Listen to Both Sides and be Enlightened! Hierarchical Modality Fusion Network for Entity and Relation Extraction

Anonymous ACL submission

Abstract

Multimodal named entity recognition and relation extraction (MNER and MRE) is a fundamental and crucial branch in multimodal learning. However, existing approaches for MNER and MRE mainly suffer from 1) error sensitivity when images contain irrelevant concepts not mentioned in texts; and 2) large modality gap between image and text features, especially hierarchical visual features. To deal with these issues, we propose a novel **H**ierarchical **M**odality fusion **NeT**work (**HMNeT**) for visual-enhanced entity and relation extraction, aim to reduce the modality gap and achieve more effective and robust performance. Specifically, we innovatively leverage hierarchical pyramidal visual features to conduct multi-layer internal integration in Transformer. We further present a dynamic gated aggregation strategy to decide modality integration according to different images. Extensive experiments on three benchmark datasets demonstrate the effectiveness of our method, and achieve state-of-the-art performance[1].

1 Introduction

Named entity recognition (NER) and relation extraction (RE) are important tasks in information extraction, due to its research significance in natural language processing (NLP) and wide applications, such as structural extraction (Hosseini, 2019; Qin et al., 2021) from massive news and web product information. Currently, with the rapid development of multimodal learning, multimodal NER (MNER) and Multimodal RE (MRE) methods (Moon et al., 2018; Zheng et al., 2021) have been proposed to enhance linguistic representations with the aid of visual clues from images. It significantly extends the text-based models by taking images as additional inputs, since the visual contexts help to resolve ambiguous multi-sense words.

[1]Code and datasets will be released for reproducibility.

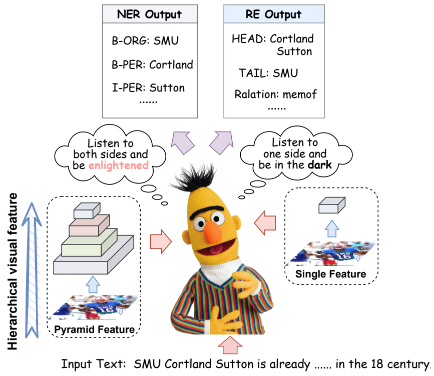


Figure 1: Motivation for robust and effective hierarchical modality fusion.

Traditional methods for MNER and MRE (Zhang et al., 2018; Moon et al., 2018; Yu et al., 2020; Zhang et al., 2021a; Zheng et al., 2021) have demonstrated the effectiveness of modality fusion. However, they mostly neglect two critical issues in modality fusion networks. The first issue is **error sensitivity**, where existing models are sensitive to wrong images since texts and images paris (e.g., tweets) could be irrelevant. As discussed in Vempala and Preotiuc-Pietro (2019), the images conveying abstract concepts instead of illustrating what is in the text are categorized as the "Image is irrelevant to the text" type, which is not rare in real-world data. Therefore, an effective method should be derived to learn *robust* multimodal representations for MNER and MRE tasks. The second issue is the **large modality gap** of image and text features. Previous models (Yu et al., 2020; Zheng et al., 2021) usually use the final output of Convolution Neural Networks (CNNs) with extremely abstract information as the visual representation, which ignore hierarchical pyramidal feature encoded in the different blocks of the visual backbone. Actually, linking such high-level visual features with semantic textual features demands a giant leap for the models to fill in the modality gap.

Intuitively, CNNs contain the pyramidal feature hierarchy, which contain semantics from low to high levels. Meanwhile, previous studies (Jawahar et al., 2019) illustrate that BERT (Devlin et al., 2019) encodes a rich hierarchy of linguistic information from the bottom to the top. This observation inspires us to make each layer of Transformer (Vaswani et al., 2017) aware of hierarchical visual features to make a more enlightened and comprehensive forecasting decision as shown in Figure 1. As the proverb says, ***"Listen to both sides and be enlightened, listen to one side and be in the dark."***, we speculate that MNER and MRE would benefit more from ***hierarchically dense learning signals like pyramidal visual features*** instead of single output feature of visual backbone[2], which can reduce the modality gap and also be more robust for the irrelevant image-text cases.

Thus, to tackle the above issues, we propose a novel **H**ierarchical **M**odality fusion **Ne**T**work** (**HMNeT**) for visual-enhanced entity and relation extraction. Specifically, we propose to make textual features of each layer broadly aware of hierarchical visual features through its self-attention module, thus reducing the modality gap and improving robustness. To automatically decide visual features of which block are suitable for Transformer, we design a dynamic gate for each layer to generate image-dependent paths, so that a variety of aggregated hierarchical visual features can be considered for further improvement. Overall, the major contributions of our paper can be summarized as follows:

- We present a hierarchical modality fusion framework towards MNER and MRE, incorporating hierarchical pyramidal visual features as visual prompts to generate effective and robust textual representation. To the best of our knowledge, it is the first work to leverage hierarchical pyramidal visual features for multimodal learning.

- We utilize the exploitation of dynamic gates to fully leverage the hierarchical visual features. Thus, textual representation of each layer in Transformer can be aware of corresponding hierarchical visual features adaptively.

