

Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer

Jianfei Yu¹, Jing Jiang², Li Yang³, and Rui Xia^{1,*}

¹ School of Artificial Intelligence, Nanjing University of Science & Technology, China

² School of Information Systems, Singapore Management University, Singapore

³ DBS Bank, Singapore

{jfyu, rxia}@njjust.edu.cn, jingjiang@smu.edu.sg, liyang@dbb.com

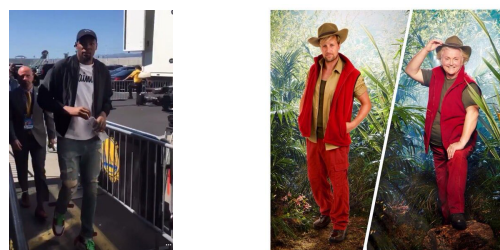
Abstract

In this paper, we study Multimodal Named Entity Recognition (MNER) for social media posts. Existing approaches for MNER mainly suffer from two drawbacks: (1) despite generating word-aware visual representations, their word representations are insensitive to the visual context; (2) most of them ignore the bias brought by the visual context. To tackle the first issue, we propose a multimodal interaction module to obtain both image-aware word representations and word-aware visual representations. To alleviate the visual bias, we further propose to leverage purely text-based entity span detection as an auxiliary module, and design a Unified Multimodal Transformer to guide the final predictions with the entity span predictions. Experiments show that our unified approach achieves the new state-of-the-art performance on two benchmark datasets.

1 Introduction

Recent years have witnessed the explosive growth of user-generated contents on social media platforms such as Twitter. While empowering users with rich information, the flourish of social media also solicits the emerging need of automatically extracting important information from these massive unstructured contents. As a crucial component of many information extraction tasks, named entity recognition (NER) aims to discover named entities in free text and classify them into pre-defined types, such as person (*PER*), location (*LOC*) and organization (*ORG*). Given its importance, NER has attracted much attention in the research community (Yadav and Bethard, 2018).

Although many methods coupled with either discrete shallow features (Zhou and Su, 2002; Finkel et al., 2005; Torisawa et al., 2007) or continuous deep features (Lample et al., 2016; Ma and Hovy,



(a). [Kevin Durant PER] enters [Oracle Arena LOC] wearing off — White x [Jordan MISC]
(b). Vote for [King of the Jungle MISC] — [Kian PER] or [David PER] ?

Figure 1: Two examples for Multimodal Named Entity Recognition (MNER). Named entities and their entity types are highlighted.

2016) have shown success in identifying entities in formal newswire text, most of them perform poorly on informal social media text (e.g., tweets) due to its short length and noisiness. To adapt existing NER models to social media, various methods have been proposed to incorporate many tweet-specific features (Ritter et al., 2011; Li et al., 2012, 2014; Limsopatham and Collier, 2016). More recently, as social media posts become increasingly multimodal, several studies proposed to exploit useful visual information to improve the performance of NER (Moon et al., 2018; Zhang et al., 2018; Lu et al., 2018).

In this work, following the recent trend, we focus on multimodal named entity recognition (MNER) for social media posts, where the goal is to detect named entities and identify their entity types given a {sentence, image} pair. For example, in Fig. 1.a, it is expected to recognize that *Kevin Durant*, *Oracle Arena*, and *Jordan* belong to the category of person names (i.e., *PER*), place names (i.e., *LOC*), and other names (i.e., *MISC*), respectively.

While previous work has shown success of fusing visual information into NER (Moon et al., 2018; Zhang et al., 2018; Lu et al., 2018), they still suffer from several limitations: (1) The first obstacle

*Corresponding author.

lies in the non-contextualized word representations, where each word is represented by the same vector, regardless of the context it occurs in. However, the meanings of many polysemous entities in social media posts often rely on its context words. Take Fig. 1.a as an example, without the context words *wearing off*, it is hard to figure out whether *Jordan* refers to a shoe brand or a person. (2) Although most existing methods focus on modeling inter-modal interactions to obtain word-aware visual representations, the word representations in their final hidden layer are still based on the textual context, which are insensitive to the visual context. Intuitively, the associated image often provides more context to resolve polysemous entities, and should contribute to the final word representations (e.g., in Fig. 1.b, the image can supervise the final word representations of *Kian* and *David* to be closer to persons than animals). (3) Most previous approaches largely ignore the bias of incorporating visual information. Actually, in most social media posts, the associated image tends to highlight only one or two entities in the sentence, without mentioning the other entities. In these cases, directly integrating visual information will inevitably lead the model to better recognize entities highlighted by images, but fail to identify the other entities (e.g., *Oracle Arena* and *King of the Jungle* in Fig. 1).

To address these limitations, we resort to existing pre-trained contextualized word representations, and propose a unified multimodal architecture based on Transformer (Vaswani et al., 2017), which can effectively capture inter-modality interactions and alleviate the visual bias. Specifically, we first adopt a recently pre-trained contextualized representation model (Devlin et al., 2018) as our sentence encoder, whose multi-head self-attention mechanism can guide each word to capture the semantic and syntactic dependency upon its context. Second, to better capture the implicit alignments between words and images, we propose a multimodal interaction (MMI) module, which essentially couples the standard Transformer layer with cross-modal attention mechanism to produce an image-aware word representation and a word-aware visual representation for each input word, respectively. Finally, to largely eliminate the bias of the visual context, we propose to leverage text-based entity span detection as an auxiliary task, and design a unified neural architecture based on Transformer. In particular, a conversion matrix is

designed to construct the correspondence between the auxiliary and the main tasks, so that the entity span information can be fully utilized to guide the final MNER predictions.

Experimental results show that our Unified Multimodal Transformer (UMT) brings consistent performance gains over several highly competitive unimodal and multimodal methods, and outperforms the state-of-the-art by a relative improvement of 3.7% and 3.8% on two benchmarks, respectively.

