Multimodal Relation Extraction with Cross-Modal Retrieval and Synthesis

Xuming Hu[1], **Zhijiang Guo**[2†], **Zhiyang Teng**[3], **Irwin King**[4], **Philip S. Yu**[1,5]
[1]Tsinghua University, [2]University of Cambridge, [3]Nanyang Technological University,
[4]The Chinese University of Hong Kong, [5]University of Illinois at Chicago
[1]hxm19@mails.tsinghua.edu.cn [2]zg283@cam.ac.uk
[3]zhiyang.teng@ntu.edu.sg [4]king@cse.cuhk.edu.hk [5]psyu@uic.edu

Abstract

Multimodal relation extraction (MRE) is the task of identifying the semantic relationships between two entities based on the context of the sentence image pair. Existing retrieval-augmented approaches mainly focused on modeling the retrieved textual knowledge, but this may not be able to accurately identify complex relations. To improve the prediction, this research proposes to retrieve textual and visual evidence based on the object, sentence, and whole image. We further develop a novel approach to synthesize the object-level, image-level, and sentence-level information for better reasoning between the same and different modalities. Extensive experiments and analyses show that the proposed method is able to effectively select and compare evidence across modalities and significantly outperforms state-of-the-art models. Code and data are available[1].

1 Introduction

Relation extraction aims to detect relations among entities in the text and plays an important role in various applications (Zhang et al., 2017; Soares et al., 2019). Early efforts mainly focus on predicting the relations based on the information from one single modality i.e., text. Recently, multimodal relation extraction (MRE) has been proposed to enhance textual representations with the aid of visual clues from images (Zheng et al., 2021a; Chen et al., 2022; Wang et al., 2022). It extends the text-based approaches by providing visual contexts to address the common ambiguity issues in identifying relations. Figure 1 shows an example from the MNRE dataset (Zheng et al., 2021b). To infer the relation between entities *Ang Lee* and *Oscar*, the model needs to capture the interactions from visual relations between objects in an image to textual relations in a sentence. The visual relation "holding" between two objects helps to detect the relation **awarded** between two textual entities.

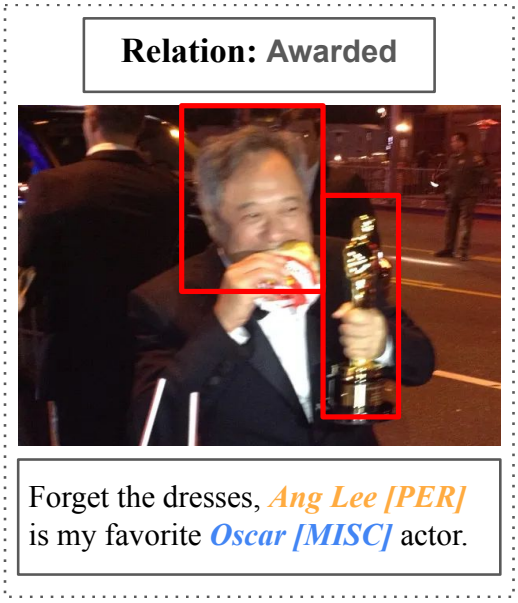

Figure 1: Example from MNRE. Entities are highlighted. Objects are denoted by the bounding boxes.

Most existing efforts focus on modeling the visual and textual content of the input. Zheng et al. (2021a) constructed the textual and visual graphs, then identify the relations based on graph alignments. Chen et al. (2022) presents a hierarchical visual prefix fusion network to incorporate hierarchical multi-scaled visual and textual features. Li et al. (2023a) proposes a fine-grained multimodal alignment approach with Transformer, which aligns visual and textual objects in representation space. Wang et al. (2022) first proposes retrieval-augmented multimodal relation extraction. The given image and sentence are used to retrieve textual evidence from the knowledge base constructed based on Wikipedia. Unlike previous retrieval-based models, we not only retrieve texts but also retrieve visual and textual evidence related to the object, sentence, and entire image. A novel strategy is used to combine evidence from the ob-

[1]https://github.com/THU-BPM/MRE
[†]Corresponding Author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 303–311

