Cross-Modal Attention Network for Temporal Inconsistent Audio-Visual Event Localization

Hanyu Xuan,[1] Zhenyu Zhang,[1] Shuo Chen,[1] Jian Yang,[1,2] Yan Yan*[1]
[1]PCA Lab, the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[2]Jiangsu Key Lab of Image and Video Understanding for Social Security
{xuanhanyu, zhangjesse, shuochen, csjyang, yyan}@njust.edu.cn

Abstract

In human multi-modality perception systems, the benefits of integrating auditory and visual information are extensive as they provide plenty supplementary cues for understanding the events. Despite some recent methods proposed for such application, they cannot deal with practical conditions with temporal inconsistency. Inspired by human system which puts different focuses at specific locations, time segments and media while performing multi-modality perception, we provide an attention-based method to simulate such process. Similar to human mechanism, our network can adaptively select ***"where" to attend***, ***"when" to attend*** and ***"which" to attend*** for audio-visual event localization. In this way, even with large temporal inconsistent between vision and audio, our network is able to adaptively trade information between different modalities and successfully achieve event localization. Our method achieves state-of-the-art performance on AVE (Audio-Visual Event) dataset collected in the real life. In addition, we also systemically investigate audio-visual event localization tasks. The visualization results also help us better understand how our model works.

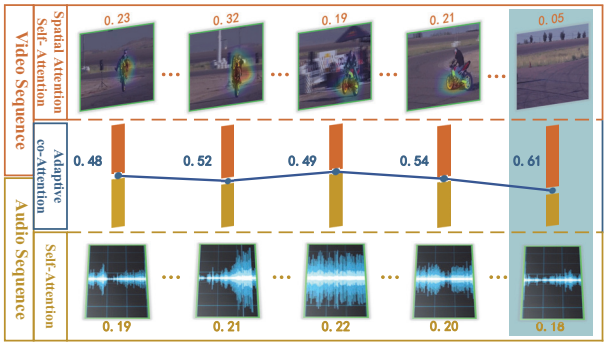


Figure 1: The diagram of audio-visual sequence of temporal inconsistency. At the last time segment, the event only occurs in audio but does not occur in the video. Our network can cope with such challenge by automatically learning to pay higher attention on specific visual regions, specific time segments and specific modality where the event occurs.

Introduction

As a proxy to the broader audio-visual scene understanding problem for real-life videos, audio-visual event localization task aims to match both audio and video components for identifying the simultaneous event of interest. Similar to the human's *Multi-modality Perception* (Smith and Gasser 2005) process, the benefits of integrating auditory and visual information are extensive as they provide plenty supplementary cues for better perception.

Generally, auditory and visual events tend to occur together *as they have consistency on time axis*, not always but often: lips moving when talking, the running cars accompany with engine noise and so on. In such cases, these events are concurrent and interactive because there is a common origin. Such *temporal consistency* between audition and vision allows us to perceive better. As a result, many works are motivated to make machine obtain similar multi-modality ability on perception, e.g., Lip Reading (Chung et al. 2017), Sound Synthesis (Owens et al. 2016a) and so on. The consistency assumption behind these works relies on the specific audio-visual scene where sound-maker should exist in the captured visual appearance simultaneously.

To this end, some works try to broaden the field of application to real-life videos. For example, (Owens and Efros 2018; Parekh et al. 2018) attempted to learn joint cross-modal representations considering sound and corresponding visual images as supervisory signal. Nevertheless, these works strongly depend on temporal consistency of visual/audio information, thus may suffer degradation in practical condition where such assumption cannot be satisfied. For audio-visual event localization, (Tian et al. 2018; Lin, Li, and Wang 2019) attempted to locate the events by multi-modal fusion. Since the simple fusion strategy assumes the features or prediction scores of a modality are explicitly complementary to one another, temporal inconsistency may mislead the results of event localization.

The temporal inconsistency is ubiquitous in the real-life videos. On the one hand, audio and video signals are managed by independent workflow in a typical multimedia presentation (Khosravan, Ardeshir, and Puri 2018), which

*Corresponding author: Yan Yan

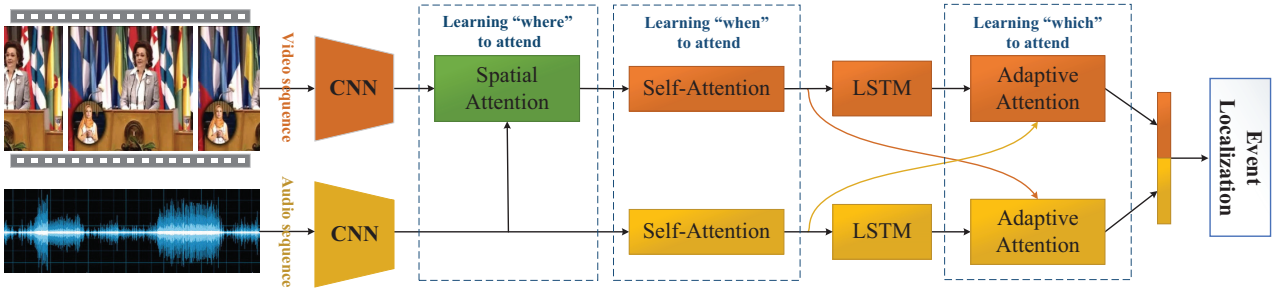


Figure 2: The diagram of the proposed cross-modal network for audio-visual event localization. The two modality-specific CNNs first process the corresponding audio and video into respective feature maps, then these feature maps are fed into three modules that automatically learn "*where*" (Spatial Attention Module), "*when*" (Global Context-aware self-Attention Module) and "*which*" (Cross-Modal Adaptive co-Attention Module) to attend for event localization.

opens up the issue of inconsistency. The sound-maker may be even out of the screen so that we cannot see it in the video, e.g., the voiceover of photographer. On the other hand, the visual scene may contain multiple objects which could be sound-makers or not, and the corresponding audio is a kind of multi-source mixture. Simply performing the global corresponding verification without having an insight into the complex scene components could result semantic irrelevant matching.

