ITA: Image-Text Alignments for Multi-Modal Named Entity Recognition

Xinyu Wang‡, Min Gui‡, Yong Jiang∥, Zixia Jia‡, Nguyen Bach‡, Tao Wang,
Zhongqiang Huang‡, Fei Huang‡, Kewei Tu‡∥

‡School of Information Science and Technology, ShanghaiTech University
∥Shanghai Engineering Research Center of Intelligent Vision and Imaging

Abstract

Recently, Multi-modal Named Entity Recognition (MNER) has attracted a lot of attention. Most of the work utilizes image information through region-level visual representations obtained from a pretrained object detector and relies on an attention mechanism to model the interactions between image and text representations. However, it is difficult to model such interactions as image and text representations are trained separately on the data of their respective modality and are not aligned in the same space. As text representations take the most important role in MNER, in this paper, we propose Image-text Alignments (ITA) to align image features into the textual space, so that the attention mechanism in transformer-based pretrained textual embeddings can be better utilized. ITA first aligns the image into regional object tags, image-level captions and optical characters as visual contexts, concatenates them with the input texts as a new cross-modal input, and then feeds it into a pretrained textual embedding model. This makes it easier for the attention module of a pretrained textual embedding model to model the interaction between the two modalities since they are both represented in the textual space. ITA further aligns the output distributions predicted from the cross-modal input and textual input views so that the MNER model can be more practical in dealing with text-only inputs and robust to noises from images. In our experiments, we show that ITA models can achieve state-of-the-art accuracy on multi-modal Named Entity Recognition datasets, even without image information.1

1: Yong Jiang and Kewei Tu are the corresponding authors.
2: This work was done when Xinyu Wang, Min Gui and Nguyen Bach were at Alibaba Group.
3: Our code is publicly available at https://github.com/Alibaba-NLP/KB-NER/ITA.

1 Introduction

Named Entity Recognition (NER) (Sundheim, 1995) has attracted increasing attention in natural language processing community. It has been applied to a lot of domains such as news (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003), E-commerce (Fetahu et al., 2021), social media (Strauss et al., 2016; Derczynski et al., 2017) and bio-medicine (Doğan et al., 2014; Li et al., 2016). Several recent studies focus on improving the accuracy of NER models through utilizing image information (MNER) in tweets (Zhang et al., 2018; Moon et al., 2018; Lu et al., 2018). Most approaches to MNER use the attention mechanism to model the interaction between image and text representations (Yu et al., 2020; Zhang et al., 2021a; Sun et al., 2021), in which image representations are from a pretrained feature extractor, i.e. ResNet (He et al., 2016), and text representations are extracted from pretrained textual embeddings, i.e. BERT (Devlin et al., 2019). Since these models are separately trained on datasets of different modalities and their feature representations are not aligned, it is difficult for the attention mechanism to model the interaction between the two modalities.

Recently, pretrained vision-language (V+L) models such as LXMERT (Tan and Bansal, 2019), UNITER (Chen et al., 2020) and Oscar (Li et al., 2020b) have achieved significant improvement on several cross-modal tasks such as image captioning, VQA (Agrawal et al., 2015), NLVR (Young et al., 2014) and image-text retrieval (Suhr et al., 2019). Most pretrained V+L models are trained on image-text pairs and simply concatenate text features and image features as the input of pretraining. There are, however, two problems. First, texts in these datasets mainly contain common nouns
instead of named entities\(^2\) which leads to an inductive bias over common nouns and images. Second, despite its important role in pretraining V+L models, the image modality only plays an auxiliary role in MNER for disambiguation, and can sometimes even be discarded. These problems make pretrained V+L models perform weaker than pretrained language models for MNER.

Pretrained textual embeddings such as BERT, XLM-RoBERTa (Conneau et al., 2020) and LUKE (Yamada et al., 2020) have achieved state-of-the-art performance on various NER datasets through simple fine-tuning of pretrained textual embeddings. Since most of the transformer-based pretrained textual embeddings are trained over long texts, recent work (Akbik et al., 2019; Schweter and Akbik, 2020; Yamada et al., 2020; Wang et al., 2021) has shown that introducing document-level contexts can significantly improve the accuracy of a NER model. The attention mechanism in transformer-based pretrained textual embeddings can utilize contexts to improve the token representation of a sequence. Moreover, pretrained V+L models such as Oscar and VinVL (Zhang et al., 2021b) can use object tags detected in images to significantly ease the alignments between text and image features. Therefore, the images in MNER can be converted to texts as well so that the image representations can be aligned to the space of text representations. As a result, the attention module of the pretrained textual embeddings have the capability to easily model the interactions between aligned image and text representations, without introducing a new attention module. In this paper, we propose ITA, a simple but effective framework for Image-Text Alignments. ITA converts an image into visual contexts in textual space by multi-level alignments. We concatenate the NER texts with the visual contexts as a new cross-modal input view instead of the text (T) input view. We feed the I+T input into a pretrained textual embeddings model to get stronger token representations of the input sentence. Then the token representations are fed into a linear-chain CRF layer to get the label sequence \(y\). To further improve the model accuracy of both input views, we use the cross-view alignment module to align the output distributions of I+T and T view to improve the accuracy of textual input view, based on cross-view alignment that minimizes the KL divergence over the probability distributions of the two views.

