Supplementary Material:
Learning to Localize Sound Source in Visual Scenes

Arda Senocak[1] Tae-Hyun Oh[2] Junsik Kim[1] Ming-Hsuan Yang[3] In So Kweon[1]
Dept. EE, KAIST, South Korea[1]
MIT CSAIL, MA, USA[2]
Dept. EECS, University of California, Merced, CA, USA[3]

Summary

The contents of this supplementary material include details of our architecture. Sample visualization results (sound plays, interactive sound localization, video localization results) can be found in the supplementary video. Dataset is available for download.

1. Architecture details

Optimization We optimize our network by using the stochastic gradient descent and back propagation. We use the Adam optimizer [3] with fixed learning rate 0.0001. We train with a batch size of 30 for 6 epochs and the training time takes almost a day using an implementation in TensorFlow [1].

Architecture We adapt our architecture from the VGG16 model [4] for the visual CNN and the SoundNet [2] for the sound CNN. For the visual CNN, we resize a frame to 320×320 as input.

For the architecture design, we use the following notations: `c3s1-k` denotes a 3×3 convolution (for sound, 3×1) and ReLU layer with k filters and stride 1, and `pool3s2` denotes 3×3 max-pooling (for sound, 3×1) with stride 2.

The sound CNN consists of: `c64s2-16`, `pool8s1`, `c32s2-32`, `pool8s1`, `c16s2-64`, `c8s2-128`, `c4s2-256`, `pool4s1`, `c4s2-512`, `c4s2-1024`, and `c8s2-1000`

The visual CNN consist of: `c3s1-64`, `c3s1-64`, `pool2s2`, `c3s1-128`, `c3s1-128`, `pool2s2`, `c3s1-256`, `c3s1-256`, `c3s1-256`, `pool2s2`, `c3s1-512`, `c3s1-512`, `c3s1-512`, `pool2s2`, `c3s1-512`, `c3s1-512`, and `c3s1-512`

2. Additional Qualitative Analysis

We present additional qualitative results in this supplementary material and video. We recommend to refer to the supplementary video for video results.

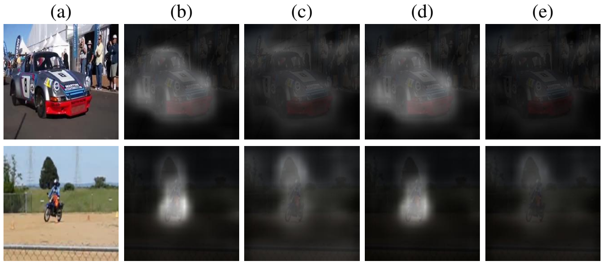

Figure 1. **Ambient sound results.** We show some examples of frames with ambient sounds. (a) sampled input frame. (b) location response against object indicating sound in *Softmax* only attention mechanism. (c) location response against ambient sound in *Softmax* only attention mechanism. (d) location response against object indicating sound in *ReLU+Softmax* attention mechanism. (e) location response against ambient sound in *ReLU+Softmax* attention mechanism. The proposed network gives noticeably distinguished confidences between object-like and ambient sounds

Ambient Sound Analysis We analyze the our proposed method with non-object/ambient sounds as well, *e.g*., environmental sounds, wind sounds, background activities, and narration. We feed the frames with one of these ambient sounds into our network to see how it reacts. Figure 1 shows that the proposed method gives noticeably low confidence scores to ambient sound, and high reaction to the object indicating sound. From the Figure 1 (c) and (e) we can observe the ReLU+softmax mechanism performs better with ambient sounds. This is due to the effect of ReLU operation clipping the negative values in attention map to zero in the training phase.

Our attention map is an outcome of inner products between normalized vectors. Therefore the values are ranged between -1 and 1. The negative values in the attention map indicate low or negative correlations while the positive values are likely to be sound source location. Without ReLu, softmax maps low or negative correlation responses to positive values. However, for ReLU applied case, all the negative values are converted to 0 making the attention range between 0 and 1. As a result, ReLU+softmax mechanism is better at suppressing uncorrelated sound responses.

We visualize the results of attention response before softmax to show values in absolute (*i.e.*, non-relative) scale. The responses of ambient sound is relatively weak than object sounds. To make a clear visual comparison, we use gray scale heatmaps in Figure 1.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 1

[2] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *Neural Information Processing Systems*, 2016. 1

[3] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1