Fine-Grained Distillation for Long Document Retrieval

Yucheng Zhou1, Tao Shen2, Xiubo Geng3, Chongyang Tao3, Jianbing Shen1, Guodong Long2, Can Xu3, Daxin Jiang3†

1SKL-IOTSC, CIS, University of Macau
2AAII, FEIT, University of Technology Sydney
3Microsoft 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 across 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 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 SKL-IOTSC(UM)-2021-2023. MYRG-CRG2022-00013-IOTSC-ICI grant and SRG2022-00023-IOTSC grant.
† Corresponding author.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

†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.

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 adopted 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 helps to derive fine-grained representations without cross-granular conflict.

- 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 granularity 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
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., $D = \{d_i\}_{i=1}^{|D|}$) where each $d_i$ denotes a document, large-scale retrieval is to fetch top-relevance documents (i.e., $D'$) by a retriever (e.g., $M$) for a text query $q$. This requires $M$ to calculate every relevance score $s_i^q$ between the $q$ and $\forall d_i \in D$, where $i \in [1, |D|]$. 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.,

$$s_{be} := M_{be}(q, d_{be}) := \langle \text{Enc}(q|\theta(q)), \text{Enc}(d|\theta(d)) \rangle, \forall d \in D,$$

where $\langle \cdot, \cdot \rangle$ denotes a non-parametric dot-product, $\text{Enc}(|\theta(\cdot))$ denotes a $\theta(\cdot)$-parameterized encoder that embeds a piece of text into a dense vector, and $\theta_{be} = \theta(q) \cup \theta(d)$ 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 $\theta_{be}$ 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 find the negatives. Providing $d_+$ and $\mathbb{N}$, we derive a score distribution over them,

$$p_{be} := P(d | q, \{d_+\} \cup \mathbb{N}; \theta_{be}) = \frac{\exp(M_{be}(q, d_{be})/\tau)}{\sum_{d' \in \{d_+\} \cup \mathbb{N}} \exp(M_{be}(q, d'_{be})/\tau)},$$

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

$$L^{(ce)} = -\sum_q \log P(d = d_+ | q, \{d_+\} \cup \mathbb{N}; \theta_{be}) = -\sum_q \log p_{be}.$$  

(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.,

$$s_{ce} := M_{ce}(q, d_{ce}) =$$

$$\text{Transformer-cls}([\text{CLS}] q [\text{SEP}] d [\text{SEP}] | \theta_{ce}),$$

where $s_{ce} \in \mathbb{R}$ and $\theta_{ce}$ 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.

![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.](image-url)
score. Note that, \( \theta^{(ce)} \) 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 \( P^{(ce)} \) as in Eq.(2). Last, the loss function of this distillation is

\[
L^{(kd)} = \text{KL-Div}(p^{(be)} \| p^{(ce)}). 
\]

So, the final training loss for the bi-encoder learning with distillation is written as \( AL^{(ce)} + L^{(bd)} \).

**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 Enc(·)), 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.,

\[
v^d = \text{Transfm-Enc}([\text{CLS}] d [\text{SEP}] \mid \theta^{(d)})_{[\text{CLS}]}
\]

\[
= \text{FFN}(\sum_{i \in [1,|d|]} \sigma(\alpha([\text{CLS}] \rightarrow d_i)h'_i) 
\]

where \( i \) denotes the token index in \( d \), \( h'_i \) denotes a hidden state for token \( d_i \) from the previous layer, \( \sigma \) denotes a non-linear function and usually \text{softmax}, \( \alpha([\text{CLS}] \rightarrow d_i) \) denotes an attention probability from [CLS] to \( d_i \), and \( \text{FFN} \) denotes post-processes including MLP and residual connection defined in the Transformer. The attention scores are calculated between global embedding \( h'_{[\text{CLS}]} \) 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

\[
v^x = \text{Enc}(x \mid d; \theta^d) 
\]

\[
= \text{FFN}(\sum_{i \in [b^x, e^x]} \sigma(\alpha([\text{CLS}] \rightarrow d_i)h'_i). 
\]

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

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

\[
v^x = \text{Enc}(x \mid d; \theta^d), j \in [0,M], k \in [1,K^j],
\]

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 \( d = x^d_0 \). Then, we rewrite Eq.(1) to score multi-grained pieces as

\[
\alpha_j^d := \mathcal{N}^{(be)}(q, x^j_k \mid d; \theta^{(be)}) = \langle u, v^x \rangle
\]

\[
= \text{Enc}(q \mid \theta^{(q)}), \text{Enc}(x^j_k \mid d; \theta^{(d)}) >.
\]

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

\[
P^{(be)}_{j,k} := P(x^j_k \mid q, \{x^j_{k+1}\} \cup \mathbb{N}^j_k; \theta^{(be)}) = \exp(\mathcal{M}^{(be)}(q, x^j_k \mid d; \theta^{(be)}/\tau) / \tau / \sum_{x^j_k \in \mathbb{N}^j_k} \exp(\mathcal{M}^{(be)}(q, x^j_k \mid d; \theta^{(be)/\tau}))),
\]

