LoRA 기반 LLM 미세 조정 기법을 통합한 연구 논문 분류에 관한 연구

딥또 비스와스[1], 변태영[2], 길준민[3], 조재춘[4*]
[1,2]대구가톨릭대학교 컴퓨터소프트웨어학과, [3]제주대학교 컴퓨터공학과, [4]제주대학교 컴퓨터교육과
[1]dipto.biswas94@gmail.com, [2]tybyun@cu.ac.kr, [3]jmgil@jejunu.ac.kr, [4]jjo@jejunu.ac.kr

A Study on Research Paper Classification Using Fine-Tuning LLM Based on LoRA

Dipto Biswas[1], Tae-Young Byun[2], Joon-Min Gil[3], Jaechoon Jo[4*]
[1,2]Dept. of Computer Software Eng., Daegu Catholic Univ., [3]Dept. of Compuer Engineering, Jeju National Univ., [4]Dept. of Computer Education, Jeju Nation Univ.

요 약

대형 언어 모델(Large Language Models, LLM)은 텍스트 분류를 포함한 다양한 자연어 처리(Natural Language Processing, NLP) 작업에서 뛰어난 성능을 입증해 왔다. 그러나 이 모델을 특정 작업에 효과적으로 적용하려면 정교한 미세 조정(fine-tuning)과 고품질의 라벨링된 데이터셋이 필요하며, 이는 전문 학술 분야에서는 흔치 않다. 본 논문에서는 대형 언어 모델인 Meta AI의 LlaMA-3(3B)를 LoRA(Low-Rank Adaptation)을 활용하여 미세 조정함으로써, 정확하고 효율적인 연구 논문 분류(Research Papers Classification, RPC)를 가능하게 하는 접근 방식을 제안한다. LoRA는 사전학습된 모델을 소수의 파라미터만 업데이트함으로써 적은 비용으로도 높은 성능을 유지하면서 모델을 적응시킬 수 있게 한다. 본 논문에서는 LoRA 기반 미세 조정 기법을 RPC에 적용하고, 분류 시나리오에서 Top-n 기법을 활용하여 모델의 성능을 평가한다. 특히, 검색 질의의 맥락에 적합한 사용자의 맞춤형 추천을 제공하기 위해 의미론적 관계를 활용하는 Top-n 라벨링 기법을 제안한다. 실험 결과, LoRA를 적용한 LlaMA-3 모델이 기존의 기본 미세 조정 기법보다 높은 분류 성능을 보였으며, 이는 대규모 학술 연구 환경이나 자원이 제한된 시스템에서 제안된 분류 및 추천 방식이 적합한 방식임을 보여준다.

Ⅰ. 서 론

최근 연구 논문의 발행이 기하급수적으로 증가함에 따라, 연구자들이 관련 논문을 정확하게 식별하고 접근하는 것이 점점 더 어려워지고 있다[1]. 디지털 라이브러리 플랫폼이나 학술 저널 저장소는 연구 분야나 발행일 기준으로 콘텐츠를 관리하고 있으며, 기존의 전통적인 방식들은 학술 문헌의 복잡성과 최근의 기술적 특성을 효과적으로 반영하지 못한다. 기존의 인덱싱 및 필터링 시스템은 의미론적 이해가 부족하여 비효율적인 추천 결과를 초래하며, 정보 과부하 문제를 유발하고 있다. 이로 인해 연구자들은 방대한 양의 문헌을 수동으로 탐색해야 하며, 이는 많은 시간 소모를 유발하므로 비효율성을 초래한다. 이러한 정보 검색의 비효율성은 연구자의 생산성과 학술 연구의 전반적인 품질에도 부정적인 영향을 미친다. 특히, 융합 연구가 증가함에 따라 전통적인 분류 체계는 서로 다른 연구 주제 간의 미묘한 관계를 정확하게 반영하지 못하고 있으며, 여러 학문 분야에 걸친 주요 연구의 식별을 더욱 어렵게 만들고 있다. 연구자들은 핵심 논문을 놓치거나 관련성이 낮은 콘텐츠를 검토하느라 불필요한 시간을 소모할 수 있다. 이러한 문제를 해결하기 위해, 본 연구는 LLaMA-3 모델의 의미론적 이해 능력과 LoRA의 경량 적응성을 결합하여 이를 RPC 작업에 적용하고자 한다. 또한, 의미적 관계를 포착하고 검색 질의의 맥락에 맞는 정밀한 사용자의 중심 추천을 제공하기 위해 Top-n 라벨링 기법을 도입한다. 본 연구의 제안 기법을 통해 사용자의 선호도에 맞는 연구 논문을 분류하고 추천하는데 활용하고자 한다.

