 데이터 이동을 최소화 한다. 그러나 여전히 최적화된 행렬 곱셈(GEMM) 연산 만큼 빠르지 않다는 한계가 존재하여, 본 연구에서는 FlashAttention2[2]를 적용한다.

FlashAttention2는 1을 기반으로 다음과 같은 더 높은 병렬성과 연산 최적화로 뛰어난 성능을 제공한다.

1. Non-Matmul FLOPs 최적화 : Softmax, Normalization 등의 Non-Matmul 연산을 최적화하여, 불필요한 연산과 메모리 접근을 줄인다.
2. GPU 병렬성 강화 : 여러 개의 GPU 스레드 블록을 활용하여 연산을 분산함으로써, 병렬 처리 능력을 크게 강화한다. 이는 연산 병목을 완화하고 처리량을 향상시킨다.
3. 공유 메모리 최적화 : 중간 계산 결과를 공유 메모리에 저장하여 메모리 대역폭 병목을 줄이고, 연산 경로를 단축시킨다.

2.4 Backbone Model

예측 모델을 구축하기 위해 Backbone 모델로 Meta의 Llama3.1-8B-Instruct[10] 모델을 채택하였다. 해당 모델은 80억 개의 비교적 작은 파라미터 수 대비 높은 성능을 제공하며, FlashAttention2[2]와 호환을 지원하여 높은 유연성을 갖춘 구조를 지닌다. 또한 instruction 기반 미세 조정에 최적화되어 특정 하위 작업에 적합한 학습을 쉽게 수행할 수 있다. 사전 학습된 Llama3.1-8B-Instruct의 가중치를 동결하여 학습 가능하지 않도록 설정하고, 미세조정 방식으로 FA-LoRA[1]를 도입하였다. 기존 LoRA의 최적화 방식과 동일하게 W_q (쿼리), W_v (값) 레이어에 어댑터를 적용하였다. 이러한 기존 LoRA 대신 FA-LoRA 방식은 학습

가능한 파라미터 수를 기존 80억개에서 약 524만개로 대폭 감소시킨다. 추가적으로 FlashAttention2를 적용함으로써 기존 Self-Attention에 비해 연산량 감소로 속도 향상과 메모리 절감을 하였다. FA-LoRA를 적용한 후 데이터 타입에 따른 메모리 절감 효과를 분석한 결과, FP32(32-bit Floating Point) 자료형 적용 시 약 32GB 메모리 절약 가능하고, FP16(16-bit Floating Point) 자료형 적용 시 약 16GB 메모리 절약 가능하다. 이러한 최적화는 메모리 사용량을 최소화하는데 매우 큰 기여를 한다.

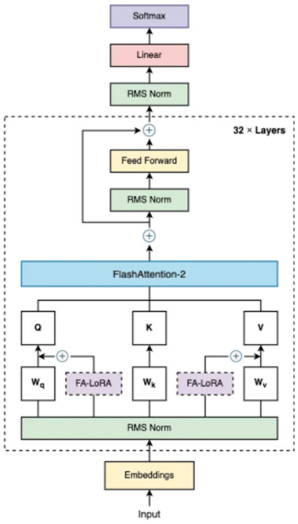

그림 1. 모델 구조

Ⅲ. 구현

3.1 데이터셋 전처리

본 연구에서는 MIMIC-IV 데이터셋을[11] 활용하였다. MIMIC-IV는 미국 Beth Israel Deaconess Medical Center 에서 2008년부터 2019년 까지 수집된 대규모 의료 데이터셋으로, 중환자실 환자 기록을 포함하고 있으며, 생체 신호, 투여된 약물, 입원 및 퇴원 정보 등 다양한 임상 데이터를 제공한다.

진단에 알맞는 약물 추천 모델을 구축하기 위해 사용할 파일 선택하여 다음과 같은 데이터 처리 단계를 거쳤다.

