Audio-Visual Event Localization in the Wild

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu
University of Rochester

1. Introduction

In this paper, we study a family of audio-visual event temporal localization tasks as a proxy to the broader audio-visual scene understanding problem for unconstrained videos. We pose and seek to answer the following questions: (Q1) Does inference jointly over auditory and visual modalities outperform inference over them independently? (Q2) How does the result vary under noisy training conditions? (Q3) How does knowing one modality help model the other modality? (Q4) How do we best fuse information over both modalities? (Q5) Can we locate the content in one modality given its observation in the other modality? Notice that the individual questions might be studied in the literature, but we are not aware of any work that conducts a systematic study to answer these collective questions as a whole.

In particular, we define an *audio-visual event* as an event that is both visible and audible in a video segment, and we establish three tasks to explore aforementioned research questions: 1) supervised audio-visual event localization, 2) weakly-supervised audio-visual event localization, and 3) event-agnostic cross-modality localization. The first two tasks aim to predict which temporal segment of an input video has an audio-visual event and what category the event belongs to. The weakly-supervised setting assumes that we have no access to the temporal event boundary but an event tag at video-level for training. Q1-Q4 will be explored within these two tasks. In the third task, we aim to locate the corresponding visual sound source temporally within a video from a given sound segment and vice versa, which will answer Q5.

We propose both baselines and novel algorithms to solve the above three tasks. For the first two tasks, we start with a baseline model treating them as a sequence labeling problem. We utilize CNN to encode audio and visual inputs, adapt LSTM [6] to capture temporal dependencies, and apply Fully Connected (FC) network to make the final predictions. Upon this baseline model, we introduce an audio-guided visual attention mechanism to verify whether audio can help attend visual features; it also implies spatial locations for sounding objects as a side output. Furthermore, we investigate several audio-visual feature fusion methods and propose a novel dual multimodal residual fusion network that achieves the best fusion results. For weakly-supervised learning, we formulate it as a Multiple Instance Learning (MIL) [10] task, and modify our network structure via adding a MIL pooling layer to handle the problem. To address the harder cross-modality localization task, we propose an audio-visual distance learning network that measures the relativeness of any given pair of audio and visual content. It projects audio and visual features into subspaces with the same dimension. Contrastive loss [5] is introduced to learn the network.

Observing that there is no publicly available dataset directly suitable for our tasks, we collect a large video dataset that consists of 4143 10-second videos with both audio and video tracks for 28 audio-visual events and annotate their temporal boundaries. Videos in our dataset are originated from YouTube, thus they are unconstrained. Our extensive experiments support the following findings: modeling jointly over auditory and visual modalities outperforms modeling independently over them, audio-visual event localization in a noisy condition can still achieve promising results, the audio-guided visual attention can well capture semantic regions covering sounding objects and can even distinguish audio-visual unrelated videos, temporal alignment is important for audio-visual fusion, the proposed dual multimodal residual network is effective in addressing the fusion task, and strong correlations between the two modalities enable cross-modality localization.

2. Dataset and Problems

***Audio-Visual Event* Dataset** To the best of our knowledge, there is no publicly available dataset directly suitable for our purpose. Therefore, we introduce the *Audio-Visual Event* (AVE) dataset , a subset of AudioSet [4], that contains 4143 videos covering 28 event categories and videos in AVE are temporally labeled with audio-visual event boundaries. Each video contains at least one 2s long *audio-visual event*. The dataset covers a wide range of audio-visual events (*e.g.*, man speaking, woman speaking, dog barking, playing guitar, and frying food *etc.*) from different domains, e.g., human activities, animal activities, music performances, and vehicle sounds.We provide examples from different categories and show the statistics in Fig. 1. Each event category contains a minimum of 60 videos and a maximum of 188 videos, and 66.4% videos in the AVE contain audio-visual events that span over the full 10 seconds.

Fully and Weakly-Supervised Event Localization The goal of event localization is to predict the event label for each video segment, which contains both audio and visual tracks, for an input video sequence. Concretely, for

Figure 1. The AVE dataset. Some examples in the dataset are shown. The distribution of videos in different categories and the distribution of event lengths are illustrated.

a video sequence, we split it into T non-overlapping segments $\{V_t, A_t\}_{t=1}^{T}$, where each segment is 1s long (since our event boundary is labeled at second-level), and V_t and A_t denote the visual content and its corresponding audio counterpart in a video segment, respectively. Let $\boldsymbol{y}_t = \{y_t^k | y_t^k \in \{0,1\}, k = 1, ..., C, \sum_{k=1}^{C} y_t^k = 1\}$ be the event label for that video segment. Here, C is the total number of AVE events plus one background label. Different than the supervised setting, in the weakly-supervised manner we have only access to a video-level event tag, and we still aim to predict segment-level labels during testing. The weakly-supervised task allows us to alleviate the reliance on well-annotated data for modelings of audio, visual and audio-visual.

