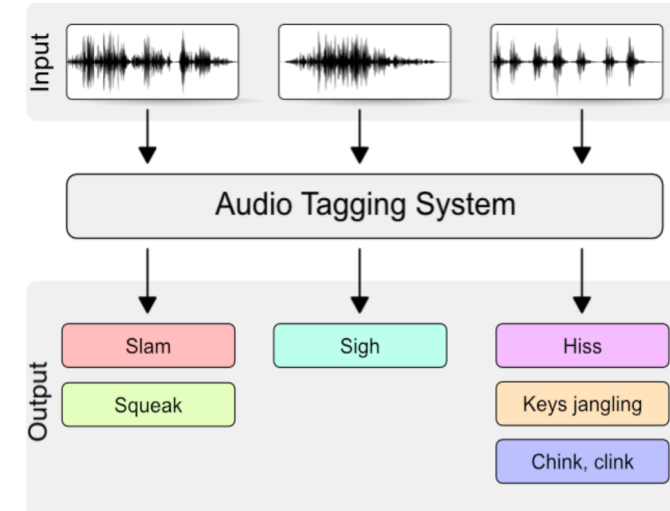


Group 38: Audio Clip Tagging

Boyang Zhang Jared Leitner Sam Thornton
{boz083, jjleitne, sjthornt}@ucsd.edu

Abstract

Automatic sound recognition has many potential applications including labeling of video/audio content and real-time sound detection. While image classification is a heavily researched topic, sound identification is less mature. In this study, we take advantage of the robust machine learning techniques developed for image classification and apply them on the sound recognition problem. Raw audio data provided by Kaggle is first converted to a spectrogram representation in order to apply these techniques. We constructed multiple convolutional neural networks (CNNs) in order to test the performance of different architectures. We achieve a LWLARP score and top-5 accuracy of 0.813 and 88.9% when predicting 80 sound classes using our top model.



