Zero-Shot Dense Retrieval with Momentum Adversarial Domain Invariant Representations

Ji Xin[1]*, **Chenyan Xiong**[2], **Ashwin Srinivasan**[2], **Ankita Sharma**[2], **Damien Jose**[2], **Paul N. Bennett**[2]
[1] University of Waterloo [2] Microsoft
ji.xin@uwaterloo.ca
chenyan.xiong,ashwinsr,ankita.sharma,
dajose,paul.n.bennett@microsoft.com

Abstract

Dense retrieval (DR) methods conduct text retrieval by first encoding texts in the embedding space and then matching them by nearest neighbor search. This requires strong locality properties from the representation space, e.g., close allocations of each small group of relevant texts, which are hard to generalize to domains without sufficient training data. In this paper, we aim to improve the generalization ability of DR models from source training domains with rich supervision signals to target domains without any relevance label, in the zero-shot setting. To achieve that, we propose Momentum adversarial Domain Invariant Representation learning (MoDIR), which introduces a momentum method to train a domain classifier that distinguishes source versus target domains, and then adversarially updates the DR encoder to learn domain invariant representations. Our experiments show that MoDIR robustly outperforms its baselines on 10+ ranking datasets collected in the BEIR benchmark in the zero-shot setup, with more than 10% relative gains on datasets with enough sensitivity for DR models' evaluation. Source code is available at https://github.com/ji-xin/modir.

1 Introduction

Rather than matching texts in the bag-of-words space, Dense Retrieval (DR) methods first encode texts into a dense embedding space (Lee et al., 2019; Karpukhin et al., 2020; Xiong et al., 2021) and then conduct text retrieval using efficient nearest neighbor search (Chen et al., 2018; Guo et al., 2020; Johnson et al., 2021). With pre-trained language models and dedicated fine-tuning techniques, the learned representation space has significantly advanced the first stage retrieval accuracy of many language systems, including web search (Xiong et al., 2021), grounded generation (Lewis et al., 2020), open domain question answering (Karpukhin et al., 2020; Izacard and Grave, 2020), etc.

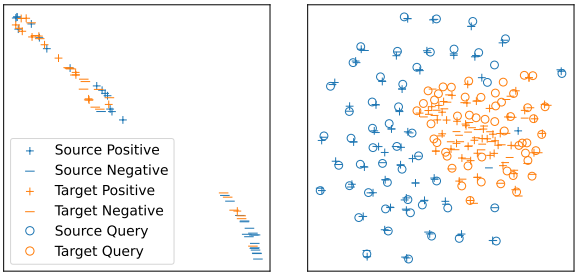


Figure 1: T-SNE plots of embedding space of a BERT reranker for q–d pairs (left) and ANCE dense retriever for queries/documents (right). Both models are trained on web search and transferred to medical search.

Purely using the learned embedding space for retrieval has raised concerns on the generalization ability, especially in scenarios without dedicated supervision signals. Many have observed diminishing advantages of DR models in various datasets if they are not fine-tuned with task-specific labels, i.e., in the zero-shot setup (Thakur et al., 2021). However, in many scenarios outside commercial web search, zero-shot is the norm. Obtaining training labels is difficult, expensive, and sometimes infeasible, especially in special domains (e.g., medical) where annotation requires strong expertise or is even prohibited because of privacy constraints. The lack of zero-shot ability hinders the democratization of advancements in dense retrieval from data-rich domains to everywhere else. Many equally, if not more important, real-world search scenarios still rely on unsupervised exact match methods that have been around for decades, e.g., BM25 (Robertson and Jones, 1976).

Within the search pipeline, the generalization of first stage DR models is notably worse than

* Work partly done during Ji's internship at Microsoft.

Findings of the Association for Computational Linguistics: ACL 2022, pages 4008 - 4020

subsequent reranking models (Thakur et al., 2021). Reranking models, similar to many classification models, only require a decision boundary between relevant and irrelevant query–document pairs (q–d pairs) in the representation space. In comparison, DR needs good local alignments across the entire space to support nearest neighbor matching, which is much harder to learn.

In Figure 1, we use t-SNE (van der Maaten and Hinton, 2008) to illustrate this difference. We show learned representations of a BERT-based reranker (Nogueira and Cho, 2019) and a BERT-based dense retriever (Xiong et al., 2021), in zero-shot transfer from web (Bajaj et al., 2016) to medical domain (Voorhees et al., 2021). The representation space learned for reranking yields two manifolds with a clear decision boundary; data points in the target domain naturally cluster with their corresponding classes (relevant or irrelevant) from the source domain, leading to good generalization. In comparison, the representation space learned for DR is more scattered. Target domain data points are grouped separately from those of the source domain; it is much harder for the learned nearest neighbor locality to generalize from source to the isolated target domain region.

In this paper, we present **Mo**mentum Adversarial **D**omain **I**nvariant **R**epresentations learning (MoDIR), to improve the accuracy of zero-shot dense retrieval (ZeroDR). We first introduce an auxiliary domain classifier that is trained to discriminate source embeddings from target ones. Then the DR encoder is not only updated to encode queries and relevant documents together in the source domain, but also trained adversarially to confuse the domain classifier and to push for a more domain invariant embedding space. To ensure stable and efficient adversarial learning, we propose a *momentum* method that trains the domain classifier with a momentum queue of embeddings saved from previous iterations.

