# ZERO-SHOT DENSE RETRIEVAL WITH MOMENTUM ADVERSARIAL DOMAIN INVARIANT REPRESENTATIONS

**Anonymous authors**
Paper under double-blind review

## 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, i.e, the close allocations of each small group of relevant texts, which is 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 relevant labels, in the zero-shot setting. To achieve that, we propose Momentum adversarial Domain Invariant Representation learning (MoDIR), which introduces a momentum method in the DR training process to train a domain classifier on the source versus target, 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 from the BEIR benchmark in the zero-shot setup, with more than 10% relative gains on datasets where the evaluation of DR models is sensitive enough. Source code of this paper will be released.

## 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., 2019b; 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 pretrained language models and dedicated fine-tuning techniques, the learned representation space has significantly advanced the first stage retrieval accuracy of many systems, including web search (Xiong et al., 2021), open domain question answering (Karpukhin et al., 2020; Izacard & Grave, 2020), grounded generation (Lewis et al., 2020), etc.

Retrieval purely with learned embedding space has raised concerns on their generalization ability, especially in scenarios without the luxury of dedicated supervision signals. Many have observed diminishing advantage 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 and sometimes infeasible, for example, in the medical domain where annotation requires strong expertise or is even prohibitive 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 are still relying on unsupervised word-based exact match methods like BM25, which are developed decades ago (Robertson & Jones, 1976).

Even within the search system, generalization ability of first stage DR models is notably worse than 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 in the entire space to support nearest neighbor matching, which is much harder for representation learning.

In Figure 1, we use t-SNE (van der Maaten & Hinton, 2008) to illustrate this challenge, with learned representations of a standard BERT-based reranker (Nogueira & Cho, 2019) and a BERT-based dense retriever (Xiong et al., 2021), in zero-shot transfer from Web (Bajaj et al., 2016) to Med (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

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 nearly impossible for the learned nearest neighbor locality from the source domain to generalize 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 generalization ability 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.

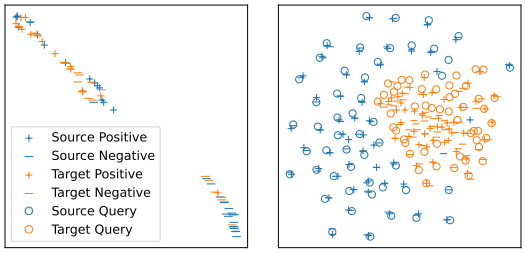



Figure 1: T-SNE plots of embedding space of a BERT reranker for q–d pairs and ANCE dense retriever for queries and documents. All models are trained on MS MARCO web search as the source domain (blue) and applied on TREC-COVID medical search as the target domain (orange).

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 significantly improves the zero-shot accuracy of ANCE (Xiong et al., 2021), a recent state-of-the-art DR model trained with web search data. Without using any target domain training labels, the improvements from MoDIR are stable, robust, and also significant on tasks where evaluation labels have sufficient coverage for DR (Thakur et al., 2021). Our studies also verify the necessity of our momentum approach, without which, the domain classifier fails to capture the domain gaps, and the adversarial training does not learn domain invariant representations, resulted in little improvement in ZeroDR.

Our further analyses reveal several interesting behaviors of MoDIR and its learned embedding space. During the adversarial training process, the target domain embeddings are gradually pushed towards the source domain, eventually absorbed as a subgroup of the source. In the learned representation space, our manual examinations found various cases where a target domain query is located closely to source queries resembling similar information needs. This indicates 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.

The rest of this paper as organized as follows: Next section presents how MoDIR learns domain invariant representations for ZeroDR; Section 3 and Section 4 discuss our experimental settings and evaluation results; We recap related works in Section 5 and conclude in Section 6.

## 2 Training Domain Invariant Representations for Dense Retrieval

In this work, we aim to improve the zero-shot ability of DR in 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 data but not any labels. This is common when applying DR in real-world scenarios: in target domains such as medical, example queries and documents are available but relevance annotations require domain expertise, while in the source domain such as web, training signals are available at large scale (Ma et al., 2020; Thakur et al., 2021).

