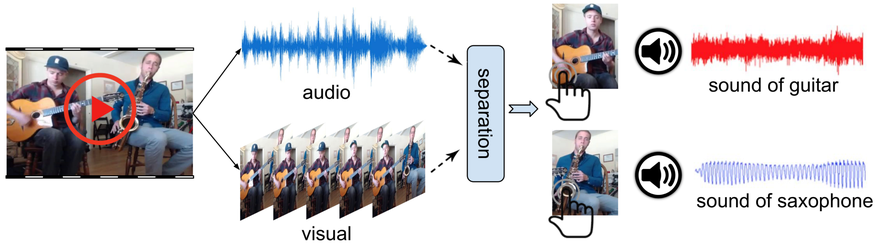
Learning to Separate Object Sounds by Watching Unlabeled Video

Ruohan Gao[1] Rogerio S. Feris[2] Kristen Grauman[1,3]

[1]The University of Texas at Austin, [2]IBM Research, [3]Facebook AI Research

Audio-Visual Source Separation

Goal: audio-visual object source separation in videos



Traditional audio-visual approaches:

- Detect low-level correlations within a single video
- Learn from clean *single audio source* examples

[Darrell et al. 2000; Fisher et al. 2001; Rivet et al. 2007; Barzelay & Schechner 2007; Parekh et al. 2017; Pu et al. 2017 ...]

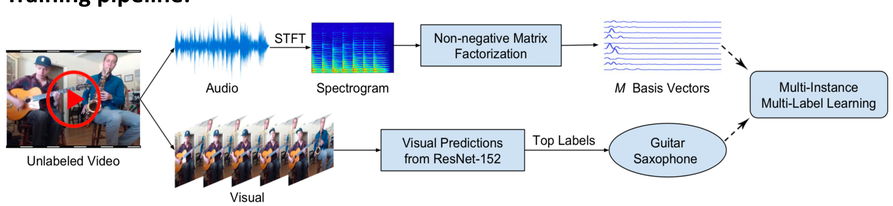
Our idea: We leverage visual objects to learn from *unlabeled* video with *multiple* audio sources to obtain a repertoire of objects and their sounds.

woof **meow** **ring** **clatter**

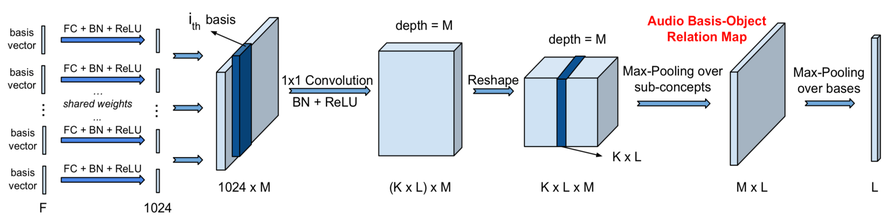
Approach

We learn what different objects sound like from a batch of unlabeled, multi-sound-source videos through deep multi-instance multi-label learning (MIML).

Training pipeline:



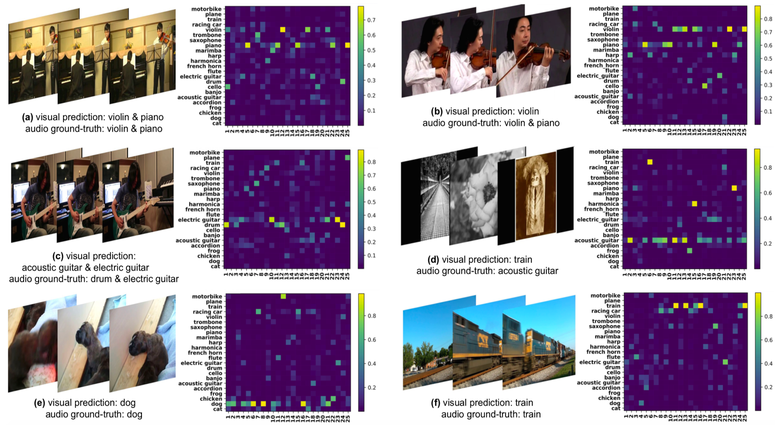
Deep multi-instance multi-Label network:



The deep MIML network takes a bag of audio basis vectors for each video as input, and gives a bag-level prediction of the objects present in the video.

