Discriminative Language Model via Self-Teaching for Dense Retrieval

Lu Chen
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China
chenlu19z@ict.ac.cn

Ruqing Zhang
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China
zhangruqing@ict.ac.cn

Jiafeng Guo*
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China
guojiafeng@ict.ac.cn

Yixing Fan
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China
fanyixing@ict.ac.cn

Xueqi Cheng
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China
cxq@ict.ac.cn

ABSTRACT

Dense retrieval (DR) has shown promising results in many information retrieval (IR) related tasks, whose foundation is high-quality text representations for effective search. Taking the pre-trained language models (PLMs) as the text encoders has become a popular choice in DR. However, the learned representations based on these PLMs often lose the discriminative power, and thus hurt the recall performance, particularly as PLMs consider too much content of the input texts. Therefore, in this work, we propose to pre-train a discriminative language representation model, called DiscBERT, for DR. The key idea is that a good text representation should be able to automatically keep those discriminative features that could well distinguish different texts from each other in the semantic space. Specifically, inspired by knowledge distillation, we employ a simple yet effective training method, called self-teaching, to distill the model's knowledge constructed when training on the sampled representative tokens of a text sequence into the model's knowledge for the entire text sequence. By further fine-tuning on publicly available retrieval benchmark datasets, DiscBERT can outperform the state-of-the-art retrieval methods.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Retrieval models and ranking**; **Document representation**.

*Jiafeng Guo is the corresponding author.

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10.
https://doi.org/10.1145/3511808.3557582

KEYWORDS

Dense Document Retrieval, Discriminative Representation, Self-Teaching

ACM Reference Format:
Lu Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022. Discriminative Language Model via Self-Teaching for Dense Retrieval. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3511808.3557582

1 INTRODUCTION

Recently, dense retrieval (DR) has received much attention in many information retrieval (IR) related tasks [12, 16]. DR generally leverages a Dual-Encoder to learn the low-dimensional representations of queries and documents, and then conducts retrieval with simple similarity metrics. In essence, high-quality sequence representations are the foundation of DR for performing efficient search.

The pre-trained language models (PLMs) like BERT [5] have become the de-facto implementation of the text encoders in DR. Despite its wide usage, current PLMs have clear limitations that relate to their discriminative power of representations. Previous observations [15] showed that the PLMs are not effective at encoding the semantics of the entire input sequence, especially in DR where text sequences are mostly long. That is, introducing much information will probably confuse the PLMs and map almost all sequences into a small area, resulting in high similarity [24].

Therefore, we propose to pre-train a discriminative language representation model, named DiscBERT, for DR. DiscBERT is a variant of BERT in which the key idea is to guide the model to automatically focus on those informative words of an input sequence. Inspired by knowledge distillation (KD) [9], we first employ a teacher-student framework, where the teacher encodes those informative content of a sequence and the student encodes the whole sequence. Specifically, we introduce two representative word selection mechanisms to generate discriminative representations in the teacher.

Then, we devise a novel training method, called self-teaching, to train the student to achieve comparable discriminative representations to the teacher. Different from standard KD training, the student model in our method is at the same time also the teacher model, and thus only one model is optimized during training. This contributes to the savings of computational resources. Specifically, we minimize the KL divergence between the self-attention distributions and the Euclidean distance between the [CLS] representation of the teacher and student. The pre-trained model DiscBERT could then be fine-tuned on a variety of downstream retrieval tasks.

We pre-train DiscBERT on Wikipedia, and then fine-tune the dual-encoder model [26] with DiscBERT as the encoder on ad-hoc document retrieval tasks, including MS MARCO Document Ranking and TREC 2019 DL Track benchmark. Empirical experimental results demonstrate the effectiveness of DiscBERT.