Our method, MoDIR, improves ZeroDR in UDA setup by encouraging the DR models to learn a domain invariant representation space to facilitate the generalization from source to target. The rest of this section describes how MoDIR enforces domain invariance for a *dense retrieval model* (Sec. 2.1) using a *momentum domain classifier* (Sec. 2.2) to distinguish the two domains, and to *adversarially train* (Sec. 2.3) the DR model for more domain invariant representations.

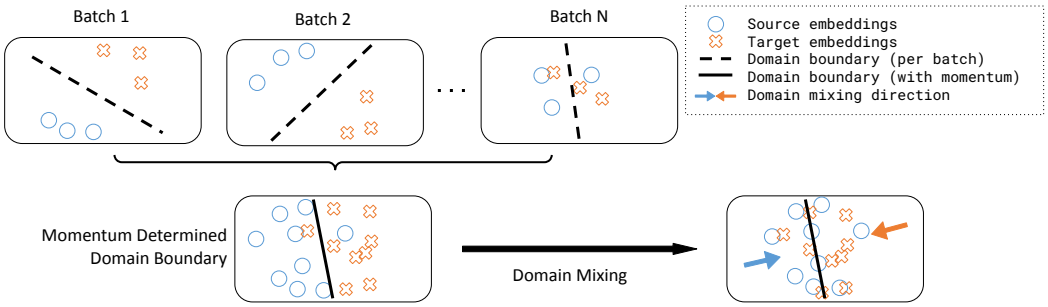



Figure 2: Momentum adversarial training provides a more accurate and robust estimation of the domain boundary in dense retrieval's embedding space.

### 2.1 Training the Dense Retrieval Model

The standard design of DR is to use a dual-encoder model (Lee et al., 2019b; Karpukhin et al., 2020), where an encoder $g$ takes as input a query or a document and encodes it into a dense vector, and then the relevance score of a query–document 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 $\theta_g$ is the collection of parameters of $g$ and sim is a similarity function that supports efficient nearest neighbor search (Johnson et al., 2021), for example, cosine similarity or dot product.

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 DR 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 standard ranking loss function. In this paper, without loss of generality, we inherit the settings of ANCE (Xiong et al., 2021) that sample negatives $x^{s-}$ using the DR model being trained. Other components are also kept the same with ANCE: $g$ is fine-tuned from RoBERTa-base (Liu et al., 2019) and outputs the embedding of the last layer's [CLS] token, $L_R$ is the Negative Log Likelihood (NLL) loss, and sim is the dot product.

### 2.2 Estimating the Domain Boundary with Momentum Domain Classifier

To capture the domain differences between source and target and enable adversarial learning for domain invariance, MoDIR introduces a domain classifier $f$ on top of the DR model's query and document embeddings to predict their probability of being source or target. We simply use a linear layer on top of data embeddings $\mathbf{e}$ as the model architecture of $f$:

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

The linear layer is often sufficient to distinguish both domains in the high-dimensional representation space. The challenge is more on the training side. As illustrated in Figure 1, the representation space learned for DR focuses more on locality than forming manifolds. Learning $f$ using a large number of data points enumerated after each DR model update is costly, while updating $f$ per data batch may result in an unstable estimation of domain boundary given the scatted representation space.

We introduce momentum learning to balance the efficiency and robustness of the domain classifier learning. As shown in Figure 2, we maintain a *momentum queue* $Q$ that includes embeddings from multiple past batches as the training data for $f$ in each iteration. Specifically, for each source domain training data $x^s$, we sample q–d pairs $x^t$ from the target domain, and add their embeddings to $Q$. The momentum queue $Q$ at step $k$ includes embeddings from source and target for all recent $n$ batches:

$$Q_k = \{\mathbf{e}_{q^s}, \mathbf{e}_{d^s}, \mathbf{e}_{q^t}, \mathbf{e}_{d^t} | (q^s, d^s, q^t, d^t) \in B_{k-n+1:k}\}, \tag{4}$$

where $B_{k-n+1:k}$ are the data from the past $n$ batches, with $n$ as the *momentum step*. We ensure the 1:1 ratio between source and target and also 1:1 between positive and negative source data.

