CM-PIE: CROSS-MODAL PERCEPTION FOR INTERACTIVE-ENHANCED AUDIO-VISUAL VIDEO PARSING

Yaru Chen[1,4], *Ruohao Guo*[2], *Xubo Liu*[1], *Peipei Wu*[1], *Guangyao Li*[3], *Zhenbo Li*[4], *Wenwu Wang*[1]

[1]Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey, United Kindom
[2]School of Intelligence Science and Technology, Peking University, China
[3]Gaoling School of Artificial Intelligence, Renmin University of China, China
[4]College of Information and Electrical Engineering, China Agricultural University, China

ABSTRACT

Audio-visual video parsing is the task of categorizing a video with weak labels at the segment level, and predicting them as audible or visible events. Recent methods have leveraged the attention mechanism to capture the semantic correlations among the whole video across the audio-visual modalities. However, these approaches may overlook the importance of individual segments and their interrelations within a video, typically relying on a single modality when learning features. In this paper, we propose a novel interactive-enhanced cross-modal perception method (CM-PIE), which can learn fine-grained features by applying a segment-based attention module. In addition, a cross-modal aggregation block is introduced to jointly optimize the semantic representation of audio and visual signals by enhancing inter-modal interactions. Experimental results show that our model offers improved parsing performance on the Look, Listen, and Parse (LLP) dataset compared to other methods.

Index Terms— Segment-Based Attention, Cross-Modal Aggregation, Audio-Visual Video Parsing, Weakly-Supervised Learning

1. INTRODUCTION

Humans perceive multisensory signals through sight, hearing, touch and more, acquiring multi-modal information when they explore the environment. Enabling machines to fuse multi-modal information like humans is a valuable research topic in scene perception and understanding [1, 2]. As two basic modalities, audio and visual play a vital role in machine perception and understanding of scenes [3, 4]. Some researchers used audio and visual signals to capture the comprehensive scene information, which can improve model performance and generalization [5, 6]. However, the above methods usually assume audio and visual signals are temporally aligned, which, however, may not be the case, and thereby leading to inaccuracies in parsing the video.

To solve this problem, Tian et al [7] proposed the audio-visual video parsing (AVVP) task for a more fine-grained scene understanding. As shown in Fig. 1, AVVP aims to locate the temporal boundaries of event categories within a video with weak labels, and annotate them as audible, visible, or a combination of both. This task involves two challenges: One is to predict the event by extracting useful information from every segment. The other is to aggregate the cross-modality information to parse audio and visual events based on weak labels.

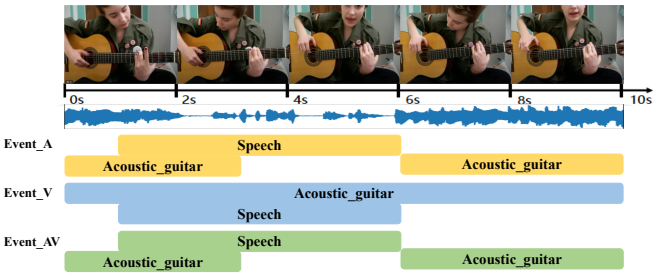

Fig. 1. Example of the AVVP task. Taking the audio and visual data as input, the task is to determine the event categories, their temporal boundaries and the modality of the event. Note that it is possible for audio events and visual events to be asynchronous (e.g. acoustic guitar).

In [7], a method combining hybrid attention networks (HAN) and a multi-modal multiple instance learning (MMIL) is used to aggregate multi-modal temporal contexts, together with the identification and suppression of noisy labels for each modality. Subsequently, Yu et al [8] proposed a method to capture and integrate multimodal pyramid features in different temporal scales. Afterward, Chen et al [9] explored common and specific characteristics between 2D and 3D visual features, and visual and audio features separately. While the above methods have achieved promising improvements, there are still some limitations: 1) The existing approaches explore all features holistically from a whole video, but overlook the importance of individual segments in a video and the relationship among them. 2) The previous methods may cause modality bias [10] due to their ineffective fusion of the

information from different modalities.

To address the above issues, we propose a novel interactive-enhanced cross-modal perception method that leverages the advantage of audio and visual modality. In detail, two stages are involved. Firstly, we propose a segment-based attention (SA) module, which can effectively learn the importance of each segment and capture the relationship between different segments in the whole video. Secondly, we design a cross-modal aggregation (CMA) block, which can enrich feature representation and enhance the ability of the model to parse video. Experimental results on the benchmark dataset show that our method achieves significant improvements as compared with the baseline method, where the event-level audio-visual event metric is improved from 48.0% to 51.3%.

