COCO-DR: Combating Distribution Shifts in Zero-Shot Dense Retrieval with Contrastive and Distributionally Robust Learning

Yue Yu[1*] **Chenyan Xiong**[2] **Si Sun**[3] **Chao Zhang**[1] **Arnold Overwijk**[2]
[1] Georgia Institute of Technology [2] Microsoft [3] Tsinghua University
{yueyu, chaozhang}@gatech.edu, s-sun17@mails.tsinghua.edu.cn
{chenyan.xiong, arnold.overwijk}@microsoft.com

Abstract

We present a new zero-shot dense retrieval (ZeroDR) method, COCO-DR, to improve the generalization ability of dense retrieval by combating the distribution shifts between source training tasks and target scenarios. To mitigate the impact of document differences, COCO-DR continues pretraining the language model on the target corpora to adapt the model to target distributions via COtinuous COtrastive learning. To prepare for unseen target queries, COCO-DR leverages implicit Distributionally Robust Optimization (iDRO) to reweight samples from different source query clusters for improving model robustness over rare queries during fine-tuning. COCO-DR achieves superior average performance on BEIR, the zero-shot retrieval benchmark. At $\text{BERT}_{\text{Base}}$ scale, $\text{COCO-DR}_{\text{Base}}$ outperforms other ZeroDR models with 60× larger size. At $\text{BERT}_{\text{Large}}$ scale, $\text{COCO-DR}_{\text{Large}}$ outperforms the giant GPT-3 embedding model which has 500× more parameters. Our analysis show the correlation of COCO-DR's effectiveness in combating distribution shifts and improving zero-shot accuracy. Our code and model can be found at https://github.com/OpenMatch/COCO-DR.

1 Introduction

Learning to represent and match queries and documents by embeddings, dense retrieval (DR) achieves strong performances in scenarios with sufficient training signals (Bajaj et al., 2016; Kwiatkowski et al., 2019). However, in many real world scenarios, obtaining relevance labels can be challenging due to the reliance on domain expertise, or even infeasible because of the strict privacy constraints. Deploying dense retrieval in these scenarios becomes zero-shot (ZeroDR, Thakur et al. (2021)), which requires first training DR models on source tasks and then generalizing to target tasks with zero in-domain supervision (Izacard et al., 2022; Ni et al., 2021; Neelakantan et al., 2022).

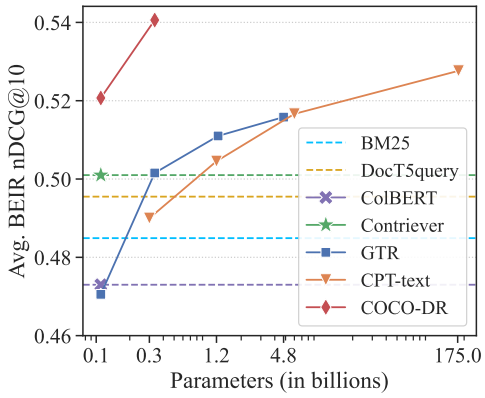


Figure 1: The average nDCG@10 of COCO-DR versus large scale models on the 11 BEIR tasks selected in Neelakantan et al. (2022). X-axis is in log scale.

ZeroDR poses great challenges to the generalization ability of DR models under the distribution shift between source and target data (Gulrajani and Lopez-Paz, 2021; Wiles et al., 2022), as it requires the alignment between queries and their relevant documents in the embedding space. It is much harder to generalize than standard classification or ranking tasks, where a robust decision boundary is sufficient (Xin et al., 2022).

In this work, we first analyze the distribution shifts in zero-shot dense retrieval. We illustrate the significant distribution shifts in both query intent and document language from the source to target tasks. After that, we show the strong correlation between the distribution shifts and the reduced zero-shot accuracy of dense retrieval models, which confirms the negative impact of distribution shifts on the generalization ability of dense retrieval.

We then present COCO-DR, a ZeroDR model that combats the distribution shifts between source and target tasks. In many ZeroDR scenarios, even though relevancy labels or queries are unavailable, the target corpus is often available pre-deploy (otherwise there is nothing to index) (Xin et al., 2022;

[*] Work partly done during Yue's internship at Microsoft.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 1462–1479
December 7-11, 2022

Wang et al., 2022). We thus design COCO-DR to perform COntinuous COntrastive pretraining (COCO) on the target corpora, which treats two text sequences from the same document as positive pairs and sequences from different documents as negative pairs. This enables COCO-DR to mitigate document distribution shifts by improving the alignment and uniformity of sequence representations for target tasks.

The distribution shift on the query intent, however, is more challenging as there only exists a few, if any, example queries available under ZeroDR scenarios. COCO-DR introduces an implicit distributionally robust optimization (iDRO) method when fine-tuning on the source retrieval labels. Specifically, it first clusters the source queries into groups based on their learned embeddings. Then, it dynamically reweights the losses on these query clusters by using the gradient similarity among groups. This improves model robustness on less represented query groups in the source, thus implicitly boosts the generalization ability of the DR model on unseen target queries.

COCO-DR is conceptually simple but empirically powerful. On 18 retrieval tasks included in BEIR, the standard ZeroDR benchmark (Thakur et al., 2021), COCO-DR outperforms state-of-the-art domain adaptation methods (Wang et al., 2022) which leverage per-task generated pseudo labels and cross-encoder teachers. COCO-DR also outperforms large scale models with orders of magnitude more parameters. As shown in Figure 1, at only $\text{BERT}_{\text{base}}$ scale with 110M parameters, COCO-DR outperforms GTR_{XXL} (Ni et al., 2021) and CPT_{L} (Neelakantan et al., 2022), which use $\sim$50$\times$ more parameters. At $\text{BERT}_{\text{Large}}$ scale, COCO-DR surpasses CPT_{XL} (Neelakantan et al., 2022), the largest DR model to date (175B parameters) on its selected tasks, only using 0.17% of its parameters.

Our analysis confirms that the better generalization ability of COCO-DR comes from its ability to combat the distribution shifts. Continuous contrastive learning helps the pretrained model better capture target corpora' sequence representation, leading to better generalization ability of models after fine-tuning. Training with iDRO helps COCO-DR achieve robust performances on source query clusters that share similar search intents to target queries, which then lead to better jgeneralization to corresponding target tasks.

In the rest of this paper, we discuss related work in Section 2, analyze the distribution shift in Section 3, and present COCO-DR in Section 4. Our experiments are discussed in Section 5 and we conclude in Section 6.

2 Related Work

Earlier research has explored various ways to learn representations for retrieval (Deerwester et al., 1990; Huang et al., 2013). Recently, with pretrained language models (Lee et al., 2019), hard training negative selection (Karpukhin et al., 2020; Xiong et al., 2021), and retrieval-oriented pretraining (Lu et al., 2021; Gao and Callan, 2022), dense retrieval has shown strong advantages over sparse retrieval methods, although the advantages are more observed in supervised settings than zero-shot scenarios (Thakur et al., 2021).

One research direction to improve zero-shot dense retrieval is bringing in domain adaption techniques. Xin et al. (2022) employ domain invariant learning to narrow the representation gap between source and target domains. Ma et al. (2021) and Wang et al. (2022) generate pseudo labels for each target task to train in-domain DR models. These techniques employ one specially trained retrieval model for each target task and improve zero-shot retrieval accuracy.

Another way to improve ZeroDR is to scale up model size and source training data. Ni et al. (2021) and Neelakantan et al. (2022) leverage models with billions of parameters (T5-XXL and GPT-3) and large-scale training data to increase the generalization capacity of DR model. Izacard et al. (2022) and Xu et al. (2022) enlarge the size of training data with retrieval-oriented pretraining tasks. As illustrated in Figure 1, the benefit of scale follows the scaling law of language models (Kaplan et al., 2020): A linear increment of zero-shot accuracy requires exponentially more training data and model parameters.

Combining dense models with sparse retrieval yields better zero-shot retrieval performances on BEIR (Formal et al., 2022; Xu et al., 2022). The reranking models, using stronger cross-encoders, can be used as teachers to improve the robustness of dense retrieval models (Wang et al., 2022).

More generally speaking, continuous pretraining and distributionally robust optimization (DRO) are two techniques for improving model generalization on other applications. Continuous pretraining BERT's masked language modeling tasks

on target domain corpora have shown benefits on both language tasks (Gururangan et al., 2020) and the reranking step of search systems (Wang et al., 2021b). The benefits of DRO are more ambivalent (Gulrajani and Lopez-Paz, 2021; Wiles et al., 2022) and are more observed when explicit group partitions are available (Oren et al., 2019; Sagawa et al., 2020; Zhou et al., 2021).

3 Distribution Shifts in Dense Retrieval

In this section, we first introduce the preliminaries of dense retrieval. Then we discuss the standard zero-shot dense retrieval settings and study the impact of distribution shifts on ZeroDR accuracy.