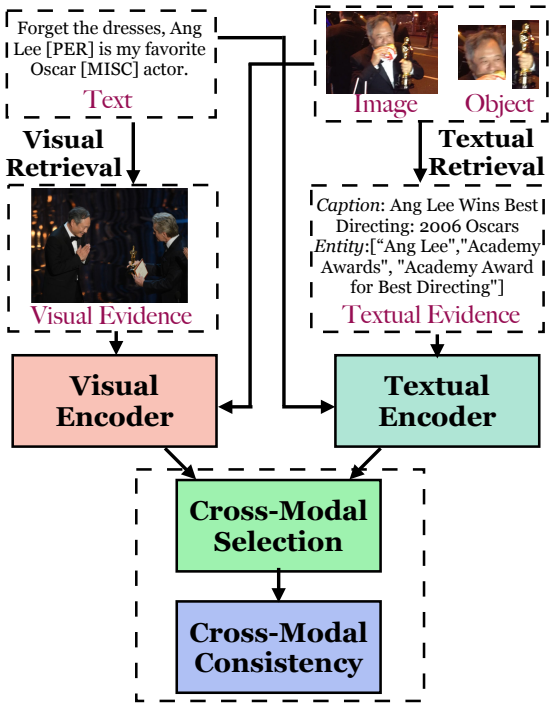

Figure 2: Overview of the model.

ject, sentence, and image levels in order to make better reasoning across modalities. Our key contributions are summarized as follows:

- We use cross-modal retrieval for obtaining multimodal evidence. To improve prediction accuracy, we further synthesize visual and textual information for relational reasoning.
- We evaluate our method on the MRE benchmark. Extensive experimental results validate the effectiveness of the proposed approach.

2 Methodology

2.1 Cross-Modal Retrieval

This module aims to retrieve visual evidence based on the input text (sentence, entities), and textual evidence based on the input image and objects.

Textual evidence We first obtain the local visual objects with top m salience by using the visual grounding toolkit (Yang et al., 2019): $V_{obj} = \{V^1_{obj}, V^2_{obj}, \cdots, V^m_{obj}\}$. Then we retrieve V_{img} and V_{obj} using Google Vision APIs[2] to obtain textual evidence, which returns a list of entities E_{entity} that describe the content of the V_{img} and V_{obj} and provide a more effective explanation for the visual content. In addition to $Entity$, the APIs could return the images' URLs and the containing pages' URLs. We propose a web crawler to search the images' URLs in the containing pages' and then return the captions $E_{caption}$ if found. Note that E_{entity} and $E_{caption}$ contain 10 entities and captions obtained for each V_{img} and V_{obj} as retrieval textual evidence.

Visual Evidence We use the textual content T of the post to retrieve the visual evidence. More specially, we leverage the Google custom search API[3] to retrieve the 10 images E_{image} for the textual content in each post.

2.2 Cross-Modal Synthesis

Given the retrieved visual and textual evidence, this module aims to synthesize multimodal information for relation extraction.

2.2.1 Visual Encoder

The visual encoder module encodes the visual content V_{img}, V_{obj} and retrieved visual evidence E_{image} of the post. First, we adopt the ResNet (He et al., 2016) which is pretrained on the ImageNet dataset (Deng et al., 2009) to obtain the visual embedding $\boldsymbol{h}_v \in \mathbb{R}^{n\times d}$, where n and d represents the number of images and the hidden dimension. To fuse the cross-modal visual and textual information, we employ a learnable linear layer $\boldsymbol{h}_v = \mathbf{W}_\phi \boldsymbol{h}_v + \mathbf{b}_\phi$

2.2.2 Textual Encoder

The textual module encodes the textual content T and retrieved textual evidence E_{entity}, $E_{caption}$ of the post. For each sentence $X = [x_1, x_2, .., x_M]$ in the textual content T where two entities $[E_1]$ and $[E_2]$ are mentioned, we follow the labeling schema adopted in Soares et al. (2019) and argument X with four reserved tokens $[E_1]$, $[/E_1]$, $[E_2]$, $[/E_2]$ to mark the beginning and the end of each entity mentioned in the sentence:

$$\begin{aligned} X = &[x_1, ..., [E_1], x_i, ..., x_{j-1}, [/E_1], \\ &..., [E_2], x_k, ..., x_{l-1}, [/E_2], ..., x_M], \end{aligned} \quad (1)$$

as the input token sequence. We adopt BERT (Devlin et al., 2019) as an encoder and obtain the textual embedding $\boldsymbol{h}_t \in \mathbb{R}^{(M+4)\times d}$, where M and d represents the number of tokens in s and the hidden dimensions. Thanks to informative visual

[2] https://cloud.google.com/vision/docs/detecting-web

[3] https://developers.google.com/custom-search/v1

embeddings, we can better capture the correlation between visual content and textual information.

