RIVA: A Pre-trained Tweet Multimodal Model Based on Text-image Relation for Multimodal NER

Lin Sun[1], **Jiquan Wang**[2], **Yindu Su**[2], **Fangsheng Weng**[1], **Yuxuan Sun**[1], **Zengwei Zheng**[1], **and Yuanyi Chen**[1]
[1] Department of Computer Science, Zhejiang University City College, Hangzhou, China
{sunl,zhengzw,chenyuanyi@zucc.edu.cn}@zucc.edu.cn
[2] College of Computer Science and Technology, Zhejiang University, Hangzhou, China
{wangjiquan,yindusu}@zju.edu.cn

Abstract

Multimodal named entity recognition (MNER) for tweets has received increasing attention recently. Most of the multimodal methods used attention mechanisms to capture the text-related visual information. However, unrelated or weakly related text-image pairs account for a large proportion in tweets. Visual clues unrelated to the text would incur uncertain or even negative effects for multimodal model learning. In this paper, we propose a novel pre-trained multimodal model based on Relationship Inference and Visual Attention (RIVA) for tweets. The RIVA model controls the attention-based visual clues with a gate regarding the role of image to the semantics of text. We use a teacher-student semi-supervised paradigm to leverage a large unlabeled multimodal tweet corpus with a labeled data set for text-image relation classification. In the multimodal NER task, the experimental results show the significance of text-related visual features for the visual-linguistic model and our approach achieves SOTA performance on the MNER datasets.

1 Introduction

Social media such as Twitter has become a part of many people's everyday lives. It is an important source for various applications such as open event extraction (Wang et al., 2019), social knowledge graph (Hosseini, 2019). Named entity recognition (NER) for tweets is the first key task of these applications. NER has achieved excellent performance on news articles; however, the recognition results in tweets are still not satisfactory (Akbik et al., 2018; Akbik et al., 2019). One of the reasons is that tweets are short messages and the context for inference is insufficient. Recent works on tweets based on multimodal learning have been increasing (Moon et al., 2018; Lu et al., 2018; Zhang et al., 2018; Arshad et al., 2019). The researchers attempted to improve the performance of NER in tweets with the aid of visual clues.

Most of the multimodal NER (MNER) methods used attention weights to extract visual clues related to the NEs (Lu et al., 2018; Zhang et al., 2018; Arshad et al., 2019) . The visual attention-based models always assume that the images in tweets are related to the texts, such as words in the text are represented in the image, e.g., Figure 1(a) shows a successful visual attention example from Lu et al. (2018). In fact, texts and images in tweets have diverse relations. Vempala and Preoţiuc-Pietro (2019) categorized text-image relationship according to whether "Image adds to the tweet meaning". "Image adds to the tweet meaning" represents the role of image to the semantics of text in tweets. The type of "Image does not add to the tweet meaning" account for approximately 56% in the Vempala's Bloomberg dataset. In addition, we test a classifier regarding whether "Image adds to the tweet meaning" on a large randomly collected corpus, Twitter100k (Hu et al., 2017), the proportion of the classified negatives is approximately 60%. Figure 1(b) shows a failure visual attention example of "Image does not add to the tweet meaning". Therefore, visual features represented by attention weights do not always have positive effects on training visual-linguistic models. Moreover, the unrelated visual clues may increase the possibility of wrong connection to the text inference.

Proceedings of the 28th International Conference on Computational Linguistics, pages 1852–1862
Barcelona, Spain (Online), December 8-13, 2020

(a) **[PER Radiohead]** offers old and new at first concert in four years...

(b) Nice image of **[PER Kevin Love]** and **[PER Kyle Korver]** during 1st half #NBAFinals #Cavsin9 # **[LOC Cleveland]**.

Figure 1: Visual attention examples in multimodal NER in (Lu et al., 2018). (a) Successful visual attention example, (b) Failure visual attention example.

In this paper, we consider inferring the text-image relationship to address the problem of inappropriate visual clues fused in the multimodal model. The text-image relationship is defined as whether the image's content can contribute additional information beyond the text. The contributions of this paper can be summarized as follows:

- We propose a novel pre-trained multimodal language model (LM) based on Relationship Inference and Visual Attention (RIVA) for tweets. A gated visual context based on text-image relation is presented. The results show that the RIVA model can better utilize visual features than other visual attention models.
- We employ a teacher-student-based semi-supervised method to learn text-image relation on a large unlabeled corpus of tweets with a labeled dataset and generate a significantly improved performance on the text-image relation classification. The ablation study justifies the significance of accurate classification on text-image relation for the tweet-based multimodal model.
- We propose a multitask framework of text-image relation classification and next word prediction (NWP) for pre-training a multimodal LM of tweets. The RIVA model is trained on a large multimodal corpus of tweets. We demonstrate the performance of the RIVA model in the multimodal NER task and achieve the state-of-the-art results.