3.1 Preliminaries on Dense Retrieval

In dense retrieval, the query q and document d are represented by *dense* vectors (Huang et al., 2013) and the relevance score $f(q, d; \theta)$ is often calculated by simple similarity metrics, *e.g.*, dot product (Lee et al., 2019):

$$f(q, d; \theta) = \langle g(q; \theta), g(d; \theta) \rangle \,. \tag{1}$$

Here $g(\cdot; \theta)$ denotes the text encoder and θ is the collection of parameter of g, which is often initialized by BERT (Devlin et al., 2019). The learning objective for dense retrieval can be expressed as

$$\theta^* = \arg\min_{\theta} \ell(\theta) = \\ - \mathbb{E}_{q \sim p(\cdot)} \mathbb{E}_{d^+ \sim p_{\text{pos}}(q)} \mathbb{E}_{d^- \sim p_{\text{neg}}(q)} \log p_\theta \left(d^+ \mid q, d^- \right), \tag{2}$$

where $p(\cdot)$ is the distribution of queries, and d^+ and d^- are sampled from the distribution of positive and negative document for q (denoted as $p_{\text{pos}}(q)$ and $p_{\text{neg}}(q)$), respectively. In practice, the negative documents can either be BM25 negatives (Karpukhin et al., 2020) or mined by DR models from the past episode (Xiong et al., 2021).

During training, we aim to maximize the probability of selecting the ground-truth document d^+ over the negative document d^- as

$$p_\theta(d^+ | q, d^-) = \frac{\exp\left(f(q, d^+; \theta)\right)}{\exp\left(f(q, d^+; \theta)\right) + \exp\left(f(q, d^-; \theta)\right)}, \tag{3}$$

This dense retrieval configuration has shown strong empirical performances in a wide range of supervised scenarios, where the training and testing data are drawn from the same distributions, and a large amount of relevance labels are available (Karpukhin et al., 2020; Xiong et al., 2021; Qu et al., 2021).

3.2 ZeroDR and Distribution Shifts

Unlike supervised settings, the empirical advantages of dense retrieval are more ambivalent in zero-shot scenarios (Thakur et al., 2021). We first discuss the common setups of ZeroDR and then investigate the impact of distribution shifts on zero-shot performance of dense retrieval models.

ZeroDR Task. A retrieval task is considered zero-shot if no task-specific signal is available. Unless in large commercialized scenarios like web search, zero-shot is often the norm, *e.g.*, when building search systems for a new application, in domains where annotations require specific expertise, or in personalized scenarios where each user has her own corpus.

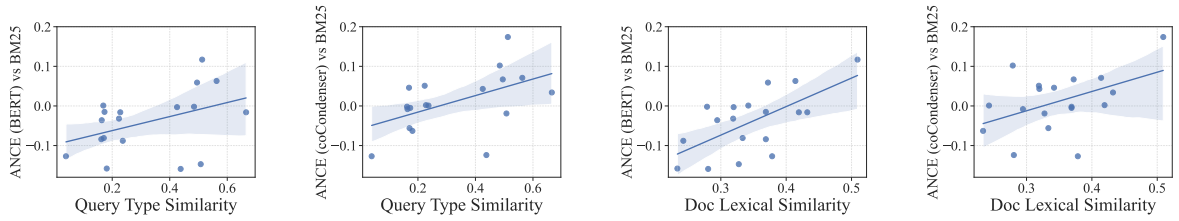
Besides relevance labels, the availability of in-domain queries is also a rarity—often only a few example queries are available. The most accessible in-domain information is the *corpus*, which is a prerequisite to build search systems. Sparse retrieval needs to pre-build the inverted index before serving any query; dense retrieval systems have to pre-compute the document embeddings.

These properties of zero-shot retrieval lead to a common ZeroDR setup where models can leverage the target corpus to perform unsupervised domain adaptation, but their supervised training signals only come from the source retrieval task, namely MS MARCO (Xin et al., 2022; Wang et al., 2022).

In this paper, we follow the standard practice in recent ZeroDR research, with MS MARCO passage retrieval (Bajaj et al., 2016) as the source retrieval task, the tasks collected in the BEIR benchmark (Thakur et al., 2021) as the zero-shot target, and the corpora of BEIR tasks available at training time for unsupervised domain adaptation.

Distribution Shifts. Before discussing our ZeroDR method, we first study the distribution shifts between the source training task (MARCO) and the zero-shot target tasks (BEIR).

Following the analysis in Thakur et al. (2021), we use pairwise weighted Jaccard similarity (Ioffe, 2010) to quantify the distribution shifts both at the query side and the document side. The document distribution shift is measured directly at the lexicon level, by the similarity of their unigram word distributions. The query distribution shift is measured on the distribution of query types, using the nine-type categorization from Ren et al. (2022) (more details in Appendix C.1). As shown in (Ren et al., 2022), search intent types are more representative



(a) Q, ANCE (BERT) (b) Q, ANCE (coCondenser) (c) Doc, ANCE (BERT) (d) Doc, ANCE (coCondenser)

Figure 2: Distribution shifts and zero-shot retrieval performances of ANCE trained on MS MARCO. X-axes are the similarity between MS MARCO and BEIR. Y-axes are NDCG@10 differences on BEIR.

than lexicon for short queries.

Figure 2 plots the distribution shifts from MARCO to BEIR tasks and the corresponding performance differences between dense retrieval and sparse retrieval. We use BM25 as the sparse retrieval method and ANCE starting from pretrained BERT (Xiong et al., 2021) and coCondenser (Gao and Callan, 2022) as representative DR models.

The average similarity between MS MARCO and BEIR tasks are 32.4% and 34.6% for queries and documents, indicating the existence of significant distribution shifts from MARCO to BEIR. Furthermore, these shifts are correlated with the performance degradation of dense retrieval models, as DR models perform much worse than BM25 on BEIR tasks that are less similar to MS MARCO. The contrastive learning on MARCO does not address this challenge; ANCE initialized from coCondenser still underperforms BM25 on BEIR tasks where distribution shifts are severe.

4 COCO-DR Method

To combat the distribution shifts from training source to zero-shot targets, COCO-DR introduces two training techniques: COntinuous COntrastive pretraining (COCO) and implicit Distributionally Robust optimization (iDRO). The first *continuously pretrains* the language model on target corpora to handle document distribution shifts. The latter improves the model robustness during *fine-tuning*, which then lead to better generalization for unseen target queries. This section describes these two components in detail.

4.1 Continuous Contrastive Pretraining

Sequence Contrastive Learning (SCL) aims to improve the alignment of similar text sequences in the pretrained representations and the uniformity of unrelated text sequences (Meng et al., 2021), which benefits supervised dense retrieval (Gao and Callan, 2022; Ma et al., 2022). In zero-shot settings, however, SCL-pretrained models still suffer from the distribution shifts, as observed in Figure 2.

COCO addresses this challenge via continuously pretraining the language model on the target corpora, using the contrastive learning settings widely adopted in recent research (Ni et al., 2021; Gao and Callan, 2022; Neelakantan et al., 2022).

Specifically, for each document d_i in target corpora, we randomly extract two disjoint sequences $s_{i,1}$ and $s_{i,2}$ from d_i to form the positive pair in:

$$\mathcal{L}_{\text{co}} = \sum_{i=1}^{n} \ell(s_{i,1}, s_{i,2})) \quad (4)$$
$$= \sum_{i=1}^{n} -\log \frac{\exp(\langle g(s_{i,1}), g(s_{i,2})\rangle)}{\sum_{j=1,2}\sum_{s^- \in B} \exp(\langle g(s_{i,j}), g(s^-)\rangle)}.$$

The contrastive loss with sequence representations $g(s)$ and in batch negatives $s^- \in B$.

This contrastive learning is used in combination with language modeling (Gao and Callan, 2022) to continuous pretrain on target corpora (Gururangan et al., 2020). It adapts the language models to target corpora before fine-tuning on source labels, to reduce the impact of document distribution shifts.

4.2 Distributionally Robust Optimization

The query distribution shifts are more challenging, as often target queries are only available, if any, at a small amount. For example, applying COCO on a few queries is unlikely useful.

To address this challenge, we exploit the assumption from distributional robust optimization (DRO): a model trained to be more robust on the source domain is likely to better generalize to unseen data (Sagawa et al., 2020; Wiles et al., 2022). In addition, as explicit target domain/group information is unavailable, we perform implicit DRO (iDRO) to improve models' robustness regarding to source query clusters during fine-tuning.

iDRO Loss. Specifically, we first cluster source queries using K-Means (Lloyd, 1982) on their embedding similarities (dot-product) from COCO, and then optimize the following iDRO loss:

$$\mathcal{L}_{\text{iDRO}}(\theta) = \sum_{i=1}^{K} \alpha_i \omega_i \ell_i(\theta), \tag{5}$$

$$\alpha_i \propto [\ell_i(\theta)]^{\beta}; \beta \geq 0. \tag{6}$$

It weights the per cluster dense retrieval loss $\ell_i(\theta)$ in Eqn. 2 of K total clusters using two parameters. The first one, α_i, up-weights clusters with higher training loss, with the emphasize on harder clusters defined by hyperparameter β. The second one $\boldsymbol{\omega} \in \mathbb{R}^K$ is learned to maximize the loss decreases on all clusters, which we derive a closed form solution in the rest of this section.

