Visually Indicated Sounds

April 2021

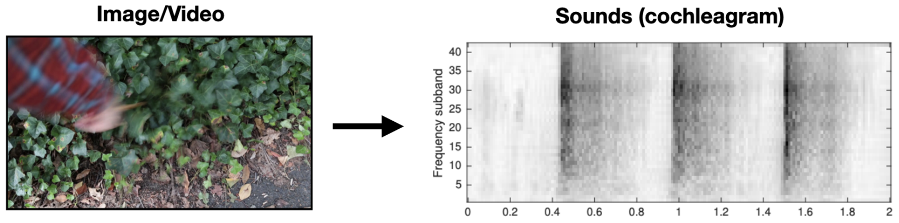

Figure 1: The task you will try to solve: predict the sounds that should accompany an input image or video.

When you see photo of a beach, it's easy to imagine the sound of wind and waves that would naturally accompany such a scene. This ability to correlate sensory experience across modalities is considered to be crucial to the development of perceptual systems that can reason coherently about all the different types of information they take in [3]. Being able to predict sounds from images or silent videos also makes for really fun applications. For example, you could add sounds to a movie from the silent films era, or you could add a soundtrack to a photo album of you on vacation. Recently there has been quite a bit of interest in the computer vision community on several versions of this problem [1, 2, 4].

In this project, your goal will be to take an image or video as input, and predict a sound as output. For training data, you could use any source of videos with audio tracks, where you simply split the video into the visual frames as the inputs to your system and the audio as the output. To represent sounds, you may want to consider the cochleograms used in [1], or you could use a spectrogram representation.

Some good datasets to consider are:

1. Drumsticks hitting objects: `http://andrewowens.com/vis/`
2. Talking people: `https://www.robots.ox.ac.uk/ vgg/data/voxceleb/`
3. Scenes with audio: `https://www.robots.ox.ac.uk/ vgg/data/vggsound/`

As a twist, you could instead try predicting images from sounds. Given audio of birdsong, can you predict what the bird might look like? Feel free to be creative and come up with your own variants on the problem!

References

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

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

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

[4] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018.