- We evaluate our method on MNER and MRE tasks. Our experimental results on three benchmark datasets validate the effectiveness and superiority of our proposed method.

[2]Our method is suitable for various visual backbones, which refer to the feature extracting network used in CV, such as VGG (Simonyan and Zisserman, 2015), ResNet (He et al., 2016), etc.

2 Related work

Multimodal Entity and Relation Extraction As the crucial components of information extraction, named entity recognition (NER) and relation extraction (RE) have attracted much attention in the research community (Liu et al., 2019; Zhang et al., 2021b; Liu et al., 2021; Chen et al., 2021b,a). Previous studies typically focus on textual modality and standard text. As multimodal data become increasingly popular on social media platforms, early research focusing on textual modality and standard text is limited. Recently, several studies have focused on the MNER and MRE task, aiming to leverage the associate images to better identify the named entities and their relation contained in the text.

In the early stages, Zhang et al. (2018),Lu et al. (2018), (Moon et al., 2018) and Arshad et al. (2019) propose to encode the text through RNN and the whole image through CNN, then designing implicit interaction to model information between two modalities to explore multimodal NER tasks. Recently, Yu et al. (2020); Zhang et al. (2021a) propose to leverage regional image features to represent objects in the image to exploit fine-grained semantic correspondences based on Transformer and visual backbones.

While most of the current methods ignore the facts that irrelevant image-text instances may mislead the final prediction, one exception is that Sun et al. (2021), which proposes to learn a text-image relation classifier to enhance multimodal BERT to reduce the interference from irrelevant images while requiring extensive annotation for the irrelevance of image-text pairs.

Pre-trained Multimodal Representation The pre-trained multimodal BERT has recently achieved significant performance gains in many multimodal tasks (e.g., visual question answering). We summarize and compare the existing visual-linguistic BERT models in two aspects as follows: 1) **Architecture**. The single-stream structures consist of Unicoder-VL (Li et al., 2020), VisualBERT (Li et al., 2019), VL-BERT (Su et al., 2020), and UNITER (Chen et al., 2020b), where the image and text tokens were combined into a sequence and fed into BERT to learn contextual embeddings. The two-streams structures, LXMERT (Tan and Bansal, 2019) and ViLBERT (Lu et al., 2019), separate visual and language processing into two streams that interact through cross-modality or co-attentional transformer layers. 2) **Pretraining tasks**. The pretraining tasks mainly include masked language modeling (MLM), masked region classification (MRC), and image-text matching (ITM). However, most of these techniques are pre-trained on image captioning (Sharma et al., 2018; Chen et al., 2015) or visual question answering datasets where multimodal interactions are required. Applying these techniques to the MNER and MRE task may not result in a good performance, since **MNER and MRE mainly focus on leveraging visual information to enhance the text rather than conducting prediction on the image side**.

3 Methodology

As illustrated in Figure 2, we present a novel hierarchical modality fusion network for multi-modal entity and relation extraction. It is worth noting that our method can also be applied to other visual-enhanced tasks towards text.

3.1 Collection of Visual Clues

Language and vision provide complementary information. On the one hand, the image associated with a sentence maintains several visual objects related to the entities in the sentence, further providing more semantic knowledge to assist information extraction. On the other hand, the global image features may express abstract concepts, which play the role of a weak learning signal. Thus, we collect multiple visual clues for multimodal entity and relation extraction, which involves taking the regional image as the vital information and the global images as the supplement.

Given an image, we first conduct object detection with Fast-RCNN (Ren et al., 2015) and merely choose the top m salient objects with the higher object classification scores as the valid visual objects for assisting the semantic extraction based on the text further processing. Then, we rescale the global image and object image to 224 × 224 pixels as the **global image** $\mathcal{I}$ and **visual objects** $\mathcal{O} = \{o_1, o_2, ..., o_m, \}$.

3.2 Pyramidal Visual Feature

The feature fusion method effectively leveraging features from different blocks in the backbone model is widely used to improve the performance (Wang et al., 2019; Kim et al., 2018; Lin et al., 2017) of models in CV. Inspired by such practices, we take the first step to pay attention to the application of pyramid features in the field of multi-modality. We propose to fuse hierarchical image features into each Transformer layer; thus, leveraging a feature pyramid is essential. Typically, given an image, we encode it with a backbone model and generate a list of **pyramidal feature maps** $\{F_1, F_2, F_3, \ldots, F_c\}$ with different scales, then map them with $M_\theta(\cdot)$ as follows:

$$V_c = Conv_{1\times1}(F_c), \tag{1}$$
$$V_i = Conv_{1\times1}(Pool(F_i)), \ i = 1, 2, \dot{} c - 1, \tag{2}$$

where i denotes the i-th block the backbone model, c is the number of blocks in the visual backbone model (here is 4 for ResNet), *Pool* represents the pooling operation to generate the features respectively with the same spatial sizes. The 1×1 convolutional layer is leveraged to map the pyramidal visual features to match the embedding size of the Transformer.

