Information Processing and Management 60 (2023) 103264

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

TSVFN: Two-Stage Visual Fusion Network for multimodal relation extraction

Qihui Zhao [a], Tianhan Gao [a], Nan Guo [b,*]

[a] Software College, Northeastern University, 195 Chuangxin Road, Hunnan District, Shenyang 110169, China
[b] School of Computer Science and Engineering, Northeastern University, 195 Chuangxin Road, Hunnan District, Shenyang 110169, China

ARTICLE INFO

Keywords:
Multimodal relation extraction
Two-stage visual fusion
Multimodal pretrained model
Multimodal graph
Graph convolution networks
Vision transformer
Information extraction

ABSTRACT

Multimodal relation extraction is a critical task in information extraction, aiming to predict the class of relations between head and tail entities from linguistic sequences and related images. However, the current works are vulnerable to less relevant visual objects detected from images and are not able to sufficiently fuse visual information into text pre-trained models. To overcome these problems, we propose a **T**wo-**S**tage **V**isual **F**usion **N**etwork (**TSVFN**) that employs the multimodal fusion approach in vision-enhanced entity relation extraction. In the first stage, we design multimodal graphs, whose novelty lies mainly in transforming the sequence learning into the graph learning. In the second stage, we merge the transformer-based visual representation into the text pre-trained model by a multi-scale cross-model projector. Specifically, two multimodal fusion operations are implemented inside the pre-trained model respectively. We finally accomplish deep interaction of multimodal multi-structured data in two fusion stages. Extensive experiments are conducted on a dataset (MNRE), our model outperforms the current state-of-the-art method by 1.76%, 1.52%, 1.29%, and 1.17% in terms of accuracy, precision, recall, and F1 score, respectively. Moreover, our model also achieves excellent results under the condition of fewer samples.

1. Introduction

Relation extraction (RE) is an important and fundamental task in natural language processing, whose performance affects many downstream tasks, such as question and answer systems (drissiya El-allaly, Sarrouti, En-Nahnahi, & Ouatik El Alaoui, 2021; Han et al., 2020; Zaporojets, Deleu, Develder, & Demeester, 2021). Most previous studies focused on text-based RE, which extracted relations through sentences and documents (Liu, Tan, & Dong, 2021; Wen, Liu, Ouyang, Lin, & Chung, 2021; Zaporojets et al., 2021). However, many entity relations exist among text and images in the real world. Text-based approaches fail to extract multimodal relations owing to different data input. In contrast to text-based RE, multimodal RE cannot obtain results directly from textual information and usually needs to combine information from other modalities (e.g., images). This leads to complex and variable semantics and contexts. Therefore, models of multimodal RE need not only to capture text information around the target entity, but also to comprehend context in other modalities associated with the target entity. In recent years, multimodal information extraction tasks have received extensive attentions, where Multimodal named entity recognition (MNER) and multimodal relation extraction (MRE) (Zheng, Feng, et al., 2021) are promoted very fast recently.

Multimodal relation extraction aims to predict relations between entities according to language modal and vision modal. Fig. 1 shows the example of MRE. The critical point of this task is how to use visual information to enhance the semantic and relation

* Corresponding author.
E-mail addresses: 1910459@stu.neu.edu.cn (Q. Zhao), gaoth@mail.neu.edu.cn (T. Gao), guonan@mail.neu.edu.cn (N. Guo).

https://doi.org/10.1016/j.ipm.2023.103264
Received 4 October 2022; Received in revised form 6 December 2022; Accepted 31 December 2022
Available online 9 January 2023

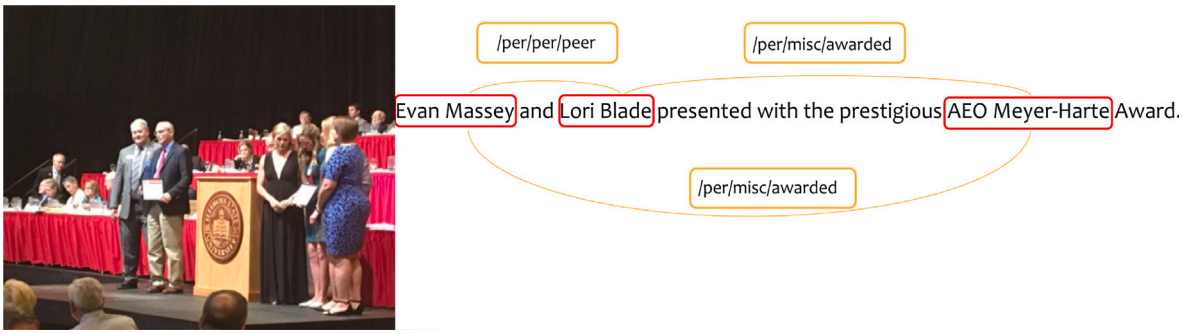

Fig. 1. An example of multimodal relation extraction. the image is on the left, the right side contains the target sentence with three entities' relation triples, <Evan Massey, /per/per/peer, Lori Blade>, <Evan Massey, /per/misc/awarded, AEO Meyer-Harte>, and <Lori Blade, /per/misc/awarded, AEO Meyer-Harte>.

representation of the entities. Zheng, Wu, et al. (2021) first propose the multimodal relation extraction task and build a related dataset, which lays a solid foundation for MRE. Zheng, Feng, et al. (2021) further design a semantic and structural alignment mechanism by aligning scene graphs with textual and structural features. This alignment operation is to learn text and vision modal representations, which enables the correct relation prediction. However, this method only utilizes a layer of attention calculation to achieve shallow alignment between images and text, the results are vulnerable to less relevant image. Besides, this method simply uses the concatenation of text features and image features to predict relations, which also introduces more noise. Xiang et al. (2022) add the vision prefix to the attention calculation of each layer to fuse visual information in the BERT (Devlin, Chang, Lee, & Toutanova, 2019) model. While the flaws of Xiang et al. (2022) are: (1) The fusion of image features into each block of the BERT is straightforward, resulting in insufficient integration of the two modalities. (2) resnet50 (He, Zhang, Ren, & Sun, 2016) is not as expressive as the transformer-based image encoder. (3) the image features extracted using Resnet50 (pyramidal structure) require linear projections on the vector space, causing a certain degree of information loss.

In this paper, we propose a **T**wo **S**tage **V**isual **F**usion **N**etwork (**TSVFN**) to solve the problems mentioned above. In the first stage, we construct a multimodal graph and use a Graph Neural Networks (GNN) (Gori & Scarselli, 2005; Kipf & Welling, 2017) based approach to process the fusion and alignment of cross-modal information. In the second stage, (1) the image feature obtained from the first stage are input into the vision transformer (Vit) (Dosovitskiy et al., 2021); (2) the visual knowledge vectors are extracted from each Vit block as the output; (3) the visual knowledge vectors are tackled into the multi-level cross-modal projector; (4) two approaches and vision-language alignment vector are designed to fuse visual knowledge into the BERT; (5) the text features embedding obtained in the first stage are input to the text modality model (BERT) to fuse the visual knowledge; Finally, the output of the BERT model is utilized to predict the relations between entities.

To test the effectiveness of the proposed model in this paper, we conduct rich experiments on a multimodal relation extraction dataset. We divide the dataset into two versions to verify the model's performance on a full dataset and a small number of datasets, respectively. The experiments show that our model's result on both datasets achieves the best performance. The main contributions of our model are as follows.

(1) A two-stage multimodal fusion framework is proposed, which innovatively combines the powerful modeling capabilities of graph neural networks and transformers networks to fully fuse critical information between visual and textual modalities, enabling the improved performance of MRE.

(2) We design two approaches and vision-language alignment vector to fuse external visual information into the text pre-training model in the second stage of multimodal fusion and verify the effectiveness through different experiments.

(3) In terms of accuracy, precision, recall, and micro F1 scores on MNRE, our model outperforms the current state-of-the-art method by 1.76%, 1.52%, 1.29%, and 1.17%, respectively. Moreover, our model also achieves excellent results under fewer samples.

2. Related works

2.1. Unimodal relation extraction

Unimodal relation extraction refers to identifying relations between entities in a single modality (textual modality). Relation extraction (Wang, Lu, Yin, & Qin, 2021) is an essential upstream task for many natural language processing tasks (knowledge graphs, reading comprehension, question and answer systems, etc.). Earlier approaches (Carlson et al., 2010) improve the performance by extensive manual feature engineering, which is a considerable and time-consuming workload. Wang (2008) designs complex kernel functions to construct SVM classifiers. With the development of deep learning, some methods that do not require feature engineering have been proposed. PCNN (Zeng, Liu, Chen, & Zhao, 2015) is one of the early masterpieces of deep learning-based model, which proposes the piecewise max pooling layer and introduces multi-instance learning, thus alleviating the prediction error caused by the poor quality of the dataset. In recent years, language pre-trained models have been developed significantly, and many works based on them have emerged (Wu & He, 2019; Zhou, Geng, Shen, Long, & Jiang, 2022). ERNIE (Zhang et al., 2019) and KnowBERT (Peters

et al., 2019) take the knowledge of entities from the knowledge graph and inject it into the pre-trained model, which makes the model learn the entity knowledge. K-Adapter (Liu et al., 2020) introduces a plug-in adaptor that fuses factual and linguistic knowledge into the pre-trained model. LUKE (Yamada, Asai, Shindo, Takeda, & Matsumoto, 2020) design a new pre-trained objective of masked language model and use an entity-aware self-attention mechanism. MTB (Soares, FitzGerald, Ling, & Kwiatkowski, 2019) fine-tunes pre-trained model using relation-based objective (matching-the-blanks objective) that automatically chooses whether two relation instances share the same entities. PURE (Zhong & Chen, 2021) offers a simple and effective method based on an entity type-marker to improve the task's performance. Although the unimodal relation extraction task has achieved noticeable results, it is clear that it can no longer satisfy the increasing multi-granularity of information.

