Multimodal Chinese Event Detection on Vision-Language Pre-training and Glyphs

Qianqian Si, Zhongqing Wang, Peifeng Li and Qiaoming Zhu

School of Computer Science and Technology, Soochow University, Suzhou, China
20215227051@stu.suda.edu.cn, {wangzq, pfli, qmzhu}@suda.edu.cn

Abstract. When using visual information to complement textual data for event extraction, current approaches primarily focus on processing text and images independently using different pre-trained models and then fusing the feature information from different modalities. How- ever, pre-training and fine-tuning schemes have been extended to the joint domain of vision and language, leading to the development of vision-language pre-trained models (VLPs). These models are extensively trained on text and its corresponding images and then fine-tuned for vision-language tasks. In this paper, we propose a method for event detection in Chinese glyphs and VLP models. Since Chinese characters are hieroglyphs, some radical features of the trigger words play a certain and auxiliary role in the detection of text trigger words. We convert the text in the ACE Chinese corpus into text images, and transport the text and images into the Vision-Language model to obtain multimodal features for event detection. Experimental results on the ACE 2005 Chinese corpus show that our proposed model outperforms the SOTA baselines

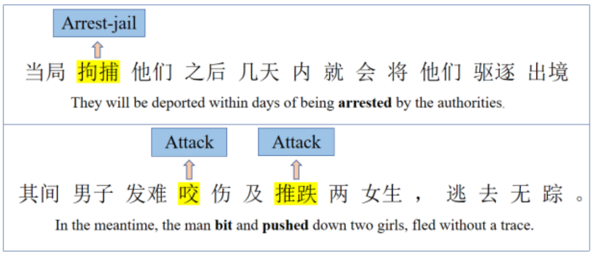
Keywords: VLP, Chinese glyphs, event detection

1 Introduction

Event extraction is a crucial area within the field of natural language processing, carrying significant research importance. The fundamental benefit of event extraction techniques is their ability to transform semi-structured and unstructured data into structured event descriptions, which can then facilitate the development of advanced downstream applications. Event extraction is typically divided into two sub-tasks: event detection and argument extraction. Event detection aims to recognize specific types of event triggers and is a crucial step in event extraction. For example, the event detection system should be able to detect the “Attack” event triggered by “bite” and the “Attack” event triggered by “push down”.

John *(Artifact)* **came back** *(Transport)* from *Cuba* *(Origin)*.

In this example, the task of event detection (i.e., trigger detection) involves identifying “came back” as a trigger mention and assigning the event type "Move- ment" to this identified trigger mention. The task of argument extraction involves identifying the entities “John” and “Cuba” as the arguments of this Movement event mention and assigning the roles “Artifact” and “Origin” to them, respectively.



当局 拘捕 他们 之后 几天 内 就 会 将 他们 驱逐 出境

They will be deported within days of being **arrested** by the authorities.

其间 男子 发难 咬 伤 及 推跌 两 女生 ， 逃 去 无 踪 。

In the meantime, the man **bit** and **pushed** down two girls, fled without a trace.

Fig. 1. Example of triggers with similar radicals

This paper focuses on Chinese event extraction. A Chinese sentence typically consists of multiple Chinese characters. Chinese characters are pictographs, and the earliest Chinese oracle bones were derived from the shapes of objects to represent their meanings. Commonly, a Chinese character consists of semantic- phonetic compounds and pictophonetic compounds. The former conveys the general meaning of the compound character, while the latter indicates the pronunciation of the compound character. In most cases, the semantic indicator refers to the radical under which the character is listed in dictionaries. Therefore, each Chinese character is an attempt to combine sound, image, and idea in a mutually reinforcing manner. These character features can sometimes be used as criteria for classifying the attributes of Chinese characters. For example, many verbs will involve words related to human body parts because most actions are performed by humans. These features are integrated into the creation of Chinese characters.

In light of this, we propose that the radical features of trigger words can also serve as supplementary features for trigger word recognition. Inspired by the characteristics of Chinese characters, we found that many trigger words in a specific event type share relatively similar radical features. In view of this, the radical features of trigger words may be used as auxiliary features to aid in detecting trigger words in the text. For example, as shown in Fig. 1, the event trigger "推跌" (push) contains the radical of "扌" (hand), and the event trigger "拘捕" (arrest) contains the radical of "扌" (hand). These two event triggers have the same radical and refer to the same type of event. We hope that these glyph features will be helpful for recognizing trigger words. Unlike the time-consuming and challenging process of searching for high-quality images for text, creating glyph-based images for Chinese text (e.g., screenshots) is easy. Therefore, it is easy and fast to create a multimodal event dataset of texts and their glyph-based images, derived from existing text-style event datasets.

