

RpBERT: A Text-image Relation Propagation-based BERT Model for Multimodal NER

Lin Sun^{1*}, Jiquan Wang^{2,1*}, Kai Zhang³, Yindu Su^{2,1}, Fangsheng Weng¹

¹ Department of Computer Science, Zhejiang University City College, Hangzhou, China

² College of Computer Science and Technology, Zhejiang University, Hangzhou, China

³ Department of Computer Science and Technology, Tsinghua University, Beijing, China
sunl@zucc.edu.cn, wangjiquan@zju.edu.cn, drogozhang@gmail.com, yindusu@zju.edu.cn

Abstract

Recently multimodal named entity recognition (MNER) has utilized images to improve the accuracy of NER in tweets. However, most of the multimodal methods use attention mechanisms to extract visual clues regardless of whether the text and image are relevant. Practically, the irrelevant text-image pairs account for a large proportion in tweets. The visual clues that are unrelated to the texts will exert uncertain or even negative effects on multimodal model learning. In this paper, we introduce a method of text-image relation propagation into the multimodal BERT model. We integrate soft or hard gates to select visual clues and propose a multitask algorithm to train on the MNER datasets. In the experiments, we deeply analyze the changes in visual attention before and after the use of text-image relation propagation. Our model achieves state-of-the-art performance on the MNER datasets. The source code is available online¹.

Introduction

Social media platforms such as Twitter have become part of the everyday lives of many people. They are important sources for various information extraction applications such as open event extraction (Wang, Deyu, and He 2019) and social knowledge graph construction (Hosseini 2019). As a key component of these applications, named entity recognition (NER) aims to detect named entities (NEs) and classify them into predefined types, such as person (PER), location (LOC) and organization (ORG). Recent works on tweets based on multimodal learning have been increasing (Moon, Neves, and Carvalho 2018; Lu et al. 2018; Zhang et al. 2018; Arshad et al. 2019; Yu et al. 2020). These researchers investigated to enhance linguistic representations with the aid of visual clues in tweets. Most of the 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). For example, Fig. 1(a) shows a successful visual attention example in (Lu et al. 2018). In fact, texts and images in tweets could also be irrelevant. Vempala and Preotiuc-Pietro (2019) categorized text-image relations according to whether the “Image adds to the tweet meaning”. The “Image does not add



(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 of MNER from (Lu et al. 2018). The left column is a tweet’s image and the right column is its corresponding attention visualization. (a) Successful case, (b) failure case.

to the tweet meaning” type accounts for approximately 56% of instances in Vempala’s text-image relation classification (TRC) dataset. In addition, we trained a classifier of whether the “Image adds to the tweet meaning” on a large randomly collected corpus, Twitter100k (Hu et al. 2017), and the proportion of classified negatives was approximately 60%. The attention-based models would also produce visual attention although the text and image are irrelevant, and such visual attention might exert negative effects on the text inference. Fig. 1(b) shows a failure visual attention example. The visual attention focuses on the wall and ground, resulting in tagging “[ORG] Cleveland” with the wrong label “LOC”.

In this paper, we consider inferring the text-image relation to address the problem of inappropriate visual attention clues in multimodal models. The contributions of this paper can be summarized as follows:

- We propose a novel text-image relation propagation-based multimodal BERT model. We investigate the soft and hard ways of propagating text-image relations through the model by training. A training procedure for

*Equal contribution.

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹<https://github.com/Multimodal-NER/RpBERT>

the multiple tasks of text-image relation classification and downstream NER is also presented.

- We provide insights into the visual attention by numerical distributions and heat maps. Text-image relation propagation can not only reduce the interference from irrelevant images but also leverage more visual information for relevant text-image pairs.
- The experimental results show that the failure cases in the related works are correctly recognized by our model, and the state-of-the-art performance is achieved in this paper.

Related Work

Multimodal NER Moon et al. (2018) proposed a modality-attention module at the input of an 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 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 into 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 determine whether the fusion feature was helpful in improving 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 was a weighted sum of the 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 the attention-guided models is that the extracted visual contextual cues do not match the text for irrelevant text-image pairs. The authors of (Lu et al. 2018; Arshad et al. 2019) showed failed examples in which unrelated images provided misleading visual attention and yielded prediction errors.