2.2. Multimodal relation extraction

With the explosive growth of information on the Internet, images and videos have also become a rich resource. Multimodal relation extraction takes advantage of these large-scale corpora and works to extract relations from them. As a vivid way to convey information, images and videos can contain a lot of knowledge. On the one hand, humans like to use images to express some common sense knowledge instead of saying it explicitly. On the other hand, the combination of multimodal corpus shows promising results in many tasks. This phenomenon highlights the importance of obtaining relations from other resources like images rather than only text. Zheng, Wu, et al. (2021) present the first multimodal relation extraction dataset (MNRE) based on social media. This is the first dataset in the field of multi-modal relation extraction and is also the dataset used in this paper. MEGA (Zheng, Feng, et al., 2021) proposes a semantic-structural alignment mechanism that uses visual information to complement the missing text semantics, thus helping the model more accurately identify entities' relations in the text. There are three differences between our approach and that of MEGA. First of all, the method used in this paper is a two-stage text and image fusion and alignment method, and MEGA only uses only a single-stage method. Then, our method adopt the Vision Transformer-based image feature extractor, which is not only superior in performance to the image feature extractor used by MEGA, but also more consistent with our model because of its structure. Finally, our approach fully incorporates the image information into a language model, which is also far superior to MEGA's shallow attention alignment approach. HVPNet (Xiang et al., 2022) offers an approach to the transformer's attention computation by fusing visual information, which uses pyramidal feature maps and participates in the computation of attention in blocks assigned to BERT through gate structures. The method proposed in this paper has some intuitive similarities with HVPNet, and its improvement points compared to HVPNet are as follows: (1) HVPNet uses ResNet as the image feature extractor, while this paper uses Vision transformer as the backbone. Its advantage lies in the similar structure to the BERT, which can to some extent promote the fusion and alignment of visual information. (2) HVPNet uses a simple multimodal project method, while this paper uses a Multi-scaled Cross-Model Projector, which is experimentally proven to have a more excellent performance. (3) This paper designs two bert-based multimodal fusion methods and introduces a vision-language alignment vector, which can better improve the robustness of multimodal RE. (4) Our method uses a two-stage fusion approach, which can more fully fuse the information between different modalities. Recently, pre-trained models based on multimodal data have also been enhanced, especially vision-language models. These pre-trained models can also be employed in multimodal relation extraction tasks, but they often do not achieve excellent results. There are two reasons for this: First, most pre-trained models are based on the image-text matching task to train the models, which has a gap with multimodal relation extraction. Second, most existing pre-trained models pay equal attention to text and images. At the same time, in multimodal relation extraction tasks, textual modality tends to occupy a dominant position, and image information takes more of a secondary role.

2.3. Graph neural networks

Graph neural networks (GNN) (Gori & Scarselli, 2005; Micheli, 2009) have received increasing scholarly attention in recent years and are originally proposed with the intention of using deep neural networks to process graph-structured data. There are many implementations of GNN, such as graph convolutional networks (Kipf & Welling, 2017), gated graph neural networks (Li, Tarlow, Brockschmidt, & Zemel, 2016), and graph attention networks (Velickovic et al., 2018). In the multimodal domain, there are also some studies based on multimodal GNN approaches (Khademi, 2020). HetEmotionNet (Jia et al., 2021) proposes a dual-stream heterogeneous graph recurrent neural network classify multimodal emotion data. MM-GCN (Wei et al., 2019) uses the interaction behavior of users and short videos to guide the learning of representations in each modality, mainly using the information transfer idea to learn modality-specific user representations and short video representations. MF-NMT (Yin et al., 2020) offers a fusion approach based on multimodal graph structure to improve the performance of multimodal machine translation. Receiving inspiration from the above, we use a graph-based multimodal fusion mechanism in the first stage of visual fusion. We employ a text graph based on a dependent syntactic tree to construct multimodal graphs and introduce global nodes and GCN to fuse information from two modalities. There is a paucity of studies using a graph-based approach to modeling multimodal language sequences, especially images and text. Our approach in this paper effectively merges information from both images and text.

2.4. Multimodal pre-training

Recently, pre-trained models for vision and language have started to gain popularity due to the transformers-based BERT being adapted to receive modal data other than text. A series of studies (Li, Yatskar, Yin, Hsieh, & Chang, 2019; Lu, Batra, Parikh, &

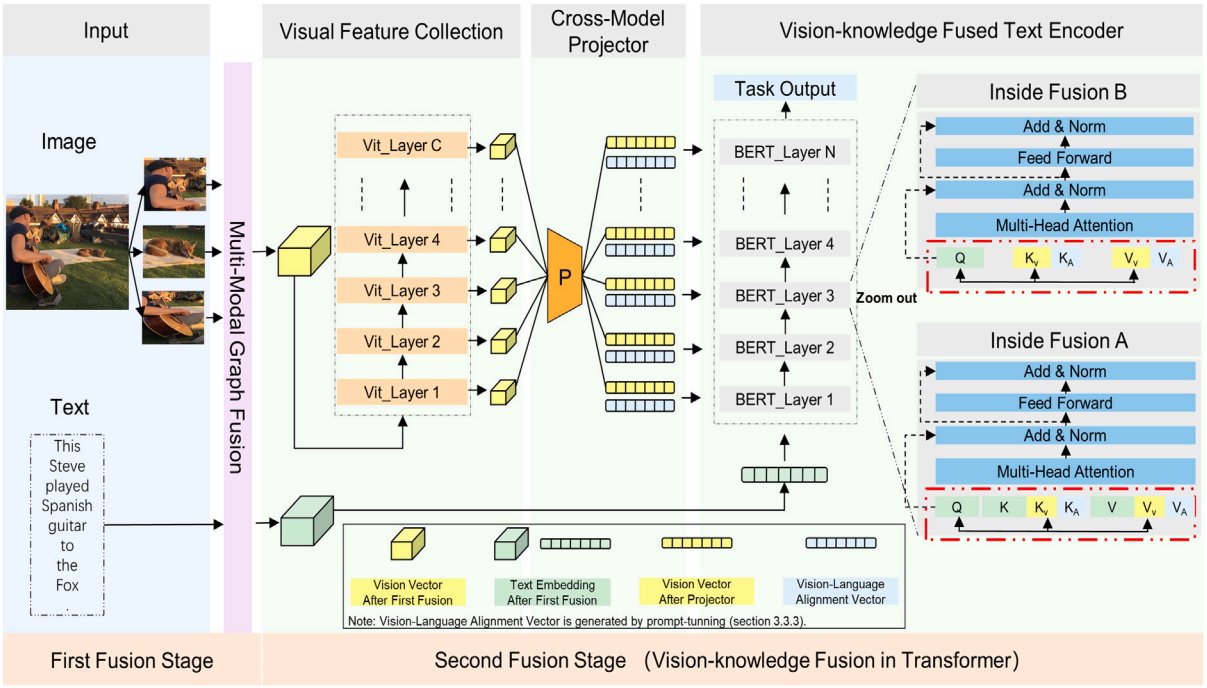

Fig. 2. Illustration of our TSVFN in multimodal relation extraction.

Lee, 2019; Su et al., 2020; Tan & Bansal, 2019) have applied the BERT to multimodal representation. This leads to our motivation to extract visual knowledge from a multimodal pre-trained model, which can unleash the power of the pre-trained model to more fine-grained integration of text and other modalities in multimodal tasks. Another trend in multimodal pre-trained is contrastive learning. CLIP is the most typical constrained pretraining model (Radford et al., 2021). It uses a visual transformer (Vit) (Dosovitskiy et al., 2021) or ResNet (He et al., 2016) as an image encoder, a transformer model as a text encoder, and trains both encoders jointly with contrast loss (Carse, Carey, & McKenna, 2021). CLIP can achieve excellent performance in image-text retrieval. In this paper, we also use the image encoder in the CLIP pre-trained model to extract image features that incorporate textual information.

3. Method

In this section, we formulate the MRE task and illustrate the overall architecture and the details of our framework.

3.1. Research objective

The objective of our paper is to address three problems of multimodal relation extraction: (1) The multimodal relational extraction dataset contains a lot of graphically unrelated information, and the image modality needs to be fused more precisely with the entities in the text modality. (2) In the text-based pre-training model, the knowledge of image modality needs to be effectively fused into it. (3) How to combine the respective advantages of graph data and sequence data to fully fuse and align the information of both modalities.

Task Formulation: Given a sentence S and its image V, the aim of MRE is to classify relations of the entity pairs E from S. We formulate the MRE task as a classification problem. Let $S = (s_1, s_2, \ldots, s_n)$ denote input words, $E = (e_{t1}, e_{h1}), (e_{t2}, e_{h2}), \ldots, (e_{tm}, e_{hm})$, and $y = (y1, \ldots, ym)$ be the corresponding label, where y_m in Y and Y is the pre-defined label set of MRE dataset.

Overall Architecture: As shown in Fig. 2, our framework have four main components: (1) Multimodal (MM) Graph Fusion: the first stage of fusion; (2) Visual Feature Collection: employ vision transformer (use image feature of the first stage fusion's output) (3) Vision-knowledge Fused Text Encoder: the second stage of fusion and we adopt two approaches (inside fusion A and B) to cooperate vision knowledge. (4) Classifier: utilize softmax as the relation classification's output.