In this paper, we propose a multimodal event detection method VLGMM, which based on Chinese glyphs and a Vision-Language model. Specifically, we use various fonts to convert text into images, and then extract the image features through a bridge framework that establishes a connection between the top layers of unimodal encoders and each layer of the cross-modal encoder. This facilitates the efficient alignment and fusion of visual and textual representations at various semantic levels of pre-trained unimodal encoders in the cross-modal encoder. Finally, we fuse them with the textual features acquired through the dynamic multi-pooling mechanism to form the feature vectors for multimodal event detection. Experimental results on the ACE 2005 Chinese corpus demonstrate that our proposed model outperforms the baselines. In summary, the contribution of this paper lies in the following two points:

- We propose a multimodal event detection method based on Chinese text and the glyph of Chinese characters.
- We apply the Vision-Language model to the task of event detection.

2 Related Work

2.1 Text-based Event Detection

Models based on neural networks have a strong ability to learn feature representation and have a broad range of applications. They can automatically extract event features from natural language without complicated feature engineering or extensive manual intervention.

In Chinese event detection, Zeng et al. [1] used a bidirectional RNN to extract sentence features and an RNN to extract lexical features. This approach helps mitigate the impact of Chinese word segmentation errors and greatly improves the performance of Chinese event detection. Wu et al. [2] proposed a neural network model based on the attention mechanism and semantic features. The model generated word vectors by combining word vector information and the attention mechanism.

In English event detection, Chen et al. [3] proposed a Dynamic Multi-Pooling Convolutional Neural Network (DMCNN), which can retain more valuable in- formation by retaining the maximum pooling value. Nguyen et al. [4] performed joint event extraction using RNNs. The emergence of pre-trained language models, their ability to express semantic information has attracted the attention of researchers in the field of event extraction. Yang et al [5] proposed Pre-trained Language Models for Event Extraction (PLMEE), which applied pre-trained language models to directly capture word features directly and achieved a large performance gain.

2.2 Multimodal Event Detection

Relying solely on text-based information for event detection will result in the loss of valuable visual information contained in images, which is crucial for event detection. Images may contain relevant information that is not mentioned in the text.

However, the lack of multimodal event datasets hinders the development of multimodal event detection. Wang et al. introduced a new dataset for multimodal event detection (MEED) [6] to address the existing gaps. The dataset defines event types and parameter roles for multimodal data and utilizes controlled text generation to produce text modalities based on visual event extraction datasets. Tong et al. [7] formed a multimodal dataset using the ACE dataset, searching for images with high similarity to each news item in the ACE dataset on various news websites to form the corresponding picture dataset for each sentence in ACE. Li et al. [8] proposed a weakly aligned structured representation method WASE based on the text dataset ACE and image dataset imSitu. This method enables the information of the two modalities to be represented in the same multimodal semantic space by converting the given sentences and pictures into a graph structure.

Considering that most of the trigger words are verbs or nouns, which have relatively similar radicals in Chinese, we propose a method to convert each sentence in the ACE Chinese corpus into a text picture. Therefore, a multimodal dataset with a strict correspondence between text and pictures can be easily formed. In current research on multimodal event extraction, most researchers use separate models to process text and images to obtain feature representations before fusion, which ignores the alignment and fusion between different modalities, resulting in losing a lot of key information. In this paper, we use an efficient vision-language pre-trained model to obtain information fusion between text and images for event detection.

2.3 Multimodal Event Detection

Following the taxonomy proposed by ViLT [9], most Vision-Language models are TWO-TOWER architecture. They feed last-layer representations of pre- trained unimodal encoders into the top cross-modal encoder and can be differentiated by the depth of the textual, visual, and cross-modal encoders. CLIP [10] and ALIGN [11] are representative models that directly perform a shallow fusion (e.g., dot product) of last-layer representations of equally expressive pre-trained unimodal encoders in the cross-modal encoder. The remaining models perform deep fusion in the multi-layer transformer-based cross-modal encoder but choose pre-trained unimodal encoders with varying levels of expressiveness. Numerous works like VisualBert [12] and VL-BERT [13] adopt various types of deep vision models (e.g., Faster R-CNN [14], ResNet [15] or ViT) as their visual encoder to obtain region, grid, or patch features, and concatenate them with word embedding to feed into their top cross-modal encoder. Unlike the previous model, some researchers have proposed BRIDGETOWER[16] which build multiple bridge layers to connect the top layers of unimodal encoders for cross-modal fusion. This does not affect the interaction in the unimodal encoders and enables different semantic levels of visual and textual representations to interact thoroughly and mildly at cross-modal encoder. This allows for effective information fusion between text and images during the encoding process.

