Visually-Guided Named Entity Recognition by Grounding Words with Images via Dense Retrieval

Wang Yao[1]　Naoki Yoshinaga[2]
[1]The University of Tokyo　[2]Institute of Industrial Science, the University of Tokyo
yao-w@tkl.iis.u-tokyo.ac.jp　ynaga@iis.u-tokyo.ac.jp

Abstract

We can verbally communicate with others since words are grounded with real-world meanings such as entities. In this study, instead of multimodal natural language processing (NLP) that assumes supplemental images along with text, we propose to ground each word in the text with images via dense retrieval (vokenization) to solve text-only NLP tasks using multimodal NLP models. We focus on named entity recognition (NER) and modify the existing NER model to refer to the retrieved images. We evaluated our methodology on two multimodal named-entity recognition datasets and confirmed the advantage of the proposed framework.

1 Introduction

Humans learn language from experience based on sense including feeling, smelling, and especially looking. Multimodal natural language processing (MNLP) has been thereby studied to consider multimodal input in existing NLP tasks such as machine translation [1], multimodal NER [2], and video captioning [3]. Multimodal NLP is motivated to address two challenges in NLP: lexical ambiguity, and out of vocabulary words. With images, people can easily understand the meanings of footballs (soccer or American football) which are sometimes unclear from text-only input. Images can also provide rich knowledge of rare and unseen words.

Current studies in multimodal NLP, for example, in multimodal machine translation, show that visual information is indeed useful for improving the quality of outputs (here, translation) [4]. However, recent research shows that visual content is not as useful as expected and it might play the role of regularization [5]. The current multimodal NLP suffers from two limitations: data sparsity and the utility of images on the task. The biggest multimodal dataset is still far smaller than the text-only counterpart. For example, in multimodal machine translation, multi30k [1] (29k training pairs) vs. ParaCrawl （51M training pairs for v5 fr-en [6]). The utility of the supplied images needs to be doubted as well. In real-world situations, the irrelevant text-image pairs account for a very large proportion. The supplied images do not always include all information in the text and sometimes images are just used to be ironic or make a funny.

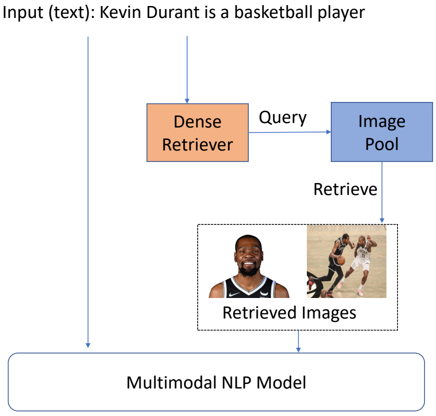

Figure 1　Applying multimodal NLP models to text-only NLP tasks via image retrieval.

In this paper, we purposed to exploit using dense image retrieval [7] to ground each word in the text with images to provide visual information for text-only NLP tasks (Figure 1). Considering image pool as an extra knowledge base, our framework can be viewed as a recently popular knowledge-based NLP that utilizes instance-wise knowledge via retrieval [8].

We evaluate our multimodal framework on the multimodal NER (MNER) task. We first train a dense retriever based on the text-image pair of the training split in the datasets, twitter-2015 [9], and twitter-2017 [2]. Then we use dense retrieval to retrieve images as visual input to the multimodal NER to train a multimodal NER model.

2 Related Work

In this section, we first review multimodal named entity recognition. We next introduce pre-trained language and vision models followed by cross-modal dense retrieval, both of which are used in our method.

Multimodal Named Entity Recognition (MNER): MNER is a new challenge generated from NER. Formally, given a sentence S and its associated image V as input, the goal is to classify entities in S to the predefined types.

Moon et al. [10] first introduced the MNER task and proposed a modality attention module at the input of a NER network. To solve the problem of irrelevance between image and text, Sun et al. [11] purposed a text-image relation propagation-based BERT model which reaches the state-of-the-art performance.

