On Learning Association of Sound Source and 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]

1. Introduction

The sight (vision) and hearing (audition) senses are the most important sources that humans use to understand their surroundings. Visual events are typically associated with sounds and they are combined. For instance; When we see that a car is moving, we hear the engine sound at the same time, *i.e.*, co-occurrence. Humans observe tremendous number of combined visual-audio examples and learn the correlation between them throughout life-long observations unconsciously. Because of the correlation between the sound and the visual events, humans can understand the object or the event that causes sound and can localize the sound source even without separate education. Naturally, videos and their corresponding sounds also come together in a synchronized way.

Given a plenty of video and sound clip pairs, can a machine model learn to associate the sound with visual scene to reveal the sound source location without any supervision in a way similar to human perception to localize sound sources in visual scenes?

In this paper, we are interested in exploring whether computational models can learn the spatial correspondence between visual and audio information by leveraging the correlation between visuals and sound based on simply watching and listening to videos in unsupervised way. We address this challenge by designing our model with a two-stream network architecture (sound and visual networks [2]) where each network leverages each modality and a localization module which contains attention mechanism [4] as in Figure 2.

Experiment setup. We first experimented unsupervised setup by making the network just watch typical Youtube videos with their paired sounds. Also, we devise a way to add simple supervised loss, so that we can feed human knowledges of localization information seamlessly. We used both unsupervised and supervised data according to the availability of the annotated data, *i.e.*, semi-supervised setting. To enable semi-supervised learning and to evaluate the proposed models and the localization results, we introduce a dataset which consists of the sound source locations for given sound and visual pairs. To the best of our knowledge, there is no publicly available dataset that addresses the problem of learning based sound localization.

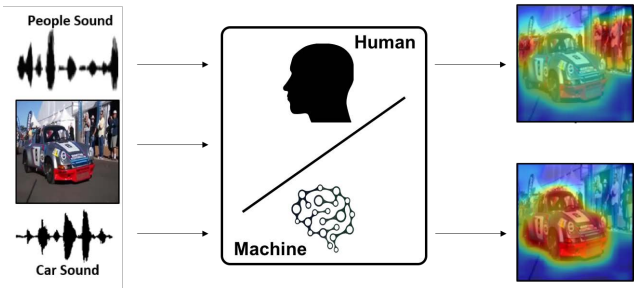

Figure 1. **Where do these sounds come from?** In this paper, we demonstrate how to learn to localize the sources (objects) of the sounds from the sound signals.

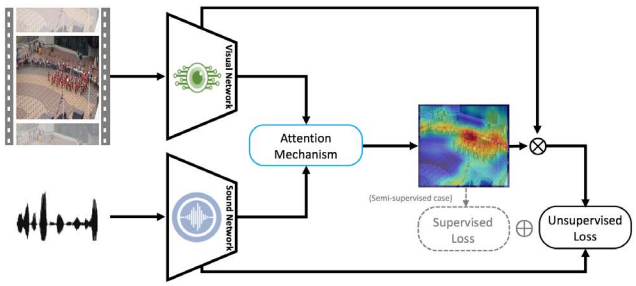

Figure 2. **Network Architecture** This architecture tackles the problem of sound source localization with unsupervised learning. The network uses frame and sound pairs to learn to localize sound source. Each modality is processed in its own network. After correlating the information from the sound and the visual network, attention mechanism localizes the sound source. By adding supervised loss component into this architecture, it can be converted to a unified architecture which can work as supervised or semi-supervised learning as well.

