논문 번호 : GEP-0668

# LLM 기반 의료진 보조 처방 시스템

김유나, 이성진, 정은성
홍익대학교 소프트웨어융합학과

## 서론

### 연구 배경

최근 의료진 부족으로 응급 상황 대응과 필수 진료 서비스 제공에 어려움이 발생하고 있으며, 의료 취약 지역에는 공공의료 인력 부족이 장기간 지속되어 필수 의료 서비스 제공에 영향을 미친다는 분석 존재.
또한, 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 각인.

### 연구 목적

의료진 부담을 줄여줄 수 있는 자동화된 의료 시스템의 필요성을 인식했으며, 이에 본 연구는 환자 진단에 알맞는 처방 약물 리스트를 예측하여 의료진 부담을 줄여주는 AI 시스템을 개발. 추가적으로 LLM의 최적화를 위해 FA-LoRA[1]와 FlashAttention2[2]을 적용하는 방법을 제안.

1. 의료진을 보조하는 AI 시스템 구축 : 환자의 진단 정보를 입력 받아 적절한 처방 약물 리스트를 예측함으로써, 의료진의 진단 업무를 보조하는 AI 시스템을 제안.
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 구현.

## 본론

### 관련 연구

- 기존 연구 : 약물의 반응, 내성, 상호작용 예측에 집중하며, 통계적으로 추적된 처방 빈도 경향만 분석. LLM 같은 모델을 사용하거나 진단 정보만으로 처방 약물을 예측하는 방식은 드뭄.
  1. GIST PANCDR : 유전자 발현 + 약물 그래프 → 암 환자 약물 반응 예측 [3]
  2. ImpriMed : 개인 맞춤 항암제 추천 AI [4]
  3. 아주대 의료원 : 진단·약물 기록 활용 → 항생제 내성 예측 [5]
  4. KAISR DeepDDI : 약물-약물·약물-음식 상호작용 예측 [6]
  5. CDSS : 환자별 처방 패턴 분석 [7]
- 본 연구 차별점 : LLM을 활용해 진단 기반 약물 리스트를 생성하고, 의료진이 이를 검토하여 처방을 결정하는 보조 시스템 구조에 초점을 맞춤

### FA-LoRA(Frozen-A Low-Rank Adaptation) [1]

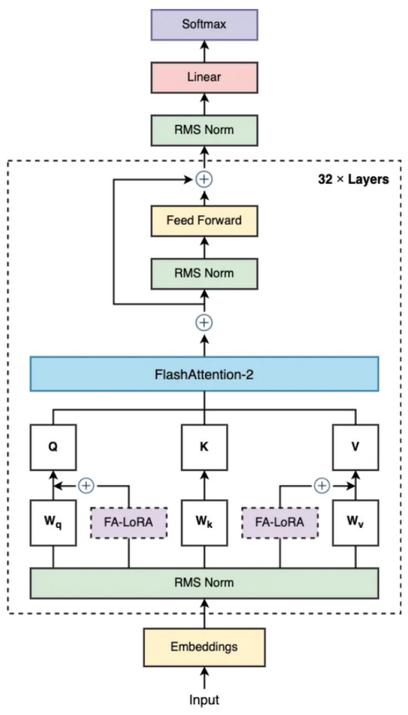
- 컴퓨팅 자원이 제한된 환경에서 LLM의 모든 파라미터를 동시에 미세 조정(fine-tuning)하기 위한 기법.
- 기존 LoRA를 한 단계 더 발전시켜 한층 가벼운 형태
- 모델의 전체 파라미터를 직접 업데이트하는 대신, 저차원(rank가 낮은) 행렬 A와 B를 추가로 삽입하고, 두 저차원 행렬 중 A 행렬을 고정(frozen) 하고, B 행렬만을 학습 가능한 파라미터로 유지
- 장점
  1. 더 작은 파라미터 수로도 동일하거나 유사한 성능을 달성.
  2. 메모리 사용량 및 연산량을 더욱 줄여, 학습과 추론 속도 모두 향상.
  3. Attention head, Feed-Forward Network(FFN) 등 특정 계층만 부분적으로 미세조정하여 전체 모델의 안정성과 효율성을 보장.