Cross-Modality Localization In the cross-modality localization task, given a segment of one modality (auditory/visual), we would like to find the position of its synchronized content in the other modality (visual/auditory). Concretely, for visual localization from audio (A2V), given a l-second audio segment $\hat{A}$ from $\{A_t\}_{t=1}^{T}$, where $l < T$, we want to find its synchronized l-second visual segment within $\{V_t\}_{t=1}^{T}$. Similarly, for audio localization from visual content (V2A), given a l-second video segment $\hat{V}$ from $\{V_t\}_{t=1}^{T}$, we would like to find its l-second audio segment within $\{A_t\}_{t=1}^{T}$.

3. Overview of Proposed Methods

Audio-Visual Event Localization Network: Our network mainly consists of five modules: feature extraction, audio-guided visual attention, temporal modeling, multimodal fusion and temporal labeling (see Fig. 2(a)). The feature extraction module utilizes pre-trained CNNs to extract visual features $v_t = [v_t^1, ..., v_t^k] \in \mathbb{R}^{d_v \times k}$ and audio features $a_t \in \mathbb{R}^{d_a}$ from each V_t and A_t, respectively. Here, d_v denotes the number of CNN visual feature maps, k is the vectorized spatial dimension of each feature map, and d_a denotes the dimension of audio features. We use an audio-guided visual attention model to generate a context vector $v_t^{att} \in \mathbb{R}^{d_v}$. Two separate LSTMs take v_t^{att} and a_t as inputs to model temporal dependencies in the two modalities respectively. For an input feature vector F_t at time step t, the LSTM updates a hidden state vector h_t and a memory cell state vector c_t, where F_t refers to v_t^{att} or a_t in our model. For evaluating the performance of the proposed attention mechanism, we compare to models that do not use attention; we directly feed global average pooling visual features and audio features into LSTMs as baselines. To better incorporate the two modalities, we introduce a multimodal fusion network. The audio-visual representation h_t^* is learned by a multimodal fusion network with audio and visual hidden state output vectors h_t^v and h_t^a as inputs. This joint audio-visual representation is used to output event category for each video segment. For this, we use a shared FC layer with the Softmax activation function to predict probability distribution over C event categories for the input segment and the whole network can be trained with a multi-class cross-entropy loss.

Audio-Guided Visual Attention: Given that attention mechanism has shown superior performance in many applications such as neural machine translation [3] and image captioning [12, 9], we use it to implement our audio-guided visual attention. The attention network will adaptively learn which visual regions in each segment of a video to look for the corresponding sounding object or activity. Concretely, we define the attention function f_{att} and it can be adaptively learned from the visual feature map v_t and audio feature vector a_t. At each time step t, the visual context vector v_t^{att} is computed by:

$$v_t^{att} = f_{att}(a_t, v_t) = \sum_{i=1}^{k} w_t^i v_t^i \ , \tag{1}$$

where w_t is an attention weight vector corresponding to the probability distribution over k visual regions that are attended by its audio counterpart. The attention weights can be computed based on MLP with a Softmax activation function. The attention map visualization results show that the audio-guided attention mechanism can adaptively capture the location information of sound source, and it can also improve temporal localization accuracy.

Audio-Visual Feature Fusion: To combine features coming from visual and audio modalities, we introduce a Dual Multimodal Residual Network (DMRN). Given audio and visual features h_t^a and h_t^v from LSTMs, the DMRN will compute the updated audio and visual features:

$$h_t^{a'} = tanh(h_t^a + f(h_t^a, h_t^v)) \ , \tag{2}$$

$$h_t^{v'} = tanh(h_t^v + f(h_t^a, h_t^v)) \ , \tag{3}$$

where $f(\cdot)$ is an additive fusion function, and the average of $h_t^{a'}$ and $h_t^{v'}$ is used as the joint representation h_t^* for labeling the video segment. Here, the update strategy in DMRN can both preserve useful information in the original modality and add complimentary information from the other modality.

