MTA4DPR: Multi-Teaching-Assistants Based Iterative Knowledge Distillation for Dense Passage Retrieval

Qixi Lu Endong Xun Gongbo Tang

Abstract

Dense Passage Retrieval (DPR) has achieved remarkable success in open-domain question answering. Knowledge distillation (KD) is widely adopted to enhance the performance of DPR by transferring knowledge from a cross-encoder to a dual-encoder. However, existing KD-based DPR methods usually rely on a single teacher model and perform only one round of distillation, which may limit the effectiveness of knowledge transfer. In this paper, we propose MTA4DPR, a novel multi-teaching-assistants based iterative knowledge distillation framework for DPR. Specifically, we first construct multiple assistant models with diverse architectures and capabilities to provide complementary supervision signals. Then, we design a fusion strategy to integrate the knowledge from multiple assistants effectively. Furthermore, we introduce an iterative distillation mechanism that progressively refines the student model by repeatedly leveraging the enhanced knowledge from updated assistants. Extensive experiments on several benchmark datasets demonstrate that our method significantly outperforms existing KD-based DPR approaches. Detailed analyses further verify the effectiveness of each component in our framework.

1 Introduction

Although PLM/LLM-based Dense Passage Retrieval (DPR) models [**?**, **?**] have superior performance, those models' inference efficiency and deployment costs are still cumbering their wide applications. To obtain an efficient and effective DPR model, researchers are paying more attention to knowledge distillation. Previous studies [**?**, **?**, **?**] have proved the effectiveness of knowledge distillation in DPR. However, the performance gap between the teacher and the distilled student often remains significant, especially when the teacher is a very good one.

In this paper, we hypothesize that incorporating assistants into knowledge distillation can help improve students' performance, just as teaching assistants in universities can assist students in learning course content. In addition, inspired by curriculum learning [**?**], we also believe that multiple iterations can further narrow the gap between the teacher and the students since the latter is capable of learning from more challenging data and more effective assistants as the iterations go on. Therefore, we introduce MTA4DPR, a multi-teaching-assistants based iterative distillation method. Specifically, MTA4DPR transfers knowledge from the teacher to the student with the help of multiple assistants iteratively. For each iteration, we first use off-the-shelf teacher/assistant DPR models to generate datasets for training and evaluation. Then, we use a fusion module to generate a series of fused assistants. After that, we train the student to learn from the teacher with the help of the best assistant selected among all fused and original assistants by our selection module, as illustrated in Figure 1. At the end of each iteration, we evaluate the student's performance and replace the worst-performing assistant with it if it outperforms any existing assistants. What's more, we also incorporate data that the student predicted incorrectly in the previous iteration into the newly constructed dataset, by which the difficulty of each iteration's dataset is increased. In

Figure 1: MTA4DPR Framework. MTA4DPR transfers knowledge from the teacher to the student with the help of the best assistant. The Fusion Module is used to generate fused assistants from the original assistants, and the Selection Module is used to select the best assistant among all original and fused assistants. The dotted arrows indicate that the corresponding procedures are not involved in the backpropagation of the training.

this way, as the training iterates, the student can learn from more performant assistants and more difficult data.

The experimental results on MSMARCO, TREC DL 2019 and 2020 and Natural Questions show the effectiveness of our method. Our 66M student model achieves the state-of-the-art performance among models with same parameters on multiple datasets, and is competitive when compared with larger, even LLM-based, DPR models.

To summarize, our main contributions are:

1. We propose a novel distillation method MTA4DPR, which improves the student's retrieval performance with the help of assistant models.

2. The experimental results show the effectiveness of our proposed method, achieving very competitive results even when compared with larger, even LLM-based, DPR models.

3. Not constrained by model structures and tasks, MTA4DPR is orthogonal to existing distillation methods and can be combined with other distillation pipelines to further improve the performance.

