Assessing quality of music generated by the Music Transformer with an American Folk dataset

github.com/gregwinther/folk_transformer

Sebastian G. Winther-Larsen *Institute of Informatics, University of Oslo*
Tom F. Hansen *Institute of Informatics, University of Oslo*
Bjørn Iversen *Institute of Informatics, University of Oslo*

***Abstract*—Here we will write the abstract.**

***Index Terms*—music generation, transformer model, evaluation, transfer learning, Americana**

I. Motivation and introduction

The Transformer model, implementing the attention principle [1], is today recognised as the best performing sequential machine learning model, surpassing RNN-based models in most cases, mainly argued by its better abilities to remember long term coherence, shorter training time and applicability in transfer learning. While originally used primarly for NLP, which today has mature implementations, the architecture can also be applied for other sequential models, such as music generation [2]. The music transformer developed in the Magenta project is trained on the Maestro dataset [3]. By setting a primer – a start music sequence, the model generates new music with good results along the same lines as the training set. With other primers than regular and systematic classical music in Maestro, the quality of the output is varying.

Motivated by generating more irregular music, the main aim of the paper is to describe a framework for the quality assessment of different approaches to the generation of music with the transformer model. By building, what is supposed to be the same music generator, but with different approaches, the evaluation framework will assess the quality of each approach. It is not known to the authors scientific papers describing such a comparison of music transformers, other than the normal comparison of music generated by RNNs and transformers [2]. At the core of such an architecture is, for each approach; 1) in order to make a fair comparison, a detailed description of model topology, its tuning parameters, and the size and structure of the dataset used for training, and 2) an evaluation format that combines a form of quantitative and qualitative evaluation technique.

To this end, we wish to employ the transformer music to a subgenre of music to which such a model has not been extensively applied. While initial findings from applying the transformer to jazz music has shown some limitations [4], while applying LSTM networks to Blues has been moderately succesful [5] and applying the transformer model to pop music seems to work well [6]. From a music theory standpoint this is very sensible - classical music often has formal rules, the epitome of which is the fugue [7]; and pop music follows some very clear norms [8]. While even Free Jazz has *some* rules, it readily falls into the category of the type of rhytmic music with the least amount of structure, per the definition that it is "characterized by the absence of set chord patterns or time patterns"[9].

American roots music, encompassing spirituals, cajun music, cowboy music, work songs, but also early blues such as Dixieland; from now on referred to as "Americana", presents itself as hitherto unexplored territory. It also provides a nice stepping stone towards more "unstructured" music as it often allows for improvisationg, but otherwise retains a relatively rigid structure [10]. We have therefor collected a dataset of MIDI files of Americana music, which we will use in our quality assessment framework.

II. Related work

The transformer model is considered state of the art in music generation, surpassing RNN-based models in the last few years. Both are sequential models, but the attention principle at the core of the transformer facilitates remembering coherence over longer sections of sequences and highlights especially important sections. From generating music of 10's of seconds with RNN, it is now possible to generate minutes of realistic music [2]. Still there is a lot of unresolved challenges, like generating long sections (over some minutes), in highly irregular compositions and multi channel (many instruments) signals. To combat these challenges the improvement of the transformer model has high focus in the research community. Some of the most recent attempts are the Transformer-XL [11] model and the Reformer [12] as a particularly promising candidate. A Reformer model claim to be trained on a standard computer with a single GPu. In this analysis we will utilize the original transformer architecture, as this is the model-architecture in the music transformer from Google.

A. Music transformers

In the few existing papers considering music generation with the *transformer* we want to emphasize some important related works besides the beforementioned music transformer. In the blues transformer

nevn blues transformer nevn LakhNES nevn jazz transformer pop music transformer Foley music - MIDI from video

However there is little work on generating intentionally the same music generator with different approaches. Another new approach is to use transfer learning from an existing model.

B. Evaluation of ML-generated music

In the evaluation of models of music generation based on machine learning we would like to point out the works by [13], [14] and [4].

These works describe either a qualitative evaluation or a qualitative evaluation. In our attempt we combine these 2 approaches and establish a common framework.