Different from these methods, we fully model the correlations and relationships between audio and video signals in order to weaken the interference of temporal inconsistency (see Fig.1) and improve understanding in audio-visual scene. For this purpose, we may refer to some studies in psychophysical and physiological, e.g, *Ventriloquism Effect* (Pick, Howard, and Templeton 1967) and *SIFI* (Sound Induced Flash Illusion) (Shams, Kamitani, and Shimojo 2000), where our brain tends to give different attention to different visual regions, time segments and sensory channels rather than focusing on all. That is to say, the human system is able to selectively capture valuable information of events from multi-media information.

Inspired by such mechanism, we propose a novel cross-modal attention framework to fully explore potential hidden correlations of same-modal and cross-modal signals. Our network takes both audio and visual sequence at each time segment as inputs and exploits global and local multi-modal correlations in the manner of Seq2Seq. We design three different attention modules to dig out "where", "when" and in "which" sensor the most event-related information should be. As a result, our method automatically learns to pay higher attention on specific visual areas, specific time segments and specific modality where the event occurs. In such way, our method can filter out the event-unrelated information and utilize event-valuable information to perform localization. In addition, we also systemically investigate three audio-visual event localization tasks: supervised, weakly-supervised and unsupervised cross-modal localization. And the visualization results also help us better understand how our model works. In summary, our contributions can be highlighted as follows:

- A novel and interpretable framework is proposed to detect where, when and on which media the event occurs and perform high-quality event localization;
- A novel spatial, sequential and cross-modal adaptive attention module is designed to capture most event-related information;
- State-of-the-art performance on widely-used dataset of audio-visual event localization is achieved.

Related Works

Audio/Video Event Localization aims to detect and temporally locate audio/video events in an acoustic/visual scene. For audio event localization, Hidden markov models (Heittola et al. 2013), gaussian mixture models (Mesaros, Heittola, and Virtanen 2016) and RNN (Parascandolo, Huttunen, and Virtanen 2016) are widely used. For video event localization, most works use a temporal sliding window approach, where each window is considered as an action candidate subject to classification, e.g., deep action proposal network (Escorcia et al. 2016), temporal convolutional network (Lea et al. 2017) and so on. These methods focus on audio or visual signals, *we simultaneously consider two types of heterogeneous data from different modalities*.

Audio-Visual Representations Learning aims to learn joint multi-modal representations using audio and video. Some works (Owens et al. 2016a; 2016b) attempt to learn enhanced visual representations considering sound as supervisory signal by virtue of its natural temporal correspondence. Some works (Aytar, Vondrick, and Torralba 2016; Gao, Feris, and Grauman 2018) attempt to learn powerful sound representations considering correspondent visual frames as supervision. Some works (Owens and Efros 2018; Parekh et al. 2018) attempt to learn joint cross-modal representations considering sound and corresponding visual image as supervisory signal in an unsupervised manner. Although the above works have shown promised cross-modal learning capacity, they often are troubled by temporal inconsistent audio-video pairs. Unlike these works, *we have no any prior assumption about temporal inconsistency of audio-video pairs*.

Audio-Visual Event Localization aims to match both audio and visual components for identifying the simultaneous event of interest. (Tian et al. 2018) use dual multi-modal residual network to fuse audio-visual features. (Lin, Li, and Wang 2019) directly concatenate audio and visual features as the input of LSTM. These methods only consider temporal relationship with audio or visual signals while ignoring potential hidden correlations between same-modal and cross-modal signals. In this paper, we use attention mechanism to model same-modal and cross-modal correlations.

Method

Problem Formulation

We define an audio-visual event as an event that is both audible and visible in the video. Concretely, we split a video sequence into T non-overlapping segments $\{\boldsymbol{v}_t, \boldsymbol{a}_t\}_{t=1}^{T}$, where each segment is 1s long (since the event boundary is labeled at second-level), $\boldsymbol{v}_t = [\boldsymbol{v}_t^1, \ldots, \boldsymbol{v}_t^k] \in \mathbb{R}^{d_v \times k}$ and $\boldsymbol{a}_t \in \mathbb{R}^{d_a}$ respectively denote the visual content and its corresponding audio counterpart in the video segment, *t* is the time segment of one video.

Supervised event localization. In supervised settings, the second-level event labels are given as $\boldsymbol{y}_t = \{\boldsymbol{y}_t^m | \boldsymbol{y}_t^m \in \{0,1\}, \sum_{m=1}^{C} \boldsymbol{y}_t^m = 1\}$, where C is the total number of event categories plus one background class. The non-background categories are determined only when audio and visual events are jointly observed.

Weakly-supervised event localization. In weakly-supervised settings, we can only access to video-level event labels, given by averaging the second-level event labels $\boldsymbol{Y} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{y}_t$. The weakly-supervised setting can not only weaken the dependency on the well-annotated labels, but also evaluate the robustness of our framework.