2 RELATED WORK

First-stage Retrieval. For the first-stage retrieval, early studies mainly use the term-based methods including BM25 [18] and query likelihood [13]. Such exact lexical match retrievers are highly efficient and universal in practice, but suffer from the vocabulary mismatch problem. In recent years, with the development of representation learning techniques, many researchers have turned to the DR models to solve the semantic matching problem. DR is a representation-based method which encodes queries and documents as dense vectors separately by a dual-encoder architecture, and then calculates their relevance. Specifically, researchers have explored the PLMs, e.g., BERT, for the first-stage retrieval and showed they can achieve state-of-the-art performance [4, 26]. For example, RepBERT [26] employed a BERT-based dual-encoder and leveraged the inner product as relevance score.

Knowledge Distillation. KD [9] is to train a small student model by mimicking a larger teacher model's behavior for achieving comparable performance. Some works transfer knowledge through alignment of soft label distributions [9, 20] or intermediate representations [11],while some explore ensemble teacher models [6] or meta learning [29] to improve the performance of the teacher. Self-distillation is a special kind of KD, which transfers knowledge within one model. Some works transfer knowledge from deeper portion of the networks into the shallow ones [27, 28], or from the word embedding layer to the output layer [8]. Some works take the student models at previous training steps as teachers [23].

3 OUR APPROACH

In this section, we introduce the proposed discriminative pre-trained language model, named DiscBERT, for dense retrieval.

3.1 Model Overview

Formally, given an input sequence $X = \{x_1, x_2, ..., x_N\}$ with N tokens, DiscBERT aims to learn its powerful representation $\mathbf{H}$. Inspired by KD, we employ a teacher-student framework, including 1) the teacher, to obtain discriminative representations by encoding the representative tokens of a sequence, and 2) the student, to learn how to achieve comparable discriminative representations to the teacher when using the whole sequence as the input.

Then, we introduce a self-teaching training approach to train the student to mimic the behavior of the teacher, i.e., automatically

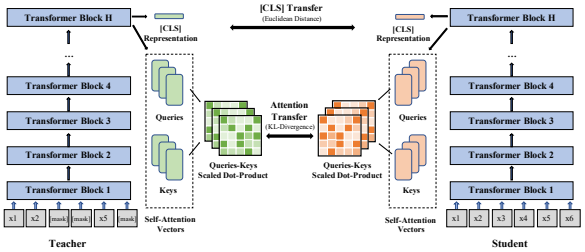


Figure 1: Overview of discriminative language model pre-training process. The input to the teacher is the representative tokens of a sequence, while the input to the student is a whole sequence. The student model in our self-teaching training is at the same time also the teacher model.

attending to the representative tokens of an input sequence. Here, the student itself is also the teacher.The overall architecture of DiscBERT is depicted in Figure 1, and we will detail it as follows.

3.2 Teacher-Student Framework

Here, we employ a teacher-student framework, including a teacher component and a student component.

Teacher. The teacher model aims to encode the discriminative features of an input sequence to guide the training of the student. The key idea is to identify those representative words which can adequately reveal the essential topic of the sequence. For this purpose, we introduce two types of representative word selection mechanisms, namely k-max pooling and term distribution sampling.

Specifically, both of them are based on the idf value [19], representing the informativity (i.e., representativeness) of each token, where higher idf value indicates higher informativity, and vice versa. The underlying idea is that if a token occurs in more documents of the collection, it has lower informativity, i.e., it is of little use in determining the discrimination of two documents. The two types of representative word selection mechanisms are as follows.

- **K-max Pooling** over an input sequence X returns the set of tokens with $k\%$ maximum idf values.
- **Term Distribution Sampling** aims to sample words from a sequence term distribution. Typically, a term may appear multiple times within the same sequence. We construct a vocabulary $\{w_m\}_{m=1}^{M}$ with M distinct terms for each sequence X, and then compute the term distribution $P(w_m|\theta_X)$,

$$P(w_m|\theta_X) = \frac{\exp(idf(w_m))}{\sum_{k=1}^{M} \exp(idf(w_m))}, \tag{1}$$

where the softmax function ensures the term distribution is valid. We sample one term w_m each time according to $P(w_m|\theta_X)$ and pick all the tokens identical to w_m over different positions. The process is repeated iteratively until $k\%$ words are selected.

