ADAM: Dense Retrieval Distillation with Adaptive Dark Examples

Chongyang Tao[1*†], **Chang Liu**[2*], **Tao Shen**[3], **Can Xu**[2], **Xiubo Geng**[4], **Binxing Jiao**[4] and **Daxin Jiang**[4]
[1]SKLSDE Lab, Beihang University [2]Peking University
[3]AAII, FEIT, University of Technology Sydney [4]Microsoft
[1]chongyang@buaa.edu.cn [2]{changliu,canxu}@pku.edu.cn
[3]tao.shen@uts.edu.cn [4]{xigeng,bxjiao,djiang}@microsoft.com

Abstract

To improve the performance of the dual-encoder retriever, one effective approach is knowledge distillation from the cross-encoder ranker. Existing works prepare training instances by pairing each query with one positive and a batch of negatives. However, most hard negatives mined by advanced dense retrieval methods are still too trivial for the teacher to distinguish, preventing the teacher from transferring abundant *dark knowledge* to the student through its soft label. To alleviate this issue, we propose ADAM, a knowledge distillation framework that can better transfer the dark knowledge held in the teacher with Adaptive Dark exAMples. Different from previous works that only rely on one positive and hard negatives as candidate passages, we create dark examples that all have moderate relevance to the query by strengthening negatives and masking positives in the discrete space. Furthermore, as the quality of knowledge held in different training instances varies as measured by the teacher's confidence score, we propose a self-paced distillation strategy that adaptively concentrates on a subset of high-quality instances to conduct our dark-example-based knowledge distillation to help the student learn better. We conduct experiments on two widely-used benchmarks and verify the effectiveness of our method.

1 Introduction

Information retrieval (IR) that aims to identify relevant passages for a given query is an important topic for both academic and industrial areas, and has powered many downstream tasks such as open-domain QA (Chen et al., 2017) and knowledge-grounded conversation (Dinan et al., 2018). Typically, IR systems usually follow the retrieve-and-re-rank paradigm (Hofstätter et al., 2020; Huang et al., 2020; Zou et al., 2021) where a fast retriever first retrieved a bundle of relevant passages from a large-scale corpus through pre-built indices and then a more sophisticated ranker comes to re-rank these candidate passages to further obtain more accurate retrieval results.

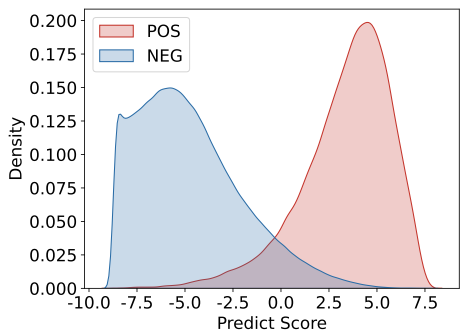


Figure 1: Distributions of the prediction for the cross-encoder of R^2anker (Zhou et al., 2023) over MS-MARCO. POS and NEG mean the distribution of positive and hard negatives respectively. The hard negatives are provided by RocketQAv2 (Ren et al., 2021c).

Under this paradigm, recent years have witnessed a growing number of works that utilize pre-trained language models (PLMs) (Qu et al., 2021; Gao and Callan, 2021b) as retrievers and rankers to build IR systems. Among these efforts, there are two commonly adopted architectures: cross-encoder (Devlin et al., 2019a) that measure the relevance of a query-passage pair through jointly modeling their deep interactions; dual-encoder (Karpukhin et al., 2020; Qu et al., 2021) that encodes queries and passages separately into dense representations and calculate the similarity. Although dual-encoders are efficient for billions of indices, they suffer from inferior performance compared with cross-encoders since they can't capture the fine-grained semantic relevance between the query and the passage due to the absence of their deep interactions (Luan et al., 2021a). To help dual-encoders achieve better retrieval performance, a common practice is to draw on the

[*] Equal Contribution.
[†] Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2024, pages 11639–11651

powerful but cumbersome cross-encoder through knowledge distillation (Yang et al., 2020; Zhang et al., 2022; Ren et al., 2021c; Zeng et al., 2022; Lin et al., 2023). Along this line of research, various techniques are proposed to improve the knowledge transfer including data curriculum (Lin et al., 2023; Zeng et al., 2022), on-the-fly distillation (Zhang et al., 2022; Ren et al., 2021c) and new distillation objectives (Lu et al., 2022; Menon et al., 2022).

Though effective, we argue that existing dense retrieval distillation methods may not fully exploit the dark knowledge deeply held by the teacher. In knowledge distillation (Xu et al., 2018; Lin et al., 2023), the student learns not just the highest-scored class from the soft labels provided by the teacher, but also the entire probability distribution over classes, as this contains comprehensive fine-grained information referred to as "dark knowledge". However, we empirically find that for existing distillation methods, the soft labels (i.e., the probability distributions over one positive and multiple negatives for a query) given by the teacher are too "sharp", despite they already adopted hard negatives (Ren et al., 2021c). As illustrated in Figure 1, we draw the score distributions of the positive and negative pairs using a pre-trained cross-encoder teacher. It can be observed that the scores for most hard negatives are quite low (concentrated in (-7.5, -2.5)) and distributed far from the positives that have high scores. A similar observation is also drawn by Menon et al. (2022). This phenomenon indicates that even the hard negatives mined by the dense retriever are still too trivial for a well-trained cross-encoder teacher to distinguish, losing most of the utile dark knowledge.

To alleviate this issue, we propose ADAM, a knowledge distillation framework that can better exploit dark knowledge deeply held in the teacher by distillation with adaptive dark examples. Our method originated from the intuition that a good soft label for the retriever to learn should be more smooth, which implies that the provided query-passage pairs should diversely distribute from highly-relevant pairs to loosely-relevant pairs from the view of the teacher. To fill the gap between highly-relevant pairs and loosely-relevant pairs existing in current negative sampling methods, we propose two approaches to construct dark examples that all have moderate relevance to the query. The first approach is to make negatives more relevant to the query by strengthening the negatives with the positive passage. The second approach is to make positives less relevant to the query by replacing some randomly selected tokens with mask tokens. Considering that the newly created passages have moderate relevance to the query, we believe it is more appropriate to call them dark examples instead of negatives. With these dark examples added, we successfully make the score distribution smoother as shown in Figure 3(b), so that we can transfer more useful dark knowledge from the teacher. Moreover, since the soft label for different query-positive-negatives have different "sharpness" which we consider as an indication of how well the dark knowledge has been exploited, we further propose a self-paced distillation strategy that adaptively selects those examples whose soft labels are sharp to conduct our dark-example-based distillation to better transfer the dark knowledge.

