MULTI-SAMPLE CONTRASTIVE NEURAL TOPIC MODEL AS MULTI-TASK LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent representation learning approaches to polish global semantics of neural topic models optimize the weighted linear combination of the evidence lower bound (ELBO) of the log-likelihood and the discriminative objective that contrasts instance pairings. However, contrastive learning on the individual level might capture noisy mutual information that is irrelevant to the topic modeling task. Moreover, there is a potential conflict between the ELBO loss that memorizes input details for better reconstruction quality, and the contrastive term which attempts to generalize representations among inputs. To address the issues, we firstly hypothesize that useful features should be shared among multiple input samples. For that reason, we propose a novel set-based contrastive learning method for neural topic models to employ the concept of multi-sample representation learning. Secondly, because the solution of the linear combination approach might not satisfy all objectives when they compete, we explicitly cast contrastive topic modeling as gradient-based multi-objective optimization, with the goal of achieving a Pareto stationary solution. Extensive experiments demonstrate that our framework consistently produces higher-performing neural topic models in terms of topic coherence, topic diversity, and downstream performance.

1 INTRODUCTION

As one of the most prevalent methods for document analysis, topic modeling has been utilized to discover topics of document corpora, with applications spanning from sentiment analysis (Brody & Elhadad, 2010; Naskar et al., 2016; Zhao et al., 2020), to language generation (Jelodar et al., 2019; Nguyen et al., 2021), and to recommender system (Cao et al., 2017; Zhu et al., 2017; Gong et al., 2018; Jelodar et al., 2019). As a conventional technique, Latent Document Analysis (LDA) (Blei et al., 2003) and its extensions perform Gibb sampling and mean field approximation to capture interpretable topic concepts. In recent years, with widespread successes of the Variational Autoencoder (VAE) (Kingma & Welling, 2013), neural network for topic modeling has been proposed, dubbed as Neural Topic Model (Miao et al., 2016), to inherit the encoder-decoder architecture of VAE. Exploiting the standard Gaussian as the prior distribution, neural topic models have not only produced expressive global semantics but also achieved high degree of flexibility and scalability to large-scale document collections (Wang et al., 2021c; 2020; Gupta et al., 2020).

It is well-known that neural topic model training procedure optimizes the evidence lower bound of the log data likelihood. The lower bound jointly lowers the reconstruction and the latent loss. Whereas the first component enhances the reconstruction quality, the second one indirectly pressures the reconstructor by regularizing its inputs, the latent representations, with distributional constraints. Thus, the joint optimization induces a trade-off between the two losses (Higgins et al., 2016; Lin et al., 2019a; Asperti & Trentin, 2020). One resolution is to introduce the optimal weight on the latent term to balance the contradiction, which requires computationally exhaustive and possibly prohibitively expensive manual hyperparameter tuning (Rybkin et al., 2021). To circumvent the problem, previous topic modeling research integrates contrastive learning with its pretext task, instance discrimination, as an auxiliary objective (Le & Akoglu, 2019; Nguyen & Luu, 2021; Li et al., 2022), since input contrastion has been proven as an effective regularizer to promote generalizability (Tang et al., 2021; Kim et al., 2021; Lee et al., 2022). By iteratively discriminating different samples, contrastive learning is able to directly adapt to the influence upon the inputs without depending on manually external assessment.

Figure 1: Illustration of the intense low-level feature influence on produced topics. We record the cosine similarity of the *input* with the *document instance 1* and *document instance 2*'s topic representations generated by NTM+CL (Nguyen & Luu, 2021) and our neural topic model. Although the *input* and *instance 2* both portray the *electric* topic, the similarity for the (*input*, *instance 1*) pair is higher, because the *input* and *instance 1* share the number of non-zero entries and maximum-minimum frequency ratio.

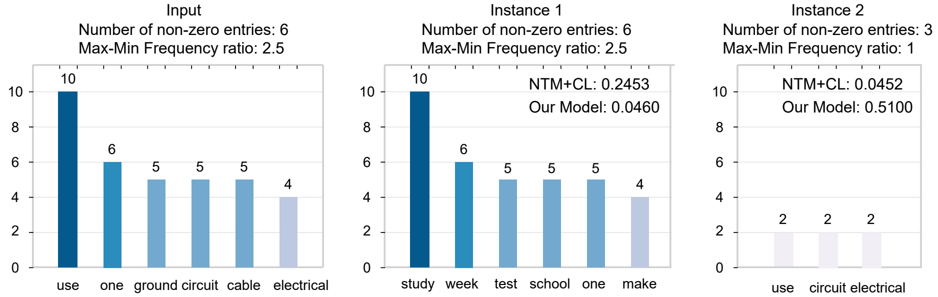


Table 1: Illustration of the effects of peculiar words on topic representations. We record the cosine similarity of the *input* and the *document instances*' topic representations generated by NTM+CL (Nguyen & Luu, 2021) and our neural topic model. As shown with underlined numbers, inserting unusual term, such as *zeppelin* or *scardino*, unexpectedly raises the similarity of two unrelated textual sequences, even surpassing the similarity of the semantically close pair.

Input	Document Instance	Cosine Similarity	
		NTM+CL	Our Model
shuttle lands on planet	job career ask development	0.0093	0.0026
	star astronaut planet light moon	0.8895	0.9178
shuttle lands on planet *zeppelin/scardino*	job career ask development *zeppelin/scardino*	0.9741/0.9413	0.0064/0.0080
	star astronaut planet light moon	0.1584/0.2547	0.8268/0.7188

However, in order to discriminate input instances, contrastive learning mainly involves mutual information, which arouses two significant challenges for neural topic models. First of all, instance discrimination could intensely concentrate on low-level mutual features (Tschannen et al., 2019), for instance, the non-zero entry quantity and the ratio of maximum-minimum frequency, which are illustrated in Figure 1 as incurring irrelevant topic representations. Second, as the contrastive objective bears a regularizing impact to improve generalization of hidden representations, generalizable but inefficient features to topic modeling could emerge, especially when the amount of mutual information exceeds the minimal sufficient statistics (Tian et al., 2020). For example, the discrimination task could go beyond mutual topic themes to encode the common existence of peculiar words or phrases, as depicted in Table 1. As a result, there is a need to control the encoding of the contrastive objective to rightly balance the influence of mutual information on the neural topic model training.

To address the first issue, we hypothesize that useful signals for neural topic models should be shared among a group of input documents (Hjelm et al., 2018). For example, if the model aims to seize the sport topic, it has to detect the common sport theme in multiple documents. Intuitively, such shared semantics define sufficient information to encourage topic models to attain efficacious features, and the learned representations should reflect the multi-sample motivation. To this end, we propose a novel Set-based Contrastive Learning for Neural Topic Models. The set-based concept is realized by arbitrarily drawing a mini-batch and dividing samples into equivalent sets in each iteration. For set representation extraction, we employ the class of symmetric functions to accomplish order-invariant encodings. Furthermore, with a view to maximizing the training efficiency, we propose to permute the mini-batch a number of times to increase the set quantity.

Regarding the second issue, an apparent resolution is to predetermine linear weights for the ELBO and contrastive losses, where the weights function as the inductive bias to manage their effects. Nonetheless, in general scenarios, if the tasks conflict with each other, the linear-combination formulation proves to be a deficient method since there does not exist a parameter set satisfying all objectives (Liang et al., 2021; Mahapatra & Rajan, 2020). Unfortunately, those schemes apply for contrastive neural topic models, since the contrastive element could encode excessive mutual information that competes with useful features learned by the ELBO loss (Tian et al., 2020). To this end, we propose to formulate the training of the neural topic model as a multi-objective optimization problem. The optimization takes into account the gradients of the contrastive and the ELBO loss to find the Pareto stationary solution, which optimally balances the trade-off among objectives (Sener & Koltun, 2018). Our proposed approach regulates the encoder so that it polishes the overall encoded topics of neural topic models, as measured by topic interpretability in the empirical study.

In a nutshell, the contributions of our paper are:

- We propose a novel self-supervised constrastive learning task of set discrimination for neural topic models.
- We reformulate our contrastive-aided topic modeling as a multi-task learning problem and propose to adapt multi-objective optimization algorithm to find a Pareto solution that moderates the effects of multiple objectives on the topic model parameter update.
- Extensive experiments on four popular topic modeling datasets demonstrate that our approaches can enhance contrastive neural topic models in terms of topic coherence, topic diversity, and downstream performance.

