On Analyzing the Role of Image for Visual-Enhanced Relation Extraction (Student Abstract)

Lei Li[1,2], **Xiang Chen**[1,2], **Shuofei Qiao**[1,2], **Feiyu Xiong**[3], **Huajun Chen**[1,2], **Ningyu Zhang**[1,2*]

[1]Zhejiang University & AZFT Joint Lab for Knowledge Engine, China
[2]Hangzhou Innovation Center, Zhejiang University, China
[3]Alibaba Group, China
{leili21,xiang_chen,shuofei,huajunsir,zhangningyu}@zju.edu.cn,
feiyu.xfy@alibaba-inc.com

Abstract

Multimodal relation extraction is an essential task for knowledge graph construction. In this paper, we take an in-depth empirical analysis that indicates the inaccurate information in the visual scene graph leads to poor modal alignment weights, further degrading performance. Moreover, the visual shuffle experiments illustrate that the current approaches may not take full advantage of visual information. Based on the above observation, we further propose a strong baseline with an implicit fine-grained multimodal alignment based on Transformer for multimodal relation extraction. Experimental results demonstrate the better performance of our method. Codes are available at https://github.com/zjunlp/DeepKE/tree/main/example/re/multimodal.

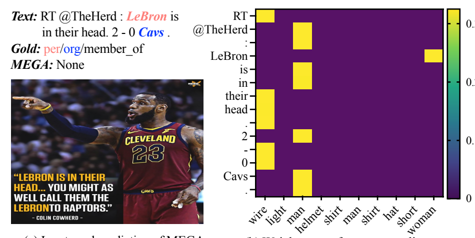
Introduction

Relation extraction (RE) aims to identify the semantic relations given two entities in a sentence, which plays an essential role in knowledge graph construction. However, existing mainstream RE methods (Soares and et al. 2019) are text-based and may suffer a sharp performance decline with social media texts since those sentences lack contexts. It is intuitive to supplement the missing semantic information with visual content to improve the performance.

Recently, Zheng and et al (2021a,b) introduce visual-enhanced relation extraction, also known as **multimodal relation extraction** (MRE), which aims to classify relations between two entities with the assistance of visual contents. The SOTA method MEGA (Zheng and et al 2021b) presents an efficient strategy to find the mapping from visual to textual contents, finally improving the MRE performance.

In this paper, we take a deeper look at the internal mechanism of MRE and obtain interesting empirical findings. We find that not all visual information contributes to performance improvement. Moreover, the empirical analysis illustrates that the missing and misleading information in the scene graph may interfere with the decision-making, and the model is not as effective as the pure text-based RE models for some relations. Building on the above discovery, we work on developing a more realistic visual-enhanced RE model. We propose an **I**mplicit **F**ine-grained multimodal **A**lignment approach with Trans**former** (IFAformer), which aligns visual and textual objects in representation space.

*Corresponding author
Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



(a) Inputs and prediction of MEGA (b) Weight scores for structure alignment

Figure 1: Analysis of the alignment weight scores between visual objects and text words in the MEGA method.

Analysis of the Role of Image for MRE

Since the improvement of MEGA is relatively small compared with text-based model MTB in MNRE dataset, we cannot help but question: *Is the ceiling for the improvement of the performance of visual-enhanced RE so low?* We conduct *visual shuffle experiments* to further explore the criticality of visual information, where we randomly shuffle the image-text pairs to compare the performance with the standard dataset. Then we analyze the structure alignment map of MEGA in Figure 1. We propose the following analysis discussion:

Stable or Invalid Performance? As shown in Table 1, although the image-text pairs are mismatched, the performance of MEGA does not drop at all. Besides, we notice that the original MEGA (trained on standard) can obtain stable performance for the mismatched image-text pairs. Our findings raise doubts about the ability of the model to leverage visual guidance with matched visual features.

Is Scene Graph Aligned with Tokens? As shown in Figure 1, some visual objects are duplicated (man, shirt) and incorrect (wire), which guides incorrect alignment weights for MEGA. We argue these may be attributed to two aspects: (1) visual objects generated by scene graph are general and straightforward ones, while entities in the text are specific; (2) coarse-grained alignment cannot fill the semantic gap between visual and text. Overall, the coarse-grained alignment of scene graph may fail to leverage visual guidance for RE.

Method	Baseline (standard)				Shuffle (train)				Shuffle (test)			
	Acc	Precision	Recall	F1	Acc	Precision	Recall	F1	Acc	Precision	Recall	F1
MTB	75.34	63.28	65.16	64.20	75.34	63.28	65.16	64.20	75.34	63.28	65.16	64.20
MEGA	76.15	64.51	68.44	66.41	76.15	65.35	64.53	64.94	75.40	63.30	66.56	64.89
IFAformer	92.38	82.59	80.78	81.67	74.23	60.84	67.97	64.21	47.71	29.82	29.22	29.52

Table 1: Visual shuffle experiment on MNRE. Baseline (standard) is the standard unshuffle setting. Shuffle (train) refer to randomly shuffling the images of the training set. IFAformer is the version of Vanilla IFAformer with Visual Objects.