Pretrained multimodal BERT The pretrained model BERT has achieved great success in natural language processing (NLP). The latest presented visual-linguistic models based on the BERT architecture include VL-BERT (Su et al. 2019), ViLBERT (Lu et al. 2019), VisualBERT (Li et al. 2019), UNITER (Chen et al. 2020), LXMERT (Tan and Bansal 2019), and Unicoder-VL (Li et al. 2020). We summarize and compare the existing visual-linguistic BERT models in three aspects as follows: 1) **Architecture**. The structures of Unicoder-VL, VisualBERT, VL-BERT, and UNITER were the same as that of vanilla BERT. The image and text tokens were combined into a sequence and fed into BERT to learn contextual embeddings. LXMERT and ViLBERT separated visual and language processing into

two streams that interacted through cross-modality or co-attentional transformer layers respectively. 2) **Visual representations**. The image features could be represented as region-of-interest (RoI) or block regions. All the above pretrained models used Fast R-CNN (Girshick 2015) to detect objects and pool RoI features. The purpose of RoI detection is to reduce the complexity of visual information and perform the task of masked region classification with linguistic clues (Su et al. 2019; Li et al. 2020). However, for the irrelevant text-image pairs, the non-useful and salient visual features could increase the interference with the linguistic features. Moreover, object recognition categories are limited and many NEs have no corresponding object class, such as company trademark and scenic location. 3) **Pretraining tasks**. The models were trained on image caption datasets such as the COCO caption dataset (Chen et al. 2015) or Conceptual Captions (Sharma et al. 2018). The pretraining tasks mainly include masked language modeling (MLM), masked region classification (MRC) (Chen et al. 2020; Tan and Bansal 2019; Li et al. 2020; Su et al. 2019), and image-text matching (ITM) (Chen et al. 2020; Li et al. 2020; Lu et al. 2019). The ITM task is a binary classification, which defines the pairs in the caption dataset as positives and the pairs generated by replacing the image or text in a paired example with other randomly selected samples as negatives. It assumed that the text-image pairs in the caption datasets were highly related; however, this assumption could not be established in the text-image pairs of tweets.

Visual features are always directly concatenated with linguistic features (Yu and Jiang 2019) or extracted by attention weights in the latest multimodal models, regardless of whether the images contribute to the semantics of the texts, resulting in failed MNER examples shown in Table 7. Therefore, in this work, we explore a multimodal variant of BERT to perform MNER for tweets with different text-image relations.

The Proposed Approach

In this section, we introduce a text-image Relation propagation-based BERT model (RpBERT) for multimodal NER, which is shown in Fig. 2. We illustrate the RpBERT architecture and then describe its training procedure in detail.

Model Design

Our RpBERT extends vanilla BERT to a multitask framework of text-image relation classification and visual-linguistic learning for MNER. First, similar to most visual-linguistic BERTs, we adapt vanilla BERT to multimodal inputs. The input sequence of RpBERT is designed as follows:

$$[\text{CLS}] \underbrace{w_1 \dots w_n}_T [\text{SEP}] \underbrace{v_1 \dots v_m}_V, \quad (1)$$

where $[\text{CLS}]$ stands for text-image relation classification, $[\text{SEP}]$ stands for the separation between text and image features, $T=\{w_1, \dots, w_n\}$ denotes a sequence of linguistic features, and $V=\{v_1, \dots, v_m\}$ denotes a sequence of visual features. The word token sequence is generated by the