3.3 Dynamic Gated Aggregation

Although the visual backbone and Transformer both have the trait of having low-level features at the bottom block and high-level semantic at the top block, it is not trivial to decide which block in the visual backbone is adopted to incorporate into each layer in Transformer. To address this challenge, we propose constructing the densely connected routing space, where hierarchical visual features are connected with each transformer layer.

3.3.1 Dynamic Gate Module

We conduct routine processes through a dynamic gate module, which can be viewed as a procedure of path decision. The motivation of the dynamic gate aims at predicting a normalized vector, which represents how much to execute the visual feature of each block. In the dynamic gate, $g_i^{(l)} \in [0, 1]$ denotes the path probability from the i-th block of visual backbone to the l-th layer of Transformer. It is calculated as $g^{(l)} = \mathbb{G}^{(l)}(V) \in \mathbb{R}^c$, where $\mathbb{G}^{(l)}(\cdot)$ denotes the gating function according to the l-th layer in Transformer, c represents the numbers of the block in backbone. We first produces the logits $\alpha_i^{(l)}$ of the gate signals:

$$\alpha^{(l)} = f(W_l(\frac{1}{c}\sum_{i=1}^{c} P(V_i))), \tag{3}$$

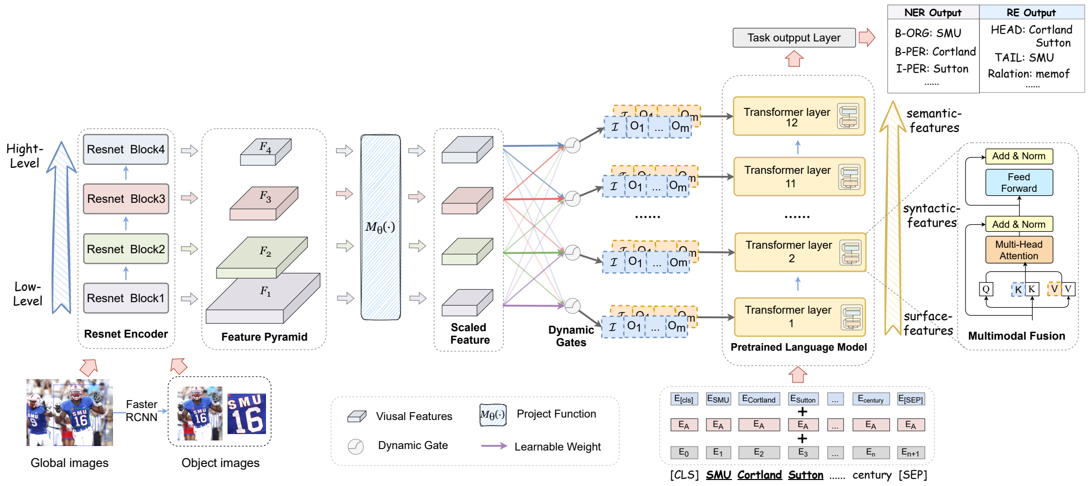


Figure 2: The overall architecture of our hierarchical modality fusion network.

where $f(\cdot)$ denotes the activate function Leaky_ReLU, P represents the global average pooling layer. The input features V_i with a shape of (d_i, h_i, w) from the i-th block in the visual backbone model are firstly squeezed by an average pooling operation and added the features from multiple blocks to generate the average vectors. Then we reduce the feature dimension by c with the MLP layer W_l. We further consider a soft gate via generating continuous values as path probabilities. Afterward, we generate the probability vector $g^{(l)}$ for the l-th layer of Transformer as follows:

$$g^{(l)} = Softmax(\alpha^{(l)}) \tag{4}$$

3.3.2 Aggregated Hierarchical Visual Feature

Based on the above dynamic gate $g^{(l)}$, we can derive the final aggregated hierarchical visual feature V_{gated} to match the l-th layer in Transformer, as:

$$V_{gated}^{(l)} = g^{(l)} V^{(l)}. \tag{5}$$

Formally, to further fully exploit the features of global and local images, the multi-granularity visual features $\tilde{V}_{gated}^{(l)}$ corresponding to the l-th layer of Transformer is obtained by the following concat operation,

$$\tilde{V}_{gated}^{(l)} = [V_{gated}^{(l,I)}; V_{gated}^{(l,o_1)}; \ldots; V_{gated}^{(l,o_m)}], \tag{6}$$

which will be adopted to enhance into layer-level representations of textual modality.