1. 데이터 병합 : 환자 ID 및 입원 ID를 기준으로 사용할 환자의 신체 정보, 진단 정보, 처방 약물 데이터를 병합하였다. 단일 입원 기간 동안 여러 진단과 다수의 약물이 함께 기록되어, 각 진단에 맞는 약물을 1:1로 매칭하는 방식엔 어려움이 존재하였다. 대신 해당 입원 기간 전체의 진단 목록과 약물 목록을 하나의 샘플로 구성하여 학습에 활용하였다.
2. 데이터 전처리 : 모델 학습을 위한 일관된 데이터셋을 구성하기 위해 환자의 개인정보와 진단을 결측치 처리, 중복 제거, 이상치 처리 등의 전처리 작업을 수행하였다.
3. 입력 데이터 변환 : 최종적으로 정제된 데이터를 모델에 입력할 수 있도록 모델 프롬프트 형식에 맞춰 변환하였다.

3.2 평가 방식

LLM의 성능을 평가하기 위해서는 입력에 따른 출력의 변화와 그 결과의 정확성을 분석하는 것이 중요하다. 특히 prompt와 context가 평가에서 중요한 구성요소이며, 다양한 성능 지표를 활용한 다각적 평가가 필요하다. 본 연구에서는 Human Evaluation과 LLM Evaluation 두가지를 정확도와 적합성으로 병행하여 보다 객관적이고 효율적인 평가를 진행한다.

Human Evaluation은 사람이 직접 평가 기준을 정하여 분석하는 방식이다. 본 연구에서는 테스트 데이터셋의 진단에 따른 약물 리스트를 정답으로 가정하고 평가하였다. 평가 기준은 첫째, 원하는 형식에 맞게 응답을 생성하였는가? 둘째, 출력된 약물 리스트가 정답의 리스트와 비교하여 얼마나 유사하게 판단하였는가? 두가지로 기준을 세웠다.

각 샘플에 대해 출력된 약물들 중 정답과 일치하는 약물의 비율을 계산하고, 이를 평균하여 전체 정확도를 산출하였다. 이는 높은 객관성을 보일 수 있으나, 시간이 오래 걸리며, 의료 전문가가 참여해야 하는 경우 비용적인 문제가 발생한다. 또한 적절한 약물을 제시했더라도 정답 데이터셋에 없는 약물일 경우 부정확하게 평가되어 정확도가 과소평가 되는 한계가 존재한다. 이러한 한계를 보완하고자 기존 평가 방식에 LLM Evaluation 방식을 병행한다.

LLM Evaluation에 활용할 LLM으로는 Hugging Face 에서 제공하는 의료 질문 답변 Leaderboard에서 약 83%로 높은 성능을 보인 GPT-4를[12] 활용한다. 또한, 현진 의사들과 동일한 환자 사례를 기반으로 진단을 수행한 연구에서, 진단의 정확도와 근거 제시 측면에서 의사보다 높은 점수를 기록하는 사례가 존재하여, 최종적으로 GPT-4를 선정하였다.

GPT-4를 활용하여 '각 샘플별로 적절한 약물 처방인지 검토하고, 적절한 처방 개수, 적절하지 않은 처방 개수로 분류하여 알려줘' 라고 GPT-4가 각

샘플의 처방 리스트를 면밀하게 평가할 수 있도록 프롬프트를 구성하였다. 이후 GPT-4의 응답을 바탕으로, 각 샘플의 전체 처방 중 적절하다고 판단된 약물의 비율을 계산하고, 이를 평균하여 전체 적합성 점수를 산출하였다.

최종적으로 본 연구는 Human Evaluation, LLM Evaluation 을 병행하여, 모델의 정확성과 의료적 타당성인 적합성을 함께 검증한다.

3.3 실험환경

RTX 6000 Ada 48GB를 사용하였으며, 모델은 학습률은 1e-3, 배치 크기는 4, FA-LoRA의 alpha는 16, rank값은 16,32, max_length는 2048로 학습 진행 하였으며, 옵티마이저는 기존 SGD를 활용하였다.

