SCS Honors Undergraduate Thesis Prospectus

Eric Zhu
Advisors: Taylor Berg-Kirkpatrick (LTI) and Matt Gormley (MLD)

September 5, 2017

1 Abstract

We study the generation of polyphonic music using generative music language modeling. Currently, in both natural language and music tasks, generated samples lack coherence. For the musical domain, this yields unpleasant samples to human listeners. In particular, we aim to add inductive bias by defining or learning new data representations to capture long-term structure in music. We expect that this will result in perplexity gains as well as better-sounding generated samples.

2 Problem and significance

2.1 RNN language models

Music language modeling attempts to learn a probability distribution over sequences of notes. In particular, given a sequence of a symbolic representation of notes $x = (x_1, x_2, ...x_n)$, we want to learn its relative likelihood

$$p(x) = p(x_1, x_2, ...x_n) = p(x_1)p(x_2|x_1)...p(x_n|x_1, ..., x_{n-1})$$

One popular approach to unsupervised generative language modeling used for natural language and symbolic music is a recurrent neural network (RNN) [1, 3]. RNNs can represent arbitrarily-sized structured inputs in a fixed-size vector. At each step, we have a state vector $s_i = R(s_{i-1}, x_i)$ and an output vector $y_i = O(s_i)$, representing the state of the RNN after observing inputs $x_{1:i}$, where R and O are model-defined functions. Using this, the conditional probability of observing event e after the sequence $x_{1:i}$ can be defined as

$$p(e = j|x_{1:i}) = softmax(y_i\mathbf{W} + \mathbf{b})[j]$$

without making any Markov assumptions.

To generate original, plausible music using an RNNLM, we sample from the softmax distribution $p(e|x_{1:t})$ at each timestep t, conditioned on the previously-generated outputs [4].

2.2 Feature learning

The state-of-the-art RNN "performances" have very expressive local structure and characteristics, but fail to exhibit any coherence or long-term structure – it sounds like “noodling”. Because music is innately highly structured (repetition, metrical hierarchy, non-hierarchical long-timescale structure) [5], being able to model this more effectively should result in large perplexity gains.

Our problem can be seen through the lens of representation learning [6]. Simple piano roll representations of music that are fed into a deep LSTM *entangle* different explanatory factors of variation behind the data, requiring that the model perform many interleaved tasks at once. The RNN/LSTM formulation lends itself to focusing on modeling the variance in local structure.

3 Research contribution

We aim to add inductive bias or a prior to the language model to better disentangle the factors of variation in a piece, allowing it perform better on the LM task and, more interestingly, generate coherent samples. The approach of this thesis work will be to iterate quickly on the following models, further pursuing those that show the most promise.

3.1 Hierarchy

One approach that has seen success in natural language is to model different, predetermined levels of sequential correlation with a hierarchical RNN or LSTM [7, 8]. This involves “stacked” LSTMs that take as input the embedding outputs from lower-level LSTMs. For language, LSTMs are defined over words, then sentences, and finally paragraphs. In music, hierarchy is mainly metrical. We will begin by creating an analogous model for music over measures. If this model sees success, we will shift to working without predefined boundaries [9].

3.2 Repetition

Repetition and regularity are fundamental features in musical form, regardless of genre or culture. We can encode repetition, as well as variation of components, in a self-similarity matrix [10]. This matrix can be used as a prior or as a way to define where a recurrent model should attend to. We also expect that generated samples will have somewhat structured self-similarity matrices.

3.3 Separation into elements

Music can be analyzed by its constituent parts or elements: pitch, volume, duration, etc. [11]. We can explicitly (or implicitly) separate the representation of these dimensions to generate each “parameter” independently of the others, or dependent on others having previously been generated.

3.4 Latent, global representation

We can train a variational autoencoder [12] to learn a latent distribution from which samples can be generated. See `ericzhu.org/research` for a summary of my independent study work from Spring 2017. We can consider furthering this work with hierarchical [13] or disentangled [14] latent representations .

4 Proposed research plan

For the first half of the fall semester, I plan on deciding upon a good initial data representation for polyphonic piano roll music, then building a preprocessing pipeline to convert MIDI files to that representation. We will begin with a dataset of 254 MIDI files from the Ragtime Press MIDI Archive. This will also involve extracting other data such as measure locations. With this data, we can build baseline models (LSTM, LSTM with attention). For the second half of the fall semester, I will work on building the simplest possible models of the directions menitoned above. Depending on their success, I will iterate on the most successful models. The spring semester will involve building more complex models based off of insights we gain from the fall semester models' performances. I also plan on writing and potentially publishing papers on the work.

References

[1] Elman, Jeffrey L. "Finding structure in time." Cognitive science 14.2 (1990): 179-211.

[2] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[3] Mikolov, Tomáš, et al. "Extensions of recurrent neural network language model." Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, 2011.

[4] Eck, Douglas, and Jasmin Lapalme. "Learning musical structure directly from sequences of music." University of Montreal, Department of Computer Science, CP 6128 (2008).

[5] Lerdahl, Fred, and Ray Jackendoff. A generative theory of tonal music. MIT press, 1985.

[6] Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." IEEE transactions on pattern analysis and machine intelligence 35.8 (2013): 1798-1828.

[7] El Hihi, Salah, and Yoshua Bengio. "Hierarchical recurrent neural networks for long-term dependencies." Advances in neural information processing systems. 1996.

[8] Li, Jiwei, Minh-Thang Luong, and Dan Jurafsky. "A hierarchical neural autoencoder for paragraphs and documents." arXiv preprint arXiv:1506.01057 (2015).

[9] Chung, Junyoung, Sungjin Ahn, and Yoshua Bengio. "Hierarchical multiscale recurrent neural networks." arXiv preprint arXiv:1609.01704 (2016).

[10] Foote, Jonathan. "Visualizing music and audio using self-similarity." Proceedings of the seventh ACM international conference on Multimedia (Part 1). ACM, 1999.

[11] Burton, Russell. "The elements of music: What are they, and who cares?." Music: Educating for life. ASME XXth National Conference Proceedings. Australian Society for Music Education, 2015.

[12] Kingma, D.P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).

[13] Goyal, Prasoon, et al. "Nonparametric Variational Auto-encoders for Hierarchical Representation Learning." arXiv preprint arXiv:1703.07027 (2017).

[14] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, “Controllable text generation,” arXiv preprint arXiv:1703.00955, 2017.