ITA can be summarized in four aspects:

1. Object Tags as Local Alignment: ITA locally extracts object tags and its corresponding attributes of image regions from an object detector.

2. Image Captions as Global Alignment: ITA summarizes what the image is describing through predicting image captions from an image captioning model.

3. Optical Character Alignment: ITA extracts the texts presented in the image via optical character recognition (OCR).

4. Cross-View Alignment: we calculate the KL divergence between the output distributions of two input views.

We show in experiments that ITA can significantly improve the model accuracy on MNER datasets and achieve the state-of-the-art. The cross-view alignment module can significantly improve both the cross-modal and textual input views, and bridge the performance gap between the two views.

2 Approaches

We consider the NER task as a sequence labeling problem. Given a sentence \(w = \{w_1, \cdots, w_n\}\) with \(n\) tokens and its corresponding image \(I\), an sequence labeling model aims to predict a label sequence \(y = \{y_1, \cdots, y_n\}\) at each position. In our framework, we focus on incorporating visual information to improve the representations of the input tokens by aligning visual and textual information effectively. We use a visual context generator to convert the image \(I\) into texts forming visual contexts \(w' = \{w'_1, \cdots, w'_m\}\) with \(m\) tokens. We then concatenate the input text and visual contexts as a cross-modal text+image (I+T) input view instead of the text (T) input view. We feed the I+T input into a pretrained textual embeddings model to get stronger token representations of the input sentence. Then the token representations are fed into a linear-chain CRF layer to get the label sequence \(y\). To further improve the model accuracy of both input views, we use the cross-view alignment module to align the output distributions of I+T and T.

\(^2\)https://visualgenome.org/data_analysis/statistics
input views during training. The architecture of our framework is shown in Figure 1.

2.1 NER Model Architecture

We use a neural model with a linear-chain CRF layer, a widely used approach for the sequence labeling problem (Huang et al., 2015; Akbik et al., 2018; Devlin et al., 2019). The input is fed into a transformer-based pretrained textual embeddings model and the output token representations \( \{r_1, \ldots, r_n\} \) are fed into the CRF layer:

\[
p_\theta(y|w) = \frac{\prod_{i=1}^{n} \psi(y_{i-1}, y_i, r_i)}{\sum_{y' \in \mathcal{Y}(w)} \prod_{i=1}^{n} \psi(y'_{i-1}, y'_i, r_i)}
\]

where \( \theta \) is the model parameters, \( \mathcal{Y}(w) \) is the set of all possible label sequences given the input \( w \). Given the gold label sequence \( y \) in the training data, the objective function of the model for the input view is:

\[
\mathcal{L}_T(\theta) = -\log p_\theta(y|w)
\]  

(1)

The loss can be calculated using Forward algorithm.

2.2 Image-text Alignments

The transformer-based pretrained textual embeddings have strong representations over texts. Therefore, ITA converts the image information into textual space through generating texts from the image so that the learning of the self-attention in the transformer-based model can be significantly eased compared with simply using image features from an object detector. We propose a local (LA), a global (GA) and an optical character alignment (OCA) approaches for alignments.

Object Tags as Local Alignment  Given an image, the image information can be decomposed into a set of objects in local regions. The object tags of each region textually describe the local information in the image. To extract the objects, we use an object detector OD to identify and locate the objects in the image:

\[ a, o = OD(I); \]

where \( a = \{a_1, a_2, \ldots, a_l\} \) and \( o = \{o_1, o_2, \ldots, o_l\} \)

The attribute predictions from the object detector contain multiple attribute tags \( a_i \) for each object \( o_i \). We linearize and sort the objects in a descending order based on the confidences of the detection model. For each object, we heuristically keep 0 to 3 attributes with confidence scores above a threshold \( m \). We linearize the attributes and put the attributes before the corresponding objects since the attributes are the adjectives describing the object tags. As a result, we take the predicted \( l \) object tags \( o \) and their attribute tags \( a \) from the object detector as the locally aligned visual contexts \( w^{LA} \):

\[ w^{LA} = \{a_1, o_1, a_2, o_2, \ldots, a_l, o_l\} \]

Image Captions as Global Alignment  Though the local alignment can localize the image into
objects, the objects cannot fully describe the of the whole image. Image captioning is a task that predicts the meaning of an image. Therefore, we align the image into $k$ image captions by an image captioning model $IC$:

$$\{w^1, w^2, \ldots, w^k \} = IC(I)$$

where $\{w^1, w^2, \ldots, w^k \}$ are captions generated from beam search with $k$ beams. We concatenate the $k$ captions together with a special separate token $[X]$ to form the aligned global visual contexts $w^{GA}$:

$$w^{GA} = [w^1, [X], w^2, [X], \ldots, [X], w^k]$$

The exact label (e.g. “[SEP]” in BERT) of the special $[X]$ token depends on the selection of embeddings.

