Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

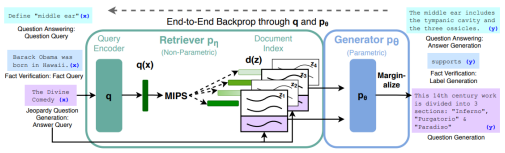
- 2020년
- RAG가 처음 소개된 논문
- LLM의 추론 능력을 향상 시키기 위해 쓰이는 현재의 사용법과 달리 처음에는 파인튜닝의 한 방법으로서 소개됨

https://arxiv.org/pdf/2005.11401

Introduction

- 기존 LLM의 한계: 메모리 확장성, 최신성/Hallucination
- RAG: parametric memory와 non-parametric memory 결합
  - parametric memory: 사전 학습된 seq2seq Transformer 모델(BART).
  - non-parametric memory: dense vector index로 표현된 Wikipedia
- REALM, ORQA: differentiable retriever를 포함한 masked language model이므로 제한적 결과 ↔ RAG: end-to-end로 open-domain 질문 답변 등에서 뛰어난 성능

Methods



x: input sequence

z: retrieve text

y: generated target sequence

Models

1. RAG-Sequence Model: z 전체를 사용해 생성
2. RAG-Token Model: z를 토큰 단위만큼 생성에 사용

Retriever: DPR (Dense Passage Retriever)

- facebook AI의 Dense Passage Retrieval (DPR) 기반
- TriviaQA [24] questions and Natural Question의 답변이 포함된 retrieve documents으로 학습
- top-k 문서만 사용
- non-parametric memory

Generator: BART

- facebook AI에서 개발한 seq2seq transformer(BERT+GPT) BART 기
- retriever로부터 검색된 문서들을 입력으로 최종 텍스트 생성
- parametric memory

Training

- end-to-end 방식으로 retriever와 generator를 함께 학습
- retrieval에 대해서는 unsupervision
- input/output 쌍에 대해 negative marginal log-likelihood 최소화

Decoding

1. RAG-Sequence: 일반적인 beam decoder
2. RAG-Token: 각 z에 대해 여러번의 beam search 수행

Experiments & Results

Table 1: Open-Domain QA Test Scores. For TQA, left column uses the standard test set for Open-Domain QA, right column uses the TQA-Wiki test set. See Appendix D for further details.

	Model	NQ	TQA	WQ	CT
Closed Book	T5-11B [52]	34.5	- /50.1	37.4	-
	T5-11B+SSM[52]	36.6	- /60.5	44.7	-
Open Book	REALM [20]	40.4	- / -	40.7	46.8
	DPR [26]	41.5	**57.9**/ -	41.1	50.6
	RAG-Token	44.1	55.2/66.1	**45.5**	50.0
	RAG-Seq.	**44.5**	56.8/**68.0**	45.2	**52.2**

Table 2: Generation and classification Test Scores. MS-MARCO SotA is [4], FEVER-3 is [68] and FEVER-2 is [57] *Uses gold context/evidence. Best model without gold access underlined.

Model	Jeopardy B-1	Jeopardy QB-1	MSMARCO R-L	MSMARCO B-1	FVR3 Label Acc.	FVR2 Label Acc.
SotA	-	-	**49.8***	**49.9***	**76.8**	**92.2***
BART	15.1	19.7	38.2	41.6	64.0	81.1
RAG-Tok.	**17.3**	**22.2**	40.1	41.5	72.5	89.5
RAG-Seq.	14.7	21.4	40.8	44.2		

Open-domain Question Answering

- Natural Questions, TriviaQA
- 기존 모델(REALM, ORQA) 대비 성능 향상 큼
- RAG-Token이 RAG-Sequence보다 더 높은 정확도

Abstractive Question Answering

- NarrativeQA
- 긴 문맥에서 요약된 답변을 생성하는 데 뛰어남

Jeopardy Question Generation

- 주어진 정답으로부터 Jeopardy 스타일의 질문 생성
- 창의적이고 지식에 기반한 질문 생성 가능

Fact Verification

- FEVER
- hallucination 줄일 수 있음

Additional Results

Generation Diversity

- 특정 분야의 질문 생성에서 더 창의적
- RAG-Token가 RAG-Sequence보다 적은 다양성

Retrieval Ablations

- retriever 답변에 중요한 역할

- BM25 retriever FEVER와 같은 작업에서 RAG의 dense retriever를 능가

Index hot-swapping

- 학습 후에도 새로운 정보를 쉽게 반영할 수 있음
- 2016년과 2018년의 서로 다른 인덱스를 사용하여 실험 → 최신 정보에 맞는 정확한 답변을 제공할 수 있음
- 인덱스 불일치 시 정확도가 급격히 떨어짐

Effect of Retrieving more documents

- RAG-Token: 검색된 문서 수가 10개일 때 최적화
- 문서 수와 Rouge-L 점수 비례, Bleu-1 점수와 반비례 → 검색된 문서 수가 성능에 미치는 영향 큼

Discussion

- open domaim QA에서 sota
- human eval에서도 RAG 모델이 BART보다 더 사실적이고 구체적인 답변을 생성한다고 평가
- 동적 지식 검색을 통해 성능을 지속적으로 향상시킬 수 있음