1.1 Dense Retrieval

Dense Passage Retrieval (DPR) [**?**] has become a fundamental component in open-domain question answering systems. Unlike traditional sparse retrieval methods such as BM25, DPR leverages pre-trained language models to encode queries and passages into dense vectors and computes their similarities in a shared embedding space. Following the success of DPR, numerous works [**?**, **?**, **?**] have proposed various improvements, including better pre-training strategies, data augmentation techniques, and more effective training objectives. Recent advances further explore large language models (LLMs) for retrieval tasks [**?**], achieving strong performance but at the cost of increased computational complexity and deployment overhead.

1.2 Knowledge Distillation

Knowledge distillation (KD) [**?**] is a widely adopted technique for compressing large models into smaller ones while preserving performance. In the context of DPR, KD typically involves transferring knowledge from a powerful cross-encoder or large dual-encoder to a smaller student model. Previous studies [**?**, **?**, **?**] have demonstrated that KD can effectively enhance the retrieval capability of compact models by utilizing soft labels, intermediate representations, or ranking signals from teacher models. However, most existing KD-based DPR methods rely on a single teacher and perform only one round of distillation, which may limit the diversity and richness of the transferred knowledge.

To address these limitations, some works investigate iterative distillation or multi-teacher frameworks in other domains, showing that multiple rounds of supervision or diverse teacher signals can improve student performance. Nevertheless, such strategies have not been fully explored in DPR.

In this work, we extend the idea of multi-teacher and iterative distillation to dense retrieval, aiming to further bridge the performance gap between teacher and student models.

2 Methodology

2.1 Preliminary

2.1.1 Task Description

Given a query q and a corpus of passages $\mathcal{P} = p_1, p_2, \ldots, p_N$, the goal of dense passage retrieval is to retrieve the most relevant passages for q from $\mathcal{P}$. Typically, a dual-encoder framework is adopted, where the query encoder $E_q(\cdot)$ and the passage encoder $E_p(\cdot)$ map queries and passages into a shared embedding space. The similarity between a query and a passage is computed by the dot product of their embeddings.

2.1.2 Dual-Encoders and Cross-Encoders

In the dual-encoder architecture, the relevance score between a query q and a passage p is defined as:

$$s(q,p) = E_q(q)^\top E_p(p). \tag{1}$$

During training, given a positive passage p^+ and a set of negative passages $\mathcal{N} = p_1^-, \ldots, p_m^-$, the objective is to maximize the probability of p^+ being more relevant than negatives:

$$P(p^+|q) = \frac{\exp(s(q,p^+))}{\exp(s(q,p^+)) + \sum_{p^- \in \mathcal{N}} \exp(s(q,p^-))}. \tag{2}$$

The corresponding loss function is:

$$\mathcal{L}_{\text{CE}} = -\log P(p^+|q). \tag{3}$$

In contrast, cross-encoders concatenate the query and passage as input and compute the relevance score by jointly encoding them, which generally yields better performance but incurs higher computational cost.

2.1.3 Knowledge Distillation for DPR

Knowledge distillation for DPR typically transfers knowledge from a teacher model (e.g., cross-encoder or large dual-encoder) to a student dual-encoder. Let $s_T(q,p)$ and $s_S(q,p)$ denote the scores predicted by the teacher and student models, respectively. The distillation objective can be formulated as:

$$\mathcal{L} * \text{KD} = \sum * p \in p^+, \mathcal{N} \text{KL}(\sigma(s_T(q,p)), ||, \sigma(s_S(q,p))), \tag{4}$$

where $\sigma(\cdot)$ denotes the softmax function over candidate passages.

The overall training loss combines the supervised loss and the distillation loss:

$$\mathcal{L} = \alpha \mathcal{L} * \text{CE} + (1-\alpha)\mathcal{L} * \text{KD}, \tag{5}$$

where α is a hyper-parameter controlling the trade-off between the two losses.

2.2 The MTA4DPR Framework

MTA4DPR is a multi-teaching-assistants based iterative knowledge distillation framework for dense passage retrieval. The overall pipeline consists of multiple iterations. In each iteration, we first prepare training and evaluation data using the teacher and assistant models. Then, we construct fused assistants through a fusion strategy. Next, we select the best assistant and distill knowledge from the teacher to the student with its help. Finally, we update the assistant pool based on the student's performance.