The remainder of this paper is organized as follows. The next section introduces the SA module and CMA block we proposed for efficient video parsing. Section 3 presents the experimental settings and the evaluation results. Conclusion and future directions are given in Section 4.

2. PROPOSED METHOD

2.1. Problem Statement

The AVVP task aims to identify the event of every segment into audio event, visual event and audio-visual event, together with their classes. When we input an audio-visual video sequence with T seconds, we regard the video sequence as divided into T segments with each segment lasting for one second long, expressed as $S = \{A_t, V_t\}_{t=1}^{T}$, where A and V denote the audio and visual segment pairs in time t. We use $y_t^a \in \mathbb{R}^C$, $y_t^v \in \mathbb{R}^C$ and $y_t^{av} \in \mathbb{R}^C$ to represent the audio, visual and audio-visual event labels at time t, where C is the number of event categories. The audio-visual event occurs when the audio event and visual event happen at the same time, which means $y_t^{av} = y_t^a * y_t^v$. Noted that we only have weak labels for training, but have detailed event labels with temporal boundaries in both modalities for evaluation.

2.2. Segment-based Attention

As shown in Fig. 2, we obtain the audio and visual features by using pre-trained audio and visual encoders, denoted as $\{f_t^a\}_{t=1}^{T}$, $\{f_t^v\}_{t=1}^{T}$. These features are firstly aggregated by the hybrid attention network (HAN) [7], which utilizes self-attention and cross-modal attention to obtain the intra- and cross-modality information. These multi-head attention blocks δ_{attn} can be described as follows:

$$\delta_{attn}(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}})V \tag{1}$$

where Q, K, V are query, key and value, with d being the dimension of the vector Q. The process of obtaining features from HAN can be denoted as:

$$\hat{f}_t^{ha} = f_t^{ha} + \delta_{attn}(f_t^{ha}, F^{ha}, F^{ha}) + \delta_{attn}(f_t^{ha}, F^{hv}, F^{hv}) \tag{2}$$

$$\hat{f}_t^{hv} = f_t^{hv} + \delta_{attn}(f_t^{hv}, F^{hv}, F^{hv}) + \delta_{attn}(f_t^{hv}, F^{ha}, F^{ha}) \tag{3}$$

where f_t^{ha} and f_t^{hv} are the feature vectors at a specific time t extracted from audio and visual encoders. F^{ha} and F^{hv} stand for the feature set in the same video, which are defined as $F^{ha} = \{f_1^{ha}, ..., f_T^{ha}\} \in \mathbb{R}^{T\times d}$ and $F^{hv} = \{f_1^{hv}, ..., f_T^{hv}\} \in \mathbb{R}^{T\times d}$. $\hat{f}_t^{ha}$ and $\hat{f}_t^{hv}$ are aggregated features obtained from HAN, and d is the feature dimension which is set to 512 in this paper.

The HAN module considers holistic feature information but ignores the features from different segments [7]. To address this limitation, we propose the segment-based attention (SA) module to obtain fine-grained feature information, which is similar to channel attention mechanism [11], which can not only selectively enhance or weaken different segments to highlight important feature information, but also assist the model in capturing feature relationships among different segments.

As shown in Fig. 3, in the SA module, firstly, the average feature representation is computed along the segment dimension, then, these representations are input into a well-defined neural network to generate the segment-based attention weight matrix W_t^a and W_t^v:

$$W_t^a = \varphi(\sum_{s=1}^{T}(\phi_t^a)) \tag{4}$$

$$W_t^v = \varphi(\sum_{s=1}^{T}(\phi_t^v)) \tag{5}$$

where ϕ_t^a and ϕ_t^v denote audio and visual features within a video, whose dimensions are (b, s, d), representing batch size, segment index, and dimension, respectively, and φ indicates a neural network that includes several linear layers and activation functions (e.g. $ReLU$ and $Sigmoid$). These attention weights enable us to modulate the feature representations of each segment according to their local importance. Afterward, we multiply the input features with the attention weights to obtain refined features $\tilde{f}_t^a$ and $\tilde{f}_t^a$ in terms of their importance,

$$\tilde{f}_t^a = \phi_t^a * W_t^a \tag{6}$$

$$\tilde{f}_t^v = \phi_t^v * W_t^v \tag{7}$$

Subsequently, these features are input into the HAN module to do further aggregation to get $\hat{f}_t^{sa}$ and $\hat{f}_t^{sv}$.