Pretrained Language and Vision Model: Incorporating different modalities to improve representation learning is very popular after pretraining and self-supervision become a fashion. In computer vision, Clip [12] proposed the idea of language supervision; training on image-text pair via contrastive learning can help a lot on zero-shot image classification. In NLP, although research by Tan et al. [7] shows that vision-and-language pre-trained models now can hardly get improvement in text-only tasks, some research like imagination [13] in machine translation also shows the potential of using visual information to benefit text-only tasks.

Cross-modal Dense Retrieval: Dense retrieval is a key technique in the text-only Question Answering (QA) task [14]. The idea of dense retrieval is to generate a new representation for both query and key in a shared subspace. Cross-modal dense retrieval is very similar to text-only dense retrieval. The difference is that the gap between different modalities is much bigger. To solve this problem, Tan et al. [7] proposed vokenization to ground words instead of sentences with images via dense retrieval.

3 Preliminaries

This section introduces how vokenization ground words with real-word images in our retrieval-based multimodal NLP.

3.1 Vokenization

The main idea of vokenization [7] is to learn a cross-matching model that can map text and image to the shared subspace from a text-image dataset and then build a "vokenizer" that can ground a given word with images using maximum inner product search (MIPS). The cross-matching model consists of two independent encoders for image encoder_{img} and text encoder_{lang}. The model takes an image v as input and a sentence s composed of a sequence of tokens $w_1, w_2, ..., w_3$. The output $r_\theta(w_i, v; s)$ is the relevance score between the token $w_i \in S$ and image v. The relevance socre is the inner product of the language features $f_\theta(w_i; s)$ and the visual feature $g_\theta(v)$. The formulation is:

$$r_\theta(w_i, v; s) = f_\theta(w_i; s)^\top g_\theta(v) \tag{1}$$

Since two encoders do not project image and text into the same dense space, the cross-matching model applies two multi-layer perceptrons (MLP) w_{mlp} and v_{mlp}. Then the language features $f_\theta(w_i; s)$ and visual features $g_\theta(v)$ can be computed by

$$\begin{aligned} f_\theta(w_i; s) &= w_{mlp}(\text{encoder}_{lang}(w_i))) \\ g_\theta(v) &= v_{mlp}(\text{encoder}_{img}(v)) \end{aligned} \tag{2}$$

The model is then optimized by maximizing the relevance score of these aligned token-image pairs in image-text datasets. Given this cross-matching model, the retrieval problem could be formulated as a maximum inner product search.

4 Methodology

We proposed to first annotate text-only NLP tasks with images via vokenization, and then use the retrieved images to turn the original input into multimodal input. Then we add two modifications to an existing text-only model to effectively utilize the retrieved images. We concretely explain the modifications using a BERT-CRF NER model [15] as an example, since in this study we evaluate our framework on NER.

4.1 Retrieving images for words in inputs

First, we use the text-image pairs in the training split to train the vokenizer we have explained in Section 3.1. We retrieve images for words in the task inputs using the trained vokenizer and some pool of images (in this case, all images

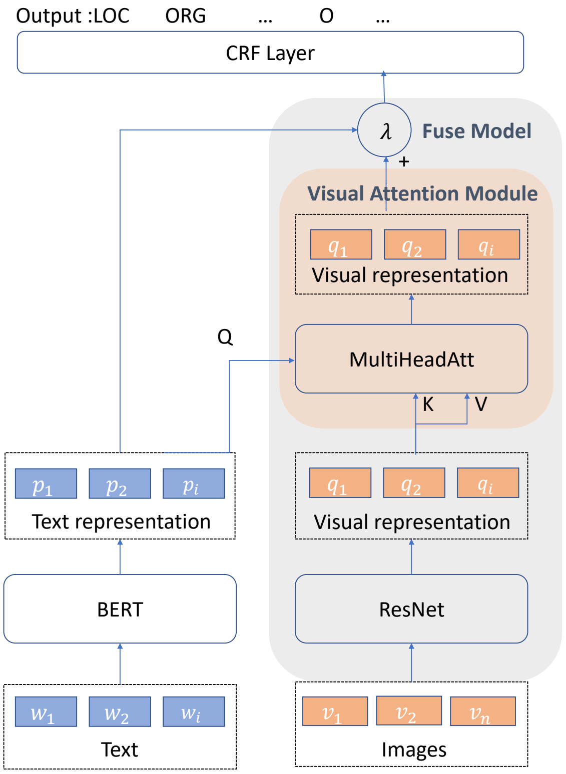

Figure 2 Multimodal BERT-CRF NER model with multiple image inputs.

