Adaptive Hyper-parameter Learning for Deep Semantic Retrieval

Mingming Li[1] and **Chunyuan Yuan**[1] and **Huimu Wang**[1]
Peng Wang[2] and **Binbin Wang** [1] and **Jingwei Zhuo**[1,*] and **Lin Liu** [1] **and Sulong Xu** [1]
[1] JD.com, Beijing, China
[2]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
{limingming65, yuanchunyuan1, wanghuimu1@jd.com wangpeng2022@iie.ac.cn
{wangbinbin77, zhuojingwei1, liulin1, xusulong}@jd.com

Abstract

Deep semantic retrieval has achieved remarkable success in online E-commerce applications. The majority of methods aim to distinguish positive items and negative items for each query by utilizing margin loss or softmax loss. Despite their decent performance, these methods are highly sensitive to hyper-parameters, e.g., the margin and the temperature, which measure the similarity of negative pairs and affect the distribution of items in metric space. How to design and choose adaptively parameters for different pairs is still an open challenge. Recently several methods have attempted to alleviate the above problem by learning each parameter through trainable/statistical methods in the recommendation. We argue that those are not suitable for retrieval scenarios, due to the agnosticism and diversity of the queries. To fully overcome this limitation, we propose a novel adaptive metric learning method that designs a simple and universal hyper-parameter-free learning method to improve the performance of retrieval. Specifically, we first propose a method that adaptive obtains the hyper-parameters by relying on the batch similarity without fixed or extra-trainable hyper-parameters. Subsequently, we adopt a symmetric metric learning method to mitigate model collapse issues. Furthermore, the proposed method is general and sheds a highlight on other fields. Extensive experiments demonstrate our method significantly outperforms previous methods on a real-world dataset, highlighting the superiority and effectiveness of our method. This method has been successfully deployed on an online E-commerce search platform and brought substantial economic benefits.

1 Introduction

In recent years, pre-trained deep semantic retrieval models have made significant progress and application, particularly in the e-commerce field, with the research and application of deep learning technology. Compared with traditional lexical-based methods, the deep semantic retrieval model has the advantages of high accuracy, low mismatch, and strong generalization. The classical deep semantic retrieval methods could be split into two categories, sparse-based retrieval (Bai et al., 2020; Shen et al., 2022; Formal et al., 2021; Gao et al., 2021), and dense-based retrieval (Zhang et al., 2020; Khattab and Zaharia, 2020; Zhan et al., 2021; Qiu et al., 2022; Wang et al., 2023). Although there are differences in representation, they all learn the deep model through an end-to-end paradigm using contrastive learning methods. Specifically, they first represent the query and item into dense/spare vectors in metric space. Then, they adopt softmax loss or margin loss to distinguish the positive item and the negative item. The hyper-parameters of margin or temperature measure the disagreement of similarity of query and candidate items.

Nevertheless, we argue that this paradigm also has several limitations as it fails to meet query requirements. For instance, different queries have different metrics for candidate items. Some queries are general words, and the distance between the query and candidate items may be smaller than precise words. Reflected in the metric space, this query should be a smaller margin in margin loss, and vice versa. Analogously, the temperature of the general query should be large, which has a greater entropy for the similarity of query and items. However, the classical Deep retrieval method fails to address these issues due to the fixed hyper-parameters. As we know, hyper-parameter searches are highly time-cost expensive to find the optimum by grid approach and have a significant impact on performance.

Designing and choosing adaptive parameters for different pairs remains an open challenge. Recently several methods considered to alleviate the above problem by learning each parameter through train-

*Corresponding Author.

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 775–782
December 6-10, 2023 ©2023 Association for Computational Linguistics

able/statistical methods in recommender systems. Typically, (Li et al., 2020) first presents a concept of user/item bias to measure the margin in metric space for margin loss, which could be trained by background optimization. (Ma et al., 2020) proposes to learn an adaptive margin for different users via bi-level optimization (Jiao et al., 2022), where a proxy function is built to explicitly update the margin generation-related parameters. However, those are not available for retrieval scenarios, due to the agnosticism and diversity of the query. Similarly, (Chen et al., 2023) also finds that the temperatures play an important role from the perspective of limitation of normalization and develop an adaptive fine-grained strategy for the temperature with satisfying four desirable properties including adaptivity. This is the first work to explicitly talk about the problem and attempt to study how to adaptively set the proper temperature for recommender systems. We argue that this method is complex and unavailable in retrieval systems because the query is unknown in the real world not like the field of recommendation, where users are given.