Afterwards, we keep the selected representative tokens and replace all the other tokens by the special mask token [MASK] in the input sequence. Specifically, the input of the teacher network is given by $X_t = \{x_1, [MASK], x_3, ..., [MASK], x_N\}$. Note we employ the attention mask over all the [MASK] tokens to omit their effects. Each token in X_t is represented by summing its distributed, segmentation, and positional embeddings.

Student. The student component takes the original sequence X as the input and we expect it can obtain similar discriminative representation to the teacher.

3.3 Knowledge Transfer with Self-Teaching

To minimize the knowledge difference between the representation obtained from the sampled representative tokens and that obtained from the corresponding whole sequence, we introduce two types of knowledge transfer mechanisms, i.e., self-attention distribution transfer and [CLS] representation transfer.

Self-Attention Distribution Transfer. Self-attention mechanism is crucial for BERT, and the top layers of BERT encode semantic features [10, 22]. We propose to utilize the self-attention distributions of teacher's last Transformer layer, to help the training of the student. Specifically, we minimize the KL divergence between the self-attention distributions of the teacher and student:

$$\mathcal{L}_{kl} = \frac{1}{A_h N} \sum_{a=1}^{A_h} \sum_{n=1}^{N} D_{KL}(\mathbf{A}^T_{L,a,n} || \mathbf{A}^S_{L,a,n}) + D_{KL}(\mathbf{A}^S_{L,a,n} || \mathbf{A}^T_{L,a,n}), \quad (2)$$

where A_h denotes the number of self-attention heads, and $\mathbf{A}^T_{L,a,n}$ and $\mathbf{A}^S_{L,a,n}$ are the attention distributions of the last Transformer layer (i.e., L-th) for the teacher and student, respectively. Note that we consider both the forward and reverse KL divergence.

[CLS] Representation Transfer. The [CLS] representation of teacher represents the discriminative understanding of the input sequence X. Beyond the token-level mimicry given by self-attention distribution transfer, we introduce to transfer the [CLS] representation at the sequence-level. Specifically, we propose to minimize the Euclidean distance (i.e., L2) between the final [CLS] representation of the teacher and student, i.e.,

$$\mathcal{L}_{euclidean} = L2(\mathbf{H}^T_{[CLS]}, \mathbf{H}^S_{[CLS]}), \quad (3)$$

where $\mathbf{H}^T_{[CLS]}$ and $\mathbf{H}^S_{[CLS]}$ denotes the final hidden state of [CLS] achieved by the teacher and student model respectively.

Based on the above knowledge transfer mechanisms, we re-train the advanced contextual language model BERT via minimizing the combination of the KL divergence and the Euclidean distance, i.e.,

$$\mathcal{L} = \mathcal{L}_{kl} + \mathcal{L}_{euclidean}. \quad (4)$$

The pre-trained model is named as DiscBERT for short. The sequence representation $\mathbf{H}$ can be computed as the average of the token's representation or the [CLS] representation. DiscBERT could then be fine-tuned on a variety of downstream dense retrieval tasks.

3.4 Discussion

Intuitively, PLMs have higher discrimination on the representative tokens than on the whole sequence. Hence, to improve the discrimination on the whole sequence, we can inject the PLMs knowledge learned from the corresponding sampled representative tokens.

The self-teaching training approach differs from existing self-distillation methods in two aspects: (1) The teacher and student share exactly the same architecture and parameters. Thus, the model can be bootstrapped, and further improves the discriminative ability.(2) We transfer knowledge between the same parts within the model.The discriminative knowledge is achieved from the carefully devised input,and it can be seamlessly adapted to different PLMs.