in the training dataset). We then obtain multimodal input with a sentence S composed by tokens $< w_1, w_2, ..., w_i >$ and a series of retrieved images $x < v_1, v_2, ..., v_i >$ that correpond to every word w_i one by one.

4.2 Modification to Multimodal Model

Here, we introduce two types of modifications on BERT-CRF [15] to take a series of retrieved images as input: 1) Fuse Model 2) Visual Attention Model (Figure 2). In the following, we explain these two modifications.

4.2.1 Fuse Model

BERT-CRF model is a variant of BERT by adding a CRF layer after representation learning.

$$\begin{aligned} f_\theta(w_i; s) &= \text{BERT}(s) \\ \text{Output} &= \text{CRF}(f_\theta(w_i; s)) \end{aligned} \tag{3}$$

We first extract text and image representation with pre-trained models from the input series of images. In this case, we take ResNet [16] as a pre-trained vision model.

$$g_\theta(v_u; x) = \text{ResNet}(x) \tag{4}$$

To alleviate the visual bias for different words, we applied a gate $\lambda \in [0, 1]$ to weigh the expected importance of retrieved image representation for each word.

Table 1 The detailed information of two MNER datasets.

Dataset	Train	Dev	Test
twitter-2015	4000	1000	3257
twitter-2017	3373	723	723

We then learn word representation by

$$f_\theta(w_i; s) = (1 - \lambda) \cdot f_\theta(w_i; s) + \lambda \cdot g_\theta(v_u; x)) \tag{5}$$

Finally, we fed the fused word representation to the CRF layer.

4.2.2 Visual Attention Model

We further integrate a visual attention module based on multi-head attention [17] to the fuse model.

$$Attention(Q, K, V) = \text{Softmax}(\frac{QK^\top}{\sqrt{d_k}}) \tag{6}$$

Query vector is generated from the word representations and both key and value are generated from visual representations.

5 Experiment and Analysis

We evaluate our methodology on the NER task based on the BERT-CRF model.

We choose two multimodal NER datasets@[2, 9] to test our methodology. Following the common settings in [18], we labeled them as twitter-2015 and twitter-2017. Table 4 shows the detailed information.

The multimodal NER dataset provides image-text pair that can test the difference between our joint model using a given image and using vokenized images. We take the whole image part as candidates for the vokenizer.

For implementation detail, we trained our cross-matching model for vokenizer using ResNet152 and BERT as the backbone of the image encoder and the text encoder. We then add two linear multi-layer perceptrons mlp_0 and mlp_1 which have two fully-connected layers with 256-dimensional intermediate outputs (followed by ReLU activation) and 64-dimensional final outputs, which mean we project two types of vectors to a 64-dim dense space.

To do fast retrieval, we built the vokenizer with FAISS [19]. We trained our cross-matching model on the training split of the dataset. In other words, we use only the image-text pairs in the training split. In validation and test,

Table 2 Results of NER and our retrieval-based multimodal NER.

Method	LOC	ORG	MISC	PER	Pre.	Recall	F_1
twitter-2015							
BERT+CRF	79.28	58.25	35.51	**85.27**	69.57	73.85	71.62
(+Fuse Model)	**79.69**	**59.03**	**36.00**	85.19	**69.60**	**74.50**	**71.93**
(+Visual Attention)	79.10	58.11	34.60	84.41	68.90	73.55	71.10
twitter-2017							
BERT+CRF	83.36	82.60	65.54	91.42	83.34	86.30	84.70
(+Fuse model)	**84.15**	**83.90**	**65.77**	91.76	83.95	**87.02**	**85.45**
(+Visual Attention)	83.26	83.45	65.40	**92.51**	**84.67**	86.40	85.37

Table 3 Ablation test.

