Fine-Grained Distillation for Long Document Retrieval

Yucheng Zhou[1*], Tao Shen[2], Xiubo Geng[3], Chongyang Tao[3], Jianbing Shen[1], Guodong Long[2], Can Xu[3], Daxin Jiang[3†]

[1]SKL-IOTSC, CIS, University of Macau
[2]AAII, FEIT, University of Technology Sydney
[3]Microsoft Corporation
yucheng.zhou@connect.um.edu.mo, {tao.shen, guodong.long}@uts.edu.au,
{xigeng,chongyang.tao,can.xu,djiang}@microsoft.com, jianbingshen@um.edu.mo

Abstract

Long document retrieval aims to fetch query-relevant documents from a large-scale collection, where knowledge distillation has become de facto to improve a retriever by mimicking a heterogeneous yet powerful cross-encoder. However, in contrast to passages or sentences, retrieval on long documents suffers from the *scope hypothesis* that a long document may cover multiple topics. This maximizes their structure heterogeneity and poses a granular-mismatch issue, leading to an inferior distillation efficacy. In this work, we propose a new learning framework, fine-grained distillation (FGD), for long-document retrievers. While preserving the conventional dense retrieval paradigm, it first produces global-consistent representations crossing different fine granularity and then applies multi-granular aligned distillation merely during training. In experiments, we evaluate our framework on two long-document retrieval benchmarks, which show state-of-the-art performance.

Introduction

Large-scale retrieval, as a fundamental task in information retrieval (IR), has attracted increased interest from industry and academia in the last decades, as it plays an indispensable role in a wide range of real-world applications, such as web engines (Fan et al. 2022), question answering (Karpukhin et al. 2020) and dialogue systems (Yu et al. 2021). Given a text query, it aims to fetch top-relevant documents[1] from a huge collection (Cai et al. 2021). As the collection usually scales up to millions or billions, a retrieval method must satisfy the efficiency or latency requirement of online deployment to calculate the relevance score between a query and every document.

[*] Work is done during internship at Microsoft. This work was supported in part by FDCT grants 0154/2022/A3, 0102/2023/RIA2 and SKLIOTSC(UM)-2021-2023, MYRG-CRG2022-00013-IOTSC-ICI grant and SRG2022-00023-IOTSC grant.

[†] Corresponding author.

[1]Each entry of the collection can be any text granularity (e.g., sentence, passage, document) but we take 'document' to denote 'entry of collection' in this paper for clear writing.

Document: … The European Renaissance was a time of massive economic and cultural growth following the stagnation of the Middle Ages. Beginning in Italy in the 14th century, the movement spread to all parts of the continent during the next 300 years. The outstanding cultural and artistic heritage of the Renaissance can still be seen today in many of the great cities of the period, including Florence and Venice in Italy, Bruges in Belgium and Toledo in Spain. It All Began in Florence Florence is the city where the Renaissance began, and where it reached its peak in the 15th and 16th centuries under the patronage of the powerful Medici family. Some of the greatest names in Renaissance art are associated with the city, including Leonardo da Vinci, Botticelli and Michelangelo. The poet Dante, the political theorist Machiavelli and the scientist Galileo also lived and worked in Florence. Buildings like the Pitti Palace, Uffizi Gallery and Florence Cathedral are among the masterpieces of Renaissance architecture. The Legacy of Venice To present-day tourists, Venice is renowned for its picturesque canals and its lack of motorized vehicles …

Query1: 3 people who were important of the european renaissance time period

Query2: what are the italian renaissance cities

Figure 1: A case for *scope hypothesis* in long document. The document contains multiple topics, and the relevance to a query may vary across different parts of the document.

Recently, pre-trained language models (PLMs), e.g., BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), DeBERTa (He et al. 2021), have dominated the field of IR in deep representation learning literature, as they are readily adapted to capture token-wise correlations and produce generic representations by fine-tuning. In the common practice of PLMs, a pair of text pieces (i.e., a query and every document in our task) should be concatenated to pass into the models (Devlin et al. 2019) for fine-grained relevance measurement – known as *cross-encoder* that performs very competitively – however cannot meet the efficiency requirement due to combinatorial explosion in terms of online PLM inference (Zhang et al. 2022a; Ren et al. 2021). In contrast, a *bi-encoder* (a.k.a. dual-encoder or two-tower) leverages the PLMs to embed queries and documents individually into a single vector in the same dense semantic space, and then query-document relevance can be derived by a lightweight metric (e.g., dot-product) (Reimers and Gurevych 2019). The bi-encoder enables offline document embeddings and satisfies the online efficiency requirement, so it has become the de facto model choice for PLM-based large-scale retrievers. However, the bi-encoder is vulnerable to information bottleneck by the single dense vector and thus lags behind the cross-encoder considerably (Wang et al. 2022; Gao and Callan 2022).

To narrow the performance gap against cross-encoder, a

recently advanced technique to train bi-encoder is distilling list-wise relevance score distributions from cross-encoder during contrastive learning (Zhang et al. 2022a; Ren et al. 2021). This technique merely affects the training process of a bi-encoder and has been proven to improve the generalization ability of bi-encoder (Menon et al. 2022), leading to better retrieval quality without any sacrifice of inference efficiency.

Nonetheless, such a distillation technique to improve bi-encoder has proven effective merely in the scenarios where the targeted text pieces are usually short semantic units (e.g., sentences (Liu et al. 2022b) and passages (Ren et al. 2021)) with an almost single topic. In contrast, long document retrieval usually targets super-long documents with up to thousands of words (cf. 65 words per passage (Nguyen et al. 2016)). Considering the *scope hypothesis* (Robertson and Zaragoza 2009) that a long document may cover multiple topics (see a case in Figure 1), distilling knowledge from a cross-encoder to bi-encoder is prone to become less effective. This is likely because modeling long documents maximizes their heterogeneity in terms of visibility – cross-encoder explicitly models the query-dependent salience part (e.g., a sentence) whereas bi-encoder directly models the whole into a query-agnostic dense bottleneck – thus such a brute-force distillation suffering from the granularity mismatching. In our pilot experiments, the brute-force distillation can only bring 0.1% gain on long document retrieval after extensive tuning, in contrast to $> 1\%$ gain frequently observed in passage retrieval (Ren et al. 2021).

