TriSampler: A Better Negative Sampling Principle for Dense Retrieval

Zhen Yang, Zhou Shao, Yuxiao Dong, Jie Tang*

Department of Computer Science and Technology, Tsinghua University, Beijing, China
yangz21@mails.tsinghua.edu.cn, {shaozhou,yuxiaod,jietang}@tsinghua.edu.cn

Abstract

Negative sampling stands as a pivotal technique in dense retrieval, essential for training effective retrieval models and significantly impacting retrieval performance. While existing negative sampling methods have made commendable progress by leveraging hard negatives, a comprehensive guiding principle for constructing negative candidates and designing negative sampling distributions is still lacking. To bridge this gap, we embark on a theoretical analysis of negative sampling in dense retrieval. This exploration culminates in the unveiling of *the quasi-triangular principle*, a novel framework that elucidates the triangular-like interplay between query, positive document, and negative document. Fueled by this guiding principle, we introduce **TriSampler**, a straightforward yet highly effective negative sampling method. The keypoint of TriSampler lies in its ability to selectively sample more informative negatives within a prescribed constrained region. Experimental evaluation show that TriSampler consistently attains superior retrieval performance across a diverse of representative retrieval models.

Introduction

Recently, dense retrieval has gained tremendous attention for its remarkable performance across a spectrum of real-world downstream applications, such as open-domain question answer (Karpukhin et al. 2020), web search (Xiong et al. 2020), and conversational search (Yu et al. 2021). Within the domain of dense retrieval, the focal point lies in the retrieval models' ability to effectively distinguish pertinent documents for a given query from the vast pool of non-relevant documents within the corpus. This task, while crucial, is confronted with the challenge of managing an extensive set of negative documents. Attempting to leverage the entirety of these negatives is often impractical, which highlights the significance of negative sampling.

Indeed, negative sampling emerges as a vital technique in tackling this challenge. By advisably selecting a subset of negative documents, negative sampling allows for a more efficient and effective training process. This strategic management of negative samples not only mitigates the computational burden but also enhances the model's capacity to discern meaningful patterns, thereby contributing to the overall success of dense retrieval methods.

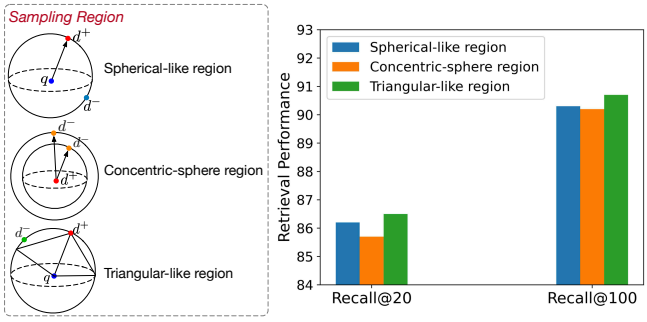


Figure 1: Insight experiments to illustrate the significance of the quasi-triangular principle.

The pursuit of optimizing negative sampling techniques for dense retrieval has been a topic of considerable investigation. Previous efforts (Karpukhin et al. 2020; Kalantidis et al. 2020; Zhan et al. 2021; Qu et al. 2020; Chuang et al. 2020) have delved into a range of methods to sample negatives at scale, encompassing strategies like in-batch negatives, random negatives, hard negatives, and debiased negatives. Drawing inspiration from the domain of contrastive learning (Oord, Li, and Vinyals 2018; He et al. 2020; Chen et al. 2020), dense retrieval models adopt in-batch negatives. This approach, which represents a specialized instance of random negatives, streamlines training efficiency by reusing samples within the same batch, thus eliminating the need for additional sampling operations. Despite its merits, results in several works (Faghri et al. 2017; Kalantidis et al. 2020; Robinson et al. 2020; Karpukhin et al. 2020; Gao et al. 2021) indicate that such easy random negatives may not provide sufficient information for model training, potentially resulting in suboptimal retrieval performance. To address this issue, hard negative sampling methods (Karpukhin et al. 2020; Xiong et al. 2020; Zhan et al. 2021; Qu et al. 2020; Sun et al. 2022) have emerged as effective strategies for performance enhancement. These strategies focus on identifying and sampling top-k hard negatives based on the current model or leveraging an auxiliary retrieval model. However, the utilization of hard negatives presents a pivotal challenge: the risk of incorporating false negatives, which can detrimentally impact the retrieval performance (Schroff, Kalenichenko, and Philbin 2015; Chuang et al. 2020; Qu et al. 2020; Zhou et al. 2022).

*Jie Tang is the Corresponding Author.

While prior studies have employed various negative sampling methods to achieve promising retrieval results, a general principle for guiding negative sampling is yet to be established. There is a distinct need for a clear and quantifiable approach that thoroughly elucidates that relationships among queries, positive documents, and negative documents. It is necessary to propose an explicit negative sampling methodology that adeptly selects more informative negative samples within a constraint region, effectively capturing the essential relationships.

Thus, we propose a general negative sampling principle called *the quasi-triangular principle*, which mandates the constraint of sampled negatives within a triangular-like region. Specifically, this principle simulates the pairwise relationship among a training triple (q, d^+, d^-), emphasizing the selection of negative samples from a region that mirrors the geometric characteristics of a triangle. To delve deeper into this negative sampling principle, we design two extended experiments: (1) sampling negatives from a spherical-like region centered on the query, with the radius defined by the similarity to the positive document; (2) sampling negatives within a concentric-sphere region centered on the positive document, aiming to find negatives that are related but not overly similar to it. As depicted in Figure 1, the outcome of constraining negatives within a triangular-like region notably elevates retrieval performance. This constrained sampling space empowers the retrieval model to more effectively distinguish relevant (positive) and irrelevant (negative) documents, bolstering the model's discriminative capabilities. Insights gleaned from these experiments suggest that confining negatives within a triangular-like region can augments retrieval performance.

