MCVAE: Margin-based Conditional Variational Autoencoder for Relation Classification and Pattern Generation*

Fenglong Ma[1], Yaliang Li[2], Chenwei Zhang[3], Jing Gao[1], Nan Du[4], Wei Fan[4]
[1]University at Buffalo, Buffalo, NY; [2]Alibaba Group, Bellevue, WA
[3]University of Illinois at Chicago, Chicago, IL; [4]Tencent Medical AI Lab, Palo Alto, CA
[1]{fenglong, jing}@buffalo.edu; [2]yaliang.li@alibaba-inc.com; [3]czhang99@uic.edu; [4]{ndu, davidwfan}@tencent.com

ABSTRACT

Relation classification is a basic yet important task in natural language processing. Existing relation classification approaches mainly rely on distant supervision, which assumes that a bag of sentences mentioning a pair of entities and extracted from a given corpus should express the same relation type of this entity pair. The training of these models needs a lot of high-quality bag-level data. However, in some specific domains, such as medical domain, it is difficult to obtain sufficient and high-quality sentences in a text corpus that mention two entities with a certain medical relation between them. In such a case, it is hard for existing discriminative models to capture the representative features (i.e., common patterns) from diversely expressed entity pairs with a given relation. Thus, the classification performance cannot be guaranteed when limited features are obtained from the corpus. To address this challenge, in this paper, we propose to employ a generative model, called conditional variational autoencoder (CVAE), to handle the pattern sparsity. We define that each relation has an individually learned latent distribution from all possible sentences expressing this relation. As these distributions are learned based on the purpose of input reconstruction, the model's classification ability may not be strong enough and should be improved. By distinguishing the differences among different relation distributions, a margin-based regularizer is designed, which leads to a margin-based CVAE (MCVAE) that can significantly enhance the classification ability. Besides, MCVAE can automatically generate semantically meaningful patterns that describe the given relations. Experiments on two real-world datasets validate the effectiveness of the proposed MCVAE on the tasks of relation classification and relation-specific pattern generation.

ACM Reference Format:
Fenglong Ma[1], Yaliang Li[2], Chenwei Zhang[3], Jing Gao[1], Nan Du[4], Wei Fan[4]. 2019. MCVAE: Margin-based Conditional Variational Autoencoder for Relation Classification and Pattern Generation. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308558.3313436

1 INTRODUCTION

Relation classification is a core task in natural language processing (NLP), which aims to identify semantic relations between a pair of entities in the given sentence. Relation classification can be widely used for multiple NLP tasks, such as knowledge graph completion and question answering. Traditional relation classification models [13, 27, 39] usually require high-quality annotated data, which makes them impossible to be applied to large-scale datasets.

Recently, distant supervision approach [26] has been proposed, which assumes that if a pair of entities has a relation in a knowledge base, then all the sentences that contain the two entities express this relation. This approach makes it easy to collect multiple sentences for different entity pairs and their relations, but it is hard to control the quality of the collected data. To remove the effect of "noisy" sentences, many approaches are proposed [8, 14, 15, 23, 31, 32, 37, 40]. All these studies are discriminative models, which aim to extract the accurate feature representations of relations to improve the prediction performance. However, if the quality of the input data is not very high, then they may not correctly capture the characteristics of the target, which will lead to unsatisfactory performance.

The quality of the collect data mainly depends on both the diversity of relation expressions and the quantity of sentences mentioning each entity pair. For a relation, it may be related to multiple entity pairs in the knowledge base. The extracted sentences mentioning those pairs may use different expressions to describe the semantic meanings of the relation. *If only a small part of commonly used expressions appear in the training data, it is hard for existing discriminative models to recognize the other possible expressions in the testing procedure.* This issue stems from the drawback of discriminative models, i.e., they are not designed to facilitate the **generalization ability**.

On the other hand, discriminative models focus on capturing those expressions with high occurrences among different sentences, i.e., recognizing relation patterns. In common domains, there may exist enough sentences with mentioned entity pairs on the same relation that can be used for the discovery of clear patterns. However, in specific domains, such as medical domain, given a pair of medical entities extracted from a medical knowledge base, to recognize the relation between the two entities, we need to extract sentences from a given corpus. Unfortunately, the text corpus used for medical sentence extraction is usually not big. Though we can extract sentences for multiple entity pairs, the number of clean patterns among these sentences is limited, which leads to the *sparsity* of relation patterns and further decreases the quality of the dataset. Therefore, how to design a model that can handle the **diversity of relation expressions and sparsity of relation patterns** in the training dataset is a challenge for relation classification.

*This work was done when the first and third authors were interns and the second author worked at Tencent Medical AI Lab.