Dynamic Cluster Weighting. An ideal choice of $\boldsymbol{\omega}^t$ at training step t would provide biggest reduction on the training loss of all query clusters, but is difficult to obtain. To derive a closed form solution of ω^t, we approximate the loss reduction using first order Taylor expansion:

$$\ell_{\text{g}} = \sum_{i=1}^{K} \left(\ell_i(\theta - \eta \nabla_\theta \mathcal{L}_{\text{iDRO}}(\theta)) - \ell_i(\theta)\right) \tag{7}$$

$$\approx -\eta \sum_{i=1}^{K} \sum_{j=1}^{K} \alpha_i \alpha_j \omega_i^t \left(\nabla_\theta \ell_i(\theta)\right)^{\intercal} \nabla_\theta \ell_j(\theta). \tag{8}$$

Eqn. 7 is the loss reduction on all clusters, after a stochastic gradient descent operation with step size η. Eqn. 8 is its first order expansion.

In addition, we avoid potential rapid change of cluster weights for optimization stability, by adding a KL divergence regularization between $\boldsymbol{\omega}$ at different steps. This leads to the following optimization target:

$$\min_{\omega^{(t)}} \ \ell_{\text{g}} + \tau \mathcal{D}_{\text{KL}}(\boldsymbol{\omega}^{(t)} || \boldsymbol{\omega}^{(t-1)}), \tag{9}$$

$$\text{s.t.} \quad \sum_{i=1}^{K} \omega_i^{(t)} = 1. \tag{10}$$

The strength of KL regularization is controlled by hyperparameter τ. By using Lagrangian multiplier (details in Appendix E), the optimal weight for each group ω_i^{t*} can be calculated as

$$\omega_i^{t*} = \frac{\omega_i^{(t-1)} \exp\left(\frac{1}{\tau} \sum_{j=1}^{K} r_{ij}\right)}{\sum_{i=1}^{K} \omega_i^{(t-1)} \exp\left(\frac{1}{\tau} \sum_{j=1}^{K} r_{ij}\right)}; \tag{11}$$

$$r_{ij} = [\ell_i(\theta)\ell_j(\theta)]^{\beta} \left(\nabla_\theta \ell_i(\theta)\right)^{\intercal} \nabla_\theta \ell_j(\theta). \tag{12}$$

Intuitively, the optimal solution considers the gradient and loss similarity between different groups r_{ij}. It favors clusters sharing more 'common needs' (Piratla et al., 2022) with others to improve the model robustness across all clusters.

COCO and iDRO operate at different training stages of dense retrieval. COCO continuously pretrains the language model to adapt to the target documents, while iDRO improves the robustness of dense retrieval in the fine-tuning stage for better generalization on unseen queries. The two together forms COCO-DR that aims to improve zero-shot retrieval accuracy by combating the distribution shift from both the query and the document side.

5 Experiments

In this section, we first describe our experiment setups and evaluate COCO-DR. Then we analyze the efficacy of COCO and iDRO.

5.1 Experimental Setups

Our experiments use the tasks collected in BEIR (Thakur et al., 2021), a recent standard benchmark for zero-shot dense retrieval. The dataset details are in Appendix A.

Baselines. We consider various baselines, including standard sparse and dense retrieval models on BEIR. We also follow (Wang et al., 2022) to further compare COCO-DR with dedicated ZeroDR approaches based on unsupervised domain adaptation: these models are first pretrained on the target corpus and then fine-tuned on MS MARCO. We list the details of baselines in Appendix B.

Implementation Details. For COCO-DR, we use the same architecture as BERT (Devlin et al., 2019) and consider both *Base* and *Large* size in our experiments. The architecture of COCO-DR$_{\text{Base}}$ is the same as BERT$_{\text{Base}}$: 12 layer Transformer, 768 hidden size. Similarly, the architecture of COCO-DR$_{\text{Large}}$ model is the same as BERT$_{\text{Large}}$, using 24 layer and 1024 hidden size. Our implementation uses PyTorch (Paszke et al., 2019) with Hugging Face Transformers (Wolf et al., 2020) and OpenMatch (Liu et al., 2021) codebase.

In COCO stage, we initialize our model with Condenser (Gao and Callan, 2021), and continuously pretrain the model for 8 epochs (around 200K steps) on the corpus of BEIR and MS MARCO. We optimize the model using AdamW (Loshchilov and Hutter, 2019) with a peak learning rate 1e-4/1e-5 for Base/Large, weight decay 0.01, and linear learn-

	Sparse	Dense									Late-Inter.	COCO-DR (Ours)	
	BM25	DPR	ANCE	Contriever	GenQ†	GPL$^{\dagger,\#}$	GTR$_{XL}^{\ddagger}$	GTR$_{XXL}^{\ddagger}$	CPT$_{L}^{\ddagger,\sharp}$	CPT$_{XL}^{\ddagger,\sharp}$	ColBERT	Base	Large
Parameters#	—	110M	110M	110M	66M*18	66M*18	1.2B	4.8B	6B	175B	110M	110M	335M
MS MARCO	0.228	0.354	0.388	0.407	0.408	—	0.439	0.442	—	—	0.401	0.419	0.424
TREC-COVID	0.656	0.575	0.654	0.596	0.619	0.700	0.584	0.501	0.642	0.649	0.677	0.789	**0.804**
BioASQ	0.465	0.232	0.306	—	0.398	0.442	0.317	0.324	—	—	**0.474**	0.429	0.449
NFCorpus	0.325	0.210	0.237	0.328	0.319	0.345	0.343	0.342	0.380	**0.407**	0.305	0.355	0.354
NQ	0.329	0.398	0.446	0.498	0.358	0.483	0.559*	**0.568***	—	—	0.524	0.505	**0.547**
HotpotQA	0.603	0.371	0.456	0.638	0.534	0.582	0.591	0.599	0.648	**0.688**	0.593	0.616	0.641
FiQA-2018	0.236	0.274	0.295	0.329	0.308	0.344	0.444	0.467	0.452	**0.512**	0.317	0.307	0.329
Signal-1M	**0.330**	0.238	0.249	—	0.281	0.276	0.268	0.273	—	—	0.274	0.271	0.285
TREC-NEWS	0.398	0.366	0.382	—	0.396	0.421	0.350	0.346	—	—	0.393	0.403	**0.432**
Robust04	0.408	0.344	0.392	—	0.362	0.437	0.479	**0.506**	—	—	0.391	0.443	0.482
ArguAna	0.414	0.414	0.415	0.446	0.493	**0.557**	0.531	0.540	0.469	0.435	0.233	0.493	0.515
Touché-2020	**0.367**	0.208	0.240	0.230	0.182	0.255	0.230	0.256	0.309	0.291	0.202	0.238	0.263
Quora	0.789	0.842	0.852	0.865	0.830	0.836	0.890	**0.892**	0.677	0.638	0.854	0.867	0.872
DBPedia-entity	0.313	0.236	0.281	0.413	0.328	0.384	0.396	0.408	0.412	**0.432**	0.392	0.391	0.407
SCIDOCS	0.158	0.107	0.122	0.165	0.143	0.169	0.159	0.161	—	—	0.145	0.160	**0.178**
Fever	0.753	0.589	0.669	0.758	0.669	0.759	0.717	0.740	0.756	0.775	0.771	0.751	**0.793**
Climate-Fever	0.213	0.176	0.198	0.237	0.175	0.235	**0.270**	0.267	0.194	0.223	0.184	0.211	0.247
SciFact	0.665	0.475	0.507	0.677	0.644	0.674	0.635	0.662	0.744	**0.754**	0.671	0.709	0.722
CQADupStack	0.299	0.281	0.296	0.345	0.347	0.357	0.388	**0.399**	—	—	0.350	0.370	0.393
Avg CPT Sub	0.484	0.397	0.437	0.502	0.464	0.516	0.511	0.516	0.517	0.528	0.473	0.521	**0.541**
Avg	0.428	0.352	0.389	—	0.410	0.459	0.453	0.458	—	—	0.431	0.462	**0.484**

Table 1: nDCG@10 on the BEIR benchmark. The best result for each task is marked **bold**, and the best result among *fair* baselines (using BERT-base or smaller models as the backbone) is underlined. Avg CPT Sub is the average performance on 11 BEIR tasks used in Neelakantan et al. (2022). *: Unfair comparison, NQ is used in training for GTR. †: Train an independent model for each task. ‡: Larger Model, more training data. #: Use cross-encoders reranking teachers. ♯: Can only be accessed with paid APIs.

ing rate decay. The model is trained with 8 Nvidia A100 80GB GPUs and FP16 mixed-precision training. The batch size for each GPU is set to 200. Maximum number of tokens per sequence is 128.

The iDRO stage trains on MARCO passage retrieval with AdamW, 5e-6 learning rate, linear learning rate schedule, and batch size 64 for each GPU. Following Xiong et al. (2021), the model is first trained using BM25 negatives and then on self-negatives from the DR model. We update the query clusters with K-Means ($K = 50$) when refreshing negative samples. The running time for COCO and iDRO are around 1.5 days each for COCO-DR$_{\text{Base}}$ and around 3 days for COCO-DR$_{\text{Large}}$.

Evaluation Details. When evaluating on the BEIR benchmark, we use sequences of 64 tokens for the questions and 128 for the documents in all datasets except TREC-NEWS, Robust04, SciFact and ArguAna. In particular, we set the document length to 256 for TREC-NEWS, Robust04 and SciFact as they have larger document length on average. For ArguAna, we set both question and document length to 128 as it has longer queries.

