Automatic Music Transcription for Monophonic Piano Music via Image Recognition

Chen Cen (ccen@stanford.edu)

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 an image recognition problem using Convolutional Neural Networks (CNN). The monophonic 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.

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 spectrum for each time step of the track. The images along with the ground truth of what note do they represent are combined into a pickle data package which will be fed into the training model directly.

To test different transformation methods, we tested several methods to improve the 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 ratio 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%.

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 ratio 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%.

In the end, we have successfully developed a machine learning model that can automatically transcribe piano tracks into music scores.

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 an image recognition problem using Convolutional Neural Networks (CNN). The monophonic 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.

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 spectrum for each time step of the track. The images along with the ground truth of what note do they represent are combined into a pickle data package which will be fed into the training model directly.

To test different transformation methods, we tested several methods to improve the 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 ratio 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%.

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 ratio 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%.

In the end, we have successfully developed a machine learning model that can automatically transcribe piano tracks into music scores.

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 spectrum for each time step of the track. The images along with the ground truth of what note do they represent are combined into a pickle data package which will be fed into the training model directly.

To test different transformation methods, we tested several methods to improve the 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 ratio 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%.

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 ratio 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%.

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 ratio 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%.

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 ratio 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%.

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 ratio 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%.