Aiding Intra-Text Representations with Visual Context for Multimodal Named Entity Recognition

Omer Arshad*, Ignazio Gallo*, Shah Nawaz*, and Alessandro Calefati*

* Department of Theoretical and Applied Science, University of Insubria, Varese, Italy

***Abstract*—With the massive explosion of social media platforms such as Twitter and Instagram, people everyday share billions of multimedia posts, containing images and text. Typically, text in these posts is short, informal and noisy, leading to ambiguities which can be resolved using images. In this paper we will explore text-centric Named Entity Recognition task on these multimedia posts. We propose an end to end model which learns a joint representation of a text and an image. Our model extends multi-dimensional self-attention technique, where now image helps to enhance relationship between words. Experiments show that our model is capable of capturing both textual and visual contexts with greater accuracy, achieving state-of-the-art results on Twitter multimodal Named Entity Recognition dataset.**

***Index Terms*—Multimodal Named Entity Recognition; Self-Attention; Gated Fusion;**

I. Introduction

Recent years have seen a surge in multimodal data containing various media types. Typically, users combine image, text, audio or video data to express views on a social media platforms. The combination of these media types has been extensively studied to solve various tasks including classification [1], [2], [3], [4], cross-modal retrieval [5], [6] semantic relatedness [7], [8] and Visual Question Answering (VQA) [9], [10]. Recently, works [11], [12] and [13] combined text and image in a multimodal approach for text Named Entity Recognition (NER) problem [14]. Typically, the text component of a NER multimodal problem is challenging due to informal language, slang and typos, etc. [14]. These attributes make the task more challenging, compared to traditional NER. Moreover there are some ambiguous cases that can only be resolved with visual context as shown in Fig. 1. If we consider only the text in the first example "My daughter got 1 place in Apple valley Tags gymnastics", *Apple* is recognized as the name of an *Organization*, but, in this tweet, *Apple* should be labeled as *Location*. Similarly, the text in the second example "Apple's latest iOS update is bad for advertisers", *Apple* is wrongly recognized as the name of an *Organization*. In both cases, the disambiguation of the text is a non-trivial task, without considering the visual context.

In this work, we propose a novel neural network architecture which leverages on visual context to recognize named entities. We combine character and word embeddings to handle characteristics of NER textual component. In addition, self-attention mechanism is extended to capture relationships between two words and image regions, unlike previous works [11], [13] which used only single words to capture visual attention. Finally, we introduce a gated multimodal fusion module to select information dynamically from textual and visual features. Intuitively, our model captures two forms of interactions: intra-modal and cross-modal interactions.[1] We achieved state-of-the-art results on NER multimodal dataset [11]. In addition, we performed extensive experiments to show the effectiveness of the proposed model. Our main contributions are:

- introduction of an end-to-end model based on attention only that jointly learns intra and cross-modal dependencies, enhancing relationship of two words;
- state-of-the-art results on NER multimodal dataset [11].

My daughter got 1 place in [Apple valley **LOC**] Tags gymnastics

[Apple **ORG**] 's latest [iOS **OTHERS**] update is bad for advertisers

Fig. 1: Two NER multimodal examples show how some entities in the text can be correctly tagged in combination with visual information. Looking only at the text, the word *Apple* is ambiguous in the text description on the left, because it can be interpreted as *Location* (LOC) or as *Organization* (ORG).

II. Related work

In this section, we summarize relevant background on previous works on attention techniques and multimodal NER. **Attention Techniques.** Attention techniques allow models to focus on parts of visual or textual inputs of a task. Visual attention models selectively pay attention to small regions of an image to extract features. On the other hand, textual attention techniques find semantic or syntactic alignments in handling long-term dependencies. Attention techniques have

[1] Intra-modal interactions deal within same modality, whereas cross-modal captures interaction between modalities.

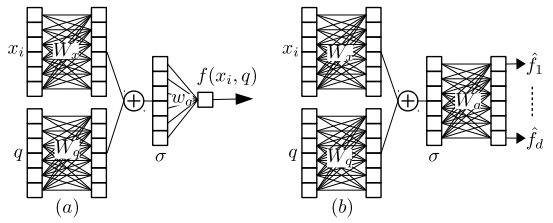


Fig. 2: (a) A neural representation of the additive attention described in Eq. (1), for two word embeddings q and x_i. (b) The additive multi-dimensional attention described in Eq. (5).

been extensively employed to vision and text related tasks, such as Image Captioning [15], VQA [10], Cross-Modal Retrieval [16] NER [11], [12], [13].