Hyperparameters. The main hyperparameters in COCO-DR includes the number of groups K, the temperature parameter τ and the importance factor β. We keep $\beta = 0.25$ in COCO-DR and study the effect of N and τ in Sec. 5.3.

5.2 Overall Results

Table 1 shows the results on BEIR. Due to space limits, we only present the strongest baselines—other reported numbers are directly comparable, if they follow the standard ZeroDR settings on BEIR.

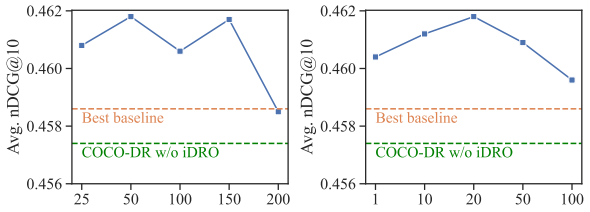
COCO-DR$_{\text{Base}}$ outperforms all previous methods on the average retrieval accuracy of all BEIR tasks, with large margin improvements over previous systems at BERT$_{\text{Base}}$ scale. It is also competitive and often better than models with significantly more parameters. COCO-DR$_{\text{Base}}$ achieves better average performance than GTR$_{XXL}$ and CPT$_{L}$ despite only using around 2% of their parameters. With more parameters, COCO-DR$_{\text{Large}}$ outperforms the giant CPT$_{XL}$ model (175B) by 2.5%, when evaluated on a subset of 11 datasets used in their experiment. It is worth noting that CPT$_{XL}$ can only be accessed with paid APIs. One inference for 18 BEIR tasks costs around 1.4 million dollars[1]. Scaling up models is not the only solution for zero-shot capacity. Better methodologies to tackle the distribution shifts can also improve the generalization of dense retrieval models, while being much "greener" (Schwartz et al., 2020).

COCO-DR also outperforms GPL, the strong domain adaptation model for ZeroDR (Wang et al., 2022). Note that GPL leverages a query generation model to produce pseudo relevance labels for each BEIR task, uses a cross-encoder to filter the pseudo

[1]The embedding model price ($0.2 per 1k tokens) at https://openai.com/api/pricing as of Oct. 2022.

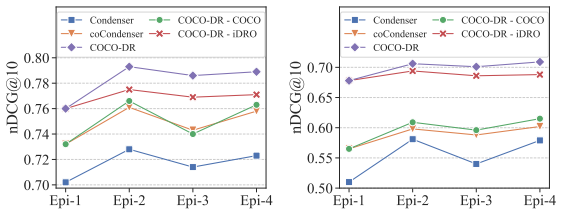
Method (→)	COCO-DR Base			COCO-DR Large			coCondenser			Condenser	
Dataset (↓)	Full	-iDRO	-COCO	Full	-iDRO	-COCO	Base (2022)	Base	Large	Base	Large
TREC-COVID	**0.789**	0.771	0.763	**0.804**	0.797	0.745	0.715	0.758	0.745	0.728	0.780
BioASQ	**0.429**	0.424	0.353	0.449	**0.450**	0.413	0.318	0.341	0.410	0.330	0.381
NFCorpus	**0.355**	0.354	0.333	**0.354**	0.353	0.349	0.307	0.326	0.350	0.282	0.317
NQ	0.505	0.503	**0.506**	**0.547**	0.536	0.519	0.494	0.503	0.516	0.472	0.492
HotpotQA	**0.616**	0.610	0.592	0.641	**0.644**	0.614	0.566	0.584	0.616	0.572	0.591
FiQA-2018	0.307	0.302	**0.312**	**0.329**	0.322	0.328	0.285	0.303	0.326	0.254	0.280
Signal-1M	0.271	0.275	**0.281**	0.285	0.285	**0.296**	0.274	0.274	0.295	0.266	0.284
TREC-NEWS	0.403	0.398	**0.426**	**0.432**	0.426	0.413	0.389	0.400	0.416	0.375	0.423
Robust04	0.443	0.443	**0.446**	**0.482**	0.467	0.466	0.399	0.442	0.461	0.385	0.418
ArguAna	**0.493**	0.479	0.473	**0.515**	0.513	0.488	0.411	0.460	0.484	0.439	0.469
Touché-2020	0.238	0.238	**0.257**	**0.263**	0.258	0.249	0.190	0.240	0.246	0.236	0.244
Quora	0.867	**0.868**	0.862	**0.872**	0.869	0.865	0.863	0.860	0.862	0.855	0.852
DBPedia-entity	**0.391**	0.389	0.382	**0.407**	0.401	0.388	0.356	0.364	0.386	0.362	0.364
SCIDOCS	0.160	**0.161**	0.154	**0.178**	0.176	0.171	0.140	0.150	0.171	0.143	0.161
Fever	0.751	**0.757**	0.739	**0.793**	0.783	0.741	0.678	0.751	0.724	0.725	0.736
Climate-Fever	**0.211**	0.209	0.202	**0.247**	0.240	0.233	0.184	0.208	0.226	0.206	0.216
SciFact	**0.709**	0.688	0.615	**0.722**	0.709	0.696	0.600	0.602	0.686	0.581	0.661
CQADupStack	**0.370**	0.365	0.349	**0.393**	0.385	0.367	0.330	0.342	0.363	0.313	0.343
Avg	**0.462**$^{\dagger,\ddagger,\flat,\natural}$	0.457	0.447	**0.484**$^{\dagger,\ddagger,\flat,\natural}$	0.478	0.463	0.417	0.440	0.460	0.418	0.445

Table 2: Ablation study of COCO-DR without iDRO (-iDRO) or continuous contrastive (-COCO). Apart from (2022), all the results are based on our own implementations. Superscripts indicate statistically significant results with p-value < 0.01 over -iDRO†, -COCO‡, coCondenser$^{\flat}$, Condenser$^{\natural}$.



(a) Effect of K (b) Effect of τ

Figure 3: Average NDCG@10 on BEIR of COCO-DR with different hyperparameters. The best baseline is GPL according to table 1.



(a) TREC-COVID (b) SciFact

Figure 4: The performance of COCO-DR and its variants over different training stages on TREC-COVID and SciFact. Epi-1 stands for the result after BM25 warmup, and Epi-2,3,4 are results of training with self-negative (ANCE). More results are in Appendix G.

labels, and trains one retrieval model for each task. COCO-DR does not rely on any of these techniques and uses one single model for all tasks. Its only modifications are on the model pretraining and fine-tuning strategies. More detailed comparisons with other domain adaptation approaches are in Sec. 5.4.

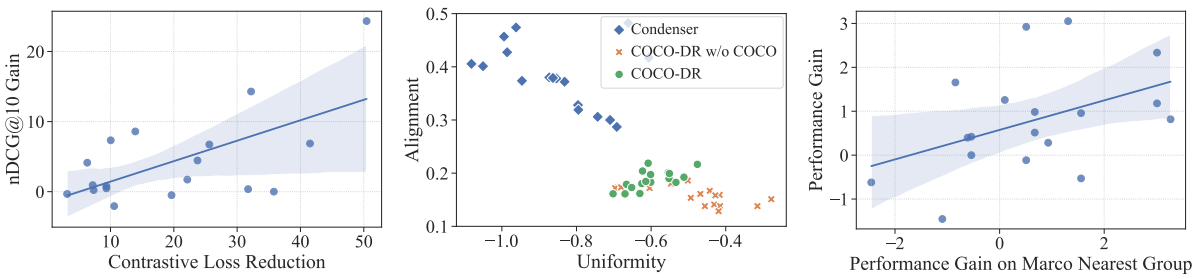
5.3 Ablation Study

We perform two groups of ablations on COCO-DR's hyperparameters and components.

Hyperparameters. Figure 3 shows the effect of two main hyperparameters, K for K-Means clustering and τ for temperatures in iDRO. When K becomes very large, the performance decreases as there exist fragmented clusters that are not close to any target BEIR tasks. As a result, focusing on these clusters hurts the average performance on BEIR tasks. When τ is too big, the weight for each group will be the same. On the contrary, if τ is too small, the model focuses too much on a few specific groups. Nevertheless, iDRO is robust and outperforms the best baseline in most studied hyperparameter regions.

Designed Components. Table 2 shows the performance of COCO-DR variations and the pretraining baselines. COCO and iDRO improve the average performance on BEIR datasets by 3.9% and 1.1% relatively. The stronger relative gains from COCO is expected, as it leverages the available in-domain corpora, while iDRO is designed for a harder challenge: to improve model generalization ability w.r.t. unseen target queries solely using training signals from the source.

Compared with coCondenser which is pretrained on MS MARCO only (-COCO) and uses the standard DR loss during finetuning (-iDRO), each design individually leads to improvements over a ma-



(a) COCO loss v.s. BEIR gain (b) ℓ_{uniform} and ℓ_{align} (c) BEIR v.s. MS MARCO nearest groups

Figure 5: Left: The relation between the gain of COCO v.s. the gain on BEIR tasks. Middle: ℓ_{uniform} & ℓ_{align} plot for COCO-DR and its variants on BEIR tasks. Right: The relation between the gain on BEIR tasks v.s. the gain on nearest MS MARCO groups.

jority of (COCO on 16; iDRO on 14) the 18 tasks included in BEIR. These two focus on different distribution shifts and operate at different stages of the training pipeline. Combining them in COCO-DR provides the best overall effectiveness.

