EasyChair Preprint

№ 11812

Cross-Modal Attention Network for Audio-Visual Event Localization

Han Liang, Jincai Chen, Jiangfeng Zeng, Tianming Jiang, Zheng Cheng and Ping Lu

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

January 19, 2024

Cross-modal Attention Network for Audio-Visual Event Localization

Han Liang[1], Jincai Chen[1,2], Jiangfeng Zeng[3] *, Tianming Jiang[3], Zheng Cheng[3], and Ping Lu[2]

[1] Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, China
{hanliang,jcchen}@hust.edu.cn
[2] School of Computer Science and Technology, Huazhong University of Science and Technology, China
luping06@hust.edu.cn
[3] School of Information Management, Central China Normal University, China
{jfzeng,tmjiang,chengzheng}@mails.ccnu.edu.cn

Abstract. Audio-visual event localization (AVEL) has been a hot research topic of computational scene analysis and machine perception, whose aim is to forecast the precise temporal segment within a video encompassing an audio-visual event, along with its corresponding categorical classification. A pivotal factor for achieving accurate audio-visual event localization lies in the effective fusion of multi-modal features. In this paper, we propose an innovative cross-modal attention network based on the self-attention mechanism, whose primary objective revolves around the extraction of important audio and visual features, subsequently fusing them effectively to yield a highly efficient representation. Specifically, we propose a Dynamic Intra- and Inter-modality Attention (DIIA) module, which cyclically facilitates the exchange of dynamic information within and across the domains of audio and visual modalities. Furthermore, we utilize audio features as guidance to direct the model to focus on the event-relevant visual regions. We validate our proposed method on the AVE Dataset and the extensive experiments demonstrate its superiority over state-of-the-art methods in supervised AVE settings.

Keywords: Audio-visual event localization · Cross-modal · Dynamic attention · Intra- and Inter-modality attention.

1 Introduction

Event localization is key for intelligent agents that perceive and understand the environment and become an increasingly important research area, whose aim is to determine whether the input video has an event and predict what category the event belongs to. It has a range of significant applications such as automatic surveillance and monitoring [1–3], improved human-machine interaction [4], and

* Corresponding author

media retrieval [5, 6]. Most of the early studies mainly focus on sound event localization and achieve promising performance. However, visual association can also provide valuable clues for recognizing and understanding the acoustic activities occurring around us. Inspired by this, Tian et al. [7] realized event localization is a multi-modal task and first introduced the audio-visual event localization (AVEL) task, which is addressed in this paper.

The primary objective of the AVEL task is to pinpoint the temporal boundaries of events within video sequences and entail categorizing these events. These video sequences encompass both audio and visual tracks. Illustrated in Fig. 1, for instance, when confronted with a video and the event category being "dog bark," the AVEL task necessitates the prediction of which specific video segments have the pertinent audio and visual signals corresponding to the occurrence of a dog bark.

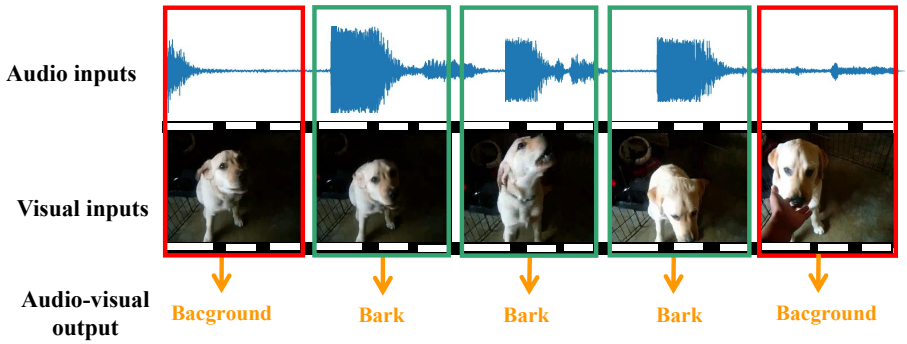

Fig. 1. Illustration of audio-visual event localization (AVEL) task. In the case of a video with a "dog bark" audio-visual event category, the goal of the AVEL task is to determine which video segments contain both the auditory and visual information of the "dog bark". For labeling the segments, segment-wise label is assigned as "dog bark" only when the event is both audible and visible. Otherwise, the segment is labeled as "background".

