검색-증강 언어 모델에 관한 개요

오민해, 이정우*
서울대학교, *서울대학교

minhae.oh@cml.snu.ac.kr, *junglee@snu.ac.kr

Survey of Retrieval-Augmented Language Model

Minhae Oh, Jungwoo Lee*
Seoul National Univ., *Seoul National Univ.

요 약

본 논문은 검색-증강 언어 모델(REALM)에 대한 종합적인 개요를 설명한다. 검색-증강 언어 모델은 최신 자연어처리 기술의 중요한 발전으로, 기존 언어 모델의 한계를 넘어선 새로운 접근 방식을 도입한다. 본 논문에서는 검색-증강 요약 모델의 구조와 다양한 검색 모델 유형에 대해 설명한다.

Ⅰ. 서 론

언어 모델은 자연어처리(NLP) 분야에서 많은 발전을 가져왔다. 최근의 대규모 언어 모델은 높은 정확도와 다양성으로 언어의 이해 및 생성 능력을 보였다. 이러한 모델들은 문맥 이해, 의미 추론, 질의응답, 글쓰기 등 다양한 분야에서 뛰어난 성능을 보인다.

하지만, 대규모 언어모델은 편향과 모델의 복잡성에 관련된 한계점이 있다. 훈련 데이터에 의해 편향된 결과물을 출력할 수도 있으며, 이는 때로 부정확하거나 부적절한 결과로 이어지기도 한다. 또한, 모델의 규모의 복잡성이 높아 훈련 및 사용에 많은 비용이 발생한다.

검색-증강 언어모델은 동적인 접근 방식을 도입하여, 외부 정보에 접근하여 사용함으로써, 최신 또는 전문 지식이 필요한 작업에서 이러한 한계점을 극복하는 모델로 제안되었다.

Ⅱ. 본론

검색-증강 언어 모델은 두개의 하위 모델인 검색기(retriever)와 언어 모델로 구성된다. Wikipedia 와 같은 대규모 코퍼스에서 상위 k 개의 관련 문서를 검색기를 사용하여 검색한 후, 그 문서들이 쿼리와 함께 언어 모델에 입력하여 질의응답에서 위키피디아 기사생성에 이르는 작업을 수행한다. 검색기는 언어 모델의 복잡도를 줄일 수 있게 도와주며, 효과적인 검색 결과를 통해 언어 모델의 성능을 향상시키도록 학습이 진행된다. 언어 모델은 BERT, RoBERTa, T5 와 같은 모델들이 많이 사용되며 검색기 모델은 어휘적 검색 (Lexical Retrieval), 희소 검색 (Sparse Retrieval), 밀집 검색 (Dense Retrieval), 후기 상호작용 (Late-Interaction), 재 순위 모델 (Re-ranking model)과 같은 모델로 분류가 된다.

어휘적 모델의 대표가 되는 BM25 는 TF-IDF 토큰 가중치를 가지고 두 고차원 벡터 사이의 토큰 매칭을 기반으로 하는 전통적인 키워드 매칭 기법이다. 즉, 쿼리와 가장 키워드 매칭이 높은 문서를 대규모 코퍼스에서 추출을 한다. 희소 검색모델인 DeepCT 는 BERT 모델을 기반으로 하여 용어 가중치 빈도를 학습한 후, T5 모델을 사용하여 합성 쿼리를 생성하여 기존 문서에 추가한다. 밀집 검색 모델인 DPR 은 단순한 키워드 매칭을 넘어서 의미 관계를 포착한다. 이를 위해 사전 학습된 BERT 와 같은 딥러닝 모델을 쿼리와 코퍼스의 임베딩 모델로 주로 사용한다. ColBERT 와 같은 후기 상호작용 모델(late-interaction model)은

쿼리와 코퍼스 사이의 더 복잡한 상호작용을 통해 관련된 문서들이 추출된다. 쿼리와 코퍼스 내의 문서들을 토큰 임베딩으로 인코딩 한 후 쿼리와 문서의 내적 값을 합산하여 관련도가 계산된다. 마지막으로 재 순위 모델(re-ranking model)은 BM25 와 같은 기존의 다른 검색기 모델을 사용하여 관련 문서를 100 개 가량 추출한 후 그 문서들을 언어모델을 사용하여 관련성 및 중요도를 재 평가하는 방식이다. 이를 통해 중요도가 높은 문서를 더 많이 참고하도록 유도할 수 있다.

Ⅲ. 결론

검색-증강 언어 모델은 자연어처리 분야에서 중요한 발전을 이뤄낸 모델로, 다양한 데이터 소스에서 지식을 검색하고 사용할 수 있다는 장점이 있다. 이는 전통적인 언어 모델이 제한된 훈련 데이터에 의존하는 것과 비교해 훨씬 넓은 범위의 데이터를 활용할 수 있게 한다. 추후 다양한 유형의 데이터 소스와 언어에 대한 학습 적응력을 향상시키고, 대규모 코퍼스로부터의 문서 검색 속도와 정확성을 높이는 모델에 대한 연구가 필요할 것으로 보인다. 이는 대규모 언어 모델과 외부 지식 코퍼스 간의 데이터 격차를 감소시키며 자연어 분야의 많은 발전을 가져올 것으로 기대된다.

ACKNOWLEDGMENT

This work is in part supported by National Research Foundation of Korea (NRF, 2021R1A2C2014504(34)), Institute of Information & communications Technology Planning & Evaluation (IITP- 2021-0-01059(33), IITP-2021-0-00106(33)) grant funded by the Ministry of Science and ICT (MSIT), INMAC, and BK21 FOUR program.

참 고 문 헌

[1] Guu, Kelvin, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. “Retrieval Augmented Language Model Pre-Training.”

[2] Karpukhin, Vladimir, Barlas Oğ uz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-Tau Yih. “Dense Passage Retrieval for Open-Domain Question Answering.”

[3] Khattab, Omar, and Matei Zaharia. “ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT.”

[4] Thakur, Nandan, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. “BEIR: A Heterogenous Benchmark for Zero-Shot Evaluation of Information Retrieval Models.”