A Contrastive Pre-training Approach to Learn Discriminative Autoencoder for Dense Retrieval

Xinyu Ma
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China
maxinyu17g@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 information retrieval. In essence, DR requires high-quality text representations to support effective search in the representation space. Recent studies have shown that pre-trained autoencoder-based language models with a weak decoder can provide high-quality text representations, boosting the effectiveness and few-shot ability of DR models. However, even a weak autoregressive decoder has the bypass effect on the encoder. More importantly, the discriminative ability of learned representations may be limited since each token is treated equally important in decoding the input texts. To address the above problems, in this paper, we propose a contrastive pre-training approach to learn a discriminative autoencoder with a lightweight multilayer perception (MLP) decoder. The basic idea is to generate word distributions of input text in a non-autoregressive fashion and pull the word distributions of two masked versions of one text close while pushing away from others. We theoretically show that our contrastive strategy can suppress the common words and highlight the representative words in decoding, leading to discriminative representations. Empirical results show that our method can significantly outperform the state-of-the-art autoencoder-based language models and other pre-trained models for dense retrieval.

CCS CONCEPTS

• **Information systems → Retrieval models and ranking**.

KEYWORDS

Dense Retrieval, Discriminative Autoencoder, Contrastive Pre-training

ACM Reference Format:
Xinyu Ma, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022. A Contrastive Pre-training Approach to Learn Discriminative Autoencoder 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/10.1145/3511808.3557527

1 INTRODUCTION

Recently, dense retrieval (DR) has achieved great success on many information retrieval (IR) related tasks, such as web search [22, 23], open-domain Question Answering (QA) [10, 17] and fact verification [4]. DR models generally employ pre-trained language models as text encoder to obtain dense representations for queries and documents. Then, retrieval with simple similarity metrics can be conducted effectively in the representation space. Effective search is based on high-quality text representation learning [13].

Despite the effectiveness of BERT-like language models [6] on learning word representations, they are not good at producing text sequence representations [11, 18, 20]. Recent studies have demonstrated that autoencoder-based language models can significantly advance the effectiveness and few-shot ability of DR models [13]. The basic idea is to train a weak autoregressive decoder that reconstructs the input text only from the encoder's encodings. In this way, the encoder creates a bottleneck to provide high-quality text sequence representations. However, even a weak autoregressive decoder has the bypass effect in which the decoder may ignore the representation and predict the next token only based on previous tokens. More importantly, the decoder treats each token equally important but common words like in, the, and of, are the majority part of the text. Therefore, the discriminative ability of dense representations may be limited since the representation will focus more on the common words and thus is not differential with other representations.

[*] Jiafeng Guo is the corresponding author.

CIKM '22, October 17–21, 2022, Atlanta, GA, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557527

To address the above problems, in this paper, we propose a contrastive pre-training approach to learn a discriminative autoencoder with a lightweight multi-layer perception (MLP) decoder. Specifically, rather than reconstructing texts in an autoregressive fashion, the MLP decoder generates word distributions of input texts in a non-autoregressive fashion to avoid the bypass effect. We then introduce a novel contrastive learning method to pull the word distributions of two masked versions of one text close while pushing away from others. We theoretically show that our contrastive strategy can suppress the common words and highlight the representative words when decoding, leading to discriminative representations. Empirical results verified the effectiveness of our proposed discriminative autoencoder over the state-of-the-art autoencoder-based language models and other pre-trained models.

2 RELATED WORK

In this section, we briefly review the recent studies on designing pre-training methods tailored for dense retrieval. Early practice in this direction is [3], which proposed three pre-training tasks to resemble the downstream passage retrieval in open-domain QA. Specifically, Inverse Cloze Task (ICT) is a commonly-adopted task, where the basic idea is to predict the context for a randomly sampled sentence from the Wikipedia page. The most related work with ours is SEED [13], which pre-trains an autoencoder with a 3-layer weak Transformer decoder while restrict its attention spans. Another line of research is to design new model architectures [7] for dense retrieval. Researchers have also investigated to leverage contrastive learning method to learn sequence representations [8, 14, 18]. But these methods are not suitable for learning high-quality document representations as we shown in Section 4.2.

3 OUR METHOD

In this section, we describe our Contrastive Pre-training a Discriminative AutoEncoders (CPDAE) for dense retrieval.