To translate the innovative *quasi-triangular principle* into actionable practice, we introduce a straightforward and efficient negative sampling methodology named **TriSampler**. This method encapsulates a two-fold approach, involving the construction of negative candidates and the implementation of a specialized negative sampling distribution. The construction process is meticulously designed to capture the essence of the triangular-like relationships within the data, setting the foundation for an informative selection of negative samples. The implementation of distribution leverages two well-designed sampling distributions to guide the selection of negatives within the predefined triangular-like candidate region. The seamless integration of these two critical components solidifies TriSampler as a general negative sampler.

Contributions: In this paper, we present a general principle for negative sampling, referred to as the *quasi-triangular principle*, centered on the idea of constraining sampled negatives within a triangular-like region that encapsulates the relationships between queries, positive documents, and negative documents. To implement this principle, we propose a simple and effective negative sampling method called TriSampler. This methodology rests on two pillars: the construction of negative candidates and the implementation of a specialized negative sampling distribution. Empirical validation across four diverse retrieval benchmarks demonstrate that TriSampler can achieve better retrieval performance compared to other negative sampling methods.

Related Work

Dense retrieval. Dense retrieval (Lee, Chang, and Toutanova 2019; Karpukhin et al. 2020; Xiong et al. 2020; Khattab and Zaharia 2020) shows tremendous success in many downstream tasks (e.g. open-domain QA and web search) compared with the traditional sparse retrieval models (e.g. TF-IDF and BM25). The primary paradigm is to model semantic interaction between queries and passages based on the learned representations. Most dense retrieval models leverage the pretrained language models to learn latent semantic representations for both queries and passages. Lee, Chang, and Toutanova (2019) first proposed the dual-encoder retrieval architecture based on BERT, paving the way for a new retrieval approach. In order to model fine-grained semantic interaction between queries and passages, Poly-encoder (Humeau et al. 2019), ColBERT (Khattab and Zaharia 2020), and ME-BERT (Luan et al. 2021) explored multi-representation dual-encoder to enhance retrieval performance. Besides, knowledge distillation has become a vital technique to enhance the capacity of the dual-encoder by distilling knowledge from a more powerful reader to a classical retriever (Qu et al. 2020; Ren et al. 2021b; Lin, Yang, and Lin 2020; Hofstätter et al. 2021).

Recently, massive works have investigated task-related pre-training methods for dense retrieval models (Gao and Callan 2021a,b; Wang, Reimers, and Gurevych 2021; Ren et al. 2021a; Oğuz et al. 2021; Meng et al. 2021). Condenser (Gao and Callan 2021a) proposed the Condenser architecture to enforce the late backbone layers to aggregate the whole information. coCondenser (Gao and Callan 2021b) leveraged contrastive learning to incorporate a query-agnostic contrastive loss. PAIR (Ren et al. 2021a) and DPR-PAQ (Oğuz et al. 2021) also designed special tasks in pre-training to enhance retrieval models. Additionally, jointly training retrieval models with the rerank model can bring about better performance. Sachan et al. (2021) proposed an end-to-end training method to jointly or individually model the retrieved documents. Zhang et al. (2021) adopted adversarial training to model the retriever and the reranker.

Negative sampling in dense retrieval. Several recent works (Karpukhin et al. 2020; Xiong et al. 2020; Qu et al. 2020; Zhan et al. 2021) demonstrate that hard negative sampling plays a crucial role in enhancing dense retrieval. Previous studies on negative sampling can be roughly categorized into three categories: (1) random sampling is the simplest way to obtain negatives. As an efficient random sampling method, in-batch negatives are widely used in dense retrieval models (Karpukhin et al. 2020; Zhan et al. 2021). Such an approach is sub-optimal because random negatives have been proven to be too easy for learning effective models. RocketQA (Qu et al. 2020) adopted cross-batch negatives to increase the number of random negatives, resulting in better performance. (2) hard negative sampling can improve model generalization and accelerate convergence. DPR (Karpukhin et al. 2020) additionally integrated hard negative passages from BM25 into in-batch negatives for dense passage re-

trieval. ANCE (Xiong et al. 2020) verified that global hard negatives obtained from the current retrieval model can significantly enhance the retrieval performance. ADORE (Zhan et al. 2021) proposed a dynamic negative sampling method to train retrieval models. ANCE-Tele (Sun et al. 2022) combined past iterations by a momentum queue and future iterations by a lookhead operation to select hard negatives for stable training. (3) debiased hard negative sampling can efficiently alleviate false negatives. RocketQA (Qu et al. 2020) utilized a well-trained cross-encoder to select hard negatives for the dual-encoder training. SimANS (Zhou et al. 2022) proposed ambiguous negatives to reweight the relevant score with the positives. Different from the abovementioned methods, our TriSampler aims to sample negatives within a triangular-like region based on a general *quasi-triangular principle*, which constraints the range of negative candidates and provides more informative negatives for model training.

Understanding Negative Sampling

In this section, we commence by providing an overview of the foundational concepts for dense retrieval. Subsequently, we delve into a comprehensive analysis of the pivotal role that negative sampling plays within the context of dense retrieval.