AVEL can be viewed as a cross-modality learning task, which aims to train a single model by using the audio and visual modalities simultaneously. The main problem of AVEL is how to extract and combine the information effectively carried by audio and visual modalities. Recently, many works [7–12] have explored various methods to track this task.

These works model the temporal information within the modality only considering their own modality, but some information from the other modality is not considered. Such cases motivate us to develop a new framework called Dynamic Intra- and Inter-modality Attention (DIIA) for precise event localization through efficient multi-modality feature fusion. The complete architecture is depicted in Fig.2. Our DIIA framework utilizes self-attention and cross-modal co-attention

mechanisms to merge relevant information within and across the audio and visual modalities, resulting in effective feature fusion.

2 Related works

In recent years, notable advancements [8–15] have emerged in the field of AVEL. To systematically investigate this task, Tian et al. [7] took a pioneering step by curating an AVE dataset that encompasses a diverse range of events and introduced an audio-guided visual attention mechanism, a tool designed to direct the model's focus towards informative visual regions. Furthermore, they proposed a dual multimodal residual network for efficient fusion of audio and visual information. Wu [8] enhanced the representation of high-level event information across extended video durations by a dual attention matching module. Simultaneously, the integration of a global cross-check mechanism allowed for the extraction of local temporal details, enhancing the model's grasp of temporal relationships. Lin [9] introduced a cross-modality co-attention network that employed an audio-visual transformer to facilitate the exploitation of both intra- and inter-frame information. Xu [10] proposed the utilization of an audio-guided spatial-channel attention module and a relation-aware module. This framework effectively captured intra- and inter-modality relations, further enriching the model's capacity to discern event patterns. Zhou [11] proposed a positive sample propagation module that evaluated the relationship between audio-visual pairs using a similarity map. Yu et al. [15] developed a multimodal parallel network, a novel approach that leverages the power of two parallel subnetworks to independently capture global and local semantics information to significantly amplify both classification and localization.

Motivated by the need for more efficient multi-modality feature fusion, we propose a novel Dynamic Intra- and Inter-modality Attention (DIIA) framework, which integrates self-attention and cross-modal co-attention mechanisms. These mechanisms work harmoniously to facilitate robust information fusion, both within individual audio and visual modalities and across them. This novel approach not only elevates the effectiveness of feature integration but also significantly contributes to the overarching goal of more comprehensive and accurate multi-modal analysis. The overall architecture is illustrated in Fig. 2.

3 Method

3.1 Preliminaries

The objective of AVEL is twofold: predicting the temporal boundaries of audio-visual events and categorizing each segment's event type in a given video sequence containing both audio and visual tracks. Generally, an input video is split into non-overlapping segments of equal duration labeled as $\{A_t, V_t\}_{t=1}^{T}$, with A_t representing audio content and V_t indicating visual content. Each video segment is assigned an event label denoted as y_t, shown as a binary vector with

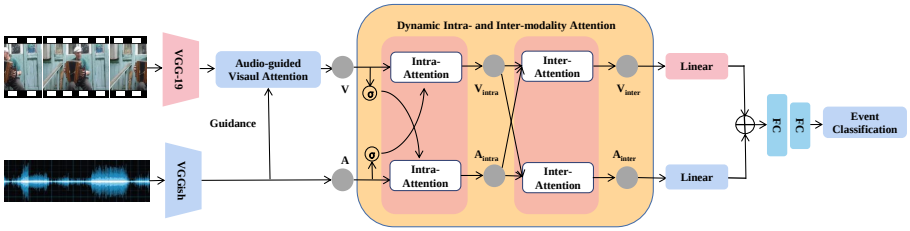

Fig. 2. Proposed framework for audio-visual event localization.

