Borrowing pretrained encoders for a differentiable retrieval index over image corpuses

Neha Desaraju
University of Texas at Austin
ndesaraju@utexas.edu

Abstract

Tradition retrieval methods, such as dual encoders, require extensive training and data collection and are not as efficient or performant as end-to-end differentiable retrieval systems. In addition, they are difficult to extend to multimodal corpuses. Based on the recent successes of architectures such as the Differentiable Search Index (DSI), which introduces a differentiable architecture with few query-document training data points needed over a text-only corpus, we propose CLIP-DSI, which involves initializing a CLIP image and text encoder with pretrained weights to obtain embeddings. These embeddings can then be used to predict image ids using a finetuned language model. While we do not achieve results close to the baseline, we demonstrate that it may be possible to utilize this paradigm in novel ways and suggest improvements to our architecture and training methods in order to make up for this gap.

Introduction

Currently, common retrieval methods include dual encoders or an embedding model (Menon et al., 2022), where the embeddings generated between a query and documents in a corpus are ranked based on distance. In the past year, there have been many advances in differentiable end-to-end retrieval systems, which generally perform better than dual encoder systems and tend to be more efficient at inference time, because the target documents are output by the model itself (ranking and reranking is no longer necessary). In addition, training dual encoder or embedding models on custom corpuses requires lots of query-document pairs, and are thus data-inefficient. Methods such as those proposed by Tay et al. (2022) involve finetuning a pretrained language model on a document corpus itself, such that the model stores a compression of the corpus itself in its weights.

Tay, et al. finetune the T5 language model (Raffel et al., 2019) on (document, document id) pairs, which forces the model to learn an internal representation on the corpus, along with a few (query, document id) pairs. They explore three ways of generating useful document ids, including one "naive" method, which simply involves using random document ids. This method works surprisingly well, outperforming traditional search baselines such as BM25 (Robertson and Zaragoza, 2009). (Other methods they discuss include generation of semantic document ids as a separate unsupervised task by a heirarchical clustering algorithm.)

This new differentiable search index (DSI) paradigm has led to some, but overall sparse, improvements in end-to-end retrieval on image or multimodal corpuses. Some approaches (Zhang et al., 2023) support image-to-id queries; that is, an image serves as the query and the task is to find image ids from the corpus that are similar. This approach also involves using a pretrained image encoder, such as VQ-VAE, and training an autoregressive encoder-decoder model on top using the resulting tokens to generate image ids. However, the standard for text-to-image retrieval remains jointly-trained dual encoder models such as CLIP (Radford et al., 2021), which generates separate image and text embeddings that can be compared, and documents can be ranked using a dot product similarity score between embeddings. CLIP in particular has also been shown to perform well in generating captions (Mokady et al., 2021), where a GPT2 decoder (Brown et al., 2020) is finetuned on top of a pretrained, frozen CLIP vision encoder. It has also been shown that CLIP can be an excellent encoder for a variety of tasks that do not have pretrained decoders, such as in segmentation (with both the image and text serving as input) (Lüddecke and Ecker, 2022). In such cases, the decoder is typically a simple decoder trained from scratch for that particular task.

BLIP-2 (Li et al., 2023) is another model that successfully implements the frozen multimodal encoder to text decoder paradigm. What is notable about BLIP-2 is that it freezes both the encoder and language decoder, and only trains a "Q-former", which they describe as a projection mapping between the latent space of the encoder and the target space of the decoder, in order to match the dimensions.

Methodology

In this paper, we propose **CLIP-DSI** [1], which operates over a corpus of images using a frozen CLIP encoder that generates image and text embeddings (the latter for the queries themselves). We finetune a T5 model (Raffel et al., 2019) that uses the resulting embedding as input to generate the given document ids. Assuming the text and image embeddings are close to each other, we hypothesize that the T5 model should successfully generate ids given either type of embedding. Figure 1 illustrates the basic architecture of the model.

Because we did not incorporate a mapping between the dimensions of the encoder and language model as BLIP does, we must ensure the output dimensions of the CLIP embeddings and the input dimensions of the T5 are the same. For that reason, we used the small T5 model along with the CLIP-32 model.

For simplicity, we use the naive method of image id generation as proposed in the original DSI paper, though we presume that the other, more complex methods proposed (such as separately learning a semantic tree to generate ids) would perform better or be much more efficient during training. Beam search constrained to integer token outputs is used to find the top ranked model outputs during inference.

Model	**Top1 %**	**Top5 %**	**Top10 %**
CLIP	0.74	0.94	0.98
CLIP-DSI	0.13	0.26	0.29

Table 1: Comparison between our model and the baseline.

