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

Anonymous ACL submission

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 **I**mage-**t**ext **A**lignments (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 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 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 extracted from a pretrained object detector, 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[1] 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

[1] https://visualgenome.org/data_analysis/statistics

(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 **I**mage-**T**ext **A**lignments. 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 and then feed it into a pretrained textual embedding model to improve the token representations of NER texts, which are fed into a linear-chain CRF (Lafferty et al., 2001) layer for prediction. In practice, a MNER model should be robust when there is only text information, as images may be unavailable or can introduce noise. Sometimes it is even undesirable to use images as image feature extraction can be inefficient in online serving. Therefore, we further propose to utilize the cross-modal input 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 $\boldsymbol{w} = \{w_1, \cdots, w_n\}$ with n tokens and its corresponding image I, an sequence labeling model aims to predict a label sequence $\boldsymbol{y} = \{y_1, \cdots, y_n\}$ 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 $\boldsymbol{w}' = \{w'_1, \cdots, w'_m\}$ with m tokens. We 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 $\boldsymbol{y}$. 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** 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

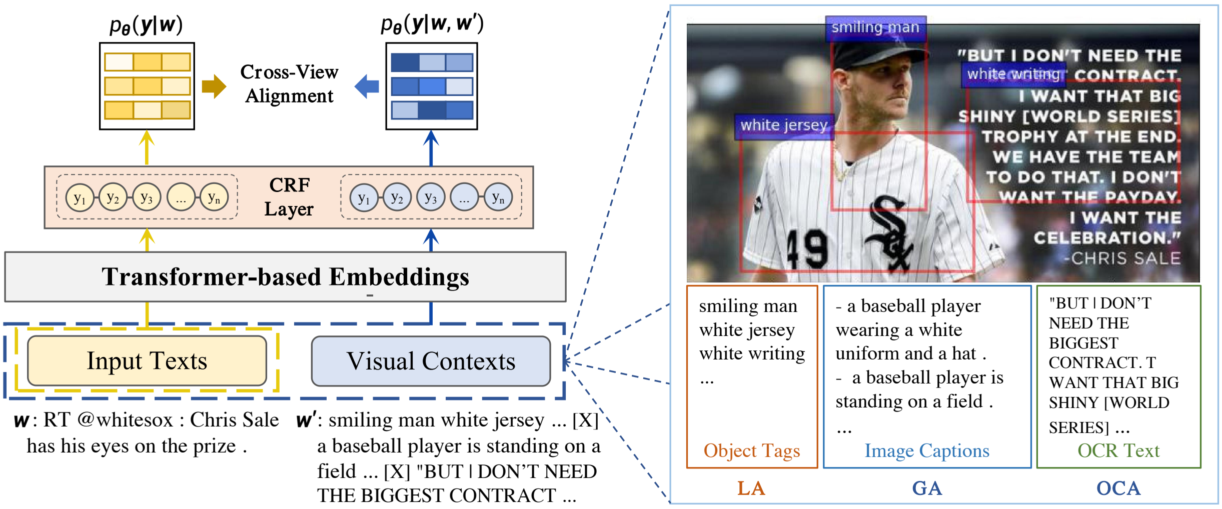

Figure 1: The architecture of ITA. ITA aligns an image into object tags, image captions and texts from OCR. ITA takes them as visual contexts and then feeds them together with the input texts into the transformer-based embeddings. In the cross-view alignment module, ITA minimizes the distance between the output distribution of cross-modal inputs and textual inputs.

