Dense Passage Retrieval for Open-Domain Question Answering

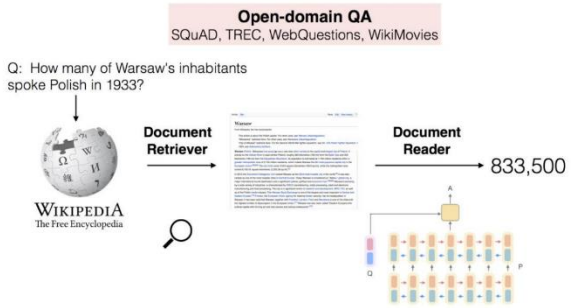
EMNLP 2020 | Tag: Open-Domain Question Answering, Retrieval

0. Summary

- Open-Domain Question Answering에서 2개의 BERT encoder로 구성된 Retrieval 방식 제안(Dense Passage Retrieval)

1. Introduction & Background

> Open-Domain Question Answering이란?



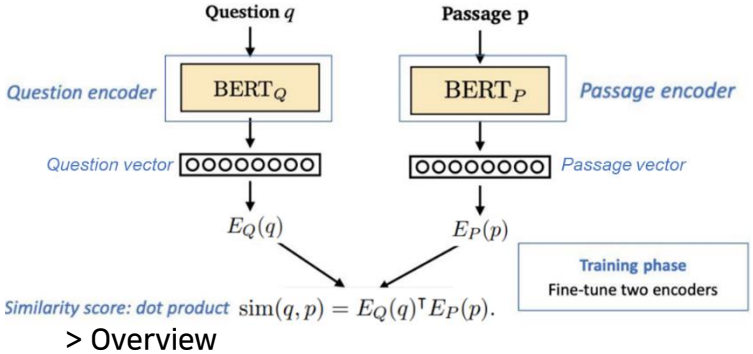
- Two stage(Retriever과 Reader)로 구성되어 있음
- **Retriever**: DB의 많은 문서 중 Question과 연관 있는 Top-k개 문서를 찾음(search)
- **Reader**: 검색된 Top-k 문서에서 Question에 대한 답을 찾음

> 기존 방식 문제점

- TF-IDF, BM25: High dimensional(문장 안의 단어마다 가중치를 계산해야 했음)
- ORQA: Computationally intensive, question encoder, reader만 fine-tuned됨 (context encoder는 X)

→ additional pretraining 없이 Question & passage 쌍으로만 dense embedding 모델 훈련이 가능할까?

2. DPR(Dense Passage Retrieval)



> Overview

- Question vector에 가장 가까운 k개의 passage를 Retrieve함
- **Similarity function**: dot product
  → 내적을 크게 하는 방향으로 Question, Passage encoder 학습
- E_P, E_Q: d(dimensions of vectors) = 768 (BERT)

> Training

- **Negative sampling**: 연관되지 않은 question passage pair
  → ***Negative sample 고르는 방법***
  1) 랜덤 선택
  2) BM25 기준으로 Top-k개 선택
  3) Gold 사용 (다른 question의 positive passage 중 고름)
  → 논문에서는 In-batch GOLD + BM25 negative sample 1개 추가하는 방식이 가장 성능 좋았음
- **목표**: 연관된 question과 passage(positive)간의 거리는 좁히고 연관되지 않은 것(negative)의 거리는 멀게 하는 것
- **Training data**: $\mathcal{D} = \{\langle q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^- \rangle\}_{i=1}^m$
  → 1 question q_i, 1 relevant(positive) passage p^+_i, n irrelevant(negative) passages $p^-_{i,j}$, m passages
- **Loss function**:

$$L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-) = -\log \frac{e^{\text{sim}(q_i, p_i^+)}}{e^{\text{sim}(q_i, p_i^+)} + \sum_{j=1}^n e^{\text{sim}(q_i, p_{i,j}^-)}}$$

  → Positive passage와 question, negative passage와 question과의 similarity score를 가져와서 Softmax 이 확률값을 NLL loss에 적용하여 학습 (세미나자료 참고)

3. 결과

- SQuAD dataset을 제외하고는 DPR이 BM25보다 성능이 좋았음
  → 왜 SQuAD에서는 낮나?
  1) Annotator에서 passage를 먼저 보고 질문 생성
  2) Only 500+ Wikipedia articles(biased)
- **Ablation study – Sample efficiency**
  : 1000개로만 학습시켜도 BM25의 성능을 능가함
  → 적은 question-passage pair로만 학습시켜도 High quality의 Dense Retriever 학습 가능
- **Ablation study – In-batch negative training**
  → 1) k>=20일때는 Random/BM25/GOLD 성능 비슷,
  2) In-batch 사용시 성능 더 좋음
  3) batch size가 늘어날수록 정확도 올라감
  → negative sample의 개수도 많아지게 됨
  4) Gold에 single BM25 negative passage 추가 + batch size 크게 하는 것이 성능 가장 좋았음
- **Question Answering**
  : SQuAD dataset을 제외하고는 DPR을 사용한 모델이 BM25를 사용했을 때보다 ODQA 자체의 성능을 높임

끝!!
질문 있으시면 정민지에게 DM 주세요!!
+) 추가 정보가 필요하다면 220117자 세미나 자료를 참고하세요