3.1 Model Architecture

Basically, the model architecture of our CPDAE is composed of a Transformer encoder and a MLP decoder.

3.1.1 Encoder. The encoder aims to encode the input text into low-dimensional dense representations. We use Transformer [21] as the encoder. Given an input text $d_i = ([CLS], w_1, ..., w_n)$, a special token [CLS] is added to the front of d_i to represent the whole text. For each d_i, we follow the masking strategy in BERT [6] to randomly mask its several tokens twice, to obtain two masked versions $d_i = \{d_i^{'}, d_i^{''}\}$. We take the [CLS] representation of the last Transformer layer as the whole text representation,

$$\mathbf{h}_{[CLS]} = Encoder(d_i),\ \mathbf{h} \in \mathbb{R}^H, \tag{1}$$

where H is the hidden size.

3.1.2 Decoder. The MLP decoder is to recover the input text solely from the text sequence representations. Specifically, the decoder includes two layers of feed-forward neural network (FFN) with a non-linear activation function Gelu [9] and a LayerNorm function [1]. Then the MLP decoder maps the text representation to word distributions,

$$\mathbf{z} = Decoder(\mathbf{h}_{[CLS]}),\ \mathbf{z} \in \mathbb{R}^{|V|}, \tag{2}$$

where V is the vocabulary.

3.2 Contrastive Pre-training

Our contrastive pre-training includes three pre-training objectives: reconstruction loss, contrastive loss and masked language modeling (MLM) loss.

3.2.1 Reconstruction Loss. Our non-autoregressive reconstruction is to predict which words in the vocabulary appear in the input text by generating a word distribution. Specifically, given the prediction vector $\mathbf{z}$ in Eq. (2), we apply the *Sigmoid* function for each value z^j in $\mathbf{z}$ separately to obtain a valid probability, i.e.,

$$\hat{z}^j = Sigmoid(z^j), j = 1, ..., |V|, \tag{3}$$

where $\hat{z}^j$ ranges from 0 to 1 and indicates the probability of j-th word in the vocabulary V appearing in the input text. The reconstruction loss is formulated as a multi-label classification problem and computed with the cross-entropy function,

$$\mathcal{L}_{REC} = -\sum_{j=1}^{|V|} (y^j \log \hat{z}^j + (1 - y^j) \log \hat{z}^j), y^j \in [0, 1]. \tag{4}$$

y^j=1 denotes the j-th word appearing in the input and vice versa.

3.2.2 Contrastive Loss. The contrastive loss is applied on word distributions $\tilde{\mathbf{z}}$ which is normalized from $\hat{z}$ in Eq. (3). Two word distributions of masked versions of one text are pulled close while pushing away from other word distributions. We use Jensen–Shannon divergence function (JS) [12] to compute the similarity between word distributions.

Given a mini-batch, the contrastive loss over $2m$ masked sequences is defined as follows,

$$\mathcal{L}_{CL} = -\sum_{i=1}^{m} \log \frac{exp(-JS(\tilde{\mathbf{z}}_i^{'}, \tilde{\mathbf{z}}_i^{''}))}{\sum_{k=1}^{2m} \mathbb{1}_{[k \neq i]} exp(-JS(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_k))}, \tag{5}$$

where $(\tilde{\mathbf{z}}_i^{'}, \tilde{\mathbf{z}}_i^{''}))$ are two masked version of one text d_i. We also propose a variant which directly contrasts the dense representations $\mathbf{h}$, and this variant is denoted as CPDAE$_R$.

3.2.3 MLM Loss. Similar to existing works [13, 16], we also adopt the MLM [6] to build good word representations. We omit its details here and refer the reader to the original BERT paper [6].

The final loss is the total sum of MLM loss, reconstruction loss and contrastive loss, which is formulated as,

$$\mathcal{L}_{total} = \mathcal{L}_{REC} + \mathcal{L}_{MLM} + \lambda\mathcal{L}_{CL}, \tag{6}$$

where λ is a hyper-parameter.

3.3 Theoretical Analysis

We mathematically show why our contrastive pre-training based on word distributions can learn a discriminative autoencoder.