2.3. Cross-modal Aggregation

Two cross-modal aggregation (CMA) blocks are used to facilitate the learning of the correlations between audio and visual features. As shown in formula (2) and (3), previous method [7, 9] only used single modality as the input of vector K and V, which leads to sub-optimal cross-modal fusion results. In contrast, we concatenate audio and visual features, and then, the single-modality features are brought closer to the fused features. This can mitigate the impact of modality

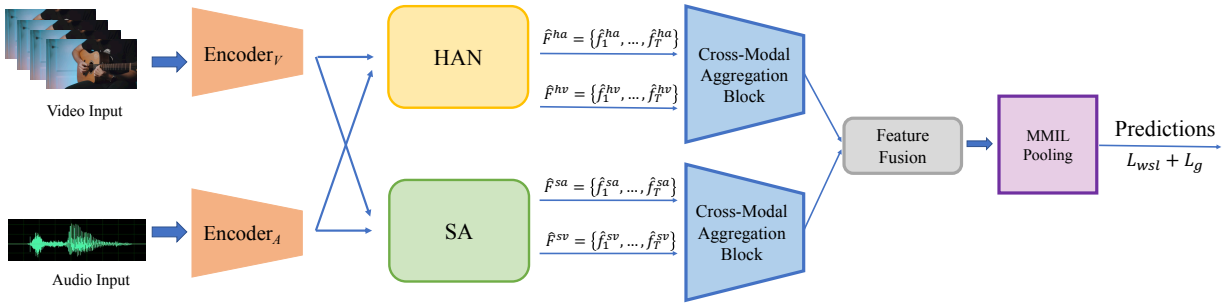

Fig. 2. The pipeline of our proposed interactive-enhanced cross-modal perception model (CM-PIE). It uses pre-trained encoders to extract audio and visual features. Firstly, we used two attention-based modules to learn fine-grained information. Then two cross-modal aggregation blocks are used to improve feature representation. Finally, multi-modal fusion is exploited and using MMIL Pooling to get the video-level event prediction.

bias and enhance the effectiveness of cross-modal information fusion:

$$\hat{g}_t^{ha} = \delta_{attn}(\hat{f}_t^{ha}, \hat{F}^{ha} \oplus \hat{F}^{hv}, \hat{F}^{ha} \oplus \hat{F}^{hv}) \quad (8)$$

$$\hat{g}_t^{hv} = \delta_{attn}(\hat{f}_t^{hv}, \hat{F}^{ha} \oplus \hat{F}^{hv}, \hat{F}^{ha} \oplus \hat{F}^{hv}) \quad (9)$$

$$\hat{g}_t^{sa} = \delta_{attn}(\hat{f}_t^{sa}, \hat{F}^{sa} \oplus \hat{F}^{sv}, \hat{F}^{sa} \oplus \hat{F}^{sv}) \quad (10)$$

$$\hat{g}_t^{sv} = \delta_{attn}(\hat{f}_t^{sv}, \hat{F}^{sa} \oplus \hat{F}^{sv}, \hat{F}^{sa} \oplus \hat{F}^{sv}) \quad (11)$$

where $\hat{F}^{ha}$ and $\hat{F}^{hv}$ are the sets of aggregated audio and visual feature obtained from HAN as defined in the previous section, $\hat{F}^{sa}$ and $\hat{F}^{sv}$ are the sets of the audio and visual features obtained from SA, $\hat{g}_t^{ha}$ and $\hat{g}_t^{hv}$ are the features obtained after applying the CMA block, and $\oplus$ means concatenating operation. In the same way, $\hat{g}_t^{sa}$, $\hat{g}_t^{sv}$ are the aggregation features derived from the SA module. Subsequently, we perform feature fusion operations as follows:

$$\tilde{g}_t^{a} = Mean(\hat{g}_t^{ha}, \hat{g}_t^{sa}) \quad (12)$$

$$\tilde{g}_t^{v} = Mean(\hat{g}_t^{hv}, \hat{g}_t^{sv}) \quad (13)$$

where *Mean* denotes taking the average of the two vectors element-wise. With the feature $\tilde{g}_t^{a}$ and $\tilde{g}_t^{a}$, we can obtain the segment-wise event prediction, which can be turned into video-level predictions by a pooling method, such as MMIL Pooling [7] based on a shared fully-connected layer and an activation function.

