Curriculum Learning for Dense Retrieval Distillation

Hansi Zeng [1] Hamed Zamani [2] Vishwa Vinay [2]

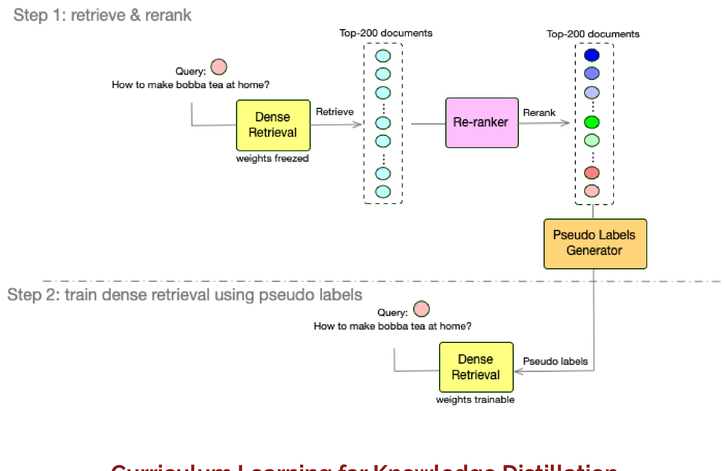
[1]University of Massachusetts Amherst [2]Adobe Research

Overview

We create a curriculum learning based generic optimization framework called CL-DRD that controls the difficultly level of training data produced by the re-ranker (teacher). CL-DRD iteratively optimizes the dense retrieval model (student) by increasing the difficulty of knowledge distillation data made available to it.

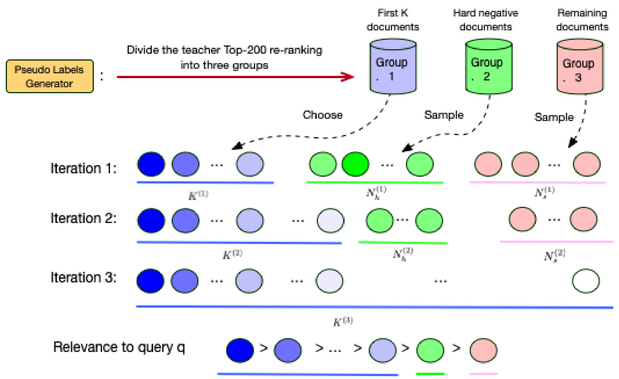
Knowledge Distillation for Dense Retrieval

The procedure of distilling the knowledge from re-ranker (teacher) to the retriever (student).

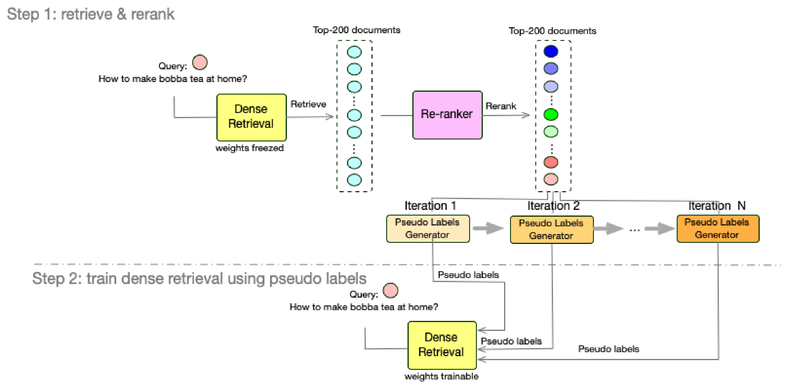


Curriculum Learning for Knowledge Distillation

- **Challenge**: The re-rankings of the teacher give us a great flexibility to generate pseudo relevance labels for knowledge distillation. The most straightforward way is the "identical generator" where the student model would learn the exact top-200 re-ranking of the teacher. However, student and teacher models have different architectures and capacities, the identical generator might lead to the sub-optimal performance.
- **Solution**: Motivated by the curriculum learning, We iteratively optimize the student model by controlling the difficulty level of pseudo relevance labels generated in each iteration. The pseudo label generator for each iteration is illustrated in the following figure:



The pseudo labels generated in early iterations can be easy and the coarse approximations of exact teacher re-rankings. After the student make progress and gradually builds up its capacity, the pseudo labels in late iterations would be more closer to the exact teacher re-rankings with higher difficulty level. The whole CL-DRD (Curriculum Learning for Dense Retrieval Distillation) is illustrated in the following figure:



Experiments

Datasets: We train our models on MS MARCO-Train dataset, and evaluate their performance on MS MARCO-Dev, TREC'DL-19 and TREC'DL-20 datasets.
The CL-DRD Models: We augment two dense retrieval models by using CL-DRD: (1) **TAS-B + CL-DRD** (single-vector retrieval model). (2) **ColBERTv2 + CL-DRD** (multi-vectors retrieval model). The performance comparison with other baseline models is as follows:

Model	KD	MS MARCO DEV MRR@10	MS MARCO DEV MAP@1k	TREC-DL'19 nDCG@10	TREC-DL'19 MAP@1k	TREC-DL'20 nDCG@10	TREC-DL'20 MAP@1k
Sparse Retrieval							
BM25	-	.187	.196	.497	.290	.487	.288
DeepCT	-	.243	.250	.550	.341	.556	.343
docT5query	-	.272	.281	.642	.403	.619	.407
Multi-Vector Dense Retrieval							
ColBERT	✗	.360	-	-	-	-	-
ColBERTv2	✓	.384	.389	**.733**	.464	.712	.473
ColBERTv2 + CL-DRD (Ours)	✓	.394	**.398**	.727	**.472**	**.717**	**.487**
Single-Vector Dense Retrieval							
ANCE	✗	.330	.336	.648	.371	.646	.408
ADORE	✗	.347	.352	.683	.419	.666	.442
RocketQA	✓	.370	-	-	-	-	-
TCT-ColBERT	✓	.335	.342	.670	.391	.668	.430
Margin-MSE	✓	.325	.331	.699	.405	.645	.416
TAS-B	✓	.344	.351	.717	.447	.685	.455
TAS-B + CL-DRD (Ours)	✓	**.382**	**.386**	**.725**	**.453**	**.687**	**.465**

Ablation Study: We plot the performance of TAS-B + CL-DRD after each curriculum iteration on three datasets.