The main contributions of this paper can be summarized as follows:

- We propose a Multimodal Transformer model for the task of MNER, which empowers Transformer with a multimodal interaction module to capture the inter-modality dynamics between words and images. To the best of our knowledge, this is the first work to apply Transformer to MNER.
- Based on the above Multimodal Transformer, we further design a unified architecture to incorporate a text-based entity span detection module, aiming to alleviate the bias of the visual context in MNER with the guidance of entity span predictions from this auxiliary module.

2 Methodology

In this section, we first formulate the MNER task, and give an overview of our method. We then delve into the details of each component in our model.

Task Formulation: Given a sentence S and its associated image V as input, the goal of MNER is to extract a set of entities from S , and classify each extracted entity into one of the pre-defined types.

As with most existing work on MNER, we formulate the task as a sequence labeling problem. Let $S = (s_1, s_2, \dots, s_n)$ denote a sequence of input words, and $\mathbf{y} = (y_1, y_2, \dots, y_n)$ be the corresponding label sequence, where $y_i \in \mathcal{Y}$ and \mathcal{Y} is the pre-defined label set with the BIOES tagging schema (Sang and Veenstra, 1999).

2.1 Overall Architecture

Fig. 2.a illustrates the overall architecture of our Unified Multimodal Transformer, which contains three main components: (1) representation learning for unimodal input; (2) a Multimodal Transformer for MNER; and (3) a unified architecture with auxiliary entity span detection (ESD) module.

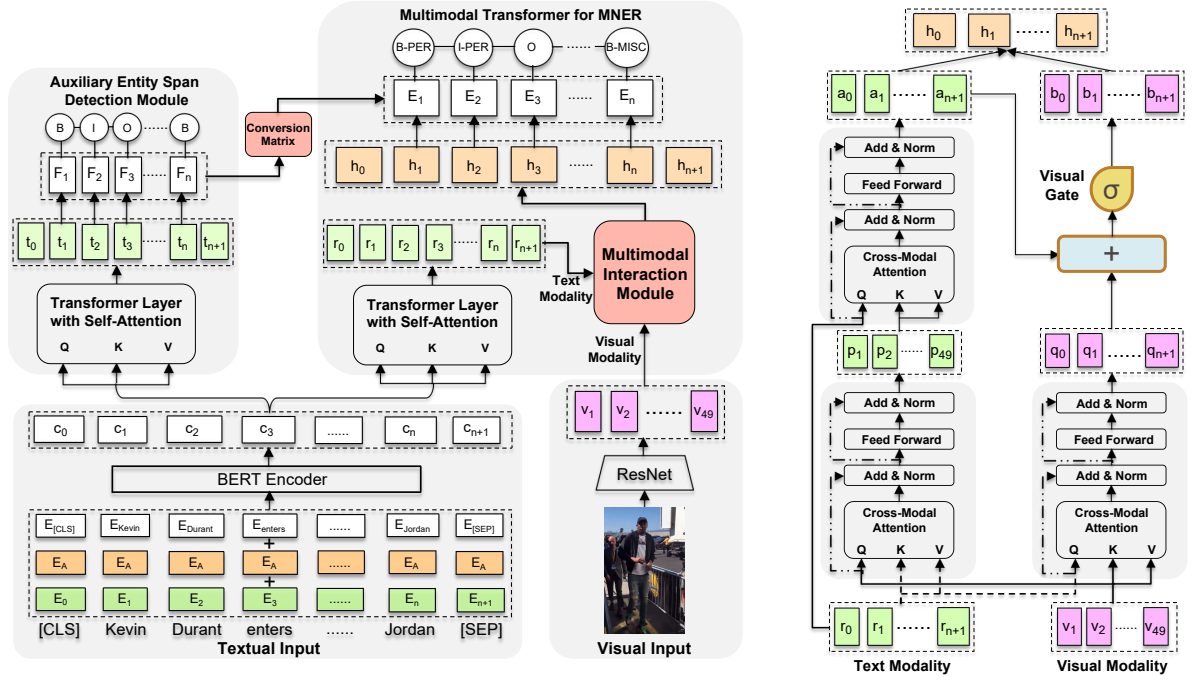


Figure 2: (a). Overall Architecture of Our Unified Multimodal Transformer. (b). Multimodal Interaction (MMI) Module.

As shown at the bottom of Fig. 2.a, we first extract contextualized word representations and visual block representations from the input sentence and the input image, respectively.

The right part of Fig. 2.a illustrates our Multimodal Transformer model for MNER. Specifically, a Transformer layer is first employed to derive each word’s textual hidden representation. Next, a multimodal interaction (MMI) module is devised to fully capture the inter-modality dynamics between the textual hidden representations and the visual block representations. The hidden representations from MMI are then fed to a conditional random field (CRF) layer to produce the label for each word.

To alleviate the visual bias in MNER, we further stack a purely text-based ESD module in the left part of Fig. 2.a, where we feed its hidden representations to another CRF layer to predict each word’s entity span label. More importantly, to utilize this for our main MNER task, we design a conversion matrix to encode the dependency relations between corresponding labels from ESD to MNER, so that the entity span predictions from ESD can be integrated to get the final MNER label for each word.