$\{\boldsymbol{r}_1, \cdots, \boldsymbol{r}_n\}$ are fed into the CRF layer:

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

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

$$\mathcal{L}_{\mathrm{T}}(\theta) = -\log p_\theta(\hat{\boldsymbol{y}}|\boldsymbol{w}) \tag{1}$$

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:

$$\boldsymbol{a}, \boldsymbol{o} = \mathbf{OD}(I); \text{where}$$
$$\boldsymbol{a} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_l\} \text{ and } \boldsymbol{o} = \{o_1, o_2, \cdots, o_l\}$$

The attribute predictions from the object detector contain multiple attribute tags $\boldsymbol{a}_i$ 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 $\boldsymbol{o}$ and their attribute tags $\boldsymbol{a}$ from the object detector as the locally aligned visual contexts $\boldsymbol{w}^{\mathrm{LA}}$:

$$\boldsymbol{w}^{\mathrm{LA}} = \{\boldsymbol{a}_1, o_1, \boldsymbol{a}_2, o_2, \cdots, \boldsymbol{a}_l, o_l\}$$

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**:

$$\{\boldsymbol{w}^1, \boldsymbol{w}^2, \cdots, \boldsymbol{w}^k\} = \mathbf{IC}(I)$$

where $\{\boldsymbol{w}^1, \boldsymbol{w}^2, \cdots, \boldsymbol{w}^k\}$ 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 $\boldsymbol{w}^{\text{GA}}$:

$$\boldsymbol{w}^{\text{GA}}{=}[\boldsymbol{w}^1, \text{[X]}, \boldsymbol{w}^2, \text{[X]}, \cdots, \text{[X]}, \boldsymbol{w}^k]$$

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:

$$\boldsymbol{w}^{\text{OCA}} = \textbf{OCR}(I)$$

where $\boldsymbol{w}^{\text{OCA}}$ are the texts extracted by the **OCR** model. Note that $\boldsymbol{w}^{\text{OCA}}$ 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 $\hat{\boldsymbol{w}} = [\boldsymbol{w}; \boldsymbol{w}']$, where $\boldsymbol{w}'$ can be one of $\boldsymbol{w}^{\text{LA}}$, $\boldsymbol{w}^{\text{GA}}$, $\boldsymbol{w}^{\text{OCA}}$ 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 $\{\boldsymbol{r}'_1, \cdots, \boldsymbol{r}'_n\}$. The token representations are fed into the CRF layer to get the probability distribution $p_\theta(\boldsymbol{y}|\hat{\boldsymbol{w}})$. Similar to Eq. 1, the objective function of the model for the **I+T** input view is:

$$\mathcal{L}_{\text{I+T}}(\theta) = -\log p_\theta(\hat{\boldsymbol{y}}|\hat{\boldsymbol{w}}) \quad (2)$$

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 noise 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:

$$\mathcal{L}_{\text{CVA}}(\theta){=}\text{KL}(p_\theta(\boldsymbol{y}|\hat{\boldsymbol{w}})||p_\theta(\boldsymbol{y}|\boldsymbol{w})) \quad (3)$$

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 $p_\theta(\boldsymbol{y}|\boldsymbol{w})$ in Eq. 3. Therefore, Eq. 3 is equivalent to calculating the cross-entropy loss over the two distributions:

$$\mathcal{L}_{\text{CVA}}(\theta){=}\sum_{\boldsymbol{y}\in\mathcal{Y}(\boldsymbol{x})} p_\theta(\boldsymbol{y}|\hat{\boldsymbol{w}}) \log p_\theta(\boldsymbol{y}|\boldsymbol{w}) \quad (4)$$

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

$$p_\theta(y_k|*){\propto}\sum_{\{y_0,\dots,y_{k-1}\}}\prod_{i=1}^{k}\psi(y_{i-1}, y_i, \boldsymbol{r}_i^*)$$

$$\times \sum_{\{y_{k+1},\dots,y_n\}}\prod_{i=k+1}^{n}\psi(y_{i-1}, y_i, \boldsymbol{r}_i^*)$$

$$\mathcal{L}_{\text{CVA}}(\theta){=}\sum_{i=1}^{n} p_\theta(y_i|\hat{\boldsymbol{w}}) \log p_\theta(y_i|\boldsymbol{w})) \quad (5)$$

where $\boldsymbol{r}_i^*$ represents either $\boldsymbol{r}_i$ or $\boldsymbol{r}'_i$.

