Dense Passage Retrieval for Open-Domain Question Answering - 2020

Thiago Coelho Vieira

1. Main Concepts

1. **Open Domain QA**: is a task that answers factoid questions using a large collection of documents.
2. **Passage retrieval**: effectively reduces the search space for answer extraction, but also identifies the support context for users to verify the answer
3. **Sparse Vectors**: contains mostly zero values, and only a few non-zero values. Sparse vectors are commonly used to represent documents in a text corpus, where each dimension represents a word in the vocabulary. **TFIDF** and **BOW**. Matches keywords efficiently with an inverted index and can be seen as representing the question and context in highdimensional, sparse vectors (with weighting).
4. **Dense Vectors**: contains non-zero values for every dimension, and are often generated using techniques such as **word embeddings**, which capture the semantic meaning of words in a language. Can also be learnable by task-specific goal representation.
5. **Retriever model**: component that can select a small set of relevant texts
6. **Reader model**: component to extract the answer from the relevant texts

2.1 Contribution

1. **Goal**:
   ```
   can we train a better dense embedding model using only pairs of questions and passages (or answers), without additional pretraining?
   ```
   1. the goal of the dense passage retriever (DPR) is to index all the passages in a low-dimensional and continuous space.
   2. uses a dense encoder which maps any text passage to a embedding vectors and builds an index for all the passages that we will use for retrieval.
2. Effective result on using dense vector for document retrieval

2.2 Architecture

1. dual-encoder (bi-encoder) with BERT
    1. encoder 1 = question
    2. encoder 2 = possible answers
    3. output = embedding from [CLS] 768dim
    4. cost function = max dot product between output vectors (ablation study showed that other similarity functions perform comparably)