Thus, the above limitation motivates us to design a simple and hyper-parameter-free method to enhance the performance of deep retrieval. Toward this end, we present a unified solution to address the problem for both softmax loss and margin loss. More precisely, we first analyze the limitation of the original method from the metric space. Subsequently, we present a heuristic method, which generates the margins/temperatures by the inner product of batch samples adaptively. To prevent the collapse of the training process i.e., all items' embedding have high similarity, we adopt the symmetric metric learning method to push the positive items far away from the negative items. We conduct extensive experiments in the real-world e-commerce field, and the results demonstrate the effectiveness of the proposed.

The contributions of this paper can be summarized as follows:

- This is the first work to present a unified solution for hyper-parameters-free learning in deep semantic retrieval.
- We propose a novel adaptive learning method that designs a parameters-free component to adjust the distance of items in metric space and aligns the query and item embedding space via symmetric metric learning.
- We conduct extensive experiments on a real-world dataset. Experimental results show that our model achieves significant improvement over the classical models.
- This method has been successfully deployed on an online E-commerce search platform and brought substantial economic benefits.

2 PRELIMINARIES

In this section, we first give the formulation of the retrieval task and then present the heuristic method for two classical paradigms, i.e., margin loss and softmax loss. Finally, we will give a detailed description of the proposed hyper-parameters-free methods.

2.1 Formulation

In the E-commerce field, the dense retrieval problem can be formulated as follows (Wang et al., 2023). Suppose there is a set of query Q and a set of items V, and all query-item interactions(clicked or ordered) are noted as $I = (q, v)|q \in Q, v \in V$. Given a query q, The algorithm of dense retrieval is to recall the K most relevant items from the large collection of N items. The dimension of representation of query or item is denoted as D. For a clear definition, throughout the rest of this paper, bold lowercase letters represent vectors.

2.2 Representation Learning

The most classical model of representation learning is the dual-encoder based model (Qu et al., 2021; Karpukhin et al., 2020; Ren et al., 2021; Li et al., 2021; Huang et al., 2020; Qiu et al., 2022), i.e., query encode and item encode, which represent queries and products with embeddings. Considering the personalized effect, the user's identifying information will be included, denoted as:

$$\mathbf{q} = f(q, u) \in \mathcal{R}^D, \mathbf{v} = f(v) \in \mathcal{R}^D \quad (1)$$

where f is the mapping function, such as MLP, Bert; the u denotes the user's features and v denotes the item's features. The similarity of query and item is calculated by inner-product:

$$s(q, v) = \langle \mathbf{q}, \mathbf{v} \rangle \quad (2)$$

2.3 Loss Function

The object of the dual-encode model is mainly trained by negative sampling or batch-negatives (Zhang et al., 2020; Li et al., 2021). Specifically,

given a sample list with one positive item and n negative items, i.e., $< q, v, v_1^-, v_2^-, \ldots, v_n^- >$, the goal is to push the negative item away from the query and pull the positive item close to the query. The mathematical formula could be described as:

$$s(q, v) > \max\left(s(q, v_1^-), \ldots, s(q, v_n^-)\right) \quad (3)$$

There are two classical loss functions for training in E-commerce, i.e., margin loss and softmax loss.

2.3.1 Margin Loss

The margin loss aims to distinguish the positive item and negative item by a margin δ, which is a fixed hyper-parameter that controls the decision boundary in the metric space. The formulation of margin loss could be denoted as follows:

$$\mathcal{L}_{margin} = \sum_{i}^{n} \left[s(q, v_i^-) - s(q, v) + \delta\right]_+ \quad (4)$$

where $[*]_+ = max(*, 0)$.

2.3.2 Softmax Loss

The softmax loss could achieve great training stability and align well with the ranking metric. It usually achieves better performance than others and thus attracts much attention in retrieval. The formulation is denoted as:

$$\mathcal{L}_{soft} = -\log \frac{\exp^{s(q,v)/\tau}}{\exp^{s(q,v)/\tau} + \sum_{i=1}^{n} \exp^{s(q,v_i^-)/\tau}} \quad (5)$$

where τ is the temperature (Wang and Liu, 2021; Li et al., 2021), smoothing the overall fitted distribution of the training data. A small value means that the model completely fits the supervisory signals and is more focused on the hard negative items, and vice versa.