Weakly-Supervised Event Localization: To address the weakly-supervised event localization, we formulate it as a MIL problem and extend our framework to handle noisy training condition. Since only video-level labels are available, we infer label of each audio-visual segment pair in the training phase, and aggregate these individual predictions into a video-level prediction by MIL pooling as in [11]:

$$\hat{m} = g(m_1, m_2, ..., m_T) = \frac{1}{T}\sum_{t=1}^{T} m_t \ , \tag{4}$$

where $m_1, ..., m_T$ are predictions from the last FC layer of our audio-visual event localization network, and $g(\cdot)$ averages over all predictions. The probability distribution of event category for the video sequence can be computed using $\hat{m}$ over the Softmax. During testing, we can predict the event category for each segment according to computed m_t.

Cross-Modality Localization: To address the cross-modality localization problem, we propose an audio-visual distance learning network (AVDLN) as illustrated in Fig. 2(b). Our network can measure the distance $D_\theta(V_i, A_i)$ for a given pair of V_i and A_i. At test time, for visual localization from audio (A2V), we use a sliding window method.

Let $\{V_i, A_i\}_{i=1}^N$ be N training samples and $\{y_i\}_{i=1}^N$ be their labels, where V_i and A_i are a pair of 1s visual and audio segments, $y_i \in \{0, 1\}$. Here, $y_i = 1$ means that V_i and A_i are synchronized. The AVDLN will learn to measure distances between these pairs. In practice, we use the Euclidean distance as the metric and contrastive loss [5] to optimize the AVDLN.

4. Experimental Results

Table 1 compares different variations of our proposed models on supervised and weakly-supervised audio-visual event localization tasks. Table 2 shows event localization performance of different fusion methods. Figures 3 illustrates generated audio-guided visual attention maps.

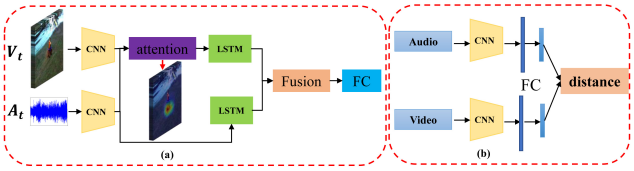

Figure 2. (a) Audio-visual event localization framework with audio-guided visual attention and multimodal fusion. One timestep is illustrated, and note that the fusion network and FC are shared for all timesteps. (b) Audio-visual distance learning network.

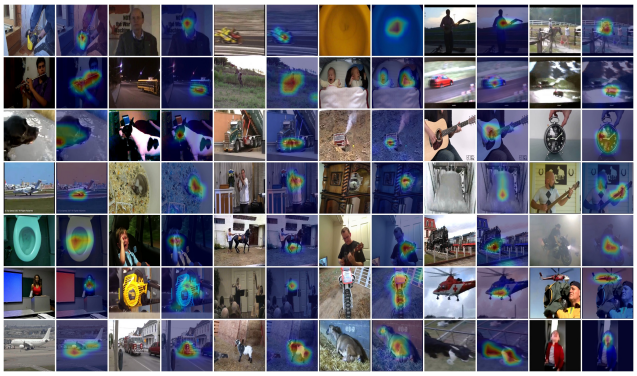

Figure 3. Qualitative visualization of audio-guided visual attention. The semantic regions containing many different sound sources, such as barking dog, crying boy/babies, speaking woman, guitar *etc*, can be adaptively captured by our attention model.

Table 1. Event localization prediction accuracy (%) on AVE dataset. A, V, V-att, A+V, A+V-att denote that these models use audio, visual, attended visual, audio-visual and attended audio-visual features, respectively. W-models are trained in a weakly-supervised manner. Note that audio-visual models all fuse features by concatenating the outputs of LSTMs.

Models	A	V	V-att	A+V	A+V-att	W-A	W-V	W-V-att	W-A+V	W-A+V-att
Accuracy	59.5	55.3	58.6	71.4	**72.7**	53.4	52.9	55.6	63.7	66.7

Audio and Visual: From Tab. 1, we observe that A outperforms V and W-A is also better than W-V. It demonstrates that audio features are more powerful to address audio-visual event localization task on the AVE dataset. However, when we look at each individual event, using audio is not always better than using visual. We observe that V is better than A for some events (*e.g.* car, motocycle, train, bus). Actually, most of these events are outdoor. Audios in these videos can be very noisy: several different sounds may be mixed together (*e.g.* people cheers with a racing car), and may have very low intensity (*e.g.* horse sound from far distance). For these conditions, visual information will give us more discriminative and accurate information to understand events in videos. A is much better than V for some events (*e.g.* dog, man and woman speaking, baby crying). Sounds will provide clear cues for us to recognize these events. For example, if we hear barking sound, we know that there may

Table 2. Event localization prediction accuracy (%) of different feature fusion methods on AVE dataset. These methods all use same audio and visual features as inputs. Top-2 results in each line are highlighted.