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 $q$-$x^j_k$ relationships exactly. So, we obtain the cross-encoder’s relevance scores by
\[ s^{(ce)}_{j,k} := \mathcal{M}^{(ce)}(q, x^j_k; \theta^{(ce)}), \tag{11} \]
It is straightforward to get multi-granular score distribution by cross-encoder, i.e.,
\[ p^{(ce)}_{j,k} := P(x^j_k | q, \{x^j_{k-1}\} \cup \mathbb{N}^j_k; \theta^{(ce)}). \]
Lastly, we can define the training loss of our multi-granular aligned distillation as
\[ L^{(fd)} = \sum_{j \in [1,M]} \frac{1}{K_j} \sum_{k \in [1,K_j]} \text{KL-Div}(p^{(be)}_{j,k} \parallel p^{(ce)}_{j,k}), \tag{12} \]
where KL-Div(\|\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 $L^{(fd)}$ in § with the above $L^{(fd)}$, we get the final training loss of our FGD, i.e.,
\[ L^{(fgd)} = \lambda L^{(ce)} + L^{(fd)}. \tag{13} \]

Remarks 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 a heavy network will lead to massive training computation overheads. On the side of the student bi-encoder, as the overheads grow quadratically with sequence length (i.e., $O(n^2)$), applying cross-encoder to sub-granularity (e.g., passage and sentence) only results in a complexity of $O(n \log n)$. 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 $d_\ast$) 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 $\mathbb{N}^j_k$ are needed to sample at each $j$-th granularity. Notably, we cannot get the precise gold label(s) at every sub-document granularity $x^j_{k+}$ in Eq.(10) except for the gold document $d_\ast$ (i.e., $j = 0$). As a weakly-supervised remedy (Yang et al. 2022), we regard each $x^j_{k+}$ 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,
\[ \mathbb{N}^j_k = \{x^j_{k-}|x^j_{k-} \sim \mathcal{M}^{(be)}(q, x^j_{k-}; d; \theta^{(be)}) \}
\& x^j_k \in \mathbb{N}^{j-1}_k \}, \tag{14} \]
\[ \mathbb{N}^0 = \mathbb{D} \setminus \{d_\ast\}. \]
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 our all 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.
Table 2: Ablation study. ‘FG’ is ‘fine-grained’, ‘w/o ALL’ is equivalent to ‘ED-MLM’ at stage 2 (stg2).

<table>
<thead>
<tr>
<th>Method</th>
<th>MARCO Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MRR@100</td>
</tr>
<tr>
<td>FGD (stg2)</td>
<td>0.440</td>
</tr>
</tbody>
</table>

**Distillation Medium**

<table>
<thead>
<tr>
<th>Method</th>
<th>MARCO Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MRR@100</td>
</tr>
<tr>
<td>FGD w/o pass-distill</td>
<td>0.435</td>
</tr>
<tr>
<td>FGD w/o sent-distill</td>
<td>0.435</td>
</tr>
<tr>
<td>FGD w/ doc-distill</td>
<td>0.436</td>
</tr>
</tbody>
</table>

**Fine-grained Representation**

<table>
<thead>
<tr>
<th>Method</th>
<th>MARCO Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MRR@100</td>
</tr>
<tr>
<td>FGD w/ FG pooling</td>
<td>0.426</td>
</tr>
<tr>
<td>FGD w/ All pooling</td>
<td>0.425</td>
</tr>
<tr>
<td>FGD w/ Separate Pieces</td>
<td>0.429</td>
</tr>
</tbody>
</table>

**Baseline Methods**

<table>
<thead>
<tr>
<th>Method</th>
<th>MARCO Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MRR@100</td>
</tr>
<tr>
<td>only doc-distill</td>
<td>0.428</td>
</tr>
<tr>
<td>w/o ALL</td>
<td>0.427</td>
</tr>
</tbody>
</table>

Table 3: Results with the other retriever or ranker.

<table>
<thead>
<tr>
<th>Method</th>
<th>MARCO Dev Doc</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MRR@100</td>
</tr>
<tr>
<td>Previous SoTA</td>
<td>0.422</td>
</tr>
<tr>
<td>FGD</td>
<td>0.440</td>
</tr>
<tr>
<td>FGD + multi</td>
<td>0.444</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>MARCO Dev Doc</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MRR@100</td>
</tr>
<tr>
<td>FGD (global-consistent)</td>
<td>0.440</td>
</tr>
<tr>
<td>- FGD w/ RGAT</td>
<td>0.441</td>
</tr>
<tr>
<td>- FGD w/ FG pooling</td>
<td>0.426</td>
</tr>
</tbody>
</table>

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

$\gamma = 0.4$. As listed in Table 4, our multi-vector retrieval can bring 0.4% absolute improvement to hit 44.4% MRR@100.

**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


Ma, X.; Guo, J.; Zhang, R.; Fan, Y.; Li, Y.; and Cheng, X. 2021. B-PROP: Bootstrapped Pre-training with Representa-
Effective Retrieval via Lightweight Late Interaction. \textit{CoRR}, abs/2112.01488.


