

MICO: Selective Search with Mutual Information Co-training

Zhanyu Wang^{*,†}

Purdue University

{wang4094}@purdue.edu

Xiao Zhang[†] and Hyokun Yun and Choon Hui Teo and Trishul Chilimbi

Amazon

{zhxao, yunhyoku, choonhui, trishulc}@amazon.com

Abstract

In contrast to traditional exhaustive search, selective search first clusters documents into several groups before all the documents are searched exhaustively by a query, to limit the search executed within one group or only a few groups. Selective search is designed to reduce the latency and computation in modern large-scale search systems. In this study, we propose MICO, a Mutual Information CO-training framework for selective search with minimal supervision using the search logs. After training, MICO does not only cluster the documents, but also routes unseen queries to the relevant clusters for efficient retrieval. In our empirical experiments, MICO significantly improves the performance on multiple metrics of selective search and outperforms a number of existing competitive baselines.

1 Introduction

In information retrieval (IR), searching over all the documents is quite costly at a large scale (Risvik et al., 2013). Selective search (Kulkarni, 2013; Kulkarni and Callan, 2015) divides documents into a number of shards (clusters) to allow an incoming query to search over only a small number of these shards that are most relevant to this query. It tries to preserve the quality of search results similar to that of searching over the whole corpus, with less computation and shorter retrieval latency, achieved by limiting an incoming query to search within only a few chosen document shards. Sharding is the process of breaking a massive collection of documents into smaller chunks called shards. Randomly assigning documents to shards, a popular approach, cannot guarantee documents relevant to a specific query restrained within one shard or only a few shards. With the observation that semantically similar documents tend to appear together in

the search results of the same query (van Rijsbergen, 1979), researchers in recent years have devised multiple machine learning-based algorithms for selective search. These algorithms are tailored to benefit the document retrieval process by dividing a large volume of the corpus in a certain way so that semantically related documents are allocated in the same cluster and then routing new queries to the most relevant clusters subsequently. In this way, the retrieval system can return the search results with both efficiency and accuracy, as shown in Figure 1. These approaches significantly differ from the traditional sharding methods that either randomly split the collection into several groups (Barroso et al., 2003) or allocate them based on simple handcrafted rules (Gravano et al., 1999), and have achieved comparable performance to exhaustive search over the entire document corpus.

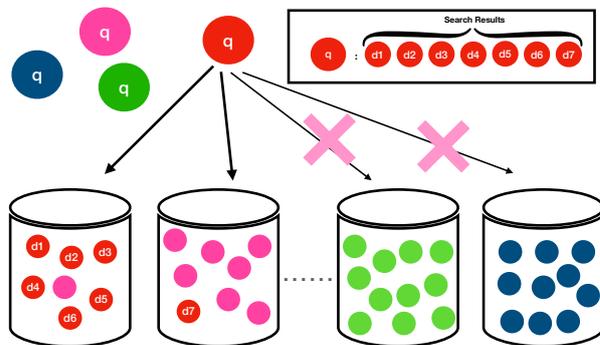


Figure 1: Machine learning based selective search first divides documents into different shards based on their semantic similarities and then routes queries to the shards containing documents most relevant to this query.

Nevertheless, having led giant strides in selective search, these approaches fail to address a few critical practical issues when scalable IR systems for a large number of documents in the industry are needed. A major concern is that these approaches heavily rely on a multi-stage process: firstly, complex clustering procedures are applied

^{*} Work done while doing an internship at Amazon.

[†] Equal contribution.

to allocate documents into different shards, while later additional elaborate algorithms are used to choose a small number of shards for incoming queries. The multi-stage process needs not only additional infrastructure support to bridge the output of the previous step as the input of the next step but also is prone to error propagation. A natural attempt to mitigate the aforementioned difficulties is to construct a model for both document allocation and subsequent query routing, with an objective function to be optimized in an end-to-end paradigm. This approach ensures applying gradient-based learning to modules or components of the system coherently.

To this end, we introduce Mutual Information CO-training (MICO), a novel end-to-end approach for selective search. We argue that by maximizing the mutual information of the query routing variable and document assignment variable of a relevant query-document pair, we can leverage co-training in an end-to-end training fashion to overcome the aforementioned difficulties. This approach first creates a bipartite graph by exploiting the query-document relations and then treats query and document as two different views of the same example while maximizing the mutual information between the cluster variable of these two views. We show an overview of the MICO framework in Figure 2 and summarize our contributions:

1. We propose MICO, Mutual Information CO-training, a novel approach for selective search. MICO treats the query and the document as two different views of the same training example: the query-document pair in the log, maximizing the mutual information between the cluster assignment variables of each view. The cluster assignment variables are parameterized via two separate distributions conditioned on the query and the document.
2. We design MICO ready for practical use as it is being trained end-to-end for both document sharding and subsequent query routing. While the trained document allocation module assigns documents into different clusters, the trained query routing module routes new queries to the target cluster(s) to retrieve the most relevant documents at a low cost. To the best of our knowledge, MICO is the first attempt to deal with document sharding and query routing in selective search in an end-to-end fashion.
3. We show significantly improved performance

on two IR data sets with MICO on multiple important metrics empirically in selective search. MICO beats competitive baselines regarding query coverage while minimizing search costs.

2 Related Works

There are quite a few studies on selective search on document sharding and query routing. Most of them aggregate similar documents into the same shards by measuring the semantic relevance among them in the collection to facilitate the subsequent search by limiting the query to be executed only within a few shards (Aly et al., 2013; Kulkarni, 2013; Kulkarni and Callan, 2010, 2015). Only a few studies on document sharding utilize the search log, where the information within the queries used to retrieve the documents are tailored to assist the clustering (Puppini et al., 2006; Poblete and Baeza-Yates, 2008; Dai et al., 2016). But they either merely extract simple features from the query to form rules or just employ the term frequency information in the queries as weights in the documents clustering process, neglecting both the strong connection between queries and documents and the semantic information hidden inside the query. Besides, all the previous studies rely on a multi-stage process for document clustering and the subsequent query routing, where an additional shard selection procedure has to be applied during the query routing stage (Si and Callan, 2003; Thomas and Shokouhi, 2009; Kulkarni et al., 2012; Aly et al., 2013). Besides, earlier studies primarily focus on ensuring semantically similar documents are allocated in the same cluster, neglecting the imbalance among the generated clusters, which weakens their application in practice. Furthermore, only a few studies take load-balancing into account when allocating documents (Kim et al., 2016; Dai et al., 2016). But these load balancing algorithms are rule-based procedures (by first ordering the shards by sizes and then regrouping them), in addition to the document sharding step.