Ⅱ. 시스템 모델 및 방법론

본 연구에서 제안하는 LLM 기반 RPC는 구조화된 파이프라인으로 시작되며, 다음과 같은 과정을 따른다. 먼저, 제목, 초록, 키워드(선택적으로 전체 텍스트나 메타데이터 포함) 등의 정보를 포함한 논문 데이터셋 D가 입력으로 주어진다. 각 논문 $p \in D$에 대해 텍스트 입력 $RPA = \{p_1, p_2, ..., p_n\}$는 연구 논문 초록의 정규화 및 전처리를 의미한다. 데이터처리 단계에서는 NLTK 모듈을 이용하여 원시 텍스트에서 문장 부호를 제거하고 불용어, 표제를 추출하며, 전체 단어를 명사형으로 변환하여 표현의 명확성, 일관성 및 일반화를 도모한다. 이를 통해 추상적 개념이나 행위를 명확한 개체로 표현할 수 있어 정보를 더욱 형식적으로 구조화할 수 있다. 전처리 함수 P는 다음과 같이 표현된다.

$$P(RPA) \rightarrow RPALSUP' \quad (1)$$

여기서 $RPALSUP'$는 전처리된 연구 논문 초록을 나타낸다.

그 다음 단계에서는 $RPALSUP'$를 LlaMA-3 모델에 맞는 SentencePiece 토크나이저로 토크화하고, 고정된 최대 토큰 길이 L (예: $L = 512$)로 자르거나 패딩 처리한다. 이로 인해 입력 시퀀스 X는 다음과 같이 정의된다.

$$X = \{x_1, x_2, ..., x_L\}, \; x_i \in V \quad (2)$$

여기서 V는 어휘 공간을 의미한다. 입력 시퀀스 X는 임베딩되어 LlaMA-3 모델에 입력된다.

언어 모델로 사용하는 LlaMA-3는 다국어 및 다양한 도메인의 대규모 말뭉치에 대해 사전학습된 디코더 기반 트랜스포머 모델이다. 이 모델은 다층의 트랜스포머 블록으로 구성되어 있으며, 각 블록은 멀티-헤드 자기 주의(Multi-Head Self-Attention, MHSA)와 위치별 피드포워드 네트워크(Position-wise Feedforward Network, FFN)를 포함하고 있다. 입력 X에 대해 모델은 다음과 같이 문맥 임베딩 H를 생성한다.

$$H = Transformer(X) = LN(X + MHSA(X) + FFN(X)) \quad (3)$$

여기서 $MHSA$는 멀티-헤드 자기 주의, FFN은 위치별 피드포워드 네트워크, LN은 계층 정규화(Layer Normalization)를 의미한다.

LlaMA-3의 전체 파라미터를 조정하는 대신, 본 연구에서는 LoRA(Low-Rank Adaptation)[2]를 활용하여 어텐션(attention) 계층에 소규모 저랭크 행렬을 삽입해 선택적으로 미세 조정한다. 구체적으로, LoRA는 쿼리 W_Q와 값 W_V 행렬에 대해 훈련 가능한 랭크 분해 행렬 $A \in RLSUPd \times r$와 $B \in RLSUPr \times d$를 삽입한다. 업데이트된 가중치 행렬 $\widetilde{W}$는 다음과 같이 표현된다.

$$\widetilde{W} = W + \triangle W = W + \alpha \cdot AB \quad (4)$$

여기서 W는 고정된 사전학습 가중치, $\triangle W = AB$는 저랭크 적응 파라미터, α는 스케일링 계수이다.