3 Methodology

3.1 Task Definition

The task definition of event detection in this paper is following that of the ACE event extraction task. Formally, let the multimodal dataset represent as $D = \{S_1, S_2, \dots, S_n\}$, where n is the number of event samples， S_i is the i-th sample. Let $S_i = \{text, img\}$, where $text$ represents the text and img represents the image. We need to learn a model $f: S_i \rightarrow Y(S_i \in D)$ to classify each sample into the predefined categories $Y = \{0,1\}$, which is the ground-truth label of the sample S_i (0 denotes non-trigger and 1 denotes trigger).

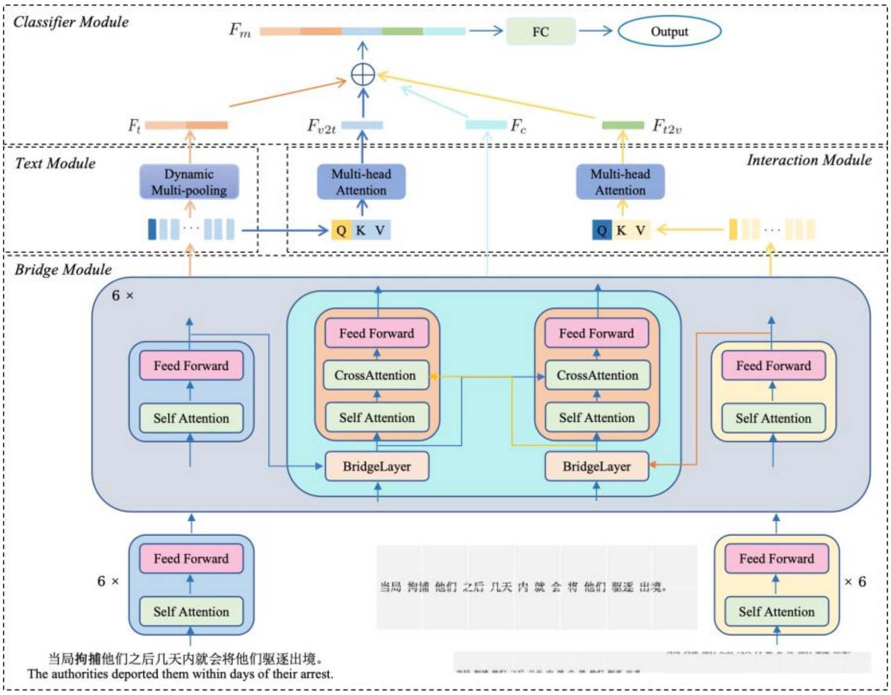


Fig. 2. Structure of the VLGMM model.

3.2 Bridge Module

We use bert-base-chinese as text encoder, Vision Transformer as visual encoder and a cross encoder to fuse the two modalities. Hendricks et al. [19] examined various attention mechanisms used in current Transformers-based cross-modal encoders and found that the co-attention mechanism performed exceptionally well. This mechanism utilizes distinct parameters for each modality. As an example, in the textual component of the cross-modal encoder, the queries for each MSA block originate from the textual modality, while the keys and values stem from the visual modality. Consequently, the model adheres to this co-attention approach. We categorize the layer l_{th} cross-modal encoder as encompassing both a visual and textual section. Each section consists of an MSA block, a Multi-head Cross-Attention (MCA) module, and an FFN module. The model defines the interaction between layers as follows.

$$\widetilde{Z_l^T} = Z_{l-1}^T \tag{3}$$

$$\widetilde{Z_l^V} = Z_{l-1}^V \tag{4}$$

$$Z_l^V, Z_l^T = Encoder_l^Z\left(\widetilde{Z_l^V}, \widetilde{Z_l^V}\right), l = 1, \dots L_Z \tag{5}$$