### FlashAttention 2 [2]

- Transformer의 Self-Attention은 모든 토큰 간 연산을 수행하기 때문에, 시퀀스 길이가 늘어날수록 연산량과 메모리 사용량이 제곱 비율로 증가한다.
- 이로 인해 시퀀스가 두 배가 되면 계산량은 네 배로 증가하며, 학습과 추론 속도 저하 및 병목 현상이 발생하여, 해결하고자 FlashAttention 2 기법을 도입.
- FlashAttention 2
  - Self-Attention 연산을 블록 단위로 수행하며, GPU의 글로벌 메모리 접근을 최소화하여 보다 빠른 공유 메모리를 적극 활용하여, 데이터 이동을 최소화.
  - FlashAttention 1의 한계를 보완해, 다음과 같은 더 좋은 성능을 제공.
    1. Non-Matmul FLOPs 최적화 : Softmax, Normalization 등의 Non-Matmul 연산을 최적화하여, 불필요한 연산과 메모리 접근을 줄임.
    2. GPU 병렬성 강화 : 여러 개의 GPU 스레드 블록을 활용하여 연산을 분산함으로써, 병렬
    3. 처리 능력을 크게 강화한다. 이는 연산 병목을 완화하고 처리량을 향상.
    4. 공유 메모리 최적화 : 중간 계산 결과를 공유 메모리에 저장하여 메모리 대역폭 병목을 줄이고, 연산 경로를 단축.

### Backbone Model

- Backbone 모델 : Meta의 Llama3.1-8B-Instruct[8]
- 모델 선택 이유
  1. 80억 개 파라미터로 상대적으로 가벼우면서도 높은 성능 제공
  2. FlashAttetnion 2 호환으로 연산 효율성 및 유연성 확보
  3. Instruction 기반 미세조정에 최적화되어 하위 작업 맞춤 학습 용이
- 최적화 방식
  1. FA-LoRA 도입 : 사전학습 가중치 동결하여 학습 가능하지 않도록 설정하며, 기존 LoRA와 동일하게 q, v 레이어에 어댑터 적용하여 최종적으로 기존 80억개 파라미터에서 약 524만개로 파라미터 대폭 감소.
  2. FlashAttention 2 도입 : 기존 Self-Attention 대비 연산량, 메모리 감소 및 속도 향상하여 FP32(32-bit)은 약 32GB 절약, FP16(16-bit)는 약 16GB 절약하여 메모리 절감 효과 가져옴.
- 핵심 성과
  최적화된 미세조정 과 연산 개선으로 메모리 사용량을 최소화하여 효율적인 학습, 추론 환경을 구축.




## 실험 구현

### 데이터셋 전처리

- 사용 데이터셋 : MIMIC-IV[9], 미국 Beth Israel Deaconess Medical Center 에서 2008년부터 2019년 까지 수집된 대규모 의료 데이터셋으로, 다양한 임상 데이터를 제공.
- 데이터 전처리 방식
  1. 데이터 병합 : 환자 ID 및 입원 ID를 기준으로 사용할 환자의 신체 정보, 진단 정보, 처방 약물 데이터를 병합한다. 단일 입원 기간 동안 여러 진단과 다수의 약물이 함께 기록되어, 각 진단에 맞는 약물을 1:1로 매칭하는 방식엔 어려움이 존재하여, 해당 입원 기간 전체의 진단 목록과 약물 목록을 하나의 샘플로 구성하여 학습에 활용.
  2. 데이터 전처리 : 모델 학습을 위한 일관된 데이터셋을 구성하기 위해 환자의 개인정보와 진단을 결측치 처리, 중복 제거, 이상치 처리 등의 전처리 작업을 수행.
  3. 입력 데이터 변환 : 최종적으로 정제된 데이터를 모델에 입력할 수 있도록 모델 프롬프트 형식에 맞춰 변환.

