Visualization of Deep Networks for Musical Instrument Recognition

Charis Cochran
Drexel University, USA
crc356@drexel.edu

Youngmoo Kim
Drexel University, USA
ykim@drexel.edu

ABSTRACT

We present a visualization tool for Convolutional Neural Networks focused on the task of instrument recognition. This tool allows you to visualize the network response layer by layer to a specific input sample as an array of animated activation plots corresponding to nodes, or filters, in the network, as seen in Figure 2.

1. INTRODUCTION

The recognition of instruments from audio, particularly in ensemble mixtures, remains a challenging and important problem fundamental to the field of music information retrieval. Early solutions to this problem focused heavily on designing task specific input features [2]. These features were very well defined, however, their performance does not come close to state-of-the-art deep learning approaches such as convolutional neural networks [3], multi-task approaches [4], and transfer learning [5]. However, the reported results of these black-box networks generally focus on overall performance across a dataset and ignore underlying instrument class performance disparities, which may overlook deeper issues with these approaches. Recently these types of deep learning approaches have become de facto standards for solving a wide variety of problems in the field of MIR. Still the underlying feature representations learned by these networks are not well understood in deep learning problems at large and even less in audio and spectrogram input specific cases. Our goal is to apply deep network and CNN analysis tools to the problem of predominant instrument recognition and create an analysis tool widely applicable and useful for MIR specific deep learning models.

2. CNN VISUALIZATION TOOL SPECIFICATIONS

Visualization of CNN networks has been a topic of interest especially in computer vision [6], but this task has not been extensively investigated or adapted to audio and music specific cases. Deconvolution and auralisation [6] of learned musical features has shown that demystifying CNN networks for MIR specific tasks can aid in providing a better understanding of network performance and classification, and in choosing network hyper-parameters to improve performance [6]. The tool we present seeks to understand MIR specific deep learning networks in general by providing animated visualizations that can be compared with input audio. This would provide a better understanding of prediction results and possible learned musical features the network might be latching onto in the case of correct and incorrect predictions.

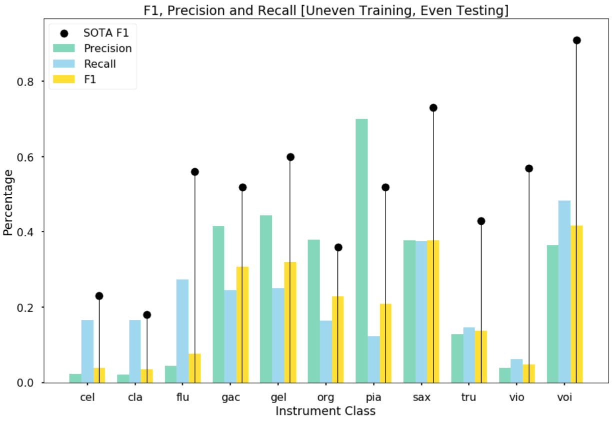

Figure 1. Model Performance As Compared to State of the Art – This figure shows how our model based on the architecture in [1] preform as compared to the state-of-the-art model in that paper. The model achieves similar results which makes it a good candidate for this feature visualization problem.

We implemented this visualization in Google Colab using python matplotlib and a pretrained Keras model. [1] The architecture for this model is based on a state-of-the-art model in [1], trained and tested using the benchmark IRMAS data set [7]. We were able to achieve comparable results using this architecture, as seen in Figure 1.

3. INITIAL RESULTS

In both the state-of-the-art system and our model we noticed similar class performance disparities that could not be explained simply by training or testing distributions, as is the case with the performance on the clarinet class, in particular. When comparing the performance on flute and

Attribution: Charis Cochran, Youngmoo Kim, "Visualization of Deep Networks for Musical Instrument Recognition", *Extended Abstracts for the Late-Breaking Demo Session of the 21st Int. Society for Music Information Retrieval Conf.*, Montréal, Canada, 2020.