3 Approach

In this section, we will first talk about the limitation of the loss function above-mentioned and then will give a general heuristic method in accordance with the measure assumptions. Finally, we describe the complete method, the symmetric metric learning method in detail.

3.1 Limitation

The loss function mentioned above depends on the hyper-parameters, which play a significant role in performance. Specific experiments will be discussed in the following section. Unfortunately, traditional methods suffer from the problem of choosing hyper-parameters adaptively. Additionally, in personalization scenarios, each pair requires a specific margin and temperature value, making it even more challenging to learn or select the appropriate value.

While other fields, such as recommender systems, have addressed this issue through bi-level or statistical learning, we argue those methods are not suitable for retrieval scenarios. Retrieval scenarios involve input queries that are different from recommendations because input queries from online systems are abundant and agnostic. Therefore, a parameter-free method is necessary to generate the specific value.

To this end, we first present a heuristic method that computes the value by inner product and then propose a symmetric metric learning method to alleviate the problem of collapse in the training process.

3.2 Heuristic Method

In the metric space, the position of the hardest negative items is very close to the positive item, while easy or random negative items remain far away from positive items. We need to distinguish the negative items in fine-grained. Given a pair $< q, v, v_i^- >$, if v_i^- is the hardest negative, the similarity of query and positive v should be higher, in other words, the margin should be smaller in margin loss. Similarly, for the hardest negative, the temperature τ also should be smaller in the softmax loss.

Without generality, we adopt the inner product to measure the similarity of items in this paper. According above metric assumption, given a pair $< q, v, v_i^- >$, the corresponding margin δ_q^i could be computed as follows:

$$\delta_i^q = \alpha * (1 - < v, v_i^- >) + \delta_0 \quad (6)$$

where α and δ_0 could be trainable or global constant parameters, scaling the value for different datasets. It is easy to find that, when $\alpha = 0$, the heuristic method is equivalent to the original model with fixed margin δ_0.

Because multi-pairs will share the positive and negative items, e.g., $< q_1, v, v_i^- >, < q_2, v, v_i^- > , \ldots, < q_m, v, v_i^- >$, in practice, we could share the margin value in one batch, i.e., $\delta_i = \delta_i^q$.

Though the heuristic method is simple and free-trained, we discuss that it will suffer from model collapse during the training process, resulting in all items being clustered together in a metric space. From the perspective of gradient, we can know that the adaptive margin will affect the update direction of positive and negative items.

$$\frac{\partial \mathcal{L}_{margin}}{\partial v} = -q - \alpha * \sum_i v_i^-; \frac{\partial \mathcal{L}_{margin}}{\partial v_i^-} = q - \alpha * v \tag{7}$$

To remit the problem of model collapse, the straightforward method is to adopt the stop gradient strategy eliminating the effect of margin.

$$\delta_i = \alpha * (1 - < sg(v), v_i^- >) + \delta_0 \tag{8}$$

where sg(*) is the operation of stop_gradient. The final loss function could be formulated as follows:

$$\mathcal{L}_{adap\_margin} = \sum_i^n \left[s(q, v_i^-) - s(q, v) + \delta_i \right]_+ \tag{9}$$

3.3 Symmetric Metric Learing

Although the above function is efficient in some scenarios, there is still a risk of collapse due to bad initialization, such as $s(q, v) <= s(q, v_i)$). Essentially, the distance between the query and the item $s(q, v)$ is too large, while the distance of item pairs $s(v, v_i)$ is smaller, reflecting the problem of misalignment of the two spaces, i.e., the query's space and the item's space.

To avoid this problem completely, we introduce an additional symmetric metric learning loss based on rank loss. More specifically, it exchanges the anchor of a given sample list, i.e., v as the anchor, q as the positive item, which aims to push the negative item away from the positive item, denoted as:

$$s(q, v) > \max\left(s(v, v_1^-), \ldots, s(v, v_n^-)\right) \tag{10}$$

Similar as Equation 9, the approximate function is:

$$\mathcal{L}_{symm\_margin} = \sum_i^n \left[s(v, v_i^-) - s(q, v) + \delta_i^{'} \right]_+ \tag{11}$$

where $\delta_i^{'} = \alpha * (1 - < sg(q), v^- >) + \delta_0$. Since this is an auxiliary task, it could be set to δ_0 (i.e., $\alpha = 0$) for simplicity.