2 Related Works

Neural Topic Models. We consider neural topic models proposed by (Krishnan et al., 2018; Miao et al., 2016; Srivastava & Sutton, 2017; Dieng et al., 2020; Burkhardt & Kramer, 2019; Card et al., 2018; Nguyen & Luu, 2021) as the closest line of related works to ours. Recent research has further sought to constrain the capacity of the latent channel in VAE, thus encouraging the model to learn more general and disentangled factors. In particular, to advance visual latent information bottleneck, (Higgins et al., 2016) propose to elevate the latent objective weight, (Mathieu et al., 2019) decompose latent representations, and (Ren et al., 2021) discover traversal directions in generative factors of unsupervised models. Despite all of those achievements in visual modeling, little effort has been put into polishing the generalizability of latent representations in neural topic models.

Contrastive Representation Learning. Contemporary contrastive learning mechanisms follow the sample-wise approach. Several proposed methods to elegantly generate samples comprise using the momentum encoder to estimate positive views (He et al., 2020), the memory bank to store negative ones (Wu et al., 2018), and modifying original salient/non-salient entries to create negative/positive samples (Nguyen & Luu, 2021). Nevertheless, in instance-based contrastive learning, each input is assumed to perceive a disparate category that different ones are concluded to be separate, regardless of some level of potential analogy between them. To address the issue, a surge of interest presents the cluster concept into contrastive representation learning, which groups hidden features into clusters and performs discrimination among groups (Guo et al., 2022; Li et al., 2020; 2021; Caron et al., 2020; Wang et al., 2021b). On the other hand, our method operates upon the input level, directly divides raw documents into random sets, and only aggregates features on-the-fly during training.

Multi-Objective Optimization for Multi-Task Learning (MTL). Latest myriad efforts have been interested in Multi-Objective Optimization (MOO) for Multi-Task Learning, which can address the quandary of optimizing competing objectives for deep neural networks (Mahapatra & Rajan, 2020; Lin et al., 2019b; 2020). Within the framework, the most prominent strategy advocates the gradient-based approach, leveraging the back-propagated gradients to determine a descent direction for all loss values. Developed by (Désidéri, 2012), the gradient based MOO has been applied to image classification (Mahapatra & Rajan, 2020; Liang et al., 2021), semantic segmentation (Liang et al., 2021), sequential decision making (Roijers et al., 2013), and multi-agent learning (Ghosh et al., 2013; Pirotta & Restelli, 2016; Parisi et al., 2014). In our circumstance, we propose a novel interpretation of the contrastive neural topic model from the multi-objective optimization viewpoint, in order to resolve the potential conflict of redundant mutual information with topic modeling signals.

3 Background

3.1 Neural Topic Models

Contemporary neural topic model (NTM) inherits the framework of Variational Autoencoder (VAE) where latent variables are construed as topics. Suppose the linguistic corpus possesses V unique words, i.e. vocabulary size, each document is represented as a word count vector $\mathbf{x} \in \mathbb{R}^V$ and a latent distribution over K topics $\mathbf{z} \in \mathbb{R}^K$. The NTM makes an assumption that $\mathbf{z}$ is generated from a prior distribution $p(\mathbf{z})$ and $\mathbf{x}$ is from the conditional distribution $p_\phi(\mathbf{x}|\mathbf{z})$ parameterized by a decoder ϕ. The goal of the model is to discover the document-topic distribution given the input. The discovery is implemented as the estimation of the posterior distribution $p(\mathbf{z}|\mathbf{x})$, which is approximated by the variational distribution $q_\theta(\mathbf{z}|\mathbf{x})$, modelled by an encoder θ. Inspired by VAEs, the NTM is trained to minimize the following objective based upon the Evidence Lower BOund (ELBO):

$$\min_{\theta,\phi} \mathcal{L}_{\text{ELBO}} = -\mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})} \left[\log p_\phi(\mathbf{x}|\mathbf{z})\right] + \mathbb{KL} \left[q_\theta(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\right] \tag{1}$$

The first term above denotes the expected log-likelihood or the reconstruction error. The second term is the Kullback-Leibler divergence that regularizes $q_\theta(\mathbf{z}|\mathbf{x})$ to be close to the prior distribution $p(\mathbf{z})$. By enforcing the proximity of the latent distribution to the prior, the regularization term puts a constraint on the latent bottleneck to bound the reconstruction capacity, thus inducing a trade-off between the two terms in the ELBO-based objective.

3.2 Contrastive Representation Learning

Contrastive formulation maximizes the agreement by learning similar representations for different views of the same input (called positive pairs), and ensures the disagreement via generating dissimilar representations of disparate inputs (called negative pairs). Technically, given a B-size batch of input documents $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_B\}, \mathbf{x}_i \in \mathbb{R}^V$, firstly data augmentation $t \sim \mathcal{T}$ is applied to each sample $\mathbf{x}_i$ to acquire the augmented view $\mathbf{y}_i$, then form the positive sample $(\mathbf{x}_i, \mathbf{y}_i)$, and pairs that involve augmentations of distinct instances $(\mathbf{x}_i, \mathbf{y}_j), i \neq j$, become negative samples. Subsequently, a prominent approach is to train a projection function f by minimizing the InfoNCE loss (Oord et al., 2018):

$$\mathcal{L}_{\text{nce}} = -\mathbb{E}_{x_i \sim X} \left[\log \frac{e^{f(\mathbf{x}_i, \mathbf{y}_i)}}{\sum_{j=1}^{N} e^{f(\mathbf{x}_i, \mathbf{y}_j)}}\right], \tag{2}$$

where f defines a similarity mapping $\mathbb{R}^V \times \mathbb{R}^V \rightarrow \mathbb{R}$, estimated as follows:

$$f(\mathbf{x}, \mathbf{y}) = \frac{g_\varphi(\mathbf{x})^T g_\varphi(\mathbf{y})}{\parallel g_\varphi(\mathbf{x}) \parallel \parallel g_\varphi(\mathbf{y}) \parallel} / \tau, \tag{3}$$

where g_φ denotes the neural network parameterized by φ, τ the temperature to rescale the score from the feature similarity.

3.3 Gradient-based Multi-Objective Optimization for Multi-Task Learning

A MTL problem considers a tuple of M tasks with a vector of non-negative losses:

$$\min_\theta \mathcal{L}(\theta) = (\mathcal{L}_1(\theta), \mathcal{L}_2(\theta), \ldots, \mathcal{L}_M(\theta))^T, \tag{4}$$

where $\mathcal{L}_i$ denotes the loss of the i-th task, θ the parameter of the shared hypothesis. In most circumstances, no single parameter set can accomplish the minimal values for all loss functions. Instead, the more prevailing solution is to seek the Pareto stationary solution, which sustains optimal trade-offs among objectives (Sener & Koltun, 2018; Mahapatra & Rajan, 2020). Theoretically, a Pareto stationary solution satisfies the Karush-Kuhn-Tucker (KKT) conditions as follows,

Theorem 1. *Let θ^* denote a Pareto point. Then, there exists non-negative scalars $\{\alpha_i\}_{i=1}^M, i \in \{1, 2, \ldots, M\}$ such that*

$$\sum_{i=1}^{M} \alpha_i = 1 \quad \textit{and} \quad \sum_{i=1}^{M} \alpha_i \nabla \mathcal{L}_i(\theta^*) = 0 \tag{5}$$

In consequence, the conditions lead to the following optimization problem:

$$\min_{\{\alpha_i\}_{t=1}^M} \left\{ \left\| \sum_{i=1}^{M} \alpha_i \nabla_{\theta^{\text{shared}}} \mathcal{L}_i(\theta^{\text{shared}}, \theta^i) \right\|_2^2 \middle| \sum_{i=1}^{M} \alpha_i = 1, \alpha_i \geq 0 \quad \forall i \right\} \tag{6}$$

Either the solution to the above problem does not exist, i.e. the outcome meets the KKT conditions, or the solution supplies a descent direction that decreases all per-task losses (Désidéri, 2012). Hence, the procedure would become gradient descent on task-oriented parameters followed by solving (6) and adapting $\sum_{i=1}^{M} \alpha_i \nabla_{\theta^{\text{shared}}} \mathcal{L}_i(\theta^{\text{shared}}, \theta^i)$ as the gradient update upon shared parameters.