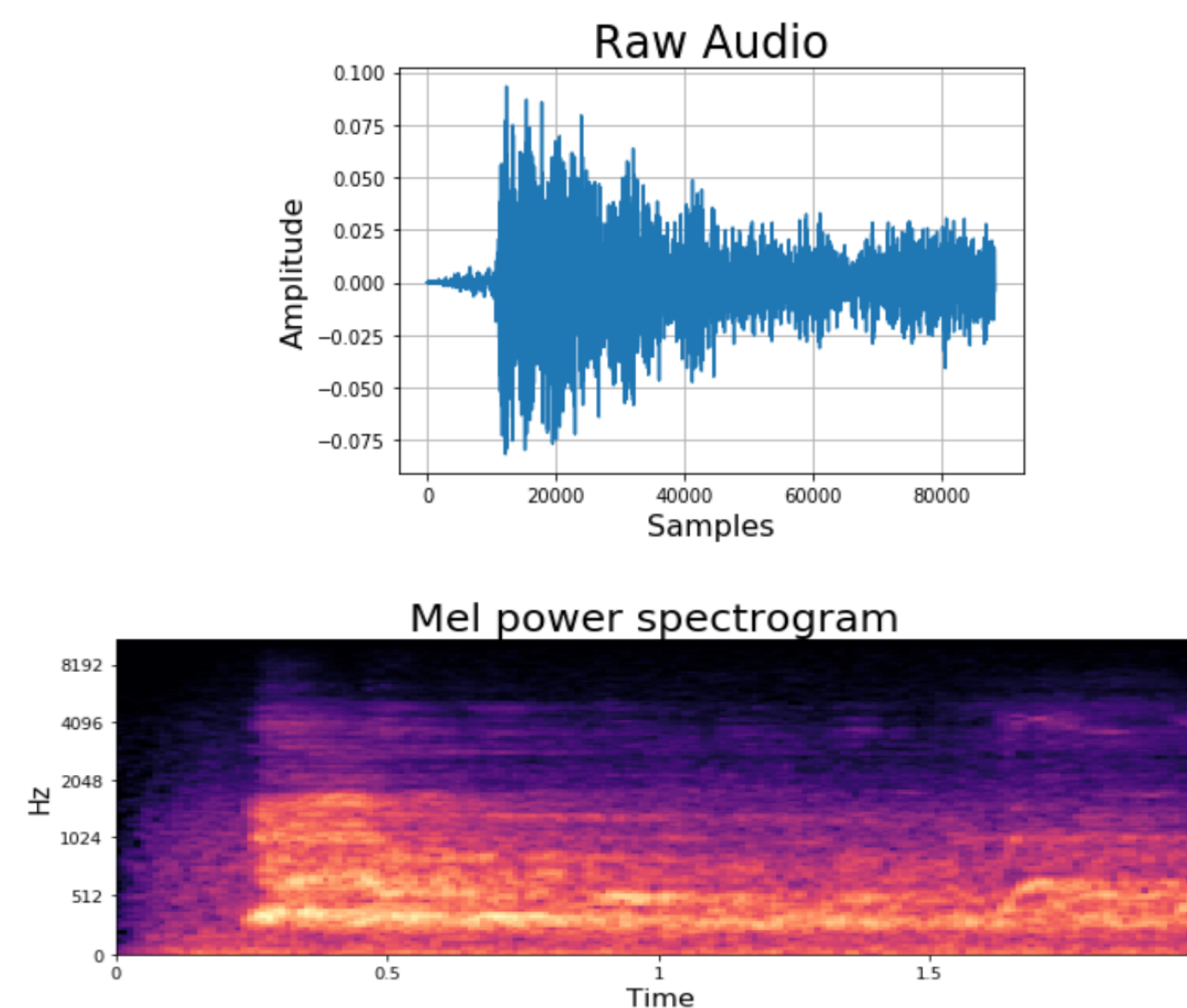
Data

The training dataset for this project was provided by Kaggle as part of an ongoing competition. There are 24,470 labeled audio clips which include 80 total classes. The sources for this data are as follows:

1. Freesound Dataset (FSD)
 - The FSD is a collection of crowdsourced annotations of 297,144 audio clips.
 - A subset (4,970) of these audio clips comprise the curated dataset.
2. Yahoo Flickr Creative Commons 100M dataset (YFCC)
 - The YFCC dataset contains 99,206,564 photos and 793,436 videos.
 - The soundtracks of a subset (19,800) of the videos comprise the noisy dataset.