3.4 Multi-layer Internal Integration

Since we attempt to push each layer of the Transformer to view the hierarchical visual features, it is intuitively to leverage the self-attention module of the Transformer rather than extra cross-modal attention independent of visual and textual representation encoders. In particular, given an input sequence $X = \{x_1, x_2, ..., x_n\}$, the contextual representations $\boldsymbol{H}^{l-1} \in \mathbb{R}^{n \times d}$ is first projected into the query/key/value vector:

$$\boldsymbol{Q}^l = \boldsymbol{H}^{l-1}\boldsymbol{W}_l^Q, \boldsymbol{K}^l = \boldsymbol{H}^{l-1}\boldsymbol{W}_l^K, \boldsymbol{V}^l = \boldsymbol{H}^{l-1}\boldsymbol{W}_l^V. \tag{7}$$

As for aggregated hierarchical visual features $\tilde{V}_{gated}^{(l)}$, we use a set of linear transformations $W_l^{\phi} \in \mathbb{R}^{d \times 2 \times d}$ for l-th layer to project them into the same embedding space[3] of textual representation in self-attention module. Besieds, we define the operation of visual prompt $\phi_k^l, \phi_v^l \in \mathbb{R}^{hw(m+1) \times d}$ as:

$$\{\phi_k^l, \phi_v^l\} = \tilde{V}_{gated}^{(l)} W_l^{\phi}, \tag{8}$$

where $hw(m+1)$ denotes the length of the visual sequences, m is the number of visual objects detected by the object detection algorithm. As shown in Figure 2, different from previous co-attention methods, we regard hierarchical visual features as visual prompts at each fusion layer and sequentially conduct multi-modal attention to update all textual states. In this way, the final textual states encode both the context and the cross-modal semantic information simultaneously. Formally, the visual fusion are calculated as follows:

$$Attention^l = softmax(\frac{\boldsymbol{Q}^l[\phi_k^l; \boldsymbol{K}^l]^T}{\sqrt{d}})[\phi_v^l; \boldsymbol{V}^l]. \tag{9}$$

[3]Remarkably, the key and value in the self-attention module contain the different information in two types of semantic space, here 2 means that we apply two sets of transformation parameters to project aggregated visual features to match the state update process, respectively.

3.5 Classifier

Baesd on above description, we get the final representation of BERT, $H^L = U(X, \tilde{V}^{(l)}_{gated})$, where $U(\cdot)$ denotes the operation of multi-layer internal integration. Finally, we conduct different classifier layers for NER and RE, respectively.

Named Entity Recognition. Follow previous works (Moon et al., 2018; Yu et al., 2020), we also adopt the CRF decoder to perform the NER task. Formally, we feed the final hidden vectors $\boldsymbol{H}^L =$ of BERT to the CRF model. For a sequence of tags $y = \{y_1, \dots, y_n\}$, the probability of the label sequence y and the objective of NER are defined as follows (Lample et al., 2016a):

$$p(y|H^L) = \frac{\prod_{i=1}^{n} S_i(y_{i-1}, y_i, H^L)}{\sum_{y' \in Y} \prod_{i=1}^{n} S_i(y'_{i-1}, y'_i, H^L)},$$
$$\mathcal{L}_{ner} = -\sum_{i=1}^{M} log(p(y^{(i)}|U(X^{(i)}, \tilde{V}_{gated}))). \tag{10}$$

where Y is the pre-defined label set with the BIO tagging schema, and $S(\cdot)$ are potential functions. Details can be referred in (Lample et al., 2016a).

Relation Extraction. An RE dataset can be denoted as $\mathcal{D}_{re} = \{(X^{(i)}, r^{(i)})\}_{i=1}^{M}$, the goal of RE is to predict the relation $r \in \mathcal{Y}$ between subject entity and object entity. Specifically, a `[CLS]` head is utilized to compute the probability distribution over the class set $\mathcal{Y}$ with the softmax function $p(r|X) = \texttt{Softmax}(\mathbf{W}\mathbf{H^L}_{\texttt{[CLS]}})$, and the parameters of $\mathcal{L}$ and $\mathbf{W}$ are fine-tuned by minimizing the cross-entropy loss over $p(r|X)$ on the entire $\mathcal{X}$ as follows:

$$\mathcal{L}_{re} = -\sum_{i=1}^{M} log(p(r^{(i)}|U(X^{(i)}, \tilde{V}_{gated}))). \tag{11}$$

4 Experiments

In this section, we conduct experiments to evaluate our method on two multimodal information extraction tasks, MNER and MRE. Specifically, we adopt ResNet50 (He et al., 2016) as visual backbone and BERT-base (Devlin et al., 2019) as textual encoder. Results on three datasets demonstrate that our HMNeT outperforms a number of unimodal and multimodal approaches.

4.1 Datasets

We adopt three datasets in our experiments: Twitter-2015 (Zhang et al., 2018) and Twitter-2017 (Lu et al., 2018) for MNER, MNRE (Zheng et al., 2021) for MRE. Statistical details of datasets and experimental details are provided in Appendix A and B.

4.2 Compared Baselines

We compare our HMNeT with several baseline models for a comprehensive comparison to demonstrate the superiority of our HMNeT. Our comparison mainly focuses on three groups of models: the text-based models, previous SOTA MNER and MRE models, and the variants of our models.

Text-based models: we first consider a group of representative text-based models: 1) *CNN-BiLSTM-CRF* (Ma and Hovy, 2016), 2) *HBiLSTM-CRF* (Lample et al., 2016b) and 3) *BERT-CRF* for NER. The following models are specific for RE: 4) *PCNN* (Zeng et al., 2015); 5) *MTB* (Soares et al., 2019) is an RE-oriented pretraining model based on BERT.