A natural language text is composed of a large portion of common words and a small portion of representative/informative words. Suppose $S \subset V$ denote the common words in the V and $\complement VS$ denote the rest words. According to Equation (5), $\mathcal{L}_{CL}$ aims to minimize $-JS(P, Q)$ between word distributions from different input texts, and maximize $-JS(P, Q)$ between word distributions from the same input texts. The $-JS(P, Q)$ can be rewritten to the following:

$$-JS(P,Q) = -\sum_{x\in|V|} p(x)log(p(x)) - \sum_{x\in|V|} q(x)log(q(x)) + \sum_{x\in|V|} (p(x)+q(x))log(p(x)+q(x)). \tag{7}$$

where we ignore $log2$ as it is a constant.

We will discuss two situations, i.e., word $x \in S$ and word $x \in \complement VS$. For the word $x \in S$, $p(x)$ is equal to $q(x)$ as the common words appear in almost every document with the same high probabilities a. Thus, Equation (7) can be reduced to:

$$\begin{aligned}-JS(P,Q)_{x\in S} &= -\sum_{x\in S} 2p(x)log(p(x)) + \sum_{x\in S} 2p(x)log(2p(x)) \\ &= \sum_{x\in S} 2p(x)log2.\end{aligned} \tag{8}$$

Therefore, given a mini-batch of $2m$ examples, $p(x)$ will be lowered as there are $2m-2$ word distributions needed to be minimized and only 1 positive word distributions needed to be maximized. So we conclude that (1) **word distribution based contrastive pre-training will suppress the probability of common words when decoding**.

For the word $x \in \complement VS$, the contrastive loss needs maximize Eq. (8) between two word distributions from the same text. For word distributions from other texts, $q(x)$ is close to 0 since representative words only occur in $p(x)$. The contrastive loss needs minimize Equation (7) which can be reduced to,

$$-JS(P,Q) = 0. \tag{9}$$

In summary, (2) **word distribution based contrastive pre-training will highlight the probability of representative/informative words when decoding**.

4 EXPERIMENTS

In this section, we conduct experiments to demonstrate the effectiveness of our proposed model.

4.1 Experimental Settings

Here, we introduce the pre-training corpus, downstream tasks, baseline methods, and implementation details.

4.1.1 Pre-training Corpus and Downstream Tasks. We use the English Wikipedia as our pre-training corpus following previous works [3, 6, 13, 15, 16]. We conduct experiments on several public dense retrieval benchmarks, including MS MARCO Passage Ranking (MARCO Dev Passage) [2], MS MARCO Document Ranking (MARCO Dev Document) [2], TREC 2019 Passage Ranking (TREC2019 Passage) [5] and TREC 2019 Document Ranking (TREC2019 Document) [5].

4.1.2 Baselines. We adopt the traditional sparse retrieval models and pre-trained models as baselines. For traditional spare retrieval models, we choose the strong **BM25** method [19]. We also list several representative results according to the TREC overview paper [5]. For the pre-trained models, besides **BERT** [6], the main baseline is the state-of-the-art autoencoder-based language models **SEED** [13]. We also consider two contrastive learning methods, including **ICT** [3], and **SimCSE** [8]. We also consider the state-of-the-art pre-trained models on the re-ranking, i.e., **PROP** [15], but we pre-train it with a bi-encoder for a fair comparison.

4.1.3 Implementation Details.

- **Pre-training details**. We use BERT to initialize our encoder. The output hidden size of the MLP decoder is set to 30522 which is the size of BERT's vocabulary. We use a learning rate of 5e-5 and Adam optimizer with a linear warm-up technique over the first 10% steps. We pre-train on Wikipedia for 3 epochs with the batch size m as 64. λ is set to 0.1 in the Eq. (6).
- **Fine-tuning details**. The decoder is only used in pre-training and is dropped during fine-tuning. Following previous works [7, 13, 22, 23], we employ a bi-encoder architecture based on the encoder of CPDAE and use a pairwise loss for fine-tuning. We use a learning rate of 5e-6, a batch size of 64, and pair each positive example with 7 negative examples. For two passage ranking datasets (i.e., MACRO Dev Passage and TREC2019 Passage), we train the model with static hard negatives using the BM25 warm-up model following [23]. For two document ranking datasets (i.e., MARCO Dev Doc and TREC2019 Doc), we use the model fine-tuned on the passage ranking task as the starting point following [13, 22, 23]. We then iteratively mine the static hard negatives using the current model twice and fine-tune the model for 1 epoch in each iteration.