4 Methodology

In this section, we introduce the details and articulate the analysis of our proposed set-based contrastive learning approach. Thereupon, we delineate the multi-objective optimization framework for contrastive neural topic models.

4.1 Set-based Contrastive Learning for Neural Topic Model

In this work, we take a step further by generalizing previous instance discrimination to the set representation learning by designating the network to capture order-invariant features. Future work can be extended to investigate the efficacy of other data structures in addition to the set.

Sample augmentation. Firstly, given a mini-batch of B samples, we perform data augmentation $t \sim \mathcal{T}$ to generate a set of B counterpart samples $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_B\}$, where $\mathbf{y}_i = t(\mathbf{x}_i)$ for $i \in \{1, 2, \dots, B\}$. We develop the Embedding-based Sampling (ES) method for the inputs of the neural topic model. In particular, given each word $j \in \{1, 2, \dots, V\}$ in $\mathbf{x}_i$ whose $x_{i,j} > 0$, we extract S words $\{w_{j,1}, w_{j,2}, \dots, w_{j,S}\}$, whose pretrained embeddings are spatially nearest to the one of word j. Subsequently, we randomize from $\{w_{j,s}\}_{s=1}^{S}$ and substitute $x_{i,j}$ into the $w_{j,s}$-th entry of $\mathbf{y}_i$. Such substitution creates a new document with semantically similar words to the original one, thus preserving the original semantics and validating the veracity of our data augmentation approach.

Algorithm 1: Pseudo-code of Multi-task Set-based Contrastive Neural Topic Model.

```
# Dataset $D$, batch size $B$, encoder
 parameter $\theta$, decoder parameter $\phi$,
 augmentation $t \sim \mathcal{T}$, learning rate $\eta$,
 temperature $\tau$
for minibatch in D :
    # neural topic model step
    for $i = 1$ to B :
        $\mathbf{y}_i = t(\mathbf{x}_i), \mathbf{z}_i = f_\theta(\mathbf{x}_i),\ \mathbf{z}'_i = f_\theta(\mathbf{y}_i)$
    $\mathcal{L}_{\texttt{ELBO}} = \frac{1}{B}\sum_{i=1}^{B} \mathcal{L}_{\texttt{VAE}}(\mathbf{x}_i)$
    # set-based contrastive learning
    let $\pi$ = permutation matrix of size
     $P \times B$
    let $N = \frac{BP}{K}$, let $H$ = []
    for $i = 1$ to N :
        $\kappa = \pi[:, (i-1)K + 1 : iK]$
        $\mathbf{h} = \phi\left(\mathbf{z}_{\kappa_1}, \mathbf{z}_{\kappa_2}, ..., \mathbf{z}_{\kappa_K}\right)$
        $\mathbf{h}' = \phi\left(\mathbf{z}'_{\kappa_1}, \mathbf{z}'_{\kappa_2}, ..., \mathbf{z}'_{\kappa_K}\right)$
        $H$.append($\mathbf{h}$, $\mathbf{h}'$)
    $s_{i,j} = \frac{H_i^\top H_j}{\|H_i\|\|H_j\|}/\tau; 1 \le i, j \le 2N$
    $\mathcal{L}_{\texttt{nce}} =$
     $-\frac{1}{2N}\sum_{l=1}^{N} \log\left(\frac{e^{s_{2l-1,2l}}}{\sum_{j=1}^{2N} e^{s_{2l-1,j}}} \cdot \frac{e^{s_{2l,2l-1}}}{\sum_{j=1}^{2N} e^{s_{2l,j}}}\right)$
    # Multi-objective optimization
    $\alpha = \texttt{solver}(\nabla_\theta \mathcal{L}_{\texttt{nce}}, \nabla_\theta \mathcal{L}_{\texttt{ELBO}})$
    $G_\theta = \alpha \nabla_\theta \mathcal{L}_{\texttt{nce}} + (1-\alpha)\nabla_\theta \mathcal{L}_{\texttt{ELBO}}$
    $G_\phi = \nabla_\phi \mathcal{L}_{\texttt{ELBO}}$
    $\theta = \theta - \eta \cdot G_\theta,\ \phi = \phi - \eta \cdot G_\phi$
return $\theta$ and $\phi$
```

Feature Extraction. For a mini-batch, each set is constructed by grouping every K input samples. Subsequently, we proceed to extract the set's representation as follows,

$$\mathbf{h} = \varphi(f_\theta(\mathbf{x}_1), f_\theta(\mathbf{x}_2), \dots, f_\theta(\mathbf{x}_K)), \tag{7}$$

$$\mathbf{h}' = \varphi(f_\theta(\mathbf{y}_1), f_\theta(\mathbf{y}_2), \dots, f_\theta(\mathbf{y}_K)), \tag{8}$$

where f_θ denotes the θ-parameterized encoder of the neural topic model, φ the symmetric pooling operator. We exert this formulation for all of the experiments and report its variants with different pooling functions in the Appendix.

Input Permutation. A plain process to generate sets of K incidents would be to distribute every K samples into a set. Unfortunately, this method yields deficient performance because of the following reasons. Firstly, the number of sets is limited as this naive approach only produces $\lfloor \frac{B}{K} \rfloor$ sets, thus $\lfloor \frac{B}{K} \rfloor$ positive and $2\left(\lfloor \frac{B}{K} \rfloor - 1\right)$ negative pairs for each mini-batch. Secondly, the utilization of input samples is sub-optimal because each sample is only considered once for the set construction.

To cope with those dilemmas, we propose to permute the input indices to augment both positive and negative set pairs. Procedurally, in the beginning, we extract hidden features of B input samples. Next, we create a permutation matrix of size $P \times B$ in which each row is a list of permuted indices from 1 to B, and inherently P is the number of times we shuffle the indices. Thereafter, each set can be established by assembling K elements corresponding to the permutation matrix. Eventually, we conduct pooling to acquire features of the augmented and original sample sets.

Our proposed approach brings about the burgeoning number of positive and negative pairs through raising the set number to $\lfloor \frac{BP}{K} \rfloor$ sets, thus maximizing the capacity of the set-based contrastive representation learning. Empirical analysis shows the effectiveness of our permutation step.

4.2 Explanatory Analysis: Merits of Set-based Contrastive Learning for Neural Topic Model

As indicated in the literature, contrastive discrimination approach will provide valuable representations if the sampling strategy can form hard negatives and the representation learning can conquer shared information (Zhang & Stratos, 2021; Hjelm et al., 2018). Here we hypothetically and theoretically prove that our set-based contrastive learning enjoys both merits to better polish topic representations of the neural topic model than the instance-based technique.

Hard Negative Sampling. First of all, we devise a permutation method to enable an input document to occur in multiple sets. As a result, sets are composed of similar instances and partially overlap with one another, making differentiating among them more difficult.

Secondly, in some circumstances, hard negatives drawn by individual discrimination systems are indeed false negatives because they belong to the same document semantic category (Robinson et al., 2020), leading to semantic collapse. Interestingly enough, our method is able to mitigate such representation collapse. Particularly, assume that the predefined topic number is C, and the topics are balancingly distributed, then the cardinalities of semantic categories for individual and set discrimination schemes are C and C^K, respectively. Correspondingly, the likelihood of incorrectly contrasting samples from the same topic category for set discrimination $p_{\text{set}}(\text{false negative}|\text{hard negative}) = 1/(C^K - 1)$ (every set has one augmented sample), whereas for individual discrimination $p_{\text{ind}}(\text{false negative}|\text{hard negative}) = 1/C \geq 1/(C^K - 1)$ $(K, C \geq 2)$, i.e. semantic collapse is harder to take place for set-based than individual-based contrastive learning.

That being the case, set discrimination supplies more correctly hard negatives and emerges to be more daunting than the instance one, consequently leading to better robustness of the topic representation learning (Wang et al., 2021a; Ge et al., 2021; Kalantidis et al., 2020).