Previous SOTA models: besides, we further consider another group of previous SOTA multi-modal approaches for MNER and MRE: 1) *AdapCoAtt-BERT-CRF* (Zhang et al., 2018); 2) *OCSGA* (Wu et al., 2020); 3) *UMT* (Yu et al., 2020); 4) *UMGF* (Zhang et al., 2021a), the newest SOTA for MNER, which proposes a unified multi-modal graph fusion approach for MNER. 5) *BERT+SG* is proposed in Zheng et al. (2021) for MRE, which concatenate the textual representation from BERT with visual features generated with scene graph (SG) tool (Tang et al., 2020). 6) *MEGA* (Zheng et al., 2021), the newest SOTA for MRE, which develops a dual graph for multi-modal alignment to capture this correlation between entities and objects for better performance. 7) *VisualBERT*(Li et al., 2019), different from the above SOTA methods mainly based on co-attention, VisualBERT is a single-stream structure, which is a strong baseline for comparison. And the results of VisualBERT listed in our paper is referred from Chen et al. (2020a)

Variants of Our Model: we set the ablation experiments to explore the effectiveness of our design. We conduct on the same parameter settings of HMNeT for each variant model for a fair comparison.

HMNeT-Single: This model is an variant of our model without the pyramid structure, which maps the visual features derived from 4-th block of ResNet to the last layer corresponding to BERT.

Modality	Methods	*Twitter-2015*			*Twitter-2017*			*MNRE*		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Text	CNN-BiLSTM-CRF	66.24	68.09	67.15	80.00	78.76	79.37	-	-	-
	HBiLSTM-CRF	70.32	68.05	69.17	82.69	78.16	80.37	-	-	-
	BERT-CRF	69.22	74.59	71.81	83.32	83.57	83.44	-	-	-
	PCNN	-	-	-	-	-	-	62.85	49.69	55.49
	MTB	-	-	-	-	-	-	64.46	57.81	60.86
Text+Image	AdapCoAtt-BERT-CRF	69.87	74.59	72.15	85.13	83.20	84.10	-	-	-
	OCSGA	74.71	71.21	72.92	-	-	-	-	-	-
	UMT	71.67	75.23	73.41	85.28	85.34	85.31	-	-	-
	UMGF	74.49	75.21	74.85	86.54	84.50	85.51	-	-	-
	BERT+SG	-	-	-	-	-	-	62.95	62.65	62.80
	MEGA	-	-	-	-	-	-	64.51	68.44	66.41
	VisualBERT	68.84	71.39	70.09	84.06	85.39	84.72	63.25	66.80	65.00
	HMNeT-Single	72.61	74.35	73.48	84.61	84.42	84.51	78.30	75.63	76.94
	HMNeT-Flat	73.76	75.32	74.54	84.43	86.42	85.41	79.32	78.20	78.75
	HMNeT-1V3	**74.25**	75.45	74.85	85.42	86.85	86.13	82.48	80.16	81.30
	HMNeT-OnlyObj	74.07	76.23	75.15	85.58	87.52	86.55	81.57	80.94	81.25
	HMNeT	73.85	**78.23**	**75.98**	**85.84**	**87.93**	**86.87**	**83.64**	**80.78**	**81.85**

Table 1: Performance comparison of different competitive baseline approaches for NER and RE.

HMNeT-Flat: This is another variant of our model without the pyramid structure. Specifically, we assign the output of the 4-th block of ResNet as the visual features and then map the visual features to each layer corresponding to BERT to conduct image-text fusion.

HMNeT-1V3: As ResNet and BERT have four blocks and 12 layers, respectively thus, it is intuitive to directly map visual features in one block to the three layers in BERT. We denote this variant as *HMNeT-1V3* to compare with our final version with dynamic gate mechanism.

HMNeT-OnlyObj: Visual objects are considered as fine-grained image representations. We conduct ablation by only adopting the object-level features in this model to validate the effect of the object features.

4.3 Overall Performance Comparison

4.3.1 Main Results

The experimental results of HMNeT and all baselines on three testing sets are presented in Table 1. From the experimental results, we can observe that:

Firstly, we can find that incorporating the visual features is generally helpful for NER and RE tasks by comparing the SOTA multimodal approaches with their corresponding text-based baselines. Despite previous multimodal approaches can generally achieve better performance, the enormous improvement of F1 score for NER is only about 2.0% (compare UMGF with BERT-CRF), which for RE is about 5.55% (compare MEGA with MTB). This observation reveals that the performance improvement of images on text-based NER tasks is relatively limited compared with RE tasks.

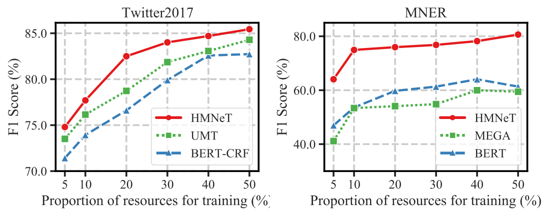


Figure 3: Performances on low-resource setting on MNER and MRE task.