Preliminary for Dense Retrieval

Previous dense retrieval works (Karpukhin et al. 2020; Xiong et al. 2020) aim to distinguish the most relevant documents $\mathcal{D}^+$ from a large document corpus $\mathcal{D}$ for a given query q. Typically, these retrieval models leverage negative sampling method to sample several negatives to substitute the entire corpus for model training, thus significantly reducing training costs. The objective function for dense retrieval can be simplified as:

$$\mathcal{L} = \sum_{q} \sum_{d^+ \in \mathcal{D}^+} \sum_{d^- \in \mathcal{D}^-} l(s(\mathbf{h}_q, \mathbf{h}_{d^+}), s(\mathbf{h}_q, \mathbf{h}_{d^-})) \quad (1)$$

where $l(\cdot)$ represents a loss function, such as cross entropy or hinge loss, $s(\cdot)$ denotes the dot product used to measure the similarity metric, $\mathbf{h}_q$ and $\mathbf{h}_d$ represent query embedding and document embedding that are encoded by a query encoder and a document encoder respectively. The pre-trained language models (PLMs) (Devlin et al. 2018; Liu et al. 2019; Zhang et al. 2019) serve as dual-encoder and the representations of the [CLS] token are leveraged as embeddings. In subsequent research endeavors, it has been observed that integrating the average of both the first and last layers embeddings confers a significant performance advantage (Li et al. 2020; Su et al. 2021).

The construction of negative candidates $\mathcal{D}^-$ depends on either the current retrieval model or sparse retrieval model like BM25. Subsequently, the final negatives are sampled by applying distinct negative sampling distributions tailored to each specific approach.

Analysis for Negative Sampling

A representative dense retrieval model is trained on a set of training triples $\{(q, d^+, \{d^-\}_{i=1}^n)\}$ where (q, d^+) is a positive query-document pair and $\{d^-\}_{i=1}^n$ are the sampled negative irrelevant documents. A conventional contrastive loss for dense retrieval can be formulated as:

$$\mathcal{L} = -\log \frac{exp(s^+)}{exp(s^+) + \sum_{i=1}^n exp(s_i^-)} \quad (2)$$

where s^+ denotes the positive similarity score between the query and the corresponding positive document $s(\mathbf{h}_q, \mathbf{h}_{d^+})$, s^- represents the negative similarity score between the query and a negative document $s(\mathbf{h}_q, \mathbf{h}_{d^-})$.

The gradient of the aforementioned contrastive loss can be separated into two distinct components concerning s^+ and s_j^-:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial s^+} &= -\frac{\sum_{i=1}^n exp(s_i^-)}{exp(s^+) + \sum_{i=1}^n exp(s_i^-)}, \\ \frac{\partial \mathcal{L}}{\partial s_j^-} &= \frac{exp(s_j^-)}{exp(s^+) + \sum_{i=1}^n exp(s_i^-)} \end{aligned} \quad (3)$$

According to Equation 3, it becomes evident that the gradient with respect to the negative document is proportional to the negative similarity score $exp(s_j^-)$. Notably, when employing randomly sampled negatives, their extremely low similarity scores result in gradients that are nearly negligible, rendering a minimal impact on model training. Conversely, negatives extracted from the top K nearest irrelevant documents yield comparatively higher similarity scores, thereby expediting the convergence of the retrieval model. However, the gradient with respect to the positive document becomes bounded to a fixed value when the negative similarity scores significantly exceed the positive ones. This necessitates the imposition of specific constraints on the relationship between positives and negatives. An uncomplicated relationship, such as $s^+ \approx s^-$ when negatives are drawn from a spherical-like region, can be effective. Such strategically selected negatives infuse more informative signals into the training process, mitigating the occurrence of either zero or fixed-value gradients.

The above analysis clearly demonstrates that negatives adhering to the constraint region $s^+ \approx s^-$ contribute to the elimination of overly hard or overly easy negatives. Furthermore, our approach involves integrating the similarity score between the positive document and negative samples into the negative sampling process. The specific interplay among queries, positive documents, and negative documents will be expounded upon in the next Section.

Method

As elaborated in the abovementioned Section, a promising negative sampling method should adhere to the constraint $s^+ \approx s^-$, implying that negatives are optimally drawn from a spherical-like region. However, considering the expansive expanse of the entire spherical region, it becomes evident that negatives located far from the positive document might not yield valuable insights, given the model's inherent capability to distinguish between positive and negative documents.

In response to this challenge, we introduce the *quasi-triangular principle*, which strategically confines the sampled negatives within a delimited triangular-like region, effectively carving out a subregion of the larger spherical space. By adopting this principle, we narrow the scope of negative

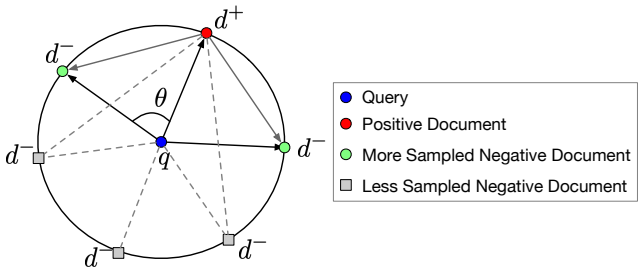


Figure 2: The proposed quasi-triangular principle for negative sampling in dense retrieval.

sampling, focusing on a region that balances the need for informative negatives with the practicality of model training. Based on this principle, we develop a simple and effective negative sampling method TriSampler.

The Principle of Negative Sampling

Here, we propose *the quasi-triangular principle* to simulate the pairwise relationship inherent in a training triple (q, d^+, d^-), ultimately enhancing negative sampling in dense retrieval. This principle operates by constraining the domain of sampled negatives within a triangular-like region, deviating from the broader spherical-like region that has traditionally been employed.