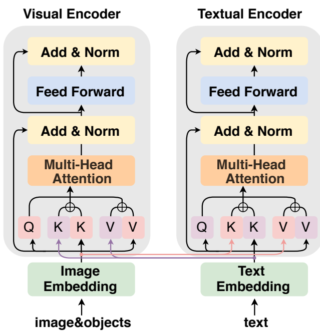


Figure 2: Implicit Fine-grained Multimodal Alignment.

Methods	Acc	Precision	Recall	F1
PCNN*	73.36	69.14	43.75	53.59
BERT*	71.13	58.51	60.16	59.32
MTB*	75.34	63.28	65.16	64.20
BERT+SG+Att.	74.59	60.97	66.56	63.64
ViLBERT	74.89	64.50	61.86	63.61
MEGA	76.15	64.51	68.44	66.41
Vanilla IFAformer	87.75	69.90	68.11	68.99
w/o Text Attn.	76.21	66.95	61.72	64.23
w/ Visual Objects	**92.38**	**82.59**	**80.78**	**81.67**

Table 2: The overall performance of baselines on MNRE.

A Strong Baseline: IFA

Based on above observation, we try to mitigate the erroneous graph alignment and propose a strong baseline IFAformer.

Multimodal Representation. As shown in Figure 2, we adopt a transformer-based dual-stream architecture to encode multi-modal inputs. For visual representation, we leverage raw image as the global image and employ visual grounding and object detection to obtain fine-grained subgraphs as local visual objects. For textual representation, we tokenize the text into token sequence, then feed it into the textual encoder.

Implicit Fine-grained Multimodal Alignment. We apply an implicit token-object alignment via multi-granularity signals at each layer of encoders to capture correlation between visual objects and entities. Given the hidden features $\boldsymbol{H}_t^l, \boldsymbol{H}_v^l \in \mathbb{R}^{n\times d}$ at the l-th layer of text and visual encoder, respectively. We project them into query/key/value vector:

$$\boldsymbol{Q}^l, \boldsymbol{K}^l, \boldsymbol{V}^l = \boldsymbol{x}\boldsymbol{W}_q^l, \boldsymbol{x}\boldsymbol{W}_k^l, \boldsymbol{x}\boldsymbol{W}_v^l; \boldsymbol{x} \in \{\boldsymbol{H}_t^l, \boldsymbol{H}_v^l\} \quad (1)$$

where $\boldsymbol{W}_q^l, \boldsymbol{W}_k^l, \boldsymbol{W}_v^l \in \mathbb{R}^{d\times d_h}$ are attention projection parameters. Then the hidden features at (l+1)-th layer through the multi-head attention can be calculated as follows:

$$\begin{aligned}\boldsymbol{H}_t^{l+1} &= \mathrm{Attn}(\boldsymbol{Q}_t^l, [\boldsymbol{K}_v^l, \boldsymbol{K}_t^l], [\boldsymbol{V}_v^l, \boldsymbol{V}_t^l]) \\ \boldsymbol{H}_v^{l+1} &= \mathrm{Attn}(\boldsymbol{Q}_v^l, [\boldsymbol{K}_t^l, \boldsymbol{K}_v^l], [\boldsymbol{V}_t^l, \boldsymbol{V}_v^l])\end{aligned} \quad (2)$$

Lastly, we use the output of textual encoder to do prediction.

Experiments

The overall results can be seen in Table 2. We observe that our Vanilla IFAformer is superior to all text-based models and the newest SOTA model MEGA. From Table 1, we can find the performance of IFAformer drops significantly when disrupting the image-text pairs, revealing that our model indeed utilizes the visual information for multimodal relation extraction. In addition, we conduct an ablation study and observe that: (1) *w/o Text Attn.*: Removing textual information in the visual encoder will reduce performance, revealing that fusing multiple information plays a vital role in MRE. (2) *w/ Visual Objects.*: Incorporating fine-grained visual objects as described in the Method section significantly improve performance and outperforms MEGA 15.26% F1 scores, indicating the significance of the fine-grained visual features for MRE.

Conclusion

In this paper, we study MRE and take an in-depth empirical analysis that indicates the pain points of current MRE methods. We further propose a strong baseline IFAformer. Experimental results demonstrate the effectiveness.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.62206246), Zhejiang Provincial Natural Science Foundation of China (No. LGG22F030011), Ningbo Natural Science Foundation (2021J190), Yongjiang Talent Introduction Programme (2021A-156-G), and CAAI-Huawei MindSpore Open Fund.

References

Soares, L. B.; and et al. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. In *ACL (1)*.

Zheng, C.; and et al. 2021a. MNRE: A Challenge Multimodal Dataset for Neural Relation Extraction with Visual Evidence in Social Media Posts. In *ICME*.

Zheng, C.; and et al. 2021b. Multimodal Relation Extraction with Efficient Graph Alignment. In *ACM Multimedia*.