3.4 실험 결과

표2 는 FA-LoRA의 rank 값을 16,32로 설정하고, 학습 steps을 1950, 2925, 3900, 4870으로 달리하며 측정한 결과를 보여준다.

rank	steps	출력률	정확도	적합성
16	1950	60%	13.0%	81.1%
	2925	45%	22.9%	79.9%
	3900	55%	19.7%	81.1%
	4870	55%	15.7%	85.5%
32	1950	55%	20.7%	98.2%
	2925	55%	18.5%	87.5%
	3900	50%	17.4%	85.1%
	4870	55%	21.6%	91.0%

표2. 모델의 rank 별 정확도 비교한 표

실험 결과, rank값에 관계없이 출력률은 약 55%로 응답을 출력하였다. 정확도 측면에서는 rank16 에서 steps 2925일 때가 가장 높은 정확도를 보였지만, 평균적으로는 rank32에서 더 높은 정확도가 나타났다. 이는 rank32가 전체적으로 더 안정적이고 좋은 성능을 보이고 있음을 의미하며, 적합성 역시 rank 32에서 더 높은 값을 기록했다.

정확도와 적합성 간의 차이가 상당히 두드러진 진다. 이는 동일 성분의 다양한 약물을 추천하는 경향이 있으나, 정답 데이터셋은 하나의 약물만 처방했기 때문이다. 예를 들어, 통증을 완화시키기 위한 동일한 성분의 약물을 여러 개 추천하는 경우가 다수 존재한다. 하지만 실제 정답 데이터셋에는 그 중 하나의 약물만 처방되었기 때문에, 모델이 추천한 여러 약물 중 하나만 정답으로 간주되며, 이로 인해 정확도가 낮게 측정되는 경향이 있다.

또한, 현재 사용 중인 Llama3.1-8B-Instruct 모델이 의료 지식에 대한 사전 학습이 어느 정도 되어 있어, 학습되지 않은 약물임에도 진단에 적합하다고 판단되면 이를 추천하는 경향이 있다고 판단된다. 이는 모델이 실제로 진단에 적절한 약물을 추천하더라도, 정답 데이터셋에는 포함되지 않게 때문에 정확도가 낮게 나오는 원인 중 하나이다.

결론적으로 정확도도 중요한 부분이지만 본 프로젝트에서 적합성이 필요한 이유는, 모델이 의료진을 대체하는 것이 아닌 보조적인 역할을 한다는 점이다. 모델이 제공하는 진단에 적합성 높은 약물 추천은 의료진이 최종 결정을 내리기 위한 참고자료로 활용될 수 있다. 따라서, 모델이 제공하는 약물이 진단에 적합하지 않다면 그 추천이 무효화되겠지만, 모델이 제공하는 약물이 적합성 기준에서 좋은 평가를 받았다면, 의료진이 그 추천을 참고하여 더 나은 판단을 내릴 수 있도록 돕는 보조적인 역할을 하게 될 것으로 기대한다.

Ⅳ. 결론 및 향후 연구 방향

본 연구는 의료진 부족 문제를 보완하기 위한 방안으로, 환자의 진단 정보를 기반으로 적절한 약물 리스트를 추전하는 LLM 기반 의료진 보조 시스템을 제안한다. 이에 모델 최적화를 위해 FA-LoRA 기법을 적용하여 기존 80억 파라미터에서 기존 LoRA를 적용하여 약 1,360만 개(전체 대비 약 16.9%)였던 학습 가능한 파라미터 수가, FA-LoRA 적용 시 약 524만 개(전체 대비 약 6.5%)로 감소함을 확인하였다. 결론적으로, 학습 가능한 파라미터 수는 전체의 약 6.5% 수준으로 줄이면서도 성능을 유지하였으며, FlashAttention2를 추가 적용함으로써 연산량과 메모리 사용량을 크게 절감하였다.