Figure 4 zooms in the performances of COCO-DR and its variations on two BEIR tasks, TREC-COVID and SciFact, at different fine-tuning stages on the source task. It shows that COCO also helps stabilize the fine-tuning step on MS MARCO and reduces the oscillation between different training iterations. The benefit of iDRO is strong on biomedical tasks as shown in Figure 4, as MS MARCO indeed has relevent search intents in the BioMed domain. In Section 5.4 and 5.5, we analyze the benefits of the two designs in detail.

5.4 Influence of COCO Pretraining

To further understand the benefit of continuous contrastive pretraining, we perform three experiments on it, including: (1) comparison with other unsupervised domain adaptation (UDA) approaches, (2) the correlations between pretraining and zero-shot, and (3) the pretrained sequence representations.

Comparison with UDA methods. In Table 3 we compare COCO-DR with methods besides dense retrieval on the five domain specific tasks used in the experimental settings of Wang et al. (2022).[1]

COCO-DR outperforms all previous approaches, even those used a *reranking* model upon first stage retrieval. The latter previously was viewed as the "generalization upper bound" since they use strong cross-encoder models that have access to term-level matching signals (Wang et al., 2022). Previous methods that conducted contrastive pretraining such as ICT (Lee et al., 2019) and SimCSE (Gao et al., 2021) underperformed simple BM25 in zero-shot retrieval. These results corroborate the necessity of continuous contrastive learning.

Model	FiQA	SciFact	TREC-Covidv2	CQAD-upStack	Robust04	Avg.
Sparse Retrieval						
BM25 (2009)	0.239	0.661	0.601	0.315	0.387	0.461
Domain Adaptation Methods						
UDALM (2021)	0.233	0.336	0.571	0.246	0.263	0.330
MoDIR (2022)	0.296	0.502	0.660	0.297	—	—
Retrieval-Oriented Pretraining						
SimCSE (2021)	0.267	0.550	0.683	0.290	0.379	0.434
ICT (2019)	0.270	0.585	0.697	0.313	0.374	0.448
MLM (2019)	0.302	0.600	0.695	0.304	0.388	0.458
TSDAE (2021a)	0.293	0.628	0.761	0.318	0.394	0.479
Condenser (2021)	0.270	0.627	0.654	0.306	0.345	0.440
Condenser (ours)	0.250	0.617	0.732	0.334	0.411	0.469
In-Domain Generated Pseudo Labels						
QGen (2021)	0.287	0.638	0.724	0.330	0.381	0.472
GPL (2022)						
w/ DistillBERT (2019)	0.328	0.664	0.726	0.345	0.414	0.495
w/ TSDAE (2021a)	0.344	0.689	0.746	0.351	0.430	0.512
Reranking with Cross-Encoders, considered as "upper bound" (2022)						
Cross Encoder (MiniLM (2020))						
w/ BM25	0.331	0.676	0.712	0.368	0.467	0.511
w/ TSDAE+GPL (2022)	**0.364**	0.683	0.714	0.381	**0.483**	0.525
Our Method						
COCO-DR$_{\text{Base}}$ w/o iDRO	0.302	0.688	0.785	0.365	0.443	0.517
COCO-DR$_{\text{Base}}$	0.307	0.709	**0.807**	0.370	0.443	0.527†
COCO-DR$_{\text{Large}}$	0.329	**0.722**	**0.807**	**0.393**	0.482	**0.547**†

Table 3: Comparison to domain adaptation methods on the BEIR tasks used in (Wang et al., 2022). † indicates statistically significant results over the strongest baseline without using reranking models (GPL w/ TSDAE).

Pretraining versus Zero-Shot. In Figure 5(a) we plot the reduction of the sequence contrastive learning loss after using COCO pretraining on BEIR corpora (versus pretraining only on MARCO corpus), as well as the corresponding zero-shot improvements on each BEIR task. There is a notable correlation between them. On BioASQ, COCO reduces contrastive loss by 50% which yields 22% gains in zero-shot. Note that the pretrained models are fine-tuned *solely* on MS MARCO, but they provide attributable gains in zero-shot afterward.

Pretrained Representations. Following Wang

[1] We omit BioASQ here as Wang et al. (2022) evaluated on its subset that is not public.

Target TREC-COVID Query	MS MARCO Nearest Query
does SARS-CoV-2 have any subtypes, and if so what are they? (+0.174)	different types of hiv virus (+0.041)
how long can the coronavirus live outside the body (+0.057)	how long does hep c live outside body (+0.056)
what are best practices in hospitals and at home in maintaining quarantine? (+0.045)	define medical quarantine (+0.055)
is remdesivir an effective treatment for COVID-19 (+0.025)	how are antiviral drugs effective in treating infection? (+0.031)
what are the impacts of COVID-19 among African-Americans that differ from the rest of the U.S. population? (+0.030)	what ethnic group does sickle cell anemia affect (+0.026)

Table 4: Case study: Examples for nearest source queries of a target query in TREC-COVID and their performance gains after COCO-DR training. The number in brackets denotes the nDCG@10 gain from iDRO.

and Isola (2020), we use *alignment* and *uniformity* to illustrate the quality of learned representations on BEIR corpora (details in Appendix H). Figure 5(b) plots the results of COCO-DR on BEIR corpora with different pretraining components, before finetuning. Without contrastive learning, Condenser representations are not well aligned, which results in degeneration on target tasks. Contrastive learning on MS MARCO does not capture the sequence representations on BEIR, COCO-DR w/o COCO has low uniformity. COCO-DR provides a balanced alignment and uniformity which leads to better generalization (Wang and Isola, 2020).

5.5 Influence of Implicit DRO

The assumption of iDRO is that it improves the model robustness on rare query clusters in *source*, which helps generalize to unseen *target*. To verify this, we find MARCO query clusters closest to queries in each BEIR task (based on average dot product in COCO-DR embeddings). Then we plot the improvements of iDRO on BEIR tasks (zero-shot NDCG@10) and on their closest source clusters (training loss) in Figure 5(c).

From the figure, we observe the connections between the two sides: iDRO improved the training loss on the majority (12 out of 18) of source query clusters closest to BEIR. Moreover, such improvements have been successfully propagated to the BEIR tasks, as there exists a clear positive correlations among the performance gain on the MS MARCO and the corresponding target tasks. In Table 4, we show example query pairs with this connection on TREC-COVID to further support this argument. There are resemblance of the search intents between the source and target queries. The improvements of iDRO on the source queries thus also lead to the gains on unseen queries in BEIR.

6 Conclusion

COCO-DR improves ZeroDR accuracy by combating the distribution shifts using continuous contrastive learning and implicit distributionally robust optimization. COCO helps models better capture the sequence representations of target corpora in pretraining. Implicit DRO improves model robustness by reweighting query clusters in fine-tuning.

COCO-DR achieves strong zero-shot performance while maintaining a lightweight system with one unified model for all 18 target tasks. Different than prior works that scaling up the DR model to billions of parameters (*e.g.* CPT-text), we provide a more efficient and sustainable way to improve the zero-shot generalization ability. Our analyses observed clear correlations on COCO-DR's ability to mitigate the distribution shifts and to generalize. Better ZeroDR accuracy is observed on tasks where continuous contrastive learning has a lower pretraining loss, and where iDRO successfully identifies and improves source query clusters similar to target queries.

Limitations

In this work, we propose COCO-DR to combat the distribution shift issue for zero-shot dense retrieval. Despite the strong performance of our two key designs (COCO and iDRO), we mainly verify their efficacy from their empirical performance on BEIR tasks. More theoretical analyses are required to gain deeper understandings of these two designs. For COCO, more powerful tools are needed to establish the connection between contrastive pretraining and the performance on ZeroDR target tasks. For iDRO, the key assumption is that the robustness over rare query clusters will lead to better zero-shot performance on target out-of-domain tasks. However, there are no theoretical groundings to connect these two terms for DR models. These analyses will go beyond our empirical observations and reveal the true inner workings of COCO-DR.

Acknowledgements

We would like to thank Ji Xin and Nandan Thakur for their help on getting access to non-public

datasets of the BEIR benchmark. We also thank anonymous reviewers for their feedback. Yue Yu and Chao Zhang were partly supported by NSF IIS-2008334, IIS-2106961, and CAREER IIS-2144338.

References

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

Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2020. Overview of Touché 2020: Argument Retrieval. In *Working Notes Papers of the CLEF 2020 Evaluation Labs*, volume 2696 of *CEUR Workshop Proceedings*.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *European Conference on Information Retrieval*, pages 716–722. Springer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. CLIMATE-FEVER: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. From distillation to hard negative sampling: Making sparse neural ir models more effective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2353–2359, New York, NY, USA. Association for Computing Machinery.

Luyu Gao and Jamie Callan. 2021. Condenser: a pre-training architecture for dense retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 981–993, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Luyu Gao and Jamie Callan. 2022. Unsupervised corpus aware language model pre-training for dense passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2843–2853, Dublin, Ireland. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ishaan Gulrajani and David Lopez-Paz. 2021. In search of lost domain generalization. In *International Conference on Learning Representations*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. 2017. DBpedia-Entity v2: A test collection for entity search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 1265–1268, New York, NY, USA. Association for Computing Machinery.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 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*, page 113–122. Association for Computing Machinery.