Thereby, we aim to improve the knowledge distillation from a cross-encoder to a long-document retriever by circumventing the granularity mismatching problem. Instead of knowledge distillation at the long-document level, we propose a brand-new bi-encoder learning framework, dubbed fine-grained distillation (FGD), for large-scale retrieval over long documents. It operates on multi-vector distillation crossing fine granularity merely in the training phase while keeping single-vector retrieval during inference. To derive fine-grained representations without cross-granular conflict, we first propose a global-consistent granularity embedding method, which enables dynamic contextualization visibility (e.g., passage, sentence) over a long document. Then, we present a local-coordinating score distilling strategy, which replaces global (i.e., document-level) distillation, for long-document retriever training. In addition, to empower our distillation strategy, we propose a hierarchical negative mining technique to produce hard negatives throughout granularity.

In the experiments, we conduct an extensive evaluation of our proposed framework on two document retrieval benchmark datasets, i.e., MS-Marco document retrieval (Nguyen et al. 2016) and TREC 2019 Deep Learning track (Craswell et al. 2020). The experimental results show that our method achieves state-of-the-art performance compared with other strong competitors. In addition, we verify the generality of our framework by evaluating it on different long document retrievers paired with different cross-encoder teachers.

There are main contributions of our paper:

- We propose a novel fine-grained distillation (FGD) framework for large-scale retrieval over long documents. FGD is trained with multi-vector distillation crossing fine granularity while keeping single-vector retrieval in inference.
- We introduce a global-consistent granularity embedding method, which helps to derive fine-grained representations without cross-granular conflict.
- We present a local-coordinating score distilling strategy and a hierarchical negative mining technique to produce hard negatives throughout granularity, which further empowers our distillation strategy.
- Our proposed method achieves state-of-the-art performance on two long-document retrieval benchmarks.

Related Work

Retriever Training with Distillation. To improve dense passage retrieval, a trend is to conduct distillation from a cross-encoder-based ranker to a dense retriever, where the ranker can be well-trained in advance (Lin, Yang, and Lin 2021; Zhou et al. 2023) or updated along with the bi-encoder (Zhang et al. 2022a). In contrast to the conventional setting, distillation in retrieval does not focus on model compression but aims to distill features from different retriever architectures to learn knowledge from different semantic perspectives (Menon et al. 2022). In distillation in retrieval, a well-trained ranker is widely used as the teacher model to produce weak labels on large-scale unlabeled query-document pairs (Ren et al. 2021; Lu et al. 2022). To explain this, Menon et al. (2022) conduct a theoretical study to prove the distillation alleviates over-fitting of the bi-encoder training. However, these methods only investigate improving passage retrievers by distillation, regardless of the inherent scope hypothesis and the granular-mismatch issue in long-document retrieval.

Multi-granular Representation. In information retrieval, multi-granular representation learning attracts increased interest from the community as it can either break the information bottleneck to represent text from multiple views (Zhang et al. 2022b) or represent fine-grained semantic units for specific tasks (Lee et al. 2021; Zheng et al. 2020). In contrast, we aim to leverage multi-granular representation as the medium to distill fine-grained relevance information during learning, thus without extra overheads to calculate or/and store multiple vectors during inference.

Multi-granular Distillation. Recently, knowledge distillation, as a critical technique for model compression in NLP, has been widely applied to PLMs for compact ones (Chen et al. 2017). These works focus on contextualized embedding or attention maps for individual tokens, regardless of various semantic units with vital contextual information. Motivated by this, Liu et al. (2022a) propose multi-granularity knowledge distillation to exploit information of multi-granularity language units for model compression. However, this work is only applicable to homogeneous distillation, i.e., the same model family, which is however the opposite of our target. Therefore, we present a new strategy to distill fine-grained information between the heterogeneous structures.

Hard Negative Mining. Hard negative mining (Khattab and Zaharia 2020; Zhang et al. 2022a; Qu et al. 2021) has been proven very effective in contrastive learning for text

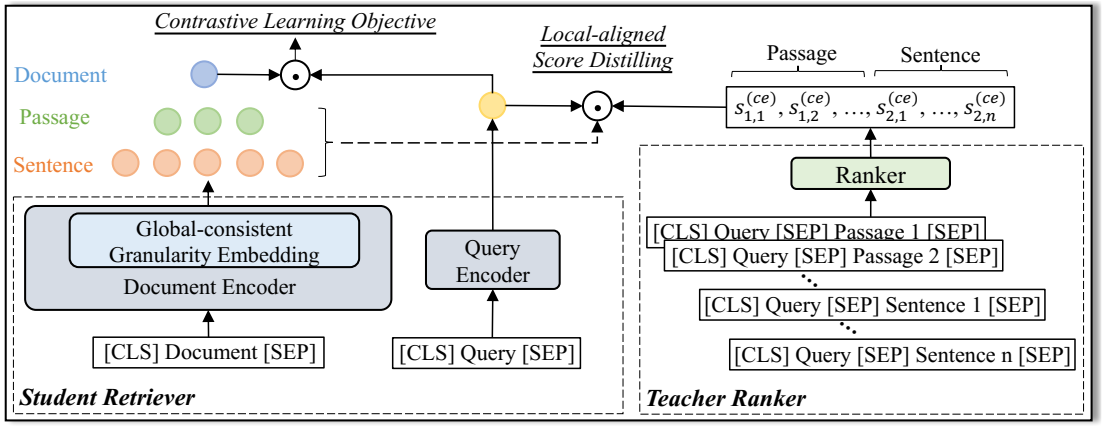

Figure 2: An overview of our fine-grained distillation (FGD) for long-document retrieval. FGD uses a global-consistent granularity embedding method to produce representations for different granularity in a document. Then, FGD applies a local-aligned score distilling strategy to learn from a cross-encoder teacher at each granularity level. Finally, a hierarchical hard negative mining provides negative samples across granularity.

representation of retrievers. In contrast to random or in-batch negatives, it finds more challenging negatives for a pair of an anchor (i.e., query) and its positive example. They compose effective contrastive samples to help model learn against contextual nuance between the positive and negatives. At the early stage, a large number of works employ the off-the-shelf BM25 retriever to fetch negative from a large collection (Karpukhin et al. 2020), which greatly boosts the retrievers. Furthermore, recent works (Gao and Callan 2021, 2022) leverage a retriever to sample retriever-specific hard negatives for each query, which are considered the most challenging negatives. But, the involved mining techniques are focused only on sequence-level hard negatives, partially incompatible with our goal. Thereby, we adapt previous methods to efficiently mine fine-grained negatives with minor modifications.