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313436

To tackle this challenge, instead of using discriminative models to capture the representative features of relations, we design a *generative* model, named MCVAE, which employs a Margin-based Conditional Variational AutoEncoder. The benefit of applying a generative model to solve a classification problem is that the generalization ability of generative models is significantly better than that of discriminative ones. Furthermore, such an ability helps generative models to handle the issue of pattern sparsity. Particularly, the proposed MCVAE consists of four modules: An **encoder** is used to learn the embeddings of input sentences. By concatenating the embeddings with the given relation, the encoder outputs the mean vector and the diagonal vector of covariance matrix for constructing the latent space of the given relation. A **decoder** aims to reconstruct the input sentence using a sample drawn from the constructed latent space. Motived by [22], a margin-based **regularizer** is designed to model the differences among different constructed latent spaces of relations and further increase the ability of classification. Finally, a **generator** is adopted to generate sentences and patterns for the given relation type based on the trained model. Experimental results on two datasets show that the proposed MCVAE outperforms existing relation classification models and produces meaningful patterns which describe the given relations. Here, we highlight our contributions as follows:

- To the best of our knowledge, this is the first work to design a generative model for solving the problems of expression diversity and pattern sparsity in the relation classification task.
- The proposed model not only accurately identifies the relation types of the input sentences with the designed margin-based regularizer, but also generates meaningful sentences and patterns for the given relations using the conditional variational autoencoder.
- Experiments on two real world datasets demonstrate the effectiveness of the proposed MCVAE for both relation classification and relation-specific sentence generation.

2 RELATED WORK

In this section, we review existing work from two aspects: relation classification and deep generative models.

Relation Extraction or Classification [1, 3, 10, 20, 21, 28, 29, 33, 34, 42] is an important sub-task of Information Extraction (IE). IE can be done in unsupervised [2, 7, 9] or semi-supervised domain [4, 36], and even in the form of OpenIE [16], where facts are extracted from the data along with the relation phrases, i.e., without predefined ontology or relation classes. Supervised relation classification aims to identify the relation between two entities with the given text. Many models are developed to solve this problem, such as deep learning based models with distant supervision technique [8, 14, 15, 23, 31, 32, 37, 40]. Distant supervision is proposed in [26] to collect large-scale datasets for training, but inevitably there is noise in the collected data. To reduce the effect of the noise in the data, many studies cast the problem of relation classification as a multiple instance learning problem [14, 32, 40]. Recently, sentence-level attention mechanism over multiple instances are proposed to assign lower weights to those noisy sentences [15, 23, 24]. Moreover, reinforcement learning-based sentence selection approaches [8, 31] are introduced to further improve the performance.

Deep generative models, such as generative adversarial networks (GAN) [11], have attracted much attention recently. Different from GAN, which generates data based on the arbitrary noise, variational autoencoder (VAE) tries to model the underlying probability distribution of data by constructing the latent variables, which makes it possible to generate new samples from the learned distributions. Actually, there are many VAE-based models [18, 38] that can be used to generate different kinds of data, such as images [12, 30], natural languages [5, 25] and structured medical entity pairs [41].

The aforementioned models either only focus on classification or generation. Different from them, the proposed model can handle both classification and generation simultaneously. The generative ability is obtained by the designed conditional variational autoencoder, while the ability of classification is achieved by the designed margin-based regularizer.

3 THE PROPOSED MODEL: MCVAE

In this paper, we introduce a novel margin-based conditional variational autoencoder (MCVAE) for distant supervised relation classification task. It is a fact that the relation between a pair of entities can be expressed by many ways (i.e., expression diversity). Thus, using a continuous vector to model a relation with multiple expressions is not enough. Moreover, the number of sentences used for relation classification may be small, which results in the problem of pattern sparsity. To solve these challenges, we propose to use a *generative model* for modeling the diverse expressions of relations. The benefits of the proposed MCVAE model are three-fold: (1) The relations are embedded into distributions, instead of vectors, which directly increases the ability of generalization; (2) given a relation, the proposed model is able to generate sentences on this relation, which may contain the frequently used patterns for this relation; and (3) the model can guarantee the prediction performance by modeling the differences among relation distributions. Next, we will introduce the details of the proposed MCVAE architecture.

3.1 Model Overview

Figure 1 shows the architecture of the proposed MCVAE. It consists of four modules: *encoder*, *decoder*, *regularizer*, and *generator*. The **encoder** module takes a relation r and a sentence $\mathbf{x} = \{w_i\}$ as input data, where $\{w_i\}$ represents a set of ordered words. For each word w_i, an embedding layer is employed to map the discrete word to a continuous vector representation. An RNN is then used to encode the whole sentence (i.e., a set of word embeddings) to a hidden vector. Combining the hidden vector with the input relation r, we can construct a latent space represented by the means and the diagonal vectors of covariances. The latent variable $\mathbf{z}$ can be drawn from the learned latent space, which is one of the inputs of the decoder, and the other input is the label of the relation r.

The **decoder** tries its best to reconstruct the original input $\mathbf{x}$ based on another RNN. Through training the reconstruction loss and the KL-divergence as described in CVAE, we can learn all the parameters. However, a drawback of the naive model is that the constructed latent spaces for different relations are indistinguishable. In other words, the model does not have the ability for classifying the

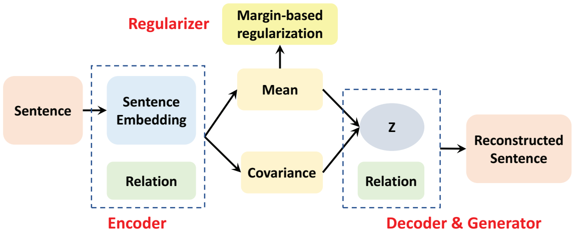

Figure 1: The Proposed MCVAE Model.

input data. To gain such an ability, we design a novel margin-based **regularizer**. The intuition behind the regularizer is that we want to increase the margins among latent spaces of different relations.

