A Two-Stage Framework for Multiple Sound-Source Localization

Rui Qian[1], Di Hu[2], Heinrich Dinkel[1], Mengyue Wu[1], Ning Xu[3], Weiyao Lin[1]
[1]Shanghai Jiao Tong University, [2]Baidu Research, [3]Adobe Research

1. Introduction

Humans usually perceive the world through information in different modalities, *e.g.*, vision and hearing. By leveraging the relevance and complementary between audio and vision, humans can clearly distinguish different sound sources and infer which object is making sound. In contrast, machines have been proven capable of separately processing audio and visual information using deep neural networks. But can they benefit from joint audiovisual learning?

Recent works mainly focus on establishing multi-modal relationship based on temporally synchronized audio and visual signals [1, 3, 8]. This synchronization works effectively for simple scenes [2, 9], *i.e.*, the single-source conditions. However, in unconstrained videos, various sounds are usually mixed, where the scene-level supervision is too coarse to provide the precise alignment between each sound and visual source pair. To tackle this problem, [6, 7] establish audiovisual clusters to associate sound-object pairs, but require pre-determined number of clusters, which is difficult in unconstrained scenarios, thus greatly affecting alignment performance. [2, 9, 11] further apply audiovisual learning into sound localization, but mainly focus on simple scenes, usually unable to find source-specific objects from mixed audio. [13] constructs a pretext task then localizes sound through energy of each pixel.

To sum up, existing dominant methods mostly lack the ability to analyze complex audiovisual scenes, and fail to effectively utilize the latent alignment between sound and visual source pairs in unconstrained videos. This is because there are majorly two challenges in complex audiovisual scene analysis: one is how to distinguish different sound-sources, the other is how to ensure the established sound-object alignment is fairly satisfactory without one-to-one annotations. To address these challenges, we develop a two-stage audiovisual learning framework. At the first stage, we employ a multi-task framework consisting of classification and audiovisual correspondence to provide the reference of audiovisual content for the second stage. At the second stage, based on the classification predictions, we use the operation of *Class Activation Mapping* (CAM) [14, 10, 4] to extract class-specific feature representations as the potential sound-object pairs (Fig. 1), then perform alignment in a coarse-to-fine manner, where the coarse correspondence based on category is evolved into the fine-grained matching in both video- and category-level.

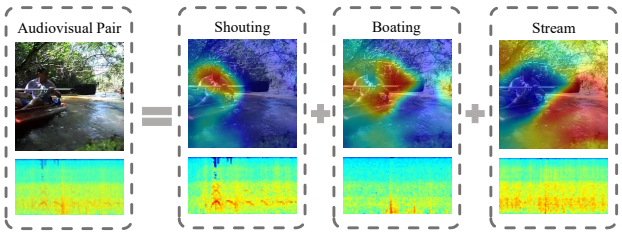

Figure 1. Our model separates a complex audiovisual scene into several simple scenes, which simplifies a complex scenario and generates several one-to-one audiovisual associations.

Our main contributions can be summarized as follows: (1) We develop a two-stage audiovisual learning framework to deal with the complex scenes. (2) We propose to establish audiovisual alignment in a coarse-to-fine manner. (3) We achieve state-of-the-art results on public dataset. In the multi-source conditions, according to our proposed class-specific localization metric, our method shows considerable performance compared with several baselines.

2. Approach

Our two-stage framework is illustrated in Fig. 2. At the first stage, we adopt multi-task learning for classification and video-level audiovisual correspondence. At the second stage, we use Grad-CAM [10] modules to disentangle class-specific features on both modalities, based on which we further perform fine-grained audiovisual alignment.

