MPT: Multimodal Prompt Tuning for Event Detection

Anonymous ACL submission

Abstract

Event Detection is a key and challenging subtask of event extraction, which has serious trigger word ambiguity. Existing studies mainly focus on contextual information in text, while there are naturally many images in news articles that need to be explored. We believe that images not only reflect the core events of the text but also help to trigger word disambiguation. In this paper, we propose a new bi-recursive multimodal Prompt Tuning (MPT) model for deep interaction between images and sentences to achieve aggregation of modal features. MPT uses pre-trained CLIP to encode and map sentences and images into the same multimodal semantic space and uses alternating dual attention to select information features for mutual enhancement. Then, a soft prompt method of multimodal guidance is proposed, and the multimodal information obtained by fusion is used to guide the downstream event detection task. Our superior performance compared to six state-of-the-art baselines and further ablation studies, demonstrate the importance of image modality and the effectiveness of the proposed architecture.

1 Introduction

Events describe state changes of participating entities. The Event Detection (ED) task is one of the essential tasks in the Information Extraction field. Event triggers are the most representative words or phrases in events, and they are usually composed of verbs or nouns (Doddington et al., 2004). There is a one-to-one correspondence between events and event triggers, so the ED task is equivalent to identifying and classifying event triggers.

As shown on the left side of Figure 1, since the confront refers to the occurrence of the event meet, it should be marked as the event trigger word of the meet. Event detection has important implications for various natural language processing tasks such as text summarization, auto summarization, machine question and answer (QA), etc.

Figure 1: The exact trigger word triggers two different events, but the semantic distinction can be perceived through the image.

ED is a challenging task because trigger words must be representative, and the localization of these trigger words is often ambiguous in relative terms. A word can trigger different events, and the surrounding context often doesn't have enough information to disambiguate them. For example, in Figure 1, the trigger word confront triggers different events due to its meaning in different contexts: meet and attack. Existing approaches address this problem by introducing a global context throughout the article or by introducing some additional linguistic resources.

In the event world, a complete interpretation of an event is often completed through multiple media (text, images, videos). The multimodal form of images accompanying news articles and natural language is becoming increasingly common in the media industry. Among them, image information has been proven to have the ability to disambiguate text by providing information gain and semantic coreference, and this disambiguation feature is just suitable for ED tasks.

This fit is reflected in the following two aspects:The accompanying images usually reflect the core events of the texts. As shown in Figure 1, the first example contains two candidate verbs of the event trigger: was and confronted, where confront is more representative in texts and is also the main content of the image. (2) Since trigger words

provide complementary information, such as physical information, image style, or action, which is difficult to describe with words, images help to disambiguate trigger words. (Tong et al., 2020) also demonstrated the role of images in ED, using the global information of images to disambiguation entities, and they significantly improved the performance of ED.

In this paper, we introduce the original images of news articles into the ED scene. Nevertheless, at the same time, the following difficulties need to be overcome: first, although there is more and more research on multimodal information, there is still no recognized method for incorporating image modalities into NLP tasks; secondly, the alignment between multimodal information Dimensionality needs to be considered, in ED scenarios, images should help the model identify specific events, but should these images map to events, or specific words, sentences, or entities like (citations) There are also considerations; finally, in the ED task In addition to the additional visual information introduced, a large amount of noise will also be added, and the effect of the additional noise is often more significant than the disambiguation gain provided by the multimodal information. Therefore, how to introduce multimodal information, how to determine the alignment dimension of the imported information and text information, how to use the semantic information of multimodal information as a practical guide, and how to assist the model in classifying are also problems that we need to solve after introducing visual information.

To address these issues, We propose a bi-recursive multimodal Prompt Tuning model to deeply interact between images and text for modal feature fusion and use the fusion information as the soft prompt for downstream tasks in ED. Specifically, two types of modal features are first integrated through an alternating dual attention mechanism; by proposing the MPT method, multimodal information is used to classify and guide traditional ED tasks. The novel alternating dual attention has a two-wheel structure for deep interaction between text and image modalities, which can repeatedly merge useful event-related images and texts and fuse the final multimodal information. As a kind of semantic guidance for downstream tasks of ED, the provided semantic information can alleviate the problem of trigger word ambiguity.