Our experiments evaluate the generalization ability of dense retrieval with MoDIR using 15 retrieval tasks from the BEIR benchmark (Thakur et al., 2021). On these retrieval tasks from various domains including biomedical, finance, scientific, etc., MoDIR improves the zero-shot accuracy of two standard models, DPR (Karpukhin et al., 2020) and ANCE (Xiong et al., 2021). On tasks where evaluation labels have sufficient coverage for DR (Thakur et al., 2021), MoDIR's improvements are robust and significant, despite not using any target domain training labels. We also verify the necessity of the proposed momentum approach, without which the domain classifier fails to capture the domain gaps, and the adversarial training does not learn domain invariant representations, resulting in little improvement in ZeroDR.

We conduct further analyses to reveal interesting properties of MoDIR and its learned embedding space. During the adversarial training process, the target domain embeddings are gradually pushed towards the source domain and eventually absorbed as a subgroup of the source. In the learned representation space, our manual examinations find various cases where a target domain query is located close to source queries with similar information needs. This indicates that ZeroDR's generalization ability comes from the combination of information overlaps of source/target domains, and MoDIR's ability to identify the right correspondence between them.

2 Related Work

In this section, we recap related work in dense retrieval and adversarial domain adaptation.

Dense Retrieval Different from sparse first stage retrieval models, dense retrieval with Transformer-based models (Vaswani et al., 2017) such as BERT (Devlin et al., 2019) conducts retrieval in the dense embedding space (Lee et al., 2019; Chang et al., 2020; Guu et al., 2020; Karpukhin et al., 2020; Luan et al., 2021). Compared with its sparse counterparts, DR improves retrieval efficiency and also provides comparable or even superior effectiveness for in-domain datasets.

One important research question for DR is how to obtain meaningful negative training instances. DPR (Karpukhin et al., 2020) uses BM25 to find stronger negatives in addition to in-batch random negatives. RocketQA (Qu et al., 2021) uses cross-batch negatives and also filters them with a strong reranking model. ANCE (Xiong et al., 2021) uses an asynchronously updated negative index built from the being-trained DR model to retrieve global hard negatives.

Recently, challenges of ZeroDR have attracted much attention (Thakur et al., 2021; Zhang et al., 2021; Li and Lin, 2021). One way to improve ZeroDR is query generation (Liang et al., 2020; Ma et al., 2021), which first trains a doc2query model in the source domain and then applies the NLG model on target domain documents to generate queries.

The target domain documents and generated queries form weak supervision labels for DR models. Our method differs from them and focuses on *directly* improving the generalization ability of the learned representation space.

Adversarial Domain Adaptation Unsupervised domain adaptation (UDA) has been studied extensively for computer vision applications. For example, maximum mean discrepancy (Long et al., 2013; Tzeng et al., 2014; Sun and Saenko, 2016) measures domain difference with a pre-defined metric and explicitly minimizes the difference. Following the advent of GAN (Goodfellow et al., 2014), adversarial training for UDA is proposed: an auxiliary domain classifier learns to discriminate source and target domains, while the main classifier model is adversarially trained to confuse the domain classifier (Ganin and Lempitsky, 2015; Bousmalis et al., 2016; Tzeng et al., 2017; Luo et al., 2017; Vu et al., 2020; Vernikos et al., 2020; Tang and Jia, 2020). The adversarial method does not require pre-defining the domain difference metric, allowing more flexible domain adaptation. MoDIR builds upon the success of UDA methods and introduces a new momentum learning technique that is necessary to learn domain invariant representations in the ZeroDR setting.

3 Training Domain Invariant Representations for Dense Retrieval

In this work, we aim to improve generalization in ZeroDR under the unsupervised domain adaptation setting (UDA) (Long et al., 2016). Given a source domain with sufficient training signals, the goal is to transfer the DR model to a target domain, with access to its queries and documents, but without any relevance label. This is the common case when applying DR in real-world scenarios: in target domains (e.g., medical), example queries and documents are available but annotating relevance is expensive and may require domain expertise; on the other hand, in the source domain (e.g., web search), training signals are available at large scale (Ma et al., 2020; Thakur et al., 2021).

Our method, MoDIR, improves ZeroDR in the UDA setup by encouraging the DR models to learn a domain invariant representation space that facilitates the generalization from source to target. In this section, we describe (1) how to train a vanilla *dense retrieval model*, (2) how to train a *momentum domain classifier* to distinguish the two domains, and (3) how to *adversarially train* the DR model for domain invariant representations.

3.1 Training the Dense Retrieval Model

The standard design of DR is to use a dual-encoder model (Lee et al., 2019; Karpukhin et al., 2020), where an encoder g takes as input a query/document and encodes it into a dense vector. The relevance score of a q–d pair $x = (q, d)$ is computed using a simple similarity function:

$$r(x) = \text{sim}(g(q; \theta_g), g(d; \theta_g)), \tag{1}$$

where θ_g is the collection of parameters of g and sim is a vector similarity function.