4 EXPERIMENTAL SETTINGS

4.1 Datasets

- **Pre-training Corpus**. We leverage a large public document collection, i.e., English Wikipedia, for pre-training, which contains 3.2 million well-formed Wiki-articles.
- **Downstream Tasks**. We fine-tune DiscBERT on two ad-hoc document retrieval tasks: (1) **MS MARCO Document Ranking (MS MARCO)** [2] is a large-scale benchmark dataset with about 0.37 million training queries and 5.2 thousand dev queries; and (2) **TREC 2019 Deep Learning Track (TREC DL)** [17] replaces the test set on MS MARCO with a novel set produced by TREC with more comprehensive labels.

4.2 Evaluation Methodology

Following [16, 25], we report mean reciprocal rank (MRR) at position 10 and 100 on the dev set of MS MARCO, normalized discounted cumulative gain (NDCG) at position 10 and recall at position 100 on the test set of TREC DL.

4.3 Baselines

- **Term-based Retrieval Model**: **BM25** [18] is a classical probabilistic retrieval model.
- **Sparse Retrieval Model**: **HDCT** [3] hierarchically aggregates passage-level term weights into document-level representations. **SparTerm** [1] learns sparse text representations with a BERT-based importance predictor and a gating controller.
- **Dense Retrieval Model**: **BERT** refers to a BERT-based dual-encoder.**BERT**$_{IDF}$ takes only the representative words (via term distribution sampling) of a sequence as the input of BERT. **BERT**$_{TS}$ has the same teacher-student architecture as DiscBERT, but the teacher and student components do not share the parameters. We apply the k-max pooling and term distribution sampling mechanisms for BERT$_{TS}$, denoted as BERT$_{TS-max}$ and BERT$_{TS-dis}$.

4.4 Implementation Details

- **Model Architecture**. We use the Transformer encoder architecture the same as BERT$_{base}$ ($A_h = 12, L = 12$) [5]. Specifically, we use Huggingface Transformers library[1] for implementations[2].
- **Pre-training Process**. We use batch size 512, learning rate 5e-5, and adopt the parameters of BERT$_{base}$ released by Google[3] for initialization. For representative word selection, we set $k = 80$. DiscBERT with k-max pooling and term distribution sampling are denoted as DiscBERT$_{max}$ and DiscBERT$_{dis}$ respectively.
- **Fine-tuning Process**. We employ the dual-encoder architecture with DiscBERT as the encoder. Here, the sequence representation is computed by averaging each token's representation following [7, 26]. We set the batch size as 10, the gradient accumulation step as 3, and the learning rate as 5e-5.

5 RESULTS AND ANALYSIS

5.1 Baseline Comparison

Table 1 shows the evaluation results on the the MS MARCO and TREC DL. We can observe that: (1) Sparse and dense retrieval generally outperform term-based retrieval, for the reason that they alleviate the vocabulary mismatch problem by capturing the deep semantic relationship between queries and documents. However, they underperform term-based models in terms of R@100 on the TREC DL. One possible reason is that neural models may be biased by the false negatives that are unlabeled yet relevant [25].

[1] https://github.com/huggingface/transformers
[2] The code is available at https://github.com/ict-bigdatalab/DiscBERT
[3] https://github.com/google-research/bert

Table 1: Comparisons between DiscBERT and the baselines on the two document ranking datasets. Two-tailed t-tests demonstrate the improvements of DiscBERT$_{dis}$ over the baseline are statistically significant ($*$ indicates p-value < 0.05).

