Sound Source Localization

Joel Huang, Justin Ng
University of Michigan
500 S State St, Ann Arbor, MI 48109
jhze@umich.edu, justinjn@umich.edu

1. Introduction

Videos can be used for cross-modal learning due to their synchronized audio and image streams. This allows for deep neural networks to be trained on unlabeled video data, which is highly accessible in modern day due to the popularization of camera-equipped smartphone devices and video sharing platforms such as YouTube and TikTok. Some applications of audio-visual learning include image and sound prediction and retrieval.

Our aim is to train a self-supervised network that can identify the location of the sound source in a video. For example, in a video of a man playing a guitar, the model should be able to identify the part of the image where the man's hand touches the guitar as the sound source. We achieve this by reimplementing the Audio-Visual Object Localization (AVOL) network described in the paper "Objects that Sound" by Relja Arandjelovi and Andrew Zisserman [1].

A network capable of such a task may possess high level semantic knowledge of objects and their sound properties, which may provide a good basis for transfer learning to other downstream tasks. This would demonstrate the potential of using self-supervised bimodal learning for high level vision tasks that capitalizes on the abundance of online video.

This report discusses our implementation of AVOL-net, including the dataset used, model architecture, and results. We also discuss modifications we have made to the network to reduce training time. We chose to implement this architecture as its contrastive learning methods and convolutional feature networks tie strongly to this class's syllabus and fit the scope of our project, as well as our own interest in multi-modal learning.

2. Related Work

SoundNet by Yusuf Aytar et al. [2] uses a deep convolutional network to learn audio feature representations, taking both waveform and image data as inputs. In their case, audio-visual learning is used for scene classification of natural sounds with high accuracy. However, this network does not attempt to perform any image segmentation or object localization.

AVOL-net itself is an extension of L3-net by Relja Arandjelovi and Andrew Zisserman [1], which learns to identify true and false pairs of images and audio spectrograms, where a true pair consists of an image frame and audio spectrogram from the same video clip. The training for AVOL-net is very similar, though modifications allow for a grid-based map to be generated enabling us to identify regions in the image with high correspondence to the audio input.

EZ-VSL by Shentong Mo et al. [3] shares the same objective as AVOL-net with the goal of localizing sounding objects within videos. The architecture similarly computes pixel-based similarities between audio and visual embeddings to generate localization maps, however their method proposes adding an additional object recognition network to generate a final localization prediction. This combined network showed greater accuracy than using audio-visual features alone, but at the cost of greater complexity.

Arda Senocak et al. [4] of "Learning to Localize Sound Source in Visual Scenes" propose using semi-supervised learning to improve the sound localization task. They claim training on unclassified data alone may lead to the pigeon superstition issue, where the model relates incorrect visual and audio features. For example, it may identify the train tracks as the source of a train sound, rather than a train itself. The limitation of this method is that it requires at least part of the dataset to be labeled and reduces the appeal of bimodal learning as a means to capitalize on the abundance of unlabeled online video data.

3. Method

We used the AudioSet-Instruments dataset, which consists of 10 second video clips of instruments being played from YouTube. We have chosen to limit the dataset to 25 classes for the scope of the implementation. The dataset has several notable limitations. Many of the videos are of poor quality as they do not show the source of audio in the image frame (for example, album art with music) or have artificially inserted audio data. In addition, the dataset is not balanced and contains an unequal distribution of classes, which may create a bias for certain classes when predicting correspondences. Furthermore, as the videos are publicly sourced, some of them have been removed or unlisted over time, reducing the reproducibility of work based on the dataset.

In total approximately ~37,000 video clips were used and partitioned into training, validation, and test sets in the ratio of 90:8:2.

The Audio-Visual Object Localization Net takes in pairs of image frames and 1 second audio log-spectrograms converted from videos. By converting the audio into a spectrogram, both the image frame and audio can be treated and processed as images. The architecture can be seen below in Figure 1.

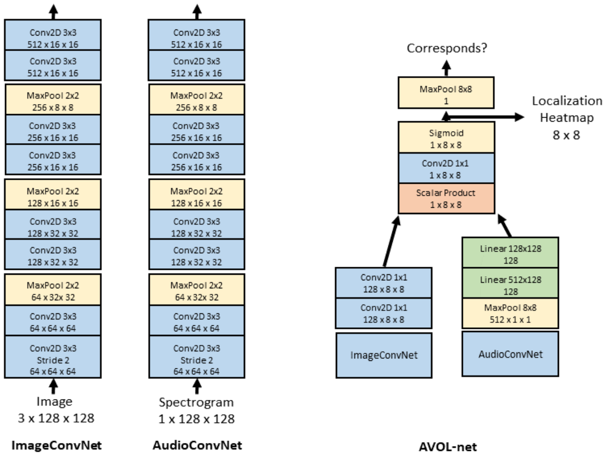

Figure 1: Audio Visual Object Localisation (AVOL-net) Architecture. Each convolution block consists of a 2D convolution with a 3x3 kernel, followed by batch normalization and a ReLU activation layer.