The training of DR uses labeled q-d pairs in the source domain $x^s = (q^s, d^s)$. With relevant q–d pair as x^{s+} and irrelevant pair as x^{s-}, the encoder g is trained to minimize the *ranking loss* L_R:

$$\min_{\theta_g} \sum_{x^{s+}, x^{s-}} L_R(r(x^{s+}), r(x^{s-})), \tag{2}$$

where L_R is a ranking loss function. Our model follows its baseline DPR/ANCE to sample irrelevant documents using BM25 or global hard negatives. Without loss of generality, other modeling designs are kept the same with ANCE: g is fine-tuned from $\text{RoBERTa}_{\text{BASE}}$ (Liu et al., 2019); the output query/document embeddings are the hidden states of the last layer's `[CLS]` token; L_R is the Negative Log Likelihood (NLL) loss; sim is the dot product.

3.2 Estimating the Domain Boundary with Momentum Domain Classifier

To capture domain differences and enable adversarial learning for domain invariance, MoDIR introduces a domain classifier f to predict the probability of a query/document embedding $\mathbf{e}$ being source or target, and we use a linear classifier as f:

$$f(\mathbf{e}) = \text{softmax}(W_f \mathbf{e}). \tag{3}$$

The linear classifier has sufficient capacity to distinguish the two domains in the high-dimensional representation space—the main challenge is on training. As illustrated in Figure 1, DR's representation space focuses more on locality than forming manifolds, and therefore it is more difficult to learn the domain boundary in this case. If we simply update f using the same amount of data points as g, f fails to accurately estimate the domain boundary; on the other hand, if we naïvely feed in more

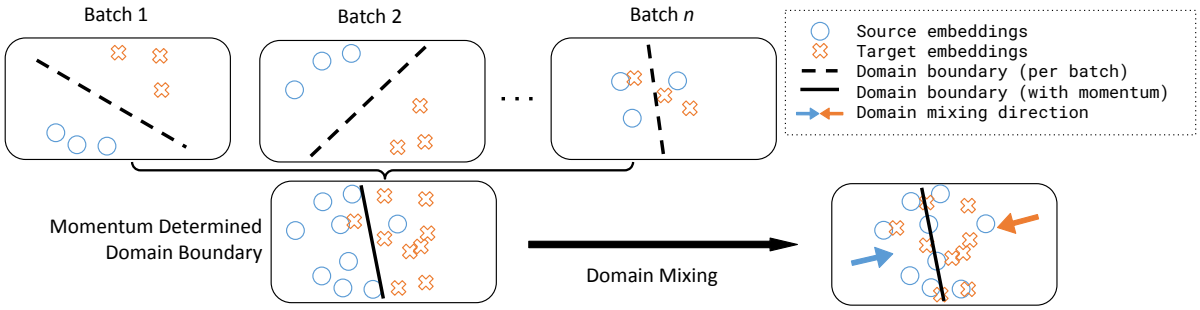


Figure 2: Momentum adversarial training provides a more accurate and robust estimation of the domain boundary in dense retrieval's embedding space.

data points for f, all these data points need to be encoded by the expensive encoder g, which makes the training process infeasibly slow.

To achieve the balance between accuracy and efficiency, we introduce the momentum method for the domain classifier, as shown in Figure 2. We maintain a *momentum queue* Q that records embeddings from multiple previous batches as the additional training data for f. Specifically, at each step, in addition to source domain training data x^s, we sample q–d pairs x^t from the target domain, and add embeddings of x^s and x^t to Q. The momentum queue Q at step k includes embeddings $\mathbf{e}_q/\mathbf{e}_d$ from source and target queries/documents for all recent n batches:

$$Q_k = \{\mathbf{e}_q, \mathbf{e}_d | (q,d) \in B_{k-n+1:k}\}, \quad (4)$$

where $B_{k-n+1:k}$ is the collection of all data points from the past n batches, including both source and target ones, and n is the *momentum step*. For simplicity of sampling, we use the 1:1 ratio between source/target data and also between positive/negative source data.

To ensure efficiency of the momentum method, all embeddings $\mathbf{e}$ from Q are *detached* from the encoder g. Take the query q^s as an example,

$$\mathbf{e}_{q^s} = \Phi(g(q^s; \theta_g)), \quad (5)$$

where Φ is the *stop-gradient* operator, i.e., gradients of $\mathbf{e}_{q^s}$ are not back propagated to θ_g. Since the linear classifier f is significantly smaller and faster than the transformer-based encoder g, this enables efficient training for f.

At each iteration, f is updated by repetitively minimizing the following discrimination loss L_D, computed with *all* embeddings from Q:

$$\min_{W_f} L_D(\mathbf{e}; f), \quad \mathbf{e} \in Q, \quad (6)$$

$$L_D(\mathbf{e}; f) = \begin{cases} -\log f(\mathbf{e}), & \mathbf{e} \text{ from source}, \\ -\log(1 - f(\mathbf{e})), & \mathbf{e} \text{ from target}, \end{cases} \quad (7)$$

where L_D is a standard classification loss. In this way, at each iteration, the domain classifier f is trained with more signals than the encoder g (the entire Q versus only one batch), ensuring accurate estimation of the domain boundary. The detached embeddings from Q also ensures training efficiency.