Category	Method	MS MARCO (dev)		TREC DL (test)	
		MRR@10	MRR@100	NDCG@10	R@100
Term-based Retrieval	BM25 [16, 25]	0.249	0.279	0.519	**0.395**
Sparse Retrieval	HDCT [3]	0.300	-	-	-
	SparTerm [1]	0.290	-	-	-
Dense Retrieval	BERT [16, 25]	0.288	-	0.510	0.223
	BERT$_{IDF}$	0.289	0.299	0.517	0.249
	BERT$_{TS-max}$	0.291	0.301	0.514	0.225
	BERT$_{TS-dis}$	0.307	0.318	0.524	0.232
DiscBERT	DiscBERT$_{max}$	0.303	0.313	0.503	0.237
	DiscBERT$_{dis}$	**0.318***	**0.328***	**0.538***	0.250

(2) *BERT$_{IDF}$* outperforms BERT, showing that only encoding the representative words does help the retrieval process. (3) *BERT$_{TS}$* performs better than BERT$_{IDF}$, showing the effectiveness of the teacher-student framework. The reason may be that such framework provides a mild way to highlight the representative words instead of absolutely ignoring those common words, which are useful in composing a document.

When we look at our method, we can find that: (1) *DiscBERT* achieves the best performance. The improvements over BERT indicate that the self-teaching training method indeed bootstraps the teacher's discriminative power as discussed in Section 3.4. (2) Among the two of our models, DiscBERT$_{dis}$ performs better than DiscBERT$_{max}$, indicating that flexibly learning the representative words can better achieve discriminative knowledge from fine-grained distinction of each word. (3) In the future work, we would like to apply the proposed teacher-student framework and self-teaching method to more complex PLMs such as TCT-ColBERTv2 [14] (not investigated here), which may well lead to even stronger and more discriminative dense retrievers.

Table 2: Performance comparison of DiscBERT$_{dis}$ for different numbers (percentages) of sampled representative tokens.

Sampled token percentage (%)	90	80	60
MRR@10 on MS MARCO	0.314	**0.318**	0.304
NDCG@10 on TREC DL	0.522	0.538	**0.541**

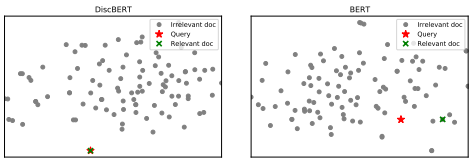
5.2 Analysis on the Sampled Token Percentage

We analyze the effect of the percentage of sampled representative tokens k% for retrieval. The results are shown in Table 2. (1) In general, the retrieval performance gradually increases when the number of sampled representative tokens decreases. Too much content increases the risk of introducing more noisy information which is harmful for retrieval. In practice, this hyper-parameter depends on the quality of the information repository and the ability of the search system. (2) DiscBERT$_{dis}$ with different numbers of sampled representative tokens can still outperform BERT, again indicating the effectiveness of the self-teaching training method.

5.3 Visual Analysis

To better understand what can be learned by DiscBERT, we conduct visual analyses in the MS MARCO dev set. Table 3 visualizes the [CLS]-token attention weights for a document. We can find that: (1) Before fine-tuning, DiscBERT can emphasize representative words well due to the self-teaching method. BERT pays similar attention to most of the words and favors common words.This is reasonable for BERT since these common words appear frequently.(2) After fine-tuning, BERT still favors common words, though the attention weights of some representative words are improved. DiscBERT can remain focusing on the representative words, demonstrating that the discriminative power can be kept in the fine-tuning process. Figure 2 visualizes the query and document representations using t-SNE [21]. We sample a query and achieve its top-100 candidate documents given by DiscBERT and BERT after fine-tuning. We can find that DiscBERT maps the relevant document in the semantic space closer to the query while far away from others.

Table 3: The visualization of [CLS]-token attention weights of BERT and DiscBERT. Darker color indicates higher attention.



(a) Encoded by DiscBERT (b) Encoded by BERT

Figure 2: The t-SNE plot of a query (QID 320792) from MS MARCO dev set and its candidate document representations.