The conceptual foundation of this principle is depicted in Figure 2, which illustrates the planar projection of a sphere. In this context, the angular parameter θ characterizes the extent of the triangular-like region. Mathematically, θ is defined as:

$$\theta = |\arccos(\frac{s(\mathbf{h}_q,\mathbf{h}_{d^+})}{||\mathbf{h}_q||\cdot||\mathbf{h}_{d^+}||}) - \arccos(\frac{s(\mathbf{h}_q,\mathbf{h}_{d^-})}{||\mathbf{h}_q||\cdot||\mathbf{h}_{d^-}||}) \quad (4)$$

Within the triangular-like region, the boundary for negatives is set at $\theta = 60°$. This judicious constraint presents a noteworthy departure from the broader spherical-like region. This distinction serves to concentrate the sampled negatives in closer proximity to the positive documents. Consequently, the triangular-like region is thoughtfully designed to ensure that negatives exhibit a substantial degree of similarity with both the query and the positive document. This strategic arrangement proves pivotal in tackling issues associated with two key types of negatives: those that are too close to the query (potentially leading to false negatives) and those situated far from both the query and the positive (often resulting in uninformative negatives).

Negative Candidates

In order to enhance the informativeness of negative candidates, we seamlessly integrate the *quasi-triangular principle* into the process of constructing negative candidates, denoted as $\mathcal{D}_q^-$. This is carried out within the bounds of the established triangular-like region, specific to each query q. Specifically, we first sample the top-ranked irrelevant documents in terms of the query based on the current retrieval model, which is widely used in previous hard negative selection methods (Xiong et al. 2020; Zhan et al. 2021; Zhang et al. 2021; Zhou et al. 2022). Subsequently, we determine the relevant scores between the positive document and the selected top-ranked irrelevant documents. Building upon the aforementioned relevant scores $s(\mathbf{h}_q, \mathbf{h}_{d^-})$ and $s(\mathbf{h}_{d^+}, \mathbf{h}_{d^-})$ where $d^- \in \mathbf{TopK}_{s(q,\mathcal{D}^-)}$, we establish criteria for the construction of a more informative negative candidate set in alignment with the *quasi-triangular principle*. This involves two pivotal constraints:

- Negative candidates should conform to the first range constraint $s(\mathbf{h}_q, \mathbf{h}_{d^+}) \approx s(\mathbf{h}_q, \mathbf{h}_{d^-})$, effectively filtering out too hard or too easy negatives;
- Negative candidates should be in line with the second range constraint $s(\mathbf{h}_{d^+}, \mathbf{h}_{d^-}) \geq s(\mathbf{h}_q, \mathbf{h}_{d^-})$, fostering the inclusion of negatives that are notably informative.

Negative Sampling Distribution

The primary goal of negative sampling method is to design an effective distribution that effectively samples high-quality negatives from the negative candidates. Based on *the quasi-triangular principle*, we formulate the first distribution to adhere to the range constraint $s(\mathbf{h}_q, \mathbf{h}_{d^+}) \approx s(\mathbf{h}_q, \mathbf{h}_{d^-})$. This distribution takes the form:

$$p_{d^-}^{(q)} \propto exp(-\frac{1}{4} * (s^- - s^+)^2) \quad (5)$$

where s^- and s^+ represent $s(\mathbf{h}_q, \mathbf{h}_{d^-})$ and $s(\mathbf{h}_q, \mathbf{h}_{d^+})$ respectively. This distribution embodies the aim of mitigating excessively hard negatives (i.e., false negatives) while concurrently reinforcing the first range constraint $s^+ \approx s^-$. Implementing this distribution exclusively over the top-ranked negative candidates $\mathbf{TopK}_{s(q,\mathcal{D}^-)}$ yields the transitional negatives $\tilde{\mathcal{D}}_q^-$.

Turning to the second range constraint, we devise a novel distribution to sample the final negatives for the training of the retrieval model. Specifically, we allocate higher sampling probabilities to negatives that are in close proximity to the positive document within the triangular-like region. This distribution can be succinctly represented as:

$$p_{d^-} \propto \text{ReLU}(s(\mathbf{h}_{d^+}, \mathbf{h}_{d^-}) - s(\mathbf{h}_q, \mathbf{h}_{d^-})) \quad (6)$$

where this distribution is conducted on transitional negatives $\tilde{\mathcal{D}}_q^-$.

The key insight of using the RuLU function is that it can exclude negatives that are not in the triangular-like region and further guarantee that negatives that are closer to positive possess higher sampling probabilities. By applying these distributions, we holistically adhere to the *quasi-triangular principle*, successfully selecting a set of informative negative samples for retrieval model training.

Discussion

Algorithm 1 represents the overall training process of TriSampler. In the ensuing discussion, we delve into the connection and discrimination between TriSampler and previous negative sampling methods.

- **TriSampler vs RandNS.** RandNS (Huang et al. 2020) is a basic method that randomly samples negatives from a huge set of negative candidates. TriSampler relies on the *quasi-triangular principle* to sample more informative negatives within the triangular-like region. Different from RandNS that assigns equal weights for each negative, TriSampler leverages a well-designed distribution to sample negatives.

Algorithm 1: Algorithm of TriSampler

Input: Positive query-documents $\{(q, \mathcal{D}^+)\}$, document corpus $\mathcal{D}$.

Output: Negative documents $\hat{\mathcal{D}}_q^-$.

1: Build ANN index on $\mathcal{D}$.
2: Generate the top-ranked negative candidates $\mathbf{TopK}_{s(q,\mathcal{D}^-)}$ from $\mathcal{D}$.
3: Sample transitional negatives $\tilde{\mathcal{D}}_q^-$ from $\mathbf{TopK}_{s(q,\mathcal{D}^-)}$ with distribution $p_{d^-}^{(q)}$.
4: Sample final negatives $\hat{\mathcal{D}}_q^- = \{d^-\}_{i=1}^n$ based on distribution p_{d^-} from $\tilde{\mathcal{D}}_q^-$.