2.2 Unimodal Input Representations

Word Representations: Due to the capability of giving different representations for the same word in different contexts, we employ the recent contextualized representations from BERT (Devlin et al.,

2018) as our sentence encoder. Following Devlin et al. (2018), each input sentence is preprocessed by inserting two special tokens, i.e., appending [CLS] to the beginning and [SEP] to the end, respectively. Formally, let $S' = (s_0, s_1, \dots, s_{n+1})$ be the modified input sentence, where s_0 and s_{n+1} denote the two inserted tokens. Let $\mathbf{X} = (\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{n+1})$ be the word representations of S' , where \mathbf{x}_i is the sum of word, segment, and position embeddings for each token s_i . As shown in the bottom left of Fig. 2.a, \mathbf{X} is then fed to the BERT encoder to obtain $\mathbf{C} = (c_0, c_1, \dots, c_{n+1})$, where $c_i \in \mathbb{R}^d$ is the generated contextualized representation for \mathbf{x}_i .

Visual Representations: As one of the state-of-the-art CNN models for image recognition, Residual Network (ResNet) (He et al., 2016) has shown its capability of extracting meaningful feature representations of the input image in its deep layers. We therefore keep the output from the last convolutional layer in a pretrained 152-layer ResNet to represent each image, which essentially splits each input image into $7 \times 7 = 49$ visual blocks with the same size and represents each block with a 2048-dimensional vector. Specifically, given an input image V , we first resize it to 224×224 pixels, and obtain its visual representations from ResNet, denoted as $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{49})$, where \mathbf{u}_i is the 2048-dimensional vector representation for the i -th visual block. To project the visual representations into the same space of the word representations,

we further convert \mathbf{U} with a linear transformation: $\mathbf{V} = \mathbf{W}_u^\top \mathbf{U}$, where $\mathbf{W}_u \in \mathbb{R}^{2048 \times d}$ is the weight matrix¹. As shown in the bottom right of Fig. 2.a, $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{49})$ is the visual representations generated from ResNet.

2.3 Multimodal Transformer for MNER

In this subsection, we present our proposed Multimodal Transformer for MNER.

As illustrated on the right of Fig. 2.a, we first add a standard Transformer layer over \mathbf{C} to obtain each word’s textual hidden representation: $\mathbf{R} = (\mathbf{r}_0, \mathbf{r}_1, \dots, \mathbf{r}_{n+1})$, where $\mathbf{r}_i \in \mathbb{R}^d$ denotes the generated hidden representation for \mathbf{x}_i .

Motivation: While the above Transformer layer can capture which context words are more relevant to the prediction of an input word \mathbf{x}_i , they fail to consider the associated visual context. On the one hand, due to the short length of textual contents on social media, the additional visual context may guide each word to learn better word representations. On the other hand, since each visual block is often closely related to several input words, incorporating the visual block representation can potentially make the prediction of its related words more accurately. Inspired by these observations, we propose a multimodal interaction (MMI) module to learn an image-aware word representation and a word-aware visual representation for each word.

2.3.1 Image-Aware Word Representation

Cross-Modal Transformer (CMT) Layer: As shown on the left of Fig. 2.b, to learn better word representations with the guidance of associated images, we first employ an m -head cross-modal attention mechanism (Tsai et al., 2019), by treating $\mathbf{V} \in \mathbb{R}^{d \times 49}$ as queries, and $\mathbf{R} \in \mathbb{R}^{d \times (n+1)}$ as keys and values:

$$\text{CA}_i(\mathbf{V}, \mathbf{R}) = \text{softmax}\left(\frac{[\mathbf{W}_{q_i} \mathbf{V}]^\top [\mathbf{W}_{k_i} \mathbf{R}]}{\sqrt{d/m}}\right) [\mathbf{W}_{v_i} \mathbf{R}]^\top;$$

$$\text{MH-CA}(\mathbf{V}, \mathbf{R}) = \mathbf{W}' [\text{CA}_1(\mathbf{V}, \mathbf{R}), \dots, \text{CA}_m(\mathbf{V}, \mathbf{R})]^\top,$$

where CA_i refers to the i -th head of cross-modal attention, $\{\mathbf{W}_{q_i}, \mathbf{W}_{k_i}, \mathbf{W}_{v_i}\} \in \mathbb{R}^{d/m \times d}$, and $\mathbf{W}' \in \mathbb{R}^{d \times d}$ denote the weight matrices for the query, key, value, and multi-head attention, respectively. Next, we stack another three sub-layers on top:

$$\tilde{\mathbf{P}} = \text{LN}(\mathbf{V} + \text{MH-CA}(\mathbf{V}, \mathbf{R})); \quad (1)$$

$$\mathbf{P} = \text{LN}(\tilde{\mathbf{P}} + \text{FFN}(\tilde{\mathbf{P}})), \quad (2)$$

¹Bias terms are omitted to avoid confusion in this paper.

where FFN is the feed-forward network (Vaswani et al., 2017), LN is the layer normalization (Ba et al., 2016), and $\mathbf{P} = (\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_{49})$ is the output representations of the CMT layer.

Coupled CMT Layer: However, since the visual representations are treated as queries in the above CMT layer, each generated vector \mathbf{p}_i is corresponding to the i -th visual block instead of the i -th input word. Ideally, the image-aware word representation should be corresponding to each word.

To address this, we propose to couple \mathbf{P} with another CMT layer, which treats the textual representations \mathbf{R} as queries, and \mathbf{P} as keys and values. As shown in the top left of Fig. 2.a, this coupled CMT layer generates the final image-aware word representations, denoted by $\mathbf{A} = (\mathbf{a}_0, \mathbf{a}_1, \dots, \mathbf{a}_{n+1})$.

2.3.2 Word-Aware Visual Representation

To obtain a visual representation for each word, it is necessary to align each word with its closely related visual blocks, i.e., assigning high/low attention weights to its related/unrelated visual blocks. Hence, as shown in the right part of Fig. 2.b, we use a CMT layer by treating \mathbf{R} as queries and \mathbf{V} as keys and values, which can be considered as a symmetric version of the left CMT layer. Finally, it generates the word-aware visual representations, denoted by $\mathbf{Q} = (\mathbf{q}_0, \mathbf{q}_1, \dots, \mathbf{q}_{n+1})$.