The major contributions of this work are:

- We propose a method to introduce visual information into ED tasks, using multimodal information as soft prompt to guide downstream tasks, which is the earliest in ED tasks;
- We propose a multimodal cue-based ED learning model called multimodal injection prompt learning fine-tuning, which utilizes multimodal information as soft cue fine-tuning to optimize ED tasks;
- We evaluate the quality of the constructed language model-based image-augmented ED dataset. We conduct a series of experiments on benchmark and compare it with six state-of-the-art baseline models. The results, as well as further studies, demonstrate the effectiveness of our model.

2 Related Work

Event Detection (ED) Existing ED works mainly focus on single mode, and ED models can be divided into sequential labelling and conditional generation models.

Chen et al. (2015) first designed ED as a sequence labelling task and used CNN and RNN to model sentence-level features. Liu et al. (2018) used GCN to emphasize semantic dependence. Lin et al. (2020) constructed an end-to-end information extraction system to extract globally optimal event structures using global features and beam search. Conditional generation approaches encode sentences using generative pre-trained language models such as BART (Lewis et al., 2019) and T5 (Raffel et al., 2020). There are also sentence-level (Nguyen and Grishman, 2018) and document-level (Duan et al., 2017) event extraction tasks of different granularity in the above two paradigms.

However, events do not exist only in textual modality. More and more multidimensional modal supervision also provides a broader space for constructing downstream tasks in ED fields.

Multimodal Learning Multimodal learning aims to build models that can integrate information from different modalities, such as images, video, and audio. Recently, multimodal learning has been widely used to deal with NLP problems such as NER and machine translation (Rahman et al., 2020). These methods strengthen the understanding of short and coarse texts from the perspective of visual context and propose different modality attention to inte-

