Self-Supervised Segmentation and Source Separation on Videos

Andrew Rouditchenko*[1], Hang Zhao*[1], Chuang Gan[2], Josh McDermott[1], Antonio Torralba[1]
[1]MIT [2]MIT-IBM Watson AI Lab
{roudi,hangzhao,jhm}@mit.edu, {ganchuang,torralba}@csail.mit.edu

1. Introduction

Semantic segmentation of images [11, 3] and sound source separation in audio [8, 4, 1] are two important and popular tasks in the computer vision and computational audition communities. Traditional approaches have relied on large, labeled datasets, but recent work has leveraged the natural correspondence between vision and sound to apply supervised learning without explicit labels. In this paper, we develop a neural network model for visual object segmentation and sound source separation that learns from natural videos through self-supervision. The model is an extension of recently proposed work that maps image pixels to sounds [9]. This paper is a workshop edit of Rouditchenko et al. 2019 [5].

In the Mix-and-Separate framework proposed in [9], neural networks are trained on videos through self-supervision to perform image segmentation and sound source separation. However, following training, the model could only be applied to videos with synchronized audio, limiting their use in real applications where synchronized data are not available. Here we seek to enable a system that can perform segmentation and separation tasks using test input containing only video frames or sound mixtures. We introduce a learning approach that disentangles concepts learned by neural networks, enabling independent inference of images and audio mixtures without needing to combine visual and auditory features.

We evaluate performance on image-only and audio-only tasks, which was not possible using the previous model. Furthermore, we substantially extend the scale of previous work [9] by training on a video dataset of naturally occurring audio-visual events with 28 event categories and over 4000 videos [6]. The results show that we can achieve promising semantic segmentation and source source separation performance.

2. Self-Supervised Cross-Modal Training

Our approach adopts the Mix-and-Separate framework used in [9], which first generates a synthetic sound separation training set by mixing the audio signals from two different videos, and then trains a neural network to separate the audio mixture conditioned on the visual input corresponding to one of the audio signals. As shown in Fig. 2, the framework we use consists of three components: an image analysis network, an audio analysis network, and an audio synthesizer network. The learning method is self-supervised because the neural networks do not require labelled data for training.

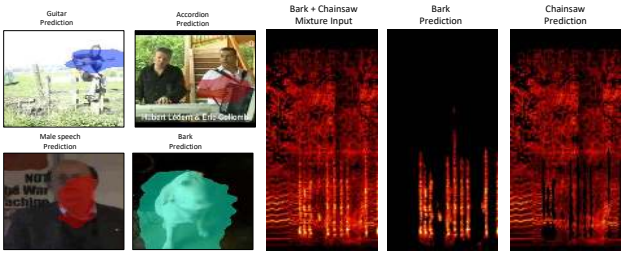

Figure 1. Joint audio-visual training and independent image and audio inference. After our neural networks are trained jointly on images and sounds from videos using our proposed learning method, they can be used independently for image-only semantic segmentation and audio-only source separation.

2.1. Disentangling Internal Representations

We designed a learning schedule with the sigmoid and softmax activation functions to disentangle the learned internal representations before the audio synthesizer network combines audio and visual features. Our technique anneals the temperature parameter in the softmax activation function in order to push output activations towards one-hot vectors. As the temperature parameter T in the softmax activation function changes from high to low, the shape of the output distribution changes from uniform to one-hot:

$$y_k = \frac{\exp(\frac{\phi_k}{T})}{\sum_{i=1}^{n} \exp(\frac{\phi_i}{T})}, \quad (1)$$

where y_k is the value of the k_{th} visual feature channel after activation, T is the temperature, and ϕ_i is the value of the i_{th} visual feature channel before activation.

The model is initially trained using a sigmoid activation on the visual feature vector ϕ, which leads to diverse activa-