Methodology

Task Definition. Considering a large-scale collection with numerous long documents (i.e., $\mathbb{D} = \{d_i\}_{i=1}^{|\mathbb{D}|}$ where each d_i denotes a document), large-scale retrieval is to fetch top-relevance documents (i.e., $\mathbb{D}^q$) by a retriever (e.g., $\mathcal{M}$) for a text query q. This requires $\mathcal{M}$ to calculate every relevance score s_i^q between the q and $\forall d_i \in \mathbb{D}$, where $i \in [1, |\mathbb{D}|]$. In the remaining, we will omit the superscript 'q' for clean demonstration if no confusion is caused.

Bi-encoder Learning with Distillation

To meet the efficiency requirement of large-scale retrieval, a de facto scheme (Gao and Callan 2021, 2022; Wang et al. 2022) is to leverage a bi-encoder for the relevance score. It encodes each query and document individually into dense semantic space and derives the score usually by a lightweight metric (e.g., dot-product, cosine similarity), i.e,

$$s^{(\text{be})} := \mathcal{M}^{(\text{be})}(q, d|\theta^{(\text{be})}) =< \boldsymbol{u}, \boldsymbol{v} >:= \quad (1)$$
$$< \text{Enc}(q|\theta^{(\text{q})}), \text{Enc}(d|\theta^{(\text{d})}) >, \exists d \in \mathcal{D},$$

where $< \cdot, \cdot >$ denotes a non-parametric dot-product, $\text{Enc}(\cdot|\theta^{(*)})$ denotes a $\theta^{(*)}$-parameterized encoder that embeds a piece of text into a dense vector, and $\theta^{(\text{be})} = \theta^{(\text{q})} \cup \theta^{(\text{d})}$ parameterize the bi-encoder where the query and document encoders can be tied in terms of parameters.

Then, the training of retrieval-related models (e.g., bi-encoder learning $\theta^{(\text{be})}$ here) is usually formulated as a contrastive learning problem. That is, only a positive document d_+ is given as a golden label for the query q, while a set of negative documents $d_- \in \mathbb{N}$ also should be mined in light of (q, d_+) for contrastive learning (Gao and Callan 2022). Basically, a BM25 or a trained retriever is usually employed to mine the negatives. Providing d_+ and $\mathbb{N}$, we derive a score distribution over them,

$$\boldsymbol{p}^{(\text{be})} := P(\text{d}|q, \{d_+\} \cup \mathbb{N}; \theta^{(\text{be})}) = \quad (2)$$
$$\frac{\exp(\mathcal{M}^{(\text{be})}(q, d|\theta^{(\text{be})})/\tau)}{\sum_{d' \in \{d_+\} \cup \mathbb{N}} \exp(\mathcal{M}^{(\text{be})}(q, d'|\theta^{(\text{be})})/\tau)},$$

where $\forall d \in \{d_+\} \cup \mathbb{N}$ and τ denotes the temperature set to 1. Next, the training loss of contrastive bi-encoder learning can be simply written as

$$L^{(\text{cl})} = -\sum_q \log P(\text{d} = d_+|q, \{d_+\} \cup \mathbb{N}; \theta^{(\text{be})})$$
$$= -\sum \log \boldsymbol{p}^{(\text{be})}_{[\text{d}=d_+]}. \quad (3)$$

To improve the bi-encoder's generalization ability and boost its retrieval qualities, a common practice is to distill score distributions from a cross-encoder to the bi-encoder retriever. In general, a cross-encoder is frequently defined as a classifier that a Transformer encoder followed by a one-way-out multi-layer perceptron (MLP), i.e.,

$$s^{(\text{ce})} := \mathcal{M}^{(\text{ce})}(q, d|\theta^{(\text{ce})}) = \quad (4)$$
$$\text{Transformer-cls}(\texttt{[CLS]}\, q\, \texttt{[SEP]}\, d\, \texttt{[SEP]}\,|\,\theta^{(\text{ce})}),$$

where $s^{(\text{ce})} \in \mathbb{R}$ and $\theta^{(\text{ce})}$ parameterizes this cross-encoder. Here, q and d concatenated with special tokens are passed into the self-attention encoder to enable token-level interaction, capture fine-grained nuance, and produce a precise relevance

score. Note that, $\theta^{(\text{ce})}$ can be either well-trained in advance (Gao and Callan 2022; Zhou et al. 2023) or updated along with the bi-encoder (Ren et al. 2021; Zhang et al. 2022a), while we opt for the former but without loss of generality. Next, we can also obtain $\boldsymbol{p}^{(\text{ce})} \coloneqq P(\text{d}|q, \{d_+\} \cup \mathbb{N}; \theta^{(\text{ce})})$ as in Eq.(2). Last, the loss function of this distillation is

$$L^{(\text{kd})} = \text{KL-Div}(\boldsymbol{p}^{(\text{be})} \| \boldsymbol{p}^{(\text{ce})}). \tag{5}$$

So, the final training loss for the bi-encoder learning with distillation is written as $\lambda L^{(\text{cl})} + L^{(\text{kd})}$.

Global-consistent Granularity Embedding

Although the bi-encoder learning with distillation has been proven very effective in passage retrieval (Wang et al. 2022) or sentence matching (Reimers and Gurevych 2019), its efficacy will be diminished when directly applied to long-document retrieval due to granularity mismatch. This is because the cross-encoder defined in Eq.(4) is able to focus only on the q-relevant topic of d via its fine-grained self-attention mechanism, regardless of other topics in the scope hypothesis. By comparison, the bi-encoder defined in Eq.(1) is constrained by its representation bottleneck (i.e., fixed-length low-dimensional vector by $\text{Enc}(\cdot)$), so it can only produce q-agnostic d representations as a whole.

