Cross-modal Multi-task Learning for Multimedia Event Extraction

Jianwei Cao[1], Yanli Hu[1*], Zhen Tan[1], Xiang Zhao[2]

[1]National Key Laboratory of Information Systems Engineering, National University of Defense Technology, China
[2]Laboratory for Big Data and Decision, National University of Defense Technology, China
{caojianwei, huyanli, tanzhen08a, xiangzhao}@nudt.edu.cn

Abstract

Multimedia event extraction aims to jointly extract event structural knowledge from multiple modalities, thus improving the comprehension and utilization of events in the growing multimedia content (e.g., multimedia news). A key challenge in multimedia event extraction is to establish cross-modal correlations during training without multimedia event annotations. Considering the complexity and cost of annotation across modalities, the multimedia event extraction task only provides parallel annotated data for evaluation. Previous works attempt to learn implicit correlations directly from unlabeled image-text pairs, but do not yield substantially better performance for event-centric tasks. To address this problem, we propose a cross-modal multi-task learning framework **X-MTL** to establish cross-modal correlations at the task level, which can simultaneously address four key tasks of multimedia event extraction: trigger detection, argument extraction, verb classification, and role classification. Specifically, to process inputs from different modalities and tasks, we utilize two separate modality-specific encoders and a modality-shared encoder to learn joint task representations, and introduce textual and visual prompt learning methods to enrich and unify task inputs. To resolve task conflict in cross-modal multi-task learning, we propose a pseudo label based knowledge distillation method, combined with dynamic weight adjustment method, which can effectively lift the performance to surpass the separately-trained models. On the **M**ulti**M**edia **E**vent **E**xtraction benchmark M2E2, experimental results show that X-MTL surpasses the current state-of-the-art (SOTA) methods by 4.1% for multimedia event mention and 8.2% for multimedia argument role.

Introduction

With the rapid growth of multimedia content, multimodal information extraction (Yu et al. 2020; Zheng et al. 2021) has gradually gained attention. However, most related studies focus on leveraging additional modalities to enhance the extraction of information from text, without fully exploring and utilizing the rich information contained in those modalities. Only a few studies (Li et al. 2020; Wang et al. 2023) have investigated methods for jointly extracting information from multiple modalities, which face significant challenges such as modality gap and annotation scarcity. In this paper, we conduct research on the multimedia event extraction task, which aims to jointly extract event structural knowledge from multiple modalities, thus improving the comprehension and utilization of events in the multimedia content (e.g., multimedia news).

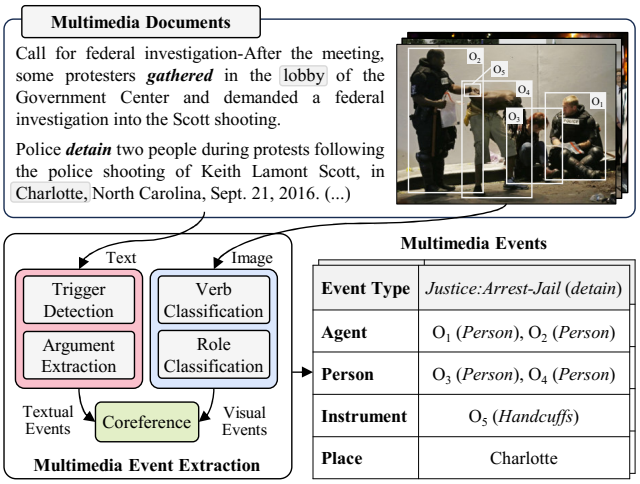

Figure 1: The overview of multimedia event extraction.

A basic multimedia event extraction pipeline typically comprises three parts, as shown in Figure 1. (1) Textual event extraction to identify event triggers and arguments from text; (2) Visual event extraction to identify verbs (i.e., triggers) and roles (i.e., arguments) from image; (3) Coreference to integrate events from multiple modalities that refer to the same real-world event into the multimedia events. Considering the complexity and cost of annotation across modalities, the task only provides parallel annotated data for evaluation, while training is conducted on independent annotated data, such as ACE 2005 (Walker et al. 2006) and imSitu (Yatskar, Zettlemoyer, and Farhadi 2016).

Hence, how to establish cross-modal correlations without multimedia event annotations is one of the significant challenges in multimedia event extraction. Previous works attempt to learn correlations from additional image-text pairs. For example, WASE (Li et al. 2020) and UniCL (Liu, Chen, and Xu 2022) use an image-text pair dataset to align modal-

*Corresponding author

ities and learn shared representations. CAMEL (Du et al. 2023) uses generative models to synthesize image-text pairs, expanding unimodal datasets to learn synergistic representations. Although these methods have achieved good results, the implicit cross-modal correlations learned from unlabeled image-text pairs often lack event knowledge, and may be insufficient to support the comprehension of complex event-specific concepts.

To address this problem, we propose a cross-modal multi-task learning framework X-MTL, which simultaneously addresses four key tasks of multimedia event extraction: trigger detection, argument extraction, verb classification, and role classification. By establishing cross-modal correlations at the task level, X-MTL can leverage event knowledge from multiple modalities and tasks to improve generalization and achieves higher task efficiency than previous methods.

In order to process inputs from different modalities and tasks, we propose a hard parameter sharing based cross-modal multi-task model to learn joint task representations, which consists of two separate modality-specific encoders and a modality-shared encoder. Meanwhile, we introduce textual and visual prompt learning methods to enrich and unify task inputs, which effectively enhances the representational capacity of model.