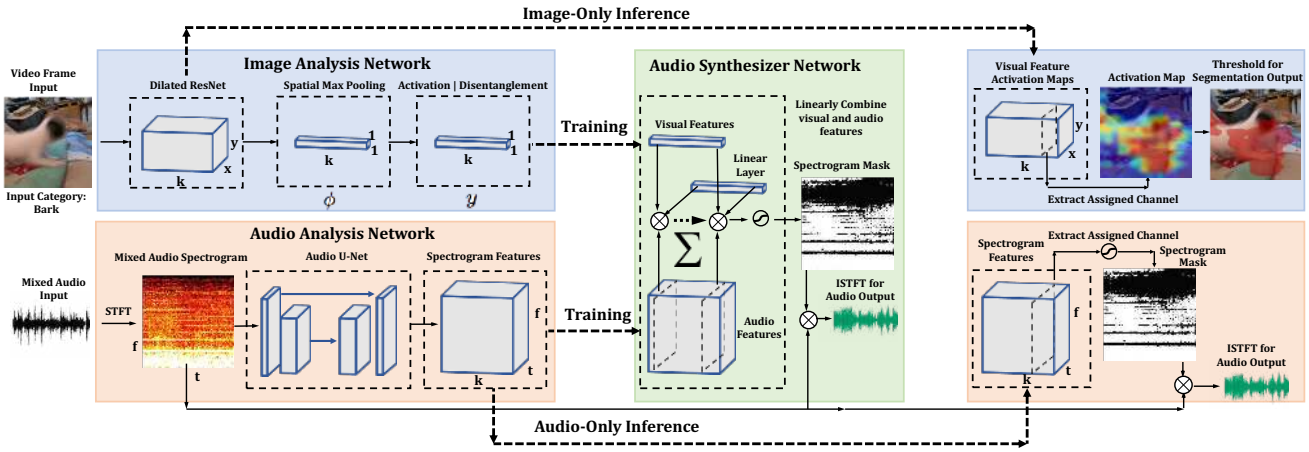

Figure 2. Joint audio-visual training and independent image and audio inference. After training on synthetic mixtures of videos, the image analysis network performs image-only segmentation and the audio analysis network performs audio-only source separation.

tions and helps with convergence to an initial solution. The sigmoid activation is then replaced with the softmax activation, and the temperature is gradually decreased, pushing the visual feature vector toward a one-hot vector, and causing the visual and audio feature representations to become sparse and disentangled.

2.2. Category to Channel Assignment

After training the networks without labels, we then use the video labels in the dataset to match categories to network feature channels. We use a matching algorithm [2] which assigns each dataset category to a network feature channel according to how strongly the visual feature vector is activated. For example, the dataset category, "cars," could correspond to the first network channel, "male speech" to the second network channel, and so forth.

The assignment of input categories to network feature channels allows independent image and audio processing without needing the audio-synthesizer network to combine the features. For object segmentation, the last spatial max pooling layer of the image analysis network is removed to preserve activation feature maps. Given an input video frame, the activation map in the channel assigned to the video's category is selected, upsampled to the input size, and thresholded to obtain a predicted segmentation. Given an audio mixture, the audio analysis network outputs spectrogram features. The channels assigned to the two source video categories are selected, and used as spectrogram masks to recover each respective source.

3. Experimental Results

The learning schedule was implemented with two stages: a training stage with a fixed sigmoid activation function and a fine-tuning stage with a softmax activation function and varying schedules for the temperature parameter. The custom schedules varied the initial temperature, the decay rate, and the epochs at which the temperature was decayed, which proved to be important.

Model Name	Learning Schedule			Performance		
	Temp.	Decay	Epochs	SDR	SIR	IOU
Sigmoid Only	-	-	-	0.865	6.04	0.204
Softmax Only	1.0	0.3	10, 20	0.172	3.37	0.207
Schedule A	10.0	0.5	4, 8, 12, 16	-0.536	4.52	0.112
Schedule B	1.5	0.75	4, 8, 12, 16	0.341	6.23	0.152
Schedule C	1.0	0.3	4, 8	0.716	6.21	**0.232**
Schedule D	1.0	0.3	3, 6, 9, 12	-1.88	2.82	0.205
Schedule E	1.0	0.5	5, 10, 15	**1.03**	**6.37**	0.225
NMF [8]	-	-	-	0.196	3.94	-
CAM [10]	-	-	-	-	-	0.190

Table 1. Sound Separation (SDR, SIR) and semantic segmentation (IoU) performance.

In Table 1, we show the performance of the proposed model ("Schedule") with several different schedule settings for the softmax fine-tuning stage, as well as of the baseline models. The sigmoid training stage is identical for $A-E$ of our proposed model. The sound source separation metrics are the Signal to Distortion Ratio (SDR) and the Signal to Interference Ratio (SIR) [7], and the semantic segmentation metric is Intersection over Union (IoU). The best sound separation performance was achieved by our proposed model, "Schedule E," and the best segmentation performance was achieved by our proposed model, "Schedule C," although "Schedule E" performed nearly as well. Qualitative results are shown in Fig. 1. Qualitatively, the model succeeds in separating the sound from different sources to a large extent, which is visible in the source spectrogram recovery. Despite the low resolution of the activation maps, the locations of the predicted segmentations were fairly accurate. Our model is also more interpretable than the baseline models because it is sparsely activated. Overall, we achieve promising results.

References

[1] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE, 2016.

[2] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[3] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. CVPR*, pages 3431–3440, 2015.

[4] Josh H McDermott. The cocktail party problem. *Current Biology*, 19(22):R1024–R1027, 2009.

[5] Andrew Rouditchenko, Hang Zhao, Chuang Gan, Josh McDermott, and Antonio Torralba. Self-supervised audio-visual co-segmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019.

[6] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[7] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4):1462–1469, 2006.

[8] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing*, 15(3):1066–1074, 2007.

[9] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[10] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proc. CVPR*, 2016.

[11] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proc. CVPR*, 2017.