2.2.1 Data Preparation

At the beginning of each iteration, we use the teacher and assistant models to retrieve candidate passages for each query. These retrieval results are used to construct the training dataset. Additionally, we incorporate the queries that the student predicted incorrectly in the previous iteration to increase the difficulty of the dataset.

Let $\mathcal{D}^{(t)}$ denote the dataset in the t-th iteration. It consists of tuples $(q, p^+, \mathcal{N})$ generated based on the retrieval results of teacher and assistants. The dataset is progressively updated as iterations proceed.

2.2.2 Fusion Strategy

To leverage complementary knowledge from multiple assistants, we design a fusion strategy to generate fused assistants. Given a set of assistant models $A_1, A_2, \ldots, A_K$, we combine their predicted scores for each query-passage pair:

$$s_F(q, p) = \sum_{k=1}^{K} w_k s_{A_k}(q, p), \tag{6}$$

where w_k denotes the weight assigned to the k-th assistant, and $s_{A_k}(q, p)$ is the score predicted by assistant A_k. The fused assistant provides diversified supervision signals for distillation.

2.2.3 Assistant Selection

After generating fused assistants, we evaluate all original and fused assistants on a validation set. The assistant with the best retrieval performance is selected as the teaching assistant for the current iteration. This selection mechanism ensures that the student always learns from the most effective assistant.

At the end of each iteration, if the updated student outperforms the worst-performing assistant in the pool, we replace that assistant with the student. In this way, the assistant pool is dynamically updated, and the overall quality of assistants is progressively improved.

2.2.4 The Student Model Optimization

In each iteration, the student model is optimized by learning from the teacher with the guidance of the selected assistant. The final loss function integrates supervised learning, distillation from the teacher, and additional guidance from the assistant:

$$\mathcal{L}^{(t)} = \alpha \mathcal{L} * \mathrm{CE} + \beta \mathcal{L} * \mathrm{KD}^T + \gamma \mathcal{L}^A_{\mathrm{KD}}, \tag{7}$$

where $\mathcal{L} * \mathrm{KD}^T$ and $\mathcal{L} * \mathrm{KD}^A$ denote the distillation losses from the teacher and the selected assistant, respectively, and α, β, γ are hyper-parameters.

Through iterative training with progressively refined assistants and increasingly challenging data, the student model gradually narrows the performance gap with the teacher.

3 Experiments and Analysis

3.1 Experimental Settings

Datasets. We conduct experiments on MSMARCO passage ranking, TREC DL 2019, TREC DL 2020 and Natural Questions (NQ). MSMARCO is a large-scale benchmark dataset for passage retrieval. TREC DL 2019 and 2020 are evaluation datasets built upon MSMARCO. NQ is a widely used open-domain question answering dataset.

Evaluation Metrics. For MSMARCO, we report MRR@10. For TREC DL 2019 and 2020, we report NDCG@10. For NQ, we report Recall@20.

Implementation Details. We use a large dual-encoder model as the teacher and several DPR models with different architectures and sizes as assistants. The student model is initialized with a smaller pre-trained language model. All models are implemented with the same training framework. We train the student model for multiple iterations as described in Section 3. The hyper-parameters are selected based on the validation set.

3.2 Main Results

Table 1 presents the main results on MSMARCO, TREC DL 2019 and 2020. Our method consistently outperforms existing knowledge distillation baselines. The 66M student model achieves state-of-the-art performance among models with similar parameter sizes.

Table 2 reports the results on the NQ dataset. MTA4DPR significantly improves the retrieval performance compared with other distillation methods and remains competitive compared with larger models.

3.3 Ablation Study

We conduct ablation studies to evaluate the effectiveness of each component in MTA4DPR. The results are shown in Table 3. Removing the fusion strategy or the assistant selection module leads to noticeable performance degradation. The iterative distillation mechanism also contributes significantly to the final performance.

3.4 Analysis

3.4.1 Multi-iteration Retrieval Performance

Table 4 shows the retrieval performance of the student model across different iterations. We observe that the student performance improves steadily as the number of iterations increases, demonstrating the effectiveness of iterative distillation.