We conduct experiments on two benchmarks, including MS-MARCO (Nguyen et al., 2016) and TREC Deep Learning 2019 (Craswell et al., 2020). In both benchmarks, the model is required to select the best response from a candidate pool. Evaluation results indicate that our method is significantly better than existing models on two benchmarks. To sum up, our contributions is three-fold:

- Propose to augment dark examples including reinforced negatives and noisy positives for more effective knowledge distillation in IR;
- Propose to adaptively concentrate on high-confidence training instances to better transfer knowledge;
- Empirical verify of the effectiveness of the proposed approach on two public datasets.

2 Related Works

There are two lines of research related to our work: dense retriever and knowledge distillation.

Dense Retriever. To overcome the vocabulary and semantic mismatch problems existing in conventional term-based approaches such as BM25 (Robertson and Zaragoza, 2009), researchers began to build neural retrievers upon pre-trained language models (Devlin et al., 2019b; Liu et al., 2019). In this way, the whole input text can be represented as a dense vector in a low-dimensional space (e.g., 768) and efficient retrieval can be achieved by approximate nearest neighbor search (ANN) algorithms such as FAISS (Johnson et al., 2019). To learn a good dense retriever, various attempts have been made including hard negative mining (Karpukhin et al., 2020; Luan et al., 2021a; Qu et al., 2021; Xiong et al., 2021; Zhan et al., 2021a), retrieval-oriented pre-training (Lee et al., 2019; Gao and Callan, 2021a,b), knowledge distillation (Ren et al., 2021c; Zhang et al., 2022; Lu et al., 2022; Zhang et al., 2023), etc. We mainly focus on knowledge distillation in this paper.

Knowledge Distillation. Knowledge distillation (Hinton et al., 2015; Xu et al., 2024) aims to transfer the knowledge from a powerful teacher model to a student model to help it learn better. To achieve this goal, the student model is provided with the teacher's outputs as the supervision signal that it is enforced to mimic. There are multiple types of supervision signals for the student to learn, including the teacher's output logits (Hinton et al., 2015), intermediate representations (Romero et al., 2014), relations of representations (Park et al., 2019), etc. In the context of dense retrieval distillation, researchers basically adopt the cross-encoder as the teacher and use the teacher's probability distribution over candidate passages as the supervision signal. On this basis, several studies (Ren et al., 2021c; Zhang et al., 2022; Lu et al., 2022) explored on-the-fly distillation to jointly optimize the teacher and the student, Zeng et al. (2022) and Lin et al. (2023) combined knowledge distillation with curriculum strategies to gradually improve the student. Different from existing work, we focus on the quality of knowledge held in the teacher's soft label and propose to distill with adaptive dark examples to better transfer the dark knowledge to the student.

3 Methodology

In this section, we first introduce the preliminaries in dense retrieval distillation, then present our dark example augmentation method and adaptive distillation with dynamic data selection.

3.1 Preliminary

Task Description In this work, we study the learning of the dense retriever following the general setting of dense retrieval in existing work (Qu et al., 2021; Ren et al., 2021c; Zhang et al., 2022). Formally, there is a training set $\mathcal{D} = \{(q_i, \mathbb{P}_i)\}_{i=1}^{n}$ where q_i is the query and $\mathbb{P}_i$ is the set of candidate passages. Commonly, $\mathbb{P}_i$ consists of a positive passage p_i^+ and m negative passages $\mathbb{P}_i^- = \{p_{i,j}^-\}_{j=1}^{m}$ constructed by random negative sampling (Henderson et al., 2017; Gillick et al., 2018) or hard negative mining (Xiong et al., 2020; Karpukhin et al., 2020; Qu et al., 2021). Based on $\mathcal{D}$, we aim to learn a retriever that can select the most relevant passage from the whole candidate pool.

Dual-Encoders A typical text retrieval system adopts the retrieve-and-rank paradigm, where the retriever is responsible for collecting a bubble of candidate passages and the ranker further re-ranks them. Considering the trade-off between efficiency and accuracy, dual-encoders (Karpukhin et al., 2020; Qu et al., 2021; Cai et al., 2022) are often chosen as the retriever while cross-encoders (Devlin et al., 2019b) are usually adopted as the ranker.[1]

The dual-encoder-based retriever $\mathsf{Enc}_{\mathsf{de}}$ is responsible for encoding the given query q_i and each of the candidate passage p_j into dense vectors $\mathsf{Enc}_{\mathsf{de}}(q_i), \mathsf{Enc}_{\mathsf{de}}(p_j) \in \mathbb{R}^h$. Then the relevance score for q_i and p_j is simply calculated as the inner product of their representations:

$$\mathcal{R}_{\mathsf{de}}(q_i, p_j) = \mathsf{Enc}_{\mathsf{de}}(q_i)^\top \cdot \mathsf{Enc}_{\mathsf{de}}(p_j). \quad (1)$$

To fulfill this goal, the retriever is typically trained with supervised contrastive loss:

$$\mathcal{L}_{\mathsf{sup}} = -\log \frac{\exp^{\mathcal{R}_{\mathsf{de}}(q_i, p_i^+)}}{\exp^{\mathcal{R}_{\mathsf{de}}(q_i, p_i^+)} + \sum\limits_{p_{i,j}^- \in \mathbb{P}_i^-} \exp^{\mathcal{R}_{\mathsf{de}}(q_i, p_{i,j}^-)}}.$$

where p_i^+ is the labeled positive document paired with q_i and $\mathbb{P}_i^-$ denotes the set of candidate documents for q_i which is typically constructed during training by random negative sampling or hard negative mining methods.

Cross-Encoders The cross-encoder ranker $\mathsf{Enc}_{\mathsf{ce}}$ is in charge of calculating the matching score of q_i and p_j more accurately as it can model their fine-grained interactions, and re-ranking the retrieved candidate passages provided by the retriever to improve the retrieval results. Concretely, given a query q_i and a passage p_j, the input is formed as the concatenation of q and p with [CLS] in the beginning and [SEP] as their separation and is fed into transformer (Vaswani et al., 2017). The representation of [CLS] in the top layer is used to calculate the relevance score with a projection head $f(\cdot)$:

$$\mathcal{R}_{\mathsf{ce}}(q_i, p_j) = f(\mathsf{Enc}_{\mathsf{ce}}([\mathtt{CLS}], q_i, [\mathtt{SEP}], p_j)). \quad (2)$$

[1]We will use retriever and dual-encoder interchangeably.