Common Feature Learning. We prove the common feature learning property of our proposed set-based contrastive method from two perspectives. First of all, the set-based InfoNCE loss favors mutual information that is prevalent among documents, according to the following theorem:

Theorem 2. *Let $\mathbf{x}_i, \mathbf{x}_j, \mathbf{y}_i, \mathbf{y}_j$ be the two elements of a set, and their corresponding augmented variants, i.e. $t(\mathbf{x}_i) = \mathbf{y}_i, t(\mathbf{y}_j) = \mathbf{y}_j$; $\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}'_i, \mathbf{z}'_j$ be the topic representations of $\mathbf{x}_i, \mathbf{x}_j, \mathbf{y}_i, \mathbf{y}_j$, i.e. $f(\mathbf{x}_i) = \mathbf{z}_i, f(\mathbf{x}_j) = \mathbf{z}_j, f(\mathbf{y}_i) = \mathbf{z}'_i, f(\mathbf{y}_j) = \mathbf{z}'_j$; $\mathbf{z}$ and $\mathbf{z}'$ be set features pooled from $(\mathbf{z}_i, \mathbf{z}_j)$ and $(\mathbf{z}'_i, \mathbf{z}'_j)$, respectively. Define $I(\mathbf{x}, \mathbf{y})$ to be the mutual information between two variables $\mathbf{x}$ and $\mathbf{y}$, then:*

$$\mathcal{L}_{\textit{InfoNCE}}(\mathbf{z}, \mathbf{z}') \leq I(\mathbf{z}, \mathbf{x}_i) + I(\mathbf{z}, \mathbf{x}_j) + I[(\mathbf{x}_i, \mathbf{x}_j)|\mathbf{z}]. \tag{9}$$

We provide the proof of the theorem (2) in the Appendix. As it can be observed, the upper bound of the set-based InfoNCE loss resembles the objective in Deep InfoMax (Hjelm et al., 2018), which captures common information across multiple inputs.

Second of all, despite aggregating several documents into a set, set representations uses the same dimension with instance representations. This bottlenecks set representations in the original boundary, thus discouraging the topic encoder to learn specific features to a single instance, since sparing space for instance-specific features will decrease the capacity for mutual features with other instances.

4.3 Multi-Objective Optimization for Neural Topic Model

Problem Statement. For a given neural topic model, we seek to find preference vectors $\boldsymbol{\alpha}$ that produces adept gradient update policy $\sum_i \alpha_i \nabla f_i$. For the contrastive neural topic model, because only the neural topic encoder receives the gradients of both contrastive and topic modeling objectives, we focus on modulating the encoder update while keeping the decoder learning intact. Hence, motivated by formulation (6), our optimization problem becomes:

$$\min_{\alpha} \left\{ \left\| \alpha \nabla_\theta \mathcal{L}_{\text{nce}}(\theta) + (1 - \alpha) \nabla_\theta \mathcal{L}_{\text{ELBO}}(\theta, \phi) \right\|_2^2 \middle| \alpha \geq 0 \right\} \tag{10}$$

which is a convex quadratic optimization with linear constraints. Erstwhile research additionally integrates prior information to compose solutions that are preference-specific among tasks (Mahapatra & Rajan, 2020; Lin et al., 2019b; 2020). However, for neural topic modeling domain, such information is often unavailable, so we exclude it from our problem notation.

Optimization Solution. Our problem formulation involves convex optimization of multi-dimensional quadratic function with linear constraints. Formally, we derive the analytical solution $\hat{\alpha}$ as:

$$\hat{\alpha} = \text{solver}[\nabla_\theta \mathcal{L}_{\text{nce}}(\theta), \nabla_\theta \mathcal{L}_{\text{ELBO}}(\theta, \phi)] = -\left[\frac{\left(\nabla_\theta \mathcal{L}_{\text{nce}}(\theta) - \nabla_\theta \mathcal{L}_{\text{ELBO}}(\theta, \phi)\right)^T \nabla_\theta \mathcal{L}_{\text{nce}}(\theta)}{\left\| \nabla_\theta \mathcal{L}_{\text{nce}}(\theta) - \nabla_\theta \mathcal{L}_{\text{ELBO}}(\theta, \phi) \right\|_2^2} \right]_+ \tag{11}$$

where $[.]_+$ denotes the ReLU operation. After accomplishing $\hat{\alpha}$, we plug the vector to control the gradient for the neural topic encoder, while maintaining the update upon the neural topic decoder whose value is determined by the back-propagation of the reconstruction loss. The pseudo-code of our Multi-Task Learning framework for Set-based Contrastive Neural Topic Model is depicted in Algorithm 1.

5 Experiments

In this section, we conduct experiments and empirically demonstrate the effectiveness of the proposed methods for neural topic models. We provide the experimental setup and report numerical results along with qualitative studies. Examples of the produced topics can be found in the Appendix.

5.1 Experimental Setup

Benchmark Datasets. We adopt popular benchmark datasets spanning various domains, vocabulary sizes, and document lengths for experiments: (i) **20Newsgroups (20NG)** (Lang, 1995), one of the most well-known datasets for topic model evaluation, consisting of 18000 documents with 20 labels; (ii) **IMDb** (Maas et al., 2011), the dataset of movie reviews, belonging to two sentiment labels, i.e. positive and negative; (iii) **Wikitext-103 (Wiki)** (Merity et al., 2016), comprising 28500 articles from the Good and Featured section on Wikipedia; (iv) **AG News** (Zhang et al., 2015), consisting of news titles and articles whose size is 30000 and 1900 for training and testing subsets, respectively.

Evaluation Metrics. To evaluate the topic coherence of generated topics, we follow previous works (Hoyle et al., 2020; Wang et al., 2019; Card et al., 2018; Nguyen & Luu, 2021) to extract the top 10 words, utilize the testing split of each dataset as the reference corpus to calculate the Normalized Pointwise Mutual Information (NPMI) score for every topic, then obtain the topics' mean value. Furthermore, because the topic quality also depends on the usefulness for downstream tasks, we estimate the text classification performance of the produced topics, leveraging the F1 score as the evaluation measure.

Baseline Models. We compare our proposed approaches with the following state-of-the-art neural topics models: (i) **(NTM)** (Miao et al., 2016), a topic model that inherits the encoder-decoder paradigm of the VAE architectue and standard Gaussian as the prior distribution; (ii) **(ETM)** (Dieng et al., 2020), a NTM which models the topic-word distribution with topic and word embeddings; (iii) **(DVAE)** (Burkhardt & Kramer, 2019), a NTM whose Dirichlet is the prior for both topic and word distributions; (iv) **BATM** (Wang et al., 2020), a GAN-based neural topic model which is composed of an encoder, a generator, and a discriminator; (v) **W-LDA** (Nan et al., 2019), an application of Wasserstein autoencoder as a topic model; (vi) **SCHOLAR** (Card et al., 2018), a NTM that incorporates external variables and uses logistic normal as the prior distribution; (vii) **SCHOLAR + BAT** (Hoyle et al., 2020), a SCHOLAR model applying knowledge distillation in which BERT is the teacher and the neural topic model is the student; (viii) **NTM + CL** (Nguyen & Luu, 2021), a NTM which jointly trains evidence lower bound and the individual-based contrastive learning objective.

Implementation Details. For implementation, we designate MaxPooling as the set representation extraction function, Word2Vec (Mikolov et al., 2013) as pretrained embedding for our embedding-based augmentation, and experiment with different set cardinality $K \in \{1, 2, 3, 4, 5, 6\}$ and permutation matrix size $P \in \{1, 2, 4, 8, 16, 32\}$. More implementation details of neural topic models can be found in the Appendix.

Table 2: Topic Coherence and Topic Diversity results on 20NG and IMDb datasets.

