Audio-Visual Event Localization via Recursive Joint Co-Attention

Bin Duan[1] Hugo Latapie[2] Gaowen Liu[2] Yan Yan[1]
[1]Illinois Institute of Technology [2]Cisco

Abstract

The major challenge in audio-visual event localization task lies in how to fuse information from multiple modalities effectively. Recent works have shown that the attention mechanism is beneficial to the fusion process. In this paper, we propose a novel joint attention mechanism with multimodal fusion methods for audio-visual event localization. Particularly, we present a concise yet valid architecture that effectively learns representations from multiple modalities in a joint manner. Initially, visual features are combined with auditory features and then turned into joint representations. Next, we make use of the joint representations to attend to visual features and auditory features, respectively. With the help of this joint co-attention, new visual and auditory features are produced, and thus both features can enjoy mutually improved benefits from each other. It is worth noting that the joint co-attention unit is recursive meaning that it can be performed multiple times for obtaining better joint representations progressively. Extensive experiments on the public AVE dataset have shown that the proposed method achieves significantly better results than the state-of-the-art methods.

1. Introduction

In this paper, we focus on the audio-visual event localization task. As shown in Fig. 1, an Audio-Visual Event (AVE) is defined in a video sequence that is both audible and visible. The audio-visual event localization task consists of two sub-tasks, one of which is to predict the event label while the other is to predict which segment of the video sequence has an audio-visual event of interest. As in the AVE definition, localizing an AVE must deal with heterogeneous information from both audio and visual modalities. Moreover, recent works [9, 16, 17] show that the performance after fusion outperforms the one that only uses a single modality. Although these approaches present interesting explorations, how to smartly fuse representations from both modalities is still a challenging task.

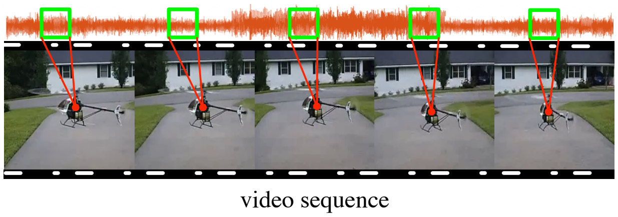

Figure 1. Audio-Visual Event (AVE) is an event both audible and visible. e.g., a person can see a helicopter in the visual sequence (the bottom row) and also hear the helicopter's engine sound in the audio sequence (the top row).

Multimodal fusion provides a global view of multiple representations for a specific phenomenon. To tackle the AVE localization problem, existing methods [9, 16] either fuse cell states out of LSTMs [16], or fuse both hidden states and cell states from LSTMs [9]. Both aforementioned approaches exploit a plain multimodal fusion strategy, where the fusion results might be unstable as it is hard to guarantee good quality of the information used for the fusion, e.g., some noisy information from the background segments may also be included. Therefore, a more robust fusion strategy is needed for better representations. Wu *et al.* [17] introduce a cross-modal matching mechanism that exploits global temporal co-occurrences between two modalities and excludes the noisy background segments from the sequence. Intuitively, having global features to interact with local features would help to localize the event, but it needs additional supervision to manually filter the background segments.

Motivated by the popular attention techniques [2, 5, 11, 14, 15, 18, 19], we propose a new Joint Co-Attention (JCA) mechanism which develops on the basis of self-attention and co-attention. We utilize the joint representation to generate the attention masks for two uni-modalities while previous methods [10, 12] independently generate attention mask for each other. In our approach, instead of using features from one single modality, each attention mask is generated using features from both modalities and thus it is more informative. As a result, each modality is attended not only by the features from itself (self-attended), but also by the features from the other modality (co-attended). Extensive experiments show the superiority of our proposed joint co-attention learning framework.

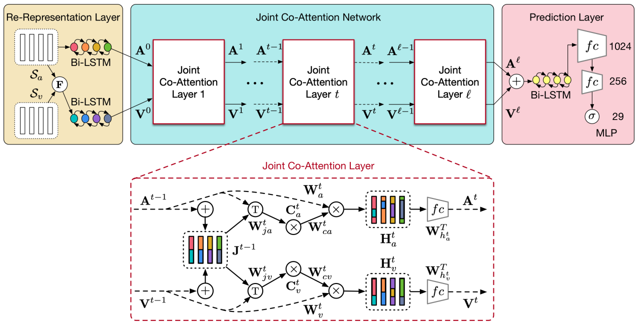

Figure 2. The overall structure of the proposed framework. We split it into three parts, i.e., sequence feature re-representation layer, joint co-attention network and category prediction layer. For the symbols, $\oplus$ denotes concatenation, Ⓕ denotes early fusion of audio feature and visual feature, $\textcircled{\sigma}$ denotes the softmax function, Ⓣ is transpose operator, and $\otimes$ is matrix multiplication operator.

