Music and Instrument Classification using Deep Learning Technics

https://youtu.be/ueg3AUBI0fo

Lara Haidar-Ahmad {larahdr@stanford.edu}

CS230 Deep Learning, Stanford University

Abstract

This paper presents our implementation of a multi-class classifier that identifies instruments in music streams. Our model consists of a CNN which's input is an audio stream that we pre-process to extract the mel-spectogram, and outputs the dominance or non-dominance of pre-selected instruments. We focus our study on 3 instruments, and thus classify audio streams in one of 4 classes: "Piano", "Drums", "Flute" or "Other". We obtained a precision of 70%, a recall of 65%, and a F1-score of 64%. As future work, we aim to implement and compare the results of more deep learning model architectures such as RNN, RCNN, CRNN, in addition to adding more instruments.

Dataset and Features

Audioset [1] : Audio ontology with human labeled 10-second sound clips from YouTube videos.

Dataset consists in 2,400 audio samples per class, divided with the same distribution into 8,000 training samples, 800 validation samples and 800 test samples.

Data augmentation method: Adding white noise

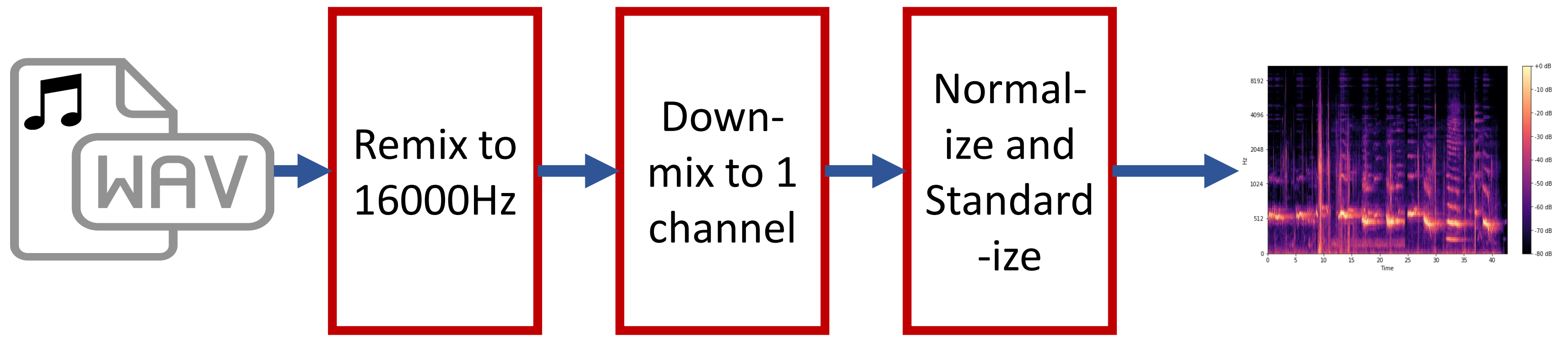

Figure 1. Raw Audio Pre-processing Pipeline

The model inputs are mel-spectrograms generated following features:
Sample rate = 16,000Hz ; n_fft = 400 ; window length = 400 ; hop length = 200; number of filter banks = 128.

Model and Results

The Model is a CNN for which the layers are presented in Figure 2:

Loss function : Cross-Entropy.
Optimizer : Adam.

Note: Cross-Entropy uses Softmax, so a Softmax layer was not added to the model.