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:

$$\mathcal{L}_{\text{ITA}} = \mathcal{L}_{\text{CVA}} + \mathcal{L}_{\text{T}} + \mathcal{L}_{\text{I+T}}$$

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 Twitter datasets[2] containing 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

[2] Twitter-15 and 17 datasets are available at https://github.com/jefferyYu/UMT.

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 image, we use VinVL (Zhang et al., 2021b), which is a pre-trained V+L model based on a newly pretrained large-scale object detector based on the 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[3] 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[4] (Smith, 2007), which is an open source OCR engine. We use the default configuration of the engine to extract texts in the image[5].

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 5×10^{-5} with a batch size of 16. For XLMR embeddings, we use a learning rate of 5×10^{-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] github.com/microsoft/Oscar

[4] github.com/tesseract-ocr/tesseract

[5] Please refer to Appendix for more statistics.

Train Modal	Approach	Twitter-15 Eval Modal T	Twitter-15 Eval Modal I+T	Twitter-17 Eval Modal T	Twitter-17 Eval Modal I+T	SNAP Eval Modal T	SNAP Eval Modal I+T
	BERT-CRF						
T	**BERT-CRF**	74.79	-	85.18	-	85.98	-
I+T	**ITA-LA**	-	75.18	-	85.67	-	86.26
	ITA-GA	-	75.17	-	85.75	-	86.72
	ITA-OCA	-	75.01	-	85.64	-	86.52
	ITA-All	-	75.15	-	85.78	-	86.79
	ITA-LA$_{+CVA}$	75.26	75.20	85.72	85.62	86.51	86.41
	ITA-GA$_{+CVA}$	75.45	75.52	85.96	**85.85**	86.42	86.39
	ITA-OCA$_{+CVA}$	75.26	75.30	85.73	85.79	86.64	86.59
	ITA-All$_{+CVA}$	**75.67**	**75.60**	**85.98**	85.72	**86.83**	**86.75**
	XLMR-CRF						
T	**XLMR-CRF**	77.37	-	88.73	-	89.39	-
I+T	**ITA-LA**	-	77.64	-	89.29	-	89.68
	ITA-GA	-	77.78	-	89.32	-	89.78
	ITA-OCA	-	77.94	-	89.31	-	89.64
	ITA-All	-	77.81	-	89.62	-	90.10
	ITA-LA$_{+CVA}$	77.87	77.93	89.45	**89.90**	89.85	89.91
	ITA-GA$_{+CVA}$	78.03	78.02	89.41	89.62	89.85	90.09
	ITA-OCA$_{+CVA}$	77.57	77.59	89.32	89.55	89.90	89.84
	ITA-All$_{+CVA}$	**78.25**	**78.03**	**89.47**	89.75	**90.02**	**90.15**

Table 1: A comparison of ITA and our baseline.

Approach	Twitter-15	Twitter-17	SNAP
Reported F1 of Previous Approaches			
OCSGA♣	72.92	-	-
UMT†	73.41	85.31	-
RIVA‡	73.80	-	86.80
RpBERT$_{base}$♠	74.40	-	87.40
UMGF◇	74.85	85.51	-
Our Reproductions			
BERT-CRF	74.79	85.18	85.98
UMT	72.83	84.88	-
UMGF	74.42	85.27	-
Ours: **ITA-All$_{+CVA}$**	**76.01**	**86.45**	**87.44**

Table 2: A comparison of our approaches and state-of-the-art approaches. ♣: Wu et al. (2020); †: results are from Yu et al. (2020); ‡: Sun et al. (2020), ♠: Sun et al. (2021), note that **RpBERT$_{base}$** uses the test set to select the best model; ◇: results are from Zhang et al. (2021a).

