A Span-based Multimodal Variational Autoencoder for Semi-supervised Multimodal Named Entity Recognition

Baohang Zhou[1,2], **Ying Zhang**[1,2,*], **Kehui Song**[1,2], **Wenya Guo**[1,2],
Guoqing Zhao[3], **Hongbin Wang**[3], **Xiaojie Yuan**[1,2]
[1] College of Computer Science, Nankai University, Tianjin, China
[2] Tianjin Key Laboratory of Network and Data Security Technology, Tianjin, China
[3] Mashang Consumer Finanace Co, Ltd
{zhoubaohang,zhangying,songkehui,guowenya}@dbis.nankai.edu.cn
{guoqing.zhao02,hongbin.wang02}@msxf.com, yuanxj@nankai.edu.cn

Abstract

Multimodal named entity recognition (MNER) on social media is a challenging task which aims to extract named entities in free text and incorporate images to classify them into user-defined types. The existing semi-supervised named entity recognition methods focus on the text modal and are utilized to reduce labeling costs in traditional NER. However, the previous methods are not efficient for semi-supervised MNER. Because the MNER task is defined to combine the text information with image one and needs to consider the mismatch between the posted text and image. To fuse the text and image features for MNER effectively under semi-supervised setting, we propose a novel span-based multimodal variational autoencoder (SMVAE) model for semi-supervised MNER. The proposed method exploits modal-specific VAEs to model text and image latent features, and utilizes product-of-experts to acquire multimodal features. In our approach, the implicit relations between labels and multimodal features are modeled by multimodal VAE. Thus, the useful information of unlabeled data can be exploited in our method under semi-supervised setting. Experimental results on two benchmark datasets demonstrate that our approach not only outperforms baselines under supervised setting, but also improves MNER performance with less labeled data than existing semi-supervised methods.

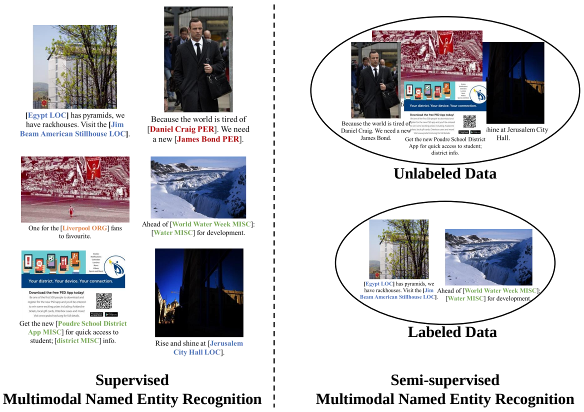
[Egypt LOC] has pyramids, we have rackhouses. Visit the [Jim Beam American Stillhouse LOC].
Because the world is tired of [Daniel Craig PER]. We need a new [James Bond PER].
One for the [Liverpool ORG] fans to favourite.
Ahead of [World Water Week MISC]: [Water MISC] for development.
Get the new [Poudre School District App MISC] for quick access to student; [district MISC] info.
Rise and shine at [Jerusalem City Hall LOC].
Because the world is tired of Daniel Craig. We need a new James Bond.
Get the new Poudre School District App for quick access to student; district info.
hine at Jerusalem City Hall.
Unlabeled Data
[Egypt LOC] has pyramids, we have rackhouses. Visit the [Jim Beam American Stillhouse LOC].
Ahead of [World Water Week MISC]: [Water MISC] for development.
Labeled Data
Supervised Multimodal Named Entity Recognition
Semi-supervised Multimodal Named Entity Recognition

Figure 1: The settings comparison between supervised and semi-supervised multimodal named entity recognition. For the labeled data, the named entities and their types are highlighted in brackets and different colors.

1 Introduction

Multimodal named entity recognition (MNER) has become a fundamental task to extract named entities from unstructured texts and images on social media (Moon et al., 2018). Compared with traditional named entity recognition (NER), MNER on social media poses the unique challenge that bridging the semantic gap between the posted texts and images is critical to extracting named entities. Therefore, the existing MNER models utilized cross-modal attention module to fuse the text and image features (Yu et al., 2020; Zhang et al., 2021). Besides, Xu et al. (2022) proposed cross-modal matching and alignment modules to make the representations of the texts and images more consistent. And to retain the useful image information for MNER, Liu et al. (2022) exploited the two-stage model to refine uncertain labels by fusing the features from the texts and images.

To reduce labeling costs in MNER, semi-supervised learning is widely utilized to exploit the useful information of unlabeled data in text modal. Unlike the supervised setting with adequate labeled data, there are small amount of labeled data and large amount of unlabeled one in semi-supervised setting as shown in Figure 1. Intuitive semi-supervised learning methods including: self-training (ST) (Yarowsky, 1995) and entropy minimization (EM) (Grandvalet and Bengio, 2004) use the pseudo labels generated by NER models for unlabeled data to train models. The NER task can be modeled as the sequence labeling problem, and SeqVAT (Chen et al., 2020) was proposed to combine virtual adversarial training

[*] Corresponding author.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6293 - 6302
December 7-11, 2022 ©2022 Association for Computational Linguistics

(VAT) (Miyato et al., 2019) with conditional random field (CRF) (Lafferty et al., 2001) for semi-supervised sequence labeling. However, the existing semi-supervised NER methods are not efficient for MNER under semi-supervised setting. Because the previous methods are only focused on the text modal and MNER needs considering the semantic correlation between the texts and images of both labeled and unlabeled data.