The modality gap amplifies conflicts between different tasks during cross-modal multi-task training. To alleviate this issue, we propose a pseudo label based knowledge distillation method to transfer knowledge from separately-trained models to the multi-task model, and use dynamic weight adjustment method to balance the training process. This method effectively lifts the performance to surpass the separately-trained models. Notably, we use an image-text pair dataset to generate pseudo labels, which preserve knowledge from separately-trained models as well as introducing additional cross-modal event knowledge.

On the multimedia event extraction benchmark M2E2, experimental results show that X-MTL has achieved significant improvements, surpassing the current state-of-the-art (SOTA) methods by 4.1% for multimedia event mention and 8.2% for multimedia argument role. Our contributions can be summarized as follows:

- We propose X-MTL, a cross-modal multi-task learning framework that establishes task-level cross-modal correlations to improve multimedia event extraction performance without relying on multimedia event annotations.
- To alleviate the task conflict during cross-modal multi-task training, we propose a pseudo label based knowledge distillation method, which effectively lifts the overall performance to surpass the separately-trained models.
- X-MTL demonstrates higher task efficiency and significantly outperforms current SOTA methods on the multimedia event extraction benchmark M2E2.

Related Work

Multimedia Event Extraction. Traditional event extraction (Lin et al. 2020; Ma et al. 2022) mainly focuses on textual data. Although some previous studies (Yu et al. 2020; Zheng et al. 2021) have incorporated images to enhance performance, the outputs remain textual. To better utilize multimedia data, literature (Li et al. 2020) proposes the multimedia event extraction task and develops the first benchmark M2E2. Unlike other similar tasks, the multimedia event extraction task does not offer parallel annotated data for training, effective methods are required to establish cross-modal correlations for improving task performance.

Most previous studies learn cross-modal correlations from image-text pairs. WASE (Li et al. 2020) utilizes image-text pairs to align structured semantic representations of images and texts into a common space. UniCL (Liu, Chen, and Xu 2022) leverages contrastive learning to create a shared space for texts and images, improving their similar representations. CAMEL (Du et al. 2023) utilizes generative models to complement unimodal datasets and learn the correlations from synthetic image-text pairs.

However, the correlations learned from unlabeled image-text pairs are implicit and not event-centric. To further learn event knowledge from image-text pairs, CLIP-Event (Li et al. 2022) incorporates structured event information into vision-language pre-training, enabling the model to comprehend complex event-specific concepts, but the dual encoder architecture limits it in handling complex tasks. UMIE (Sun et al. 2024) employs instruction tuning on a large language model for unified multimodal information extraction, to learn task-related knowledge from similar tasks. Moreover, JMMT (Chen et al. 2021) focuses on the video modality, directly modeling event-related video-article pairs for joint text and video event and argument extraction. This paper employs cross-modal multi-task learning to jointly model four tasks of multimedia event extraction across multiple modalities, to establish cross-modal correlations at the task level.

Multi-task Learning. Multi-task learning (Caruana 1997; Zhang and Yang 2022) effectively utilizes task-specific and shared information to simultaneously address related tasks, leading to more generalized representations and improved model robustness. However, most related studies focus on a specific modality, limiting broader application. Some works explore transformer-based unified models for handling tasks across multiple modalities. ViT-BERT (Li et al. 2021) employs a transformer encoder and multiple task-specific heads to handle text-only and image-only tasks simultaneously. UniT (Hu and Singh 2021) and OFA (Wang et al. 2022) extend the encoder-decoder architecture to different modalities, enabling it to handle a broader range of tasks, including multimodal tasks.

Nevertheless, these studies are typically in the general domain and require a lot of data for training. In this paper, we propose a cross-modal multi-task modeling and training framework for multimedia event extraction task, which significantly improves task performance and efficiency.

Problem Formulation

The multimedia event extraction task aims to extract event structural knowledge from multimedia documents. Each multimedia document consists of a set of sentences $S = \{s_1, s_2, \ldots\}$ and images $M = \{m_1, m_2, \ldots\}$, where each