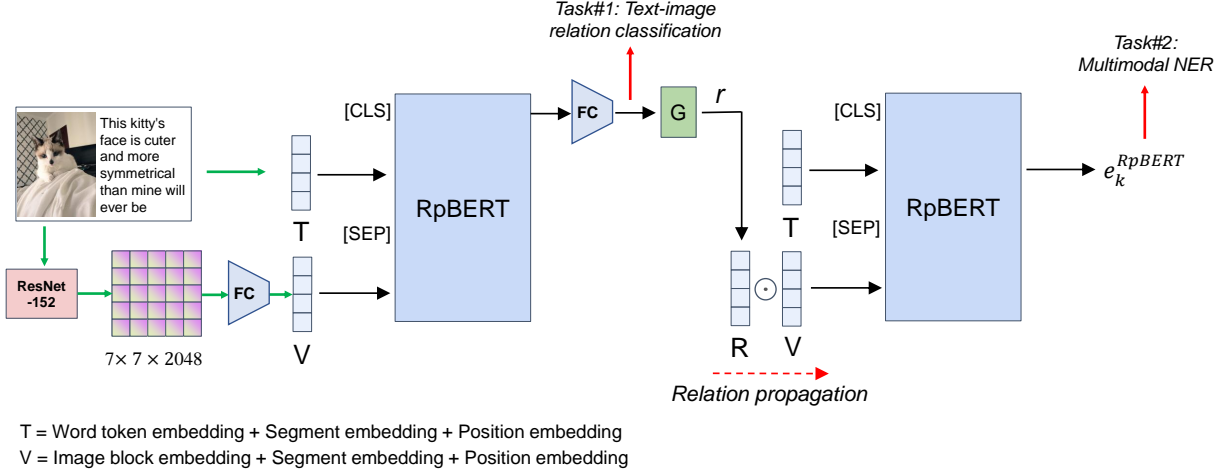


Figure 2: The RpBERT architecture overview. Two RpBERTs share the same structure and parameters.

BERT tokenizer, which breaks an unknown word into multiple word-piece tokens. Unlike the latest visual-linguistic BERT models (Su et al. 2019; Lu et al. 2019; Li et al. 2020), we represent visual features as block regions instead of RoIs. The visual features are extracted from the image by ResNet (He et al. 2016). The output size of the last convolutional layer in ResNet is $7 \times 7 \times d_v$, where 7×7 denotes 49 block regions in an image. The extracted features of block regions $\{f_{i,j}\}_{i,j=1}^7$ are arranged into an image block embedding sequence $\{b_1 = f_{1,1}W^v, \dots, b_{49} = f_{7,7}W^v\}$, where $f_{i,j} \in \mathbb{R}^{1 \times d_v}$ and $W^v \in \mathbb{R}^{d_v \times d_{BERT}}$ to match the embedding size of BERT, and $d_v = 2048$ when working with ResNet-152. Following the practice in BERT, the input embeddings of tokens are the sum of word token embeddings (or image block embeddings), segment embeddings, and position embeddings. The segment embeddings are learned from two types, where A denotes text tokens and B denotes image blocks. The position embeddings of word tokens are learned from the word order in the sentence, but all positions are the same for visual tokens.

The output of the token [CLS] is fed to a fully connected (FC) layer as a binary classifier for Task#1 of text-image relation classification. Additionally, we use the probability gate G shown in Fig. 2 to yield probabilities $[\pi_0, \pi_1]$. The text-image relevant score r is defined as the probability of being positive,

$$r = \pi_1. \quad (2)$$

We use the relevant score r to construct a visual mask matrix R in Fig. 2,

$$R = (x_{i,j} = r)_{49 \times d_{BERT}}. \quad (3)$$

The text-image relation is propagated to RpBERT via $R \odot V$, where \odot is the element-wise multiplication. For example, if $\pi_1 = 0$, then all visual features are discarded. Finally, e_k^{RpBERT} , the outputs of the tokens T with visual clues, are fed to the NER model for Task#2 training.

Relation Propagation

We investigate two kinds of relation propagation, soft and hard, by different probability gates G :

- **Soft relation propagation:** In soft relation propagation, the output of G can be viewed as a continuous distribution. The visual features are filtered according to the strength of the text-image relation. The gate G is defined as a softmax function:

$$G_s = \text{softmax}(x). \quad (4)$$

- **Hard relation propagation:** In hard relation propagation, the output of G can be viewed as a categorical distribution. The visual features are either selected or discarded based on 0 or 1. The gate G is defined as follows:

$$G_{h1} = [\text{softmax}(x) > 0.5], \quad (5)$$

where $[\cdot]$ is the Iverson bracket indicator function, which takes a value of 1 when its argument is true and 0 otherwise. As G_{h1} is not differentiable, an empirical way is to use a straight-through estimator (Bengio, Léonard, and Courville 2013) for propagating gradients back through the network. Besides, Jang et al. (2017) proposed Gumbel-Softmax to create a continuous approximation to the categorical distribution. Inspired by this, we define the gate G as Gumbel-Softmax in Eq. (6) for hard relation propagation.

$$G_{h2} = \text{softmax}((x + g)/\tau), \quad (6)$$

where g is a noise sampled from Gumbel distribution and τ is a temperature parameter. As the temperature approaches 0, samples from the Gumbel-Softmax distribution become one-hot and the Gumbel-Softmax distribution becomes identical to the categorical distribution. In the training stage, the temperature τ is annealed using the schedule of 1 to 0.1.

In the experimental results, we compare the performances of G_s , G_{h1} , and G_{h2} in Table 4.

