Few-Shot Relation Extraction with Hybrid Visual Evidence

Jiaying Gong, Hoda Eldardiry
Virginia Tech
Blacksburg, VA, USA
{gjiaying, hdardiry}@vt.edu

Abstract

The goal of few-shot relation extraction is to predict relations between name entities in a sentence when only a few labeled instances are available for training. Existing few-shot relation extraction methods focus on uni-modal information such as text only. This reduces performance when there is no clear contexts between the name entities described in text. We propose a multi-modal few-shot relation extraction model (MFS-HVE) that leverages both textual and visual semantic information to learn a multi-modal representation jointly. The MFS-HVE includes semantic feature extractors and multi-modal fusion components. The MFS-HVE semantic feature extractors are developed to extract both textual and visual features. The visual features include global image features and local object features within the image. The MFS-HVE multi-modal fusion unit integrates information from various modalities using image-guided attention, object-guided attention, and hybrid feature attention to fully capture the semantic interaction between visual regions of images and relevant texts. Extensive experiments conducted on two public datasets demonstrate that semantic visual information significantly improves performance of few-shot relation prediction.

Keywords: few-shot learning, relation extraction, multi-modal fusion

1. Introduction

Relation extraction aims to predict the relation between two name entities in a sentence. To alleviate the reliance on high-quality annotated data, few-shot learning has drawn more attention, requiring only a few labeled instances for training to adapt to new tasks. Existing few-shot relation extraction methods can be roughly divided into two categories. One category involves methods only using plain text data, without any auxiliary information. For example, meta-learning models prototypical networks (Gao et al., 2019), siamese neural networks (Yuan et al., 2017) are trained with only a few examples for each class to extract relations. The other category introduces external data sources such as relation information (Liu et al., 2022a,b), concepts of entities (Yang et al., 2021), side information (Gong and Eldardiry, 2021), external datasets (Geng et al., 2020), and graphs (Qu et al., 2020), to compensate the limited information in the above methods, to enhance the performance in few-shot relation extraction.

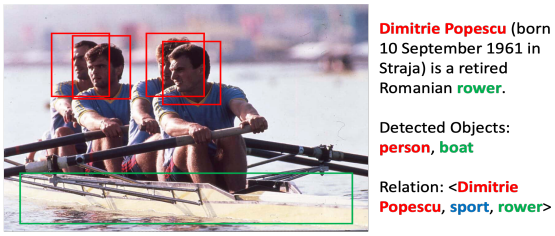

Figure 1: An example of multi-modal relation extraction based on visual information.

However, these methods mainly explore single-modality text-based data and may suffer a significant performance decline when texts lack contexts. For example, in Figure 1, given two name entities 'Dimitrie Popescu' and 'rower', it is difficult for text-based models to detect the relation 'sport' without other supplementary information because the word 'sport' or other similar words does not appear in the text. As a result, uni-modal models will incorrectly extract the relation 'winner' or 'candidate' of the two name entities according to the short given textual sentence. Even models using external information such as knowledge graphs or related words with similar meanings still can not correctly extract the relation due to the limited information in short given textual sentences.

Therefore, we question that *Can visual information be a good external source to supplement the missing contexts in textual sentences for few-shot relation extraction?* In the above case, we can easily classify the relation into 'winner' from the guidance of an image showing that a person is holding a trophy. Utilizing visual information to support contextual information for texts involves multi-modal learning. However, fusing information from different modalities is also a challenging task. First, simply concatenating textual and visual features without considering semantic information may even have a negative impact on the performance as shown in Sec. 4.4. For example, in Figure 1, the multiple people's faces in the background are noise for the image with the relation 'sport'. Second, existing multi-modal models (Sec. 2.2) mainly focus on fusing global visual features with text without considering the semantic information of visual objects in images. In Figure 1, visual objects such as

LREC-COLING 2024, pages 7232–7247

'person' and 'boat' contain essential information to the relation 'sport'.

To address these challenges, we propose a **M**ultimodal **F**ew-**S**hot model based on **H**ybrid **V**isual **E**vidence (MFS-HVE) for relation extraction. We first generate the representations through the textual feature extractor in Sec. 3.2.1 and the visual feature extractor in Sec. 3.2.2. We consider the visual representations from both the local perspective in low resolution (Sec. 3.3.2) and the global perspective in high resolution (Sec. 3.3.1). To be more specific, a local feature vector is the embedding of the objects detected from the image, and a global feature vector is the embedding of the whole image. Because local features only focus on objects, global features can overcome the problem of sparsity with more information; however, they may probably contain noise (irrelevant information). We integrate both local features and global features to solve the problem of sparsity and noise.

Secondly, inspired by the cross-modal attention mechanism (Yu et al., 2021), we propose a multi-modal fusion unit including image-guided attention, object-guided attention, and hybrid feature attention to integrating semantic information from different modalities at both global and local levels. From the global perspective, image-guided attention based on the scaled dot-product attention (Vaswani et al., 2017) combines global feature vectors from the image with texts to capture the semantic interaction between visual regions of images and texts. From the local perspective, object-guided attention fuses objects detected from the image with relevant name entities from the textual sentences. Then the hybrid feature attention fuses all textual and visual information, including global image features and local object features. The hybrid feature attention generates a weight vector, multiplied by the multi-modal representations.

Finally, we concatenate text features, image-guided features, and object-guided features through a cross-modality encoder to generate the final multi-modal representations. Each relation representation is calculated based on the prototypical networks (Snell et al., 2017). Next, based on the prototypical networks (Snell et al., 2017), we compute the mean value of all multi-modal support vectors as the prototype to represent each relation. Because of the hierarchical structure of the detected objects and name entities discussed in Sec. 3.3.3, hyperbolic distance is calculated between multi-modal query representations and prototypes to predict the relation. We conduct extensive experiments on two public datasets MNRE (Zheng et al., 2021a) and FewRel (Han et al., 2018) to evaluate whether semantic visual information can supplement the missing contexts in textual sentences for few-shot relation extraction. FewRel is a uni-modal dataset containing only text, we crawl the image automatically by icrawler [1] for each instance to provide visual information, which can facilitate future research on multi-modal few-shot relation extraction. Details are introduced in Sec. 4.1.By comparing MFS-HVE with some state-of-the-art uni-modal few-shot relation extraction models and some multi-modal fusion methods with the same feature extractors, we show that, in general, models with multi-modal information perform better than the text-only models. We also conduct ablation studies and parameter sensitivity studies to learn the impact of each attention and function. The contributions of this paper can be summarized as:

- We propose the first approach (MFS-HVE) for multi-modal few-shot relation extraction. Existing models for few-shot relation extraction only focus on a single data modality.
- MFS-HVE combines information from different modalities through image-guided attention, object-guided attention, and hybrid feature attention to integrating semantic visual information and textual information.
- We conduct extensive experiments on two public datasets. The experimental results show that introducing visual information can supplement the missing contexts in textual sentences for the few-shot relation extraction task.

