Multimedia Event Extraction with LLM Knowledge Editing

Jiaao Yu, Yijing Lin, Zhipeng Gao, Xuesong Qiu, Lanlan Rui*
State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications
{yujiaao, yjlin, gaozhipeng, xsqiu, llrui}@bupt.edu.cn.

Abstract

Multimodal event extraction task aims to identify event types and arguments from visual and textual representations related to events. Due to the high cost of multimedia training data, previous methods mainly focused on weakly alignment of excellent unimodal encoders. However, they ignore the conflict between event understanding and image recognition, resulting in redundant feature perception affecting the understanding of multimodal events. In this paper, we propose a multimodal event extraction strategy with a multi-level redundant feature selection mechanism, which enhances the event understanding ability of multimodal large language models by leveraging knowledge editing techniques, and requires no additional parameter optimization work. Extensive experiments show that our method outperforms the state-of-the-art (SOTA) baselines on the M2E2 benchmark. Compared with the highest baseline, we achieve a 34% improvement of Precision on event extraction and a 11% improvement of F1 on argument extraction.

1 Introduction

In the real world, the representation of events frequently encompasses a multitude of expression modalities such as images and text. This catalyzes the evolution of Multimedia Event Extraction (MEE) task, which aims to extract the event type and argument information from the event's text description and associated image.

Due to the high cost of building multimedia data sets, existing multimedia event extraction dataset M2E2 only provides data for testing, which restricts the fine-tuning on multimodal large language model (MLLM) in multimedia event extraction task. Recent MEE methods mainly focus on the weak alignment of features obtained from well-pretrained unimodal encoders. According to their cross-modal alignment strategy, they can be divided into graph-based methods, which align the visual objects and textual entities with a knowledge graph structure (Li et al., 2020, 2022; Liu et al., 2024); template-based methods, which align candidate items from different modalities with the template uniformly (Seeberger et al., 2024); and fine-tune-based methods, which align the features under different modal by parameter-efficient fine-tuning on multimodal large language models (Du et al., 2023; Sun et al., 2024).

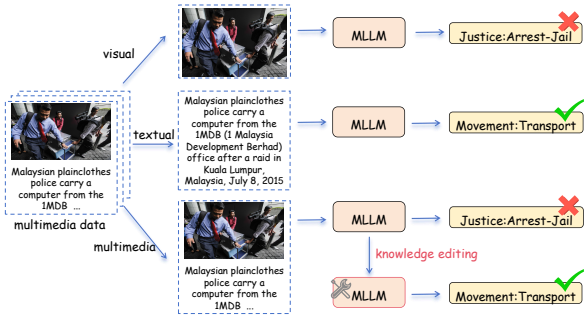

Figure 1: Example of multimedia data on M2E2, the overunderstanding of visual features led to the incorrect prediction of MLLM on image-only task, which further effect the prediction on multimedia task.

While these methods have demonstrated favorable performance, they overlook the inherent disparities in visual encoding requirements between the image recognition task and the event extraction task. Notably, a fundamental trade-off exists between these two tasks: the image recognition task desires to perceive the image features as much as possible, while the event extraction task prefers to focus on those important features and neglect redundant content. Due to the fact that the visual encoder utilized by previous methods are trained from image recognition tasks, there may exist redundant feature perception capabilities which affect event understanding of MLLMs. Taking Fig. 1 as an example, the image data from M2E2 are orig-

*Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 4116–4124
November 4-9, 2025 ©2025 Association for Computational Linguistics

inally intended to show a person running away with a computer. However, due to the forward-leaning posture of the body which resembles being detained, it is wrongly predicted as an arrest when only inputting visual features, which further misleads the judgment of multimodal large language models.

