

More Robust Dense Retrieval with Contrastive Dual Learning

Yizhi Li^{1*}, Zhenghao Liu^{1*}, Chenyan Xiong², and Zhiyuan Liu¹
Tsinghua University¹, Microsoft Research²

yizhi.li@hotmail.com; liuzhenghao0819@gmail.com; chenyan.xiong@microsoft.com; liuzy@tsinghua.edu.cn

ABSTRACT

Dense retrieval conducts text retrieval in the embedding space and has shown many advantages compared to sparse retrieval. Existing dense retrievers optimize representations of queries and documents with contrastive training and map them to the embedding space. The embedding space is optimized by aligning the matched query-document pairs and pushing the negative documents away from the query. However, in such training paradigm, the queries are only optimized to align to the documents and are coarsely positioned, leading to an anisotropic query embedding space. In this paper, we analyze the embedding space distributions and propose an effective training paradigm, Contrastive Dual Learning for Approximate Nearest Neighbor (DANCE) to learn fine-grained query representations for dense retrieval. DANCE incorporates an additional dual training object of query retrieval, inspired by the classic information retrieval training axiom, query likelihood. With contrastive learning, the dual training object of DANCE learns more tailored representations for queries and documents to keep the embedding space smooth and uniform, thriving on the ranking performance of DANCE on the MS MARCO document retrieval task. Different from ANCE that only optimized with the document retrieval task, DANCE concentrates the query embeddings closer to document representations while making the document distribution more discriminative. Such concentrated query embedding distribution assigns more uniform negative sampling probabilities to queries and helps to sufficiently optimize query representations in the query retrieval task. Our codes are released at <https://github.com/thunlp/DANCE>.

CCS CONCEPTS

• Information systems → Retrieval models and ranking.

KEYWORDS

Neu-IR; Dense Retrieval; Dual Learning; Contrastive Training

ACM Reference Format:

Yizhi Li^{1*}, Zhenghao Liu^{1*}, Chenyan Xiong², and Zhiyuan Liu¹. 2021. More Robust Dense Retrieval with Contrastive Dual Learning. In *Proceedings of the 2021 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '21)*, July 26, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3471158.3472245>

* indicates equal contribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICTIR '21, July 26, 2021, Virtual Event, Canada.

© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8611-1/21/07...\$15.00
<https://doi.org/10.1145/3471158.3472245>

1 INTRODUCTION

The recent dense retrieval provides an opportunity to conduct semantic matches and serves lots of applications, such as open domain question answering [2], conversational search [32, 45], and fact verification [34]. Instead of using discrete bag-of-word matching in document retrieval, dense retriever encodes queries and documents into a high-dimensional embedding space and conducts text matching in the embedding space, overcoming the vocabulary mismatch problem of sparse retrieval.

Existing dense retrieval models aim to learn effective representations of queries and documents and build an embedding space for the document retrieval task. With the support of approximate nearest neighbor (ANN) search [11, 15], dense retrievers can efficiently retrieve documents by conducting semantic matching in the embedding space. To encode queries and documents as dense representations, dense retrievers usually employ the BERT-Siamese architecture to provide fully learnable, well pretrained, and effective representations for queries and documents to construct embedding space for retrieval [16, 40, 41].

To optimize the embedding space for document retrieval, existing dense retrieval models usually sample negative documents and use them to contrastively train dense retrievers to learn query and document representations [16, 40, 41]. Wang et al. [36] prove that contrastive representation learning optimizes neural models and keeps two properties of the document embedding space, “alignment” and “uniformity”. The two properties, “alignment” and “uniformity”, help to keep a homogeneous and isotropic embedding space [16, 40, 41] – which is that “alignment” assigns similar embedding features to query and its related documents and conducts better clustering for similar document representations; “uniformity” encourages encoders to maintain maximal information for documents. It helps to form a uniform document embedding space and better classifies the confusable documents according to queries. The training object in existing dense retrieval methods [16, 40, 41] is to train dense retrievers on the document retrieval task and focuses on optimizing document embeddings. However, such document retrieval dominated training paradigm only optimizes the queries aligned to the relevant documents while discarding the uniformity nature in contrastive learning, leading to a non-smooth and non-uniform query embedding space. Thus, some work [46] solely optimizes the query representations by training with full retrieved documents instead of negative sampled documents and fixing the document representations.

In this paper, we enhance the query representations learning with a new training paradigm, Contrastive Dual Learning for Approximate Nearest Neighbor (DANCE). DANCE uses a model-level dual training object [38] to contrastively train BERT for optimizing the query and document embedding spaces. It consists of two optimization directions, document retrieval and query retrieval,

which are the main training task and dual training task, respectively. Query retrieval learns the query likelihood [18] to retrieve related queries for documents, and mainly focuses on optimizing query representations in consideration of the document representations near to them. Intuitively, DANCE can learn more appropriate query representations and better keep “alignment” and “uniformity” of the query embedding space, thus improving the robustness of the learned representation space. Besides, DANCE also normalizes the representations to avoid overfitting during training [3]. We map embeddings into a unit hyperspherical space and calculate the similarity between embeddings according to their angles.

Experiments on the MS MARCO [1] document retrieval task show the effectiveness of DANCE. DANCE mainly focuses on optimizing the query representations by training dense retrievers with the additional query retrieval task. It helps to build a smooth and uniform embedding space for text retrieval by pushing the query embeddings closer to document embeddings. Moreover, DANCE scatters the document distribution to make them more discriminative. Our analyses find that the queries are assigned with more balanced and uniform probabilities to be recalled in the contrastive training of query retrieval task, which sufficiently trains the query representations. For the documents recalled more easily in the contrastive training, they are usually distributed in a more concentrated area in the embedding space, which are confusable. DANCE shows better performance on these documents by learning more fine-grained query representations, which helps to distinguish the off-topic and unrelated documents during retrieval.

2 RELATED WORK

Based on whether term-level interactions are modeled between query and documents beyond their final encodings, Neural IR (Neu-IR) methods can be categorized into representation-based or interaction-based [10, 28].

Interaction-based models enjoy fine-grained modeling of term-level interactions between query and documents; thus they are typically more effective though more expensive and usually used as re-rankers since that requires scoring candidate documents according to the given query [7, 10, 14, 27, 29, 39]. Representation-based ones, often encode queries and documents as low-dimensional dense representations without explicit term-level matches. The representation-based models can achieve more efficient retrieval with the document representation precomputing and the support of approximate nearest neighbor (ANN) [9, 16, 17, 19, 26, 40]. The representation-based models help to achieve an efficient dense retrieval, benefiting many downstream tasks by providing more accurate evidence, such as fact verification, conversational dense retrieval, and open domain question answering [16, 22, 32, 41].

