A Recurrent Neural Network as a Generative Model for Sheet Music

Eric Wolf （种伟神）

Mid-Project Report

Topic exploration

During my preparations for the project, I explored multiple interesting areas before eventually settling on automated sheet music composition.

Text simplification

The field of machine translation using machine learning techniques is currently evolving rapidly. This is evidenced by Google's most recent successes in developing a recurrent neural network for translation[1], which outperformed their existing production system. Since last month, Google is using neural machine translation in production[2]. Meanwhile, text simplification, i.e. the conversion of a sentence or document to a sentence or document of the same language, but with a smaller vocabulary and simplified structure, seems to have been neglected in current research. Traditional approaches to text simplification involve replacing or removing words or sentence fragments using elaborate rulesets and have not yet moved on to using machine learning techniques such as recurrent neural networks.

The models currently used in machine translation typically consist of an LSTM recurrent neural network for sequence-to-sequence learning that is symmetrically split into an encoder and decoder section, augmented by an attention module to determine what parts of the original and translated sentence correspond[3]. I hypothesize that the same models that perform well in machine translation could yield similar results in text simplification. Since simplification is a simpler task than translation, the model could probably be simplified while still maintaining usable results.

To train a text simplification model, a parallel corpus of corresponding non-simplified and simplified sentences is required. Most publications make use of an automatically sentence-aligned corpus derived from Wikipedia's Simple English articles[4]. However, upon personal inspection, the data quality of the Wikipedia corpus turned out to be wanting. This is confirmed by Wei Xu et al.[5], who argue that Newsela, a company providing English resources in various degrees of difficulty for language learners, can offer a much higher quality corpus. Unfortunately, Newsela did not respond to my requests for their data in time.

As a last resort, I considered making up for a lack of training data by devising a loss function incorporating both sentence complexity and semantic equivalence. However, as semantic similarity is itself a problem that is still very much subject to research, I decided against this and did not pursue text simplification further.

[1] Wu, Yonghui, et al. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." *arXiv preprint arXiv:1609.08144* (2016).

[2] "A Neural Network for Machine Translation, at Production Scale." https://research.googleblog.com/2016/09/a-neural-network-for-machine.html

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the International Conference on Learning Representations (ICLR).*

[4] Woodsend, Kristian, and Mirella Lapata. "WikiSimple: Automatic Simplification of Wikipedia Articles." In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*. 2011.

[5] Xu, Wei, Chris Callison-Burch, and William B. Dolan. "SemEval-2015 Task 1: Paraphrase and semantic similarity in Twitter (PIT)." *Proceedings of SemEval* (2015).

Paintings
The Web Gallery of Art[6] provides a free database of about 40'000 high-resolution pictures of paintings and sculptures, which could be sufficiently large to train a neural network. Besides the well-known style transfer from paintings to photographs[7], possible applications include classification by epoch or painter, building a generative model, detecting forgeries[8] or detecting potential artistic influence[9]. The Web Gallery of Art's dataset has not been used much in machine learning research so far and could therefore give rise to some interesting results. However, most of the possible applications are not fundamentally different from ideas tried and tested on other datasets. Furthermore, the size of the dataset is at the lower bound of the images required to train a neural network. Therefore, I also decided against working with the Web Gallery of Art's fine art dataset.

Music composition
Since the rise of LSTM recurrent neural network, remarkable demonstrations have shown that these networks can generate plausible prose, C code or latex documents[10]. However, little work has been done so far to apply RNNs to music generation. I seek to apply RNNs with LSTM or GRU cells to a dataset of annotated music in the MIDI format with the goal of generating aurally pleasing compositions.

Related works
First attempts to train an RNN on music go back as far as 1994, when Michael C. Mozer achieved modest success training a small recurrent neural network without LSTM gates on classical music[11]. In order to generate harmonically sounding composition, he tried to model similarly sounding notes as similar representations in the feature space.
More recently, Douglas Eck and Jürgen Schmidhuber used LSTM gates to compose music using a relatively simple RNN, encoding notes in a binary vector which allows for multiple notes to be played simultaneously. However, their representation is still limited, as it does not differentiate between multiple short notes and one long note of the same pitch.
Others have tried to reuse a character-based RNN to generate music[12]. This model is simple to set up, but has severe limitations as it does not necessarily produce a valid rhythm.
A completely different direction is taken by Google's DeepMind team. Their primary objective for WaveNet[13] was the development of a text-to-speech system by modeling sound as raw audio waveform. Their results are remarkable, but require very expensive computations. As a byproduct, WaveNet was also trained on music and produced reasonable-sounding results of short duration. However, as music generation was not DeepMind's main focus, it is hard to judge WaveNet's performance on music generation in comparison to other publications.
A few months ago, Google's Brain team has released their TensorFlow Magenta framework[14], which is intended for artistic uses of machine learning, especially in the field of music generation. They attempt to solve the problem of modeling long-term dependencies by augmenting an RNN with an attention mechanism as first

[6] Kren, Emil, and Daniel Marx. "Web Gallery of Art." (1996).