To address this problem, we propose a multimedia event extraction framework that enhances the event understanding ability of MLLMs by editing the knowledge of the visual layer. Specifically, we first design a multilevel redundant neuron selection mechanism with information entropy and L1-normalization, and then conduct a mask matrix based knowledge editing strategy. The evaluations on M2E2 show that our method significantly outperforms all SOTA methods for multimedia event extraction. The main contributions of this paper are as follows:

- We propose a multimedia event extraction framework based on LLM knowledge editing which enhances the understanding ability of MLLMs on event structure features, avoids the reliance on the quality of training data or the consistency of data distribution brought by strategies such as fine-tuning.
- We propose a multi-level knowledge editing strategy which can effectively mitigate the influence of redundant neurons. Compared with the existing popular knowledge editing strategies, our method does not introduce any additional parameters and has extremely low computational and storage costs.
- The experiments on M2E2 benchmark show that our model achieves SOTA performance. We outperform the SOTA baselines by 34% Precision on multimedia event extraction and by 11% F1 on multimedia argument extraction.

2 RELATED WORK

2.1 Multimedia Event Extraction

Multimedia event extraction task (MEE) mainly includes event extraction and event argument extraction. Existing MEE methods can be divided into graph-based methods, template-based methods, and fine-tune-based methods according to their cross-modal alignment strategy. Graph-based methods typically supervise visual objects using textual entities through association algorithms grounded in knowledge graph structures. Following this idea, WASE (Li et al., 2020) applies an attention-based graph encoder to link the entities and objects extracted in different modals. Clip-event (Li et al., 2022) proposes an event graph alignment via optimal transport and evaluates their similarity at different granularities. MGIM (Liu et al., 2024) represents the multimedia information in a graph structure and performs coarse-grained alignment. Template-based methods identify the arguments according to the similarity calculated between candidates and roles in templates. Guided by this design paradigm, MMUTF (Seeberger et al., 2024) designs a unified template filling framework that encodes the corresponding event roles in templates to match the candidate textual or visual arguments. Fine-tune-based methods achieve cross-modal alignment by updating some parameters. For example, CAMEL (Du et al., 2023) proposes an incremental training strategy that complements missing images and text with a bidirectional cross-modality data augmentation. UMIE (Sun et al., 2024) adaptively learns the correlation between textual embedding and visual objects with a gated attention mechanism.

Compared to these methods, our method is the first to apply knowledge editing technique to alleviate the redundancy feature understanding problem on MEE, obtaining more accurate event structure understanding with less cost overhead.

2.2 Instruction following

Instruction following aims to help LLMs to understand various instructions and produce the corresponding responses, so that the model can be suitable for downstream tasks. Existing instruction following technology can be divided into three types: natural-language-inference-oriented instructions (Yin et al., 2019; Xu et al., 2023; Zhong et al., 2021), which unifies various classification problems into an inference task by constructing hypotheses to explain the labels; LLM-oriented instructions (Gao et al., 2020; Bach et al., 2022a,b), which utilizes templates to convert the origin inputs into machine-friendly fill-in-blank questions; and human-oriented instructions (Mishra et al., 2021; Gupta et al., 2022; Yin et al., 2022), which splits the task into definitions, demonstrations, instances, and other information with obvious descriptive or paragraph-style based on the way people under-