With the development of the deep neural network, pretrained language models, such as BERT [25] and RoBERTa [8], have been widely used in both representation-based and interaction-based Neu-IR models. Dense retrievers also employ the BERT-Siamese architecture to encode queries and documents as dense representations to conduct an embedding space for retrieval. To learn query and document representations, dense retrievers contrastively train BERT on the document retrieval task with related (positive) document and sampled unrelated (negative) documents for the given

queries. While some work [17] proposes a late query–document interaction paradigm to improve the retrieval effectiveness, many others focus on learning better representations with contrastive training. They propose various negative sampling methods during contrastive training, including in-batch negatives, BM25 retrieved negatives and random selected negatives [20, 40, 41]. Approximate Nearest Neighbor Negative Contrastive Learning (ANCE) [40] further comes up with a training method that asynchronously updates the document index and samples negatives from queries’ nearest areas in the document embedding space to avoid diminishing gradient norms during training [40]. Some work [31] also improves such training paradigm by alleviating the affect from incomplete relevant labels.

Recent work in computer vision further discusses how the embedding space is optimized with contrastive training [3]. The contrastive learning for dense retrievers optimizes the embedding space to satisfy two properties, “alignment” and “uniformity” [36], which aim to align matched query-document pairs and make the embedding space uniform by pushing the negative documents away from the given queries, respectively. Several work demonstrates that mapping the representations into a unit hyperspherical space, where all embeddings are represented as unit vectors, helps to keep a smooth embedding space and brings improvement for various tasks [3, 4, 24, 35, 36]. Chen et al. [3] introduce the normalized temperature-scaled cross entropy loss as the standard contrastive training loss. Different from the standard cross entropy loss, it uses the temperature scaling technique to adjust the sharpness of the softmax distribution, which plays an important role in optimizing the normalized embedding space [3]. Moreover, Chen et al. [4] further adjust the temperature and balance the influence of “alignment” and “uniformity” to learn a better embedding space with contrastive training.

In existing dense retrievers, the document embeddings are trained more sufficiently with the contrastive training of document retrieval task. Thus, some work fixes the learned document embeddings and focuses on optimizing the query representations. Such technology is leveraged to achieve better retrieval performance [46] or conduct an effective retriever in language modeling [12] and question answering system [21], demonstrating the importance of learning tailored query embeddings in dense retrievers.

To learn more fine-grained query representations, the query retrieval task provides an opportunity, which is inspired by the IR training axiom query likelihood [18]. Contrastively training dense retriever on the query retrieval task further optimizes the query embedding space more smooth and uniform, which is a dual task for the document retrieval task and can be incorporated in the model optimization with dual learning [13, 23, 37, 38, 43]. Dual learning is a method to train models of symmetric structures and has shown promising performance in some NLP tasks, such as machine translation and image caption [13, 33, 43]. Xia et al. [38] further categorize dual learning algorithms as task-level dual learning and model-level dual learning according to whether the prime and dual model share components. They illustrate that model-level dual learning performs better compared to task-level dual learning [38]. Query retrieval and document retrieval are in the symmetric training directions and recent dense retrievers [40] usually share the same

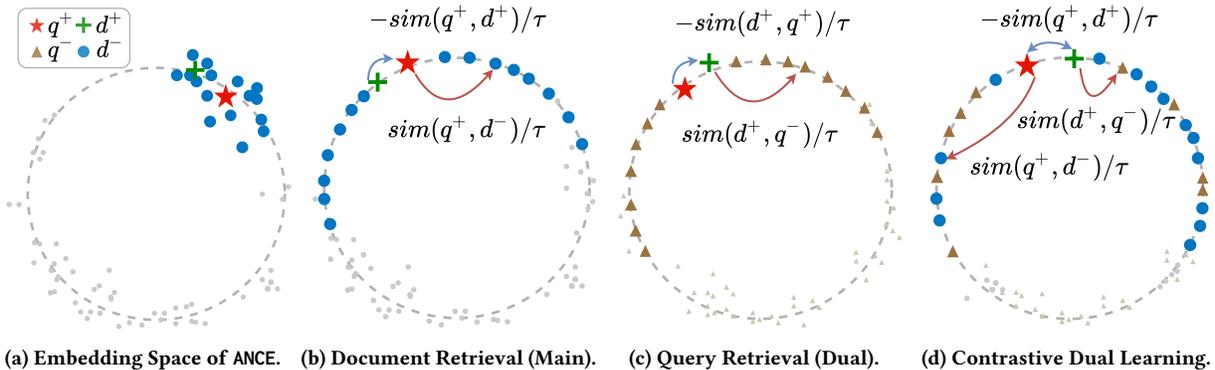


Figure 1: The Illustration of Contrastive Dual Learning for Approximate Nearest Neighbor (DANCE) Training Paradigm.

BERT encoder for queries and documents. Thus, for dense retrievers, training with query retrieval and document retrieval can be regarded as a model-level dual learning paradigm.

3 METHODOLOGY

This section describes our proposed training paradigm for dense retrieval, Contrastive Dual Learning for Approximate Nearest Neighbor (DANCE), as shown in Figure 1. We first introduce the preliminaries of dense retrieval (Sec. 3.1), and then theoretically analyze how the contrastive training optimizes the query and document representations (Sec. 3.2). Finally, we describe our contrastive dual learning mechanism in DANCE (Sec. 3.3).

3.1 Preliminary

Given a query q and a document collection $D = \{d_1, \dots, d_j, \dots, d_n\}$, dense retrievers calculate the ranking score $f(q, d)$ by learning the representations for documents similar to the intuition of document likelihood $p(d|q)$.

Dense retrievers [16, 40, 41] leverage the same BERT encoder to get the representations of queries and documents:

$$H(q) = \text{BERT}([\text{CLS}] \circ q \circ [\text{SEP}]); \quad (1)$$

$$H(d) = \text{BERT}([\text{CLS}] \circ d \circ [\text{SEP}]), \quad (2)$$

where \circ is the concatenation operation. “[CLS]” and “[SEP]” are special tokens in BERT. The representations of the first token “[CLS]” are used as representations of query q and document d , which are denoted as $H_0(q)$ and $H_0(d)$, respectively. Then the similarity score $f(q, d)$ of query q and document d can be calculated with their dense representations:

$$f(q, d) = \text{sim}(H_0(q), H_0(d)), \quad (3)$$

where $\text{sim}(\cdot)$ is the similarity function to estimate the relevance between two embeddings. The efficient calculation of $\text{sim}(\cdot)$ for large scale dataset in our method is provided by FAISS¹ and the dot product is usually used as the similarity function [16, 40, 41].

