DEEPDREAM APPLIED TO AN INSTRUMENT RECOGNITION CNN

Charis Cochran
Drexel University
crc356@drexel.edu

Youngmoo Kim
Drexel University
ykim@drexel.edu

ABSTRACT

Explainability for the behavior of deep learning models has been a topic of increasing interest, especially in computer vision; however, it has not been as extensively investigated or adapted for audio and music. In this paper, we explore feature visualizations which give insight into learned models based on optimizing inputs to activate certain nodes. To do this we apply DeepDream [1], an algorithm used in the visual domain for exaggerating features which activate specific nodes. We used a model trained on the problem of predominant instrument recognition, which is based on the state-of-the-art (SOTA) model described in Han et.al. [2]. From initial results in optimizing test samples towards any target instrument using DeepDream, we find that the instrument models are highly sensitive to small imperceptible perturbations in the input spectrograms which can consistently influence the model to classify a sample towards the target with 100% accuracy. Additionally, when starting with noise we found that DeepDream creates consistent patterns across instrument classes which are visually distinguishable, but still indistinguishable when sonified. Both of these results indicate that learned instrument models are very fragile.

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 [3], however recent and state-of-the-art models rely on deep networks [4], [5], [6]. Recently 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. As a result of this, explainability has been a growing field of research with two main areas: visualization of learned features and attribution of input to classification [7], [1]. Visualization looks at how to identify features that network nodes may have learned usually by means of optimization of the input space. DeepDream is one method for feature visualization, originally used in the visual domain, to exaggerate features from some original input to the model, which activates the nodes [1]. Our goal is to apply the DeepDream algorithm to explore the learned feature space of our instrument recognition model.

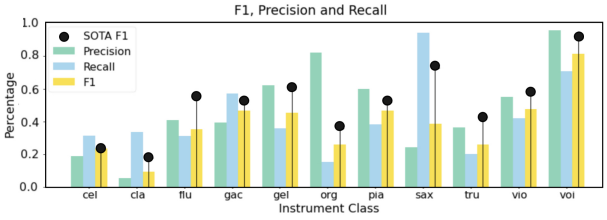

Figure 1. Network Performance as Compared to the State of the Art (SOTA)

2. DEEP DREAM EXPERIMENTAL SETUP

DeepDream is an algorithm that exaggerates possible features and patterns learned by deep learning algorithms [1]. Using a static feed forward target network, the output of a node or number of nodes in question is computed by simply taking the intermediate outputs of these nodes. The sum of these outputs is the total loss which is to be maximized by gradient ascent with respect to the input image. [1]

We performed two main classes of experiments; the first set of experiments was performed on a subset of the IRMAS data set [8] evenly sampled from the 11 different instrument classes. The network performance on these original samples is seen in Figure 1. Then we used our implementation of the DeepDream algorithm to create two classes of modified test samples such that in the first set each modified spectrgram maximizes classification confidence for the correct instrument label, and in the second set each spectrogram maximizes confidence for a randomly selected incorrect instrument class. After testing network performance on these modified sets, we also calculated the Mean Squared Error (MSE) between the original and modified spectrograms and between correctly and incorrectly optimized spectrograms. Additionally, we sonified various modified samples from both sample sets to compare to the original test samples. The second set of experiments was done using randomly generated noise samples

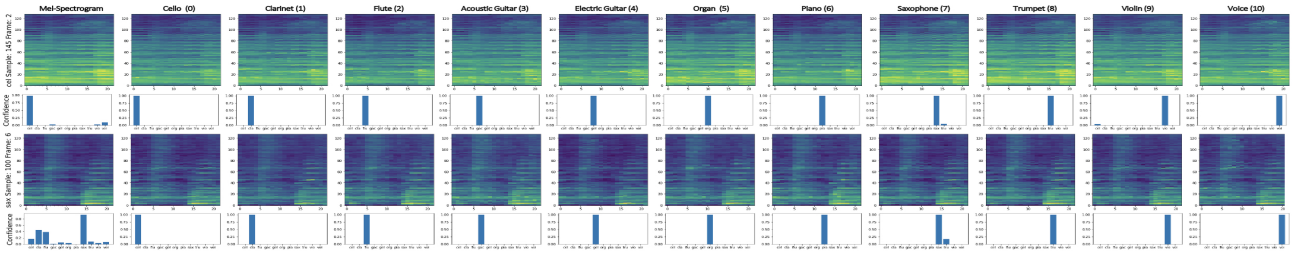

Figure 2. Results of DeepDream on Samples from IRMAS Test Set - Two original samples (cello - top ; saxophone - bottom) that have been modified using DeepDream to optimize for other instrument classes. Network output varies greatly across these modified inputs however the inputs themselves do not. When sonified, input variation is indistinguishable.