2. Related Work

2.1. Few-shot Relation Extraction

Recent studies of few-shot relation extraction focused on metric-based representative methods. For example, the prototypical network learns a prototype for each relation via instance embeddings (Gao et al., 2019; Ye and Ling, 2019; Baldini Soares et al., 2019; Hui et al., 2020). Siamese neural network learns the metric of relational similarities between pairs of instances (Yuan et al., 2017; Gao et al., 2020). Additional data sources are also used to help improve the performance in few-shot learning. Meta information such as relation information (Dong et al., 2020; Liu et al., 2022a,b; Zhenzhen et al., 2022; Dou et al., 2022; Li and Qian, 2022; Zhang and Lu, 2022), concepts of entities (Wang et al., 2020b; Yang et al., 2021), additional auxiliary information (Gong and Eldardiry, 2021), knowledge from cross domains (Geng et al., 2020), data augmentation (Qin and Joty, 2022; Gong and Eldardiry, 2023), and global graphs of all relations (Qu et al., 2020) are considered as prior information to establish connections between instance-based information and conceptual

[1] https://icrawler.readthedocs.io/en/latest/

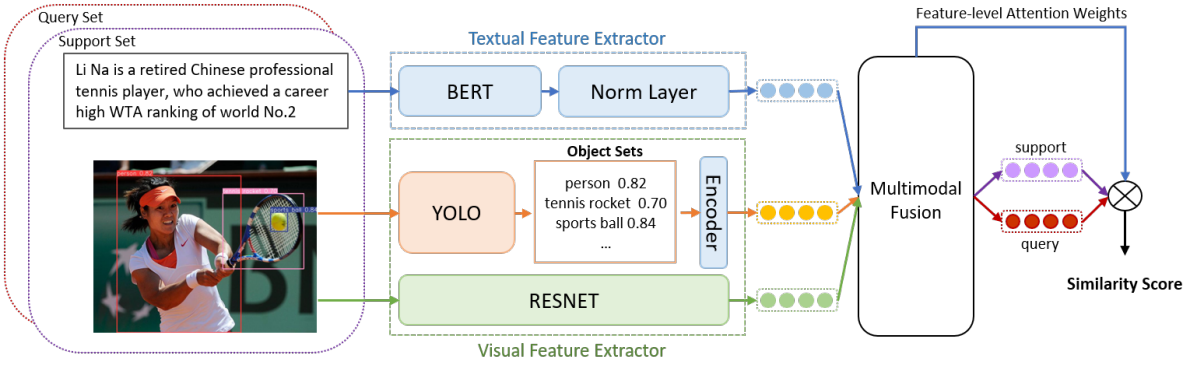

Figure 2: The overview of MFS-HVE. Details of multi-modal fusion is introduced in Sec. 3.3 and Figure 3

semantic-based information. However, the above studies only explore uni-modal text data. Different from these studies, we propose utilizing different data modalities, including both textual information and visual information, to supplement the missing semantics in texts.

2.2. Few-Shot Multi-Modal Fusion

Few-shot multi-modal fusion extracts relevant information from different modalities and integrates information collaboratively. MNRE is the first dataset developed for multimodal relation extraction (Zheng et al., 2021c). Existing few-shot multi-modal fusion has been studied in the areas of visual question answering (Tsimpoukelli et al., 2021; Najdenkoska et al., 2023; Jiang et al., 2023), image caption (Alayrac et al., 2022; Moor et al., 2023), action recognition (Wanyan et al., 2023; Ni et al., 2022), sentiment analysis (Yang et al., 2022), and so on. Studies have demonstrated that the performance of these tasks can be improved by fusing information from different modalities in few-shot learning (Lin et al., 2023). Inspired by these works, we consider fusing visual information for few-shot relation extraction to provide the missing context in texts. The only work on few-shot relation extraction focuses on social relation extraction, in which relations describe connections only between people (Wan et al., 2021a). Besides, the dataset in (Wan et al., 2021a) is not in English and includes a limited number of classes, and is therefore not sufficient to conduct 10-way-K-shot learning experiments. Considering the above limitations, we focus on few-shot general relation extraction that is conducted on (1) a re-splitted MNRE dataset to satisfy few-shot learning, and (2) a subset of the FewRel dataset, where we collected corresponding images, to explore relation extraction in FSL.

3. Methodology

Figure 2 shows the architecture of MFS-HVE for few-shot relation extraction. It consists of Semantic Feature Extractors and Multi-Modal Fusion. We describe these parts in detail below, starting with problem formulation.

3.1. Problem Definition

We follow the N-way-K-shot definition and settings of few-shot learning from (Gao et al., 2019) to conduct our experiments. The N-way-K-shot setting means N classes with K examples of each. Typically K is no more than 10. There is no overlap between the classes in training data and testing data. For the multi-modal few-shot relation extraction task, we tend to classify the relation between two name entities based on text and image inputs. Let the input dataset represented by a set of tuples $(x_i, h_i, t_i, y_i, r_i)$, where x_i is a sentence, h_i is a head entity, t_i is a tail entity, y_i is the corresponding image and r_i is the relation between h_i and t_i. Our goal is to train a few-shot learning model M to learn the representation function for the above tuples so that when randomly given support set with N relations and corresponding K tuples (NK tuples in total) as well as a query set with the same N relations and Q tuples, the model M can predict the relations in the query set base on the given support set. M is learned by minimizing the semantic distance between the input embedding from the support set and the embedding from the query set. At test time, we use a different set of relations and evaluate performance on the query set, given the support set.

3.2. Semantic Feature Extractor

Each instance contains a text message and a corresponding image for relation extraction. The text is the input for the textual feature extractor, and the image is the input for the visual feature extractor.

3.2.1. Textual Feature Extractor

For the textual feature extractor, we use a pre-trained language model BERT (Devlin et al., 2019) as the sentence encoder to generate the contextual representation. Two unique tokens [CLS] and [SEP] are appended to the first and last positions. The input text message is first tokenized into word pieces, and the positions of the name entities are marked by four special tokens [SEP] at the start and end of each entity mentioned in the relation statement of (Baldini Soares et al., 2019). Then output representation of the textual feature extractor r_t can be formulated as follows:

$$v_i = f_\phi(x_i, h, t) \tag{1}$$

$$r_t = tanh(W \cdot v_i + b) \tag{2}$$

where v_i is the output of sentence encoder, f_ϕ is BERT encoder, x_i is the input sentence, and h and t are head and tail entities, respectively. A fully-connected layer is added after BERT encoder, where $W \in \mathbb{R}^{256\times768}$ and $b \in \mathbb{R}^{256}$ are trainable.

3.2.2. Visual Feature Extractor

Object Feature Representation Object-level features are considered as the semantic information of the objects appearing in the image instead of the features of the whole image. For relation extraction tasks, a relation happens between the two name entities. Different from other multi-modal representation tasks, semantic information of the objects appearing in the images is of great importance. To extract objects from images, we utilize the pre-trained object detection model Yolo (Bochkovskiy et al., 2020) to recognize the objects in the images. We consider the top K frequent objects detected in the images to be the object labels because, in most cases, only the salient objects in the images are related to the name entities. Then, we transform the object labels into object embeddings to augment the semantic information of the two name entities and address the problem of semantic disparity of different modalities. The representation of object-level features r_o can be expressed as:

$$o = g_\phi(y_i) \tag{3}$$

$$r_o = f_\phi(o_0) \oplus \cdots \oplus f_\phi(o_k) \tag{4}$$

where g_ϕ denotes the object detection model, y_i is the input image, $\{o_0, o_1, \cdots, o_k\} \in o$, indicating the objects detected in the image, f_ϕ is the object embedding encoder and $\oplus$ denotes concatenation.