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 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$_{+\text{CVA}}$** 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** 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$_{+\text{CVA}}$** over 5 runs. We also report our reproduced results of **UMT**, **UMGF** on the corresponding datasets. The results show that **ITA-All$_{+\text{CVA}}$** outperforms all of the previous approaches. On the SNAP dataset, **RpBERT$_{\text{base}}$** is competitive with **ITA-All$_{+\text{CVA}}$**. However, we find that in the official code of RpBERT, the model checkpoint is selected by the F1 scores on the test set[6] during training while we only use the development set during training. As a result, we believe **RpBERT** is not directly comparable with our models. We would also like to rerun the official code of RpBERT with our setting, but we were unable to run its code correctly. 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.

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 3:

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 noise of random visual contexts makes the model accuracy drop slightly comparing with our **BERT-CRF** baseline, which shows the improvement of our approach is from the visual contexts rather than extending the input sequence length the embeddings.

Approach	Twitter-15 Eval Modal T	Twitter-15 Eval Modal I+T	Twitter-17 Eval Modal T	Twitter-17 Eval Modal I+T	SNAP Eval Modal T	SNAP Eval Modal I+T
ITA-Random	-	74.67	-	84.98	-	85.82
BERT-CRF$_{+\text{ImgFeat}}$	-	74.70	-	84.99	-	85.90
VinVL-CRF	-	60.58	-	75.55	-	74.53
BERT+VinVL-CRF	-	74.89	-	85.19	-	86.14
ITA-Joint	74.88	75.22	85.31	85.60	86.06	86.34
REFERENCES						
RpBERT w/o Rp	-	72.60	-	-	-	86.20
ITA-All$_{+\text{CVA}}$	**75.50**	**75.41**	**85.89**	**85.84**	**86.83**	**86.75**

Table 3: A comparison of other variants of MNER models. IF: using image features as visual contexts.

ITA-Joint: It is an ablated model of **ITA-All$_{+\text{CVA}}$**. We train the **ITA-All** model for both input views without the CVA loss in Eq. 5. The model accuracy is improved moderately with only the **T** input view while our **ITA-All$_{+\text{CVA}}$** can improve both input views significantly, which shows the effectiveness of the CVA module of ITA.

BERT-CRF$_{+\text{ImgFeat}}$: Instead of **ITA**, we can directly feed the image region 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+$_{+\text{ImgFeat}}$** 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

[6] github.com/Multimodal-NER/RpBERT

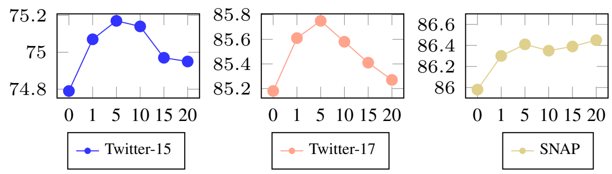

Figure 2: A relation between the number of captions input to the MNER model and model accuracy. The x-axis is the number of captions. The y-axis is the averaged F1 score on the test set.

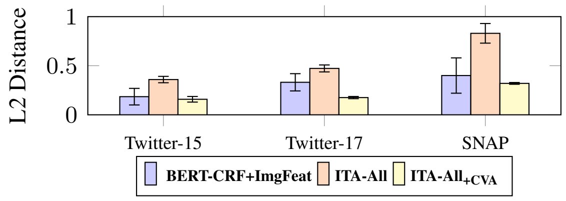

Figure 3: Averaged L2 distance between the token representations without image input ($\boldsymbol{r}_i$) and with image input ($\boldsymbol{r}'_i$). The error bars mean the standard deviation over 5 runs.

