By: Marwan Harajli
Last Edit:02/04/2019
Subject: Capstone Project 2 Proposal.

Music Composition with Deep Learning

Problem Statement:
Given a data set of "music", we investigate learning what constitutes music (using Deep Learning tools), and generating new unheard music. More specifically, we limit the scope to solo piano pieces (no other instrument).

Project Importance:
This project is a great way of getting exposed to and familiar with various neural network architectures (LSTM, ConvNet, GAN net, etc.). Furthermore, the project presents an opportunity in dealing with a problem that does not clearly fall within either of supervised or unsupervised learning. Quantifying performance is consequently complex and requires some thought (which I think separates it from the typical data science project where accuracy and F1 scores reign supreme). Finally, the project (subjectively) has a certain "cool" factor that I as an amateur musician am attracted to: a computer is creating music!

Potential Client:
Potential clients include:

- Musicians who want to add fills to their to their tracks or want to jam over some computer generated backing track.
- Members of the film industry who want to add background music to their movie scenes.
- People just interested in creating music similar to some musician's style.

The Dataset:
The internet is full of piano midi files. An example can be found here, where we have piano compositions from the classical era from many artists.

Methodology:
Given a dataset of midi piano solos, we use music21 to parse this data and obtain a representation of each song. This representation constitutes a stream of events where each event has the following characteristics:

- An offset: the time it started relative to the start of the stream.
- A duration, the time the event lasts for.
- The type of event: is it a note, or a chord (chord is a combination of notes played together).
- The pitches of the note (or notes if it is a chord) played. This indicates what key/keys on the piano is/are being played.
- A velocity: how hard the piano key is hit to play this note.

- And many other characteristics that enrich a musical piece.

To keep it simple, we consider only a subset of the above characteristics. More specifically we ignore velocity (and lose some of the dynamics in the compositions), and assume that the duration of any event is exactly the next event's offset minus its very own offset (so the note/chord cannot last beyond the start of the next event nor can it end before the next event starts).

With that, we obtain a vocabulary of note/chord/duration possibilities. We are ready to learn what constitutes music. We propose the following approaches:

1. Train an LSTM (or stacked LSTM) to guess the (n+1)th event from a stream of n events. Here n is a parameter to be tuned, as is the specifics of the network architecture. Having a trained network, we then feed the network a stream of n events (where each of the n events is a word in our vocabulary) and let it generate the n+1th events. We then take events 2 to n+1 (assuming the index starts at 1) and let the model generate the n+2nd event. We keep doing this until we have generated a long enough sequence of events. We can then use music21 to convert our stream into a midi file and listen to the result. Quantifying performance here is tricky: on one hand we want to make sure that the resulting sounds are nice, but on the other hand we want to make sure they are not replicas of songs in the dataset. In this context, having a network with very low loss is not necessarily a good thing (it could mean we are overfitting and are just memorizing the songs). Dropout layers are investigated to increase generalizability of the model.

2. Train a Generative Adversarial Network. We consider each song as an image (see figure below). These images explain what notes are being played at each duration. We then create two networks (a generator and discriminator) where the generator's goal is to generate images that represent songs, and the discriminator goal is to differentiate real songs (ones from the dataset), from generated songs.

3. Try other mixed models (like a combination of conv nets and LSTMs, see reference 1).

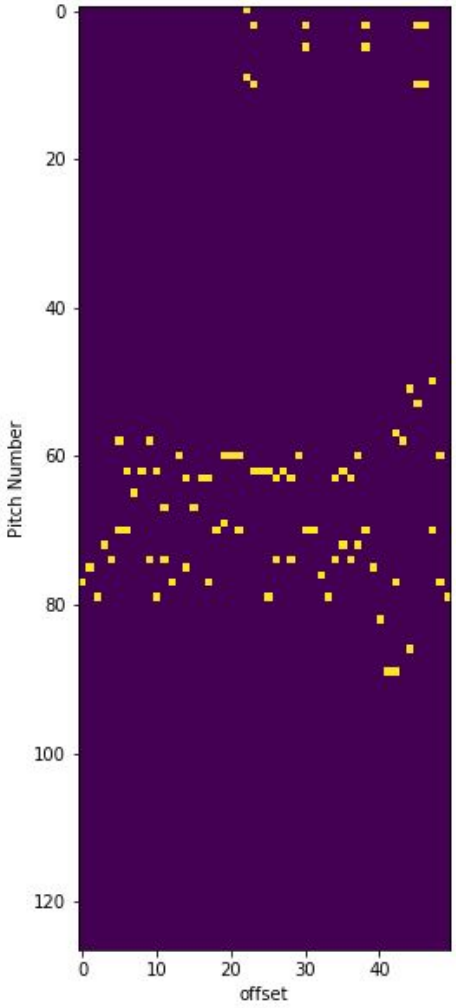

Figure 1: Image representation of music.

Deliverables:

A milestone report will be submitted that highlights the process of obtaining the vocabulary of a dataset of piano songs. A preliminary run of the LSTM model is also presented where temporal features are ignored (i.e. all pitches and chords have the same duration). Generated music snippets as well as model information (like loss values and architecture) are to be presented.

A final report is to be presented with more sophisticated models, and a discussion regarding the performance of each.

The files for this project can be found in the following GitHub repository:

https://github.com/harajlim/MusicGeneration

References:

1-http://www.hexahedria.com/2015/08/03/composing-music-with-recurrent-neural-networks/
2-https://medium.com/datadriveninvestor/music-generation-using-deep-learning-85010fb982e2
3-https://magenta.tensorflow.org/performance-rnn