Model	Image	twitter-2015	twitter-2017
Fuse Model	Original	70.98	84.93
Fuse Model	Vokenized	**71.93**	**85.45**
+VisualAtt	Original	71.28	85.22
+VisualAtt	Vokenized	71.10	85.37

Table 4 Comparison to the SOTA multimodal NER model.

Model	Image	Twitter-2015
Fuse Model	Original	70.98
RpBERT [11]	Original	**74.4**

we only use retrieved images as visual information rather than given images in the original dataset.

5.1 Results

In Table 2, we compared different methods on both twitter-2015 and twitter-2017 datasets with F_1 scores of different types of entities, total precision, recall, and weighted average F_1 score. Based on the result, we can find that in most cases, our methods can slightly improve the text-only model. In our cases, adding an attention mechanism cannot improve the performance of the multimodal NER model. Simply using the fuse model (Section 4.2.1) is more effective.

5.2 Analysis

To investigate the effectiveness of using the images retrieved by the vokenizer, we also consider using the given images to train our model. Table 3 shows the results. In both datasets, the model using vokenized images can outperform the model using the original image which means that vokenization may help solve the irrelevance between image and text.

We also compared our model with the given images with the state-of-the-art multimodal NER model. Table 4 shows that our proposed method still has an over 3% gap between the state-of-the-art MNER model, which means our method may not be good enough to incorporate visual information with textual information.

6 Conclusions

In this paper, we first point out two challenges of multimodal NLP: data sparsity and the utility of images on the task. To overcome these challenges, We introduce using a dense retriever to ground words in inputs of text-only NLP tasks with images. Moreover, to use more than one image (usually, normal multimodal models assume a single image), we also proposed two easy multimodal models which can incorporate a series of images to guide the final prediction. The experimental results demonstrate that our method could consistently provide useful information in the two different multimodal NER datasets.

There are several future directions for this work. First, considering the performance comparison with the state-of-the-art MNER model, our model is still not that good to capture and incorporate visual information. Second, in this paper, we only consider using about 6k images in the MNER dataset as candidates for very limited retrieval. For a better study, we may collect a larger image-text dataset (e.g., Twitter dataset) to better understand the effectiveness of vokenization. Last but not the least, since the result only shows the difference in metrics, we have to dig more into the cases in the future study.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP21H03494 and JST CREST Grant Number JPMJCR19A4.

References

[1] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. In **Proceedings of the 5th Workshop on Vision and Language**, pp. 70–74. Association for Computational Linguistics, 2016.

[2] Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng 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)**, pp. 1990–1999, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[3] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In **Proceedings of the IEEE international conference on computer vision**, pp. 706–715, 2017.

[4] Quanyu Long, Mingxuan Wang, and Lei Li. Generative imagination elevates machine translation. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 5738–5748, Online, June 2021. Association for Computational Linguistics.

[5] Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 6153–6166, Online, August 2021. Association for Computational Linguistics.

[6] Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. ParaCrawl: Web-scale acquisition of parallel corpora. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4555–4567, Online, July 2020. Association for Computational Linguistics.

[7] Hao Tan and Mohit Bansal. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2066–2080, Online, November 2020. Association for Computational Linguistics.

[8] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics.

[9] Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. Adaptive co-attention network for named entity recognition in tweets. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 32, No. 1, Apr. 2018.

[10] Seungwhan Moon, Leonardo Neves, and Vitor 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)**, pp. 852–860, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[11] Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. RpBERT: A text-image relation propagation-based BERT model for multimodal NER, 2021.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In **ICML**, 2021.

[13] Desmond Elliott and Ákos Kádár. Imagination improves multimodal translation. In **Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 130–141, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[14] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 6086–6096, Florence, Italy, July 2019. Association for Computational Linguistics.

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

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 770–778, Los Alamitos, CA, USA, jun 2016. IEEE Computer Society.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Advances in neural information processing systems**, pp. 5998–6008, 2017.

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

[19] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. **arXiv preprint arXiv:1702.08734**, 2017.