Visual Gate: As pointed out in some previous studies (Zhang et al., 2018; Lu et al., 2018), it is unreasonable to align many function words such as *the*, *of*, and *well* with any visual block. Therefore, it is important to incorporate a visual gate to dynamically control the contribution of visual features. Following the practice in previous work, we design a visual gate by combining the information from the above word representations \mathbf{A} and visual representations \mathbf{Q} as follows:

$$\mathbf{g} = \sigma(\mathbf{W}_a^\top \mathbf{A} + \mathbf{W}_q^\top \mathbf{Q}), \quad (3)$$

where $\{\mathbf{W}_a, \mathbf{W}_q\} \in \mathbb{R}^{d \times d}$ are weight matrices, and σ is the element-wise sigmoid function. Based on the gate output, we can obtain the final word-aware visual representations as $\mathbf{B} = \mathbf{g} \cdot \mathbf{Q}$.

2.3.3 CRF Layer

To integrate the word and the visual representations, we concatenate \mathbf{A} and \mathbf{B} to obtain the final hidden representations $\mathbf{H} = (\mathbf{h}_0, \mathbf{h}_1, \dots, \mathbf{h}_{n+1})$, where $\mathbf{h}_i \in \mathbb{R}^{2d}$. Following Lample et al. (2016), we then feed \mathbf{H} to a standard CRF layer, which defines the

probability of the label sequence \mathbf{y} given the input sentence S and its associated image V :

$$P(\mathbf{y}|S, V) = \frac{\exp(\text{score}(\mathbf{H}, \mathbf{y}))}{\sum_{\mathbf{y}'} \exp(\text{score}(\mathbf{H}, \mathbf{y}'))}; \quad (4)$$

$$\text{score}(\mathbf{H}, \mathbf{y}) = \sum_{i=0}^n T_{y_i, y_{i+1}} + \sum_{i=1}^n E_{\mathbf{h}_i, y_i}; \quad (5)$$

$$E_{\mathbf{h}_i, y_i} = \mathbf{w}_{\text{MNER}}^{y_i} \cdot \mathbf{h}_i, \quad (6)$$

where $T_{y_i, y_{i+1}}$ is the transition score from the label y_i to the label y_{i+1} , $E_{\mathbf{h}_i, y_i}$ is the emission score of the label y_i for the i -th word, and $\mathbf{w}_{\text{MNER}}^{y_i} \in \mathbb{R}^{2d}$ is the weight parameter specific to y_i .

2.4 Unified Multimodal Transformer

Motivation: Since the Multimodal Transformer presented above mainly focuses on modeling the interactions between text and images, it may lead the learnt model to overemphasize the entities highlighted by the image but ignore the remaining entities. To alleviate the bias, we propose to leverage text-based entity span detection (ESD) as an auxiliary task based on the following observation. As ResNet is pre-trained on ImageNet (Deng et al., 2009) for the image recognition task, its high-level representations are closely relevant to the final predictions, i.e., the types of contained objects. This indicates that the visual representations from ResNet should be quite useful for identifying types of the detected entities, but are not necessarily relevant to detecting entity spans in the sentence. Therefore, we use purely text-based ESD to guide the final predictions for our main MNER task.

Auxiliary Entity Span Detection Module: Formally, we model ESD as another sequence labeling task, and use $\mathbf{z} = (z_1, \dots, z_n)$ to denote the sequence of labels, where $z_i \in \mathcal{Z}$ and $\mathcal{Z} = \{\mathbf{B}, \mathbf{I}, \mathbf{O}\}$.

As shown in the left part of Fig. 2.a, we employ another Transformer layer to obtain its specific hidden representations as $\mathbf{T} = (\mathbf{t}_0, \mathbf{t}_1, \dots, \mathbf{t}_{n+1})$, followed by feeding it to a CRF layer to predict the probability of the label sequence \mathbf{z} given S :

$$P(\mathbf{z}|S) = \frac{\exp(\sum_{i=0}^n T_{z_i, z_{i+1}} + \sum_{i=1}^n \mathbf{w}_{\text{ESD}}^{z_i} \cdot \mathbf{t}_i)}{\sum_{\mathbf{z}'} \exp(\sum_{i=0}^n T_{z'_i, z'_{i+1}} + \sum_{i=1}^n \mathbf{w}_{\text{ESD}}^{z'_i} \cdot \mathbf{t}_i)},$$

where $\mathbf{w}_{\text{ESD}}^{z_i} \in \mathbb{R}^d$ is the parameter specific to z_i .

Conversion Matrix: Although ESD is modeled as an auxiliary task separated from MNER, the two tasks are highly correlated since each ESD label should be only corresponding to a subset of labels in MNER. For example, given the sentence in Fig. 2.a, if the first token is predicted to be the

Entity Type	TWITTER-2015			TWITTER-2017		
	Train	Dev	Test	Train	Dev	Test
Person	2217	552	1816	2943	626	621
Location	2091	522	1697	731	173	178
Organization	928	247	839	1674	375	395
Miscellaneous	940	225	726	701	150	157
Total	6176	1546	5078	6049	1324	1351
Num of Tweets	4000	1000	3257	3373	723	723

Table 1: The basic statistics of our two Twitter datasets.

beginning of an entity in ESD (i.e., have the label \mathbf{B}), it should be also the beginning of a typed entity in MNER (e.g., have the label $\mathbf{B-PER}$).