Then we can calculate the ranking loss L with both positive document d^+ and negative document d^- for the given query q to

contrastively train BERT:

$$L = \sum_q \sum_{d^+ \in D^+} l(q, d^+, D^-), \quad (4)$$

where D^+ is the positive document collection for the given query q that provided by annotation data. $l(q, d^+, D^-)$ is the contrastive training loss function, which is the same as the state-of-the-art dense retrievers [40]:

$$l(q, d^+, D^-) = -\log \frac{e^{f(q^+, d^+)}}{e^{f(q^+, d^+)} + \sum_{d^- \in D^-} e^{f(q^+, d^-)}}, \quad (5)$$

where D^- is the collection of negative documents for query q sampled with different retrieval methods, such as BM25 [16] and dense retriever itself [40].

3.2 Alignment and Uniformity

Dense retrievers are contrastively trained to encode queries and documents with sampled negative documents as well as keep beneficial properties of the embedding space. As shown in recent research [36], contrastive loss keeps “alignment” and “uniformity” of the document embedding space for retrieval. Such optimization progress is visualized in Figure 1a and Figure 1b.

Following previous work [3], we normalize the embeddings with L2 norm to keep representations of queries and documents in the unit hyperspherical space and calculate the similarity between the query q and document d :

$$f_{\text{norm}}(q, d) = \text{sim}(\|H_0(q)\|_2, \|H_0(d)\|_2), \quad (6)$$

where $-1 \leq f_{\text{norm}}(q, d) \leq 1$ and $\|\cdot\|_2$ is the L2 normalization operation. The similarity function $f_{\text{norm}}(q, d)$ only focuses on estimating the relevance between query q and document d according to the angle of their representations. Then we use the normalized temperature-scaled cross entropy [3] to replace the regular cross entropy function in Eq. 5:

$$L_{\text{norm}} = -\log \frac{e^{f_{\text{norm}}(q, d^+)/\tau}}{e^{f_{\text{norm}}(q, d^+)/\tau} + \sum_{d^- \in D_{\text{ANN}}^-} e^{f_{\text{norm}}(q, d^-)/\tau}}, \quad (7)$$

where D_{ANN}^- is the negative document collection. These documents are sampled from the ones distributed near to the query q in the

¹<https://github.com/facebookresearch/faiss>

embedding space, which is the same as previous work [40]. τ is the temperature hyperparameter used to control the sharpness of the softmax distribution.

Following Wang et al. [36], the normalized temperature-scaled loss can be transformed as:

$$L_{\text{norm}} = \underbrace{-f_{\text{norm}}(q, d^+)/\tau}_{\text{alignment}} + \underbrace{\log(e^{f_{\text{norm}}(q, d^+)/\tau} + \sum_{d^- \in D_{\text{ANN}}^-} e^{f_{\text{norm}}(q, d^-)/\tau})}_{\text{uniformity}}, \quad (8)$$

where the loss function encourages the model to optimize embedding distributions and keeps two important properties of the embedding space:

- **Alignment:** the query q is distributed closer to the document d than the negative document d^- in the embedding space;
- **Uniformity:** the embedding distribution is encouraged to be uniform on the hyperspherical space by pushing the negative documents d^- away from the given query q .

During contrastive training, dense retriever keeps alignment and uniformity of the embedding space, which helps to learn discriminative representations for documents.

3.3 Contrastive Dual Learning

To further learn a uniform and smooth embedding space for queries, we train dense retrievers with a model-level dual learning task [38], which optimizes dense retrievers with query retrieval task and document retrieval task. They are inspired by query likelihood $P(q|d)$ and document likelihood $P(d|q)$.

Query likelihood and document likelihood learning are symmetric. And the query likelihood $P(q|d)$ can also be approximated to the relevance between query and document:

$$P(d|q) = P(q|d) \cdot P(d), \quad (9)$$

where the document distribution $P(d)$ can be regard uniformly. To learn the query likelihood with contrastive training, we use a dual loss function L_{dual} :

$$L_{\text{dual}} = -\log \frac{e^{f_{\text{norm}}(d, q^+)/\tau}}{e^{f_{\text{norm}}(d, q^+)/\tau} + \sum_{q^- \in Q_{\text{ANN}}^-} e^{f_{\text{norm}}(d, q^-)/\tau}}, \quad (10)$$

where Q_{ANN}^- is the negative query collection and sampled from the documents distributed near to the document d in the embedding space. Similar to Eq. 8, the contrastive learning for query likelihood also keeps alignment and uniformity for the query embedding space:

$$L_{\text{dual}} = \underbrace{-f_{\text{norm}}(d, q^+)/\tau}_{\text{alignment}} + \underbrace{\log(e^{f_{\text{norm}}(d, q^+)/\tau} + \sum_{q^- \in Q_{\text{ANN}}^-} e^{f_{\text{norm}}(d, q^-)/\tau})}_{\text{uniformity}}, \quad (11)$$

Finally, we add the prime training loss L_{norm} and the dual training loss L_{dual} to conduct our contrastive dual learning loss and train dense retrievers:

$$L_{\text{final}} = L_{\text{norm}} + \lambda L_{\text{dual}}, \quad (12)$$

where λ is used to weight the training loss of the dual task.

4 EXPERIMENTAL METHODOLOGY

This section describes the dataset, evaluation metrics, baselines, and experimental details of our implementation.

Dataset. In all experiments, we use MS MARCO [1] to evaluate model performance. The dataset consists of massive anonymized questions sampled from Bing’s search query logs and texts extracted from 3,563,535 Bing retrieved web pages. For each query, the documents that have at least one related passage are recognized as relevant ones. We focus on the full ranking setting of the document retrieval task and retrieve 100 documents from the whole collection of 3,213,835 documents for each query to directly evaluate the retrieval performance. We keep the same official data partitions, the training set, development set and evaluation set contain 367,013 queries, 5,193 queries and 5,793 queries, respectively.