Unsupervised cross-modal event localization aims to locate events of interest in the visual scene with corresponding audio signals. It can locate which regions of visual frames the events of interest occur in. Different from classic localization setting, the location labels are not given during the train phase.

The diagram of our method is shown in Fig2. The two modality-specific CNNs first process the video and corresponding audio sequences into respective feature maps and then these features are fed into three modules. First of all, *Spatial Attention Module* utilizes audio signals as a guidance to locate "where" the regions of visual frames that the events of interest occur in. What's more, *Global Context-aware Self-Attention Module* explores same-modal hidden correlations to adaptively learns "when" to attend among series of video segments through introducing global context representation of entire input sequence. At last, *Cross-Modal Adaptive co-Attention Module* explores cross-modal hidden correlations to determine "which" the modality attend to through "modality sentinel" vector. These modules will be introduced respectively below.

Spatial Attention Module

As sound implies plentiful information about its source but also its location, it can be used to locate the source in videos with audio-visual contingency. Some methods inspire us to utilize audio signals as a means of guidance when searching the event location. Similar to (Tian et al. 2018), we use an attentional module to combine visual feature and its supplementary audio information for learning ***"where" to attend.***

We define the spatial attention module for computing the visual spatial feature vector $\boldsymbol{v}_t^{att} \in \mathbb{R}^{d_v}$ which is defined as $\boldsymbol{v}_t^{att} = \sum_{i=1}^{k} \boldsymbol{w}_t^i \boldsymbol{v}_t^i$, where $\boldsymbol{w}_t \in \mathbb{R}^k$ is a weight vector corresponding to the probability distribution over *k* visual regions that are attended by its audio counterpart. In order to get the spatial attention weight $\boldsymbol{w}_t$, we firstly use *MLP* with nonlinearity to project $\boldsymbol{a}_t$ and $\boldsymbol{v}_t$ to the same dimension. Different from (Tian et al. 2018), we use the simple normalized inner product operation rather than other *MLP* layer to get final $\boldsymbol{w}_t$. Such operation does not contain any additional learning parameters, and is intuitive that the inner product measures the cosine similarity between audio and visual features.

This weight vector $\boldsymbol{w}_t$ reflects in which regions of video frames the events of interest occurred in the form of a probability distribution. As a result, we can achieve cross-modal event localization through up-sampling $\boldsymbol{w}_t$ to the image size using bilinear interpolation. Compared to the cross-modal localization method (Owens and Efros 2018) using global average pooling(*GAP*) and strongest class activation map(*CAM*) response (Zhou et al. 2016), our module not only locate the spatial regions of the event of interest (see Fig.5), but also improve audio-visual event temporal localization accuracy through *adaptively selecting which visual features are more useful* in a weighted-average manner rather than global averaging (see Tab.2).

Global Context-aware self-Attention Module

For real-life videos, the most events of interest often occur at certain segments rather than entire audio/video sequence. Such hidden correlations between each segment of same-modal sequence are critical for temporal event localization. Self-attention mechanism is often used for modeling such hidden correlations. However, classic self-attention method treats the input sequence as the bag-of-word tokens and each token individually performs attention over the bag-of-word tokens (Yang et al. 2019). In other word, classic self-attention methods only capture local direct correlations between segments. In order to adaptively learn ***"when" to attend***, global rather than local correlations need be captured. For this purpose, we design a global context-aware self-attention method.

We introduce a global context vector $\boldsymbol{C}$, used mean operation to summarize the representation of input sequence, to represent global meaning of entire sequence. Considering the convenience of describing our method, we use the token $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T] \in \mathbb{R}^{d_x \times T}$ to indicate the input audio/video sequence, where T is the length of sequence, $\boldsymbol{x}_t \in \{\boldsymbol{v}_t^{att}, \boldsymbol{a}_t\}$ is the features of sequence segment at t time, $d_x \in \{d_v, d_a\}$ is the feature dimension. $\mathbf{C}$ can be expressed as $\boldsymbol{C} = [\boldsymbol{c}, \ldots, \boldsymbol{c}] \in \mathbb{R}^{d_x \times T}$, where $\boldsymbol{c} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t$.

Formally, given an input sequence $\boldsymbol{X}$, the hidden states in the output are constructed by attending to the states of input. Specifically, the $\boldsymbol{X}$ is first transformed into candidate query

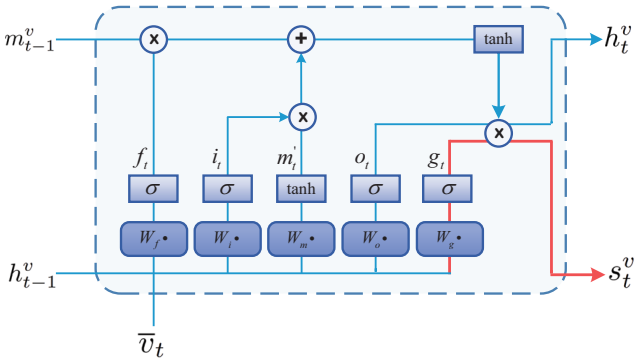


Figure 3: A illustration of LSTM extension with sentinel vector s_t^v. s_t^v is an additional latent representation of the LSTM's memory.

