Review on “Look, Listen and Learn”

Authors: R. Arandjelovic, A. Zisserman

# Summary

In this paper, the authors investigate audio-visual correspondence (AVC) by introducing a novel learning task: determining whether an audio clip corresponds to an image in a video via a binary classification. They propose that this correspondence can be learned in an unsupervised manner and show that this approach achieves state-of-the-art on sound classification benchmarks. Furthermore, it is shown that the learned image features are also comparable to state-of-the-art on self-supervised classification approaches on ImageNet.

The AVC task is described as a co-training one where there are two views of the data providing complementary information. Herein, the two videos are visual and audio data. Given that the approach taken to solve this problem is unsupervised, the authors specify that the training is subject to noise and unclean audio sources. However, despite this, performance is still competitive.

The model (Look, Listen and Learn or can be divided into three sections: the visual network, the audio network and the fusion network. The first in the list is inspired by VGG and produces a single 512-dimensional feature vector from a three-channel color image. The audio network also produces a 512-dimensional feature vector using a similar architecture but does so using 1-dimensional log-scaled spectrograms as input (these represent 1 second of audio). The fusion network concatenates the two vectors and then passes them through fully connected layers for classification on whether they correspond to one another.

To train the networks, the Flickr-SoundNet and Kinetic-Sounds datasets are used. Using the labels in the Kinetic-Sounds dataset, the authors are able to show that a supervised model undergoing pre-training achieves the same results (74% accuracy on Kinetic-Sounds) as the unsupervised model. The model also achieves 78% accuracy on Flickr-SoundNet correspondence. Extending these results to audio scene classification with the ESC-50 and DCASE datasets, a linear SVM is used for classification once features have been extracted. The proposed model achieves state-of-the-art performance with accuracies of 79.3% and 93% on each respective task. In self-supervised image classification, the model achieves an accuracy of 32.3% which is competitive with current state-of-the-art.

Investigating what is being learned by the network, the authors also discover that it is possible for the model to localize objects in an image frame and perform fine-grained recognition tasks. It is suspected that this confirms that the model is indeed learning semantic features and is mostly evaluated using qualitative evidence.

# Main Contributions

1. Proposed an alternative way of thinking about the audio-visual correspondence learning task
2. Proposed the model to take advantage of AVC
3. Achieved state-of-the-art or comparable results on audio scene classification and ImageNet demonstrating the effectiveness of co-training
4. Demonstrated that the model is learning semantic information by providing sound-source localization heatmaps

# Evaluation of Paper

This paper builds upon prior work in audio-visual correspondence and begins with simple motivating factors. While tackling this, the authors yield some impressive results. Conceptually, what their model does is very simple yet the results seem to be quite powerful. The paper is also easy to read and follow making it quite accessible. The visuals aids clarify what the model is capable of; however, I do not think it was necessary to provide images corresponding to the learned audio features in lieu of providing the images of the actual log-spectrograms. I also didn’t fully grasp what the purpose of the Fully Normalized Score (NMI) was, or how to interpret this metric. To my knowledge, it seems to be a method to evaluate whether the learned features are separable and provide semantic meanings.

The evaluation methodology is standard in terms of providing classifications and then comparing said results to historical baselines. More interestingly, the sound-source localizations are more representative in showing what the model is learning and why the unsupervised approach is impressive.

It is worth noting that with the amount and dimensionality of the data, 16 GPUs need to be used to facilitate their experimentation. This is in part due to the massive volume of data associated with an unsupervised learning task. While having more data is often good, it sometimes imposes these restrictions that make it difficult to reproduce results. Furthermore, unsupervised data also has the potential of being very noisy and there’s no easy way to deal with that other than some selective pruning of the data. For example, as the authors mentioned, an audio montage would minimize the model’s ability to learn natural sounds. An interesting extension of this project would be to redesign the model for concurrency such that it can learn temporal patterns between a sequence of events and the corresponding sounds. I believe this may even be valuable for exploring temporal reasoning in video.