AutoRAG를 이용한 음성인식 기반 맞춤형 로컬 챗봇 시스템의 성능 개선에 관한 연구

김성진[1], 임재훈[1], 유동관[2]
[1]동서울대학교 컴퓨터소프트웨어학과 학부생
[2]동서울대학교 컴퓨터소프트웨어학과 교수
clapa0216@naver.com, pungddang@naver.com, dgyoo@du.ac.kr

Research on performance improvement of voice recognition-based customized local chatbot system using AutoRAG

Sung-jin Kim[1], Jae-hoon Lim[1], Sae-Hun Yeom[2]
[1]Dept. of Computer Software, Dong-seoul University

요 약

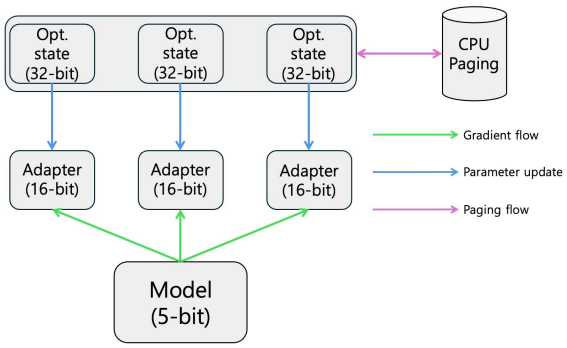
본 논문은 오픈소스 LLM(Large Language Model)인 Llama3를 기반으로 음성 인터페이스를 갖춘 맞춤형 로컬 챗봇 시스템을 개발하였다. 이 시스템은 PEFT(Parameter Efficient Fine-Tuning)와 AutoRAG(Auto Retrieval-Augmented Generation)로 최적화된 RAG(Retrieval-Augmented Generation) 방식을 결합한 하이브리드 접근법을 통해 Llama3를 전이학습 하였다. Ollama를 사용하여 로컬 환경에서 챗봇을 구현하였으며, LangServe와 Ngrok을 활용해 배포하였다. Raspberry Pi 5에 구현하여 모바일 환경으로 동작 가능하게 하였고 음성인식 기능을 추가하여 사용자 편의성을 높였다. 연구한 모델의 성능 평가는 총 18 종류의 데이터셋에 대해 각 질문당 5회씩, 총 90회의 질문으로 정확도를 측정하였다. 실험결과, PEFT 학습 모델과 Advanced RAG를 결합한 시스템이 가장 우수한 성능을 나타냈다.

1. 서론

최근 공공 기관과 기업들은 맞춤형 챗봇 구현을 위해 ChatGPT와 Gemini 같은 폐쇄형 LLM의 API 서비스를 주로 활용해 왔다. 그러나 이러한 접근법은 기업 정보 보안 문제와 API 사용료에 따른 재정적 부담을 야기했다. 본 논문에서는 이러한 문제를 해결하고자 무료 라이선스의 오픈 소스 Llama 3 모델을 기반으로 한 맞춤형 로컬 챗봇 시스템을 제안한다. 제안된 시스템은 PEFT와 RAG를 결합한 하이브리드 방식을 채택하고, AutoRAG 프레임워크를 통해 RAG시스템의 성능을 최적화했다. 또한 Raspberry Pi 5를 활용하여 모바일 환경에서 음성 인터페이스를 구현함으로써, 사용자의 편의성을 향상시킨 효율적이고 안전한 맞춤형 로컬 챗봇 솔루션을 개발하였다.[1]

