Review on “SoundNet: Learning Sound Representations from Unlabeled Video”

Authors: Y. Aytar, C. Vondrick, A. Torralba

# Short Summary

In this paper, the authors propose a novel method of learning sound representations through jointly learning from unlabeled video and its corresponding audio. As this data is readily available online in bulk (model training used one year of continuous natural video), it allows for creating deeper models less prone to overfitting. Their proposed solution takes inspiration from Transfer Learning and teacher-student models; herein, the teacher works with vision to train the student on audio.

Video is inputted as frames of an image and audio is only minorly preprocessed (down sampled to 22kHz with mono channel). That is, the audio data is taken learned in its raw waveform. The network performs one-dimensional convolutions (for translational invariance on sound) and is fully convolutional allowing for variable length audio. The output layer of the network is also convolution producing multiple outputs over the timesteps in the video. For loss, KL divergence is optimized because the outputs of the vision network can be interpreted as a distribution of categories; both scene (Places) and object (ImageNet) networks are considered during the optimization (K = 2). Ultimately, to perform sound classification, a small amount of audio data is used to train a linear multi-class (one vs all) SVM where the internal representation of a hidden layer in the network acts as the features for training this classifier.

Following training, SoundNet exceeds state-of-the-art on DCASE, ESC-10 and ESC-50 acoustic scene classification tasks by around 10% classification accuracy. This datasets contain short recording of natural sound spread across a variety of categories and groups. Performance, while still below the human baseline, is beginning to approach human-levels at this task by leveraging the amount of data in the network. The authors also completed an ablation analysis demonstrating that KL divergence far outperforms L2 loss, using both vision networks improves performance over using a single one, and that the unlabeled video is a key factor in improving performance. Their results also showed that both VGG and AlexNet as underlying structures provided distinct benefits in different cases.

Finally, the authors provide some visualizations to show that both vision and audio features contain useful discriminatory information, and that the learned filters are in fact learning waveforms that correspond to specific categories.

# Main Contributions

1. Proposed SoundNet model architecture and methodology to take advantage of unlabeled video
2. Demonstrated that their model exceeds state-of-the-art at acoustic scene classification by a large margin
3. Performed an ablation analysis to provide insight into their model selection
4. Provided visualizations to demonstrate what the model filters are learning

# High-Level Evaluation of Paper

Having done some work with music information retrieval in the past, I found this paper quite fascinating. Not just in terms of the novel approach to solving an audio problem, but in terms of the additional work that went into empirically justifying decision decisions and visualizing what the model was learning. Having some experience with sound, I did notice a few key issues in the paper.

The authors did not justify many of their decisions nor provide any intuition behind why they felt that something they proposed would work best. For example, why were raw waveforms used instead of Mel-spectrograms as an audio representation? Has there been any recent research to suggest that the former is a better than the latter? I feel that this is a critical aspect of working with sound that should have been explained. This also extends into their choice of using a one-dimensional convolutional model. Using a spectrogram could have allowed them to take of frequency content and two-dimensional spectrograms which have also shown impressive performance in MIR tasks. I am not claiming that the authors’ decisions were incorrect, but rather I would have liked to hear more about why they made such decisions.

While the final model obtains impressive performance on the two acoustic scene classification challenges, I wonder what the model is truly learning. The authors claim that some of the filters respond aggressively to certain inputs, but what aspect of said inputs is the model learning. If it can differentiate between a child and adult talking, is this due to speech patterns or voice pitch? Humans intuitively use both, but the model only has access to a waveform. Auralization of the learned features would be an interesting direction to explore next.