Image Feature Representation The global image features are extracted from ResNet18 (He et al., 2016). We use features from the last layer to produce the global vector. We then transform each feature vector into a new vector with the same dimension as the representation of the textual features using a single-layer perception. The representation of image-level features r_i is:

$$v_i = h_\phi(y_i) \tag{5}$$

$$r_i = tanh(W \cdot v_i + b) \tag{6}$$

where h_ϕ denotes the image encoder, y_i is the input image, $W \in \mathbb{R}^{256\times512}$ and $b \in \mathbb{R}^{256}$ are trainable weights and bias.

3.3. Multi-Modal Fusion

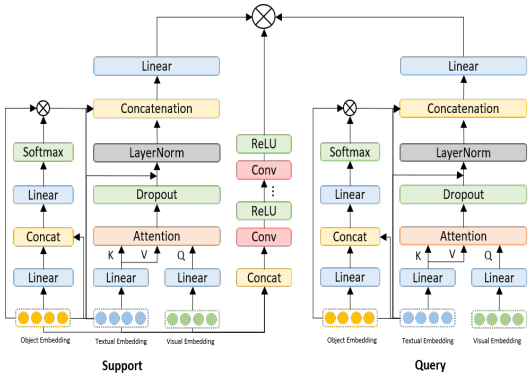

Figure 3: Detailed structure of multi-modal fusion.

The architecture of our proposed multi-modal fusion is shown in Figure 3, including image-guided attention, object-guided attention, and feature-level attention.

3.3.1. Image-Guided Attention

A cross-modal attention layer can provide a more sophisticated fusion between different modalities (Sun et al., 2020). Hence, we design a cross-attention layer module that combines the images and texts to capture the semantic interaction between visual regions of images and texts. As shown in Figure 3, the cross-modal attention layer is image-guided attention, which is calculated by combining Key-Value pairs from one modality with the Query from another modality. Specifically, the multi-modal representation is computed based on a modified version of the Scaled Dot-Product Attention (SA) (Vaswani et al., 2017). The attended feature for images $\hat{f}_i = GA(q_i, k_t, v_t)$ is obtained by reconstructing q_i using all samples in v_t for their

normalized cross-modal similarity to q_i. The image-guided attention unit is:

$$GA(q,k,v) = softmax((\frac{(W_Q q)(W_K k)^T}{\sqrt{d_k}})W_V v) \quad (7)$$

where W_Q, W_K, W_V are trainable query, key and value parameters and d_k is the dimension of key vectors. Note that queries are from visual images, while keys and values are from text.

For each instance, the textual representation $r_t \in \mathbb{R}^{n \times d_t}$ is obtained through Equation 2 and image representation is obtained through Equation 6. We first input the textual representation r_t and image representation r_i into fully connected layers, respectively. Then, the image-guided attention unit models the pairwise relationship between the paired sample $< r_t, r_i >$, where r_i guided the attention learning for r_t. The new image-guided feature vector related to r_t based on the cross-modal attention can be expressed as:

$$\hat{r_i} = LayerNorm(r_t + GA(r_i, r_t, r_t)) \quad (8)$$

where LayerNorm is used to stabilize the training.

3.3.2. Object-Guided Attention

Name entities in the textual sentence are always related to some objects detected from the input image. As shown in Figure 3, we propose an object-guided attention unit to fuse relevant words (name entities) and visual regions (objects). Given a textual feature r_t obtained from Equation 2 and a local object feature r_o obtained from Equation 4, we feed these features into a single neural network layer followed by a softmax function to generate the attention distribution over the objects:

$$v_{r_t} = tanh(W_{r_t} r_t \oplus (W_{r_o} r_o + b_{r_o})) \quad (9)$$

$$\alpha_{r_t} = softmax(W_{a_t} v_{r_t} + b_{a_t}) \quad (10)$$

where $r_t \in \mathbb{R}^d$, $r_o \in \mathbb{R}^d$, W_{r_o}, W_{r_t}, W_{a_t}, b_{r_t} and b_{a_t} are all trainable weights and bias. $\oplus$ denotes concatenation. Based on the attention distribution a_t, the new object vector $\hat{r_o}$ related to r_t is:

$$\hat{r_o} = \sum \alpha_{r_t} r_o \quad (11)$$

3.3.3. Hybrid Feature Attention

As shown in the middle of Figure 3, the hybrid feature attention fuses text information, global image-guided visual information, and local object-guided information, highlighting the important dimensions in the joint feature space to alleviate feature sparsity. For few-shot relation extraction, only a few instances in the support set are used for training so that the features extracted from the support set suffer from the problem of data sparsity. The feature-level attention generation block contains one concatenation layer, two or three 2D convolutional layers, and two or three activation functions, which can pay more attention to those more discriminative features when computing the space distance.

For space distance, studies show that hyperbolic spaces, where suitable curvatures match the characteristics of data, can lead to more generic embedding spaces (Gao et al., 2021; Liu et al., 2020). In the example shown in Figure 1, the detected object 'person' is the hypernym of the name entity 'Magic Johnson' in the text. Thus, we adopt hyperbolic distance with feature-level attention in our networks to preserve such hierarchical structure:

$$d(s_1, s_2) = \alpha_i \cdot cosh^{-1}(1 + 2\frac{\|s_1 - s_2\|^2}{(1 - \|s_1\|^2)(1 - \|s_2\|^2)}) \quad (12)$$

where α_i is the score vector for relation r_i calculated via the hybrid feature attention shown in Figure 3. By multiplying the hybrid feature attention weight by the support and query embeddings, we make the distance metrics better fit the given support sets and relations.

3.4. Model Training

The objective of training MFS-HVE is to minimize the distance between each instance embedding L_{multi} and the relation embedding $P_{multi}(S)$. A cross-modality encoder concatenates the three vectors: sentence embedding r_t, object-guided textual embedding $\hat{r_o}$, and image-guided textual embedding $\hat{r_i}$, to yield the multi-modal representation. Then, a fully connected layer is added to refine the multi-modal representation. The final multi-modal instance embedding L_{multi} is:

$$L_{multi} = tanh(W_{multi} \cdot (r_t \oplus \hat{r_o} \oplus \hat{r_i}) + b_{multi}) \quad (13)$$

where W_{multi} and b_{multi} are trainable.

Given support set S in the N way K shot setting, we compute a prototype for each of the N relations R in S based on the multi-modal representations L_{multi} of K tuples. To be more specific, the prototype representation $P_{multi}(S)$ for R is shown as:

$$P_{multi}(S) = \frac{1}{K} \sum_{i=1}^{K} L_{multi} \quad (14)$$

To predict the final relation among N ways, hyperbolic distance d as shown in Equation 12 is calculated between a query instance and each prototype $P_{multi}(S)$. Then, a softmax function is applied over the distance vector to generate a probability distribution on relations. More precisely, the

probabilities of the relations for a query instance q are computed as:

$$Pr(y = r_i|q) = \frac{exp(-d((L_{multi}), P_m(S)))}{\sum_{i=1}^{|R|} exp(-d((L_{multi}), P_i(S)))} \quad (15)$$

where $d(\cdot)$ is the hyperbolic distance.

4. Experiments

We conducted experiments with ablation studies, case studies, and parameter sensitivity experiments on two public datasets: MNRE and FewRel, to show that integrating semantic visual information with object-level and global feature-level attention mechanisms can help improve the performance.