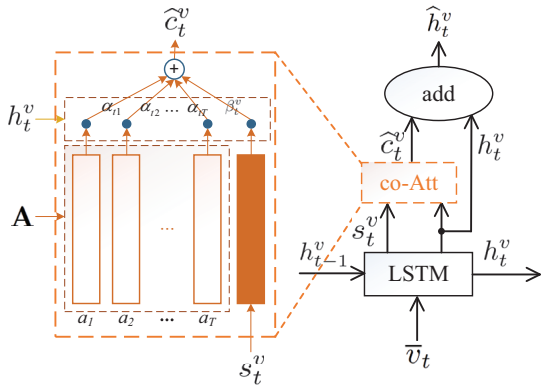


Figure 4: A illustration of cross-modal adaptive co-attention module generating visual modality sentinel β_t^v and the new context vector $\widehat{c}_t^v$.

states $\boldsymbol{Q}_X$, candidate key states $\boldsymbol{K}_X$ and value states $\boldsymbol{V}_X$. Similarly, the context vector $\boldsymbol{C}$ is transformed into candidate query states $\boldsymbol{Q}_C$ and candidate key states $\boldsymbol{K}_C$. Through two candidate query states $\boldsymbol{Q}_X/\boldsymbol{Q}_C$ and two candidate key states $\boldsymbol{K}_X/\boldsymbol{K}_C$, we can get the query states $\boldsymbol{Q}$ and key states $\boldsymbol{K}$ that are ultimately used to calculate the self-attention weights:

$$\begin{bmatrix}\boldsymbol{Q}\\ \boldsymbol{K}\end{bmatrix} = (1 - \begin{bmatrix}\boldsymbol{\lambda}_Q\\ \boldsymbol{\lambda}_K\end{bmatrix}) \begin{bmatrix}\boldsymbol{Q}_X\\ \boldsymbol{K}_X\end{bmatrix} + \begin{bmatrix}\boldsymbol{\lambda}_Q\\ \boldsymbol{\lambda}_K\end{bmatrix} \begin{bmatrix}\boldsymbol{Q}_C\\ \boldsymbol{K}_C\end{bmatrix}, \quad (1)$$

where $\boldsymbol{\lambda}_Q, \boldsymbol{\lambda}_K \in \mathbb{R}^T$ are weight matrices used to explicitly quantify how much each segment and the context vector contribute to the prediction of self-attention weight. The $\boldsymbol{\lambda}_Q, \boldsymbol{\lambda}_K$ can be calculated as:

$$\begin{bmatrix}\boldsymbol{\lambda}_Q\\ \boldsymbol{\lambda}_K\end{bmatrix} = \sigma(\begin{bmatrix}\boldsymbol{Q}_X\\ \boldsymbol{K}_X\end{bmatrix} \begin{bmatrix}\boldsymbol{V}_Q^X\\ \boldsymbol{V}_K^X\end{bmatrix} + \begin{bmatrix}\boldsymbol{Q}_C\\ \boldsymbol{K}_C\end{bmatrix} \begin{bmatrix}\boldsymbol{V}_Q^C\\ \boldsymbol{V}_K^C\end{bmatrix}), \quad (2)$$

where $\boldsymbol{V}_Q^X, \boldsymbol{V}_K^X, \boldsymbol{V}_Q^C, \boldsymbol{V}_K^C \in \mathbb{R}^{d_x}$ are associated trainable parameters, $\sigma()$ is the logistic sigmoid function.

The final output of our global context-aware self-attention model $\widehat{\boldsymbol{X}}$ is constructed by $\widehat{\boldsymbol{X}} = \boldsymbol{w}_c \boldsymbol{V}_X$. $\boldsymbol{w}_c$ is the weights of our self-attention method, and can be expressed as $\boldsymbol{w}_c = softmax(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}})$, where $\sqrt{d}$ is the scaling factor.

On the one hand, the global context-aware self-attention weights can be viewed as a probability distribution that reflects the contribution of each individual segment for event localization (see Fig.1). And introduction of the global representation also can improve event localization performance (see Tab.2). On the other hand, compared to the classic self-attention method (Lin et al. 2017), a large magnitude of $\boldsymbol{Q}_X$ and $\boldsymbol{K}_X$ may push the *softmax* function into regions where it has extremely small gradients. $\boldsymbol{\lambda}_Q, \boldsymbol{\lambda}_K$ *can also be treated as factors to regulate the magnitude of* $\boldsymbol{Q}$ *and* $\boldsymbol{K}$.

Cross-Modal Adaptive co-Attention Module

For sequence data such as audio and video, how to better model the dependencies between segments of sequence is crucial, as the event itself is temporally related and consistent in a audio/video counterpart. Normal convolutional layer is suitable to learning spatial information, but has limitation on learning temporal features. Further, as the event location may move and change sharply, the spatially long-range dependency is also crucial for localizing. As a result, we choose to use LSTM for modeling the temporal dependencies.

While modeling the temporal dependencies, some works (Lu et al. 2017; 2018) attempt to use "visual sentinel" to determine whether the generated words depend on visual signals or language models for image captioning. Motivated from them, we introduce a new LSTM extension, which produces an additional ***"modality sentinel"*** vector instead of a single hidden state. Our module can explore cross-modal hidden correlations to determine when to rely on visual signals and when to rely on the audio signals.

