Multimodal Representation with Embedded Visual Guiding Objects for Named Entity Recognition in Social Media Posts

Zhiwei Wu*
Changmeng Zheng*
zhiwei.w@qq.com
sethecharm@mail.scut.edu.cn
Key Laboratory of Big Data and Intelligent Robot (South China University of Technology), Ministry of Education
School of Software Engineering, South China University of Technology
Guangzhou, Guangdong, China

Yi Cai
Key Laboratory of Big Data and Intelligent Robot (South China University of Technology), Ministry of Education
School of Software Engineering, South China University of Technology
Guangzhou, Guangdong, China
ycai@scut.edu.cn

Junying Chen
Key Laboratory of Big Data and Intelligent Robot (South China University of Technology), Ministry of Education
School of Software Engineering, South China University of Technology
Guangzhou, Guangdong, China
jychense@scut.edu.cn

Ho-fung Leung
Department of Computer Science and Engineering, The Chinese University of Hong Kong
Hong Kong, China
lhf@cuhk.edu.hk

Qing Li
Department of Computing, The Hong Kong Polytechnic University
Hong Kong, China
qing-prof.li@polyu.edu.hk

ABSTRACT

Visual contexts often help to recognize named entities more precisely in short texts such as tweets or snapchat. For example, one can identify "Charlie" as a name of a dog according to the user posts. Previous works on multimodal named entity recognition ignore the corresponding relations of visual objects and entities. Visual objects are considered as fine-grained image representations. For a sentence with multiple entity types, objects of the relevant image can be utilized to capture different entity information. In this paper, we propose a neural network which combines object-level image information and character-level text information to predict entities. Vision and language are bridged by leveraging object labels as embeddings, and a dense co-attention mechanism is introduced for fine-grained interactions. Experimental results in Twitter dataset demonstrate that our method outperforms the state-of-the-art methods.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; • **Computing methodologies** → **Information extraction**.

*Both authors contributed equally to this research.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413650

KEYWORDS

multimodal named entity recognition; modality gap; fine-grained image representations

ACM Reference Format:
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. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413650

1 INTRODUCTION

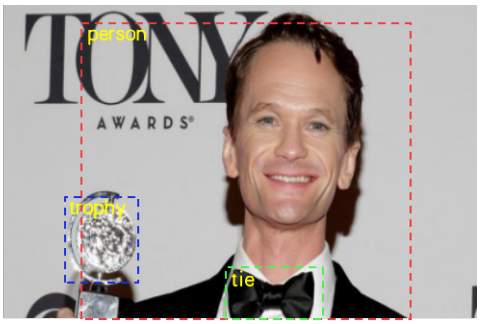
Named entity recognition (NER) is a task which locates and classifies named entities into predefined categories such as location, organization and person name. Previous works on NER either rely on hand-crafted features [8, 14] or leverage neural networks on distributed representation of texts [15, 17]. And most works on NER concern about the newswire domain where language expressions are formal and complete [1, 26].

Unlike newswire domain, texts in social media provide abundant user-generated information for understanding events, opinions and preferences of groups and individuals. Despite the impressive progress for newswire domain entity recognition, the methods there exhibit the following limitations for social media posts:

- Texts in social media are usually short, which is difficult in providing adequate information for determining entity types.
- Texts in social media are usually ambiguous because they contain slangs and polysemies.

Figure 1: An Example Multimodal Social Media Post from Twitter, where "Alibaba" is the Name of a Person Rather Than an Organization.

*Sentence: Actor **Neil Patrick Harris[PER]** will host the 2015 **Oscars[MISC]**.*



Selected objects in the image:
Person, Trophy, Tie
Relative objects and entities:
[PER] - Person [MISC] - Tie, Trophy

Figure 2: An Example of the Twitter Dataset. The Visual Object with Label "person" will Lead to the Detection of "Neil Patrick Harris" as PER Category, and Objects with "tie, trophy" will Lead to the Extraction of "Oscars" as the Name of an Award (MISC).

Such limitations pose major challenges in social media NER. On the other hand, in social media people like to share their life with texts and relevant images. Such visual information can assist extracting entities in the named entity recognition task. For example, in Figure 1, the term "Alibaba" appearing in tweets could be recognized as multiple types of entities such as "Organization" and "Person". Without the help of image information, identifying it in a correct entity category is non-trivial and difficult.