Co-training (Blum and Mitchell, 1998; Ganchev et al., 2008) was initially proposed for semi-supervised training, where two classifiers are trained separately on two distinct views of the same input data and forced to have similar predictions. These two classifiers are designed to constrain each other to make coherent decisions, and the objective function penalizes the disagreement between them. Dasgupta et al. (2001) established the PAC generalization bounds for co-training for multiple classes

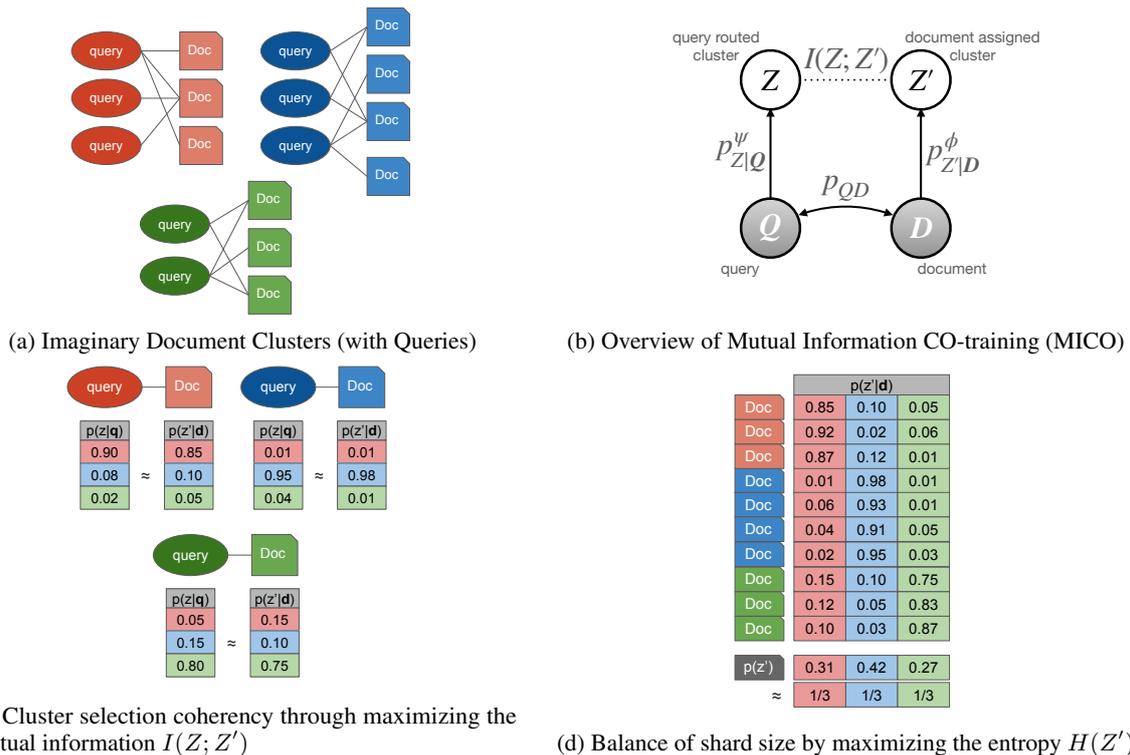


Figure 2: Figure 2a on the left illustrates the query-document pairs with the clusters they belong to, denoted by red, blue, and green, respectively. Figure 2b on the right is an overview of our Mutual Information CO-training (MICO) approach: the query and the document are treated as two views of the same example, and two different distributions parameterize the cluster assignment variables of the two views respectively. In Figure 2c and Figure 2d, we illustrate cluster selection with an example where the predicted cluster index distribution is a vector for each query and the document.

classification, extending the theoretical bound on only two classes in the original study (Blum and Mitchell, 1998). However, as pointed out by Pierce and Cardie (2001), the success of co-training heavily relies on the high-quality labeled data. This is also true in an attempt to apply co-training for unsupervised learning (Collins and Singer, 1999), where sophisticated rules which are manually extracted are used as seeds for bootstrapping.

Information Maximization (IM) (Gomes et al., 2010; Bridle et al., 1992), promoting the idea that the output shall retain as much information as the input variable, has shed light on learning without direct supervision signals. In language processing, the celebrated brown clustering (Brown et al., 1992) maximizes the mutual information between random bigrams. The information bottleneck approach (Tishby et al., 2000) proposes an algorithm that discovers the representative coding of the input by capturing its relevant structure and proves its convergence. Gomes et al. (2010) observe the degeneration and class imbalance when using mutual information for clustering and propose adding a

regularization term for remedy. Combined with the rejuvenated deep neural networks, IM has shown robust performance on clustering (Hu et al., 2017) with data augmentation.

More recent works have studied the application of information maximization in relational data. Co-training has been jointly used with information maximization in speech processing (McAllester, 2018). Stratos (2019) models the contextual information and the target class via mutual information maximization for word class induction, showing that the variational lower bound is more robust against the bias from noise. It is built upon the theoretical findings that the mutual information estimation through any high-probability lower bound has limitations that measuring and maximizing mutual information from finite data is a challenging training objective (McAllester and Stratos, 2020). This study is further extended in an adversarial training setup to circumvent computing the global marginal distribution (Stratos and Wiseman, 2020). Our work on selective search is inspired by these studies, treating the query and the document as two

views of one sample and respecting the resulting cluster size balance. It is trained in an end-to-end paradigm without direct supervision signals.

In the IR community, whether dense or sparse retrieval is better is still a prolonged debate (Lin, 2021; Luan et al., 2021), and recent studies have shown the improved performance of interpolating both (Li et al., 2022), implying designing an appropriate approach to combine them is a promising direction. We would like to note selective search is the strategy for distributing the information retrieval workload, and is complementary and agnostic to innovations in algorithms related to dense or sparse retrieval. In fact, our selective search approach can be used in conjunction with dense, sparse or combined retrieval.

3 MICO: Mutual Information CO-training

Our assumption is that for a relevant query-document pair in the search, the cluster index assignments z of the query \mathbf{q} and z' of the document \mathbf{d} are two views of that pair (\mathbf{q}, \mathbf{d}) , thus z is equal to z' . MICO, consisting of a document allocation module and a query routing module, is trained based on this assumption. After training, MICO assigns the document to the cluster based on the prediction made by the document allocation module and routes new incoming queries to the cluster based on the prediction by the query routing module.

Our goal is to learn a model from the double-view examples in the search logs. The search logs are historical data that contains the relevance information between the query and its relevant documents, but does not carry explicit information of how documents shall be allocated into different shards, neither the query routing information of which shard the query shall be dispatched to. Thus, we have minimal supervision signal when training the model to learn to cluster documents and queries.

The desired model maximizes the document coverage (recall) and minimizes the search cost simultaneously. We force the cluster selection of the query and the relevant document to be coherent during training, formally as $p(z|\mathbf{q}) \approx p(z'|\mathbf{d})$. Here we force the two distributions to be consistent if the query and the document are relevant to reflect the coherent cluster assignment. Meanwhile, to avoid certain resulting clusters being oversized,

causing the search in these shards to be inefficient, we consider balancing the document shard sizes. This concept is illustrated in Figure 3. This balance is achieved by assigning the documents to clusters as uniform as possible (formally $p(z') \approx \text{Uniform}$ where $p(z')$ is the distribution of cluster indices z'). MICO achieves cluster selection coherency between queries and documents by minimizing the mutual information between $p(z|\mathbf{q})$ and $p(z'|\mathbf{d})$, and balances shard size by adding an entropy regularization term $p(z')$.