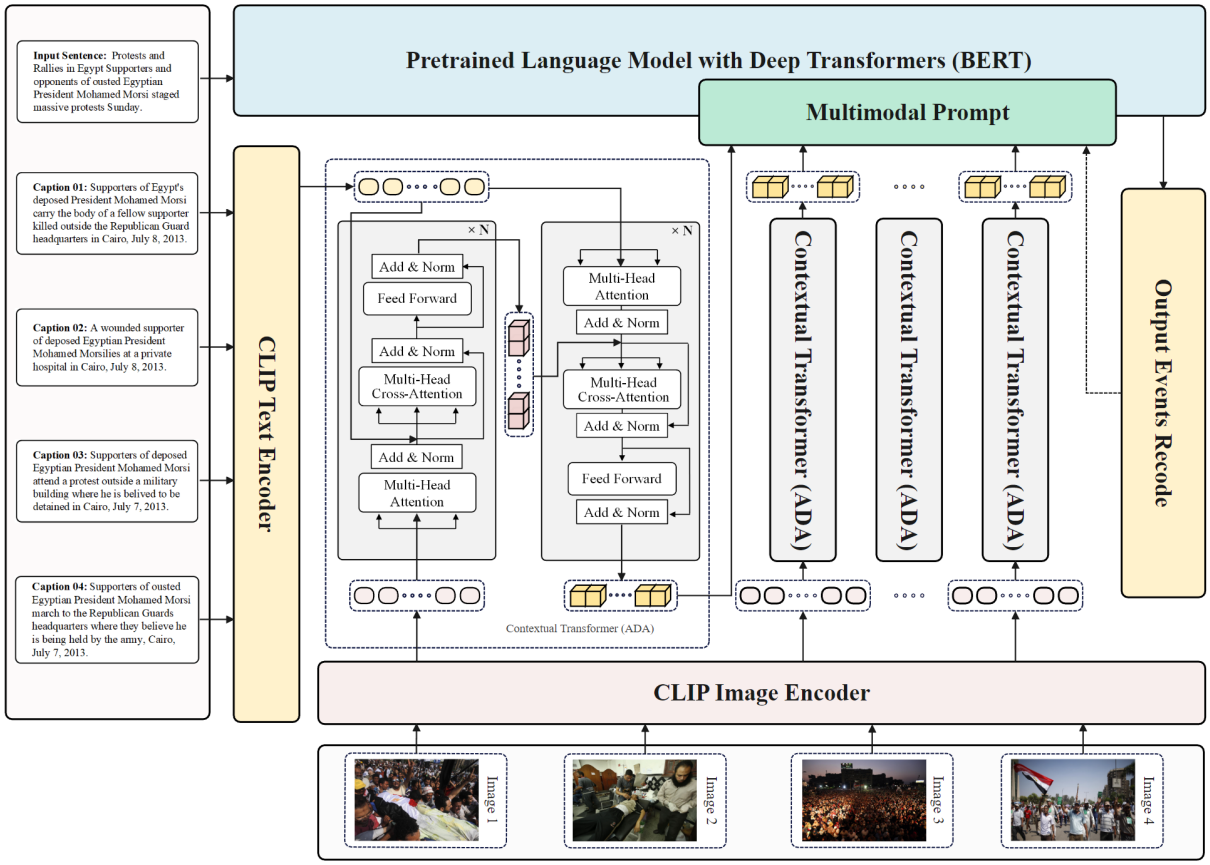

Figure 2: The overall structure of the model.

grating information from different heterogeneous sources.

Zhang et al. (2017) incorporates image modality into the ED task by visualizing entities in sentences. Tong et al. (2020) press multiple images in the disambiguation process through the attention mechanism, get the global image modality and then integrate it into the ED task. Li et al. (2020) and Li et al. (2022b), respectively, correspond to the alignment of entities and the alignment of events and their argument structures in multimodal ED tasks by constructing entity and argument relationships.

Prompt-based Learning Methods

Prompt-based learning methods use cues to guide pre-trained language models to generate results, so the quality of cue templates is crucial. Current prompt-based learning templates include: manually setting discrete prompt templates and building trainable continuous prompt templates.

Schick and Schütze (Schick and Schütze, 2020) transferred the text classification task into a cloze-filling task by using manual prompt templates. Li and Liang (Li and Liang, 2021) used trainable prefix tokens as prompts and added soft tokens in each layer of the language model.

Both methods do not introduce task-related knowledge and cannot optimize prompts and external knowledge. Until KiPT (Li et al., 2022a), a knowledge injection method was proposed to inject event-related semantic knowledge into the prompt template, WordNet, token's part of speech (POS) mechanism and other related external knowledge as the prompt learning fine-tuning strategy to optimize the prompt. Inspired by his work, we decided to inject visual information into ED tasks as external knowledge and a prompt to guide deep learning models.

3 Methodology

Figure 2 shows our Multimodal Prompt Tuning model. MPT has three components. Feature extraction first extracts text and image features from a large-scale pretrained CLIP network. Second, the multimodal ensemble enables two-round deep interaction between text and image modalities through a novel Alternating Dual Attention (ADA). Finally, event prediction uses Soft-Prompt to map the final multimodal representation to the event type semantic space to guide Bert to complete event detection.

3.1 Feature Extraction

In this section, we will elaborate on the details of the feature extraction layer. Since events exist not only in text modalities but also in image modalities, we extract features from text and image modalities. A contrastive learning multimodal pre-trained CLIP model has demonstrated the potential to learn open-set visual concepts. CLIP (Radford et al., 2021)is built with two encoders, one for images and one for text, as shown in Figure 3. The image encoder can be either ResNet or ViT (Liu et al., 2021) to convert images to feature vectors. The text encoder is a Transformer that takes as input a sequence of word tokens and again generates a vectorized representation. CLIP employs a contrastive loss during training to learn the joint embedding space of the two modalities. Specifically, for a small batch of image-text pairs, CLIP maximizes the cosine similarity of each image to the matched text while minimizing the cosine similarity to all other unmatched texts and each text similarly. Calculate the loss. After training, CLIP can be used for zero-shot image recognition. Let x be the image features generated by the image encoder, and $\{\boldsymbol{w}_i\}_{i=1}^{K}$ be a set of weight vectors generated by the text encoder, each representing a category (assuming there are K categories in total). In particular, each $\mathrm{W_i}$ is derived from a hint, such as "a photo of a class", where the "class" tag is populated with the ith class name. Then the predicted probability is:

$$p(y \mid \boldsymbol{x}) = \frac{\exp\left(\operatorname{sim}\left(\boldsymbol{x}, \boldsymbol{w}_y\right)/\tau\right)}{\sum_{i=1}^{K} \exp\left(\operatorname{sim}\left(\boldsymbol{x}, \boldsymbol{w}_i\right)/\tau\right)}$$