4.2 Baseline Comparison

The performance comparisons between CPDAE and the baselines are shown in Table 1. We have the following observations: (1) Dense retrieval models generally outperform the traditional sparse retrieval models by a large margin on most of the datasets. This is mainly because dense retrieval models could well capture the semantics meanings of queries and documents and can over the vocabulary mismatch problem. (2) SimCSE, ICT and PROP show slight improvements over BERT, indicating that these pre-training methods may not be optimal for dense retrieval. ICT only pulls a random sentence close to its context in the representation space, while the random sentence may be semantic similar to other texts and thus not be distinguishable from different texts. (4) SEED performs the best among all the baseline, indicating the autoencoder-based language models can produce high-quality dense representation via reconstruction.

We find that CPDAE can generally outperform baseline methods significantly, including general pre-trained language model BERT, other contrastive learning methods, and autoencoder-based language models. The better results demonstrate the effectiveness of our contrastive loss to encode discriminative text sequence representations. CPDAE performs better than CPDAE$_R$ even though contrasting dense representations in the representation space is more straightforward. But the performance of contrasting dense representations heavily depends on the data augmentation while we only use a weak randomly masking.

4.3 Ablation Analysis

We conduct an ablation analysis to investigate the effect of the proposed contrastive loss (CL) in our CPDAE. We also compare with a weighted reconstruction loss (IDF-REC) which weights each

Table 1: Comparisons between CPDAE and the baselines. Two-tailed t-tests demonstrate the improvements of CPDAE to the baselines are statistically significant ($p \leq 0.05$). $*$, †, ‡ indicates significant improvements over BERT, ICT, and SEED, respectively. Results not available or not applicable are marked as '-'.

Model	MARCO Dev Passage		TREC2019 Passage		MARCO Dev Doc		TREC2019 Doc	
	MRR@10	Recall@1000	NDCG@10	Recall@1000	MRR@100	Recall@100	NDCG@10	Recall@100
BM25	0.187	0.857	0.501	0.745	0.277	0.808	0.519	**0.395**
Best TREC Trad[5]	-	-	0.554	-	-	-	0.549	-
BERT	0.335	0.957	0.661	0.769	0.389	0.877	0.594	0.301
SimCSE	0.335	0.955	0.662	0.766	0.391	0.879	0.598	0.302
ICT	0.339	0.955	0.670	0.775	0.396	0.882	0.605	0.303
PROP	0.337	0.951	0.673	0.771	0.394	0.884	0.596	0.298
SEED	0.342^{*}	0.963	0.679^{*}	$0.782^{*\dagger}$	0.396	0.902^{*}	0.605^{*}	0.307
CPDAE_R	$0.350^{*\dagger}$	$0.965^{*\dagger}$	$0.686^{*\dagger}$	$0.789^{*\dagger}$	0.402^{*}	**0.909***†	0.609^{*}	0.311^{*}
CPDAE	**0.355***†‡	**0.968***†	**0.696***†‡	**0.799***†‡	**0.408***†‡	$0.907^{*\dagger}$	**0.615***†‡	$0.315^{*\dagger}$

Table 2: Ablation studies of the contrastive loss (CL) in CPDAE. IDF-REC: using IDF to weight reconstruction loss. Best results are marked bold.

	MARCO Dev Passage		MARCO Dev Doc	
	MRR@10	R@1000	MRR@100	R@100
SEED	0.342	0.963	0.396	0.902
w/o CL	0.344	0.963	0.397	0.905
IDF-REC	0.347	0.962	0.399	0.906
w/ CL	**0.355**	**0.968**	**0.408**	**0.907**

Table 3: Fine-tuning with limited supervised data. Performance is measured by Recall@1000 and Recall@100 for MARCO Dev Passage and MARCO Dev Doc, respectively.

	MARCO Dev Passage			MARCO Dev Doc		
	0.1k	1k	10k	0.1k	1k	10k
BERT	0.636	0.803	0.891	0.512	0.692	0.784
SEED	0.659	0.827	0.914	0.523	0.717	0.835
CPDAE	**0.708**	**0.855**	**0.923**	**0.573**	**0.821**	**0.868**

token loss with IDF value in Eq. (4). As shown in Table 2, we can find that: (1) By removing the CL, the performance of CPDAE (w/o CL) is slightly better than SEED, indicating the effectiveness of the novel reconstruction loss with a non-autoregressive decoder by avoiding the bypass effect. (2) IDF-REC and CPDAE (w/o CL) have a similar performance while both perform significantly worse than CPDAE (w CL), again demonstrating the proposed contrastive loss could help learn discriminative representations.