2. PEFT(Parameter Efficient Fine-Tuning)

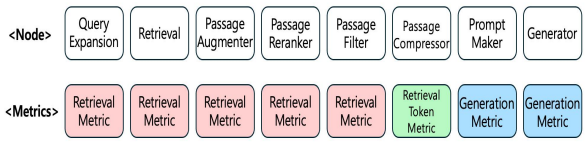
PEFT는 대규모 언어 모델의 효율적인 학습을 위해 개발된 방법론이다. 모델의 성능 향상에 따라 파라미터 수가 급증하면서 Full Fine-Tuning에 소요되는 비용과 시간이 크게 증가하여 실제 구현의 어려움이 대두되었다. PEFT는 이러한 문제를 해결하기 위해 대부분의 파라미터를 고정시키고 소수의 파라미터만을 조정하여 효과적인 학습을 가능케 한다. 본 연구에서는 PEFT 기법 중 하나인 QLoRA(Quantized Low-Rank Adaptation)를 채택하였다. 본 논문에서는 Meta 사의개발한 Llama 모델의 3번째 버전으로 오픈소스, 프리 라이선스로 제공되는 대규모 언어 모델인 Llama3 8B를 이용해 한국어 Fine-tuning된 모델을 5-bit Quantization한 모델을 사용하였다.QLoRA의 동작도는 그림 1과 같다.[2]



(그림 1) QLoRA방식의 동작 블록도

3. AutoRAG(Auto Retrieval-Augmented Generation)

수 많은 RAG 시스템 종류 중 어떤 시스템이 자신의 데이터에 적합한지 알아내는 것이 상당히 오래 걸리며, 복잡하고 어렵다. 이러한 문제를 해결하기 위해 Markr.AI 사가 개발한 AutoRAG[3] 프레임워크가 등장하게 되었다. 본 논문에서는 AutoRAG를 사용하여 보다 최적화된 시스템을 찾기 위한 방법으로 Passage reranker 모듈을 적용한 Advanced RAG를 사용한다. AutoRAG의 구성은 그림 3과 같다. AutoRAG를 사용하여 탐색한 동서울대학교 컴퓨터소프트웨어학과의 데이터셋에 가장 최적화된 Advanced RAG 시스템의 모듈의 규격은 표1과 같다.



(그림 2) AutoRAG의 구성 블록도

<표 1> 최적화 Advanced RAG 시스템의 모듈 규격

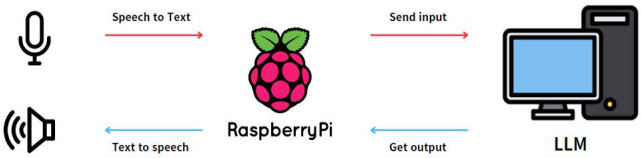
metric	module	module parameter	
Retreiver	BM25	top_k	5
Passage reranker	Koreranker	top_k	3
Generator	PEFT llama3	temperature	0.1

4. 개방형 LLM인 Llama3

Llama3은 Meta 사에서 개발한 Llama 모델의 3번째 버전으로 오픈소스, 프리 라이선스로 제공되는 대규모 언어 모델이다. Llama3은 비교적 적은 매개변수로 빠른 출력의 생성이 가능한 8B 모델과 복잡합 작업에 능숙한 70B 모델로 구분된다.[4]

5. 맞춤형 로컬 챗봇 시스템의 전체 구성도

본 논문에서 구현한 맞춤형 로컬 챗봇 시스템은 클라이언트와 서버로 이원화된 구조를 채택하고 있다. 클라이언트 부분은 음성 인식 및 합성 기능을 중심으로 설계되었으며, 모바일 환경에서의 동작을 구현하기 위해 교육용 임베디드 시스템인 Raspberry Pi 5를 플랫폼으로 선택하였다. 서버 부분은 맞춤형 모델의 로컬 실행을 위해 Ollama, LangServe, 그리고 Ngrok을 통합적으로 활용하여 구축되었다. 맞춤형 로컬 챗봇 시스템의 구성도는 그림 3과 같다. 사용한 Raspberry Pi 5의 구현 환경은 표2과 같다.