Firstly, we extend the LSTM to obtain a new component – sentinel vector, an additional latent representation of the decoder's memory, provides a fallback option to the decoder through co-attention mechanism. As shown in Fig.3, the mathematical definitions are as follows:

$$\boldsymbol{s}_t^v, \boldsymbol{h}_t^v = \textbf{LSTM}(\overline{\boldsymbol{v}}_t, \boldsymbol{h}_{t-1}^v, \boldsymbol{m}_{t-1}^v), \quad (3)$$

where $\overline{\boldsymbol{v}}_t \in \mathbb{R}^{d_v}$ is the visual features of current input at time step t, $\boldsymbol{h}_t^v$ and $\boldsymbol{s}_t^v$ are hidden state and sentinel vector as an output of the LSTM unit respectively. Notably, $\overline{\boldsymbol{v}}_t$ represents $\boldsymbol{v}_t^{att}$ obtained by spatial attention module or $\boldsymbol{v}_t$ dealt with GAP operation. And the expression of $\boldsymbol{s}_t^v$ is computed as:

$$\boldsymbol{s}_t^v = \boldsymbol{g}_t^v \odot tanh(\boldsymbol{m}_t^v), \quad (4)$$

where $\boldsymbol{g}_t^v = sigmiod(\boldsymbol{W}_g[h_{t-1}^v; \overline{\boldsymbol{v}}_t] + \boldsymbol{b}_g)$, $\odot$ represents the element-wise product.

Based on the sentinel vector $\boldsymbol{s}_t^v$, we propose an adaptive co-attention module to compute the new context vector. As shown in Fig.4, the new context vector is defined as $\widehat{\boldsymbol{c}}_t^v$, which is modeled as a mixture of the sentinel vector and the attended audio features. It trades off how much new auxiliary information from another modality(audio) is being considering. The mixture model is defined as follows:

$$\widehat{\boldsymbol{c}}_t^v = \boldsymbol{\beta}_t^v \boldsymbol{s}_t^v + (1 - \boldsymbol{\beta}_t^v)\boldsymbol{c}_t^a, \quad (5)$$

where $\boldsymbol{\beta}_t^v$ is the visual sentinel gate that produces a scalar in the range [0,1] at time step t. A value of 1 implies that only the sentinel information is used and 0 means only context information from another modality is used to locate audio-visual event. The context vector $\boldsymbol{c}_t^a$ represents new supplementary information from another modality, which is defined as:

$$\boldsymbol{c}_t^a = g(\boldsymbol{A}, \boldsymbol{h}_t^v) = \sum_{i=1}^{T} \boldsymbol{\alpha}_{ti} \boldsymbol{a}_{ti}, \tag{6}$$

where g is the attention function, $\mathbf{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_T]$, $\boldsymbol{a}_t \in \mathbb{R}^{d_a}$ is the audio feature at time step t. We feed $\boldsymbol{A}$ and $\boldsymbol{h}_t^v$ through a single layer neural network followed by a *softmax* function to generate the attention distribution over T audio segments. Co-attention weight $\boldsymbol{\alpha}_t$ is equal to the result of softmax $\boldsymbol{z}_t$, where the expression of $\boldsymbol{z}_t$ as follows:

$$\boldsymbol{z}_t = \boldsymbol{W}_h^T tanh(\boldsymbol{W}_a \boldsymbol{A} + \boldsymbol{W}_g \boldsymbol{h}_t) \tag{7}$$

where $\boldsymbol{W}_a, \boldsymbol{W}_h \in \mathbb{R}^{d_a \times T}$, $\boldsymbol{W}_g \in \mathbb{R}^{d \times d}$ are parameters to be learnt.

To compute $\boldsymbol{\beta}_t^v$, we add an additional element $\boldsymbol{z}_t$ to $\boldsymbol{\alpha}_t$. This element indicates how much "attention" the network is placing on the sentinel (as opposed to the features from another modality). The new weight can be formulated as:

$$\widehat{\boldsymbol{\alpha}}_t = softmax[\boldsymbol{z}_t; \boldsymbol{W}_h^T tanh(\boldsymbol{W}_s \boldsymbol{s}_t + \boldsymbol{W}_g \boldsymbol{h}_t)], \tag{8}$$

where $[\ ;\ ]$ indicates concatenation. $\widehat{\boldsymbol{\alpha}}_t \in \mathbb{R}^{T+1}$ is the attention distribution over both another modality feature as well as the sentinel vector. We interpret the last element of this vector to be the visual sentinel gate value, $\boldsymbol{\beta}_t^v = \widehat{\boldsymbol{\alpha}}_t[T+1]$. Finally, we combine obtained adaptive context vector $\widehat{\boldsymbol{c}}_t^v$ with the state vector of LSTM to obtain a cross-modal visual representation vector $\widehat{\boldsymbol{h}}_t^v$, i.e., $\widehat{\boldsymbol{h}}_t^v = \widehat{\boldsymbol{c}}_t^v + \boldsymbol{h}_t^v$.

Similarly, we can get the audio sentinel gate value $\boldsymbol{\beta}_t^a$ and cross-modal audio representation $\widehat{\boldsymbol{h}}_t^a$. The concatenate representation of $\widehat{\boldsymbol{h}}_t^v$ and $\widehat{\boldsymbol{h}}_t^a$ is used for final event localization. Notably, in order to make our module adaptively select ***"which to attend"*** with a certain probability, we normalize $\boldsymbol{\beta}_t^v$ and $\boldsymbol{\beta}_t^a$ by *softmax*.