Secondly, our method is superior to the newest SOTA models UMGF and MEGA, which improves 1.14%, 1.36%, and 15.44% F1 scores for Twitter-2015, Twitter-2017, and MNRE datasets, respectively. While previous multimodal methods all merely leverage the highest-level features of two modalities based on extra co-attention networks, which belong to the two-stream structure. This results indicate that elaborately establishing hierarchical fusion with pyramidal visual features is beneficial for multimodal tasks.

Finally, we also compare with VisualBERT, which is a pre-trained multimodal BERT with a single-stream structure. We notice that even as the pre-trained multimodal model, VisualBERT leaves much to be desired in MNER and MRE tasks, which performs worse than UMGF and MEGA, let alone our methods. We hold that VisualBERT is truly dissatisfactory since the datasets and pre-training process are less relevant to information extraction tasks.

4.3.2 Low-resource Scenario

Figure 3 shows the performance of our method in a low-resource scenario compared with several baselines. By analyzing this results, we can ob-

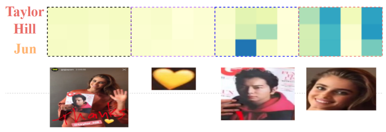
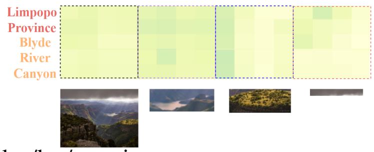
	Relevant Image-text Pair	Weak Relevant Image-text Pair	Irrelevant Image-text Pair
	Taylor Hill holding Jun 's GQ japan lol.	Cold front over Blyde River Canyon in Limpopo Province , South Africa.	President Bush when he sees the lights of America .
	Text-Images Attention of HMNeT 		
	Gold Relations: per/per/couple	loc/loc/contain	per/loc/place_of_residence
BERT:	per / per /couple ✗	misc / misc /part_of ✗	per / loc /place_of_residence ✔
VisualBERT:	per / per /peer ✔	misc / misc /part_of ✗	misc / loc /held_on ✗
MEGA:	per / per /peer ✔	per / per /peer ✗	misc / loc /held_on ✗
HMNeT(Ours):	per / per /peer ✔	loc / loc /contain ✔	per / loc /place_of_residence ✔

Table 2: The first row shows the split of the relevance of image-text pairs, and the several middle rows indicate representative samples together with their entity-object attention in the test set of MNRE datasets, and the bottom four rows show predicted relation of different approaches on these test samples.

serve: 1) UMT and MEGA consistently outperform the compared baselines in the low-resource scenario; the improvement indicates that incorporating the visual features is still helpful for NER and RE tasks in low-resource scenarios. 2) Moreover, it can be observed that the performance of HMNeT still outperforms the other baselines. It further proves the effectiveness and robustness of our proposed method. This may be attributed to letting BERT listen to hierarchical visual features rather than only the final high-level features, thus, effectively injecting visual knowledge.

4.3.3 Cross-task Scenario

Table 3 shows performance comparison of HMNeT and UMGF in a cross-task scenario for versatility analysis. For the first part, Twitter2017 → MNRE denotes that the trained model on Twitter-2017 is further used to train and test on MNRE. For the second part, MNRE → Twitter-2017 represents that the trained model on Twitter-2017 is used to further train and test on Twitter-2017. From this Table, we can observe that our HMNeT significantly outperforms UMGF by a more considerable margin. It is worth noting that our method can achieve further improvement in a cross-task scenario, while UMGF performs worse than previous results on the corresponding dataset. This justifies that our HMNeT is robust to automatically reduce the interference of visual information of irrelevant picture; thus, more image-text data may facilitate learning better parameters for modality fusion. Besides, it is also interesting to extend our work to multi-task learning or multi-modal pre-training and we leave these for further works.

Methods	*Twitter-2017 → MNRE*	*MNRE → Twitter-2017*
UMGF	63.85 → 62.90 ↓ (0.95)	85.51 → 84.35 ↓ (1.16)
HMNeT	81.85 → 82.50 ↑ (0.75)	86.87 → 87.13 ↑ (0.26)

Table 3: Performance comparison of HMNeT and UMGF in cross-task scenario.

4.4 Detailed Model Analysis

Ablation Study. In this part, we conduct extensive experiments with the variants of our model to further analyze the effectiveness of our model. Table 1 shows the results of the variant set. We observe that:

(1) ***Multi-layer Internal Integration.*** To gain insights into our design of multi-layer fusion, we conduct ablation studies incrementally to compared previous SOTA models with the following variants: 1) HMNeT-Single and 2) HMNeT-Flat. On the one hand, compared with HMNeT and HMNeT-Flat, the performance of HMNeT-Single degrades dramatically on all criteria of three datasets. On the other hand, HMNeT-Flat is comparable to previous SOTA models and even perform much better than MEGA in multimodal RE task. Note that these empirical findings indicate that layer-wise visual knowledge guidance (Allow every layer of BERT to see high-level visual features.) is beneficial.