2 Related Work

Multimodal NER for social media posts has been investigated in recent years. Moon et al. (2018) proposed a modality-attention module at the input of the NER network. The module computed a weighted modal combination of word embeddings, character embeddings, and visual features. Lu et al. (2018) presented a visual attention model to find the image regions that were related to the content of the text. The attention weights of the image regions were computed by a linear projection of the sum of the text query vector and regional visual representations. The extracted visual context features were incorporated to the word-level outputs of the biLSTM model. Zhang et al. (2018) designed an adaptive co-attention network (ACN) layer, which was between the LSTM and CRF layers. The ACN contained a gated multimodal fusion module to learn a fusion vector of the visual and linguistic features. The author designed a filtration gate to decide whether the fusion feature was helpful to improve the tagging accuracy of each token. The output score of the filtration gate was computed by a sigmoid activation function. Arshad et al. (2019) also presented a gated multimodal fusion representation for each token. The gated fusion is a weighted sum of visual attention feature and token alignment feature. The visual attention feature was calculated by the weighted sum of VGG-19 (Simonyan and Zisserman, 2014) visual features and the weights were the additive attention scores between a word query and image features. Overall, the problem of attention-guided visual feature is that the incorrect visual context clues could be extracted when the images and texts are not relevant. They pointed out that the unrelated images caused the wrong attention and prediction errors, and showed the failed examples. Although Zhang et al. (2018) attempted to use a sigmoid-based gate to filter the unrelated visual feature, the improvement was marginal when we tested the available source code by the authors.

The pre-trained models such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019) have achieved great success in natural language processing (NLP). These models are trained on large unlabeled corpus by self-supervised learning tasks. The latest pre-trained multimodal models, visual-linguistic BERT, were presented in (Su et al., 2019; Lu et al., 2019; Li et al., 2019). The authors extended the popular BERT architecture to multimodal models for vision and language. Li et al. (2019) pre-trained VisualBERT on the COCO image caption dataset (Chen et al., 2015). The VisualBERT consisted of a stack of Transformer layers that implicitly aligned elements of the text and regions in the image with self-attention. Lu et al. (2019) and Su et al. (2019) trained the models on a larger caption dataset, Conceptual Captions (Sharma et al., 2018) consisting of 3.3 million images annotated with captions. Lu et al. (2019) presented co-attentional transformer layers in ViLBERT to learn interactively between visual and linguistic features. Two pre-training tasks: masked learning and alignment prediction, were applied. The alignment task was a binary classification which defined the pairs in the caption dataset as positives and the pairs generated by replacing the image or text with each other as negatives. Su et al. (2019) designed four types of embeddings in VL-BERT and input the sum to the BERT architecture. The attention module took both visual and linguistic features as input. The input element was either word embeddings or Fast R-CNN (Girshick, 2015) features for Regions of Interest (RoIs) in an image. Su et al. performed the text-image relation prediction and the data setting was the same as the alignment task in (Lu et al., 2019). The ablation study of the training tasks in (Su et al., 2019) showed that the task of text-image relation would decrease the accuracy of all downstream tasks, and the authors guessed the reason was that the unmatched image and caption pairs were introduced.

3 The RIVA Model

The proposed pre-trained multimodal model for tweets, called RIVA, is shown in Figure 2. The RIVA model contains three parts: 1) text-image relation gating network (RGN), 2) attention-guided visual context network (VCN), and 3) visual-linguistic contextual network (VLCN). The RGN is based on binary classification of text-image relation and outputs a relevance score s^G between text and image. The score s^G is served as a gating control in the path from VCN to VLCN. The VCN is a visual-linguistic attention-based network, which attempts to extract the local visual information relevant to the text. The visual contextual output of VCN is fed to the input of long short-term memory (LSTM) network to guide the learning of VLCN. The VLCN is a visual-linguistic language model that performs the next word prediction (NWP) task learning. The detailed descriptions of the model are presented in the following subsections.

