Silent video to sound

Computer Vision Student

October 2018

Have you ever taken an airplane without earphones and tried to watch the television without sound? In this project, you will create a computer vision system to recover sounds from silent video.

There are a couple of ways you could do this. Since most videos have sound, you can train a neural network to generate waveforms from pixels [5, 4]. To make the sounds realistic, you will probably need to use a generative adversarial network [3] so that sounds are on the natural manifold. The datasets in [5, 4] are probably good places to start, but you can always collect more complex videos.

You could also learn to embed sounds and video into a common space, and then just retrieve the nearest sounds [2]. With an embedding, you won't need to learn natural sound priors.

Finally, you could also go the other way. Given a sound, can you generate the video? Our previous work might get you started [1], but there is still much to do.

References

[1] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, pages 892–900, 2016.

[2] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*, 2017.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative ad-

versarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[4] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2405–2413, 2016.

[5] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. *arXiv preprint arXiv:1712.01393*, 2017.