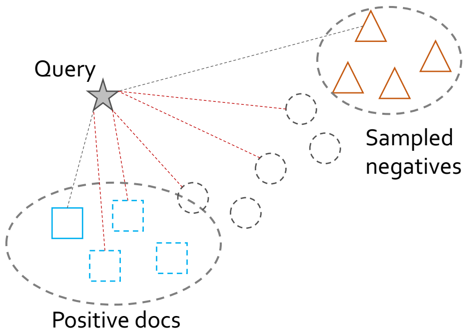


Figure 2: Illustration of dark examples. The solid rectangle and triangles mean the gold passage and the negative passages respectively. Dotted rectangles and circles denote noisy positives and mixed samples respectively.

Knowledge Distillation in IR As cross-encoders are more capable of measuring the relevance of q_i and p_j than dual-encoders but at a cost of computational inefficiency, it's promising to transfer the knowledge from the strong cross-encoders to the weak dual-encoders through knowledge distillation (Zhang et al., 2022; Ren et al., 2021c; Zeng et al., 2022; Lu et al., 2022; Lin et al., 2023). In dense retrieval distillation, as both the positive passage p_i^+ and the negatives $\mathbb{P}_i^-$ can be treated uniformly, we use $\mathbb{P}_i = \{p_i^+\} \cup \mathbb{P}_i^-$ to denote the whole candidate set of passages. The relevance score of q_i and each $p_j \in \mathbb{P}_i$ can be calculated using a dual-encoder $\mathsf{Enc}_{\mathsf{de}}$ and a cross encoder $\mathsf{Enc}_{\mathsf{ce}}$ using Eq. 1 and Eq. 2. Then, the probability distributions over candidate passages of the dual-encoder and the cross-encoder $\boldsymbol{p}_{de,i}, \boldsymbol{p}_{ce,i} \in \mathbb{R}^{|\mathbb{P}_i|}$ are calculated by normalizing the relevance scores over $\mathbb{P}_i$, where each element is calculated as:

$$\hat{\mathcal{R}}^j_{de,i} = \frac{\exp^{\mathcal{R}_{\mathsf{de}}(q_i,p_j)}}{\sum_{p_k \in \mathbb{P}_i} e^{\mathcal{R}_{\mathsf{de}}(q_i,p_k)}}$$
$$\hat{\mathcal{R}}^j_{ce,i} = \frac{\exp^{\mathcal{R}_{\mathsf{ce}}(q_i,p_j)}}{\sum_{p_k \in \mathbb{P}_i} \exp^{\mathcal{R}_{\mathsf{ce}}(q_i,p_k)}}. \quad (3)$$

To distill the knowledge from the cross-encoder to the dual-encoder, the distribution of the cross-encoder $\hat{\mathcal{R}}_{ce,i}$ is considered as the soft label that guides the learning of the dual-encoder by minimizing the KL-divergence between $\hat{\mathcal{R}}_{ce,i}$ and $\hat{\mathcal{R}}_{de,i}$:

$$\mathcal{L}_{kd} = -\sum_{(q_i,\mathbb{P}_i)\in\mathcal{D}} \text{KL-Div}(\hat{\mathcal{R}}_{ce,i}||\hat{\mathcal{R}}_{de,i}) \quad (4)$$

3.2 Dark Examples Construction

When transferring the knowledge from the cross-encoder teacher to the dual-encoder student using Eq. 4, the set of candidate passages $\mathbb{P}_i$ plays a vital role. Previous works in dense retrieval distillation (Zhang et al., 2022; Ren et al., 2021c; Zeng et al., 2022; Lu et al., 2022; Lin et al., 2023) simply follow the supervised learning setting where they utilize $\mathbb{P}_i = \{p_i^+\} \cup \mathbb{P}_i^-$ as the candidate set. However, by empirical analyses on Fig. 1, we have found that the negative set $\mathbb{P}_i^-$ produced by existing hard negative mining approaches (Qu et al., 2021) is too trivial for the cross-encoder teacher, which makes the soft label provided by the cross-encoder teacher too sharp at the positive passage and therefore prevents the student from learning utile dark knowledge hidden in the distribution of other passages (i.e., negatives).

We suppose smoother soft labels naturally obtained (instead of scaled by softmax temperature) can be better knowledge carriers that transfer the dark knowledge. Given the teacher and the query, we point out that the natural way to smoothen the soft label is to operate on the set of candidate passages, or more precisely, to replace the original set of candidate passages $\mathbb{P}_i$ that are either too relevant or too irrelevant from the teacher's view with new ones $\tilde{\mathbb{P}}_i$ whose relevance to the query cannot be easily tell apart by the cross-encoder teacher.

To construct the new set of candidate passages that satisfy this desired characteristic, we propose two dual approaches that operate on the original positive passage p_i^+ and the negative set $\mathbb{P}_i^-$ respectively. We name the newly constructed passages in $\tilde{\mathbb{P}}_i$ dark examples to demonstrate that can no longer be simply categorized into positives and negatives as they have moderate relevance to the query. An illustration of dark examples is shown in Figure 2. It should be noticed that it is the specific setting of knowledge distillation where the supervision signal is derived from the teacher's soft label instead of human labels that make it possible to learn from dark examples.

Sampled Negatives. Early works (Henderson et al., 2017; Gillick et al., 2018) randomly choose negative passages by considering the passages of other query-passage pairs within the same mini-batch as the negatives. More recently, researchers use BM25 (Karpukhin et al., 2020) or dual-encoders (Xiong et al., 2020) to select hard negatives globally from the whole candidate passages with the fast retrieval method (Qu et al., 2021; Ren et al., 2021c). We will compare the effectiveness of random negatives (denoted as Rand) and hard negatives (denoted as Hard) with our method (denoted as Dark) in experiments.

Dark Examples with Reinforced Negatives The reasonable way to create dark examples based on $\mathbb{P}_i^-$ is to make hard negatives harder, or in other words, more relevant to the query. To achieve this goal, it is non-trivial to accurately edit the semantics of a negative passage towards increasing its relevance to the query with controllable text generation techniques. Instead, we propose a rather simple yet effective approach that mixes up query-relevant content with negative passages to directedly strength their relevance to the query. Based on this motivation, we consider mixing up hard negatives with the positive passage[2]. Formally, given a training example $(q_i, p_i^+, \mathbb{P}_i^-)$, we concatenate p_i^+ with each of the negative passage $p_{i,j}^-$ to form the set of dark examples for q_i:

$$\mathcal{N}_i^{rein} = \{p_i^+ [\texttt{SEP}] p_{i,j}^-\}_{j=1}^m. \quad (5)$$

Here, we choose to mix-up passages at the lexical level instead of the embedding space (Guo et al., 2019) because our method can produce valid language inputs and can preserve the relevant cues while introducing some less-relevant content. We also tried mixing-up negatives with the positive in the embedding space but found this kind of mix-up resulted in low-quality predictions of the cross-encoder teacher since it has never seen samples based on mixed embeddings during training.