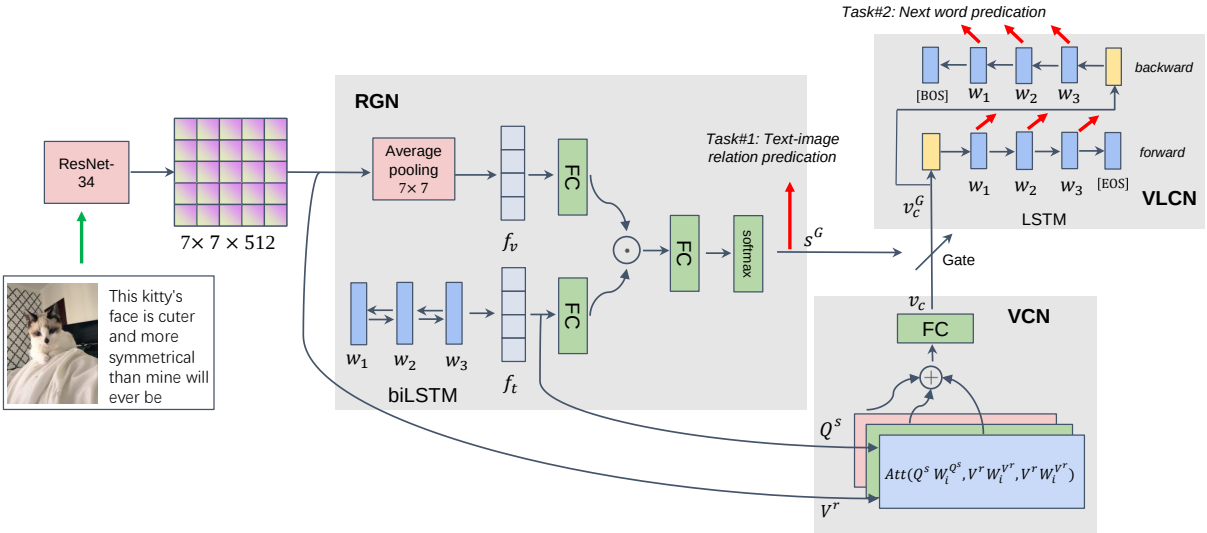


Figure 2: The neural architecture of RIVA.

3.1 The RGN and Semi-supervised Learning

In the RGN, the text-image relation classification is performed by a fully connected (FC) layer based on a fusion of linguistic and visual features. The linguistic feature of tweets is learned from a bidirectional LSTM (biLSTM) network. The input of biLSTM is a concatenation of word and character embeddings (Lample et al., 2016). We encode a Twitter text into a vector $f_t \in \mathcal{R}^{1\times d_t}$, which is a concatenation of the forward and backward outputs of biLSTM. The visual feature f_v is extracted from the image by ResNet (He et al., 2016). The output size of the last convolutional layer in ResNet is $7 \times 7 \times d_v$. We use average pooling over 7×7 regions and represent the full image as a d_v-dimensional vector f_v, where $d_v = 512$ when working with ResNet-34. The element-wise multiplication of the linguistic and visual features, $f_t \odot f_v$, is followed by a FC and softmax layer to yield a score s_G for binary classification and visual context gating. The training task on the RGN is described as follows:

Task#1: Text-image relation classification: We employ the "Image Task" data of (Vempala and Preoţiuc-Pietro, 2019) for text-image relation classification. This classification attempts to identify whether the image's content contributes additional information beyond the text. The types of text-image relation and statistics in the Bloomberg's dataset are shown in Table 1.

To leverage a large unlabeled multimodal tweet corpus, we employ a teacher-student semi-supervised paradigm (Yalniz et al., 2019). First, we train a teacher model on the Bloomberg's dataset. The teacher model is a separate network, which has the same architecture as the RGN. Second, we predict a large unlabeled tweet corpus, Twitter100k (Hu et al., 2017), using the teacher model. We pick the tweets with higher category scores (> 0.6) to construct a new pseudo-labeled training data, denoted as "pseudo-labeled 100k". In the training of the RIVA model, the RGN acts as a student model. It is firstly trained on the "pseudo-labeled 100k" data and then fine-tuned by the "Bloomberg" labeled data to reduce noisy labeling errors.

Let $x_i =< text_i, image_i >$ be a text-image pair of tweet. The loss $\mathcal{L}_{task_1}$ of relation classification on "Bloomberg" and "pseudo-labeled 100k" data is calculated by cross entropy:

$$\mathcal{L}_{task_1}^{Bloomberg} = - \sum_{x_i \in Bloomberg} log(p(x_i)), \tag{1}$$

$$\mathcal{L}_{task_1}^{pseudo\text{-}labeled\ 100k} = - \sum_{x_i \in pseudo\text{-}labeled\ 100k} log(p(x_i)), \tag{2}$$

where $p(x)$ is the probability for correct classification, and computed by a softmax layer.

