别再双塔了！谷歌提出DSI索引，检索效果吊打双塔，零样本超BM25！

原创 jxyxiangyu 夕小瑶的卖萌屋 2022-02-21 11:55

收录于话题
#谷歌 1 #信息检索 1 #NLP 1

卖萌屋今日学术精选

这篇论文展示了信息检索可以用一个Transformer来完成，其中，关于语料库的所有信息都被编码在Transformer模型的参数中。

论文标题：
Transformer Memory as a Differentiable Search Index
链接：
https://arxiv.org/abs/2202.06991

作者提出了可微搜索索引（Differentiable Search Index，DSI)的概念，这是一种新的搜索范式，它可以学习出一个Query-to-DocID的文本检索模型，将用户Query直接映射到相关的DocID节点上；换句话说，DSI模型直接使用其模型参数来回答用户查询，极大地简化了整个检索过程。

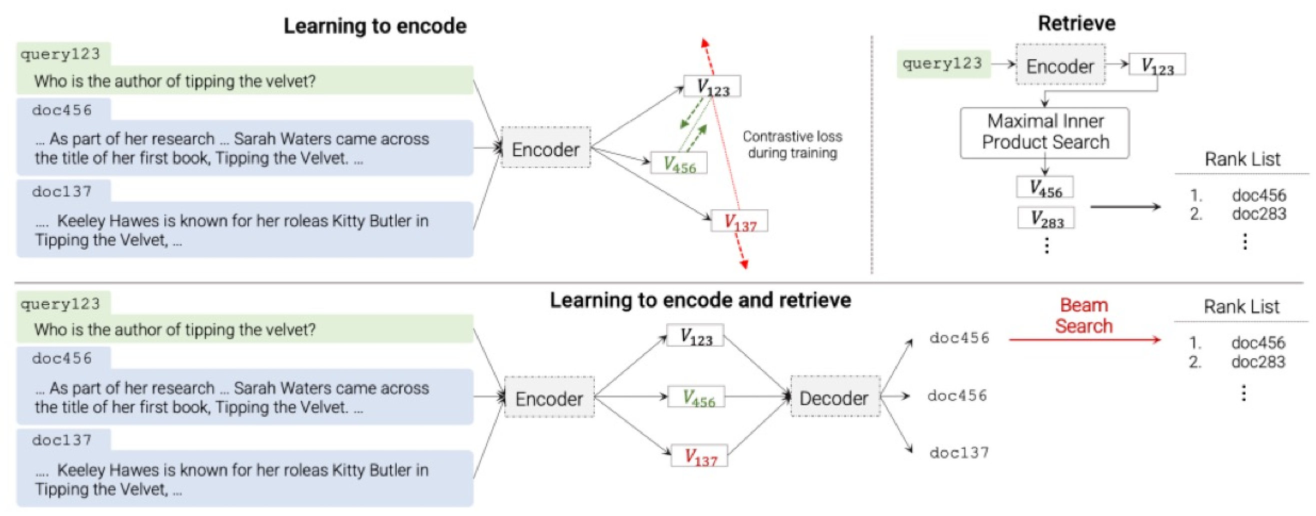

Figure 1: Top, an overview of dual encoders (DE), that has two independent steps, encoding and retrieval. In DEs, the queries q_i and documents d_j are mapped to vectors $\mathbf{v}_i$ and $\mathbf{v}_j$ in a common space. In the retrieval stage, the documents relevant to a given query can be found by performing a maximal inner product search (MIPS) to find the k documents d_j with largest inner product with the query $\mathbf{v}_i^T\mathbf{v}_j$. Bottom, the same task solved with a differentiable search index (DSI), which unifies the encoding and retrieval. In DSI, the docids (represented as strings) are generated directly from the query using a seq-to-seq model, and a rank list of k document is created incorporating the beam search mechanism during decoding.