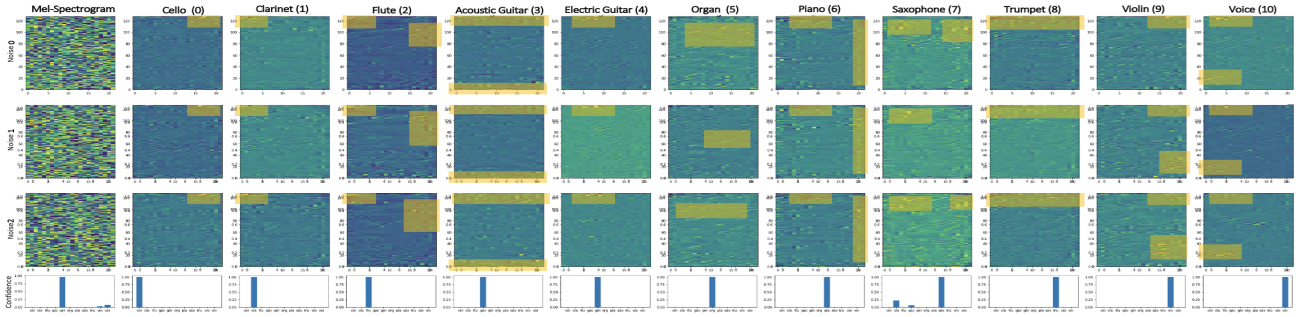

Figure 3. Result of DeepDream on Randomly Generated Noise - Random noise samples (left column) are "dreamed" to activate different instruments. Highlighted areas show consistency in patterns that arise in "dreamed" noise. Network output for the three samples in a column is averaged in the last row of that column.

with the same size as our input spectrograms. Using the DeepDream algorithm we processed these noise samples to maximally activate the confidence of each instrument class.

3. RESULTS

When the original test samples from the IRMAS data set were modified using DeepDream, we were able to bring the network performance for each class to positive and negative extremes since in each "dreamed" example the modified input was classified with above 90% confidence to the target class with 100% accuracy whether the modified input had been "dreamed" to the correct class or a randomly selected incorrect class. We also calculated the MSE between the original and modified test sets and found that the modified samples only differed from the original on average by about 2 decibels so that the spectrograms are mostly indistinguishable, as seen in Figure 2. Furthermore, upon listening to sonified examples and comparing them to the original test samples the changes implemented by DeepDream are still imperceptible. Additionally, we looked at the MSE between correct and incorrect modified examples and found greater difference between these optimized samples with each other than between the original spectrogram and any modified spectrogram for most instrument class pairs. This shows that the model is indeed learning to separate different instrument representations and the classes where the MSE may be lower could show where the instrument models are closer together. Looking at optimized randomly generated samples, similar patterns arise in each of the instrument classes after "dreaming" the noise. In many cases these patterns arise as small checkered or streaking areas in specific time frequency locations, as seen in Figure 3. Upon listening to the sonified versions of these examples, we found that the noise was still indistinguishable from the input noise by ear, however in the spectrogram space there are very clear differences between original and "dreamed" inputs. So, these time frequency patterns resulting from the various "dreamed" noise samples may give us some insight into the patterns that were added to the test samples to drastically change the network performance while keeping the input sample very close to the original.

4. CONCLUSIONS AND FUTURE WORK

We have shown that using the DeepDream algorithm we are able to manipulate the network output, while changing the input in a nearly imperceptible way. Based on the greater MSE between modified samples with other modified sample and the consistency of patterns generated from different noise samples, we can see that the network is learning a well separated latent representation of the examples. However, since these specific defining features have been shown to be imperceptible to human listeners it is highly likely that the model has latched on to statistical features, which have much more to do with the very limited training set and less to do with the underlying question of predominant instrument recognition. We hope that in the future we may be able to not only use these methods to highlight the network's issues but also to inform the training and possibly the structure of the network.

5. REFERENCES

[1] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, 2017, https://distill.pub/2017/feature-visualization.

[2] 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.

[3] 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.

[4] 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.

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

[6] A. Molgora, "Musical instrumentsrecognition: A transferlearning approach," Ph.D. dissertation, POLITECNICO DI MILANO, 2017.

[7] S. Becker, M. Ackermann, S. Lapuschkin, K.-R. Müller, and W. Samek, "Interpreting and explaining deep neural networks for classification of audio signals," 2019.

[8] 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.

[9] M. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," 13.

[10] A. Eronen and A. Klapuri, "Musical instrument recognition using cepstralcoefficients and temporal features," in *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, 2000.

[11] A. Eronen, "Musical instrument recognition using ica-based transform of features and discriminatively trained hmms," in *Seventh International Symposium on Signal Processing and Its Applications*, 2003.

[12] B. Toghiani-Rizi and M. Windmark, "Musical instrument recognition using theirdistinctive characteristicsin artificial neural networks," *CoRR*, 2016.

[13] C. Joder, S. Essid, and G. Richard, "Temporal integration for audio classification withapplication to musical instrument classification," *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, 2009.

[14] F. Fuhrmann and P. Herrera, "Polyphonic instrument recognition for exploring semantic similarities in music," in *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx-10), Graz, Austria , September 6-10, 2010*, 2010.

[15] J. D. Deng, C. Simmermacher, and S. Cranefield, "A study on feature analysis formusical instrument classification," in *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS*, 2008.

[16] M. S. Keunwoo Choi, Gy¨orgy Fazekas, "Explaining deep convolutional neural networks on music classification," *CoRR*, vol. abs/1607.02444, 2016.