(2) ***Dynamic Gated Aggregation.*** To validate the impact of our proposed dynamic gate mechanism, we carry out experiments by introducing two variants: 1) HMNeT-Flat, crudely conducting multi-layer fusion with single visual feature; and 2) HMNeT-1V3, intuitively leveraging hierarchical visual features from low-level to high-level blocks. We observe that HMNeT with dynamic

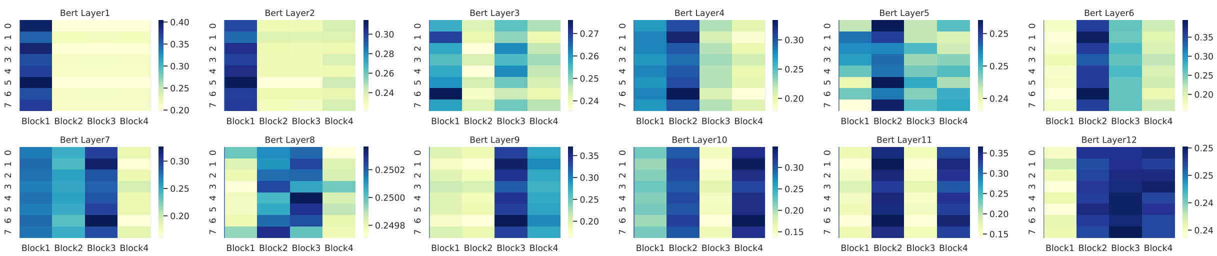


Figure 4: Visualization of dynamic gate learned on MNER task. Each subgraph denotes one layer in BERT, and the ordinate and abscissa respectively represent the instance id in a batch and the block id of ResNet.

gate achieves the best performance consistently compared with the other variants. Although the HMNeT-1V3 performs slightly lower than the version of dynamic gate, it still outperforms the crude variant HMNeT-Flat. It reveals that the dynamic gate can automatically learn appropriate weights for different visual clues, enabling the model to explore possible optimal visual pyramidal features polymerization for each Transformer layer.

(3) ***Visual Clues Term.*** As recent SOTA models such as UMT, UMGF, and MEGA all adopt visual objects to enhance textual representation, we conduct experiments by ablating global images to explore the impact of the visual clues. As expected, we find that HMNeT-OnlyObj performs slightly worse than HMNeT, which is consistent with the observation of previous works. This can be attributed to that abstract clues maybe not be associated with the text in information extraction tasks. In other words, this empirical finding demonstrates the flexibility of our methods to infuse visual clues with different granularity.

Case Analysis for Image-text Relevance To validate the effectiveness and robustness of our method, we conduct case analysis for image-text relevance as indicated in Table 2 We notice that VisualBERT, MEGA, and our method can recognize the relation for the relevant image-text pair. We can further find that the attention between relevant entities and objects is significant. While in the situation that image represents the abstract semantic that is weak relevant to the text, only our method success in prediction due to HMNeT captures the more hierarchical features. It should be noted that another two multimodal baselines fail in irrelevant image-text pairs while text-based BERT and ours still predict correctly. These observations reveal that our model can learn more robust multimodal representation dynamically, which is essential for the noise of uncorrelated image-text samples.

Gate Visualization We hypothesis that the key component of HMNeT achieving the superior performance is the dynamic gated aggregation in multi-layer internal integration, which can adaptively assign different modality integration paths for different input images. To this end, we randomly sample eight images in a batch and visualize their gate vectors learned by HMNeT according to 12 layers of BERT in Figure 4. Note that HMNeT-1V3 perform a little worse than our HMNeT, and the optimized gate vectors follow the trend of matching low-level textual semantics with low-level visual semantics and matching high-level textual semantics with high-level visual semantics. Meanwhile, the modality fusion obtained by dynamic gate learning may provide some valuable insights for efficient visual-language approaches in the future.

5 Conclusion and Future Work

In this paper, inspired by the proverb ***"Listen to both sides and be enlightened, listen to one side and be in the dark."***, we propose a hierarchical modality fusion framework towards multimodal NER and RE to reduce modality gap and bias of irrelevant image-text pairs, which is the first work leveraging hierarchical pyramidal visual features to conduct multi-layer internal integration in Transformer. Concretely, we propose a multi-layer internal integration network for modality fusion, and design a dynamic gated aggregation strategy to extract hierarchical visual features automatically. Extensive experimental results on three benchmarks have demonstrated the effectiveness and robustness of our proposed method.

In the future, we plan to 1) explore more applications of hierarchical modality fusion framework in multimodal representation learning, making it more flexible and extensible; 2) apply the reverse version of our approach to boost visual representation with text for CV; 3) extend our approach to multitask multimodal pre-training.

References

Omer Arshad, Ignazio Gallo, Shah Nawaz, and Alessandro Calefati. 2019. Aiding intra-text representations with visual context for multimodal named entity recognition. *ArXiv preprint*, abs/1904.01356.

Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. 2020a. A caption is worth A thousand images: Investigating image captions for multimodal named entity recognition. *CoRR*, abs/2010.12712.

