Automatic Music Transcription for Monophonic Piano Music via Image Recognition

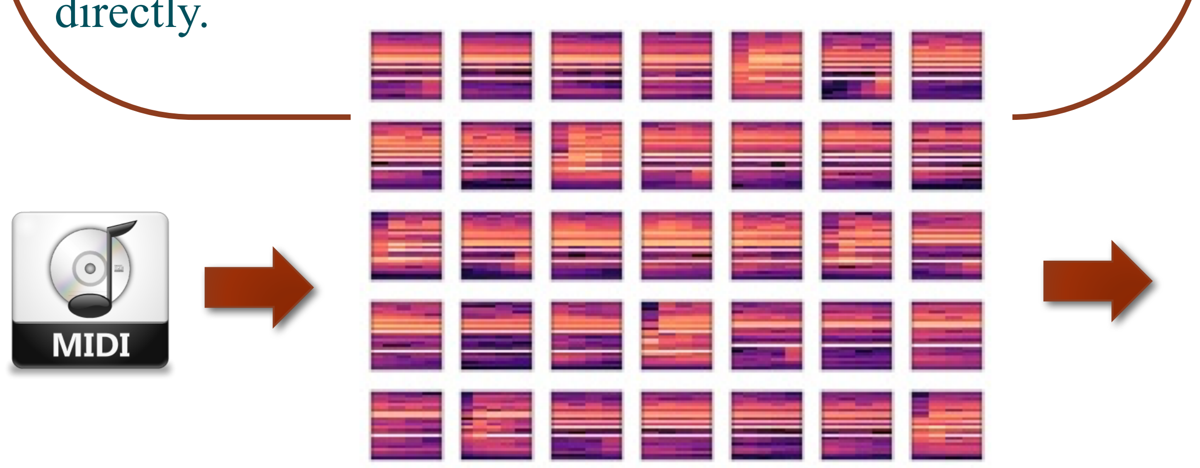
Chen Cen (ccen@stanford.edu) An Jiang (jianga@stanford.edu)

Introduction

In this project, we tried to tackle the problem of Automatic Music Transcription (AMT) using a new approach that transforms the music note detection problem into a image recognition problem using Convolutional Neural Networks (CNN). The monotonic piano soundtracks are processed into images of spectrums with constant-Q transform. Then we use an image recognition algorithm with Tensorflow to build and train a CNN that can differentiate and label the corresponding piano note from the spectrums. In the end, we have successfully developed a machine learning model that can automatically transcribe piano tracks into music scores.

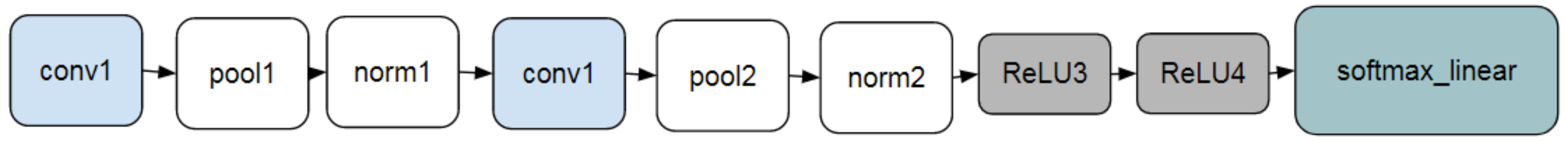
Data

We used MIDI Aligned Piano Sounds (MAPS) dataset which is a database for MIDI-annotated piano recordings. It contains real piano recordings and corresponding MIDI file and labels indicating the on and off time of each note. For the model training purpose, we only utilized the recordings of single isolated notes that are in the MIDI code range [21:108]. The tracks are mapped to frequency spectrums with constant-Q transform and we generate a 32x32 thumbnail of the spectrogram for each time step of the track. The images along with the ground truth of what note do they represent are combined into a cPickle data package which will be fed into the training model directly.

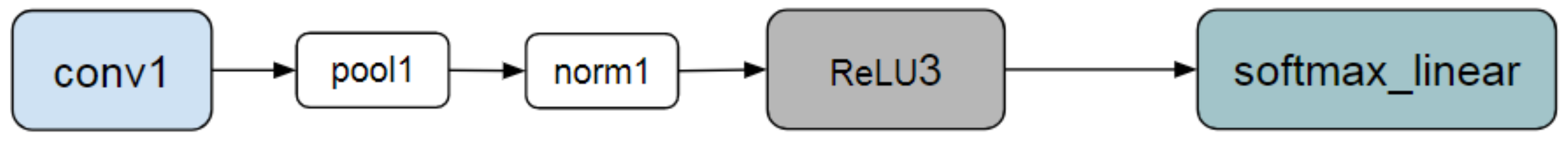


Model

Our initial training model applies from the paper from our reference. Which was originally used to solve the CIFAR-10 problem. It is a multi-layer architecture consisting of alternating convolutions and nonlinearities. Specifically, it consists of two CNN layers with two fully connected layer with ReLU activation function, followed by a 128-way softmax output. The blue box is the convolutional layer processes data only for its receptive field and reduce number of parameters to learn, The following pooling layer combines the outputs of neuron clusters at one layer into a single neuron in the next layer. The norm layer is a special layer introduced in the paper to aid generalizations and reduce over-fit. Initially we trained our model with this 2 layer architecture, but later we can achieve similar accuracy with just one layer and trains faster, thus we have the simplified CNN model.