The modality sentinel encourages the network to adaptively select attending to the audio features vs. the video features for event localization. Notably, as shown in Fig.4, *the generated context vector could be considered as the residual information of current hidden state, which diminishes the uncertainty or complements the informativeness of current hidden state.*

Experiments

Dataset and Implementation Details

Dataset AVE Dataset (Tian et al. 2018), which is a subset of AudioSet (Gemmeke et al. 2017), contains 4143 samples covering 28 event categories, e.g., dog barking, man speaking, chainsaw logging and airplane flying. The samples in AVE Dataset are filled with temporal inconsistency, abrupt view changes and different types of noise. These event categories cover a wide range of real-life scene, e.g., music performances, main street, public speaking and so on. Each event category contains a minimum of 60 samples and a maximum of 188 samples. Each sample in the AVE is temporally labeled with audio-visual event boundaries and contains at least one $2s$ long audio-visual event.

We use the same settings as *Tian* (Tian et al. 2018). We divide the AVE dataset into three parts, i.e., 80% for training, 10% for validation and 10% for testing. To better verify the effectiveness and robustness of our modules, we also extend our model in a weakly-supervised manner.

Table 1: The event localization accuracy (%) on AVE dataset in both supervised and weakly supervised settings, which are separated by the symbol '/'. **A** and **V** denote these models only use audio and visual features as inputs, respectively. **V-att**, **S-att**, **Co-att** denote three modules which are elaborated *Method* section.

Index	A	V	V-att	S-att	Co-att	Accuracy
1	✓	×	×	×	×	59.5/56.5
2	×	✓	×	×	×	55.3/53.7
3	✓	✓	×	×	×	71.4/69.0
4	✓	✓	✓	×	×	72.8/70.9
5	✓	✓	×	✓	×	72.4/71.2
6	✓	✓	×	×	✓	73.2/72.1
7	✓	✓	✓	✓	×	73.5/71.8
8	✓	✓	×	✓	✓	74.8/73.4
9	✓	✓	✓	×	✓	75.6/73.9
10	✓	✓	✓	✓	✓	**77.1/75.7**

Implementation Details For visual and audio representation, we respectively adopt ResNet-151 network pre-trained on ImageNet and VGG-like network pre-trained on AudioSet. Specifically, we extract pool5 feature maps from sampled 16 RGB frames for each 1s video segment. Respectively, we extract 512×7×7-D visual representation and 128-D audio representation for each 1s audio segment and 1s visual segment.

Quantitative Analysis

Ablation Studies In this section, we first explore whether multi-modal information can help us better perceive the environment. Then we discuss the impact of different modules elaborated in the *Method* section. In addition, we show the effect of combining two different modules on the results of event localization.

In order to ensure the comparability of the experimental results, all models have the same setting, e.g., the same number of fully connected layers. The attention modules are implemented on the feature maps of the middle layer in our framework. The addition or removal of each module means that the corresponding attention is calculated or not. The number of introduced parameters can be ignored. In other words, it is the computational rather than complexity increasing with three introduced modules.

Tab.1 shows the experimental result about ablation studies. As shown in the first-three rows, we can observe that the performance of instantaneous usage of audio and video

Table 2: The event localization accuracy (%) comparison with state-of-the-art methods in both supervised and weakly supervised settings, which are separated by the symbol '/'.

Method	Accuracy
AVEL(Tian et al. 2018)	74.7/73.1
AVSDN(Lin, Li, and Wang 2019)	75.4/74.2
CAM(Owens and Efros 2018)	72.3/68.8
our model - global context vector	75.7/74.3
our model - sentinel vector	76.1/74.9
our model	**77.1/75.7**

information as input is better than just using audio or visual data. It also validates that combining audio and video modalities is significantly beneficial for understanding the events. As shown in the 4*th* to 6*th* rows, we only use one of the three modules in order to explore the impact of the independent module. We observer the cross-modal adaptive co-attention module can improve accuracy significantly. Such results also demonstrate the need for the model to adaptively determine when to rely on visual signals and when to rely on audio signals for real-life event understanding. As shown in the 7*th* to 9*th* rows, we can observe that combining two different modules can improve performance to some extent. Such results also further proves the validity of three modules.

Comparison Results To test the effectiveness and robustness of our framework, we compare the accuracy of audio-visual event localization with two state-of-the-art methods. The results of comparison are shown in the first-two rows of Tab.2. In addition, we compare two different strategies for locating audio-visual event in video frames, i.e., strongest *CAM* response used by (Owens and Efros 2018) and spatial attention module used in our framework. In order to generate map of strongest CAM response, *Owens and Efros* use a global manner to average image-level features and get an accuracy of 72.3/68.8. Our spatial attention module, which use a weighted manner rather than global manner, achieves an accuracy of **77.1/75.7**. The weighting manner can not only adaptively capture the spatial visual location information of the events of interest (see Fig.5), but also adaptively learn which region in the video needs to be concerned.

Furthermore, as shown in the row 4 and 5 of Tab.2, the introduction of global context vector can better explore the hidden correlations between same-modal segments and the sentinel vector of LSTM can better model the dependencies between cross-modal segments.

In the *Supplementary Material*, we also compare the accuracy of each event category with AVEL (Tian et al. 2018). Our method significantly improves performance for most categories of audio-visual events localization.

Qualitative Analysis

Spatial Localization As shown in Fig.5, we compare the results of cross-modal event localization generated by AVEL and our model. AVEL seems to tend to focus on small region in the image, while our model can focus on the general shape of the objects that trigger the event of interest. The larger regions mean that our model can capture more event-related information in the visual frames. Our model only needs to find the event-related information in a limited scope when focusing on specific time segments or modality where the event occurs. In other words, the specific rather than the entire scope makes spatial attention module much easier to capture the event-related information. More results are shown in the *Supplementary Material*.