Xiang Chen, Ningyu Zhang, Lei Li, Xin Xie, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2021a. Lightner: A lightweight generative framework with prompt-guided attention for low-resource NER. *CoRR*, abs/2109.00720.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2021b. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. *CoRR*, abs/2104.07650.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020b. UNITER: universal image-text representation learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer.

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

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

Hawre Hosseini. 2019. Implicit entity recognition, classification and linking in tweets. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, page 1448. ACM.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3651–3657. Association for Computational Linguistics.

Seung-Wook Kim, Hyong-Keun Kook, Jee-Young Sun, Mun-Cheon Kang, and Sung-Jea Ko. 2018. Parallel feature pyramid network for object detection. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V*, volume 11209 of *Lecture Notes in Computer Science*, pages 239–256. Springer.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016a. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016b. Neural architectures for named entity recognition. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 260–270. The Association for Computational Linguistics.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. 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, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11336–11344. AAAI Press.

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

Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. 2017. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 936–944. IEEE Computer Society.

Kun Liu, Yao Fu, Chuanqi Tan, Mosha Chen, Ningyu Zhang, Songfang Huang, and Sheng Gao. 2021. Noisy-labeled NER with confidence estimation. In *Proceedings of NAACL*, pages 3437–3445. Association for Computational Linguistics.

Yijin Liu, Fandong Meng, Jinchao Zhang, Jinan Xu, Yufeng Chen, and Jie Zhou. 2019. GCDT: A global context enhanced deep transition architecture for sequence labeling. In *Proceedings of ACL*, pages 2431–2441, Florence, Italy. Association for Computational Linguistics.

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.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity recognition for short social media posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 852–860, New Orleans, Louisiana. Association for Computational Linguistics.

Yujia Qin, Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. 2021. ERICA: improving entity and relation understanding for pre-trained language models via contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3350–3363. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2556–2565. Association for Computational Linguistics.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2895–2905. Association for Computational Linguistics.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pretraining of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: A text-image relation propagation-based BERT model for multimodal NER. 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, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13860–13868. AAAI Press.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3713–3722. Computer Vision Foundation / IEEE.

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

Alakananda Vempala and Daniel Preotiuc-Pietro. 2019. Categorizing and inferring the relationship between the text and image of twitter posts. In *Proceedings* of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2830–2840. Association for Computational Linguistics.

Tiancai Wang, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. 2019. Learning rich features at high-speed for single-shot object detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 1971–1980. IEEE.

Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-Fung Leung, and Qing Li 0001. 2020. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 1038–1046. ACM.

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

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1753–1762. The Association for Computational Linguistics.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021a. Multimodal 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, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14347–14355. AAAI Press.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021b. Document-level relation extraction as semantic segmentation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3999–4006. ijcai.org.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5674–5681. AAAI Press.

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

A Detailed Statistics of Dataset

Dataset	Train	Dev	Test	Avg length (characters)
Twitter-2015	4,000	1,000	3,257	95
Twitter-2017	4,290	1,432	1,459	64

Table 4: Size of the datasets in numbers of tweets.

Dataset	# Sent.	# Ent.	# Rel.	# Img.
TACRED	53,791	152,527	41	-
MNRE	14,796	20,178	31	10,089

Table 5: Comparison of MNRE with existing sentence-level Relation Extraction dataset TACRED (Sent.: sentence, Ent.: entity, Rel.: relation,Img.: image.

B Experimental Details

This section details the training procedures and hyperparameters for each of the datasets. Considering the instability of the few-shot learning, we run each experiment 5 times on the random seed [1, 49, 1234, 2021, 4321] and report the averaged performance. We utilize Pytorch to conduct experiments with 1 Nvidia 3090 GPUs. All optimizations are performed with the AdamW optimizer with a linear warmup of learning rate over the first 10% of gradient updates to a maximum value, then linear decay over the remainder of the training. And weight decay on all non-bias parameters is set to 0.01. We set the number of image objects m to 3. We describe the details of the training hyper-parameters in the following sections.

B.1 Standard Supervised Setting

In the MNER task, we fix the batch size as 8 and search for the learning rates in varied intervals [1e-5, 3e-5]. We train the model for 30 epochs and do evaluation after the 16th epoch. In the MRE task, we fix the batch size as 32 and learning rates as 1e-5. We train the model for 12 epochs and do evaluation after the 8th epoch. In the two tasks, we all choices the model performing the best on the validation set and evaluate it on the test set.

B.2 Low-Resource Setting

For different instances per class, we sample five times on the random seed [1, 2, 49, 4321, 1234] and report the averaged performance. For all models, we fix the batch size as 8 and search for the learning rates in varied intervals [3e-5, 5e-5]. We train the model for 30 epochs and do evaluation after the 16th epoch. We choose the model performing the best on the validation set and evaluate it on the test set.

B.3 Cross-Task Setting

In the MNER task and RE task, we all use ResNet and BERT-base as the backbone, we transfer the same parameters except the classifier layer and CRF layer when we do cross-task. In further training, we fix the batch size as 8 and search for the learning rates in varied intervals [1e-5, 3e-5]. We train the model for12 epochs and do evaluation after the 8th epoch. We choose the model performing the best on the validation set and evaluate it on the test set.