2.2.3 Cross-Modal Selection

Given the encoded multimodal evidence and inputs $\boldsymbol{h}_t^l \in \mathbb{R}^{(M+4)\times d}$, $\boldsymbol{h}_v^l \in \mathbb{R}^{n\times d}$. The module selects visual/textual evidence and compares it against the input image/sentence. Inspired by Vaswani et al. (2017), we leverage multi-head attention to perform the cross-modal selection. We first project the presentations as query, key, and value vectors:

$$\boldsymbol{Q}^l, \boldsymbol{K}^l, \boldsymbol{V}^l = \boldsymbol{x}\boldsymbol{W}_q^l, \boldsymbol{x}\boldsymbol{W}_k^l, \boldsymbol{x}\boldsymbol{W}_v^l; \boldsymbol{x} \in \left\{\boldsymbol{h}_t^l, \boldsymbol{h}_v^l\right\}, \quad (2)$$

where $\boldsymbol{W}_q^l, \boldsymbol{W}_k^l, \boldsymbol{W}_v^l \in \mathbb{R}^{d\times d_h}$ represent attention projection parameters. We then obtain the hidden features at $(l+1)$-th layer via multi-head attention:

$$\begin{aligned}\boldsymbol{h}_t^{l+1} &= \text{Attn}\left(\boldsymbol{Q}_t^l, \left[\boldsymbol{K}_v^l, \boldsymbol{K}_t^l\right], \left[\boldsymbol{V}_v^l, \boldsymbol{V}_t^l\right]\right), \\ \boldsymbol{h}_v^{l+1} &= \text{Attn}\left(\boldsymbol{Q}_v^l, \left[\boldsymbol{K}_t^l, \boldsymbol{K}_v^l\right], \left[\boldsymbol{V}_t^l, \boldsymbol{V}_v^l\right]\right).\end{aligned} \quad (3)$$

Note that the textual features $\boldsymbol{h}_t$ come from two types: The first is the textual content in the post with two entities, so we get the relational features of the $[E_1]$and $[E_2]$ positions. The other is retrieved textual evidence, since it does not have entities, we obtain representations of the CLS position:

$$\begin{aligned}\boldsymbol{h}_{t,content} &= \text{Avg.}(\boldsymbol{h}_{t,[E_1]}, \boldsymbol{h}_{t,[E_2]}), \\ \boldsymbol{h}_{t,retrieved} &= \boldsymbol{h}_{t,[CLS]}.\end{aligned} \quad (4)$$

where $\boldsymbol{h}_t = \{\boldsymbol{h}_{t,content}, \mathbf{h}_{t,retrieved}\} \in \mathbb{R}^d$ is the representation of the textual content and retrieved textual evidence for each post, where d is the embedding size 768. Similarly, we use a learnable linear layer $\boldsymbol{h}_t = \mathbf{W}_\theta \boldsymbol{h}_t + \mathbf{b}_\theta$ to change the dimension d from 768 to 2048 and employ the multi-head attention in Eq. 2, 3, and 4 to update the visual content and retrieved visual evidence.

2.2.4 Cross-Modal Consistency

This module aims to evaluate the consistency between the retrieved textual and visual evidence and the original post. A natural idea is to leverage the textual and visual content in the original post to update the retrieved textual and visual evidence. We could obtain the updated evidence $\boldsymbol{h}_{t,retrieved}$ and $\boldsymbol{h}_{v,retrieved}$ with $\boldsymbol{h}_{t,content}$ and $\boldsymbol{h}_{v,content}$ as:

$$\begin{aligned}\boldsymbol{h}_{t,r.} &= \text{softmax}(\frac{\boldsymbol{h}_{t,c.}\mathbf{W}_t \times (\boldsymbol{h}_{t,r.}\mathbf{W}_t^{'})^T}{\sqrt{d_t}})\boldsymbol{h}_{t,r.}, \\ \boldsymbol{h}_{v,r.} &= \text{softmax}(\frac{\boldsymbol{h}_{t,c.}\mathbf{W}_v \times (\boldsymbol{h}_{v,r.}\mathbf{W}_v^{'})^T}{\sqrt{d_v}})\boldsymbol{h}_{v,r.},\end{aligned} \quad (5)$$

where $\mathbf{W}_t, \mathbf{W}_t^{'} \in \mathbb{R}^{768\times 768}$ and $\mathbf{W}_v, \mathbf{W}_v^{'} \in \mathbb{R}^{2048\times 2048}$ are trainable projection matrices and d_t, d_v are hyperparameters.

