Improved Optical Music Recognition

Justin Greet, Stanford University

Stanford University

Introduction

Optical music recognition (OMR) is the problem of converting a scanned image of sheet music into a symbolic representation like MusicXML. There are many practical applications for such a solution.

Previous approaches include the following, neither of which produces strong results:

- Conventional vision methods like edge detection that don't employ machine learning
- Unsupervised clustering around note similarity

Problem Statement

Given an image of a single measure taken from a piece of sheet music, the goal is to correctly identify and classify the pitch and duration of each note, and the duration of each rest. The order of notes and rests should also be captured. The general approach proposes regions for locations of notes then runs them through a convolutional neural network (CNN) for classification.

Evaluation at a high-level is straightforward. We compare the proposed notes and rests from our algorithm and compare them to those contained in the input image. If the input image can be perfectly reconstructed then the algorithm succeeded.

Dataset

We generate labelled images of measures of sheet music using the Lilypond music engraving program. Noise is introduced along the following dimensions:

- Image size
- Amount of whitespace around each note
- Amount/type of notes contained in each measure

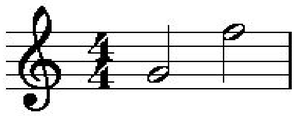
10,000 images of measures were created to train the network, which resulted in about 150 images for each of the 60 different kinds of notes. A further 1,000 images of "background" (containing no single note) were also introduced.

Methods

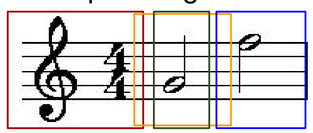
The method we employed to classify each measure of music can be broken down into 5 stages:

1. Gather the input image of a measure of sheet music.
2. Propose 1,000 random regions along the horizontal axis of the input image
3. Run each region through a CNN to identify it as one of 60 types of notes or rests, or as background.
4. Reject regions classified as background and those that have an intersection over union higher than the learned factor of .13 with another region classified as the same type of note with a higher score.
5. Order the remaining regions by the horizontal position of their bounding boxes and output the notes they contain.

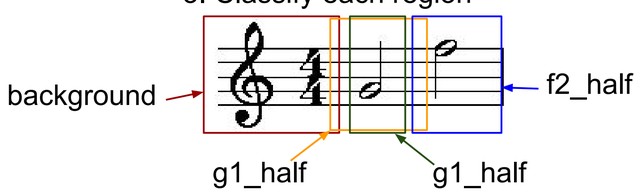
1. Gather input image



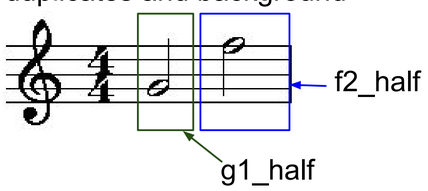
2. Propose regions



3. Classify each region



4. Reject duplicates and background

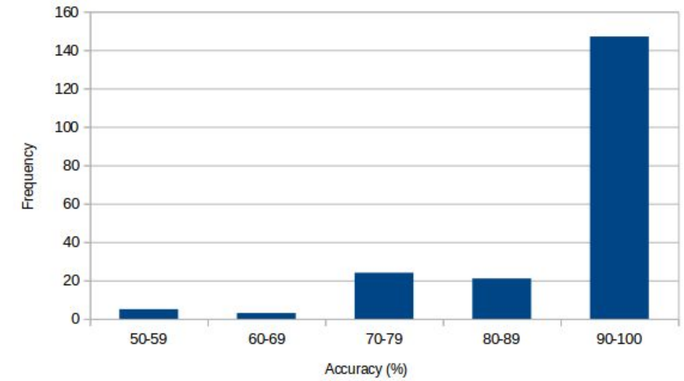


5. Output: g1_half, f2_half

Evaluation

Given an input measure *X* with notes and rests *R* = {r_1 , r _2 , ..., r_n } and an output measure *Y* with notes and rests *S* = {s_1, s_2, …, s_n} we calculate a percentage accuracy by comparing the pitch and duration of each note or rest in *R* to the note or rest at same index in *S*. The CNN achieved 98.9% accuracy on the note identification task after 5 epochs of training. For 200 test measure images, we achieved 100% accuracy on 147 of them.

Prediction accuracy across 200 test measures



Conclusion

The results of the experiment are very promising. Our CNN was close to perfect accuracy. Because we didn't achieve similar accuracy on measures overall, this suggests there was something wrong with the method used for region proposals. It would be worth experimenting with other methods for region proposal like selective search.

Our initial hypothesis of using a CNN for individual note identification paired with random region proposals for optical music recognition seems to be confirmed.

Though the network was trained and evaluated using generated data, one major area for future work is getting it to work on images of real-world sheet music. The question that gets raised is whether or not we can introduce the right kind of noise during data generation to be able to achieve this.