3.4 Overall Loss

Now we summarize the optimization complete objective of origin margin loss and symmetric loss as follows:

$$\mathcal{L} = \mathcal{L}_{adap\_margin} + w * \mathcal{L}_{symm\_margin} \tag{12}$$

where w is the weight of symmetric loss.

Along the same perspective, we can give the total objective for the softmax loss paradigm, denoted as:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{adap\_soft} + w * \mathcal{L}_{symm\_soft} = \\ &- \frac{1}{N} \sum \log \frac{\exp^{s(q,v)/\tau_0}}{\exp^{s(q,v)/\tau_0} + \sum_{i=1}^n \exp^{s(q,v_i^-)/\tau_i}} \\ &- \frac{w}{N} \sum \log \frac{\exp^{s(q,v)/\tau_0}}{\exp^{s(q,v)/\tau_0} + \sum_{i=1}^n \exp^{s(v,v_i^-)/\tau_i^{'}}} \end{aligned} \tag{13}$$

where $\tau_i = \delta_i$, $\tau_i^{'} = \delta_i^{'}$.

4 Experiments

In this section, we conduct extensive experiments to evaluate the performance of the proposed method and investigate the effect of different components by ablation studies.

Datasets and Metrics

We collect search logs of user clicks and purchases for 60 days from an online E-commerce website, where the size of the dataset is 5 billion. We choose the standard retrieval quality metric batch-top@K to measure the results based on the batch samples, and Recall@K to measure the results based on the full corpus, where $K \in \{1, 2, 5, 10, 50\}$ and $\{1, 50, 500, 1000\}$, respectively.

To evaluate the online performance, we choose the classical metrics(Wang et al., 2023; Li et al., 2021; Yuan et al., 2023), such as UV-value (revenue per Unique Visitor), and UCVR (Oderlines/UV), to measure the results of the A/B test. We also measure the performance by the number of items after passing the relevance module, and participating in the pranking phase, denoted as Num_{prank}.

Baselines

In the industrial field, the most widely used work could be divided into two backbones: DSSM (Huang et al., 2013) and a pre-trained model based on Bert (Devlin et al., 2018). Without loss of generality, we select DSSM and DPSR (Zhang et al., 2020) (considering the personalized information) as baselines with the backbone of DSSM,

Table 1: Offline experimental results on recall@K (K is set to 1, 50, 500, 1000)

Methods	recall@1	recall@50	recall@500	recall@1000
Backbone DSSM[a]				
DSSM[a] (Huang et al., 2013)	0.0069	0.1789	0.5706	0.6806
DSSM+MMSE (Wang et al., 2023)	0.0228	0.4063	0.7517	0.8067
DPSR (Zhang et al., 2020)	0.0076	0.1946	0.5771	0.6817
DPSR+MMSE	0.0237	0.3993	0.7447	0.8014
LTR (Liu et al., 2009)	0.0061	0.1558	0.4782	0.5835
Backbone Bert[b]				
RSR (Qiu et al., 2022)	0.0094	0.1980	0.5486	0.6496
RSR+M_1	0.0076	0.1932	0.5622	0.6583
RSR+M_2	0.0107	0.2108	0.6025	0.6993
RSR+MMSE (Wang et al., 2023)	0.0099	0.2201	0.6145	0.7104
Ours	**0.0137**	**0.2637**	**0.6156**	**0.7122**

[a]The vocabulary size and batch size of backbone DSSM is set to 400k, 1024, while Bert is 20k, 350.

and RSR (Qiu et al., 2022) and variant version with multi-objective learning (RSR + MMSE) as the representative of the backbone of Bert. Noting that the strong baselines are **RSR**, deep personalized and semantic retrieval, devoted to tackling the personalized problem of different users, which had been deployed in the online system, severing hundreds of millions of users.

It is worth noting that the DSSM, DPSR, RSR, and MMSE methods all used the softmax loss paradigm during training since the performance of softmax loss is better than margin loss. Thus, our method also adopts softmax loss for fair comparison.

Our method is easily extensible and could be adapted in various versions of dual-encode (e.g, MGDSPR (Li et al., 2021), Colbert (Khattab and Zaharia, 2020)), and multi-objective learning RSR+MMSE (Wang et al., 2023).