Multitask Training for MNER

In this section, we present how to train RpBERT for MNER. The training procedure involves multitask learning of text-image relation classification and MNER, represented by solid red arrows in Fig. 2. The two tasks are described in detail as follows:

Task#1 Text-image relation classification (TRC): We employ the “Image Task” splits of the TRC dataset (Vempala and Preoțiuc-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 relations and statistics of the TRC dataset are shown in Table 1.

Let $\mathcal{D}_1 = \{a^{(i)}\}_{i=1}^N = \{< text^{(i)}, image^{(i)} >\}_{i=1}^N$ be a set of text-image pairs for TRC training. The loss \mathcal{L}_1 of binary relation classification is calculated by cross entropy:

$$\mathcal{L}_1 = - \sum_{i=1}^N \log(p(a^{(i)})), \quad (7)$$

where $p(x)$ is the probability for correct classification and is calculated by softmax.

Task#2 MNER via relation propagation: In this stage, we use the mask matrix R to control the additive visual clues. The input sequence of RpBERT is $[CLS] \ T \ [SEP] \ R \odot V$. We denote the output of T as e_k^{RpBERT} . To perform NER, we use biLSTM-CRF (Lample et al. 2016) as a baseline NER model. The biLSTM-CRF model consists of a bidirectional LSTM and conditional random fields (CRF) (Lafferty, McCallum, and Pereira 2001). The input e_k of biLSTM-CRF is a concatenation of word and character embeddings (Lample et al. 2016). CRF uses the biLSTM hidden vectors of each token to tag the sequence with entity labels. To evaluate the RpBERT model, we concatenate e_k^{RpBERT} as the input of biLSTM, i.e., $[e_k; e_k^{RpBERT}]$. For out-of-vocabulary (OOV) words, we average the outputs of BERT-tokenized subwords not only to generate an approximate vector but also to align the broken words with the input embeddings of biLSTM-CRF.

In biLSTM-CRF, named entity tagging is trained on a standard CRF model. We feed the hidden vectors $H = \{h_t = [\bar{h}_t^{LSTM}; \bar{h}_t^{LSTM}]\}_{t=1}^n$ of biLSTM to the CRF model. For a sequence of tags $y = \{y_1, \dots, y_n\}$, the probability of the label sequence y is defined as follows (Lample et al. 2016):

$$p(y|x) = \frac{e^{s(x,y)}}{\sum_{y' \in Y} e^{s(x,y')}}, \quad (8)$$

where Y is all possible tag sequences for the sentence x and $s(x, y)$ are feature functions modeling transitions and emissions. Details can be referred in (Lample et al. 2016). The objective of Task#2 is to minimize the negative log-likelihood over the training data $\mathcal{D}_2 = \{(x^{(i)}, y^{(i)})\}_{i=1}^M$:

$$\mathcal{L}_2 = - \sum_{i=1}^M \log(p(y^{(i)}|x^{(i)})). \quad (9)$$

Combining Task#1 and Task#2, the complete training procedure of RpBERT for MNER is illustrated in Algorithm 1.

Algorithm 1 Multitask training procedure of RpBERT for MNER.

Input: The TRC dataset and MNER dataset.

Output: θ_{RpBERT} , θ_{ResNet} , θ_{FCs} , θ_{biLSTM} , and θ_{CRF} .

```

1: for all epochs do
2:   for all batches in the TRC dataset do
3:     Forward text-image pairs through RpBERT;
4:     Compute loss  $\mathcal{L}_1$  by Eq. (7);
5:     Update  $\theta_{FCs}$  and finetune  $\theta_{RpBERT}$  and  $\theta_{ResNet}$  using  $\nabla \mathcal{L}_1$ ;
6:   end for
7:   for all batches in the MNER dataset do
8:     Forward text-image pairs through RpBERT;
9:     Compute the visual mask matrix  $R$ ;
10:    Forward text-image pairs with relation propagation through RpBERT and biLSTM-CRF;
11:    Compute loss  $\mathcal{L}_2$  by Eq. (9);
12:    Update  $\theta_{biLSTM}$  and  $\theta_{CRF}$  and finetune  $\theta_{RpBERT}$  and  $\theta_{ResNet}$  using  $\nabla \mathcal{L}_2$ ;
13:   end for
14: end for

```

θ_{RpBERT} , θ_{ResNet} , θ_{FCs} , θ_{biLSTM} , and θ_{CRF} represent the parameters of RpBERT, ResNet, FCs, biLSTM, and CRF, respectively. In each epoch, the procedure first performs Task#1 to train the text-image relation on the TRC dataset and then performs Task#2 to train the model on MNER dataset. In the test stage, we execute lines 8-10 of Algorithm 1 and decode the valid sequence of labels using Viterbi algorithm (Lafferty, McCallum, and Pereira 2001).

