2024 년 봄 POSTECH 컴퓨터공학과 과제연구 연구제안서

Conformalized Retrieval for RAG

학　　번: 20210792
이　　름: 최준하
연구 지도교수: 박상돈

연구 목적 (Problem statement)

최근 대규모 언어 모델(LLM)은 다양한 자연어 처리(NLP) 작업에서 획기적인 성과를 거두었습니다. LLM 은 대량의 텍스트 데이터를 학습하여 텍스트 생성, 번역, 질의응답 등 다양한 작업을 수행할 수 있습니다. 하지만 LLM 은 학습되지 않은 정보에 대한 query 에 대해 거짓된 정보를 출력하는 hallucination 문제가 있습니다. 이를 해결하기 위한 방법 중 하나로 RAG 모델이 쓰입니다. RAG(Retrieval Augmented Generation) 모델은 검색 결과를 기반으로 텍스트를 생성하는 LLM 모델입니다. RAG 모델은 검색 엔진(Retriever)과 생성 엔진(Generator)으로 구성됩니다. Retriever 는 질의와 관련된 텍스트를 검색하고, Generator 는 검색 결과를 기반으로 새로운 텍스트를 생성합니다.

본 연구는 RAG 모델에서 Retriever 의 성능을 개선하기 위한 Top-K 설정 연구를 수행합니다. Top-K 설정은 Retriever 가 질의와 관련된 상위 K 개의 텍스트만 검색하도록 설정하는 방법입니다. 적절한 Top-K 설정을 통해 Retriever 의 성능을 개선하고, 이는 결과적으로 RAG 모델의 전체적인 성능 향상으로 이어질 수 있습니다.

연구 배경 (Motivation and background)

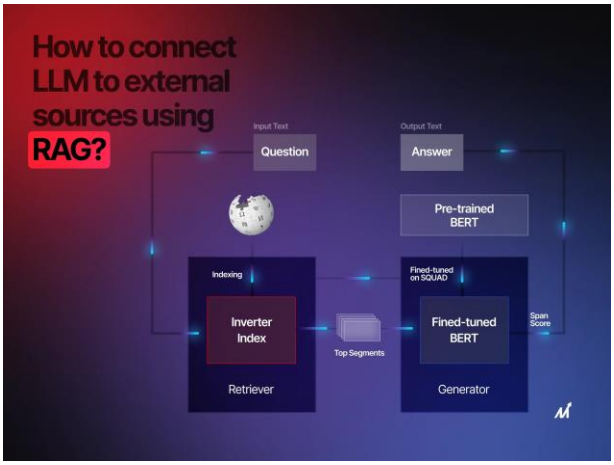

그림 1 RAG 의 구조를 간략히 표현한 그림[1]

RAG 모델은 기존 LLM 모델보다 더 나은 성능을 보여주지만, 여전히 개선할 점이 많습니다. 특히 Retriever 의 성능은 RAG 의 전체적인 성능에 큰 영향을 주는데, Retriever 가 질의와 관련 없는 텍스트를 검색하면 Generator 가 잘못된 정보를 기반으로 텍스트를 생성하기 때문에 RAG 모델의 정확도가 떨어지는 문제가 발생합니다.

기존 연구에서 Retriever 의 성능을 개선하기 위해 Query expansion, document ranking 등의 방법을 사용했으나 성능이 완벽히 개선되지 않았습니다. 본 연구는 conformal prediction 과 같은 방법으로 top-K 를 설정하고 Retriever 의 성능을 개선하려 합니다.

연구 방법 (Research proposal)

본 연구는 Huggingface 에 공개된 facebook 의 RAG 코드를 바탕으로 합니다. 본 코드의 retriever 부분을 수정할 예정입니다. 다음과 같은 방법으로 연구를 진행할 예정입니다.

[1] 그림 출처: How to Connect LLM to External Sources Using RAG? (markovate.com)

1. 다양한 prediction 기법을 사용해 top-K 를 설정합니다. 예측값에 대한 신뢰 수준을 함께 제공하는 정합 추론(conformal prediction)과 같은 방법을 사용할 수 있습니다.

2. 설정된 top-K 로 코드를 evaluation 합니다. evaluation 은 huggingface 에 공개된 RAG 의 eval 함수로 수행합니다.

3. 기존의 연구와 eval 점수를 비교하여 결과를 도출합니다.

기대 효과 (Expected output)

본 연구는 RAG 모델에서 Retriever 의 성능을 개선할 수 있습니다. 이는 더 나아가 RAG 모델의 전체적인 성능 향상으로 이어질 수 있으며, 이는 곧 LLM 모델의 발전을 의미합니다. LLM 모델은 현재 chat-GPT 의 선풍적인 인기와 함께 다양한 분야에서 사용되고 있습니다. 의료 분야에서는 LLM 을 사용해 의료 기록, 연구 논문 및 환자의 데이터를 분석하여 진단이나 치료 계획, 그리고 약물 개발에 사용합니다. 또한 LLM 은 고객 서비스 프로그램에 사용됩니다. 대규모 고객 문의를 처리하고 적절히 대응하여 인건비를 줄이고, 인간 고객 지원 담당자의 부담을 크게 줄였습니다. LLM 을 통해 기업은 고객 서비스 효율을 향상시키고 고객에게 더 개인화된 경험을 제공할 수 있습니다. LLM 은 콘텐츠 제작에도 사용될 수 있습니다.

이외에도 텍스트를 생성하고 질의응답을 하는 등 다양한 곳에 사용되는 NLP, 정보 검색 및 분석, 그리고 지능형 튜터나 개인 맞춤형 학습과 같은 교육 및 학습 분야까지 다양한 분야에 도움을 줄 수 있을 것입니다.

참고 문헌

RAG: Retrieval Augmented Generation for Knowledge-Intensive NLP Tasks: https://arxiv.org/abs/2005.14165

Improving Retrieval for Text Generation: https://arxiv.org/abs/2104.14732

RAG code: transformers/examples/research_projects/rag at main · huggingface/transformers (github.com)

연구 추진 일정

3/11 ~ 3/20: RAG 공부 및 코드 분석

3/20 ~ 4/6: conformal prediction 을 포함한 Top-K 예측에 사용할 방법 공부

4/7 ~ 4/20: RAG 코드에 적용 및 evaluation 진행

4/21 ~ 4/25: 연구 진행 보고서 작성 및 중간 발표 비디오 제작

4/27 ~ 5/25: RAG 코드에 적용 및 evaluation 진행

5/26 ~ 5/27: 포스터 제작

5/28 ~ 5/31: 최종 보고서 작성