3.2 The VCN

The output size of the last convolutional layer in ResNet has a shape of $7 \times 7 \times d_v$, where 7×7 denotes 49 regions in an image. Let $V^r = \{v_{i,j}^r\}$ be the region features of a given image, where $i = 1, \ldots, 7, j = 1, \ldots, 7, v_{i,j}^r \in \mathcal{R}^{1\times d_v}$. We employ Scaled Dot-Product Attention (Vaswani et al., 2017) to capture the local visual features related to the linguistic context. Scaled Dot-Product Attention is generally defined as follows:

$$Att(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{3}$$

where matrices Q, K and V consist of queries, keys and values, d_k is the dimension of keys. In this work, we use a linguistic query vector $Q^s = f_t$ as a query and region feature V^r as both keys and values. Q^s and V^r are transformed into the same dimension by the linear projections, $W^{Q^s} \in \mathcal{R}^{d_t\times d_v}$ and $W^{V^r} \in \mathcal{R}^{d_v\times d_v}$. Therefore, the computation of the visual-linguistic attention can be formulated as $Att(Q^sW^{Q^r}, V^rW^{V^r}, V^rW^{V^r})$. We also extend single attention to multi-head attention as in (Vaswani et al., 2017). Finally, the output of local visual context v_c is defined as follows:

$$head_i = Att(Q^sW_i^{Q^s}, V^rW_i^{V^r}, V^rW_i^{V^r}) \tag{4}$$

$$v_c = head_1 \oplus ... \oplus head_h, \tag{5}$$

where $W_i^{Q^s} \in \mathcal{R}^{d_t\times d_h}$, $W_i^{V^r} \in \mathcal{R}^{d_v\times d_h}$, $d_h = d_v/h$.

3.3 The VLCN

We use biLSTM to learn visual-linguistic contextual embeddings on a large multimodal Twitter dataset, Twitter100k. The architecture of VLCN is similar to that of image caption generation (Vinyals et al., 2015). Given a visual vector $v_c^G = s^G \cdot v_c$ and a sequence of T words $\{w_t\}, t = 1, \ldots, T$, a forward LSTM predicts w_t on the history $(w_1, \ldots, w_{t-1})$ with v_c^G at $t = 0$. In this work, we add a backward LSTM to predict w_t on the history $(w_{t+1}, \ldots, w_T)$ with v_c^G at $t = T + 1$. To align tokens in the forward and backward directions, we add the beginning [BOS] and end [EOS] tokens in the word sequence: ([BOS], $w_1, \ldots, w_T$, [EOS]). We replace [BOS] with visual features in the forward prediction and [EOS] in the backward prediction. The word input of LSTM is a concatenation of word and character embeddings, the same as biLSTM in the RGN. The details of the NWP task are described as follows:

Task#2: Next word prediction: The VLCN module computes the probability of the sequence by modeling the probability of the next word w_t in both forward and backward directions. The probability of the sentence is:

$$\begin{aligned} &p(w_1, w_2, \ldots, w_T | v_c^G) = \\ &\prod_{t=1}^{T} p(w_t | v_c^G, w_1, w_2, \ldots, w_{t-1}) p(w_t | w_{t+1}, w_{t+2}, \ldots, w_T, v_c^G), \end{aligned} \tag{6}$$

where the probability $p(w_t|\cdot)$ can be computed on the hidden output of LSTM followed by a FC and a softmax layer.

The log probability of $p(w_1, w_2, \ldots, w_T | v_c^G)$ can be implemented by cross entropy loss on the predicted words. Therefore, the training task is to minimize the objective $\mathcal{L}_{task_2}$ of the forward and backward directions:

$$\begin{aligned} \mathcal{L}_{task_2} = &- \sum_{t=1}^{T} log(p(w_t | v_c^G, w_1, w_2, \ldots, w_{t-1})) \\ &- \sum_{t=1}^{T} log(p(w_t | w_{t+1}, w_{t+2}, \ldots, w_T, v_c^G)) \end{aligned} \tag{7}$$

Finally, combining with Task#1, the complete training procedure of the RIVA model is illustrated in Algorithm 1. θ_{RGN}, θ_{VCN}, and θ_{VLCN} represent the parameters of the RGN, VCN, and VLCN, respectively. In each epoch, the algorithm firstly performs both Task#1 and Task#2 to train the RIVA model on the "pseudo-labeled 100k" data and secondly, performs Task#1 to finetune the RGN on the "Bloomberg" labeled data.

