Lead Sheet Generation with Musically Interdependent Networks

Benjamin Genchel[1], Alexander Lerch[2]

Georgia Tech Center for Music Technology
[1] bgenchel3@gatech.edu
[2] alexander.lerch@gatech.edu

In the proposed work, we introduce a novel architecture for generating lead sheets using Recurrent Neural Networks (RNN) towards the goal of modeling music as a set of explicitly represented, interdependent components. The architecture at present consists of two Long Short Term Memory (LSTM) RNNs; one models note pitch, and the other note duration. Each network is conditioned on the other's output, and both are individually conditioned on the harmony playing at the time of a note's generation. We compare this system with the performance of two baseline systems, each of which also consists of two separate LSTM RNNs: The first has neither individual chord conditioning nor inter-conditioning between the models, modeling the pitch and duration sequences in isolation, while the second only provides the individual chord conditioning.

RNNs have become a staple in sequence modeling, with many variants applied to the task of music generation[8]. In the domain of symbolic music generation, there have been many attempts to apply RNNs, however, these typically assume a strict hierarchical structure[1] or tonality and rhythm, by using time slices as input for example. Here, we explicitly model the interplay between harmony, pitch, and duration in a flattened structure, giving equal weight to each component.

We trained each system on a small initial dataset of 60 lead sheets from the Charlie Parker Omnibook, provided by researchers from IRCAMs Creative Dynamics of Improvised Interaction project[3]. Each system was trained for 100 epochs using standard supervised n-gram prediction [4], using the ADAM optimizer [7] with Negative Log Likelihood Loss (NLL).

Our initial results have been positive, with our system achieving lower losses than the baseline systems. In an informal listening test conducted among peers, we found that music produced by our system was more natural sounding and pleasant in comparison to the baseline systems. We also performed some statistical analyses comparing the output distributions of each model with the training data, and surprisingly found that the output of our model was less similar to the training data in each measure than the unconditioned baseline and chord conditioned baseline. This will need further investigation. A demo of each system generating a melody over chords from Charlie Parker's Yardbird Suite is available online [5]

In the future, we plan to train our system on a larger data set consisting of a jazz lead sheets extracted from the Wikifonia project. We additionally hope to introduce modern techniques and innovations into the training process such as scheduled sampling [6]. Further, our eventual hope is to develop a network that generates the harmony co-conditioned on the other two networks as well to form a system capable of fully generating a lead sheet from scratch.

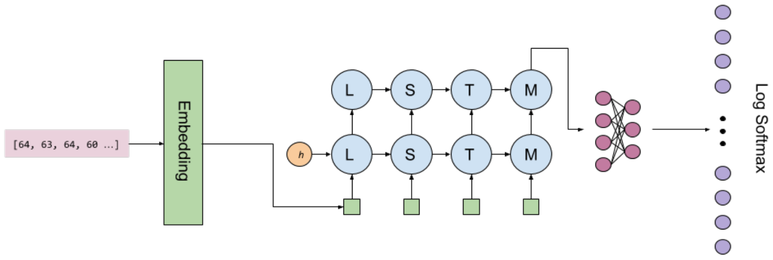

Fig. 1: This figure displays a diagram of the model architecture used for both the pitch and duration network in the unconditioned baseline system. To further clarify, the input in this diagram shows MIDI pitch numbers, but this is simply changed to duration tags when this same architecture is used for the duration network. Pitch and duration symbols are first dictionary embedded before being fed to the LSTM layer. The 'h' here refers to the initial hidden and cell states for the LSTM layer.