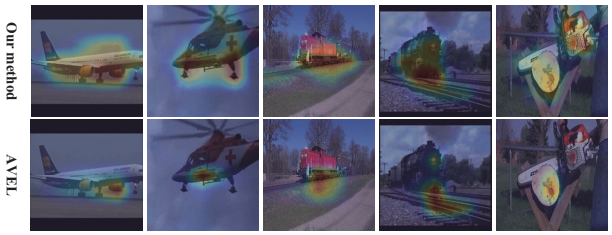


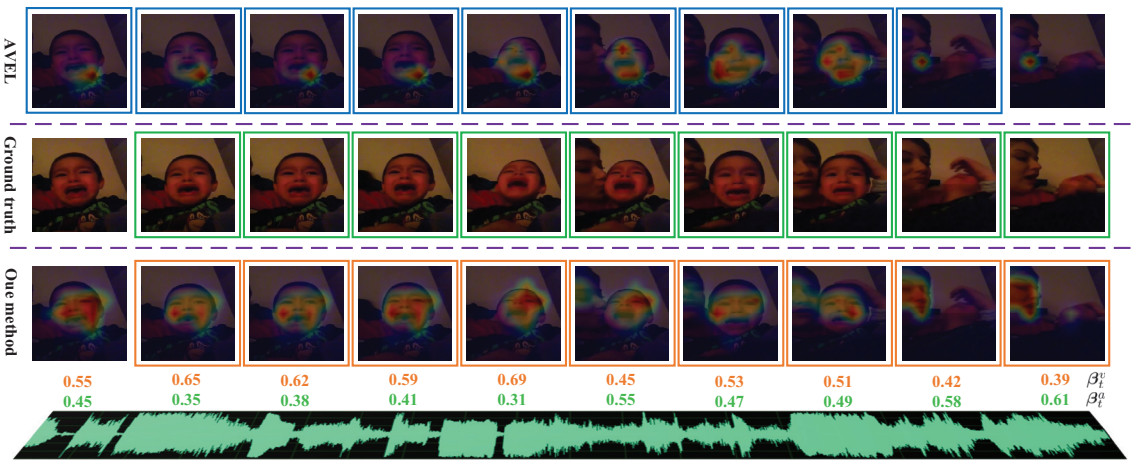
Figure 5: Spatial localization comparison with AVEL.

Temporal Localization We visualise the results of cross-modal attention network proposed by us with AVEL. As shown in Fig.6, the green boxes indicate the ground truth given in the AVE dataset. The blue boxes and orange boxes respectively indicate localization result of AVEL and our network. At the same time, we also count the weight of sentinel gate β_t^v and β_t^a, which is used to adaptively trade information between different modalities in a weighted manner, as shown in the middle of the audio and video sequence.

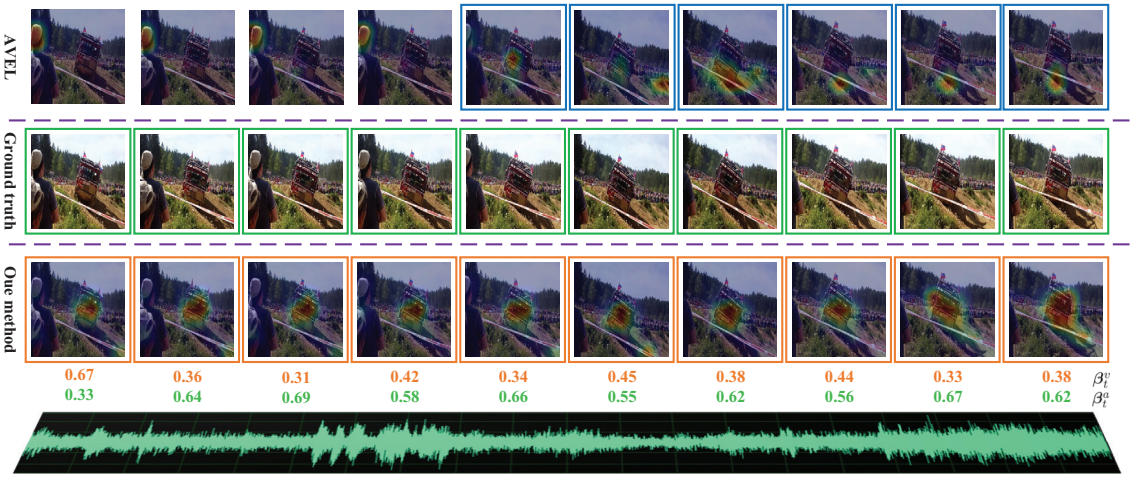
Discussion

How does our model work? Since there are often multiple objects in the real-life videos, these objects may cause different degrees of interference when we locate the event of interest. For example, the crowd beside the truck interfere with the temporal localization of the truck event. The audio-guide attention makes our network only focus on specific regions where the event of interest (truck) occurs in, thus avoiding the influence of the crowd. As shown in Fig.6(b), AVEL gets error results of event localization due to focus on error regions. In addition, as shown in the last two columns of Fig.6(a), when our model focuses on error visual regions, modality sentinel will act as a role in correcting errors of spatial attention maps by assigning smaller weights to visual signals.