To encode such inter-task correspondence, we propose to use a conversion matrix $\mathbf{W}^c \in \mathbb{R}^{|\mathcal{Z}| \times |\mathcal{Y}|}$, where each element $\mathbf{W}_{j,k}^c$ defines the conversion probability from \mathcal{Z}_j to \mathcal{Y}_k . Since we have some prior knowledge (e.g., the label \mathbf{B} can only convert to a label subset $\{\mathbf{B-PER}, \mathbf{B-LOC}, \mathbf{B-ORG}, \mathbf{B-MISC}\}$), we initialize \mathbf{W}^c as follows: if \mathcal{Z}_j is not corresponding to \mathcal{Y}_k , $\mathbf{W}_{j,k}^c$ is set to 0; otherwise, $\mathbf{W}_{j,k}^c$ is set to $\frac{1}{|\mathcal{C}_j|}$, where \mathcal{C}_j denotes a subset of \mathcal{Y} that is corresponding to \mathcal{Z}_j .

Modified CRF Layer for MNER: After obtaining the conversion matrix, we further propose to fully leverage the text-based entity span predictions to guide the final predictions of MNER. Specifically, we modify the CRF layer for MNER by incorporating the entity span information from ESD into the emission score defined in Eqn. (6):

$$E_{\mathbf{h}_i, y_i} = \mathbf{w}_{\text{MNER}}^{y_i} \cdot \mathbf{h}_i + \mathbf{w}_{\text{ESD}}^{z_i} \cdot \mathbf{t}_i \cdot \mathbf{W}_{z_i, y_i}^c. \quad (7)$$

2.5 Model Training

Given a set of manually labeled training samples $D = \{S^j, V^j, \mathbf{y}^j, \mathbf{z}^j\}_{j=1}^N$, our overall training objective function is a weighted sum of the sentence-level negative log-likelihood losses for our main MNER task and the auxiliary ESD task²:

$$\mathcal{L} = -\frac{1}{|D|} \sum_{j=1}^N (\log P(\mathbf{y}^j | S^j, V^j) + \lambda \log P(\mathbf{z}^j | S^j)),$$

where λ is a hyperparameter to control the contribution of the auxiliary ESD module.

3 Experiments

We conduct experiments on two multimodal NER datasets, comparing our Unified Multimodal Transformer (UMT) with a number of unimodal and multimodal approaches.

²We obtain \mathbf{z}^j by removing the type information in \mathbf{y}^j .

Modality	Methods	TWITTER-2015							TWITTER-2017						
		Single Type (F1)				Overall			Single Type (F1)				Overall		
		PER.	LOC.	ORG.	MISC.	P	R	F1	PER.	LOC.	ORG.	MISC.	P	R	F1
Text	BiLSTM-CRF	76.77	72.56	41.33	26.80	68.14	61.09	64.42	85.12	72.68	72.50	52.56	79.42	73.43	76.31
	CNN-BiLSTM-CRF	80.86	75.39	47.77	32.61	66.24	68.09	67.15	87.99	77.44	74.02	60.82	80.00	78.76	79.37
	HBiLSTM-CRF	82.34	76.83	51.59	32.52	70.32	68.05	69.17	87.91	78.57	76.67	59.32	82.69	78.16	80.37
	BERT	84.72	79.91	58.26	38.81	68.30	74.61	71.32	90.88	84.00	79.25	61.63	82.19	83.72	82.95
	BERT-CRF	84.74	80.51	60.27	37.29	69.22	74.59	71.81	90.25	83.05	81.13	62.21	83.32	83.57	83.44
Text+Image	GVATT-HBiLSTM-CRF	82.66	77.21	55.06	35.25	73.96	67.90	70.80	89.34	78.53	79.12	62.21	83.41	80.38	81.87
	AdaCAN-CNN-BiLSTM-CRF	81.98	78.95	53.07	34.02	72.75	68.74	70.69	89.63	77.46	79.24	62.77	84.16	80.24	82.15
	GVATT-BERT-CRF	84.43	80.87	59.02	38.14	69.15	74.46	71.70	90.94	83.52	81.91	62.75	83.64	84.38	84.01
	AdaCAN-BERT-CRF	85.28	80.64	59.39	38.88	69.87	74.59	72.15	90.20	82.97	82.67	64.83	85.13	83.20	84.10
	MT-BERT-CRF (Ours)	85.30	81.21	61.10	37.97	70.48	74.80	72.58	91.47	82.05	81.84	65.80	84.60	84.16	84.42
	UMT-BERT-CRF (Ours)	85.24	81.58†	63.03†	39.45†	71.67	75.23	73.41†	91.56†	84.73†	82.24	70.10†	85.28	85.34	85.31†

Table 2: Performance comparison on our two TWITTER datasets. † indicates that **UMT-BERT-CRF** is significantly better than GVATT-BERT-CRF and AdaCAN-BERT-CRF with p-value < 0.05 based on paired t-test.

3.1 Experiment Settings

Datasets: We take two publicly available Twitter datasets respectively constructed by Zhang et al. (2018) and Lu et al. (2018) for MNER. Since the two datasets mainly include multimodal user posts published on Twitter during 2014-2015 and 2016-2017, we denote them as TWITTER-2015 and TWITTER-2017 respectively. Table 1 shows the number of entities for each type and the counts of multimodal tweets in the training, development, and test sets of the two datasets³. We have released the two datasets pre-processed by us for research purpose via this link: <https://github.com/jefferyYu/UMT>.

Hyperparameters: For each unimodal and multimodal approach compared in the experiments, the maximum length of the sentence input and the batch size are respectively set to 128 and 16. For our UMT approach, most hyperparameter settings follow Devlin et al. (2018) with the following exceptions: (1) the word representations \mathbf{C} are initialized with the cased $BERT_{base}$ model pre-trained by Devlin et al. (2018), and fine-tuned during training. (2) we employ a pre-trained 152-layer ResNet⁴ to initialize the visual representations \mathbf{U} and keep them fixed during training. (3) For the number of cross-modal attention heads, we set it as $m=12$. (4) The learning rate, the dropout rate, and the tradeoff parameter λ are respectively set to $5e-5$, 0.1, and 0.5, which can achieve the best performance on the development set of both datasets via a small grid search over the combinations of $[1e-5, 1e-4]$, $[0.1, 0.5]$, and $[0.1, 0.9]$.

