Dense Passage Retrieval for Open-Domain Question Answering

Sparse retrieval methods struggle with semantic meaning. This paper uses dense retrieval method with a pre-trained [BERT](#).

Beats 9-19% improvement.

Datasets:

- Natural Questions (NQ)
- TriviaQA
- WebQuestions (WQ)
- CuratedTREC (TREC)

Interesting ideas / notes

- they considered 3 different types of negatives: random, top passages returned by BM25 which don't contain the answer but match most question tokens, and gold. You can see that BM25 actually does better when it comes to Top-5, however once you get to top-100 its benefit just goes away
- In-batch negatives (though that seems analogous to just contrastive loss from CLIP).

Questions

What are the authors trying to do? Articulate their objectives.
The authors introduce a method to improve the performance of passage retrieval via dense retrieval, by leveraging the BERT architecture to encode text in a dense representation (an embedding) via a dual encoder framework, where the relevance of a question and answer pair is simply computed by the dot product of their embeddings.

How was it done prior to their work, and what were the limits of current practice?
Traditionally, TF-IDF or BM25 techniques were used, which are very classical (non-learning) based methods that count terms. These methods efficiently match keywords with an inverted index and represent questions and contexts in high-dimensional, sparse vectors. However, these could not correctly capture the semantic meaning of text (ex: "bad guy" and "villain" essentially mean the same thing), hence sometimes resulting in suboptimal retrieval results. Dense retrieval methods were explored in past literature, however it was generally believed that learning a good dense vector representation required a large number of labeled pairs of question and contexts.

What is new in their approach, and why do they think it will be successful?
In their approach, they introduce a Dense Passage Retriever (DPR), and leverage a BERT pre-trained model (which has already a good semantic understanding of text), and a dual-encoder structure, one to encode the question and one to encode the passages. The loss is trained on minimizing the negative log likelihood over an in-batch softmax of question passage pair, where the value is given by the dot product of the two embeddings. Because DPR can leverage pretraining from the BERT model, it requires much less question-context pairs to a good vector representation and correctly rank relevant passages.

What are the mid-term and final "exams" to check for success? (i.e., How is the method evaluated?)
They use several datasets, including Natural Questions (NQ), TriviaQA, WebQuestions (WQ), and CuratedTREC (TREC), and compare the performance of the dense passage retrieval technique against sparse retrieval techniques (BM25), and find that DPR significantly outperforms the sparse retrieval technique on the top-20 and top-100 accuracy scores. The authors also experiment with using a hybrid approach of BM25 + DPR, which uses a linear combination of the scores as a new ranking function.