By training the proposed model MCVAE, we can generate sentences based on the given the relation types with the generator. The **generator** has the same structure with the decoder. However, instead of reconstructing the original inputs, it directly generates samples from the learned latent spaces of relations. The details of different modules are introduced in the following sections.

3.2 Encoder

Traditional relation classification models embeds a "bag" of sentences, which contains the same entity pair, into a common vector representation for the given entity pair. Different from existing models, we consider the sentence-level embeddings. We believe that each extracted sentence with distant supervision more or less expresses the relation between two entities.

For each word $w_i \in \mathbf{x}$, a word embedding layer is used to map it to a continuous vector $\mathbf{v}_i = f_{\text{EMB}}(w_i) \in \mathbb{R}^d$, where d is the size of the word embeddings, and $\mathbf{v}_i$ is the input of the encoder RNN. In particular, we employ the Gated Recurrent Unit (GRU) [6] as the encoder network in the proposed model. The output of the encoder is a hidden vector $\mathbf{h} \in \mathbb{R}^g$ produced by the GRU, which is defined as follows:

$$\mathbf{h} = f_{\text{ENC}}(\mathbf{x}) = \text{GRU}_{\text{ENC}}(\mathbf{V}; \varphi), \tag{1}$$

where g is the dimensionality, $\mathbf{V} \in \mathbb{R}^{s \times d}$ is the matrix of all the word embeddings, s is the length of the input $\mathbf{x}$, and φ is the parameter set of the GRU.

Based on the latent representation of the input $\mathbf{x}$ and its relation r, we can construct the latent space $Q_\phi(\mathbf{z}|\mathbf{x}, r)$ of the given relation r, which is represented by the mean μ and σ, and $\mathbf{z}$ is a sample drawn from the latent space. Towards this goal, we first concatenate the hidden vector $\mathbf{h}$ and the one-hot vector $\mathbf{r} \in \mathbb{R}^n$, and then use two separate linear functions to learn μ and σ as follows:

$$\mu = [\mathbf{h}; \mathbf{r}] \cdot \mathbf{W}_\mu + \mathbf{b}_\mu, \tag{2}$$

$$\sigma = [\mathbf{h}; \mathbf{r}] \cdot \mathbf{W}_\sigma + \mathbf{b}_\sigma, \tag{3}$$

where n denotes the number of relations, $\mathbf{W}_\mu$, $\mathbf{W}_\sigma \in \mathbb{R}^{(g+n) \times m}$ and $\mathbf{b}_\mu$, $\mathbf{b}_\sigma \in \mathbb{R}^m$ are the parameters to be learned, and m is the number of dimensions. With the learned μ and σ, we can obtained the latent space of the given relation r. Next, we will introduce how to reconstruct the input sentence x with the learned μ and σ.

3.3 Decoder

Given μ and σ, we can directly sample a latent variable $\mathbf{z} \in \mathbb{R}^m$ from the constructed relation space $\mathcal{N}(\mu, \sigma)$. However, the direct sampling dose not make the whole model differentiable, which leads to the failure of existing optimization approaches to compute the gradients. To address this problem, we follow the method in [19] and apply the reparameterization trick. It works as follows: Instead of directly sampling from $\mathcal{N}(\mu, \sigma)$, we first sample from a standard normal distribution $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and then obtain a reparameteried $\mathbf{z} = \mu + \sigma \odot \epsilon$, where $\odot$ represents the element-wise multiplication. Since sampling from ϵ dose not depend on the network, it makes the proposed model differentiable again.

The decoder aims to reconstruct the input x with a recurrent neural network, which means that the input and output must be sequences. To decode the input sentence x, we need to add two special symbols: $\langle sos \rangle$ (the start of the sequence) and $\langle eos \rangle$ (the end of the sequence). The input of the decoder is $\langle sos \rangle + \mathbf{x}$, and the target is $\mathbf{x} + \langle eos \rangle$.

Since reconstructing relation-specific sentences is the goal of the designed decoder, the relation information should be considered in the decoding procedure, and the decoder can be denoted as $P_\theta(\mathbf{x} + \langle eos \rangle | \mathbf{z}, r)$. To this end, we first concatenate the latent variable $\mathbf{z}$ and the one-hot relation vector $\mathbf{r}$, and then convert it to a new vector $\mathbf{h}' \in \mathbb{R}^g$ with a linear function. The mapped vector $\mathbf{h}'$ is considered as the initial hidden state of the decoder RNN. Similar to the encoder, we need to embed the input words into continuous vector representations. Here, we use dropout technique [35] to remove the irrelevant information on the word embeddings of the decoder input, and then these embeddings $\mathbf{V}' \in \mathbb{R}^{(s+1) \times d}$ are treated as the inputs of the decoder RNN. Mathematically, we use the following formulation to represent the decoder RNN:

$$\mathbf{C} = f_{\text{DEC}}(\langle sos \rangle + \mathbf{x}) = \text{GRU}_{\text{DEC}}(\mathbf{V}', \mathbf{h}'; \varphi'), \tag{4}$$