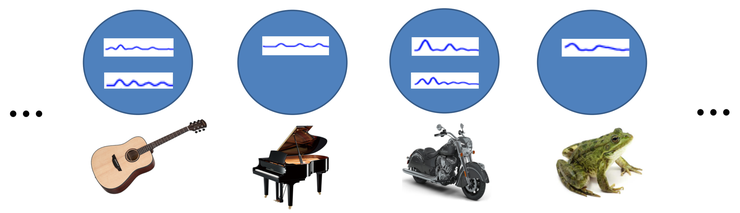
Audio basis-object relation discovery:

We collect high-quality representative bases for each object category by inspecting the audio basis-object relation map.

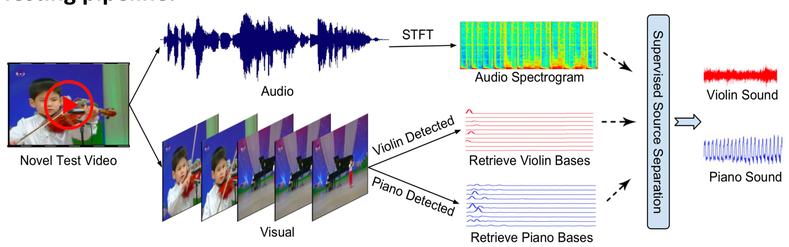


The MIML network successfully learns the prototypical spectral patterns of different sounds, and is capable of associating audio-bases with object categories.

Disentangling output: Group of audio basis vectors per object class.



Testing pipeline:



Semi-supervised NMF: The retrieved audio bases are used to "guide" NMF-based audio source separation.

$$\mathbf{V}^{(q)} \approx \tilde{\mathbf{V}}^{(q)} = \mathbf{W}^{(q)}\mathbf{H}^{(q)}$$
$$= \left[\mathbf{W}_1^{(q)} \cdots \mathbf{W}_j^{(q)} \cdots \mathbf{W}_J^{(q)}\right] \left[\mathbf{H}_1^{(q)} \cdots \mathbf{H}_j^{(q)} \cdots \mathbf{H}_J^{(q)}\right]^T$$

J: number of detected objects (potential sound sources).
The spectrogram for each detected object: $\mathbf{V}_j^{(q)} = \mathbf{W}_j^{(q)}\mathbf{H}_j^{(q)}$

Results

Dataset: AudioSet [Gemmeke *et al.* 2017] as the source of unlabeled videos. 193k video clips of musical instruments, animals, and vehicles.

Visually-aided audio source separation: (in SDR)

	Instrument Pair	Animal Pair	Vehicle Pair	Cross-Domain Pair
Upper-Bound	2.05	0.35	0.60	2.79
K-means Clustering	-2.85	-3.76	-2.71	-3.32
MFCC Unsupervised	0.47	-0.21	-0.05	1.49
Visual Exemplar	-2.41	-4.75	-2.21	-2.28
Unmatched Bases	-2.12	-2.46	-1.99	-1.93
Gaussian Bases	-8.74	-9.12	-7.39	-8.21
Ours	**1.83**	**0.23**	**0.49**	**2.53**

Our method achieves large gains, and it also has the capability to match the separated source to meaningful acoustic objects in the video.

Visually-aided audio denoising: (in NSDR)

	Wooden Horse	Violin Yanni	Guitar Solo	Average
Sparse CCA (Kidron et al. 2005)	4.36	5.30	5.71	5.12
JIVE (Lock et al. 2013)	4.54	4.43	2.64	3.87
Audio-Visual (Pu et al. 2017)	8.82	5.90	**14.1**	9.61
Ours	**12.3**	**7.88**	11.4	**10.5**

Our method is better than / competitive with all prior audio-visual methods, including Pu et al . 2017, which is tailored to noise separation.

Concurrent work on audio-visual source separation:

Owens & Efros, ECCV 2018; Ephrat et al., SIGGRAPH 2018; Zhao et al. ECCV 2018; Afouras et al. arXiv 2018

We study a broader set of object-level sounds including instruments, animals, and vehicles. We are the first to learn from uncurated "in the wild" videos that contain multiple objects and multiple audio sources.

Please visit our project page for video results/code/data:
http://vision.cs.utexas.edu/projects/separating_object_sounds/