Evaluation Metrics. Following the official evaluation of MS MARCO [1], we use evaluation metrics NDCG@10 and MRR@100 in our experiments, where MRR@100 is regarded as our main evaluation. The evaluation of MS MARCO document ranking is a black-box testing. All the results of submitted runs are available on the leaderboard². Statistic significance is tested by permutation test with $p < 0.05$.

Baselines. Our baselines include two kinds of models: sparse retrieval models and dense retrieval models. The state-of-the-art dense retriever ANCE (FirstP) is regarded as the main baseline in our experiments.

The sparse retrieval baselines include docT5query [30, 42] and HDCT [6], both of which inherit the discrete bag-of-word matches of the classical information retrieval method BM25 and focus on improving the document representations to achieve better retrieval performance for the sparse retriever. Different from vanilla BM25, docT5query [30] improves document representations in sparse retrieval indexing by expanding documents with generated queries. HDCT uses pretrained language model, BERT [8] to predict the term weights in passages and then improves document bag-of-words representations by combining the passage term weights.

Other baselines are dense retrievers. They encode both queries and documents as dense representations and form an embedding space to conduct effective document retrieval [16, 26, 40] with ANN search tools like ScaNN³ [11] and FAISS [15]. DPR first leverages contrastive training methods to train BERT [8] as the encoder for queries and documents. It proposes several negative document sampling methods, such as in-batch negatives, BM25 negatives and random negatives. DPR w. BM25-Rand Neg is compared in our experiments, which achieves the best retrieval performance among variants of DPR. To better train dense retrievers, ANCE [40] uses RoBERTa [25] to learn representations of queries and documents and then samples negatives from the documents located near to the queries in the embedding space. Such training method avoids the

²<https://microsoft.github.io/msmarco/>

³<https://github.com/google-research/google-research/tree/master/scann>

Table 1: Overall Performance. All models are evaluated on the MS MARCO document retrieval task. Superscript † indicates statistically significant improvement over ANCE (FirstP)†.

Model	Dev		Eval
	NDCG@10	MRR@100	MRR@100
docT5query [30, 42]	-	0.327	0.291
HDCT [6]	-	0.300	-
DE-BERT [26]	-	0.288	-
ME-BERT [26]	-	0.330	-
DE-HYBRID [26]	-	0.313	0.287
ME-HYBRID [26]	-	0.346	0.310
DPR w. BM25-Rand Neg [16, 26]	0.362	0.312	-
ANCE (FirstP) [40]	0.437	0.373	0.334
DANCE (FirstP)	0.447 [†]	0.383 [†]	0.341

diminishing gradients during training and helps to achieve competitive retrieval performance. Besides, two dense retrieval models, DE-BERT and ME-BERT, from the previous work [26] are compared in our experiments. DE-BERT is similar to DPR, and it uses the “[CLS]” hidden states to represent queries and documents. Different from DE-BERT, ME-BERT represents documents with multiple embeddings of the tokens of different positions learned by pretrained language models. The two models DE-HYBRID and ME-HYBRID linearly incorporate retrieval scores from sparse retriever to DE-BERT and ME-BERT with learnable weights, which are also compared.

Implementation Details. In our experiments, we follow our main baseline ANCE [40] and contrastively train DANCE. We globally sample negatives from the whole collection of queries and documents to train DANCE and asynchronously update ANN indexes with the latest saved checkpoint.

We conduct a two-stage training progress to train DANCE on the full-ranking document retrieval task of MS MARCO, including the embedding normalization stage and the dual training stage. In the embedding normalization stage, we start with the checkpoint of well-trained dense retriever ANCE⁴, which uses RoBERTa [25] as encoder. Then we map the representations of queries and documents into a hyperspherical embedding space and normalize them with L2 normalization. To tune the sharpness of softmax distribution and better train dense retrievers, we replace the standard cross entropy loss with the normalized temperature-scaled loss and set the temperature τ as 0.01. Then we leverage the dual training paradigm to train DANCE, where the additional dual task (query retrieval task) is incorporated and the dual training loss weight λ is set to 0.1. The two models for ablation study, ANCE w. Norm and ANCE w. Dual, trained with individual steps are also evaluated in our experiments.

Our experiments mainly focus on the FirstP setting proposed in previous work [5] to evaluate retrieval effectiveness. In FirstP setting, queries and documents are truncated and padded to the sequences with the maximum lengths of 64 and 512, respectively. DANCE is optimized with LAMB [44] optimizer with warming up step of 3000 and learning rate of 5e-6. The training batch sizes are set to 4 and 210 for training and inference, respectively. The gradient accumulation step is set to 2. For other experiments, we keep the same with ANCE [40] and use the IndexFlatIP in the toolkit

⁴<https://github.com/microsoft/ANCE/>

Table 2: The Ranking Performance of Ablation Models of DANCE on MS MARCO Document Retrieval Task. Superscripts †, ‡, § indicate statistically significant improvements over ANCE (FirstP)†, ANCE w. Norm (FirstP)‡, and ANCE w. Dual§, respectively.

Model	Dev	
	NDCG@10	MRR@100
ANCE (FirstP)	0.437	0.373
ANCE w. Norm (FirstP)	0.443 [†]	0.380 [†]
ANCE w. Dual (FirstP)	0.444 [†]	0.381 [†]
DANCE (FirstP)	0.447^{†‡}	0.383^{†‡}

FAISS [15] to build the index for query and document embeddings during retrieval. Our final model is trained with 110k steps in the normalization step and 80k in the contrastive dual training. We train our models with 8 GeForce RTX 2080 Ti GPUs of 11GB with half-precision setting and the inference program runs on 4 same GPUs.

5 EVALUATION RESULT

In this section, we present six groups of experiments on the overall performance of DANCE, the embedding distributions of queries and documents learned by DANCE, the retrieval effectiveness in different testing scenarios and case studies.

5.1 Overall Performance

The performance of DANCE (FirstP) and baselines on MS MARCO document retrieval task is shown in Table 1.

DANCE (FirstP) outperforms the sparse retrievers docT5query and HDCT, showing the effectiveness of well-trained dense retrieval models by conducting semantic matches in document retrieval. Among all the dense retrievers, both DANCE (FirstP) and ANCE (FirstP) show much better performance by choosing more valuable negative documents to contrastively train dense retrievers. Benefited by the dual training paradigm, DANCE (FirstP) further improves ANCE (FirstP) with 0.7% MRR@100 score. The improvement demonstrates that the additional query retrieval task can help dense retrievers learn more tailored representations of queries and documents.