Language and vision provide complementary information. Meanwhile, an image related to a sentence can have different visual objects related to different entities in the sentence. For example, in Figure 2, the sentence contains two entities with two different types: one PER entity and one MISC entity. The detected visual object with label "person" is related to the PER entity "Neil Patrick Harris". The objects "tie, trophy" which are most relevant to awards are corresponding to the MISC entity "Oscar". Recent works on multimodal NER extract the features of the whole image which have only one semantic label [16, 18, 28]. Their methods simply combine the image features with the representation of each word in sentences. In such a case, their methods only reflect the relations between the whole image (rather than objects) and only one entity. The corresponding relations of multiple visual objects and different entities are ignored. As a result, the visual features of the whole image with only one semantic label may mislead their models to identify different type of entities into the same type. For example, the two entities "Neil Patrick Harris" (PER) and "Oscars" (MISC) in Figure 2 will be both identified into the same PER category incorrectly. To address this, it is necessary to leverage different visual objects (i.e., object-level features) to assist extracting entities with different types.

Since features of different modalities (vision and language) usually have inconsistent distribution and representation, simple concatenation [16, 18, 28] of their features may bring semantic disparity. The labels of visual objects have semantics of images and at the same time, they are in the same vector space as texts. Therefore, we consider to utilize the visual object labels and convert them into the embedding vector space as texts, so as to bridges the vision and language.

In this paper, we propose a neural network which incorporates object-level visual information with textual representations for NER in social media posts. We use a pre-trained object detector [10] to extract the visual objects. To address the problem of semantic disparity of different modality, we transform the object labels into word embeddings. Considering that previous co-attention models [28] only learn the correlation between each image and each text, which ignores the inner connections of visual objects or textual entities, we extend them into a dense co-attention network (DCAN). In particular, DCAN models the self-attention of objects or entities, as well as the guided attention between objects and entities. Finally, the fusion features of visual objects and textual entities are sent to a CRF layer to output entity labels.

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

- We propose a multimodal representation which combines object-level visual information and textual representations for NER in social media posts. Our method considers the corresponding relations of visual objects and textual entities, while previous works only reflect the relation of the whole image and only one entity. Object-level features contribute to extract entities with different types.
- We introduce a dense co-attention network to fuse the visual and textual representations. Our dense co-attention module can model the correlations between visual objects and textual entities, as well as inner connections of objects or entities, which helps extract entities precisely.
- We conduct experiments on the multimodal social media NER dataset, and the experimental results demonstrate that our model outperforms previous state-of-the-art methods.

2 RELATED WORK

2.1 NER in social media

NER has drawn attention of NLP researchers because several downstream NLP tasks rely on it [9, 29]. Neural models have been proposed and achieve state-of-the-art performance in variable datasets and domains. Recently, NER in social media domain has raised concerns since texts in social media are explosively growing and provide abundant user-generated information for various applications such as the identification of natural disasters [22, 24], cyber attack detection [12, 21] and breaking news aggregation [19]. Ritter et al. [20] propose a T-NER system which uses LabeledLDA to exploit Freebase dictionaries as a source of distant supervision. However, their method only identifies whether a span is an entity or not. Moon et al. [18] and Zhang et al. [28] leverage the visual information to help extract entities. However, image information is not fully exploited in their methods for the reason that one single image vector trained with only one label cannot assist recognizing multiple entities. Our model introduces object level representations, for focusing the attention on effective regions in images and entity-relevant objects can help extract the entities precisely.

2.2 Multimodal Representation

A lot of research has shown that combining textual and visual representation as multimodal representations can improve the performance of semantic tasks [4, 6]. Based on current literature, posterior combination strategies are most commonly used. The simplest way of combining visual and textual representations is concatenation [5, 13]. However, simple concatenation may bring semantic drift due to the vector space discrepancy of vision and language. Collell et al. [6] propose to learn a mapping function from text to vision. The outputs of the mapping themselves are used in the multimodal representations. Still, the mapping function is a bottleneck when the text and image are not relevant. In our approach, we propose to leverage the object labels as embeddings, and we bridge the vision and language by concatenating object embeddings with textual representations.

2.3 Attention Mechanism

Attention mechanism is widely used in a variety of deep learning tasks [2, 23, 28]. For multimodal NER (MNER) task, Moon et al. [18] propose a modality attention which focus on image, word and character level representation. Lu et al. [16] propose a visual attention module which take the salient visual regions into account. Zhang et al. [28] propose an adaptive co-attention model where text and image attentions are captured simultaneously. However, these attention models learn inter-correlations between two modalities (image and text), and neglect the intra-connections of visual objects or textual entities. This become a bottleneck for understanding fine-grained relationships of multimodal features. In contrast, we propose a dense co-attention mechanism to establish the complete interaction between visual objects and textual entities, so as to improve the NER performance.