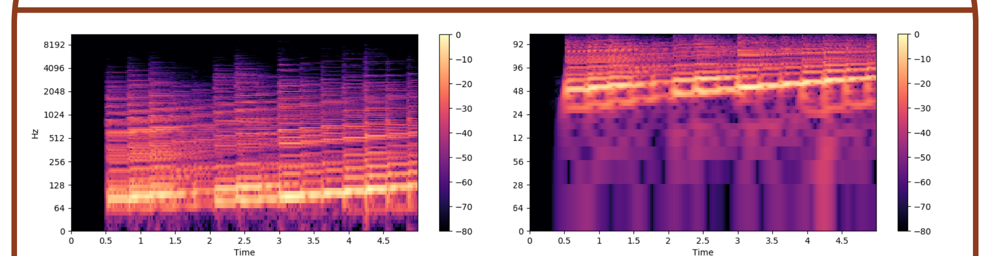


Original CNN

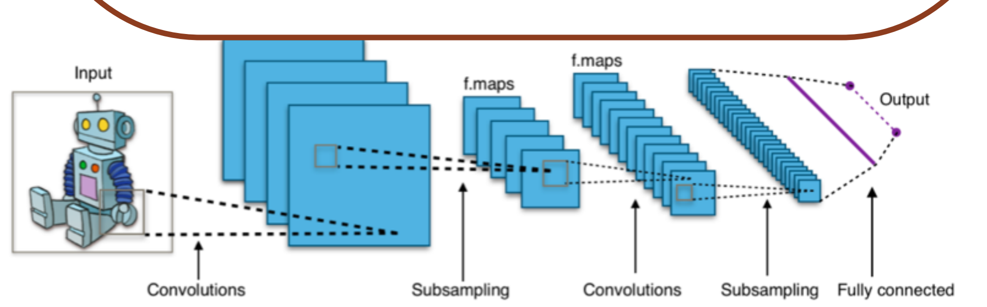


Simplified CNN

Sound Visualization



Spectrograms are the most standard way of visualizing sounds, which maps the music notes to frequency domain. However, most of the notes are located in low frequency range. Constant-Q transform on the right provides better time-frequency resolutions at these range compared to Short-Time Fourier Transform (STFT) on the left. From the plots above we can see the bottom quarter of the spectrogram is extended for Constant-Q and make the individual notes more extinguishable.



Discussion

Initially we trained our data set with exact same architecture with the original cuda-convet model. One key finding is that when we remove the original algorithm's image compressing process, the overall training accuracy increased by 5%. We believe the reason behind this is that the images processed by the original model are real world objects. Given the complexity of these images, compressing them won't loss too much information, images in one category will still be distinguishable from other categories. However our inputs are relatively simple images, they already look, therefore further compressing them would loss more information, making the trained model less accurate. Based on this observation, we later on simplified our model architecture to only one layer of CNN and one layer of local ReLU, the new model achieves similar results of the original model .

Result

We have conducted several experiments on the image processing procedure, learning parameters and CNN structure to improve the performance of the model. All the data below are obtained by training the neural network with the same dataset with more than 8000 images after 5000 steps while the data are split for training and testing with the ration of 4:1.

We can see that the original network used for image classification is not performing so badly though still not as good as the accuracy achieved by the original article (86%). Thus, we have tested several methods to improve the result: to test different transformation methods, we preprocessed the tracks with STFT; to make the cross entropy converges faster, we tried to modify the decay rate of the learning rate; to reduce the noise of the input, we skipped the random distortion of the images before feeding to the model; to enhance the quality of the dataset, we increased the resolution of the images. As a result, improving the quality of the images have increased the test accuracy the most, to an extend of 82.3%.

Future

There are still room for improvements on dataset processing and the CNN model to further increase the accuracy. The neural network can be modified with more convolutional layers and less pooling layers to solve the note classification problem better. Furthermore, we can explore the polyphonic music transcription problem using the same methodology.

Reference

[1] Convolutional neural networks: https://www.tensorflow.org/tutorials/deep_cnn
[2] MAPS database: Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle, V. Emiya, R. Badeau, B. David, IEEE Transactions on Audio, Speech and Language Processing, 2010.
[3] Ole Martin Bjorndalen. Mido - midi objects for python.
[4] Cen Chen and An Jiang. Cs229 project github page.
[5] Alex Krizhevsky. The cifar-10 dataset, 2009.
[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6):84–90, 2017.
[7] Bernd Krueger. Classical piano midi page.
[8] Brian McFee, Matt McVicar, Oriol Nieto, Stefan Balke, Carl Thome, Dawen Liang, Eric Battenberg, Josh Moore, Rachel Bittner, Ryuichi Yamamoto, Dan Ellis, Fabian-Robert Stoter, Douglas Repetto, Simon Waloschek, CJ Carr, Seth Kranzler, Keunwoo Choi, Petr Viktorin, Joao Felipe Santos, Adrian Holovaty, Waldir Pimenta, and Hojin Lee. librosa 0.5.0, February 2017.
[9] Eita Nakamura, Kazuyoshi Yoshii, and Shigeki Sagayama. Rhythm transcription of polyphonic piano music based on merged-output hmm for multiple voices. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(4):794–806, 2017.
[10] Graham E. Poliner. Classification-based music transcription, 2008.
[11] Daylin Troxel. 17th International Society for Music Information Retrieval Conference.

Stanford University