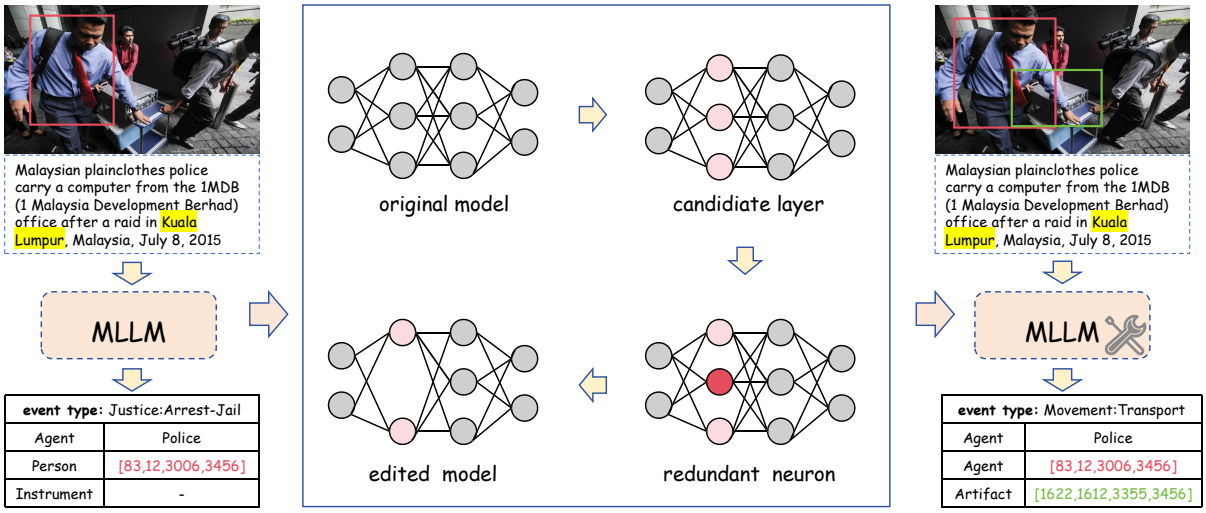

Figure 2: The framework of our method, we first select the candidate layer with entropy and Gaussian smoothing, then identify the redundant neurons on visual layers, and finally conduct knowledge editing for MLLMs.

stand the problem. These techniques can significantly improve the model's ability to understand downstream tasks through different instruction design strategies, and have been widely used in few-shot and zero-shot classification scenarios.

2.3 Knowledge editing

Knowledge editing aims to update, correct, optimize, or supplement existing knowledge. Common LLM knowledge editing methods can be divided into external-memorization-based methods, reparameterization-based methods, and local-modification-based methods. In order to obtain an accurate representation of knowledge in downstream tasks, external-memorization-based methods update knowledge by maintaining a memory which retrieves the most relevant cases (Zheng et al., 2023; Mitchell et al., 2022; Zhong et al., 2023), or introducing additional trainable parameters while maintaining the pre-trained backbone unchanged (Pfeiffer et al., 2020; Lei et al., 2023; Houlsby et al., 2019); reparameterization-based methods construct low-dimensional reparameterization of original parameters for training and equivalently transforms it back for inference (Lin et al., 2024; Hu et al., 2022; Wu et al., 2024); local-modification-based methods aim to locate and optimize relevant parameters for specific knowledge learning in LLM (Guo et al., 2020; Zhang et al., 2024; He et al., 2023). Our method is designed based on the idea of local modification, but compared with the previous local-modification-based work, we propose a knowledge editing technique that does not require back propagation, which improves the performance on downstream tasks with extremely small computational overhead.

3 Model

Let $d = (t, v)$ represent a multimedia document consisting of a text-image pair that describes a specific event, where t is the text and v is the corresponding image. The multimedia event extraction task contains the following two components.

Event Extraction: Given a multimedia document d, Event Extraction (EE) requires predicting the event type y_e from the candidates C.

Event Argument Extraction: Given a multimedia document d and event type y_e, Event Argument Extraction (EAE) aims to identify event arguments (entities or objects) with pre-defined event roles (all participants and attributes) associated with y_e.

Different from the currently popular methods of introducing additional parameters to fine-tune the pre-trained model for downstream tasks (Pfeiffer et al., 2020; Houlsby et al., 2019; Hu et al., 2022), we propose a large language model knowledge editing approach that does not require fine-tuning. As mentioned above, we focus on the visual perception layer. Our approach consists of two stages: redundant neuron selection and knowledge editing. We first propose a multilevel redundant neuron selection mechanism for MLLMs, and then use a mask-matrix-based editing strategy to edit redundant visual perception knowledge.