³The TWITTER-2017 dataset released by Lu et al. (2018) is slightly different from the one used in their experiments, as they later remove a small portion of tweets for privacy issues.

⁴<https://download.pytorch.org/models/resnet152-b121ed2d.pth>.

3.2 Compared Systems

To demonstrate the effect of our Unified Multimodal Transformer (UMT) model, we first consider a number of representative text-based approaches for NER: (1) *BiLSTM-CRF* (Huang et al., 2015), a pioneering study which eliminates the heavy reliance on hand-crafted features, and simply employs a bidirectional LSTM model followed by a CRF layer for each word’s final prediction; (2) *CNN-BiLSTM-CRF* (Ma and Hovy, 2016), a widely adopted neural network model for NER, which is an improvement of *BiLSTM-CRF* by replacing each word’s word embedding with the concatenation of its word embedding and CNN-based character-level word representations; (3) *HBiLSTM-CRF* (Lample et al., 2016), an end-to-end hierarchical LSTM architectures, which replaces the bottom CNN layer in *CNN-BiLSTM-CRF* with an LSTM layer to obtain the character-level word representations; (4) *BERT* (Devlin et al., 2018), a multi-layer bidirectional Transformer encoder, which gives contextualized representations for each word, followed by stacking a softmax layer for final predictions; (5) *BERT-CRF*, a variant of *BERT* by replacing the softmax layer with a CRF layer.

Besides, we also consider several competitive multimodal approaches for MNER: (1) *GVATT-HBiLSTM-CRF* (Lu et al., 2018), a state-of-the-art approach for MNER, which integrates *HBiLSTM-CRF* with the visual context by proposing a visual attention mechanism followed by a visual gate to obtain word-aware visual representations; (2) *AdaCAN-CNN-BiLSTM-CRF* (Zhang et al., 2018), another state-of-the-art approach based on *CNN-BiLSTM-CRF*, which designs an adaptive co-attention network to induce word-aware visual representations for each word; (3) *GVATT-BERT-CRF*

Methods	TWITTER-2015			TWITTER-2017		
	P	R	F1	P	R	F1
UMT-BERT-CRF	71.67	75.23	73.41	85.28	85.34	85.31
w/o ESD Module	70.48	74.80	72.58	84.60	84.16	84.42
w/o Conversion Matrix	70.43	74.98	72.63	84.72	84.97	84.85
w/o Image-Aware WR	70.33	75.44	72.79	83.83	85.94	84.87
w/o Visual Gate	71.34	75.15	73.19	85.31	84.68	84.99

Table 3: Ablation Study of Unified Multimodal Transformer.

and *AdaCAN-BERT-CRF*, our two variants of the above two multimodal approaches, which replace the sentence encoder with *BERT*; (4) *MT-BERT-CRF*, our Multimodal Transformer model introduced in Section 2.3; (5) *UMT-BERT-CRF*, our unified architecture by incorporating the auxiliary entity span detection module into Multimodal Transformer, as introduced in Section 2.4.

All the neural models are implemented with PyTorch, and all the experiments are conducted on *NVIDIA RTX 2080 Ti* GPUs.

3.3 Main Results

In Table 2, we report the precision (**P**), recall (**R**), and F1 score (**F1**) achieved by each compared method on our two Twitter datasets.

First, comparing all the text-based approaches, we can clearly observe that *BERT* outperforms the other compared methods with a significant margin on both datasets. Moreover, it is easy to see that empowering *BERT* with a CRF layer can further boost the performance. All these observations indicate that the contextualized word representations are indeed quite helpful for the NER task on social media texts, due to the context-aware characteristics. This agrees with our first motivation.

Second, comparing the state-of-the-art multimodal approaches with their corresponding unimodal baselines, we can find that the multimodal approaches can generally achieve better performance, which demonstrates that incorporating the visual context is generally useful for NER. Besides, we can see that although *GVATT-HBiLSTM-CRF* and *AdaCAN-CNN-BiLSTM-CRF* can significantly outperform their unimodal baselines, the performance gains become relatively limited when replacing their sentence encoder with *BERT*. This suggests the challenge and the necessity of proposing a more effective multimodal approach.

Third, in comparison with the two existing multimodal methods, our Multimodal Transformer *MT-BERT-CRF* outperforms the state-of-the-art by 2.5% and 2.8% respectively, and also achieves bet-

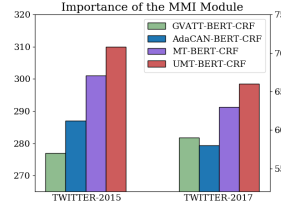


Figure 3: The number of entities (shown in y-axis) that are **incorrectly** predicted by BERT-CRF, but **get corrected** by each multimodal method

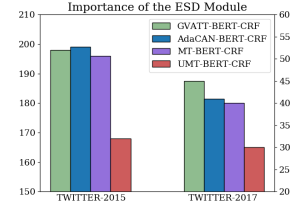


Figure 4: The number of entities (shown in y-axis) that are **correctly** predicted by BERT-CRF, but **wrongly predicted** by each multimodal method

ter performance than their BERT variants. We conjecture that the performance gains mainly come from the following reason: the two multimodal methods only focus on obtaining word-aware visual representations, whereas our *MT-BERT-CRF* approach targets at generating both image-aware word representations and word-aware visual representations for each word. These observations are in line with our second motivation.