Experiments

Datasets

In the experiments, we use three datasets to evaluate the performance. One is the TRC dataset, and the other two are MNER datasets of Fudan University and Snap Research. The detailed descriptions are as follows:

• TRC dataset of Bloomberg LP (Vempala and Preoțiuc-Pietro 2019)

In this dataset, the authors annotated tweets into four types of text-image relation, as shown in Table 1. “Image adds to the tweet meaning” is centered on the role of the image to the semantics of the tweet while “Text is presented in image” focuses on the text’s role. In the RpBERT model, we treat the text-image relation for the image’s role 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 Preoțiuc-Pietro 2019). We use this dataset to perform learning Task#1 of RpBERT.

	R_1	R_2	R_3	R_4
Image adds to the tweet meaning	✓	✓	×	×
Text is presented in image	✓	×	✓	×
Percentage (%)	18.5	25.6	21.9	33.8

Table 1: Four relation types in the TRC dataset.

- **MNER dataset of Fudan University (Zhang et al. 2018)**

The authors sampled the tweets with images collected through Twitter’s API. In this dataset, the NE types are Person, Location, Organization, and Misc. The authors labeled 8,257 tweet texts using the BIO2 tagging scheme and used a 4,000/1,000/3,257 train/dev/test split.

- **MNER dataset of Snap Research (Lu et al. 2018)**

The authors collected the data from Twitter and Snapchat, but Snapchat data are not available for public use. The 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 the BIO tagging scheme and used a 4,817/1,032/1,033 train/dev/test split.

Settings

We use the 300-dimensional fastText Crawl (Mikolov et al. 2018) word vectors in biLSTM-CRF. All images are reshaped to a size of 224×224 to match the input size of ResNet. We use ResNet-152 to extract visual features and finetune it with a learning rate of $1e-6$. The FC layers in our model are a linear neural network followed by ReLU activation. The architecture of RpBERT is the same as that of BERT-Base, and we load the pretrained weights from BERT-base-uncased model to initialize our RpBERT model. We train the model using Adam (Kingma and Ba 2014) optimizer with default settings. Table 2 shows the hyperparameter values in the RpBERT and biLSTM-CRF models. We use F1 score as evaluation metric for TRC and MNER.

Hyperparameter	Value
LSTM hidden state size	256
+RpBERT	1024
LSTM layer	2
mini-batch size	8
char embedding dimension	25
optimizer	Adam
learning rate	$1e-4$
learning rate for finetuning RpBERT and ResNet	$1e-6$
dropout rate	0.5

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

Result of TRC

Table 3 shows the performance of RpBERT in text-image relation classification on the test set of the TRC data. In terms of the network structure, Lu et al. (2018) represented the multimodal feature as a concatenation of linguistic features from LSTM and visual features from InceptionNet (Szegedy et al. 2015). The result shows that the BERT-based visual-linguistic model significantly outperforms that of Lu et al. (2018), and F1 score of RpBERT on the test set of the TRC data increases by 7.1% compared to Lu et al. (2018).

Results of MNER

Table 4 illustrates the improved performance by visual clues, such as biLSTM-CRF vs. biLSTM-CRF with image and

	Lu et al. (2018)	RpBERT
F1 score	81.0	(+7.1) 88.1

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

	Fudan Univ.	Snap Res.
biLSTM-CRF	(+0.0) 69.9	(+0.0) 80.1
biLSTM-CRF w/ image at $t = 0$	(+0.1) 70.0	(+0.5) 80.6
Zhang et al. (2018)	(+0.8) 70.7	-
Lu et al. (2018)	-	(+0.6) 80.7
biLSTM-CRF + BERT	(+0.0) 71.6	(+0.0) 85.1
biLSTM-CRF + RpBERT _{G_{h1}}	(+2.2) 73.8	(+1.3) 86.4
biLSTM-CRF + RpBERT _{G_{h2}}	(+2.6) 74.2	(+1.5) 86.6
biLSTM-CRF + RpBERT _{G_s}	(+2.8) 74.4	(+2.3) 87.4

Table 4: Comparison of the improved performance by visual clues in F1 score (%).