**Optical Character Alignment** Some image contain text when they are created to enrich the semantic information that the images want to convey. In order to better understand this type of image, we use an OCR model to identify and extract the texts in the image:

$$w^{OCA} = OCR(I)$$

where $w^{OCA}$ are the texts extracted by the OCR model. Note that $w^{OCA}$ may be an empty text if there is no text in the image. We concatenate the input sentence and our aligned visual contexts to form the I+T input view $\hat{w} = [w; w']$, where $w'$ can be one of $w^{I+T}, w^{GA}, w^{OCA}$ or the concatenation of all (we denote it as All). The transformer-based embeddings are fed with the I+T input view and then output image-text fused token representations for each token $\{r'_1, \ldots, r'_n\}$. The token representations are fed into the CRF layer to get the probability distribution $p_0(y|\hat{w})$. Similar to Eq. 1, the objective function of the model for the I+T input view is:

$$L_{I+T}(\theta) = - \log p_0(y|\hat{w})$$ (2)

**Cross-View Alignment** There are several limitations in incorporating images into NER prediction: 1) the images may not available in testing; 2) aligning images to texts requires several pipelines in pre-processing instead of an end-to-end manner, which is so time-consuming that it is not applicable to some time-critical scenes such as online serving; 3) the noises in the image can mislead the MNER model to make wrong predictions. To alleviate these issues, we propose Cross-View Alignment (CVA), which targets at reducing the gap between the I+T and T input views over the output distributions so that the MNER model can better utilize the textual information in the input. During training, CVA minimizes the KL divergence over the probability distribution of I+T and T input views:

$$L_{CVA}(\theta) = KL(p_0(y|\hat{w})||p_0(y|w))$$ (3)

Since the I+T input view has additional visual information in the input and we want the T input view to match the accuracy of I+T input view, we only back-propagate through $p_0(y|w)$ in Eq. 3. Therefore, Eq. 3 is equivalent to calculating the cross-entropy loss over the two distributions:

$$L_{CVA}(\theta) = \sum_{y \in Y(x)} p_0(y|\hat{w}) \log p_0(y|w)$$ (4)

As the set of all possible label sequences $Y(x)$ is exponential in size, we calculate the posterior distributions of each position $p_0(y_i|w)$ and $p_0(y_i|\hat{w})$ through forward-backward algorithm to approximate Eq. 4:

$$p_0(y_k|*) \propto \sum_{\{y_0, \ldots, y_{k-1}\}} \prod_{i=1}^{k} \psi(y_{k-1}, y_i, r_i^*) \times \sum_{\{y_{k+1}, \ldots, y_n\}} \prod_{i=k+1}^{n} \psi(y_{i-1}, y_i, r_i^*) \quad (5)$$

where $r_i^*$ represents either $r_i$ or $r_i'$.

**Training** During training, we jointly train T and I+T input views with the training objective in Eq. 1 and 2 together with the CVA alignment training objective in Eq. 5. As a result, the final training objective for ITA is:

$$L_{ITA} = L_{CVA} + L_T + L_{I+T}$$

3 Experiments

We conduct experiments on two MNER datasets. To show the effectiveness of our approaches, we use two embedding settings and compare our approaches with previous multi-modal approaches.
3.1 Settings

Datasets We show the effectiveness of our approaches on Twitter-15, Twitter-17 and SNAP. The Twitter datasets contain 4,000/1,000/3,357, 3,373/723/723 and 4,290/1,432/1,459 sentences in train/development/test split respectively. The Twitter-15 dataset is constructed by Zhang et al. (2018). The SNAP dataset is constructed by Lu et al. (2018) and the Twitter-17 dataset is a filtered version of SNAP constructed by Yu et al. (2020).

Model Configuration For token representations, we use BERT base model to fairly compare with most of the recent work (Yu et al., 2020; Zhang et al., 2021a; Sun et al., 2021). Recently, XLM-RoBERTa has achieved state-of-the-art accuracy on various NER datasets by feeding the input together with contexts to the model. To further utilize the visual contexts in transformer-based embeddings, we use XLM-RoBERTa large (XLMR) model as another embedding in our experiments. To extract object tags and image captions of the images, we use VinVL (Zhang et al., 2021b), which is a pretrained V+L model based on a newly pretrained large-scale object detector based on ResNeXt-152 C4 architecture. We use the object detection module of VinVL to predict object tags and their corresponding attributes. The number of object tags and attributes varies over the images and is no more than 100. We set the threshold $m$ to be 0.1 for keeping the attributes of each object. For image captions, we use VinVL large model finetuned on MS-COCO (Lin et al., 2014) captions with CIDEr optimization (Rennie et al., 2017). In our experiments, we use a beam size of 5 with at most 20 tokens for prediction and keep all the 5 captions as the visual contexts. For OCR, we use Tesseract OCR (Smith, 2007), which is an open source OCR engine. We use the default configuration of the engine to extract texts in the image.