We adopt Cliptext as the text feature extractor. We feed the input sentence $\mathrm{S} = \{\mathrm{W}_1, \mathrm{W}_2, \ldots, \mathrm{W}_\mathrm{N}\}$ into Cliptext and use the sequence output as the sentence representation $\mathrm{H}_0 = \{\mathrm{H}_1, \mathrm{H}_2, \ldots, \mathrm{H}_\mathrm{N}\}$.

$$H_0 = \text{ Cliptext } (S)$$

ClipVision was an effective image representation () Given multiple images $\mathrm{p} = \{\mathrm{p}_1, \mathrm{p}_2, \ldots, \mathrm{p}_\mathrm{k}\}$ in a news article, we feed each image $\mathrm{p_i}$ into ClipVision and then take the last residual The block output serves as the image has hidden representation U_i.

$$u_i = \text{ClipVision } (p_i)$$

To map the image to the same latitude space as the text, we employ a sigmoid function to generate the final image representation:

$$m_i = \sigma\left(W_u u_i + b_u\right)$$

3.2 Multimodal Integration

In this section, we illustrate the steps of ADA. We first obtain an image-augmented text representation through a recursive multi-image encoder. In each step, we propose a novel Alternating Double Attention (ADA) method that first refines the image representation with textual information and then conversely performs deep interaction. We then aggregate the image-augmented text representation and Bert's raw output with a residual network to obtain the final multimodal representation.

Alternate double attention (ADA module) As shown in Figure 4, ADA has a dual structure: using text information to guide image attention and then using image information to guide text attention. Since image and text information influence each other, we adopt a binary structure. The focus area of the same image is different under different text backgrounds. Likewise, the same word can describe different events in different visuals.

Specifically, ADA is a two-round multi-head attention module. We first introduce the first round and then the second round. For the first round, the goal of ADA is to update the image representation with textual information. Formally, we use three fully connected layers to map the text representation Ht to the first two inputs of the scaled dot product attention module and the image representation Mt to the third input, denoted V, K, and Q, respectively. Then we compute the attention by querying k with q. We rescale the attention value by dividing the dimension of K to avoid vanishing gradients (citations). Next, we do a dot product of the learned attention with the third input V to obtain the weighted image representation z.

$$\begin{aligned}
\mathbf{s} &= \frac{\mathbf{q} \cdot \mathbf{k}}{\sqrt{d_k}} \\
\alpha_i &= \frac{s_i}{\sum_{i=1}^{L} s_i} \\
z &= \alpha \mathbf{v}^{\mathrm{T}}
\end{aligned}$$

We repeat the above steps u times and use linear transformation to get the final attention-corrected image representation h

$$\begin{aligned}
Z &= [z_1; z_2; \ldots; z_u] \\
h &= W_h Z + b_o
\end{aligned}$$

Finally, the query signal Q_P is directly sent to the attention-corrected output $\mathrm{H_P}$ using the residual block, and the refined representation m_t' of the image is obtained in the t-th step

$$m_t' = h + q$$

Reducing the above calculation process to Ω, the calculation process of the first round can be expressed as:

$$m_t' = \Omega\left(m_t, H_t\right)$$

ADA aims to update the text representation with image information, the operation in the middle is the same as the first round, but the input is different. We swap inputs in the scaled dot product attention module by mapping $\mathrm{H_t}$ to the third input and $\mathrm{M_t}$ to the first and second inputs. We formulate the second round process as follows:

$$H_{t+1} = \Omega\left(m_t', H_t\right)$$

3.3 Multimodal-to-language prompting

Including the description of the visual context can make the text more accurate. In this paper, we use the multimodal features obtained by fusion to guide the classification of text features. In general, we can use the cross-attention mechanism in the Transformer decoder [citation needed] to simulate the interaction between multimodal information and languages. We propose two different context-aware cueing strategies, as shown in Figure 5. One strategy we consider is pre-language model cueing, or simply pre-model cueing. We pass the features$[\overline{\mathbf{z}}, \mathbf{z}]$to the Transformer decoder to encode the visual context:

$$\mathbf{v}_{\mathrm{pre}} = \text{ Trans Decoder } (\mathbf{q}, [\mathbf{z}, \mathbf{z}])$$