To overcome the above disadvantages of the existing methods, we propose the **s**pan-based **m**ultimodal **v**ariational **a**uto**e**ncoder (SMVAE)[1] for semi-supervised multimodal named entity recognition. The previous MNER models fused the sentence-level features and image ones for predicting sequence labels and had the difficulty to model mulitmodal features of unlabeled data under semi-supervised setting. Because the semantic correlation between sentences and images should be focused on the specific tokens. Therefore, the proposed method splits the texts into span-level tokens, and combines the span-level features of texts with image features for predicting labels of all spans in each text. SMVAE utilizes modal-specific VAEs to model latent representations of images and span-level texts respectively, and acquires the multimodal features by applying product-of-experts (PoE) (Hinton, 2002) on the latent representations of two modals. The prediction probabilities and multimodal features are exploited to reconstruct the input features for implicitly modeling the correlation between span label and multimodal features. Therefore, the useful information of unlabeled multimodal data can be exploited to improve the performance on MNER. The contributions of this manuscript can be summarized as follows:

1. We analyze that the existing semi-supervised NER methods are not efficient for MNER under semi-supervised setting. To the best of our knowledge, we are the first one to focus on the semi-supervised MNER problem.

2. For semi-supervised MNER, we propose the span-based multimodal variational autoencoder to implicitly model the correlation between span label and multimodal features which takes advantage of unlabeled multimodal data effectively.

3. We compare the proposed model with the semi-supervised methods and state-of-the-art MNER models on two benchmark datasets under semi-supervised setting. The experimental results demonstrate that our model outperforms the baseline approaches.

2 Related Work

2.1 Multimodal Named Entity Recognition

Moon et al. (2018) firstly extended the traditional text-based named entity recognition (NER) to the multimodal named entity recognition (MNER) by taking the images into account. The vital challenge of MNER is to fuse the text features with image features. Moon et al. (2018) proposed to utilize long short term memory networks (LSTM) to extract text features and convolution neural networks (CNN) to extract image features, and combine them with the modality attention module to predict sequence labels. Zhang et al. (2018) proposed an adaptive co-attention network to control the combination of text and image representations dynamically. To extract the image regions that are most related to the text, Lu et al. (2018) utilized the attention-based model to fuse the text and image features. Yu et al. (2020) proposed the uniform multimodal transformer that enhances the interactions of text and image modalities for the MNER task. With the development of multimodal knowledge graph, Chen et al. (2021) exploited the image attributes and semantic knowledge to improve the performance of MNER model. Considering to avoid the influence of mismatch between texts and images, Xu et al. (2022) proposed the cross-modal alignment and matching modules to fuse the text and image representations consistently. Besides, Liu et al. (2022) designed a two-stage model to combine the text features with image ones for refining uncertain labels.

The above studies are under the supervised setting, and we focus on the semi-supervised MNER to reduce the labeling costs. Unlike the supervised learning with adequate labeled data, the semi-supervised learning is focused on utilizing the useful information of unlabeled data.

2.2 Semi-supervised Learning for Named Entity Recognition

For traditional named entity recognition, the labeled data is not always adequate because of the labeling costs. Therefore, semi-supervised learning is an important way to improve NER model

[1] When ready, the code will be published at `https://github.com/ZovanZhou/SMVAE`.

performance without enough labeled data. Two widely used semi-supervised learning methods self-training (ST) (Yarowsky, 1995) and entropy minimization (EM) (Grandvalet and Bengio, 2004) has been proved the effectiveness on NER (Chen et al., 2020). Clark et al. (2018) proposed the cross-view training method to make the predictions consistently when utilizing the partial or full input. Considering to combine virtual adversarial training (VAT) (Miyato et al., 2019) with conditional random field (CRF) (Lafferty et al., 2001) for semi-supervised sequence labeling, Chen et al. (2020) proposed SeqVAT model to improve the robustness and accuracy on NER model.

The existing methods are focused on the text modal, and semi-supervised MNER is proposed to take advantage of unlabeled multimodal data. Therefore, we make efforts on semi-supervised MNER to improve the performance of the model without adequate labeled multimodal data.

3 Model

Before getting into the details of the proposed model, we introduce the notations for semi-supervised MNER. The labeled and unlabeled datasets are denoted as D_l and D_u respectively. The unlabeled dataset D_u with $|D_u|$ samples is formulated as $\{(\mathbf{S}_i^u, \mathbf{V}_i^u)\}_{i=1}^{|D_u|}$. And the labeled dataset D_l with $|D_l|$ samples is defined as $\{(\mathbf{S}_i^l, \mathbf{V}_i^l, \mathbf{y}_i)\}_{i=1}^{|D_l|}$ where $\mathbf{S}_i^l$ and $\mathbf{V}_i^l$ are the text and image of i-th sample, and $\mathbf{y}_i$ is the task defined label for MNER.