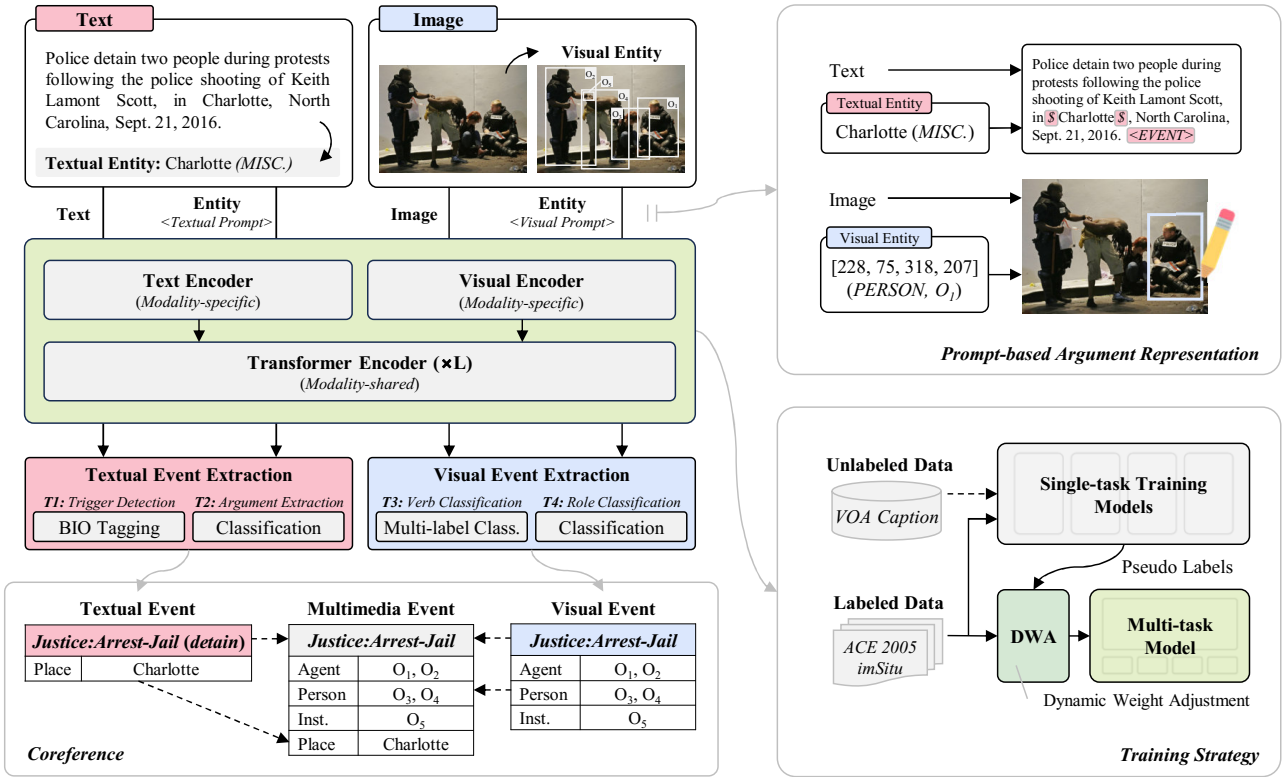

Figure 2: The overview of our proposed cross-modal multi-task learning framework X-MTL.

sentence consists of a sequence of tokens $s = \{w_1, w_2, \ldots\}$. Each document contains a set of textual entities $T = \{t_1, t_2, \ldots\}$ and visual entities $O = \{o_1, o_2, \ldots\}$, which typically refer to individually unique objects in the real world. The task mainly comprises three key parts:

Textual event extraction. Given a set of sentences S in the multimedia document, we extract a set of textual events e_t and their corresponding textual arguments a_t. Each textual event e_t has an event type y_e and a trigger word w, obtained through the **trigger detection** task. Each textual argument a_t has a role type y_a and corresponding textual entity t, obtained through the **argument extraction** task.

Visual event extraction. Given a set of images M in the multimedia document, we extract a set of visual events e_v and their corresponding visual arguments a_v. Each visual event e_v has an event type y_e and a corresponding image m. Each visual argument a_v has a role type y_a and corresponding visual entity o. Since the visual event extraction task is adapted from situation recognition, we refer to its subtasks as **verb classification** and **role classification**, following the definitions in situation recognition.

Coreference. The textual and visual events that refer to the same real-world event can be further combined into multimedia events, where their arguments are merged. The remaining events without corresponding relations are categorized as text-only events or image-only events.

The multimedia event extraction task only provides multimedia event annotations for evaluation. Three datasets are used for training: (1) ACE 2005 (Walker et al. 2006) includes textual event annotations from the event extraction task; (2) imSitu (Yatskar, Zettlemoyer, and Farhadi 2016) includes visual event annotations from the situation recognition task; (3) VOA Caption (Li et al. 2020) includes image-text pairs crawled from VOA news.

Methodology

Cross-modal multi-task learning framework X-MTL mainly consists of three parts: cross-modal task-shared model, task-specific output heads, and training strategy.

Cross-modal Task-shared Model

Transformer (Vaswani et al. 2017) demonstrates powerful capabilities in handling multiple modalities of data. Our proposed cross-modal task-shared model is based on the transformer encoder architecture and comprises two parts: modality-specific encoder and modality-shared encoder. It can simultaneously process inputs from different modalities and tasks, and output task-specific representations.

Modality-specific encoder. Two transformer-based encoders, BERT (Devlin et al. 2019) and CLIP vision encoder (Radford et al. 2021), are employed to process different modalities of inputs. For input sentence s or image m, we can extract their representations H_s or H_m.

$$H_s = \text{BERT}(s), \quad H_m = \text{CLIP}_{\text{vision}}(m), \tag{1}$$

where H_s denotes the representation of each token in the sentence s; H_m denotes the representation of each patch in

the image m. We utilize an linear layer to align the dimension of representation.

Modality-shared encoder. We stack multiple transformer encoder blocks to further extract task-specific representation Z_k from sentence representation H_s or image representation H_m. Through parameter sharing, the modality-shared encoder can learn knowledge from different modalities and tasks, enhancing the model's generalization. Each block consists of a multi-head self-attention (MHSA) mechanism and a feed-forward neural network (FFN). Layer normalization (LN) and skip connection are used to improve model performance and convergence.

$$H_{s/m}^{(0)} = H_s \text{ or } H_m, \quad Z_k = H_{s/m}^{(L)}, \tag{2}$$

$$\tilde{H}_{s/m}^{(l)} = \text{MHSA}(\text{LN}(H_{s/m}^{(l-1)})) + H_{s/m}^{(l-1)}, \tag{3}$$

$$H_{s/m}^{(l)} = \text{FFN}(\text{LN}(\tilde{H}_{s/m}^{(l)})) + \tilde{H}_{s/m}^{(l)}, \tag{4}$$

where $H_{s/m}^{(l)}$ denotes the output representation of the l-th blocks among a total of L blocks, the tilde denotes the intermediate representation; $k \in \{td, ae, vc, rc\}$ denotes the the four tasks in multimeida event extraction.