Note that $\mathbf{e}$ is the *detached* embedding, for example, of the query $q^s$:

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

where $\Phi$ is the *stop-gradient* operator, i.e., gradients of $\mathbf{e}_{q^t}$ will not be back propagated to $\theta_g$. This enables efficient momentum learning as only the embedding vectors are maintained in $Q$.

At each iteration, the domain classifier is updated by minimizing the following discrimination loss:

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

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

where $L_D$ is a standard classification loss. In this way, the domain classifier is trained with signals from multiple batches, leading to a faster and more robust estimation of the domain boundary.

### 2.3 Adversarial Learning for Domain Invariant Representations

With an estimated domain boundary from the domain classifier $f$, MoDIR then adversarially trains the encoder $g$ to generate domain invariant representations that $f$ cannot distinguish. Here we choose the Confusion loss widely used in domain adaptation (Tzeng et al., 2017):

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

where $x \in \{x^s, x^t\}$ is a q-d pair from either source or target, as the confusion loss aims to push for random classification probability for any data points. It reaches the minimum when the embeddings are domain invariant and the domain classifier predict 50%-50% probability for all data.

To push for domain invariance, we freeze the domain classifier and update parameters of the encoder:

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

We use the hyperparameter $\lambda$ to balance the learning of DR ranking in the source domain (Equation (2)) and the learning of domain invariance (Equation (9)).

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 while capturing the relevance matches in the source domain.

## 3 Experimental Setups

**Datasets** We choose the MS MARCO passage dataset (Bajaj et al., 2016) as the source domain dataset and choose the 15 publicly available datasets gathered in the BEIR benchmark (Thakur et al., 2021) as target domain datasets. 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 standard unsupervised domain adaptation setup (Long et al., 2016). Details of the datasets can be found in Appendix A.

**Evaluation for DR** Target domain datasets do not always have an ideal coverage for relevance labels. The annotation procedure of many datasets requires some retrieval models to generate candidates for labeling, which were mainly sparse models back at their time of construction. This makes their evaluation biased towards sparse models and the evaluation on dense retrieval models less sensitive, often with high Hole rates (a *hole* is a predicted q–d pair without annotation) for dense

Table 1: Overall performance and label coverage (hole rate) in the tasks collected in BEIR. Relative improvements of MoDIR over its base DR model ANCE is shown in percentages. Datasets are ordered by ANCE's Hole rate. Lower hole rate indicates more robust evaluation.