3.4.2 The impact of selection methods

Table 5 compares different assistant selection strategies. Selecting the best-performing assistant based on validation performance yields better results than random or fixed selection strategies.

3.4.3 The impact of the number of layers and the embedding sizes of student models

Table 6 presents the results of student models with different numbers of layers and embedding sizes. Our method consistently improves performance across various configurations.

3.4.4 The impact of the performance of assistant models

Table 7 analyzes how the quality of assistant models affects the final performance. Using stronger assistants generally leads to better student performance.

3.4.5 The composition of the best assistant

Figure 2 illustrates the composition of the best assistant selected during different iterations. We find that fused assistants are frequently selected, indicating that the fusion strategy effectively integrates complementary knowledge.

3.4.6 The complexity of the training process

Table 8 reports the training complexity of MTA4DPR compared with baseline methods. Although our method introduces additional computation for assistant fusion and selection, the overall training cost remains acceptable.

3.4.7 The computational costs of MTA4DPR

Table 9 presents the computational costs during inference. The student model maintains efficient inference speed while achieving competitive retrieval performance.

4 Conclusion

In this paper, we propose MTA4DPR, a novel multi-teaching-assistants based iterative knowledge distillation framework for dense passage retrieval. By introducing multiple assistants, a fusion strategy, and an iterative distillation mechanism, our method effectively enhances the student model's retrieval performance. Extensive experiments on MSMARCO, TREC DL 2019 and 2020, and Natural Questions demonstrate that MTA4DPR significantly outperforms existing knowledge distillation baselines and achieves competitive performance compared with larger models. The detailed analyses further verify the effectiveness of each component in our framework.

Limitations

Although MTA4DPR achieves strong performance improvements, it also introduces additional computational overhead due to the use of multiple assistants and the iterative training process. The fusion and selection procedures require extra evaluation steps, which increase the overall training time compared with single-teacher distillation methods. Moreover, the effectiveness of our framework depends on the diversity and quality of assistant models. If the assistants lack sufficient diversity or are of low quality, the performance gains may be limited. In future work, we will explore more efficient assistant construction and selection strategies to further reduce the computational cost while maintaining performance improvements.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments and suggestions. This work was supported by [ILLEGIBLE].

Ethics Statement

This work focuses on improving dense passage retrieval through knowledge distillation techniques. We do not introduce new datasets and conduct experiments on publicly available benchmarks. The proposed method does not involve human subjects or sensitive personal information. Nevertheless, as with other retrieval systems, potential biases present in the training data may affect model behavior. We encourage future research to further investigate bias mitigation and responsible deployment of retrieval models.

Licenses

The datasets used in this paper, including MSMARCO, TREC DL 2019, TREC DL 2020, and Natural Questions, are publicly available and subject to their respective licenses. We use these datasets strictly in accordance with their licensing terms.

References

References

[1] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of EMNLP*.

[2] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialu Liu, Paul N. Bennett, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *Proceedings of ICLR*.

[3] Sheng-Chieh Ren, Xiangyang Liu, Shujie Liu, and others. 2021. RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of NAACL*.

[4] Luyu Gao and Jamie Callan. 2021. Condenser: a Pre-training Architecture for Dense Retrieval. In *Proceedings of EMNLP*.

[5] [ILLEGIBLE]

[6] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. In *Proceedings of NIPS Workshop*.

[7] [ILLEGIBLE]

[8] [ILLEGIBLE]

[9] [ILLEGIBLE]

Algorithm 1 MTA4DPR Training Process

Initialize teacher model T, assistant pool $\mathcal{A} = A_1, A_2, \ldots, A_K$, and student model S
for iteration $t = 1$ to T **do**
  Construct dataset $\mathcal{D}^{(t)}$ using T and $\mathcal{A}$
  Generate fused assistants via the fusion module
  Evaluate all assistants on the validation set
  Select the best assistant A^*
  Train student S with knowledge distillation from T guided by A^*
  Evaluate updated student S
  if S outperforms the worst assistant in $\mathcal{A}$ **then**
    Replace the worst assistant with S
  end if
end for
return trained student model S

[10] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum Learning. In *Proceedings of ICML*.