5.2 Ablation Study

In this part, we conduct ablation studies to further explore the roles of different components in DANCE on MS MARCO document retrieval task, as shown in Table 2.

The two individual modules of DANCE, hyperspherical normalization (Norm) and contrastive dual learning (Dual), are evaluated in this experiment, which are two optimization strategies used in DANCE to train dense retrievers. Norm and Dual focus on optimizing the embedding space from different aspects. The Norm module normalizes the embedding space and maps query and document representations into a unit hyperspherical space with L2 normalization, making the similarity calculation mainly focusing on the angle



Figure 2: Document Retrieval Performance of ANCE w. Dual and DANCE during Training. The model performance is evaluated on the development set.

between two vectors. And the Dual module incorporates the additional training object, query retrieval task, in the training process and mainly optimizes query embedding space for text retrieval.

We first evaluate the retrieval performance of two individual modules of DANCE, Norm and Dual, on the document retrieval task. Compared with baseline ANCE (FirstP), ANCE w. Norm (FirstP) and ANCE w. Dual (FirstP) achieve about 0.7% and 0.8% improvements of MRR@100 score on the development set, demonstrating their effectiveness on learning a more tailored embedding space for text retrieval from different aspects. With both individual modules incorporated into the dense retriever, the performance of DANCE (FirstP) is further improved and gets the best ranking performance among all models.

In our experiments, we find that Norm can alleviate the overfitting problem during training. As shown in Figure 2, DANCE shows a more stable performance compared with ANCE w. Dual in the training process. Some dense retrieval models also face the unstable training problem, which encourages them to use BM25 negatives to warm up training [40] and carefully optimize the embedding space locally with lots of tricks [20]. The Norm module conducts more stable training progress and may shed some light to deal with these problems. To evaluate the effectiveness of contrastive dual training, we regard the ANCE w. Norm (FirstP) model as our main baseline in the following experiments.

5.3 Learned Embedding Space of DANCE

This set of experiments further explores the embedding distributions of queries and documents learned by DANCE.

Pairwise Distance of Learned Embeddings. As shown in Table 3, we first evaluate the mean and variance of the cosine distances between query-query pairs, document-document pairs, and query-document pairs in the hyperspherical embedding spaces learned by ANCE w. Norm and DANCE. We use all queries from both training and development sets in this experiment.

By evaluating the change of pairwise distances before and after adding the Dual module, we find that DANCE shows a statistical difference of embedding distributions of ANCE w. Norm and DANCE. First, compared with ANCE w. Norm, the average distance of document-document pairs becomes larger in the embedding space learned by DANCE. It shows the document embeddings learned by DANCE are more scattered, benefiting the document retrieval

Table 3: The Statistics of Embedding Distance on MS MARCO Document Retrieval Task. The mean value and variance value of embedding distances between query-query pairs, document-document pairs and query-document pairs in the embedding space are calculated. The mean value represents the density of embedding distribution. The variance value represents the uniformity and smooth of the embedding space.

Distance Pair	Mean Distance		Variance Change
	ANCE w. Norm	DANCE	
Doc-Doc	0.179	0.191	- 1.22e-5
Que-Que	0.299	0.281	- 5.00e-4
Que-Doc	0.251	0.243	- 3.71e-4

in the embedding space. Meanwhile, the mean distance of query-document pairs shows a contrary trend. DANCE reduces it and generally concentrates the query embeddings closer to the document embedding population, apparently leading to a smaller mean distance of query-query pairs. As shown in the next experiment, such document embedding distribution derives from the concentrated query embedding distribution and the contrastive training on document retrieval task.

For all three kinds of pairwise distances, the consistently reducing variance value in DANCE further manifests the learned embedding spaces of queries and documents are more uniform and smooth, which is one of the sources of the effectiveness of DANCE.

Document Embedding Visualization. In this experiment, we visualize the document embedding space via t-SNE in Figure 3.

We choose one matched query-document pair from the development set of MS MARCO document retrieval dataset and select some documents to plot the embedding space. These documents are chosen from top-ranked documents of the corresponding model, top-retrieved documents from BM25, and random sampling. The embedding spaces of three models, ANCE, ANCE w. Norm and DANCE, are shown in Figure 3a, 3b, and 3c, respectively.

DANCE shows its ability to position queries and documents appropriately in the embedding space when training with the contrastive dual training paradigm. First, compared with ANCE w. Norm, DANCE can better align the query-document pairs by pulling the query embedding closer to the related document. Meanwhile, the surrounding unrelated documents are pushed away during contrastive training on the document retrieval task, making them more scattered and discriminative. Such document embedding distribution intuitively keeps the “uniformity” of embedding space and helps to distinguish confusable documents during retrieval.

Besides, different with vanilla ANCE, ANCE w. Norm only normalizes the representations of queries and documents and restricts them in a hyperspherical space. As expected, ANCE w. Norm concentrates the embedding space and shares almost the same embedding distribution of ANCE.

Query Embedding Distribution. In the rest of this set of experiments, we select the same relevant query-document pair as previous experiments and retrieve top-ranked negative queries with corresponding models to visualize the query embedding distributions of ANCE w. Norm and DANCE.

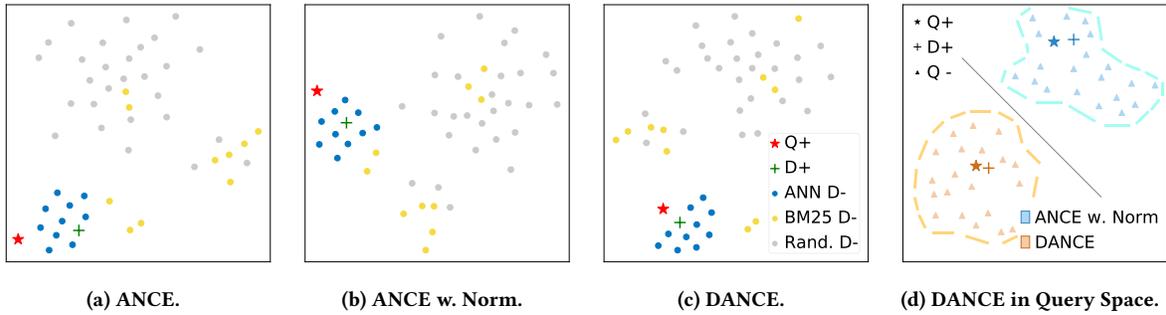


Figure 3: Embedding Space Distributions of Different Dense Retrieval Models. In Figure 3a, Figure 3b and Figure 3c, we choose the same case from the development set to visualize the embedding distribution of ANCE, ANCE w. Norm and DANCE, respectively. For the three figures on the left, visualized embeddings consist of a related query-document pair, top-10 negative documents retrieved by a dense retriever, top-10 negative documents retrieved by BM25 model and 30 documents that are randomly sampled from the whole document collection. In Figure 3d, to visualize the query embedding space, a related query-document pair, top-20 negative queries retrieved by ANCE w. Norm and DANCE are plotted.