4.4 Low-Resource Setting and Visual Analysis

To further illustrate the effectiveness of CPDAE, we simulate a low-resource setting on the MARCO Dev Passage and MARCO Dev Doc respectively. We randomly sample a limit number of queries (i.e., 0.1k, 1k, 10k) from the original training set to fine-tune the pre-trained models. Each experimental result is reported as the average of three runs with different sampled queries. As shown in Table 3, we can see that CPDAE outperforms BERT and SEED on all the datasets using the same number of limited supervised data. This result demonstrates that CPDAE can provide more discriminative text representations than BERT and SEED.

Table 4: Visualization of the word distributions generated by the MLP decoder in CPDAE. Darker color indicates higher probability.

Probability: 0 , 0.1 , 0.2 , 0.3 , 0.4 , >0.5

Nanotechnology raises many of the same issues as any new technology , including concerns about the toxicity and environmental impact of nanomaterials , and their potential effects on global economics , as well as speculation ...

To illustrate how CPDAE improves retrieval performance, we visualize the normalized word distributions generated by the MLP decoder in Table 4. We randomly sample a short piece of text from Wikipedia and sum up the normalized output probabilities of all the subwords of a whole word. As shown in Table 4, we can see that CPDAE can suppress the common words and highlight the informative words as shown in the theoretical analysis 3.3.

5 CONCLUSION

In this paper, we present a contrastive pre-training method to learn a discriminative autoencoder for dense retrieval. We propose to employ a non-autoregressive MLP decoder to avoid the bypass effect and apply contrastive learning to the word distributions produced by the decoder. We theoretically show that our contrastive strategy can suppress the common words and highlight the representative words, leading to discriminative representations. Experiments at four public dense retrieval benchmarks show that our method could achieve significant improvements over the baselines. For future work, we would like to investigate the data augmentation techniques and apply our method to other IR scenarios.

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] Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. *ArXiv* abs/1607.06450 (2016).

[2] Daniel Fernando Campos, Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, Li Deng, and Bhaskar Mitra. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *ArXiv* abs/1611.09268 (2016).

[3] Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training Tasks for Embedding-based Large-scale Retrieval. *ArXiv* abs/2002.03932 (2020).

[4] Jiangui Chen, Ruqing Zhang, J. Guo, Yixing Fan, and Xueqi Cheng. 2022. GERE: Generative Evidence Retrieval for Fact Verification. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2022).

[5] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Fernando Campos, and Ellen M. Voorhees. 2020. Overview of the TREC 2020 Deep Learning Track. *ArXiv* abs/2102.07662 (2020).

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv* abs/1810.04805 (2019).

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

[8] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *EMNLP*.

[9] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian Error Linear Units (GELUs). *arXiv: Learning* (2016).

[10] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. *ArXiv* abs/2004.04906 (2020).

[11] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the Sentence Embeddings from Pre-trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.733

[12] Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* 37 (1991), 145–151.

[13] 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 *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2780–2791.

[14] Xinyu Ma, J. Guo, Ruqing Zhang, Yixing Fan, and Xueqi Cheng. 2022. Pre-train a Discriminative Text Encoder for Dense Retrieval via Contrastive Span Prediction. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2022).

[15] Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, Xiang Ji, and Xueqi Cheng. 2021. PROP: Pre-training with Representative Words Prediction for Ad-hoc Retrieval. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining* (2021).

[16] Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, Yingyan Li, and Xueqi Cheng. 2021. B-PROP: Bootstrapped Pre-training with Representative Words Prediction for Ad-hoc Retrieval. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2021).

[17] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. In *NAACL*.

[18] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *ArXiv* abs/1908.10084 (2019).

[19] Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3 (2009), 333–389.

[20] Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316* (2021).

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

[22] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. *ArXiv* abs/2007.00808 (2021).

[23] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing Dense Retrieval Model Training with Hard Negatives. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2021).