	LOC P	LOC R	LOC F1	ORG P	ORG R	ORG F1	PER P	PER R	PER F1	OTHER P	OTHER R	OTHER F1
	Twitter-15											
Base	80.0	83.8	81.8	65.9	61.0	63.3	84.2	86.8	85.4	44.2	44.2	44.1
ITA	81.1	84.2	82.6	68.8	60.6	64.4	84.0	87.2	85.6	44.9	44.6	44.8
Δ	1.1	0.4	0.8	2.8	-0.4	1.1	-0.2	0.4	0.1	0.8	0.5	0.6
	Twitter-17											
Base	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	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											
Base	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	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 4: A comparison between our ITA (**ITA-All$_{+\mathbf{CVA}}$** 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.

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 $\boldsymbol{r}_i$ and $\boldsymbol{r}'_i$ 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 the image region features make the alignment become difficult and unstable while our visual contexts can significantly ease the cross-modal alignments. Moreover, with CVA, the L2 distance becomes much smaller and stable as CVA aligns the two input views to reduce the dependence on images, which shows the MNER model can better utilize the textual information with CVA.

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 4. 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.[7]

[7] In Appendix, we show several cases to show the effectiveness of ITA to affect NER prediction.

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 Preoţiuc-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 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 adds 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.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. L. Zitnick, Devi Parikh, and Dhruv Batra. 2015. Vqa: Visual question answering. *International Journal of Computer Vision*, 123:4–31.

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019. Pooled contextualized embeddings for named entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 724–728, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Tao Chen, Xiangnan He, and Min-Yen Kan. 2016. Context-aware image tweet modelling and recommendation. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1018–1027.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

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

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Besnik Fetahu, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. Gazetteer enhanced named entity recognition for code-mixed web queries. In *SIGIR '21*, SIGIR '21, New York, NY, USA. Association for Computing Machinery.

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

Zhiheng Huang, W. Xu, and Kailiang Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *ArXiv*, abs/1508.01991.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation*, 2016.

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

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020a. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*.

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

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024.

Stefan Schweter and Alan Akbik. 2020. Flert: Document-level features for named entity recognition. *arXiv preprint arXiv:2011.06993*.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. 2016. Results of the WNUT16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 138–144, Osaka, Japan. The COLING 2016 Organizing Committee.

Alane Suhr, Stephanie Zhou, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *ACL*.

Lin Sun, Jiquan Wang, Yindu Su, Fangsheng Weng, Yuxuan Sun, Zengwei Zheng, and Yuanyi Chen. 2020. RIVA: A pre-trained tweet multimodal model based on text-image relation for multimodal NER. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1852–1862, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: A text-image relation propagation-based bert model for multimodal ner. In *AAAI*.

Beth M. Sundheim. 1995. Named entity task definition, version 2.1. In *Proceedings of the Sixth Message Understanding Conference*, pages 319–332.

Hao Hao Tan and M. Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

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

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Improving named entity recognition by external context retrieving and cooperative learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1800–1812, Online. Association for Computational Linguistics.

Yue Wang, Jing Li, Michael Lyu, and Irwin King. 2020. Cross-media keyphrase prediction: A unified framework with multi-modality multi-head attention and image wordings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3311–3324, Online. Association for Computational Linguistics.

Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. 2016. What value do explicit high level concepts have in vision to language problems? In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 203–212.

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.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Peter Young, Alice Lai, M. Hodosh, and J. Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. In *AAAI*.

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

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021a. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021b. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.

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

A 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. Our code is based on transformers[8] and flair[9]. We will release our code upon acceptance.

A.2 Details of OCA

Table 5 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.

For OCA, we study how the extracted texts can help 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.4 Discussion

In our paper, we use the captioning and object detection model based on MSCOCO and visual genome. The model performance could be 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.