Method	20NG				IMDb			
	$C = 50$		$C = 200$		$C = 50$		$C = 200$	
	NPMI	TD	NPMI	TD	NPMI	TD	NPMI	TD
NTM	0.283±0.004	0.734±0.009	0.277±0.003	0.686±0.004	0.170±0.008	0.777±0.021	0.169±0.003	0.690±0.015
ETM	0.305±0.006	0.776±0.022	0.264±0.002	0.623±0.002	0.174±0.001	0.805±0.019	0.168±0.001	0.687±0.007
DVAE	0.320±0.005	0.824±0.017	0.269±0.003	0.786±0.005	0.183±0.004	0.836±0.010	0.173±0.006	0.739±0.005
BATM	0.314±0.003	0.786±0.014	0.245±0.001	0.623±0.008	0.065±0.008	0.619±0.016	0.090±0.004	0.652±0.008
W-LDA	0.279±0.003	0.719±0.026	0.188±0.001	0.614±0.002	0.136±0.007	0.692±0.016	0.095±0.003	0.666±0.009
SCHOLAR	0.319±0.007	0.788±0.008	0.263±0.002	0.634±0.006	0.168±0.002	0.702±0.014	0.140±0.001	0.675±0.005
SCHOLAR + BAT	0.324±0.006	0.824±0.011	0.272±0.002	0.648±0.009	0.182±0.002	0.825±0.008	0.175±0.003	0.761±0.010
NTM+CL	0.332±0.006	0.853±0.005	0.277±0.003	0.699±0.004	0.191±0.004	0.857±0.010	0.186±0.002	0.843±0.008
Our Model	**0.340±0.005**	**0.913±0.019**	**0.291±0.003**	**0.905**±0.004	**0.200±0.007**	**0.916±0.008**	**0.197±0.003**	**0.892±0.007**

Table 3: Topic Coherence and Topic Diversity results on Wiki and AG News datasets.

Method	Wiki				AG News			
	$C = 50$		$C = 200$		$C = 50$		$C = 200$	
	NPMI	TD	NPMI	TD	NPMI	TD	NPMI	TD
NTM	0.250±0.010	0.817±0.006	0.291±0.009	0.624±0.011	0.197±0.015	0.729±0.007	0.205±0.002	0.797±0.002
ETM	0.332±0.003	0.756±0.013	0.317±0.009	0.671±0.006	0.204±0.004	0.736±0.006	0.055±0.002	0.683±0.005
DVAE	0.404±0.006	0.815±0.009	0.359±0.007	0.721±0.011	0.271±0.016	0.850±0.013	0.182±0.001	0.746±0.005
BATM	0.336±0.010	0.807±0.005	0.319±0.005	0.732±0.012	0.256±0.019	0.754±0.008	0.144±0.003	0.711±0.006
W-LDA	0.451±0.012	0.836±0.007	0.308±0.007	0.725±0.014	0.270±0.033	0.844±0.014	0.135±0.004	0.708±0.004
SCHOLAR	0.429±0.011	0.821±0.010	0.446±0.009	0.860±0.014	0.278±0.024	0.886±0.007	0.185±0.004	0.747±0.002
SCHOLAR + BAT	0.446±0.010	0.824±0.006	0.455±0.007	0.854±0.011	0.272±0.027	0.859±0.008	0.189±0.003	0.750±0.001
NTM+CL	0.481±0.005	0.841±0.012	0.462±0.006	0.831±0.016	0.279±0.025	0.890±0.007	0.190±0.002	0.790±0.006
Our Model	**0.496±0.010**	**0.959±0.010**	**0.491±0.006**	**0.938±0.011**	**0.325±0.012**	**0.934±0.012**	**0.221±0.003**	**0.885±0.004**

5.2 Main Results

Overall Results. Table 2 and 3 show the topic quality results with topic number $C = 50$ and $C = 200$. We run all models five times with different random seeds and record the mean and standard deviation of the results. As it can be seen, we have the remarks that our method achieves the best topic coherence on three benchmark datasets. We outperform the state-of-the-art NTM+CL (Nguyen & Luu, 2021) and other baselines in both $C = 50$ and $C = 200$ scenarios. We also conduct significance tests to prove the significant and robust improvement of our multi-task set-based contrastive framework, which can be found in the Appendix.

5.3 Document Classification

For the downstream experiment, we extract latent distributions inferred by neural topic models in the $C = 50$ setting and train Random Forest classifiers with the number of decision trees of 100 to predict the class of each input. 20NG and IMDb datasets are designated for the experiment. As shown in Table 4, our model accomplishes the best performance over other neural topic models, substantiating the refined usefulness of the topic representations yielded by our model.

5.4 Ablation Study

Objective Control Strategy. We perform an ablation study and analyze the importance of our proposed multi-objective training framework. In detail, we remove the Pareto stationary solver and then conduct the training with the approaches of linear combination, in which we attach the weight α to the contrastive objective with $\alpha \in \{0.25, 0.5\}$, the Uncertainty Weighting (UW) (Kendall et al., 2018), gradient normalization (GradNorm) (Chen et al., 2018), Projecting Conflicting Gradients (PCGrad) (Yu et al., 2020), and Random Weighting (RW) (Lin et al., 2021). Without the multi-objective framework, the contrastive neural topic model suffers from sub-optimal performance, as shown in Table 5. Hypothetically, this means that in the fixed linear and heuristics weighting systems, the optimization procedure could not efficiently adapt to the phenomenon that the contrastive objective overwhelms the topic modeling task. Hence, contrasting samples might supply excessive mutual information which eclipses the useful features for topic learning.

Sample Set Size. Here we investigate the influence of the set size upon the topic representations of the neural topic model. Procedurally, for $P = 8$, we gradually increase the number of elements, execute the training, and plot the results, which are l2-normalized for visibility, in Figure 2. We observe that the initial growth of the set cardinality results in the performance improvement until it surpasses the threshold at which the burgeon ceases and the topic quality starts decreasing. In-

Table 4: Text classification results with topic representations.

Method	20NG	IMDb
BATM	30.8	66.0
SCHOLAR	52.9	83.4
SCHOLAR + BAT	32.2	73.1
NTM+CL	54.4	84.2
Our Model	**57.0**	**86.4**

Table 5: Ablation Results on the Objective Control Strategy.

Method	Wiki	
	$C = 50$	$C = 200$
UW	0.482±0.009	0.474±0.005
GradNorm	0.485±0.007	0.469±0.006
PCGrad	0.489±0.011	0.472±0.005
RW	0.492±0.009	0.476±0.007
Linear-$\alpha = 0.25$	0.483±0.011	0.487±0.007
Linear-$\alpha = 0.50$	0.491±0.015	0.485±0.004
Our Model	**0.496**±**0.010**	**0.491**±**0.006**

Figure 2: Topic Coherence with different set size.

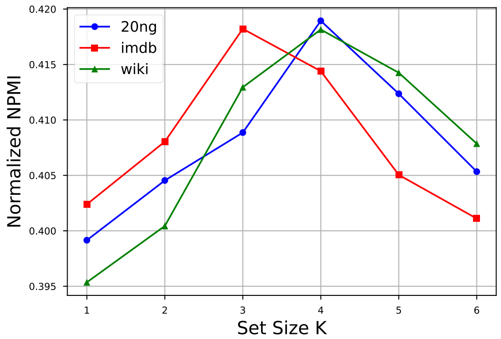
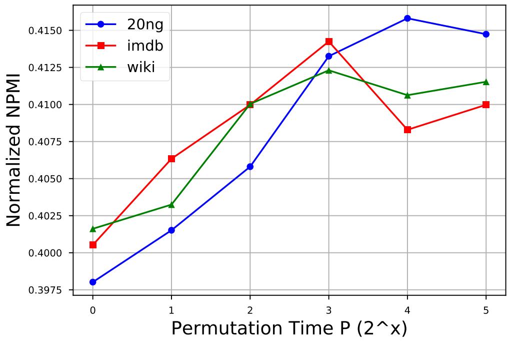


Figure 3: Topic Coherence with different permutation matrix size.



tuitively, we reason that escalating the set size leads to more challenging representation learning because the model needs to gauge common features of more documents. In consequence, only if representation learning is sufficiently difficult can the network preserve auspicious information to benefit the topic distribution outputs.

Permutation Matrix Size. We now extend our study to the effect of the matrix permutation size, i.e. the number of permutation times, on the learned network. Similar to Figure 2, we report the normalized performance of the set-based contrastive neural topic model with varying numbers of permutation $P \in \{1, 2, 4, 8, 16, 32\}$ for $K = 4$ in Figure 3. We realize that set-based discrimination becomes more effective when the permutation matrix grows because there are elevated numbers of positive and negative sets. However, when there is overwhelming contrastive information, our multi-task framework produces adaptive weights to constrain the impact of the discrimination objective, inducing the convergence of the topic coherence quality.