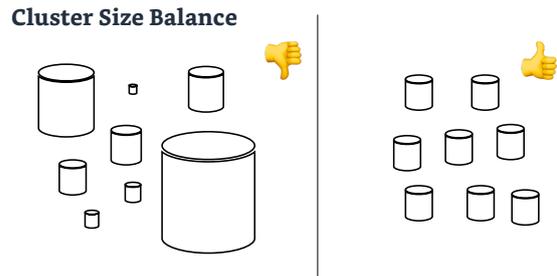


Figure 3: We use entropy regularization to ensure the desired cluster size balance, thus document shard size balance is achieved by maximizing the entropy $H(Z')$.

After training, the model allocates all the documents into shards; and for a new coming query, since we have cast query routing as a clustering problem, the query routing module generates a score for an incoming query during the inference time across all the shards, and will route the query to the shards that have the highest scores.

MICO consists of three main components: One component maps queries to clusters, denoted by the conditional distribution $p^\psi(z|\mathbf{q})$, parameterized by ψ ; Another one projects documents to clusters, denoted by the conditional distribution $p^\phi(z'|\mathbf{d})$, parameterized by ϕ ; The third one, denoted by $g^\theta(z')$ parameterized by θ , uses a variational distribution to approximate an intractable distribution $p^\phi(z')$. All these three components are modeled via neural networks to account for the complexity.

3.1 Tractable Loss of MICO

As the mutual information $I(Z; Z')$ of the query and the document cluster assignment and the entropy regularization $H(Z')$ of the document assignment are intractable (details are in the Appendix Section B.2), we approximate them using $H(Z; Z')$ and $H^+(Z')$ respectively, i.e., we define the loss

function of MICO as

$$\mathcal{L} = H(Z'; Z) - \beta H^+(Z') \quad (1)$$

where

$$\begin{aligned} H(Z'; Z) &= \mathbb{E}_{(\mathbf{q}, \mathbf{d}) \sim p_{QD}} [H(p^\phi(z'|\mathbf{d}); p^\psi(z|\mathbf{q}))] \\ &= \mathbb{E}_{(\mathbf{q}, \mathbf{d}) \sim p_{Q, D}} \left[\sum_{z=z'} -p^\phi(z'|\mathbf{d}) \log(p^\psi(z|\mathbf{q})) \right], \text{ and} \\ H^+(Z') &= \sum_{z'} -\mathbb{E}_{\mathbf{d} \sim p_D} [p^\phi(z'|\mathbf{d}) \log(g^\theta(z'))], \end{aligned}$$

with β as a tunable hyper-parameter.

The optimization objective is to find

$$\phi^*, \psi^* = \arg \min_{\phi, \psi} \left\{ \max_{\theta} \mathcal{L} \right\}. \quad (2)$$

We solve this minimax optimization problem by interleaving updating (ϕ, ψ) and θ during training. This is similar to training generative adversarial networks (GAN) in deep learning or the Actor-Critic model in reinforcement learning.

In the rest of this section, we provide brief explanations of the loss above. We rewrite \mathcal{L} as $-(I^+(Z; Z') + (\beta - 1)H^+(Z'))$ where $I^+(Z; Z') = -H(Z'; Z) + H^+(Z')$, $I^+(Z; Z')$ and $H^+(Z')$ are approximated upper bounds of $I(Z; Z')$ and $H(Z')$ respectively. The intractable term $H(Z')$ is approximated by $H^+(Z')$, by replacing $\log(p^\phi(z'))$ with $\log(g^\theta(z'))$ ($\log(p^\phi(z'))$ is intractable due to the nature it is evaluated on full dataset): we maximize $H^+(Z')$ w.r.t. θ to force $g^\theta(z')$ to be close to $p^\phi(z')$. Hence, minimizing \mathcal{L} is equivalent to maximizing $I^+(Z; Z')$ and $H^+(Z')$ when $\beta > 1$. In contrast to the proposed loss, direct application of the cross-entropy $H(Z'; Z)$, corresponding to $\beta = 0$, is inappropriate here, as it will result in cluster imbalance: the model tends to assign all the documents into one cluster and route all queries to this shard.

The choice of β in the loss \mathcal{L} in Equation 1 controls the trade-off between the mutual information $I^+(Z; Z')$ and the entropy $H^+(Z')$, which in turn balances shard sizes while retaining the success of selective search. $H(Z')$ is maximized if and only if when $\mathbb{E}_{\mathbf{d} \sim p_D} [p^\phi(z'|\mathbf{d})]$ is a uniform distribution over Z' , i.e., the cluster sizes are all equivalent. Larger β helps regularize our document allocation module to better balance the cluster sizes which leads to reduced search costs and latency.

3.2 MICO-q: MICO with Query Consistency

Multiple relevant documents can correspond to the same single query in the search log. We hypothesize that forcing the coherency of cluster assignment among these documents shall help the model yield a better document allocation module. Therefore, we use a cross-entropy H^q among those documents as an additional regularization term. With this intuition, we develop a variant of MICO termed MICO-q, whose loss function is

$$\begin{aligned} \mathcal{L}^q &= -\beta H^+(Z') + \frac{1}{1 + \gamma} (H(Z'; Z) + \gamma H^q), \text{ with} \\ H^q &= \mathbb{E}_{\mathbf{q} \sim p_Q, \mathbf{d}_1, \mathbf{d}_2 \sim p_{D|Q}} H(p^\phi(z|\mathbf{d}_1); p^\phi(z|\mathbf{d}_2)), \end{aligned}$$

where γ is another tunable hyper-parameter.

4 Experiments

4.1 Data sets

We evaluate the proposed framework on the following two data sets, representing two different domains: the E-commerce search logs (ECSL) data set and the Cross-Lingual Information Retrieval (CLIR) data set derived from Wikipedia (Sasaki et al., 2018). We process the data to ensure the data sets contain only distinct documents and divide the queries as non-overlapped training, development, and testing set, following standard IR research procedures. We show the detailed data statistics in Table 5 in Appendix.

ECSL E-commerce search logs (ECSL) data set is created by sampling queries and product description documents on an English commercial shopping website. The relevance of a document to a query is derived from user actions on the search results. We consider three types of user actions on documents with increasing levels of relevance to the query: *impression* (the document is among the search results), *click* (the document link in the search results is clicked), and *purchase* (the product of the document from the search result is purchased). The former relation always entails the latter. We randomly split the queries into training, validation, and testing sets with the ratio of 8:1:1.

CLIR In the CLIR set, queries are in English, extracted as the first sentences from English wiki pages, with the title words removed. Two different types of relevance exist: *DL* stands for directly relevant, meaning the documents are the foreign-language pages having an inter-language link to

the English pages; *PL* stands for partially relevant, denoting the documents having mutual links to and from the *DL* document. To provide a comprehensive study on the robustness of the proposed framework, we select the documents in two high-resource languages: French (*fr*), Italian (*it*) and in two low-resource languages: Tagalog (*ta*), Swahili (*sw*). Queries are randomly split into training, validation, and testing sets with the ratio of 3:1:1.

4.2 Experimental Setup

In this study, we preset the number of shards to be 64 for the ECSL data set and 10 for the CLIR data set, according to the number of documents. To make a fair comparison with other baselines, we use the TF-IDF feature with the most frequent words as the feature vector for both the queries and the documents.