| | **Hole@10** | | | **nDCG@10** | | | | |
|---|---|---|---|---|---|---|---|---|
| | BM25 | ANCE | MoDIR | BM25 | DPR-(NQ/MARCO) | | ANCE | MoDIR |
| TREC-COVID | 10.6% | 22.4% | 19.2% | 0.616 | 0.332 | 0.561 | 0.654 | 0.676 (+3.4%) |
| Touché | 29.8% | 56.9% | 53.5% | 0.605 | 0.127 | 0.243 | 0.284 | 0.315 (+10.9%) |
| DBPedia | 41.3% | 65.8% | 65.0% | 0.288 | 0.263 | 0.236 | 0.281 | 0.284 (+1.1%) |
| NFCorpus | 74.1% | 83.1% | 82.6% | 0.297 | 0.189 | 0.208 | 0.237 | 0.244 (+3.0%) |
| Quora | 88.7% | 87.1% | 87.0% | 0.742 | 0.248 | 0.842 | 0.852 | 0.856 (+0.5%) |
| BioASQ | 80.7% | 89.5% | 89.1% | 0.514 | 0.127 | 0.232 | 0.306 | 0.320 (+4.6%) |
| HotpotQA | 87.7% | 90.9% | 90.7% | 0.601 | 0.391 | 0.371 | 0.456 | 0.462 (+1.3%) |
| FEVER | 92.6% | 91.2% | 91.1% | 0.648 | 0.562 | 0.589 | 0.669 | 0.680 (+1.6%) |
| FiQA | 93.4% | 91.5% | 91.5% | 0.239 | 0.112 | 0.275 | 0.295 | 0.296 (+0.3%) |
| ArguAna | 92.7% | 92.6% | 92.6% | 0.441 | 0.175 | 0.414 | 0.415 | 0.418 (+0.7%) |
| NQ | 94.9% | 92.6% | 92.6% | 0.310 | 0.474 | 0.398 | 0.446 | 0.442 (−0.9%) |
| SciFact | 91.5% | 92.8% | 92.9% | 0.620 | 0.318 | 0.478 | 0.507 | 0.502 (−1.0%) |
| SCIDOCS | 92.2% | 93.8% | 93.7% | 0.156 | 0.077 | 0.108 | 0.122 | 0.124 (+1.6%) |
| Climate-FEVER | 95.7% | 94.1% | 93.9% | 0.179 | 0.148 | 0.176 | 0.198 | 0.206 (+4.0%) |
| CQADupStack | 94.8% | 94.9% | 94.9% | 0.316 | 0.153 | 0.281 | 0.296 | 0.297 (+0.3%) |

models (Xiong et al., 2021; Thakur et al., 2021). In fact, ANCE underperforms sparse methods such as BM25 on TREC-COVID with the original annotation, but after adding extra labels based on ANCE's prediction, its scores greatly improve, achieving the state of the art (Thakur et al., 2021). Nevertheless, TREC-COVID is the dataset with the lowest hole rates on DR models since participating systems include dense ones, and is one of the best to measure the progress of ZeroDR.

**Model Validation in ZeroDR** In the ZeroDR setting, there is no access to relevance labels in the target domain during training/validation. Therefore, choosing the optimal hyperparameters is impossible without directly tuning on the test set. In our experiments, most of our hyperparameters are kept the same with ANCE. We also use exactly the same experimental setting and evaluate checkpoints after a fixed number of training steps (10k) for all target domain datasets. This evaluation setup may not yield the optimal empirical results, but it is the closest to ZeroDR in the real world. Please refer to Appendix B for detailed hyperparameters.

**Baselines** As a first stage retrieval method, MoDIR's baselines include BM25 (Robertson & Jones, 1976), DPR (Karpukhin et al., 2020), and ANCE (Xiong et al., 2021). The original DPR is trained on NQ (Kwiatkowski et al., 2019). We train another DPR model on MARCO to eliminate training dataset differences. The nDCG scores of BM25, DPR-NQ, and ANCE are taken from the BEIR paper (verified to be consistent with our runs). DPR-MARCO and MoDIR are from our own.

BEIR also reports results of other retrieval methods, such as docT5query (Nogueira et al., 2020), TAS-B (Hofstätter et al., 2021), GenQ (Ma et al., 2021), ColBERT (Khattab & Zaharia, 2020), etc. However, they are not directly comparable with MoDIR since they may include stronger supervision signals, data augmentation, and/or expensive late interaction, and are thus orthogonal with MoDIR and can be combined for better empirical results. Our main baseline is ANCE, which MoDIR is built upon and is also shown to be the state of the art on TREC-COVID (Thakur et al., 2021).

## 4 Results and Analyses

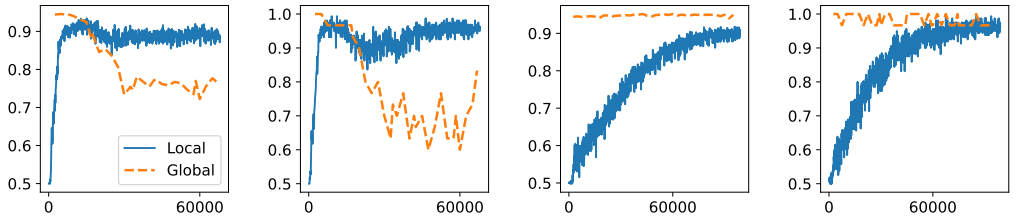
This section evaluates the effectiveness of MoDIR, its momentum training, and the benefits of domain invariant representations.