The l^{th} layer of the cross-modal encoder produces output representations for both the visual and textual components, denoted as Z_l^V and Z_l^T, Meanwhile, the input to each part is represented by $\widetilde{Z_l^V}, \widetilde{Z_l^T}$. The cross-modal encoder comprises L_Z layers. In this study, we introduce multiple bridge layers to establish connections between the topmost layers of unimodal encoders and each layer of the cross-modal encoder.

$$\widetilde{\mathrm{Z_l^V}} = \mathrm{BridgeLayer_l^V}\left(\mathrm{Z_{l-1}^V}, \mathrm{V_k W_V} + \mathrm{V^{type}}\right) \tag{6}$$

$$\widetilde{\mathrm{Z_l^T}} = \mathrm{BridgeLayer_l^T}\left(\mathrm{Z_{l-1}^T}, \mathrm{T_k W_T} + \mathrm{T^{type}}\right) \tag{7}$$

where k is the index of layer representations of unimodal encoders. In this paper, $L_V = L_T = 12, L_Z = 6$. We incorporate the representations from the top 6 layers of unimodal encoders, specifically for k ranging from 7 to 12. The definition of bridge layer adopts the most straightforward approach.

$$BridgeLayer(x, y) = LayerNorm(x + y) \tag{8}$$

For each text and image pair, after being processed by the bridge module, we denote the text features output by the text encoder as f_{text}, the image features output by the visual encoder as f_{image}, and the output of the cross-modal encoder as F_c. $\mathrm{f_{text}}$ and f_{image} will be used to obtain external interaction features based on the attention mechanism in the Interaction Module.

3.3 Text Module

The process of extracting text features is represented by the dotted box in the upper-left part of Fig**.** 2. We use BERT to encode the representation of the input sentence. We denote the sentence as $X = (x_1, x_2, \dots, x_t, \dots, x_n)$, where $x_i (1 \leq i \leq n)$ represents the word vector representation of the i_{th} token in the sentence, and x_t represents the word vector representation of the token that needs to be judged as the trigger word. Here, $X \in R^{n*h}, h = 768$. Then the words of the sentence are arranged based on the position of the current candidate trigger word in the sentence.

The vector indicates that X is divided into two parts, $X_{1:t-1}$ and $X_{t:n}$, and then processed by the maximum pooling mechanism to obtain the two components $f_{1:t-1}$ and $f_{t:n}$ of the text feature as follows.

$$f_{1:t-1} = maxpooling(X_{1:t-1}) \tag{9}$$

$$f_{t:n} = maxpooling(X_{t:n}) \tag{10}$$

The above two parts are directly concatenated to obtain the text feature representation F_t as follows, where.$\mathrm{f_{1:t-1}} \in \mathrm{R^h}, \mathrm{F_t} \in \mathrm{R^{2*h}}, \mathrm{h} = 768$

$$F_t = concat([f_{1:t-1}, f_{t:n}]) \tag{11}$$

3.4 Interaction Module

The purpose of this module is to extract the interaction features and F_{t2v} and F_{v2t}. The steps are as follows: For each text-and-image pair, the textual content is typically represented in the picture. The shared information of the graphic-text is more prominent, and this information usually constitutes the core content of the event, requiring our focused attention. We extract this information using the attention mechanism. f_{text} obtained by the bridge module is denoted as $(x_0, x_1, x_2, \dots, x_{M-1}, x_M, x_{M+1})$. The first and last parts of the vector representation (CLS and SEP) to obtain the text sequence vector representation $X = (x_1, x_2, \dots, x_{M-1}, x_M)$, where x_0 represents the textual semantics. The obtained f_{image} representation is denoted as $(m_0, m_1, m_2, \dots, m_{N-1}, m_N)$, remove the identifier $[class]$ to get the image sequence vector representation $M = (m_1, m_2, \dots, m_{N-1}, m_N)$, m_0 represents the text semantics. The text semantic vector x_0 is utilized as the query vector, while the vector representation Mof the picture sequence serves as the key vector and the value vector. The text-guided image feature F_{t2v} is obtained through the multi-head attention mechanism. Similarly, the image vector representation m_0 is used as the query vector, and the vector representation X of the text sequence is used as the key vector and the value vector to obtain the image-guided text feature F_{v2t} through the multi-head attention mechanism. The formula is as follows.