3 THE PROPOSED METHOD

In this section, we present a novel neural model which combines object-level image representations and character-level textual representations. The overall architecture is shown as Figure 3. Our model is built upon a classic Bi-LSTM-CRF network, the object-aware gated attention module is applied before the CRF layer. For each input image, we extract objects with an object detector. The fine-grained object level features are utilized to assist recognizing different type of entities. The image features are in different vector space with textual embeddings, so we utilize the object labels. The object labels are transformed into object embeddings which are then fused with textual representations. We design a dense co-attention module to learn the inter- and intra-connections between visual objects and textual entities.

3.1 Feature Extractor

3.1.1 Visual Feature Extractor. Object-level features are considered as bottom-up attention in several multimodal tasks [2]. Different from previous multimodal representation methods, we bridge the vision and language by transforming object labels into object embeddings. In order to extract the objects from images, we utilize the pre-trained Mask RCNN [10] object detection model to recognize the objects in images. In most cases, only the salient objects of images are related to the entities mentioned in tweets. So we only consider the top K objects with the higher object classification scores as object labels , denoted as $v = (v_0, v_1, \ldots, v_k)$. Then, the object labels are transformed into object embeddings:

$$\widetilde{\mathbf{v}}_i = \mathbf{e}^v(v_i), \tag{1}$$

where $\mathbf{e}^v$ denotes an object embedding lookup table. Therefore, a object embedding can be represented as $\widetilde{\mathbf{v}} = \{\widetilde{\mathbf{v}}_0, \widetilde{\mathbf{v}}_1, \ldots, \widetilde{\mathbf{v}}_k\}$.

We map the object embeddings into new vectors with the same dimensions as the textual vector using a single layer perceptron for calculation convenience:

$$\mathbf{v} = tanh(W_I\widetilde{\mathbf{v}} + b_I), \tag{2}$$

where $\mathbf{W}_I$ and $\mathbf{b}_I$ are trainable parameters. The object embeddings are initialized with pre-trained word embeddings.

3.1.2 Textual Feature Extractor. Following the success of Lample et al. [15], Ma and Hovy [17] , we represent each word in a sentence by combining character embedding into word embedding. Given an input sentence with n words $s = (w_0, w_1, ..., w_n)$, each input word w_i is first embedded in latten space by word embedding:

$$\mathbf{x}_i^w = \mathbf{e}^w(x_i), \tag{3}$$

where $\mathbf{e}^w$ denotes a word embedding lookup table. We use pre-trained word embedding (the same setting as Zhang et al. [28]) to initialize it.

We capture the orthographic and morphological features of the word by integrating character representations. Previous works on social media posts have demonstrated that character embeddings can alleviate the serious OOV problem[18]. Denoting the representation of characters within w_i as $\mathbf{x}_i^c$, the embedding of each character within word w_i is denoted as $\mathbf{e}^c(c_j)$. $\mathbf{e}^c$ is the character embedding lookup which is initialized randomly. Then we feed

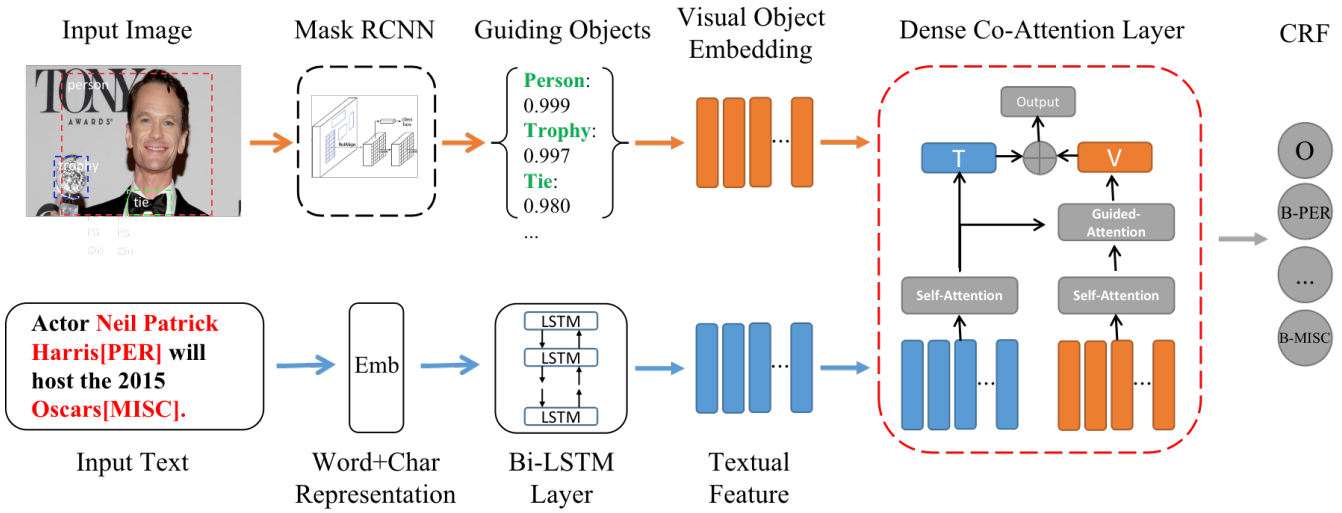


Figure 3: The Overall Architecture of Our Model. Our Model Combines the Object-level Visual Features and Character-level Word Representations to Predict Entities. A Dense Co-attention Module is Applied to Find Relevant Objects and Entities and Filter Out Irrelevant Visual Information.