2. Joint Co-Attention Network

An Audio-Visual Event (AVE) is defined as an event that is both visible and audible [16]. As in [16, 17], for a given audio-visual video sequence $\mathcal{S} = (\mathcal{S}_a, \mathcal{S}_v)$, while $\mathcal{S}_a$ denotes the audio portion and $\mathcal{S}_v$ denotes the visual portion. The video sequence $\mathcal{S}$ is split into N non-overlapping yet continuous segments where each segment is typically one second long. For each segment, a label $y \in \{0, 1\}$ is given, while 0 indicates the segment is background and 1 indicates that is an AVE. The sequence features, i.e., $\mathcal{S}_a$ and $\mathcal{S}_v$ are extracted using a pre-trained CNN. We denote the extracted segment-level feature as s_a^t and s_v^t corresponding to the audio and visual modality respectively, where $t \in \{1, 2, \cdots, N\}$. Our network is built on the basis of fixed s_a^t and s_v^t, and the architecture is shown in Fig. 2.

The proposed joint co-attention layer attends to visual features and audio features simultaneously. It takes the audio representation $\mathbf{A}$ and the visual representation $\mathbf{V}$ as inputs and concatenates two representations as the joint representation $\mathbf{J}$. We employ $\mathbf{J}$ to co-attend to $\mathbf{A}$ and $\mathbf{V}$, respectively. It is worth noting that we only preserve $\mathbf{J} \rightarrow \mathbf{A}$ (i.e., joint feature attend to audio feature) and $\mathbf{J} \rightarrow \mathbf{V}$ (i.e., joint feature attend to visual feature), the inverse directions of $\mathbf{A} \rightarrow \mathbf{J}$ and $\mathbf{V} \rightarrow \mathbf{J}$ are abandoned for simplicity, which is different from the original co-attention mechanism [10]. One property of JCA is mutual attention, that is, it can attend to features from two different modalities simultaneously. Another special property of JCA is stackability, i.e., we can stack several JCAs so that we can recursively perform the process multiple times.

Primary Idea for Joint Co-Attention. Recent studies [10, 12] explore the co-attention theory in Visual Question Answering (VQA). The text sequence representations and the visual sequence representations attend mutually to obtain new representations. Inspired by this, we explore a mode that allows representation from one modality not only attending to the other representation from the other modality but also attending to the representation from its original modality. Given audio representation $\mathbf{A} \in \mathbb{R}^{N \times d_a}$, and visual representation $\mathbf{V} \in \mathbb{R}^{N \times d_v}$, the joint representation $\mathbf{J} \in \mathbb{R}^{N \times d}$ is acquired by the concatenation of $\mathbf{A}$ and $\mathbf{V}$, i.e, $\mathbf{J} = [\mathbf{A}; \mathbf{V}]$ where $d = d_a + d_v$. We take audio feature $\mathbf{A}^\ell$ as an example to elaborate the process of joint co-attention. Here, we denote $\mathbf{A}^1$ as the initial state of audio feature and $\mathbf{A}^\ell$ as the audio feature after ℓ-th joint co-attention layer. First, the $(\ell-1)$-th layer's audio representation $\mathbf{A}^{\ell-1}$ is concatenated with $\mathbf{V}^{\ell-1}$ to obtain joint representation $\mathbf{J}^{\ell-1}$; next, we employ the $\mathbf{J}^{\ell-1}$ to attend to $\mathbf{A}^{\ell-1}$ and finally obtain the ℓ-th layer's audio feature $\mathbf{A}^\ell$. Similarly, the new visual feature $\mathbf{V}^\ell$ is obtained.