Algorithm 1 Training procedure of the RIVA model.

Require: The "pseudo-labeled 100k" data, the "Bloomberg" labeled data, θ_{RGN},θ_{VCN}, and θ_{VLCN}.

for all epochs **do**
  for all batches **do**
    Forward text-image pairs of "pseudo-labeled 100k"
    Compute loss $\mathcal{L}^{pseudo\text{-}labeled\ 100k} = \mathcal{L}_{task_1}^{pseudo\text{-}labeled\ 100k} + \mathcal{L}_{task_2}^{pseudo\text{-}labeled\ 100k}$
    Update θ_{RGN},θ_{VCN}, and θ_{VLCN} using $\nabla \mathcal{L}^{pseudo\text{-}labeled\ 100k}$
  end for
  for all batches **do**
    Forward text-image pairs of "Bloomberg"
    Compute loss $\mathcal{L}_{task_1}^{Bloomberg}$
    Update θ_{RGN} using $\nabla \mathcal{L}_{task_1}^{Bloomberg}$
  end for
end for

4 Experiments

4.1 Datasets

- **Twitter100k dataset (Hu et al., 2017)**: This dataset is comprised of 100,000 image-text pairs randomly crawled from Twitter. An image-text pair contains an image and a text appearing in one piece of tweet. Approximately 1/4 of the images are highly correlated to their respective texts. The authors studied weakly supervised learning for cross-media retrieval on this dataset. In this paper, we use this dataset as a large unlabeled multimodal corpus, and to perform Task#1 and Task#2 of the RIVA model.

- **Bloomberg's text-image relation dataset (Vempala and Preoţiuc-Pietro, 2019)**: In this dataset, the authors annotated tweets into four types of text-image relation, shown in Table 1. "Text is presented in image" is centered on the role of text to the semantics of tweet while "Image adds to the tweet meaning" focuses on the image's role. In the RIVA model, we treat text-image relation as binary classification task between $R_1 \cup R_2$ and $R_3 \cup R_4$. We follow the same split of 8:2 for train/test sets as in (Vempala and Preoţiuc-Pietro, 2019). We use this dataset for training the teacher model and fine-tuning the student model in semi-supervised learning of Task#1.

	Image adds to the tweet meaning	Text is presented in image	Percentage (%)
R_1	$\surd$	$\surd$	18.5
R_2	$\surd$	$\times$	25.6
R_3	$\times$	$\surd$	21.9
R_4	$\times$	$\times$	33.8

Table 1: Four types of text-image relation in Bloomberg's dataset.

- **MNER Twitter dataset of Fudan University (Zhang et al., 2018)** : The authors sampled the tweets with images collected through Twitter's API. In this dataset, NE types are Person, Location, Organization, and Misc. The authors labeled 8,257 tweet texts using BIO2 tagging scheme and used a 4,000/1,000/3,257 train/dev/test split.

- **MNER Twitter dataset of Snap Research (Lu et al., 2018)**: The authors collected the data from Twitter and Snapchat, but Snapchat data is not available for public use. NE types are Person, Location, Organization, and Misc. Each data instance contains one sentence and one image. The authors labeled 6,882 tweet texts using BIO tagging scheme and used a 4,817/1,032/1,033 train/dev/test split.

4.2 Using RIVA for Multimodal NER Task

We use a baseline NER model, biLSTM-CRF (Lample et al., 2016), to test our pre-trained multimodal model. The biLSTM-CRF model consists of a bidirectional LSTM and conditional random fields (CRF) (Lafferty et al., 2001). The token embeddings e_k are fed to the input of biLSTM. The CRF uses the biLSTM hidden vectors h_t of each token to tag the sequence with entity labels. When using the RIVA model, the text-image pairs are input. We concatenate the hidden output of the forward and backward LSTMs in the VCLN for each token as visual-linguistic context embeddings e_k^{RIVA}. For NER task, we replace the token embeddings e_k with $\left[e_k; e_k^{RIVA}\right]$.