Doris Hoogeveen, Karin M. Verspoor, and Timothy Baldwin. 2015. CQADupStack: A benchmark data set for community question-answering research. In *Proceedings of the 20th Australasian Document Computing Symposium*, ADCS '15, New York, NY, USA. Association for Computing Machinery.

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

Sergey Ioffe. 2010. Improved consistent sampling, weighted minhash and l1 sketching. In *2010 IEEE International Conference on Data Mining*, pages 246–255. IEEE.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Constantinos Karouzos, Georgios Paraskevopoulos, and Alexandros Potamianos. 2021. UDALM: Unsupervised domain adaptation through language modeling. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2579–2590, Online. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 39–48, New York, NY, USA. Association for Computing Machinery.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1907.11692*.

Zhenghao Liu, Kaitao Zhang, Chenyan Xiong, Zhiyuan Liu, and Maosong Sun. 2021. Openmatch: An open source library for neu-ir research. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2531–2535, New York, NY, USA. Association for Computing Machinery.

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Shuqi Lu, Di He, Chenyan Xiong, Guolin Ke, Waleed Malik, Zhicheng Dou, Paul Bennett, Tie-Yan Liu, and Arnold Overwijk. 2021. Less is more: Pretrain a strong Siamese encoder for dense text retrieval using a weak decoder. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2780–2791, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2021. Zero-shot neural passage retrieval via domain-targeted synthetic question generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1075–1088, Online. Association for Computational Linguistics.

Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, and Xueqi Cheng. 2022. Pre-train a discriminative text encoder for dense retrieval via contrastive span prediction. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 848–858, New York, NY, USA. Association for Computing Machinery.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. WWW'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Yu Meng, Chenyan Xiong, Payal Bajaj, saurabh tiwary, Paul N. Bennett, Jiawei Han, and Xia Song. 2021. COCO-LM: Correcting and contrasting text sequences for language model pretraining. In *Advances in Neural Information Processing Systems*.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. 2022. Text and code embeddings by contrastive pretraining. *arXiv preprint arXiv:2201.10005*.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y Zhao, Yi Luan, Keith B Hall, Ming-Wei Chang, et al. 2021. Large dual encoders are generalizable retrievers. *arXiv preprint arXiv:2112.07899*.

Yonatan Oren, Shiori Sagawa, Tatsunori B. Hashimoto, and Percy Liang. 2019. Distributionally robust language modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4227–4237, Hong Kong, China. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. 2022. Focus on the common good: Group distributional robustness follows. In *International Conference on Learning Representations*.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qifei Wu, Yuchen Ding, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2022. A thorough examination on zero-shot dense retrieval. *arXiv preprint arXiv:2204.12755*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks. In *International Conference on Learning Representations*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green ai. *Commun. ACM*, 63(12):54–63.

Ian Soboroff, Shudong Huang, and Donna Harman. 2018. Trec 2018 news track overview.

Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. 2020. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352.

Axel Suarez, Dyaa Albakour, David Corney, Miguel Martinez, and José Esquivel. 2018. A data collection for evaluating the retrieval of related tweets to news articles. In *European Conference on Information Retrieval*, pages 780–786. Springer.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):1–28.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. TREC-COVID: Constructing a pandemic information retrieval test collection. *SIGIR Forum*, 54(1).

Ellen M Voorhees et al. 2004. Overview of the trec 2004 robust retrieval track. In *Trec*, pages 69–77.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021a. TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States. Association for Computational Linguistics.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Wei Chen, and Tie-Yan Liu. 2013. A theoretical analysis of ndcg ranking measures. In *Proceedings of the 26th annual conference on learning theory (COLT 2013)*, volume 8, page 6. Citeseer.

Yu Wang, Jinchao Li, Tristan Naumann, Chenyan Xiong, Hao Cheng, Robert Tinn, Cliff Wong, Naoto Usuyama, Richard Rogahn, Zhihong Shen, et al. 2021b. Domain-specific pretraining for vertical search: Case study on biomedical literature. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3717–3725.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dj Dvijotham, and Ali Taylan Cemgil. 2022. A fine-grained analysis on distribution shift. In *International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ji Xin, Chenyan Xiong, Ashwin Srinivasan, Ankita Sharma, Damien Jose, and Paul Bennett. 2022. Zero-shot dense retrieval with momentum adversarial domain invariant representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4008–4020, Dublin, Ireland. Association for Computational Linguistics.

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 *International Conference on Learning Representations*.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2022. LaPraDoR: Unsupervised pretrained dense retriever for zero-shot text retrieval. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3557–3569, Dublin, Ireland. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Chunting Zhou, Xuezhe Ma, Paul Michel, and Graham Neubig. 2021. Examining and combating spurious features under distribution shift. In *International Conference on Machine Learning*, pages 12857–12867. PMLR.

A Datasets Details

Target domain datasets used in our experiments are collected in the BEIR benchmark (Thakur et al., 2021)[1] and include the following domains:

- Bio-Medical Information Retrieval: TREC-COVID (Voorhees et al., 2021), NFCorpus (Boteva et al., 2016), and BioASQ (Tsatsaronis et al., 2015).
- Open-domain Question Answering (QA): HotpotQA (Yang et al., 2018), NQ (Kwiatkowski et al., 2019), and FiQA (Maia et al., 2018).
- Argument Retrieval: Webis-Touché2020 (Bondarenko et al., 2020) and ArguAna (Wachsmuth et al., 2018).
- News Retrieval: TREC-NEWS (Soboroff et al., 2018) and Robust04 (Voorhees et al., 2004).
- Tweet Retrieval: Signal-1m (Suarez et al., 2018).
- Duplicate Question Retrieval: Quora (Thakur et al., 2021) and CQADupStack (Hoogeveen et al., 2015).
- Entity Retrieval: DBPedia (Hasibi et al., 2017)
- Citation Prediction: SCIDOCS (Cohan et al., 2020)
- Fact Checking: SciFact (Wadden et al., 2020), FEVER (Thorne et al., 2018), and Climate-FEVER (Diggelmann et al., 2020)

We list the statistics of the BEIR benchmark in Table 5.

Metric To measure the effectiveness of search algorithms or retrieval models, the benchmark uses Normalized Discounted Cumulative Gain (nDCG@10) (Wang et al., 2013) as the evaluation metric. The higher value indicates better performance.

B Baselines

We use the baselines from the current BEIR leaderboard (Thakur et al., 2021) and recent papers. For the main experiments, the baselines can be divided into four groups: dense retrieval, dense retrieval with generated queries[2], lexical retrieval, and late interaction.

B.1 Baselines for Main Experiments

Dense Retrieval For dense retrieval, the baselines are the same dual-tower model as ours. We consider **DPR** (Karpukhin et al., 2020), **ANCE** (Xiong et al., 2021), **Contriever** (Izacard et al., 2022), and two recently-proposed giant model, namely **GTR** (Ni et al., 2021) and **CPT-text** (Neelakantan et al., 2022) in this paper.

- **DPR** uses a single BM25 retrieval example and in-batch examples as hard negative examples to train the model. Different from the original paper (Thakur et al., 2021) that train the DPR on QA datasets, we train DPR on MS MARCO (Bajaj et al., 2016) Dataset for *fair comparison*. Notice that this also lead to better results according to Xin et al. (2022).
- **ANCE** constructs hard negative examples from an ANN index of the corpus. The hard negative training instances are updated in parallel during fine-tuning of the model. The model is a RoBERTa (Liu et al., 2019) model trained on MS MARCO for 600k steps.
- **Contriever** conducts unsupervised contrastive pretraining with data augmentations and momentum queues on Wikipedia and CC-Net (Wenzek et al., 2020) corpora for 500k steps.
- **GTR** initializes the dual encoders from the T5 models (Raffel et al., 2019). It is first pre-trained on Community QA[3] with 2 billion question-answer pairs then fine-tuned on NQ and MS Marco dataset.
- **CPT-text** initializes with the large GPT models (Brown et al., 2020), and pre-trained on web-scale Internet data with neighboring pieces of text as positive pairs for the contrastive objective.

Dense Retrieval with Generated Queries

- **GenQ** first fine-tunes a T5-base (Raffel et al., 2019) model on MS MARCO for 2 epochs and then generate 5 queries for each passage as additional training data for the target domain to

[1] https://github.com/beir-cellar/beir

[2] We separate them from dense retrieval since they usually rely on Seq2seq models to generate pseudo query-document pairs, and they train a model for each dataset *independently* instead of using a single model for all datasets.

[3] Unfortunately, this corpus is not publicly available.