Methods	Additive	MP	Gated	MB	GMU	GMB	Concat	MRN	DMRN
Early Fusion	59.9	67.9	67.9	69.2	**70.5**	**70.2**	61.0	69.8	68.0
Late Fusion	71.3	71.4	70.5	70.5	71.6	71.0	**72.7**	70.8	**73.1**
Decision Fusion	**70.5**	64.5	65.2	64.6	67.6	67.3	69.7	63.8	**70.4**

be a dog. We also observe that A+V is better than both A and V, and W-A+V is better than W-A and W-V. From the above results and analysis, we can conclude that auditory and visual modalities will provide complementary information for us to understand events in videos.

Audio-Guided Visual Attention: The quantitative results (see Tab. 1) show that V-att is much better than V (a 3.3% absolute improvement) and A+V-att outperforms A+V by 1.3%. We show qualitative results of our attention method in Fig. 3. We observe that a range of semantic regions in many different categories and examples can be attended by sound, which validates that our attention network can learn which visual regions to look at for sounding objects. An interesting observation is that the audio-guided visual attention tends to focus on sounding regions, such as man's mouth, head of crying boy *etc*, rather than whole objects in some examples.

Audio-Visual Fusion: We compare our fusion method: DMRN with several network-based multimodal fusion methods: Additive, Maxpooling (MP), Gated, Multimodal Bilinear (MB), and Gated Multimodal Bilinear (GMB) in [7], Gated Multimodal Unit (GMU) in [2], Concatenation (Concat), and MRN [8]. Three different fusion strategies: early, late and decision fusions are explored. Here, early fusion methods directly fuse audio features from pre-trained CNNs and attended visual features; late fusion methods fuse audio and visual features from outputs of two LSTMs; and decision fusion methods fuse the two modalities before Softmax layer. Table 2 shows audio-visual event localization prediction accuracy of different multimodal feature fusion methods on AVE dataset. Our DMRN model in the late fusion setting can achieve better performance than all compared method. We also observe that late fusion is better than early fusion and decision fusion. The superiority of late fusion over early fusion demonstrates that temporal modeling before audio-visual fusion is useful. We know that auditory and visual modalities are not completely aligned, and temporal modeling can implicitly learn certain alignments between the two modalities, which is helpful for the audio-visual feature fusion task. The decision fusion can be regard as a type of late fusion but using lower dimension (same as the category number) features. The late fusion outperforms the decision fusion, which validates that processing multiple features separately and then learning joint representation using a middle layer rather than the bottom layer is an efficient fusion way.

Full and Weak Supervision: Obviously, supervised models are better than weakly supervised ones, but quantitative comparisons show that weakly-supervised approaches achieve promising event localization performance, which demonstrates the effectiveness of the MIL frameworks, and validates that the audio-visual event localization task can be addressed even in a noisy condition.

Cross-Modality Localization: We compare our method: AVDLN with DCCA [1] on cross-modality localization tasks: A2V (visual localization from audio segment query) and V2A (audio localization from visual segment query). The prediction accuracy of AVDLN and DCCA on the A2V and V2A tasks are 44.8/36.6 and 34.8/34.1, respectively. Our AVDL outperforms DCCA over a large margin both on A2V and V2A tasks. Even using the strict evaluation metric (which counts only the exact matches), our models on both subtasks: A2V and V2A, show promising results, which further demonstrates that there are strong correlations between audio and visual modalities.

Acknowledgement: This work was supported in part by NSF IIS 1741472, IIS 1813709, and CHE 1764415. This article solely reflects the opinions and conclusions of its authors and not the funding agents.

References

[1] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *ICML*, pages 1247–1255, 2013. 4

[2] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González. Gated multimodal units for information fusion. In *ICLR workshop*, 2017. 4

[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015. 2

[4] J. F. Gemmeke, D. P. 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. 1

[5] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *Proc. CVPR*, 2006. 1, 3

[6] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 1

[7] D. Kiela, E. Grave, A. Joulin, and T. Mikolov. Efficient large-scale multi-modal classification. In *AAAI*, 2018. 4

[8] J.-H. Kim, S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang. Multimodal residual learning for visual qa. In *NIPS*, pages 361–369, 2016. 4

[9] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 2017. 2

[10] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *Advances in neural information processing systems*, pages 570–576, 1998. 1

[11] J. Wu, Y. Yu, C. Huang, and K. Yu. Deep multiple instance learning for image classification and auto-annotation. In *Proc. CVPR*, pages 3460–3469, 2015. 3

[12] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICLR*, 2015. 2