2.3 Classifier

We concatenate the resulting representations to form the final multimodal representations and leverage a feed-forward neural network to predict the relation:

$$\boldsymbol{h}_{final} = \text{FFNN}([\boldsymbol{h}_{t,c.}; \boldsymbol{h}_{t,r.}; \boldsymbol{h}_{v,c.}; \boldsymbol{h}_{v,r.}]), \quad (6)$$

where $\boldsymbol{h}_{final}$ is then fed into a linear layer followed by a softmax operation to obtain a probability distribution $p \in \mathbb{R}^m$ over m relation labels.

3 Experiments and Analyses

3.1 Experimental Setup

We evaluate the model on MNRE (Zheng et al., 2021b), which contains 12,247/1,624/1,614 samples in train/dev/test sets, 9,201 images, and 23 relation types. Following prior efforts, we adopt Accuracy, Precision, Recall, and F1 as the evaluation metrics. For fair comparisons, all baselines and our method use ResNet50 (He et al., 2016) as the visual backbone and BERT-base (Devlin et al., 2019) as the textual encoder. We computed the Accuracy and Macro F1 as the evaluation metric. The hyper-parameters are chosen based on the development set. Results are reported with mean and standard deviation based on 5 runs. For the textual encoder of the retrieval-based model, we use the BERT-Base default tokenizer with a max-length of 128 to preprocess data. For the visual encoder of the retrieval-based model, we use ResNet 50 to encode the visual images. We scale the image proportionally so that the short side is 256, and crop the center to $224 * 224$. For the feed-forward neural network of the classifier, we set the layer dimensions as h_R-1024-verification_labels, where $h_R = 768 * 2 + 2048 * 2$. We use BertAdam with 3e-5 learning rate, warmup with 0.06 to optimize the cross-entropy loss and set the batch size as 16.

3.2 Baselines

We adopt two types of baselines:

Text-based Baselines only encode text content: (1) PCNN (Zeng et al., 2015), (2) BERT (Devlin et al., 2019), and (3) MTB (Soares et al., 2019).

Methods		Accuracy	Precision	Recall	F1
Text Based	PCNN	73.15	62.85	49.69	55.49
	BERT	74.42	58.58	60.25	59.40
	MTB	75.69	64.46	57.81	60.86
Multi modal	UMT	77.84	62.93	63.88	63.46
	UMGF	79.27	64.38	66.23	65.29
	BSG	77.15	62.95	62.65	62.80
	MEGA	80.05	64.51	68.44	66.41
	VBERT	73.97	57.15	59.48	58.30
	MoRe	79.87	65.25	67.32	66.27
	Iformer	92.38	82.59	80.78	81.67
	HVPnet	92.52	82.64	80.78	81.85
Ours		**93.54**±0.16	**85.03**±0.14	**84.25**±0.17	**84.64**±0.16
w/o Object Evi.		92.37±0.16	83.02±0.14	82.36±0.18	82.69±0.15
w/o Image Evi.		92.83±0.15	83.44±0.18	83.15±0.15	83.29±0.17
w/o Visual Evi.		92.72±0.17	82.78±0.19	83.63±0.24	83.20±0.21
w/o Selection		92.75±0.16	82.81±0.14	83.44±0.16	83.12±0.16
w/o Consistency		92.68±0.15	83.40±0.13	82.71±0.16	83.05±0.15

Table 1: The overall performance on MNRE.

Multi-modal Baselines encode both text and image contents: (1) UMT (Yu et al., 2020) adopts the multimodal interaction module to obtain the token representations incorporated with visual information and visual representations. (2) UMGF (Zhang et al., 2021) adopts a unified multi-modal graph fusion method. (3) BSG (Zheng et al., 2021a) adopts the textual representation from BERT and the visual characteristics produced by the scene graph (SG). (4) MEGA (Zheng et al., 2021b) adopts a dual graph, which could align multi-modal features between entities and objects to improve performance. (5) VBERT (Li et al., 2019) adopts the single-stream structure which is different from the attention-based methods. (6) MoRe (Wang et al., 2022) obtains more textual information by retrieving images and titles, thereby improving the accuracy of relation classification and named entity recognition. (7) Iformer (Li et al., 2023a) increases the amount of information in the image by detecting the objects. (8) HVPnet (Chen et al., 2022) treats visual representations as visual prefixes that can be inserted to guide textual representations of error-insensitive prediction decisions.