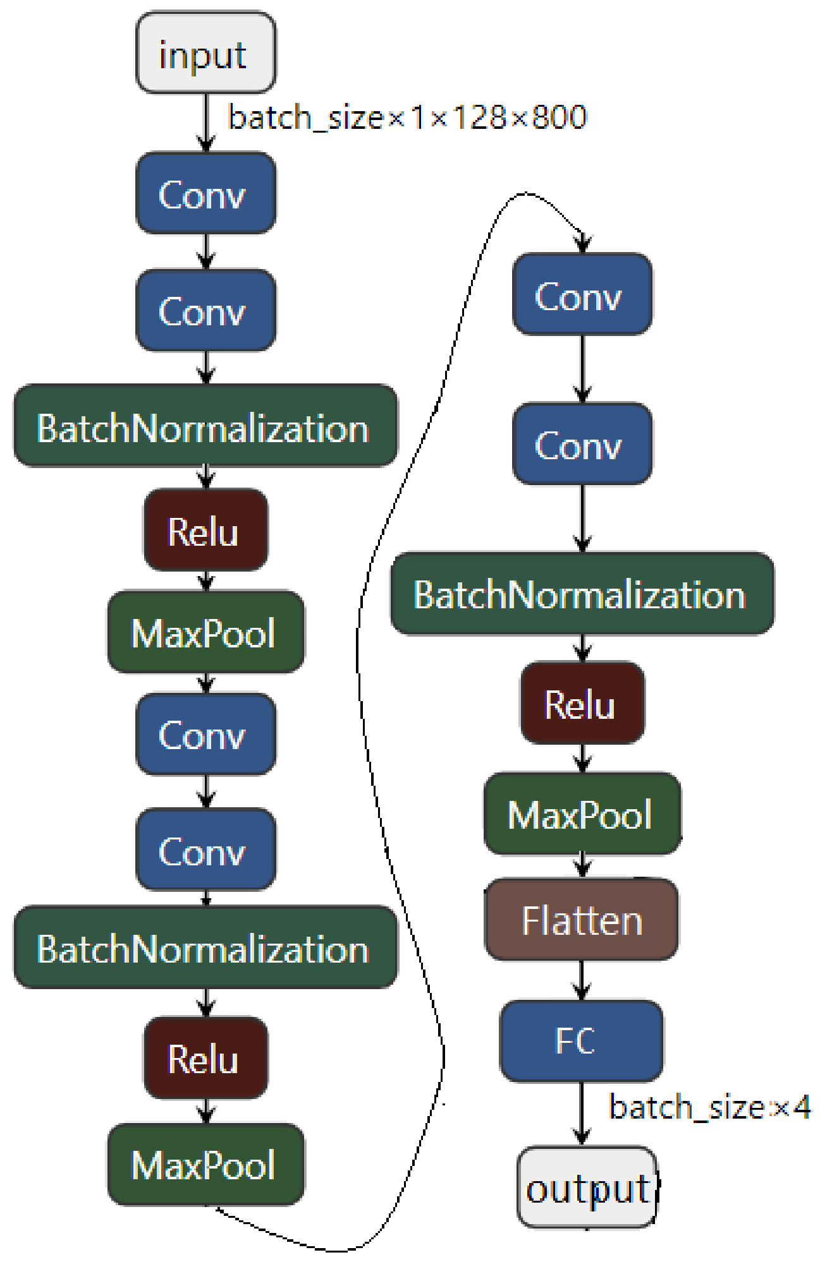

Figure 2. Model Architecture

actual \ predicted	Piano	Drums	Flute	Other
Piano	134	2	40	24
Drums	19	74	2	105
Flute	30	0	160	10
Other	41	5	6	148

Table 1. Confusion Matrix for the 4 Classes on the Test Dataset

	Piano	Drums	Flute	Other	Total
Precision (%)	59.82	91.36	76.82	51.57	69.92
Recall (%)	67.00	37.00	80.00	74.00	64.50
F1-score (%)	63.21	52.67	78.43	60.78	63.77

Table 2. Precision, Recall and F1-score for the 4 classes on the Test Dataset

	Other Sounds (%)	Unlabled instrument (%)	Mislabeled (%)	Sounds like predicted class (%)	Other (%)
Piano	10.00	16.67	-	43.33	30.0
Drums	23.33	36.67	10.00	-	30.00
Flute	13.33	30.00	13.33	16.67	26.67
Other	-	76.67	-	10.00	13.33

Table 3. Error Analysis per class on the Test Dataset

Conclusion

- Our results show an average precision of 70%, an average recall of 65% recall, and an average F1-score of 64%.
- We compare our results to a similar study in [1]. Their CNN attains a micro precision of 66% , a micro recall of 56%, and a micro f1-score of 54.1%.
- We obtain higher values for these metrics; however we should note that this study classifies audio streams into 11 classes which can be more challenging than dealing with 4 classes. On the other hand, our model was trained on more diverse data, that was extracted from YouTube, which can be more challenging to classify since it can contain any kind background noises and data closer to 'real world data'.

Future Work

- Clean data and label unlabeled instruments in streams.
- Support more instrument and split the current classes into subclasses (e.g. split piano into "electric keyboard" and "classical piano", etc.)
- Optimize and setup a hyper param tuning system to automatize the process.
- Support identifying more than one instrument in a stream.
- Try new architectures : RNN, RCNN, CRNN.

Links and References

Github Repository: https://github.com/lara-hdr/music-classifier.

[1] Google (2019), AudioSet. https://research.google.com/audioset.
[2] Toghiani-Rizi, B., & Windmark, M. (2017). Musical Instrument Recognition Using Their Distinctive Characteristics in Artificial Neural Networks. arXiv preprint arXiv:1705.04971.
[3] Han, Y., Kim, J., Lee, K., Han, Y., Kim, J., & Lee, K. (2017). Deep convolutional neural networks for predominant instrument recognition in polyphonic music. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 25(1), 208-221.

Additional related work and resources are referenced in the final report.