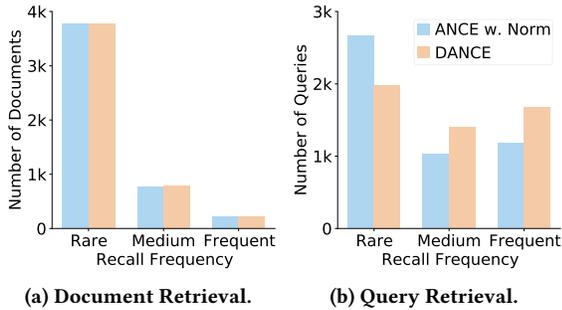


Figure 4: Number Distributions of Queries and Documents with Different Recall Frequency. All documents and queries in the development set are divided into three groups according to the recall frequency, which include rare, medium and frequent.

As shown in Figure 3d, DANCE forms a more smooth and uniform query embedding space by better keeping the “alignment” and “uniformity”. First, same as our previous observation, DANCE positions the related query closer to the document in the learned embedding space, showing better “alignment” between matched query-document pairs. Second, DANCE learns a more homogeneous query embedding distribution derives from contrastive training on the additional query retrieval task, which helps to better maintain the “uniformity” of the query embedding space.

5.4 Embedding with Different Recall Frequency

In this subsection, we further explore how the proposed contrastive dual training paradigm optimizes the embedding distribution. We conduct two experiments to study the recall frequencies of queries and documents during contrastive training and the change of the embedding distributions with different recall frequencies.

To estimate the probabilities of queries and documents that are sampled as negative ones in contrastive training, we use the recall

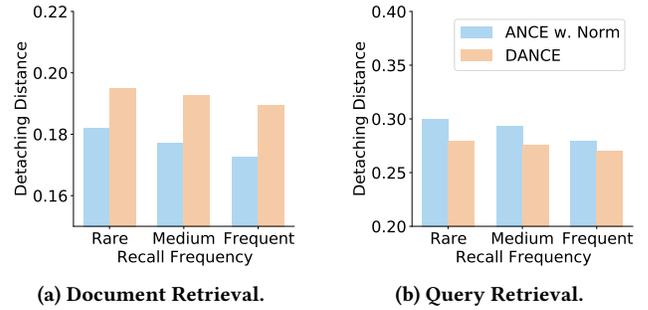


Figure 5: Detaching Distances of Queries and Documents with Different Recall Frequency. The recall times of queries and documents are calculated by query retrieval task and document retrieval task. All queries and documents are divided into three groups according to the recall frequency, which include rare, medium and frequent.

frequency for approximation. It calculates the times of queries and documents that appear in the top-100 retrieved candidates in the corresponding retrieval tasks, query retrieval, and document retrieval. All queries and documents in the development set are used in our experiments and divided into three groups according to the recall frequency, which includes rare (recalled once), medium (recalled twice), and frequent (recalled more than twice).

In the first experiment, we plot the number distributions of documents and queries along with different recall frequencies in Figure 4. The main difference between ANCE w. Norm and DANCE is on the query recall frequency distribution. As shown in Figure 4b, DANCE assigns more uniform recall frequencies to queries, which balances the sampling probabilities of queries during contrastive training on the query retrieval task. Such sampling mechanism optimizes query representations more sufficiently, especially for those long-tailed queries. Next, we further study how the recall frequency influences embedding distributions.

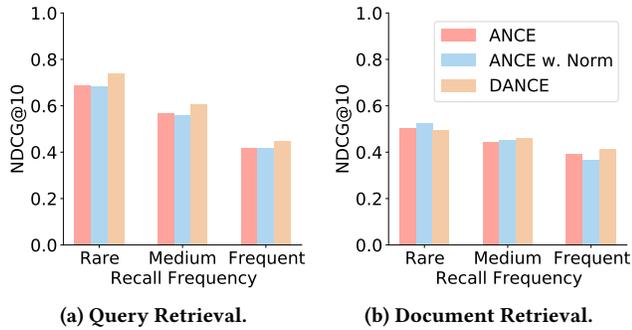


Figure 6: Ranking Performance of Queries and Documents with Different Recall Frequency. The results of query retrieval task and document retrieval task are shown in Figure 6a and Figure 6b, respectively. All queries and documents are divided into three groups according to the recall frequency, which includes rare, medium and frequent.

In the second experiment, we first introduce the detaching distance to serve for embedding distribution analysis. The detaching distances of queries and documents are calculated by the mean distances between one query or document and others in the group of query or document. The queries or documents with larger detaching distances indicate that they are located in a more scattered area and far from the query or document cluster.

Then we plot the detaching distances of queries and documents with different recall frequencies in Figure 5. Overall, the queries and documents with higher recall frequency usually have smaller detaching distances. The main reason is that the semantic meanings behind these representations are more confusable, leading to similar representations, concentrated embedding distributions and higher probabilities to be recalled in the contrastive training. Compared with ANCE w. Norm, DANCE achieves larger detaching distances of document pairs and smaller detaching distances of query pairs. It again demonstrates that DANCE learns more scattered and discriminative document embeddings and concentrates the query embeddings in the embedding space, which are observed in previous experiment (Sec. 5.3). Then we also conduct following experiments to study how the embedding distribution learned by DANCE affects the ranking effectiveness.

5.5 Effectiveness in Different Scenarios

We conduct these experiments to evaluate the ranking performance of DANCE in two testing scenarios, recall frequency and detaching distance. The NDCG@10 scores on both query retrieval task and document retrieval task are shown. Three models, ANCE, ANCE w. Norm and DANCE are evaluated in our experiments.