Where $\mathbf{q} \in \mathbb{R}^{N \times C}$ is a set of learnable queries, and $\mathbf{v}_{\mathrm{pre}} \in \mathbb{R}^{N \times C}$ is the extracted visual context. We replace P in the formula with the visual context V to form the input of the text encoder. Since the input of the text encoder is modified, we refer to this approach as the bootstrap pre-prompt model.

The multimodal information output $\mathrm{V_{pre}}$ for each Transformer is constructed as soft prompts. Then, build a prompt template with input x, prompt(x), and the target event record y.

Template : $\mathrm{Prompt}(x)[x]$, Events: $[y]$

It is necessary to optimize $\mathrm{Prompt}(x)$ of multimodal information injection by training for the following two reasons: (1) Some rule-based algorithms are used in the construction of knowledge injection $K(x)$. However, these rules may be ineffective or even wrong in some cases, so these rules need to be softened by training; (2) Soft tokens are randomly initialized virtual tensors without original semantics. They need to be trained to approximate the distribution of actual words in order to serve as a cue to the language model. Therefore, we propose knowledge injection prompt tuning to optimize $\mathrm{Prompt}(x)$. Given a pre-trained language model and its vocabulary v, the input t of the cued template is:

$$\begin{aligned} T &= \left[H^k; H^s; e(x)\right] \\ &= \left\{h_1^k, \ldots, h_{|K|}^k, h_1^s, \ldots, h_p^s, e\left(x_1\right), \ldots, e\left(x_n\right)\right\} \end{aligned}$$

Where $e\left(x_i\right)$ represents the embedding of the input token, h_i^s and h_i^k stand for knowledge injection and embedded hints for soft tokens, respectively. Note thath_i^s and h_i^k are initialized using an embedding of the actual token from the $\mathcal{LM}$ vocabulary $\mathcal{V}$, while h_i^s represents a tensor initialized randomly. The conditional probability of the output of this event record can be obtained by generating the language model $\mathcal{LM}$. Finally, given the Golden event record y, the gradient update is performed using the following log-likelihood loss function:

$$L = -\sum_{(x,y)\in\mathcal{D}} \log\left(y \mid H^k, H^s, e(x), \theta_{\mathcal{LM}}\right)$$

where $\mathcal{D}$ stands for the whole training dataset, and $\theta_{\mathcal{LM}}$ stands for the $\mathcal{LM}$'s parameters.

4 Experiments

In this section, we evaluate the proposed dataset and approach by extensive experiments. We first give a description of dataset and hyperparameters in the experiment. We then will compare our results with several existing SOTA approaches on the same benchmarks to show the effectiveness of our image dataset and the superiority of the proposed approach. Next, we conduct experiments to answer three questions: 1) the quality of images, 2) whether to use images and 3) how to use images. Finally, we analyze when and how the images are helpful in ED by a case study.

4.1 Experiment Setup

Datasets. We employ open data set Multimedia Event Extraction (M2E2) and a new data set based

on the partition of the ScienceQA Dataset.Their statistics are shown in Table 1.

Implementation Details of MPT. Specifically, we used the CLIP pre-training model as the encoder, and we directly used ViT-B (Dong et al., 2022) as the visual encoder. For the locale prompt, we use a context length of 8. The Transformer decoder used to extract the visual context consists of 6 layers and we set the number of headers to 4. We fixed the text encoder during the training to preserve the natural language knowledge learned from the large-scale pre-training. In order to reduce computational costs, both image embedding and text embedding are projected to a lower dim(256) in front of the Transformer module. A modification was made compared to the CLIP default configuration: we used AdamW instead of the default SGD, inspired by the latest advances in Visual Transformers. We utilize the Text and Vision Transformers of "ViT-B/32" to initialize our encoders. The batch size is 128. We set the learning rate as $1e-6$ with a linearly-decaying schedule. We train 20 epochs with Adam as the optimizer, and select the best model based on the image-retrieval performance on VOANews testing dataset. The optimal transport plan is obtained within $k = 50$ iterations. To get the bounding box embeddings from CLIP visual backbone, we extract grid features and perform average pooling on the grids covered by the bounding box. For CLIP-ViT-B models, we reshape the patch representation of the final layer into grid features. For CLIP-ResNet models, we use the grid features from the last layer before the pooling. The model is trained on 4 Tesla V100 GPUs with 16GB DRAM.