Split (→)					Train	Dev	Test			Avg. Word Lengths	
Task (↓)	Domain (↓)	Dataset (↓)	Title	Relevancy	#Pairs	#Query	#Query	#Corpus	Avg. D / Q	Query	Document
Passage-Retrieval	Misc.	MS MARCO	✗	Binary	532,761	—-	6,980	8,841,823	1.1	5.96	55.98
Bio-Medical	Bio-Medical	TREC-COVID	✓	3-level	—-	—-	50	171,332	493.5	10.60	160.77
Information	Bio-Medical	NFCorpus	✓	3-level	110,575	324	323	3,633	38.2	3.30	232.26
Retrieval (IR)	Bio-Medical	BioASQ	✓	Binary	32,916	—-	500	14,914,602	4.7	8.05	202.61
Question	Wikipedia	NQ	✓	Binary	132,803	—-	3,452	2,681,468	1.2	9.16	78.88
Answering	Wikipedia	HotpotQA	✓	Binary	170,000	5,447	7,405	5,233,329	2.0	17.61	46.30
(QA)	Finance	FiQA-2018	✗	Binary	14,166	500	648	57,638	2.6	10.77	132.32
Tweet-Retrieval	Twitter	Signal-1M (RT)	✗	3-level	—-	—-	97	2,866,316	19.6	9.30	13.93
News	News	TREC-NEWS	✓	5-level	—-	—-	57	594,977	19.6	11.14	634.79
Retrieval	News	Robust04	✗	3-level	—-	—-	249	528,155	69.9	15.27	466.40
Argument	Misc.	ArguAna	✓	Binary	—-	—-	1,406	8,674	1.0	192.98	166.80
Retrieval	Misc.	Touché-2020	✓	3-level	—-	—-	49	382,545	19.0	6.55	292.37
Duplicate-Question	StackEx.	CQADupStack	✓	Binary	—-	—-	13,145	457,199	1.4	8.59	129.09
Retrieval	Quora	Quora	✗	Binary	—-	5,000	10,000	522,931	1.6	9.53	11.44
Entity-Retrieval	Wikipedia	DBPedia	✓	3-level	—-	67	400	4,635,922	38.2	5.39	49.68
Citation-Prediction	Scientific	SCIDOCS	✓	Binary	—-	—-	1,000	25,657	4.9	9.38	176.19
	Wikipedia	FEVER	✓	Binary	140,085	6,666	6,666	5,416,568	1.2	8.13	84.76
Fact Checking	Wikipedia	Climate-FEVER	✓	Binary	—-	—-	1,535	5,416,593	3.0	20.13	84.76
	Scientific	SciFact	✓	Binary	920	—-	300	5,183	1.1	12.37	213.63

Table 5: Statistics of datasets in the BEIR benchmark. The table is taken from the original BEIR benchmark paper (Thakur et al., 2021).

continue to fine-tune the TAS-B (Hofstätter et al., 2021) model.

- **GPL** is a recent work that improve the perforance of GenQ with cross-encoder reranking. It first generates queries for documents from the target domain, then use an additional cross-encoder (Wang et al., 2020) to rank each (query, document)-pair and then train a dense retrieval model on these generated, pseudo-labeled queries[1].

Lexical Retrieval Lexical retrieval is a score function for token matching calculated between two high-dimensional sparse vectors with token weights.

- **BM25** (Robertson et al., 2009) is the most commonly used lexical retrieval function. We use the BM25 results reported in Thakur et al. (2021) for comparison.

Late Interaction We also consider a late interaction baseline, namely **ColBERT** (Khattab and Zaharia, 2020). The model computes multiple contextualized embeddings for each token of queries and documents, and then uses a maximum similarity function to retrieve relevant documents. This type of matching requires significantly more disk space for indexes and has a higher latency.

[1]In the original paper, they have tried on multiple backbones including DistillBERT (Sanh et al., 2019), TSDAE (Wang et al., 2021a), TAS-B (Hofstätter et al., 2021) for evaluations, and we select the best model that based on TAS-B for comparison in our main experiments.

B.2 Additional Domain Adaptation Baselines

We further compare COCO-DR with additional baselines focus on domain adaptation to specialized domains including **UDALM** (Karouzos et al., 2021), **MoDIR** (Xin et al., 2022), **SimCSE** (Gao et al., 2021), **ICT** (Lee et al., 2019), **MLM** (Liu et al., 2019), **TSDAE** (Wang et al., 2021a), and **Condenser** (Gao and Callan, 2021). Note that these models are first pre-trained on the target corpus and then fine-tuned on the MS MARCO dataset.

- **UDALM** is a domain adaptation method that originally designed for sentiment analysis. It applies the multi-task training to jointly learn from the target task and the MLM task.
- **MoDIR** is a momentum-based method to ensure stable and efficient adversarial learning for domain adaptation.
- **SimCSE** is a simple approach proposed for sentence similarity calculation. Specifically, it regards the document text twice with different dropout as the positive sample pairs to enable contrastive learning.
- **ICT** selects one sentence from a whole document as the pseudo query to that document for pre-training.
- **MLM** random masks 15% tokens in a text and designs a cloze-style test for pre-training the model.
- **TSDAE** leverages an additional denoising autoencoder to pre-train the dense retriever model

Dataset (↓)	Query Intent Similarity	Document Lexical Similarity	ANCE ($\text{BERT}_{\text{Base}}$) v.s. BM25	ANCE (coCondenser) v.s. BM25
TREC-COVID	0.4845	0.2789	-0.002	+0.102
BioASQ	0.4380	0.2806	-0.159	-0.124
NFCorpus	0.2367	0.2426	-0.088	+0.001
NQ	0.5127	0.5092	+0.117	+0.174
HotpotQA	0.5078	0.3275	-0.147	-0.019
FiQA-2018	0.4950	0.3721	+0.059	+0.067
Signal-1M	0.1708	0.3334	-0.081	-0.056
TREC-NEWS	0.2280	0.4194	-0.016	+0.002
Robust04	0.6656	0.4323	-0.016	+0.008
ArguAna	0.1690	0.3421	+0.001	+0.046
Touché-2020	0.0391	0.3785	-0.127	-0.127
Quora	0.5629	0.4141	+0.063	+0.071
DBPedia-entity	0.2235	0.3189	-0.032	+0.051
SCIDOCS	0.1636	0.2945	-0.036	-0.008
Fever	0.1621	0.3689	-0.084	-0.002
Climate-Fever	0.1732	0.3689	-0.015	-0.014
SciFact	0.1809	0.2335	-0.158	-0.092
CQADupStack	0.4254	0.3196	-0.003	+0.043

Table 6: Detailed statistics for (1) query intent similarity and document lexical similarity between MS MARCO and BEIR tasks (2) the performance gap between ANCE starting from $\text{BERT}_{\text{base}}$ and coCondenser and BM25. The positive value indicates ANCE performs better than BM25.

with 60% random tokens deleted in the input document.

- **Condenser** improves the representation of `[CLS]` token by enforcing it to aggregate with the token embedding. In this way, the head model can then condition on late `[CLS]` to make LM predictions to enforce `[CLS]` to capture the global meaning of the input text.

C Details for Similarity Calculation

In this section, we provide more details on how to calculate the distribution shifts between the source training task (MS MARCO) and the zero-shot target tasks (BEIR). We first define the types of queries used in Section 3.2, and then give more details about the calculation of the weighted Jaccard similarity (Ioffe, 2010) used in this study.

C.1 Types of Queries

We adopt the same method as (Ren et al., 2022) to partition the training queries into 9 types: for queries starting with the following 7 words, 'what', 'when', 'who', 'how', 'where', 'why', 'which', they fall into the corresponding category. Besides, queries starting with the first word `is/was/are/were/do/does/did/have/has/had/should/can/could/would/am/small`', are classified as Y/N queries. The rest of the queries belong to declarative queries.

C.2 Calculation of Weighted Jaccard Similarity

We follow (Thakur et al., 2021) to use the weighted Jaccard similarity $J(S,T)$ to measure the unique word overlap for all words present in the source dataset S and the target dataset T.

Denote S_k as the frequency of word k in the source dataset S and T_k for the target dataset T respectively. The weighted Jaccard similarity $J(S,T)$ between S and T is defined as:

$$J(S,T) = \frac{\sum_k \min(S_k, T_k)}{\sum_k \max(S_k, T_k)}, \qquad (13)$$

where the sum is over all unique words k present in dataset S and T.

D Statistics for Query and Document Similarities

Table 6 lists the exact pairwise weighted Jaccard similarity between MS MARCO and different BEIR tasks. For tasks comes from biomedical domains (e.g. BioASQ, NFCorpus) and scientific domains (e.g. SCIDOCS, SciFact), the lexical overlap between them and MS MARCO is small. For these datasets, ANCE can hardly outperform BM25. On the other hand, for those tasks which ANCE outperforms BM25 by a wide margin (e.g. NQ, Quora), they tend to have a larger weighted Jaccard similarity score with MS MARCO.