5.5 Analysis

Effects of Set-based and Instance-based Contrastive Learning on Topic Representations. As mentioned, Table 1 and Figure 1 include the similarity scores of the input with different document samples. Different from NTM+CL (Nguyen & Luu, 2021), our model does pull together latent representations of semantically close topics. For example, in Figure 1, *instance 2* still maintains high equivalence level, despite its noticeable distinction with the input in terms of the low-level properties. Furthermore, in Table 1, inserting unfamiliar words does not alter the similarity scale between the input with two document samples. These results indicate that our set-based contrastive learning focuses on encoding topic information and avoids low-level vector features.

6 Conclusion

In this paper, we propose a novel and well-motivated Multi-Task Set-based Contrastive Neural Topic Model. Our contrastive approach incorporates the set concept to encourage the learning of common instance features and constructs hard negative samples that partially overlap. We propose to adapt multi-objective optimization to efficiently polish the encoded topic representations. Extensive experiments indicate that our framework accomplishes state-of-the-art performance in terms of producing both high-quality and useful topics.

References

Andrea Asperti and Matteo Trentin. Balancing reconstruction error and kullback-leibler divergence in variational autoencoders. *IEEE Access*, 8:199440–199448, 2020.

Normand J Beaudry and Renato Renner. An intuitive proof of the data processing inequality. *Quantum Information & Computation*, 12(5-6):432–441, 2012.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Samuel Brody and Noemie Elhadad. An unsupervised aspect-sentiment model for online reviews. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pp. 804–812, 2010.

Sophie Burkhardt and Stefan Kramer. Decoupling sparsity and smoothness in the dirichlet variational autoencoder topic model. *J. Mach. Learn. Res.*, 20(131):1–27, 2019.

Da Cao, Xiangnan He, Liqiang Nie, Xiaochi Wei, Xia Hu, Shunxiang Wu, and Tat-Seng Chua. Cross-platform app recommendation by jointly modeling ratings and texts. *ACM Transactions on Information Systems (TOIS)*, 35(4):1–27, 2017.

Dallas Card, Chenhao Tan, and Noah A Smith. Neural models for documents with metadata. 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pp. 794–803. PMLR, 2018.

Jean-Antoine Désidéri. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5-6):313–318, 2012.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020.

Songwei Ge, Shlok Mishra, Chun-Liang Li, Haohan Wang, and David Jacobs. Robust contrastive learning using negative samples with diminished semantics. *Advances in Neural Information Processing Systems*, 34:27356–27368, 2021.

Shaona Ghosh, Chris Lovell, and Steve R Gunn. Towards pareto descent directions in sampling experts for multiple tasks in an on-line learning paradigm. In *2013 AAAI Spring Symposium Series*, 2013.

Yeyun Gong, Qi Zhang, and Xuanjing Huang. Hashtag recommendation for multimodal microblog posts. *Neurocomputing*, 272:170–177, 2018.

Yuanfan Guo, Minghao Xu, Jiawen Li, Bingbing Ni, Xuanyu Zhu, Zhenbang Sun, and Yi Xu. Hcsc: Hierarchical contrastive selective coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9706–9715, 2022.

Pankaj Gupta, Yatin Chaudhary, Thomas Runkler, and Hinrich Schuetze. Neural topic modeling with continual lifelong learning. In *International Conference on Machine Learning*, pp. 3907–3917. PMLR, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2018.

Alexander Miserlis Hoyle, Pranav Goel, and Philip Resnik. Improving neural topic models using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Hamed Jelodar, Yongli Wang, Mahdi Rabbani, and SeyedValyAllah Ayobi. Natural language processing via lda topic model in recommendation systems. *arXiv preprint arXiv:1909.09551*, 2019.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33:21798–21809, 2020.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.

Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9619–9628, 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Rahul Krishnan, Dawen Liang, and Matthew Hoffman. On the challenges of learning with inference networks on sparse, high-dimensional data. In *International conference on artificial intelligence and statistics*, pp. 143–151. PMLR, 2018.

Ken Lang. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, pp. 331–339. Elsevier, 1995.

Tuan Le and Leman Akoglu. Contravis: contrastive and visual topic modeling for comparing document collections. In *The World Wide Web Conference*, pp. 928–938, 2019.

Doyup Lee, Sungwoong Kim, Ildoo Kim, Yeongjae Cheon, Minsu Cho, and Wook-Shin Han. Contrastive regularization for semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3911–3920, 2022.

Jiacheng Li, Jingbo Shang, and Julian McAuley. Uctopic: Unsupervised contrastive learning for phrase representations and topic mining. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6159–6169, 2022.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*, 2020.

Yujia Li, Richard Zemel, Marc Brockschmidt, and Daniel Tarlow. Gated graph sequence neural networks. In *Proceedings of ICLR'16*, 2016.

Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8547–8555, 2021.

Jian Liang, Kaixiong Gong, Shuang Li, Chi Harold Liu, Han Li, Di Liu, Guoren Wang, et al. Pareto domain adaptation. *Advances in Neural Information Processing Systems*, 34:12917–12929, 2021.

Baijiong Lin, Feiyang Ye, and Yu Zhang. A closer look at loss weighting in multi-task learning. *arXiv preprint arXiv:2111.10603*, 2021.

Shuyu Lin, Stephen Roberts, Niki Trigoni, and Ronald Clark. Balancing reconstruction quality and regularisation in elbo for vaes. *arXiv preprint arXiv:1909.03765*, 2019a.

Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, and Sam Kwong. Pareto multi-task learning. *Advances in neural information processing systems*, 32, 2019b.

Xi Lin, Zhiyuan Yang, Qingfu Zhang, and Sam Kwong. Controllable pareto multi-task learning. *arXiv preprint arXiv:2010.06313*, 2020.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

Debabrata Mahapatra and Vaibhav Rajan. Multi-task learning with user preferences: Gradient descent with controlled ascent in pareto optimization. In *International Conference on Machine Learning*, pp. 6597–6607. PMLR, 2020.

Emile Mathieu, Tom Rainforth, Nana Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning*, pp. 4402–4412. PMLR, 2019.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. 2016.

Yishu Miao, Lei Yu, and Phil Blunsom. Neural variational inference for text processing. In *International conference on machine learning*, pp. 1727–1736. PMLR, 2016.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Navid Naderializadeh, Joseph F Comer, Reed Andrews, Heiko Hoffmann, and Soheil Kolouri. Pooling by sliced-wasserstein embedding. *Advances in Neural Information Processing Systems*, 34: 3389–3400, 2021.

Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. Topic modeling with wasserstein autoencoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6345–6381, 2019.

Debashis Naskar, Sidahmed Mokaddem, Miguel Rebollo, and Eva Onaindia. Sentiment analysis in social networks through topic modeling. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 46–53, 2016.

Thong Nguyen and Anh Tuan Luu. Contrastive learning for neural topic model. *Advances in Neural Information Processing Systems*, 34:11974–11986, 2021.

Thong Nguyen, Anh Tuan Luu, Truc Lu, and Tho Quan. Enriching and controlling global semantics for text summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9443–9456, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Simone Parisi, Matteo Pirotta, Nicola Smacchia, Luca Bascetta, and Marcello Restelli. Policy gradient approaches for multi-objective sequential decision making. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pp. 2323–2330. IEEE, 2014.

Matteo Pirotta and Marcello Restelli. Inverse reinforcement learning through policy gradient minimization. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Xuanchi Ren, Tao Yang, Yuwang Wang, and Wenjun Zeng. Learning disentangled representation by exploiting pretrained generative models: A contrastive learning view. In *International Conference on Learning Representations*, 2021.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2020.

Diederik M Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

Oleh Rybkin, Kostas Daniilidis, and Sergey Levine. Simple and effective vae training with calibrated decoders. In *International Conference on Machine Learning*, pp. 9179–9189. PMLR, 2021.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*, 2017.