3.1 Multilevel redundant neuron selection

Relevant studies show that the pre-trained trunk exhibits different feature patterns at different parameter positions (Chatterji et al., 2019; Kumar et al., 2022; Naseer et al., 2021). In order to improve the ability of large language models to understand event features, a redundant neuron recognition strategy is proposed in this paper. Specifically, given the multimodal large language model $\theta = \{W^1, W^2, ..., W^n\}$, where $W^k = \{w_{ij}^k | i \in I, j \in O\}$ represents the network parameter set of layer k of the large language model, I and O represent the number of input feature and output feature dimensions of layer k respectively. We first convert the network weight parameter into a probability distribution, and the formula is expressed as:

$$P_{ij}^k = \frac{\exp(w_{ij}^k)}{\sum_{b=1}^{O}\sum_{a=1}^{I}\exp(w_{ab}^k)} \tag{1}$$

We use entropy to measure the importance of neural network layers, which can be expressed as:

$$H_i = -\sum_{b=1}^{O}\sum_{a=1}^{I} P_{ij}^k \log_2(P_{ij}^k) \tag{2}$$

$$H_i^{avg} = \sum_{k=-K}^{K} \mathcal{G}(k;\sigma) \cdot H_{i+k} \tag{3}$$

where $\mathcal{G}(k;\sigma)$ is a Gaussian kernel with bandwidth σ, K is the kernel size. The consideration behind this is that if the probability distribution of weights within a neural network layer is very different, and such difference is more obvious than the weight distribution of the neighboring layer, it means that there is a certain preference from input features to outputs in that layer, resulting in the layer being sensitive to feature learning. We select the layer of neural network with the most average weight distribution according to the entropy value, which is represented as θ^r, and consider that redundant information transfer in this layer is the cause of incorrect knowledge learning.

Next, we evaluated the differences in the information value of neurons within θ^r. We think that neurons represent the ability to perceive different feature dimensions, and we want to identify those neurons that are used to capture redundant features. For each weight belonging to θ^r, we use the L1-norm to identify redundant neurons and information transfer pathways. Neuron c_i importance can be expressed as:

$$c_i^r = \sum_{j=1}^{O} |w_{ij}^r| \tag{4}$$

We use $\delta \in (0, 1)$ as the threshold for judging the redundancy of the target layer to avoid the case of layer collapse, and use s to represent the corresponding confidence importance boundary of δ in θ^r. Neurons that are determined to be redundant can be expressed as:

$$C^r = \{c_i^r \mid c_i^r < s\} \tag{5}$$

3.2 Mask matrix based knowledge editing strategy

In order to modify the understanding of event knowledge of the MLLMs, we apply a mask matrix to mask the redundant neurons. The mask matrix can be expressed as:

$$M_{ij} = \begin{cases} 0, c_j^r \in C^r \\ 1, c_j^r \notin C^r \end{cases} \forall i \in I, j \in O \tag{6}$$

The network weight parameters after the mask can be expressed as:

$$\tilde{W}^r = w_{ij}^r * M_{ij} \tag{7}$$

M_{ij} indicates the state information of the mask matrix at (i, j). For each parameter in $\tilde{W}^r$, it is masked if it represents the weight connected to the neuron in C^r, otherwise it is retained as original.

Event Type	**Argument Role**
Movement:Transport	Agent, Artifact, Vehicle, Destination, Origin
Conflict:Attack	Attacker, Target, Instrument, Place
Conflict.Demonstrate	Entity, Police, Instrument, Place
Justice:ArrestJail	Agent, Person, Instrument, Place
Contact:PhoneWrite	Entity, Instrument, Place
Contact:Meet	Participant, Place
Life:Die	Agent, Instrument, Victim, Place
Transaction:TransferMoney	Giver, Recipient, Money

Table 1: Event types and corresponding argument roles in multimedia dataset M2E2.