4.3 Settings

We use the 100-dimensional GloVe (Pennington et al., 2014) word vectors in the RIVA model and 300-dimensional FastText Crawl (Mikolov et al., 2018) word vectors in the biLSTM-CRF model, since the vocabulary size of the pre-trained LMs is not large, e.g., $30K$ in BERT and $130K$ in ELMo. All images are reshaped to a size of 224×224 to match the input of ResNet. We use ResNet-34 to extract visual features and finetune it with a learning rate of 1e-4. The FC layers in Figure 2 are a linear neural network followed by a GELU activation (Hendrycks and Gimpel, 2016). We train the model on a machine with NVIDIA Tesla K80 (GPU) and Intel Xeon Silver 4114 Processor 2.2 GHz (CPU). The training

of the RIVA model takes approximately 32 hours for 35 epochs on one GPU kernel. Table 2 shows the hyperparameter values in the RIVA and biLSTM-CRF models.

RIVA		biLSTM-CRF		Misc.	
biLSTM hidden size of RGN	256	biLSTM hidden size	256	dropout rate	0.5
biLSTM layer of RGN	2	+RIVA	512	optimizer	SGD
LSTM hidden size of VLCN	256	+BERT	1024	learning rate	2e-2
LSTM layer of VLCN	2	biLSTM layer	1	learning rate for pretrained models	1e-4
number of heads h in VCN	8	batch size	8	char embedding dimension	25
batch size	16			clip gradient norm	5.0

Table 2: Hyperparameters of the RIVA and biLSTM-CRF models.

4.4 Performance of Text-image Relation Classification

Table 3 shows the performance of the RGN for text-image relation classification on the test set of Bloomberg's data. In terms of network structure, Lu et al. (2018) represented the multimodal feature as a concatenation of the linguistic and visual features while the RGN employs element-wise multiplication. The merit of element-wise multiplication is that the parameter gradients in one modality can be influenced more by the data of another modality and achieves collaborative learning on the multimodal data. F1 score of the RGN on Bloomberg's data increases by 4.7% compared to Lu et al. (2018). Combining with "pseudo-labeled 100k", the performance of the RGN achieves an improvement of 1.1%.

	Lu et al. (2018)	RGN		
		Bloomberg	pseudo-labeled 100k	pseudo-labeled 100k with Bloomberg finetuning
F1 score	81.0	(**+4.7**) 85.7	(**+5.2**) 86.2	(**+5.8**) 86.8

Table 3: Comparison of text-image relation classification in F1 score (%).

4.5 Results of the RIVA Model

Table 4 illustrates the improved performance of the RIVA model compared to biLSTM-CRF. "biLSTM-CRF (text)" performs the NER task on the sequence of word embeddings. "biLSTM-CRF (image+text)" adds the visual feature at the beginning of the embedding sequence to inform the biLSTM-CRF model about the image content. "biLSTM-CRF (text) + RIVA (image+text)" denotes that the text-image pairs are input to the RIVA model, as clarified in Section 4.2. "biLSTM-CRF (text) + RIVA (text)" denotes that texts are the only input of the RIVA model, i.e., the RGN and VCN are ablated. The results show that "+ RIVA (image+text)" achieves an increase of 1.8% and 2.2% compared to "biLSTM-CRF (text)" on the Fudan Univ. and Snap Res. datasets, respectively. In terms of the role of visual features, the improvement of "biLSTM-CRF (image+text)" compared to "biLSTM-CRF (text)" in F1 score is on average 0.4% while the performance of "+ RIVA (image+text)" increases by an average of 1.5% compared to "+ RIVA (text)". This indicates that the RIVA model can better utilize visual features to enhance the context of tweets.

	Fudan Univ.	Snap Res.
biLSTM-CRF (text)	(**+0.0**) 69.7	(**+0.0**) 80.1
biLSTM-CRF (image+text)	(**+0.2**) 69.9	(**+0.5**) 80.6
biLSTM-CRF (text) + RIVA (text)	(**+0.3**) 70.0	(**+0.8**) 80.9
biLSTM-CRF (text) + RIVA (image+text)	(**+1.8**) 71.5	(**+2.2**) 82.3

Table 4: Comparison of the improved performance of the RIVA model in F1 score (%).