Learning to Joint Co-Attend. Fusion is one of the key challenges for multimodal learning [1]. Following recent studies [10, 12] in VQA, we specifically derive the fusion to fit our audio-visual event localization task. After calculating the joint representation matrix $\mathbf{J}$, we use it to attend to different uni-modal representations as $\mathbf{C}_a = \mathrm{Tanh}\left(\frac{\mathbf{A}^{\mathrm{T}}\mathbf{W}_{ja}\mathbf{J}}{\sqrt{d}}\right)$, where $\mathbf{C}_a$ is the joint-audio affinity matrix, T denotes transpose operation, and $\mathbf{W}_{ja} \in \mathbb{R}^{N \times N}$ is a learnable weight matrix ($\mathbf{W}_{ja}$ is implemented as fully-connected layer). Following the same rule, the joint-visual affinity matrix $\mathbf{C}_v$ can be written as $\mathbf{C}_v = \mathrm{Tanh}\left(\frac{\mathbf{V}^{\mathrm{T}}\mathbf{W}_{jv}\mathbf{J}}{\sqrt{d}}\right)$, where $\mathbf{W}_{jv} \in \mathbb{R}^{N \times N}$ is also a learnable weight matrix. After calculating the joint uni-modal affinity matrices $\mathbf{C}_a$ and $\mathbf{C}_v$, we then calculate the attention probabilities map $\mathbf{H}_a, \mathbf{H}_v$ of two modalities as, $\mathbf{H}_a = \mathrm{ReLU}\left(\mathbf{W}_a\mathbf{A} + \mathbf{W}_{ca}\mathbf{C}_a^{\mathrm{T}}\right)$ and $\mathbf{H}_v = \mathrm{ReLU}\left(\mathbf{W}_v\mathbf{V} + \mathbf{W}_{cv}\mathbf{C}_v^{\mathrm{T}}\right)$, where $\mathbf{H}_a \in \mathbb{R}^{k \times d_a}, \mathbf{H}_v \in \mathbb{R}^{k \times d_v}$ represent the attention probabilities map of audio modality and visual modality, respectively. $\mathbf{W}_a, \mathbf{W}_v \in \mathbb{R}^{k \times N}$, $\mathbf{W}_{ca}, \mathbf{W}_{cv} \in \mathbb{R}^{k \times d}$ are learnable weight matrices.

After obtaining the attention map $\mathbf{H}_a$ and $\mathbf{H}_v$, we recompute the new audio representation and new visual representation by

$$\mathbf{A}^\ell = g(\mathbf{A}^{\ell-1}, \mathbf{W}_{h_a^\ell}^{\mathrm{T}}\mathbf{H}_a^\ell), \mathbf{V}^\ell = g(\mathbf{V}^{\ell-1}, \mathbf{W}_{h_v^\ell}^{\mathrm{T}}\mathbf{H}_v^\ell), \quad (1)$$

where $\mathbf{W}_{h_a^\ell}, \mathbf{W}_{h_v^\ell} \in \mathbb{R}^{k \times N}$ are learnable weight matrices in the ℓ-th layer. $\ell-1$ represents the features produced by the $\ell-1$-th layer. In our case, g is a summation function.

Fusion by Fusion. Multimodal fusion can generate more robust representation using the features from multiple modalities that are collected for the same phenomenon. Earlier studies [9, 16, 17] particularly exploit the method in an audio-visual dual-modality setting either directly fusing the features or using cross dot product operation. Different from them, we consider multimodal fusion as a recursive process, where we fuse audio representation $\mathbf{A}$ and visual representation $\mathbf{V}$ recursively to obtain more robust representations. Following Eq. (1), we generalize this recursive process as

$$\mathbf{A}^{\ell} = g(\cdots g(\mathbf{A}^{0}, \mathbf{W}_{h_a^1}^{\mathrm{T}} \mathbf{H}_a^1) \cdots , \mathbf{W}_{h_a^{\ell}}^{\mathrm{T}} \mathbf{H}_a^{\ell}), \tag{2}$$

$$\mathbf{V}^{\ell} = g(\cdots g(\mathbf{V}^{0}, \mathbf{W}_{h_v^1}^{\mathrm{T}} \mathbf{H}_v^1) \cdots , \mathbf{W}_{h_v^{\ell}}^{\mathrm{T}} \mathbf{H}_v^{\ell}), \tag{3}$$

where ℓ represents the amount of times that the joint co-attention is repeated. After fusing ℓ times, we will obtain two more robust representations for audio and visual modality, respectively. The final fused audio and visual representations will be fed into the MLP prediction layer to predict the AVE category.

3. Experiments

Audio-Visual Event Dataset. The Audio-Visual Event (AVE) dataset by [16] is a subset of AudioSet [6]. It consists of $4,143$ video clips that involve 28 event categories. We adopt the split technology of [16] where train/validation/test sets are $3,309/402/402$ video clips, respectively. While training, the model has no access to the test portion to better evaluate the model's generalization ability. For the AVE dataset, it contains comprehensive audio-visual event types, in general, instrument performances, human daily activities, vehicle activities, and animal actions. To be more specific, for more detailed event categories, take instrument performances as an example, AVE dataset contains accordion playing, guitar playing, and ukulele playing, etc. A typical video clip is 10 seconds long and is labeled with the start point and endpoint at the segment level to clarify whether the segment is an audio-visual event.