### 평가 방식

- 본 연구에서는 Human Evaluation과 LLM Evaluation 두가지를 정확도와 적합성으로 병행하여 보다 객관적이고 효율적인 평가를 진행.

1. Human Evaluation
   - 사람이 직접 평가 기준을 정해 분석하는 방식
   - 테스트 데이터셋의 진단에 따른 처방 약물 리스트를 정답으로 가정하고 평가
   - 평가 기준
     1. 형식에 맞게 응답을 생성했는가?
     2. 출력 약물 리스트가 정답 리스트와 얼마나 유사한가?
   - 평가 방법 : 각 샘플에서 정답과 일치하는 약물의 비율을 계산해 평균 정확도 산출.
   - 장점 : 높은 객관성
   - 단점 : 시간, 비용 소모. 적절한 처방이나 정답 데이터셋에 없는 약물은 과소평가된다.
   - 단점을 보완 : LLM Evaluation 방식을 병행해 평가.
2. LLM Evaluation
   - LLM이 LLM을 평가하는 방식
   - Hugging Face의 의료 QA 리더보드에서 약 83%의 높은 성능을 기록한 GPT-4는, 실제 환자 사례를 기반으로 의사와 동일한 조건에서 진단을 수행한 연구에서도, 진단 정확도와 근거 제시 측면에서 일부 경우 의사보다 높은 평가를 받은 바 있다.
   - 평가방법
     1. 프롬프트로 '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 구성.
     2. GPT-4의 응답을 바탕으로, 각 샘플의 전체 처방 중 적절하다고 판단된 약물의 비율을 계산하고, 이를 평균하여 전체 적합성 점수를 산출.

### 실험환경

- RTX 6000 Ada 48GB
- learning rate : 1e-3
- batch size : 4
- epochs : 5
- FA-LoRA의 alpha : 16, rank : { 16, 32 }
- max_length : 2048
- optimizer : 기존 SGD를 활용
- steps : 업데이트 누적 횟수 (batch 단위로 학습할 때마다 1씩 증가)

### 실험 결과

- 여기서 steps는 전체 학습 중 모델 파라미터가 누적 업데이트된 횟수를 의미하며, checkpoint 저장 시점마다(예: steps 1950, 2925...) 모델을 추출해 별도로 추론 및 평가를 진행했다.
- 출력률 : rank 값과 관계없이 약 55% 유지
- 정확도 : rank 16, steps 2925에서 최고 정확도를 보였으며, 평균적으로는 rank32가 더 높은 정확도와 더 안정적이고 일관된 성능 보인다.
- 적합성 : rank 32에서 더 높은 정확성 기록

| 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% |
| | 4875 | 55% | 21.6% | 91.0% |

- 정확도와 적합성 차이 이유
  1. 모델은 진단에 적합한 유사한 성분의 다양한 약물을 추천하지만, 정답 데이터셋은 단일 약물만을 정답으로 간주되어, 정확도가 낮게 측정되는 경향 존재.
  2. Llama3.1-8B-Instruct 모델는 의료 지식에 대한 사전 학습이 어느 정도 되어있어, 학습되지 않은 약물임에도 잔단에 적합하면, 약물을 추천하는 경향 존재한다. 이는 정답 데이터셋에 존재하지 않은 약물이기에 정확도가 낮게 나오는 원인중 하나이다.
- 결론
  정확도도 중요하지만, 본 연구에서는 정합성이 중요하다. 이는 모델이 의료진을 대체하는 것이 아니라 진단에 적합한 약물 추천을 제공하여, 의료진의 최종 결정을 돕는 보조적인 역할이기 때문이다. 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.

### 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] Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." arXiv preprint arXiv:2302.13971 (2023).
[9] 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