To break the bottleneck during distillation, we propose to perform knowledge distillation over fine-grained text pieces instead of the whole document. However, an open question remains about how to derive consistent embeddings across granularity. In particular, to produce consistent embeddings, previous methods directly apply mean-pooling over contextual embeddings for different granularity, which however becomes inferior when the document length goes extremely long and has proven less effective in our pilot experiments. This is the reason why most previous document retrieval works rely on `[CLS]` embedding paradigm (Ma et al. 2022; Xiong et al. 2021; Zhan et al. 2021b).

Thereby, to better align with the prevalent `[CLS]` embedding paradigm, we present a global-consistent granularity embedding method. Specifically, '`[CLS]` embedding' denotes using the contextual embedding of `[CLS]` to represent the whole sequence, which is equivalent to applying a self-attention pooling (Lin et al. 2017; Shen et al. 2018) to the penultimate layer, i.e.,

$$\begin{aligned}\boldsymbol{v}^d &= \text{Transfm-Enc}(\texttt{[CLS]}\, d\, \texttt{[SEP]}|\theta^{(\text{d})})_{\texttt{[CLS]}} \\ &= \text{FFN}(\sum\nolimits_{i\in[1,|d|]} \sigma(\alpha_{\texttt{[CLS]}\leftarrow d_i}) h'_i)\end{aligned} \tag{6}$$

where i denotes the token index in d, h'_i denotes a hidden state for token d_i from the previous layer, σ denotes a non-linear function and usually softmax, $\alpha_{\texttt{[CLS]}\leftarrow d_i}$ denotes an attention probability from `[CLS]` to d_i, and FFN denotes post-processes including MLP and residual connection defined in the Transformer. The attention scores are calculated between global embedding $h'_{\texttt{[CLS]}}$ and each token embedding h'_i by the attention module in the last layer of the Transformer (Vaswani et al. 2017). Then, following such global-aware attention pooling, we can leverage the off-the-shelf attention scores to produce global-consistent embeddings across granularity. Formally, given an arbitrary text span $x \in d$ with the token indices $[b^x, e^x]$, its global-consistent embedding can be written as

$$\begin{aligned}\boldsymbol{v}^x &= \text{Enc}(x|d;\theta^{(\text{d})}) \\ &\coloneqq \text{FFN}(\sum\nolimits_{i\in[b^x,e^x]} \sigma(\alpha_{\texttt{[CLS]}\leftarrow d_i}) h'_i).\end{aligned} \tag{7}$$

Consequently, we can readily derive representation for various granularity, e.g., passages and sentences, via $\text{Enc}(x|d;\theta^{(\text{d})})$.

Remark on Propagation. Besides the mean-pooling methods (Reimers and Gurevych 2019), a recent trend to get multi-granular representation is employing graph neural network (GNN) (Wu et al. 2021) for deep embedding propagation (Zheng et al. 2020). Both of them focus on fine-grained representations rather than document-level ones and target the final applications of the representations, e.g., open-domain and context-based question answering. Standing with a distinct motivation, we still focus on the single document-level bottleneck but leverage fine-grained representations as the intermediate for knowledge distillation. This necessitates the paradigm of original global `[CLS]` representation, which requires consistency between document-level and fine-grained representations without complicated embedding propagation.

Local-aligned Score Distilling

After applying $\text{Enc}(x|d;\theta^{(\text{d})})$ to fine-grained text piece in d, we obtain fine-grained representations, respectively. That is

$$\boldsymbol{v}^{x^j_k} = \text{Enc}(x^j_k|d;\theta^{(\text{d})}), j \in [0,M], k \in [1,K^j], \tag{8}$$

where j is the index of granularity, M denotes the total number of granularity, i is the index of text piece in j-th granularity, and K^j denotes the number of total text pieces in j-th granularity. Here, $j = 0$ denotes the granularity at the document level, leading to $K^0 = 1$ and $d = x^0_1$. Then, we rewrite Eq.(1) to score multi-grained pieces as

$$\begin{aligned}s^{(\text{be})}_{j,k} &\coloneqq \mathcal{M}^{(\text{be})}(q, x^j_k|d;\theta^{(\text{be})}) = < \boldsymbol{u}, \boldsymbol{v}^{x^j_k} > \\ &\coloneqq < \text{Enc}(q|\theta^{(\text{q})}), \text{Enc}(x^j_k|d;\theta^{(\text{d})}) > .\end{aligned} \tag{9}$$

Next, following Eq.(2), we can also derive multi-granular score distributions as

$$\boldsymbol{p}^{(\text{be})}_{j,k} \coloneqq P(\text{x}^j_k|q, \{x^j_{k+}\} \cup \mathbb{N}^j_k; \theta^{(\text{be})}) = \frac{\exp(\mathcal{M}^{(\text{be})}(q, x^j_k|d;\theta^{(\text{be})})/\tau)}{\sum_{x'^j_k \in \{x^j_{k+}\}\cup\mathbb{N}^j_k} \exp(\mathcal{M}^{(\text{be})}(q, x'^j_k|d;\theta^{(\text{be})})/\tau)}, \tag{10}$$

where $\mathbb{N}^j_k$ is a set of negative samples in j-th granularity, which we dive into in the next sub-section.

After, we could apply the cross-encoder to each pair of q and x^j_k and its negative pairs for multi-granular distributions. It is noteworthy that differing from the bi-encoder, the score between the q and each x^j_k by cross-encoder is based solely on x^j_k, independent of the other parts in d. This is because, in contrast to our bi-encoder that takes global-consistent fine-grained representations to align document-level bottleneck

learning, the cross-encoder here aims to provide precise relevance scores to describe q-x_k^j relationships exactly. So, we obtain the cross-encoder's relevance scores by

$$s_{j,k}^{(\text{ce})} \coloneqq \mathcal{M}^{(\text{ce})}(q, x_k^j|\theta^{(\text{ce})}). \tag{11}$$

It is straightforward to get multi-granular score distribution by cross-encoder, i.e., $\boldsymbol{p}_{j,k}^{(\text{ce})} \coloneqq P(\text{x}_k^j|q, \{x_{k+}^j\} \cup \mathbb{N}_k^j; \theta^{(\text{ce})})$.