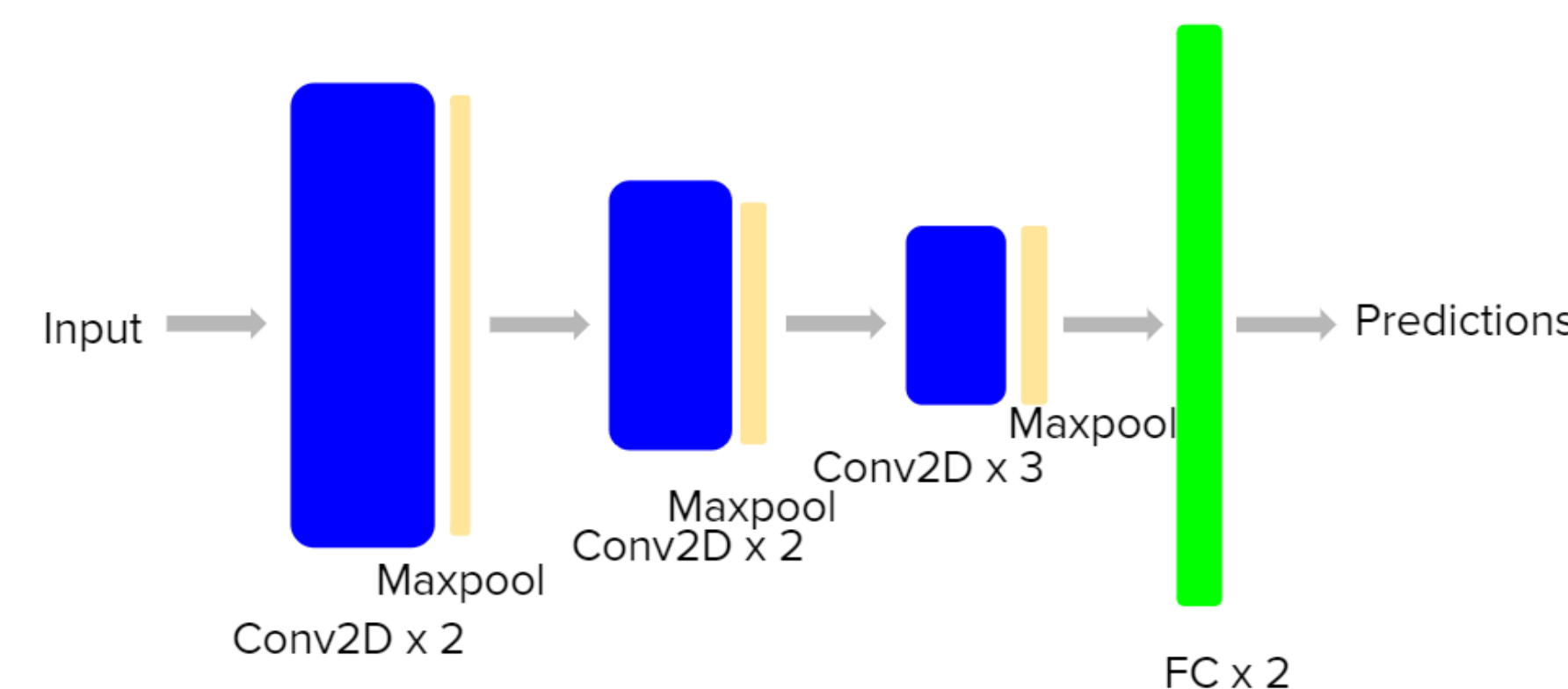
Features

The input audio data are sampled at 44.1kHz and range from 0.2-30s in length. We remove the silent sections of the clips and then trim/augment the clips to 2s. The raw audio waveform is then passed through filter banks to obtain the Mel log power spectrogram. The input to the model has shape 128 x 128, indicating 128 filter banks used and 128 time steps per clip.



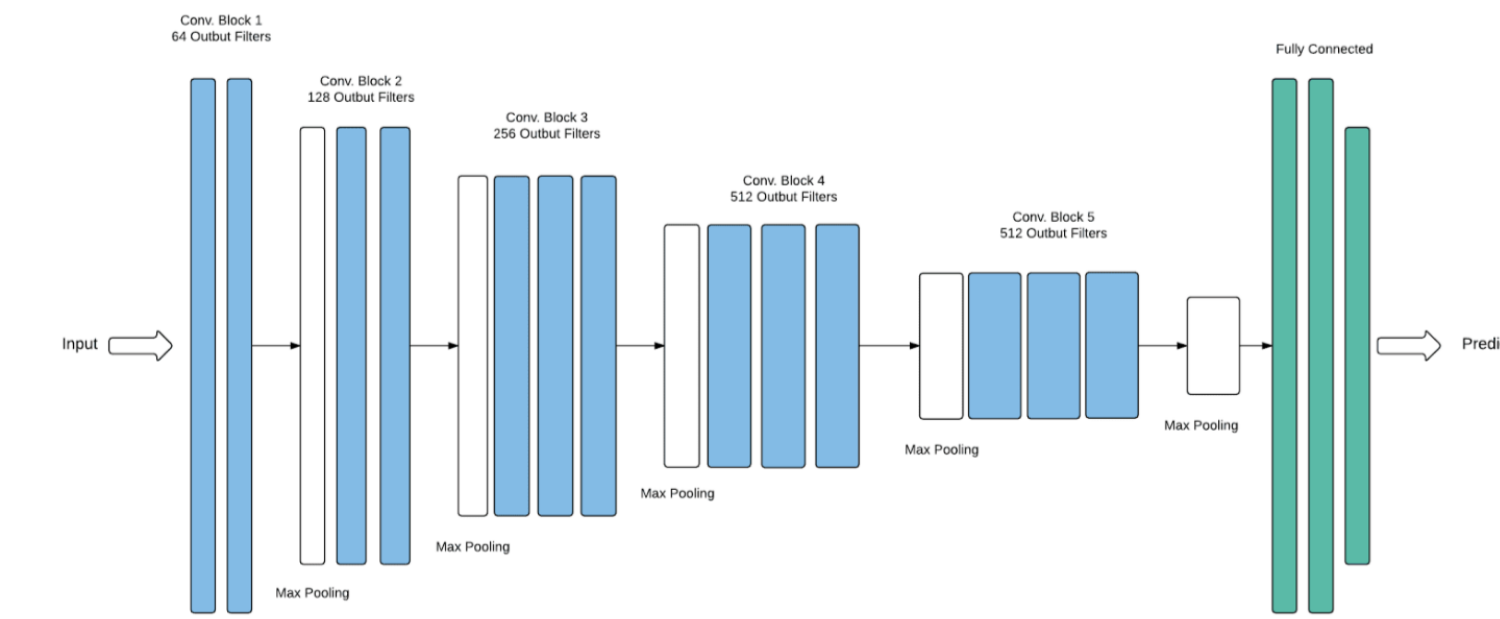
Model 1

We design and train several CNN based models, including architectures that have been successfully used for image classification applications. Our best performing model is our own deep CNN architecture.