3. EXPERIMENTAL RESULTS

3.1. Experiment Setup

Dataset. The LLP dataset [7] is used to evaluate our method. This dataset has 11849 videos with 25 categories taken from YouTube, containing various scenes and species. The dataset has 10000 videos with weak labels as the training set, 1200 videos and 649 videos as the testing set and the validation set with fully annotated labels. Each video has 10 segments and each segment lasts 1 second.

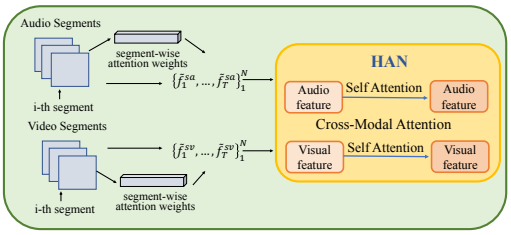

Fig. 3. Details of segment-based attention (SA) module. This module can learn the importance of each segment by calculating the segment-wise attention weights.

Implementation Details. We use pre-trained VGGish [13] to get 128-D audio features, and use 2D and 3D ResNet [14, 15] to extract 512-D visual features. The concatenation of 2D and 3D visual features is subsequently fed through a multi-layer perceptron (MLP) to generate segment-wise representations.

Evaluation Metrics. Following [7], we evaluate the performance of the proposed methods using F-scores, calculated at both segment-level and event-level. For the segment-level performance, we compute the F-score for each segment. The F-scores for the Audio, Visual and AV columns in Table 1 were calculated by comparing the predictions with the ground truth annotations for audio, visual and audio-visual sequence, and averaged for all the video sequences. With Ty@AV, we evaluate the overall performance by averaging the results from Audio, Visual, and AV columns. With Ev@AV, we evaluate the performance of the models for event classification by averaging of the F-scores calculated for each event, i.e. comparing the prediction results for each event and its ground truth annotation along the segments in the video sequences. For the event-level performance, we first concatenate the

Table 1. Comparison with the state-of-the-art methods on the LLP dataset in terms of F-scores. The event-level F-scores are calculated using a threshold of mIoU = 0.5.

Method	Segment-level					Event-level				
	Audio	Visual	AV	Ty@AV	Ev@AV	Audio	Visual	AV	Ty@AV	Ev@AV
AVE [5]	47.2	37.1	35.4	39.9	41.6	40.4	34.7	31.6	35.5	36.5
AVSDN [12]	47.8	52.0	37.1	45.7	50.8	34.1	46.3	26.5	35.6	37.7
HAN [7]	60.1	52.9	48.9	54.0	55.4	51.3	48.9	43.0	47.7	48.0
MM-Pyramid [8]	60.9	54.4	51.8	55.1	**57.6**	52.7	50.0	44.4	49.9	50.5
CM-CS+HAN [9]	57.1	**57.6**	**52.9**	**55.9**	53.9	49.4	**54.2**	**46.7**	**50.1**	47.8
CM-PIE (Ours)	**61.7**	55.2	50.1	55.7	56.8	**53.7**	51.3	43.6	49.5	**51.3**

positive segments in a sequential order, then calculate the F-score for each event, finally take the average F-score for all the events. The event-level evaluations also take into account the accuracy of the onset and offset time of event occurrence.

3.2. Comparison with State-of-the-art Method

We compare our method with several popular baselines, including HAN [7], MM-Pyramid [8] and CM-CS [9]. We also compare our method to some modified audio-visual event localization methods, including AVE [5] and AVSDN [12]. As shown in Table 1 , the proposed method shows improved performance due to adding the SA module and CMA block. The outcomes demonstrate a significant performance improvement of our model over the baseline method HAN across all evaluation metrics. Compared with HAN, our model is improved in both single and multi-modal metrics. For example, it achieves up to a 1.6% improvement in Audio & Segment-level and 2.3% enhancement in Visual & Segment-level. What's more, our model achieves a 3.3% improvement in the Ev@AV & Event-level metric. This confirms that accurate event localization at different segments can be achieved by learning useful information from important segments and the fusion of features of different modalities, thereby improving video parsing performance. Our method achieves state-of-the-art results on some indicators, such as Audio & Segment-level and Ev@AV & Event-level, and the model also shows promising results on other performance indicators.