上图展示了经典的双塔模型（Dual Encoder）+最大内积检索（MIPS）的经典检索范式，与本文提出的可微搜索索引（DSI）的范式的区别。后者统一了模型的训练与检索。

实验结果

首先作者在不同规模的NQ数据集上，检验了DSI模型的supervised learning能力。

Table 3: Experimental results on NQ document retrieval. DSI outperforms BM25 and Dual Encoder baselines. Among all the Docid representation methods, Semantic String Docids perform the best.

				NQ10K		NQ100K		NQ320K	
Model	Size	Params	Method	Hits@1	Hits@10	Hits@1	Hits@10	Hits@1	Hits@10
BM25	-	-	-	12.4	33.5	20.9	46.4	11.6	34.4
T5	Base	220M	Dual Encoder	16.2	48.6	18.7	55.2	20.5	58.3
T5	Large	800M	Dual Encoder	18.8	55.7	22.3	60.5	22.4	63.3
T5	XL	3B	Dual Encoder	20.8	59.6	23.3	63.2	23.9	65.8
T5	XXL	11B	Dual Encoder	22.1	61.6	24.1	64.5	24.3	67.3
DSI	Base	250M	Atomic Docid	13.0	38.4	23.8	58.6	20.7	40.9
DSI	Large	800M	Atomic Docid	31.3	59.4	17.1	52.3	6.9	27.3
DSI	XL	3B	Atomic Docid	40.1	76.9	19.0	55.3	28.1	61.9
DSI	XXL	11B	Atomic Docid	39.4	77.0	25.3	**67.9**	24.0	55.1
DSI	Base	250M	Naive String Docid	28.1	48.0	18.7	44.6	6.7	21.0
DSI	Large	800M	Naive String Docid	34.7	60.5	21.2	50.7	13.3	19.9
DSI	XL	3B	Naive String Docid	44.7	66.4	24.0	55.1	16.7	58.1
DSI	XXL	11B	Naive String Docid	46.7	**77.9**	**27.5**	62.4	23.8	55.9
DSI	Base	250M	Semantic String Docid	33.9	57.3	19.0	44.9	27.4	56.6
DSI	Large	800M	Semantic String Docid	37.5	65.1	20.4	50.2	35.6	62.6
DSI	XL	3B	Semantic String Docid	41.9	67.1	22.4	52.2	39.1	66.8
DSI	XXL	11B	Semantic String Docid	**48.5**	72.1	26.9	59.5	**40.4**	**70.3**

从上表可以看到，DSI模型经过finetune之后，强势吊打了BM25基线和同样finetune之后的T5模型。

此外，作者还在NQ数据集上检验了DSI模型的zero-shot能力。

Table 4: Experimental results on Zero-Shot NQ document retrieval. DSI outperforms BM25, T5 embeddings and SentenceT5, the state-of-the-art for unsupervised similarity modeling. Among Docid representation method, the Atomic Docid performs the best on zero-shot learning.

			NQ10K		NQ100K		NQ320K	
Model	Size	Method	Hits@1	Hits@10	Hits@1	Hits@10	Hits@1	Hits@10
BM25	-	-	12.4	33.5	20.9	46.4	11.5	33.7
T5	XXL	Dual Encoder	0.3	1.3	1.9	8.0	1.1	5.9
SentenceT5	Large	Dual Encoder	17.6	50.7	17.4	50.8	16.9	51.0
DSI	XXL	Atomic Docid	25.7	60.1	**23.0**	**57.3**	**25.1**	**56.6**
DSI	XXL	Naive String Docid	43.4	67.4	17.4	41.5	9.2	22.6
DSI	XXL	Semantic String Docid	**43.9**	**68.8**	11.4	26.6	13.9	31.1

众所周知，BM25是zero shot方面非常高的一个基线，从上表可以看出，DSI的zero shot能力也显著优于BM25。

实验表明，给定适当的设计选择，DSI不仅显著优于双塔模型为代表的强基线模型，此外，DSI展示了很强的泛化能力，在zero-shot实验中显著优于BM25基线。