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

Lin Sun$^1$, Jiquan Wang$^{2,1}$*, Kai Zhang$^3$, Yindu Su$^{2,1}$, Fangsheng Weng$^1$
$^1$ 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, 31801131@stu.zucc.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.

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 Preoțiuc-Pietro (2019) categorized text-image relations according to whether the “Image adds to tweet meaning”. The “Image does not add to 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 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 the multiple tasks of text-image relation classification and downstream NER is also presented.

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.

(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].
- 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 the same as that of vanilla BERT. The image and text were 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), 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:

$$[CLS] \overline{w_1 \ldots w_n} [SEP] \overline{v_1 \ldots v_m},$$

where $[CLS]$ stands for text-image relation classification, $[SEP]$ stands for the separation between text and image features, $T={w_1, \ldots, w_n}$ denotes a sequence of linguistic features, and $V={v_1, \ldots, v_m}$ denotes a sequence of visual features. The word token sequence is generated by the 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.
In this section, we present how to train RpBERT for MNER. Multitask Training for MNER

In the training stage, the temperature is annealed using the schedule of 1 to 0.1. The gate $G$ is defined as follows:

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

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.

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

where $g$ is a noise sampled from Gumbel distribution and $\tau$ 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 $\tau$ 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 leaning of text-image relation classification and MNER, represented by

- **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:

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

- **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:

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

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.

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

where $g$ is a noise sampled from Gumbel distribution and $\tau$ 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 $\tau$ 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.
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 Preotu-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 \( D_1 = \{a^{(i)}\}_{i=1}^N = \{<text^{(i)}, image^{(i)}><\}_{i=1}^N \) be a set of text-image pairs for TRC training. The loss \( L_1 \) of binary relation classification is calculated by cross entropy:

\[
L_1 = -\sum_{i=1}^N \log(p(a^{(i)})), \tag{7}
\]

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 \( \otimes \) V. We denote the output of T as \( e_k^{R\text{BERT}} \). 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 \( e_k^{R\text{BERT}} \) as the input of biLSTM, i.e., \([e_k; e_k^{R\text{BERT}}]\). 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 \( H = \{h_t = [h_t^{\text{LSTM}}, h_t^{\text{LSTM}}]\}_{t=1}^Y \) of biLSTM to the CRF model. For a sequence of tags \( y = \{y_1, \ldots, y_n\} \), the probability of the label sequence \( y \) is defined as follows (Lample et al. 2016):

\[
p(y|x) = \frac{e^{c(y|x)}}{\sum_{y' \in Y} e^{c(y'|x)}}, \tag{8}
\]

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 \( D_2 = \{(x^{(i)}, y^{(i)})\}_{i=1}^M \):

\[
L_2 = -\sum_{i=1}^M \log(p(y^{(i)}|x^{(i)})) \tag{9}
\]

Combining Task#1 and Task#2, the complete training procedure of RpBERT for MNER is illustrated in Algorithm 1. \( \theta_{R\text{BERT}}, \theta_{R\text{Net}}, \theta_{FCs}, \theta_{bi\text{LSTM}}, \) and \( \theta_{CRF} \) 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 Preotu-Pietro 2019)**

In this dataset, the authors annotated tweets into four types of text-image relation, as shown in Table 1. “Image adds to 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 \( R_1 \cap R_2 \) and \( R_3 \cup R_4 \). We follow the same split of 8:2 for train/test sets as in (Vempala and Preotu-Pietro 2019). We use this dataset to perform learning Task#1 of RpBERT.

<table>
<thead>
<tr>
<th>Image adds to tweet meaning</th>
<th>Text is presented in image</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \checkmark )</td>
<td>( \checkmark )</td>
<td>18.5</td>
</tr>
<tr>
<td>( \times )</td>
<td>( \times )</td>
<td>25.6</td>
</tr>
<tr>
<td>( \times )</td>
<td>( \checkmark )</td>
<td>21.9</td>
</tr>
<tr>
<td>( \checkmark )</td>
<td>( \times )</td>
<td>33.8</td>
</tr>
</tbody>
</table>

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 $224 \times 224$ 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.

### 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 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 $e_k^{R_{PBERT}}$ 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 2.3% 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 $e_k^{R_{PBERT}}$ 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.
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 text 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:

$$S_{TV} = \frac{1}{LH} \sum_{l=1}^{L} \sum_{h=1}^{H} \sum_{i=1}^{n_l} \sum_{j=1}^{m_i} \text{Att}^{l,h}(w_i, v_j),$$

(10)

where $\text{Att}^{l,h}(w_i, v_j)$ 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 test set from 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 adding relation propagation, the trend is towards leveraging more visual cues in stronger text-image relations.