### 4.1 Effectiveness of Proposed Methods

Table 1 shows the overall ZeroDR accuracy of MoDIR and baselines on the BEIR benchmark (Thakur et al., 2021). MoDIR improves ANCE's overall effectiveness in the ZeroDR setting. On datasets with low hole rates (good label coverage), the gains are significant; on datasets with high hole rates, which are less sensitive to DR model improvement, the gains are less significant but still stable.

Table 2: Ablation studies on TREC-COVID and Touché. Underlined scores are generated by the default experimental setting, and bold scores are the highest in the row. Scores are nDCG@10.

| | **w/o Momentum** | | **MoDIR Variants w/ Momentum** | | | | | **ANCE** |
|---|---|---|---|---|---|---|---|---|
| **Adversarial Loss** | Confusion | | Minimax | GAN | Confusion | | | |
| **Momentum Step** $N$ | 1 | 1k | 1k | 1k | 100 | 1k | 5k | |
| TREC-COVID | 0.650 | 0.664 | 0.666 | 0.641 | 0.649 | **<u>0.676</u>** | 0.600 | 0.654 |
| Touché | 0.294 | 0.309 | 0.322 | 0.325 | 0.294 | <u>0.315</u> | **0.333** | 0.284 |




(a) w/ Mom.: document (b) w/ Mom.: query (c) w/o Mom.: document (d) w/o Mom.: query

Figure 3: Global and Local Domain-Acc when momentum is/isn't used at different training steps.

Moreover, results of MoDIR are obtained *without* hyperparameter tuning or checkpoint selection, and therefore present a fair comparison in the realistic ZeroDR setting.

### 4.2 Ablation Studies

Our ablation studies evaluate the importance of the momentum method and the effects of other experimental setups. We use the two datasets with the best label coverage, TREC-COVID and Touché, and show the results in Table 2. MoDIR's default setting is underlined.

We first evaluate the accuracy of MoDIR *without* the momentum method, i.e., we do not maintain the momentum queue, but simply update the domain classifier with embeddings of the current batch. Without momentum, the improvement over ANCE diminishes.

We then evaluate MoDIR with other two choices of adversarial loss: Minimax and GAN (Tzeng et al., 2017). GAN loss is less stable as expected (Tzeng et al., 2017). Minimax performs comparatively to Confusion. MoDIR can be applied with other domain adaptation training methods.

We also vary the momentum step $N$ without changing the rest experimental settings. We find that $N$ mainly impacts the balance between learning the domain invariance and the retrieval loss, and is an important hyperparameter for MoDIR.

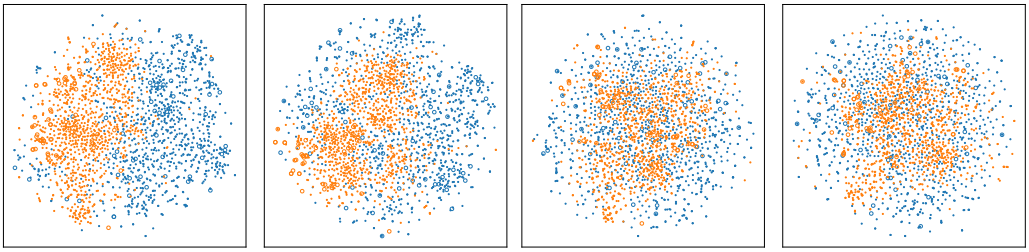
### 4.3 Convergence of Adversarial Training with Momentum