Finally, comparing all the unimodal and multimodal approaches, it is clear to observe that our Unified Multimodal Transformer (i.e., *UMT-BERT-CRF*) can achieve the best performance on both datasets, outperforming the second best methods by 1.14% and 1.05%, respectively. This demonstrates the usefulness of the auxiliary entity span detection module, and indicates that the auxiliary module can help our Multimodal Transformer alleviate the bias brought by the associated images, which agrees with our third motivation.

3.4 Ablation Study

To investigate the effectiveness of each component in our Unified Multimodal Transformer (UMT) architecture, we perform comparison between the full UMT model and its ablations with respect to the auxiliary entity span detection (ESD) module and the multimodal interaction (MMI) module.

As shown in Table 3, we can see that all the components in UMT make important contributions to the final results. On the one hand, removing the whole ESD module will significantly drop the performance, which shows the importance of alleviating the visual bias. In particular, discarding the conversion matrix in the ESD module also leads to the performance drop, which indicates the usefulness of capturing the label correspondence between the auxiliary module and our main MNER task.

On the other hand, as the main contribution of





Importance of the MMI Module	Importance of the ESD Module	Importance of Associated Images	Noise of Associated Images
			
A. Review of [Wolf Hall MISC] ¹ , Episode 1 : Three Card Trick (bit.ly/1BHnWNb) #[WolfHall MISC] ²	B. [Kevin Love PER] ¹ was more excited about [GameofThrones MISC] ² than beating the [Hawks ORG] ³	C. My mum took some awesome photos of @ [iamrationale PER] ¹ and @ [bastilledan PER] ² .	D. Ask [Siri MISC] ¹ what 0 divided by 0 is and watch her put you in your place.
BERT-CRF: 1-LOC✗, 2-LOC✗	1-PER✓, 2-MISC✓, 3-ORG✓	1-MISC✗, 2-ORG✗	1-MISC✓
AdaCAN-BERT-CRF: 1-LOC✗, 2-LOC✗	1-PER✓, 2-NONE✗, 3-ORG✓	1-PER✓, 2-PER✓	1-PER✗
MT-BERT-CRF: 1-MISC✓, 2-MISC✓	1-PER✓, 2-NONE✗, 3-ORG✓	1-PER✓, 2-PER✓	1-PER✗
UMT-BERT-CRF: 1-MISC✓, 2-MISC✓	1-PER✓, 2-MISC✓, 3-ORG✓	1-PER✓, 2-PER✓	1-PER✗

Table 4: The second row shows several representative samples together with their manually labeled entities in the test set of our two TWITTER datasets, and the bottom four rows show predicted entities of different methods on these test samples.

our MMI module, Image-Aware Word Representations (WR) demonstrates its indispensable role in the final performance due to the moderate performance drop after removal. Besides, removing the visual gate also results in minor performance drop, indicating its importance to the full model.

3.5 Further Analysis

Importance of MMI and ESD Modules: To better appreciate the importance of two main contributions (i.e., MMI and ESD modules) in our proposed approaches, we conduct additional analysis on our two test sets. In Fig. 3 and Fig. 4, we show the number of entities that are wrongly/correctly predicted by BERT-CRF, but correctly/wrongly predicted by each multimodal method⁵.

First, we can see from Fig. 3 that with the MMI module, our *MT-BERT-CRF* and *UMT-BERT-CRF* approaches correctly identify more entities, compared with the two multimodal baselines. Table 4.A shows a specific example. We can see that our two methods correctly classify the type of *Wolf Hall* as *MISC* whereas the compared systems wrongly predict its type as *LOC*, probably because our MMI module enforces the image-aware word representations of *Wolf Hall* to be closer to drama names.

Second, in Fig. 4, it is clear to observe that compared with the other three methods, *UMT-BERT-CRF* can significantly decrease the bias brought by the visual context due to incorporating our auxiliary ESD module. In Table 4.B, we show a concrete example: since *Game of Thrones* is ignored by the image, the two multimodal baselines fail to identify them; in contrast, with the help of the auxiliary

ESD module, *UMT-BERT-CRF* successfully eliminates the bias.

Effect of Incorporating Images: To obtain a better understanding of the general effect of incorporating associated images into our MNER task, we carefully examine our test sets and choose two representative test samples to compare the prediction results of different approaches.

First, we observe that most improvements gained by multimodal methods come from those samples where the textual contents are informal or incomplete but the visual context provides useful clues. For example, in Table 4.C, we can see that without the visual context, *BERT-CRF* fails to identify that the two entities refer to two singers in the concert, but all the multimodal approaches can correctly classify their types after incorporating the image.

Second, by manually checking the test set of our two datasets, we find that in around 5% of the social media posts, the associated images might be irrelevant to the textual contents due to two kinds of reasons: (1) these posts contain image memes, cartoons, or photos with metaphor; (2) their images and textual contents reflect different aspects of the same event. In such cases, we observe that multimodal approaches generally perform worse than *BERT-CRF*. A specific example is given in Table 4.D, where all the multimodal methods wrongly classify *Siri* as *PER* because of the unrelated face in the image.

4 Related Work

As a crucial component of many information extraction tasks including entity linking (Derczynski et al., 2015), opinion mining (Maynard et al., 2012), and event detection (Ritter et al., 2012), named

⁵Note that here we use strict matches (i.e., correct span and type predictions).

entity recognition (NER) has attracted much attention in the research community in the past two decades (Li et al., 2018).