Experimental Results

We trained on a subset of the **Flickr30k** dataset, which consists of 30k images and five human-generated captions for each. Our training set consists of a corpus of 400 images from this dataset, all of which are represented exactly once as image-id pairs in the training data. Tay, et al. note that the best performance is achieved when only a fraction (around 20%) of the training data consists of queries. As such, 100 of the images in our corpus also have a query-id pair in the training data (making the training data 500 total points), and of the 300 that do not have query-id representations, we select 50 of them to be used in our validation set (which exclusively consists of query-id pairs, and no images). We trained for around 500k steps. Our results compared to the CLIP baseline are outlined in Table 1.

On our validation set, CLIP-only retrieval — matching and ranking query and image embeddings — performs extremely well. Comparatively, our model does not perform nearly as well on the validation set. We noticed that the accuracy seemed to increase very suddenly, at around 400k steps, but did not notice that there were any significant improvements after this, so it is unlikely that increasing the number of training iterations alone would suggest an improvement in model performance.

In addition, we saw that using numbers between 1-400 for our dataset (as opposed to random numbers) helped the model to train faster initially, but we do not have any significant longterm data to support this. This may be because the model simply learns that a good guess is typically between 1 and 400, the size of our dataset, and somewhat constrains future guesses. We did not test if choosing completely random numbers for image ids would change the results in the long term, but this is an area for future study.

[1] The implementation can be found here: https://github.com/estaudere/dsi/tree/main/clip-dsi.

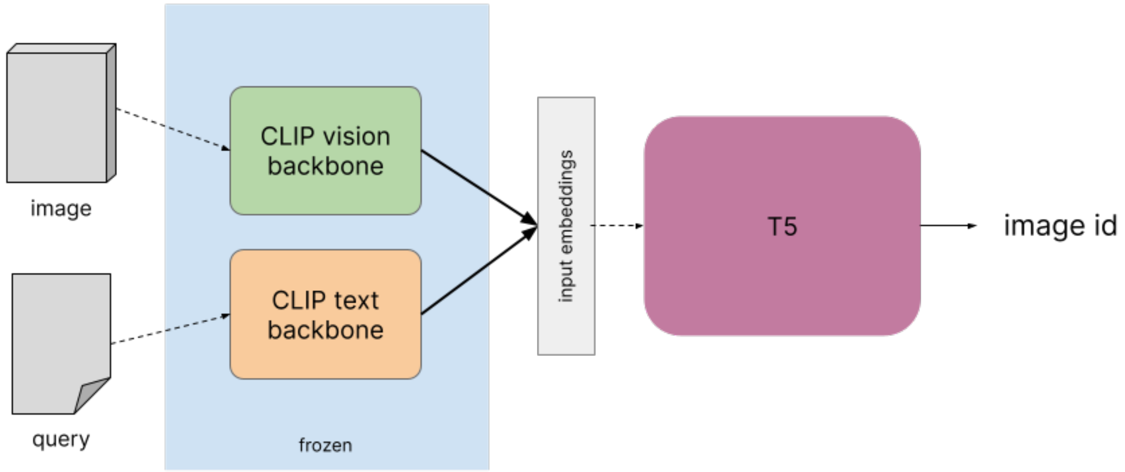

Figure 1: CLIP-DSI model architecture.

Conclusion

While our method did not perform at or better than the current benchmark for retrieval, we point out that this architecture may be a much more ideal method when finetuning on more precise corpuses or tasks; for example, a corpus of medical images. In addition, this architecture automatically lends itself to retrieval on multimodal corpuses; the prerequesite for the implementation of such a model would be a suitable encoder for each type of input.

There are several areas of improvement on this model. For one, it would be experimentally pertinent to train on a much larger corpus; the evaluations for DSI involve training on 10k, 100k, and 320k training samples from the Natural Questions dataset. This methodolgy may outperform CLIP on a larger corpus of images, where the language model can accurately learn small differences in embeddings that the CLIP model on its own is unable to differentiate.

Secondly, deciding on the ratio of image training pairs to query training pairs is an important piece of training the model. Since our multimodal model must learn both language and image representations rather than just text, it may be worthwhile to experiment with flipping the ratio. Our model is also limited in the amount of information it can store. By projecting into a lower embedding dimension and passing this into our language model, we lose crucial information in the latent space of the encoder. However, we must either use a pretrained language model whose internal dimensions match CLIP's, as in CLIPCap (Mokady et al., 2021), or implement a mapping function between the latent space of the encoder directly to the language model's decoder, as done in BLIP.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

Timo Lüddecke and Alexander S. Ecker. 2022. Image segmentation using text and image prompts.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Seungyeon Kim, Sashank J. Reddi, and Sanjiv Kumar. 2022. In defense of dual-encoders for neural ranking. In *International Conference on Machine Learning*, volume 39.

Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. Clipcap: Clip prefix for image captioning.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

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. 2022. Transformer memory as a differentiable search index.

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. 2023. Irgen: Generative modeling for image retrieval.