BERT vs. RpBERT. The inputs of “biLSTM-CRF” and “biLSTM-CRF + BERT” are text only, while those of other models are text-image pairs. “biLSTM-CRF w/ image at $t = 0$ ” means that the image feature is placed at the beginning of LSTM before the word sequence, similar to the model in (Vinyals et al. 2015). “biLSTM-CRF + RpBERT” means that the contextual embeddings e_k^{RpBERT} with visual clues are concatenated as the input of biLSTM-CRF, as clarified in the section of “Multitask Training for MNER”. The results show that the best “+ RpBERT_{G_s}” achieves increases of 4.5% and 7.3% compared to “biLSTM-CRF” on the Fudan Univ. and Snap Res. datasets, respectively. In terms of the role of visual features, the increase of “+ RpBERT_{G_s}” achieves approximately 2.5% compared to “+ BERT”, which is larger than those of the biLSTM-CRF based multimodal models such as Zhang et al. (2018) and Lu et al. (2018) compared to biLSTM-CRF. This indicates that the RpBERT model can better leverage visual features to enhance the context of tweets.

In Table 5, we compare performance with the state-of-the-art method (Yu et al. 2020) and visual-linguistic pretrained models which codes are available, such as VL-BERT (Su et al. 2019), ViLBERT (Lu et al. 2019), and UNITER (Chen et al. 2020). Similar to e_k^{RpBERT} in RpBERT, we take out the contextual embeddings of word sequence in visual-linguistic models and concatenate them with the token embeddings e_k as the input embedding of biLSTM-CRF. For example, “biLSTM-CRF + VL-BERT” means that the output of word sequence in VL-BERT is concatenated as the input of biLSTM-CRF, i.e., $[e_k; e_k^{VL-BERT}]$. The results show that RpBERT_{G_s} outperforms all pretrained models. Additionally, we test RpBERT using the structure of BERT-Large, which has 24 layers and 16 attention heads. “biLSTM-CRF + RpBERT-Large_{G_s}” achieves state-of-the-art results on the MNER datasets and outperforms the current best results (Yu et al. 2020) by 1.5% on the Fudan Univ. dataset and 2.5% on the Snap Res. dataset.

	Fudan Univ.			Snap Res.		
	Image adds	Image doesn't add	Overall	Image adds	Image doesn't add	Overall
biLSTM-CRF + RpBERT _{G_s}	74.6	74.1	74.4	87.7	86.9	87.4
- w/o Rp	(-0.5) 74.1	(-3.1) 71.0	(-1.8) 72.6	(-0.7) 87.0	(-2.3) 84.6	(-1.2) 86.2

Table 6: Performance comparison in F1 score (%) when the relation propagation (Rp) is ablated.

	Fudan Univ.	Snap Res.
Arshad et al. (2019)	72.9	-
Yu et al. (2020)	73.4	85.3
biLSTM-CRF + VL-BERT	72.4	86.0
biLSTM-CRF + ViLBERT	72.0	85.7
biLSTM-CRF + UNITER	72.7	86.1
biLSTM-CRF + RpBERT _{G_s}	74.4	87.4
biLSTM-CRF + RpBERT-Large _{G_s}	74.9	87.8

Table 5: Performance comparison with other models in F1 score (%).

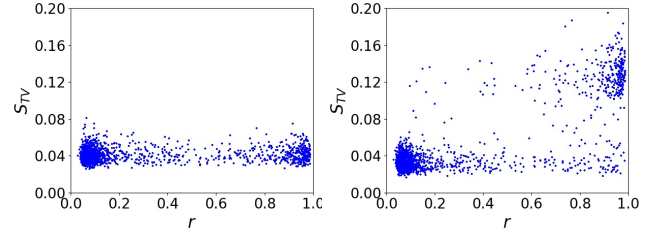
Ablation Study

In this section, we report the results when ablating the relation propagation in RpBERT, or equivalently performing only Task#2 in the training of RpBERT. Table 6 shows that the overall performance without relation propagation (“w/o Rp”) decreases by -1.8% and -1.2% on the Fudan Univ. and Snap Res. datasets, respectively. In addition, we divide the test data into two sets, “Image adds” and “Image doesn’t add”, by the text-image relation classification, and compare the impact of the ablation on the data of different relation types. The performances on all relation types are improved with relation propagation. More importantly, regarding the “Image doesn’t add” type, “w/o Rp” lowers the F1 scores by a large margin, -3.1% on the Fudan Univ. dataset and -2.3% on the Snap Res. dataset. This justifies that the text-unrelated visual features exert large negative effects on learning visual-linguistic representations.