Dark Examples with Noisy Positive Different from the above approach that creates dark examples by making hard negatives harder, we also consider the opposite direction: making the positive passage p^+ not that relevant to the query by introducing noise. We achieve this goal by input-masking (Devlin et al., 2019b). Given the positive passage p_i^+ for the query q_i, we randomly sample a subset of tokens from p_i^+ and replace them with the special token [MASK] with the masking ratio m_r:

$$\mathcal{N}_i^{mask} = \{\texttt{MASK}_{m_r}(p_i^+)\}_{m_r}. \quad (6)$$

To generate noisy positives with more diverse relevant to the query, we use masking with a variety of masking ratios.

[2]We also tried to make the hard negatives even harder by mixing up hard negatives with the query following Kalantidis et al. (2020), however, we found little change in performance.

3.3 Distillation with Adaptive Dark Examples

We have elaborated our motivation and approach to create dark examples, the remaining question is how to conduct effective knowledge distillation with dark examples. Existing knowledge distillation methods using all the labeled data without distinction, which we argue is sub-optimal. As knowledge distillation relies on the teacher's prediction as the supervision signal, the "quality" of knowledge held in the teacher's soft label naturally varies among different training examples. We assume that those training examples that the teacher is more confident than others are better carriers of knowledge for three reasons: (1) These instances are far from the decision boundaries of the model, and thus the corresponding passages are more likely to be true positives and true negatives, avoiding data noise. (2) Only the knowledge held in the instances that the teacher can cope with well are reliable and worth to be learned by the student. (3) The teacher's soft label for the high-confidence instances is too sharp, which indicates the dark knowledge held in these reliable instances has not been well exploited.

Therefore, we propose to adaptively concentrate on these high-confidence training instances during the training process to conduct our dark-example-based knowledge distillation. Formally, for a training instance, we can calculate the log-probability of the positive passage p_i^+ against negatives $\mathbb{P}_i^-$ with the teacher as the confidence score:

$$\mathcal{C}(q_i) = \log \frac{\exp^{\mathcal{R}_{ce}(q_i, p_i^+)}}{\exp^{\mathcal{R}_{ce}(q_i, p_i^+)} + \sum_{p_{i,j}^- \in \mathbb{P}_i^-} \exp^{\mathcal{R}_{ce}(q_i, p_{i,j}^-)}}. \quad (7)$$

Suppose the training process consists of T epochs, in each epoch t, we can sort a batch of training instances $\mathcal{B}_t$ in ascending order based on the confidence scores. Then we adaptively select the subset of instances $\hat{\mathcal{B}}_t$ in the batch that have the highest confidence scores with the ratio $(1 - \frac{t}{2*T})$ to construct dark examples:

$$\tilde{\mathcal{B}}_t = \mathop{\arg\max}_{q_i \in \mathcal{B}_t, \tilde{\mathcal{B}}_t \subset \mathcal{B}_t, \|\tilde{\mathcal{B}}_t\| = (1 - \frac{t}{2*T}) \times b} \mathcal{C}(q_i). \quad (8)$$

where b is the batch size for training.

Thereby, we have two sets in each step of the t-th training epoch: the original training batch $\mathcal{B}_t$ and the subset with the highest confidence that has both original candidate passages and our created

Methods	PLM	KD	MS-MARCO Dev MRR@10	MS-MARCO Dev R@50	MS-MARCO Dev R@1000	TREC DL 19 NDCG@10	TREC DL 19 R@100
Sparse retrieval							
BM25 (anserini) (Yang et al., 2017a)	-	-	18.7	59.2	85.7	50.6	-
doc2query (Nogueira et al., 2019b)	-	-	21.5	64.4	89.1	-	-
DeepCT (Dai and Callan, 2019b)	$\text{BERT}_{\text{base}}$	-	24.3	69.0	91.0	55.1	-
docTTTTTquery (Nogueira et al., 2019a)	-	-	27.7	75.6	94.7	-	-
UHD-BERT (Jang et al., 2021)	$\text{BERT}_{\text{base}}$	-	29.6	77.7	96.1	-	-
COIL-full (Gao et al., 2021)	$\text{BERT}_{\text{base}}$	-	35.5	-	96.3	70.4	-
UniCOIL (Lin and Ma, 2021)	$\text{BERT}_{\text{base}}$	-	35.2	80.7	95.8	-	-
SPLADE-max (Formal et al., 2021)	$\text{BERT}_{\text{base}}$	-	34.0	-	96.5	68.4	-
$\text{Unifier}_{\text{lexicon}}$ (Shen et al., 2023)	$\text{coCon}_{\text{base}}$	✓	39.7	-	98.1	73.3	-
Dense retrieval							
DPR-E (Ren et al., 2021c)	$\text{ERNIE}_{\text{base}}$	-	32.5	82.2	97.3	-	-
ANCE (single) (Xiong et al., 2020)	$\text{RoBERTa}_{\text{base}}$	-	33.0	-	95.9	65.4	44.5
TAS-Balanced (Hofstätter et al., 2021a)	$\text{BERT}_{\text{base}}$	✓	34.0	-	-	71.2	-
ME-BERT (Luan et al., 2021b)	$\text{BERT}_{\text{large}}$	-	34.3	-	-	-	-
ColBERT (Khattab and Zaharia, 2020a)	$\text{BERT}_{\text{base}}$	-	36.0	82.9	96.8	67.0	-
ColBERT v2 (Santhanam et al., 2021)	$\text{BERT}_{\text{base}}$	✓	39.7	86.8	98.4	72.0	-
ADORE+STAR (Zhan et al., 2021b)	$\text{RoBERTa}_{\text{base}}$	-	34.7	-	-	68.3	-
Condenser (Gao and Callan, 2021a)	$\text{BERT}_{\text{base}}$	-	36.6	-	97.4	-	-
RocketQA (Qu et al., 2021)	$\text{ERNIE}_{\text{base}}$	-	37.0	85.5	97.9	-	-
PAIR (Ren et al., 2021a)	$\text{ERNIE}_{\text{base}}$	-	37.9	86.4	98.2	-	-
CoCondenser (Gao and Callan, 2022)	$\text{BERT}_{\text{base}}$	-	38.2	-	98.4	-	-
RocketQAV2 (Ren et al., 2021c)	$\text{BERT}_{\text{base}}$	✓	38.8	86.2	98.1	-	-
AR2 (Zhang et al., 2022)	$\text{BERT}_{\text{base}}$	✓	39.5	-	98.6	-	-
CL-DRD (Zeng et al., 2022)	DistilBERT	✓	38.2	-	-	72.5	45.3
ERNIE-Search (Lu et al., 2022)	$\text{BERT}_{\text{base}}$	✓	40.1	87.7	98.2	-	-
RetroMAE (Xiao et al., 2022)	$\text{BERT}_{\text{base}}$	✓	39.3	87.0	98.5	-	-
$\text{Unifier}_{\text{dense}}$ (Shen et al., 2023)	$\text{coCon}_{\text{base}}$	✓	38.8	-	97.6	71.1	-
bi-SimLM (Wang et al., 2023)	$\text{BERT}_{\text{base}}$	✓	39.1	87.3	98.6	69.8	-
PROD (Lin et al., 2023)	ERNIE-2.0-BASE	✓	39.3	87.1	98.4	73.3	48.4
InDi (Cohen et al., 2024)	$\text{coCon}_{\text{base}}$	-	38.8	86.6	98.5	-	-
Rand KD (*Teacher = RocketQAV2*)	$\text{BERT}_{\text{base}}$	✓	38.1	86.9	98.2	-	-
Hard KD (*Teacher = RocketQAV2*)	$\text{BERT}_{\text{base}}$	✓	39.1	87.6	98.5	-	-
ADAM (*Teacher = RocketQAV2*)	$\text{BERT}_{\text{base}}$	✓	39.8	88.1	98.6	72.1	50.3
Rand KD (*Teacher = R^2anker*)	$\text{BERT}_{\text{base}}$	✓	38.1	86.0	97.9	-	-
Hard KD (*Teacher = R^2anker*)	$\text{BERT}_{\text{base}}$	✓	40.0	87.6	98.1	-	-
ADAM (*Teacher = R^2anker*)	$\text{BERT}_{\text{base}}$	✓	<u>41.0</u>	<u>88.5</u>	98.5	<u>73.4</u>	<u>49.8</u>