According to the conventional MNER studies (Moon et al., 2018), the input text is denoted as $\mathbf{S} = \{w_1, w_2, \ldots, w_{N_s}\}$ and the corresponding label sequence is $\mathbf{y} = \{y_1, y_2, \ldots, y_{N_s}\}$ for MNER. For instance, given a sentence $\mathbf{S} = \{\text{Anyway}, \text{the}, \text{best}, \text{Benz}, \text{in}, \text{the}, \text{world}\}$, the label sequence is annotated as $\mathbf{y} = \{\text{O}, \text{O}, \text{O}, \text{B-PER}, \text{O}, \text{O}, \text{O}\}$ with BIO2 tagging schema (Tjong Kim Sang and Veenstra, 1999). Unlike the existing MNER models that combine the whole sentence features with image features directly, we focus on the fine-grained correlation between the phrases of sentence and the image. Therefore, the span-level representations of each phrase in the sentence are utilized to predict the labels. And the label for the input text $\mathbf{S}$ is reformulated as named entity set $\mathbf{y} = \{y_k\}_{k=1}^{N_e}$ where y_k is a tuple $(l_k, r_k, \bar{y})$ and N_e is the number of named entities. (l_k, r_k) is the span of an entity that corresponds to the phrase $\mathbf{S}_{(l_k, r_k)} = \{w_{l_k}, w_{l_k+1}, \ldots, w_{r_k}\}$ and $\bar{y}$ is the named entity type. For instance, the label for sentence $\mathbf{S} = \{\text{Anyway}, \text{the}, \text{best}, \text{Benz}, \text{in}, \text{the}, \text{world}\}$ is formulated as $\mathbf{y} = \{(4, 4, \text{PER})\}$.

The SMVAE model is shown in Figure 2. For the multimodal data, we use BERT (Devlin et al., 2019) as text encoder to obtain the representations of sentences and ResNet (He et al., 2016) as visual encoder to obtain the regional representations of images. The proposed SMVAE consists of two modal-specific VAEs to acquire the latent representations of the two modality features. And we obtain the multimodal representations to predict the labels by applying product-of-experts (PoE) (Hinton, 2002) on the latent representations of two modalities. The latent representations and the labels are combined to reconstruct the input features in the modal-specific VAE for modeling the correlation between span label and multimodal features implicitly. Therefore, the unlabeled data can be exploited to improve the performance of MNER.

3.1 Multimodal Feature Extraction

Given the multimodal data as input, we need to preprocess them and map them into the dense representations for deep neural networks as shown in Figure 2. We denote the input text with N_s words as $\mathbf{S} = \{w_1, w_2, \ldots, w_{N_s}\}$. With the impressive performance of pre-trained language models, we utilize BERT (Devlin et al., 2019) to map the discrete words of sentence into the dense distributed representations. Before feeding the text into BERT, we should insert special tokens [CLS] and [SEP] into the start and end of the text. And the extended text is formulated as $\mathbf{S}' = \{w_0, w_1, \ldots, w_{N_s+1}\}$ where w_0 and w_{N_s+1} represent the special tokens respectively. The text feature extraction process can be simplified as $\mathbf{B} = \text{BERT}(\mathbf{S}') = \{b_i\}_{i=0}^{N_s+1}$. Considering to capture the contextual information further, we use BiLSTM networks for extracting hidden representations of the text. The extraction process can be defined as $\mathbf{H}^g = \text{BiLSTM}(\mathbf{B}; \theta_g) = \{\mathbf{h}_i^g\}_{i=0}^{N_s+1}$ and $\mathbf{H}^e = \text{BiLSTM}(\mathbf{B}; \theta_e) = \{\mathbf{h}_i^e\}_{i=0}^{N_s+1}$ where θ_g and θ_e are trainable weights in BiLSTM networks. As mentioned above, we focus on the span features and exploit them to predict the entities in the text. The spans of the text can be formulated as $\{\mathbf{S}_{(i,j)} | 1 \leq i \leq j \leq N_s\}$ where $\mathbf{S}_{(i,j)} = \{w_i, w_{i+1}, \ldots, w_j\}$. And the global representations of spans are denoted as $\{\mathbf{c}_{(i,j)}^g | 1 \leq$

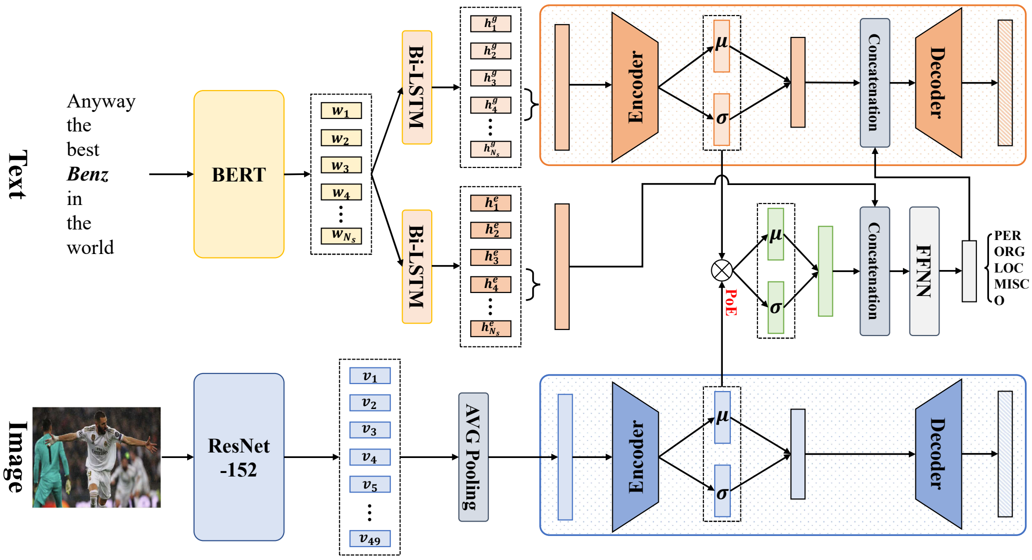

Figure 2: The overall architecture of span-based multimodal variational autoencoder for semi-supervised MNER.