Shixiang Tang, Peng Su, Dapeng Chen, and Wanli Ouyang. Gradient regularized contrastive learning for continual domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2665–2673, 2021.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020.

Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2019.

Rui Wang, Deyu Zhou, and Yulan He. Atm: Adversarial-neural topic model. *Information Processing & Management*, 56(6):102098, 2019.

Rui Wang, Xuemeng Hu, Deyu Zhou, Yulan He, Yuxuan Xiong, Chenchen Ye, and Haiyang Xu. Neural topic modeling with bidirectional adversarial training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 340–350, 2020.

Weilun Wang, Wengang Zhou, Jianmin Bao, Dong Chen, and Houqiang Li. Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14020–14029, 2021a.

Xudong Wang, Ziwei Liu, and Stella X Yu. Unsupervised feature learning by cross-level instance-group discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12586–12595, 2021b.

Yiming Wang, Ximing Li, and Jihong Ouyang. Layer-assisted neural topic modeling over document networks. In *IJCAI*, pp. 3148–3154, 2021c.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836, 2020.

Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. An end-to-end deep learning architecture for graph classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Wenzheng Zhang and Karl Stratos. Understanding hard negatives in noise contrastive estimation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1090–1101, 2021.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

He Zhao, Dinh Phung, Viet Huynh, Trung Le, and Wray Buntine. Neural topic model via optimal transport. In *International Conference on Learning Representations*, 2020.

Konglin Zhu, Lin Zhang, and Achille Pattavina. Learning geographical and mobility factors for mobile application recommendation. *IEEE Intelligent Systems*, 32(3):36–44, 2017.

A Proof of Theorem 2

Theorem 2. *Let $\mathbf{x}_i, \mathbf{x}_j, \mathbf{y}_i, \mathbf{y}_j$ be the two elements of a set, and their corresponding augmented variants, i.e. $t(\mathbf{x}_i) = \mathbf{y}_i, t(\mathbf{y}_j) = \mathbf{y}_j$; $\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}'_i, \mathbf{z}'_j$ be the topic representations of $\mathbf{x}_i, \mathbf{x}_j, \mathbf{y}_i, \mathbf{y}_j$, i.e. $f(\mathbf{x}_i) = \mathbf{z}_i, f(\mathbf{x}_j) = \mathbf{z}_j, f(\mathbf{y}_i) = \mathbf{z}'_i, f(\mathbf{y}_j) = \mathbf{z}'_j$; $\mathbf{z}$ and $\mathbf{z}'$ be set features pooled from $(\mathbf{z}_i, \mathbf{z}_j)$ and $(\mathbf{z}'_i, \mathbf{z}'_j)$, respectively. Define $I(\mathbf{x}, \mathbf{y})$ to be the mutual information between two variables $\mathbf{x}$ and $\mathbf{y}$, then:*

$$\mathcal{L}_{\textit{InfoNCE}}(\mathbf{z}, \mathbf{z}') \leq I(\mathbf{z}, \mathbf{x}_i) + I(\mathbf{z}, \mathbf{x}_j) + I[(\mathbf{x}_i, \mathbf{x}_j)|\mathbf{z}]. \tag{12}$$

Proof. The probabilistic graphical model of our problem formulation can be denoted as follows,

$$\begin{tikzcd}
& (X_1, X_2) \arrow[dl] \arrow[dr] & & & & (X_1, X_2) \arrow[dr] & \\
(Z_1, Z_2) \arrow[d] & & (Z'_1, Z'_2) \arrow[d] & \text{, which is I-equivalent to} & (Z_1, Z_2) \arrow[ur] & & (Z'_1, Z'_2) \arrow[d] \\
Z & & Z' & & Z \arrow[u] & & Z'
\end{tikzcd}$$

Adapting Data Processing Inequality (Beaudry & Renner, 2012) to the above right graph, we have:

$$I((X_1, X_2); Z) \geq I(Z, Z') \geq \text{InfoNCE}(Z, Z'), \tag{13}$$

where $I(X, Y)$ represents the joint mutual information between two variables X and Y. Equipped with the notion of entropy of a random variable, the joint and conditional mutual information quantities can be derived as:

$$I((X_1, X_2), Z) = H(X_1, X_2) - H(X_1, X_2|Z), \tag{14}$$

$$I((X1, X2)|Z) = H(X_1|Z) + H(X_2|Z) - H(X_1, X_2|Z), \tag{15}$$

Moreover, the random construction of the pair of X_1 and X_2 makes them independent with each other. Thus,

$$H(X_1, X_2) = H(X_1) + H(X_2) \tag{16}$$

Combining this with Eq. (14) and (15) yields:

$$\begin{aligned}
I((X_1, X_2); Z) &= H(X_1) + H(X_2) - H(X_1, X_2|Z) \\
&= H(X_1) + H(X_2) - H(X_1|Z) - H(X_2|Z) + I(X_1, X_2|Z) \\
&= I(X_1, Z) + I(X_2, Z) + I(X_1, X_2|Z),
\end{aligned}$$

which completes the proof. □

B Significance Tests

Table 6: Significance test results measured on the performance of NTM+CL (Nguyen & Luu, 2021) and our model.

Metric	20NG		IMDb		Wiki	
	$C = 50$	$C = 200$	$C = 50$	$C = 200$	$C = 50$	$C = 200$
p-value	0.0230	0.0422	0.0408	0.0009	0.0406	0.0234

We conduct significance tests and report the p-values in Table 6. All of the p-values are smaller than 0.05, showing our multi-task set-based contrastive framework significantly outperforms the standard contrastive learning method.

Table 7: Topic Coherence results with different pooling functions.

Method	IMDb	
	$C = 50$	$C = 200$
MinPooling	0.190±0.005	0.186±0.004
MeanPooling	0.194±0.006	0.195±0.010
SumPooling	0.191±0.007	0.188±0.005
GAP (Li et al., 2016)	0.197±0.003	0.196±0.005
FSP (Zhang et al., 2018)	0.199±0.004	0.195±0.007
PSWE (Naderializadeh et al., 2021)	0.194 ±0.006	0.191±0.003
MaxPooling	**0.200**±**0.007**	**0.197**±**0.003**

C Choice of Pooling Functions.

We experiment with different pooling operations and compare their performance with our choice of MaxPooling function on the IMDb dataset. In detail, we adapt MeanPooling, SumPooling, Global Attention Pooling (GAP) (Li et al., 2016), Featurewise Sort Pooling (FSP) (Zhang et al., 2018), and Pooling by Sliced-Wasserstein Embedding (PSWE) Naderializadeh et al. (2021). We report the outcomes of our multi-task set-based contrastive framework in Table 7. As it can be observed, MaxPooling acquires the highest topic coherence quality. We hypothesize that different from other aggregation functions, MaxPooling directly retrieves strong features, thus being able to extract important information and suppress the noise. In addition, we realize that our proposed framework consistently outperforms the individual-based contrastive neural topic model, demonstrating the robustness of set-based contrastive learning with a variety of aggregation functions.

D Topic-by-Topic Results.

We perform individual comparison upon each of our topics with the aligned one generated by the baseline contrastive neural topic model. Inspired by (Hoyle et al., 2020), we construct a bipartite graph connecting our topics with the baseline ones. We adopt competitive linking to greedily estimate the optimal weight for the matching in our bipartite graph matching. The weight of every connection is the Jensen-Shannon (JS) divergence between two topics. Each iteration will retrieve two topics whose JS score is the lowest and proceed to remove them from the topic list. The procedure is repeated until the lowest JS score surpasses a definite threshold. Figure 4 (left) depicts the aligned score for three benchmark datasets. Inspecting visually, we extract 44 most aligned topic pairs for the comparison procedure. As shown in Figure 4 (right), we observe that our approach possesses more high-NPMI topics then the NTM+CL baseline. This demonstrates the set-based contrastive method not only improves general topic quality but also manufactures better individual topics.

E Hyperparameter Settings

In Table 8, we denote hyperparameter details of our neural topic models, i.e. learning rate η, batch size B, and the temperature τ for the InfoNCE loss. For training execution, the hyperparameters vary with respect to the dataset.