Table 1: Passage retrieval results on MS-MARCO and TREC DL 19 datasets. PLM is the abbreviation of the pre-trained language Model. KD indicates whether a model is distilled by a ranker. We copy the results from original papers and leave them blank if the original paper does not report the result. The best results are in underlined fonts.

dark examples $\tilde{\mathcal{B}}_t$. We jointly optimize the student with the supervised loss (Eq. 3.1) on $\mathcal{B}_t$ and the knowledge distillation loss (Eq. 4) on $\tilde{\mathcal{B}}_t$:

$$\mathcal{L}_t = \lambda \cdot \sum_{\mathcal{B}_t \in \mathcal{D}} \sum_{(q_i, \mathbb{P}_i) \in \mathcal{B}_t} \mathcal{L}_{sup} + \sum_{\hat{\mathcal{B}}_t \in \mathcal{D}} \sum_{(q_i, \tilde{\mathbb{P}}_i) \in \tilde{\mathcal{B}}_t} \mathcal{L}_{kd}. \tag{9}$$

where $\tilde{\mathbb{P}}_i = \{\mathbb{P}_i^- \cup \mathcal{N}_i^{mix} \cup \mathcal{N}_i^{mask}\}$ is the new candidate set for q_i, and λ is a hyper-parameter as a trade-off between the supervised objective and distillation objective with adaptive dark examples.

4 Experiments

We evaluate our method on two public human-annotated real-world benchmarks, namely MS-Marco and TREC Deep Learning 2019.

4.1 Datasets and Evaluation Metrics

Consisting with previous studies on dense information retrieval (Hofstätter et al., 2021b; Xiong et al., 2021), we use popular passage retrieval datasets, MS-MARCO (Nguyen et al., 2016). The dataset contains 8.8M passages from Web pages gathered from Bing's results to real-world queries. The training set contains about 500k pairs of query and relevant passage, and the dev set consists of 6,980 queries. Based on the queries and passages in the dataset, MS-MARCO passage retrieval and ranking tasks were created. Following previous works (Zeng et al., 2022), we report the performance on MS-MARCO Dev set as well as TREC Deep Learning (DL) 2019 set (Craswell et al., 2020) which includes 43 queries. We report MRR@10 and Recall@50/1K for MS-MARCO, and nDCG@10 and Recall@100 for TREC DL 19. We also report zero-shot transfer performance (nDCG@10) on BEIR benchmark (Thakur et al., 2021).

4.2 Baselines

To make a comprehensive comparison, we choose the following state-of-the-art approaches as baselines. These methods contain both sparse and dense passage retrievers.

The sparse retrieval methods include the traditional retriever BM25 (Yang et al., 2017b) and several representative sparse retrievers, including doc2query (Lu et al., 2020), DeepCT (Dai and Callan, 2019a), docTTTTT-query (Nogueira et al., 2019a), UHD-BERT (Jang et al., 2021), COIL-full (Gao et al., 2021), UniCOIL (Lin and Ma, 2021), and SPLADE-max (Formal et al., 2021).

The dense retrieval methods produce continuous neural vectors for each passage and query. The methods include DPR-E (Qu et al., 2021), ANCE (Xiong et al., 2021), TAS-Balanced (Hofstätter et al., 2021b), ME-BERT (Luan et al., 2021a), ColBERT (Khattab and Zaharia, 2020b), ColBERT v2 (Santhanam et al., 2021), NPRINC (Lu et al., 2021), ADORE+STAR (Zhan et al., 2021a), Condenser (Gao and Callan, 2021a), RocketQA (Qu et al., 2021), PAIR (Ren et al., 2021b), CoCondenser (Gao and Callan, 2022), RoketQAV2 (Ren et al., 2021c), AR2 (Zhang et al., 2022), CL-DRD (Zeng et al., 2022), ERNIE-Search (Lu et al., 2022), RetroMAE (Xiao et al., 2022), Unifier (Shen et al., 2023), bi-SimLM (Wang et al., 2023), PROD (Lin et al., 2023) and InDi (Cohen et al., 2024). Some of them are enhanced by knowledge distillation from the ranker. For example, RoketQAV2, AR2, and ERNIE-Search introduce the on-the-fly distillation method. CL-DRD and PROD propose progressive distillation with a data curriculum to gradually improve the student.