또한 Llama3.1-8B-Instruct 모델을 기반으로 실험을 진행 하였으며, Human Evaluation과 LLM Evaluation을 기반 평가를 통해 정확도와 적합성 측면에서 데이터셋의 한계가 있음에도 유의미한 성능을 확인하였다. 특히, 적합성 평가에서 높은 결과를 확인함으로써, 본 시스템이 실제 의료 상황에서 의료진의 의사결정을 효과적으로 보조할 수 있는 가능성을 입증하였다. 이러한 결과는 FA-LoRA와 FlashAttention의 결합으로 제한된 자원 환경에서도 고성능 약물 추천 시스템을 구축하는 데 효과적 일거라는 기대를 한다.

향후 연구에서는 단일 노드 기반 실험을 넘어, 다기관 분산 환경에서 적용 가능한 연합학습 구조를 설계하여, 의료 데이터의 다양성과 실제 분산 환경에서의 적용 가능성을 높이고자 하며, 의료 데이터가 공유됐을 때의 개인정보 보호 문제를 고려하여, 분산학습 환경에서 차등 프라이버시 기반 노이즈를 추가함으로써 개인정보 보호를 강화할 예정이다. 이는 LLM 기반 의료 시스템에서도 법적, 윤리적 안정성을 확보하는데 중요한 요소가 될 것이며, 분산학습으로 인한 약물 추천에 대한 다양성과 다기관의 협력을 기대한다.

참고문헌

[1] Sun et al., "Improving LoRA in privacy-preserving federated learning," *arXiv preprint*, arXiv:2403.12313, 2024.

[2] Dao, "Flashattention-2: Faster attention with better parallelism and work partitioning," *arXiv preprint*, arXiv:2307.08691, 2023.

[3] Kim, S. H. Park, and H. Lee, "PANCDR: precise medicine prediction using an adversarial network for cancer drug response," *Brief. Bioinform.*, vol. 25, no. 2, p. bbae088, Jan. 2024

[4] S. Park, J. C. Lee, and J. M. Byun, "ML-based sequential analysis to assist selection between VMP and RD for newly diagnosed multiple myeloma," *npj Precis. Onc.*, vol. 7, no. 46, 2023.

[5] Kim, Y. H. Choi, J. Y. Choi, H. J. Choi, R. W. Park, and S. J. Rhie, "Translation of machine learning-based prediction algorithms to personalised empiric antibiotic selection: A population-based cohort study," *Int. J. Antimicrob. Agents*, vol. 62, no. 5, p. 106966, Nov. 2023

[6] Ryu, Jae Yong, Hyun Uk Kim, and Sang Yup Lee. "Deep learning improves prediction of drug–drug and drug–food interactions." *Proceedings of the national academy of sciences* 115.18, 2018

[7] H. Lee, H. Y. Jeong, M. H. Kim, M. E. Lim, D. H. Kim, Y. W. Han, Y. W. Kim, J. H. Choi, and S. H. Kim, "Trends of Clinical Decision Support System (CDSS)," *J. Biomed. Inform. Res.*, vol. 31, no. 4, pp. 77–85, Aug. 2016.

[8] E. J. Hu et al., 'LoRA: Low- Rank Adaptation of Large Lan- guage Models.' arXiv, Oct. 16, 2021. Accessed: Mar. 08, 2024. [Online]. Available: http://arxiv.org/abs/2106.09685

[9] Dao et al., "Flashattention: Fast and memory-efficient exact attention with IO-awareness," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 16344–16359, 2022.

[10] Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." *arXiv preprint arXiv:2302.13971* (2023).

[11] Johnson, A.E.W., Bulgarelli, L., Shen, L. et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 10, 1 (2023). https://doi.org/10.1038/s41597-022-01899-x

[12] U. Lee et al., 'Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in eng-lish education,' Educ Inf Technol, Oct. 2023, doi: 10.1007/s10639-023-12249-8.