Dense and Sparse Retrieval

Gaurav Rajesh Parikh

April 20, 2024

For my experiments I tried sparse, dense and hybrid approaches. I also tried different learning regimes to explore how to train the model better.

Algorithm	**Recall@5**	**Recall@20**
TF-IDF	0.552	0.819
BM 25	0.612	0.836
In Batch Fine-tuned DPR	0.707	0.888
Hard Neg DPR	0.784	0.922
5 Hard Neg DPR	0.831	0.929
10 Hard Neg DPR	0.862	0.957
Hard Neg DPR w/ Training Regime	0.836	0.940
10 Hard Negatives w/ TFIDF ranker	0.724	0.922
10 Hard Negatives w/ BM25 ranker	0.876	0.912

Table 1: Recall values for different algorithms

1 Sparse Retrieval: BM 25

The first experiment I try is to improve sparse retrieval with BM25. BM25 is an improvement over TF-IDF as it helps in avoiding situations where overly repetitive terms dominate. BM25 normalizes the term frequency based on the length of the document which means that longer documents are penalized for having higher term frequencies, preventing longer documents from having an unfair advantage in the ranking. I implemented BM25 and tried a few different values for the set of tunable parameters (k_1, b) to better suit my dataset and I settled on values 1.5 and 0.75 respectively.

2 Dense Retrieval: Multiple hard negatives

I next attempted dense retrieval using multiple hard negatives. A Dense retrieval method represents documents and queries using dense, fixed-length vectors, which we learn through the dual BERT neural network architecture. The reason that dense retrieval performs better is that it has richer semantic understanding and so the denser representations can capture these semantic relationships between terms and documents, enabling better understanding of document meaning. I believe increasing the number of negatives should perform better as by giving multiple negatives the model can learn a finer decision boundary to determine positive from negative classes. I tried first with 5 and then 10 hard negatives, selected by randomly sampling over the passages excluding those that are in the positive set. To ensure sufficient negatives, if I ran out of negative passages to sample I then randomly sampled over all passages to construct complete negative samples.

3 Sparse Ranked Dense Retrieval

TF-IDF Ranker Next, I augmented the prior method with using a base ranker to draw 10 negatives by first ranking queries with a TF-IDF ranker function and then sequentially drawing passages (ordered by the highest TFIDF score) that are not in the positive passage set. This ranking implies that passages that are assumed to be most similar to the query by the TF-IDF ranker but are actually incorrect represent “harder to distinguish” passages and so training on these should improve performance as the deep learning model learns to differentiate across these better. However, I did not observe substantial improvement in performance.

BM25 Ranker I tried the aforementioned approach now using the BM25 model instead of TF-IDF to rank. This did better than TF-IDF as expected since BM 25 does learn a better ranking so it does help to train the model on a better set of ”harder to distinguish” passages but the improvement was only minor on the test set presumably due to randomness in the runtime.

4 Learning Regime

I also implement a learning regime for each of my aforementioned methods that takes 10 negatives at the beginning and then decreases the number of negatives in every subsequent run until stabilizing on taking 5 hard negatives. The reason I expect this to do well is that initially the model is learning a broader notion of correct and incorrect - i.e. drawing a wider decision boundary and so training on multiple samples should allow for this to learn better and subsequently the model is learning the decision boundary more locally and at a finer scale and so it should do so on fewer samples. I found this approach to give me the best performance on the test set.

Conclusion Overall, I explored a variety of approaches to improve my model and found that the best model was attained by doing a combination of hard negatives and having a training regime.