3.3 Main Results

Table 1 shows the mean and standard deviation results with 5 runs of training and testing on MRNE. We first compare text-based and multi-modal baselines and observe the performance improvement after incorporating visual content, indicating that images can help reveal the potential relationship between two entities. For the multi-modal model, Iformer (Li et al., 2023a) and HVPnet (Chen et al., 2022) specifically detect the objects in the image and achieve the average 17.23% F1 and 14.15% Accuracy compared with other multi-modal baselines. Therefore, we retrieve textual and visual evidence based on the object, sentence, and whole image, and achieve an average of 2.79% F1 and 1.02% Accuracy gains compared to the best-reported model HVPnet. Thanks to the retrieved visual and textual evidence, the text and image content in the original post is further explained, which helps our model obtain valuable clues to classify the relations between two entities.

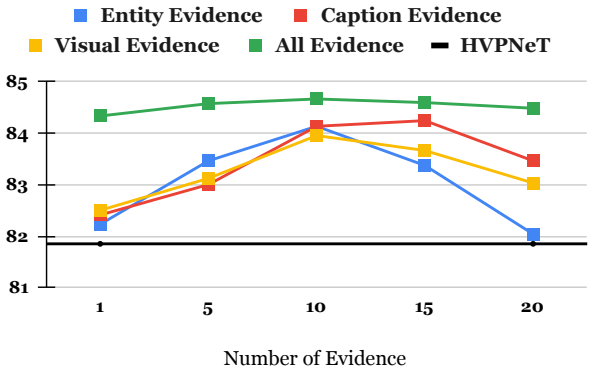

Figure 3: Comparison of different amounts of evidence.

3.4 Analysis and Discussion

Ablation Study. We conduct an ablation study to show the effectiveness of different modules of our model on the test set. Ours *w/o Object Evidence* and Ours *w/o Image Evidence* remove the descriptions of Objects and Images respectively in the retrieved textual evidence. Correspondingly, Ours *w/o Visual Evidence* removes the visual evidence for text content retrieval. The results from Table 1 demonstrate that the three types of evidence can bring 1.95%, 1.35%, and 1.44% F1 improvements, respectively. Among them, the textual evidence obtained from the object retrieval brings the greatest benefit, which is also related to the potential entity information contained in the object. The removal of the *Cross-Modal Selection* and *Cross-Modal Consistency* modules means that we no longer use the appropriate evidence selection and update the retrieved evidence with the original content, which increases the noise from irrelevant evidence and leads to 1.52% F1 and 1.59% F1 down.

Analyze the Impact of Evidence. In Figure 3, we vary the numbers of retrieved visual and textual evidence from 1 $\sim$ 20 and report the F1 on the test set. The fluctuation results indicate that both the quantity and quality of retrieved evidence affect the performance. Using less textual or vi-

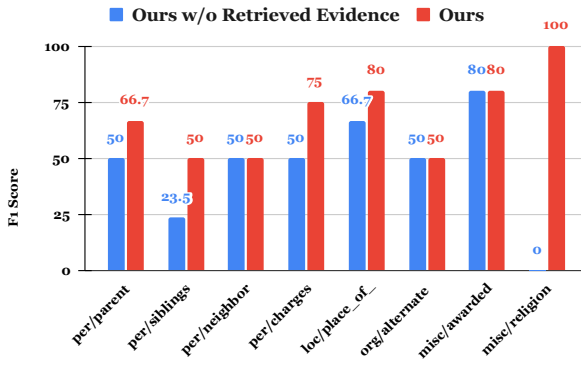

Figure 4: F1 performance changes of the tail relations.

sual evidence cannot bring enough explanation to the original post, which leads to a decrease in the quality of the model classification. Using too much evidence will introduce false or irrelevant evidence noise, affecting performance. However, no matter how much evidence is adopted, our method consistently outperforms HVPnet, which illustrates the effectiveness of adding evidence. In our model, we adopt 10 textual and visual evidence for each post to achieve the best performance. We believe the Cross-Modal Consistency module can alleviate the irrelevant noise so that the model can obtain helpful auxiliary evidence.

