자동 목차 추출을 위한 생성형 인공지능 기반 모델의 파인튜닝에 관한 연구

권수빈[1], 원일용[2]
[1]서울호서직업전문학교 컴퓨터공학과 학부생
[2]서울호서직업전문학교 컴퓨터공학과 교수
puppy9718@naver.com, clccclcc@shoseo.ac.kr

A Study on Fine-Tuning Generative AI-Based Models for Automatic Table of Contents Extraction

Su-Bin Kwon[1], Il-Young Won[2],
[1]Dept. of Computer Science, Hoseo Vocational School
[2]Dept. of Computer Science, Hoseo Vocational School

요 약

본 연구에서는 주어진 문서에서 목차 구조를 자동으로 정확하게 추출하기 위해 LLM 기반 모델인 Qwen2.5-3B-Instruct 을 사용하여 파인튜닝 한 모델과 하지 않은 모델을 비교 분석하였다. 실험 결과, 파인튜닝한 LLM 모델은 주어진 문서 환경에서 파인튜닝 전 모델보다 뛰어난 구조적 일치도를 달성하였다.

1. 서론

방대한 문서에서 필요한 정보를 빠르게 찾으려면 명확한 목차 구조가 필요하다. 그러나 실제 보고서는 비표준 형식이거나 스캔본으로 제공되는 경우가 많아, 목차 정보 추출이 쉽지 않다 [1].

본 연구는 생성형 AI의 파인튜닝을 통해 연구 보고서에서 계층적 목차구조를 효과적으로 추출하는 방법을 제안하고, 실험을 통해 효용을 검증하였다.

2장에서는 파인튜닝 및 사용한 기본 모델에 대해 설명하고, 3장에서는 실험 환경 및 실험 방법을 다룬다. 4장에서는 실험 결과와 정량적 비교를 통해 제안 방법의 효과를 분석한다.

2. 관련지식

2.1 파인 튜닝

파인튜닝은 사전학습(pre-trained)된 언어 모델을 특정 다운스트림 태스크에 맞추어 추가 학습하는 과정이다. 파인튜닝은 전체 파라미터 업데이트뿐만 아니라 LoRA, Adapter 등 파라미터 효율적 방법(PEFT)도 도입되어 효율성과 성능을 모두 추구한다 [2]. 특히 Supervised Fine-Tuning(SFT)은 입력-출력 쌍 데이터로 모델의 안전성, 도메인 특화 성능을 높이는 데 주로 사용되며, Alpaca, Qwen2.5 등 다양한 모델에 적용되고 있다[3].

2.2 Qwen2.5-3B-Instruct

본 연구에서는 Qwen2.5-3B-Instruct 모델을 기본 언어 모델로 선정하였다. Qwen2.5-3B-Instruct는 약 30억 파라미터의 instruction-tuned 언어 모델로, 29개 이상의 다양한 언어를 지원 및 실용적 성능을 보인다. 제한된 자원 환경에서도 효율적인 학습과 추론이 가능하도록 QLoRA 기반 4비트 양자화를 적용하였다. 4비트 양자화는 모델의 메모리 사용량과 연산량을 크게 줄이면서도, 성능 저하를 최소화할 수 있는 장점이 있다 [4].

3. 실험 환경 및 방법

3.1 실험환경

본 연구의 실험은 고성능 원격 서버 환경에서 수행되었다. 하드웨어는 NVIDIA RTX 4090 GPU를 사용하였으며, 소프트웨어 환경은 Python 3.10.12와 CUDA 12.1로 구성하였다. 모델 학습과 평가를 위해 다음과 같은 주요 라이브러리를 활용하였다.