2. Results

Our network learns to localize sound sources on a variety of categories without any supervision. It is interesting to note that sound sources are successfully localized in-

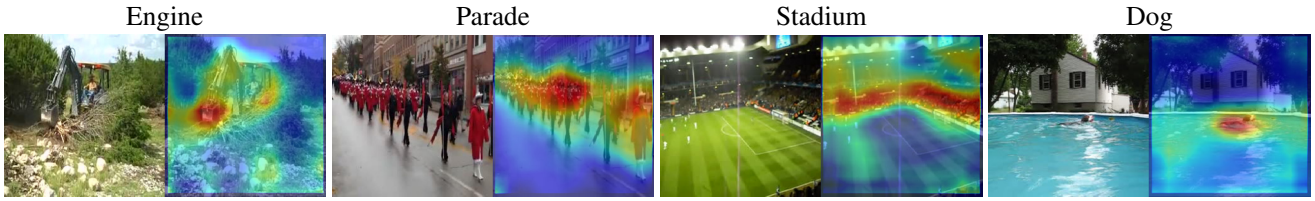

Figure 3. **Qualitative Sound Localization Results from Unsupervised Network.** We visualize some of the sound source locations. We feed image and sound pairs through our unsupervised network and it highlights the regions that sound is originated. Titles of the columns are subject and shown only for visualization purpose to give an idea about the sound context to readers: We do not use explicit labels.

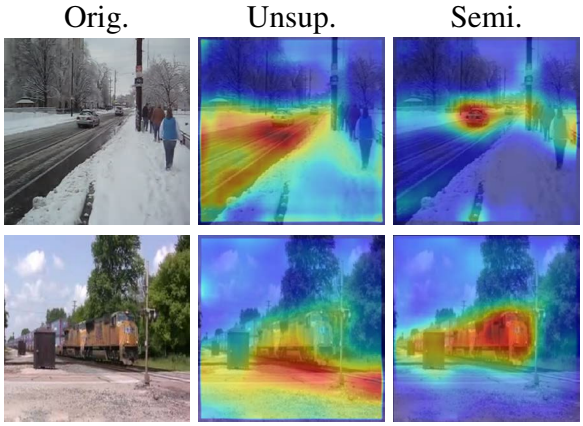

Figure 4. **Failure Cases of the Unsupervised Approach and its Correction by the Semi-supervised One.** We show some cases where proposed unsupervised network draws false conclusions. We correct this issue by providing a prior knowledge (we used human annotation in this work).

the-wild visuals. Figure 3 shows some qualitative results in unsupervised setting. While the network learns to localize sound sources in wide range of categories, from our experiments with the proposed unsupervised model, we observe some results which network draws contextually mismatched conclusions for the given visual and audio pairs, *i.e.*, pigeon superstition phenomenon in the animal learning theory [3, 1]. Since the relationship between source and result information was not trivial, the learner made a wrong decision with high confidence in that there is no way to validate and correct such superstition for the learner only with unsupervised loss. In order to validate this hypothesis, we strengthened the learner with a small number of human supervision as a clear and representative way of imposing human knowledge. By adding a supervised loss to our network, our unified architecture, which is be able to learn either in unsupervised or semi-supervised settings based on the availability of the annotated data, remedies such imperfection. Figure 4 shows the results from unsupervised and semi-supervised methods.

3. Discussion and Conclusion

In this work, we show the empirical learnability of the task, *i.e.*, the learning-based mono-channel sound source localization in visual scene. The conclusion from our experiments is that the task would not be a learnable problem in a pure unsupervised way. Furthermore, it suggests that, to pose this problem a proper way, we need additional prior knowledges that can guide the relationship between source and result information. As an addendum, it is interesting to see that the learner showed a similar behavior appeared in the animal learning context, *i.e.*, pigeon superstition.

Note that humans leverage the time difference of arrival (TDoA) by two ears to localize the sound source in 3D space, which is a significant benefit over our experiment setup. In this work, since we focus on the context association of sound and visual modalities, we isolate the TDoA factor. However, appending such another axis into the machine learner would be the interesting direction to a future work.

References

[1] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014. 2

[2] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 1

[3] B. F. Skinner. "Superstition" in the pigeon. *Journal of experimental psychology*, 1948. 2

[4] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, 2015. 1