In Fig. 3, we illustrate the comparison of RpBERT and RpBERT w/o Rp in terms of the numerical distribution between the relevant score r and S_{TV} , where S_{TV} is the average sum of visual attentions and is defined as follows:

$$S_{TV} = \frac{1}{LH} \sum_{l=1}^L \sum_{h=1}^H \sum_{i=1}^n \sum_{j=1}^m Att^{(l,h)}(w_i, v_j), \quad (10)$$

where $Att^{(l,h)}(w_i, v_j)$ is the attention between the i th word and j th image block on the h th head and l th layer in RpBERT. The samples are from the test set of the Snap Res. dataset. In Fig. 3(a), we find that the distribution of S_{TV} of RpBERT w/o Rp is close to a horizontal line and is unrelated to the relevant score r . In Fig. 3(b), most S_{TV} values of RpBERT decrease on irrelevant text-image pairs ($r < 0.5$) and increase on relevant text-image pairs ($r > 0.5$) compared to those of RpBERT w/o Rp. Quantitatively, the mean of S_{TV} decreases by 20% from 0.041 to 0.034 on irrelevant text-image pairs while it increases from 0.042 to 0.102 on relevant text-image pairs. In general, after using relation propagation, the trend is towards leveraging more visual cues in stronger text-image relations.



(a) RpBERT w/o Rp

(b) RpBERT_{G_s}

Figure 3: The numerical distribution between r and S_{TV} .

Case Study via Attention Visualization

We illustrate five failure examples mentioned in (Lu et al. 2018; Arshad et al. 2019; Yu et al. 2020) in Table 7. The common reason for these failed examples is inappropriate visual attention features. The table shows the relevant score r and overall image attentions of RpBERT and RpBERT w/o Rp. The visual attention of an image block j across all words, heads and layers is defined as follows:

$$a_j^v = \frac{1}{LH} \sum_{l=1}^L \sum_{h=1}^H \sum_{i=1}^n Att^{(l,h)}(w_i, v_j). \quad (11)$$

We visualize the overall image attentions $\{a_j^v\}_{j=1}^{49}$ by heat maps. The NER results of “+ RpBERT w/o Rp”, “+ RpBERT_{G_s}”, and the previous works are also presented for comparison.

Examples 1 and 2 are from the Snap Res. dataset, and Examples 3, 4, and 5 are from the Fudan Univ. dataset. The NER results of all examples obtained by RpBERT are correct. In Example 1, RpBERT performs correct and the visual attentions have no negative effects on the NER results. In Example 2, the visual attentions focus on the ground and result in tagging “Cleveland” with the wrong label “LOC”. In Example 3, “Reddit” is misidentified as “ORG” by the visual attentions. In Example 5, “Siri” is wrongly identified as “PER” because of the visual attentions to the human face. In Examples 2, 3, and 5, the text-image pairs are recognized as irrelevant since the values of r are small. With text-image relation propagation, much less visual features are weighted to the linguistic features in RpBERT and the NER results are correct. In Example 4, the text and image are related, i.e., $r = 0.74$. The persons are significantly concerned in (Arshad et al. 2019), resulting in the wrong label “PER” for “Mount Sherman”. RpBERT w/o Rp extends some visual attention to the mountain scene, while RpBERT increases much more visual attention to the scenery, such as sky and mountain, and thus strengthens the understanding

	1	2	3	4	5
Image					
$\{a_j^v\}_{j=1}^{49}$ of RpBERT w/o Rp					
$\{a_j^v\}_{j=1}^{49}$ of RpBERT _{G_s}					
r	0.14	0.13	0.24	0.74	0.20
+ RpBERT w/o Rp	Looking forward to editing some [ORG SBU] baseball shots from Saturday.	Nice image of [PER Kevin Love] and [PER Kyle Korver] during 1 st half # NBAFinals # CavsIn9 # [LOC Cleveland]	[ORG Reddit] needs to stop pre-tending racism is valuable debate.	[MISC PSD] Leshner teachers take school spirit to top of 14ner [LOC Mount Sherman] .	Ask [PER Siri] what 0 divided by 0 is and watch her put you in your place.
+ RpBERT _{G_s}	Looking forward to editing some [ORG SBU] baseball shots from Saturday.	Nice image of [PER Kevin Love] and [PER Kyle Korver] during 1 st half # NBAFinals # CavsIn9 # [ORG Cleveland]	[MISC Reddit] needs to stop pre-tending racism is valuable debate.	[ORG PSD] Leshner teachers take school spirit to top of 14ner [LOC Mount Sherman] .	Ask [MISC Siri] what 0 divided by 0 is and watch her put you in your place.
Previous work	Looking forward to editing some SBU baseball shots from Saturday. (Lu et al. 2018)	Nice image of [PER Kevin Love] and [PER Kyle Korver] during 1 st half # NBAFinals # CavsIn9 # [LOC Cleveland] . (Lu et al. 2018)	[ORG Reddit] needs to stop pre-tending racism is valuable debate. (Arshad et al. 2019)	[ORG PSD] Leshner teachers take school spirit to top of 14ner [PER Mount Sherman] . (Arshad et al. 2019)	Ask [PER Siri] what 0 divided by 0 is and watch her put you in your place. (Yu et al. 2020)

