Benchmarking the Music Transformer with American Folk

github.com/gregwinther/folk_transformer

Tom F. Hansen, Bjørn Iversen and Sebastian G. Winther-Larsen

October 2, 2020

1 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 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 propose a method to benchmark different approaches for generating music with the transformer model. It is not known to the authors scientific papers describing such a comparison of music transformers, other than the typical comparison of music generated by RNNs and Transformers [4]. 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 finding from applying the transformer to jazz music has shown some limitations [5], while applying LSTM networks to Blues has been moderately succesful [6] and applying the transformer model to pop music seems to work well [7]. From a music theory standpoint this is very sensible - classical music often has formal rules, the epitome of which is the fugue [8]; and pop music follows some very clear norms [9]. 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"[10].

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 [11]. We have therefor collected a dataset of MIDI files of Americana music, which we will use in our evaluation anb benchmarking.

2 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. Still there is a lot of unresolved challenges, like generating long sections (over xxx min), 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 [4] and the Reformer [12] as particularly promising candidates. In this analysis we will utilize the original transformer architecture, as this is the model-architecture in music transformer from Google.

(Mention different use cases of music generation with the transformer model. The links below will be described in short.)

- `https://www.gwern.net/GPT-2-music#transformers`
- `https://magenta.tensorflow.org/music-transformer`
- `https://medium.com/swlh/create-your-own-classical-music-with-google-magenta-transforme`
- `https://towardsdatascience.com/creating-a-pop-music-generator-with-the-transformer-586`
- `https://github.com/scpark20/Music-GPT-2`
- `https://github.com/YatingMusic/remi`
- `https://github.com/chrisdonahue/LakhNES`
- `https://github.com/jason9693/MusicTransformer-tensorflow2.0`
- `https://github.com/magenta/magenta/tree/master/magenta/models/score2perf`

3 Methods

Acting as a base and for exemplification of the benchmark architecture, Americana music is generated in 3 different model concepts:

1. Directly from music transformer (trained on the Maestro dataset) with Americana midi-files as primers. This will act as the reference model for the 2 other approaches.

2. Utilize transfer learning with music transformer as a base, and train with the full dataset of Americana midi-files.

3. Train a new transformer model only using the full Americana dataset

The hypothesis is that concept number 2 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 (`https://ieeexplore-ieee-org.ezproxy.uio.no/stamp/stamp.jsp?arnumber=1030094&tag=1`). 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`, `https:/egithub.com/slSeanWU/MusDr` and `https://link-springer-com.ezproxy.uio.no/article/10.1007/s00521-018-3849-7` 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.

3.1 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?

3.2 Model topology and tuning

3.3 Qantitative evaluation

3.4 Qualitative evaluation

4 Results

5 Discussion

6 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. Dai, Z. *et al. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context* in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019), 2978–2988.
5. 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).
6. 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.
7. Huang, Y.-S. & Yang, Y.-H. Pop music transformer: Generating music with rhythm and harmony. *arXiv preprint arXiv:2002.00212* (2020).
8. Giraud, M., Groult, R., Leguy, E. & Levé, F. Computational fugue analysis. *Computer Music Journal* **39,** 77–96 (2015).
9. Hennion, A. The production of success: an anti-musicology of the pop song. *Popular Music* **3,** 159–193 (1983).
10. Jazz., F. *Oxford Languages* Accessed 2 October 2020 (Oxford University Press, September 2020).
11. Center, A. F. *Folk Music and Song* Accessed 2 October 2020. `%5Curl%7Bhttps://www.loc.gov/folklife/guide/folkmusicandsong.html%7D`.

12. Kitaev, N., Kaiser, L. & Levskaya, A. *Reformer: The Efficient Transformer* in *International Conference on Learning Representations* (2019).