3.3 Adversarial Learning for Domain Invariant Representations

MoDIR adversarially trains the encoder g to generate domain invariant representations that are hard for f to distinguish. This is done by minimizing the adversarial loss L_M. Here we choose the widely used Confusion loss (Tzeng et al., 2017):

$$L_M(x; g, f) = -\frac{1}{2}\Big(\log f(g(q)) + \log f(g(d)) + \log(1 - f(g(q))) + \log(1 - f(g(d)))\Big), \quad (8)$$

where $x \in \{x^s, x^t\}$ is a q-d pair from either source or target domain. It reaches the minimum when the embeddings are domain invariant so that the domain classifier predict 50%-50% probability for all data. In order for the encoder to learn domain invariance, we freeze the domain classifier and update only the encoder when minimizing L_M:

$$\min_{\theta_g} \lambda \sum_{x \in \{x^s, x^t\}} L_M(x; g, f). \quad (9)$$

The hyperparameter λ balances the learning of DR ranking in the source domain (Equation (2)) and the learning of domain invariance (Equation (9)).

	Hole@10			nDCG@10				
	BM25	DPR	ANCE	BM25	DPR	DPR+MoDIR	ANCE	ANCE+MoDIR
TREC-COVID	10.6%	33.0%	22.4%	0.616	0.561	0.591(+5.3%)	0.654	0.676 (+3.4%)
Touché	29.8%	63.3%	56.9%	0.605	0.243	0.258(+6.2%)	0.284	0.315 (+10.9%)
DBPedia	41.3%	73.2%	65.8%	0.288	0.236	0.240(+1.7%)	0.281	0.284 (+1.1%)
NFCorpus	74.1%	85.2%	83.1%	0.297	0.208	0.212(+1.9%)	0.237	0.244 (+3.0%)
Quora	88.7%	87.3%	87.1%	0.742	0.842	0.848(+0.7%)	0.852	0.856 (+0.5%)
BioASQ	80.7%	92.0%	89.5%	0.514	0.232	0.247(+6.5%)	0.306	0.320 (+4.6%)
HotpotQA	87.7%	92.3%	90.9%	0.601	0.371	0.387(+4.3%)	0.456	0.462 (+1.3%)
FEVER	92.6%	92.1%	91.2%	0.648	0.589	0.607(+3.1%)	0.669	0.680 (+1.6%)
FiQA	93.4%	91.9%	91.5%	0.239	0.275	0.276(+0.4%)	0.295	0.296 (+0.3%)
ArguAna	92.7%	92.6%	92.6%	0.441	0.414	0.413(−0.2%)	0.415	0.418 (+0.7%)
NQ	94.9%	93.2%	92.6%	0.310	0.398	0.402(+1.0%)	0.446	0.442 (−0.9%)
SciFact	91.5%	93.2%	92.8%	0.620	0.478	0.476(−0.4%)	0.507	0.502 (−1.0%)
SCIDOCS	92.2%	94.4%	93.8%	0.156	0.108	0.108(+0.0%)	0.122	0.124 (+1.6%)
Climate-FEVER	95.7%	94.7%	94.1%	0.179	0.176	0.175(−0.6%)	0.198	0.206 (+4.0%)
CQADupStack	94.8%	95.2%	94.9%	0.316	0.281	0.280(−0.4%)	0.296	0.297 (+0.3%)

Table 1: Overall performance and label coverage (Hole rate) on tasks from BEIR. Relative improvements of MoDIR over its base DR model DPR/ANCE are shown in percentages. Datasets are ordered by ANCE's Hole rates, and datasets with lower Hole rates provide more accurate evaluation.

To summarize, for each training batch in the source domain, the domain classifier f and the encoder g are optimized by:

$$\min_{W_f} \quad L_D(\mathbf{e}; f), \quad \mathbf{e} \in Q, \tag{10}$$

$$\min_{\theta_g} \sum_{x^{s+}, x^{s-}} L_R(r(x^{s+}), r(x^{s-})) + \lambda \sum_{x \in \{x^s, x^t\}} L_M(x; g, f), \tag{11}$$

where f is trained to estimate the boundary between source/target and g is trained to provide domain invariant representations that also captures relevance matching in the source domain.

4 Experiments

This section describes experimental setups and evaluates the effectiveness of MoDIR. Furthermore, we dive deep into the importance of momentum training and properties of domain invariant embedding space, which provides new insights for ZeroDR.

4.1 Datasets

We choose the MS MARCO passage dataset (Bajaj et al., 2016) as the source domain dataset and choose the 15 publicly available datasets from the BEIR benchmark (Thakur et al., 2021) as target domain datasets (details in Appendix A). These datasets cover a large number of various domains, including biomedical, finance, scientific, etc. We treat each target domain dataset separately and produce an individual model for each of them, following the ZeroDR setting described in Section 3.