We compare our proposed methods with a number of competitive baselines in selective search, including **Random** assignment of documents into clusters, **K-means**, a variant of K-means called **Balanced K-means** ensuring balanced cluster sizes, an information-maximization based clustering algorithm **IMSAT** (Hu et al., 2017), **KLD** sharding algorithm (Kulkarni and Callan, 2015) and **QKLD** sharding algorithm (Dai et al., 2016). Details of these baselines are in Appendix A.2.

In our experiments, we use the *impression* relationship between the training query and the documents for training in ECSL data. We use both the *DL* and *PL* relations for training in CLIR, as they are non-overlapped documents. The expectation over the whole training set is approximated by the average over one batch during the training, t . We also provide a detailed theoretical analysis of batch training in MICO in Appendix C. Due to limited spaces, more details of the experimental setup can be found in Appendix A. Note neither the queries nor the documents have cluster labels, and the clustering effect in selective search is evaluated by the retrieval performance explained next.

4.3 Query Coverage Analysis

We first evaluate the query coverage (“recall” in some literature) of our methods and compare it with baselines. Query coverage measures the percentage of documents retrieved in the top few shards over all the documents retrieved in the exhaustive search.

Formally, the query coverage of a single query q in the top N shards is defined as $Cov_N(q) =$

$\left(\sum_{i=1}^N R_{s_i}^q\right) / R^q$, where R^q is the total number of relevant documents in the corpus per the query q , and $R_{s_i}^q$ is the number of relevant document per the query q in the i -th selected shard s_i^q per query q . Consequently, the average coverage over the query set is $Cov_N = \left(\sum_{q \in Q} Cov_N(q)\right) / |Q|$, where Q is the testing queries set.

Our implemented coverage metric is even more strict than the coverage metric in Dai et al. (2016)’s study. Instead of sorting the shards by the number of retrieved documents they contain, we rigorously sort the chosen shards of a testing query based on the prediction score generated by the query routing module, which is more intrinsic to the nature of selective search.

We show our experimental results on ECSL in Table 1, with the mean value and the standard deviation in the parenthesis over five runs on three different relevances. We limit each testing query being routed to the most relevant **one** shard and **ten** shards based on the prediction. Our MICO and its variant MICO-q beat all the baselines, except *impression* with one shard only. By only searching within the ten most relevant shards, which is 15.6% of all the shards, MICO-q can achieve almost 95% coverage on *impression*, even comparable to the exhaustive search. MICO-q performs slightly more robust than MICO when searching over more shards, which might be attributed to its additional regularization term.

We also present the evaluation results on CLIR in Table 2, where we only probe the most relevant shard for each query. Due to the queries and documents being of different languages, we build vocabularies for the queries and documents separately, denoted by sv , to generate input to the document sharding module and the query routing module. MICO (and MICO-q) beats all the baselines (except *DL* in *ta*). We also notice the performance of QKLD dramatically drops on the CLIR data set. We attribute this failure to its usage of the query information to build document shards inappropriately because queries and documents are from different languages. QKLD and MICO (including MICO-q) are the only two models to utilize the query information specifically for document sharding. Based on the observation, we argue inappropriate use of the query information will deteriorate document sharding in selectively search. MICO, on the contrary, is able to leverage the queries for better document sharding, showing its robustness

in the cross-lingual scenario.

Though MICO beats all the baselines, if the top selected shards are extremely large, this approach is inapplicable in practice, as the actual search cost will be similar to the exhaustive search. In addition, MICO fails to beat KLD and QKLD in some exceptions, potentially ascribed to the most relevant shard created by them being larger with more documents. Thus, we analyze the search cost next.

4.4 Cost Analysis

To measure the search efficiency, we use the metrics **search resource cost** and **search latency cost** (the lower, the better) introduced in the early study (Kulkarni et al., 2012) to evaluate our models and the competitors. Search cost depends highly on the size of the shards we send the query to. Search resource cost calculates the resource usage as the upper bound on the number of document evaluated for each query, defined as $C_N^{res} = \sum_{i=1}^N |s_i^q|$, where $|s_i^q|$ is the number of documents in the i -th relevant cluster to query q . Similarly, search latency cost counts the number of evaluated documents on the longest execution path for query q for its top N selected shards: $C_N^{lat} = \max_{1 \leq i \leq N} |s_i^q|$.¹

We show our evaluation results on the ECSL data set and the CLIR data set in Table 3, by restricting the search within the **five** most relevant clusters for ECSL, and the **two** most relevant for CLIR. Please note in these evaluations **Random** is the skyline, as all the documents are evenly distributed to N shards such that the shard sizes are almost equal. MICO demonstrates its supreme performance, even beating the skyline in some cases, showing its ability to control shard sizes while ensuring semantically similar documents are allocated in the same shard. MICO-q also achieves comparable performance with MICO. We interpret it as MICO and MICO-q are able to trade off document relevance in the same shard and balance of shard sizes on a sweet point. We also notice the high cost yielded by KLD and QKLD, which compromises their use in practice.

4.5 Balance Among Shard Sizes

We further investigate the performance of MICO and MICO-q via visualizing the trade-off between *query coverage* and *search resource cost* and the

¹Note the metrics in Kulkarni et al. (2012) introduce an additional term for the resource selection step. Since MICO doesn't require this additional step, that term is always evaluated to 0 with MICO (and MICO-q).

balance of resulting shard sizes on ECLS. Figure 4a shows that MICO and MICO-q always perform the best in terms of query coverage, at different levels of search resource cost. Figure 4b indicates that IMSAT is able to create well balanced shards, and MICO and MICO-q are on a par with it. To trade off query coverage and shard size balance, MICO and MICO-q are the best among all the competitors, gaining the favor in practical usage. Note that KLD and QKLD create very unbalanced shards, which explains MICO and MICO-q fail to beat them in some cases on query coverage analysis earlier: most of the documents are allocated into the same shard.

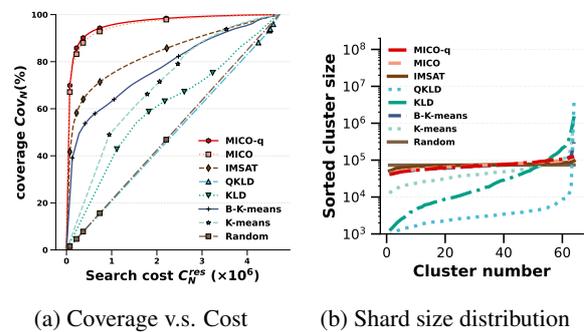


Figure 4: Figure 4a shows MICO and MICO-q are significantly better than all other methods as they have high *impression* coverage with low search cost. From bottom-left to top-right, the markers on each line represent query coverage limited within the top-1, top-3, top-5, top-10, and top-30 clusters selectively. Figure 4b shows **Random** generates the most balanced shard sizes (as a flat line), and IMSAT also creates very balanced shards. MICO and MICO-q are on a par with IMSAT. In contrast, QKLD and KLD yield very unbalanced shards.

4.6 Effect of Entropy Regularization in MICO and MICO-q

We also examine the effect of entropy regularization by adjusting its strength in MICO and MICO-q by tuning their hyper-parameter β on ECSL. From Figure 5a, we can see that β does not affect the cost-coverage curve in a significant way. From Figure 5b, setting β larger than 1 yields much more balanced cluster sizes which means that the entropy regularization is necessary to balance cluster sizes. We conjecture this difference results from the degree of homogeneity of the documents the model allocates to different shards.