where $\mathbf{C} \in \mathbb{R}^{(s+1) \times g}$, and φ' is the parameter set of the decoder RNN. To generate discrete words, we need to obtain the probability of each word using a softmax layer after a linear mapping as follows:

$$\mathbf{O} = \text{softmax}(\mathbf{C}\mathbf{W}_o + \mathbf{b}_o), \tag{5}$$

where $\mathbf{O} \in \mathbb{R}^{(s+1) \times |V|}$ is the probability matrix, $\mathbf{W}_o \in \mathbb{R}^{g \times |V|}$ and $\mathbf{b}_o \in \mathbb{R}^{|V|}$ are parameters to be learned, and $|V|$ is the vocabulary size. Cross entropy between the probability matrix $\mathbf{O}$ and the target $\mathbf{x} + \langle eos \rangle$ can be used to optimize the objective. To simplify the notations, we remove the special symbols (i.e., $\langle sos \rangle$ and $\langle eos \rangle$) in the following sections. We use $\mathbf{x}$ to represent the input and output of the decoder.

3.4 Training with Large-Margin Regularizer

Given the encoder and decoder, and following the loss of CVAE, we can define the reconstruction loss to minimize the variational lower bound:

$$\begin{aligned} &\mathcal{L}_{\text{ENC+DEC}}(\mathbf{x}, r; \theta, \phi) = \\ &- \text{KL}(Q_\phi(\mathbf{z}|\mathbf{x}, r) \| P_\theta(\mathbf{z}|r)) + \mathbb{E}[\log P_\theta(\mathbf{x}|\mathbf{z}, r)]. \end{aligned} \tag{6}$$

The first term is in charge of measuring the difference between the prior distribution $P_\theta(\mathbf{z}|r)$ and a sample distribution $Q_\phi(\mathbf{z}|\mathbf{x}, r)$ with KL-divergence loss. Actually, the unknown true distribution can be simplified by a prior distribution $P_\theta(\mathbf{z}|r)$. The second term is the reconstruction loss from the decoder. Let $P_\theta(\mathbf{z}|r) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be the prior distribution, and then the KL-divergence can be rewritten as follows:

$$\begin{aligned} &-\mathrm{KL}(Q_{\phi}(\mathbf{z}|r)\|P_{\theta}(\mathbf{z}|r)) = -\mathrm{KL}(\mathcal{N}(\mu,\sigma)\|\mathcal{N}(\mathbf{0},\mathbf{I})) \\ &= -\frac{1}{2}\sum_{i=1}^{m}(-\log(\sigma_i)+\sigma_i+\mu_i^2-1). \end{aligned} \tag{7}$$

Since the exponential term is more stable than the log term, in the experiments, we use $\log(\sigma)$ to model σ, i.e,

$$-\mathrm{KL}(Q_{\phi}(\mathbf{z}|r)\|P_{\theta}(\mathbf{z}|r)) = -\frac{1}{2}\sum_{i=1}^{m}(-\sigma_i+\exp(\sigma_i)+\mu_i^2-1). \tag{8}$$

As discussed in the decoder, the cross entropy loss can be used, i.e.,

$$\mathbb{E}[\log P_{\theta}(\mathbf{x}|\mathbf{z},r)] = -\sum_{j=1}^{s+1}\left(\mathbf{y}_j^{\top}\log(\mathbf{O}_j)+(\mathbf{1}-\mathbf{y}_j)^{\top}\log(\mathbf{1}-\mathbf{O}_j)\right), \tag{9}$$

where $\mathbf{y}_j$ is the one-hot representation of the j-th word in $\mathbf{x}+\langle eos\rangle$, and $\mathbf{O}_j$ is the estimated probability of the j-th word. Using both Eq. (8) and Eq. (9), we can optimize the model with the gradient-based optimizer, such as Adam [17]. However, this basic model does not have the ability of classification. The reason is that all the relation distributions are close to the prior distribution $\mathcal{N}(\mathbf{0},\mathbf{I})$, which means that different relation spaces are mixed together. Thus, for a given sentence, it is hard for the basic model to assign the correct category label.

To gain the ability of distinguishing different relation types, the proposed model requires to learn the differences among relation spaces. In the proposed MCVAE, we add a large-margin regularizer into the objective of $\mathcal{L}_{\text{ENC+DEC}}$. There are n relation types, and for each sentence, it has a relation type r. In the encoder, we can obtain the mean vector $\mu_r \in \mathbb{R}^m$ of the relation distribution r. To ensure the classification ability of the proposed model, μ_r should be far away from the mean vectors of other relation distributions. Meanwhile, $\mathcal{N}(\mu_r,\sigma_r)$ should be closer to the prior distribution $\mathcal{N}(\mathbf{0},\mathbf{I})$. Following these motivations, we design a margin-based regularizer, which is defined as follows:

$$\mathcal{L}_{\text{REG}} = \max(\alpha + f_{\text{D}}(\mu_r) - \beta, 0), \tag{10}$$

where $\alpha > 0$ is the predefined margin, $f_{\text{D}}()$ is the distance function (i.e., $f_{\text{D}}(\mu_r) = \sum_{i=1}^{m}\mu_{ri}^2$) to calculate the distance between the mean vector μ_r and the mean of the prior vector $\mathbf{0} \in \mathbb{R}^m$. β is defined as follows:

$$\beta = \min(f_{\text{D}}(\mu_1),\cdots,f_{\text{D}}(\mu_{r-1}),f_{\text{D}}(\mu_{r+1}),\cdots,f_{\text{D}}(\mu_n)). \tag{11}$$