Multimodal Named Entity Recognition. NER task for short and noisy texts has been extensively studied in the literature [17], [18]. Recent years have seen an interest in capturing visual context from social media posts to improve this task [11], [12], [13]. These works used bi-directional Long Short Term Memory (LSTM) networks to extract features from a sequence of words. The work in [11] captures interactions between words and an image in a bi-directional way. However, it represents this interaction in a uni-directional manner. In our work, we extended multi-dimensional attention to jointly learn intra and cross-modal dependencies.

III. Proposed Model

In this paper we propose a novel architecture, inspired from Disan [19], to learn a joint representation of text and image for multimodal NER. Our model improves the intra-text attention to learn enhanced representations exploiting the relevancy with images. Instead of learning text representations separately from textual context and then leveraging on image information [11] and [13], our model jointly learns shared semantics between intra-text representation and visual features. In next subsections, we first explain each module of our network, and then, the proposed end to end model, as shown in Fig. 3.

A. Attention

The purpose of the attention module is to compute an *alignment score* between elements coming from different sources. In Natural Language Processing, given a sequence of word embeddings $\mathbf{x} = [x_1, x_2, ..., x_n]$ and the embedding of a query q, with $x_i, q \in \mathbb{R}^{d_e}$, the alignment score between x_i and q can be calculated using the common *additive attention*:

$$f(x_i, q) = w_a \sigma(x_i W_x + q W_q) \tag{1}$$

where σ is an activation function, w_a is a vector of weights and W_q, W_x are weights matrices. A graphical representation of the additive attention is available in Fig. 2(a). Furthermore, we use a special case of attention called *self attention* in which both elements q and x_i come from the same source.

$f(x_i, q)$ is a scalar score, which determines how important x_i is to a query q. Alignment score between q and all tokens becomes:

$$a = [f(x_1, q), \ldots, f(x_n, q)] \tag{2}$$

Then, a probability distribution $p(z \mid \mathbf{x}, q)$ is calculated over a by applying softmax. This gives a measure on how much the token $x_i \in \mathbf{x}$ is important to a query q.

$$p(z \mid \mathbf{x}, q) = softmax(a) \tag{3}$$

The final output we obtain from attention vector a is the weighted sum of all tokens in $\mathbf{x}$

$$C = \sum_{i=1}^{n} p(z = i \mid \mathbf{x}, q) x_i \tag{4}$$

that is the *context vector* for a query q.

B. Multi-Dimensional Attention

Multi-dimensional attention [19] proposed for self attention, is an additional attention technique which computes the feature-wise score vector $\hat{f}(x_i, q)$ described in Eq. (5) instead of computing the scalar score shown in Eq. (1).

$$\hat{f}(x_i, q) = W_a \sigma(x_i W_x + q W_q) \tag{5}$$

where $\hat{f}(x_i, q) \in \mathbb{R}^{d_e}$ is a vector with the same length as x_i, and W_a, W_q, $W_x \in \mathbb{R}^{d_e \times d_e}$ are the weight matrices.

Softmax is applied to the output function $\hat{f}$ to compute categorical distribution $p(z \mid \mathbf{x}, q)$ over all tokens.

To find the importance of each feature k in a word embedding x_i, $\hat{f}(x_i, q)$ becomes $[\hat{f}(x_i, q)]_k$ and the categorical distribution is calculated as:

$$P_{ki} = p(z_k = i \mid \mathbf{x}, q) = softmax([\hat{f}(x_i, q)]_k) \tag{6}$$

Thus the final context becomes:

$$C = \Big[\sum_{i=1}^{n} P_{ki} x_{ki} \Big]_{k=1}^{d_e} \tag{7}$$

The *multi-dimensional attention* defined in Eq. (5) is known as "token2token" self attention. It explores dependency between elements of the same source, i.e. query q and word x_i from a single source $\mathbf{x}$.