4.1. Datasets

In our experiments, we evaluate our model [2] over two widely used datasets: MNRE (Zheng et al., 2021a), FewRel (Han et al., 2018), and a subset of FewRel, which includes only clean images. FewRel is a balanced dataset, and MNRE is an unbalanced dataset. The statistics of MNRE and FewRel datasets are shown in Table 1. We describe each dataset and dataset construction in detail in Appendix 8.1. For the MNRE dataset, we randomly re-split the original supervised MNRE dataset to ensure that there is no overlap of relations between the training set and testing set. For FewRel and FewRel$_{\text{small}}$ datasets, we follow the same training and validation set.

Table 1: The statistics of each dataset.

	#instances	#relations	avg. len.
MNRE	15,484	23	16.67
FewRel	56,000	80	24.95
FewRel$_{\text{small}}$	3,703	80	23.90

4.2. Baselines and Evaluation Metrics

We compare our model with six only text-based models: **Siamese** (Koch et al., 2015), **Proto** (Snell et al., 2017), **SNAIL** (Mishra et al., 2018), **GNN** (Satorras and Estrach, 2018), **MLMAN** (Ye and Ling, 2019), **MTB** (Baldini Soares et al., 2019) and eight text-based models with external information: **REGRAB** (Qu et al., 2020), **ZSLRC** (Gong and Eldardiry, 2021), **ConceptFERE** (Yang et al., 2021), **MapRE** (Dong et al., 2021), **HCPR** (Han et al., 2021), **GM_GEN** (Li and Qian, 2022), **FAEA** (Dou et al., 2022) and **SimpleFSRE** (Liu et al., 2022b). For multi-modal fusion baselines, we considering fusing the information from different modalities at different levels. The early fusion includes **Concatenation** (Wan et al., 2021a), and **Circulant Fusion** (Gong et al., 2023). The mid-level fusion includes **Deep Fusion** (Wang et al., 2020a), **Dual Co-Att** (Liu et al., 2021), and **Proto$_{\text{multimodal}}$** (Ni et al., 2022). We follow the same settings as (Qu et al., 2020) to run the experiments. The evaluation metric is the Accuracy (Acc.) of query instances.

4.3. Parameter Settings

Table 2: Parameter Settings

Parameter	Value
Textual Information Dimension d_t	512
Visual Information Dimension d_v	128
Object Information Dimension d_o	256
Batch Size	1
Initial Learning Rate α	0.1
Weight Decay	10^{-5}
Dropout	0.2
Sentence Max Length	128
Objects Number	2

For the hyperparameter and configuration of MFS-HVE, we implement MFS-HVE based on the PyTorch framework and optimize it with AdamW optimizer. We report the result based on a five-times run of the experiment. GPU of 16G memory is needed for the training process. The training time is around 5-6 hours depending on the computing resource. For the sentence encoder, we initialize the textual representation by pre-trained BERT (Devlin et al., 2019) and set the dimension size at 768. Then we follow (Baldini Soares et al., 2019) to combine the token encodings of the entity mentioned in the sentence. For the image encoder, we initialize the visual representation by pre-trained ResNet18 (He et al., 2016) and set the dimension size at 512. For the object encoder, we employ 50-dimensional GloVe (6B tokens, 400K vocabulary) (Pennington et al., 2014) for word embeddings of the objects detected from the image. Table 2 shows other parameters used in the experiment.

4.4. Results and Discussion

4.4.1. Main Results

The experiment results of few-shot learning on MNRE and FewRel$_{\text{small}}$ are shown in Table 3 with the average of five times run. Because some relations have less than ten instances in MNRE, it is impossible to run 5-shot experiments on MNRE.

[2] Code is available: https://github.com/gjiaying/MFS-HVE

Table 3: Results of Accuracy Comparison Among Models (%) on MNRE and $\text{FewRel}_{\text{small}}$ Datasets.

Modality	Model	MNRE		$\text{FewRel}_{\text{small}}$			
		5-Way 1-Shot	10-Way 1-Shot	5-Way 1-Shot	5-Way 5-Shot	10-Way 1-Shot	10-Way 5-Shot
Only Text	GNN (Satorras and Estrach, 2018)	29.08	22.53	46.38	70.45	28.74	62.07
	Snail (Mishra et al., 2018)	30.90	19.43	40.16	60.07	21.19	47.56
	Siamese (Koch et al., 2015)	36.08	26.50	62.74	73.92	42.17	65.05
	MLMAN (Ye and Ling, 2019)	35.08	29.06	63.47	74.47	61.86	72.58
	Proto_BERT (Snell et al., 2017)	49.75	33.57	75.64	84.64	64.17	75.27
	MTB (Baldini Soares et al., 2019)	46.02	32.35	76.38	86.27	65.27	73.81
Text+Others	ZSLRC (Gong and Eldardiry, 2021)	45.65	32.23	71.82	81.74	64.88	71.81
	ConceptFERE (Yang et al., 2021)	-	-	75.86	83.38	68.38	76.06
	REGRAB (Qu et al., 2020)	-	-	78.53	84.96	70.65	78.00
	HCRP (Han et al., 2021)	31.10	10.45	78.04	84.68	69.54	77.91
	MapRE (Dong et al., 2021)	51.92	35.20	79.44	85.60	70.71	78.84
	GM_GEN (Li and Qian, 2022)	52.58	35.82	60.04	73.74	42.22	59.23
	FAEA (Dou et al., 2022)	52.14	33.37	80.80	87.94	71.30	79.29
	SimpleFSRE (Liu et al., 2022b)	50.32	35.05	80.84	87.46	**71.67**	80.14
Text+Image	Concat (Wan et al., 2021a)	40.17	29.83	74.10	84.69	66.08	75.95
	CirculantFusion (Gong et al., 2023)	38.39	29.19	73.21	83.58	65.11	76.29
	DeepFusion (Wang et al., 2020a)	48.27	33.28	78.38	86.76	66.36	76.08
	$\text{Proto}_{\text{multimodal}}$ (Ni et al., 2022)	50.84	34.10	77.18	86.28	68.19	78.29
	Dual Co-Att (Liu et al., 2021)	52.52	35.62	77.60	87.24	68.69	78.54
	MFS-HVE	**54.88**	**36.62**	**81.32**	**89.65**	69.52	**80.55**

because we need five instances for the support set and the same number of instances for the query set. Thus, we only run 1-shot experiments on MNRE. FewRel is a public dataset with only textual information. We crawl the image relevant to each textual instance to construct a few-shot multi-modal dataset: $\text{FewRel}_{\text{small}}$, which is a subset of FewRel, including only clean images. Note that the baselines of multi-modal fusion works are implemented based on MTB (Baldini Soares et al., 2019) to have a fair comparison in few-shot relation extraction.

From Table 3, we observe that models integrating external information (labels, graphs, images, etc) perform much better than only text-based models. Models fusing semantic visual information can help improve the performance, but the performance highly depends on the fusion methods. Simply concatenating the visual information or fusing information at a coarse-grained level without considering semantic meanings such as circulant fusion may negatively impact the performance. This is probably because these methods treat all visual and textual information with equal importance (weights). However, only partial visual images contain relevant semantic meanings to the text. Directly using all information in the image may bring noise to the textual data. We further explore the robustness in Appendix 8.2. After considering image-guided textual information, object-guided textual information, and joint learning of text, image, and objects, our proposed model MFS-HVE significantly outperforms all state-of-the-art models on MNRE. More details about the performance of different attention layers of MFS-HVE are discussed in the ablation study in Section 4.4.2.