Task-specific Output Heads

Task-specific representations Z_k will be fed into task-specific output heads to complete corresponding tasks. The design of each task head is outlined as follows.

Textual Event Extraction. The first step of textual event extraction is trigger detection, which is treated as a sequence labeling task in this paper, and BIO tagging is used to label event types and trigger words. After obtaining the task-specific representation Z_{td} for trigger detection, we feed it into a CRF layer for decoding to obtain the predicted textual events. During training, we optimize the negative log-likelihood loss function to maximize the score ratio of the actual label sequence.

$$P(y_{td} \mid s) = \frac{\text{score}(y_{td}, Z_{td})}{\sum_{y'_{td} \in Y(s)} \text{score}(y'_{td}, Z_{td})}, \tag{5}$$

$$\mathcal{L}_{td} = -\log(P(\hat{y}_{td} \mid s)), \tag{6}$$

where y_{td} denotes the predicted label sequence; $Y(s)$ denotes a set of all possible label sequences; $\hat{y}_{td}$ denotes the gold label sequence.

The second step is argument extraction. We follow previous work (Li et al. 2020) and treat it as a classification task based on a set of ground-truth entities. Inspired by literature (Wu and He 2019), we use the textual prompt shown in Figure 2 to mark entity positions in the sentence and concatenate it with the predicted event, to generate the task-specific representation Z_{ae} for argument extraction. This method effectively preserves both the global and local semantic information of the textual arguments. We then perform average pooling on task-specific representation Z_{ae} and feed it into a classifier for classification. During training, the cross-entropy loss function is used to optimize task performance.

$$y_{ae} = \text{softmax}(\text{AvgPool}(Z_{ae}) \cdot A_{ae}^T + b_{ae}), \tag{7}$$

$$\mathcal{L}_{ae} = -y_{ae} \cdot \log(\hat{y}_{ae}), \tag{8}$$

where A and b are learnable parameters in linear classifier; y_{ae} denotes the predicted probability distribution; $\hat{y}_{ae}$ is the actual probability distribution, where the value for the ground-truth class is 1 and for all other classes is 0.

Visual Event Extraction. Similar to textual event extraction, visual event extraction also involves two steps: verb classification and role classification. Since an image may contain multiple events, we treat verb classification as a multi-label classification task and use the binary cross-entropy loss function to optimize model performance. Different from single-label classification task, the model's output is the sigmoid function instead of the softmax function.

$$y_{vc} = \text{sigmoid}(\text{AvgPool}(Z_{vc}) \cdot A_{vc}^T + b_{vc}), \tag{9}$$

$$\mathcal{L}_{vc} = -(y_{vc} \cdot \log(\hat{y}_{vc}) + (1 - y_{vc}) \cdot \log(1 - \hat{y}_{vc})), \tag{10}$$

where y_{vc} denotes the predicted probability for each class; $\hat{y}_{vc}$ is the actual probability, where the ground-truth classes are 1 and other classes are 0.

Following the textual argument extraction task, we classify the visual arguments from a set of visual objects, which are extracted using an offline object detector. Inspired by literature (Shtedritski, Rupprecht, and Vedaldi 2023), we use the visual prompt method to generate the task-specific representation Z_{rc} for role classification, where objects in the image are marked with special markers. This method is the same as the textual prompt method, which effectively preserves both the global and local semantic information of the visual arguments. The cross-entropy loss function is used to optimize task performance.

$$y_{rc} = \text{softmax}(\text{AvgPool}(Z_{rc}) \cdot A_{rc}^T + b_{rc}), \tag{11}$$

$$\mathcal{L}_{rc} = -y_{rc} \log(\hat{y}_{rc}), \tag{12}$$

where y_{rc} denotes the predicted probability distribution; $\hat{y}_{rc}$ is the actual probability distribution, where the value for the ground-truth class is 1 and for all other classes is 0.

Coreference. We use the same coreference method as in previous work (Du et al. 2023). A textual event and a visual event refer to the same real-world event only if they have the same predicted event type and their CLIP image-text similarity score exceeds a certain threshold. The multimedia event will inherit all arguments from both textual events and visual events.

Training Strategy

We propose a pseudo label based knowledge distillation method combined with dynamic weight adjustment to optimize the cross-modal multi-task training process.

Pseudo label based knowledge distillation. Pseudo label can effectively transfer knowledge from single-task trained models to the multi-task model, thereby maintaining the performance of multi-task model on single tasks.

First, we train four single-task models separately with the same task heads described in the previous section. Then, we utilize these models to annotate an image-text pair dataset and select pseudo labels based on a confidence threshold for each task. For sequence labeling tasks, the probability of the

predicted path is used as the confidence score. For classification tasks, the probability of the predicted class is used. An image-text pair is retained only if both the text and image have the same event prediction within the M2E2 schema and the scores exceed the threshold, which can further improve the accuracy of pseudo labels and help the model learn cross-modal knowledge.