Baselines. The baselines include: (1) Text-only models: We use the state-of-the-art model JMEE (Liu et al., 2018) and GAIL (Zhang et al., 2019) for comparison. We also evaluate the effectiveness of cross media joint training by including a version of our model trained only on M2E2 and ScienceQA, denoted as WASET. (2) Image-only models: Since we are the first to extract newsworthy events, and the most similar work situation recognition can not localize arguments in images, we use our model trained only on image corpus as baselines. Our visual branch has two versions, object-based and attention-based, denoted as WASEI obj and WASEI att (Li et al., 2020). (3) Multimedia models: To show the effectiveness of structured embedding, we include a baseline by removing the text and image GCNs from our model, which is denoted as Flat. The Flat baseline ignores edges and treats images and sentences as sets of vectors. We also compare to the state-of-the-art crossmedia common representation model, Contrastive Visual Semantic Embedding VSE-C (Shi et al., 2018), by training it the same way as WASE. We use micro-averaged Precision (P), Recall (R), and F1 score (F1) in all the following evaluations.

Table 1: Overall Performance on M2E2 and ScienceQA dataset (%)

DataSet	M2E2 (%)			ScienceQA (%)		
Method	P	R	F1	P	R	F1
BERT_QA	37.7	56.4	50.8	36.1	51.7	48.5
GDAP	35.8	55.3	40.6	35.4	47.9	41.1
VSE-C	33.3	48.2	39.3	34.7	42.4	33.2
Flat	34.1	56.4	42.5	36.5	53.9	43.4
WASE	43.1	59.2	49.9	40.6	55.3	39.2
CLIP-Event	41.3	**72.8**	52.7	42.8	**65.4**	46.1
MPT	**44.6**	66.3	**54.2**	**43.3**	58.7	**48.3**

4.2 Overall Results

Overall Performance.

Evaluation of Image Dataset.

In the main experiment, though, MPT was better than current methods. However, it cannot reflect the action mechanism of single mode or multi-mode in ED. Therefore, we need to find out the action mechanism of single mode and multi-mode in subsequent experiments and maximize the gain of multi-mode information for ED task.Since the image dataset is one of the most principal contribution of the paper, we evaluate the quality of images by a series of experiments. To validate the effectiveness of the image dataset, two questions need to be answered. The first is to what extent news articles are related to the images. It is necessary that images are closely related to their articles. Otherwise, the images are noises that may harm the understanding of texts. Secondly, how much extra information images can provide to the understanding of texts.

Effectiveness of Image Modality Knowledge .

In the section, we discuss how much improvement the image modality brings to Event Detection. We employ the same models with and without image modality on M2E2 dataset. Different from the overall part, we employ three different base encoders, including RNN and BERT, in order to show the improvement of image modality on ED is omnipresent rather than a model-related situation. We

Table 2: The evaluation of image modality.

DataSet	M2E2 (%)			ScienceQA (%)		
Method	P	R	F1	P	R	F1
LSTM	35.7	44.6	40.1	33.2	41.4	38.2
LSTM+Image	37.3	47.3	43.9	33.7	41.7	40.8
improvement	**+1.6**	**+2.7**	**+3.8**	**+0.5**	**+0.3**	**+2.6**
BERT	36.1	56.4	42.5	36.5	53.9	43.5
BERT+Image	39.4	59.2	45.9	37.6	55.3	45.3
improvement	**+3.3**	**+2.8**	**+3.4**	**+1.1**	**-0.6**	**+1.8**

Table 3: Effectiveness of multimodal fusion in MPT.