$i \leq j \leq N_s\}$ where $\mathbf{c}^g_{(i,j)} = \frac{1}{j-i+1}\sum_{k=i}^{j}\mathbf{h}^g_k$. The edge representations of spans are calculated as $\{\mathbf{c}^e_{(i,j)}|1 \leq i \leq j \leq N_s\}$ where $\mathbf{c}^e_{(i,j)} = \left[\mathbf{h}^e_i;\mathbf{h}^e_j;\mathbf{h}^e_i - \mathbf{h}^e_j;\mathbf{h}^e_i \odot \mathbf{h}^e_j\right]$ and $\odot$ is the element-wise vector product.

For the visual modality, we utilize ResNet (He et al., 2016) to extract the regional representations of images. Before feeding the image into ResNet, we resize the image to 224 × 224 pixels. The regional representations of the image $\mathbf{V} = \{v_1, v_2, \ldots, v_{49}\}$ are extracted from the last conventional layer of ResNet. We apply an average pooling layer on the regional representations, and the global feature of the image is calculated as $\mathbf{V}^g = \frac{1}{49}\sum_{i=1}^{49} v_i$.

3.2 Multimodal Variational Autoencoder

To model the latent representations of the text and image modalities, the proposed SMVAE model consists of two modal-specific VAE networks named text-VAE and image-VAE. The encoders of VAEs contain dense layers to map the input features to the mean vector μ and standard deviation vector σ. For the text modality, the global representations of spans $\mathbf{c}^g$ are fed into text-VAE to parameterize the mean vector μ_s and standard deviation vector σ_s. The true posterior $p(\mathbf{z}^s|\mathbf{c}^g)$ can be approximated by the above parameters, and the distribution of $\mathbf{z}^s$ is formulated as $\mathbf{z}^s \sim q(\mathbf{z}^s|\mathbf{c}^g) = \mathcal{N}(\mu_s, \sigma_s^2)$. Therefore, μ_s and σ_s are computed by $\mu_s = \text{FFNN}(\mathbf{c}^g;\theta^s_\mu)$, $\sigma_s = \text{FFNN}(\mathbf{c}^g;\theta^s_\sigma)$ where FFNN is short for feed-forward neural networks, and θ^s_μ and θ^s_σ are trainable parameters in the encoder of text-VAE. For the visual modality, the global image features $\mathbf{V}^g$ are also fed into the encoder of the image-VAE. And the mean vector μ_v and standard deviation vector σ_v for image latent representations are calculated as $\mu_v = \text{FFNN}(\mathbf{V}^g;\theta^v_\mu)$, $\sigma_v = \text{FFNN}(\mathbf{V}^g;\theta^v_\sigma)$ where θ^v_μ and θ^v_σ are trainable weights in the encoder of image-VAE. We exploit the above parameters to approximate the true posterior $p(\mathbf{z}^v|\mathbf{V}^g)$, and the distribution of $\mathbf{z}^v$ is formulated as $\mathbf{z}^v \sim q(\mathbf{z}^v|\mathbf{V}^g) = \mathcal{N}(\mu_v, \sigma_v^2)$.

To bridge the semantic gap between the text and image representations, we need to calculate the multimodal features for predicting the results. The previous studies treated the text and image features as equals and mapped the concatenated features of the two modalities into the same latent representations (Khattar et al., 2019). However, there is the mismatch situation of the text and image that will introduce the noise into the model for predicting the result. We exploit the modal-specific VAEs to map the features of the two modalities into the respective latent representations with independent distributions. According to the assumption that two modalities are conditionally independent given the multimodal latent representations, the latent distribution $p(\mathbf{z}^m|\mathbf{c}^g, \mathbf{V}^g)$ of multimodal representations can be simplified as the combination of two individual latent distributions $p(\mathbf{z}^m|\mathbf{c}^g)$ and $p(\mathbf{z}^m|\mathbf{V}^g)$. Therefore, we apply the product-of-experts (Hinton, 2002) (PoE) to estimate the multimodal latent distribution by $p(\mathbf{z}^m|\mathbf{c}^g, \mathbf{V}^g) \propto p(\mathbf{z}^m|\mathbf{c}^g)p(\mathbf{z}^m|\mathbf{V}^g) = q(\mathbf{z}^s|\mathbf{c}^g)q(\mathbf{z}^v|\mathbf{V}^g)$. We assume the latent representations are independent

Gaussian distributions with mean and standard deviation parameters. Therefore, the distribution of $\mathbf{z}^m$ is formulated as $\mathbf{z}^m \sim \mathcal{N}(\mu_m, \sigma_m^2)$ where $\mu_m = \frac{\mu_s\sigma_v^2+\mu_v\sigma_s^2}{\sigma_s^2\sigma_v^2}$ and $\sigma_m^2 = (\sigma_s^{-2} + \sigma_v^{-2})^{-1}$.