Lastly, we can define the training loss of our multi-granular aligned distillation as

$$L^{(\text{fkd})} = \sum_{j\in[1,M]} \frac{1}{K^j} \sum_{k\in[1,K^j]} \text{KL-Div}(\boldsymbol{p}_{j,k}^{(\text{be})} \| \boldsymbol{p}_{j,k}^{(\text{ce})}), \tag{12}$$

where $\text{KL-Div}(\cdot\|\cdot)$ denotes the KL divergence between the two distributions. It is remarkable that we do not include $j = 0$ here as the document-level relevance is only learned via contrastive learning. After replacing $L^{(\text{kd})}$ in § with the above $L^{(\text{fkd})}$, we get the final training loss of our FGD, i.e.,

$$L_{\theta^{(\text{bi})}}^{(\text{be})} = \lambda L^{(\text{cl})} + L^{(\text{fkd})}. \tag{13}$$

Please refer to Figure 2 for the illustration.

Remark on Overheads. The first thought that comes into our mind is that such extensive knowledge distillation from a heavy network will lead to massive training computation overheads. On the side of the student bi-encoder, there is only a little extra computation (i.e., applying the attention pooling multiple times with off-the-shelf attention scores as defined by Eq.(7)) in the top layer of the Transformer. On the side of the teacher cross-encoder, as the overheads grow quadratically with sequence length (i.e., $\mathcal{O}(n^2)$), applying cross-encoder to sub-granularity (e.g., passage and sentence) only results in a complexity of $\mathcal{O}(n \log n)$. Thereby, the complexity of FGD in invoking cross-encoder is even less than the traditional document-level distillation. Again, we would like to mention that we still use one single bottleneck vector to represent each document instead of multiple vectors (Santhanam et al. 2021; Humeau et al. 2020), where multi-granular embeddings only as the intermediate for distillations.

Hierarchical Hard Negative Mining

Hard negative mining has been proven very effective in achieving competitive performance by many previous works (Xiong et al. 2021; Wang et al. 2022). It leverages the best-so-far retriever to retrieve hard examples (i.e., top-relevant documents but not d_+) for each query q, which are used as negative documents $\mathbb{N}$ for the next round of retriever training.

Nonetheless, as formulated in Eq.(10), negative text pieces $\mathbb{N}_k^j$ are needed to sample at each j-th granularity. Notably, we cannot get the precise gold label(s) at every sub-document granularity x_{k+}^j in Eq.(10) except for the gold document d_+ (i.e., $j = 0$). As a weakly-supervised remedy (Yang et al. 2022), we regard each $x_k^j \in d_+$ as a positive text piece during our multi-granular aligned distillation. Thereby, we present a simple yet effective hierarchical hard negative mining technique from top to bottom. That is,

$$\mathbb{N}_k^j = \{x_{k-}^j | x_{k-}^j \sim \mathcal{M}^{(\text{be})}(q, x_k^j|d; \theta^{(\text{be})}) \tag{14}$$

$$\wedge\, x_k^j \in \mathbb{N}_*^{j-1}\}, \tag{15}$$

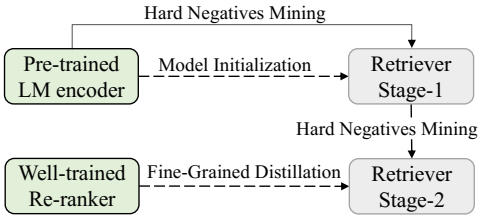

Figure 3: The pipeline of our method.

where $\mathbb{N}_*^{j-1}$ denotes all negatives in $(j-1)$-th granularity and $\mathbb{N}^0 = \mathbb{D} \setminus \{d_+\}$.

Experiments

Datasets and Metrics. In experiments, we conduct extensive evaluations of our method on the two long-document retrieval benchmark datasets: MS-Marco Doc (Nguyen et al. 2016) and TREC Deep Learning 2019 document retrieval (TREC 2019) (Craswell et al. 2020). Following previous works (Ma et al. 2022), we use official metrics MRR@100 and Recall@100 (R@100) to report evaluation results on MS-Marco dev, while using nDCG@10 and Recall@100 for TREC 2019.

Pre-training & Fine-tuning Pipeline

Following previous works (Ma et al. 2022), we detail our pre-training and fine-tuning pipelines (shown in Figure 3) for document retrieval by FGD.

Stage-0: Pre-training. Initialing a model by self-supervised pre-training has been proven effective by numerous works (Xiong et al. 2021; Zhan et al. 2021b; Ma et al. 2021, 2022), which can be categorized into two groups, i.e., *general pre-training* and *corpus-aware pre-training*. Specifically, the former is referred to as PLMs that are pre-trained on general corpora by language modeling (e.g., RoBERTa (Liu et al. 2019)). Built upon the former, the latter is proposed for continual pre-training on the collection corpus by language modeling and/or pseudo-label training (e.g., coCondenser (Gao and Callan 2022) and SimLM (Wang et al. 2022)). In this work, we test our framework on both, corresponding to RoBERTa and ED-MLM (Wang et al. 2022). In addition, following all previous works in document retrieval (Xiong et al. 2021; Zhan et al. 2021b), we also conduct a supervised pre-training on passage retrieval by default.

Stage-1: Warmup Fine-tuning. Providing the document-level hard negatives mined by the pre-trained retriever, the first fine-tuning step is based solely on the contrastive learning loss defined in Eq.(3) to warm up in retriever for document retrieval (Zhan et al. 2021b; Wang et al. 2022).

Stage-2: Continual Fine-tuning. Upon the retriever from the warmup, the hard negative mining is invoked again for more challenging negatives. In contrast to previous works merely employing the contrastive learning (Ma et al. 2022), we apply FGD by Eq.(13) for more competitive results.