C. Image Feature extraction

In order to obtain features F from an image I we use a pretrained VGG-19 model. We extracted features of different image regions from the last pooling layer which has a shape of $7 \times 7 \times 512$. To simplify the calculations, we resized it to 49×512. We have $N = 49$ regions with $d_i = 512$ as dimension for each feature vector F_j with $j \in [1, \ldots, N]$. Regional features of a given image are represented by the matrix $F = VGG(I)$.

D. Character-based representation

Text extracted from social media is usually informal. In addition, it contains many out of vocabulary words. Character level features can play a crucial role in handling such text. We use a 2D Convolutional Neural Network to learn character embeddings. A word w is transformed into a sequence of characters $c = [c_1, c_2, ..., c_n]$ where n is the word length. A convolutional operation of filter size $1 \times k$ is applied to

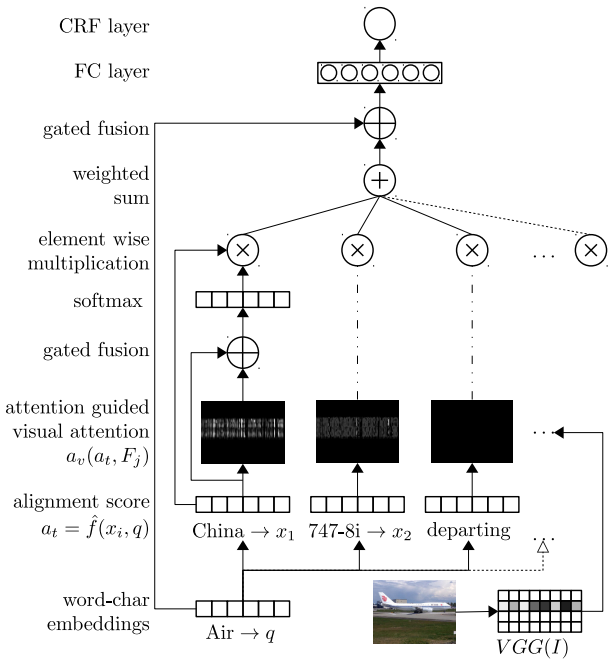


Fig. 3: End-to-end Multi-Dimensional Attention based model improving intra-modal attention using visual context for multimodal NER.

the matrix W, where $W \in \mathbb{R}^{d_e \times n}$ and d_e is the character embedding size. Then, we compute column-wise maximum operation to get the embedding for word w.

E. Attention Guided Visual Attention

Our proposed model uses alignment score between two textual tokens to compute cross-modal attention, show in Fig. 4. Given a word and query q, their alignment score is calculated using Eq. (5).

$$a_t = \hat{f}(x_i, q) \tag{8}$$

where a_t is a feature-wise score vector with same length of x_i. We calculate attention between a_t and image feature matrix F.

$$a_v(a_t, F_j) = W_v \sigma(a_t W_t + F_j W_i) \tag{9}$$

where $a_v(a_t, F_j)$ is a vector representing a single row of the visual attention scores matrix between a_t and F, $a_t \in \mathbb{R}^{d_e}$, $F_j \in \mathbb{R}^{d_i \times N}$, N is number of image regions, $W_i \in \mathbb{R}^{d_i \times d_e}$, $W_t, W_v \in \mathbb{R}^{d_e \times d_e}$ are the weight matrices. To obtain the final visual attention matrix, we compute a_v, such that $a_v \in \mathbb{R}^{d_e \times N}$

Then we normalize the scores a_v by applying Eq. (6) to get a probability distribution (columns-wise) over all regions of image.

$$P(a_v) = softmax(a_v) \tag{10}$$

The final output is a element-wise product between $p(a_v)$ and F.

$$C_v = \sum_{i=1}^{n} P_i(a_v) \odot F_i \tag{11}$$

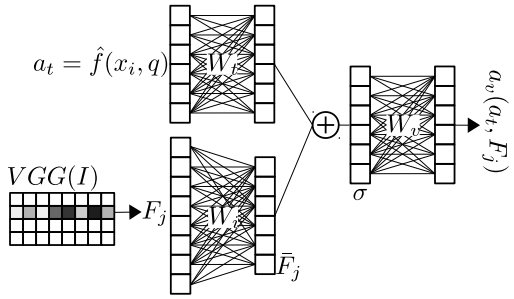


Fig. 4: Attention Guided Visual Attention neural representation. The output of this model is a single vector that will be concatenated to the others to obtain the final matrix.

where $C_v \in \mathbb{R}^{d_e}$, is a vector containing context vector for a_t.