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Fudan Univ.</th>
<th>Snap Res.</th>
</tr>
</thead>
<tbody>
<tr>
<td>biLSTM-CRF + RpBERT$_{G_a}$</td>
<td>74.6</td>
<td>87.4</td>
</tr>
<tr>
<td>- w/o Rp</td>
<td>-0.5</td>
<td>-1.8</td>
</tr>
<tr>
<td>biLSTM-CRF + ViLBERT</td>
<td>74.1</td>
<td>86.9</td>
</tr>
<tr>
<td>biLSTM-CRF + UNITER</td>
<td>72.7</td>
<td>86.1</td>
</tr>
<tr>
<td>biLSTM-CRF + BERT-Large</td>
<td>72.4</td>
<td>86.3</td>
</tr>
<tr>
<td>biLSTM-CRF + RpBERT$_{G_a}$</td>
<td>74.4</td>
<td>87.4</td>
</tr>
<tr>
<td>biLSTM-CRF + RpBERT-Large$_{G_a}$</td>
<td>74.9</td>
<td>87.8</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Relation Type</th>
<th>Fudan Univ.</th>
<th>Snap Res.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image adds</td>
<td>74.1</td>
<td>87.0</td>
</tr>
<tr>
<td>Image doesn’t add</td>
<td>71.0</td>
<td>86.9</td>
</tr>
<tr>
<td>Overall</td>
<td>74.4</td>
<td>87.4</td>
</tr>
</tbody>
</table>

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:

$$a_{ij}^{v} = \frac{1}{LH} \sum_{l=1}^{L} \sum_{h=1}^{H} \sum_{i=1}^{n_l} \sum_{j=1}^{m_i} \text{Att}^{l,h}(w_i, v_j).$$

(11)

We visualize the overall image attentions $\{a_{ij}^{v}\}_{j=1}^{19}$ by heat maps. The NER results of “+ RpBERT w/o Rp”, “+ RpBERT$_{G_a}$”, 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.


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.

<table>
<thead>
<tr>
<th>Image</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>{a^v_i}_{i=1}^{49} of RpBERT w/o Rp</td>
<td><img src="image1.png" alt="Image 1" /></td>
<td><img src="image2.png" alt="Image 2" /></td>
<td><img src="image3.png" alt="Image 3" /></td>
<td><img src="image4.png" alt="Image 4" /></td>
<td><img src="image5.png" alt="Image 5" /></td>
</tr>
<tr>
<td>{a^v_i}<em>{i=1}^{49} of RpBERT$</em>{G_s}$</td>
<td><img src="heatmap1.png" alt="Heatmap 1" /></td>
<td><img src="heatmap2.png" alt="Heatmap 2" /></td>
<td><img src="heatmap3.png" alt="Heatmap 3" /></td>
<td><img src="heatmap4.png" alt="Heatmap 4" /></td>
<td><img src="heatmap5.png" alt="Heatmap 5" /></td>
</tr>
<tr>
<td>r</td>
<td>0.14</td>
<td>0.13</td>
<td>0.24</td>
<td>0.74</td>
<td>0.20</td>
</tr>
<tr>
<td>Previous work</td>
<td>Looking forward to editing some [<strong>ORG SBU</strong>] baseball shots from Saturday.</td>
<td>Nice image of [<strong>PER Kevin Love</strong>] and [<strong>PER Kyle Korver</strong>] during 1 st half # NBAFinals # CavsIn9 # [<strong>LOC Cleveland</strong>].</td>
<td>[<strong>ORG Reddit</strong>] needs to stop pretending racism is valuable debate.</td>
<td>[<strong>MISC Reddit</strong>] teachers take school spirit to top of 14ner [<strong>LOC Mount Sherman</strong>].</td>
<td>Ask [<strong>PER Siri</strong>] what 0 divided by 0 is and watch her put you in your place.</td>
</tr>
<tr>
<td>+ RpBERT w/o Rp</td>
<td>Looking forward to editing some [<strong>ORG SBU</strong>] baseball shots from Saturday.</td>
<td>Nice image of [<strong>PER Kevin Love</strong>] and [<strong>PER Kyle Korver</strong>] during 1 st half # NBAFinals # CavsIn9 # [<strong>ORG Cleveland</strong>].</td>
<td>[<strong>MISC Reddit</strong>] needs to stop pretending racism is valuable debate.</td>
<td>[<strong>ORG PSD Lesher</strong>] teachers take school spirit to top of 14ner [<strong>LOC Mount Sherman</strong>].</td>
<td>Ask [<strong>MISC Siri</strong>] what 0 divided by 0 is and watch her put you in your place.</td>
</tr>
<tr>
<td>+ RpBERT$_{G_s}$</td>
<td>Looking forward to editing some [<strong>ORG SBU</strong>] baseball shots from Saturday.</td>
<td>Nice image of [<strong>PER Kevin Love</strong>] and [<strong>PER Kyle Korver</strong>] during 1 st half # NBAFinals # CavsIn9 # [<strong>ORG Cleveland</strong>].</td>
<td>[<strong>ORG Reddit</strong>] needs to stop pretending racism is valuable debate.</td>
<td>[<strong>ORG PSD Lesher</strong>] teachers take school spirit to top of 14ner [<strong>LOC Mount Sherman</strong>].</td>
<td>Ask [<strong>PER Siri</strong>] what 0 divided by 0 is and watch her put you in your place.</td>
</tr>
</tbody>
</table>

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 Lesher**” 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 propagation. 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