$$F_{t2v} = MHATT(x_0, M, M) \tag{12}$$

$$F_{v2t} = MHATT(m_0, X, X) \tag{13}$$

where $MHATT$ represents multi-head attention mechanism, $x_0, m_0 \in R^h, X \in R^{M*h}, X \in R^{N*h}, F_{v2t}, F_{v2t} \in R^h, h = 768.$

3.5 Classifier Module

The text feature F_t, the interaction features F_{v2t} and F_{t2v} and the multimodal feature F_c are concatenated into a new multimodal feature F_m as follows.

$$F_m = concat([F_t, F_{v2t}, F_{t2v}, F_c]) \tag{14}$$

To calculate the confidence level of each candidate trigger word, the resulting multimodal feature vector is classified by Softmax as follows.

$$y_i = softmax(W_s F_m + b_s) \tag{15}$$

where W_s is the learnable parameter and b_s is bias unit, in order to prevent the occurrence of over-fitting, we use the dropout operation before the fully connected layer, in addition, we use the cross-entropy loss function as the loss function. where $\widehat{y_i}$ is the true value of the class of the sample i, y_i is the forecast value.

$$L = -\sum_i \widehat{y_i} log(y_i) \tag{16}$$

4 Experimentation and Analysis

4.1 Datasets and Measure Metrics

In this paper, we use the ACE2005 Chinese dataset and on this basis, text-based images are generated to form the multimodal dataset required for our experiments. In the ACE

Table 1. Statistics of the dataset.

Category	Train	Dev	Test
Number of sentences	5670	359	665
Number of events	2776	191	365

2005 Chinese corpus, the trigger words are mostly verb or noun. ACE annotates 8 types and 33 sub-types (e.g., Attack, Die, and Start-Position) for event mentions that also correspond to the types and sub-types of the event triggers. There are 6694 sentences in ACE Chinese corpus, in which some sentences contain one or more triggers while others have not a trigger word in them. The training set, validation set and test set are shown in Table 1. We report the micro-average Precision (P), Recall (R) and F1-score (F1), following the standards defined in [18]. A trigger is correctly identified if its position in the document matches a reference trigger and an event type is correctly determined if the trigger's event type and position in the document match a reference trigger.

In this paper, we set the learning rate to 5e-6, the maximum length of sentence is 128, and the heads of multi-head attention is set to 8.

4.2 Results and Analysis

To study the effects of our proposed model on multimodal event detection, we compared it with the following strong baselines. The results of comparison with the experimental results are shown in Table 2.

- **DRMM** [7]: The model treats the event detection task as a sequence annotation task. using an alternate dual attention mechanism to enable textual and image representations to complement each other, aggregating features from both modalities.
- $\boldsymbol{Concat}_{\mathbf{text-image}}$: Using ResNet for extracting image features and DMBERT for extracting text features, and then fuses them for multimodal event extraction.
- **ViLT** [9]: A low parameter count, fast training, transformer-based implementation of a visual language pre-training model.
- **CLIP** [19]: A visual language pre-training model based on contrastive learning,
- **FLAVA** [20]: Using the hidden state of the unimodal encoder, a multimodal encoder is designed for modal fusion.
- **VLGMM**$_{\textit{-interaction}}$**:** By removing the interaction module to demonstrate the impact of information interaction at the highest level of unimodal coding on multimodal event detection.
- **VLGMM**$_{\textit{-bridge}}$: By removing the bridge module, we aim to illustrate the impact of information interaction before the unimodal coding at the top layer on multimodal event detection.
- **VLGMM**$_{\textit{-interaction-bridge}}$: This indicates that the model relies solely on textual modalities for event extraction and utilizes the sentence-level feature representation of DMBERT [21].

Table 2. Comparison of experimental results. We used the t-test with a 95% confidence interval for the significance test and all improvements of our model over $\text{Concat}_{text-image}$ are significant ($p < 0.01$).

Model	P	R	F1
CAEE	84.0	68.2	75.3
DRMM	74.8	80.4	77.4
ViLT	78.4	77.5	77.9
CLIP	80.0	76.6	78.3
FLAVA	82.1	77.4	79.6
$\text{Concat}_{text\text{-}image}$	81.3	77.5	79.3
$\text{VLGMM}_{\text{-}interaction}$	81.5	78.0	79.7
$\text{VLGMM}_{\text{-}bridge}$	82.7	76.8	79.6
$\text{VLGMM}_{\text{-}interaction\text{-}bridge}$	79.0	75.3	77.1
VLGMM	**81.7**	**78.6**	**80.1**

Table 3. Experimental effect of different font types.