Main Results

MS-Marco Doc. As shown in Table 1, our FGD consistently achieve the best performance across the two metrics

Method	Corpus-aware	MS-MARCO Doc Dev		TREC 2019 Doc	
		MRR@100	R@100	nDCG@10	R@100
Sparse or lexicon retriever					
BM25		0.277	0.808	0.519	0.395
DeepCT (Dai and Callan 2019)		0.320	-	0.544	-
BestTRECTrad (Craswell et al. 2020)		-	-	0.549	-
Dense retriever					
ANCE (Xiong et al. 2021)		0.377	0.894	0.610	0.273
BERT (Ma et al. 2022)		0.389	0.877	0.594	0.301
STAR (Zhan et al. 2021b)		0.390	0.913	0.605	0.313
ICT (Lee, Chang, and Toutanova 2019)	✓	0.396	0.882	0.605	0.303
PROP (Ma et al. 2021)	✓	0.394	0.884	0.596	0.298
B-PROP (Ma et al. 2021)	✓	0.395	0.883	0.601	0.305
SEED (Lu et al. 2021)	✓	0.396	0.902	0.605	0.307
RepCONC (Zhan et al. 2022)		0.399	0.911	0.600	0.305
JPQ (Zhan et al. 2021a)		0.401	0.914	0.623	-
ADORE+STAR (Zhan et al. 2021b)		0.405	0.919	0.628	0.317
SeDR (Chen et al. 2022)		0.409	0.921	0.632	0.343
COSTA (Ma et al. 2022)	✓	0.422	0.919	0.626	0.320
FGD-STAR (ours)		0.430	0.915	0.629	0.338
FGD (ours)	✓	**0.440**	**0.925**	**0.635**	**0.349**

Table 1: Comparison results on MS-Marco and TREC 2019 datasets.

on MS-Marco document retrieval benchmark. Compared to STAR (Zhan et al. 2021b), our FGD-STAR exhibits 4% MRR@100 absolute improvement and also outperforms ADORE+STAR (Zhan et al. 2021b). When coupled with corpus-aware pre-training, ED-MLM (Wang et al. 2022), FGD achieves state-of-the-art performance, surpassing carefully designed COSTA model (Ma et al. 2022).

TREC Deep Learning 2019 Doc. It is observed that our method is superior to its baselines and competitors consistently in Table 1, and achieves state-of-the-art effectiveness across different metrics on TREC Deep Learning 2019 Doc.

Ablation Study and Model Choice

To further investigate the contribution of each part in our model, we conduct an ablation study as shown in Table 2.

Distillation Medium. In our FGD framework, we opt in to distill the ranker's relevance information in fine granularity (i.e., passage and sentence), however without document-level distillation for the sake of granularity mismatch. As listed in the table, discarding fine-grained distillation at either the passage (pass) or sentence level leads to a 0.5% MRR@100 drop. Interestingly, equipping our proposed FGD with document-level distillation results in a 0.4% MRR@100 drop, which verifies severity of the granularity mismatch issue.

Fine-grained Representation. To verify the global-consistent granularity embedding in our proposed model, we evaluate our FGD with other schemes to derive the fine-grained representation for distillation medium. First, by replacing our global-consistent embedding with mean-pooling over the corresponding tokens for fine-grained representation (i.e., FGD w/ FG pooling), there is global-local inconsistency to derive representation, diminishing the model by 1.4% MRR@100. Second, to alleviate the inconsistency, we also try to obtain all embeddings at different granularity with mean-pooling (i.e., FGD w/ All pooling), but there is still a 1.5% MRR@100 gap from our main result. This is because mean-pooling become inferior when representing a long text. Third, instead of leveraging contextualized fine-grained representation in our FGD, we make the text inputs symmetric between the retriever and ranker by separating text pieces into passages and sentences (i.e., FGD w/ Separate Pieces). Despite all representations derived from `[CLS]`, results show non-contextual representations still cause 1.1% MRR@100 drop due to less effectiveness of the distillation process.

Baseline Methods. To check exact improvement brought by FGD, we ablate all our major modules. First, we remove all the fine-grained distillation but opt in mere document-level distillation, leading to 42.8% MRR@100 (-1.2%). Second, we fine-tune a retriever at stage 2 without any distillation, resulting in 42.7% MRR@100 (-1.3%).

Further Analysis

Compatibility with Other Retriever/Ranker. To verify the generality of our proposed distillation framework, we replace either retriever backbone (i.e., the student model) or ranker model (i.e., the teacher model) in our FGD framework. As shown in Table 3, when replacing ED-MLM initialization with STAR (Zhan et al. 2021b) for the retriever, the fine-tuned results still exhibit consistent improvement, i.e., 1.3% MRR@10 over its baseline, STAR retriever at stage 2. Then, when replacing the used ranker, R2anker, with a document ranker by (Gao, Dai, and Callan 2021), a remarkable improvement (+1.1% MRR@10) is still observed in contrast to its baseline, i.e., fine-tuning ED-MLM encoder w/o distillation.

Method	MARCO Dev	
	MRR@100	R@100
FGD (stg2)	**0.440**	**0.925**
Distillation Medium		
◊ FGD w/o pass-distill	0.435	0.924
◊ FGD w/o sent-distill	0.435	0.925
◊ FGD w/ doc-distill	0.436	0.924
Fine-grained Representation		
◊ FGD w/ FG pooling	0.426	0.924
◊ FGD w/ All pooling	0.425	0.923
◊ FGD w/ Separate Pieces	0.429	0.924
Baseline Methods		
◊ only doc-distill	0.428	0.923
◊ w/o ALL	0.427	0.923

Table 2: Ablation study. 'FG' is 'fine-grained', 'w/o ALL' is equivalent to 'ED-MLM' at stage 2 (stg2).

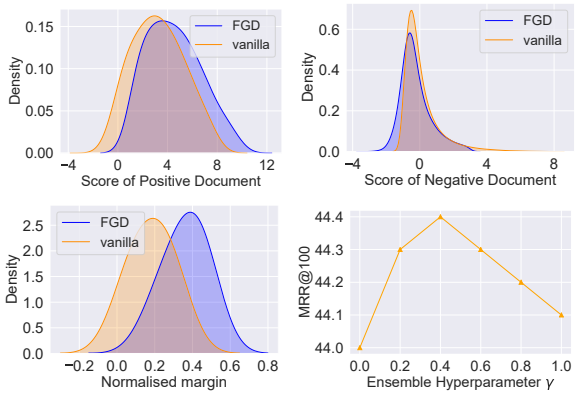

Figure 4: Comparison of our FGD (top left) and vanilla document-level distillation (top right) model on the MS-Marco test set; and the margins (bottom left) of FGD and vanilla. Different γ for multi-vector retrieval (bottom right).

Generalization Improvement by FGD. Following Menon et al. (2022), it is intuitive to leverage distributions of positive scores, negative scores, and their margins for generalization analysis. This is because a more generalizable retriever is prone to produce higher relevance scores for positive pairs, lower relevance scores for negative pairs, and larger margins for test triples (i.e., a query, its positive and negative). In Figure 4, compared to document-level distillation, the proposed FGD is more capable of distinguishing a positive pair from the negative ones, revealing its generalization.