- **TriSampler vs TopNS.** TopNS aims to sample top-k ones from all ranked negatives based on a dynamic-trained dense retrieval model (Xiong et al. 2020; Zhan et al. 2021) or a sparse retrieval model (Karpukhin et al. 2020) (BM25). Unlike TopNS which has a higher risk of false negatives, TriSampler eliminates too hard negatives via a constraint triangular-like region.
- **TriSampler vs SimANS.** SimANS (Zhou et al. 2022) designs a negative sampling distribution to sample ambiguous negatives, which avoids sampling negatives that are either too hard or too easy. Similar to SimANS, TriSampler also devises two distributions for the constraint region. The main difference between these is that TriSampler limits negatives within a triangular-like region while SimANS leverages top-ranked negatives as the sampling region.
- **TriSampler vs ANCE-Tele.** ANCE-Tele (Sun et al. 2022) combines three types of negatives (standard ANCE negatives, momentum negatives, and lookahead negatives) to form negative candidates and then randomly sample negatives from the above candidates. Different from ANCE-Tele, TriSampler constraints the sampling region within a triangular-like region and employs two specifically-designed distributions for sampling.

Experiments

Experimental Setup

Datasets. We conduct experiments on the first retrieval stage of four benchmarks: three passage retrieval datasets: MS MARCO passage (MS Pas) (Nguyen et al. 2016), Natural Questions (NQ) (Kwiatkowski et al. 2019), and TriviaQA (TQA) (Joshi et al. 2017), and a document retrieval dataset: MS MARCO document (MS Doc) (Nguyen et al. 2016). The statistics of each dataset is illustrated in Table 1.

Datasets	Training	Dev	Test	Documents
NQ	58,880	8,757	3,610	21,015,324
TQA	60,413	8,837	11,313	21,015,324
MS Pas	502,939	6,980	-	8,841,823
MS Doc	367,013	5,193	-	3,213,835

Table 1: The statistics of four retrieval datasets.

Evaluation metrics. We evaluate retrieval performance using official evaluation methodologies, such as MMR@10 and R@k. For the NQ and TQA datasets, R@20 and R@100 serve as metrics to measure whether the top-k retrieved passages contain the answer span. We evaluate the results on their dev datasets in terms of MRR@10 and R@50 for MS Pas dataset, MRR@10 and R@100 for MS Doc dataset.

Baselines. We compare TriSamplerwith previously established baselines for retrieval benchmarks. Baselines can be generally divided into the following categories.

- **Sparse Retrieval.** The compared sparse retrieval models contains BM25 (Yang, Fang, and Lin 2017) and improved variants of BM25 models that incorporate pretrained language models, such as doc2query (Nogueira, Lin, and Epistemic 2019), DeepCT (Dai and Callan 2019), docTTTTTquery (Nogueira et al. 2019), and GAR (Mao et al. 2020).
- **Dense Retrieval.** Massive dense retrieval baselines have investigated a variety of training methods to improve the retrieval performance, such as hard negative sampling (Karpukhin et al. 2020; Xiong et al. 2020; Zhan et al. 2021; Zhou et al. 2022), distillation (Qu et al. 2020; Lu et al. 2022; Ren et al. 2021b), integrating rerankers into retrievers (Zhang et al. 2021), pre-training (Ren et al. 2021a; Gao and Callan 2021b,a), etc. Among these, hard negative sampling is a particularly important strategy. DPR (Karpukhin et al. 2020), RocketQA (Qu et al. 2020), ANCE (Xiong et al. 2020), ADORE (Zhan et al. 2021), and SimANS (Zhou et al. 2022) attempt to design various negative sampling methods to obtain top-k hard negatives.

Implementation details. We implement TriSampler based on SOTA dense retrieval model AR2 (Zhang et al. 2021) and run all experiments on 8 NVIDIA Tesla A100 GPUs. Following AR2, ERNIE-2.0-base (Sun et al. 2020) serves as a backbone model to encode queries and passages. Similar to SimANS (Zhou et al. 2022), we directly utilize checkpoints in the AR2 model to continue training with the proposed TriSampler. For MS Doc dataset, the model parameters are initialized with STAR (Zhan et al. 2021). In our experiments, the ratio of positive to negative pairs is set to $1:15$, the inner product is leveraged to estimate the relevance score and Faiss (Johnson, Douze, and Jégou 2019) is adopted for efficient similarity search. We utilize the top-200 passages for NQ and TQA datasets and the top-400 documents for MS Pas and MS Doc datasets as negative candidates.

Overall Results

TriSampler achieves a better retrieval performance than most of baselines on all metrics (See Table 2 and Table 3). The improvements primarily stem from the superiority of the *quasi-triangular principle* over previous hard negative sampling methods. Since the measurement principle between query-negatives and pos_passage-negatives may share a *quasi-triangular principle*, previous methods are unable to capture this principle or even overlook the impact of pos_passage-negatives. Compared with pre-training methods for dense retrieval (SimLM and LexMAE), our TriSampler outperforms SimLM but not surpass LexMAE. LexMAE encompasses