F. Gated Fusion

In order to combine alignment score a_t of a word w and a query q with its visual attention vector C_v we use a gate function to dynamically combine alignment score and visual attention vectors.

$$G = \sigma(W^{(1)} a_t + W^{(2)} C_v + b^{(G)}) \tag{12}$$

$$O = G \odot C_v + (1 - G) \odot a_t \tag{13}$$

Throughout our network, we used gated fusion to merge different modalities.

IV. End to End Model

Our end to end network jointly learns intra and cross-modal dependencies. The architecture is shown in Fig. 3.

Given a sequence of words $\mathbf{x} = [x_1, x_2, \ldots, x_n]$, and image features F, we first compute the alignment score between x_i and a query q, where i ranges from 1 to the length n of a sentence. We use Eq. (8) to compute the alignment score a_t. In order to aid intra-text representation a_t, we compute alignment score $alignment_v$ between a_t and F using the compatibility function given in Eq. (9).

$$alignment_v = a_v(a_t, F) \tag{14}$$

We then apply the method described in Eqs. (10)-(11) on $alignment_v$ to get a weighted sum C_v. Now this C_v is a visual context vector, which will contribute to a_t. This step makes intra-modal relation a_t between x_i and q stronger because it also includes image context. Relation a_t between two words will remain the same if textual context is taken into account. But if we included visual context, this relation would become dynamic and more expressive. Thus helping intra-modal attention to learn improved relationships between words.

In order to make visual context vector C_v helpful for a_t, we combine them using gated fusion. This will dynamically select which features to select from various modalities. We applied Eqs. (12)-(13) to get a fused representation F_{rep} between a_t and C_v.

Now we have fused representation F_{rep} representing a_t between query q and x_i. Note that, this is calculated for all tokens of a sequence $\mathbf{x}$. We apply Eq. (6) to get a categorical distribution P over all n tokens of $\mathbf{x}$. Then, the element-wise product is computed between P and each token of a sequence $\mathbf{x}$ to get context vector C for query q.

$$C = \sum_{n=1}^{n} P_i \odot x_i \tag{15}$$

where $C \in \mathbb{R}^{d_e}$, n is total tokens of a sentence and d_e is vector dimension of each token. Now C is a context vector jointly learned from text and visual features. To handle attributes of text component of NER, we also used word representation x_i with C, exploiting gated fusion Eqs. (12)-(13) to obtain the final output O. Furthermore, O is passed through a fully connected layer.

$$O_{fc} = Relu(W^{(1)}O + b_o) \tag{16}$$

where $W^{(1)}$ is the learnable parameter and b_o is the bias vector. For tag prediction, final output O_{fc} is passed to Conditional Random Field (CRF) layer [20].

A. Conditional Random Field

CRF [21] are useful in tasks where output labels have a strong dependency (e.g. I-PER cannot follow B-LOC). Predicting such outputs independently is challenging without correlation between labels and their neighborhood. Given $\mathbf{x} = [x_1, x_2, ..., x_n]$ as a text sequence and $\mathbf{y} = [y_1, y_2, ..., y_n]$ as a sequence of labels for $\mathbf{x}$, possible labels sequences can be calculated using following equation.

$$p(\mathbf{y} \mid \mathbf{x}) = \frac{\prod_{i=1}^{n} \Omega_i(y_{i-1}, y_i, \mathbf{x})}{\sum_{y' \in Y} \prod_{i=1}^{n} \Omega_i(y'_{i-1}, y'_i, \mathbf{x})} \tag{17}$$

Where $\Omega_i(y_{i-1}, y_i, \mathbf{x})$ and $\Omega_i(y'_{i-1}, y'_i, \mathbf{x})$ are potential functions. We use maximum conditional likelihood to learn best parameters that maximize the log-likelihood.

$$L(p(\mathbf{y} \mid \mathbf{x})) = \sum_{i} logp(\mathbf{y} \mid \mathbf{x})) \tag{18}$$

V. Datasets

We used multimodal NER dataset [11]. It contains 4 types of entities {Person, Location, Organization and Misc.} collected from 8257 tweets, containing 4000/1000/3257 samples for training/val/test sets respectively. Table I shows number of samples per entity. Figs. 1, 5 show some ambiguous examples.