Fonts	P	R	F1
Original Song	81.7	78.5	80.1
Cursive	80.3	76.9	78.6
Clerical	81.6	76.5	79.0
Running	79.9	77.3	78.5
Traditional	82.6	77.1	79.8

Compared to the DMBERT, VLGMM proposed in this chapter improves the precision, recall, and F1 score by 2.7, 3.2, and 3.0, respectively. This result demonstrates the effectiveness of the image modality based on Chinese glyphs. $Concat_{text-image}$ is our previous work, which obtained multimodal features for event detection with a simple fusion of text features extracted by DMBERT and image features extracted by ResNet. Compared with $Concat_{text-image}$, VLGMM improves the precision, recall, and overall F1 score by 0.4, 1.1, and 0.8, respectively. This demonstrates the effectiveness of the fusion method proposed in this chapter.

It is noteworthy that VLGMM achieves the best performance on two classification tasks compared to the current state-of-the-art visual language pre-training models. All three models ViLT, CLIP and FLAVA, belong to the type of bimodal structure summarized in the previous sections and all utilize the last layer of the output state of a unimodal encoder to fuse the information between the two modalities. Among them, VLGMM improves the F1 value by 2.2 on the event detection task compared to ViLT. VLGMM enhances the F1 value by 1.8 on the task in this paper compared to the CLIP model, which is based on comparative learning and achieves multimodal representation by maximizing the cosine similarity of the positive samples. VLGMM boosts the F1 value by 1.8 on this paper's task compared to the unimodal image and textual representations using the multimodal Transformer encoder with projection. In comparison to the FLAVA model, which employs cross-attention to merge two modalities on the projected single-model image and text representation using the multimodal Transformer encoder, VLGMM enhances the F1 value by 0.5 on the task in this paper. These aforementioned multimodal models all adhere to the two-tower structural model, where the fusion of the two modalities occurs solely at the final layer of the single-modal encoder. While VLGMM model fuses the features of different coding layers through the cross-modal encoder before the last layer of the unimodal encoder, which can efficiently align and fuse the visual and textual representations of different semantic levels in the cross-modal encoder. In addition, the VLGMM model is designed with an interaction module that enables text-only features and image-only features, output from the unimodal encoder, to interact at the last layer. This preserves the original unimodal features without noise interference. Comparison results with these state-of-the-art visual language pretraining models show that the method proposed in this paper captures important information in text and images more efficiently, resulting in a superior joint multimodal representation.

In order to demonstrate the effectiveness of the Interaction layer and Bridge layer, we conducted a set of ablation experiments, the results of which are shown in the Table 2 for VLGMM*-interaction* and VLGMM*-bridge* If we rely solely on the bridge layer or the interaction layer, the performance is not as good as when the two are combined.

We also separately experimented with five different fonts. The experimental results for the different fonts are shown in Table 3, experimental results on various fonts demonstrated the effectiveness of glyph information for event detection.

4.3 Ablation Study

Since both unimodal encoders possess 12 layers, the number of cross-modal layers can vary from 1 to 12. illustrates the outcomes of employing varying cross-modal layer counts in the Bridge Module. Notably, we observed that increasing the number of cross-modal layers does not uniformly enhance performance, potentially due to:

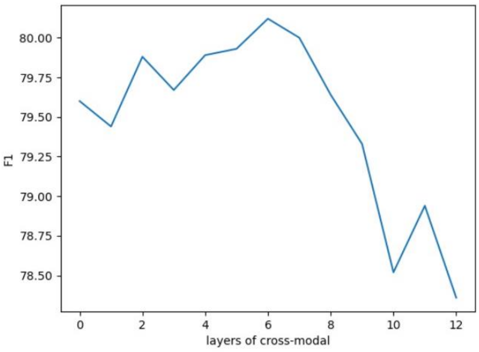


Fig. 3. Analysis on different layers of cross-model layer

- more cross-modal layers are more difficult to train and are more data-hungry.
- unimodal representations of top layers are beneficial to cross-modal alignment and fusion, while unimodal representations of bottom layers may be less useful and even detrimental.