Method	NQ		TQA		MS Pas	
	R@20	R@100	R@20	R@100	MRR@10	R@50
BM25 (Yang, Fang, and Lin 2017)	59.1	73.7	66.9	76.7	18.7	59.2
doc2query (Nogueira et al. 2019)	-	-	-	-	21.5	64.4
DeepCT (Dai and Callan 2019)	-	-	-	-	24.3	69.0
docTTTTTquery (Nogueira, Lin, and Epistemic 2019)					27.7	75.6
GAR (Mao et al. 2020)	74.4	85.3	80.4	85.7	-	-
DPR (Karpukhin et al. 2020)	78.4	85.4	79.3	84.9	-	-
ME-BERT (Luan et al. 2021)	-	-	-	-	33.8	-
Joint top-k (Sachan et al. 2021)	81.8	87.8	81.3	86.3	-	-
Individual top-k (Sachan et al. 2021)	84.0	89.2	83.1	87.0	-	-
RocketQAv2 (Ren et al. 2021b)	83.7	89.0	-	-	38.8	86.2
PAIR (Ren et al. 2021a)	83.5	89.1	-	-	37.9	86.4
DPR-PAQ (Oğuz et al. 2021)	84.0	89.2	-	-	31.1	-
Condenser (Gao and Callan 2021a)	83.2	88.4	81.9	86.2	36.6	-
coCondenser (Gao and Callan 2021b)	84.3	89.0	83.2	87.3	38.2	-
ANCE-Tele (Sun et al. 2022)	84.9	89.7	83.4	87.3	39.1	-
ERNIE-Search (Lu et al. 2022)	85.3	89.7	-	-	40.1	-
AR2+SimANS (Zhou et al. 2022)	86.2	90.3	84.6	88.1	40.9	88.7
ColBERTv2 (Santhanam et al. 2021)	-	-	-	-	39.7	86.8
SimLM (Wang et al. 2022)	85.2	89.7	-	-	41.1	87.8
LexMAE-Stage1 (Shen et al. 2022)	-	-	-	-	39.3	-
LexMAE-Stage1 (Shen et al. 2022)	-	-	-	-	40.8	-
LexMAE (Shen et al. 2022)	-	-	-	-	42.6	-
ANCE (Xiong et al. 2020)	81.9	87.5	80.3	85.3	33.0	81.1
ANCE + TriSampler	**83.8**	**89.1**	**83.4**	**87.2**	**35.8**	**83.4**
RocketQA (Qu et al. 2020)	82.7	88.5	-	-	37.0	85.5
RocketQA + TriSampler	**85.3**	**89.6**	-	-	**38.3**	**86.0**
AR2 (Zhang et al. 2021)	86.0	90.1	84.4	87.9	39.5	87.8
AR2 + TriSampler	**86.5**	**90.7**	**85.0**	**88.5**	**41.4**	**89.1**

Table 2: Results on three retrieval benchmarks, including NQ test set, TQA test set, and MS Pas dev set. The results of baselines are directly obtained from the original papers and results not provided are marked as "-".

three stages: BM25 Negatives, Hard Negatives, Reranker-Distilled. While our TriSampler combined with AR2 exceeds the performance of LexMAE in its first two stages, it falls short in matching the effectiveness of LexMAE's complete three-stage protocol. This difference is primarily attributed to LexMAE's utilization of an advanced off-the-shelf reranker, a component that significantly enhances its overall performance. Experimental results in Table 2 show that TriSampler is a general method that can be naturally applied to various dense retrieval models. Such a method can provide more informative negatives to consistently improve downstream performance in dense retrieval.

Method	MRR@100	R@100
BM25 (Yang, Fang, and Lin 2017)	27.9	80.7
DPR (Karpukhin et al. 2020)	32.0	86.4
ANCE (Xiong et al. 2020)	37.7	89.4
STAR (Zhan et al. 2021)	39.0	91.3
ADORE (Zhan et al. 2021)	40.5	91.9
AR2 (Zhang et al. 2021)	41.8	91.4
AR2+SimANS (Zhou et al. 2022)	43.1	92.3
AR2+TriSampler	**43.8**	**93.1**

Table 3: Experimental performance on MS Doc dev set.

Why TriSampler Performs Better?

Perspective of candidates. To deepen the understanding of TriSampler, we vary the selection methods of negative candidates and conduct two extended experiments on the NQ dataset and the MS Pas dataset using the AR2 retrieval model: (1) top-k query-document ranked negative candidates $\mathcal{D}_q^- = \mathbf{TopK}_{s(q,\mathcal{D}^-)}$; (2) top-k document-document ranked negative candidates $\mathcal{D}_q^- = \mathbf{TopK}_{s(d^+,\mathcal{D}^-)}$.

As shown in Table 4, TriSampler surpasses all other variants of negative candidate selection methods, indicating the effectiveness of TriSampler. For $\mathbf{TopK}_{s(q,\mathcal{D}^-)}$ and $\mathbf{TopK}_{s(d^+,\mathcal{D}^-)}$, they seem to only account for the impact of the query or positive document on negatives, ignoring the triangular-like relationship outlined in Section . TriSampler combines these two methods based on *the quasi-triangular principle*, which alleviates the excessive reliance on the query and constrains the region of negative candidates. Consequently, TriSampler can achieve enhanced performance, suggesting that the triangular-like relationship is a valuable constraint for selecting negative candidates.

Perspective of distributions. To demonstrate the effectiveness of the negative sampling distribution proposed in TriSampler, we evaluate the retrieval performances on three

Method	NQ		MS Pas	
	R@20	R@100	MRR@10	R@50
TopK$_{s(q,\mathcal{D}^-)}$	86.2	90.3	40.9	88.7
TopK$_{s(d^+,\mathcal{D}^-)}$	85.5	90.4	40.3	88.5
TriSampler	86.5	90.7	41.4	89.1

Table 4: Various negative candidate selection methods on the NQ dataset and the MS Pas dataset.

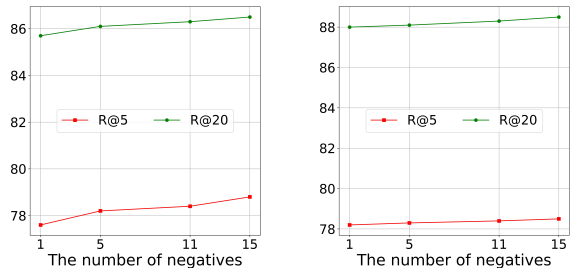
variations of TriSampler on the MS Pas dataset: (1) Uniform sampling that assigns negative candidates with equal weights; (2) TopK Sampling that leverages the relevant score as sampling weights; (3) Debiased Sampling that computes sampling weights by reducing the impact of the positive relevant score. Table 5 reveals that TriSampler outperforms the other variant negative sampling distributions. According to Equation (5), the negative sampling distribution suggested by TriSampler adheres to *the quasi-triangular principle*. This principle allocates higher sampling probabilities to negatives that are closer to the positive document within a restricted region. This observation confirms that a well-designed sampling distribution can indeed contribute to enhanced performance.