4.2 Effectiveness of MoDIR

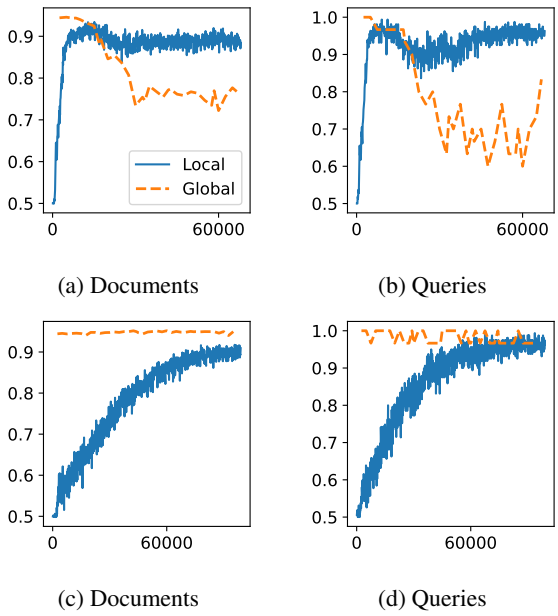
We build MoDIR on top of DPR and ANCE, but it can also be applied to other DR frameworks similarly. Table 1 shows the Hole rates and nDCG scores on the BEIR benchmark; we omit the Hole rates of MoDIR since they are very similar to its baseline DPR/ANCE's. We first discuss Hole rates and baseline selection, and then discuss effectiveness of each model.

Hole Rates and DR Evaluation A *hole* is an unlabeled q–d pair retrieved by a model, and the percentage of holes among all retrieved q–d pairs is the *Hole rate*. Datasets with high Hole rates for dense models are *less sensitive* to dense models' effectiveness (Xiong et al., 2021), and we therefore consider datasets with low Hole rates more important, since they provide more accurate measurements for ZeroDR. On the other hand, many of BEIR's datasets are annotated with candidates generated by some sparse retrieval models at the time of dataset construction, therefore the evaluation of these datasets is biased towards sparse models. Take TREC-COVID as an example, ANCE underperforms BM25 under the original annotation, but it achieves the state of the art (SOTA) after adding extra labels based on ANCE's prediction (Thakur et al., 2021).

Baselines Our baselines include BM25 (Robertson and Jones, 1976), DPR (Karpukhin et al., 2020), and ANCE (Xiong et al., 2021). The original DPR is trained on NQ (Kwiatkowski et al., 2019), but we instead train DPR on MARCO, which not only eliminates training dataset differences but also provides

Method	L_M	n	TREC-COVID	Touche
Single Repeat	Confusion	1	0.650	0.294
		1k	0.664	0.309
Momentum	Confusion	100	0.649	0.294
		1k	**0.676**	0.315
	Minimax	1k	0.666	0.322
	GAN	1k	0.641	**0.325**
Vanilla ANCE			0.654	0.284

Table 2: Ablation studies show that momentum is critical for learning domain invariant representation. Default settings are underlined and **best** scores are bold.



(a) Documents (b) Queries

(c) Documents (d) Queries

Figure 3: Global and Local Domain-Acc at different training steps with/without momentum (top/bottom).

better overall results. BEIR also reports results of other methods, such as docT5query (Nogueira et al., 2020), TAS-B (Hofstätter et al., 2021), GenQ (Ma et al., 2021), ColBERT (Khattab and Zaharia, 2020), etc. However, they are *not* directly comparable with MoDIR since they involve stronger supervision signals from rerankers (TAS-B), data augmentation from expensive sequence-to-sequence models (docT5query and GenQ), and high-latency late interaction (ColBERT). MoDIR instead directly improves the generalization ability of the representation space, and are orthogonal to these methods and can be combined for better performance.

Effectiveness Comparison From Table 1 we can see that MoDIR improves DPR and ANCE's overall effectiveness in the ZeroDR setting. On datasets with low Hole rates, where evaluation is more stable, the gains are significant; on datasets with high Hole rates, the gains are smaller but still stable. Moreover, to present a fair comparison in the realistic ZeroDR setting, results of MoDIR are obtained *without* hyperparameter tuning or checkpoint selection: in the ZeroDR setting, there is no access to relevance labels in the target domain during training/validation. For all target domain datasets, we keep most of the experimental settings the same with ANCE and evaluate checkpoints after the same number of training steps (details in Appendix B). This evaluation setup is the closest to ZeroDR in the real world, but it may not show the full potential and the best empirical results for MoDIR. We further study this in Section 4.5.

4.3 Effectiveness of Momentum Training and Ablation Studies

Our ablation studies evaluate the importance of the momentum method and the effects of other experimental setups. We compare different training setups against vanilla ANCE, using TREC-COVID and Touché which have the best label coverage (lowest Hole rates), and show the results in Table 2.

Firstly, we evaluate the effectiveness of not using the momentum queue: each iteration, the domain classifier is trained either with a *single* batch $n = 1$, or *repeat*[1] the current batch for $n = 1$k times. We can see that using a single batch fails to improve over ANCE, indicating the necessity of using more data to train the domain classifier; repeating the current batch also provides smaller improvements than using different batches from the queue. Secondly, we use a smaller momentum step $n = 100$ for momentum training, which also yields little improvement. This shows that n has to be sufficiently large for the momentum method to work, proving the necessity of our efficiency method to detach embeddings before storing them into the queue. Thirdly, we train MoDIR with two other choices of L_M from Equation (9): Minimax and GAN. GAN loss is less stable as described by Tzeng et al. (2017), while Minimax performs comparatively to Confusion. This shows that MoDIR can also be applied with other domain adaptation training methods.