2.1. Multi-Task Learning Framework

Given audio and visual (image) messages, $\{a_i, v_i\}$ from *i-th* video, we use video tags or pseudo labels from pre-trained models as supervision for classification. We denote C as the number of class and c as the *c-th* class. Multi-label binary cross entropy loss is considered for classification:

$$L_{cls} = \mathcal{H}_{bce}(\boldsymbol{y}_{a_i}, \boldsymbol{p}_{a_i}) + \mathcal{H}_{bce}(\boldsymbol{y}_{v_i}, \boldsymbol{p}_{v_i}), \quad (1)$$

where $\mathcal{H}_{bce}$ is the binary cross-entropy loss, $\boldsymbol{y}$ and $\boldsymbol{p}$ are the class labels and corresponding predicted probabilities

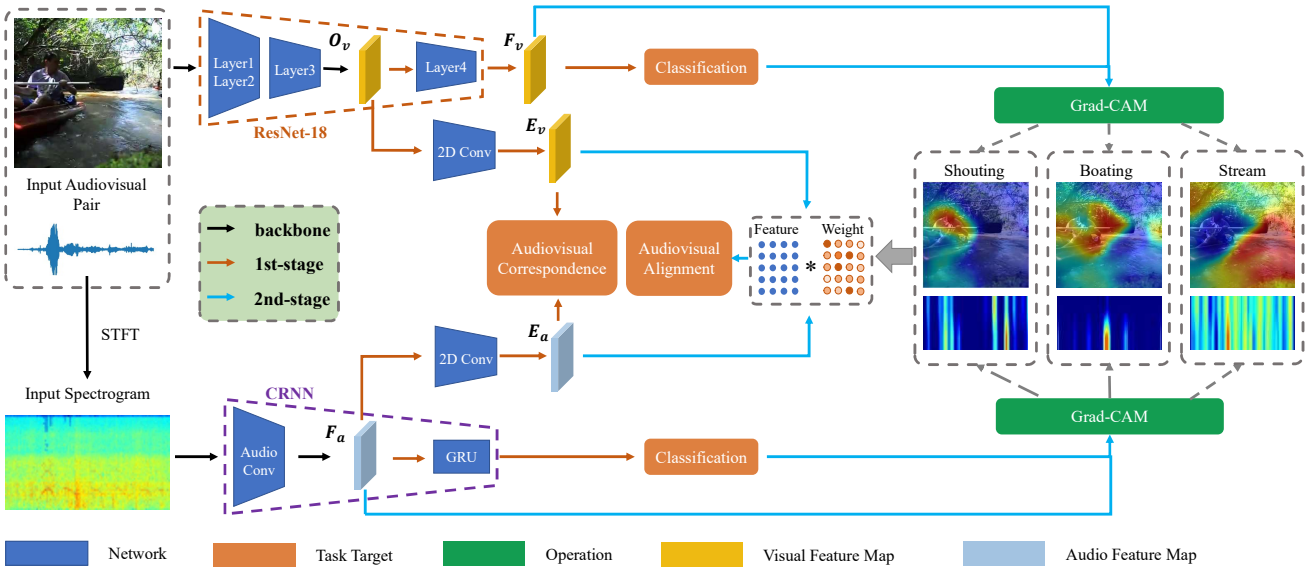

Figure 2. An overview of our two-stage audiovisual learning framework. At the first stage, the model performs classification and video-level correspondence. At the second stage, we disentangle class-aware representations and implement fine-grained audiovisual alignment.

respectively, $\boldsymbol{y} \in \{0,1\}^C$, $\boldsymbol{p} \in [0,1]^C$.

For audiovisual correspondence learning, similar to [1], we view it as a two-class classification problem. Specifically, we take F_a and O_v in Fig. 2 as inputs[1] for audiovisual correspondence network. Through a series of convolutions and concatenation, we get one 1024-D vector and pass it through two fully-connected layers of 1024-128-2. The 2-D output with softmax regression aims to determine whether audio and vision correspond. $\{a_i, v_i\}$ from *i-th* video are viewed as corresponding pair, the image of a randomly selected video, v_j, is used to construct mis-corresponding pair $\{a_i, v_j\}$. The learning objective can be written as:

$$L_{avc} = \mathcal{H}_{cce}(\boldsymbol{\delta}, \boldsymbol{q}), \tag{2}$$

where $\mathcal{H}_{cce}$ is the categorical cross entropy loss, $\boldsymbol{q} \in [0,1]^2$ is the predicted output, $\boldsymbol{\delta}$ is the class indicator, $\boldsymbol{\delta} = (0,1)$ for correspondence while $\boldsymbol{\delta} = (1,0)$ for not.