4.3 Implementation Details

Consisting with the setting of RocketQA V2 (Ren et al., 2021c), we choose the learned dual-encoder in the first step of RocketQA (Qu et al., 2021) as the initialization of our dense retriever[3]. We adopt two advanced cross-encoder rankers as our teacher model: RocketQAV2 (Ren et al., 2021c) and R^2anker (Zhou et al., 2023)[4]. We randomly select m hard negatives provided by Ren et al. (2021c) for each query. For supervised learning, a positive passage and all the selected negatives are used. While for distillation, the candidate passage set for a query consists of m original negatives, m dark examples in $\mathcal{N}_i^{mix}$, and 5 dark examples in $\mathcal{N}_i^{mask}$ with different masking ratios $m_r \in \{0.15, 0.25, 0.35, 0.45, 0.55\}$. We set the number of negatives m to 10 from $\{5, 10, 15, 20, 25, 30\}$[5]. We set the maximum lengths for queries and passages as 32 and 128. The dropout rate is set to 0.1 on the cross-encoder. In training, we use AdamW (Loshchilov and Hutter, 2017) as the optimizer to train the model. We set the batch size as 128, the peak learning rate as $5e-5$, and the warm-up steps as 100. We set the weight λ for the supervised objective as 0.01 by varying it in $\{0.001, 0.01, 0.05, 0.1, 0.5\}$.

4.4 Overall Performance

We report the overall evaluation results on MS-MARCO and TREC Deep Learning 2019 respectively. On both benchmarks, we not only show the performance of our dual-encoder retriever under knowledge distillation from two different cross-encoder teachers, but also provide comparisons between different choices of construction of candidate set $\mathbb{P}_i$. The main results are shown in Table 1. We can draw three main conclusions:

Our created dark examples improve the performance of knowledge distillation over hard negatives and random negatives. With the same cross-encoder as the teacher, we analyze the impact of how the candidate set of passages is constructed. It can be observed that using random negatives results in poor performance and the integration of hard negative mining indeed improve the performance. When equipped with our created dark examples which are even harder than existing hard negatives, our model further makes a substantial improvement over that using hard negatives.

Our framework ADAM *is compatible with different teachers.* To test the generalization ability

[3]The retriever can also be replaced with other trained retriever. We observed that using the trained model to initialize the retriever can help achieve slightly better results.

[4]The results of BM25-reranking on MS-MARCO Dev for R^2anker (Zhou et al., 2023) and RocketQAV2 (Ren et al., 2021c) are 40.1 and 40.7 respectively.

[5]We found $m = 15$ to be the optimal parameter. However, considering that our method will expand the number of negatives with the augmented dark examples, we set m=10 in our experiment.

Rep type	Sparse			Mul-vec	Dense						
Method	BM25	SPLADE	UnifieR	ColBERT	DPR	ANCE	TAS-B	CoCond	CL-DRD	RocQA	ADAM
Distillation	✗	✓	✓	✓	✗	✗	✓	✗	✓	✓	✓
TREC-COVID	65.6	71.0	71.5	73.8	33.2	65.4	48.1	71.2	58.4	67.5	73.0
NFCorpus	32.5	33.4	32.9	33.8	18.9	23.7	31.9	32.5	31.5	29.3	31.5
FiQA	23.6	33.6	31.1	35.6	11.2	29.5	30.0	27.6	30.8	30.2	31.5
ArguAna	31.5	47.9	39.0	46.3	17.5	41.5	42.9	29.9	41.3	45.1	40.3
Tóuche-2020	36.7	27.2	30.2	26.3	13.1	24.0	16.2	19.1	20.3	24.7	25.6
Scidocs	15.8	15.8	15.0	15.4	7.7	12.2	14.9	13.7	14.6	13.1	14.1
SciFact	66.5	69.3	68.6	69.3	31.8	50.7	64.3	61.5	62.1	56.8	59.4
NQ	32.9	52.1	51.4	56.2	47.4	44.6	46.3	48.7	50.0	50.5	51.9
HotpotQA	60.3	68.4	66.1	66.7	39.1	45.6	58.4	56.3	58.9	53.3	58.6
DBPedia	31.3	43.5	40.6	44.6	26.3	28.1	38.4	36.3	38.1	35.6	39.6
Fever	75.3	78.6	69.6	78.5	56.2	66.9	70.0	49.5	73.4	67.6	66.8
Climate-FEVER	21.3	23.5	17.5	17.6	14.8	19.8	22.8	14.4	20.4	18.0	21.4
AVERAGE	41.1	42.4	44.5	47.0	26.4	37.7	40.4	38.9	42.0	41.0	42.8

Table 2: Zero-shot transfer performance (nDCG@10) on BEIR benchmark. 'BEST ON' and 'AVERAGE' do not take the in-domain result into account. 'ColBERT' is its v2 version (Santhanam et al., 2021). 'CoCond' refers to CoCondenser (Gao and Callan, 2021b) and 'RocQA' means RocketQAV2 (Ren et al., 2021c).

over different teachers, we conduct experiments using two advanced cross-encoders (R^2anker and RocketQAV2) as the teacher. Consistent improvement can be observed when using our proposed dark examples for knowledge distillation with the two different teachers. Moreover, we can compare the effectiveness of the two teachers. When using random negatives, knowledge distillation with the two teachers results in comparable results. But when using hard negatives and dark examples, the model distilled by R^2anker yields significantly better performance than its counterparts. Therefore, for the remaining ablation studies and analyses, we use R^2anker as the teacher by default.

With R^2anker as the teacher, our method (the bottom line) achieves superior performance over most baselines. Our model achieves 41.00 on MRR@10 on the development set of MS-MARCO, outperforming most of the existing methods and is comparable with SimLM (Wang et al., 2023) which is obtained by a time-consuming large-scale pretraining followed with a cumbersome multi-stage supervised fine-tuning.

4.5 Ablation study

We have analyzed the overall performance on two benchmarks and proved the effectiveness of our method. Here, we conduct ablation studies to verify the indispensability of each crucial design. We provide the results of the ablation study in Table 3.