Dynamic weight adjustment. A uniform and fixed learning rate for all tasks generally yields suboptimal results in cross-modal multi-task learning. To address this, we utilize a dynamic weight adjustment (DWA) method from the literature (Liu, Johns, and Davison 2019) to adjust task weight individually based on the task-specific loss.

$$\mathbf{W}^{(t)} = K \cdot \text{softmax}(\alpha \cdot \frac{\mathbf{L}^{(t-1)}}{\mathbf{L}^{(t-2)}}), \tag{13}$$

where $\mathbf{W}^{(t)}$ and $\mathbf{L}^{(t)}$ denote the combination of weights and losses for all tasks at time step t, respectively; α denotes the temperature factor, controlling the rate of weight change; K denotes the total number of tasks, which is set to 8 after including the pseudo label tasks.

Finally, combining all tasks and pseudo label tasks, the overall training loss is as follows.

$$\mathcal{L}_{total}^{(t)} = \sum\nolimits_{k,k' \in G} \mathcal{W}_k^{(t)} \cdot \mathcal{L}_k^{(t)} + \beta \cdot \mathcal{W}_{k'}^{(t)} \cdot \mathcal{L}_{k'}^{(t)}, \tag{14}$$

where $L_k^{(t)}$ denotes the loss of task k at time step t; $k^{'}$ denotes the pseudo label task corresponding to task k; G denotes the task group including all tasks and pseudo label tasks; β is the initial weight for pseudo label tasks.

Experiments

In this section, we evaluate our proposed framework X-MTL by comparing it with SOTA methods and conduct further analysis on different training strategies and argument representation methods.

Datasets and Evaluation

Following previous work (Li et al. 2020), we use the ACE 2005, imSitu, and VOA Caption datasets to train the model, and evaluate its performance on the M2E2 benchmark. ACE 2005 is an event extraction dataset comprising 15,789 sentences, covering 33 event types and 36 semantic roles. imSitu is a situation recognition dataset comprising 126,102 images, covering 504 activity verbs and 1,788 semantic roles. We additionally use grounding object information from SWiG (Pratt et al. 2020) for training. VOA Caption dataset is an unlabeled image-text pair dataset comprising 123,078 image-text pairs. M2E2 is a multimedia event extraction benchmark comprising 6,167 sentences and 1,014 images from 245 multimedia documents. It covers 8 event types and 15 argument roles, with 1297 textual events, 391 visual events, and 309 multimedia events (coreferenced by 192 textual events and 203 visual events).

For evaluation, we use precision (P), recall (R), and F1 score (F1) as evaluation metrics. Additionally, the Intersection over Union (IoU) between the predicted and ground truth bounding box for the visual argument must exceed 0.5.

Experimental Setup

For fair comparison, we adopt the experimental setup used in previous work (Du et al. 2023). For the backbone models, we use the BERT model (bert-base-uncased) to initialize the parameters of the text encoder, and the visual transformer of the CLIP model (clip-vit-base-patch32) to initialize the parameters of the visual encoder. To detect objects for visual argument extraction, we leverage the pretrained YOLOv8 (Varghese and M. 2024) as the object detector and remove detection results with confidence below 0.8.

The number of layers in the modality-shared encoder is set to 2 with a hidden layer size of 1024. We select pseudo label with confidence greater than 0.8. During training, the temperature factor for dynamic weight adjustment is 0.5; The initial weight for pseudo label tasks is set to 0.5, while the other tasks are 1.0. The maximum text input length is 300, which covers almost all input text lengths. The threshold for event coreference in CLIP scores is 0.2.

All the experiments are conducted on NVIDIA RTX 4090 GPU using the PyTorch framework. We implement the AdamW optimizer to minimize the loss function. The learning rate for the text encoder is set to 1e-4, the visual encoder is set to 1e-5, and other parameters are set to 1e-3. The mini-batch size is 64, sampled proportionally according to the dataset size. We train each model for 10 epochs to obtain the final results.

Compared Methods

We select the following multimedia event extraction methods for comparison to verify the effectiveness of our proposed method.

- VAD (Zhang et al. 2017) introduces additional visual features from an external visual repository to enhance the performance of textual event extraction.
- WASE (Li et al. 2020) utilizes unimodal datasets to train structured semantic representations of textual and visual data separately, then leverages the VOA Caption dataset to align two modalities. WASE-T and WASE-V refer to the WASE model without modality alignment. Flat ignores the structured semantic information modeling in WASE. There are two different methods for extracting visual arguments: the attention-based (att) method employs attention heatmaps to recognize visual arguments, while the object-based (obj) method employs object detectors.
- CLIP-Event (Li et al. 2022) integrates structured event knowledge into vision-language pretraining, enhancing the vision-language model's comprehension of events and associated participant roles. It can achieve visual event extraction by image-text matching.
- UniCL (Liu, Chen, and Xu 2022) utilizes contrastive learning to create a shared space for texts and images using the VOA Caption dataset, and then leverages cross-modal representation to separately enhance the performance of textual and visual event extraction models.
- CAMEL (Du et al. 2023) utilizes image generative networks and image captioning networks to complement existing unimodal training data, allowing multimodal event

