Borrowing pretrained encoders for a differentiable retrieval index on images

Neha Desaraju
ndesaraju@utexas.edu

October 27, 2023

There has been much work in alternatives to traditional retrieval methods such as dual encoder and retrieve-then-rank, particularly ones that are differentiable end-to-end or perform better than normal query-to-document training methods, in which a constrastive loss-trained dual encoder attempts to map a $q \in Q$ to a ranked list of documents $\{d_1, d_2, ..., d_n\} \in D$. One such example of this work is the differentiable search index (DSI) [1], in which multiple methods of document id representations and finetuning of a pretrained transformers language model to map a document d to a *docid* is proposed and investigated. DSI achieves state-of-the-art indexing performance when compared to baselines such as BM25. Much work on retrieval of text documents has been done based on the work on DSI, but thus far there is very little applying this technique to multiple modalities of input corpus. In fact, some of the most important investigations done in the paper are not the archiecture itself, but its method of mapping documents to docids such that the document index is stored in the weights of the language model. In addition, the DSI model is extensible enough to be able to borrow pretrained model weights (in the paper, T5 is used).

Notably, [1] demonstrates that a naive docid representation, in which the generation of docids is performed by a sequential decoder, works quite well. Using a "naive" representation of docids, where the model aims to maximize the likelihood the generation of token t_i, $p(t_i|\text{document}, t_1, t_2, ..., t_{i-1})$. It is trained mostly on $(\text{document}, \text{docid})$ pairs, with a few $(\text{query}, \text{docid})$ examples, rather than finetuning with query-docid pairs.

In fact, I have defined the problem as training a model to predict an image id given an image, as well as predicting an image id given a text query. To do so, I propose that two pretrained encoders, the CLIP text and vision encoders, are finetuned in a manner similar to the one proposed in the DSI paper, so that they are both jointly trained to decode the correct learned image id. A similar method is utilized by IRGen [2] for the case of images themselves being used as query images for other images (the base model is a single pretrained ViT encoder), but otherwise the training methods and model setup are very similar.

I was able to recreate the original DSI paper on a much smaller corpus of 1k documents (from the Natural Questions dataset) and successfully trained it for 1000 epochs. While it currently has a comparitively low accuracy, I hope to train it for much longer (for $100N$ steps, as they do in the paper) to demonstrate that I am able to recreate the methods in the paper. Further, I also hope to test and compare my new text-image querying model to datasets such as Flickr30k and COCO Captions.

References

[1] Yi Tay, Vinh Q. Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. Transformer memory as a differentiable search index, 2022.

[2] Yidan Zhang, Ting Zhang, Dong Chen, Yujing Wang, Qi Chen, Xing Xie, Hao Sun, Weiwei Deng, Qi Zhang, Fan Yang, Mao Yang, Qingmin Liao, and Baining Guo. Irgen: Generative modeling for image retrieval, 2023.