E Details of iDRO

This section exhibits the details for deriving the optimal weight $\boldsymbol{\omega}^{(t)}$ for the training step t. Note that the overall objective can be expressed as

$$\min_{\omega^{(t)}} \ \ell_{\mathrm{g}} + \tau \mathcal{D}_{\mathrm{KL}}(\boldsymbol{\omega}^{(t)} || \boldsymbol{\omega}^{(t-1)}), \tag{14}$$

$$\text{s.t.} \quad \sum_{i=1}^{K} \omega_i^{(t)} = 1, \tag{15}$$

where τ is the temperature to control the strength of the regularization. Then, the KKT conditions can be expressed as

$$\mathcal{L} = -\sum_{i=1}^{K}\sum_{j=1}^{K} \omega_i \alpha_i \alpha_j \left(\nabla_\theta \ell_i(\theta)\right)^{\intercal} \nabla_\theta \ell_j(\theta) \tag{16}$$

$$+ \tau \sum_{i=1}^{K} \left(\log\left(\frac{\omega^{(t)}}{\omega^{(t-1)}} \right) - 1 \right) \tag{17}$$

$$+ \gamma \left(\sum_{i=1}^{K} \omega_i^{(t)} - 1 \right) \tag{18}$$

Setting the corresponding gradients to 0 gives the global optimum as

$$\frac{\partial \mathcal{L}}{\partial \omega_i^{(t)}} = -\sum_{j=1}^{K} r_{ij} + \tau \log\left(\frac{\omega^{(t)}}{\omega^{(t-1)}} \right) + \widehat{\gamma} = 0; \tag{19}$$

$$\sum_{i=1}^{K} \omega_i^{(t)} = 1, \tag{20}$$

where

$$r_{ij} = \sum_{i=1}^{K} \alpha_i \alpha_j \left(\nabla_\theta \ell_i(\theta)\right)^{\intercal} \nabla_\theta \ell_j(\theta),$$

$$\widehat{\gamma} = \gamma + \tau.$$

From the above Eqn. 19, we have

$$\omega_i^{(t)} = \omega_i^{(t-1)} \exp\left(\frac{1}{\tau} \left(\sum_{j=1}^{K} r_{ij} - \widehat{\gamma} \right) \right). \tag{21}$$

By plugging the Eqn. 21 to Eqn. 20, we obtain

$$\exp\left(\frac{\widehat{\gamma}}{\tau}\right) = \sum_{i=1}^{K} \exp\left(\frac{1}{\tau} \sum_{i=1}^{K} \omega_i^{(t-1)} r_{ij} \right). \tag{22}$$

Finally, by combining the Eqn. 21 and Eqn. 22, the weight for i-th group can be expressed as

$$\omega_i^{t*} = \frac{\omega_i^{(t-1)} \exp\left(\frac{1}{\tau} \sum_{j=1}^{K} r_{ij} \right)}{\sum_{i=1}^{K} \omega_i^{(t-1)} \exp\left(\frac{1}{\tau} \sum_{j=1}^{K} r_{ij} \right)}. \tag{23}$$

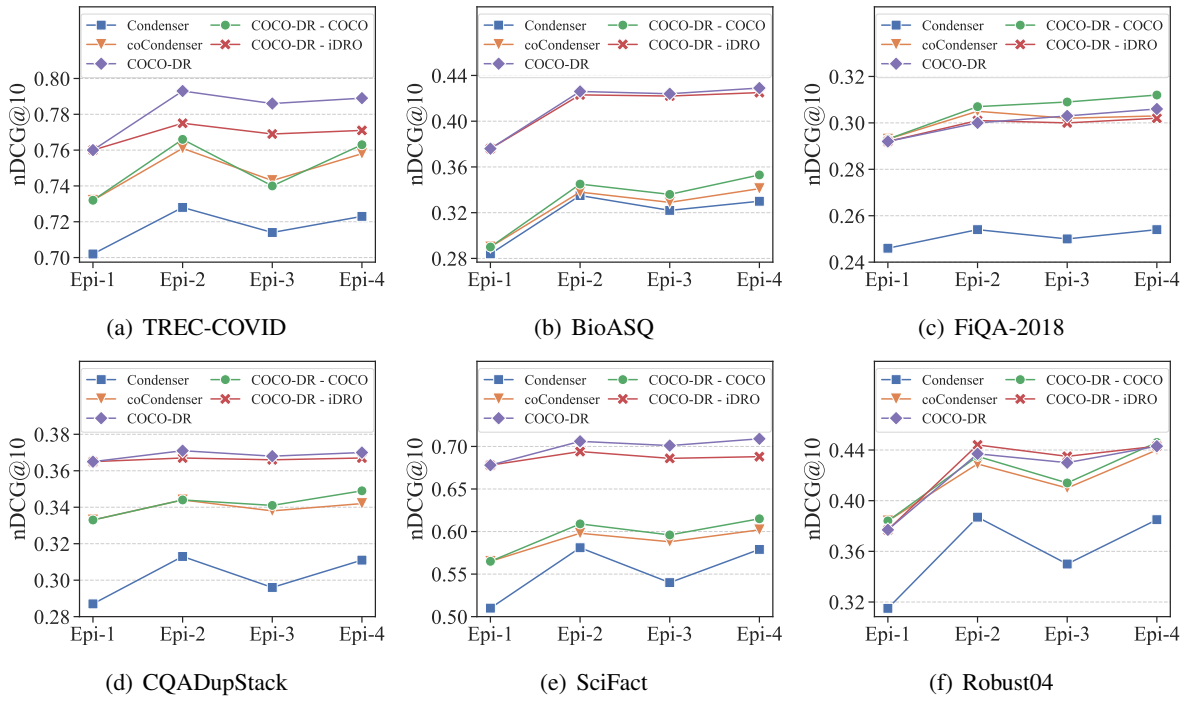
Dataset (↓)	COCO-DR	GroupDRO (2020)
TREC-COVID	0.789	**0.793**
BioASQ	**0.429**	0.411
NFCorpus	**0.355**	0.352
NQ	**0.505**	0.494
HotpotQA	**0.616**	0.609
FiQA-2018	**0.307**	0.300
Signal-1M	0.271	**0.274**
TREC-NEWS	0.403	**0.408**
Robust04	**0.443**	0.438
ArguAna	0.493	0.493
Touché-2020	0.238	**0.243**
Quora	**0.867**	0.866
DBPedia-entity	**0.391**	0.390
SCIDOCS	0.160	**0.162**
Fever	**0.751**	0.746
Climate-Fever	0.211	0.211
SciFact	0.709	**0.712**
CQADupStack	**0.370**	0.367
Avg	**0.462**	0.459

Table 7: Comparision between iDRO and GroupDRO (Sagawa et al., 2020). COCO-DR achieves better performance on the majority of BEIR tasks.

F Comparision with GroupDRO

We further compare iDRO with GroupDRO (Sagawa et al., 2020), which assigns higher weights to groups with higher training loss. Note that GroupDRO requires gold labels for group assignments which is unavailable for ZeroDR. To adopt GroupDRO in our settings, we use the cluster information derived from K-means clustering as group labels, which is the same as (Sohoni et al., 2020). To ensure fair comparison, we use the model after COCO pretraining as initialization, and use GroupDRO to reweight different groups during fine-tuning the model on MS MARCO.

Table 7 shows the performance of GroupDRO on BEIR tasks. From the results, we find that although GroupDRO achieves better performance on some specific tasks (e.g. TREC-COVID and SciFact), it fails to perform well on the majority of tasks, especially for general-domain datasets such as NQ, HotpotQA and Fever. This is because during GroupDRO training, it assigns higher weights for large-loss groups while neglecting other groups. As a result, although it will lead to better worse-group performance, it cannot improve the average performance. In contrast, iDRO leverages gradient similarities to dynamically reweight different groups to avoid sacrificing the average performance on all tasks.



(a) TREC-COVID (b) BioASQ (c) FiQA-2018

(d) CQADupStack (e) SciFact (f) Robust04

Figure 6: The performance of COCO-DR and its variants over different training stages on 6 of BEIR tasks.

G Performance on Different Training Stages of COCO-DR

Figure 6 exhibits the performance on different episodes on six BEIR tasks from different domains, used in (Wang et al., 2022). From the results, we observe that COCO is more beneficial for the biomedical domains than others such as news and finance. The more significant gain is mainly due to the limited overlap between biomedical corpus and MS MARCO, as well as the extremely large size of the biomedical corpora. For other two tasks (Robust04 and FiQA-2018), the DR models can already achieve better or comparable performance compared with BM25 when finetuning on MS MARCO only, which indicates the distribution shift issue is not severe on these datasets. Therefore, the relative gain of COCO on them is smaller.

For the iDRO part, it provides additional performance gains on 5 of 6 datasets. As these datasets are all domain specific text retrieval tasks (Wang et al., 2022), the results justify the benefits of iDRO for improving the DR model's performance on unseen target queries.

H Calculation of Alignment and Uniformity

Recently, Wang and Isola (2020) propose two terms, namely *alignment* and *uniformity* to measure the quality of representations. In particular, we denote the whole data distribution as p_{data} and the distribution of positive pairs as p_{pos}. Then, the two metrics can be calculated as

$$\ell_{\text{align}} \triangleq \mathbb{E}_{(x,x^+)\sim p_{\text{pos}}} \|f(x) - f(x^+)\|^2, \quad (24)$$

$$\ell_{\text{uniform}} \triangleq \log \mathbb{E}_{(x,y)\overset{\text{i.i.d.}}{\sim} p_{\text{data}}} e^{-2\|f(x)-f(y)\|^2}. \quad (25)$$

Notably, *alignment* is the expected distance between the representations of positive text pairs, and *uniformity* measures how well the text representations are uniformly distributed (Gao et al., 2021). In our experiments, we use the code released by the original authors to calculate these two metrics.[1]

[1]Link: https://github.com/SsnL/align_uniform