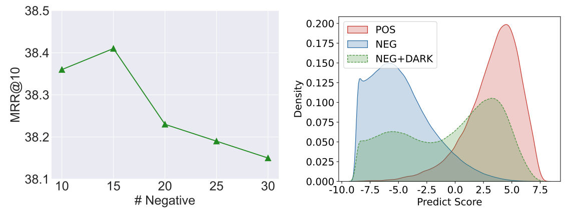
Dark examples. Recall that we propose two types of methods to construct dark examples: (1) strengthening negatives ($\mathcal{N}^{rein}$) by mixing with the positive to make negatives more relevant to the query, and (2) polluting positives (($\mathcal{N}^{mask}$)) to make positives not that relevant. We first test the individual effect of $\mathcal{N}^{rein}$ and $\mathcal{N}^{mask}$. When removing each of them individually, performance drops can be observed. And when we remove both of them, the model performs worse. This observation indicates that the incorporation of both $\mathcal{N}^{rein}$ and $\mathcal{N}^{mask}$ is beneficial to the overall performance.

Methods	MRR@10
ADAM	<u>38.99</u>
w/o. $\mathcal{N}^{rein}$ (Eq.5)	38.82
w/o. $\mathcal{N}^{mask}$ (Eq.6)	38.76
w/o. $\{\mathcal{N}^{rein}$ & $\mathcal{N}^{mask}\}$	38.64
w/o. $\{\mathcal{N}^{rein}$ & $\mathcal{N}^{mask}$ & ADA $\}$	38.61
w/o. $\{\mathcal{N}^{rein}$ & $\mathcal{N}^{mask}$ & ADA & $\mathcal{L}_{\text{sup}}\}$	38.36

Table 3: Ablation results on MS-Marco. We report the reranking performance.

Distillation with adaptive dark examples. In addition to dark examples, we also introduce a self-paced distillation algorithm that can better transfer dark knowledge with adaptive dark examples. When this strategy is removed, we create dark examples for all the training instances. It can be seen that distillation adaptively using the subset of instances that the teacher is most confident is better than using the whole training set, which is in accord with our assumption that the instances with higher confidence are a better carrier of knowledge.

Distillation with additional supervised loss. Although the teacher's soft label provides abundant



(a) The impact of m (b) Dist. of Ranker output

Figure 3: (a) The impact of m; (b) Distributions of model prediction for the R^2anker over MS-MARCO.

dark knowledge for the student to learn, we also involve the traditional supervised loss. We can observe that although the weight λ for supervised loss is quite small (i.e., 0.01), we find this term indispensable for the overall performance.

4.6 Discussions

Zero-shot transfer performance. We also curious about the transfer ability of our method and conduct experiments on the BEIR benchmark. Table 2 reports the zero-shot performance. We can find that our method surpasses existing dense approaches and three distillation methods (TAS-B, CL-DRD, and RocketQAV2). These results demonstrate that the reranker can transfer knowledge more effectively to the dense retriever using dark examples, and its out-of-distribution (OOD) adaptation ability is also well inherited by the retriever.

The impact of the number of negatives. When constructing the training set, the number of negatives plays a vital role as it also indirectly controls the number of dark examples. To explore the effect of the number of negative samples as well as to find the best choice for m, we conduct experiments on different m[6]. As illustrated in Figure 3(a), when m is small, increasing m brings a positive effect and leads to the best performance when $m = 15$. But as the curve indicates, incorporating more negatives brings no benefit, which is also in line with existing findings (Karpukhin et al., 2020). The above trend also indicates that too many trivial negatives ($m > 15$) can not always bring improvement while incorporating our dark examples can still bring improvement to the knowledge transfer. The phenomenon also reveals the importance of distillation data in IR knowledge transfer.

The impact of dark examples on the output distribution of ranker. Finally, we examine the impact of dark examples on the output distribution of the ranker. As illustrated in Figure 3(b), we draw the score distributions of the positive, negative candidates, and negative candidates plus dark examples using a teacher (R^2anker) over MS-MARCO. It can be observed that the scores for most original hard negatives are quite low and distributed far from the positives that have high scores. By incorporating these dark examples, we are able to improve the smoothness of the score distribution and prob our teacher model with a wider range of candidates that are more diversely relevant to the query. This enables us to more effectively transfer valuable "dark" knowledge from the teacher model.

[6]To better analyze the impact of the number of negative samples, we conduct the experiment on the model without adaptive dark examples.

5 Conclusion

In this paper, we propose a knowledge distillation framework that can better transfer the dark knowledge in the cross-encoder with adaptive dark examples to help the dual-encoder achieve better performance. We propose two approaches to create dark examples that are much harder for the cross-encoder teacher to distinguish than typical hard negatives to transfer more dark knowledge. Further, we propose a self-paced distillation strategy that transfers the knowledge adaptively with high-confidence training instances. Experimental results in two widely-used benchmarks verify the effectiveness of our proposed method.

Limitations

(i) *Training computation overheads*: although having the same inference complexity as any other dense retrieval models, our approach requires more computation resources during training as it expands the number of negatives with the augmented dark examples. (ii) *More analysis on noisy positives*: due to the limited computation resource, we only test and compare several typical settings of noisy positives, better strategies for constructing noisy positives (e.g., better masking methods and varying the number of noisy positives) can be explored to further improve the performance.

References

ZeFeng Cai, Chongyang Tao, Tao Shen, Can Xu, Xiubo Geng, Xin Alex Lin, Liang He, and Daxin Jiang. 2022. Hyper: Multitask hyper-prompted training enables large-scale retrieval generalization. In *The Eleventh International Conference on Learning Representations*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1870–1879.

Nachshon Cohen, Hedda Cohen Indelman, Yaron Fairstein, and Guy Kushilevitz. 2024. Indi: Informative and diverse sampling for dense retrieval.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the TREC 2019 deep learning track. *CoRR*, abs/2003.07820.

Zhuyun Dai and Jamie Callan. 2019a. Context-aware sentence/passage term importance estimation for first stage retrieval. *CoRR*, abs/1910.10687.

Zhuyun Dai and Jamie Callan. 2019b. Deeper text understanding for IR with contextual neural language modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 985–988.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE v2: Sparse lexical and expansion model for information retrieval. *CoRR*, abs/2109.10086.

Luyu Gao and Jamie Callan. 2021a. Condenser: a pre-training architecture for dense retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 981–993. Association for Computational Linguistics.

Luyu Gao and Jamie Callan. 2021b. Unsupervised corpus aware language model pre-training for dense passage retrieval. *CoRR*, abs/2108.05540.

Luyu Gao and Jamie Callan. 2022. Unsupervised corpus aware language model pre-training for dense passage retrieval. In *ACL*.