From the designed regularizer (Eq. (10)), we can observe that if $f_{\text{D}}(\mu_r) > \beta - \alpha$, then $\mathcal{L}_{\text{REG}} > 0$. In this case, the proposed model should penalize the loss of $\mathcal{L}_{\text{REG}}$, which indicates that the difference value between $f_{\text{D}}(\mu_r)$ and β is at least α. The final loss function of the proposed MCVAE used in the training procedure is formulated as follows:

$$\mathcal{L}_{\text{MCVAE}} = \mathcal{L}_{\text{ENC+DEC}} + \mathcal{L}_{\text{REG}}. \tag{12}$$

In the testing procedure, assume that there is a bag of sentences $\{\mathbf{x}_1,\cdots,\mathbf{x}_t\}$ which mentions a pair of entities. For each sentence x_i, we first use the encoder to obtain the mean vector according to Eq. (2) for each relation r, and then calculate the distance using $f_{\text{D}}(\mu_r)$. Since we have n relations, for each sentence, n distances can be obtained and are denoted as $\mathbf{d}_i = [f_{\text{D}}(\mu_1),\cdots,f_{\text{D}}(\mu_n)]$. As there are t sentences in the bag, we aggregate all the distance vectors $\mathbf{d}_1,\cdots,\mathbf{d}_t$ to obtain an average vector $\mathbf{d}_{avg} \in \mathbb{R}^n$. Finally, we use the index corresponding to the minimum value in $\mathbf{d}_{avg}$ as the predicted label $\hat{r}$.

3.5 Generator

Based on the trained model, we can generate sentences which are related to a specific relation. Here, we use a density-based sampling approach for the generator to sample $\hat{\mathbf{z}}$ directly from the distribution of relations' latent spaces as [41]. The intuition of using density-based sampling approach is that the dense area (i.e., the area around the mean vector) contains "patterns" to express the relation r. A pattern means that a phrase or a set of words frequently appears in the sentences that express the relation r. Therefore, using the patterns in the dense area of the latent space, the generator can generate multiple meaningful sentences. The benefit of applying the density-based sampling method is that we do not need to cast a sentence with its relation into the encoder for learning μ and σ for the generator.

Particularly, given a relation type r, its one-hot vector representation can be obtained, which is denoted as $\mathbf{r} \in \mathbb{R}^n$. The following step is to sample a latent variable $\hat{\mathbf{z}}$ from the latent space $P_{\theta}(\hat{\mathbf{z}}|r)$. If the model is well-trained, then the latent variable $\hat{\mathbf{z}}$ will be drawn from $P_{\theta}(\hat{\mathbf{z}}|r) \sim \mathcal{N}(\mathbf{0},\mathbf{I})$. The concatenation of $\hat{\mathbf{z}}$ and $\mathbf{r}$ and the embedding of $\langle sos\rangle$ are the inputs of the generator RNN, which has the same architecture with the decoder RNN, to generate the first word. The output of the generator RNN and the embedding of the first generated word are used to produce the second word. By repeating this procedure, the generator will generate the whole sentence.

4 EXPERIMENTS

In this section, we conduct experiments to show the effectiveness of the proposed model MCVAE.

4.1 Experimental Setting

In this subsection, we introduce the datasets, baselines, parameter settings and evaluation measures.

• **Datasets**. In our experiments, we adopt two datasets. The first one is a widely used public available dataset, called NYT[1], generated by [32]. For this dataset, we remove the sentences in the NA relation. For each entity pair, at most 500 sentences are kept. Finally, there are 136,844 sentences and 19,539 entity pairs (bags) in the training dataset. In the testing dataset, we have 6,437 sentences and 1,755 entity pairs. The total number of relation types is 54, which are extracted from Freebase.

The other dataset, named MedBook, is generated from medical books with Tencent medical knowledge graph. In the training dataset, there are 8,509 sentences and 2,213 entity pairs. The testing dataset contains 592 entity pairs with 740 sentences, and there are 14 relations in the MedBook dataset.

• **Baselines**. Since the proposed MCVAE is a generative model, and there is no existing work for relation classification task using generative models. Thus, we use the variant of the proposed model as a baseline method, which is CVAE. CVAE removes the proposed large-margin regularizer from the proposed MCVAE. The proposed model can be used for not only sentence generation but also relation

[1] http://iesl.cs.umass.edu/riedel/ecml/

classification. Therefore, we compare MCVAE with discriminative models for relation classification: PRNN+AVG and PRNN+ATT. PRNN+ATT uses attention mechanism over sentences in a bag as the way in [23]. The only difference is that we use RNN instead of CNN. The reason is that the proposed model is based on RNN, and designing a CNN-based generative model is our future work. PRNN+AVG learns the embeddings of entity pairs by averaging representations of all the sentences in the same bag.