[8] github.com/huggingface/transformers
[9] github.com/flairNLP/flair

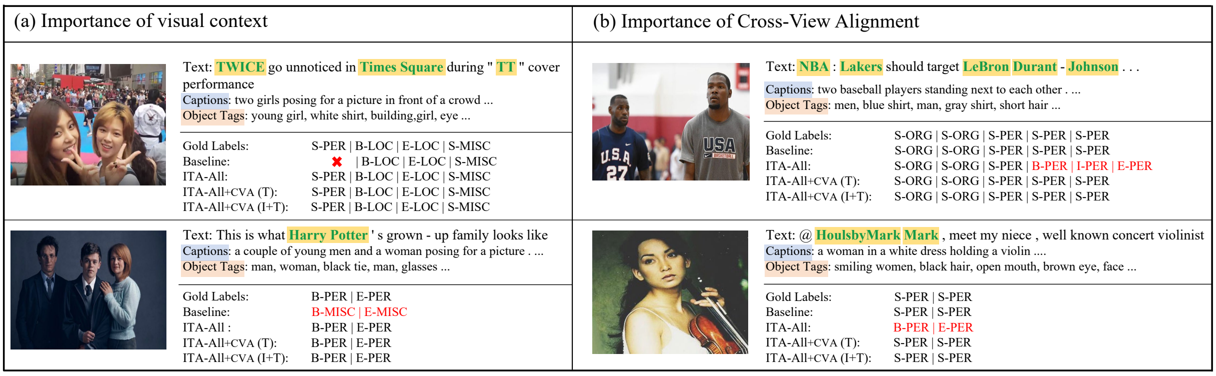

Figure 4: Examples of the positive and negative effects of images. The named entities in the text are colored. The wrongly predicted entities are marked in bold and colored in red. The missing entities are marked with ✖. We use BIOES format to represent the label spans (https://en.wikipedia.org/wiki/Inside-outside-beginning_(tagging))

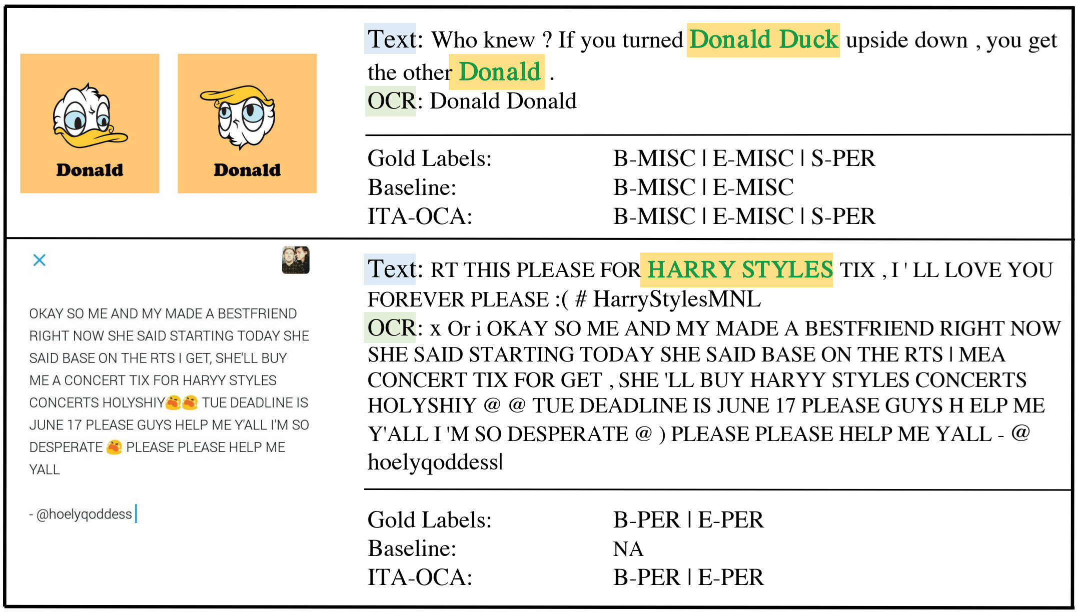

Figure 5: Examples of the positive effects of OCA. The named entities in the text are colored.

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

Table 5: A statistic about the number of sentences has OCR texts and the average length of OCR texts.