Table 2: Results of Event Extraction on M2E2, we compare our framework against 5 baselines and 3 variants, the bold numbers denote the best results of all methods.

Metrics \ Task / Baselines	Event Extraction Task								
	text-only			**image-only**			**multimedia**		
	P	R	F1	P	R	F1	P	R	F1
WASE	42.8	61.9	50.6	43.1	59.2	49.9	43.0	62.1	50.8
CLIP-EVENT	-	-	-	41.3	**72.8**	52.7	-	-	-
CAMEL	45.1	71.8	55.4	52.1	66.8	58.5	55.6	59.5	57.5
UMIE	-	-	-	-	-	-	-	-	62.1
MMUTF	45.1	71.8	55.4	55.1	59.1	57.0	47.9	63.4	54.6
Qwen2vl-7b	80.05	73.94	74.49	69.11	65.38	64.83	83.77	73.14	70.64
Variant-1	81.36	71.81	73.06	68.55	63.09	64.34	83.88	73.46	70.99
Variant-2	82.07	70.74	72.70	69.03	64.15	63.46	85.41	79.93	82.66
Variant-3	81.07	70.89	72.26	69.15	65.10	64.89	89.13	87.06	86.59
Ours	**82.10**	**76.06**	**76.01**	**69.32**	65.77	**65.30**	**89.82**	**87.70**	**87.39**

4 Experiment

In this section, we extensively evaluate our framework by comparing against existing SOTA approaches and variants of MLLM with different redundant neuron selection strategies.

4.1 Experimental settings

Datasets We evaluated our framework on the M2E2 benchmark, a large-scale multimedia event extraction dataset with 8 types of events and 15 types of arguments, as shown in Tab. 1. Specifically, M2E2 contains 245 multimedia documents with 6167 sentences and 1014 images. There are 1,297 text events and 391 visual events, of which 192 text event mentions and 203 visual event mentions are arranged into 309 multimedia events.

Baselines We compare our proposed method against a wide range of SOTA models: WASE (Li et al., 2020) uses weakly aligned structured embedding to encode the multimodal events, CLIP-Event (Li et al., 2022) performs visual event extraction with a pre-trained CLIP network, CAMEL (Du et al., 2023) learns more accurate text alignment with image generator and image captioner, UMIE (Sun et al., 2024) constructs a series of instruction following templates, and MMUTF (Seeberger et al., 2024) addresses the MEE with candidate-query matching.

Experimental Setup In this work, we use Qwen2-VL-7B as the foundational model. The experiments are conducted on an NVIDIA A100 GPU with a memory capacity of 40GB, and due to GPU memory limitations, the size of images in M2E2 is set to 512*512 resolution. In addition, we also build several MLLM-based variants with other redundant neuron selection strategies which expands the scope of knowledge editing to all fully connected layers (Variant-1), uses L1-normalization of neurons for identification (Variant-2), and focuses on the redundancy of weights rather than the neurons (Variant-3).

4.2 Results

Tab. 2 and Tab. 3 show the comparison in detail. Compared with previous SOTA methods, the original Qwen2-VL-7b has a significantly outperformance, demonstrating the strong semantic understanding ability of the pre-trained multimodal large language model for text and images. Our method inherits such ability, and enhances the event structure understanding of MLLMs with knowledge editing that makes a comprehensive performance improvement. Especially for multimedia event extraction, the selection of redundant feature makes a 6.15% Precision improvement, 14.56% Recall improvement, 16.75% F1 improvement on multimedia event extraction. Compared with Variant-1, the focus on the visual encoding layer makes a 5.94% Precision improvement, which validates our proposed hypothesis that the bottleneck of multimodal event extraction lies in the capability of image event understanding. The comparison against Variant-2

Table 3: Results of Argument Extraction on M2E2.