III. Methods

Acting as a base and for exemplification of the benchmark architecture, Americana music is generated in 2 different model concepts:

1) Utilize transfer learning with music transformer as a base, and train with the full dataset of Americana midi-files.
2) Train a new transformer model only using the full Americana dataset

The hypothesis is that concept number 1 will result in the best performing model, but an important issue is what makes up the best model and how to evaluate such a subjective "sequence-result" as music in a fair and trustworthy manner? Some will say this is an impossible task [13]. Will a transfer learning model be significantly better than only training on a single dataset, such as has been shown in other ML-applications, like image classification [**ref**], even though the MAESTRO dataset is totally different from the Americana dataset.

An attempt to sort this out is by evaluating in a quantitative and qualitative way. The quantitative, hence objective, way can shortly be described as a technical comparison of the predicted signal and the real signal. Principles by https://github.com/RichardYang40148/mgeval, [4] and will be utilized.

The qualitative part constitutes an music expert judgement, based on listening to the generated music files from the objective evaluation. In a second, and survey based part, a large number of random people is asked to rate the different music files.

Qualitative and quantitative measures will finally be summarised in a common scheme.

A. Datasets

MAESTRO [3] (MIDI and Audio Edited for Synchronous TRacks and Organization) is a dataset with over 200 hours of virtuosic piano perfomances captured with a fine alignment of approximately 3ms between note lables and audio waveforms.

The data is a produce from performances in the International Piano-e-competition. During each installment of the competiton, vrituoso pianists perform on Yamaha Disklaviers which, in addition to being concert-quality acoustic grand pianos, utilize integrated high-precision MIDI capture and playback.

Question: How many of the performances are of the same piece?

Americana The Americana dataset contains xxx hours of music in the subgenres xxx.

To make a fair comparison of the datasets we have set up some main statistical figures in table xxx. - Number of songs - Mean length of songs - +++

B. Model topology and tuning

C. Qantitative evaluation

D. Qualitative evaluation

IV. Results

Americana generator with transfer learning some tekst
Americana generator in stand alone learning some text

V. Discussion

VI. Conclusions and further development

References

1. Vaswani, A. *et al. Attention is all you need* in *Advances in neural information processing systems* (2017), 5998–6008.
2. Huang, C.-Z. A. *et al. Music transformer: Generating music with long-term structure* in *International Conference on Learning Representations* (2018).
3. Hawthorne, C. *et al. Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset* in *International Conference on Learning Representations* (2019). https://openreview.net/forum?id=r1lYRjC9F7.
4. Wu, S.-L. & Yang, Y.-H. The Jazz Transformer on the Front Line: Exploring the Shortcomings of AI-composed Music through Quantitative Measures. *arXiv preprint arXiv:2008.01307* (2020).
5. Eck, D. & Schmidhuber, J. *Finding temporal structure in music: Blues improvisation with LSTM recurrent networks* in *Proceedings of the 12th IEEE workshop on neural networks for signal processing* (2002), 747–756.
6. Huang, Y.-S. & Yang, Y.-H. Pop music transformer: Generating music with rhythm and harmony. *arXiv preprint arXiv:2002.00212* (2020).
7. Giraud, M., Groult, R., Leguy, E. & Levé, F. Computational fugue analysis. *Computer Music Journal* **39,** 77–96 (2015).
8. Hennion, A. The production of success: an anti-musicology of the pop song. *Popular Music* **3,** 159–193 (1983).
9. Jazz., F. *Oxford Languages* Accessed 2 October 2020 (Oxford University Press, September 2020).
10. Center, A. F. *Folk Music and Song* Accessed 2 October 2020. %5Curl%7Bhttps://www.loc.gov/folklife/guide/folkmusicandsong.html%7D.
11. Dai, Z. *et al. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context* 2019. arXiv: 1901.02860 [cs.LG].
12. Kitaev, N., Kaiser, Ł. & Levskaya, A. *Reformer: The Efficient Transformer* 2020. arXiv: 2001.04451 [cs.LG].
13. Eck, D. & Schmidhuber, J. *Finding temporal structure in music: blues improvisation with LSTM recurrent networks* in *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing* (2002), 747–756.
14. Yang, L.-C. & Lerch, A. On the evaluation of generative models in music. *Neural Computing and Applications* **32,** 4773–4784 (2020).