Training Configuration During training, we finetune the pretrained textual embedding model by AdamW (Loshchilov and Hutter, 2018) optimizer. In experiments we use the grid search to find the learning rate for the embeddings within $[1 \times 10^{-6}, 5 \times 10^{-4}]$. For BERT embeddings, we finetune the embeddings with a learning rate of $5 \times 10^{-5}$ with a batch size of 16. For XLMR embeddings, we use a learning rate of $5 \times 10^{-6}$ and a batch size of 4 instead. For the learning rate of the CRF layer, we use a grid search over $[0.05, 0.5]$ and $[0.005, 0.05]$ for BERT and XLMR respectively. The MNER models are trained for 10 epochs and we report the average results from 5 runs with different random seeds for each setting.

3.2 Results

In Table 1, we compare our approaches with our baselines with different training and evaluation modalities (T for the text-only input view and I+T for image and text together).
for the multi-modal input view). Results show that ITA models are significantly stronger than our BERT-CRF and XLMR-CRF baselines (Student’s t-test with \( p < 0.05 \)). For the aligned visual contexts, LA, GA and OCA are competitive in most of the cases. To show the effectiveness of CVA, we report the evaluation results of both input views in evaluation. With CVA, the accuracy of both input views can be improved, especially the T input view. CVA can improve the T input view to be competitive with I+T input view. Moreover, the combination of all the alignments ITA-All+CV A can further improve the model accuracy in most of the cases. The accuracy of the MNER models can be significantly improved if we use XLMR embeddings, which shows the importance of the text modality in MNER. With XLMR embeddings, the model accuracy can be further improved with ITA. The relative improvements over the baseline models are sometimes higher with XLMR than with BERT, which shows that the visual contexts can be further utilized with stronger embeddings.

In Table 2, we compare ITA with previous state-of-the-art approaches. For previous approaches, we report the results including OCSGA, UMT, RIVA, RpBERT, UMGF, which are the proposed approaches of Wu et al. (2020), Yu et al. (2020), Sun et al. (2020), Sun et al. (2021) and Zhang et al. (2021a) respectively. For fair comparison, we report the results of these models based on the BERT base embeddings. Moreover, since most of these previous approaches report the best model accuracy instead of the averaged model accuracy, we use the best model accuracy of ITA-All+CV A over 5 runs. We also report our reproduced results of UMT, RpBERT and UMGF on the corresponding datasets. The results show that ITA-All+CV A outperforms all of the previous approaches. On the SNAP dataset, the reported accuracy of RpBERTbase is competitive with ITA-All+CV A. However, we find that the accuracy of our reproduced RpBERTbase is significantly lower than the reported accuracy, even after careful check of the source code and hyper-parameter tuning. Moreover, the fact that our BERT-CRF baseline achieves competitive accuracy with previous state-of-the-art multi-modal approaches shows that most of the previous work has not fully explored the strength of the text representations for the task.

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Twitter-15</th>
<th>Twitter-17</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-CRF\textsuperscript{UMT}</td>
<td>71.81</td>
<td>83.44</td>
</tr>
<tr>
<td>BERT-CRF\textsuperscript{Ours}</td>
<td>74.79</td>
<td>85.18</td>
</tr>
</tbody>
</table>

**OUR REPRODUCTIONS**

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Twitter-15</th>
<th>Twitter-17</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-CRF\textsuperscript{UMT}</td>
<td>71.74</td>
<td>84.20</td>
</tr>
<tr>
<td>BERT-CRF\textsuperscript{UMT + Improved}</td>
<td>72.53</td>
<td>84.48</td>
</tr>
<tr>
<td>UMT</td>
<td>72.83</td>
<td>84.88</td>
</tr>
<tr>
<td>\textsuperscript{uMT Improved}</td>
<td>72.96</td>
<td>84.50</td>
</tr>
</tbody>
</table>

Table 3: Our reproductions of previous baselines and approaches. “Improved” means our improved models based on the UMT code base.