We take L_{mul} as final loss function for multi-task learning, λ is the hyperparameter of weighting:

$$L_{mul} = L_{cls} + \lambda L_{avc}. \tag{3}$$

After training with L_{mul}, we could achieve coarse-grained audiovisual correspondence in the category level.

2.2. Audiovisual Feature Alignment

In this section, we propose to disentangle feature representations of different categories and implement fine-grained audiovisual alignment.

[1]We choose F_a and O_v for two reasons: we can obtain more fine-grained local features and achieve easier training process.

2.2.1 Disentangle Features by Grad-CAM.

We leverage the operation of Grad-CAM [10] to perform disentangle class-specific features. For simplicity, we use $r \in \{a, v\}$ to represent audio or visual modality. Given the feature map activations of the last convolutional layer, F_r, and the output of classification branch without activation for class c, $\hat{p_r^c}$, we calculate the class-specific map W_r^c, *i.e.*,

$$W_r^c = \text{Grad-CAM}(F_r, \hat{p_r^c}). \tag{4}$$

Then we take W_r^c as weights to perform weighted global pooling over the feature map $E_r(u,v)$ to obtain class-aware representation[2], where u and v are the map entries. That is:

$$f_r^c = \frac{\sum_{u,v} E_r(u,v) W_r^c(u,v)}{\sum_{u,v} W_r^c(u,v)}. \tag{5}$$

Finally, we get C 512-D vectors as the feature representation of all the categories. $\{f_{a_i}^m | m = 1,2,...,C\}$ and $\{f_{v_i}^n | n = 1,2,...,C\}$ are the set of audio and visual class-specific feature representations for *i-th* video.

2.2.2 Fine-Grained Audiovisual Alignment.

To establish audiovisual alignment with disentangled features, we take pairs of the same class from the same video as positive pairs for alignment. As each category contains various entities (*e.g.*, the animal category contains audio and visual patterns of dogs, cats, birds *etc.*), with this sampling

[2]We find that directly using F_r with the weights W_r^c is difficult to perform alignment objective, but by performing weighted pooling on E_r, we achieve easier training and faster convergence.

strategy, we could acquire higher-quality positive pairs and establish sound-object association beyond category.

To establish cross-modality alignment, we project $f_{a_i}^m$ and $f_{v_j}^n$ into a shared embedding space via two fully-connected layers of 512-128-128. Then we compare the projected features using Euclidean distance,

$$D(f_{a_i}^m, f_{v_j}^n) = ||g_a(f_{a_i}^m) - g_v(f_{v_j}^n)||_2, \tag{6}$$

where g_a and g_v are the fully-connected layers. We then adopt contrastive loss to perform sound-object alignment. The loss function is written as[3]

$$L_{ava} = \sum_{i,j=1}^{N} \sum_{m} \sum_{n} (\delta_{i=j}^{m=n} D^2(f_{a_i}^m, f_{v_j}^n) + (1 - \delta_{i=j}^{m=n}) max(\Delta - D(f_{a_i}^m, f_{v_j}^n), 0)^2), \tag{7}$$

where $\delta_{i=j}^{m=n}$ indicates whether the audiovisual pair is positive, *i.e.*, $\delta_{i=j}^{m=n} = 1$ when $i = j$ and $m = n$, otherwise 0. Δ is a margin hyper-parameter.