<표 1> 주요 라이브러리

라이브러리	설명	버전
transformers	Hugging Face에서 만든 자연어 처리(NLP) 및 생성형 AI 모델들을 쉽게 사용하고 학습할 수 있게 해주는 라이브러리	4.52.4
peft	형 언어 모델을 적은 파라미터만 수정하여 효율적으로 미세 조정(Fine-Tuning)할 수 있게 돕는 라이브러리	0.10.0
trl	트랜스포머 모델에 강화학습 기법을 적용하는 라이브러리	0.8.6
accelerate	PyTorch 및 TensorFlow 모델을 다양한 하드웨어(GPU, TPU, 멀티 GPU 등)에서 쉽게 분산 학습할 수 있게 해주는 라이브러리	1.8.1
triton	NVIDIA가 만든 딥러닝 커널 최적화 및 GPU 커스텀 커널을 쉽게 작성할 수 있도록 하는 라이브러리	3.3.1

3.2 데이터셋

본 연구는 2021~2024년 한국문화관광연구원 연구보고서 약 250건에서 목차 포함 페이지만 추출해 데이터셋을 구성하였다. 추출한 목차 데이터를 계층구조와 페이지정보가 포함된 통일된 포맷으로 정제했다. 전체 데이터셋은 8:2로 훈련과 테스트용으로 나눴다. 또한 '제1장', '제1절', '1.1' 등의 패턴을 기반으로 목차의 단계별 깊이(장·절·항)를 구분하도록 구성하였다.

(그림 1) 데이터셋 일부 형식

3.3 하이퍼파라미터 설정

본 연구에서는 Qwen2.5-3B-Instruct 모델을 문서 목차 계층 구조 추출 태스크에 Supervised Fine-Tuning(SFT) 방식을 적용하였다. 학습에는 SFTTrainer, 4비트 QLoRA 양자화와 LoRA 기법을 적용해 메모리 효율성과 태스크 적응을 모두 확보했다. 주요 하이퍼파라미터는 표 2와 같다.

<표 2> 하이퍼파라미터

하이퍼파라미터	설명	값
per_device_train_batch_size	각 GPU나 디바이스마다 사용하는 미니배치 크기	1
gradient_accumulation_steps	그래디언트 누적 스텝 수	4
num_train_epochs	전체 훈련 데이터셋에 대해 반복하는 에포크(횟수) 수	3
learning_rate	학습률로, 모델 가중치를 업데이트할 때 적용하는 보폭 크기	2e-4
logging_steps	몇 스텝마다 학습 로그(예: 손실, 정확도 등)를 출력할지 결정하는 값	5
save_total_limit	체크포인트로 저장할 최대 개수	2
save_strategy	모델 체크포인트 저장 방식을 지정	"epoch"
fp16	16비트 반정밀도(floating point 16) 사용 여부	True
report_to	학습 로그를 보고할 대상	"none"

4. 실험 결과 및 분석

실험 결과, 테스트 데이터 50건에서 파인튜닝된 모델은 목차 구조 생성 결과와 실제 정답 간 평균 유사도(TEDS) 0.89를 달성했다. 베이스모델, GPT 계열 대비 구조적 일치도가 뛰어났으며, 실제 문서 환경에서도 신뢰할 수 있는 수준임을 확인했다.

<표 3> 평가

성능평가 / 모델 / 평가지표	파인튜닝 전 모델	파인튜닝 후 모델	chat gpt
Teds	0.27	0.89	0.79
llm	0.31	0.91	0.83

5. 결론

본 연구는 생성형 AI 기반 파인튜닝 기법으로 비정형 문서의 목차 정보를 효과적으로 구조화할 수 있음을 실험적으로 입증했다.

향후 연구에서는 적용 범위를 확대하고 새로운 형식에 대응하는 모델로 발전시킬 필요가 있다.

참고문헌

[1] Pengfei Hu et al., "Multimodal Tree Decoder for Table of Contents Extraction in Document Images", ICPR2022, Montréal,Canada, 2022, 1756 - 1762

[2] Andre Storhaug, Jingyue Li, "Parameter-Efficient Fine-Tuning of Large Language Models for Unit Test Generation: An Empirical Study", arXiv preprint, arXiv:2411.02462, 2024

[3] Long Ouyang et al., "Training language models to follow instructions with human feedback", arXiv preprint, arXiv:2203.02155, 2022

[4] Qwen, "Qwen2.5-3B-Instruct", Hugging Face, https://huggingface.co/Qwen/Qwen2.5-3B-Instruct, 2024