In Table 5, we compare performance with other biLSTM-CRF based MNER methods (Zhang et al., 2018; Lu et al., 2018) and visual-linguistic pre-trained models. To compare with VL-BERT (Su et al., 2019) and ViLBERT (Lu et al., 2019), we concatenate the RIVA embeddings with the BERT$_{Base}$ embeddings. We finetune BERT, VL-BERT, and ViLBERT models with a learning rate of 1e-4 for MNER task. We average the embeddings of BERT-tokenized subwords to generate an approximate

vector for out-of-vocabulary (OOV) words and additionally use character-level contextual embeddings in Flair (Akbik et al., 2019). The input token embeddings of LSTM-CRF is a concatenation of the original embedding and pre-trained contextual embeddings. For example, "biLSTM-CRF + RIVA + BERT" means that the input token embeddings of LSTM-CRF is $\left[e_k; e_k^{RIVA}; e_k^{BERT}\right]$. "+ RIVA + BERT" achieves an increase of 1.6% on average in F1 score compared to "+ BERT" and outperforms both "+ VL-BERT" and "+ ViLBERT" by approximately 1%. The setting of "biLSTM-CRF + RIVA + BERT + Flair" performs the best, achieving 73.8% on the Fudan Univ. dataset and 87.4% on the Snap Res. dataset.

	Fudan Univ.	Snap Res.
Zhang et al. (2018)	70.7	-
Lu et al. (2018)	-	80.7
biLSTM-CRF + RIVA	71.5	82.3
biLSTM-CRF + BERT	**(+0.0)** 71.5	**(+0.0)** 85.5
biLSTM-CRF + VL-BERT	**(+0.7)** 72.2	**(+0.6)** 86.1
biLSTM-CRF + ViLBERT	**(+0.5)** 72.0	**(+0.3)** 85.8
biLSTM-CRF + RIVA + BERT	**(+1.8)** 73.3	**(+1.3)** 86.8
biLSTM-CRF + RIVA + BERT + Flair	**(+2.3) 73.8**	**(+1.9) 87.4**

Table 5: Performance comparison with other methods and visual-linguistic pre-trained models in F1 score (%).

4.6 Ablation Study

In this section, we report the results when ablating text-image relation classification. We ablate the RGN ("$-$RGN") in the RIVA model, equivalently, the output of the VCN is directly passed to the input of biLSTM of the VLCN, i.e., $s^G = 1$. Table 6 shows that the overall performance decreases 0.7% and 0.9% on the Fudan Univ. and Snap Res. datasets, respectively, when the RGN is ablated. In addition, we divide the test data into two sets, "Image adds" and "Image doesn't add", by classification of the RGN, and compare the impact of the ablation on data of different text-image relation types. More importantly, we find that the performance has hardly changed on the data of "Image adds", but drops on the data of "Image doesn't add", -1.2% on the Fudan Univ. dataset and -1.5% on the Snap Res. dataset. This also justifies that the text-unrelated visual features have negative effects on learning visual-linguistic representations.

	Fudan Univ.			Snap Res.		
	Image adds	Image doesn't add	Overall	Image adds	Image doesn't add	Overall
biLSTM-CRF + RIVA	71.3	71.6	71.5	82.5	82.0	82.3
$-$RGN	**(-0.1)** 71.2	**(-1.2)** 70.4	**(-0.7)** 70.8	**(-0.3)** 82.2	**(-1.5)** 80.5	**(-0.9)** 81.4

Table 6: Performance comparison when the RGN is ablated.

4.7 Case Study

We illustrate four failure examples mentioned in (Lu et al., 2018) and (Arshad et al., 2019) in Table 7. The common reason for these failed examples is due to the incorrect visual attention features. The "Text" column is Twitter text with the ground truth labels. The "biLSTM-CRF + RIVA" column is the labeled results by "biLSTM-CRF + RIVA" and the "Previous work" column is the labeled results by (Lu et al., 2018) or (Arshad et al., 2019). We also show the relevance score s^G and visual attention weights in the RIVA model. The visual attention weights are computed as the average sum of softmax weights of multiple heads in Eq. (3),

$$\frac{1}{h}\sum_{i=1}^{h} softmax(\frac{(Q^s W_i^{Q^s}) \cdot (V^r W_i^{V^r})^T}{\sqrt{d_h}}). \tag{8}$$

The attention weights are visualized using heat map. High values are in yellow and low values are in blue.