4.4 Convergence of Adversarial Training with Momentum

In this experiment, we study how our momentum method helps adversarial training converge to a

[1]Concretely, for *repeat*, we update the domain classifier with the current batch's detached embeddings repetitively for n times (i.e., all using the same input embeddings).

	KNN-Source%				nDCG@10			
Checkpoint (→)	0	10k	30k	50k	0	10k	30k	50k
w/ Momentum	5.2%	6.2%	14.0%	17.2%	0.654	0.676	0.689	**0.724**
w/o Momentum	5.2%	5.4%	5.6%	5.6%	0.654	0.650	0.673	0.668

Table 3: K-Nearest Neighbor Source Percentage (KNN-Source%) and nDCG@10 scores after different number of training steps of ANCE with/without momentum, on TREC-COVID.

domain invariant embedding space. To quantify domain invariance, we use *Domain Classification Accuracy* (Domain-Acc), which includes two measurements based on the choice of domain classifier: (1) Directly take the domain classifier used in MoDIR's training (f in Section 3.2) and record its accuracy when applied to a new batch, which leads to *Local* Domain-Acc. (2) Randomly initialize a new domain classifier and train it globally on source and target embeddings, which leads to *Global* Domain-Acc. Global Domain-Acc measures the real degree of domain invariance: it is lower when embeddings of the two domains are not easily separable. Local Domain-Acc is an efficient approximation provided by the domain classifier f.

In Figure 3, we compare Global and Local Domain-Acc on the TREC-COVID dataset when training ANCE with/without momentum (without momentum is the *single* setting described in Section 4.3). With momentum, Local Domain-Acc quickly increases to be comparable with Global Domain-Acc. The domain classifier f (used in MoDIR's training) converges quickly and Global Domain-Acc starts to decrease, showing that embeddings from the two domains become less separable. Note that Local Domain-Acc does not decrease because f has seen and memorized almost all data, while Global Domain-Acc's domain classifier is always tested on unseen data for accurate results. This shows that momentum helps with the balance of adversarial training, ensuring its convergence towards a domain invariant representation space.

On the other hand, when momentum is not used, there exists a long-lasting gap between Local and Global Domain-Acc, showing that f does not capture the domain boundary well. As a result, the two domains remain (almost) linearly separable in the embedding space, as shown by the fact that Global Domain-Acc does not decrease, and the model fails to produce domain invariant representations.

4.5 Impact of Domain Invariance

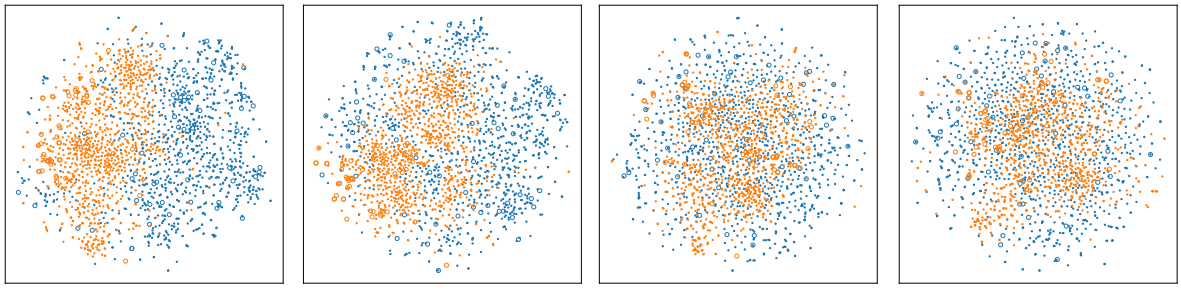
In this subsection, we study the behavior and benefits of ANCE+MoDIR in learning domain invariance. We focus on TREC-COVID as it provides the most robust evaluation for ZeroDR.

Learning Domain Invariance with Momentum We show how the momentum method gradually pushes for a domain invariant representation space. To measure how much the two domains are mixed together, we use *K-Nearest Neighbor Source Percentage* (KNN-Source%): We index source and target documents together; given a target domain query in the embedding space, we retrieve its top-100 nearest documents from the index, and calculate the percentage of source documents from the nearest neighbors; the average percentage for all target domain queries is reported. A higher KNN-Source% means that the target domain embeddings are surrounded by more source domain ones, indicating a more domain invariant representation space.

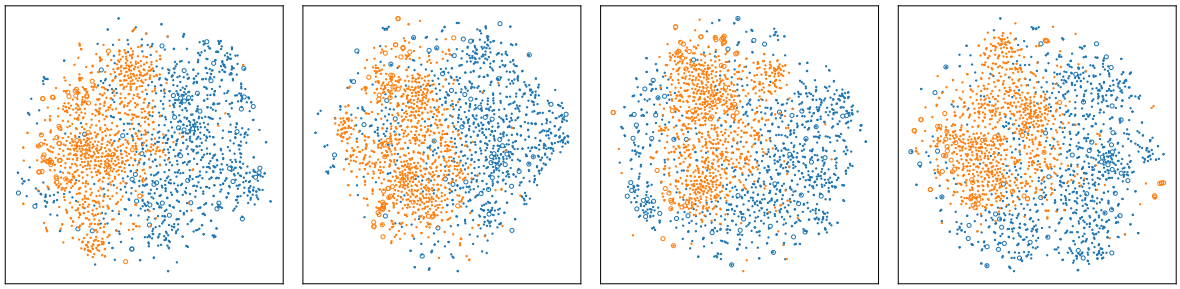
The results are in Table 3. With momentum, both KNN-Source% and nDCG gradually increase as training proceeds. This shows that when target domain embeddings are pushed towards the source domain, the ranking performance of the target domain also improves. On TREC-COVID, MoDIR eventually reaches **0.724**, which is the SOTA for first stage retrievers. On the other hand, without momentum (the *single* setting in Section 4.3), KNN-Source% and nDCG scores hardly increase.