To train the model in an end-to-end way, we utilize the reparameterization strategy (Kingma and Welling, 2014) to sample the latent representations. The latent variable $\mathbf{z}^m$ for multimodal representations can be calculated as $\mathbf{z}^m = \mu_m + \sigma_m \odot \epsilon$ where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. We utilize the multimodal features to predict the probabilities by $\hat{y} = \text{FFNN}([\mathbf{z}^m; \mathbf{c}^e]; \theta_o)$ where θ_o is the trainable weights of the prediction FFNN. Given the annotated entity set $\mathbf{y}$, the all negative instance candidates are defined as $\tilde{\mathbf{y}} = \{(l, r, \text{O}) | (l, r, \bar{y}) \notin \mathbf{y}, 1 \leq l \leq r \leq N_s, \bar{y} \in \mathcal{Y}\}$ where $\mathcal{Y}$ is the label space and O is the label for non-entity spans. To confirm the balanced class distribution of the samples in one batch, we randomly select a subset $\tilde{\mathbf{y}}'$ from the candidate set $\tilde{\mathbf{y}}$ with the same size of $\mathbf{y}$. The span-level cross entropy loss for training the model is defined as

$$\mathcal{L}_1 = \sum_{(i,j,\bar{y}) \in \tilde{\mathbf{y}}' \cup \mathbf{y}} -\bar{y} \log \hat{y}_{(i,j)} \tag{1}$$

where $\hat{y}_{(i,j)}$ is the prediction probability for the phrase $\mathbf{S}_{(i,j)}$.

The decoders of SMVAE are trained to reconstruct the representations of samples. For the text modality, the span types are correlated to the representations of spans. Therefore, we combine the true labels of labeled data or prediction probabilities of unlabeled data with the text latent representations and feed them into the decoder of text-VAE. The reconstructed representation of span is calculated as $\hat{\mathbf{c}}^g = \text{FFNN}([\mathbf{z}^s; \bar{y}]; \theta_d^s)$ for labeled data where $\mathbf{z}^s = \mu_s + \sigma_s \odot \epsilon$. The latent representations of images are fed into the decoder of image-VAE directly and the reconstructed representation is calculated as $\hat{\mathbf{V}}^g = \text{FFNN}(\mathbf{z}^v; \theta_d^v)$ where $\mathbf{z}^v = \mu_v + \sigma_v \odot \epsilon$. According to the evidence lower bound (ELBO) function of VAE (Kingma and Welling, 2014), the training loss for SMVAE on labeled data is formulated as follows:

$$\mathcal{L}_2 = \sum_{(i,j,\bar{y}) \in \tilde{y}' \cup \mathbf{y}} \|\mathbf{c}^g_{(i,j)} - \hat{\mathbf{c}}^g_{(i,j)}\|^2 + \|\mathbf{V}^g - \hat{\mathbf{V}}^g\|^2 + \text{KL}(q(\mathbf{z}^s|\mathbf{c}^g_{(i,j)})||p(\mathbf{z}^s)) + \text{KL}(q(\mathbf{z}^v|\mathbf{V}^g)||p(\mathbf{z}^v)) \tag{2}$$

where $\hat{\mathbf{c}}^g_{(i,j)}$ is the reconstructed representation of the phrase $\mathbf{S}_{(i,j)}$. For the unlabeled data, the reconstructed representation of span is calculated as $\hat{\mathbf{c}}^{g\prime} = \text{FFNN}([\mathbf{z}^s; \hat{y}]; \theta_d^s)$. Considering that there are more non-entity spans than named entity ones in a sample, we only learn the latent representations for the latter. And the training loss for unlabeled data is defined as follows:

Item	Twitter-2015			Twitter-2017		
	Train	Dev	Test	Train	Dev	Test
# Tweets	4,000	1,000	3,257	3,373	723	723
# PER entities	2,217	552	1,816	2,943	626	621
# LOC entities	2,091	522	1,697	731	173	178
# ORG entities	928	247	839	1,674	375	395
# MISC entities	940	225	726	701	150	157

Table 1: The statistical information of two MNER benchmark datasets.

$$\mathcal{L}_3 = \sum_{\substack{1 \leq i \leq j \leq N_s \\ \hat{y}_{(i,j)} \neq \text{O}}} \|\mathbf{c}^g_{(i,j)} - \hat{\mathbf{c}}^{g\prime}_{(i,j)}\|^2 + \|\mathbf{V}^g - \hat{\mathbf{V}}^g\|^2 + \text{KL}(q(\mathbf{z}^s|\mathbf{c}^g_{(i,j)})||p(\mathbf{z}^s)) + \text{KL}(q(\mathbf{z}^v|\mathbf{V}^g)||p(\mathbf{z}^v)) \tag{3}$$

3.3 Training Procedure

After acquiring the pre-processed multimodal labeled and unlabeled data, we feed them into the model to learn the latent representations of different modalities and extract the named entities. To train the model with different objectives at once, we introduce the hyper-parameter to sum Equation 1, Equation 2 and Equation 3. The overall loss function for the proposed model is defined as follows:

$$\mathcal{L} = \lambda \cdot \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3. \tag{4}$$

where λ is the hyper-parameter to balance the different losses. We feed the multimodal data into the model and acquire the loss according to Equation 4. To train the parameter weights of the model, we utilize the stochastic gradient descent (SGD) methods to update them based on the overall loss.