Type	Method	Event Mention			Argument Role		
		P	R	F1	P	R	F1
Textual	VAD	34.8	64.4	45.2	23.1	27.5	25.1
	Flat	34.2	63.2	44.4	20.1	27.1	23.1
	WASE-T	42.3	58.4	48.2	21.4	30.1	24.9
	$WASE_{att}$	37.6	66.8	48.1	27.5	33.2	30.1
	$WASE_{obj}$	42.8	61.9	50.6	23.5	30.3	26.4
	UniCL	49.1	59.2	53.7	27.8	34.3	30.7
	CAMEL	45.1	71.8	<u>55.4</u>	24.8	41.8	<u>31.1</u>
	X-MTL	49.7	65.7	**56.6**	34.6	37.6	**36.0**
Visual	Flat	27.1	57.3	36.7	4.3	8.9	5.8
	$WASE\text{-}V_{att}$	29.7	61.9	40.1	9.1	10.2	9.6
	$WASE\text{-}V_{obj}$	28.6	59.2	38.7	13.3	9.8	11.2
	$WASE_{att}$	32.3	63.4	42.8	9.7	11.1	10.3
	$WASE_{obj}$	43.1	59.2	49.9	14.5	10.1	11.9
	CLIP-Event	41.3	72.8	52.7	21.1	13.1	17.1
	UniCL	54.6	60.9	57.6	16.9	13.8	15.2
	CAMEL	52.1	66.8	<u>58.5</u>	21.4	28.4	<u>24.4</u>
	X-MTL	73.1	70.3	**71.7**	33.2	31.3	**32.2**
Multi.	Flat	33.9	59.8	42.2	12.9	17.6	14.9
	$WASE_{att}$	38.2	67.1	49.1	18.6	21.6	19.9
	$WASE_{obj}$	43.0	62.1	50.8	19.5	18.9	19.2
	UniCL	44.1	67.7	53.4	24.3	22.6	23.4
	CAMEL	55.6	59.5	57.5	31.4	35.1	<u>33.2</u>
	UMIE	-	-	<u>62.1</u>	-	-	24.5
	X-MTL	78.3	57.3	**66.2**	40.3	42.6	**41.4**

Table 1: Main results on event mention and argument role for textual event extraction (top), visual event extraction (middle), and multimedia event extraction (bottom) on the M2E2 benchmark (%). Bold indicates the optimal result, underline indicates the suboptimal result.

extraction models to learn cross-modal correlations without relying on image-text pair datasets.

- UMIE (Sun et al. 2024) is a unified multimodal information extractor derived from instruction tuning on the FLAN-T5 model (Chung et al. 2022). It employs a generative approach to jointly address three tasks: multimodal entity recognition, multimodal relation extraction, and multimedia event extraction.

Main Results

Table 1 shows the performance of X-MTL compared with several other methods on the M2E2 benchmark. Our method performs well on all six metrics of the multimodal event extraction task, surpassing current SOTA performance.

For textual event extraction, we outperform CAMEL by 1.2% on event mention and 4.9% on argument role. The performance of argument extraction is affected by the preceding event extraction. We note that CAMEL improves by 1.7% on event mention relative to UniCL, but only by 0.4% on argument role. With a comparable improvement in event extraction, X-MTL achieves greater enhancement in argument extraction, demonstrating its effectiveness in improving overall task performance.

For visual event extraction, we outperform CAMEL by 13.2% on event mention and 7.8% on argument role. It should be emphasized that we employ the M2E2 schema to refine the imSitu dataset, which leads to significant performance improvements. Specifically, in single-task training, the visual event improves from 53.2% to 66.9% and the visual argument improves from 21.0% to 30.4%. We employ the same object-based argument localization method as CAMEL and outperform attention-based methods used in $WASE_{att}$ and UniCL. This further demonstrates the effectiveness of the object-based method in argument extraction.

For multimedia event extraction, we outperform UMIE by 4.1% on event mention and surpass CAMEL by 8.2% on argument role. It demonstrates that our proposed framework X-MTL effectively learns cross-modal correlations in multimedia data. Compared to the UMIE method, which also employs multi-task learning with a larger numbers of parameters and more labeled data for training, our approach is more effective.

Analysis of Training Strategies

We utilize cross-modal multi-task learning to jointly handle four tasks of multimedia event extraction, allowing each task to benefit from others across modalities. However, this approach confronts greater challenges in training due to the modality gap, which often diminish the overall task performance. To find an effective training strategy, we first examine two commonly used multi-task training strategies: (1) Alternating training to learn tasks one by one cyclically; (2) Joint training to learn multiple tasks simultaneously.

As shown in Table 2, joint training yields better performance in cross-modal multi-task training, but neither method reaches the performance of single-task training. Alternating training performs poorly on certain tasks due to catastrophic forgetting, with visual argument declining by 9.5% compared to single-task training. ARK (Ma et al. 2023) employs exponential moving average and knowledge distillation methods to alleviate the forgetting problem, but the results were not as effective as desired. Joint training is more stable in overall performance but is affected by task conflict, with performance improving on textual event extraction while the performance of visual and multimedia event extraction decline.

Our proposed training strategy is based on joint training, consisting of two main components: pseudo label based knowledge distillation and dynamic weight adjustment. Experimental results demonstrate that this strategy can significantly enhance the performance of the multi-task model, surpassing the single-task models. Through ablation studies, we can further conclude that the pseudo label helps maintain the performance on single tasks, while dynamic weight adjustment can alleviate task conflict in joint training. The combination of these two components achieves the optimal overall performance.