3.2. Multimodal graph fusion (first fusion)

In order to be able to fully learn the feature representation between different modalities, i.e., most of the sentences in MRE grounded in the visual objects and their relationships, we construct multimodal graphs based on images and text and utilize the graph encoder GCN on the graphs. The first stage of fusion is crucial in the two-stage multimodal fusion approach, which deals with

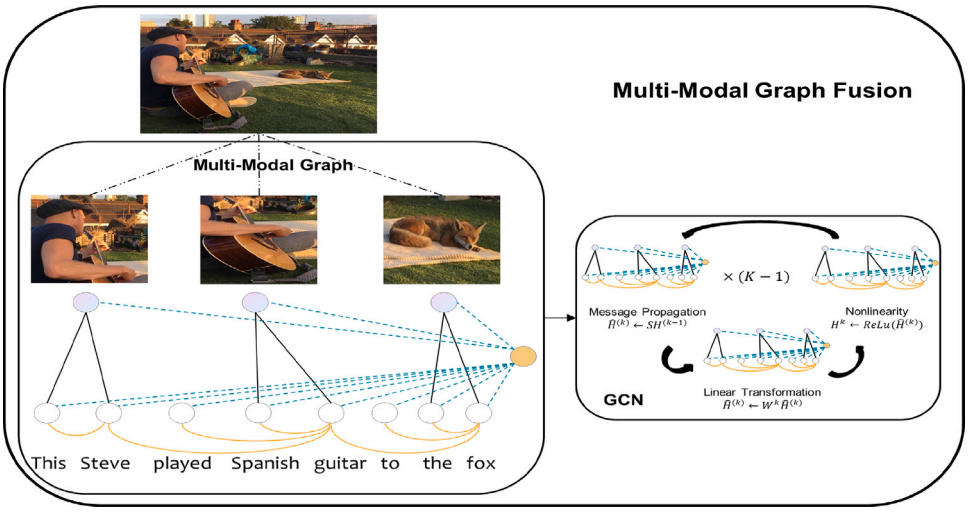

Fig. 3. The workflow of our proposed Multimodal Graph Fusion.

the degree of interaction of the low-level information between two modalities. As described in the introduction section, most prior work does not use a two-stage multi-modal fusion approach, which is susceptible to certain limitations. In contrast, our work uses GCN to first learn shallow representations of multimodalities. With a suitable adjacency matrix, GCN are able to perform parallel computations in different dimensions and are able to learn long-term sequence information by identifying and connecting relevant interaction steps.

3.2.1. Multi-modal graph construction

The multimodal relation extraction task needs to use the visual information provided in the data as an essential basis. In the first stage of fusion, we take a visual object detection tool to obtain the local visual information to better combine the feature information between the image and the relevant entities in the sentence. There are three kinds of nodes: textual word nodes, visual object nodes, and global nodes in the node-set V. Different approaches are conducted for the 3 kinds of nodes. (1) Textual word nodes: we segment all words as text nodes in one sentence. For example, in Fig. 3, there are eight text nodes since the sentence is separated into eight words. (2) Visual Object nodes: We employ the NER and Parser tool to identify the noun phrases in the target sentence. We then adopt the visual tool to detect objects based on the noun phrases to get visual object nodes. For example, in Fig. 3, we can identify three noun phrases, “This Steve”, “Spanish guitar” and “the fox”, so that three visual nodes are built. (3) Global node: We build one global node for each multimodal graph that can fully exploit the information between textual nodes and visual nodes. We design four kinds of edges in the multimodal graph: text node–visual object node, text node–text node, text node–global node (all separated word nodes connect to global node), and visual object node–global node (all detected visual object nodes connect to global node). There is also a demonstration of the multimodal graph in Fig. 3. For text node v_{xi}, we set its initial state by BERT (Devlin et al., 2019). For visual object nodes v_{oj}, we get visual features from the fully-connected layer that follows the pooling layer of Faster-RCNN (Ren, He, Girshick, & Sun, 2017). For the global node, we set the random value of normal distribution as the initial state.

3.2.2. Graph encoder

After constructing the multimodal graph, we need to fuse and extract the feature information in the first stage. To balance the relationship between model complexity and task performance, we utilize graph neural networks (GCN) to handle the first stage of fusion, since GCNs have sufficiently powerful modeling capabilities. The feature representation of the nodes in the multimodal graph after the graph encoder is as follows.

$$\boldsymbol{h}^l_{v_{x1}}, \ldots, \boldsymbol{h}^l_{v_{xi}}, \boldsymbol{h}^l_{v_{o1}}, \ldots, \boldsymbol{h}^l_{v_{oj}}, \boldsymbol{h}^l_{v_g} = G(e_{v_{x1}}, \ldots, e_{v_{xi}}, e_{v_{o1}}, \ldots, e_{v_{oj}}, e_{v_g}) \tag{1}$$

where $G()$ is the graph encoder and a graph convolution networks (GCN) (Kipf & Welling, 2017)is introduced for the strong representation ability. where $e_{v_{x1}}, \ldots, e_{v_{xi}}$ are the word representation of the input text, $e_{v_{o1}}, \ldots, e_{v_{oj}}$ means the visual representation of j visual objects, , e_{v_g} is the visual vector of the global image. $\boldsymbol{h}^l_{v_{x1}}, \ldots, \boldsymbol{h}^l_{v_{xi}}, \boldsymbol{h}^l_{v_{o1}}, \ldots, \boldsymbol{h}^l_{v_{oj}}, \boldsymbol{h}^l_{v_g}$ is the output of GCN encoder. Then we use a multi-layer perceptron with ReLU activation function to project these features onto the same space so the vector can input to the second fusion stage.

3.3. Vision-knowledge fusion in transformer (second fusion)

In the second stage of fusion, there are three major modules: (1) Visual Feature Collection: The role of this module is to convert our chosen visual knowledge (detected local visual objects and the overall image) to visual features. (2) Multi-scaled Cross-Model Projector: this module can process the feature information vector of visual knowledge into specific dimensions. (3) Vision-knowledge Fused Text Encoder: this module is to merge visual knowledge into BERT so that BERT can learn multimodal knowledge.

3.3.1. Visual feature collection

In multimodal relation extraction, text still plays a dominant role, and the role of other modalities is almost always to assist in semantic enhancement. From among the visual knowledge we select several visual objects related to the entities of the sentence. The reason for choosing the complete image as part of visual knowledge is that we also take it to provide some contextual information as well as some more implicit abstract information and may alleviate the model training over-fitting. Therefore, we collect multiple visual cues for multimodal relation extraction, which consists of using the local objects as important information and the global image as a complement. We employ the vision transformer (Vit) (Dosovitskiy et al., 2021) as the backbone to get the visual features. After the first stage of fusion operation, we get the visual object feature and the global image as the input of Vit. We also conduct the BERT in the vision-knowledge fusion, so the similarity of structures can be more convenient and effective for multimodal information fusion. Typically, we process the given visual features with Vit and generate a new list of visual hierarchical feature maps $F_1, F_2, F_3, \ldots, F_c$ from different blocks.

$$[\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots, \boldsymbol{F}_c]_{o1}, \ldots, [\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots, \boldsymbol{F}_c]_{oj}, [\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots, \boldsymbol{F}_c]_g = \mathbf{Vit}([\boldsymbol{h}_{v_{o1}}], \ldots [\boldsymbol{h}_{v_{oj}}], [\boldsymbol{h}_{v_g}]) \tag{2}$$

where $[\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots, \boldsymbol{F}_c]_{o1}, \ldots, [\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots, \boldsymbol{F}_c]_{oj}$ are the hierarchical feature maps from j visual objects. $[\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots, \boldsymbol{F}_c]_g$ are the hierarchical feature maps from the global image.

3.3.2. Multi-scaled cross-model projector

How assign visual features to the text encoder is a critical step, as random assignment may introduce too much noise into the model. To address this issue, we propose the Multi-scaled Cross-Model Projector, inspired by Multi-scaled Attention mechanism (Chen, Ling, & Zhu, 2018). We accomplish this operation through a dynamic attention mechanism to dynamically compute multiple normalized vectors to determine the degree of integration of visual features at each layer. The specific equation is as follows:

$$\begin{aligned} \alpha_c^l &= \frac{exp(MLP(A(\boldsymbol{F}_c)))}{\sum_{c=1}^{C} exp(MLP(A(\boldsymbol{F}_c)))} \\ \boldsymbol{V}^l &= \sum_{c=1}^{C} \alpha_c^l \boldsymbol{F}_c \end{aligned} \tag{3}$$

where we first change the visual features $\boldsymbol{F}_c$ to appropriate dimensions from the cth vit layer by an average pooling operation A. Then we reduce the feature dimension with the MLP layer. Formally, the vision-knowledge features $\boldsymbol{V}^l$ corresponding to BERT's lth layer is obtained by the following concatenation operation,

$$\boldsymbol{V}_f^l = [\boldsymbol{V}_g^l; \boldsymbol{V}_{o_1}^l; \ldots; \boldsymbol{V}_{o_m}^l] \tag{4}$$

where $\boldsymbol{V}_g^l$ presents the full image feature, $\boldsymbol{V}_{o_m}^l$ presents the mth objects feature, $\boldsymbol{V}_f^l$ presents the visual knowledge feature inputted in lth layer of BERT.

3.3.3. Vision-knowledge fused text encoder

The combine of hierarchical feature (from low-level to high-level) of the visual knowledge and textual information provides essential semantic information. Such relationship is indispensable to associate the visual knowledge in the image with the entities and words mentioned in the sentence. The MRE task can be achieved through incorporating visual information into the text encoder. In order to incorporate visual knowledge into a text encoder, we design an attention layer fusion in two ways: 1. Inside Fusion A; 2. Inside Fusion B.

Vision-language Alignment Vector: In this section, we also design a special vision-language alignment vector (VLA vector) for visual knowledge fusion which are added to each layer. The intuition behind our design of the VLA vector is inspired by the idea of the soft prompt (Gu, Han, Liu, & Huang, 2022). The VLA vector align knowledge in two modalities using the extra vector. Specifically, the generation of the VL vectors is an independent process, we first initialize the VLA vector randomly in the BERT embedding layer from the parameter set obeying a standard normal distribution. Then we perform other downstream multimodal task (here we use VQAv2 (Goyal, Khot, Summers-Stay, Batra, & Parikh, 2017)). We finetune and get the VLA vector while fixing BERT and Vit. Fig. 4 shows the process that we get VL Vector. Since the vector's dimension after linear projection is the same as the dimension of the self-attention matrix, it can be directly concatenated with the self-attention matrix.