DataSet	M2E2 (%)			ScienceQA (%)		
Method	P	R	F1	P	R	F1
Connection+Prefix	33.1	49.8	40.5	35.2	46.3	34.2
Connection+Prompt	32.7	50.3	42.4	36.3	47.6	36.5
ADA+Prefix	42.4	63.2	51.9	40.8	58.0	45.7
Co-Attention+Prefix	41.2	61.5	52.1	39.3	60.7	43.6
Co-Attention+Prmopt	42.3	64.8	45.9	41.9	**62.5**	46.1
MPT(ADA+Prompt)	**44.6**	**66.3**	**54.2**	**43.3**	58.7	**48.3**

employ two layers of Bi-LSTM model with hidden units 384 for each direction. As shown in Table 3, the incorporation of image modality improves the performance of Event Detection on Precision, Recall and F score independent to specific models. The three models are the most commonly used neural network models, so the results validate the significance of image modality in Event Detection. Note that the improvement for CNN and LSTM encoders is obviously bigger than that on BERT, which reflects the complementary role of images in Event Detection. When the capacity of text encoder is small, images can bring in larger improvement.

Effectiveness of Multimodal Fusion and Prompt

It is not difficult to see from the ablation experiment that direct connection of multimodal information does not necessarily disambiguation ED tasks, because although multimodal information is mapped into a multimodal semantic space by CLIP, direct cascading will add additional redundant information. The effect of joint attention is not as good as the ADA cross-attention mechanism proposed by us, because in the process of semantic coupling decoupling, ADA can carry out additional subration of noise information, which makes the semantic co-reference stronger and suitable for eliminating the ambiguity of ED. Prompt's approach is slightly better than the Prefix approach, because in MPT, we cascade the multimodal information with the natural language, and the semantic gap is smaller than the directly generated multimodal information, which is more suitable for downstream ED tasks. Prefix provides semantic coreference, but it provides a smaller number of variable parameters, and the semantic gap is larger than the direct one.

5 Conclusions

This paper proposes an ED method based on multimodal prompt guidance. Adjust the learning method by introducing a piece of multimodal information promptly. Multimodal fusion information and soft tokens are used to construct final multimodal hints that can be optimized through training. Comprehensive experiments show that the proposed method is superior to the current prompt-based ED model and has a strong baseline. The prompt-based model can introduce task-related multimodal knowledge more conveniently and efficiently through our approach. In the future, we will explore more multimodal cue mechanisms and their application in other tasks.

Acknowledgement

We thank the anonymous reviewers helpful suggestions.

References

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Shuyang Gu, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. 2022. Clip itself is a strong fine-tuner: Achieving 85.7% and 88.0% top-1 accuracy with vit-b and vit-l on imagenet. *arXiv preprint arXiv:2212.06138*.

Shaoyang Duan, Ruifang He, and Wenli Zhao. 2017. Exploiting document level information to improve

event detection via recurrent neural networks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 352–361.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Haochen Li, Tong Mo, Hongcheng Fan, Jingkun Wang, Jiaxi Wang, Fuhao Zhang, and Weiping Li. 2022a. Kipt: Knowledge-injected prompt tuning for event detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1943–1952.

Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. 2022b. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16420–16429.

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020. Cross-media structured common space for multimedia event extraction. *arXiv preprint arXiv:2005.02472*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009.

Xiao Liu, Zhunchen Luo, and Heyan Huang. 2018. Jointly multiple events extraction via attention-based graph information aggregation. *arXiv preprint arXiv:1809.09078*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Thien Nguyen and Ralph Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Haoyue Shi, Jiayuan Mao, Tete Xiao, Yuning Jiang, and Jian Sun. 2018. Learning visually-grounded semantics from contrastive adversarial samples. *arXiv preprint arXiv:1806.10348*.

Meihan Tong, Shuai Wang, Yixin Cao, Bin Xu, Juanzi Li, Lei Hou, and Tat-Seng Chua. 2020. Image enhanced event detection in news articles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9040–9047.

Tongtao Zhang, Heng Ji, and Avirup Sil. 2019. Joint entity and event extraction with generative adversarial imitation learning. *Data Intelligence*, 1(2):99–120.

Tongtao Zhang, Spencer Whitehead, Hanwang Zhang, Hongzhi Li, Joseph Ellis, Lifu Huang, Wei Liu, Heng Ji, and Shih-Fu Chang. 2017. Improving event extraction via multimodal integration. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 270–278.