VI. Experiments

We performed various experiments to evaluate the effectiveness of the proposed model on NER multimodal dataset [11]. Standard Precision, Recall and F1 scores are used as evaluation metrics.

(a) Finding [California **OTHER**] [oil **OTHER**] [spill **OTHER**]'s cause could take a month

(b) Oil pipeline break dumps crude oil on California **[LOC]** beach

Fig. 5: Some examples showing how images can help to resolve ambiguities. The word "California" in (a) and (b) is difficult to understand without looking at the images as it has different tags.

TABLE I: Details of Dataset

	Train	**Val**	**Test**
Person	2217	552	1816
Location	2091	522	1697
Organization	928	247	839
Misc	940	225	726
Total samples	4000	1000	3257

A. Baselines

1) Disan: Our model is inspired from Disan, thus we consider this approach as a baseline for the text only evaluation. We used multi-dimensional self attention method to extract context-aware representation of texts. For fair comparison, we keep the same architecture of Fig. 3, but we exclude the "Attention Guided Visual Attention" module. This to analyze differences between the proposed model and multi-dimensional self attention approaches.

2) T-NER: T-NER [22] is a tweet specific NER system which uses hand crafted features i.e. orthographic, contextual and dictionary features. It was trained and evaluated on multimodal NER dataset [11]

3) State-of-the-art models: We compare our model with previous state-of-the-art methods leveraging on LSTM networks to capture textual dependencies and visual attention to exploit cross-modal interactions.

B. Word embeddings

We used 300D fasttext crawl embeddings. It contains 2 million word vectors trained with subword information on Common Crawl (600B tokens). However, we do not apply fine-tuning on these embeddings during the training stage.

C. Character Embeddings

50D character embeddings are trained from scratch using a single layer 2D CNN with a kernel size of 1×3.

D. Optimization

We set Adam optimizer with different learning rate initialization: 0.001, 0.01 and 0.005. We achieved the best score

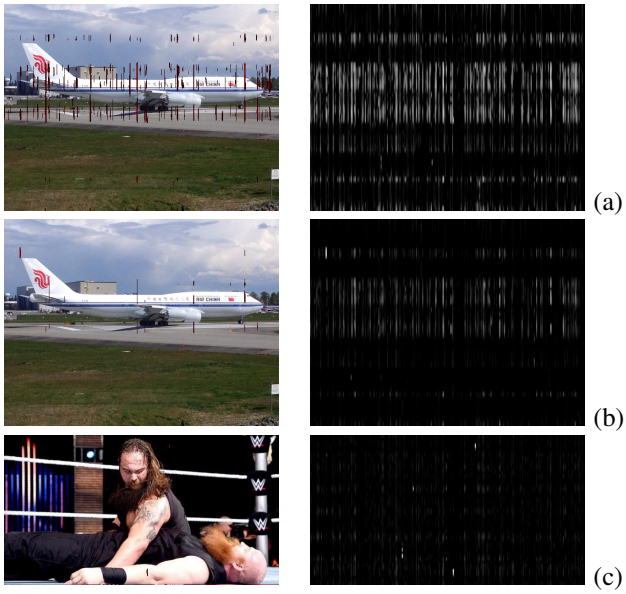


Fig. 6: Examples of "attention guided visual attention". In (a) a graphical representation of the visual attention matrix $a_v(a_t, F_j)\forall j$, for the the alignment score $a_t = \hat{f}(x_i, q) = \hat{f}(\text{Air}, \text{China})$ and in (b) a graphical representation of the visual attention matrix for the alignment score $a_t = \hat{f}(x_i, q) = \hat{f}(\text{Air}, \text{Field})$. In (c) the visual attention matrix obtained from a different image (Best viewed in color).

using the learning rate equal to 0.001. Batch size and dropout keep probability are set 20 and 0.5 respectively.