AVOL-net consists of two identical convolutional subnetworks labeled *ImageConvNet* and *AudioConvNet*. These networks generate feature encodings from the input images and audio spectrograms respectively, with their scalar product used to produce an 8x8 grid whose values represent the similarity between the audio and image at each region of the image frame. The highest value taken via the final max pooling layer to infer if the input image-audio pair is from the same video clip. The architecture we implemented is identical to the one described in the original paper. However, to speed up training we chose to downsize the images and spectrograms from 224x224 and 257x200 in the original paper to 128x128 for both.

Training is done via contrastive learning by feeding the model true and false image-spectrogram pairs. A true pair consists of an image frame and audio spectrogram from the same video clip, while a false pair takes a frame and spectrogram from different videos. This prediction is known as the audio-visual correspondence task. Binary Cross Entropy loss measures the difference between the inferred and actual correspondence values. We used the Adam optimizer to train our model weights. No labels are used for training.

Our code was mainly adapted from Kyuyeon Kim et al. [5]'s implementation. This includes the PyTorch dataset implementation, the dataset processing, and formats. Problem Sets 7 and 9 were also used as a reference for writing the training and test functions, and some of the visualization code. However, the code was heavily modified or added to, and was not copied as-is. A large amount of visualization code was also originally written, including all attached .py files used to visualize the dataset and model predictions. Original visualization code also includes the visualization for the model's weights and the receptive field of the outputs of the model.

4. Experiments

The model was trained for 200 epochs and took over 20 hours on a GTX 1060 Max-Q. An initial learning rate of 2.5e-5 was used, which was decreased by 6% every 16 epochs as suggested in [1]. A weight decay value of 1e-5 was used.

We also trained two modified versions of the original network to see if reducing the number of trainable parameters would affect performance. The first had half the number of output channels at each convolution and fully connected layer, which reduced the number of trainable parameters to a quarter of the original network. The second version had half the number of convolution layers as the original. These were trained alongside the original network on a subsample of the dataset containing 2000 image-spectrogram pairs for 100 epochs.

Our full implementation managed to achieve an accuracy of 70.9% on the validation set for the audio-visual correspondence task, meaning it was able to predict if the image and spectrogram were from the same video clip approximately 70% of the time. This is lower than the 83% accuracy achieved by the original paper. The discrepancy could be due to the size of the inputs, dataset, the learning rate, and classes used for training. Our training loss and accuracy histories for the main model can be seen below in Figure 2.

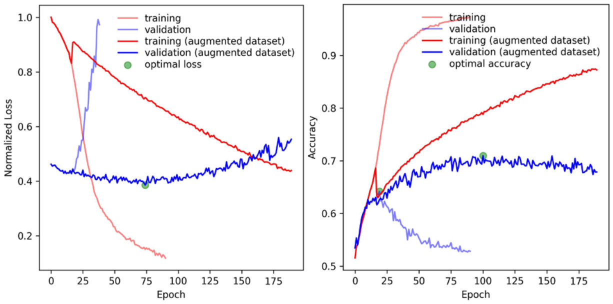

Figure 2: Loss and accuracy history over 200 training epochs. Green dots represent points at which overfitting begins

The model was first trained by taking 1 image-spectrogram pair from each video in the dataset. We achieved a validation accuracy of ~63% before the model began overfitting severely after 16 epochs as seen in Figure 2 above. Following this, we augmented our dataset by taking 9 image-spectrogram pairs from each video in the dataset and applying standard data augmentation techniques. This included adding noise, cropping, or flipping images. and resumed training. Validation accuracy appears to plateau after 100 epochs of training at about 70%.

For the modified architectures trained on 2000 clips, we found that the training and validation losses and accuracies were quite comparable between the full model, model with half the number of channels, and model with half the number of convolution layers as seen in Figure 3 below. This suggests reducing the number of training parameters does not drastically reduce the quality of the visual and audio feature embeddings generated by the convolutional subnetworks, at least for datasets of this small size.

With this small dataset, a maximum validation accuracy of ~0.6 was achieved, indicating that dataset size is a key determiner of generalizability and performance. It is highly likely that the lower performance of our full implementation compared to that of the implementation in [1] is due to the much smaller dataset we used for training.

