Notion 계정이 있는 모든 사용자가 이 페이지에 댓글을 달 수 있습니다. 댓글을 작성하려면 가입 또는 로그인하세요.

📚 RAG 논문

1. RAG란 무엇인가?

기존 언어모델의 한계

최근 대형 언어모델(GPT, BERT 등)은 대규모 데이터에서 지식을 학습하여 다양한 작업에서 좋은 성능을 보이고 있음.

하지만 이러한 모델은 내부 파라미터에 지식을 저장하므로 다음과 같은 문제점이 존재함:

- 지식을 쉽게 업데이트하거나 수정할 수 없음
- 왜 그런 답을 했는지 설명하기 어려움
- 종종 사실이 아닌 내용('환각', hallucination)을 생성함

RAG의 핵심 개념

RAG(Retrieval-Augmented Generation)는 위 문제를 해결하기 위한 모델 구조로, **외부 문서(예: 위키백과)에서 필요한 정보를 검색하고, 그 정보를 바탕으로 문장을 생성**하는 방식임.

즉, 기존 언어모델처럼 말도 잘하지만, 그 내용을 **직접 찾아서 말한다는 점**에서 차별화됨.

---

2. RAG는 어떻게 구성되어 있는가? (구조)

RAG는 크게 두 부분으로 구성됨:

① 검색기 (Retriever)

- 입력 문장(예: 질문)을 받으면 관련된 문서를 외부 데이터베이스에서 검색함
- 검색 방식은 벡터 형태로 바꾸어 가장 유사한 문서(top-k)를 선택하는 방식이며, DPR(Dense Passage Retriever)을 사용함.

- DPR(Dense Passage Retriever)은 질문과 문서를 벡터로 변환해 의미적으로 유사한 문서를 효율적으로 찾는 딥러닝 기반 검색 방식임

② 생성기 (Generator)

- 검색된 문서를 기반으로 문장을 생성함
- BART라는 언어 생성 모델을 사용하여, 검색된 문서와 입력을 함께 보고 최종 출력을 생성함

---

3. 문장 생성 방식의 두 가지 모델

RAG-Sequence

- 전체 문장을 하나의 문서를 바탕으로 생성함
- 예를 들어 "중이란?"이라는 질문에 대해 한 문서를 참고하여 처음부터 끝까지 답변을 생성함

RAG-Token

- 문장을 구성하는 각 단어마다 다른 문서를 참고할 수 있음
- 예를 들어 긴 답변이 여러 개의 사실을 포함해야 할 때 유리함

---

4. RAG는 어떤 일에 활용되는가? (실험)

1) 오픈 도메인 질문 답변 (Open-domain QA)

- 예: "오바마는 어디서 태어났나?"
- 기존 모델은 내장된 지식에 의존하지만, RAG는 위키에서 하와이를 찾아 정확히 답변함

2) 설명형 문장 생성 (MS MARCO)

- 예: "중이란?"

- 기존 모델은 부정확한 문장을 만들 가능성이 있지만, RAG는 실제 위키 내용을 참고하여 정확한 정의를 생성함

3) 퀴즈 질문 생성 (Jeopardy-style)

- 주어진 정답(예: "워싱턴")에 대해 퀴즈 문장을 생성하는 작업
- 예: "미국 대통령 이름에서 유래한 유일한 주"
- 사실에 기반한 구체적인 문장을 생성함

4) 사실 검증 (Fact Verification)

- 주어진 문장이 참인지 거짓인지, 또는 알 수 없는지를 판단
- 위키백과에서 관련 문서를 찾아 직접 판단함

---

5. 실험 결과 및 비교

기존 모델 대비 RAG의 성능
정확도 최고 기록 달성
더 정확하고 구체적인 생성
사람이 평가해도 더 나음
복잡한 파이프라인 없이도 우수한 결과

---

6. RAG의 장점은 무엇인가?

① 지식 업데이트가 쉬움

- 내부 파라미터를 다시 학습하지 않아도, **문서 인덱스만 교체하면 최신 정보 반영 가능**
- 예: "미국 대통령은 누구인가?" 같은 질문에서 최신 대통령 정보 반영 가능

② 생성 문장의 근거 제공 가능

- 생성된 문장이 **어떤 문서를 기반으로 만들어졌는지 추적 가능**
- 설명 가능하고 신뢰성 있는 인공지능 구현 가능

③ 사실 기반 생성에 유리

- 외부 문서를 참고하여 **더 정확하고 사실적인 문장 생성 가능**
- 환각 현상이 줄어듦

---

7. 실제 사용 사례 및 확장 가능성

- 최신 뉴스나 법률 정보를 다루는 **대화형 시스템**
- 의료, 금융 등 **정확한 정보가 중요한 분야**
- 교육용 **퀴즈 자동 생성** 시스템
- **논문 요약**, 위키 기반 질문 생성 등

🔗 참고 링크

- 💻 RAG 데모 체험:
  https://huggingface.co/rag/