Models	<i>impression</i>		<i>click</i>		<i>purchase</i>	
	N=1	N=10	N=1	N=10	N=1	N=10
Random	1.56 (6e-3)	15.62 (0.02)	1.49 (0.08)	15.32 (0.85)	1.45 (0.24)	14.54 (0.27)
K-means	48.98 (1.60)	79.05 (0.51)	51.90 (1.56)	81.57 (4.0)	54.49 (1.97)	83.58 (1.49)
B-K-means	39.72 (1.12)	64.56 (1.30)	43.89 (2.03)	64.25 (1.78)	49.02 (2.37)	69.59 (1.22)
IMSAT	41.68 (0.55)	71.37 (0.28)	47.48 (1.62)	79.12 (2.94)	52.41 (0.42)	79.83 (1.06)
KLD	43.46 (5.91)	69.87 (5.34)	44.94 (8.04)	71.17 (5.55)	46.77 (9.32)	70.5 (4.08)
QKLD	86.14 (8.85)	93.96 (0.77)	73.72 (7.25)	81.89 (1.2)	75.79 (7.22)	83.56 (1.57)
MICO	67.09 (0.20)	92.85 (0.12)	82.85 (1.51)	97.81 (0.19)	81.21 (0.49)	96.61 (0.14)
MICO-q	69.81 (0.34)	94.28 (0.09)	82.48 (1.91)	98.26 (0.20)	81.15 (1.23)	97.25 (0.16)

Table 1: This table shows the performance of query coverage (recall) of MICO, MICO-q, and different baselines over three different query-document relationships on the ECSL data set. We show the performance by only probing the top-1 most relevant shard and the top-10 most relevant shards given a query. The number in the parenthesis right next to the coverage is the standard deviation over five runs. We observe other than the impression relation in which QKLD has the best performance, MICO or MICO-q beat all the baselines.

Models	<i>fr</i>		<i>it</i>		<i>ta</i>		<i>sw</i>	
	<i>DL</i>	<i>PL</i>	<i>DL</i>	<i>PL</i>	<i>DL</i>	<i>PL</i>	<i>DL</i>	<i>PL</i>
Random	10.02 (0.07)	9.72 (0.16)	10.02 (0.09)	10.0 (0.35)	9.88 (0.93)	9.86 (0.48)	10.01 (0.23)	10.0 (0.67)
K-means	12.19 (1.99)	10.79 (2.04)	14.91 (2.46)	16.36 (3.55)	16.25 (2.5)	21.08 (3.46)	21.71 (4.84)	18.55 (3.65)
B-K-means	12.2 (1.82)	11.44 (1.32)	12.45 (3.13)	12.46 (4.59)	12.78 (2.85)	11.23 (3.84)	11.71 (1.29)	12.16 (1.21)
IMSAT	19.77 (9.53)	19.84 (9.68)	40.09 (8.91)	40.13 (8.88)	12.72 (4.22)	11.89 (3.62)	8.4 (2.24)	8.6 (2.63)
KLD	38.6 (6.02)	40.65 (7.58)	60.94 (5.25)	61.83 (3.12)	66.53 (8.43)	59.77 (7.18)	21.11 (3.52)	24.83 (3.81)
QKLD	17.76 (3.63)	18.82 (2.08)	18.9 (4.91)	17.45 (4.12)	23.65 (6.81)	24.4 (5.54)	12.23 (0.75)	16.45 (2.25)
MICO (<i>sv</i>)	44.93 (3.47)	53.12 (2.17)	58.08 (1.22)	65.83 (1.06)	63.55 (4.45)	60.94 (4.92)	26.0 (3.51)	28.67 (3.73)
MICO-q (<i>sv</i>)	47.9 (2.68)	48.04 (3.44)	75.27 (3.6)	75.01 (4.39)	63.91 (5.3)	61.29 (5.31)	27.42 (3.37)	28.14 (2.54)

Table 2: This table shows the performance of query coverage of MICO, MICO-q, and different baselines on two different query-document relationships on the CLIR data set by only probing the most relevant shard given a query because we only divide the documents into ten shards. The number in the parenthesis right next to the coverage is the standard deviation over multiple runs. *sv* stands for separate vocabularies for the queries and documents, as in cross-lingual retrieval, the source language and the target language have different vocabularies, and separate vocabularies perform better than unified ones empirically. MICO and MICO-q beat all the baselines except *DL* in *ta*.

4.7 Ablation Study

We further investigate the effectiveness of different neural architecture variants of MICO. Instead of parameterizing the document allocation module and the query routing module differently by two separate neural networks ($p_q^\psi(z|\mathbf{q}), p_d^\phi(z'|\mathbf{d})$), we can parameterize both modules with the same neural network ($p_q^\psi(z|\mathbf{q}), p_d^\psi(z'|\mathbf{d})$), avoiding the risk of over-parameterization. We denote this variant as **-Par**. The recent emergence of the pre-trained language models such as BERT (Devlin et al., 2019) also provides a richer representation with contextual information. We use the BERT multilingual

base model² to produce the text-level representations and feed them into MICO, either as fixed (denoted as **+BERT.fx**) or by fine-tuning the underlying BERT model (denoted as **+BERT.ft**). We show the query coverage (QC) and Search Resource Cost of these neural variants in Table 4 on ECSL, by probing the most relevant shard for each query. The results show that unifying the query module and the document module parameterization deteriorates the performance. We also notice that fixed BERT representations hurt the model performance significantly, while appropriate fine-tuning help improve the performance.

²<https://huggingface.co/bert-base-multilingual-cased>

Models	C_N^{res}					C_N^{lat}				
	ECSL	C-fr	C-it	C-ta	C-sw	ECSL	C-fr	C-it	C-ta	C-sw
K-means	2.061	14.12	11.54	1.6	1.42	1.572	6.44	5.95	0.95	1.27
Balaced K-means	0.620	8.1	6.83	0.99	0.87	0.277	2.58	2.02	0.34	0.67
IMSAT	0.370	9.57	5.89	0.93	0.61	0.082	3.57	4.78	0.58	0.53
KLD	2.17	17.48	13.15	1.93	1.09	1.41	13.26	11.43	1.72	0.74
QKLD	4.5	8.84	7.42	1.34	0.99	4.47	3.72	3.07	0.94	0.66
MICO	0.367	6.19	5.13	0.85	0.93	0.089	2.34	1.89	0.5	0.5
MICO-q	0.369	7.12	6.71	0.94	1.07	0.093	2.73	2.47	0.51	0.58
Random	0.368	7.20	5.95	0.8	0.73	0.074	2.41	1.99	0.27	0.25

Table 3: This table shows the performance of different models on the *Search Resource Cost* and the *Search Latency Cost* metrics, representing the search efficiency, with the lower the number, the better the performance. The results shown in this table are scaled by being divided by 10^6 on the ECSL data set and by 10^4 on the CLIR data set. Note in this set of experiments, we use separate vocabulary (*sv*) for MICO and MICO-q on CLIR. We observe the supreme performance of MICO, which in some cases even beats the **Random** skyline.