Inside Fusion A: Inspired by Xiang et al. (2022), we make corresponding improvements in each block of BERT. We use the hierarchical visual features and the vectors for learning visual and text alignment as the final visual vector. Then we concatenate

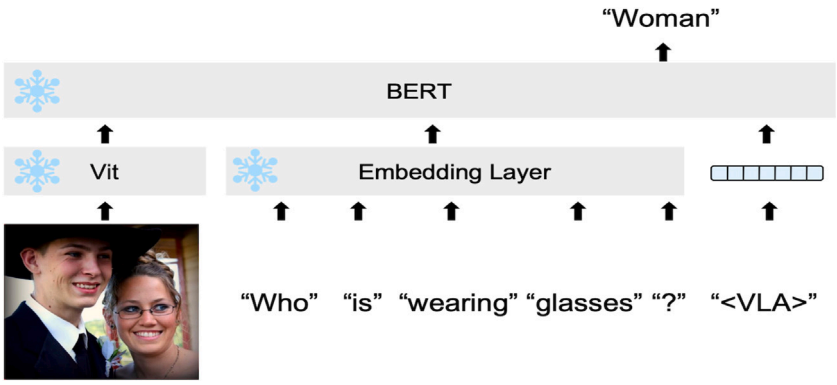

Fig. 4. The process of obtaining the VLA vector. We can obtain the VLA vector by introducing the prompt vector (which represents the alignment of image and text). Vit is to extract the representation of the image. For training, we use the VQAv2 dataset; BERT and Vit are fixed weights in training; the VLA vector is the only trainable parameter.

the visual prefix vectors into the computation process at each self-attention layer of BERT. For the text feature sequence obtained in the previous step, the query (Q)/key (K)/value (V) vector of each layer is first computed.

$$\begin{aligned} \boldsymbol{P}_k^l &= W_k^1[\boldsymbol{V}_f^l; \boldsymbol{V}_{VLA}] \\ \boldsymbol{P}_v^l &= W_v^1[\boldsymbol{V}_f^l; \boldsymbol{V}_{VLA}] \end{aligned} \tag{5}$$

The obtained hierarchical visual information feature with the VLA vector is then multiplied by a set of linear transformations $W_k^1, W_v^1 \in \mathbb{R}^{d\times 2\times d}$ to be projected into the same embedding space of the textual representation in the self-attention module. In addition, $\boldsymbol{P}_k^l, \boldsymbol{P}_v^l \in \mathbb{R}^{hw(m+2)\times d}$ represents the visual prefix vector, $hw(m+2)$ denotes the length of the visual and VLA features, and m is the number of visual objects detected by the object detection tool. Formally, the new visual prefix based attention matrix $\boldsymbol{K}'$ and $\boldsymbol{V}'$ is computed as follows and we then use the new attention matrix to substitute for the original attention matrix.

$$\boldsymbol{K}' = [\boldsymbol{P}_k^l; \boldsymbol{K}]; \boldsymbol{V}' = [\boldsymbol{P}_v^l; \boldsymbol{V}] \tag{6}$$

Inside Fusion B: Inside Fusion B differs from Inside Fusion A in that we replace the key (K) directly in the self-attention layer after the linear transformation of the visual feature vector and the VLA feature vector. It is necessary to concatenate the value (V) with the original V. We believe that this approach allows the model to be better focused with the visual features. The formula is as follows:

$$\begin{aligned} \boldsymbol{P}_k^l &= W_k^2[\boldsymbol{V}_f^l; \boldsymbol{V}_{VLA}] \\ \boldsymbol{P}_v^l &= W_v^2[\boldsymbol{V}_f^l; \boldsymbol{V}_{VLA}] \\ \boldsymbol{K}' &= \boldsymbol{P}_k^l \\ \boldsymbol{V}' &= [\boldsymbol{P}_v^l; \boldsymbol{V}] \end{aligned} \tag{7}$$

where $W_k^2, W_v^2 \in \mathbb{R}^{d/m\times d}$, $\boldsymbol{K}', \boldsymbol{V}'$ denote the new key and value, respectively. We also use the new attention matrix to substitute for the original one.

3.4. Classifier

The softmax function is introduced as the classifier. The formula is as below:

$$p(r|X) = Softmax(W\boldsymbol{H}_{[CLS]}^L) \tag{8}$$

where parameters of W is trainable. $[CLS]$ head is utilized to get the probability distribution over the class set Y. We adopt cross-entropy loss for MRE task:

$$L_{mre} = -\sum_{i=1}^{M} log(p(r^i|FO(X^i, \boldsymbol{P}_k^l, \boldsymbol{P}_v^l))) \tag{9}$$

where X^i is ith sample in the dataset, $FO()$ represents fusion operation of vision and text feature in self-attention of BERT, M represents the number of samples in the dataset.

Table 1
The statistics of MNRE dataset.

Statistics	MNRE-F	MNRE-S
Number of Word	258k	51k
Number of Sentence	9,201	1828
Number of instance	15,485	2450
Number of Entity	30,970	6123
Number of Relation	23	23
Number of Image	9,201	1828

4. Results and discussion

4.1. Dataset

We validate our model on the multimodal relation extraction dataset named MNRE collected on Twitter with topics including music, sports, etc. MNRE contains 15,484 samples, 9,201 images, and 23 relation categories. We further divide the MNRE into two datasets: MNRE-F and MNRE-S. MNRE-F is composed of a training set of 12,247, a validation set of 1624, and a test set of 1614. MNRE-S is 20% of MNRE-F with a training set of 2450, and validation and test sets of 324, respectively. The detailed statistics of the datasets are shown in Table 1.

4.2. Experiment setting and evaluation metrics

We run each experiment 5 times on the random seed [1, 88, 7788, 6688, 9876] and report the averaged performance. Pytorch is utilized to conduct experiments with 1 Nvidia 3090 GPUs. The BERT based experiments use BERT-base-uncased model from huggingface (http://www.huggingface.co./). AdamW (Loshchilov & Hutter, 2019) optimizer is used with a linear warmup of learning rate over the first 10% of gradient updates to a maximum value and weight decay is set to 0.01. The number of image objects m is set to 3 and batch size as 8. We search for the learning rates from 1e-5 to 3e-5 (MNRE-F) and from 3e-5 to 5e-5 (MNRE-S). The train epoch is 36 and evaluation is made after the 18th epoch. We choose the model performing the best on the validation set and evaluate it on the test set. Following the previous works, accuracy, precision, recall and micro F1 scores are selected as the evaluation metrics.

4.3. Compared baselines

Two classes of baseline is chosen to compare with our model. The first category is the text-based model.

PCNN (Zeng et al., 2015): PCNN incorporates Piecewise Max Pooling and Multi-instance Learning, which is one of the most commonly used baselines in relation extraction.

MTB (Soares et al., 2019): MTB borrows the idea of distant supervision to set up a pre-training task of comparing different relational vectors to be more focused on learning relation features.

PURE (Zhong & Chen, 2021): PURE adopts a simple and easy relation extraction method to improve the performance by adding the typed marker to entities.

The second category is a multimodal relation extraction model based on MNRE.

UMT (Yu, Jiang, Yang, & Xia, 2020): UMT uses a multimodal interaction module to obtain both image-aware word representations and word-aware visual representations to guide the final predictions.

UMGF (Zhang et al., 2021): UMGF employs image object detectors as interaction units of image modalities and graph neural networks to implement multimodal interaction.

BERT+SG (Zheng, Feng, et al., 2021): This model leverages BERT to obtain text features and predict multimodal relations after concatenating them with visual features extracted from the scene graph extraction tool.

BERT+SG+Att (Zheng, Feng, et al., 2021): The model considers the semantic similarity between visual graphs (scene graphs) and text content and uses an attention mechanism to calculate the semantic similarity.

MEGA (Zheng, Feng, et al., 2021): MEGA uses the structural alignment and semantic alignment between visual scene graph and text-dependent graph structure to find the most relevant visual relations with textual relations, thus improving the multimodal relation extraction performance.

VisualBERT (Li et al., 2019): VisualBERT implicitly aligns the input text and regions associated with the input image with self-attention in each block of the Transformer layer.

HVPNeT (Xiang et al., 2022): HVPNeT adds visual information to the process of attention computation at each layer of transformers and improves the performance of multimodal relation extraction.

Table 2
Accuracy (Acc), precision (Pre) and micro F1 scores (F1) of PCNN, MTB, PURE, UMT, UMGF, BERT+SG, BERT+SG+Att, VisualBERT, MEGA, HVPNeT and our model on MNRE-F.

Models	MNRE-F			
	Acc	Pre	Recall	F1
Text Models				
PCNN	72.67	62.85	49.69	55.49
MTB	72.73	64.46	57.81	60.86
PURE	73.05	64.50	58.19	61.22
Text+Image Models				
VisualBERT	69.22	57.15	59.48	58.30
UMT	74.49	62.93	63.88	63.46
UMGF	76.31	64.38	66.23	65.29
BERT+SG	74.09	62.95	62.65	62.80
BERT+SG+Att	74.59	60.97	66.56	63.64
MEGA	76.15	64.51	68.44	66.41
HVPNeT	90.95	83.64	80.78	81.85
TSVFN (inside fusion A)	92.67	**85.16**	**82.07**	**83.02**
TSVFN (inside fusion B)	**92.71**	84.97	81.99	82.98

4.4. Experimental result and discussion

TSVFN of this paper is compared with the existing benchmark methods. Table 2 shows the experimental results, from which we can observe that:

1. The visual information can enhance the representation ability of the text modality model. We can see that the performance of text-based unimodal models is lower than that of text+image models (except VisualBERT). For example, MEGA is a multimodal model built on the BERT model, and PURE is a pure text modality model based on the pre-trained model (BERT). We can find that MEGA outperforms PURE by 3.2%, 0.1%, 10.25%, and 5.19% in accuracy, precision, recall, and F1 score, respectively. This boost proves that visual information is crucial for the multimodal relation extraction task.
2. Pre-trained models can be supplemented with external knowledge by additional design (entity boundaries, visual knowledge, etc.). However, the selection of pre-trained models needs to consider their original training tasks. In the comparison models, we mainly use pre-trained models (PURE, VisualBERT, TSVFN). PU-RE achieves better results than MTB and PCNN by adding special markers to the starts and ends of entities to enrich information. VisualBERT's performance on MRE is mediocre. The main reason is that its pre-training objective function is biased towards the image task, so some critical textual information is lost. TSVFN proposed in this paper adds image knowledge to the pre-trained model with a text-side task through the inside approaches. The experimental results also show the effectiveness of TSVFN.
3. From Table 2, we can see very clearly that the model proposed in this paper achieves the best results. There are three main reasons: (1) A graph and text fusion operation is used in two-stage, which can better and more accurately extract the information associated with the text and images, making the performance of the multimodal relation extraction task improved. (2) this paper uses Vit as the image feature extractor, which has the same structure as BERT, so that it can achieve better results than other image feature extractors in the second stage of visual information fusion. (3) A VLA vector is designed, which is equivalent to a more fine-grained correction of the alignment information between text and image in the model.