F Case Studies

Table 9 shows randomly selected examples of the discovered topics by the individual-based and our set-based contrastive model. In general, our model produces topics which are more coherent and less repetitive. For example, in the 20NG dataset, our inferred topic focuses on "*computer screening*" topic with closely correlated terms such as "*monitor*", "*port*", and "*vga*", whereas topic of NTM+CL consists of "*apple*" and "*orange*", which are related to "*fruits*". For IMDb, our topic mainly involves words concerning "*martial art movies*". However, NTM+CL mix those words with "*zombies*", "*flight*", and "*helicopter*". In the same vein, there exists irrelevant words within the

Figure 4: (left) Jensen-Shannon for aligned topic pairs of NTM+CL (Nguyen & Luu, 2021) and Our Model. (right) The number of aligned topic pairs which Our Model improves upon NTM+CL(Nguyen & Luu, 2021).

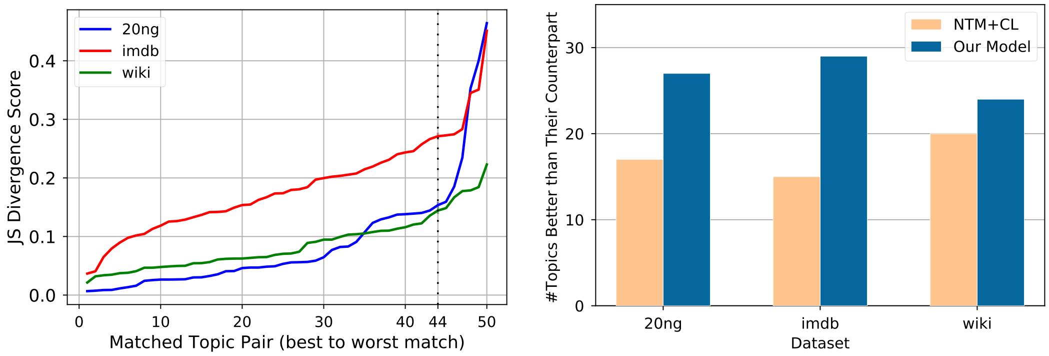


Table 8: Hyperparameter Settings for Neural Topic Model Training.

Hyperparameter	20NG		IMDb		Wiki	
	$C = 50$	$C = 200$	$C = 50$	$C = 200$	$C = 50$	$C = 200$
sample set size K	4	4	3	3	4	4
permutation matrix size P	8	8	8	8	8	8
temperature τ	0.2	0.2	0.2	0.2	0.2	0.2
learning rate η	0.002	0.002	0.002	0.002	0.001	0.002
batch size B	200	200	200	200	500	500

Table 9: Examples of the topics produced by NTM+CL (Nguyen & Luu, 2021) and Our Model.

Dataset	Method	NPMI	Topic
20NG	NTM+CL	0.2766	mouse monitor orange gateway video apple screen card port vga
	Our Model	0.3537	vga monitor monitors colors video screen card mhz cards color
IMDb	NTM+CL	0.1901	seagal ninja martial arts zombie zombies jet fighter flight helicopter
	Our Model	0.3143	martial arts seagal jackie chan kung hong ninja stunts kong
Wiki	NTM+CL	0.1070	architectural castle architect buildings grade historic coaster roller sculpture tower
	Our Model	0.2513	century building built church house site castle buildings historic listed

baseline topic in the Wiki dataset, for example "*grade*", "*roller*", and "*coaster*", while our topic entirely concentrates on "*architecture*".

G Analysis of Sampling Approaches

As indicated by $\mathcal{L}_{\text{nce}}$ in Algorithm 1, for each set i, we consider all sets $j \neq i, 1 \leq j \leq n$ as negative sets. Therefore, we do not specify any negative sampling procedure, which makes our framework simpler than (Nguyen & Luu, 2021).

As such, to examine whether the enhanced topic quality is derived from such difference in document sampling strategies, we proceed to compare our Embedding-based Sampling with Word-based Sampling (WS) (Nguyen & Luu, 2021) to draw positive and negative documents, which we utilize to construct document sets. Conducted experiments in Table 10 do not observe any significant difference with respect to sampling strategies, in terms of both topic coherence and topic diversity.

Table 10: Topic Coherence and Topic Diversity results on 20NG and IMDb datasets with different sampling strategies.

Method	20NG				IMDb			
	$C = 50$		$C = 200$		$C = 50$		$C = 200$	
	NPMI	TD	NPMI	TD	NPMI	TD	NPMI	TD
NTM+CL	0.332±0.006	0.775±0.008	0.277±0.003	0.799±0.009	0.191±0.004	0.857±0.010	0.186±0.002	0.856±0.012
Our Model (WS)	0.339±0.005	0.909±0.020	0.289±0.004	0.902±0.004	0.198±0.009	0.914±0.006	0.195±0.001	0.888±0.012
Our Model (ES)	0.340±0.005	0.913±0.019	0.291±0.003	0.905±0.004	0.200±0.007	0.916±0.008	0.197±0.003	0.892±0.007

H Further Study of Objective Control Strategies

We extend the comparison of multi-objective optimization (MOO) and other control strategies on 20NG and IMDb datasets in Table 11. In general, we find that our adaptation of gradient-based MOO balances the contrastive and ELBO losses more proficiently to obtain better topic quality.

Table 11: Topic Coherence and Topic Diversity results on 20NG and IMDb datasets with different objective control strategies.

Method	20NG				IMDb			
	$C = 50$		$C = 200$		$C = 50$		$C = 200$	
	NPMI	TD	NPMI	TD	NPMI	TD	NPMI	TD
UW	0.328±0.006	0.837±0.011	0.273±0.003	0.831±0.004	0.187±0.004	0.848±0.007	0.175±0.009	0.818±0.007
GradNorm	0.338±0.006	0.870±0.003	0.276±0.005	0.839±0.006	0.189±0.005	0.873±0.013	0.187±0.008	0.841±0.008
PCGrad	0.327±0.008	0.857±0.008	0.272±0.001	0.828±0.003	0.189±0.003	0.872±0.003	0.186±0.010	0.830±0.006
RW	0.329±0.008	0.864±0.009	0.274±0.003	0.833±0.003	0.186±0.003	0.862±0.005	0.181±0.004	0.823±0.009
Linear - $\alpha = 0.25$	0.327±0.006	0.862±0.007	0.275±0.001	0.835±0.005	0.184±0.003	0.853±0.019	0.182±0.002	0.825±0.007
Linear - $\alpha = 0.5$	0.322±0.009	0.850±0.005	0.272±0.001	0.824±0.004	0.187±0.005	0.859±0.009	0.184±0.006	0.829±0.008
Our Model	**0.340±0.005**	**0.913±0.019**	**0.291±0.003**	**0.905±0.004**	**0.200±0.007**	**0.916±0.008**	**0.197±0.003**	**0.892±0.007**

I Additional Evaluation of Topic Quality

We investigate a larger scope for our proposed set-based contrastive topic model, where we train prior baselines and our model on Yahoo Answer (Zhang et al., 2015) dataset, which consists of about 1.4 million training and testing documents. The illustrated results in Table 12 demonstrate that set-based contrastive topic model does polish the produced topics better than other topic models, proving both of its effectiveness and efficacy.

Table 12: Topic Coherence and Topic Diversity results on Yahoo Answers dataset.

Method	Yahoo Answers			
	$C = 50$		$C = 200$	
	NPMI	TD	NPMI	TD
NTM	0.290±0.003	0.897±0.012	0.161±0.001	0.718±0.004
ETM	0.207±0.001	0.808±0.017	0.158±0.002	0.709±0.007
DVAE	0.319±0.001	0.932±0.014	0.163±0.003	0.721±0.004
BATM	0.247±0.004	0.883±0.019	0.177±0.001	0.767±0.003
W-LDA	0.222±0.004	0.821±0.008	0.173±0.005	0.730±0.001
SCHOLAR	0.327±0.005	0.936±0.019	0.191±0.003	0.784±0.004
SCHOLAR+BAT	0.328±0.005	0.940±0.020	0.196±0.001	0.793±0.005
NTM+CL	0.334±0.002	0.954±0.004	0.198±0.001	0.794±0.003
Our Model	**0.348±0.004**	**0.988±0.003**	**0.208±0.002**	**0.882±0.001**