**Discussion about Textual Modules** As we have shown in Table 1 and 2, the textual baselines (i.e. BERT-CRF) of previous work are significantly lower than that of ours. In most of the previous MNER architectures, the textual modules are mainly based on the baseline architectures with some modifications. We further show the baselines of previous work are not well-trained and how the multi-modal approaches perform with stronger textual modules. In Table 3, we rerun the BERT-CRF baseline based on the released codes of UMT\textsuperscript{8}. Based on the code of UMT, we tried to improve the baseline models in the code by using the same loss function as ours\textsuperscript{9}. The accuracy of BERT-CRF models in the code are significantly improved but the UMT models based on the improved code are not improved and even get worse in Twitter-17. Therefore, we suspect the UMT model cannot be further improved even with stronger textual modules. Zhang et al. (2021a) also reported the baseline based on the implementation of Yu et al. (2020), so we suspect the UMGF model cannot be improved as well. Therefore, the under-trained textual baselines of previous work make the effectiveness of the images unclear and we show that some of the MNER models perform even weaker than our BERT-CRF model.

**3.3 Comparison with Other Variants**

To further show the effectiveness of ITA, we perform several comparisons between ITA and the following variants of the MNER model in Table 4:

**ITA-Random:** We generate random image-text pairs for the model. For each sentence, we randomly select the image in the dataset and generate the corresponding visual contexts. The noises of random visual contexts make the model accuracy

---

\textsuperscript{7}We reproduced the results based on the official code for RpBERT\textsubscript{base}: \url{https://github.com/Multimodal-NER/RpBERT}

\textsuperscript{8}\url{https://github.com/jefferyYu/UMT}

\textsuperscript{9}The details are discussed in Appendix A.5
Table 4: A comparison of other variants of MNER models.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Twitter-15</th>
<th>Twitter-17</th>
<th>SNAP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Eval</td>
<td>Eval</td>
<td>Eval</td>
</tr>
<tr>
<td></td>
<td>Modal</td>
<td>1+T</td>
<td>Modal</td>
</tr>
<tr>
<td>ITA-Random</td>
<td>-74.67</td>
<td>-84.98</td>
<td>-85.82</td>
</tr>
<tr>
<td>ITA-GA_BU</td>
<td>-75.10</td>
<td>-85.77</td>
<td>-86.51</td>
</tr>
<tr>
<td>ITA-LA_BU</td>
<td>-75.18</td>
<td>-85.59</td>
<td>-</td>
</tr>
<tr>
<td>ITA-OCA Paddle</td>
<td>-75.12</td>
<td>-85.87</td>
<td>-86.66</td>
</tr>
<tr>
<td>BERT-CRF+ImgFeat</td>
<td>-74.70</td>
<td>-84.99</td>
<td>-85.90</td>
</tr>
<tr>
<td>VinVL-CRF</td>
<td>-60.58</td>
<td>-75.55</td>
<td>-74.53</td>
</tr>
<tr>
<td>BERT+VinVL-CRF</td>
<td>-74.89</td>
<td>-85.19</td>
<td>-86.14</td>
</tr>
<tr>
<td>ITA-Joint</td>
<td>74.88</td>
<td>75.22</td>
<td>85.31</td>
</tr>
<tr>
<td></td>
<td>85.60</td>
<td>86.06</td>
<td>86.34</td>
</tr>
</tbody>
</table>

REFERENCES

- RpBERT w/o Rp
- ITA-All_{CVA} 75.50 75.41 85.89 85.84 86.83 86.75

Figure 2: A relation between the number of captions input to the MNER model and model accuracy. The x-axis is the number of captions. The y-axis is the averaged F1 score on the test set.

BERT-CRF+ImgFeat: Instead of ITA, we can directly feed the image features generated from an object detector into the BERT. We use ResNet-152 model to generate region features and then feed the features into a linear layer to project the region features into the same space of text features in the BERT. Moreover, we compare the model with RpBERT w/o Rp, which is an ablated model of RpBERT and is equivalent to BERT-CRF+ImgFeat over the usage of BERT embeddings. Sun et al. (2021) showed RpBERT w/o Rp can improve the model accuracy compared with their baseline. However, our results show that the model accuracy slightly drops comparing with our BERT-CRF, which shows that it is difficult for the attention module of BERT to learn the relations of the unaligned representations of two modalities.

VinVL-CRF: To show how the pretrained V+L models perform on the NER task, we use VinVL since it is a very recent state-of-the-art pretrained V+L model on a lot of multi-modal tasks. We feed the VinVL model with texts and images in the MNER datasets and finetune the model over the task. We take the text representations output from VinVL as the input of the CRF layer. The accuracy of the finetuned VinVL model drops significantly compared to the BERT model, which shows that the inductive bias of the pretrained V+L model hurts the model accuracy on MNER.

BERT+VinVL-CRF: As the VinVL model may lead to an inductive bias over the common nouns and the image, we jointly finetune the BERT and VinVL models and concatenate the output text representations of the two models. The accuracy is improved on a moderate scale, which shows BERT is complementary to VinVL for MNER.
3.4 Analysis

Effect of the Number of Captions Using more captions output from the captioning model can improve diversities of the visual contexts but can add noises to them as well. To better understand how the number of captions affects the model accuracy, we change the beam size and keep all the sentences output from the captioning model. The trends in Figure 2 show that the model accuracy increases until 5 captions for all the datasets and gradually drops when the number of captions further increases for Twitter-15 and 17 datasets. The observation shows that using 5 captions keeps a good balance between the diversities and correctness of the captions.