In addition, Fig. 5 shows the several models' performance in F1 scores when only using a small amount of data. We randomly select 20% of the samples from the original dataset to construct MRE-S. In this experiment, we chose MTB, PURE, UMGF, BERT+S (BERT+SG), BERT+SA (BERT+SG+Att), MEGA, and HVPNeT with our models proposed in the paper (TSVFN-IA and TSVFN-IB) for comparison experiments. We can observe that: **1.** the models that use visual knowledge (UMGF, BERT+S, BERT+SA, MEGA, HVPNeT) are better than the models that use only textual features (MTB, PURE) in the case of small amount of data. This shows that fusing visual features can improve the performance of the MRE task in low-resource situations. **2.** TSVFN outperforms the other baselines. This further demonstrates the effectiveness of our proposed method with different amounts of data resources.

4.5. Ablation study

We conduct seven sets of ablation experiments on MNRE-F to verify the effectiveness of multimodal graph fusion, multi-scaled cross-model projector, multimodal fusion strategy, VL aligned vector, image encoder, method's structure and different layers of image encoder, respectively.

4.5.1. Analysis of the effectiveness of the multimodal graph fusion

We conduct an ablation study of MNRE-F to verify the effectiveness of each part in Multimodal Graph Fusion including the dependency parse tree nodes in the Multimodal Graph, the visual object nodes, the global nodes, and the entire module. Table 3

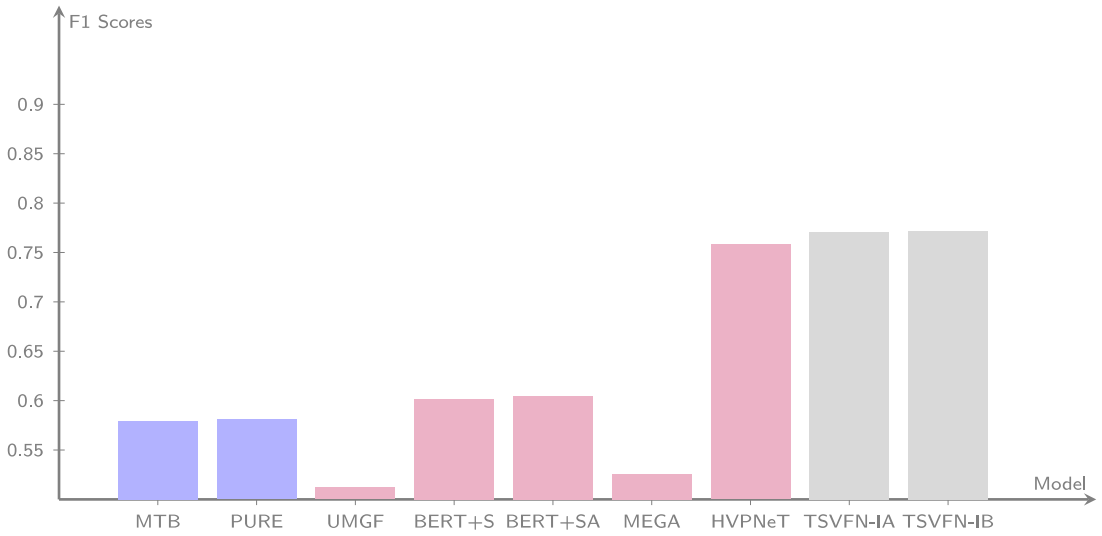

Fig. 5. Micro F1 scores (F1) of MTB, PURE, UMGF, BERT+S (BERT+SG), BERT+SA (BERT+SG+Att), MEGA, HVPNeT and our model (TSVFN-IA, TSVFN-IB) on MRE-S.

Table 3
Quantitative results of our method on the effectiveness of the multimodal graph fusion.

Model	F1
TSVFN-IA	83.02
w/o DPT node	81.65 (**−1.37**)
w/o Global node	81.93 (**−1.09**)
w/o Object node	80.83 (**−2.19**)
w/o All	81.62 (**−1.4**)

shows the F1 scores for multimodal relation extraction. The results are clear that the model performance degrades when any part is removed or replaced.

First, we use the text graph with two–two connections instead of the DPT graph, and get a 1.37% drop in performance. Then we remove the visual object node and the global node, where the performance is downgraded by 1.09% and 2.19%, respectively. It can be concluded from the results that the visual object node is able to fully fuse the information between the visual object and text. If only the whole image node is left, it will introduce too much useless information and cause dramatic degradation of the model performance. The removal of the global node results in a loss of 2.19% performance, suggesting that the global node effectively incorporates different node information from the multimodal graph. The impact of the DPT node, the visual object node, and the global node is different. However, they all contribute to the model's effectiveness, while their combination yields excellent performance. The multimodal graph fusion module also significantly contributes to the performance. Without this component, the F1 score of the model decreases by 1.4%. The multimodal graph fusion module helps to learn better the fine-grained correspondence between image objects and words in sentences, which facilitates the final relation classification.

To further demonstrate the usefulness of our multimodal graph fusion in two-stage fusion, we visualize the attention parameters separately for the model without the multimodal graph fusion and for the model with the two-stage fusion. We choose one of three visual targets and entity words from the text. We then visualize the attention weights with the visual target as the x axis (here the image features are flattened into 4 patches) and the entity words as the y axis. The visual attention weights are computed as the average sum of softmax weights of multiple heads. The attention weights are visualized using heat map. High values are in blue and low values are in white. The results are presented in Figs. 6 and 7. As can be seen in Fig. 6, the attention weights of model without the first stage of fusion are smaller than with the two stages of fusion ("Claudio"-image and "Marchisio"-image). In Fig. 7, it can be observed that the attention parameters of the image-text in the model with two-stage fusion are smaller than the model without multimodal graph fusion. This shows that the two-stage fusion method can not only effectively enhance the semantic effect of relevant images, but also mitigate the semantic effect of irrelevant images. This demonstrates the robustness and usefulness of our multimodal graph fusion approach.

4.5.2. Analysis of the effectiveness of the multi-scaled cross-model projector

We perform ablation experiments to verify the effectiveness of the Multi-scaled Cross-Model Projector. As shown in Table 4, three methods are introduced for the analysis. "Concatenate+Linear" specifically concatenates each layer of image features in Vit and

Fig. 6. A visualization example of relevant image&text. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

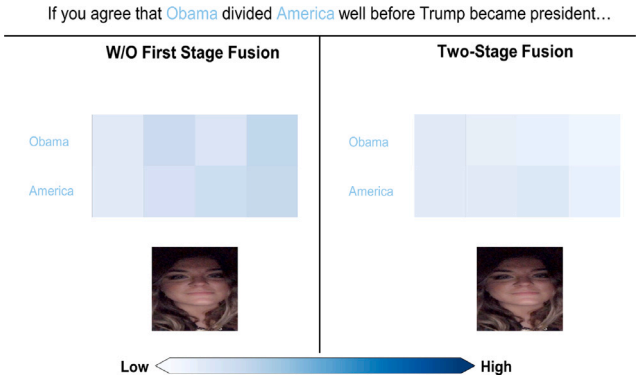

Fig. 7. A visualization example of irrelevant image&text. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Quantitative results of our method on the effectiveness of the vision-language projector method.

Model	F1
TSVFN-IA	83.02
Concatenate+Linear	79.21 (**−3.81**)
Dynamic Gated Aggregation	81.95 (**−1.07**)
Random Matching	79.61 (**−3.41**)

then passes through a linear layer. "Dynamic Gated Aggregation" borrows the method from Xiang et al. (2022) directly. "Random Matching" records the random matching strategy and fixes the matching relationship to the model. We can see that the Multi-scaled Cross-Model Projector method achieves the best results, where the F1 scores are higher than "Concatenate+Linear", "Dynamic Gated Aggregation", and "Random Matching" by 3.81%, 1.07%, and 3.41%, respectively. Therefore, Multi-scaled Cross-Model Projector is a crucial component of TSVFN to achieve superior performance. Specifically, the dynamic projector can adaptively assign different image features to the text model, thus combining visual knowledge with multi-scale information. Since the text encoder and the image encoder are based on transformer architecture, the proposed approach can better match different levels of text semantics with visual semantics.

4.5.3. Analysis of the effectiveness of the fusion strategy

In this subsection, we design an outside fusion method to compare with the inside fusion strategy in our paper. We propose vision prompt vectors to fuse visual knowledge. We concatenate vision prompt vector, VLA vector and text embedding from BERT as the input. TSVFN-O means we use outside fusion in TSVFN. The motivations for introducing the outside fusion like prompt learning are threefold. First, prompt learning has shown sound performance in large text-based unimodal models. Second, the Vit's structure which adopts for producing visual prompts has some structural similarity to text encoders, which can facilitate possible knowledge fusion between the two modalities. Third, although the naive approach of learning a single shared prompt for a task enables the fusion of visual information, it still leads to serious forgetting problems, limited by the expressive power of the single prompt.

Table 5
Quantitative results of our method on the effectiveness of the image encoder.

Models	Acc	F1
Inside Fusion A	92.67	**83.02**
Inside Fusion B	**92.71**	82.98
TSVFN-O	89.97 (**−2.74**)	80.21 (**−2.81**)

Table 6
Quantitative results of our method on the effectiveness of the vision-language alignment vector.