From the table, it can be seen that the best results were achieved using 6 layers of cross-modal connections, and 8 results were better than those without cross-modal connections VLGMM$_{\text{-}bridge}$. It further illustrates that the bridge layers can facilitate effective alignment and fusion between different semantic levels in the cross-modal encoder, integrating unimodal representations. Conclusion

5 Conclusion

In this paper, we propose a Chinese event detection model that combines cross- modal pre-training with glyph features. The model not only performs guided interaction of inter-modal information through an attention mechanism, allowing the model to focus on the more salient information shared by the graphic, but also connects the top layer of the unimodal encoder through a bridging layer. Experiments on the ACE corpus validate the effectiveness of the model proposed in this chapter. Our future work will focus on introducing different types of glyphs to represent trigger and boost event extraction

Acknowledgments. The authors would like to thank the two anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China (Nos. 62276177 and 62376181), and Key R&D Plan of Jiangsu Province (BE2021048).

References

1. Ying Zeng et al. “A Convolution BiLSTM Neural Network Model for Chinese Event Extraction”. In: *Proceedings of the 5th CCF Conference on Natural Language Processing and Chinese Computing, and 24th International Conference on Computer Processing of Oriental Languages,* Proceedings 24. 2016, pp. 275–287.
2. Yue Wu and Junyi Zhang. “Chinese Event Extraction Based on Attention and Semantic Features: A Bidirectional Circular Neural Network”. In: Future Internet 10.10 (2018), p. 95.
3. Yubo Chen et al. “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*, 2015, pp. 167–176.
4. Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. “Joint Event Extraction via Recurrent Neural Networks”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: human language technologies.* 2016, pp. 300–309.
5. Sen Yang et al. “Exploring Pre-trained Language Models for Event Extraction and Generation”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 5284–5294.

6. Shuo Wang et al. “MEED: A Multimodal Event Extraction Dataset”. In: *Proceedings of Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction*, 2021, pp. 288–294.
7. Meihan Tong et al. “Image Enhanced Event Detection in News Articles”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020, pp. 9040–9047.
8. Manling Li et al. “Cross-media Structured Common Space for Multimedia Event Extraction”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 2557–2568.
9. Wonjae Kim, Bokyung Son, and Ildoo Kim. “ViLT: Vision-and-language Transformer without Convolution or Region Supervision”. In: *Proceedings of International Conference on Machine Learning*. 2021, pp. 5583–5594.
10. Alec Radford et al. “Learning Transferable Visual Models From Natural Language Supervision”. In: *Proceedings of the International conference on machine learning*. 2021, pp. 8748–8763.
11. Chao Jia et al. “Scaling up Visual and Vision-language Representation Learning with Noisy Text Supervision”. In: *Proceedings of International Conference on Machine Learning*. 2021, pp. 4904–4916.
12. Liunian Harold Li et al. “VisualBERT: A Simple and Performant Baseline for Vision and Language”. In: *CoRR* abs/1908.03557 (2019).
13. Weijie Su et al. “VL-BERT: Pre-training of Generic Visual-Linguistic Representations”. In: *Proceedings of 8th International Conference on Learning Representations.* 2020.
14. Ross B. Girshick. “Fast R-CNN”. In: *Proceedings of 2015 IEEE International Conference on Computer Vision*. 2015, pp. 1440–1448.
15. Brett Koonce and Brett Koonce. “ResNet 50”. In: *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization* (2021), pp. 63–72.
16. Xiao Xu et al. “Bridgetower: Building Bridges between Encoders in Vision- language Representation Learning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2023, pp. 10637–10647.
17. Lisa Anne Hendricks et al. “Decoupling the Role of Data, Attention, and Losses in Multimodal Transformers”. In: *Transactions of the Association for Computational Linguistics* 9 2021, pp. 570–585.
18. Pei-Feng Li and Guo-Dong Zhou. “Three-Layer Joint Modeling of Chinese Trigger Extraction with Constraints on Trigger and Argument Semantics”. In: *Journal of Computer Science and Technology* 32 (2017), pp. 1044–1056.
19. An Yang et al. “Chinese CLIP: Contrastive Vision-Language Pretraining in Chinese”. In: *CoRR* abs/2211.01335 (2022).
20. Amanpreet Singh et al. “Flava: A Foundational Language and Vision Alignment Model”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 15638–15650.
21. Xiaozhi Wang et al. “Adversarial Training for Weakly Supervised Event Detection”. 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)*. 2019, pp. 998– 1008.