Method	MRR@10	R@50	R@1k
Uniform Sampling	39.7	87.9	98.6
TopK Sampling	40.6	88.6	98.7
Debiased Sampling	41.1	88.9	98.8
TriSampler	41.4	89.1	98.9

Table 5: Various negative sampling distributions on the MS Pas dataset.

Further Analysis

Impact of negative sample size. We further investigate the impact of negative sample size k on retrieval performance using the AR2 model. We vary k in the range of $\{1, 5, 11, 15\}$ and conduct experiments on the NQ and the TQA datasets. As depicted in Figure 3, retrieval performance consistently enhances with the increasing number of negatives, verifying the significance of negative sample size in improving performance. These experimental results align with findings from RocketQA, which also suggest that increasing the number of negatives contributes to better retrieval performance.



(a) NQ (b) TQA

Figure 3: The impact of negative sample size on the NQ dataset.

Training efficiency comparison. To explore the training efficiency of TriSampler, we test the wall-clock time cost, including the cost of training per batch Cost$_D$ and the cost of training instances construction Cost$_C$. As shown in Table 6, it is obviously observed that the training cost of TriSampler is slightly higher compared with SimANS. Although TriSampler requires more time to construct training instances, the cost is distributed across $t = 2000$ training steps, resulting in a per-batch cost of Cost$_{C/t} = 0.055s$. Thus, the overall cost for training each batch has increased only slightly. However, the total training time to reach optimal performance is reduced because TriSampler achieves faster convergency (See Figure 4). To sum up, TriSampler demonstrates improved efficiency gains in comparison to SimANS.

Method	Cost$_D$	Cost$_C$	Cost$_{C/t}$	Cost$_{all}$
AR2+SimANS	2.9s	85s	0.043s	2.943s
AR2+TriSampler	3.0s	110s	0.055s	3.055s

Table 6: Training efficiency comparison on the NQ dataset.

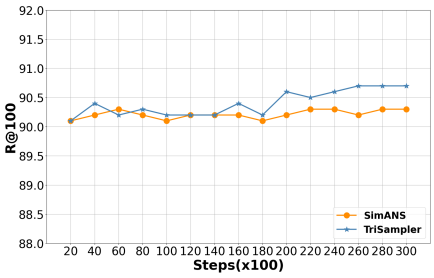


Figure 4: Training convergency curves comparison between SimANS and TriSampler on the NQ dataset.

Conclusion

In this paper, we investigate the fundamental principle that negative sampling should satisfy in dense retrieval. We first analyze negative sampling from the perspective of objective and subsequently propose a general principle to guide negative sampling, termed *the quasi-triangular principle*. This principle advocates for the confinement of sampled negatives within a region reminiscent of a triangle. Capitalizing on this principle, we propose a simple and effective negative sampling method TriSampler to sample more informative negatives within the designated constrained region. Experiments conducted across four benchmark datasets demonstrate the efficacy of TriSampler, showcasing its capacity to deliver superior retrieval performance when compared to alternative strategies.

Acknowledgments

This work is supported by the Technology and Innovation Major Project of the Ministry of Science and Technology of China under Grant 2022ZD0118600, Natural Science Foundation of China (NSFC) 62276148, Tsinghua University Initiative Scientific Research Program 20233080067, the New Cornerstone Science Foundation through the XPLORER PRIZE.

References

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chuang, C.-Y.; Robinson, J.; Lin, Y.-C.; Torralba, A.; and Jegelka, S. 2020. Debiased contrastive learning. *Advances in neural information processing systems*, 33: 8765–8775.

Dai, Z.; and Callan, J. 2019. Deeper text understanding for IR with contextual neural language modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 985–988.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

Gao, L.; and Callan, J. 2021a. Is your language model ready for dense representation fine-tuning. *arXiv preprint arXiv:2104.08253*.

Gao, L.; and Callan, J. 2021b. Unsupervised corpus aware language model pre-training for dense passage retrieval. *arXiv preprint arXiv:2108.05540*.

Gao, L.; Dai, Z.; Chen, T.; Fan, Z.; Van Durme, B.; and Callan, J. 2021. Complement lexical retrieval model with semantic residual embeddings. In *European Conference on Information Retrieval*, 146–160. Springer.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Hofstätter, S.; Lin, S.-C.; Yang, J.-H.; Lin, J.; and Hanbury, A. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 113–122.

Huang, J.-T.; Sharma, A.; Sun, S.; Xia, L.; Zhang, D.; Pronin, P.; Padmanabhan, J.; Ottaviano, G.; and Yang, L. 2020. Embedding-based retrieval in facebook search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2553–2561.

Humeau, S.; Shuster, K.; Lachaux, M.-A.; and Weston, J. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969*.

Johnson, J.; Douze, M.; and Jégou, H. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3): 535–547.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Kalantidis, Y.; Sariyildiz, M. B.; Pion, N.; Weinzaepfel, P.; and Larlus, D. 2020. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33: 21798–21809.

Karpukhin, V.; Oğuz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Khattab, O.; and Zaharia, M. 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*, 39–48.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7: 453–466.

Lee, K.; Chang, M.-W.; and Toutanova, K. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*.