Model	F1
TSVFN-O	80.21
w/o VLA vector	80.02 (**−0.21**)
TSVFN-IA	83.02
w/o VLA vector	81.35 (**−1.67**)
TSVFN-IB	82.98
w/o VLA vector	82.08 (**−0.9**)

Table 7
Quantitative results of our method on the effectiveness of the image encoder.

Model	F1
TSVFN-IA	
Vit_16	**83.02**
Resnet50	82.01 (**−1.01**)
DenseNet50	81.98 (**−1.04**)
Googlenet	80.45 (**−2.57**)

We can observe from Table 5 that the accuracy and F1 scores of our model are reduced if outside fusion is used, which indicates that outside fusion is less capable than inside fusion. Inside fusion allows for more accurate fusion of visual information into the language model, which plays a crucial role in the performance improvement of the model. We can also conclude that internal fusion can exploit visual knowledge more fully in interactive multilevel computation.

4.5.4. Analysis of the effectiveness of the vision-language aligned vector

In this section, we design experiments on the effectiveness of the added Vision-Language Aligned Vector. The VLA vector is removed from TSFVN-O, TSFVN-IA, and TSFVN-IB. It can be observed from Table 6 that the F1 scores of the models are reduced if the VL Aligned vector is removed, which indicates that the VLA Vector plays an essential role in improving the model performance. The experiments show that the VL Aligned Vector learns the alignment information of the two modalities and can fuse the information in a more fine-grained manner. Further, the VLA Vector is able to decouple Vision and Language modalities to some extent.

4.5.5. Analysis of the effectiveness of the image encoder

We also test different image encoders. We select ResNet, DesNet, and GoogleNet for comparison with the Vision Transformer in this paper. From Table 7, we can observe that using Vit as the image feature extractor significantly improves the performance of the multimodal relation extraction. Besides, taking Vit as an image encoder enables the efficient fusion of visual features into the same structure as BERT. This demonstrates the method's superiority in terms of the same structure of the text and image encoder. Moreover, Vit can extract different feature information between different levels of the image, emphasizing the differences between different levels, which facilitates cross-modal knowledge fusion. If other image feature extractors are used, their performance decreases by 1.01% (Resnet50), 1.04% (DenseNet50) and 2.57% (GoogleNet), respectively. Consequently, Vit helps to subsequently learn the intrinsic relationship between visual knowledge and the words in the sentence, that boosts the final multimodal fusion.

4.5.6. Analysis of the effectiveness of our method's structure

Although our model structure is validated in a single-variable controlled ablation experiment, we also design experiments that control both variables simultaneously. The variables we chose for this experiment are: 1. whether or not to include first-stage fusion; 2. Multi-scaled Cross-Model Projector or Dynamic Gated Aggregation; 3. Outside Fusion or Inside Fusion A; and 4. whether or not to utilize the VLA vector. Table 8 shows that we have combined ten model structures using TSVFN with the conditions described above. From Table 8, we can find that our proposed multimodal model outperforms other compared models, which indicates that each part of our method's structure is necessary and can benefit the multimodal relation extraction in our framework. In the experiments controlling for two variables, in addition to our model, W/O First Fusion + Dynamic Gated Aggregation is the best performer, which is also consistent with the main contribution of this paper, namely the effectiveness of the inside multimodal fusion. We

Table 8
Quantitative results of our method on the effectiveness of our method's structure.

Model	F1
TSVFN-IA	83.02
W/O First Fusion + Dynamic Gated Aggregation	81.56 ↓
W/O First Fusion + Outside Fusion	78.84 ↓
W/O First Fusion + W/O VLA Vector	81.17 ↓
Dynamic Gated Aggregation + Outside Fusion	80.11 ↓
Dynamic Gated Aggregation + W/O VLA Vector	81.37 ↓
Outside Fusion + W/O VLA Vector	80.02 ↓
W/O First Fusion + Dynamic Gated Aggregation + Outside Fusion	77.13 ↓
W/O First Fusion + Dynamic Gated Aggregation + W/O VLA Vector	80.25 ↓
W/O First Fusion + Outside Fusion + W/O VLA Vector	77.09 ↓
Dynamic Gated Aggregation + Outside Fusion + W/O VLA Vector	79.97 ↓

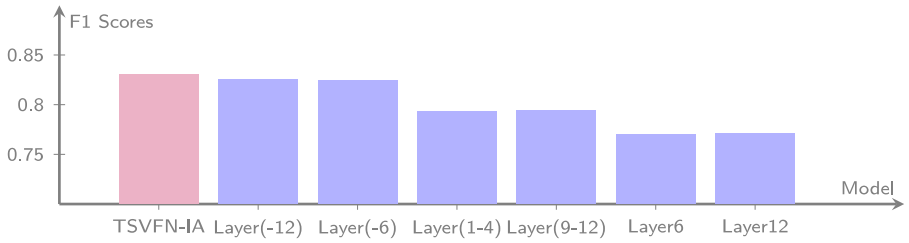

Fig. 8. Quantitative results of our method on the effectiveness of using different layers of image encoder.

can also observe that W/O First Fusion + Outside Fusion is the model with the worst results, primarily because neither the first stage multimodal graph fusion nor the second stage inside fusion is used. This model is not able to extract and align the image-text information efficiently and accurately. "W/O First Fusion + Outside Fusion + W/O VLA Vector" model achieves the worst performance when controlling for three variables. This model does not use the first stage of multimodal graph fusion and does not use either the inside fusion or the VLA vector in the second stage fusion.This structure does not allow the model to completely fuse information from the text and image modalities or to learn precise information about the alignment of the text-image information. This result also further illustrates that the structure of the method in this paper has performance advantages. The experiments described above verify that the structure of the method proposed in this paper is reasonable and efficient, and that it can improve the performance of the multi-modal relation extraction task.

4.5.7. Analysis of the effectiveness of using different layers of image encoder

To analyze the effect of using different layers of image encoder in our framework, we conduct a series of ablation studies with certain layers chosen and report the F1 scores for evaluation. Seven models are selected for comparison, including the model TSVFN-IA using all layers of the image encoder, the model Layer (−12) using all layers except layer 12, the model Layer (−6) using all layers except layer 6, the model Layer (1–4) using layers 1–4, the model Layer (5–8) using layers 5–8. the model Layer (9–12) using layers 9–12, the model Layer6 using layer 6 and the model Layer12 using the last layer of the image encoder. Fig. 8 shows the result of the this ablation study. As Fig. 8 shows, the corresponding model without the use of full layers does not perform well in terms of F1 score. The model results with 1 layer show considerable decay compared to the mode with full layers, which points to a hierarchical (from low-level to high-level) feature of visual knowledge may provide essential and different semantic information in the second stage of fusion. As shown in Fig. 8, if we remove any of one or 4 layers, the value of F1 decreases. These results indicate that the visual knowledge is distributed throughout each layer of the image encoder, and for the multi-modal fusion step, the visual knowledge embedded in each layer is critical. Our full model achieves the best trade-off between the number of image encoder layers and model performance.

4.6. Parameter sensitivity

In this section, we experimentally verify the impact of several vital hyperparameters and the results are shown in Fig. 9. First, as we emphasize before, the number of visual objects is set to ensure quality and adequacy when performing modal fusion, thus it plays a crucial role in our model. Intuitively, choosing too many visual objects means introducing too many irrelevant visual features, which makes the model relatively fragile. Too many choices of visual objects do not necessarily lead to easier recognition of the correspondence between textual and visual modalities. Meanwhile, too few visual objects may bring some opposite effects. Therefore, we conduct experiments with this hyperparameter on the MNRE-F dataset. We find that the number of visual objects should be a slightly lower value that takes into account adequate visual knowledge and a moderate introduction of noise. Then, we verify the impact of different layers of GCN in Multimodal Graph Fusion. The optimal results are achieved for the number of layers

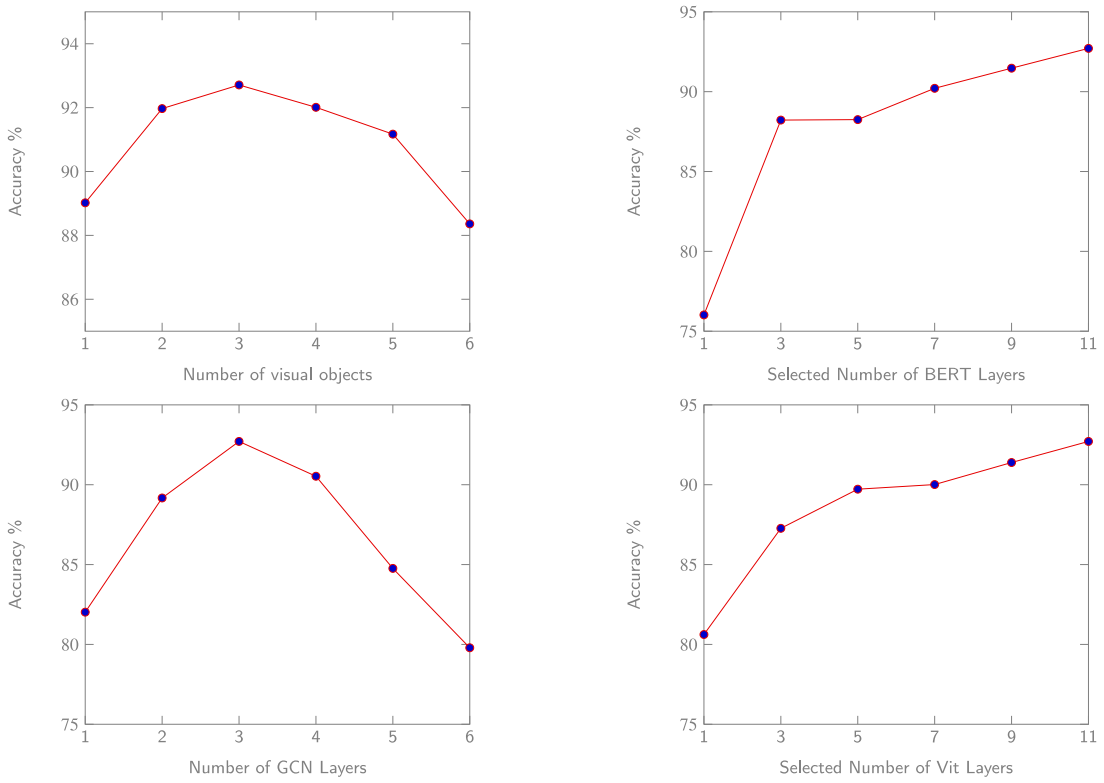

Fig. 9. Several hyper-parameters experiments.