In summary, based on the experiment results on MNRE, FewRel, and $\text{FewRel}_{\text{small}}$, we have the following findings:

1. We find that models with multi-modal information perform better than text-based models in general.

2. Multi-modal models based on high-quality visual information are more robust than text-based models when the dataset size becomes smaller.

3. The performance of multi-modal models highly depends on the fusion methods. Simple concatenation or circulant multiplication of information from different modalities may probably have a negative impact.

4. For the relation extraction task, the local object information from the image is also very important because they are related to name entities in textual sentences and help reduce the noise of global image features.

4.4.2. Ablation Study

To illustrate the effectiveness of MFS-HVE and explore the role of each attention unit in MFS-HVE, we carry out the ablation study on the datasets only with clean and high-quality visual data (MNRE and $\text{FewRel}_{\text{small}}$) because the performance of fusion with different multi-modal information is unstable

Table 4: Ablation study over MFS-HVE components (%) on MNRE and $\text{FewRel}_{\text{small}}$ datasets.

Model Component	MNRE		$\text{FewRel}_{\text{small}}$			
	5-Way 1-Shot	10-Way 1-Shot	5-Way 1-Shot	5-Way 5-Shot	10-Way 1-Shot	10-Way 5-Shot
Only Text	49.39	31.95	76.66	85.82	63.54	76.73
Image Attention	50.43	32.40	78.37	86.75	66.28	77.18
Object Attention	50.57	33.63	78.85	86.24	66.96	77.96
Image&Object Attention	52.26	35.38	80.50	88.72	69.49	79.17
MFS-HVE	**54.88**	**36.62**	**81.32**	**89.65**	**69.52**	**80.55**

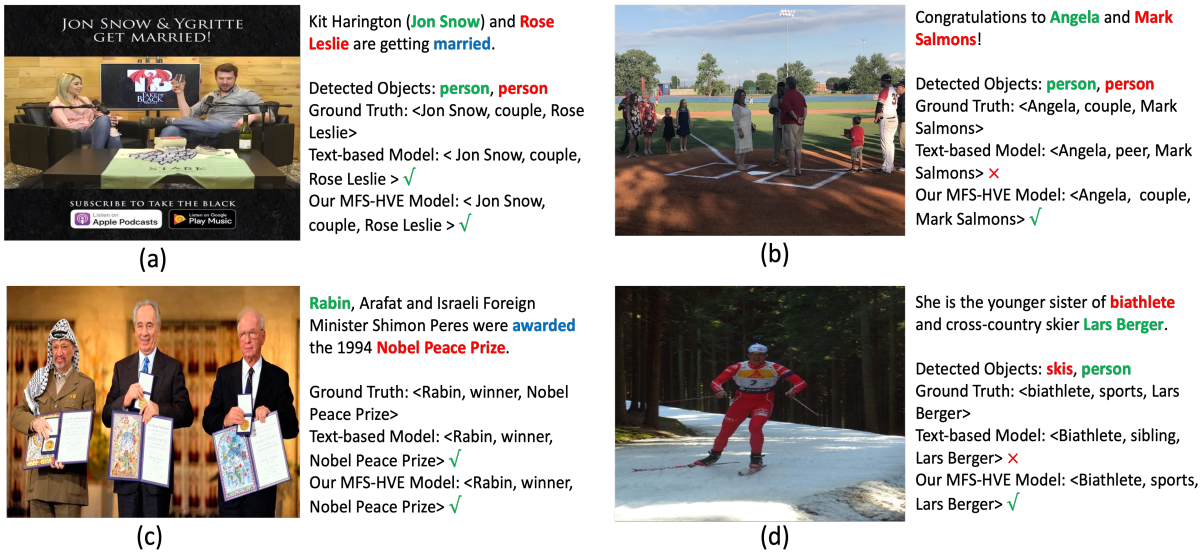

Figure 4: The examples of our proposed model MFS-HVE comparing to a text-based model on both the MNRE and FewRel datasets. We present the relation extraction results with the detected objects from the relevant image in the right column. The head entities are highlighted in green, whereas the tail entities are highlighted in red.

with noisy data.The ablation experiment results shown in Table 4 are reported by the mean value of five times the experimental results. We observe that utilizing multi-modal information performs better than uni-modal information (text). However, only using image-guided attention or object-guided attention can not achieve a great performance improvement. This is probably because considering the whole image from a global perspective may introduce noise to the text, resulting in a similar performance in few-shot settings compared with text-based models. In addition, if only object-guided textual attention is added to the model, the model still can not achieve a significant improvement. This is because not all images include the objects that are relevant to the name entities in the text. Thus, when the model jointly fuses image attention and object attention, there is a promising performance increase. The image attention overcomes the problem of sparsity, whereas the object attention reduces the noise brought by the whole image features. After adding hybrid feature attention to fuse all textual and visual information from both global and local perspectives, a significant performance gain is seen.

4.4.3. Case Study

Figure 4 shows the case study comparing our MFS-HVE model with a text-based model MTB on both MNRE and FewRel datasets. To evaluate the advantage and effectiveness of semantic visual information, we compare our model with an unimodal model, which only depends on textual information. We present four examples of two relations. For each relation, we present two cases. One case is that both the text-based model and the multimodal model MFS-HVE predict the relation correctly. The other case is that the relation is incorrectly predicted by the text-based model but correctly predicted by MFS-HVE.

Based on these examples, we observe that the text-based model only performs well when rich information is in the text. For the examples shown on the left, the text-based model can only correctly predict the relation 'couple' when relevant words or

phrases with similar meanings appear in the text, such as 'married' in the first sentence. Similarly, for the relation 'winner', the text-based model also performs well when the long textual sentence contains detailed information such as the word 'awarded'. These words relevant to the target relations provide enough semantic hints for the models with only text. However, not all cases have such long or detailed textual hints for the model. In the examples shown on the right, the textual sentences are short, without any words related to the target relation. In these cases, the text-based model can not predict the relation correctly. The text-based model predicts 'Angel' and 'Mark' are peers instead of 'couple', 'Roger Federer' is the 'participant of' the tennis tournament 'Wimbledon' instead of 'winner' of 'Wimbledon'. Nevertheless, with the guidance of informative visual evidence, more semantics are provided to the text. In the upper-right example, a wedding ceremony is shown in the image, and people objects are detected in the image. Based on this information, MFS-HVE correctly predicts the relation 'couple' instead of other relations in the MNRE dataset such as 'sibling', 'peer', 'parent', etc. Similarly, in the lower right example, MFS-HVE predicts the relation 'Roger Federer' is the 'winner' of 'Wimbledon' based on the visual information that a person is holding a tennis racket. In summary, integrating semantic visual information at both global and local levels provides more relevant information to supplement the missing contexts in textual sentences, resulting in a better and more robust performance for few-shot relation extraction.

4.4.4. Parameter Sensitivity

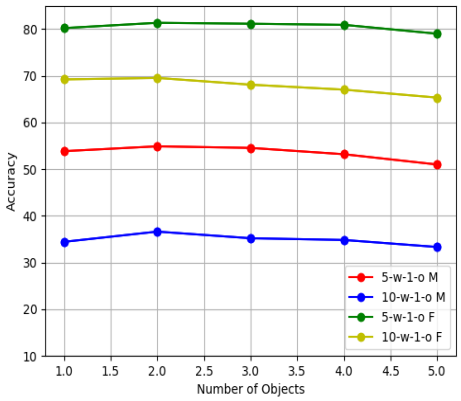

Figure 5: Effects on varying the number of embedded objects in one-shot settings on MNRE and $\text{FewRel}_{\text{small}}$ datasets.