C elements. Each element indicates if a specific event is present in the segment. Notably, C covers all distinct events in the AVE dataset, including an extra category for background.

In the supervised AVEL task, the ground truth for each audio segment A_t or visual segment V_t is known during the training phase. During the testing phase, the event label for each segment is required to be predicted by the model.

3.2 Overall Pipeline

The proposed approach comprises three modules, as depicted in Fig. 2. Specifically, the first module is responsible for feature extraction, where pre-trained CNNs are employed to extract visual and audio features. To direct the model's attention towards visually relevant regions of events, we employ the audio modality to guide the extraction of visual features in spatial and channel dimensions [7]. The second module, referred to as Dynamic Intra- and Inter-modality Attention (DIIA), is designed to capture the intra- and inter-modal relationships within and between audio and visual features. The DIIA module is capable of effectively learning these relationships. In the final classification module, the audio and visual features are fused after passing through the DIIA module, which comprises of several fully connected layers.

3.3 Dynamic Intra- and Inter-modality Attention

To capture both the correlations within each modality and between the audio and visual modalities, we introduce the DIIA module that incorporates dynamic intra-modality attention and inter-modality attention. The dynamic intra-modality attention is implemented using a self-attention mechanism. This allows the model to focus on different parts of the input data within each modality, identifying relevant patterns and relationships. By dynamically adjusting the attention weights, the model can adaptively capture dependencies and correlations within the audio and visual data streams. Simultaneously, inter-modality attention facilitates the cross-modal attention mechanism. This attention mechanism enables the model to attend to and integrate relevant information from both the audio and visual modalities. By doing so, the model can effectively

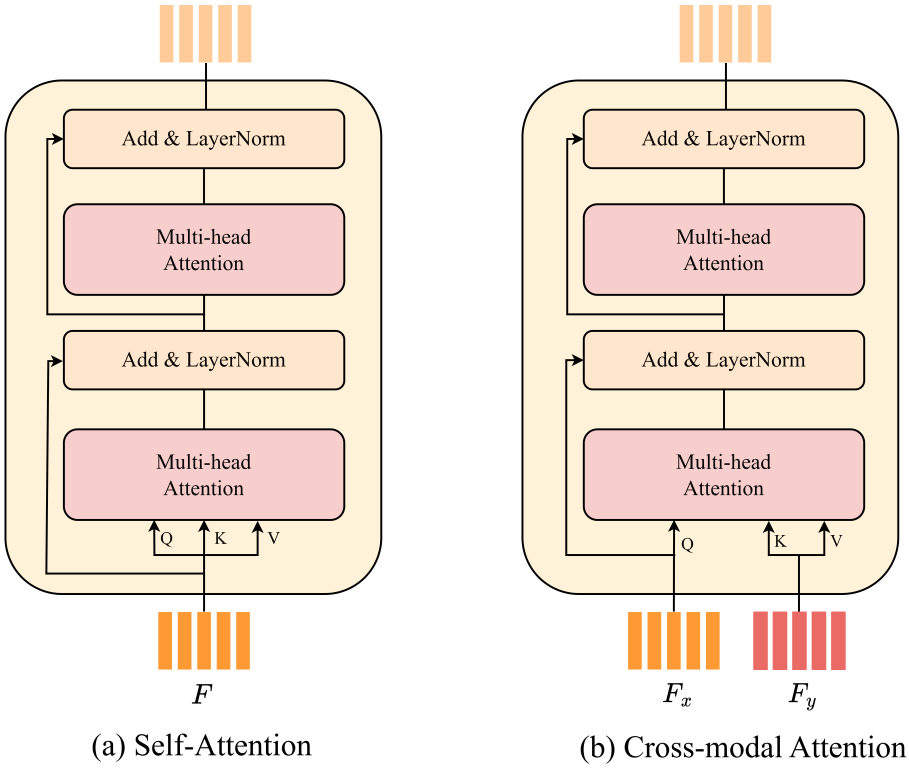

Fig. 3. The structures of the Self-Attention block and the Cross-Modal Attention block

leverage complementary cues from each modality, enhancing its ability to learn complex multimodal relationships. Fig. 3 provides a visualization of the proposed self-attention and cross-modal attention blocks, which share some structural similarities with those introduced in [16]. However, our DIIA module incorporates dynamic aspects that make it particularly suited for capturing temporal and spatial dependencies within and between modalities.