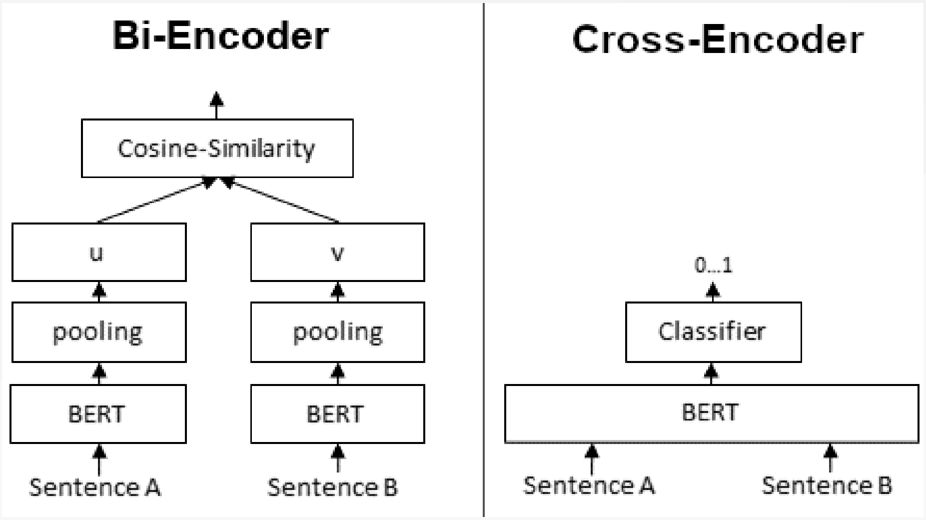


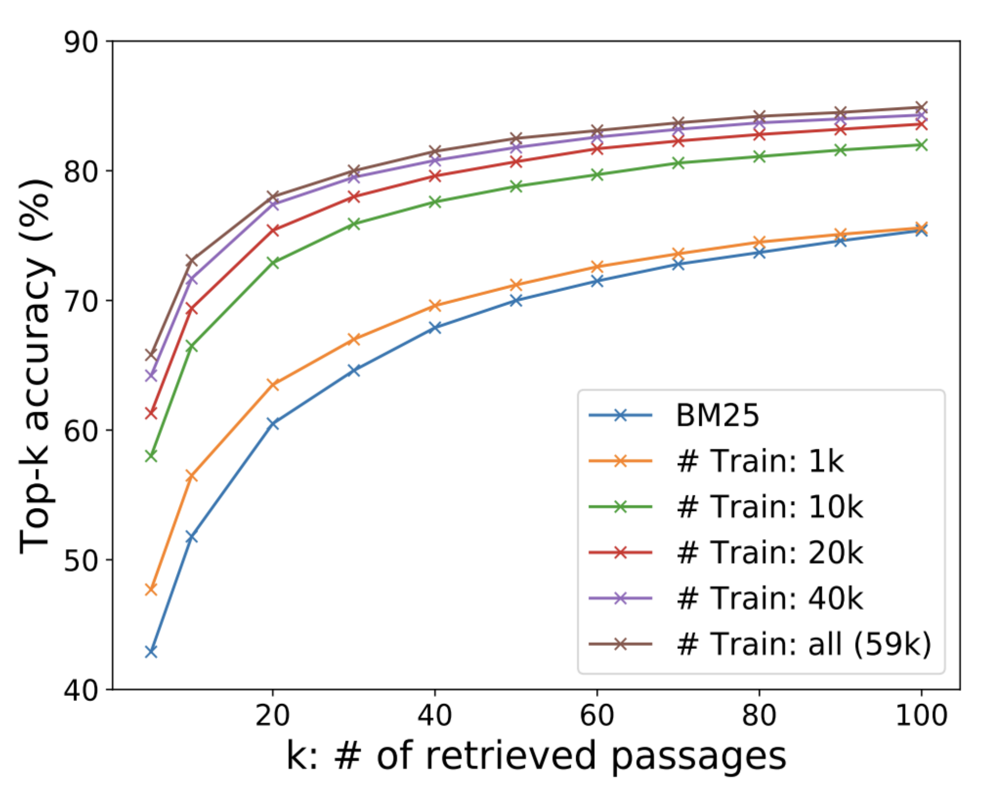
image source

3. interesting/unexpected results

- empirical analysis and ablation studies indicate that more complex model frameworks or similarity functions do not necessarily provide additional values.
- simply fine-tuning the question and passage encoders on existing question-passage pairs is sufficient to greatly outperform BM25
- **Negative Selection Strategies**
  - Random: random passages from the corpus
  - BM25: top passages that don't contain the answer but matches most of the question tokens
  - Gold: positive passages paired with other questions which appear in the training set
- **in-batch negatives** approach and reuse gold passages fro the same batch as negatives (improve computation efficiency)
- **FAISS**: an extremely efficient, open-source library for similarity search and clustering of dense vectors, which can easily be applied to billions of vectors. Used to index the passages before the inference.

4.1 Results

1. Outperforms Lucene-BM25 by 9%-19% absolute top-20
2. Outperforms BM25 by 65.2% vs. 42.9% in Top-5 accuracy



4.2 Results

1. Stablish SOTA on multiple open domain QA benchmarks on models with the best retriever precision
2. `In the context of open-domain question answering, a higher retrieval precision indeed translates to a higher end-to-end QA accuracy.`
3. Demonstrated that dense retrieval can outperform and potentially replace the traditional sparse retrieval component in open-domain question answering. **Still holds?**

Training	Retriever	Top-20 NQ	Top-20 TriviaQA	Top-20 WQ	Top-20 TREC	Top-20 SQuAD	Top-100 NQ	Top-100 TriviaQA	Top-100 WQ	Top-100 TREC	Top-100 SQuAD
None	BM25	59.1	66.9	55.0	70.9	68.8	73.7	76.7	71.1	84.1	80.0
Single	DPR	78.4	79.4	73.2	79.8	63.2	85.4	**85.0**	81.4	89.1	77.2
	BM25 + DPR	76.6	79.8	71.0	85.2	**71.5**	83.8	84.5	80.5	92.7	**81.3**
Multi	DPR	**79.4**	78.8	**75.0**	**89.1**	51.6	**86.0**	84.7	**82.9**	93.9	67.6
	BM25 + DPR	78.0	**79.9**	74.7	88.5	66.2	83.9	84.4	82.3	**94.1**	78.6

Table 2: Top-20 & Top-100 retrieval accuracy on test sets, measured as the percentage of top 20/100 retrieved passages that contain the answer. *Single* and *Multi* denote that our Dense Passage Retriever (DPR) was trained using individial or combined training datasets (all the datasets excluding SQuAD). See text for more details.