them into a bidirectional LSTM layer to learn hidden states. The forward and backward outputs are concatenated to construct character representations:

$$\mathbf{x}_i^c = [\overrightarrow{\mathbf{h}_i^c}; \overleftarrow{\mathbf{h}_i^c}], \tag{4}$$

where $\overleftarrow{\mathbf{h}_i^c}$ and $\overrightarrow{\mathbf{h}_i^c}$ denote the forward and backward outputs of bidirectional LSTM, respectively.

The total word representation $\mathbf{x}_i^t$ is obtained as the concatenation of word embeddings $\mathbf{x}_i^w$ and character embeddings $\mathbf{x}_i^c$:

$$\mathbf{x}_i^t = [\mathbf{x}_i^w; \mathbf{x}_i^c]. \tag{5}$$

We pass the total word representation $\mathbf{x}_i^t$ into a bi-directional LSTM to learn the contextual information. Specifically, the hidden state of bidirectional LSTM can be expressed as follows:

$$\overrightarrow{\mathbf{h}}_i^t = \overrightarrow{\text{LSTM}}(\mathbf{x}_i^t, \overrightarrow{\mathbf{h}}_{i-1}^t), \tag{6}$$

$$\overleftarrow{\mathbf{h}}_i^t = \overleftarrow{\text{LSTM}}(\mathbf{x}_i^t, \overleftarrow{\mathbf{h}}_{i-1}^t), \tag{7}$$

$$\mathbf{h}_i^t = [\overrightarrow{\mathbf{h}}_i^t; \overleftarrow{\mathbf{h}}_i^t]. \tag{8}$$

We feed $\mathbf{x}_i^t$ into a Dropout layer to prevent overfitting. $\overrightarrow{\mathbf{h}}_i^t$ and $\overleftarrow{\mathbf{h}}_i^t$ denote the i-th forward and backward hidden state of Bi-LSTM layer, respectively. Notationally, the final textual representations extracted from a sentence is denoted as $\mathbf{h}_i^t$.

3.2 Dense Co-attention Layer

As our multimodal network is built upon the BiLSTM and CRF framework, we design a dense co-attention layer module which combines the visual and textual features into predicting entities. The dense co-attention layer module learns to model the self-attention of objects or entities, as well as the guided attention between objects and entities, and produces a vector representation with aggregated knowledge among image and text. As shown in Figure 3, our dense co-attention layer module gets input from visual object representations and textual representations which as mentioned in last section.

One sentence may contain multiple entities with different entity types. For example, "Neil Patrick Harris" and "Oscars" are the names of person and award, respectively, in Figure 2. However, one can know that "Neil Patrick Harris" is a person name with the guiding objects "person", and "Oscars" with "tie and trophy". We apply dense co-attention mechanism here to find out the correlations of entities and visual objects, and the inner connections of objects or entities.

The dense co-attention layer is a variant of Deep Modular Co-Attention Networks[27], and is a modular composition of the self attention (SA) unit and the guided-attention (GA) unit. SA and GA are inspired by the scaled dot-product attention proposed in [25].

The input of scaled dot-product attention consists of queries of dimension d, keys of dimension d_{key} and values of dimension d_{value}. For simplicity, we set d_{key} and d_{value} to the same number d. We can calculate the attended feature as follows:

$$F = A(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V, \tag{9}$$

where, queries $Q \in \mathbb{R}^{m \times d}$, keys $K \in \mathbb{R}^{n \times d}$ and values $V \in \mathbb{R}^{n \times d}$. The attended feature $F \in \mathbb{R}^{m \times d}$ is obtained by weighted summation over all values V with respect to the attention learned form Q and K.

And we apply multi-head attention to further improve the representation capacity of the attended features. We firstly map the queries, keys and values to h different spaces, which means h paralleled 'heads'. Then, in each space, we apply an independent scaled dot-product attention function. Finally, we concatenate all head

results as follows:

$$F = MA(Q, K, V) = [head_1, head_2, ..., head_h]W^0, \quad (10)$$

$$head_j = A(QW_j^Q, KW_j^K, VW_j^V), \quad (11)$$

where, $W_j^Q, W_j^K, W_jV \in \mathbb{R}^{d \times d_h}$ are the projection matrices for the j-th head, and $W_o \in \mathbb{R}^{h*d_h \times d}$. $d_h = d/h$ is the dimension size of the output features from each head.