Multi-Head Attention (MHA) Our approach utilizes MHA to enable the model to simultaneously attend to information from diverse representation subspaces across different positions. In our specific context, we leverage MHA to implement our concept. This attention mechanism involves an intricate interplay between queries and a collection of key-value pairs, yielding a meaningful output. Importantly, all elements involved in queries, keys, values, and outputs are vectors.

$$Att(Q, K, V) = softmax(\frac{QW_Q(KW_K)^T}{\sqrt{d_k}})VW_V, \tag{1}$$

where Q, K, and V denote the query, key, and value matrices respectively; W_Q, W_K, W_V, and d_k represent learnable parameters of linear transformation

along with a scaling factor. Notably, n represents the number of heads in MHA. This equation is particularly pertinent in the context of intra-modality attention where the query, keys, and values all stem from the same input.

Subsequent to the MHA, our model integrates a fully connected feed-forward network (FFN), involving two linear transformations separated by a ReLU activation:

$$FFN(x) = ReLU(xW_1 + b_1)W_2 + b_2. \tag{2}$$

Here, W_1 and W_2 signify transformation matrices, while b_1 and b_2 denote bias terms.

Dynamic Intra-modality Attention The dynamic intra-modality attention module investigates methods for appropriately combining the knowledge acquired from two modalities, aiding in the training of the AVEL model. To accomplish this, we employ the formula below to learn the intra-relationship of audio and visual features:

$$V_{intra} = \text{FFN}(\text{MHA}(\hat{V}; \hat{V}; V), \tag{3}$$

$$A_{intra} = \text{FFN}(\text{MHA}(\hat{A}; \hat{A}; A)). \tag{4}$$

To better promote the learning of intra-relationship, we further design a conditional gate operation denoted as G. This operation is devised to update queries and keys, drawing inspiration from [17, 18]. The procedure is outlined as follows:

$$\hat{V} = (1 + G_A) \odot V, \tag{5}$$

$$\hat{A} = (1 + G_v) \odot A. \tag{6}$$

Here, the symbol $\odot$ signifies element-wise multiplication. The conditional gate operations, represented as G_V and G_A, are defined as follows:

$$G_V = \sigma(Avg\_pool(V)W_V), \tag{7}$$

$$G_A = \sigma(Avg\_pool(A)W_A). \tag{8}$$

In these equations, σ denotes the sigmoid function, while Avg_pool represents the process of average pooling.

Inter-modality Attention In the pursuit of achieving higher quality representations of audio and visual features, we design an inter-modality attention module. Taking visual features as an example, visual features are utilized as queries, and audio features play the role of both keys and values. Consequently, what emerges is a collection of attended audio features tailored to the visual features. The same as audio features. This process can be represented using the following formula:

$$V_{inter} = \text{FFN}(\text{MHA}(V_{intra}; A_{intra}; A_{intra})), \tag{9}$$

$$A_{inter} = \text{FFN}(\text{MHA}(A_{intra}; V_{intra}; V_{intra})). \tag{10}$$

The resulting V_{inter} and A_{inter} provide an enriched perspective on the intricate relationships linking audio and visual features, which significantly enhance the model's grasp of cross-modal interactions.

3.4 Classification and Objective function

Before classification, we fuse audio and vision features through simple averages. The fusion feature F_{va} can be obtained through:

$$F_{va} = \frac{1}{2}(V_{inter}W^{v}_{inter} + A_{inter}W^{a}_{inter}), \tag{11}$$

where W^{v}_{inter} and W^{a}_{inter} represent learnable parameters in the linear layers.