How ITA Eases the Cross-Modal Alignments
Previous work such as Moon et al. (2018); Sun et al. (2021) visualized modality attention in several cases to show the effectiveness of their approaches. However, visualizing the multi-layer attention in transformer-based embeddings is relatively difficult. Instead of studying special cases, we statistically calculate the averaged L2 distance between token representations \(r_i\) and \(r'_i\) from two input modalities to show how the token representations depend on image information. In Figure 3, the L2 distance ITA-All is significantly larger than that of BERT-CRF+ImgFeat. Besides, the standard deviation of BERT-CRF+ImgFeat is very large. The observations show that using 5 captions keeps a good balance between the diversities and correctness of the captions.

How Images Affect the NER Prediction To study the effectiveness of the images over each label, we show a comparison between our model and our baselines in Table 5. When the relative improvement of the F1 score is larger than 0.5, the relative improvement of precision is larger than that of recall. The observation shows that the main improvement of MNER is mainly because the images can help the model to reduce false-positive predictions for disambiguation on uncertain entities.12

4 Related Work

Multi-modal Named Entity Recognition Most of the previous approaches to MNER focus on the interaction between image and text features through attention mechanisms. Moon et al. (2018) proposed a modality attention network to fuse the text and image features before the input to the BiLSTM layer. Lu et al. (2018) additionally used a visual attention gate for the output features of the BiLSTM layer. Zhang et al. (2018) proposed an adaptive co-attention network after the BiLSTM layer to model the interaction between image and text. Recently, Wu et al. (2020) proposed OCSGA, which use object labels to model the interaction between image and object labels in an additional dense co-attention layer. Compared with the work, we show a simpler and more effective way to utilize object labels and additionally use other alignment approaches to further improve the model accuracy. Yu et al. (2020) proposed UMT, which utilized a multi-modal interaction module and an auxiliary entity span detection module for MNER. Zhang et al. (2021a) proposed UMGF, which utilizes a pretrained parser to create the graph connection between visual object tags and textual words. They used a graph attention network to fuse the textual and visual features. In order to better model whether the image is related to the text, Sun et al. (2021) proposed RpBERT, which additionally trains on a text-image relation classification dataset proposed by Vempala and Preotiuc-Pietro (2019) to prevent the negative effect of noisy images. Comparing with RpBERT, we use CVA to let the NER model better utilize the input sentences without such kinds of supervision. All of these approaches focus on fusing the image and text features through the attention mechanism but ignore the gap between the image and text features while we propose to fully utilize the attention mechanism in the pretrained textual embeddings through

12In Appendix A.3, we show several cases to show the effectiveness of ITA to affect NER prediction.
aligning image features into textual space. Besides, some cross-media research also shows the effectiveness of OCR texts (Chen et al., 2016; Wang et al., 2020) and object tags (Wu et al., 2016) have been shown. Most of the approaches introduced a new attention module over cross-modal features while in comparison ITA effectively utilizes the attention module in the pretrained textual embeddings.

**Pretrained Vision-Language Models**  Inspired by related work on language model pretraining, visual-language pretraining (VLP) has recently attracted a lot of attention (Li et al., 2019; Lu et al., 2019; Chen et al., 2020; Tan and Bansal, 2019; Li et al., 2020a; Yu et al., 2021; Zhang et al., 2021b). The pretrained V+L models are pretrained on large-scale image-text pairs and have achieved state-of-the-art accuracy over various vision-language tasks such as image captioning, VQA, NLVR and image-text retrieval. Recently, Li et al. (2020a) proposed Oscar to add object tags in pretraining so that self-attention can learn the image-text alignments easily. Following Oscar, Zhang et al. (2021b) proposed VinVL to train a large-scale object detector to improve the pretrained V+L model’s accuracy. Comparing with VLP, MNER is a totally different task. Firstly, the image-caption pairs are given in VLP and the image and text are equally important in pretraining for general representations. Therefore, using global alignment is meaningless for VLP but makes sense for MNER. In MNER, the input text is not the caption of the image and the image may not add additional information to the input text. Secondly, though captions and object tags are often utilized in VLP, how to effectively utilize the captions and object tags of the image in MNER is rarely considered. Finally, besides the local and global alignments, another aspect of ITA is the optical character alignment and cross-view alignment, which is rarely considered in VLP.

### 5 Conclusion

In this paper, we propose Image-Text Alignments for multi-modal named entity recognition, which convert images into object labels, captions and OCR texts to align the image representations into textual space in a multi-level manner and form a cross-modal input view. The model can effectively utilize attention module of the transformer-based embeddings. Considering noises, availability of images and inference speed for practical use, we propose cross-view alignment, which let the MNER models better utilize the text information in the input. In our experiments, we show that ITA significantly outperforms previous state-of-the-art approaches on MNER datasets. We also show that most of the previous work failed to train a good textual baseline while our textual baseline can easily match or even outperform previous multi-modal approaches. In analysis, we further analyze how ITA eases the cross-modal alignments and how the images affect the NER prediction.