Figure 5 shows the results of our proposed MFS-HVE model influenced by embedding a different number of objects detected from the image. By varying the object number from one to five, the results in terms of Accuracy on both MNRE and $\text{FewRel}_{\text{small}}$ are exhibited in Figure 5. We observe that the object number affects the performance of few-shot relation extraction. The model achieves the best performance when the object number is two. The performance drops when the object number increases. This is reasonable because relations always happen between two name entities. The two detected objects are usually relevant to the two corresponding name entities if the images are of high quality. Embedding only one object may lose critical information, whereas embedding lots more objects also introduces noise (irrelevant information) to the visual information.

5. Conclusion and Future Work

In this paper, we propose MFS-HVE, a multi-modal few-shot relation extraction approach leveraging semantic visual information to supplement the missing contexts in textual sentences. Our multi-modal fusion module consists of image-guided attention, object-guided attention, and hybrid feature attention that integrates information from different modalities. Experimental results demonstrate that MFS-HVE leveraging attention-based multi-modal information outperforms other uni-modal baselines and multi-modal fusion methods in few-shot relation extraction. In future work: (1) We will implement other powerful SOTA image encoders such as ViT (Dosovitskiy et al., 2021) to generate feature-level image embeddings. (2) We will explore utilizing the semantic visual information as an external source in zero-shot learning.

6. Ethical Considerations

Dataset Construction Because FewRel is a uni-modal relation extraction dataset, we obtain the images for each sentence from Wikidata, which is a free and collaborative knowledge base. Wikidata follows open data principles, which means that the data it contains is available to the public for various purposes, including research. We collect instance-related images from the wiki to make a multi-modal relation extraction dataset for experiments. Because not all instances from FewRel have relevant images, we remove the instances that do not have a corresponding relevant image on FewRel, resulting in a new dataset $\text{FewRel}_{\text{small}}$

Computing Cost Our proposed model MFS-HVE requires GPU training, which imposes a computational burden. Specifically, our model needs 5-6 hours of training on a single GPU card, which results in 0.25lbs of carbon dioxide emissions.

7. Bibliographical References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Hedi Ben-younes, Remi Cadene, Nicolas Thome, and Matthieu Cord. 2019. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8102–8109.

Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. Yolov4: Optimal speed and accuracy of object detection.

Christel Chappuis, Valérie Zermatten, Sylvain Lobry, Bertrand Le Saux, and Devis Tuia. 2022. Prompt-rsvqa: Prompting visual context to a language model for remote sensing visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1372–1381.

Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022a. 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.

Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022b. Good visual guidance makes a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. *arXiv preprint arXiv:2205.03521*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Bowen Dong, Yuan Yao, Ruobing Xie, Tianyu Gao, Xu Han, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. 2020. Meta-information guided meta-learning for few-shot relation classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1594–1605, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Manqing Dong, Chunguang Pan, and Zhipeng Luo. 2021. MapRE: An effective semantic mapping approach for low-resource relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2694–2704, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Chunliu Dou, Shaojuan Wu, Xiaowang Zhang, Zhiyong Feng, and Kewen Wang. 2022. Function-words adaptively enhanced attention networks for few-shot inverse relation classification. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 2937–2943. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Junhao Feng, Guohua Wang, Changmeng Zheng, Yi Cai, Ze Fu, Yaowei Wang, Xiao-Yong Wei, and Qing Li. 2023. Towards bridged vision and language: Learning cross-modal knowledge representation for relation extraction. *IEEE Transactions on Circuits and Systems for Video Technology*.

Ze Fu, Changmeng Zheng, Junhao Feng, Yi Cai, Xiao-Yong Wei, Yaowei Wang, and Qing Li. 2022. Drake: Deep pair-wise relation alignment for knowledge-enhanced multimodal scene graph generation in social media posts. *IEEE Transactions on Circuits and Systems for Video Technology*.

Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. 2019. Hybrid attention-based prototypical networks for noisy few-shot relation classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6407–6414.

Tianyu Gao, Xu Han, Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. 2020. Neural snowball for few-shot relation learning. In *AAAI*.

Zhi Gao, Yuwei Wu, Yunde Jia, and Mehrtash Harandi. 2021. Curvature generation in curved spaces for few-shot learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8671–8680.

Xiaoqing Geng, Xiwen Chen, Kenny Q. Zhu, Libin Shen, and Yinggong Zhao. 2020. Mick: A meta-learning framework for few-shot relation classification with small training data. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 415–424, New York, NY, USA. Association for Computing Machinery.

Jiaying Gong and Hoda Eldardiry. 2021. *Zero-Shot Relation Classification from Side Information*, page 576–585. Association for Computing Machinery, New York, NY, USA.

Jiaying Gong and Hoda Eldardiry. 2023. Prompt-based zero-shot relation extraction with semantic knowledge augmentation.

Peizhu Gong, Jin Liu, Xiliang Zhang, Xingye Li, and Zijun Yu. 2023. Circulant-interactive transformer with dimension-aware fusion for multimodal sentiment analysis. In *Asian Conference on Machine Learning*, pages 391–406. PMLR.

Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan Yuille. 2022. Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5078–5088.

Jiale Han, Bo Cheng, and Wei Lu. 2021. Exploring task difficulty for few-shot relation extraction. In *Proc. of EMNLP*.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. 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)*, pages 770–778.

Xuming Hu, Zhijiang Guo, Zhiyang Teng, Irwin King, and Philip S Yu. 2023. Multimodal relation extraction with cross-modal retrieval and synthesis. *arXiv preprint arXiv:2305.16166*.

Bei Hui, Liang Liu, Jia Chen, Xue Zhou, and Yuhui Nian. 2020. Few-shot relation classification by context attention-based prototypical networks with bert. *EURASIP Journal on Wireless Communications and Networking*, 2020.

Guangyuan Jiang, Manjie Xu, Shiji Xin, Wei Liang, Yujia Peng, Chi Zhang, and Yixin Zhu. 2023. Mewl: Few-shot multimodal word learning with referential uncertainty. *arXiv preprint arXiv:2306.00503*.

Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L. Iuzzolino, and Kazuhito Koishida. 2020. Mmtm: Multimodal transfer module for cnn fusion. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear Attention Networks. In *Advances in Neural Information Processing Systems 31*, pages 1571–1581.

Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, page 0. Lille.

Qian Li, Shu Guo, Cheng Ji, Xutan Peng, Shiyao Cui, and Jianxin Li. 2023. Dual-gated fusion with prefix-tuning for multi-modal relation extraction. *arXiv preprint arXiv:2306.11020*.

Wanli Li and Tieyun Qian. 2022. Graph-based model generation for few-shot relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 62–71, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. 2023. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19325–19337.

Shaoteng Liu, Jingjing Chen, Liangming Pan, Chong-Wah Ngo, Tat-Seng Chua, and Yu-Gang Jiang. 2020. Hyperbolic visual embedding learning for zero-shot recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang Liu, Jinpeng Hu, Xiang Wan, and Tsung-Hui Chang. 2022a. Learn from relation information: Towards prototype representation rectification for few-shot relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1822–1831, Seattle, United States. Association for Computational Linguistics.