VII. Results

Table II shows comparison of our model with baseline and previous state-of-the-art methods. We achieved best $F1$ score, outperforming previous approaches.

A. Impact of word embeddings

Word embeddings with different sizes and trained on different corpora have a strong impact on performance. We evaluate the model on two word embeddings to analyze the effect.

- Twitter Embeddings: we used embeddings trained on 30 million tweets [11]. To compare our results with this work, we used same embeddings to evaluate the improvement that comes from our proposed architecture.
- Crawl Embeddings: we used 300D Crawl embeddings [23] trained on 600B tokens to fully exploit the capability of our model,

Table II shows results with different word embeddings. Our model achieves better results using Twitter Embedding, but there is a clear performance boost with Crawl embeddings. It is interesting to note that entities "PER" and "LOC" have a minor improvement whereas there is an enhancement of 5-6% in "ORG" and "MISC". With Crawl embeddings, our model performs best to predict the "ORG" category, with a 3% difference from baseline (text only using the Disan model). Our model achieves best F1 scores using both embeddings. This clearly shows that our model performs better regardless to the word embedding.

B. Impact of cross-modal attention

Table II shows that "Attention guided visual attention" boosts Disan (intra-modal attention) performance. We present a qualitatively example in Fig 6 that shows the focus of visual attention. Given an annotated sentence:

S = [Air B-OTHER] [China I-OTHER] [747-8i I-OTHER] [departing O]... [Field O]

an alignment score (inter-modal attention) $a_t = \hat{f}(x_i, q)$ between two tokens "Air" and "China" is aided by visual context, see Fig. 6 (a). Our model can successfully focus on related image regions, strengthening the relation between two words. Whereas alignment score $a_t = \hat{f}(x_i, q)$ between "Air" and "Field" and visual context (Fig. 6 (b)) has less relation because word "Air" has higher dependency on "China" then on "Field" when labeling "Air" as "B-OTHER". Similarly, in order to verify that relation between two tokens varies according to image, we changed the image and kept same sentence. Relation between fake image and alignment score $a_t = \hat{f}(x_i, q)$ between two tokens "Air" and "China" can be seen in Fig. 6(c) with no prominent relation, proving that our model learns better relationships between words given an image.

Similarly, Fig. 7 shows some examples in which our model pays attention to related image regions. It clearly shows that our model focuses on certain parts of image which are beneficial for named entity task. Fig. 7(a) shows that attention is payed only to car to predict correct tag for "Mercedes" and "Benz". Similarly Fig. 7(b) shows that "Opera House" is a building and not a location, and our models correctly identifies it by paying attention to correct image regions.

C. Error Analysis

In Fig. 8, we show some examples where our approach fails because of the following reasons:

- Unrelated image : Text information do not match with an image, as we can see in Fig. 8(a), "Reddit" belongs to "Other" but unrelated image caused wrong attention and it results to wrong prediction "ORG".
- Wrong attention: Fig. 8(b) shows an example where text and image are aligned correctly, but wrong attention results in wrong tag prediction. Words "Mount" and "Sherman" were tagged as "Person" as most of attention is on persons, whereas expected tag was "LOC".

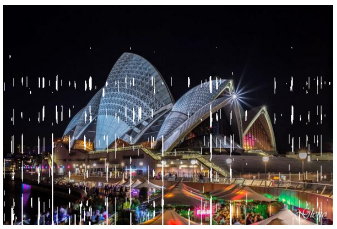
VIII. Conclusion

We introduced a novel model for multimodal NER. It extends multi-dimensional self-attention approaches by enhancing intra-text relationships using visual features. Qualitative examples show that our model successfully captures correct relations between words and images removing ambiguities caused by the text. Our model is flexible and can be further

TABLE II: Comparison of our approach with baselines and previous state-of-the-art methods.