Multi-vector Retrieval. The multi-grained representations across documents, passages, and sentences allow us to break single-vector information bottleneck and perform multi-vector retrieval. Despite unsatisfactory retrieval based on either passage- or sentence-level embeddings, we find that it is effective to complement document-level vectors with fine-grained ones, i.e.,

$$s^{(be)} := s^{(be)} + \gamma \cdot (\max_i(s^{(be)}_{1,i}) + \max_j(s^{(be)}_{2,j})).$$

As such, we tune the hyperparameter γ by grid search in Figure 4 and find model achieves optimal performance when $\gamma = 0.4$. As listed in Table 4, our multi-vector retrieval can bring 0.4% absolute improvement to hit 44.4% MRR@100.

Method	MARCO Dev	
	MRR@100	R@100
FGD (ED-MLM + psg-ranker)	**0.440**	**0.925**
Replacing the bi-encoder (student) retriever		
STAR (stg2)	0.417	0.914
FGD (STAR as student)	0.430	0.915
Replacing the cross-encoder (teacher) ranker		
ED-MLM (stg2)	0.427	0.923
FGD (doc-ranker as teacher)	0.438	0.923

Table 3: Results with the other retriever or ranker.

Method	MARCO Dev Doc	
	MRR@100	R@100
Previois SoTA	0.422	0.919
FGD	0.440	0.925
FGD + multi	**0.444**	**0.926**

Table 4: FGD with multi-vector (i.e., 'multi') retrieval.

Method	MARCO Dev Doc	
	MRR@100	R@100
FGD (global-consistent)	0.440	**0.925**
- FGD w/ RGAT	**0.441**	**0.925**
- FGD w/ FG pooling	0.426	0.924

Table 5: Comparisons on MS-Marco dev w.r.t. different methods to derive fine-grained representations.

Fine-Grained Propagation. In addition to our global-consistent granularity embedding method, we also present two fine-grained representation derivation methods: *FG pooling* that use mean-pooling to aggregate the corresponding tokens and *RGAT* that leverage position- & granularity-specific features for deep graph propagation (Busbridge et al. 2019). Although the latter excels fine-grained representation, it is likely to break our local-global representing consistency and be redundant against the deep contextualized Transformer encoder. Thus, despite complexity, results shown in Table 5 are similar to those of the more concise FGD.

Conclusion

In this work, we propose a new knowledge distillation framework for long-document retrieval, called fine-grained distillation (FGD). Integrated with the hierarchical hard negative mining technique, the proposed framework produces fine-grained representations consistent with the global document-level one and then distills multi-granular score distributions from a heterogeneous cross-encoder. The proposed framework will not affect the long-document retrieval procedure in terms of both retrieval paradigm and efficiency. The experiment results show that the proposed framework achieves a state-of-the-art quality in document retrieval and is compatible with a broad spectrum of baseline choices in terms of both the bi-encoder student and the cross-encoder teacher.

References

Busbridge, D.; Sherburn, D.; Cavallo, P.; and Hammerla, N. Y. 2019. Relational Graph Attention Networks. *CoRR*, abs/1904.05811.

Cai, Y.; Fan, Y.; Guo, J.; Sun, F.; Zhang, R.; and Cheng, X. 2021. Semantic Models for the First-stage Retrieval: A Comprehensive Review. *CoRR*, abs/2103.04831.

Chen, G.; Choi, W.; Yu, X.; Han, T. X.; and Chandraker, M. 2017. Learning Efficient Object Detection Models with Knowledge Distillation. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 742–751.

Chen, J.; Chen, Q.; Li, D.; and Huang, Y. 2022. SeDR: Segment Representation Learning for Long Documents Dense Retrieval. *CoRR*, abs/2211.10841.

Craswell, N.; Mitra, B.; Yilmaz, E.; Campos, D.; and Voorhees, E. M. 2020. Overview of the TREC 2019 deep learning track. *CoRR*, abs/2003.07820.

Dai, Z.; and Callan, J. 2019. Context-Aware Sentence/Passage Term Importance Estimation For First Stage Retrieval. *CoRR*, abs/1910.10687.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Fan, Y.; Xie, X.; Cai, Y.; Chen, J.; Ma, X.; Li, X.; Zhang, R.; and Guo, J. 2022. Pre-training Methods in Information Retrieval. *Found. Trends Inf. Retr.*, 16(3): 178–317.

Gao, L.; and Callan, J. 2021. Condenser: a Pre-training Architecture for Dense Retrieval. In *EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, 981–993. Association for Computational Linguistics.

Gao, L.; and Callan, J. 2022. Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval. In *ACL 2022, Dublin, Ireland, May 22-27, 2022*, 2843–2853. Association for Computational Linguistics.

Gao, L.; Dai, Z.; and Callan, J. 2021. Rethink Training of BERT Rerankers in Multi-stage Retrieval Pipeline. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II*, volume 12657 of *Lecture Notes in Computer Science*, 280–286. Springer.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2021. Deberta: decoding-Enhanced Bert with Disentangled Attention. In *ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Humeau, S.; Shuster, K.; Lachaux, M.; and Weston, J. 2020. Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P. S. H.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *EMNLP 2020, Online, November 16-20, 2020*, 6769–6781. Association for Computational Linguistics.

Khattab, O.; and Zaharia, M. 2020. 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*, 39–48. ACM.

Lee, J.; Sung, M.; Kang, J.; and Chen, D. 2021. Learning Dense Representations of Phrases at Scale. In *ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, 6634–6647. Association for Computational Linguistics.

Lee, K.; Chang, M.; and Toutanova, K. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In *ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 6086–6096. Association for Computational Linguistics.

Lin, S.-C.; Yang, J.-H.; and Lin, J. 2021. In-Batch Negatives for Knowledge Distillation with Tightly-Coupled Teachers for Dense Retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, 163–173. Online: Association for Computational Linguistics.

Lin, Z.; Feng, M.; dos Santos, C. N.; Yu, M.; Xiang, B.; Zhou, B.; and Bengio, Y. 2017. A Structured Self-Attentive Sentence Embedding. In *ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Liu, C.; Tao, C.; Feng, J.; and Zhao, D. 2022a. Multi-Granularity Structural Knowledge Distillation for Language Model Compression. In *ACL 2022, Dublin, Ireland, May 22-27, 2022*, 1001–1011. Association for Computational Linguistics.