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. COIL: Revisit exact lexical match in information retrieval with contextualized inverted list. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3030–3042.

Daniel Gillick, Alessandro Presta, and Gaurav Singh Tomar. 2018. End-to-end retrieval in continuous space. *arXiv preprint arXiv:1811.08008*.

Hongyu Guo, Yongyi Mao, and Richong Zhang. 2019. Augmenting data with mixup for sentence classification: An empirical study. *arXiv preprint arXiv:1905.08941*.

Matthew L. Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *CoRR*, abs/1705.00652.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020. Improving efficient neural ranking models with cross-architecture knowledge distillation. *CoRR*, abs/2010.02666.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021a. Efficiently teaching an effective dense retriever with balanced topic aware sampling. *CoRR*, abs/2104.06967.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021b. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 113–122. ACM.

Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. 2020. Embedding-based retrieval in facebook search. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 2553–2561.

Kyoungrok Jang, Junmo Kang, Giwon Hong, Sung-Hyon Myaeng, Joohee Park, Taewon Yoon, and Hee-Cheol Seo. 2021. UHD-BERT: bucketed ultra-high dimensional sparse representations for full ranking. *CoRR*, abs/2104.07198.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. 2020. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33:21798–21809.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020a. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 39–48.

Omar Khattab and Matei Zaharia. 2020b. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 39–48. ACM.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6086–6096. Association for Computational Linguistics.

Jimmy Lin and Xueguang Ma. 2021. A few brief notes on deepimpact, coil, and a conceptual framework for information retrieval techniques. *CoRR*, abs/2106.14807.

Zhenghao Lin, Yeyun Gong, Xiao Liu, Hang Zhang, Chen Lin, Anlei Dong, Jian Jiao, Jingwen Lu, Daxin Jiang, Rangan Majumder, et al. 2023. Prod: Progressive distillation for dense retrieval. In *Proceedings of the ACM Web Conference 2023*, pages 3299–3308.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101.

Jing Lu, Gustavo Hernández Ábrego, Ji Ma, Jianmo Ni, and Yinfei Yang. 2020. Neural passage retrieval with improved negative contrast. *CoRR*, abs/2010.12523.

Shuqi Lu, Chenyan Xiong, Di He, Guolin Ke, Waleed Malik, Zhicheng Dou, Paul Bennett, Tie-Yan Liu, and Arnold Overwijk. 2021. Less is more: Pre-training a strong siamese encoder using a weak decoder. *CoRR*, abs/2102.09206.

Yuxiang Lu, Yiding Liu, Jiaxiang Liu, Yunsheng Shi, Zhengjie Huang, Shikun Feng Yu Sun, Hao Tian, Hua Wu, Shuaiqiang Wang, Dawei Yin, et al. 2022. Erniesearch: Bridging cross-encoder with dual-encoder via self on-the-fly distillation for dense passage retrieval. *arXiv preprint arXiv:2205.09153*.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021a. Sparse, dense, and attentional representations for text retrieval. *Trans. Assoc. Comput. Linguistics*, 9:329–345.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021b. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Aditya Menon, Sadeep Jayasumana, Ankit Singh Rawat, Seungyeon Kim, Sashank Reddi, and Sanjiv Kumar. 2022. In defense of dual-encoders for neural ranking. In *International Conference on Machine Learning*, pages 15376–15400. PMLR.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019a. From doc2query to doctttttquery. *Online preprint*.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019b. Document expansion by query prediction. *CoRR*, abs/1904.08375.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5835–5847. Association for Computational Linguistics.

Ruiyang Ren, Shangwen Lv, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021a. PAIR: Leveraging passage-centric similarity relation for improving dense passage retrieval. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2173–2183.

Ruiyang Ren, Shangwen Lv, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021b. PAIR: leveraging passage-centric similarity relation for improving dense passage retrieval. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2173–2183. Association for Computational Linguistics.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021c. Rocketqav2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2825–2835. Association for Computational Linguistics.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2021. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *CoRR*, abs/2112.01488.

Tao Shen, Xiubo Geng, Chongyang Tao, Can Xu, Guodong Long, Kai Zhang, and Daxin Jiang. 2023. Unifier: A unified retriever for large-scale retrieval. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 4787–4799, New York, NY, USA. Association for Computing Machinery.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *CoRR*, abs/2104.08663.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2023. SimLM: Pre-training with representation bottleneck for dense passage retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2244–2258, Toronto, Canada. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. RetroMAE: Pre-training retrieval-oriented language models via masked auto-encoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 538–548, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *CoRR*, abs/2007.00808.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Kai Xu, Dae Hoon Park, Chang Yi, and Charles Sutton. 2018. Interpreting deep classifier by visual distillation of dark knowledge. *arXiv preprint arXiv:1803.04042*.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017a. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 1253–1256.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017b. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 1253–1256. ACM.

Yinfei Yang, Ning Jin, Kuo Lin, Mandy Guo, and Daniel Cer. 2020. Neural retrieval for question answering with cross-attention supervised data augmentation. *CoRR*, abs/2009.13815.

Hansi Zeng, Hamed Zamani, and Vishwa Vinay. 2022. Curriculum learning for dense retrieval distillation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1979–1983.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021a. Optimizing dense retrieval model training with hard negatives. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 1503–1512. ACM.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021b. Optimizing dense retrieval model training with hard negatives. *CoRR*, abs/2104.08051.

Hang Zhang, Yeyun Gong, Yelong Shen, Jiancheng Lv, Nan Duan, and Weizhu Chen. 2022. Adversarial retriever-ranker for dense text retrieval. In *International Conference on Learning Representations*.

Kai Zhang, Chongyang Tao, Tao Shen, Can Xu, Xiubo Geng, Binxing Jiao, and Daxin Jiang. 2023. Led: Lexicon-enlightened dense retriever for large-scale retrieval. In *Proceedings of the ACM Web Conference 2023*, pages 3203–3213.

Yucheng Zhou, Tao Shen, Xiubo Geng, Chongyang Tao, Can Xu, Guodong Long, Binxing Jiao, and Daxin Jiang. 2023. Towards robust ranker for text retrieval. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5387–5401, Toronto, Canada. Association for Computational Linguistics.

Lixin Zou, Shengqiang Zhang, Hengyi Cai, Dehong Ma, Suqi Cheng, Shuaiqiang Wang, Daiting Shi, Zhicong Cheng, and Dawei Yin. 2021. Pre-trained language model based ranking in baidu search. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pages 4014–4022. ACM.