The self-attention (SA) unit is composed of a multi-head attention and a feed-foward layer. Firstly, the multi-head attention layer takes one group input feature $X \in \mathbb{R}^{m \times d_x}$ as query, keys and values. Then, the output features of the multi-head attention layer are transformed by two fully-connected layers with ReLU activation and dropout (FC(4d)-RELU-Dropout(0.1)-FC(d)). Moreover, to facilitate optimization, we apply the residual connection[11] with normalization layer[3] to the outputs of the two layers.

The guide-attention (GA) unit has the same architecture as SA, while has two group input features $X \in \mathbb{R}^{m \times d_x}$ and $Y \in \mathbb{R}^{n \times d_y}$. And X is input as queries, while Y is input as keys and values.

Then, we use the SA and GA attention units to build our dense co-attention layer module. For each word in the sentence, the textual representations $\mathbf{h}^t \in \mathbb{R}^{n \times d_t}$ are obtained by Equation (8) and the objects matrix $\mathbf{v} \in \mathbb{R}^{k \times d_v}$ is obtained by Equation (2). We first input the textual representations and object matrix into independent SA unit to capture the inner connections of objects or entities, respectively. Then, the output features of textual SA are input to GA as queries, and the output features of object SA are input to GA as keys and values. The GA unit can model the pairwise relationship between the each paired word and object:

$$\mathbf{h}_{SA}^t = SA(\mathbf{h}^t, \mathbf{h}^t, \mathbf{h}^t), \quad (12)$$

$$\mathbf{h}_{SA}^v = SA(\mathbf{v}, \mathbf{v}, \mathbf{v}), \quad (13)$$

$$\mathbf{h}_{GA}^v = GA(\mathbf{h}_{SA}^t, \mathbf{h}_{SA}^v, \mathbf{h}_{SA}^v) \quad (14)$$

where, $\mathbf{h}_{GA}^v \in \mathbb{R}^{n \times d_v}$ is the output features of GA unit, which contain the attended object features for each word in the sentence. Finally we add the textual representation from LSTM and the output features of GA unit to generate the multimodal representation:

$$\mathbf{m} = \mathbf{h}^t + \mathbf{h}_{GA}^v. \quad (15)$$

where $\mathbf{m}$ denotes the multimodal representation of the input sentence.

3.3 CRF Layer

Label dependencies are helpful for named entity recognition task. After obtain the multimodal representation from Dense Co-attention Layer, we apply a conditional random field (CRF) layer to model the label dependencies and predict the NER label sequences.

For an input sentence $s = (w_0, w_1, \ldots, w_n)$, we get the model prediction score matrix by applying the single feed forward layer to the multimodal representation:

$$\mathbf{q} = \mathbf{m}\mathbf{W}_q, \quad (16)$$

where $\mathbf{W}_q$ is a parameter matrix, $\mathbf{q}$ is the score matrix, and $\mathbf{q}_{i,j}$ is the score of the j-th label of the i-th word in the sentence. For a possible sequence of predictions:

$$y = (y_0, y_1, \ldots, y_n), \quad (17)$$

the CRF layer will take the model prediction score matrix and the dependencies among labels to calculate the score of y:

$$\pi(s, y) = \sum_{i=0}^{n} \mathbf{q}_{i,y_i} + \sum_{i=0}^{n-1} \mathbf{A}_{y_i, y_{i+1}}, \quad (18)$$

where $\mathbf{A}$ is the parameter matrix of transition scores used to modal the dependencies among labels, and $\mathbf{A}_{i,j}$ represents the score of a transition from the label i to label j. Then, we can get the probability for the sequence y by applying a softmax to all possible tag sequences:

$$p(y \mid s) = \frac{e^{\pi(s,y)}}{\sum_{\hat{y} \in Y(s)} e^{\pi(s,\hat{y})}}, \quad (19)$$

where $Y(s)$ is the set of all possible label sequences for the input sentence s. During training phase, we maximize the log-probability of the correct label sequence:

$$L = -logp(y \mid s). \quad (20)$$

During decoding, the label sequence y^* with the highest conditional probability is selected as output label sequence:

$$y^* = \underset{\hat{y} \in Y(s)}{\text{argmax}} p(\hat{y} \mid s) \quad (21)$$

4 EXPERIMENT SETTINGS

4.1 Dataset

To provide empirical evidence for effectiveness of our model, we evaluate our method on a multimodal social media dataset from Twitter. The dataset is constructed by Zhang et al. [28]. To the best of our knowledge, it is the only available multimodal named entity recognition dataset online. It contains 8257 tweets posted by 2116 users. Each tweet includes a sentence and an image. We split the dataset into training, development and testing parts following the same setting as Zhang et al. [28]. The entity types in the dataset are **Person**, **Location**, **Organization** and **Misc**. Especially, the MISC category contains "award", "project", "sports" etc. The statistics of each entity type are listed in Table 1.