Type	Method	Textual Event	Textual Arg.	Visual Event	Visual Arg.	Multi. Event	Multi. Arg.
Single-task	Separately Training	52.8	33.2	66.9	30.4	62.7	41.1
Multi-task	Alternating Training	54.1	29.1	60.0	20.9	60.0	36.0
	ARK (Ma et al. 2023)	50.7	26.9	63.1	29.4	61.0	36.9
	Joint Training	55.6	34.0	63.6	26.6	56.8	35.4
	X-MTL (Ours)	**56.6**	**36.0**	**71.7**	**32.2**	**66.2**	**41.4**
	w/o Dynamic Weight Adjustment	55.5	35.2	67.3	31.0	61.5	39.7
	w/o Pseudo Label	53.1	34.1	62.2	25.5	59.8	39.0

Table 2: Performance comparison of different training strategies and ablation study of our proposed training strategy (%).

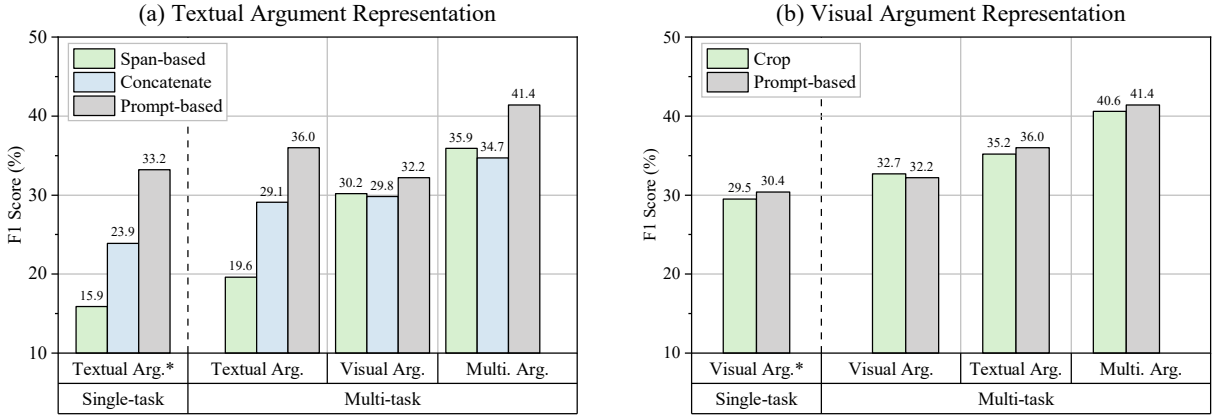

Figure 3: Performance comparison of different textual argument representation methods (left) and visual argument representation methods on model performance (%). Tasks marked with * indicate performance under single-task training, while those without are under multi-task training.

Comparison of Prompt-Based Representation

To better capture the semantic information of arguments and mitigate differences across tasks, we adopt a prompt-based representation for textual and visual arguments. To demonstrate its effectiveness, we compare this method with several different argument representation methods, and the results are presented in Figure 3.

In Figure 3(a), we compare two commonly used argument representation methods in event extraction. The span-based method directly uses the word representation of the argument as its semantic representation. The concatenate method concatenates the argument to a specified position in the text, using the global sentence representation as its semantic representation. Experimental results show that our proposed prompt-based method not only outperforms other methods in single-task training but also better improves the overall performance in multi-task training. Interestingly, while the concatenation method outperforms the span-based methods in single-task training, its enhancement for other tasks in multi-task training does not surpass span-based methods. This suggests that appropriate task design impacts the overall performance in multi-task training.

In Figure 3(b), we follow the literature (Shtedritski, Rupprecht, and Vedaldi 2023) and compare the prompt-based method with the crop method for visual argument. The crop method uses image region as representation, which retains only the local information of visual argument, while the prompt-based method preserves both local and global information. Experimental results show that different argument representation methods have relatively small effects on visual argument performance, but the prompt-based method can better enhance other tasks in multi-task training.

Conclusions and Future Work

In this paper, we investigate how to establish cross-modal correlations to improve the performance of multimedia event extraction without multimedia event annotations. We propose a cross-modal multi-task learning framework X-MTL to achieve this goal, where a cross-modal task-shared model is used to process inputs from different modalities and tasks, and a pseudo label based knowledge distillation method is used to alleviate the task conflict during cross-modal multi-task training. Experimental results demonstrate that X-MTL effectively outperforms separately-trained models, achieving significant performance improvements on the M2E2 benchmark. However, X-MTL is affected by error propagation from pseudo labels, and the overall process is relatively complex. In future work, we will explore more efficient optimization methods for cross-modal multi-task learning and broaden its application scope.

Acknowledgments

This research was supported by the National Key R&D Program of China (No. 2022YFB3103600), the National Natural Science Foundation of China (Nos. 72471237, 72371245, U23A20296, 62272469) and the Science and Technology Innovation Program of Hunan Province (No. 2023RC1007).

References

Caruana, R. 1997. Multitask Learning. *Mach. Learn.*, 28(1): 41–75.