This group of experiments evaluates the impact of momentum in adversarial training. We use *Domain Classification Accuracy* to indicate the domain invariance, which includes two measurements based on the choices of domain classifier: (1) The domain classifier is trained globally on source and target domain embeddings until convergence, which leads to *Global*-Domain-Acc. (2) We take the domain classifier during in MoDIR's training ($f$ in Section 2.2), and record its accuracy when it is applied on a new batch, which leads to *Local*-Domain-Acc. Global-Domain-Acc measures the real degree of domain invariance: it is lower when the embeddings of the two domains are not easily separable. Local-Domain-Acc is an approximation provided by the domain classifier $f$. A large gap between local and global accuracy means the domain boundary estimated by $f$ is inaccurate.

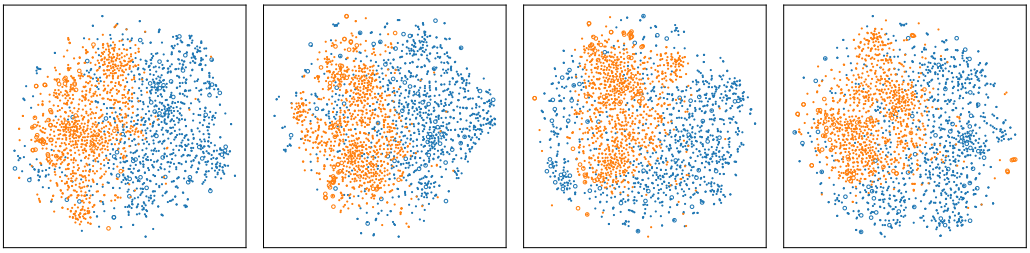
We compare Global- and Local- Domain-Acc on the TREC-COVID dataset when momentum is/isn't used in Figure 3. With momentum, Local-Domain-Acc quickly increases to be comparable with Global-Domain-Acc. The domain classifier $f$ (used in MoDIR's training) converges quickly and Global-Domain-Acc starts to decrease. Embeddings from the two domains are becoming less separable as the result of effective adversarial training. Note that Local-Domain-Acc does not

Table 3: K-Nearest Neighbor Source Percentage and nDCG@10 scores after different number of training steps: comparison between with and without momentum, on TREC-COVID.

| | 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 |



(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 with different continuous training (steps) after ANCE, with or without momentum. Blue: source (MARCO); orange: target (TREC-COVID).

decrease because $f$ has seen and memorized almost all data, while Global-Domain-Acc's domain classifier is always tested on unseen data.

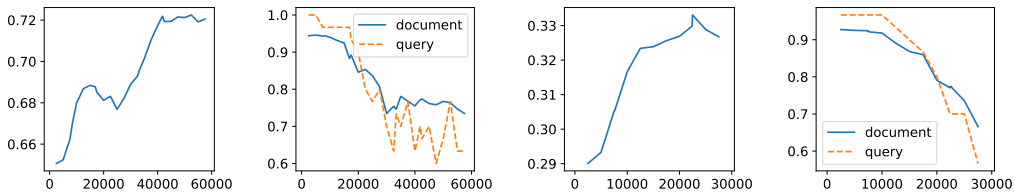
On the other hand, when momentum is not used, Local-Domain-Acc is much lower than Global-Domain-Acc for a long time, showing that $f$ does not capture the domain boundary well. As a result, the two domains remain linearly separable, as shown by Global-Domain-Acc, and the model fails to learn a domain invariant representation space.

### 4.4 Impact of Domain Invariant Representation Learning

This group of experiments studies the behavior and benefits of MoDIR in learning domain invariance. We focus on TREC-COVID in these analyses as it provides the most robust evaluation for ZeroDR.

**Learned Domain Invariance** We show how the momentum method gradually pushes the encoder to produce more domain invariant representations. 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 domain documents from the nearest neighbors. A higher KNN-Source% means that the target domain embeddings are more mixed with source domain ones, indicating a more domain invariant representation space.