Li, B.; Zhou, H.; He, J.; Wang, M.; Yang, Y.; and Li, L. 2020. On the sentence embeddings from pre-trained language models. *arXiv preprint arXiv:2011.05864*.

Lin, S.-C.; Yang, J.-H.; and Lin, J. 2020. Distilling dense representations for ranking using tightly-coupled teachers. *arXiv preprint arXiv:2010.11386*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lu, Y.; Liu, Y.; Liu, J.; Shi, Y.; Huang, Z.; Sun, S. F. Y.; Tian, H.; Wu, H.; Wang, S.; Yin, D.; et al. 2022. ERNIE-Search: Bridging Cross-Encoder with Dual-Encoder via Self On-the-fly Distillation for Dense Passage Retrieval. *arXiv preprint arXiv:2205.09153*.

Luan, Y.; Eisenstein, J.; Toutanova, K.; and Collins, M. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9: 329–345.

Mao, Y.; He, P.; Liu, X.; Shen, Y.; Gao, J.; Han, J.; and Chen, W. 2020. Generation-augmented retrieval for open-domain question answering. *arXiv preprint arXiv:2009.08553*.

Meng, Y.; Xiong, C.; Bajaj, P.; Bennett, P.; Han, J.; Song, X.; et al. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *Advances in Neural Information Processing Systems*, 34: 23102–23114.

Nguyen, T.; Rosenberg, M.; Song, X.; Gao, J.; Tiwary, S.; Majumder, R.; and Deng, L. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@ NIPs*.

Nogueira, R.; Lin, J.; and Epistemic, A. 2019. From doc2query to docTTTTTquery. *Online preprint*, 6.

Nogueira, R.; Yang, W.; Lin, J.; and Cho, K. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.

Oğuz, B.; Lakhotia, K.; Gupta, A.; Lewis, P.; Karpukhin, V.; Piktus, A.; Chen, X.; Riedel, S.; Yih, W.-t.; Gupta, S.; et al.

2021. Domain-matched pre-training tasks for dense retrieval. *arXiv preprint arXiv:2107.13602*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Qu, Y.; Ding, Y.; Liu, J.; Liu, K.; Ren, R.; Zhao, W. X.; Dong, D.; Wu, H.; and Wang, H. 2020. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2010.08191*.

Ren, R.; Lv, S.; Qu, Y.; Liu, J.; Zhao, W. X.; She, Q.; Wu, H.; Wang, H.; and Wen, J.-R. 2021a. PAIR: Leveraging passage-centric similarity relation for improving dense passage retrieval. *arXiv preprint arXiv:2108.06027*.

Ren, R.; Qu, Y.; Liu, J.; Zhao, W. X.; She, Q.; Wu, H.; Wang, H.; and Wen, J.-R. 2021b. Rocketqav2: A joint training method for dense passage retrieval and passage re-ranking. *arXiv preprint arXiv:2110.07367*.

Robinson, J.; Chuang, C.-Y.; Sra, S.; and Jegelka, S. 2020. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*.

Sachan, D. S.; Patwary, M.; Shoeybi, M.; Kant, N.; Ping, W.; Hamilton, W. L.; and Catanzaro, B. 2021. End-to-end training of neural retrievers for open-domain question answering. *arXiv preprint arXiv:2101.00408*.

Santhanam, K.; Khattab, O.; Saad-Falcon, J.; Potts, C.; and Zaharia, M. 2021. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *arXiv preprint arXiv:2112.01488*.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Shen, T.; Geng, X.; Tao, C.; Xu, C.; Huang, X.; Jiao, B.; Yang, L.; and Jiang, D. 2022. Lexmae: Lexicon-bottlenecked pretraining for large-scale retrieval. *arXiv preprint arXiv:2208.14754*.

Su, J.; Cao, J.; Liu, W.; and Ou, Y. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Sun, S.; Xiong, C.; Yu, Y.; Overwijk, A.; Liu, Z.; and Bao, J. 2022. Reduce Catastrophic Forgetting of Dense Retrieval Training with Teleportation Negatives. *arXiv preprint arXiv:2210.17167*.

Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Tian, H.; Wu, H.; and Wang, H. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8968–8975.

Wang, K.; Reimers, N.; and Gurevych, I. 2021. Tsdae: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning. *arXiv preprint arXiv:2104.06979*.

Wang, L.; Yang, N.; Huang, X.; Jiao, B.; Yang, L.; Jiang, D.; Majumder, R.; and Wei, F. 2022. Simlm: Pre-training with representation bottleneck for dense passage retrieval. *arXiv preprint arXiv:2207.02578*.

Xiong, L.; Xiong, C.; Li, Y.; Tang, K.-F.; Liu, J.; Bennett, P.; Ahmed, J.; and Overwijk, A. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Yang, P.; Fang, H.; and Lin, J. 2017. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, 1253–1256.

Yu, S.; Liu, Z.; Xiong, C.; Feng, T.; and Liu, Z. 2021. Few-shot conversational dense retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 829–838.

Zhan, J.; Mao, J.; Liu, Y.; Guo, J.; Zhang, M.; and Ma, S. 2021. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1503–1512.

Zhang, H.; Gong, Y.; Shen, Y.; Lv, J.; Duan, N.; and Chen, W. 2021. Adversarial retriever-ranker for dense text retrieval. *arXiv preprint arXiv:2110.03611*.

Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; and Liu, Q. 2019. ERNIE: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.

Zhou, K.; Gong, Y.; Liu, X.; Zhao, W. X.; Shen, Y.; Dong, A.; Lu, J.; Majumder, R.; Wen, J.-R.; Duan, N.; et al. 2022. Simans: Simple ambiguous negatives sampling for dense text retrieval. *arXiv preprint arXiv:2210.11773*.