[1] Example Colab Notebook: http://bit.ly/CNNVisualization

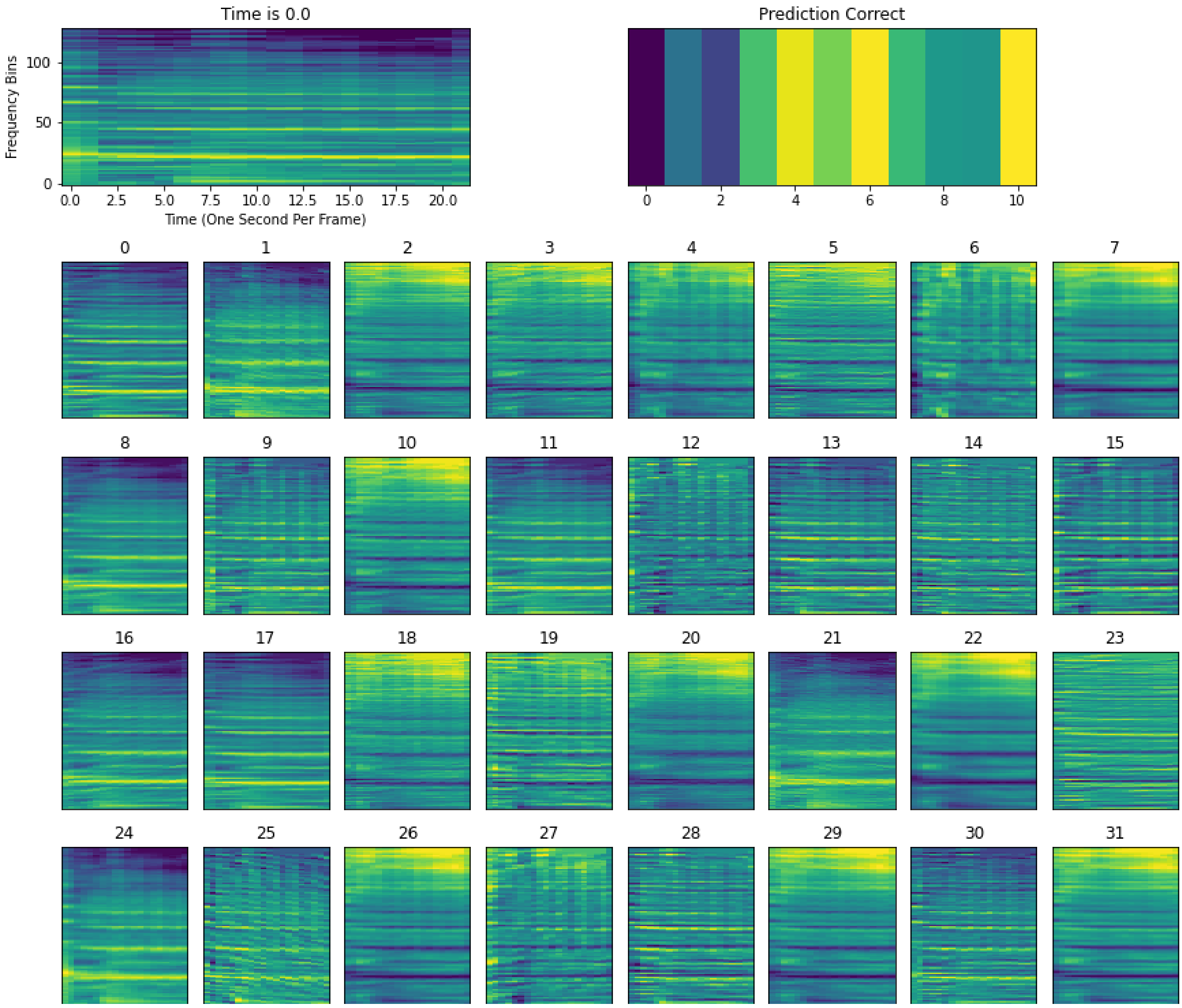

Figure 2. Single Frame of CNN Visualization Tool - This figure shows a single frame of the animated plots generated from the proposed visualization tool. Each animated visualization shows the original input Mel-spectrogram (top left), the final classification layer (top right), and an array of numbered nodes, or filter, activations within the CNN. This frame corresponds to the activations from the first convolutional layer.

clarinet samples, these classes have similar training and testing distributions in the IRMAS data set [7], but the model performs significantly better on flute samples than on clarinet. From the initial results we can see some possible areas of interest in the network with clear differences in activation bias on instrument class or prediction correctness.

4. FUTURE WORK

In the future, we are looking to extend the tool to other CNN visualization and deconvolution techniques to further improve our understanding of learned features in MIR specific deep networks as no such tool exists within current libraries for deep learning.

5. REFERENCES

[1] Y. Han, J. Kim, and K. Lee, “Deep convolutional neural networks for predominant and instrument recognition in polyphonic music,” in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

[2] M. S. Nagawade and V. R. Ratnaparkhe, “Musical instrument identification using mfcc,” *2nd IEEE International Conference On Recent Trends in Electronics Information Communication Technology*, 2017.

[3] D. Kim, T. T. S. andSooYoung Cho, G. Lee, and C.-B. Sohn, “A single predominant instrument recognition of polyphonicmusic using cnn-based timbre analysis,” *International Journal of Engineering Technology*, 2018.

[4] Y.-N. Hung, Y.-A. Chen, and Y.-H. Yang, “Multitask learning for frame-level instrument recognition,” in *ICASSP*, 2019.

[5] A. Molgora, “Musical instrumentsrecognition: A transferlearning approach,” Ph.D. dissertation, POLITECNICO DI MILANO, 2017.

[6] M. S. Keunwoo Choi, Gyorgy Fazekas, “Explaining deep convolutional neural networks on music classification,” *CoRR*, vol. abs/1607.02444, 2016.

[7] J. J. Bosch, J. Janer, F. Fuhrmann, and P. Herrera, “A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals,” in *13th International Society for Music Information Retrieval Conference*, 2012.