(그림 3) 맞춤형 로컬 챗봇 시스템의 구성도

<표 2> 임베디드 시스템(Raspberry PI 5)의 환경

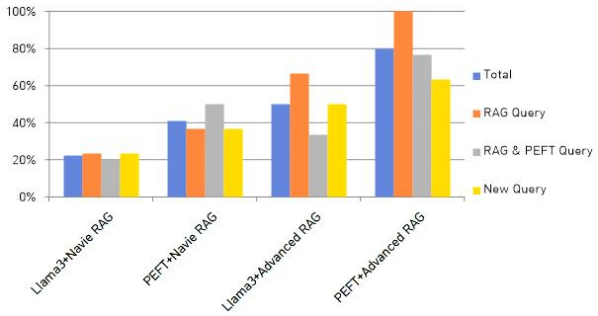
항목		내용
H/W	CPU	BCM2712 (2.4GHz)
	GPU	VideoCore VII (800MHz)
	MEMORY	SDRAM 4267
	SD card	micro 카드 슬롯, SDR104 고속 모드 지원
S/W	O/S	Debian GNU/Linux 12
	Library	langchain_community = 0.2.7 speechrecognition = 3.10.7 langserve = 0.2.2, pyttsx3 = 2.90

6. 맞춤형 로컬 챗봇 시스템의 서버모듈의 성능평가

본 연구의 성능평가는 동서울대학교 컴퓨터소프트웨어학과를 위한 안내 챗봇 시스템 개발을 목표로 수행되었다. 실험에 사용된 LLM은 Llama3의 8B 모델과 이를 PEFT 학습한 모델이며, RAG 시스템으로는 기본적인 Naive RAG와 AutoRAG를 통해 최적화된 Advanced RAG를 채택하였다. 이를 통해 총 4가지 조합의 시스템을 구성하여 실험을 진행하였다. 평가 데이터셋은 RAG 시스템의 성능을 중점적으로 검증하기 위해 RAG에 사용된 PDF 문서에서 추출한 6개의 질문, PDF와 PEFT 데이터셋에서 공통으로 나타나는 내용에서 6개의 질문, 새로운 내용을 다루는 6개의 질문으로 구성했다. 각 질문에 대해 5회씩 질문을 반복하여, 총 90회의 질의응답을 수행하였다. 성능평가 결과는 표3과 같고, 결과 그래프는 그림 5와 같다.

<표 3> 챗봇서버 모듈의 각 방식별 성능평가결과

RAG / LLM	Naive RAG	Advanced RAG
Llama3	22.2	50.0
PEFT	41.1	80.0



(그림 4) 각 방식별 성능 평가 결과 그래프

Advanced RAG시스템과 PEFT학습 LLM을 사용한 시스템이 80.0%로 가장 좋은 성능을 보여줬다. 각 RAG의 성능 차이와 PEFT 모델과의 시너지를 그래프를 통해서 전체적으로 점진적인 성능의 향상을 증명하였다.

7. 결론

본 논문에서는 오픈소스 LLM인 Llama3 8B를 활용하여 음성 인터페이스를 갖춘 맞춤형 로컬 챗봇 시스템을 개발하였다. 클라이언트-서버 구조로 설계된 이 시스템은 모바일 환경에서의 사용성을 고려하여 음성 기반으로 동작하도록 하였다. 성능 평가를 위해 맞춤형으로 변환되지 않은 일반 LLM과 PEFT 학습된 LLM, 그리고 Naive RAG와 Advanced RAG의 조합으로 네 가지 시스템을 비교 평가하였다. 실험 결과, PEFT 학습된 LLM과 Advanced RAG를 결합한 시스템이 가장 우수한 성능을 보였다. 본 논문에서 오픈소스 모델을 활용한 효율적이고 안전한 맞춤형 로컬 챗봇 구현의 방법을 제시하였다.

참고문헌

[1] https://bitnine.tistory.com/586

[2] Lingling Xu et al, “Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment” arXiv:2312.12148, 2023

[3] https://docs.auto-rag.com/structure.html

[4] Portakal, E. (2024, May 10). Llama 3 by Meta Ai 리뷰. TextCortex. https://textcortex.com/ko/post/llama-3-review