Music Generation with LSTMs

Joyce Xu
Stanford University
jexu@stanford.edu

Sam Xu
Stanford University
samx@stanford.edu

Eric Tang
Stanford University
etang21@stanford.edu

November 18, 2018

Abstract

Music composition is an extremely creative task, and consequently very difficult for AI models to successfully perform. We hope to generate music using audio embeddings from deep neural networks, and in doing so, further our understanding of computational creativity. We also hope to visualize and explore relationships between music embeddings using tools such as Principal Component Analysis and t-sne visualization. In this milestone, we generate baseline compositions using Logistic Regression, a traditional classification algorithm, adapted for application to the composition.

I. Motivation

Music exhibits both short-term structure, such as the relation between successive chords and notes, and long-term structure, such as a song's overall key, tempo, and melody. Previous results in music generation have seen limited success, largely due to difficulties in capturing music's complex long-term structure. Recently, however, the use of novel network architectures, such as Long Short-Term Memory networks (LSTMs) and variational autoencoders (VAEs) have opened new possibilities for music generation [1].

Notably, these networks have also been used to generate vector embeddings of musical notes and chords. These embeddings can be used to train compositional models, feeding in richly structured data to generate successive notes. In our work, we begin with simple baseline Logistic Regression models, trained on piano notes and chords in various jazz and classical pieces. We plan to gradually incorporate these more complicated embeddings as we move towards neural network-based architectures, and begin using a more musically diverse and expanded corpus to improve our model's performance.

Relevant audio processing tools and datasets also improve our ability to tackle this problem. The music21 library, developed at MIT, offers a suite of tools for audio processing. The NSynth dataset contains over 300,000 annotated musical notes for over one thousand unique instruments, as well as high-quality music files [2].

II. Methods

i. Dataset

We begin our exploration with a 2 datasets: one of 26 Beethoven compositions, and the other of 18 Mozart compositions, both stored in the popular MIDI format. MIDI files contain information about each note's pitch, duration, and other data. This gives us a rich starting place for our exploration.

III. Preliminary Experiments

i. Baseline: Logistic Regression

i.1 Data Processing

We begin processing our data into a set of note sequences, each with a one-hot encoding. In our preliminary experiments, we reduce all chords to a single note, and convert all notes to a one-hot encoded vector representation. We use a sliding window approach to extract our data: we move across each song, capturing 11 notes per window, and use the first 10 notes as

our training data and the last note as the "class" we are trying to predict. Each training input is a vector of concatenated one-hot encodings, representing a sequence of notes. Each target output is a label representing the next note. With our 26 songs, this yields a total of 74,593 training examples, which split into an 85/15 train/dev split.

i.2 Model

Using these concatenated vector inputs and label outputs, we then trained a logistic regression classifier to predict subsequent notes given a sequence of notes. Our baseline treats the problem as a multiclass classification problem: there are 78 possible successor notes in our database, and our model emits the most probable successor to a given sequence.

i.3 Evaluation

Evaluating on our development model, the logistic regression model obtains 28.5% accuracy in predicting the a sequence's successive note on the Beethoven dataset, and an accuracy of 29.9% for the Mozart dataset. For comparison, the most common note (C4 and D5 respectively) comprises only 3.2% of notes in the Beethoven corpus and only 5.8% of notes in the Mozart dataset.

i.4 Error Analysis

We examined the logistic regression model for one common heuristic it might use to predict the next note: simply predicting the previous note in the sequence. It appears to only do this 1.8% of the time, which reassures us that our model is (hopefully) learning something vaguely valuable?

i.5 Generation

In order to turn our logistic regression model into a music generator, we feed the model a seed string of notes, then repeatedly ask our model to predict a subsequent note. With the subsequent note, we have a new a string of notes which can serve as our model's next generation seed.

IV. Next Steps

i. Incorporating Chords and Rhythm

First and foremost, we intend to extend our models to handle and predict more complicated musical structures, such as chords and rhythms. Right now, our current model only deals with one-hot encoding of single notes. We can handle chords by reframing our problem as a multilabel problem with inputs vectors marked as 1 for every note present, which we are already in the middle of transitioning to.

Similarly, because we are reducing all our training data into sequences of notes, we are currently throwing away all rhythmic information in the data. At generation time, we end up generating one note (or chord) at a time, each with the same length, which does not mimic any rhythm patterns in real music. We have two proposed solutions.

First, at each step, we could have the neural network simultaneously predict the next note and the duration of the next note, and train on both losses simultaneously (tuning the weight of the latter loss to an optimal performance). Another idea would be to discretize the space of note lengths, and at each smallest unit, predict either the next note, or predict a "hold" beat or "rest" beat. If we predict "hold," we would continue with the previous note, and if we predict "rest", there would be no output for that beat. While this idea would require more data preprocessing and significantly simplifying some input songs and rhythmic patterns, it may be much easier to learn, as we only add 2 more classes to the label space instead of a whole separate target to learn and optimize for.

ii. Improving the Model

Firstly, we intend to develop more powerful models to handle the complext structure in musical data. Using the sequence extraction methods described above, we plan to next feed

these one-hot encoded sequences into recurrent neural network architectures. We will likely begin with a vanilla Recurrent Neural Nework (RNN), then move to tuning a Long Short-Term Memory network (LSTM), and possibly try generative models such as Variational Autoencoders as well. Improving the performance of these networks will likely require a good deal of hyperparameter tuning and a more diverse training corpus.

Next, we aim to incorporate the music vector embeddings mentioned earlier, developed in previous work on audio processing and composition. We can then feed these embeddings into our models in place of the existing one-hot encodings, which should enhance our models with richer and more accessible information.

Third, we hope to expand our musical range beyond our existing work on the 26-work Beethoven corpus. We may augment our dataset with more classical works, or consider finding a larger corpus in the genres of Jazz or electronica.

Finally, we hope to evaluate our ultimate composition using human evaluators. Previous papers in music generation have used a small number of human evaluators to compare the performance of new models with previous models. This can be measured either by evaluators' personal preference, or how "human" they rate the music to be. We plan to recruit several human testers to compare the acoustic quality of our baseline, neural network model, and human standard.

V. Contributions

Eric handled data preprocessing to extract note sequences from MIDI files and set up the logistic regression baseline. Joyce and Sam have run an LSTM baseline to compose using our Mozart dataset, and are currently testing other instrumental and synthesized datasets.

References

[1] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. A hierarchical latent vector model for learning long-term structure in music. 2018.

[2] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. Neural audio synthesis of musical notes with wavenet autoencoders, 2017.