4 Experiments

4.1 Datasets and Experiment Settings

We compare the proposed model with the existing methods on the two widely used MNER datasets including: Twitter-2015 (Lu et al., 2018) and Twitter-2017 (Zhang et al., 2018). Each sample in the datasets is collected from Twitter and contain the text-image pair. There are four types of named entities including: Person (PER), Location (LOC), Organization (ORG) and others (MISC) that are annotated in the text. The detailed statistical information

Methods	Twitter-2015							Twitter-2017						
	Single Type (F1)				Overall			Single Type (F1)				Overall		
	PER	LOC	ORG	MISC	P	R	F1	PER	LOC	ORG	MISC	P	R	F1
Text														
ST	73.30	46.90	16.62	0.77	56.85	46.05	50.88	83.53	48.35	53.11	17.84	63.81	62.39	63.09
EM	76.29	50.12	8.52	0.78	61.30	46.27	52.73	81.69	51.95	48.80	1.47	69.45	59.32	63.99
SeqVAT	74.17	58.21	17.58	8.04	60.92	49.32	54.51	84.82	60.19	53.87	11.11	65.26	66.45	65.85
Multimodal														
UMT+ST	76.30	58.41	23.63	7.52	55.49	53.86	54.66	81.03	60.16	56.58	13.95	67.07	60.92	63.85
UMT+EM	72.91	**65.85**	28.51	13.92	52.59	58.03	55.17	79.94	58.74	54.02	18.00	62.84	62.84	62.84
UMT+SeqVAT	70.36	63.94	28.01	**12.89**	52.17	**60.42**	56.00	76.82	61.11	55.48	19.75	61.03	63.88	62.42
MAF+ST	77.18	52.44	12.77	0.52	57.14	51.18	54.12	82.31	51.49	52.35	9.76	70.81	58.18	63.88
MAF+EM	76.20	54.74	26.09	6.39	50.47	57.30	53.67	77.64	60.90	52.45	14.88	56.32	64.62	60.19
MAF+SeqVAT	74.00	63.54	29.96	10.34	52.67	58.41	55.39	81.66	**61.81**	57.03	19.61	64.15	65.43	64.79
Ours	**78.33**	65.44	**38.04**	7.90	**68.92**	55.76	**61.65***	**87.40**	58.33	**69.76**	**32.86**	**79.27**	**69.36**	**73.98***

Table 2: Performance comparison on two MNER datasets under semi-supervised settings. The numbers with $*$ indicate the improvement of our model over all baselines is statistically significant with $p \leqslant 0.05$ under t-test.

of two datasets is shown in Table 1. To compare our model with baselines under the **semi-supervised setting**, we split the original training set of each dataset into two parts: labeled dataset D_l and unlabeled one D_u. To assume that we are working with a small amount of labeled data, we randomly select 100 samples from the original training set as D_l and the remaining ones as D_u. And we run the semi-supervised experiment with five random seeds, each with a different split of labeled and unlabeled datasets, and report the mean performance on test data.

In the proposed model, we utilize the `BERT-base`[2] version of pre-trained language model BERT (Devlin et al., 2019) to extract text features, and use `ResNet152` (He et al., 2016) to extract image features. The size of hidden layers is set to 768, and the dimension of latent variables is set to 100 for modal-specific VAEs. We set the learning rate to 1e-5 and batch size to 8 for training the model. And the hyper-parameter λ in Equation 4 is set to $e^{(1-\frac{|D_l|}{|D_l|+|D_u|})}$. During the training process, we firstly train the model with the labeled and unlabeled set 100 epochs at most and test it on the development set. According to the early stopping strategy, we stop training the model when the F1 score on the development set does not increase within 10 epochs, and evaluate the best model on the test set. All experiments are accelerated by NVIDIA GTX 2080 Ti devices.

4.2 Compared Methods

Considering that there is no previous studies on semi-supervised MNER, we compare the proposed model with the widely used semi-supervised NER methods. The self-training (ST) and entropy minimization (EM) has been demonstrated the effectiveness on the semi-supervised NER (Chen et al., 2020). Besides, the existing state-of-the-art method SeqVAT combine virtual adversarial training (VAT) (Miyato et al., 2019) with conditional random field (CRF) (Lafferty et al., 2001) for semi-supervised sequence labeling (Chen et al., 2020). Therefore, we utilize BERT stacked with BiLSTM and CRF layers as the baseline model while applying ST, EM and SeqVAT methods based on it.

The above baseline methods are only for text modality. Besides, we also combine the effective MNER models with the above semi-supervised learning methods as semi-supervised MNER baselines. The uniform multimodal transformer (UMT) (Yu et al., 2020) was proposed to enhance the interactions of text and image modalities for the MNER task and achieved impressive performance. Xu et al. (2022) proposed the general matching and alignment for MNER (MAF) to fuse the text and image representations consistently and gained the best performance. Therefore, the semi-supervised MNER baselines are the combinations of the above MNER models and semi-supervised NER methods including: UMT+ST, UMT+EM, UMT+SeqVAT, MAF+ST, MAF+EM and MAF+SeqVAT.

4.3 Experimental Results

We compare SMVAE with the baseline methods on two benchmark datasets under semi-supervised setting, and report the metrics of F1 score (F1) for every single type and overall precision (P), recall (R) and F1 score (F1). The detailed experi-