Implementation Details

To ensure a fair comparison among different methods, we keep the feature, vocabulary size, the dimension of query/item, and parameters of PQ the same as (Wang et al., 2023). Specifically, we set the dimension as 128, batch size as 350, and n-list of IVF-PQ as 32768, and the indexing construction is used in the Faiss ANNS library [1]. The default temperature τ of softmax is 1/30, the margin is set to 0.1. α is set to 0.5, δ_0 is set to 0.01, and τ_0 is set to 1/30. The default value of w is set to 0.05. The Adam optimizer is employed with an initial learning rate of 5e-5. The maximum length of query and product sequences are 30, and 100, respectively.

[1]https://github.com/facebookresearch/faiss

4.1 Experiment Results

The experimental results are shown in Table 1. From the results, we can conclude that the proposed method achieves a significant improvement over the baselines. Specifically, Our method performs better than the RSR and RSR's variations (such as RSR + M1, and RSR +M2) in terms of recall@K. It is particularly noteworthy that our method is similar to RSR which is also finetuned based on a pre-trained Bert model in the clicked dataset. Therefore, comparing the performance of RSR and ours, we can find that the proposed components, i.e., hyper-parameters-free and symmetric metric learning loss, make potential improvements gained for retrieval. This also demonstrates that the design of objective loss is significant for the training process. Compared with RSR's variation, especially the MMSE which measures full sample space, we can see that the more finely grained would bring great benefits, motivating us to extend ours to multi-objective learning in the future.

4.2 Impact of Hyper-paremeters

As we mentioned in the introduction, we discuss that hyper-parameters will have a huge effect on performance. In this subsection, to prove that we conduct several experiments based on RSR to investigate the impact of different margins δ and temperature τ for margin loss and softmax loss. According to experimental results, we find that the metric of batch-top@K is positively correlated with recall@K. Thus, for quick validation, we only use

Table 2: The impact of different margin m and temperature τ for margin-loss and softmax loss, the batch size is 512.

method	batch-top@1	batch-top@2	batch-top@5	batch-top@10	batch-top@50
margin loss m = 0.01	0.7502	0.8610	0.9376	0.9689	0.9976
margin loss m = 0.1	0.7510	0.8621	0.9382	0.9691	0.9977
margin loss m = 0.5	0.7016	0.8292	0.9257	0.9529	0.9973
margin loss m = 1.0	0.1811	0.2871	0.4904	0.6827	0.9617
softmax loss $1/\tau = 1$	0.1510	0.2481	0.4475	0.6430	0.9521
softmax loss $1/\tau = 10$	0.7361	0.8532	0.9436	0.9675	0.9973
softmax loss $1/\tau = 50$	0.7536	0.8633	0.9385	0.9691	0.9977

Table 3: Online performance of A/B tests. The improvements are averaged over 10 days in 2023. p-value is obtained by t-test over the RSR+MMSE.

Metric	Num_{prank}	UCVR	UV-value
Gain	+2.0%	+0.450%	+0.353%
p-value[b]	-	0.0238	0.2492

[b] Small p-value means statistically significant.

batch-top@k as the metric. Results are shown in Table 2. As we have shown, the performance is highly sensitive to hyper-parameters, both margin and temperature. For the margin loss, the smaller m is intended for better performance and the τ has a similar phenomenon for softmax loss. This could be explained by the more attention on hard negative items, the more performance improvement. This conclusion is consistent with previous work (Li et al., 2020; Jiao et al., 2022; Chen et al., 2023).

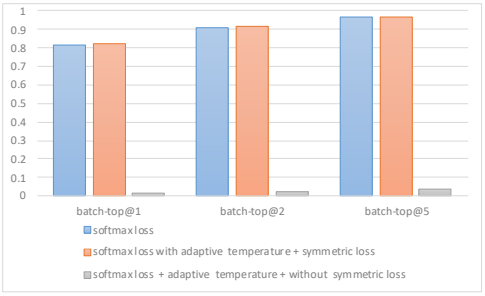

Figure 1: The impact of different components on performance in terms batch-top@K, K $\in \{1, 2, 5\}$.

4.3 Ablation Study

In this subsection, we investigate the impact of different components on performance in terms of batch-top@K. There are two components, adaptive temperature, and symmetric loss. As shown in Figure 1, we can observe that the method without the symmetric loss component will lead to extremely poor results, which validates the above discussion of the model collapse phenomenon. In other words, symmetric metric learning is indispensable for the training process. What's more, the adaptive temperature is also a useful component for retrieval.