Subsequently, the fused features undergo a series of transformations through two fully connected (FC) layers, which are then followed by the application of a softmax function. This process yields the classifier prediction, denoted as O_{tc}, indicating the model's assessment of the segment event category.

In the context of evaluating the model's prediction, the classifier's output O_{tc} is compared to the ground truth label Y_{tc}, which is used to determine how accurately the model's predictions align with the actual event categories. To quantify this alignment, the cross-entropy loss is employed as the chosen objective function. Mathematically, the cross-entropy loss (L_{CE}) is calculated as the negative average of the logarithmic differences between the ground truth labels and the classifier's predictions. The formula for the cross-entropy loss is as follows:

$$L_{CE} = -\frac{1}{TC}\sum_{t=1}^{T}\sum_{c=1}^{C} Y_{tc}log(O_{tc}), \tag{12}$$

where T is the temporal segments and C is the number of the event categories.

4 Experiments

4.1 Data Description and Evaluation Metrics

Dataset. AVE dataset is a collection of videos derived from AudioSet [19], encompassing 28 categories of events from various domains, including but not limited to speeches by men or women, barking dogs, racing cars, guitar playing, and church bells. Each video in the dataset lasts for 10 seconds and is temporally labeled with event boundaries. The dataset is split into three parts for training, validation, and testing, respectively, following the same distribution as [7].

Evaluation Metrics. The primary goal of the Audio-Visual Event Localization (AVEL) task is to accurately assign each video segment to its respective event category. Drawing from prior studies [7–9], we adopt the overall accuracy (Acc) as a crucial performance metric to evaluate our model's effectiveness in this task. The overall accuracy metric provides a comprehensive assessment by taking into consideration several factors, including true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The computation of the overall accuracy is given by the following formula:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}. \tag{13}$$

In this equation, TP represents the count of event segments accurately identified, TN stands for the count of non-event segments correctly identified, FP indicates the count of non-event segments incorrectly labeled as events, and FN denotes the count of event segments mistakenly classified.

4.2 Experimental Settings

We employ the same extraction method following the previous works[8–11]. Specifically, we use a VGG-like network [20] pre-trained on AudioSet to extract acoustic features with 128 dimensions for each segment. For visual features, we extract features of size $7 \times 7 \times 512$ for each segment using the VGG-19 network [21], pre-trained on ImageNet [22]. In the training phase, the Adam optimizer is utilized with a batch size of 128. We initialize the learning rate to 7×10^{-4} and apply a gradual decay strategy at epochs 10, 20, and 30, where the learning rate is reduced to 0.5. We implement it in PyTorch [23].

Table 1. Performance Comparison with Existing Approaches on AVE Dataset

Model	Fully-Supervised Acc
Audio	59.5
Visual	55.3
Audio + visual	71.4
AVEL [7]	72.7
DAM [8]	74.5
AVFB [24]	74.8
AVSDN [25]	75.4
CMRAN [10]	77.4
PSP [11]	77.8
Ours	**78.4**

4.3 Experimental Results and Analysis

We evaluate our proposed approach against several recent methods that adopt the same features for fully-supervised event detection on the AVE dataset, including AVEL [7], DAM [8], AVFB [24], AVSDN [25], CMRAN [10], and PSP [11]. Table 1 presents the experimental results of our method and the compared methods.

Our proposed model achieved superior results in supervised event detection when compared to the single-modality baselines proposed in [7], indicating the effectiveness of capturing audio-visual interactions. The results also demonstrate that modeling both intra- and inter-modality interactions is important for achieving better performance, as shown by the outperformance of methods that exploit

only one of these interaction types. In particular, our proposed DIIA enables the dynamic fusion of multi-modal features with both intra- and inter-modality information, resulting in the highest accuracy of 78.4% among all evaluated methods in supervised AVE settings.