Methods for NER: In the literature, various supervised learning approaches have been proposed for NER. Traditional approaches typically focus on designing various effective NER features, followed by feeding them to different linear classifiers such as maximum entropy, conditional random fields (CRFs), and support vector machines (Chieu and Ng, 2002; Florian et al., 2003; Finkel et al., 2005; Ratnov and Roth, 2009; Lin and Wu, 2009; Passos et al., 2014; Luo et al., 2015). To reduce the feature engineering efforts, a number of recent studies proposed to couple different neural network architectures with a CRF layer (Lafferty et al., 2001) for word-level predictions, including convolutional neural networks (Collobert et al., 2011), recurrent neural networks (Chiu and Nichols, 2016; Lample et al., 2016), and their hierarchical combinations (Ma and Hovy, 2016). These neural approaches have been shown to achieve the state-of-the-art performance on different benchmark datasets based on formal text (Yang et al., 2018).

However, when applying these approaches to social media text, most of them fail to achieve satisfactory results. To address this issue, many studies proposed to exploit external resources (e.g., shallow parser, Freebase dictionary, and orthographic characteristics) to incorporate a set of tweet-specific features into both traditional approaches (Ritter et al., 2011; Li et al., 2014; Baldwin et al., 2015) and recent neural approaches (Limsopatham and Collier, 2016; Lin et al., 2017), which can obtain much better performance on social media text.

Methods for Multimodal NER (MNER): As multimodal data become increasingly popular on social media platforms, several recent studies focus on the MNER task, where the goal is to leverage the associate images to better identify the named entities contained in the text. Specifically, Moon et al. (2018) proposed a multimodal NER network with modality attention to fuse the textual and visual information. To model the inter-modal interactions and filter out the noise in the visual context, Zhang et al. (2018) and Lu et al. (2018) respectively proposed an adaptive co-attention network and a gated visual attention mechanism for MNER. In this work, we follow this line of work. But different

from them, we aim to propose an effective multimodal method based on the recent Transformer architecture (Vaswani et al., 2017). To the best of our knowledge, this is the first work to apply Transformer to the task of MNER.

5 Conclusion

In this paper, we first presented a Multimodal Transformer architecture for the task of MNER, which captures the inter-modal interactions with a multimodal interaction module. Moreover, to alleviate the bias of the visual context, we further proposed a Unified Multimodal Transformer (UMT), which incorporates an entity span detection module to guide the final predictions for MNER. Experimental results show that our UMT approach can consistently achieve the best performance on two benchmark datasets.

There are several future directions for this work. On the one hand, despite bringing performance improvements over existing MNER methods, our UMT approach still fails to perform well on social media posts with unmatched text and images, as analyzed in Section 3.5. Therefore, our next step is to enhance UMT so as to dynamically filter out the potential noise from images. On the other hand, since the size of existing MNER datasets is relatively small, we plan to leverage the large amount of unlabeled social media posts in different platforms, and propose an effective framework to combine them with the small amount of annotated data to obtain a more robust MNER model.

Acknowledgments

We would like to thank three anonymous reviewers for their valuable comments. This research is supported by the National Research Foundation, Singapore under its International Research Centres in Singapore Funding Initiative, and the Natural Science Foundation of China under Grant 61672288. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

- Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135.
- Hai Leong Chieu and Hwee Tou Ng. 2002. Named entity recognition: a maximum entropy approach using global information. In *Proceedings of COLING*.
- Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.
- Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2461–2505.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255.
- Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke Van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. 2015. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL*.
- Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named entity recognition through classifier combination. In *Proceedings of NAACL*.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of CVPR*, pages 770–778.
- Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.
- John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*.
- Chenliang Li, Aixin Sun, Jianshu Weng, and Qi He. 2014. Tweet segmentation and its application to named entity recognition. *IEEE Transactions on knowledge and data engineering*, 27(2):558–570.
- Chenliang Li, Jianshu Weng, Qi He, Yuxia Yao, Anwitaman Datta, Aixin Sun, and Bu-Sung Lee. 2012. Twiner: named entity recognition in targeted twitter stream. In *Proceedings of SIGIR*.
- Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2018. A survey on deep learning for named entity recognition. *arXiv preprint arXiv:1812.09449*.
- Nut Limsopatham and Nigel Collier. 2016. Bidirectional LSTM for named entity recognition in twitter messages. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*.
- Bill Yuchen Lin, Frank F Xu, Zhiyi Luo, and Kenny Zhu. 2017. Multi-channel bilstm-crf model for emerging named entity recognition in social media. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*.
- Dekang Lin and Xiaoyun Wu. 2009. Phrase clustering for discriminative learning. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*.
- 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 ACL*, pages 1990–1999.
- Gang Luo, Xiaojiang Huang, Chin-Yew Lin, and Zaiqing Nie. 2015. Joint entity recognition and disambiguation. In *Proceedings of EMNLP*.
- Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of ACL*.
- Diana Maynard, Kalina Bontcheva, and Dominic Rout. 2012. Challenges in developing opinion mining tools for social media. In *Proceedings of the @ NLP can u tag usergeneratedcontent*.
- Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity recognition for short social media posts. In *Proceedings of NAACL*.
- Alexandre Passos, Vineet Kumar, and Andrew McCallum. 2014. Lexicon infused phrase embeddings for named entity resolution. In *Proceedings of CoNLL*.
- Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of CoNLL*.

- Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of ACL*.
- Alan Ritter, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from twitter. In *Proceedings of SIGKDD*.
- Erik F Sang and Jorn Veenstra. 1999. Representing text chunks. In *Proceedings of EACL*.
- Kentaro Torisawa et al. 2007. Exploiting wikipedia as external knowledge for named entity recognition. In *Proceedings of EMNLP-CoNLL*.
- Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of ACL*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NIPS*, pages 5998–6008.
- Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of COLING*.
- Jie Yang, Shuailong Liang, and Yue Zhang. 2018. Design challenges and misconceptions in neural sequence labeling. In *Proceedings of COLING*.
- Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of AAAI*, pages 5674–5681.
- GuoDong Zhou and Jian Su. 2002. Named entity recognition using an hmm-based chunk tagger. In *Proceedings of ACL*.