연구 논문에 대한 입력 시퀀스를 받은 사전학습된 LlaMA 모델은 문맥 임베딩으로 인코딩한다. 이후 풀링 전략을 통해 최종 토큰의 임베딩

에서 전역 표현 $h \in RLSUPd$를 추출하고, 이를 LoRA가 적용된 선형 분류기에 입력하여 로짓(logits) $z \in RLSUPC$을 생성한다.

$$z = W_{cls}h + b + \alpha \cdot ABh \quad (5)$$

여기서 $W_{cls} \in RLSUPC \times d$, $b \in RLSUPC$는 분류기 층의 학습 가능한 파라미터이다. W_{cls}는 임베딩 h를 클래스 수 C에 해당하는 공간으로 투영하며, b는 각 클래스에 대한 편향 벡터이다.

로짓은 소프트맥스 함수로 변환되어 각 클래스에 대한 확률 분포를 형성한다.

$$\hat{y_i} = softmax(z_i) = \frac{eLSUPz_i}{\sum_{j=1}^{C} LSUPeLSUPz_j}, \ \forall_i \in \{1,2,...,C\} \quad (6)$$

여기서 z_i는 i-번째 클래스에 대한 로짓이며, $\hat{y_i}$는 예측 확률, C는 클래스의 총 개수이다. 클래스는 각각 Top-n 레이블을 기준으로 5와 10으로 설정하였다.

최종적으로, 예측된 레이블은 가장 높은 확률을 가진 클래스로 정의된다.

$$\hat{y} = \arg\max_i \hat{y}_i \quad (7)$$

본 연구에서는 단일 레이블 다중 클래스 분류에 적합한 범주형 교차 엔트로피(Categorical Cross-Entropy, CCE) 손실 함수를 사용하여 모델을 최적화한다. 한 샘플에 대한 손실은 다음과 같이 정의된다.

$$L_{CCE} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \quad (8)$$

여기서 $y_i \in \{0,1\}$은 원-핫 인코딩된 정답 벡터이며, $\hat{y}_i$는 예측된 클래스 확률이다. 이 손실 함수는 상호 배타적인 클래스 예측을 유도하며, 단일 레이블 분류 작업에 효과적인 최적화를 가능하게 한다.

성능 평가를 위해서는 [3]의 성능 지표를 기반으로 소프트맥스 함수에서 가장 높은 점수를 갖는 클래스를 예측 결과로 선택하며, 정확도(accuracy), 정밀도(precision), 재현율(recall), F1-점수(F1-score) 등의 지표를 통해 라벨이 부착된 테스트 세트에서 모델의 성능을 평가한다.

Ⅲ. 실험 환경 및 결과

(1)데이터셋: 본 연구에서 활용한 데이터셋은 표 1과 같다. 표 1의 데이터셋은 "HuggingFace" 플랫폼의 API를 활용하여 수집되었다. 각 데이터셋은 논문 ID, 제목, 저자, 초록, 키워드, 출판 연도 및 추가적인 서지 정보 등으로 구성된 메타데이터 분야를 포함하고 있으며, 다양한 저널에 걸쳐 여러 학문 분야를 폭넓게 포괄하고 있다. 본 연구에서는 최신 연구 동향을 반영하기 위해 2023년에 출판된 논문들만을 선정하였다.

표 1. 본 연구의 실험에 활용한 데이터셋.

	arXiv	Elsevier
논문	35,000	32,072
문장	232,993	278,437
단어	1,217,714	1,496,324
학습 세트	28,000	25,657
테스트 세트	7,000	6,415
Top-n 라벨	5, 10	5, 10
평균 길이	278.65	370.72

(2) 모델 하이퍼파라미터: 이 설정에서는 "open_LlaMA_3b" 모델을 LoRA 구성 방식을 통해 전체 파라미터의 0.16%인 5,324,800개의 파라미터만 학습하도록 미세 조정하였다. 학습률은 5×10^{-5}, 학습 배치 크기는 8, 평가 배치 크기는 16이며, 총 5회의 에포크 동안 학습이 진행되었다. 과적합 방지를 위해 가중치 감쇠 $\lambda = 0.01$이 적용되었고, LoRA 관련 하이퍼파라미터로는 랭크 $r = 16$, 스케일링 계수 $\alpha = 16$, 드롭아웃 비율 0.5이다. 수치적 안정성을 위한 $\epsilon = 1\times10^{-6}$이 사용되며, 학습 상태는 10 스텝마다 기록된다. 이 구성은 전체 모델이 아닌 저차원 행렬에 집중해 효율적인 미세 조정을 가능하게 한다.