Metrics \ Task / Baselines	Argument Extraction Task								
	text-only			**image-only**			**multimedia**		
	P	R	F1	P	R	F1	P	R	F1
WASE	23.5	30.3	26.4	14.5	10.1	11.9	19.5	18.9	19.2
CLIP-EVENT	-	-	-	21.1	13.1	17.1	-	-	-
CAMEL	24.8	41.8	31.1	21.4	**28.4**	24.4	31.4	35.1	33.2
UMIE	-	-	-	-	-	-	-	-	24.5
MMUTF	33.6	**44.2**	38.2	23.6	18.8	20.9	39.9	20.8	27.4
Qwen2vl-7b	50.10	36.19	37.83	27.30	27.63	27.31	47.13	42.70	43.04
Variant-1	48.07	36.25	36.81	25.42	26.95	25.67	46.45	40.63	42.77
Variant-2	50.69	36.19	37.95	25.08	24.58	25.88	48.02	42.25	44.08
Variant-3	51.45	37.05	38.22	26.64	27.63	25.90	47.47	41.03	43.39
Ours	**51.92**	37.48	**39.34**	**27.30**	27.63	**27.31**	**48.11**	**43.89**	**44.25**

demonstrates the superiority of the entropy-based redundant identification mechanism over directly judging based on L1 values. And the comparison against Variant-3 may due to the unstable weight masking causes the broken of structured sparsity, which may causes incomplete input received by downstream neurons, affecting feature expression ability. For multimedia argument extraction task, our method still achieve a 8.21% Precision improvement, 8.79% Recall improvement, 11.05% F1 improvement compared with the nearest SOTA performance. We attribute this as our appropriate feature extraction granularity which revises the understanding of roles in events by MLLM. Note that we do not utilize any additional image-text paired datasets or cross-modal data augmentation for fine-tuning in this work. Incorporating cross-modal information from the document's context and external datasets, as well as customized Prompt Engineering for each task, might further improve our method's performance.

4.3 Case Study

We selected four representative cases to illustrate the effect of multimodal large language models after knowledge editing, as shown in Fig. 3.

In case-1, the main reason for the original MLLM's error lies in the presence of multiple triggers strongly associated with different candidate event types in the text: the simultaneous existence of "Protestor" and "detained" leads to a misunderstanding of the event from a textual perspective. Our method does not focus on achieving more accurate textual identification but corrects the erroneous event understanding by optimizing the main subject's actions in visual modality. Another type of error is shown in case-2: the event "Confilct:Attack" is not reflected in the corresponding image. We attribute this to the temporal misalignment between the text and image information. Despite the painstaking efforts of M2E2 to clean the mismatched text-image data, such temporal misalignment still affects model's understanding of the event content. In contrast, our method alleviates the over-interpretation of image information by filtering redundant image features, thereby achieving an appropriate fusion of event text and image features.

Case-3 shows the mistake of identifying the argument of event type "Conflict:Demonstrate" as "Islamist terroism". In fact, it illustrates a typical problem of multimedia textual argument extraction with original LLM. Due to the lack of understanding of the event structure, the pre-trained model interprets the object of the event as the subject of the event. Although our model only edits the visual layer, it still significantly enhances the ability to extract textual arguments, which are also observed on text-only MAE. We infer that it may be because the moderate visual sparsity can induce text-side capability compensation, resulting in an overall performance boost. Case-4 gives an example of correcting incomplete visual event argument. Compared with the original MLLM, our method provides more accurate visual event object identi-