Chen, B.; Lin, X.; Thomas, C.; Li, M.; Yoshida, S.; Chum, L.; Ji, H.; and Chang, S. 2021. Joint Multimedia Event Extraction from Video and Article. In *EMNLP (Findings)*, 74–88. Association for Computational Linguistics.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; Webson, A.; Gu, S. S.; Dai, Z.; Suzgun, M.; Chen, X.; Chowdhery, A.; Narang, S.; Mishra, G.; Yu, A.; Zhao, V. Y.; Huang, Y.; Dai, A. M.; Yu, H.; Petrov, S.; Chi, E. H.; Dean, J.; Devlin, J.; Roberts, A.; Zhou, D.; Le, Q. V.; and Wei, J. 2022. Scaling Instruction-Finetuned Language Models. *CoRR*, abs/2210.11416.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*, 4171–4186. Association for Computational Linguistics.

Du, Z.; Li, Y.; Guo, X.; Sun, Y.; and Li, B. 2023. Training Multimedia Event Extraction With Generated Images and Captions. In *ACM Multimedia*, 5504–5513. ACM.

Hu, R.; and Singh, A. 2021. UniT: Multimodal Multitask Learning with a Unified Transformer. In *ICCV*, 1419–1429. IEEE.

Li, M.; Xu, R.; Wang, S.; Zhou, L.; Lin, X.; Zhu, C.; Zeng, M.; Ji, H.; and Chang, S. 2022. CLIP-Event: Connecting Text and Images with Event Structures. In *CVPR*, 16399–16408. IEEE.

Li, M.; Zareian, A.; Zeng, Q.; Whitehead, S.; Lu, D.; Ji, H.; and Chang, S. 2020. Cross-media Structured Common Space for Multimedia Event Extraction. In *ACL*, 2557–2568. Association for Computational Linguistics.

Li, Q.; Gong, B.; Cui, Y.; Kondratyuk, D.; Du, X.; Yang, M.; and Brown, M. 2021. Towards a Unified Foundation Model: Jointly Pre-Training Transformers on Unpaired Images and Text. *CoRR*, abs/2112.07074.

Lin, Y.; Ji, H.; Huang, F.; and Wu, L. 2020. A Joint Neural Model for Information Extraction with Global Features. In *ACL*, 7999–8009. Association for Computational Linguistics.

Liu, J.; Chen, Y.; and Xu, J. 2022. Multimedia Event Extraction From News With a Unified Contrastive Learning Framework. In *ACM Multimedia*, 1945–1953. ACM.

Liu, S.; Johns, E.; and Davison, A. J. 2019. End-To-End Multi-Task Learning With Attention. In *CVPR*, 1871–1880. Computer Vision Foundation / IEEE.

Ma, D.; Pang, J.; Gotway, M. B.; and Liang, J. 2023. Foundation Ark: Accruing and Reusing Knowledge for Superior and Robust Performance. In *MICCAI (1)*, volume 14220 of *Lecture Notes in Computer Science*, 651–662. Springer.

Ma, Y.; Wang, Z.; Cao, Y.; Li, M.; Chen, M.; Wang, K.; and Shao, J. 2022. Prompt for Extraction? PAIE: Prompting Argument Interaction for Event Argument Extraction. In *ACL (1)*, 6759–6774. Association for Computational Linguistics.

Pratt, S. M.; Yatskar, M.; Weihs, L.; Farhadi, A.; and Kembhavi, A. 2020. Grounded Situation Recognition. In *ECCV (4)*, volume 12349 of *Lecture Notes in Computer Science*, 314–332. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Shtedritski, A.; Rupprecht, C.; and Vedaldi, A. 2023. What does CLIP know about a red circle? Visual prompt engineering for VLMs. In *ICCV*, 11953–11963. IEEE.

Sun, L.; Zhang, K.; Li, Q.; and Lou, R. 2024. UMIE: Unified Multimodal Information Extraction with Instruction Tuning. In *AAAI*, 19062–19070. AAAI Press.

Varghese, R.; and M., S. 2024. YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, 1–6.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NIPS*, 5998–6008.

Walker, C.; Strassel, S.; Medero, J.; and Maeda, K. 2006. ACE 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57: 45.

Wang, J.; Li, Z.; Yu, J.; Yang, L.; and Xia, R. 2023. Fine-Grained Multimodal Named Entity Recognition and Grounding with a Generative Framework. In *ACM Multimedia*, 3934–3943. ACM.

Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, 23318–23340. PMLR.

Wu, S.; and He, Y. 2019. Enriching Pre-trained Language Model with Entity Information for Relation Classification. In *CIKM*, 2361–2364. ACM.

Yatskar, M.; Zettlemoyer, L.; and Farhadi, A. 2016. Situation Recognition: Visual Semantic Role Labeling for Image Understanding. In *CVPR*, 5534–5542. IEEE Computer Society.

Yu, J.; Jiang, J.; Yang, L.; and Xia, R. 2020. Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer. In *ACL*, 3342–3352. Association for Computational Linguistics.

Zhang, T.; Whitehead, S.; Zhang, H.; Li, H.; Ellis, J. G.; Huang, L.; Liu, W.; Ji, H.; and Chang, S. 2017. Improving Event Extraction via Multimodal Integration. In *ACM Multimedia*, 270–278. ACM.

Zhang, Y.; and Yang, Q. 2022. A Survey on Multi-Task Learning. *IEEE Trans. Knowl. Data Eng.*, 34(12): 5586–5609.

Zheng, C.; Feng, J.; Fu, Z.; Cai, Y.; Li, Q.; and Wang, T. 2021. Multimodal Relation Extraction with Efficient Graph Alignment. In *ACM Multimedia*, 5298–5306. ACM.