Evaluation Metrics. We follow [9, 16, 17] and adopt the global classification accuracy obtained from the last prediction layer as the evaluation metric. For an input video sequence, our goal is to predict the category label for each segment. It is worth noting that the background category contains 28 backgrounds since each event category can have its own background so that it is hard to predict.

Experimental Details. Following [16, 17], we adopt pre-trained CNN models to extract features for each audio and visual segment. Specifically, we exploit the VGG19 [13] network pre-trained on ImageNet [4] as the backbone to extract segment-level visual features. Meanwhile, for the audio segment, we extract the segment feature using a Vggish network [7] which is pre-trained on AudioSet [6]. For a fair comparison, we use the same extracted features (i.e, audio and visual features) as used in [16,17]. In the training stage, the only supervision we exploit is the annotation labels for the temporal segments.

State-of-the-Art Comparison. Results compared with the leading methods are reported in Table 1. We take a similar model architecture as in [16] and run single modality models as our baselines, which only take audio features or visual features during the experiments. First, to validate the proposed method can enable efficient interactions between audio features and visual features, we compare with a state-of-the-art temporal labeling network, i.e, ED-TCN [8], which can integrate information from multiple temporal segments. Next, to verify the effectiveness of our fusion strategy of audio feature and visual feature, we compare with two methods, i.e, Audio-Visual [16] and AVSDN [9]. Both methods utilize a straightforward fusion strategy, where fuses the audio and visual features out of LSTMs by concatenation. Lastly, to evaluate that our method is tolerant with less supervision, we compare our method with DAM [17], which needs additional supervision to exclude event-irrelevant segments during training.

Table 1. Results of comparisons with the state-of-the-art methods on the AVE dataset. For a fair comparison, * is obtained by exploiting the same pre-trained audio and visual features. While the task is hard, it can still be observed that our model outperforms the existing methods.

Method	**Accuracy (%)**
Audio_Only (Vggish [7])	59.5
Visual_Only (Vgg19 [13])	55.3
ED-TCN [8]	46.9
Audio-Visual [16]	71.4
AVSDN* [9]	72.6
Full-Audio-Visual [16]	72.7
DAM [17]	74.5
Ours	**76.2**

Comparison Analysis. Due to the absence of interactions between audio modality and visual modality, our proposed model can easily surpass the performance of the baselines. In addition, by comparing with ED-TCN, our model enables more effective interactions between two modalities. Thus, it can be testified that interactions or fusion can boost the task performance and our model is more superior on enabling interactions between two different modalities. Unsurprisingly, by fusing the two different features using our joint co-attention mechanism, our model outperforms Audio-Visual and AVSDN using a plain fusion strategy. Moreover, even without additional effort to exclude event-irrelevant segments, our model can learn useful representations from noisy inputs and contribute to better performance.

Framework Decoupling Analysis. Results are showed in Table 2. First, the overall performance of the proposed framework outperforms the state-of-the-art method [17] which needs additional supervision. Among all the observed declines, Bi-LSTM has the highest impact. That confirms the effectiveness of the Bi-LSTM part. For alternatives to early fusion, neither the global average pooling nor the global max pooling surpasses our full model.

Among the experiment results with two different co-attention mechanisms, i.e., original co-attention method [12] and our joint co-attention method, our joint co-attention method excels the original co-attention method

Table 2. Ablation studies on the proposed framework. Uni-modal Bi-LSTM is the LSTM in sequence feature re-representation layer while Joint Bi-LSTM is the one in the prediction layer. * denotes we remove the residual embedding of LSTMs while † denotes that we adopt the primary co-attention mechanism into the proposed framework.