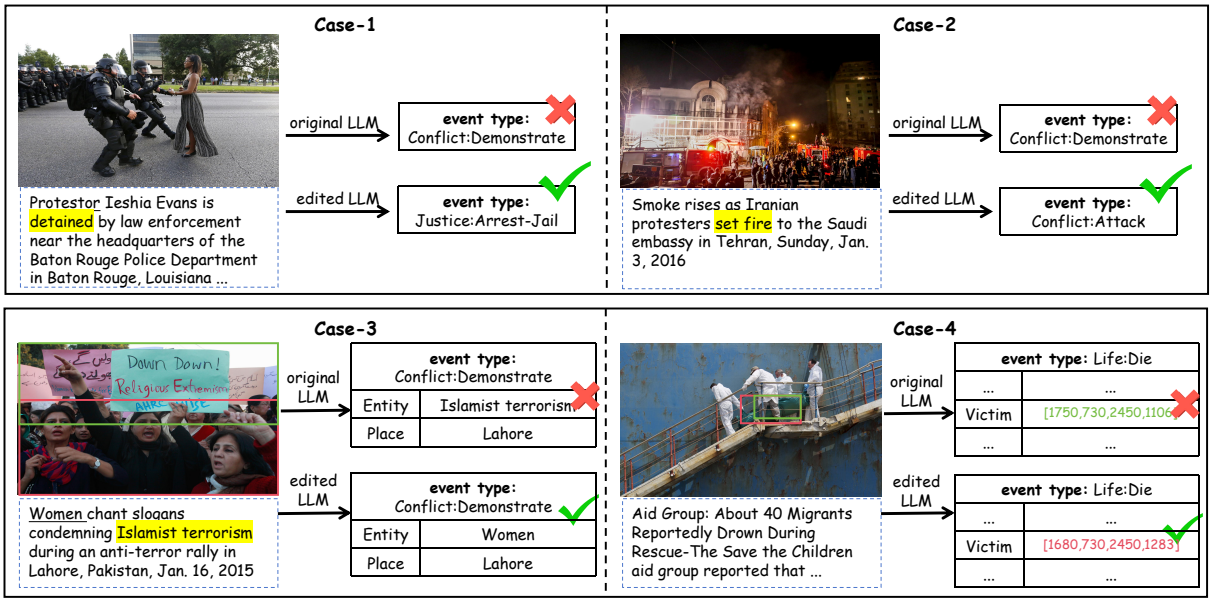

Figure 3: Cases on M2E2, the above two shows the error correction cases in the MEE task, and the following two correct errors for text argument and image argument in the MAE task.

fication with appropriate size boundaries. We attribute this to our knowledge editing strategy which focus on eliminating redundant visual features. It not only solves the feature dilution caused by excessive visual layer channels, but also reduces the interference of noise on subsequent layers, making the model more inclined to focus on some key features such as lines and colors, ultimately alleviating the boundary blurring problem of the original MLLM.

5 Conclusion

In this study, we propose a multimedia event extraction method based on the knowledge editing of large language models. We first analyze the reasons for the poor performance of multimodal large language models in multimedia event extraction tasks, and then design a multi-level redundant neuron selection mechanism. Finally, based on the identified redundancy, we edited the knowledge of LLM to enhance the understanding of event structure. Compared to previous work, our method does not require fine-tuning and has extremely low computational and storage costs. Substantial experiments on the M2E2 benchmark show that our method significantly improves the LLM's understanding of multimedia event structures, and has become a new SOTA with a 34% Precision improvement on MEE and a 11% F1 performance improvement on MAE. Future work mainly focus on the research of domain adaptation capabilities for multimodal large language models.

6 Acknowledgement

This work is supported by National Natural Science Foundation of China (62471051, 92467203 and 62372050), Beijing Natural Science Foundation of China (L251038, QY24203, L244010 and 4232029), CCF-Huawei Populus Grove Fund (TC202418), Fellowship of China National Postdoctoral Program for Innovative Talents (BX20240045), China Postdoctoral Science Foundation General Program (2025M773481).

Limitations

Although our work achieved a significant improvement in accuracy in multimodal event extraction and argument extraction tasks through knowledge editing of the visual layer, this editing is not as sensitive to image only argument extraction as other subtasks. This may be due to our multimodal large language model's tendency towards object level bounding-boxes rather than instance level, and we leave more validation work to future experiments.