[2] https://github.com/google-research/bert

Methods	Twitter-2015							Twitter-2017						
	Single Type (F1)				Overall			Single Type (F1)				Overall		
	PER	LOC	ORG	MISC	P	R	F1	PER	LOC	ORG	MISC	P	R	F1
Text														
BERT	84.72	79.91	58.26	38.81	68.30	74.61	71.32	90.88	84.00	79.25	61.63	82.19	83.72	82.95
BERT-CRF	84.74	80.51	60.27	37.29	69.22	74.59	71.81	90.25	83.05	81.13	62.21	83.32	83.57	83.44
BERT-BiLSTM-CRF	84.32	79.31	61.66	37.53	71.03	73.57	72.27	90.29	84.55	80.97	64.85	83.20	84.68	83.93
Multimodal														
GVATT-BERT-CRF	84.43	80.87	59.02	38.14	69.15	74.46	71.70	90.94	83.52	81.91	62.75	83.64	84.38	84.01
AdaCAN-BERT-CRF	85.28	80.64	59.39	38.88	69.87	74.59	72.15	90.20	82.97	82.67	64.83	85.13	83.20	84.10
MT-BERT-CRF	85.30	81.21	61.10	37.97	70.48	74.80	72.58	91.47	82.05	81.84	65.80	84.60	84.16	84.42
UMT-BERT-CRF	85.24	81.58	63.03	39.45	71.67	75.23	73.41	91.56	84.73	82.24	70.10	85.28	85.34	85.31
UMGF	84.26	**83.17**	62.45	42.42	**74.49**	75.21	74.85	91.92	85.22	83.13	69.83	**86.54**	84.50	85.51
UAMNer	85.14	81.66	62.46	40.95	73.02	74.75	73.87	91.86	85.71	84.25	68.73	86.17	86.23	86.20
MAF	84.67	81.18	**63.35**	41.82	71.86	75.10	73.42	91.51	**85.80**	**85.10**	68.79	86.13	86.38	86.25
Ours	**85.82**	81.56	63.20	**43.67**	74.40	**75.76**	**75.07**	**91.96**	81.89	84.13	**74.07**	85.77	**86.97**	**86.37**

Table 3: Performance comparison on two MNER datasets under supervised settings. The MNER models are trained with the training set of Twitter-2015 and Twitter-2017.

mental results on Twitter-2015 and Twitter-2017 are shown in Table 2. Our model can achieve the best results on most metrics, and the overall F1 scores of the proposed model increase 5.6% and 9.2% over baselines on two datasets respectively. The traditional semi-supervised NER method SeqVAT can achieve the best results over other baselines on the text modality, indicating that CRF combined with VAT for sequence modeling can improve the performance of models effectively. Therefore, the semi-supervised MNER methods including UMT+SeqVAT and MAF+SeqVAT can also gain the best overall F1 scores over than other baselines. Besides, the semi-supervised MNER methods can always gain better results than NER methods on Twitter-2015 but not on Twitter-2017. This situation verifies that the MNER baselines are not adapted to the low-resource setting and can not always make use of the mulitmodal features effectively under this setting. Our model can outperform the semi-supervised MNER baselines because we utilize the span features fused with image ones, and exploit modal-specific VAEs to jointly model multimodal latent representations and span labels for taking advantage of unlabeled data. Although SeqVAT can improve the robustness of sequence models, SMVAE can learn the multimodal latent representations and implicit correlation between it and labels that benefits semi-supervised MNER.

4.4 Further Discussion

To dig into the model, we conduct the analysis for presenting it in different aspects. We discuss the effect of the labeled data percent to the original training set and latent variable dimension. To demonstrate the effectiveness of SMVAE, we compare it with the superior MNER models under supervised setting and conduct ablation study to verify the usefulness of multimodal VAE.

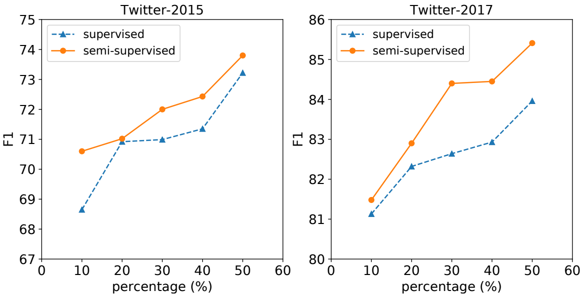

Figure 3: The performance of SMVAE under different settings vs. percent of the labeled data D_l to the original training set $D_l \cup D_u$.

Effect of Labeled Dataset Size. We explore the SMVAE performance with the percent of labeled data to the original training data under different settings. As shown in Figure 3, the "supervised" indicates SMVAE is trained with the labeled data under supervised learning, and the "semi-supervised" means that SMVAE is trained with the labeled and unlabeled data under semi-supervised learning. Under the same percent of labeled data, the performance of semi-supervised SMVAE can outperform the supervised learning results which demonstrates the effectiveness of SMVAE taking advantage of unlabeled data. And with the increase of labeled data, the performance of the proposed model can achieve better results on two datasets.

Supervised Setting. To verify the effectiveness of the proposed model with adequate labeled data, we compare SMVAE with state-of-the-art MNER models under supervised setting. The training sets of two datasets are used to train the model and eval-

Settings	Methods	Twitter-2015			Twitter-2017		
		P	R	F1	P	R	F1
Supervised	Ours	74.40	75.76	75.07	85.46	87.42	86.43
	w/o MVAE	73.17	75.51	74.32	85.16	87.05	86.09
Semi-supervised	Ours	68.92	55.76	61.65	79.27	69.36	73.98
	w/o MVAE	67.48	54.56	60.51	77.28	67.21	71.89

Table 4: The ablation study for SMVAE under different settings. "w/o MVAE" indicates that we turn off the multimodal VAE (MVAE) including text-VAE and image-VAE, and train the model for MNER.

uate it on the test set. The conventional MNER models including GVATT-BERT-CRF (Lu et al., 2018), AdaCAN-BERT-CRF (Zhang et al., 2018), UMT-BERT-CRF (Yu et al., 2020) designed the interaction module to fuse text and image modalities. Besides, Zhang et al. (2021) proposed UMGF model to combine the fine-grained image information with text one in the constructed graph way and achieved the impressive performance. Recently, MAF (Xu et al., 2022) and UMANer (Liu et al., 2022) were proposed to make the text and image aligned, and fuse them in a consistent way. As shown in Table 3, SMVAE outperforms the baselines on most metrics, and the overall F1 scores of it increase 0.22% and 0.12% over best baselines on two datasets respectively. We find that all MNER models are better than text-based NER models, indicating that the image information on social media posts is helpful to extract named entities in text. Compared with above discriminative models, our model can learn the modal-specific latent representations, and fuse the text and image modality by applying PoE on them to estimate multimodal latent features for tackling MNER.