The results are shown in Table 3. With momentum, KNN-Source% gradually increases with more training steps. Target domain queries are surrounded by more source domain documents, and the nDCG@10 score of TREC-COVID also improves. On TREC-COVID, MoDIR eventually reaches a state-of-the-art 0.724 for first stage retrievers. When momentum is not used, KNN-Source% and nDCG scores hardly increase.




(a) TREC-COVID nDCG@10 (b) TREC-COVID global-Domain-Acc (c) Touché nDCG@10 (d) Touché global-Domain-Acc

Figure 5: Target domain effectiveness and Global-Domain-Acc at different training steps ($x$-axis).

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

| | | |
|---|---|---|
| Target | what are the transmission routes of coronavirus? | nDCG@10 gain: 0.23 |
| Source Before | • what is the coronavirus<br>• what are symptoms of coronavirus | • incubation period for coronavirus |
| Source After | • countries where guinea worm is transmitted<br>• 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<br>• shingles vaccination needed for those without chickenpox | • is there a vaccine for tuberculosis |
| Source After | • what makes rna<br>• what is the mmr vaccine called | • what is used to make mrna |

We also use t-SNE (van der Maaten & 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 despite adversarial learning, the same as observed in Table 3.

**ZeroDR Accuracy w.r.t. Domain Invariance** The last experiment studies the correlation between ZeroDR accuracy and domain invariance. We use Global-Domain-Acc as the indicator of domain invariance and plot it with the corresponding ZeroDR accuracy during training in Figure 5.

Global-Domain-Acc starts at near 100%. The source and target domain are linearly separable with our one-layer domain classifier. It decreases as training proceeds, as the learned representation space is more domain invariant, and the ZeroDR accuracy improves alongside. This shows that when the embeddings are more domain invariant, ranking performance in the target domain improves. We also record the DR accuracy on the source domain (MARCO). It slightly decreases, but not by much (about $0.5\%$). This indicates that the high dimensional embedding space has sufficient capacity to capture relevance matching in the source domain while learning domain invariant representations.

### 4.5 Case Study

We show two examples of queries from TREC-COVID and their nearest MARCO queries before and after MoDIR training in Table 4. In the first case, MoDIR shifts the focus from "coronavirus" to "transmission", and potentially retrieves more documents about the transmission of diseases, thereby improving the nDCG score. In the second case, it also shifts the focus from "vaccine" to "mRNA". However, since the mRNA vaccine is relatively new[1] with few appearances in the MARCO dataset, MoDIR fails to improve model effectiveness 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 information that covers the relevance needs of the target domain; if there is no such information, as in the second example, generalization becomes a challenge. Where 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.

[1] `https://en.wikipedia.org/wiki/MRNA_vaccine`.

## 5 Related Work

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

**Dense Retrieval** Compared to conventional sparse methods in first stage retrieval, dense retrieval (DR) with Transformer-based models (Vaswani et al., 2017) such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) conduct retrieval in the dense embedding space (Lee et al., 2019a; 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 of the most important research questions 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 of being trained DR model to retrieve global hard negatives.

Recently, the challenges of DR models' generalization ability in ZeroDR has attracted much attention (Thakur et al., 2021; Zhang et al., 2021; Li & Lin, 2021). One way to improve ZeroDR is by synthetic query generation (Liang et al., 2020; Ma et al., 2021), which first trains a doc2query model that learns to generate queries in the source domain given their relevant documents, and then applies the NLG model on target domain documents to generate queries. The target domain documents and generated queries form weak supervision labels in the target domain to train DR models. Our method differs from these approaches and focuses on 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 & Saenko, 2016) measures domain difference with a pre-defined metric and explicitly minimize 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 & Lempitsky, 2015; Bousmalis et al., 2016; Tzeng et al., 2017; Luo et al., 2017). The adversarial method does not require pre-defining the domain difference metric, allowing more flexible domain adaptation. MoDIR builds upon the success of these UDA methods and introduces a new momentum learning technique that is necessary to learn domain invariant representations in the ZeroDR setting.