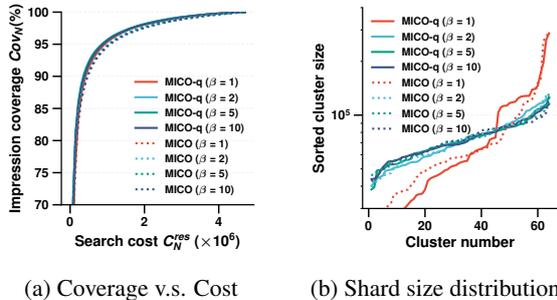


Figure 5: This figure shows the results of MICO and MICO-q with different entropy parameter $\beta \in \{1, 2, 5, 10\}$. (a) The overall performance is hardly affected by β . (b) While larger β induces better balance on the cluster sizes, the difference between $\beta = 5$ and $\beta = 10$ is very small. When setting $\beta = 1$, the objective degenerates to use mutual information only (without the additional entropy regularization term) and the resulting shard sizes are significantly more unbalanced than with the other choices of β .

5 Conclusion

In this study, we present MICO, a Mutual Information CO-training framework for document sharding and query routing in selective search with minimal supervision. Our contributions are: First, during training, MICO maximizes the mutual information between the cluster assignment variables of the query and the document, forcing the prediction of two views of the query-document pair coherently. Second, this design enables it to be trained end-to-end for both document sharding (clustering) and subsequent query routing, featuring its practical use when vast volumes of documents are required

Model	QC	$C_{N=1}^{res}$
MICO (-Par)	66.08	0.365
MICO	67.09	0.367
MICO (+BERT.fx)	41.67	0.367
MICO (+BERT.ft)	76.41	0.375

Table 4: This table shows MICO with neural architecture variants. BERT with fine-tuning achieves better performance than the original MICO, while the other variants yield deteriorated performance. The search cost is slightly higher with the best-performing system. We attribute that the refined representations cause the model to weigh more on semantic similarity than cluster balance.

to be searched simultaneously. Third, we show improved performance of MICO empirically on multiple important metrics in selective search.

Our future research direction includes further investigation of the regularization for more balanced sizes among all the shards, the utilization of complex neural language models to enrich the document and the query representation for higher query coverage via better clustering results, the reduction of potential noise between a query and its associated document in the data set, and the detection policy to decide when the model needs to be retrained if constant update of the model is needed when deployed. We are also interested in potential extension of this approach to multi-modal data, e.g., image search, music search and software search where the non-traditional documents are in the non-textual formats.

References

- Robin Aly, Djoerd Hiemstra, and Thomas Demeester. 2013. [Taily: shard selection using the tail of score distributions](#). In *The 36th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '13, Dublin, Ireland - July 28 - August 01, 2013*, pages 673–682. ACM.
- Luiz André Barroso, Jeffrey Dean, and Urs Hölzle. 2003. Web search for a planet: The google cluster architecture. *IEEE micro*, 23(2):22–28.
- Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100.
- John Bridle, Anthony Heading, and David MacKay. 1992. [Unsupervised classifiers, mutual information and phantom targets](#). In *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann.
- Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n -gram models of natural language. *Computational Linguistics*, 18(4):467–479.
- Michael Collins and Yoram Singer. 1999. [Unsupervised Models for Named Entity Classification](#). In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.
- Zhuyun Dai, Chenyan Xiong, and Jamie Callan. 2016. [Query-biased partitioning for selective search](#). In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, pages 1119–1128. ACM.
- Sanjoy Dasgupta, Michael L. Littman, and David A. McAllester. 2001. [PAC generalization bounds for co-training](#). In *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pages 375–382. MIT Press.
- 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.
- Kuzman Ganchev, João V. Graça, John Blitzer, and Ben Taskar. 2008. Multi-view learning over structured and non-identical outputs. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI'08*, page 204–211, Arlington, Virginia, USA. AUAI Press.
- Ryan Gomes, Andreas Krause, and Pietro Perona. 2010. [Discriminative clustering by regularized information maximization](#). In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, pages 775–783. Curran Associates, Inc.
- Luis Gravano, Héctor García-Molina, and Anthony Tomasic. 1999. [Gloss: Text-source discovery over the internet](#). *ACM Trans. Database Syst.*, 24(2):229–264.
- Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. 2017. [Learning discrete representations via information maximizing self-augmented training](#). In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1558–1567. PMLR.
- Yubin Kim, Jamie Callan, J. Shane Culpepper, and Alistair Moffat. 2016. [Load-balancing in distributed selective search](#). In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16*, page 905–908, New York, NY, USA. Association for Computing Machinery.
- Anagha Kulkarni. 2013. *Efficient and Effective Large-scale Search*. PhD thesis, Carnegie Mellon University.
- Anagha Kulkarni and Jamie Callan. 2010. [Document allocation policies for selective searching of distributed indexes](#). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10*, page 449–458, New York, NY, USA. Association for Computing Machinery.
- Anagha Kulkarni and Jamie Callan. 2015. [Selective Search: Efficient and Effective Search of Large Textual Collections](#). *ACM Trans. Inf. Syst.*, 33(4). Place: New York, NY, USA Publisher: Association for Computing Machinery.
- Anagha Kulkarni, Almer S. Tigelaar, Djoerd Hiemstra, and Jamie Callan. 2012. [Shard ranking and cutoff estimation for topically partitioned collections](#). In *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, pages 555–564. ACM.
- Hang Li, Shuai Wang, Shengyao Zhuang, Ahmed Mourad, Xueguang Ma, Jimmy Lin, and Guido Zuccon. 2022. [To interpolate or not to interpolate: Prf, dense and sparse retrievers](#). In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22*, page 2495–2500, New York, NY, USA. Association for Computing Machinery.

- Jimmy Lin. 2021. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In *Proc. of SIGIR*, page 7.
- 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.
- David McAllester. 2018. Information Theoretic Co-Training. Eprint: 1802.07572.
- David McAllester and Karl Stratos. 2020. Formal limitations on the measurement of mutual information. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 875–884. PMLR.
- David Pierce and Claire Cardie. 2001. Limitations of co-training for natural language learning from large datasets. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*.
- Barbara Poblete and Ricardo Baeza-Yates. 2008. Query-sets: Using implicit feedback and query patterns to organize web documents. In *Proceedings of the 17th International Conference on World Wide Web, WWW '08*, page 41–50, New York, NY, USA. Association for Computing Machinery.
- Diego Puppin, Fabrizio Silvestri, and Domenico Laforenza. 2006. Query-driven document partitioning and collection selection. In *Proceedings of the 1st International Conference on Scalable Information Systems, InfoScale '06*, page 34–es, New York, NY, USA. Association for Computing Machinery.
- Knut Magne Risvik, Trishul Chilimbi, Henry Tan, Karthik Kalyanaraman, and Chris Anderson. 2013. Maguro, a system for indexing and searching over very large text collections. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 727–736. ACM.
- Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh, and Kentaro Inui. 2018. Cross-lingual learning-to-rank with shared representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 458–463, New Orleans, Louisiana. Association for Computational Linguistics.
- Luo Si and Jamie Callan. 2003. Relevant document distribution estimation method for resource selection. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, SIGIR '03*, page 298–305, New York, NY, USA. Association for Computing Machinery.
- Karl Stratos. 2019. Mutual information maximization for simple and accurate part-of-speech induction. 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 1095–1104, Minneapolis, Minnesota. Association for Computational Linguistics.
- Karl Stratos and Sam Wiseman. 2020. Learning discrete structured representations by adversarially maximizing mutual information. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9144–9154. PMLR.
- Paul Thomas and Milad Shokouhi. 2009. Sushi: Scoring scaled samples for server selection. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09*, page 419–426, New York, NY, USA. Association for Computing Machinery.
- Naftali Tishby, Fernando C. N. Pereira, and William Bialek. 2000. The information bottleneck method. *Proc. 37th Annual Allerton Conference on Communications, Control and Computing, 1999*, pages 368–377.
- C. J. van Rijsbergen. 1979. *Information Retrieval*, 2nd edition. Butterworth-Heinemann, London.