We also use t-SNE (van der Maaten and Hinton, 2008) to visualize the learned representation space at different training steps in Figure 4. Before training with MoDIR, the two domains are well separated in the representation space learned by ANCE. With more MoDIR training steps, the target domains are pushed towards the source domain and gradually becomes a subset of it. Without momentum, the two domains remain separated, which is consistent with observations from Table 3.

ZeroDR Effectiveness VS Domain Invariance We study the correlation between ZeroDR ranking effectiveness and domain invariance. We use Global Domain-Acc as the indicator of domain invariance and plot it with the corresponding ZeroDR nDCG scores during training in Figure 5.

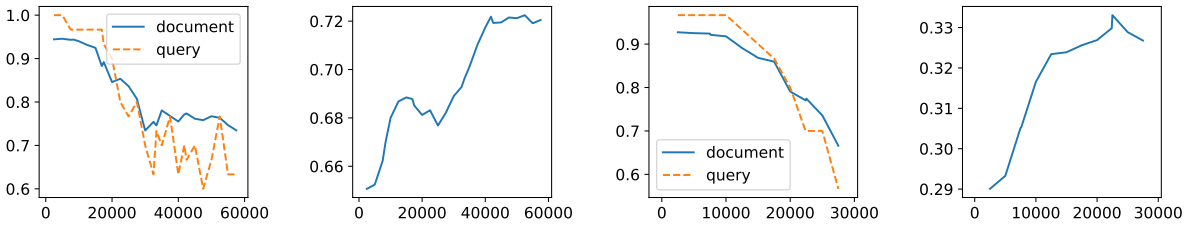


(a) MoDIR (0) (b) MoDIR (10k) (c) MoDIR (30k) (d) MoDIR (50k)



(e) w/o Mom. (0) (f) w/o Mom. (10k) (g) w/o Mom. (30k) (h) w/o Mom. (50k)

Figure 4: T-SNE of the representation space after different training steps (in the parentheses), with/without momentum. Blue: source (MARCO); orange: target (TREC-COVID).



(a) Global Domain-Acc (b) nDCG@10 (c) Global Domain-Acc (d) nDCG@10

Figure 5: Global Domain-Acc and target domain ZeroDR nDCG scores at different training steps: TREC-COVID (left two) and Touché (right two).

Global Domain-Acc starts at near 100% and decreases as training proceeds, showing that source and target embeddings are almost linearly separable at the beginning but are gradually pushed together. ZeroDR accuracy improves as Global Domain-Acc decreases, showing that domain invariance is the source of ZeroDR's improvements. We also record that the DR accuracy on the source domain (MARCO) decreases by no more than 0.5%. This indicates that the high dimensional embedding space has sufficient capacity to learn domain invariant representations while maintaining relevance matching in the source domain.

4.6 Case Study

We show two cases of queries from TREC-COVID and their nearest MARCO queries before and after MoDIR training in Table 4. In the first case, MoDIR pays more attention to "transmission", and potentially retrieves more documents about the transmission of diseases, thereby improving the nDCG score; documents about "coronavirus" are also likely to be retrieved by MoDIR since it is a very noticeable word. In the second case, it focuses on "mRNA" more than "vaccine". However, since the mRNA vaccine is relatively new[2] with few appearances in the MARCO dataset, the shift in focus fails to improve MoDIR for this query.

These examples help reveal the source of generalization ability on ZeroDR. For the DR models to be able to generalize, the source domain itself needs to include relevance information that resembles the target domain's needs; if there is no such information,

[2]The first mRNA vaccine was approved in 2020, according to https://en.wikipedia.org/wiki/MRNA_vaccine.

Target	what are the transmission routes of coronavirus?	nDCG@10 gain: 0.23
Source Before	• what is the coronavirus • what are symptoms of coronavirus	• incubation period for coronavirus
Source After	• countries where guinea worm is transmitted • through which body system are cancer cells able to travel to different locations in the body?	• what is the most common method of hiv transmission
Target	what is known about an mRNA vaccine for the SARS-CoV-2 virus?	nDCG@10 gain: −0.12
Source Before	• is there a vaccine for hepatitis • shingles vaccination needed for those without chickenpox	• is there a vaccine for tuberculosis
Source After	• what makes rna • what is the mmr vaccine called	• what is used to make mrna

Table 4: Case study: nearest source queries of a target query before and after MoDIR training.

as in the second example, generalization becomes a hard challenge. When the source domain has such coverage, MoDIR is able to align target queries to source ones with similar information needs in its domain invariant representation space, and such alignments enable DR models to generalize.