For these events in the door (e.g., dog, man and woman speaking, baby crying), there are several different sounds may be mixed together. In this case, the audio signals may have very low intensity. Comparatively speaking, visual information will give us more discriminative and accurate information to understand events this moment. Empirically, we find our network seems to rely more on the video signals in some indoor scenes, as shown in Fig.6(a). For these events (e.g., car, motocycle, train, bus), sounds will provide clear cues. Our network seems to rely more on the audio signals in some outdoor scenes, as shown in Fig.6(b). These two rules seem to be universally applicable to most noisy



(a) Baby Cry(indoor scene)



(b) Truck(outdoor scene)

Figure 6: Visualization analysis. The green boxes indicate the ground truth value given in the AVE dataset. The blue boxes and orange boxes respectively indicate localization result of AVEL and our method. The numbers in the middle of the audio and video sequence are statistical weights of sentinel gate β_t^v and β_t^a.

outdoor scenes. As mentioned in *Ablation Studies*, this empirical finding may explain why the cross-modal adaptive co-attention module can significantly improve accuracy relative to the other two modules.

How to cope with temporal inconsistency? As mentioned in *Introduction* section, the temporal inconsistency is ubiquitous in the real-life videos. Simple fusion strategy, assumed the features or prediction scores of a modality are explicitly complementary to one another, will mislead the results of event temporal localization. In other word, the complementarity between audio and video signals is an interference for the task of audio-visual event localization. As a result, it is necessary to explore potential hidden correlations between cross-modal signals, rather than simple feature fusion. For this purpose, modality sentinel is introduced by our network.

As shown in the last time segment of Fig.1, the event only occurs in audio but does not occur in video. At this time, our network is more inclined to learn visual and audio modality sentinel such that β_t^v is greater than β_t^a. Available from Eq.5, the final concatenate representation will be more dependent on the visual information that no event occur in.

Conclusion

Inspired by human multi-modal perception mechanism, in this paper we propose a novel cross-modal attention framework consisted of three attention modules to fully explore same-modal and cross-modal potential hidden correlations. In addition, we also systemically investigate audio-visual event localization tasks: supervised, weakly-supervised and cross-modal localization. Our model also achieves competitive results on AVE Dataset. The results of the visualization can help us better understand how our model works.

Acknowledgement: This work was supported by the National Science Fund of China under Grant No. 61806094,

No. U1713208 and Program for Changjiang Scholars.

References

Aytar, Y.; Vondrick, C.; and Torralba, A. 2016. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*, 892–900.

Chung, J. S.; Senior, A.; Vinyals, O.; and Zisserman, A. 2017. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3444–3453. IEEE.

Escorcia, V.; Heilbron, F. C.; Niebles, J. C.; and Ghanem, B. 2016. Daps: Deep action proposals for action understanding. In *European Conference on Computer Vision*, 768–784. Springer.

Gao, R.; Feris, R.; and Grauman, K. 2018. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 35–53.

Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780. IEEE.

Heittola, T.; Mesaros, A.; Eronen, A.; and Virtanen, T. 2013. Context-dependent sound event detection. *EURASIP Journal on Audio, Speech, and Music Processing* 2013(1):1.

Khosravan, N.; Ardeshir, S.; and Puri, R. 2018. On attention modules for audio-visual synchronization. *arXiv preprint arXiv:1812.06071* 1.

Lea, C.; Flynn, M. D.; Vidal, R.; Reiter, A.; and Hager, G. D. 2017. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 156–165.

Lin, Z.; Feng, M.; Santos, C. N. d.; Yu, M.; Xiang, B.; Zhou, B.; and Bengio, Y. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Lin, Y.-B.; Li, Y.-J.; and Wang, Y.-C. F. 2019. Dual-modality seq2seq network for audio-visual event localization. *arXiv preprint arXiv:1902.07473*.

Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 375–383.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2018. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7219–7228.

Mesaros, A.; Heittola, T.; and Virtanen, T. 2016. Tut database for acoustic scene classification and sound event detection. In *2016 24th European Signal Processing Conference (EUSIPCO)*, 1128–1132. IEEE.

Owens, A., and Efros, A. A. 2018. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 631–648.

Owens, A.; Isola, P.; McDermott, J.; Torralba, A.; Adelson, E. H.; and Freeman, W. T. 2016a. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2405–2413.

Owens, A.; Wu, J.; McDermott, J. H.; Freeman, W. T.; and Torralba, A. 2016b. Ambient sound provides supervision for visual learning. In *European conference on computer vision*, 801–816. Springer.

Parascandolo, G.; Huttunen, H.; and Virtanen, T. 2016. Recurrent neural networks for polyphonic sound event detection in real life recordings. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6440–6444. IEEE.

Parekh, S.; Essid, S.; Ozerov, A.; Duong, N. Q.; Pérez, P.; and Richard, G. 2018. Weakly supervised representation learning for unsynchronized audio-visual events. In *CVPR Workshops*, 2518–2519.

Pick, H. L.; Howard, I. P.; and Templeton, W. B. 1967. Human spatial orientation. *The American Journal of Psychology* 80(3):476–478.

Shams, L.; Kamitani, Y.; and Shimojo, S. 2000. Illusions: What you see is what you hear. *Nature* 408(6814):788.

Smith, L., and Gasser, M. 2005. The development of embodied cognition: Six lessons from babies. *Artificial life* 11(1-2):13–29.

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

Yang, B.; Li, J.; Wong, D. F.; Chao, L. S.; Wang, X.; and Tu, Z. 2019. Context-aware self-attention networks. *arXiv preprint arXiv:1902.05766*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.