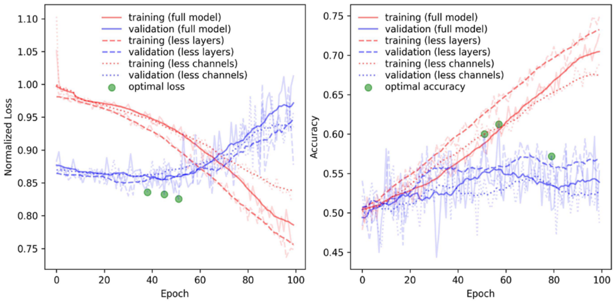

Figure 3: Loss and accuracy comparison of original and modified architectures over 100 epochs on 2000 clips. Moving average is overlaid on top of the actual loss and accuracy

For the main task of sound localization, we found that our model was able to generate fairly accurate predictions for the locations of sounding objects. Some examples can be seen below in Figure 4, with the predicted areas highlighted in green.

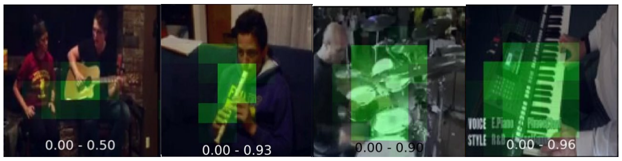

Figure 4: Sound location predictions for true image-audio pairs

The model also does not just select the locations of instruments or salient objects in the image frame, which compose the bulk of the AudioSet-Instruments dataset. Instead, it predicts areas with relation to the accompanying audio sample. In the example in Figure 5, an image of a band was paired with audio from a saxophone solo, then a drum solo. The model accurately highlights only the saxophone instrument in the first case, and the drum set in the second case.

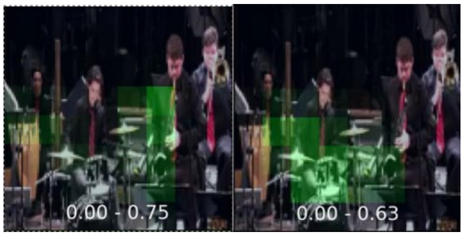

Figure 5: Predictions for a band of instruments with saxophone (left) and drum audio (right)

It can also be interesting to observe predictions where the source of sound is not in the frame of the image. Several videos in the test set consisted of music where the image was the album cover or music scores. In these cases, the model predicted seemingly arbitrary regions as having high audio-visual correlation. The fact that these predictions tend to be in the middle of the frames may suggest a bias towards the location of sounding objects from the training dataset.

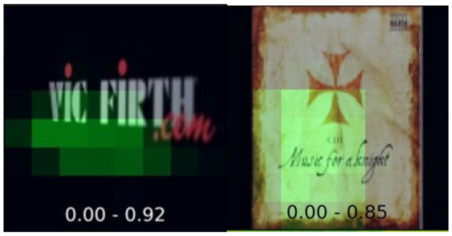

Figure 6: Predictions for true pairs where the source is not in the image frame

We did not test the accuracy of the sound localization task quantitatively. While it is possible to label a set of videos with marked regions at sounding objects, identifying the ground truths can be quite subjective. In some cases, the sound is clearly produced by a single source that is in frame. However, in many cases, the audio track contains a wide variety of sounds that come from sources both in and out of the frame. Some cases also contain no sound sources. Other complications include frames which include sound sources and sound amplifiers (speakers). The actual sound source can thus be rather subjective. As a result, a quantitative test of accuracy may be uninformative.

The generalizability of semantic representations learned by the network may be more reliably tested by measuring its performance on a more specific downstream transfer learning task.

5. Conclusion

To conclude, our reproduction of AVOL-net confirms the network can accurately localize sound sources, with an accuracy of ~70% for predicting audio-visual correspondence. It demonstrates an understanding of different sound sources and their visual appearances and can localize them in the image frames. The model's performance is also not significantly affected by small changes in the CNN's size or depth. This suggests that there is room for optimization in the model's specific implementation for specific audio-visual correspondence tasks or datasets.

To test the robustness of the architecture, future work could involve testing with another dataset, such as the Flickr sourced SoundNet so that more comprehensive comparisons can be made. Future work could also involve more concretely defined and measurable audio-visual correspondence tasks that would allow more quantitative performance metrics.

More generally, the model's success also demonstrates that bimodal learning allows CNNs to self-supervise by finding correlations between different modalities, and in the process learn meaningful semantic representations, thus capitalizing on the abundance of unlabeled online video.

6. References

[1] Relja Arandjelović, Andrew Zisserman. Objects that Sound. In ECCV, 2018

[2] Yusuf Aytar, Carl Vondrick, Antonio Torralba. SoundNet: Learning Sound Representations from Unlabeled Video. In NIPS, 2016.

[3] Shentong Mo, Pedro Morgado. Localizing Visual Sounds the Easy Way. 2022.

[4] Arda Senocak, Tae-Hyun Oh, et al. Learning to Localize Sound Sources in Visual Scenes: Analysis and Applications. In TPAMI, 2021.

[5] Kyuyeon Kim, Hyeongyeol Ryu, Yeonjae Kim. All that Sound, 2020. Github Repository. URL: https://github.com/kyuyeonpooh/objects-that-sound/blob/master/material/CS570_Final_Report_Team7.pdf