5 Conclusion

This paper focuses on the problem of audio-visual event localization, and we present a novel cross-modal attention network that utilizes the self-attention mechanism to extract informative features from both the audio and visual modalities. The core component of our model is the Dynamic Intra- and Inter-modality (DIIA) module, which enables dynamic information flow within and across modalities. Unlike existing approaches, our DIIA can flexibly adjust the intra-modal attention and capture complex relationships within the audio or visual modality. Experimental results demonstrate the effectiveness of our proposed method.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grant No. 62272178, No. 62102159 and the Humanities and Social Science Fund of Ministry of Education of China under Grant No. 21YJC870002 and the Fundamental Research Funds for the Central Universities under Grant No. CCNU22QN017 and Wuhan Knowledge Innovation Special project under Grant No. 2022010801020287.

References

1. Greco, A., et al.: DENet: a deep architecture for audio surveillance applications. Neural Comput. Appl. 1-12 (2021)
2. Foggia, P., et al.: Audio Surveillance of Roads: A System for Detecting Anomalous Sounds. IEEE Trans. Intell. Transp. Syst. 17(1), 279-288 (2016)
3. Mnasri, Z., et al.: Anomalous sound event detection: A survey of machine learning based methods and applications. Multimedia Tools Appl. 81(4), 5537–5586 (2022)
4. Cech, J., et al.: Active-speaker detection and localization with microphones and cameras embedded into a robotic head. In: 13th IEEE-RAS International Conference on Humanoid Robots (Humanoids), pp. 203-210. (2013)
5. Jhuo, I. H., Lee, D. T.: Video Event Detection via Multi-modality Deep Learning. In: 22nd International Conference on Pattern Recognition, pp. 666-671. (2014)
6. Wang, Y., et al.: Exploring audio semantic concepts for event-based video retrieval. In: IEEE international conference on acoustics, speech and signal processing, pp. 1360-1364. (2014)
7. Tian, Y., et al.: Audio-visual event localization in unconstrained videos. In: Proceedings of the European conference on computer vision, pp. 247-263. (2018)
8. Wu, Y., et al.: Dual attention matching for audio-visual event localization. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 6291-6299. (2019)
9. Lin, Y., Wang, Y.: Audiovisual transformer with instance attention for audio-visual event localization, In: Proceedings of the Asian Conference on Computer Vision. (2020).

10. Xu, H., et al.: Cross-modal relation-aware networks for audio-visual event localization. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 3893–3901. (2020)
11. Zhou, J., et al.: Positive sample propagation along the audio-visual event line. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8432-8440. (2021)
12. Zheng, X., Wei, Y.: Audio-Visual Event and Sound Source Localization Based on Spatial-Channel Feature Fusion. In: 7th International Conference on Signal and Image Processing, pp. 106-110. (2022)
13. Brousmiche, M., Dupont, S., Rout. J.: Intra and inter-modality interactions for audio-visual event detection. In Proceedings of the 1st International Workshop on Human-centric Multimedia Analysis, pp. 5-11. (2020)
14. Ramaswamy, J.: What Makes the Sound?: A Dual-Modality Interacting Network for Audio-Visual Event Localization. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4372-4376. (2020)
15. Yu, J., Cheng, Y., Feng, R.: MPN: Multimodal Parallel Network for Audio-Visual Event Localization. In: IEEE International Conference on Multimedia and Expo, pp. 1-6. (2021)
16. Vaswani, A., et al.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing System, pp. 6000–6010. (2017)
17. Liu, F., et al.: Simnet: Stepwise image-topic merging network for generating detailed and comprehensive image captions. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 137–149. (2018)
18. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks, In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132-7141. (2018)
19. Gemmeke, J. F., et al.: Audio Set: An ontology and human-labeled dataset for audio events. s. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 776-780. (2017)
20. Hershey, S., et al.: CNN architectures for large-scale audio classification, In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 131-135. (2017)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd International Conference on Learning Representation, (2014)
22. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis., 115, 211-252 (2015)
23. Paszke, A., et al.: Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32, 8024–8035 (2019).
24. Ramaswamy, J., Das, S.: See the sound, hear the pixels. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2970-2979. (2020)
25. Lin, Y. B., et al.: Dual-modality Seq2Seq Network for Audio-visual Event Localization. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2002-2006. (2019)