In addition, we found in the experiment that the M2E2 dataset itself has some missing argument labels. In fact, the extracted events can be further subdivided into events with clear subject-object-relationships (e.g. Justice:Arrest-Jail) and events with unclear subject-object-relationships (e.g. Contact:Meet). Normally, they should have different mining difficulties, but due to the limitations of

multimodal datasets, we have to leave these tasks for further research.

References

Stephen H Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, et al. 2022a. Promptsource: An integrated development environment and repository for natural language prompts. *arXiv preprint arXiv:2202.01279*.

Stephen H Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, et al. 2022b. Promptsource: An integrated development environment and repository for natural language prompts. *arXiv preprint arXiv:2202.01279*.

Niladri S Chatterji, Behnam Neyshabur, and Hanie Sedghi. 2019. The intriguing role of module criticality in the generalization of deep networks. *arXiv preprint arXiv:1912.00528*.

Zilin Du, Yunxin Li, Xu Guo, Yidan Sun, and Boyang Li. 2023. Training multimedia event extraction with generated images and captions. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5504–5513.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Demi Guo, Alexander M Rush, and Yoon Kim. 2020. Parameter-efficient transfer learning with diff pruning. *arXiv preprint arXiv:2012.07463*.

Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey P Bigham. 2022. Instructdial: Improving zero and few-shot generalization in dialogue through instruction tuning. *arXiv preprint arXiv:2205.12673*.

Haoyu He, Jianfei Cai, Jing Zhang, Dacheng Tao, and Bohan Zhuang. 2023. Sensitivity-aware visual parameter-efficient fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11825–11835.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*.

Tao Lei, Junwen Bai, Siddhartha Brahma, Joshua Ainslie, Kenton Lee, Yanqi Zhou, Nan Du, Vincent Zhao, Yuexin Wu, Bo Li, et al. 2023. Conditional adapters: Parameter-efficient transfer learning with fast inference. *Advances in Neural Information Processing Systems*, 36:8152–8172.

Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. 2022. 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*.

Yang Lin, Xinyu Ma, Xu Chu, Yujie Jin, Zhibang Yang, Yasha Wang, and Hong Mei. 2024. Lora dropout as a sparsity regularizer for overfitting control. *arXiv preprint arXiv:2404.09610*.

Yang Liu, Fang Liu, Licheng Jiao, Qianyue Bao, Long Sun, Shuo Li, Lingling Li, and Xu Liu. 2024. Multi-grained gradual inference model for multimedia event extraction. *IEEE Transactions on Circuits and Systems for Video Technology*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. 2022. Memory-based model editing at scale. In *International Conference on Machine Learning*, pages 15817–15831. PMLR.

Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. 2021. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*.

Philipp Seeberger, Dominik Wagner, and Korbinian Riedhammer. 2024. Mmutf: Multimodal multimedia event argument extraction with unified template filling. *arXiv preprint arXiv:2406.12420*.

Lin Sun, Kai Zhang, Qingyuan Li, and Renze Lou. 2024. Umie: Unified multimodal information extraction with instruction tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19062–19070.

Xun Wu, Shaohan Huang, and Furu Wei. 2024. Mixture of lora experts. *arXiv preprint arXiv:2404.13628*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

Wenpeng Yin, Jia Li, and Caiming Xiong. 2022. Contintin: Continual learning from task instructions. *arXiv preprint arXiv:2203.08512*.

Zhi Zhang, Qizhe Zhang, Zijun Gao, Renrui Zhang, Ekaterina Shutova, Shiji Zhou, and Shanghang Zhang. 2024. Gradient-based parameter selection for efficient fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28566–28577.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? *arXiv preprint arXiv:2305.12740*.

Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. 2021. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. *arXiv preprint arXiv:2104.04670*.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. 2023. Mquake: Assessing knowledge editing in language models via multi-hop questions. *arXiv preprint arXiv:2305.14795*.