APPENDIX

A Experimental Setup

A.1 Data Processing

We use the `gensim` package to preprocess all the queries and documents into TF-IDF features in three steps: 1. Tokenize the sentences into words and remove stop words; 2. Create a dictionary with most frequent words (20,000 for ECSL, and 3,000 for CLIR); 3. Transform the queries and documents into TF-IDF vectors.

Data sets	Queries				Documents	
	#Train	#Dev	#Test	Avg($ L $)	#Doc	Avg($ L $)
ECSL	339k	38k	55K	4	4.7M	16
CLIR-FR	14.8K	4.9K	4.9K	21	240K	177
CLIR-IT	14.8K	4.9K	4.9K	21	198K	169
CLIR-SW	12.8K	4.1K	4.1K	18	24k	82
CLIR-TA	14.3K	4.7K	4.7K	19	27k	79

Table 5: This table shows the statistics of the ECSL data set and the CLIR data set, including the numbers of queries, documents and their average length.

A.2 Baselines

Balanced K-means is built on top of K-means with approximated Hungarian algorithm to ensure the formed cluster sizes are well balanced. In K-means and Balanced K-means, documents and training queries are used together to form clusters, and we select the nearest clusters to the query vector as the top relevant clusters to route the query to. In IMSAT, the same model is used for both document assignment and query routing, after being trained on both queries and documents. KLD (Kullback-Liebler Divergence) only uses document information to form document clusters while QKLD (Query-based Kullback-Liebler Divergence) leverages the query information to form clusters. In KLD and QKLD, an additional shard selection algorithm is used to choose the relevant shards for new incoming queries.

A.3 MICO

During training, we set the batch size as 256. We optimize the objective using Adam with learning rate 0.03 for ϕ and ψ , and with learning rate 0.1 for θ . We model both $p_q^\psi(z|\mathbf{q})$ and $p_d^\phi(z|\mathbf{d})$ with neural networks with one hidden layer (dimension 20) and one softmax layer. We set the hyper-parameter for entropy regularization as $\beta = 10$. To stabilize the training process, we set gradient clipping as 10.0.

A.4 MICO-q

We set the hyper-parameters the same as in MICO except: 1. We set the gradient clipping as 1.0; 2. We update θ for 4 steps instead of 1; 3. We set the entropy regularization strength as $\beta = 3$; 4. We set the query consistency parameter as $\gamma = 3$.

A.5 Computing Environment

We use an AWS EC2 instance (p3.2xlarge) consisting of 8 vCPUs (Intel Xeon E5-2686 v4) with 64GB memory, with one NVIDIA GPU card (Tesla V100-SXM2-16GB) for running our experiments. The computing environment is with Python 3.8, PyTorch 1.5, CUDA 9.2.

B Details of MICO

B.1 Preliminary

We introduce the notations and the basic concept of mutual information briefly here. We denote a random variable X 's probability function by $p_X(x)$. We shorten it to $p(x)$ or p_X when its meaning is clear under its context. We use $p^\theta(x)$ when $p(x)$ is in a parametric distribution family and θ is the parameter. $p_{XY}(x, y)$ is the joint distribution of two random variables X and Y , and $p_{X|Y}(x|y)$ is the conditional distribution of X given Y , while $p_Y(y) = \sum_x p_{XY}(x, y)$ is the marginal distribution of Y under $p_{XY}(x, y)$. The expectation of a function $f(x)$ w.r.t. a random variable x is defined by $\mathbb{E}_{x \sim p_X}[f(x)] = \sum_{x \in \mathcal{X}} f(x)p_X(x)$. We denote the entropy of X by $H(X) = H(p_X) = \mathbb{E}_{x \sim p_X}[-\log(p_X(x))]$, while the conditional entropy is $H(Y|X) = \mathbb{E}_{x \sim p_X} H(p_{Y|X}) = \mathbb{E}_{(x,y) \sim p_{XY}}[-\log(p_{Y|X})]$. The cross-entropy between X and X' is defined as $H(X; X') = \mathbb{E}_{x \sim p_X}[-\log(p_{X'}(x))]$. We denote the mutual information between X and Y by $I(X; Y) = \sum_x \sum_y p_{XY} \log\left(\frac{p_{XY}}{p_X p_Y}\right) = H(X) - H(X|Y)$. It is minimized at 0 when X and Y are independent, i.e., $p_{XY} = p_X p_Y$. The chain rule for mutual information is $I(X; (Y, Z)) = I(X; Z) + I(X; Y|Z)$ where $I(X; Y|Z) = H(X|Z) - H(X|Y, Z) \geq 0$.

B.2 Information Maximization Approximation

We explain the reason of having the three components for calculating the mutual information. We seek to train the probabilistic modules $p^\psi(z|\mathbf{q})$ and $p^\phi(z'|\mathbf{d})$ through maximizing the mutual information between Z and Z' over our training data set, as