Ablation Study. To investigate the effectiveness of multimodal VAE (MVAE) module in our model under different settings, we perform comparisons between the full model and the ablation method. The overall results of the models on two datasets are shown in Table 4. We find that the results of the model without MVAE are worse than the full model SMVAE under different settings which verifies the effectiveness of MVAE for tackling MNER. Further more, the ablation model performance degradation under supervised setting is lower than that under semi-supervised setting. Because there is adequate labeled data to train the model under supervised setting and MVAE plays an important role in SMVAE under semi-supervised setting. Under the low-resource settings, SMVAE can exploit MVAE module to jointly model the implicit correlation between multimodal representations and labels for making use of unlabeled data effectively.

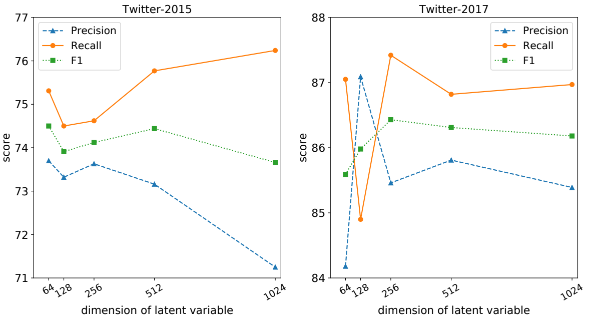

Figure 4: The performance of SMVAE under supervised setting vs. dimension of latent variable.

Effect of Latent Variable Dimension. The dimension of latent variables in MVAE is the key hyper-parameter to affect the performance of SMVAE, and we discuss the effect of it to the model under supervised setting. We set the dimension range from 64 to 1024 and take 2 times as an adjustment step. As shown in Figure 4, the performance of the model changes with the various dimensions of latent variable. When the dimension of latent variable is set higher, the performance of the model is degraded more. The multimodal latent variable represents the fusion of text and image modality, and the higher dimension means that more image information is introduced into the model. When the semantic relations of text and image in social media posts are mismatched, the latent variable with higher dimension introduces more noise into the model and affects the performance on MNER.

5 Conclusion

In this manuscript, we propose the semi-supervised multimodal named entity recognition (MNER) task and pose the critical challenge of it compared with traditional semi-supervised named entity recognition (NER). Further more, we analyze the disadvantage of the existing semi-supervised NER methods that are not sufficient to multimodal data. Therefore, we propose the span-based multimodal variational autoencoder to tackle semi-supervised MNER. The proposed model exploits multimodal VAE including two modal-specific VAEs to learn the multimodal latent representations and jointly model the implicit correlation between labels and multimodal features to make use of unlabeled multimodal data effectively. The experimental results verify that our approach not only outperforms supervised learning baselines, but also gains superior results than semi-supervised learning methods.

6 Limitations

The proposed model is limited to the length of input sentence because it needs to predict the type of all candidate spans during inference time. And the number of spans is proportional to the length of the sentence. Therefore, the inference time is increased with the length of sentence. Besides, our model has poor scalability to process more than one image, and the posted Twitter message may contain more than one image. Therefore, the future MNER model should be able to process the text with more images.

Acknowledgements

We thank the anonymous reviewers for the valuable comments on our manuscript. This research is supported by the Chinese Scientific and Technical Innovation Project 2030 (2018AAA0102100), the National Natural Science Foundation of China (No. 62272250, U1936206, U1903128, 62002178).

References

Dawei Chen, Zhixu Li, Binbin Gu, and Zhigang Chen. 2021. Multimodal named entity recognition with image attributes and image knowledge. In *Database Systems for Advanced Applications - 26th International Conference*, volume 12682 of *Lecture Notes in Computer Science*, pages 186–201. Springer.

Luoxin Chen, Weitong Ruan, Xinyue Liu, and Jianhua Lu. 2020. SeqVAT: Virtual adversarial training for semi-supervised sequence labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8801–8811, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc Le. 2018. Semi-supervised sequence modeling with cross-view training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1914–1925, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yves Grandvalet and Yoshua Bengio. 2004. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems 17*, pages 529–536.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778. IEEE Computer Society.

Geoffrey E. Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, pages 2915–2921. ACM.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Luping Liu, Meiling Wang, Mozhi Zhang, Linbo Qing, and Xiaohai He. 2022. Uamner: uncertainty-aware multimodal named entity recognition in social media posts. *Appl. Intell.*, 52(4):4109–4125.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999, Melbourne, Australia. Association for Computational Linguistics.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2019. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(8):1979–1993.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity recognition for short social media posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 852–860, New Orleans, Louisiana. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Jorn Veenstra. 1999. Representing text chunks. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 173–179, Bergen, Norway. Association for Computational Linguistics.

Bo Xu, Shizhou Huang, Chaofeng Sha, and Hongya Wang. 2022. MAF: A general matching and alignment framework for multimodal named entity recognition. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, pages 1215–1223. ACM.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3352, Online. Association for Computational Linguistics.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021. Multimodal graph fusion for named entity recognition with targeted visual guidance. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 14347–14355. AAAI Press.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5674–5681. AAAI Press.