• **Parameter Settings**. For the NYT dataset, we use a pretrained word embeddings[2]. The parameters are set as follows: For each sentence, the max length s is 70; the size of word embeddings d is 50; the hidden size of RNN g is 256; the size of the latent spaces of relations m is 16; and the margin $\alpha = 5$. The learning rate for Adam is 0.001, and the dropout rate for the decoder inputs is 0.5. The batch size is 256 on the NYT dataset and 32 on the MedBook dataset. The number of epochs is set as 50 for both datasets.

• **Evaluation Measures**. To fairly evaluate the proposed MCVAE, we utilize two measures: accuracy and weighted F1 score. The number of entity pairs for different relations in the testing dataset is significantly different, so the weighted F1 score is introduced, which considers the number of instances when calculating the score. The higher the measures, the better the performance.

4.2 Performance of Relation Classification

Table 1 shows the results for relation classification task on the two datasets. We can observe that the overall performance of the proposed MCVAE is better than that of the baselines.

Table 1: Performance of Different Methods.

Method	NYT		MedBook	
	Accuracy	Weighted F1	Accuracy	Weighted F1
PRNN+AVG	0.4650	0.2951	0.2010	0.0775
PRNN+ATT	0.5031	**0.5019**	0.1537	0.0975
CVAE	0.0501	0.0537	0.2010	0.2180
MCVAE	**0.5288**	0.4875	**0.3108**	**0.2821**

On the NYT dataset, the variant of the proposed MCVAE, i.e., CVAE, has the worst performance. Both the accuracy and weighted F1 value are largely lower than those of other approaches. The reasons are two-fold: (1) When CVAE is well trained, all the latent spaces of different relations are close to the prior distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. It means that CVAE only focuses on how to generate new sentences based on the given relations, instead of paying attention to classification. Thus, the latent spaces of different relations may be mixed together. In such a case, using the learned μ's for classifying different relations is impossible. (2) There are a lot of shared common words to describe multiple relations on this dataset. For example, the relations, */people/person/nationality*, */people/person/place_of_birth* and */people/person/place_lived*, may share many context words, which leads to the failure of CVAE on the relation classification task. However, by adding the proposed large-margin regularizer with CVAE, the proposed MCVAE can significantly improve the performance. This demonstrates that when using generative models to do the classification tasks, it is essential for them to consider the differences among different classes.

Compared with the discriminative models, the performance of the proposed MCVAE is better than that of PRNN+AVG, and is comparable with PRNN+ATT. Since the number of sentences for each entity pair is large on the NYT dataset, the attention mechanism can effectively assign greater weights to those sentences semantically expressing the given relations. Moreover, the classifier can correctly capture the characteristics of different relations. Thus, PRNN+ATT can achieve better performance compared with PRNN+AVG. Different from the learning procedure of discriminative models, the proposed MCVAE tries to enhance the ability of generalization. It means that MCVAE aims to use latent distributions to model different expressions of relations. MCVAE intends to put all the words or patterns describing the same relation together, but designs a margin-based regularizer to distinguish the differences among different distributions.

On the MedBook dataset, since the number of sentences in each bag (i.e., entity pair) is small, which leads to the bad performance of using attention mechanism. Thus, PRNN+AVG performs better than PRNN+ATT. However, both the proposed MCVAE and its variant CVAE achieve better performance than PRNN+AVG and PRNN+ATT. It is because on this dataset, for the same relation, there are many different expressions, and the occurrence of those expressions is much lower than that on the NYT dataset. Without the generalization ability being facilitated, discriminative models are less likely to capture features that are generalizable for classification during the test phase. However, for the generative models, they can collect all the related features to construct the relation distributions. Since the dataset is not big, even for CVAE, the learned relation distributions are still distinguishable. Therefore, the generative models outperform existing discriminative models. It shows the advantage of employing the designed margin-based regularizer for relation classification task.

4.3 Case Study

The benefit of the proposed MCVAE is that it not only guarantees the performance for relation classification, but also generates sentences that express the semantics of the given relations. To validate this claim, we conduct a case study on the NYT dataset. Given a relation type, the proposed MCVAE can produce sentences with the designed generator. In this case study, we show the sentences generated from the following four relations: */people/person/nationality*, */people/person/place_lived*, */business/company/founders*, and */location/us_state/capital*. The generated sentences are shown in Table 2.

From Table 2, we can observe that the proposed MCVAE indeed generates some meaningful patterns on the given relations. Although there are some noisy cases, such as the third sentence in relation */people/person/nationality* and the first sentence in relation */location/us_state/capital*, most generated sentences are reasonable. For the relation */people/person/place_lived*, there is a clear pattern extracted by the proposed MCVAE, i.e., "the former [title] of [location name]". This observation can also be found in the relation */business/company/founders*. When expressing the relation between company and founders, the proposed MCVAE can recognize the key phases: "the founder of [company name]", "chairman", "[company name] chief executive", and so on. Even for the noisy cases, we can observe the ability of generalization of the proposed MCVAE. For

[2] https://github.com/thunlp/NRE

Table 2: Case Study on the NYT Dataset.

