Self-Supervised Machine Listening

B V Nithish Addepalli, Jatin Khilnani, Ravi Choudhary, Shreyas Chandrakaladharan
Advisors — Vincent Lostanlen, Brian McFee

NYU CENTER FOR DATA SCIENCE

MARL

Overview

Evaluation of self-supervised representation learning in the time-frequency domain of machine listening for musical instrument recognition.

Motivation

Music does not have a lot of labelled data.

Annotating music demands a high level of domain specific expertise.

We want to investigate the effect of self-supervision in alleviating the need for human intervention in the design of convnets for machine listening.

Related Work

[1] uses Audio-Visual correspondence as cross-modal self-supervision to generate audio and visual embeddings.

[2] introduces a pretext task to identify the ordering of scrambled pieces of an image, much like solving a jigsaw puzzle.

[3] trains a classifier on videos to predict whether the video is playing forwards or backwards.

Problem Definition

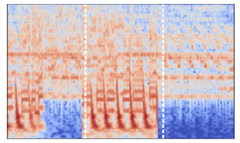
- Define two pretext tasks that try to predict the spatial structure of audio clip spectrograms. Model learns useful representations by solving these tasks.
- Downstream task: build a classifier that detects the presence of each instrument class in an audio clip.
- Establish baseline performance by training a randomly initialized classifier and compare it with classifiers that utilize the learned representations.

Data

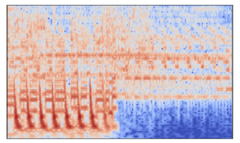
- Pretext tasks use the Free Music Archive (FMA)[4] dataset. It includes 106,574 tracks of 30 seconds each by some 16,341 artists. FMA is an unlabelled dataset.
- Downstream task is evaluated on OpenMIC-2018[5]. This dataset contains 20,000 tracks of 10s each. The samples are partially labeled for the presence or absence of 20 instrument classes.
- MP3 files from the datasets are converted into dB-scaled spectrograms using Constant-Q Transform and amplitude-to-dB conversion.

Pretext Tasks

Jigsaw: In this task, we split the spectrogram of an audio clip into three equal parts along the time axis. We shuffle these pieces randomly and train a classifier to predict whether the spectrogram is shuffled or not.

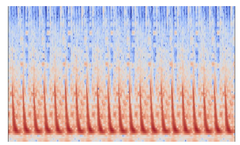


Original Spectrogram (Order: 1,2,3)

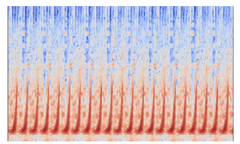


Jigsaw Spectrogram (Order: 2,1,3)

Time Reversal: In this task, we flip the spectrogram of an audio clip along the time axis. We then train a classifier to predict whether the spectrogram is flipped or not.

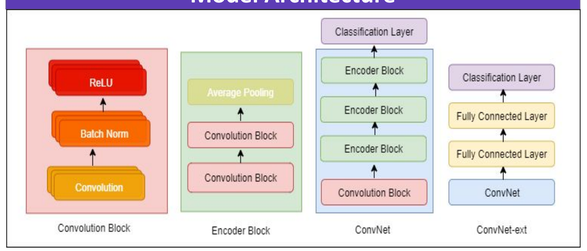


Original Spectrogram



Reversal Spectrogram

Model Architecture



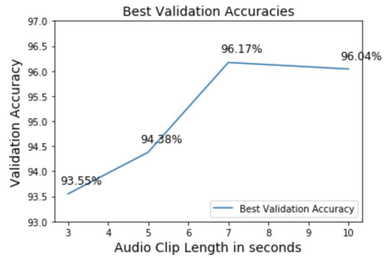
Pretext Results

Pretext Task	Architecture	Audio Clip Length (in secs)	Validation Accuracy (%)
Jigsaw	ConvNet	3 * 1	50.23
Jigsaw	ConvNet	3 * 3	51.27
Jigsaw	ConvNet-ext	3 * 1	63.63
Jigsaw	ConvNet-ext	3 * 3	**88.69**
Jigsaw	ResNet18	3 * 1	81.26
Jigsaw	ResNet18	3 * 3	82.15
Time Reversal	ConvNet	5	94.54
Time Reversal	ConvNet	10	**96.33**

*We tried ConvNet-ext and ResNet only for the Jigsaw task because ConvNet achieved low accuracy.