Analyze Performance Changes in Tail Relations. We select the tail relations with the least number of data among the 23 relation classes in MNRE, and study their F1 performance changes after adding retrieval evidence in Figure 4. Compared with the 2.79% improvement brought by the evidence on all relations, we find that almost all tail relations can get more than 22.68% F1 improvement (46.28 vs. 68.96), which shows that the retrieved evidence is more helpful for the few-shot tail relation types: It is an attractive property in real-world applications since classes of tail relations are usually more difficult to obtain training labeled data to improve.

4 Related Work

Relation extraction has garnered considerable interest in the research community due to its essential role in various natural language processing applications (Guo et al., 2019; Nan et al., 2020; Hu et al., 2021b,a). The initial efforts in this field focused on detecting relations between entities in the text, with different neural architectures (Zeng et al., 2015; Zhang et al., 2017; Guo et al., 2020) and pre-trained language models (Soares et al., 2019; Devlin et al., 2019) used to encode the textual information. Multimodal relation extraction has recently been proposed, where visual clues from images are used to enhance entity representations (Zheng et al., 2021a,b; Chen et al., 2022; Wang et al., 2022). Most existing efforts focus on fusing the visual and textual modalities effciently. Zheng et al. (2021b) constructed the dual modality graph to align multimodal features among entities and objects. Chen et al. (2022) concatenated object-level visual representation as the prefix of each self-attention layer in BERT. Li et al. (2023a) introduced a fine-grained multimodal fusion approach to align visual and textual objects in representation space. Closest to our work, Wang et al. (2022) proposed to retrieve textual information related to the entities based on the given image and sentence. Unlike prior efforts, we not only retrieve texts related to entities but also retrieve visual and textual evidence related to the object, sentence, and entire image. We further synthesize the retrieved object-level, image-level, and sentence-level information for better reasoning between the same and different modalities.

5 Conclusion and Future Work

We propose to retrieve multimodal evidence and model the interactions among the object, sentence, and whole image for better relation extraction. Experiments show that the proposed method achieves competitive results on MNRE. For future research directions, we can utilize open-source image search and caption generation tools to retrieve textual and image evidence. For example, to retrieve visual evidence, one can (1) use a web crawler to search Google Images, or (2) utilize a searchable image database: PiGallery[4], where images can be sourced from Open Image Dataset[5], which contains ∼9 million images. For retrieving textual evidence, one can use CLIP to generate image captions. Moreover, we can also apply the method of multimodal retrieval to low-resource relation extraction (Hu et al., 2020; Liu et al., 2022b; Hu et al., 2023), natural language inference (Li et al., 2023b, 2022), semantic parsing (Liu et al., 2022a, 2023), and other NLP tasks, thus realizing information enhancement based on images and retrieval.

[4] https://github.com/vladmandic/pigallery

[5] https://storage.googleapis.com/openimages/web/factsfigures_v7.html

6 Limitation

In this paper, we suggest incorporating textual and visual data from search engines for multimodal relation extraction. Despite the fact that the proposed model yields competitive results on the benchmark, it still has several limitations. Firstly, using a search engine is a feasible way to obtain related knowledge, but it also brings the issue of noisy evidence. Unrelated visual and textual evidence returned by the search engine may lead to incorrect predictions from the model. Additionally, not all the retrieved evidence is equally reliable, and sometimes sources may contradict each other. On the other hand, retrieval-augmented methods are slower than content-based counterparts, since retrieving evidence from the Internet requires extra time. Therefore, it may not satisfy some of the time-sensitive scenarios. Lastly, evidence may be presented in different forms other than texts and images. For instance, structural information such as tables, info lists, and knowledge graphs also provide important contexts for identifying semantic relations. Humans are able to extract relevant information from these heterogeneous sources for inference, while our relation extraction system can only model and reason over textual and visual evidence.

7 Acknowledgement

We thank the reviewers for their valuable comments. The work described here was partially supported by grants from the National Key Research and Development Program of China (No. 2018AAA0100204) and from the Research Grants Council of the Hong Kong Special Administrative Region, China (CUHK 14222922, RGC GRF, No. 2151185), NSF under grants III-1763325, III-1909323, III-2106758, and SaTC-1930941. Zhiyang Teng was partially supported by CAAI-Huawei MindSpore Open Fund (CAAIXSJLJJ-2021-046A).