Ⅳ. 실험 결과

표 2와 그림 1은 LlaMA-3(LoRA)가 2가지 데이터셋(arXiv, Elsevier)의 Top-5 및 Top-10 레이블 설정에서 달성한 성능 향상을 보여준다. 특히, LlaMA-3(LoRA)는 정확도(ACC)와 F1-점수에서 기존 모델 대비 10% 이상의 성능 향상을 달성하였으며, 이는 최소한의 파라미터 조정만으로 모델을 최적화하는 LoRA 기반의 미세 조정과 논문 간 의미적 관계를 효과적으로 반영하는 Top-n 라벨링 방식 때문이다. 이러한 접근 방식은 개인화된 논문 분류(RPC) 작업에서 보다 정밀한 추천을 가능하게 한다. 특히, LlaMA-3(LoRA)는 Gemini-pro(S) 및 GPT-4 대비 12~15%의 성능 향상을 보여주며, 2가지 데이터셋에 대한 유연성과 도메인 특화된 학습에 대한 적응 능력을 보여준다.

표 2. Top-5 라벨 기반 제안 모델의 실험 결과

Model \ 데이터셋	arXiv		Elsevier	
	정확도	**F1-점수**	**정확도**	**F1-점수**
Gemini-pro(S)	0.4988	0.4780	0.4667	0.4546
GPT-4	0.5267	0.5059	0.4876	0.4820
LlaMA-2(8B)	0.5363	0.5298	0.4718	0.4258
LlaMA-3(LoRA)$^{Top-5}$	**0.6732**	**0.6529**	**0.7048**	**0.6786**

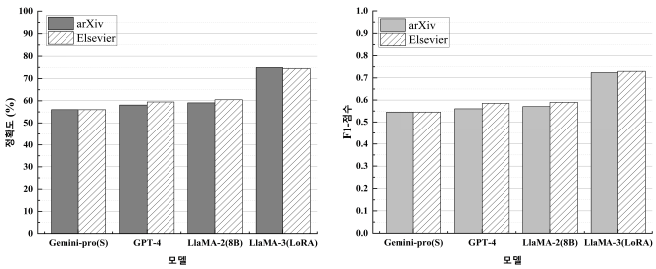

그림 1. Top-10 라벨 기반 제안 모델의 실험 결과

Ⅴ. 결론

본 연구에서는 LlaMA-3 모델을 LoRA로 미세 조정하여 연구 논문 분류(RPC)를 수행하는 접근 방식을 제시하였다. 제안한 모델은 사용자 맞춤 추천을 위한 분류 성능을 향상시키며, 실험 결과는 LoRA 기반 미세 조정이 Top-n 라벨링 방법과 결합될 때 분류 정확도를 크게 향상시킬 수 있음을 보여주었다. 제안 모델은 연구 논문 간의 의미적 관계를 효과적으로 포착하여 보다 정밀한 추천을 가능하게 해준다. 실험 결과는 LlaMA-3(LoRA)가 다양한 데이터셋에서 정확도 및 F1 점수 측면에서 기존 모델보다 높은 성능을 가짐을 보여준다. 또한, Top-n 라벨링의 도입은 보다 정확하고 사용자 중심적인 추천을 가능하게 하며, 학제 간 연구의 증가로 인한 복잡성을 해결하고 기존의 키워드 기반 분류 시스템의 한계를 극복하는데 기여할 수 있다. 한편, 모델의 적응성, 정확도, 개인화된 추천 성능을 향상시키기 위해 LoRA-FA(Feature Augmentation), 능동 학습, 멀티모달 데이터, 사용자 피드백을 통합하는 방안을 향후 연구로 수행할 예정이다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2025-16072365). 또한 2019년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 제주대학교 기초과학연구소 자율운영중점연구지원사업에서 수행된 기초연구사업임(RS-2019-NR040080).

참 고 문 헌

[1] Barbu, "Global trends in the scientific research of the health economics: a bibliometric analysis from 1975 to 2022," Health economics review, vol. 13, no. 1, p. 31, 2023.

[2] J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," Proceedings of the International Conference on Learning Representations (ICLR), 2022.

[3] G. Naidu, T. Zuva, and E. M. Sibanda, "A review of evaluation

metrics in machine learning algorithms," Computer science on-line conference, pp. 15-25, Springer, 2023.