Pretext Experiments

- Varying audio clip length in pretext tasks — **Time Reversal**



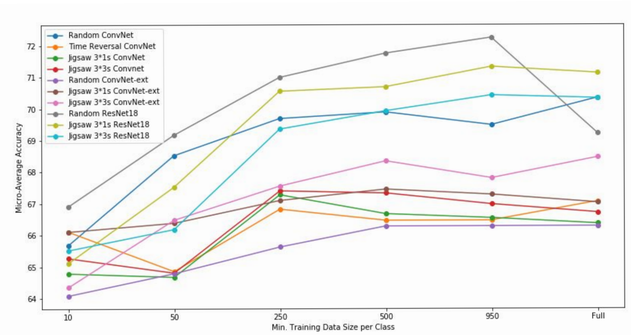
- Varying audio clip length in pretext tasks — **Jigsaw**

We experimented with total clip lengths of 10s and 3s for the jigsaw task. Task always performed better in terms of accuracy for total clip length of 10s. This is also inline with our intuition about human performance. If we are given three clips of just one second each, it is hard for us to arrange them in order.

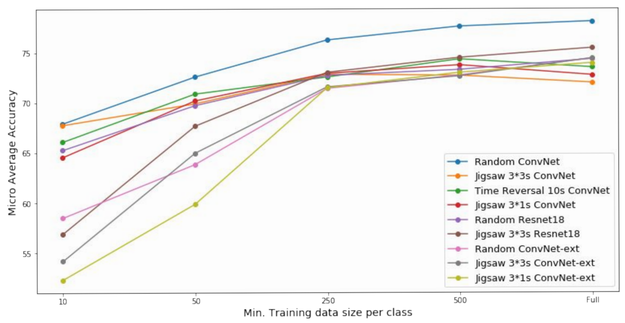
Downstream Experiments

- Varying amount of labelled data used for training downstream

Performance after freezing weights from pretext models



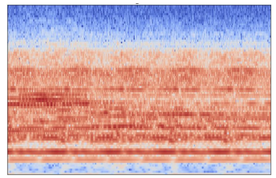
Performance after fine-tuning



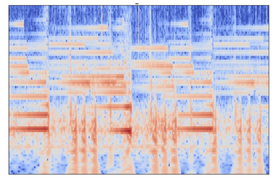
Error Analysis

Time Reversal: On analyzing the mistakes of our model, we found two interesting cases:

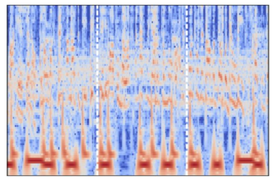
a) Model got confused by drone music
b) Model's false negative was a track already in reversed state



(a)



(b)



(c)

Jigsaw: We found that the musical pieces (c), where model was confused (prediction probability ~0.5) on whether the piece was jigsawed, were confusing even to a human listener.

Conclusion

Our pretext training does not reduce the need for labelled data.

It improves model performance (better than random) for ConvNet-ext and ResNet architectures after crossing the 250 labelled examples per class.

Prompts further research investigating different pretext tasks and alternative datasets to reduce necessity of labelled data.

Our pretext model is good at reasoning about sequence in a music clip. This is a new feature in music information retrieval.

Future Work

- Pretext invariant representation learning
- Contrastive Predictive Coding as a pretext task
- Alternate direction would be to test our models and strategies on a bioacoustics dataset like BirdVox

References

[1] R. Arandjelovi and A. Zisserman. *Objects that sound*. ECCV 2018.

[2] Mehdi Noroozi and Paolo Favaro. *Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles*. ECCV 2016.

[3] Donglai Wei et al. *Learning and Using the Arrow of Time*. CVPR 2018.

[4] Free Music Archive: https://www.freemusicarchive.org

[5] Humphrey, Durand and McFee. *OpenMIC-2018*. ISMIR 2018.