Relation	Generated Sentences
/people/person/nationality	(1) he was born in <unk> , germany , and raised in 1958 , england , and received a bachelor (2) president hamid_karzai of afghanistan said he believed that he had been <unk> by the taliban and passport (3) president bush <unk> prime minister , ariel_sharon , said that he had spoken that the united states would not be
/people/person/place_lived	(1) <unk> , the former senator from north_carolina , said he was <unk> by the (2) a former senator from nebraska , said he was <unk> by the senate speaker (3) the former governor of massachusetts , who was elected to represent the senate seat <unk>
/business/company/founders	(1) but <unk> , the founder of microsoft , has been (2) the news_corporation <unk> chairman , rupert_murdoch , objected to the news_corporation (3) but larry_page , google <unk> chief executive , said that
/location/us_state/capital	(1) showdown in ethiopia <unk> capital of addis_ababa , ethiopia (2) new_york wants to endorse the state of albany , the state of new_york , which has been broadcast in albany (3) <unk> <unk> , born in columbus , ohio , on august 10 , 2005 .

example, the first sentence in relation */location/us_state/capital* is obviously wrong. Since the target relation is about the states in USA and their capitals, the country is not “ethiopia”. Though the relation of the generated sentence is different from the target one, the relation between “addis_ababa” and “ethiopia” is indeed related to capitals.

4.4 Relation Pattern Generation

It is obvious that the generated sentences shown in Table 2 contain noisy information, and it may be hard for us to discover clear patterns that describe the given relations from the noisy generated sentences. To further enhance the understandability and readability of the discovered patterns, we conduct another experiment on the NYT dataset. In this experiment, for each entity, we first extract its types from Freebase and manually assign a type to it according to the semantic meaning of the relation. Next, we replace the entities in the sentences with their types. Finally, we run the proposed MCVAE on the new typed NYT dataset with the same parameter settings.

Table 3: Generated Patterns on the Typed NYT Dataset.

Relation	Generated Patterns
/people/person/nationality	(1) the deputy president of $LOCATION , $PEOPLE (2) president $PEOPLE of $LOCATION (3) the former prime minister of $LOCATION , $PEOPLE
/people/person/place_lived	(1) senator $PEOPLE of $LOCATION (2) representative $PEOPLE of $LOCATION (3) $PEOPLE , the $LOCATION borough president
/business/company/founders	(1) $PEOPLE , the founder of $BUSINESS (2) $PEOPLE , the chief executive of $BUSINESS (3) the company owned by $PEOPLE , $BUSINESS
/location/us_state/capital	(1) located in $LOCATION , $LOCATION (2) who was born in $LOCATION , $LOCATION (3) who lives in $LOCATION , $LOCATION

The accuracy on the typed NYT dataset is **0.7761**, and the weighted F1 value is **0.7358**, which are much greater than those on the original NYT dataset. The reason is that after mapping the entities into their types, the proposed MCVAE can effectively recognize the patterns of different relations, and the margin-based regularizer can successfully distinguish these latent distributions of relations. Similar to the case study conducted on the original NYT dataset, we list some patterns extracted by the proposed MCVAE on the new typed NYT dataset shown in Table 3. From Table 3, we can observe that the generated patterns are highly related to the semantic expressions of all the four listed relations. Based on the results in this experiment, we can safely conclude that the proposed MCVAE is able to generate meaningful and distinguishable patterns for different relation categories.

5 CONCLUSIONS

In this paper, we propose a novel generative model for relation classification task, which can handle the problems of both expression diversity and pattern sparsity. The proposed MCVAE not only guarantees the classification performance, but also generates useful patterns which describe the semantic meanings of relations. MCVAE consists of four modules: encoder, decoder, regularizer and generator. The encoder embeds each input sentence into a vector representation, which is further adopted for constructing the latent space of the given relation. The decoder is in charge of reconstructing the input sentences with a designed decoder RNN. The designed regularizer is the core module of the proposed MCVAE, which equips MCVAE with the ability of distinguishing different relation distributions. The generator is employed to generate sentences based on the given relations. Experiments on two real datasets demonstrate the effectiveness of the proposed MCVAE and the strong ability of generation.

ACKNOWLEDGMENTS

This work is supported in part by the US National Science Foundation under grants IIS-1553411 and IIS-1747614. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Azad Abad, Moin Nabi, and Alessandro Moschitti. 2017. Self-crowdsourcing training for relation extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vol. 2. 518–523.

[2] Alan Akbik, Larysa Visengeriyeva, Johannes Kirschnick, and Alexander Löser. 2013. Effective selectional restrictions for unsupervised relation extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. 1312–1320.

[3] Roberto Basili, Cristina Giannone, Chiara Del Vescovo, Alessandro Moschitti, and Paolo Naggar. 2009. Kernel-Based Learning for Domain-Specific Relation Extraction. In *Proceedings of the Congress of the Italian Association for Artificial Intelligence*. 161–171.

[4] Lidong Bing, William Cohen, Bhuwan Dhingra, and Richard Wang. 2016. Using Graphs of Classifiers to Impose Constraints on Semi-supervised Relation Extraction. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*. 1–6.

[5] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating Sentences from a Continuous Space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. 10–21.

[6] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *arXiv preprint arXiv:1409.1259* (2014).