4.4 Online A/B test

To investigate the effectiveness of the proposed method in the real-world commercial scenario, we conduct several A/B tests, and the online results are reported in Table 3. Comparing with the base model (RSR+MMSE)in the real online environment, we can note that our performance increases by 2.0% in terms of Num_{prank} and 0.45% in UCVR, respectively, which demonstrates that the designed techniques are practical gains for the online system.

5 Conclusion and Future Work

This paper addresses the challenge of adaptive hyper-parameter selection for the contrastive learning paradigm in deep retrieval fields. Toward this end, we first present a straightforward heuristic method, which uses the batch similarity of items to generate a margin or temperature, eliminating the complex trainable variables. However, our theoretical analysis reveals that this method is prone to model collapse. To prevent the above problem, we adopt the symmetric metric learning method to align the query and items in metric space. Experiments verify that our assumptions and method are simple, effective, and significantly outperform other models in real-world datasets. Moreover, we have successfully deployed this method on online search platforms, leading to significant commercial

value. It is worth noting that, this method could give a great inspiration for other fields, such as recommendation and knowledge graph learning.

In future work, we aim to explore the benefits of multi-objective learning with adaptive hyper-parameters in the full sample space.

References

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*.

Jiawei Chen, Junkang Wu, Jiancan Wu, Xuezhi Cao, Sheng Zhou, and Xiangnan He. 2023. Adap-τ: Adaptively modulating embedding magnitude for recommendation. In *Proceedings of the ACM Web Conference 2023*, pages 1085–1096.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292.

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. Coil: Revisit exact lexical match in information retrieval with contextualized inverted list. *arXiv preprint arXiv:2104.07186*.

Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. 2020. Embedding-based retrieval in facebook search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2553–2561.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338.

Yang Jiao, Kai Yang, Tiancheng Wu, Dongjin Song, and Chengtao Jian. 2022. Asynchronous distributed bilevel optimization. *arXiv preprint arXiv:2212.10048*.

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 *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Mingming Li, Shuai Zhang, Fuqing Zhu, Wanhui Qian, Liangjun Zang, Jizhong Han, and Songlin Hu. 2020. Symmetric metric learning with adaptive margin for recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4634–4641.

Sen Li, Fuyu Lv, Taiwei Jin, Guli Lin, Keping Yang, Xiaoyi Zeng, Xiao-Ming Wu, and Qianli Ma. 2021. Embedding-based product retrieval in taobao search. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3181–3189.

Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331.

Chen Ma, Liheng Ma, Yingxue Zhang, Ruiming Tang, Xue Liu, and Mark Coates. 2020. Probabilistic metric learning with adaptive margin for top-k recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on knowledge discovery & data mining*, pages 1036–1044.

Yiming Qiu, Chenyu Zhao, Han Zhang, Jingwei Zhuo, Tianhao Li, Xiaowei Zhang, Songlin Wang, Sulong Xu, Bo Long, and Wen-Yun Yang. 2022. Pre-training tasks for user intent detection and embedding retrieval in e-commerce search. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4424–4428.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne 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 *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847.

Ruiyang Ren, Shangwen Lv, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. Pair: Leveraging passage-centric similarity relation for improving dense passage retrieval. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2173–2183.

Tao Shen, Xiubo Geng, Chongyang Tao, Can Xu, Kai Zhang, and Daxin Jiang. 2022. Unifier: A unified retriever for large-scale retrieval. *arXiv preprint arXiv:2205.11194*.

Binbin Wang, Mingming Li, Zhixiong Zeng, Jingwei Zhuo, Songlin Wang, Sulong Xu, Bo Long, and Weipeng Yan. 2023. Learning multi-stage multi-grained semantic embeddings for e-commerce search.

In *Companion Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 411–415. ACM.

Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504.

Chunyuan Yuan, Yiming Qiu, Mingming Li, Haiqing Hu, Songlin Wang, and Sulong Xu. 2023. A multi-granularity matching attention network for query intent classification in e-commerce retrieval. In *Companion Proceedings of the ACM Web Conference 2023*, pages 416–420.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Jointly optimizing query encoder and product quantization to improve retrieval performance. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2487–2496.

Han Zhang, Songlin Wang, Kang Zhang, Zhiling Tang, Yunjiang Jiang, Yun Xiao, Weipeng Yan, and Wen-Yun Yang. 2020. Towards personalized and semantic retrieval: An end-to-end solution for e-commerce search via embedding learning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2407–2416.