Liu, F.; Jiao, Y.; Massiah, J.; Yilmaz, E.; and Havrylov, S. 2022b. Trans-Encoder: Unsupervised sentence-pair modelling through self- and mutual-distillations. In *ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Lu, S.; Xiong, C.; He, D.; Ke, G.; Malik, W.; Dou, Z.; Bennett, P.; Liu, T.; and Overwijk, A. 2021. Less is More: Pre-training a Strong Siamese Encoder Using a Weak Decoder. *CoRR*, abs/2102.09206.

Lu, Y.; Liu, Y.; Liu, J.; Shi, Y.; Huang, Z.; Feng, S.; Sun, Y.; Tian, H.; Wu, H.; Wang, S.; Yin, D.; and Wang, H. 2022. ERNIE-Search: Bridging Cross-Encoder with Dual-Encoder via Self On-the-fly Distillation for Dense Passage Retrieval. *CoRR*, abs/2205.09153.

Ma, X.; Guo, J.; Zhang, R.; Fan, Y.; and Cheng, X. 2022. Pre-train a Discriminative Text Encoder for Dense Retrieval via Contrastive Span Prediction. In *SIGIR '22, Madrid, Spain, July 11 - 15, 2022*, 848–858. ACM.

Ma, X.; Guo, J.; Zhang, R.; Fan, Y.; Li, Y.; and Cheng, X. 2021. B-PROP: Bootstrapped Pre-training with Representa-

tive Words Prediction for Ad-hoc Retrieval. In *SIGIR '21, Virtual Event, Canada, July 11-15, 2021*, 1318–1327. ACM.

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Kim, S.; Reddi, S. J.; and Kumar, S. 2022. In defense of dual-encoders for neural ranking. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 15376–15400. PMLR.

Nguyen, T.; Rosenberg, M.; Song, X.; Gao, J.; Tiwary, S.; Majumder, R.; and Deng, L. 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.

Qu, Y.; Ding, Y.; Liu, J.; Liu, K.; Ren, R.; Zhao, W. X.; Dong, D.; Wu, H.; and Wang, H. 2021. RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. In *NAACL-HLT 2021, Online, June 6-11, 2021*, 5835–5847. Association for Computational Linguistics.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 3980–3990. Association for Computational Linguistics.

Ren, R.; Qu, Y.; Liu, J.; Zhao, W. X.; She, Q.; Wu, H.; Wang, H.; and Wen, J. 2021. RocketQAv2: A Joint Training Method for Dense Passage Retrieval and Passage Re-ranking. In *EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, 2825–2835. Association for Computational Linguistics.

Robertson, S. E.; and Zaragoza, H. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.*, 3(4): 333–389.

Santhanam, K.; Khattab, O.; Saad-Falcon, J.; Potts, C.; and Zaharia, M. 2021. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. *CoRR*, abs/2112.01488.

Shen, T.; Zhou, T.; Long, G.; Jiang, J.; Pan, S.; and Zhang, C. 2018. DiSAN: Directional Self-Attention Network for RNN/CNN-Free Language Understanding. In *AAAI-18, New Orleans, Louisiana, USA, February 2-7, 2018*, 5446–5455. AAAI Press.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008.

Wang, L.; Yang, N.; Huang, X.; Jiao, B.; Yang, L.; Jiang, D.; Majumder, R.; and Wei, F. 2022. SimLM: Pre-training with Representation Bottleneck for Dense Passage Retrieval. *CoRR*, abs/2207.02578.

Wu, L.; Chen, Y.; Shen, K.; Guo, X.; Gao, H.; Li, S.; Pei, J.; and Long, B. 2021. Graph Neural Networks for Natural Language Processing: A Survey. *CoRR*, abs/2106.06090.

Xiong, L.; Xiong, C.; Li, Y.; Tang, K.; Liu, J.; Bennett, P. N.; Ahmed, J.; and Overwijk, A. 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.

Yang, J.; Duan, J.; Tran, S.; Xu, Y.; Chanda, S.; Chen, L.; Zeng, B.; Chilimbi, T.; and Huang, J. 2022. Vision-Language Pre-Training with Triple Contrastive Learning. In *CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 15650–15659. IEEE.

Yu, S.; Liu, Z.; Xiong, C.; Feng, T.; and Liu, Z. 2021. Few-Shot Conversational Dense Retrieval. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, 829–838. ACM.

Zhan, J.; Mao, J.; Liu, Y.; Guo, J.; Zhang, M.; and Ma, S. 2021a. Jointly Optimizing Query Encoder and Product Quantization to Improve Retrieval Performance. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, 2487–2496. ACM.

Zhan, J.; Mao, J.; Liu, Y.; Guo, J.; Zhang, M.; and Ma, S. 2021b. 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*, 1503–1512. ACM.

Zhan, J.; Mao, J.; Liu, Y.; Guo, J.; Zhang, M.; and Ma, S. 2022. Learning Discrete Representations via Constrained Clustering for Effective and Efficient Dense Retrieval. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, 1328–1336. ACM.

Zhang, H.; Gong, Y.; Shen, Y.; Lv, J.; Duan, N.; and Chen, W. 2022a. Adversarial Retriever-Ranker for Dense Text Retrieval. In *International Conference on Learning Representations*.

Zhang, S.; Liang, Y.; Gong, M.; Jiang, D.; and Duan, N. 2022b. Multi-View Document Representation Learning for Open-Domain Dense Retrieval. In *ACL 2022, Dublin, Ireland, May 22-27, 2022*, 5990–6000. Association for Computational Linguistics.

Zheng, B.; Wen, H.; Liang, Y.; Duan, N.; Che, W.; Jiang, D.; Zhou, M.; and Liu, T. 2020. Document Modeling with Graph Attention Networks for Multi-grained Machine Reading Comprehension. In *ACL 2020, Online, July 5-10, 2020*, 6708–6718. Association for Computational Linguistics.

Zhou, Y.; Shen, T.; Geng, X.; Tao, C.; Xu, C.; Long, G.; Jiao, B.; and Jiang, D. 2023. Towards Robust Ranker for Text Retrieval. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, 5387–5401. Association for Computational Linguistics.