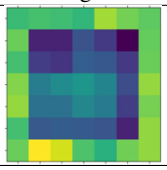
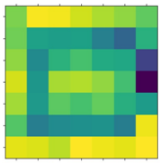
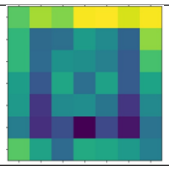
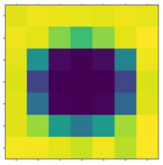
	Image	Text	s^G	Visual attention weights	biLSTM-CRF + RIVA	Previous work
1		Looking forward to editing some **[ORG SBU]** baseball shots from Saturday.	2.7e-7		Looking forward to editing some **[ORG SBU]** baseball shots from Saturday.	Looking forward to editing some SBU baseball shots from Saturday. (Lu et al., 2018)
2		Nice image of **[PER Kevin Love]** and **[PER Kyle Korver]** during 1 st half # NBAFinals # Cavsin9 # **[ORG Cleveland]**	6.0e-7		Nice image of **[PER Kevin Love]** and **[PER Kyle Korver]** during 1 st half # NBAFinals # Cavsin9 # **[ORG Cleveland]**	Nice image of **[PER Kevin Love]** and **[PER Kyle Korver]** during 1 st half # NBAFinals # Cavsin9 # **[LOC Cleveland]**. (Lu et al., 2018)
3		**[MISC Reddit]** needs to stop pretending racism is valuable debate.	1.7e-6		**[ORG Reddit]** needs to stop pretending racism is valuable debate.	**[ORG Reddit]** needs to stop pretending racism is valuable debate. (Arshad et al., 2019)
4		**[ORG PSD Lesher]** teachers take school spirit to top of 14ner **[LOC Mount Sherman]**.	0.99		**[ORG PSD Lesher]** teachers take school spirit to top of 14ner **[LOC Mount Sherman]**.	**[ORG PSD Lesher]** teachers take school spirit to top of 14ner **[PER Mount Sherman]**. (Arshad et al., 2019)
				Low  High		

Table 7: Four MNER examples in previous work and the results using the RIVA model.

Examples 1 and 2 are from the Snap Res. dataset and Example 3 and 4 are from the Fudan Univ. dataset. In Example 1, the visual attention regions are not related to the entity "SBU" and result in missed tagging of "SBU". In Example 2, the visual attention regions focus on the wall and ground and result in tagging "Cleveland" as a wrong label "LOC". In the RIVA model, the score s^G is approximately 0, therefore no visual feature is used when tagging labels and we obtain the correct results in Examples 1 and 2. In Example 3, although the text-image pair is classified as unrelated, "Reddit" is still labeled as "ORG" because of the linguistic features. In Example 4, the highly related visual attention, e.g., sky and mountain, produces the correct label of "Mount Sherman" in the RIVA model. However, in (Arshad et al., 2019), the incorrect visual attention, e.g., teachers, causes a wrong label "PER" for "Mount Sherman".

5 Conclusion

This paper concerns the problem of visual attention features in multimodal learning when images are unrelated to texts, typically in tweets. The text-unrelated visual features would incur negative rewards in multimodal NER. In the paper, we propose a pre-trained multimodal model based on text-image relationship inference. The relation of whether "Image adds to the tweet meaning" is employed and the classification achieves excellent performance in our model. The RIVA model is trained on a large multimodal corpus of tweets under a multitask framework of text-image relation classification and next word prediction. In the experiments, we show the quantitative results of the impact of text-unrelated visual attention features on NER task in the ablation study, -1.2% on the Fudan Univ. dataset and -1.5% on the Snap Res. dataset. We illustrate the failed visual attention examples that can be resolved by the RIVA model. The performance of the RIVA model is better than other visual-linguistic models and SOTA performance of MNER is achieved in this paper.

Acknowledgment

We would like to thank the anonymous reviewers for their valuable comments. This work was supported by the National Innovation and Entrepreneurship Training Program for College Students under Grant 202013021005 and in part by the National Natural Science Foundation of China under Grant 62072402 and 61802343.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019. Pooled contextualized embeddings for named entity recognition. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page 724–728.

O Arshad, I Gallo, S Nawaz, and A Calefati. 2019. Aiding intra-text representations with visual context for multimodal named entity recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 337–342.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

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.

Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Hawre Hosseini. 2019. Implicit entity recognition, classification and linking in tweets. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1448–1448.

Yuting Hu, Liang Zheng, Yi Yang, and Yongfeng Huang. 2017. Twitter100k: A real-world dataset for weakly supervised cross-media retrieval. *IEEE Transactions on Multimedia*, 20(4):927–938.

John D Lafferty, Andrew McCallum, and Fernando CN 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.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of NAACL-HLT*, pages 260–270. Association for Computational Linguistics.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

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.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on EMNLP (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July. Association for Computational Linguistics.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Alakananda Vempala and Daniel Preoţiuc-Pietro. 2019. Categorizing and inferring the relationship between the text and image of twitter posts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2830–2840.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Rui Wang, ZHOU Deyu, and Yulan He. 2019. Open event extraction from online text using a generative adversarial network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 282–291.

I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. 2019. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*.

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