[7] Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "A neural algorithm of artistic style." *arXiv preprint arXiv:1508.06576* (2015).

[8] Polatkan, Güngör, et al. "Detection of forgery in paintings using supervised learning." *2009 16th IEEE International Conference on Image Processing (ICIP).* IEEE, 2009.

[9] Saleh, Babak, et al. "Toward automated discovery of artistic influence." *Multimedia Tools and Applications* 75.7 (2016): 3565-3591.

[10] Andrej Karpathy . "The Unreasonable Effectiveness of Recurrent Neural Networks." http://karpathy.github.io/2015/05/21/rnn-effectiveness/

[11] Mozer, Michael C. "Neural network music composition by prediction: Exploring the benefits of psychoacoustic constraints and multi-scale processing." *Connection Science* 6.2-3 (1994): 247-280.

[12] "Lisl's Stis": Recurrent Neural Networks for Folk Music Generation. https://highnoongmt.wordpress.com/2015/05/22/lisls-stis-recurrent-neural-networks-for-folk-music-generation/

[13] Aäron van den Oord, Sander Dieleman et al. "WaveNet: A Generative Model for Raw Audio." 2016.

[14] "Magenta: Make Music and Art Using Machine Learning." https://magenta.tensorflow.org

introduced by Bahdanau et al.[15] that lets the RNN learn to use a weighted vector of the previous RNN outputs as additional input. Their team also attempted to enforce harmonic constraints dictated by music theory using reinforcement learning[16]. Magenta's main focus does not appear to be fully automated composition, but rather providing useful tools to human composers, which is why it allows users to prime the RNN with a starting sequence whose composition is then automatically completed.

Dataset

The Lakh MIDI dataset[17] comprises about 170.000 MIDI files, 45.000 of which have been matched to entries in the Million Song Dataset[18], which contains genre labels. The Lakh dataset should therefore be large enough to successfully train a neural network and is significantly larger than the datasets used in most previous publications.

Challenges

Modeling sheet music is significantly more challenging than modeling languages, as music is usually polyphonic (one instrument can play multiple notes at the same time) and contains multiple instruments. Different instruments may sound alike, such as trumpets and trombones, but can also be completely unrelated, such as drums. It is not clear yet what the best way to model multiple instruments is.

Human composers use a broad range of concepts from music theory while writing new pieces. Harmonics are usually determined by the relative intervals between simultaneously or subsequently played notes. In order for the RNN to comprehend these relationships, notes are often modeled relative to previous notes. However, the final output of the network has to be converted back to absolute pitch values.

A major challenge when using recurrent neural networks is the modeling of long-term structures. So far, most attempts at automated composition could not capture long-term relationships in the compositional structure such as the repetition of a chorus or very well-known music theoretical concepts such as consonance an dissonance. Similarly, when generating English text, RNNs often output plausible sentences but lack structure spanning multiple paragraphs or chapters. The structure of music, however, is simple compared to that of a book, as the content does not refer to real world concepts that have to be understood. If the RNN could learn to produce simple structures such as A-B-A, that would already be a significant achievement. I hope to solve this problem either using an attention module, which takes all previously produced notes as input, or using a separate planning stage to lay out the song structure prior to composition.

Goal

The goal of my project is to generate aurally pleasing compositions using either a given genre as input or by priming the network with a starting melody. The quality of a generative model is hard to quantify, but the long-term goal is for the result to be indistinguishable from human compositions to a human listener. Another way to evaluate the performance while configuring the model could be the comparison of songs generated at various stages of the training process.

Planned methods

I plan to use a RNN using LSTM or GRU cells. To capture long-term structures, I want to use an attention mechanism. I will experiment with different and increasingly complex input formats to ideally capture multiple polyphonic instruments. If the structure permits it, I would like to try to apply Beam Search. Lastly, if there is enough time, I would like to experiment with a GAN architecture and/or reinforcement learning.

[15] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In Proceedings of the International Conference on Learning Representations (ICLR).

[16] Jaques, Natasha, et al. "Tuning Recurrent Neural Networks with Reinforcement Learning." *arXiv preprint arXiv:1611.02796* (2016).

[17] Colin Raffel. "Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching". PhD Thesis, 2016.

[18] Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. "The Million Song Dataset". In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pages 591–596, 2011.