### Acknowledgements

This work was supported by the National Natural Science Foundation of China (61976139) and by Alibaba Group through Alibaba Innovative Research Program.

### References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. L. Zitnick, Devi Parikh, and Dhruv

<table>
<thead>
<tr>
<th></th>
<th>LOC P</th>
<th>LOC R</th>
<th>LOC F1</th>
<th>ORG P</th>
<th>ORG R</th>
<th>ORG F1</th>
<th>PER P</th>
<th>PER R</th>
<th>PER F1</th>
<th>OTHER P</th>
<th>OTHER R</th>
<th>OTHER F1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Twitter-15</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT-CRF</td>
<td>80.0</td>
<td>83.8</td>
<td>81.8</td>
<td>65.9</td>
<td>61.0</td>
<td>63.3</td>
<td>84.2</td>
<td>86.8</td>
<td>85.4</td>
<td>44.2</td>
<td>44.2</td>
<td>44.1</td>
</tr>
<tr>
<td>ITA-All_{CV A}</td>
<td>81.1</td>
<td>84.2</td>
<td>82.6</td>
<td>68.8</td>
<td>60.6</td>
<td>64.4</td>
<td>84.0</td>
<td>87.2</td>
<td>85.6</td>
<td>44.9</td>
<td>44.6</td>
<td>44.8</td>
</tr>
<tr>
<td>Δ</td>
<td>1.1</td>
<td>0.4</td>
<td>0.8</td>
<td>2.8</td>
<td>-0.4</td>
<td>1.1</td>
<td>-0.2</td>
<td>0.4</td>
<td>0.1</td>
<td>0.8</td>
<td>0.5</td>
<td>0.6</td>
</tr>
</tbody>
</table>

| **Twitter-17** |       |       |        |       |       |        |       |       |        |         |         |         |
| BERT-CRF | 85.5  | 84.4  | 84.9   | 83.5  | 83.8  | 83.7   | 90.7  | 90.8  | 90.7   | 68.9  | 65.1  | 66.9    |
| ITA-All_{CV A} | 86.0  | 83.7  | 84.8   | 83.9  | 84.2  | 84.0   | 91.9  | 90.9  | 91.4   | 73.7  | 64.3  | 68.6    |
| Δ       | 0.5   | -0.7  | -0.1   | 0.3   | 0.4   | 0.4    | 1.2   | 0.1   | 0.7    | 4.8   | -0.8  | 1.7     |

| **SNAP** |       |       |        |       |       |        |       |       |        |         |         |         |
| BERT-CRF | 82.1  | 82.8  | 82.5   | 87.8  | 86.9  | 87.3   | 91.0  | 91.5  | 91.2   | 72.3  | 75.1  | 73.7    |
| ITA-All_{CV A} | 80.3  | 81.7  | 81.0   | 87.8  | 86.5  | 87.1   | 90.1  | 91.2  | 90.6   | 70.1  | 73.2  | 71.6    |
| Δ       | 1.9   | 1.1   | 1.5    | 0.6   | 0.5   | 0.5    | 0.9   | 0.3   | 0.6    | 2.2   | 1.9   | 2.1     |

Table 5: A comparison between our ITA (ITA-All_{CV A} with I+T inputs) model and the baseline (BERT-CRF) in precision (P), recall (R) and F1. Δ represents the relevant improvement of ITA over the Baseline.


Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-Fung Leung, and Qing Li. 2020. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. ACM MM.


Peter Young, Alice Lai, M. Hodosh, and J. Hockenmaier. 2014. From image descriptions to visual denotations:


Appendix

A.1 Details of Experiment Settings
We run our code on Tesla V100 GPU with 16 GB memory. It takes about two hours to train a model. The size of model parameter is approximately equal to size of BERT/XLMR embeddings.

A.2 Details of OCA
Table 6 shows that the OCR system only finds about 26% sentences have texts in the image and the extracted texts have an average of 28 tokens. The statistics show that ITA-OCA can help to improve the model accuracy with only 26% of the samples have OCR texts.