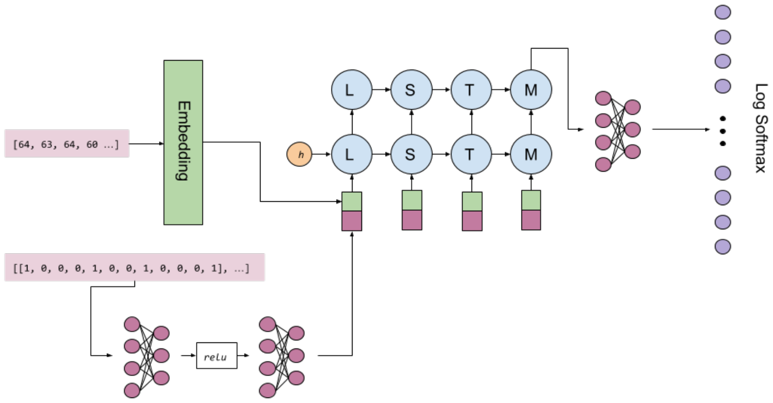

Fig. 2: This figure displays a diagram of the model used for both the pitch and duration network in the chord conditioned baseline system. As above, the

input in this diagram shows MIDI pitch numbers, but this is simply changed to duration tags when this same architecture is used for the duration network. Chords are given to the network as a length 12 binary vector of pitch classes in which the included pitches are labeled with a '1', then encoded and appended to the pitch or duration symbol before being fed to the LSTM layer.

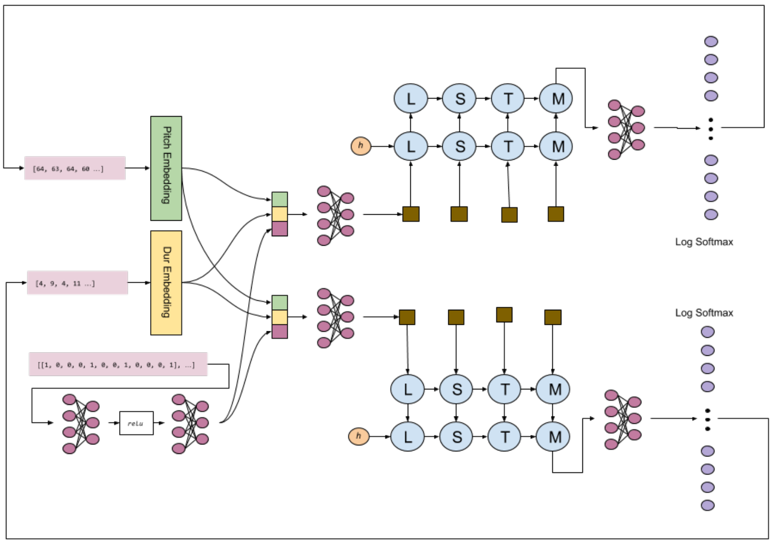

Fig. 3: This figure displays a diagram of the full system described above, showing two mirrored, interdependent LSTM networks, each concatenating the other's output to its own input along with the current chord to condition their generations. The concatenation of pitch, duration and chord sequences is encoded by each network before being fed to their respective LSTM layers.

References

[1] Hang Chu, Raquel Urtasun, and Sanja Fidler. Song from PI: A musically plausible network for pop music generation. arXiv preprint arXiv:1611.03477, 2016

[2] Elliot Waite, Douglas Eck, Adam Roberts, and Dan Abolafia. Project Magenta. `https://magenta.tensorflow.org/`

[3] DYCI2: Creative Dynamics of Improvised Interaction. `http://repmus.ircam.fr/dyci2/home`

[4] Goodfellow, Ian, et al. Deep learning. Vol. 1. Cambridge: MIT press, 2016.

[5] CSMC Demo.
`https://soundcloud.com/bgenchel/sets/csmc-demo/s-j12e0`

[6] Bengio, Samy, et al. "Scheduled sampling for sequence prediction with recurrent neural networks." Advances in Neural Information Processing Systems. 2015.

[7] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

[8] Briot, Jean-Pierre, Gatan Hadjeres, and Franois Pachet. "Deep learning techniques for music generation-a survey." arXiv preprint arXiv:1709.01620 (2017).