References

Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1607–1618, Seattle, United States. Association for Computational Linguistics.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhijiang Guo, Guoshun Nan, Wei Lu, and Shay B. Cohen. 2020. Learning latent forests for medical relation extraction. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3651–3657. ijcai.org.

Zhijiang Guo, Yan Zhang, and Wei Lu. 2019. Attention guided graph convolutional networks for relation extraction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 241–251. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Xuming Hu, Zhaochen Hong, Chenwei Zhang, Irwin King, and Philip S Yu. 2023. Think rationally about what you see: Continuous rationale extraction for relation extraction. *arXiv preprint arXiv:2305.03503*.

Xuming Hu, Lijie Wen, Yusong Xu, Chenwei Zhang, and Philip S. Yu. 2020. Selfore: Self-supervised relational feature learning for open relation extraction. In *Proc. of EMNLP*, pages 3673–3682.

Xuming Hu, Chenwei Zhang, Fukun Ma, Chenyao Liu, Lijie Wen, and Philip S. Yu. 2021a. Semi-supervised relation extraction via incremental meta self-training. In *Findings of EMNLP*, pages 487–496.

Xuming Hu, Chenwei Zhang, Yawen Yang, Xiaohe Li, Li Lin, Lijie Wen, and Philip S. Yu. 2021b. Gradient imitation reinforcement learning for low resource relation extraction. In *Proc. of EMNLP*, pages 2737–2746.

Lei Li, Xiang Chen, Shuofei Qiao, Feiyu Xiong, Huajun Chen, and Ningyu Zhang. 2023a. On analyzing the role of image for visual-enhanced relation extraction. In *In Proc. of AAAI*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Shu'ang Li, Xuming Hu, Li Lin, Aiwei Liu, Lijie Wen, and Philip S. Yu. 2023b. A multi-level supervised contrastive learning framework for low-resource natural language inference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1771–1783.

Shu'ang Li, Xuming Hu, Li Lin, and Lijie Wen. 2022. Pair-level supervised contrastive learning for natural language inference. *arXiv preprint arXiv:2201.10927*.

Aiwei Liu, Xuming Hu, Li Lin, and Lijie Wen. 2022a. Semantic enhanced text-to-sql parsing via iteratively learning schema linking graph. In *Proc. of KDD*, pages 1021–1030.

Aiwei Liu, Xuming Hu, Lijie Wen, and Philip S Yu. 2023. A comprehensive evaluation of chatgpt's zero-shot text-to-sql capability. *arXiv preprint arXiv:2303.13547*.

Shuliang Liu, Xuming Hu, Chenwei Zhang, Shu'ang Li, Lijie Wen, and Philip S. Yu. 2022b. Hiure: Hierarchical exemplar contrastive learning for unsupervised relation extraction. In *Proc. of NAACL-HLT*, pages 5970–5980.

Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1546–1557. Association for Computational Linguistics.

Livio Baldini Soares, Nicholas Fitzgerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *In Proc. of ACL*, pages 2895–2905.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Xinyu Wang, Jiong Cai, Yong Jiang, Pengjun Xie, Kewei Tu, and Wei Lu. 2022. Named entity and relation extraction with multi-modal retrieval. In *In Proc. of EMNLP*.

Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. 2019. A fast and accurate one-stage approach to visual grounding. In *In Proc. of ICCV*, pages 4683–4693.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *In Proc. of ACL*, pages 3342–3352.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *In Proc. of EMNLP*, pages 1753–1762.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *In Proc. of AAAI*, volume 35, pages 14347–14355.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 35–45. Association for Computational Linguistics.

Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. 2021a. Multimodal relation extraction with efficient graph alignment. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 5298–5306. ACM.

Changmeng Zheng, Zhiwei Wu, Junhao Feng, Ze Fu, and Yi Cai. 2021b. MNRE: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts. In *2021 IEEE International Conference on Multimedia and Expo, ICME 2021, Shenzhen, China, July 5-9, 2021*, pages 1–6. IEEE.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6

☑ A2. Did you discuss any potential risks of your work?
Section 6

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 2, Section 3

☑ B1. Did you cite the creators of artifacts you used?
Section 2, Section 3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 2, Section 3

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 2, Section 3

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3

C ☑ Did you run computational experiments?

Section 3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 3

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 2, Section 3

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.