Table 1: The Statistics of Each Entity Type in the Twitters Dataset.

Category	Train	Dev	Test	Total
Person	2217	552	1816	4583
Location	2091	522	1697	4308
Organization	928	247	839	2012
Misc	940	225	726	1881
Total Entity	6176	1546	5078	12784

4.2 Baseline Methods

We compare our methods with several state-of-the-art methods. Our experiments mainly consider two groups of models: previous state-of-the-art methods and the variants of our methods.

Previous State-of-the-art Methods: compared with NER in newswire domain, there are much fewer approaches concerning about social media domain.

Table 2: The Overall Performance of Our Models and Other State-of-the-art Methods. The Second Part is the Variants of Our Methods. Results on Rows where the Model Name is Marked with a ‡ Symbol are Reported as Published, All Other Numbers have been Computed by Us. Our Models Outperform Other Methods in All Metrics. * Indicates the Difference Against the F1 of Our Baseline Variant (OCM) is Statistically Significant by One-Tailed Paired t-test with $p < 0.01$.

Model	Prec.	Recall	F1
Stanford NER [8]‡	60.98	62.00	61.48
T-NER [20]‡	69.54	68.65	69.09
CNN+LSTM+CRF [17]‡	66.24	68.09	67.15
MNER-MA [18]	72.33	63.51	67.63
VAM [16]	69.09	65.79	67.40
AdapCoAtt Model [28]‡	72.75	68.74	70.69
BERT-NER [7]	70.65	**73.29**	71.87
OCM (Object + Character)	72.71	70.95	71.82
OCGA (Object + Character + GA)	72.22	**72.29**	72.26*
OCSGA (Object + Character + SA + GA)	**74.71**	71.21	**72.92***

- **Stanford NER:** Stanford NER is a widely used named entity recognition tool. It was proposed by Finkel et al. [8].
- **T-NER:** T-NER [20] is a method concerning about named entity recognition in tweets. T-NER exploits Freebase dictionaries as a source of distant supervision.
- **CNN+LSTM+CRF:** The model proposed by Ma and Hovy [17] is a traditional method in named entity recognition task which only considers the textual information.
- **AdapCoAtt Model:** Zhang et al. [28] proposed an adaptive co-attention network which combines the whole image features (rather than object features) and textual features in the Twitter dataset.
- **MNER-MA:** MNER-MA is a multimodal NER model proposed by moon et al. Moon et al. [18], which incorporates visual information with a modality attention module. Since they did not provide the data and code used in their paper, we reimplement their model following the same settings and validate on the Twitter dataset.
- **VAM:** The visual attention model Lu et al. [16] is another neural model for multimodal NER task. This model is composed by a BiLSTM-CRF model and a designed visual attention model. Also, we reimplement their model following the same settings of their paper since they did not provide the code and data.
- **BERT-NER:** To show the effectiveness of combining object-level features, we compare our model with contextual language models. We use the BERT BASE model Devlin et al. [7] and fine-tuning in the Twitter dataset.

Variants of Our Methods: We set ablation experiments to evaluate the contributions of each component. For fair comparison, we assign the same parameter settings for each model.

- **OCM(Object + Character):** This model is a variant of our model without the dense co-attention layer. We concatenate the object embeddings with textual representations for multimodal representations.
- **OCGA(Object + Character + GA):** This model is also a variant of our model without the self-attention unit in the dense co-attenion layer.
- **OCSGA(Object + Character + SA + GA):** The complete model which combines dense co-attention network (self-attention and guide attention) to model the correlations between visual objeccts and textual entities and the inner connections of objects or entities.

4.3 Parameter Settings

Our model is implemented by PyTorch framework[1] [2]. To initialize the word embeddings used in our model, we use the Glove pre-trained word embeddings, and the dimension is set to 200. The object embeddings are also set to 200 dimension and initialized with the pretrained word embeddings. Our model is trained with an SGD optimizer, where we set batch size for 10 and a learning rate for 0.008. Our dropout rate is 0.5 and the learning rate decay is 0.05. The number of objects is tuned from 1 to 5, and we gain the best result when setting it to 4. The number of attention heads in multi-head attention set to 2.