We first study the ranking performance with different recall frequencies, as shown in Figure 6. As expect, DANCE shows consistent improvements in the query retrieval task over ANCE and ANCE w. Norm with different recall frequencies, which derives from training with the additional training object, query retrieval task. In the document retrieval task, DANCE improves the ranking effectiveness on the queries that are frequently recalled and maintains a comparable performance on the queries of rare and medium frequencies. The

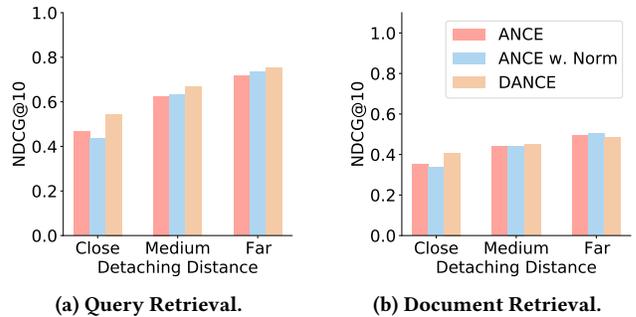


Figure 7: Ranking Performance of Queries and Documents with Different Detaching Distances. The ranking results on query retrieval task and document retrieval task are shown in Figure 7a and Figure 7b, respectively. All queries and documents are divided into three groups, close, medium and far, according to the detaching distances. These three groups have almost same numbers of queries or documents. The queries or documents with larger detaching distances are farther away from the query or document group.

improvement on frequent queries manifests that DANCE can learn more fine-grained representations for these queries and position them more appropriately in the embedding space. The main reason is that DANCE assigns more balanced optimization frequencies for queries during training with the query retrieval task, which is revealed in the previous experiment (Sec. 5.4).

In Figure 7, we evaluate retrieval effectiveness with different detaching distances. All three dense retrievers achieve better performance on both query and document retrieval tasks when the embedding distributions of queries and documents move towards detaching direction in the space. These queries and documents with larger detaching distances are more discriminative and usually positioned in the scattered area of the embedding space, which derives from the contrastive learning that pushes the negative queries and documents away from the related documents and queries, respectively. It is noteworthy that DANCE improves the query retrieval performance of the documents with closer distances by most, as shown in Figure 7a. Even though these documents are more confusable, DANCE shows its effectiveness in learning fine-grained query representations and conducting a more effective query embedding space by aligning the related query-document pairs and pushing the unrelated queries away from the given document.

5.6 Case Study

Finally, we show three cases that are selected from the development set of MS MARCO in Table 4 to analyze the ranking effectiveness.

The dense retrievers, such as ANCE, indeed show their effectiveness in conducting semantic matches and dealing with the vocabulary mismatch problem. As shown in the second case, the given query asks the safety of the place where the “chernobyl” accident happened. ANCE ranks the confusing document describing the general background of “chernobyl accident” at a top rank. Such a document is semantically related but is off-topic to the given query. In the other cases, the given queries ask about information of “Willie

Table 4: Case Studies. Three cases are selected from the development set of MS MARCO to qualitatively analyze the ranking effectiveness of DANCE.

Model	DANCE	ANCE
Query	1101870: willie mays worth	
Doc. Retrieval	relevant doc rank 1, recall frequency Medium	relevant doc rank 8, recall frequency Medium
Que. Retrieval	relevant query rank 2, recall frequency Medium	relevant query rank 6, recall frequency Rare
ID, Rank	D1658662, 1	D1411505, 1
Title	Who is Willie Mays? Biography, gossip, facts?	Mo Williams Maurice Williams Jr
Snippet	... Advertisement Willie Howard Mays Jr. (born May 6 1931) is a retired American professional baseball player who spent the majority of his major league career with the New York and San Francisco Giants before finishing with the New York Mets...	... Williams (born December 19, 1982) is an American professional basketball player who currently plays for the Portland Trail Blazers of the National Basketball Association (NBA) ...
Query	414714: is it still dangerous to go to the place of where the chernobyl happened	
Doc. Retrieval	relevant doc rank 2, recall frequency Rare	relevant doc rank 6, recall frequency Rare
Que. Retrieval	relevant query rank 1, recall frequency Medium	relevant query rank 2, recall frequency Rare
ID, Rank	D1761818, 2	D2092055, 2
Title	How radiation-safe are short-term trips to the Chernobyl Zone?	.
Snippet	... Up to now in the Zone there are places with considerably elevated and perhaps even deadly radiation. A prolonged, careless stay at such places can lead to radiation injuries of the body and, perhaps, even to chronic radiation sickness ...	Preface: The Chernobyl Accident On 26 April 1986, the most serious accident in the history of the nuclear industry occurred at Unit 4 of the Chernobyl nuclear power plant in the former Ukrainian Republic of the Soviet Union ...
Query	764139: what is ladder move	
Doc. Retrieval	relevant doc rank 3, recall frequency Rare	relevant doc rank 10, recall frequency Rare
Que. Retrieval	relevant query rank 2, recall frequency Medium	relevant query rank 20, recall frequency Rare
ID, Rank	D3220966, 3	D3268840, 3
Title	Your next career move: The ladder or the lattice?	What Is a Lateral Move in Reference to Employment?
Snippet	... Up is not the only way forward Climbing the ladder is the traditional model for career growth, taking a single pathway upward through the corporate hierarchy A lateral move in employment is when an employee transfers to a different department in the same company or to a different company – without any significant change in his salary ...

May” and “ladder move”, but ANCE ranks documents that are about “Maurice Williams” and “lateral move” with higher ranks. These documents are unrelated and off-topic to the given queries, but ANCE shows less effectiveness to distinguish the actually matched documents from such confusable documents, making the performance of dense retriever worse than matching with discrete bag-of-words.

Different from ANCE, DANCE incorporates an additional query retrieval task training object, which mainly focuses on learning query representations and optimizing the embedding space. Under the contrastive training, the query retrieval task learns query likelihood and keeps the “uniformity” and “alignment” in the embedding space. It benefits all three cases and helps to achieve better ranking performance on both query retrieval task and document retrieval task compared with ANCE. To sufficiently train the query representations, DANCE increases the recall frequency of these queries in the query retrieval task, making these queries have more probabilities to be sampled during the contrastive training. Our case studies show that DANCE can assign the matched documents with top ranks, manifesting DANCE can better “align” the matched query-document pairs from confusable document clusters by learning more fine-grained query representations.