Model 2

We also applied transfer learning to directly use frequency features learned from larger image datasets. Specifically, we used the VGG19 network pretrained on Imagenet. The last 2 convolution layers and 2 fully connected layers are retrained using our data in order to learn higher level features specific to audio data.

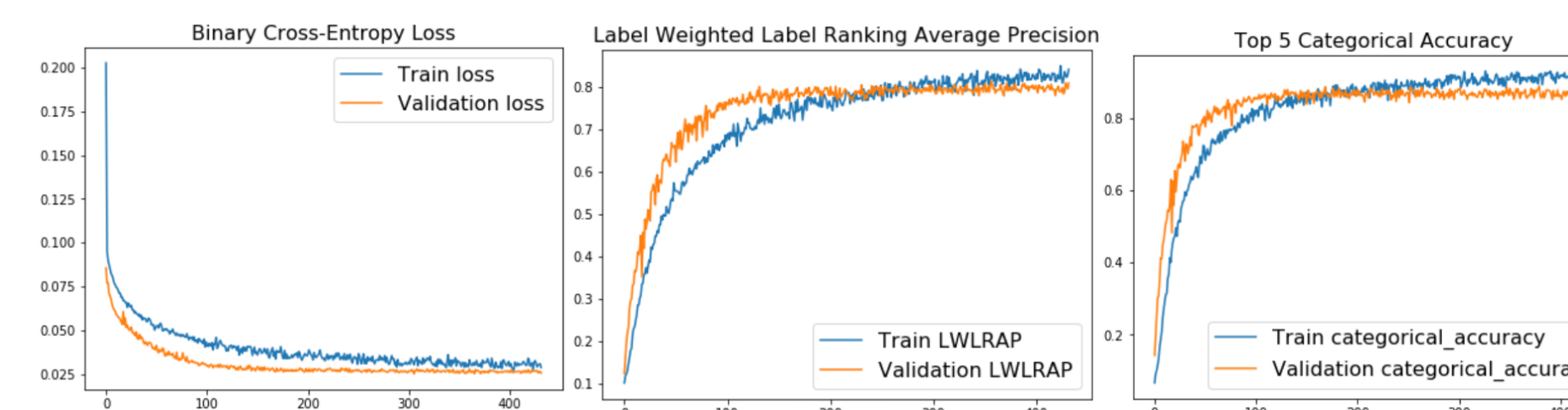


Results

The deep CNN model performs the best in terms of prediction accuracy. The VGG19 network pretrained on the Imagenet dataset is able to deliver similar performance in 100 epochs by retraining only the last block of convolution layers and fully connected layers. Both models perform worse on the noisy data. Thus, our models trained with curated data are not robust to noise.

Model	Epochs	Training LWLARP	Val. LWLARP	Training Top-5	Val. Top-5 Acc	Noisy Data LWLARP	Noisy Data Top-5
Deep CNN	432	0.850	0.813	93.2%	88.9%	0.219	28.3%
VGG19	100	0.801	0.797	91.8%	88.5%	0.215	27.7%

Loss/Accuracy curves for Deep CNN during training:



Discussion

When we began this project and were discussing how we could use the CNN's we learned about in class for audio classification, we determined it would be beneficial to transform the raw audio data to an image data format. This is because CNN models are primarily designed to work well for image classification. By transforming the audio to mel spectrograms, we had a suitable input for a CNN and could then test their performance. Since the test data for the Kaggle competition is not public, we split off part of the training data into a validation set and are using this for testing and generating results. Additionally, we decided to only use the curated dataset for training since this is a much cleaner dataset to work with. As a result, we had to stay with smaller CNN architectures that don't have too many parameters since our total training data before splitting is only 4,970. Overall, we are happy with the results we were able to obtain with this, but believe that they can be even further improved with more time.

Future Work

In order to develop a more powerful sound recognition system, larger networks could be implemented. This would require a larger amount of training data, meaning we would need to utilize the noisy dataset provided by YFCC. Techniques such as data augmentation could also be used in order to construct more training data. A future application of interest would be to use this system in conjunction with a computer vision system for automatically extracting information from video clips.

References

- [1] <https://www.kaggle.com/c/freesound-audio-tagging-2019>
- [2] Eduardo Fonseca, Jordi Pons, Xavier Favory, Frederic Font, Dmitry Bogdanov, Andres Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. "Freesound Datasets: A Platform for the Creation of Open Audio Datasets." In Proceedings of the International Conference on Music Information Retrieval, 2017.