Model	**Accuracy (%)**
Ours w/o Uni-modal Bi-LSTM	74.5
Ours w/o Joint Bi-LSTM	74.9
Ours w/o Residual Embedding*	75.2
Ours w/ GRU [3]	75.3
Ours w/ Average Pooling	75.1
Ours w/ Max Pooling	75.0
Ours w/ Co-Attention† [12]	75.4
Ours w/ Joint Co-Attention	**76.2**

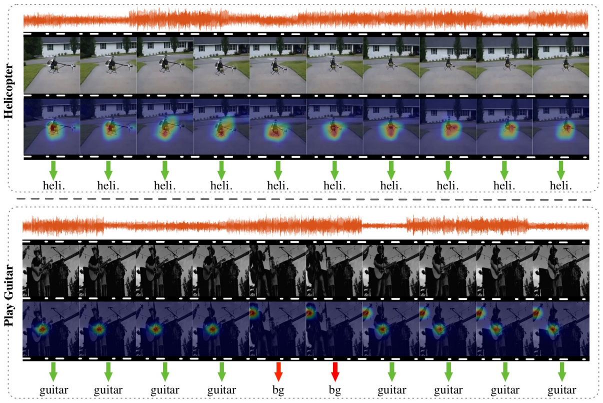

Figure 3. Two qualitative results on audio-visual localization task. The first example is helicopter hovering, i.e, 'heli.' is the abbreviation for helicopter for better layout; while second example is playing guitar, i.e., 'guitar' for short, 'bg' denotes 'background'. The green arrow represents the correct prediction whereas the red arrow denotes the wrong prediction. To visualize where they attend to, we generate images with their corresponding attention map. *Best viewed in color.*

which follows a dual-modality mutual attending way (visual features attend to audio features and audio features attend to visual features). By not only attending to the corresponding modality but also the modality of itself, our proposed joint co-attention method performs better in the audio-visual fusion task. To sum up, the ablation studies demonstrate the efficiency of our proposed framework.

3.1. Qualitative Evaluation

In this section, we show some qualitative results of our proposed framework in Fig. 3. For each row in Fig. 3, the left is the category of this audio-visual event; the top content is the waveform of input audio sequence; the middles are raw frames and frames with attention map of the input video sequence; the bottom is the audio-visual event prediction. Among the two instances in Fig. 3, the second instance is much harder as the scene is more complicated where different people are playing different instruments. In the beginning, the proposed network predicts well. However, as the singer changing his posture, the guitar can hardly be seen even with our eyes. Therefore, the network fails to predict it as playing guitar. Surprisingly, as the singer turns back to the front, our network works again, and it marks two guitars in the picture even the other guitar is indistinct.

4. Conclusion

In this paper, we investigate an interesting problem on deep audio-visual learning for the AVE task. To better cope with this multimodal learning task, we propose a novel joint co-attention mechanism with double fusion. To the best of our knowledge, this is the first time of applying the co-attention mechanism into the audio-visual event localization task. The integration with double fusion leading to better representations for the AVE task by co-attending to both audio and visual modalities. Moreover, experimental results on the AVE dataset have confirmed the superiority of the proposed framework.

Acknowledgements. This research was partially supported by NSF SCH-2123521, NeTS-2109982 and the gift donation from Cisco. This article solely reflects the opinions and conclusions of its authors and not the funding agents.

References

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *TPAMI*, 2018. 2

[2] X. Chen, Y. Bin, C. Gao, N. Sang, and H. Tang. Relevant region prediction for crowd counting. *Elsevier Neurocomputing*, 2020. 1

[3] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, 2014. 4

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3

[5] L. Ding, H. Tang, and L. Bruzzone. Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images. *IEEE TGRS*, 2020. 1

[6] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017. 3

[7] 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 *ICASSP*, 2017. 3

[8] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager. Temporal convolutional networks for action segmentation and detection. In *CVPR*, 2017. 3

[9] Y.-B. Lin, Y.-J. Li, and Y.-C. F. Wang. Dual-modality seq2seq network for audio-visual event localization. In *ICASSP*, 2019. 1, 2, 3

[10] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *NeurIPS*, 2016. 1, 2

[11] H. Nam, J.-W. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. In *CVPR*, 2017. 1

[12] D.-K. Nguyen and T. Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *CVPR*, 2018. 1, 2, 3, 4

[13] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2014. 3

[14] H. Tang, S. Bai, L. Zhang, P. H. Torr, and N. Sebe. Xinggan for person image generation. In *ECCV*, 2020. 1

[15] H. Tang, D. Xu, N. Sebe, Y. Wang, J. J. Corso, and Y. Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *CVPR*, 2019. 1

[16] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018. 1, 2, 3

[17] Y. Wu, L. Zhu, Y. Yan, and Y. Yang. Dual attention matching for audio-visual event localization. In *ICCV*, 2019. 1, 2, 3

[18] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci. Structured attention guided convolutional neural fields for monocular depth estimation. In *CVPR*, 2018. 1

[19] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. In *ICML*, 2019. 1