6 CONCLUSION

In this paper, we propose a training paradigm, Contrastive Dual Learning for Approximate Nearest Neighbor (DANCE), to train dense retrievers. DANCE introduces the additional training object, query

likelihood, in dense retriever training to learn query and document representations. Different from ANCE, DANCE concentrates the query embeddings and assigns more uniform recall frequency to queries to sufficiently optimize their representations, while the document embedding distribution is optimized to be more scattered and discriminative. Through such embedding space optimization, DANCE achieves better ranking performance than the previous state-of-the-art dense retriever ANCE, especially for the documents that are more frequently recalled during contrastive training on the document retrieval task. Even these documents have smaller detaching distances and are hard to distinguish, DANCE shows better performance on them by learning more fine-grained query representations, better aligning related query-document pairs, and forming a more uniform and smooth embedding space for retrieval tasks. The observations of our work provide some possible directions to further improve dense retriever effectiveness and sufficiently optimize the embedding space by learning more effective query and document representations during contrastive training.

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China (No. 2020AAA0106501) and Beijing Academy of Artificial Intelligence (BAAI).

REFERENCES

- [1] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. MS MARCO: A Human Generated Machine Reading Comprehension Dataset. *CoRR* abs/1611.09268 (2016). arXiv:1611.09268
- [2] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of ACL*. 1870–1879.
- [3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of ICML*. 1597–1607.
- [4] Ting Chen and Lala Li. 2020. Intriguing Properties of Contrastive Losses. *CoRR* (2020). arXiv:2011.02803
- [5] Zhuyun Dai and Jamie Callan. 2019. Deeper Text Understanding for IR with Contextual Neural Language Modeling. In *Proceedings of SIGIR*. 985–988.
- [6] Zhuyun Dai and Jamie Callan. 2020. Context-Aware Document Term Weighting for Ad-Hoc Search. In *Proceedings of WWW*. 1897–1907.
- [7] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. 2018. Convolutional Neural Networks for Soft-Matching N-Grams in Ad-hoc Search. In *Proceedings of WSDM*. 126–134.
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.
- [9] Luyu Gao, Zhuyun Dai, Zhen Fan, and Jamie Callan. 2020. Complementing Lexical Retrieval with Semantic Residual Embedding. *CoRR* abs/2004.13969 (2020). arXiv:2004.13969
- [10] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A Deep Relevance Matching Model for Ad-hoc Retrieval. In *Proceedings of CIKM*. 55–64.
- [11] Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. 2020. Accelerating large-scale inference with anisotropic vector quantization. In *Proceedings of ICML*. 3887–3896.
- [12] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Paspupat, and Mingwei Chang. 2020. Retrieval Augmented Language Model Pre-Training. In *Proceedings of ICML*. 3929–3938.
- [13] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual Learning for Machine Translation. In *Proceedings of NIPS*. 820–828.
- [14] Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. 2017. PACRR: A Position-Aware Neural IR Model for Relevance Matching. In *Proceedings of EMNLP*. 1049–1058.
- [15] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* (2019).
- [16] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of EMNLP*. 6769–6781.
- [17] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of SIGIR*. 39–48.
- [18] Victor Lavrenko and W. Bruce Croft. 2017. Relevance-Based Language Models. *SIGIR Forum* (2017), 260–267.
- [19] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In *Proceedings of ACL*. 6086–6096.
- [20] Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020. Pre-training via Paraphrasing. In *Proceedings of NeurIPS*.
- [21] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of NeurIPS*.
- [22] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of NeurIPS*. 6–12.
- [23] Jianxin Lin, Yingce Xia, Tao Qin, Zhibo Chen, and Tie-Yan Liu. 2018. Conditional image-to-image translation. In *Proceedings of CVPR*. 5524–5532.
- [24] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of CVPR*. 6738–6746.
- [25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* (2019). arXiv:1907.11692
- [26] Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, Dense, and Attentional Representations for Text Retrieval. *Trans. Assoc. Comput. Linguistics* (2021), 329–345.
- [27] Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. CEDR: Contextualized Embeddings for Document Ranking. In *Proceedings of SIGIR*. 1101–1104.
- [28] Bhaskar Mitra, Nick Craswell, et al. 2018. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval* 1 (2018), 1–126.
- [29] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *CoRR* abs/1901.04085 (2019). arXiv:1901.04085
- [30] Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019. From doc2query to docTTTTTquery. *Online preprint* (2019).
- [31] Prafull Prakash, Julian Killingback, and Hamed Zamani. 2021. Learning Robust Dense Retrieval Models from Incomplete Relevance Labels. In *Proceedings of SIGIR*.
- [32] Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W. Bruce Croft, and Mohit Iyyer. 2020. Open-Retrieval Conversational Question Answering. In *Proceedings of SIGIR*. 539–548.
- [33] Duyu Tang, Nan Duan, Tao Qin, Zhao Yan, and Ming Zhou. 2017. Question answering and question generation as dual tasks. *CoRR* (2017). arXiv:1706.02027
- [34] James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The Fact Extraction and VERification (FEVER) Shared Task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. 1–9.
- [35] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. 2017. Normface: L2 hypersphere embedding for face verification. In *Proceedings of ACM MM*. 1041–1049.
- [36] Tongzhou Wang and Phillip Isola. 2020. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. In *Proceedings of ICML*. 9929–9939.
- [37] Yingce Xia, Tao Qin, Wei Chen, Jiang Bian, Nenghai Yu, and Tie-Yan Liu. 2017. Dual Supervised Learning. In *Proceedings of ICML*. 3789–3798.
- [38] Yingce Xia, Xu Tan, Fei Tian, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2018. Model-Level Dual Learning. In *Proceedings of ICML*. 5379–5388.
- [39] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-End Neural Ad-hoc Ranking with Kernel Pooling. In *Proceedings of SIGIR*. 55–64.
- [40] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *Proceedings of ICLR*.
- [41] Wenhan Xiong, Xiang Lorraine Li, Srinivasan Iyer, Jingfei Du, Patrick S. H. Lewis, William Yang Wang, Yashar Mehdad, Wen-tau Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oğuz. 2020. Answering Complex Open-Domain Questions with Multi-Hop Dense Retrieval. *CoRR* (2020). arXiv:2009.12756
- [42] Peilin Yang, Hui Fang, and Jimmy Lin. 2018. Anserini: Reproducible ranking baselines using Lucene. *Journal of Data and Information Quality (JDIQ)* 4 (2018), 1–20.
- [43] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. 2017. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. In *Proceedings of ICCV*. 2849–2857.
- [44] Yang You, Jing Li, Sashank J. Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. In *Proceedings of ICLR*.
- [45] Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-Shot Conversational Dense Retrieval. In *Proceedings of SIGIR*.
- [46] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2021. Optimizing Dense Retrieval Model Training with Hard Negatives. In *Proceedings of SIGIR*.