Low  High

Table 7: Five failed examples in the previous works tested by “+ RpBERT_{G_s}” and “+ RpBERT w/o Rp”. Blue and black labels are correct and red ones are wrong.

of the whole picture and yields the correct labels of “PSD Leshner” and “Mount Sherman”.

Conclusion

This paper concerns the visual attention problem raised by the text-unrelated images in tweets for multimodal learning. We propose a relation propagation-based multimodal model based on text-image relation inference. The model is trained by the multiple tasks of text-image relation classification and downstream NER. In experiments, the ablation study quantitatively evaluates the role of text-image relation propaga-

tion. The heat map visualization and numerical distribution regarding the visual attention justify that RpBERT can better leverage visual information adaptively according to the relation between text and image. The failed cases mentioned in other papers are effectively resolved by the RpBERT model. Our model achieves the best F1 scores in both TRC and MNER, i.e., 88.1% on the TRC dataset, 74.9% on the Fudan Univ. dataset, and 87.8% on the Snap Res. dataset.

Acknowledgements

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 (NSFC) under Grant 62072402.

References

- Arshad, O.; Gallo, I.; Nawaz, S.; and Calefati, A. 2019. Aiding Intra-Text Representations with Visual Context for Multimodal Named Entity Recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 337–342.
- Bengio, Y.; Léonard, N.; and Courville, A. C. 2013. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. *arXiv preprint arXiv:1308.3432*.
- Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.
- Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *Computer Vision – ECCV 2020*, 104–120.
- Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Hosseini, H. 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*, 1448–1448.
- Hu, Y.; Zheng, L.; Yang, Y.; and Huang, Y. 2017. Twitter100k: A real-world dataset for weakly supervised cross-media retrieval. *IEEE Transactions on Multimedia* 20(4): 927–938.
- Jang, E.; Gu, S.; and Poole, B. 2017. Categorical Reparameterization with Gumbel-Softmax. In *International conference on learning representations*.
- Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Lafferty, J. D.; McCallum, A.; and Pereira, F. C. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 282–289.
- Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; and Dyer, C. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of NAACL-HLT*, 260–270. Association for Computational Linguistics.
- Li, G.; Duan, N.; Fang, Y.; Gong, M.; Jiang, D.; and Zhou, M. 2020. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training. In *AAAI*, 11336–11344.
- Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.
- Lu, D.; Neves, L.; Carvalho, V.; Zhang, N.; and Ji, H. 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)*, 1990–1999.
- Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, 13–23.
- Mikolov, T.; Grave, E.; Bojanowski, P.; Puhersch, C.; and Joulin, A. 2018. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.
- Moon, S.; Neves, L.; and Carvalho, V. 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)*, 852–860.
- Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 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)*, 2556–2565. Melbourne, Australia: Association for Computational Linguistics.
- Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2019. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *International Conference on Learning Representations*.
- Szegedy, C.; Wei Liu; Yangqing Jia; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9.
- Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. 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)*, 5103–5114.
- Vempala, A.; and Preotiuc-Pietro, D. 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*, 2830–2840.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.

Wang, R.; Deyu, Z.; and He, Y. 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)*, 282–291.

Yu, J.; and Jiang, J. 2019. Adapting BERT for target-oriented multimodal sentiment classification. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 5408–5414. AAAI Press.

Yu, J.; Jiang, J.; Yang, L.; and Xia, R. 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*, 3342–3352.

Zhang, Q.; Fu, J.; Liu, X.; and Huang, X. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Thirty-Second AAAI Conference on Artificial Intelligence*.