2.3. Visual Localization of Sounds.

We use learned representations to visually localize sounds by generating source-aware localization maps. To leverage the established alignment, the visual feature map E_{v_i} of testing image is firstly projected into the shared embedding space via the fully-connected layers of g_v in Eq. 6, then compared with the disentangled *c-th* class audio features $f_{a_i}^c$ through Eq. 8,

$$K_i^c(u, v) = ||g_a(f_{a_i}^c) - g_v(E_{v_i})(u, v)||_2. \tag{8}$$

The obtained $K_i^c \in \mathbb{R}^{U \times V}$ is then normalized and resized to the original image size as the final localization maps for sound source in the *c-th* class.

3. Experiments

In this work, we train and evaluate our model on SoundNet-Flickr [3], AudioSet-instrument [5], and also show some qualitative examples on AVE dataset [12].

3.1. Sound Localization on SoundNet-Flickr

We implement quantitative evaluation on 249 audiovisual paris in the subset of SoundNet-Flickr used in [11]. Consensus Intersection over Union (cIoU) and Area Under Curve (AUC) [11] are reported. We use weighted summation over class-specific localization maps of valid categories as final results, where the weights are normalized predicted probabilities. Table 1 shows the results of different methods. Despite that most audiovisual pairs in test set are of single-source, our model still outperforms all the other methods including CAM results from the first stage. It demonstrates that our fine-grained alignment effectively facilitates audiovisual learning in unconstrained videos.

[3]In practice, a threshold is considered to select valid categories.

Table 1. Quantitative localization results on SoundNet-Flickr.

Methods	cIoU@0.5	AUC
Random	7.2	30.7
Attention[11]	43.6	44.9
DMC AudioSet[6]	41.6	45.2
CAVL AudioSet[7]	50.0	49.2
Ours CAM	44.2	48.1
Ours	**52.2**	**49.6**

Table 2. Quantitative results on AudioSet. The cIoU_class threshold is 0.5 for level-1 and level-2, and 0.3 for level-3. Note that †AVC method is evaluated in a class-agnostic way.

Methods	level-1		level-2		level-3	
	cIoU_c	AUC	cIoU_c	AUC	cIoU_c	AUC
†AVC	24.8	32.0	4.27	23.6	5.3	14.9
Multi-task	20.6	29.5	2.37	17.4	10.5	17.8
Ours	**32.8**	**38.3**	**6.16**	**23.9**	**21.1**	**22.0**

3.2. Multi-Source Localization on AudioSet

To better evaluate sound localization performance in multi-source scenes, we propose to use cIoU and AUC metric in a class-aware manner. We use the detected bounding boxes with category labels of Faster RCNN to indicate the localization of sounding objects[4]. Next, we calculate cIoU scores on each valid sound source and take an average. Final cIoU_class on each frame can be calculated by

$$\text{cIoU\_class} = \frac{\sum_{c=1}^{C} \theta_c cIoU_c}{\sum_{c=1}^{C} \theta_c}, \tag{9}$$

where c indicates the class index of instruments, $\theta_c = 1$ if instrument of class c makes sounds, otherwise 0.

We compare with two baselines: (1) AVC: only using video-level audiovisual correspondence, and inferring the sound locations in a class-agnostic way. (2) Multi-task learning: using both of classification and audiovisual correspondence. We report results on three difficulty levels, *i.e.*, single-source (level-1), two-source (level-2) and three-source (level-3), in Table 2, and there are several observations. First, localizing sound in a class-agnostic way is effective with limited sounding objects, but fails with more sources. This is because the video-level correspondence is too coarse to provide sound-object association in complex scenes. Second, although AVC takes a much looser evaluation metric of class-agnostic, it is still worse than the multi-task method on level-3, which reveals classification significantly helps to distinguish sounds of different sources. Third, our method significantly outperforms two baselines and is robust on all difficulty levels. It demonstrates that our fine-grained alignment is effective to establish one-to-one association in both single-source and multi-source scenes.

[4]We have filtered out those silent detected objects.