6 CONCLUSION

In this paper, we proposed DiscBERT to enhance the discriminative power of the learned representations for DR. Specifically, we introduced a teacher-student network in combination with self-teaching for learning DiscBERT that are embedding-invariant w.r.t the representative tokens of a sequence and a whole sequence. Experimental results on several document retrieval datasets show the effectiveness of DiscBERT. For future work, we would like to investigate better word selection mechanisms and training process to improve the discriminative power of language models.

ACKNOWLEDGMENTS

This work was funded by the National Natural Science Foundation of China (NSFC) under Grants No. 62006218 and 61902381, the Youth Innovation Promotion Association CAS under Grants No. 20144310, and 2021100, the Young Elite Scientist Sponsorship Program by CAST under Grants No. YESS20200121, and the Lenovo-CAS Joint Lab Youth Scientist Project.

REFERENCES

[1] Yang Bai, Xiaoguang Li, Gang Wang, Chaoliang Zhang, Lifeng Shang, Jun Xu, Zhaowei Wang, Fangshan Wang, and Qun Liu. 2020. SparTerm: Learning Term-based Sparse Representation for Fast Text Retrieval. *arXiv preprint arXiv:2010.00768* (2020).

[2] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268* (2016).

[3] Zhuyun Dai and Jamie Callan. 2020. Context-aware document term weighting for ad-hoc search. In *WWW*.

[4] Zhuyun Dai and Jamie Callan. 2020. Context-aware term weighting for first stage passage retrieval. In *SIGIR*.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

[6] Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. 2017. Efficient Knowledge Distillation from an Ensemble of Teachers.. In *Interspeech*.

[7] Luyu Gao, Zhuyun Dai, Tongfei Chen, Zhen Fan, Benjamin Van Durme, and Jamie Callan. 2021. Complement lexical retrieval model with semantic residual embeddings. In *ECIR*.

[8] Sangchul Hahn and Heeyoul Choi. 2019. Self-Knowledge Distillation in Natural Language Processing. In *RANLP 2019*.

[9] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[10] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language?. In *ACL*.

[11] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for Natural Language Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

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

[13] Victor Lavrenko and W Bruce Croft. 2017. Relevance-based language models. In *SIGIR Forum*.

[14] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In *RepL4NLP*.

[15] Shuqi Lu, Di He, Chenyan Xiong, Guolin Ke, Waleed Malik, Zhicheng Dou, Paul Bennett, Tie-Yan Liu, and Arnold Overwijk. 2021. Less is More: Pretrain a Strong Siamese Encoder for Dense Text Retrieval Using a Weak Decoder. In *EMNLP*.

[16] Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2020. Sparse, dense, and attentional representations for text retrieval. *arXiv preprint arXiv:2005.00181* (2020).

[17] Craswell Nick, Mitra Bhaskar, Yilmaz Emine, Campos Daniel, and Ellen Voorhees M. 2021. Overview of the TREC 2019 deep learning track. (2021).

[18] Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR*.

[19] Gerard Salton and Michael J McGill. 1983. *Introduction to modern information retrieval*. mcgraw-hill.

[20] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[21] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *JMLR* (2008).

[22] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv preprint arXiv:2002.10957* (2020).

[23] Yige Xu, Xipeng Qiu, Ligao Zhou, and Xuanjing Huang. 2020. Improving bert fine-tuning via self-ensemble and self-distillation. *arXiv preprint arXiv:2002.10345* (2020).

[24] Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer. *arXiv preprint arXiv:2105.11741* (2021).

[25] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing Dense Retrieval Model Training with Hard Negatives. *arXiv preprint arXiv:2104.08051* (2021).

[26] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. RepBERT: Contextualized text embeddings for first-stage retrieval. *arXiv preprint arXiv:2006.15498* (2020).

[27] Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. 2021. Self-distillation: Towards efficient and compact neural networks. (2021).

[28] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[29] Wangchunshu Zhou, Canwen Xu, and Julian McAuley. 2022. BERT Learns to Teach: Knowledge Distillation with Meta Learning. In *ACL*.