5 RESULTS AND DISCUSSION

5.1 Overall Results

We conduct our experiments on the Twitter dataset. Table 2 shows the overall results on Twitter test set. The first part of Table 2 is the performance of previous state-of-the-art methods. All the variants of our methods outperform previous works in precision and F1 values.

Ablation Study: The second part of this table is the performance of our method and its variants. We achieve an improvement of 5.77% in F1 value compared to the method proposed by Ma and Hovy [17]. The method proposed by Ma and Hovy [17] is a baseline of our model without visual object features. This indicates that combining the visual information with textual information is useful

[1] https://pytorch.org/
[2] Code is available at https://github.com/softhuafei/Pytorch-implementation-for-OCSGA

Figure 4: The Results of Our Method (OCSGA) Comparing to CNN+LSTM+CRF [17] and AdapCoAtt Model [28] on the Twitter Test Set. Objects from Images are Detected in the Left Column, We Present the NER Results with Related Objects in the Right Column. The GroundTruth Labels are in Red and the Detected Objects are in Green. Our Model Extracts Named Entities Precisely When Named Entities are Related to Visual Objects.

in recognizing entities of social media texts. The state-of-the-art multimodal NER methods: MNER-MA [18], VAM [16] and AdapCoAtt Model [28], outperform the textual baseline by considering the image-level features. However, when we transform the visual representation into object embeddings, our models outperform the method proposed by Zhang et al. [28] by 2.23%. The fine-grained object embeddings perform better than full image representations because we can extract different types of entities with the guiding of relevant objects. When we add the dense co-attention network module into our architecture, we gain the best Precision, Recall and F1 values. Our complete model achieves an F1 value of 72.92% and outperforms other state-of-the-art methods with a large margin (2.23 percentage's improvement in F1 scores). The dense co-attention network is designed to find the correlations between visual objects and entities, as well as the inner connections of objects or entities, so that our model can focus on the valuable visual objects. We also compare our model with pre-trained language model BERT-NER[7]. Our variant model OCM gains comparable results against BERT-NER, and we show that with the Dense Co-attention Layer, our model can outperform the BERT-NER in Precision and F1 values.

Performance on Categories: We also report our results in all four categories (Table 3). Our final method gains the highest F-score value in all the four categories and outperforms two state-of-the-art systems in social media NER [20, 28]. Interestingly, our model achieves a higher degree of improvement in ORG and MISC categories. ORG and MISC entities are those covering abundant visual information. For example, we can extract the ORG entities with objects such as "chair, tv and billboard". The MISC category, as another example, contains "award" which can be identified by objects such as "tie and trophy". We evaluate the effectiveness of object representations in the Case Study section.

Table 3: Our Results on Four Categories Compared to T-NER [20] and Adap. [28] on the Twitter Test Set.

Category	Our Model (OCSGA)			T-NER	Adap.
	Prec.	Recall	F1	F1	F1
PER	82.83	86.62	**84.68**	83.64	81.98
LOC	79.88	80.02	**79.95**	76.18	78.95
ORG	62.75	51.61	**56.64**	50.26	53.07
MISC	45.74	34.71	**39.47**	34.56	34.02
Overall	74.71	71.21	**72.92**	69.09	70.69

5.2 Parameter Sensitivity

In this section, we evaluate our model on different settings of the parameters. We are concerned about the impact of Dropout because it is demonstrated as effective in most NER tasks. Specifically, the number of objects is also important in producing a better result.

Table 4: The Performance (F1 Value) of Our Model on the Twitter Test Set with and Without Dropout.

Dropout	Overall	PER	LOC	ORG	MISC
No	71.49	83.64	78.30	54.20	33.90
Yes	**72.92**	**84.68**	**79.95**	**56.64**	**39.47**

Dropout is a strong strategy in avoiding over-fitting in training periods. We show the results of our model with and without Dropout strategy in Table 4. Our model gains the state-of-the-art performance with Dropout strategy and it demonstrates the effectiveness of Dropout in NER tasks.

Table 5 describes the results of our proposed model influenced by different object numbers. We have mentioned in Section 3 that we pick up the top K detected objects according to the detection

Table 5: The Performance of Our Proposed Model on the Twitter Test Set Influenced by Different Object Numbers.

Object Num.	Precision	Recall	F1
Top-1	74.41	70.72	72.52
Top-2	74.46	70.72	72.54
Top-3	**75.02**	70.66	72.77
Top-4	74.71	**71.21**	**72.92**
Top-5	74.56	70.60	72.53

possibility of each object in images. Our model gains the best performance when setting the number of objects to 4. The results demonstrate that the performance (F1 value) improves with the increase of visual objects. Since the average number of entities is less than 4, too many visual objects may bring noise (irrelevant visual information) into the prediction of entities. That indicates a proper number of visual objects can provide more effective semantic information for predicting entities with multiple categories.

