Learning to Create Piano Performances

Sageev Oore[*]
Google Brain
osageev@gmail.com

Ian Simon
Google Brain
iansimon@google.com

Sander Dieleman
Google DeepMind
sedeilem@google.com

Douglas Eck
Google Brain
deck@google.com

Abstract

Nearly all previous work on music generation has effectively focused on creating musical scores. In contrast, we learn to create *piano performances*: besides predicting the notes to be played, we also predict expressive variations in the timing and musical dynamics (loudness). We have provided samples generated by our system to a set of professional musicians and composers, and the feedback was positive. Overall, the comments indicate that our system is generating music that, while lacking high-level structure, does indeed sound very much like human performance, and is closely reminiscent of the classical piano repertoire.

1 Introduction

While music is in many ways considered a quintessentially human activity, interest in automatic music generation has existed for centuries— at least since the Musikalisches Würfelspiel of the 1700's (Nierhaus [2009], Hedges [1978]). Nearly all previous work on music generation, both algorithmic and machine-learning-based, has focused on creating pieces that are, effectively, *scores* (Nierhaus [2009], Briot et al. [2017], Boulanger-Lewandowski et al. [2012], Lattner et al. [2017], McDonald [2017]). That is, the generated output is a sequence of notes turned on and off, with timings exactly aligned to a standard metrical grid (e.g. 8th notes, 16th notes, triplets).

In this work, we learn to create *piano performances*: besides predicting the notes to be played, we also predict expressive variations in the timing and musical dynamics (loudness). The recent work of Malik and Ek [2017] includes expressive dynamics but does not have expressive timing, and is conditioned on scores.

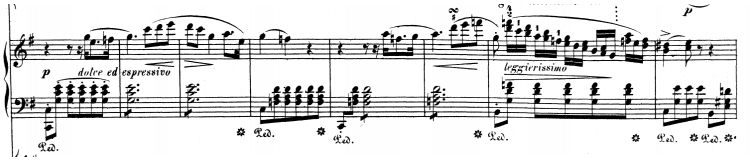

Figure 1: Excerpt from the score of Chopin's Piano Concerto No. 1.

To give a sense of the impact of the difference between a score and a performance, we provide two examples:

- a *direct rendering* of the above score (Figure 1): https://clyp.it/jhdkghso, and

[*]also Saint Mary's University and Dalhousie University

- an *expressive performance* of the same score: https://clyp.it/x24hp1pq.

We therefore propose generating directly in the domain of musical performance.

2 Data

To avoid coarse (score-level) quantization, we require a data set that consists entirely of high-resolution, high-quality examples. We use the e-Piano Competition dataset (eCompetition [2017]), which contains MIDI captures of roughly 1400 performances by skilled pianists. The pianists were playing a Disklavier, which is a real piano that also has internal sensors recording all MIDI events corresponding to the performer's actions. Thus, every example in our dataset has high-quality expressive timing and dynamics.

3 LSTM-based Recurrent Neural Network (RNN)

We use an LSTM-based RNN model, consisting of three layers of 512 cells each. We represent the MIDI data as a sequence of events drawn from a vocabulary allowing the following types of events: `Note-On`, `Note-Off`, `Set-Velocity` and `Time-Shift`. The `Time-Shift` allows the model to move forward in time by multiples of 8 ms up to 1 second. For example, to play the first couple of notes of the score in Figure 1, assuming a tempo of 60 bpm and a velocity of 80 for the first note, the sequence might begin as follows:

```
Set-Velocity (80), Note-On (C2), Note-On (C3), Time-Shift 304ms, Note-Off
(C2), Note-Off (C3), Time-Shift 200ms, ...
```

The neural network operates on a one-hot encoding over this event vocabulary.

4 Results & Informal Feedback

Generated examples are available at `https://clyp.it/user/3mdslat4`. Our system generated all MIDI events: timing and duration of notes as well as note velocities. We then used publicly-available piano samples to synthesize audio from the resulting MIDI file.

Evaluation of generative models is known to be very challenging (Theis et al. [2016]). We gave a small set of clips to professional musicians (pianists, teachers) and composers (TV, film, classical, jazz, professors) for informal comments, mentioning that the clips were generated by an automated system, and simply asking for any reactions they might have. Comments included:

"Fantastic!!!! This [. . .] absolutely blows the stuff I've heard online out of the solar system. . . . The melodic sense is still foggy . . . but it's staggering that it makes nice pauses with some arcing chord progressions [. . .] its not far from actually coming up with a worthwhile melody . . ."

". . .sounds like you fed a bunch of Mozart, Beethoven, Schubert, and Chopin into the system "

". . .a very drunken Chopin, messing around a bit with psychedelics . . .something a Russian composer would've written under the influence of Impressionism. . . . "

". . . Not liking the somewhat messy run but [. . .] it seems wrong in a human way. "

". . . I wanted to hear more, as in longer sections, to hear how the piece would unfold from there."

5 Discussion & Conclusion

The comments above are representative of a much larger set, and from that collective set, there was general agreement about a few characteristics, including the following:

1. the performances sounded very human;
2. the samples were very reminiscent of the classical piano repertoire;
3. most individual samples were not internally stylistically consistent.

Interestingly, multiple composers, who generally *were* impressed with the system, were *not* particularly impressed with its melodies: this suggests considering that the generation of good melodies, in musical context, involves more subtlety and sophistication than one might initially suspect, at least without musical background.

When musicians perform a score, they do so *expressively*: they incorporate variations in the timing and the dynamics. Some of these variations are very small, yet altogether, this expressiveness plays a huge role in our perceptual experience and understanding of music. This perceptual experience, in turn, shapes our perception of the underlying creativity. We have demonstrated an example of how we can use machine learning to work effectively in the space of expressive performance.

Acknowledgments

The authors would like to thank the members of the Magenta team.

References

N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

J.-P. Briot, G. Hadjeres, and F. Pachet. Deep learning techniques for music generation - a survey, 2017. URL https://arxiv.org/abs/1709.01620.

eCompetition. International piano-e-competition, 2017. URL http://www.piano-e-competition.com/.

S. Hedges. Dice music in the eighteenth century. *Music and Letters*, 59, 1978.

S. Lattner, M. Grachten, and G. Widmer. Imposing higher-level structure in polyphonic music generation using convolutional restricted boltzmann machines and constraints, 2017. URL https://arxiv.org/abs/1612.04742.

I. Malik and C. H. Ek. Neural translation of musical style. *CoRR*, abs/1708.03535, 2017. URL http://arxiv.org/abs/1708.03535.

K. McDonald. Neural nets for generating music, 2017. URL https://medium.com/artists-and-machine-intelligence/neural-nets-for-generating-music-f46dffac21c0.

G. Nierhaus. *Algorithmic Composition: Paradigms of Automated Music Generation*. 2009.

L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. In *ICLR*, 2016.