(3) in this paper. Moreover, too few layers lead to insufficient information transfer between nodes in the graph, while too many layers lead to over-smoothing of the nodes and loss of uniqueness in the graph. Finally, we verify the effect of the number of layers selected by the image encoder (Vit) and the text encoder (BERT), where we note them in the order of layer1–layer11. For example, if we select one layer, layer1 is selected; if we select three layers, layer1, layer2, and layer3 are selected. The model's accuracy increases as the number of selected layers increases. The results verify our hypothesis that an adequate number of layers facilitates the hierarchical fusion of multimodal models. It is also noteworthy that choosing too few layers leads to poorer performance. The reason is that visual and textual knowledge is not sufficiently fused, and the correlation between textual and visual features may be difficult to be learned.

4.7. Case study

In this section, we first conduct several visual examples to discuss the effectiveness of the visual-assisted text model and the advantages of the proposed model in this paper. Five models are selected for comparison, PURE representing the text modality model, MEGA, HVPNet, TSVFN (w/o F) and TSVFN representing the multimodal model. It should be explained that TSVFN (w/o F) represents the model discarding the first stage fusion and directly performing the second stage fusion. Some qualitative results are presented in Fig. 10, where four cases are chosen for a specific illustration. Each of these examples includes the target sentence, the head entity, the tail entity, the visual picture (containing the selected visual objects), the gold relation between the entities, and the prediction results of the compared model. We observe that purely unimodal-based relation prediction models require visual information to make incorrect predictions in the examples. For example, in case 1, which is a peer relationship between two soccer players, the pure text-based model fails to predict the correspondence, while the models incorporating visual information make correct predictions. Meanwhile, in case 2, our model and HVPNet make the correct prediction. Although the MEGA model includes visual information, it outputs incorrect relation results due to the model's deficiency in fine-grained alignment between different modalities and fusion. In cases 3 and 4, HVPNeT makes the incorrect prediction like MEGA. The HVPNeT model fuses different levels of feature information of images. However, our model chooses the image and text encoder with the same structure so that the relation between visual and text can be aligned more accurately and efficiently in multimodal feature fusion. In case 5, there is a weak association between the semantics of image and text. Among all the compared models, only our method successfully predicts the results, which also indicates that our method adequately incorporates multi-level semantic information of images. In Case 6, in the absence of any relation between image and text semantics, in this paper, the model successfully makes correct predictions with unimodal PURE, which considers text alone. Cases 5 and 6 illustrate that our model can achieve good performance not only

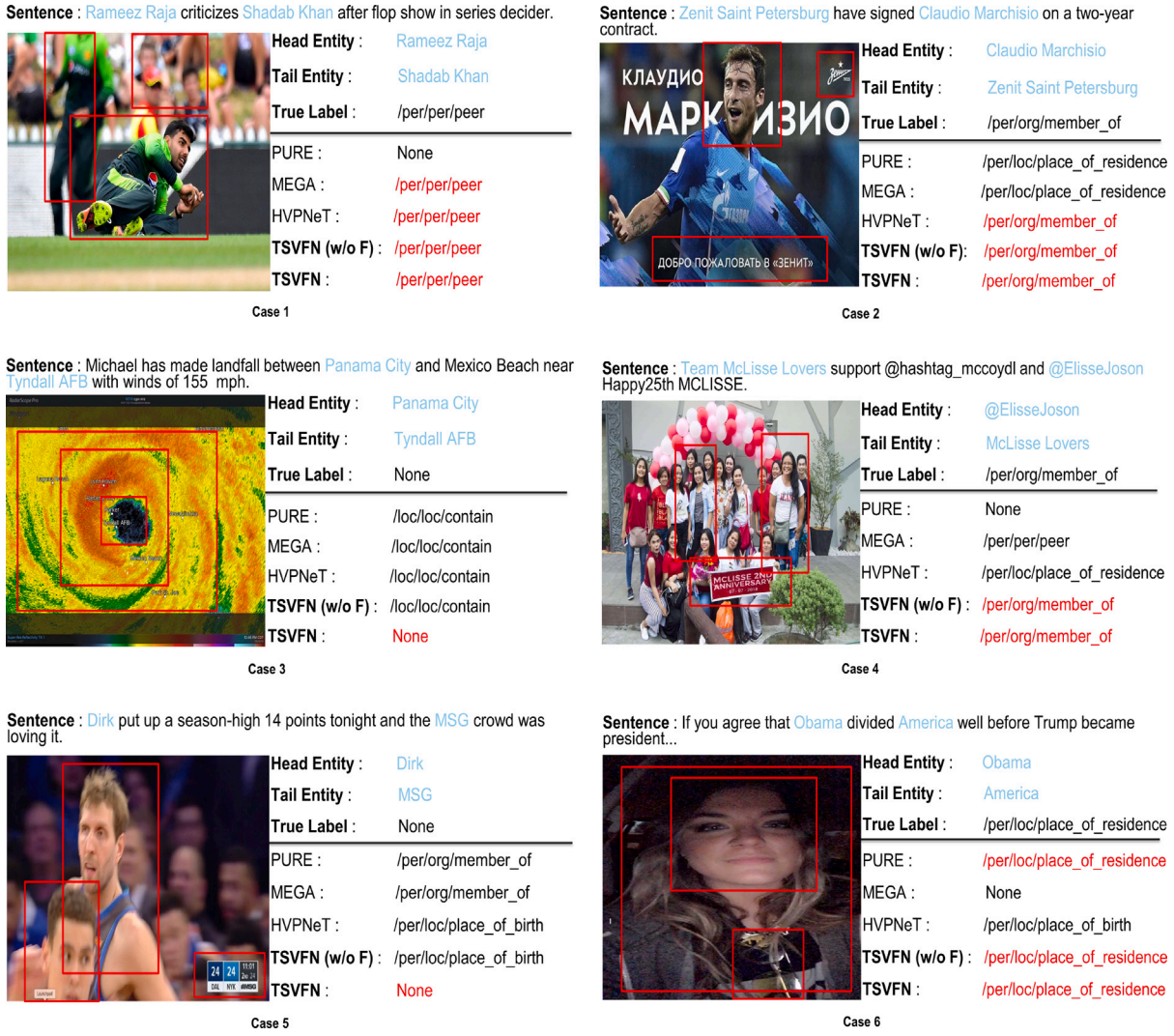

Fig. 10. Several cases predicted by four models with text and image.

when image and text are related, but also when image and text are unrelated or weakly related, in addition, the model can extract multimodal information accurately and efficiently, and hence can achieve good performance. Lastly, from the above six cases, it can be observed that TSVFN makes six predictions correctly, the TSVFN (w/o F) evaluates two of them incorrectly. This also shows that each stage of the two-stage multimodal fusion method proposed in this paper has its irreplaceable role. Intuitively, our TSVFN model can find the association between visual and textual information in a more fine-grained way and extract the relations through the accurate fusion of visual and textual information.

5. Conclusion and future work

A two-stage multimodal fusion model TSVFN for multimodal relation extraction is proposed in this paper. The first stage constructs a multimodal graph and achieves multimodal information fusion through information transfer on the graph. In the second stage, the model takes visual information from the visual pre-trained model as an aid. Three ways are designed in the text pre-trained model for information fusion of text and visual modalities. We conduct many comparative experiments to verify the effectiveness of the proposed model which achieves SOTA on the multimodal relation extraction dataset MNRE. In addition, we conduct some ablation studies and parameter sensitivity experiments to demonstrate the validity of each part of TSVFN model and the correctness of hyper-parameters selection. In the future, we will explore how to better fuse the feature information of different modalities among the transformer-based multimodal pre-trained models and how to combine graph neural networks to improve the performance.

CRediT authorship contribution statement

Qihui Zhao: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Data curation. **Tianhan Gao:** Supervision, Writing – review & editing. **Nan Guo:** Funding acquisition, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by National Natural Science Foundation of China (grant no. 52130403) and the Fundamental Research Funds for the Central Universities, China (grant no. N2017003).

References

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E. R., Jr., & Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. In M. Fox, & D. Poole (Eds.), *Proceedings of the twenty-fourth AAAI conference on artificial intelligence*. AAAI Press, URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI10/paper/view/1879.

Carse, J., Carey, F. A., & McKenna, S. J. (2021). Unsupervised representation learning from pathology images with multi-directional contrastive predictive coding. In *18th IEEE international symposium on biomedical imaging* (pp. 1254–1258). IEEE, http://dx.doi.org/10.1109/ISBI48211.2021.9434140.

Chen, Q., Ling, Z., & Zhu, X. (2018). Enhancing sentence embedding with generalized pooling. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th international conference on computational linguistics* (pp. 1815–1826). Association for Computational Linguistics, URL: https://aclanthology.org/C18-1154/.

Chen, X., Zhang, N., Li, L., Yao, Y., Deng, S., Tan, C., et al. (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* (pp. 1607–1618). Seattle, United States: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.findings-naacl.121, URL: https://aclanthology.org/2022.findings-naacl.121.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/n19-1423.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16 × 16 words: Transformers for image recognition at scale. In *9th international conference on learning representations*. OpenReview.net, URL: https://openreview.net/forum?id=YicbFdNTTy.

drissiya El-allaly, E., Sarrouti, M., En-Nahnahi, N., & Ouatik El Alaoui, S. (2021). MTTLADE: A multi-task transfer learning-based method for adverse drug events extraction. *Information Processing & Management*, *58*(3), Article 102473. http://dx.doi.org/10.1016/j.ipm.2020.102473, URL: https://www.sciencedirect.com/science/article/pii/S0306457320309626.