A.3 Case Study
Despite that images can generally help to improve the accuracy of the NER model, there are a lot of cases that the images may contain misleading information to hurt the model prediction. We study two cases for LA nad GA: 1) the entities are wrongly predicted by BERT-CRF baseline but are correctly predicted by ITA; 2) the entities are wrongly predicted by ITA without CVA but are correctly predicted by the baseline and ITA with CVA. Figure 4 shows the two cases with two samples for each. Figure 4 (a) shows the first case, which shows the importance of the visual contexts. The baseline model failed to recognize the person entities “TWICE” and “Harry Potter” possibly because the two words are usually an adverb and a book name respectively. For the I-T input view, our MNER model is able to recognize the hints such as “two girls”, “young girl”, “a couple of young men” and “woman” in the visual contexts and then correctly predict the two entities. Figure 4 (b) shows the second case, which shows how the noises from the image mislead the model predictions. There are three- and two-person entities in gold labels but the visual contexts indicate that the top right image has “two baseball players” and the bottom right image has only “a woman”. As a result, ITA without CVA only predict two and one person entities according to the visual contexts in the two samples respectively. However, with CVA, ITA takes a good balance in utilizing the textual and visual information and correctly predicts the entity labels in both T and I-T input views.

A.4 Discussion
In our paper, we use the captioning and object detection model based on MSCOCO and visual genome. The model performance could be improved if we use domain-specific models (Twitter domain). For OCA, the model accuracy may be poor if the OCR system does not support a certain language.

A.5 Loss Function Comparison with UMT
In the codes of UMT, the BERT embeddings tokenize the token in a sentence into subtokens. The codes use the first subtoken as the token representation to predict the corresponding label. However, for the other subtokens, the codes use a special label “PAD” for prediction. Therefore, the target labels are changed. For example, the original label predicted by ITA; 2) the entities are wrongly predicted by ITA without CVA but are correctly predicted by the baseline and ITA with CVA. Figure 4 shows the two cases with two samples for each. Figure 4 (a) shows the first case, which shows the importance of the visual contexts. The baseline model failed to recognize the person entities “TWICE” and “Harry Potter” possibly because the two words are usually an adverb and a book name respectively. For the I-T input view, our MNER model is able to recognize the hints such as “two girls”, “young girl”, “a couple of young men” and “woman” in the visual contexts and then correctly predict the two entities. Figure 4 (b) shows the second case, which shows how the noises from the image mislead the model predictions. There are three- and two-person entities in gold labels but the visual contexts indicate that the top right image has “two baseball players” and the bottom right image has only “a woman”. As a result, ITA without CVA only predict two and one person entities according to the visual contexts in the two samples respectively. However, with CVA, ITA takes a good balance in utilizing the textual and visual information and correctly predicts the entity labels in both T and I-T input views.

The baseline model failed to identify the latter one while ITA-OCA can successfully identify both of them. In the bottom of Figure 5, the texts in the image are mainly talking about “HARRY STYLES”, which helps the model prediction.

In the upper sample of Figure 5, there are two “Donald” words in the image. The baseline model failed to identify the latter one while ITA-OCA can successfully identify both of them. In the bottom of Figure 5, the texts in the image are mainly talking about “HARRY STYLES”, which helps the model prediction.
### (a) Importance of visual context

| Gold Labels: S-PER B-LOC E-LOC S-MISC |
| Baseline: B-LOC E-LOC S-MISC |
| ITA-All: S-PER B-LOC E-LOC S-MISC |
| ITA-All+CVA (T): S-PER B-LOC E-LOC S-MISC |
| ITA-All+CVA (I+T): S-PER B-LOC E-LOC S-MISC |

### (b) Importance of Cross-View Alignment

| Text: @HoulsbyMark Mark, meet my niece, well known concert violinist. Captions: a woman in a white dress holding a violin. Object Tags: smiling woman, black hair, open mouth, brown eye, face. |
| Gold Labels: S-PER S-PER |
| Baseline: S-PER S-PER |
| ITA-OCA: B-PER E-PER |

---

**Figure 4:** Examples of the positive and negative effects of images. The named entities in the text are colored. The wrongly predicted entities are marked in bold and colored in red. The missing entities are marked with ✗. We use BIOES format to represent the label spans (https://en.wikipedia.org/wiki/Inside-outside-beginning_(tagging))

---

**Figure 5:** Examples of the positive effects of OCA. The named entities in the text are colored.
sequence is “B-X, I-X, O, B-X, O, O” but now it becomes “B-X, PAD, PAD, I-X, O, B-X, O, PAD, O”. As a result, the exact training objective changes compared with the training objective in the paper of UMT. We improve the code by removing all the “PAD” labels and just use the first subtoken of each token as the token representation. Our improved baseline model is significantly improved, while the accuracy of UMT model in the improved code cannot be further improved.

<table>
<thead>
<tr>
<th></th>
<th>Twitter-15</th>
<th>Twitter-17</th>
<th>SNAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Num Sents w/ OCR / Total Sents</td>
<td>2,049 / 8,288 (24.72%)</td>
<td>1,197 / 4,461 (26.83%)</td>
<td>1,869 / 7,181 (26.03%)</td>
</tr>
<tr>
<td>Avg. Length</td>
<td>27.72</td>
<td>27.00</td>
<td>28.93</td>
</tr>
</tbody>
</table>

Table 6: A statistic about the number of sentences has OCR texts and the average length of OCR texts.