## 6 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 their emphases of locality 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 on ranking datasets from the BEIR benchmark demonstrate robust and significant improvements of MoDIR on the zero-shot accuracy of ANCE, a recent state-of-the-art DR model.

We conduct a series of studies to show the effects of our momentum learning in learning domain invariant representations. Without momentum, the adversarial learning is unstable as the inherent variance of the DR embedding space hinders the convergence of the domain classifier. With momentum training, the model is able to fuse the target domain data into the source domain representation space, and thus discovers related information from the source domain and improves generalization, without requiring any target domain training labels.

We view MoDIR an initial step of zero-shot dense retrieval, an area demanding democratization of the rapid advancements to many real-world scenarios. Our approach inherits the success of domain adaptation techniques and upgrades them by addressing the unique challenges of ZeroDR. How to better understand the dynamics of representation learning for DR and further improve its effectiveness, robustness, and generalization ability is a future research direction with potential impacts in both representation learning research and also real-world applications.

## 7 Reproducibility Statement

We provide the following information to ensure our proposed method is reproducible:

- All datasets are publicly available and details can be found in Section 3 and Appendix A.
- Detailed experimental setups can be found in Appendix B.
- Model validation and evaluation details are discussed in Section 3.
- Source code and model checkpoints will be made public upon acceptance.

## References

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

Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. *Overview of Touché 2020: Argument Retrieval*, pp. 384–395. 09 2020. ISBN 978-3-030-58218-0. doi: 10.1007/978-3-030-58219-7_26.

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

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

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

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

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2270–2282, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.207.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

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

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1180–1189, Lille, France, 07–09 Jul 2015. PMLR.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. Accelerating large-scale inference with anisotropic vector quantization. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3887–3896. PMLR, 13–18 Jul 2020.

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

Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. 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, pp. 1265–1268, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450350228. doi: 10.1145/3077136.3080751.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 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, pp. 113–122, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3462891.

Doris Hoogeveen, Karin M. Verspoor, and Timothy Baldwin. 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, 2015. Association for Computing Machinery. ISBN 9781450340403. doi: 10.1145/2838931.2838934.

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

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572.

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

Omar Khattab and Matei Zaharia. 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, pp. 39–48, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164. doi: 10.1145/3397271.3401075.

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. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, March 2019.

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

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2*, pp. 6086–6096. ACL, 2019b.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*, 2020.

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

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

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

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

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

Zelun Luo, Yuliang Zou, Judy Hoffman, and Li F Fei-Fei. Label efficient learning of transferable representations acrosss domains and tasks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

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

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 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*, pp. 1075–1088, Online, April 2021. Association for Computational Linguistics.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pp. 1941–1942, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. ISBN 9781450356404. doi: 10.1145/3184558.3192301.

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

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 708–718, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.63.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 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*, pp. 5835–5847, Online, June 2021. Association for Computational Linguistics.

Stephen E. Robertson and Karen Spärck Jones. Relevance weighting of search terms. *JASIS*, 27(3): 129–146, 1976. doi: 10.1002/asi.4630270302.

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

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

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 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)*, pp. 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074.

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

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

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

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. TREC-COVID: Constructing a pandemic information retrieval test collection. *SIGIR Forum*, 54(1), February 2021. ISSN 0163-5840. doi: 10.1145/3451964.3451965.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 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)*, pp. 241–251, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1023.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.609.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*, 2021.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259.

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

## A Datasets Details

Target domain datasets used in our experiments are from 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 DR encoder's modeling and training. We initialize the encoder with the publicly released ANCE checkpoint[2], and randomly initialize the domain classifier. Detailed hyperparameter choices are shown in Table 5. We also use an exponential decay routine for the hyperparameter $\lambda$ to improve training stability, where the value is reduced to a half every 10k steps.

Table 5: Detailed hyperparameter choices of MoDIR.

| Hyperparameter | Value |
|---|---|
| Same as ANCE | |
| Learning rate for $\theta_g$ | 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 $\lambda$ | 1.0 |

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