shown in Figure 2b. From this probabilistic graph, we know Z and Z' are independent when given \mathbf{Q} , i.e., $I(Z'; Z|\mathbf{Q}) = 0$. Based on the chain rule of mutual information, we also have $I(Z'; (\mathbf{Q}, Z)) = I(Z'; \mathbf{Q}) + I(Z'; Z|\mathbf{Q}) = I(Z'; Z) + I(Z'; \mathbf{Q}|Z)$. Therefore, $I(Z'; Z) = I(Z'; \mathbf{Q}) - I(Z'; \mathbf{Q}|Z)$. Since $I(Z'; Z)$ is intractable under large amount of data and $I(Z'; \mathbf{Q}|Z) \geq 0$, we maximize $I(Z'; \mathbf{Q})$ instead, which is an upper bound of $I(Z'; Z)$. With the fact $I(Z'; \mathbf{Q}) = H(Z') - H(Z'|\mathbf{Q})$, we need the marginal distribution $p^\phi(z') = \mathbb{E}_{\mathbf{d} \sim p_D}[p^\phi(z'|\mathbf{d})]$ and the conditional distribution $p^\phi(z'|\mathbf{q}) = \sum_{\mathbf{d}} p^\phi(z'|\mathbf{d})p(\mathbf{d}|\mathbf{q})$, both of which are also part of the logarithm part of the two entropy terms. Therefore, $I(Z'; \mathbf{Q})$ is again intractable if we only use $p^\phi(z'|\mathbf{d})$. We approximate $p^\phi(z')$ and $p^\phi(z'|\mathbf{q})$ by neural networks $g^\theta(z')$ and $p^\psi(z|\mathbf{q})$ respectively, which falls into the category of variational methods. As $H(Z') = H(p^\phi(z')) \leq H(p^\phi(z'); g^\theta(z'))$ and $H(Z'|\mathbf{Q}) = \mathbb{E}_{\mathbf{q} \sim p_Q} H(p^\phi(z'|\mathbf{q})) \leq \mathbb{E}_{\mathbf{q} \sim p_Q} H(p^\phi(z'|\mathbf{q}); p^\psi(z|\mathbf{q}))$, we upper-bound $H(Z'|\mathbf{Q})$ with $H(Z'; Z) := \mathbb{E}_{(\mathbf{q}, \mathbf{d}) \sim p_{Q,D}} [H(p^\phi(z'|\mathbf{d}); p^\psi(z|\mathbf{q}))]$, and introduce $H^+(Z') := \mathbb{E}_{\mathbf{d} \sim p_D} H(p^\phi(z'|\mathbf{d}); g^\theta(z'))$ as an upper bound of $H(Z')$. Note that with $H(X) = \min_Y H(X; Y)$, our approximation $\hat{I}(Z'; \mathbf{Q}) = H^+(Z') - H(Z'; Z)$ gives exact evaluation of $I(Z'; \mathbf{Q})$ when $H^+(Z')$ is minimized w.r.t. θ when $g^\theta(z') = p^\phi(z')$, and $H(Z'; Z)$ is minimized w.r.t. ψ when $p^\psi(z|\mathbf{q}) = p^\phi(z'|\mathbf{q})$.

C Theoretical Analysis of Stochastic Optimization of MICO

First we show why computing $I(Z'; Z)$ is intractable even under the stochastic optimization framework. With the MICO framework, the expectation over $p_{Q,D}$ is approximated by the average over one batch of data $\mathbf{B} = \{(\mathbf{q}_1, \mathbf{d}_1), (\mathbf{q}_2, \mathbf{d}_2), \dots, (\mathbf{q}_b, \mathbf{d}_b)\}$. If one objective is f consisting of the expectation over $p_{Q,D}$, we evaluate f on \mathbf{B} and denote the corresponding result as f_B . The core assumption in stochastic optimization is $f = \mathbb{E}_{\mathbf{B}} f_B$.

For

$$I(Z'; Z) = \mathbb{E}_{(\mathbf{q}, \mathbf{d}) \sim p_{Q,D}} \left[\sum_{z, z'} p^\phi(z'|\mathbf{d}) p^\psi(z|\mathbf{q}) \log \left(\frac{\mathbb{E}_{(\mathbf{q}, \mathbf{d}) \sim p_{Q,D}} [p^\phi(z'|\mathbf{d}) p^\psi(z|\mathbf{q})]}{\mathbb{E}_{(\mathbf{q}, \mathbf{d}) \sim p_{Q,D}} [p^\phi(z'|\mathbf{d})] \mathbb{E}_{(\mathbf{q}, \mathbf{d}) \sim p_{Q,D}} [p^\psi(z|\mathbf{q})]} \right) \right],$$

if we use

$$I_B(Z'; Z) = \frac{1}{b} \sum_{i=1}^b \left[\sum_{z, z'} p^\phi(z'|\mathbf{d}_i) p^\psi(z|\mathbf{q}_i) \log \left(\frac{\frac{1}{b} \sum_{i=1}^b [p^\phi(z'|\mathbf{d}_i) p^\psi(z|\mathbf{q}_i)]}{\frac{1}{b} \sum_{i=1}^b [p^\phi(z'|\mathbf{d}_i)] \frac{1}{b} \sum_{i=1}^b [p^\psi(z|\mathbf{q}_i)]} \right) \right],$$

it violates the assumption since $I(Z'; Z) \neq \mathbb{E}_{\mathbf{B}} [I_B(Z'; Z)]$.

In MICO, we replace $I(Z'; Z)$ with $I(Z'; \mathbf{Q})$.

Moreover, we use

$$H_B(Z'; Z) = \frac{1}{b} \sum_{i=1}^b \left[\sum_{z=z'} -p^\phi(z'|\mathbf{d}_i) \log (p^\psi(z|\mathbf{q}_i)) \right]$$

and

$$H_B(Z') = \frac{1}{b} \sum_{i=1}^b \sum_{z'} -p^\phi(z'|\mathbf{d}_i) \log \left(\frac{1}{b} \sum_{i=1}^b p^\phi(z'|\mathbf{d}_i) \right)$$

to approximate

$$H(Z'; Z) = \mathbb{E}_{(\mathbf{q}, \mathbf{d}) \sim p_{Q,D}} \left[\sum_{z=z'} -p^\phi(z'|\mathbf{d}) \log (p^\psi(z|\mathbf{q})) \right]$$

and

$$H(Z') = \sum_{z'} -\mathbb{E}_{(\mathbf{q}, \mathbf{d}) \sim p_{Q,D}} [p^\phi(z'|\mathbf{d})] \log \left(\mathbb{E}_{(\mathbf{q}, \mathbf{d}) \sim p_{Q,D}} [p^\phi(z'|\mathbf{d})] \right),$$

respectively.

Note that $H(Z', Z) = \mathbb{E}_{\mathbf{B}} [H_B(Z', Z)]$ while $H(Z') \neq \mathbb{E}_{\mathbf{B}} [H_B(Z')]$. This is due to the logarithm part in $H_B(Z')$. Without using the stochastic approximation on the logarithm, since it is expensive to evaluate the expectation of $p^\phi(z'|\mathbf{d})$ on the full dataset, we have to replace it with $g^\theta(z')$ in the logarithm term which creates an upper bound $H^+(Z')$ on $H(Z')$ as an approximation (used in the tractable loss function of MICO):

$$H^+(Z') = \sum_{z'} -\mathbb{E}_{(\mathbf{q}, \mathbf{d}) \sim p_{Q,D}} [p^\phi(z'|\mathbf{d})] \log (g^\theta(z')).$$

$$H_B^+(Z') = \frac{1}{b} \sum_{i=1}^b \sum_{z'} -p^\phi(z'|\mathbf{d}_i) \log (g^\theta(z')),$$

$$H^+(Z') = \mathbb{E}_{\mathbf{B}} [H_B^+(Z')]$$

The stochastic evaluation of \mathcal{L} is then $\mathcal{L}_B = \beta H_B^+(Z') - H_B(Z'; Z)$ and we have $\mathcal{L} = \mathbb{E}_B[\mathcal{L}_B]$, which is also the reason of why we use $H(Z'; Z)$ instead of $H(Z'|\mathbf{Q})$ in MICO.

For each batch of data, since $H(Z') = \min_{g^\theta(z')} H^+(Z')$, we first update $g^\theta(z')$ for one step to descend $H_B^+(Z')$ (corresponding to SGD), and then update $p^\psi(z'|\mathbf{q})$ and $p^\phi(z'|\mathbf{d})$ for one step to descend \mathcal{L}_B (corresponding to another SGD).