3.3. Ablation Study

We investigate the influence of each part within the proposed approach and the results are shown in Table 2. We notice that both the SA module and the CMA block can improve the experimental results in several metrics. By learning from crucial segments, we have achieved notable performance enhancements in both visual and audio-visual evaluations, especially for audio events. This module can address the potential limitation described earlier of relying solely on aggregated features from the HAN and obtain a more precise understanding of the audio content and temporal information. Furthermore, the usage of the CMA block significantly improves the results in terms of Ty@AV and Ev@AV evaluations. Given that the Ev@AV evaluation considers the F-score for all audio and visual events, the enhancement in Ev@AV further indicates the substantial improvement introduced by the proposed method in audio-visual parsing.

Table 2. Ablation study on the LLP dataset. Seg Attention denotes adding the SA module. *w/o. V* and *w/o. A* means using only the CMA block for audio features or visual features.

Segment-level					
Methods	Audio	Visual	AV	Ty@AV	Ev@AV
HAN [7]	60.1	52.9	48.9	54.0	55.4
+Seg Attention	60.7	**55.5**	48.6	54.9	56.1
w/o. V	60.4	55.3	**51.2**	55.6	56.1
w/o. A	61.5	54.8	50.0	55.4	**57.0**
CM-PIE (Ours)	**61.7**	55.2	50.1	**55.7**	56.8
Event-level					
HAN [7]	51.3	48.9	43.0	47.7	48.0
+Seg Attention	53.2	49.8	42.1	48.3	50.5
w/o. V	52.2	50.0	**43.8**	48.6	49.3
w/o. A	53.0	51.1	43.5	49.2	50.9
CM-PIE (Ours)	**53.7**	**51.3**	43.6	**49.5**	**51.3**

4. CONCLUSION

In this paper, we have presented a novel weakly-supervised audio-visual video parsing framework. Two modules are introduced to leverage the segment relationships and semantics across the modalities. The segment-based attention module extracts local features from segments, which can learn more fine-grained semantics in a video. The cross-modal aggregation block effectively reduces modality bias, facilitating the effective fusion of the cross-modal information. Our approach has achieved promising results on the LLP dataset. In future work, we will further study the relationship between different segments across the video sequence.

5. ACKNOWLEDGMENT

This work was partly supported by a research scholarship from the China Scholarship Council (CSC).

6. REFERENCES

[1] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: a survey," *Multimedia Tools and Applications*, vol. 80, pp. 9411–9457, 2021.

[2] Y. Wei, D. Hu, Y. Tian, and X. Li, "Learning in audio-visual context: A review, analysis, and new perspective," *arXiv preprint arXiv:2208.09579*, 2022.

[3] R. Gao and K. Grauman, "Co-separating sounds of visual objects," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3879–3888.

[4] G. Li, Y. Wei, Y. Tian, C. Xu, J.-R. Wen, and D. Hu, "Learning to answer questions in dynamic audio-visual scenarios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 108–19 118.

[5] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 247–263.

[6] Y. Wu, L. Zhu, Y. Yan, and Y. Yang, "Dual attention matching for audio-visual event localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6292–6300.

[7] Y. Tian, D. Li, and C. Xu, "Unified multisensory perception: Weakly-supervised audio-visual video parsing," in *Proceedings of the 16th European Conference on Computer Vision, Glasgow, UK, August 23–28, 2020*. Springer, 2020, pp. 436–454.

[8] J. Yu, Y. Cheng, R.-W. Zhao, R. Feng, and Y. Zhang, "MM-Pyramid: Multimodal pyramid attentional network for audio-visual event localization and video parsing," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 6241–6249.

[9] H. Chen, D. Zhu, G. Zhang, W. Shi, X. Zhang, and J. Li, "CM-CS: cross-modal common-specific feature learning for audio-visual video parsing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[10] B. Odegaard, D. R. Wozny, and L. Shams, "Biases in visual, auditory, and audiovisual perception of space," *PLoS Computational Biology*, vol. 11, no. 12, p. e1004649, 2015.

[11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[12] Y. Lin, Y. Li, and Y. F. Wang, "Dual-modality seq2seq network for audio-visual event localization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2002–2006.

[13] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 131–135.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[15] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.