	PER F1	LOC F1	ORG F1	MISC F1	Overall Prec.	Overall Recall	Overall F1
T-NER ([11)	83.64	76.18	50.26	34.56	69.54	68.65	69.09
Adaptive Co-Attention Network [11]	81.98	**78.95**	53.07	34.02	72.75	68.74	70.69
Disan [19] (Twitter Embedding)	82.07	76.87	55.34	32.29	71.00	70.53	70.77
Our Model (Twitter Embedding)	82.83	78.22	55.88	33.00	72.81	70.33	71.55
Disan [19] (Crawl Embedding)	83.03	77.96	56.66	**39.54**	71.65	71.89	71.77
Our Model (Crawl Embedding)	**83.98**	78.65	**59.27**	**39.54**	**73.50**	**72.33**	**72.91**

(a) [Mercedes **OTHER**] [Benz **OTHER**]



(b) photo of [Sydney **LOC**] [Opera **OTHER**] [House **OTHER**]

Fig. 7: Two examples of correct visual attention. Our model successfully highlights related image regions required to predict correct tag.

(a) [Reddit **ORG**] needs to stop pretending

(b) teachers on top of [Mount **PER**] [Sherman **PER**]

Fig. 8: Two examples of wrong visual attention: (a) shows an unrelated image and a wrong prediction, while (b) shows a related image with wrong attention and prediction.

extended to other multimodal tasks. Experiments show that our model achieved state-of-the-art results on multimodal NER dataset.

REFERENCES

[1] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, "Gated multimodal units for information fusion," *arXiv preprint arXiv:1702.01992*, 2017.

[2] D. Kiela, E. Grave, A. Joulin, and T. Mikolov, "Efficient large-scale multi-modal classification," *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 5198–5204, 2018.

[3] I. Gallo, A. Calefati, and S. Nawaz, "Multimodal classification fusion in real-world scenarios," in *Document Analysis and Recognition (ICDAR)*, vol. 5. IEEE, 2017, pp. 36–41.

[4] S. Nawaz, A. Calefati, M. K. Janjua, M. U. Anwaar, and I. Gallo, "Learning fused representations for large scale multi-modal classification," *IEEE Sensors Letters*, 2018.

[5] S. Nawaz, M. K. Janjua, A. Calefati, and I. Gallo, "Revisiting cross modal retrieval," *arXiv preprint arXiv:1807.07364*, 2018.

[6] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5005–5013.

[7] D. Kiela and L. Bottou, "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)*, 2014, pp. 36–45.

[8] C. W. Leong and R. Mihalcea, "Going beyond text: A hybrid image-text approach for measuring word relatedness," in *International Joint Conference on Natural Language Processing*, 2011, pp. 1403–1407.

[9] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *Proceedings of Empirical Methods in Natural Language Processing, EMNLP 2016*, pp. 457–468, 2016.

[10] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, vol. 3, no. 5, 2018, p. 6.

[11] Q. Zhang, J. Fu, X. Liu, and X. Huang, "Adaptive co-attention network for named entity recognition in tweets." in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 5674–5681.

[12] S. Moon, L. Neves, and V. Carvalho, "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*. Association for Computational Linguistics, 2018, pp. 852–860.

[13] D. Lu, L. Neves, V. Carvalho, N. Zhang, and H. Ji, "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)*, vol. 1, 2018, pp. 1990–1999.

[14] T. Baldwin, M.-C. de Marneffe, B. Han, Y.-B. Kim, A. Ritter, and W. Xu, "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*, 2015, pp. 126–135.

[15] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[16] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 299–307.

[17] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," *arXiv preprint arXiv:1603.01360*, 2016.

[18] G. Aguilar, S. Maharjan, A. P. L. Monroy, and T. Solorio, "A multi-task approach for named entity recognition in social media data," in *Proceedings of Noisy User-generated Text*, 2017, pp. 148–153.

[19] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, and C. Zhang, "Disan: Directional self-attention network for rnn/cnn-free language understanding," *AAAI Conference on Artificial Intelligence*, pp. 5446–5455, 2018.

[20] X. Ma and E. H. Hovy, "End-to-end sequence labeling via bi-directional lstm-cnns-crf," *CoRR*, vol. abs/1603.01354, 2016.

[21] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289.

[22] A. Ritter, S. Clark, M. Mausam, and O. Etzioni, "Named entity recognition in tweets: An experimental study." Association for Computational Linguistics, 01 2011, pp. 1524–1534.

[23] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.