Yang Liu, Jinpeng Hu, Xiang Wan, and Tsung-Hui Chang. 2022b. A simple yet effective relation information guided approach for few-shot relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 757–763, Dublin, Ireland. Association for Computational Linguistics.

Yun Liu, Xiaoming Zhang, Qianyun Zhang, Chaozhuo Li, Feiran Huang, Xianghong Tang, and Zhoujun Li. 2021. Dual self-attention with co-attention networks for visual question answering. *Pattern Recognition*, 117:107956.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999, Melbourne, Australia. Association for Computational Linguistics.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. 2018. A simple neural attentive meta-learner. In *International Conference on Learning Representations*.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. 2023. Med-flamingo: a multimodal medical few-shot learner. *arXiv preprint arXiv:2307.15189*.

Tsendsuren Munkhdalai and Hong Yu. 2017. Meta networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2554–2563. PMLR.

Ivona Najdenkoska, Xiantong Zhen, and Marcel Worring. 2023. Meta learning to bridge vision and language models for multimodal few-shot learning. *arXiv preprint arXiv:2302.14794*.

Xinzhe Ni, Hao Wen, Yong Liu, Yatai Ji, and Yujiu Yang. 2022. Multimodal prototype-enhanced network for few-shot action recognition. *arXiv preprint arXiv:2212.04873*.

Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. 2016. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, page 284–288, New York, NY, USA. Association for Computing Machinery.

Ahmed Osman and Wojciech Samek. 2019. Drau: Dual recurrent attention units for visual question answering. *Comput. Vis. Image Underst.*, 185:24–30.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Chengwei Qin and Shafiq Joty. 2022. Continual few-shot relation learning via embedding space regularization and data augmentation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2776–2789, Dublin, Ireland. Association for Computational Linguistics.

Meng Qu, Tianyu Gao, Louis-Pascal AC Xhonneux, and Jian Tang. 2020. Few-shot relation extraction via bayesian meta-learning on relation graphs. In *International Conference on Machine Learning*.

Victor Garcia Satorras and Joan Bruna Estrach. 2018. Few-shot learning with graph neural networks. In *International Conference on Learning Representations*.

Hrituraj Singh, Anshul Nasery, Denil Mehta, Aishwarya Agarwal, Jatin Lamba, and Balaji Vasan Srinivasan. 2021. MIMOQA: Multimodal input multimodal output question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5317–5332, Online. Association for Computational Linguistics.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lin Sun, Jiquan Wang, Yindu Su, Fangsheng Weng, Yuxuan Sun, Zengwei Zheng, and Yuanyi Chen. 2020. RIVA: A pre-trained tweet multimodal model based on text-image relation for multimodal NER. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1852–1862, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. 2018. Centralnet: a multilayer approach for multimodal fusion. In *ECCV Workshops*.

Hai Wan, Manrong Zhang, Jianfeng Du, Ziling Huang, Yufei Yang, and Jeff Z Pan. 2021a. Fl-msre: A few-shot learning based approach to multimodal social relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13916–13923.

Hai Wan, Manrong Zhang, Jianfeng Du, Ziling Huang, Yufei Yang, and Jeff Z. Pan. 2021b. Fl-msre: A few-shot learning based approach to multimodal social relation extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13916–13923.

Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, and Kewei Tu. 2022a. ITA: Image-text alignments for multi-modal named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3176–3189, Seattle, United States. Association for Computational Linguistics.

Xuwu Wang, Jiabo Ye, Zhixu Li, Junfeng Tian, Yong Jiang, Ming Yan, Ji Zhang, and Yanghua Xiao. 2022b. Cat-mner: Multimodal named entity recognition with knowledge-refined cross-modal attention. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. 2020a. Deep multimodal fusion by channel exchanging. *Advances in neural information processing systems*, 33:4835–4845.

Yingyao Wang, Junwei Bao, Guangyi Liu, Youzheng Wu, Xiaodong He, Bowen Zhou, and Tiejun Zhao. 2020b. Learning to decouple relations: Few-shot relation classification with entity-guided attention and confusion-aware training. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 5799–5809. International Committee on Computational Linguistics.

Yuyang Wanyan, Xiaoshan Yang, Chaofan Chen, and Changsheng Xu. 2023. Active exploration of multimodal complementarity for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6492–6502.

Aming Wu and Yahong Han. 2018. Multi-modal circulant fusion for video-to-language and backward. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 1029–1035. International Joint Conferences on Artificial Intelligence Organization.

Mingrui Wu, Xuying Zhang, Xiaoshuai Sun, Yiyi Zhou, Chao Chen, Jiaxin Gu, Xing Sun, and Rongrong Ji. 2022. Difnet: Boosting visual information flow for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18020–18029.

Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. 2020. *Multimodal Representation with Embedded Visual Guiding Objects for Named Entity Recognition in Social Media Posts*, page 1038–1046. Association for Computing Machinery, New York, NY, USA.

Bo Xu, Shizhou Huang, Ming Du, Hongya Wang, Hui Song, Chaofeng Sha, and Yanghua Xiao. 2022a. Different data, different modalities! reinforced data splitting for effective multimodal information extraction from social media posts. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1855–1864.

Bo Xu, Shizhou Huang, Chaofeng Sha, and Hongya Wang. 2022b. Maf: A general matching and alignment framework for multimodal named entity recognition. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 1215–1223, New York, NY, USA. Association for Computing Machinery.

Huaiping Yan, Erlei Zhang, Jun Wang, Chengcai Leng, and Jinye Peng. 2022. Mtffn: Multimodal transfer feature fusion network for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5.

Shan Yang, Yongfei Zhang, Guanglin Niu, Qinghua Zhao, and Shiliang Pu. 2021. Entity concept-enhanced few-shot relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 987–991, Online. Association for Computational Linguistics.

Xiaocui Yang, Shi Feng, Daling Wang, Pengfei Hong, and Soujanya Poria. 2022. Few-shot multimodal sentiment analysis based on multimodal probabilistic fusion prompts. *arXiv preprint arXiv:2211.06607*.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Multi-level matching and aggregation network for few-shot relation classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2872–2881, Florence, Italy. Association for Computational Linguistics.

Tan Yu, Yi Yang, Yi Li, Lin Liu, Hongliang Fei, and Ping Li. 2021. *Heterogeneous Attention Network for Effective and Efficient Cross-Modal Retrieval*, page 1146–1156. Association for Computing Machinery, New York, NY, USA.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jianbo Yuan, Han Guo, Zhiwei Jin, Hongxia Jin, Xianchao Zhang, and Jiebo Luo. 2017. One-shot learning for fine-grained relation extraction via convolutional siamese neural network. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 2194–2199.

Li Yuan, Yi Cai, Jin Wang, and Qing Li. 2023. Joint multimodal entity-relation extraction based on edge-enhanced graph alignment network and word-pair relation tagging. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11051–11059.

Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. 2021. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6995–7004.

Peiyuan Zhang and Wei Lu. 2022. Better few-shot relation extraction with label prompt dropout. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6996–7006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Wei Zhang, Yue Ying, Pan Lu, and Hongyuan Zha. 2020. Learning long- and short-term user literal-preference with multimodal hierarchical transformer network for personalized image caption. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9571–9578.

Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. 2021a. *Multimodal Relation Extraction with Efficient Graph Alignment*, page 5298–5306. Association for Computing Machinery, New York, NY, USA.

Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. 2021b. Multimodal relation extraction with efficient graph alignment. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5298–5306.

Changmeng Zheng, Zhiwei Wu, Junhao Feng, Ze Fu, and Yi Cai. 2021c. 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)*, pages 1–6. IEEE.

Changmeng Zheng, Zhiwei Wu, Junhao Feng, Ze Fu, and Yi Cai. 2021d. 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)*, pages 1–6.

Li Zhenzhen, Yuyang Zhang, Jian-Yun Nie, and Dongsheng Li. 2022. Improving few-shot relation classification by prototypical representation learning with definition text. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 454–464, Seattle, United States. Association for Computational Linguistics.

8. Appendix

8.1. Dataset Construction and Description

In the following, we describe each dataset in detail:

- ***MNRE (Zheng et al., 2021a).*** The MNRE dataset is a public human-annotated unbalanced multi-modal neural relation extraction

dataset. It is originally built upon Twitter15 (Lu et al., 2018), Twitter17 (Zhang et al., 2018) and crawling data from Twitter [3]. Each piece of data includes a sentence with two name entities and an image ID to correlate the text with the image. Because MNRE is a relation extraction dataset for supervised learning, there is an overlap of relations between the training and the testing dataset. For few-shot relation extraction, we randomly re-split the MNRE dataset to ensure no overlap of classes between the training and testing sets. There are 23 classes in total. After splitting the dataset, there are 13 classes for training and 10 classes for testing.

- ***FewRel (Han et al., 2018).*** The FewRel dataset is a public human-annotated balanced few-shot RC dataset consisting of 80 types of relations (64 for training and 16 for validation, another 20 for testing but it is not public), each of which has 700 instances. Because we need to combine images with the original text, so we only run experiments on the public part (64 training + 16 validation). Because FewRel is a fully uni-modal dataset, we insert an image ID to each instance to make it into a multi-modal relation extraction dataset. The image for each instance is automatically crawled by a built-in web crawler [4] on wiki data from the Google search engine.

- ***FewRel$_{small}$.*** FewRel$_{small}$ is a subset of FewRel. Because FewRel doesn't have image information, we crawl the images for FewRel. We view these images as external information, similar to auxiliary information such as label description, knowledge graphs, entity description, etc. Because images crawled for FewRel is an automatic process, some of the images are not relevant to their corresponding texts. Noise exists in the newly constructed multi-modal FewRel dataset. Noisy images are removed to ensure that FewRel$_{small}$ is a small, clean, and high-quality multi-modal few-shot relation extraction dataset. Note that we did not do any labeling work. The labels remain the same in FewRel$_{small}$ as FewRel, and we only add more information (images) for the existing dataset.

In all, FewRel is a balanced dataset. Due to the data cleaning, FewRel$_{small}$ is an unbalanced dataset. MNRE is also an unbalanced dataset.

[3] https://archive.org/details/twitterstream
[4] https://github.com/hellock/icrawler

8.2. Model Robustness

To further study the robustness of integrating visual information with textual information, we also conduct experiments on the model's performance comparison on FewRel and FewRel$_{small}$. To make fair comparisons, instead of directly reporting the performance of other state-of-the-art models, we re-implement other models with the same parameter settings as the models run on FewRel$_{small}$. Table 5 shows the results of performance decrease from dataset FewRel to FewRel$_{small}$ in few-shot settings. Because the FewRel dataset is more than ten times larger than FewRel$_{small}$, there are more training instances in FewRel. It is reasonable to expect a performance drop when the model is training on a smaller dataset. From Table5, we observe that the performance of text-based models drops significantly when the dataset tends to be smaller. This is because models usually can perform better when more data is available. In addition, we also find that models based on multi-modal information are more robust than text-based models. They have a smaller performance decrease than text-based models. Our proposed model MFS-HVE performs the best in the one-shot learning setting. We conjecture that the high-quality semantic visual information neutralizes the negative impact of little training data in FewRel$_{small}$, resulting in a more robust performance of multi-modal models.

8.3. Limitations

We view the following current limitations as some opportunities to build on in future work. First, MFS-HVE requires high-quality images for training. As shown in Table 3, MFS-HVE has a significant performance improvement compared with models using other text-based external information on MNRE. This is because MNRE is a public multi-modal dataset including clean and high-quality images. However, MFS-HVE shows a slight improvement or similar performance with models using other text-based external information on FewRel. The images crawled automatically contain much noise, which means some of the crawled images are irrelevant to the textual sentences. To further improve the performance on the FewRel dataset, human efforts or other crawling techniques are needed to get a large, clean, and high-quality image dataset.

Second, we compare MFS-HVE with five different fusion models introduced in Sec. 2.2. There are no existing multi-modal fusion models for the few-shot relation extraction task. We follow the five models' papers to implement the multi-modal fusion algorithms. To meet the requirement for few-shot learning, these fusion methods are built upon MTB (Baldini Soares et al., 2019). More latest multi-modal fusion methods are needed for

Table 5: Results of performance decrease in Accuracy(%) from FewRel to $\text{FewRel}_{\text{small}}$.

Model	5-Way 1-Shot	5-Way 5-Shot	10-Way 1-Shot	10-Way 5-Shot
GNN (Satorras and Estrach, 2018)	12.32	10.90	12.18	6.53
Snail (Mishra et al., 2018)	9.88	12.12	11.19	11.58
Siamese (Koch et al., 2015)	5.61	8.36	14.24	4.25
MLMAN (Ye and Ling, 2019)	3.30	1.97	2.70	2.25
Proto_BERT (Snell et al., 2017)	2.20	4.31	3.11	7.35
MTB (Baldini Soares et al., 2019)	3.14	1.00	3.54	3.66
ZSLRC (Gong and Eldardiry, 2021)	4.01	6.10	2.34	5.83
ConceptFERE (Yang et al., 2021)	3.56	2.96	3.34	3.76
REGRAB (Qu et al., 2020)	4.32	4.88	3.44	4.07
HCRP (Han et al., 2021)	4.36	3.00	2.76	4.24
MapRE (Dong et al., 2021)	6.29	7.24	8.47	8.80
FAEA (Dou et al., 2022)	7.97	6.78	4.55	5.59
SimpleFSRE (Liu et al., 2022b)	5.45	7.45	5.79	7.54
Concat (Wan et al., 2021a)	3.08	1.32	2.58	0.83
DeepFusion (Wang et al., 2020a)	2.14	4.72	0.38	**0.39**
CirculantFusion (Gong et al., 2023)	3.99	2.60	5.75	2.24
Dual Co-Att (Liu et al., 2021)	2.60	1.58	3.67	2.02
$\text{Proto}_{\text{multimodal}}$ (Ni et al., 2022)	2.01	3.08	3.75	2.99
MFS-HVE	**1.95**	**0.83**	**0.27**	1.32

performance comparison. To further improve the performance, more SOTA visual encoders such as ViT (Dosovitskiy et al., 2021) and large GPU memories are needed to conduct more experiments.

Finally, we want to clarify that our work focuses on few-shot relation extraction. We compare our model's performance with 14 SOTA open-code few-shot RE models and 5 different fusion models on two public English datasets. State-of-the-art multi-modal models in supervised learning for other tasks (i.e. NER, etc) or other languages besides English, are outside the scope of our paper because not all supervised models could be adapted/changed to few-shot settings as the training process is completely different.