[7] Hady Elsahar, Elena Demidova, Simon Gottschalk, Christophe Gravier, and Frederique Laforest. 2017. Unsupervised Open Relation Extraction. In *Proceedings of the European Semantic Web Conference*. 12–16.

[8] Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement Learning for Relation Classification from Noisy Data. In *AAAI*.

[9] Kata Gábor, Haïfa Zargayouna, Isabelle Tellier, Davide Buscaldi, and Thierry Charnois. 2016. Unsupervised Relation Extraction in Specialized Corpora Using Sequence Mining. In *Proceedings of the International Symposium on Intelligent Data Analysis*. 237–248.

[10] Cristina Giannone, Roberto Basili, Chiara Del Vescovo, Paolo Naggar, and Alessandro Moschitti. 2009. Kernel-based relation extraction from investigative data. In *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*. 93–100.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the Advances in neural information processing systems*. 2672–2680.

[12] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. 2015. DRAW: A Recurrent Neural Network For Image Generation. In *Proceedings of the International Conference on Machine Learning*. 1462–1471.

[13] Zhou GuoDong, Su Jian, Zhang Jie, and Zhang Min. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 427–434.

[14] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 541–550.

[15] Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2017. Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Descriptions. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. 3060–3066.

[16] Shengbin Jia, Yang Xiang, and Xiaojun Chen. 2018. Supervised Neural Models Revitalize the Open Relation Extraction. *arXiv preprint arXiv:1809.09408* (2018).

[17] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[18] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Proceedings of the Advances in Neural Information Processing Systems*. 3581–3589.

[19] Diederik P Kingma and Max Welling. 2014. Stochastic gradient VB and the variational auto-encoder. In *Proceedings of the Second International Conference on Learning Representations (ICLR)*.

[20] Shantanu Kumar. 2017. A Survey of Deep Learning Methods for Relation Extraction. *arXiv preprint arXiv:1705.03645* (2017).

[21] Tuur Leeuwenberg and Marie-Francine Moens. 2017. Structured learning for temporal relation extraction from clinical records. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. 1150–1158.

[22] Chongxuan Li, Jun Zhu, and Bo Zhang. 2018. Max-margin deep generative models for (semi-) supervised learning. *IEEE transactions on pattern analysis and machine intelligence* 40, 11 (2018), 2762–2775.

[23] Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 2124–2133.

[24] Tianyu Liu, Kexiang Wang, Baobao Chang, and Zhifang Sui. 2017. A Soft-label Method for Noise-tolerant Distantly Supervised Relation Extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1790–1795.

[25] Diego Marcheggiani and Ivan Titov. 2016. Discrete-state variational autoencoders for joint discovery and factorization of relations. *Transactions of the Association for Computational Linguistics* 4 (2016), 231–244.

[26] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*. Association for Computational Linguistics, 1003–1011.

[27] Raymond J Mooney and Razvan C Bunescu. 2006. Subsequence kernels for relation extraction. In *Proceedings of the Advances in neural information processing systems*. 171–178.

[28] Truc-Vien T Nguyen, Alessandro Moschitti, and Giuseppe Riccardi. 2009. Convolution kernels on constituent, dependency and sequential structures for relation extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. 1378–1387.

[29] Barbara Plank and Alessandro Moschitti. 2013. Embedding semantic similarity in tree kernels for domain adaptation of relation extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Vol. 1. 1498–1507.

[30] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. 2016. Variational autoencoder for deep learning of images, labels and captions. In *Proceedings of the Advances in neural information processing systems*. 2352–2360.

[31] Pengda Qin, Weiran Xu, and William Yang Wang. 2018. Robust Distant Supervision Relation Extraction via Deep Reinforcement Learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. ACL, Melbourne, Victoria, Australia.

[32] Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 148–163.

[33] Rudolf Schneider, Cordula Guder, Torsten Kilias, Alexander Löser, Jens Graupmann, and Oleksandr Kozachuk. 2016. Interactive Relation Extraction in Main Memory Database Systems. In *Proceedings of the 26th International Conference on Computational Linguistics*. 103–106.

[34] Meilun Sheng, Lin Qiu, Chenyang Wu, Haofen Wang, and Yong Yu. 2013. Effective Chinese Relation Extraction by Sentence Rolling and Candidate Ranking. In *Proceedings of the China Semantic Web Symposium and Web Science Conference*. 147–160.

[35] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

[36] Ang Sun, Ralph Grishman, and Satoshi Sekine. 2011. Semi-supervised relation extraction with large-scale word clustering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 521–529.

[37] Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. Association for Computational Linguistics, 455–465.

[38] Weidi Xu, Haoze Sun, Chao Deng, and Ying Tan. 2017. Variational Autoencoder for Semi-Supervised Text Classification. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. 3358–3364.

[39] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of machine learning research* 3, Feb (2003), 1083–1106.

[40] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 1753–1762.

[41] Chenwei Zhang, Yaliang Li, Nan Du, Wei Fan, and Philip S Yu. 2018. On the Generative Discovery of Structured Medical Knowledge. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2720–2728.

[42] Hong-Tao Zhang, Min-Lie Huang, and Xiao-Yan Zhu. 2012. A unified active learning framework for biomedical relation extraction. *Journal of Computer Science and Technology* 27, 6 (2012), 1302–1313.