Gori, G. M. M., & Scarselli, F. (2005). A new model for learning in graph domains. In *International joint conference on neural networks* (pp. 729–734). IEEE Computer Society.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE conference on computer vision and pattern recognition* (pp. 6325–6334). IEEE Computer Society, http://dx.doi.org/10.1109/CVPR.2017.670.

Gu, Y., Han, X., Liu, Z., & Huang, M. (2022). PPT: pre-trained prompt tuning for few-shot learning. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the Association for Computational Linguistics (vol. 1: Long papers)* (pp. 8410–8423). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.acl-long.576.

Han, X., Gao, T., Lin, Y., Peng, H., Yang, Y., Xiao, C., et al. (2020). More data, more relations, more context and more openness: A review and outlook for relation extraction. In *Proceedings of the 1st conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th international joint conference on natural language processing* (pp. 745–758). Suzhou, China: Association for Computational Linguistics, URL: https://aclanthology.org/2020.aacl-main.75/.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition* (pp. 770–778). IEEE Computer Society, http://dx.doi.org/10.1109/CVPR.2016.90.

Jia, Z., Lin, Y., Wang, J., Feng, Z., Xie, X., & Chen, C. (2021). HetEmotionNet: Two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition. In H. T. Shen, Y. Zhuang, J. R. Smith, Y. Yang, P. Cesar, F. Metze, & et al. (Eds.), *MM '21: ACM multimedia conference, Virtual Event* (pp. 1047–1056). ACM, http://dx.doi.org/10.1145/3474085.3475583.

Khademi, M. (2020). Multimodal neural graph memory networks for visual question answering. In D. Jurafsky, J. Chai, N. Schluter, & J. R. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 7177–7188). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.643.

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *5th international conference on learning representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference track proceedings*. OpenReview.net, URL: https://openreview.net/forum?id=SJU4ayYgl.

Li, Y., Tarlow, D., Brockschmidt, M., & Zemel, R. S. (2016). Gated graph sequence neural networks. In Y. Bengio, & Y. LeCun (Eds.), *4th international conference on learning representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, Conference track proceedings*. URL: http://arxiv.org/abs/1511.05493.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C., & Chang, K. (2019). VisualBERT: A simple and performant baseline for vision and language. arXiv:1908.03557.

Liu, X., Tan, K., & Dong, S. (2021). Multi-granularity sequential neural network for document-level biomedical relation extraction. *Information Processing & Management*, *58*(6), Article 102718. http://dx.doi.org/10.1016/j.ipm.2021.102718, URL: https://www.sciencedirect.com/science/article/pii/S0306457321002028.

Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., et al. (2020). K-BERT: Enabling language representation with knowledge graph. In *The thirty-fourth AAAI conference on artificial intelligence, AAAI 2020, the thirty-second innovative applications of artificial intelligence conference, IAAI 2020, the tenth AAAI symposium on educational advances in artificial intelligence* (pp. 2901–2908). AAAI Press, URL: https://ojs.aaai.org/index.php/AAAI/article/view/5681.

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *7th international conference on learning representations*. OpenReview.net, URL: https://openreview.net/forum?id=Bkg6RiCqY7.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32: Annual conference on neural information processing systems 2019* (pp. 13–23). URL: https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html.

Micheli, A. (2009). Neural network for graphs: A contextual constructive approach. *IEEE Transactions on Neural Networks*, *20*(3), 498–511. http://dx.doi.org/10.1109/TNN.2008.2010350.

Peters, M. E., Neumann, M., Logan, R. L., IV, Schwartz, R., Joshi, V., Singh, S., et al. (2019). Knowledge enhanced contextual word representations. In K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 43–54). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D19-1005.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. In M. Meila, T. Zhang (Eds.), *Proceedings of machine learning research: vol. 139*, *Proceedings of the 38th international conference on machine learning* (pp. 8748–8763). PMLR, URL: http://proceedings.mlr.press/v139/radford21a.html.

Ren, S., He, K., Girshick, R. B., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149. http://dx.doi.org/10.1109/TPAMI.2016.2577031.

Soares, L. B., FitzGerald, N., Ling, J., & Kwiatkowski, T. (2019). Matching the blanks: Distributional similarity for relation learning. In A. Korhonen, D. R. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th conference of the Association for Computational Linguistics* (pp. 2895–2905). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/p19-1279.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., et al. (2020). VL-BERT: Pre-training of generic visual-linguistic representations. In *Eighth international conference on learning representations*. URL: https://www.microsoft.com/en-us/research/publication/vl-bert-pre-training-of-generic-visual-linguistic-representations/.

Tan, H., & Bansal, M. (2019). LXMERT: learning cross-modality encoder representations from transformers. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 5099–5110). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D19-1514.

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In *6th international conference on learning representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018, Conference track proceedings*. OpenReview.net, URL: https://openreview.net/forum?id=rJXMpikCZ.

Wang, M. (2008). A re-examination of dependency path kernels for relation extraction. In *Third international joint conference on natural language processing* (pp. 841–846). The Association for Computer Linguistics, URL: https://aclanthology.org/I08-2119/.

Wang, H., Lu, G., Yin, J., & Qin, K. (2021). Relation extraction: A brief survey on deep neural network based methods. In Y. Li, & H. Nishi (Eds.), *ICSIM 2021: 2021 the 4th international conference on software engineering and information management* (pp. 220–228). ACM, http://dx.doi.org/10.1145/3451471.3451506.

Wei, Y., Wang, X., Nie, L., He, X., Hong, R., & Chua, T. (2019). MMGCN: multi-modal graph convolution network for personalized recommendation of micro-video. In L. Amsaleg, B. Huet, M. A. Larson, G. Gravier, H. Hung, C. Ngo, & et al. (Eds.), *Proceedings of the 27th ACM international conference on multimedia* (pp. 1437–1445). ACM, http://dx.doi.org/10.1145/3343031.3351034.

Wen, W., Liu, Y., Ouyang, C., Lin, Q., & Chung, T. (2021). Enhanced prototypical network for few-shot relation extraction. *Information Processing & Management*, *58*(4), Article 102596. http://dx.doi.org/10.1016/j.ipm.2021.102596, URL: https://www.sciencedirect.com/science/article/pii/S0306457321000959.

Wu, S., & He, Y. (2019). Enriching pre-trained language model with entity information for relation classification. In W. Zhu, D. Tao, X. Cheng, P. Cui, E. A. Rundensteiner, D. Carmel, & et al. (Eds.), *Proceedings of the 28th ACM international conference on information and knowledge management* (pp. 2361–2364). ACM, http://dx.doi.org/10.1145/3357384.3358119.

Yamada, I., Asai, A., Shindo, H., Takeda, H., & Matsumoto, Y. (2020). LUKE: Deep contextualized entity representations with entity-aware self-attention. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 6442–6454). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.523.

Yin, Y., Meng, F., Su, J., Zhou, C., Yang, Z., Zhou, J., et al. (2020). A novel graph-based multi-modal fusion encoder for neural machine translation. In D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 3025–3035). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.273.

Yu, J., Jiang, J., Yang, L., & Xia, R. (2020). Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In D. Jurafsky, J. Chai, N. Schluter, & J. R. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 3342–3352). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.306.

Zaporojets, K., Deleu, J., Develder, C., & Demeester, T. (2021). DWIE: An entity-centric dataset for multi-task document-level information extraction. *Information Processing & Management*, *58*(4), Article 102563. http://dx.doi.org/10.1016/j.ipm.2021.102563, URL: https://www.sciencedirect.com/science/article/pii/S0306457321000662.

Zeng, D., Liu, K., Chen, Y., & Zhao, J. (2015). Distant supervision for relation extraction via piecewise convolutional neural networks. In L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, & Y. Marton (Eds.), *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1753–1762). The Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/d15-1203.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). ERNIE: Enhanced language representation with informative entities. In A. Korhonen, D. R. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th conference of the Association for Computational Linguistics, volume 1: Long papers* (pp. 1441–1451). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/p19-1139.

Zhang, D., Wei, S., Li, S., Wu, H., Zhu, Q., & Zhou, G. (2021). Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *Thirty-fifth AAAI conference on artificial intelligence, AAAI 2021, Thirty-third conference on innovative applications of artificial intelligence, IAAI 2021, the eleventh symposium on educational advances in artificial intelligence* (pp. 14347–14355). AAAI Press, URL: https://ojs.aaai.org/index.php/AAAI/article/view/17687.

Zheng, C., Feng, J., Fu, Z., Cai, Y., Li, Q., & Wang, T. (2021). Multimodal relation extraction with efficient graph alignment. In H. T. Shen, Y. Zhuang, J. R. Smith, Y. Yang, P. Cesar, F. Metze, & et al. (Eds.), *MM '21: ACM multimedia conference* (pp. 5298–5306). ACM, http://dx.doi.org/10.1145/3474085.3476968.

Zheng, C., Wu, Z., Feng, J., Fu, Z., & Cai, Y. (2021). 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* (pp. 1–6). IEEE, http://dx.doi.org/10.1109/ICME51207.2021.9428274.

Zhong, Z., & Chen, D. (2021). A frustratingly easy approach for entity and relation extraction. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, & et al. (Eds.), *Proceedings of the 2021 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies* (pp. 50–61). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.naacl-main.5.

Zhou, Y., Geng, X., Shen, T., Long, G., & Jiang, D. (2022). EventBERT: A pre-trained model for event correlation reasoning. In F. Laforest, R. Troncy, E. Simperl, D. Agarwal, A. Gionis, & et al. (Eds.), *WWW '22: The ACM web conference 2022, Virtual Event* (pp. 850–859). ACM, http://dx.doi.org/10.1145/3485447.3511928.