5 Conclusion and Future Work

In this paper, we present MoDIR, a new representation learning method that improves the zero-shot generalization ability of dense retrieval models. We first show that dense retrieval models differ from classification models in that they emphasize locality properties in the representation space. Then we present a momentum-based adversarial training method that robustly pushes text encoders to provide a more domain invariant representation space for dense retrieval. Our experiments demonstrate that, compared with ANCE, a recent SOTA DR model, MoDIR's improvements are robust overall and significant on datasets where ZeroDR's evaluation is more accurate.

We conduct a series of studies to show the effects of our momentum method in learning domain invariant representations. Without momentum, the adversarial learning is unstable. The inherent variance of the DR embedding space hinders the convergence of the domain classifier. With momentum training, the model fuses the target domain data into the source domain representation space and discovers related information from the source domain, thus improving generalization of ZeroDR.

We view MoDIR as an initial step of zero-shot dense retrieval, an area that democratizes the rapid advancements in search technologies to many real-world scenarios. Our approach inherits the success of domain adaptation techniques and upgrades them by addressing the unique challenges of ZeroDR. Understanding the dynamics of dense retrieval is an important future direction for not only representation learning research but also real-world applications.

Acknowledgments

We thank anonymous reviewers for their constructive feedback.

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.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval. In *International Conference on Learning Representations*.

Qi Chen, Haidong Wang, Mingqin Li, Gang Ren, Scarlett Li, Jeffery Zhu, Jason Li, Chuanjie Liu, Lintao Zhang, and Jingdong Wang. 2018. *SPTAG: A library for fast approximate nearest neighbor search*.

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.

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*.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France. PMLR.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

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 the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3887–3896. PMLR.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

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*, SIGIR '21, page 113–122, New York, NY, USA. 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.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

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*, SIGIR '20, 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.

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 *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Minghan Li and Jimmy Lin. 2021. Encoder adaptation of dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2110.01599*.

Davis Liang, Peng Xu, Siamak Shakeri, Cicero Nogueira dos Santos, Ramesh Nallapati, Zhiheng Huang, and Bing Xiang. 2020. Embedding-based zero-shot retrieval through query generation. *arXiv preprint arXiv:2009.10270*.

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*.

Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S. Yu. 2013. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. 2016. Unsupervised domain adaptation with residual transfer networks. *arXiv preprint arXiv:1602.04433*.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Zelun Luo, Yuliang Zou, Judy Hoffman, and Li F Fei-Fei. 2017. Label efficient learning of transferable representations acrosss domains and tasks. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2020. Zero-shot neural retrieval via domain-targeted synthetic query generation. *arXiv preprint arXiv:2004.14503*.

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.

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.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.

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.

Stephen E. Robertson and Karen Spärck Jones. 1976. Relevance weighting of search terms. *JASIS*, 27(3):129–146.

Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer.

Hui Tang and Kui Jia. 2020. Discriminative adversarial domain adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5940–5947.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

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.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all

you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Giorgos Vernikos, Katerina Margatina, Alexandra Chronopoulou, and Ion Androutsopoulos. 2020. Domain Adversarial Fine-Tuning as an Effective Regularizer. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3103–3112, Online. Association for Computational Linguistics.

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).

Thuy-Trang Vu, Dinh Phung, and Gholamreza Haffari. 2020. Effective unsupervised domain adaptation with adversarially trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6163–6173, Online. Association for Computational Linguistics.

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.

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*.

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.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A multi-lingual benchmark for dense retrieval. *arXiv preprint arXiv:2108.08787*.

A Datasets Details

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

- General-domain (Wikipedia): DBPedia (Hasibi et al., 2017), HotpotQA (Yang et al., 2018), FEVER (Thorne et al., 2018), and NQ (Kwiatkowski et al., 2019).
- Bio-medical: TREC-COVID (Voorhees et al., 2021), NFCorpus (Boteva et al., 2016), and BioASQ (Tsatsaronis et al., 2015).
- Finance: FiQA (Maia et al., 2018).
- Controversial arguments: Touché (Bondarenko et al., 2020) and ArguAna (Wachsmuth et al., 2018).
- Duplicate questions: Quora (Thakur et al., 2021) and CQADupStack (Hoogeveen et al., 2015).
- Scientific: SciFact (Wadden et al., 2020), SCIDOCS (Cohan et al., 2020), and Climate-FEVER (Diggelmann et al., 2020)

B Detailed Experimental Settings

We follow the design of ANCE for the DR encoder's modeling and training. We initialize the encoder with the publicly released checkpoints: "ANCE-warmup" for DPR+MoDIR and "ANCE-passage" for ANCE+MoDIR.[3] We randomly initialize the domain classifier. Detailed hyperparameter choices are shown in Table 5. We also use an exponential decay routine for the hyperparameter λ to improve training stability, where the value is reduced continuously and shrunk to half every 10k steps.

Hyperparameter	Value
Same as ANCE	
Learning rate for θ_g	1e-6
Effective batch size	16
Maximum Query Length	64
Maximum Document Length	512
New for MoDIR	
Learning rate for W_f	5e-6
Early stopping steps	10k
Momentum step n	1k
Initial λ	1.0

Table 5: Detailed hyperparameter choices of MoDIR.

[3] https://github.com/microsoft/ANCE.