5.3 Case Study

Figure 4 shows the case study of comparing our method with the CNN+LSTM+CRF [17] model and AdapCoAtt Model [28]. Our method performs better in all the cases due to the leveraging of visual objects. To evaluate the effectiveness of the usage of visual information, we compare our method to the CNN+LSTM+CRF model which extracts entities only relying on textual representations. Our model (OCSGA) extracts the MISC entity "Golf Classic" and the PER entity "Davonta Burdine" correctly. However, the CNN+LSTM+CRF model misses the two entities without the guidance of visual information. We think the object labels "sports ball" and "person" contribute to the extraction of "Golf Classic" and "Davonta Burdine" , respectively. Interestingly, our method identifies the PER entity "Davonta Burdine" correctly although the irrelevant visual object "truck" is extracted. We think the dense co-attention network module helps our model to filter out the objects irrelevant to textual entities.

On the right side of Figure 4, we compare our method with AdapCoAtt Model [28] which utilizes the full-image features as visual representations. Our method can identify the MISC entity "Oscars" with the guiding of objects "trophy and tie" and extract the correct entity type of "Blackhawks" with objects "skis and snowboard". However, the AdapCoAtt Model [28] cannot recognize the entities correctly due to the ignorance of correspondence of visual objects and entities.

6 CONCLUSION

In this paper, we propose a novel object-aware neural model that combines visual and textual representations into predicting named entities in social media posts. Our model takes the corresponding relations of multiple visual objects and different textual entities into consideration. The vision and language are bridged by transforming the object labels into embeddings. Our dense co-attention module can take the inter- and intra-connections between visual objects and textual entities into account. The experimental results demonstrate that our model outperforms other state-of-the-art methods in terms of Precision, Recall and F values.

For future work, we plan to investigate the method of transforming textual representations into visual vector space, which is another way to tackle the space discrepancy of vision and language.

ACKNOWLEDGMENTS

This work was supported by the Fundamental Research Funds for the Central Universities, SCUT (No. D2182480), the Science and Technology Planning Project of Guangdong Province (No. 2017B050506004), the Science and Technology Programs of Guangzhou (No. 201704030076, 201707010223, 201802010027, 201902010046), the National Natural Science Foundation of China under Grant (No. 61802130), the Guangdong Natural Science Foundation under Grant (No. 2018A030310355), the Guangzhou Science and Technology Program under Grant (No. 201707010223), CUHK Direct Grant for Research (No. 134920340) and the Research Grants Council of the Hong Kong Special Administrative Region, China (Collaborative Research Fund, No. C1031-18G).

REFERENCES

[1] 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)*. 724–728.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6077–6086.

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

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 423–443.

[5] Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research* 49 (2014), 1–47.

[6] Guillem Collell, Ted Zhang, and Marie-Francine Moens. 2017. Imagined visual representations as multimodal embeddings. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[7] 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* (2018).

[8] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 363–370.

[9] Nitish Gupta, Sameer Singh, and Dan Roth. 2017. Entity linking via joint encoding of types, descriptions, and context. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2681–2690.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[11] 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*. 770–778.

[12] Rupinder Paul Khandpur, Taoran Ji, Steve Jan, Gang Wang, Chang-Tien Lu, and Naren Ramakrishnan. 2017. Crowdsourcing cybersecurity: Cyber attack detection using social media. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 1049–1057.

[13] Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 36–45.

[14] John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (2001).

[15] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360* (2016).

[16] 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). 1990–1999.

[17] Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354* (2016).

[18] Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity recognition for short social media posts. *arXiv preprint arXiv:1802.07862* (2018).

[19] Swit Phuvipadawat and Tsuyoshi Murata. 2010. Breaking news detection and tracking in Twitter. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 3. IEEE, 120–123.

[20] Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 1524–1534.

[21] Alan Ritter, Evan Wright, William Casey, and Tom Mitchell. 2015. Weakly supervised extraction of computer security events from twitter. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 896–905.

[22] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2012. Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering* 25, 4 (2012), 919–931.

[23] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*. 2440–2448.

[24] István Varga, Motoki Sano, Kentaro Torisawa, Chikara Hashimoto, Kiyonori Ohtake, Takao Kawai, Jong-Hoon Oh, and Stijn De Saeger. 2013. Aid is out there: Looking for help from tweets during a large scale disaster. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1619–1629.

[25] 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 *Advances in neural information processing systems*. 5998–6008.

[26] Jie Yang and Yue Zhang. 2018. Ncrf++: An open-source neural sequence labeling toolkit. *arXiv preprint arXiv:1806.05626* (2018).

[27] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6281–6290.

[28] 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*.

[29] Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1227–1236.