Benchmarking music transformer models with folk music

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, containing in general classical piano music. 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 results is not that good, especially for irregular and unsystematic music genres like jazz [3].

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 (refs). 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.

A dataset of midi-files from the folk-based music genre Americana has been collected, a genre that can be seen a as a stepping stone from the successfully generated regular classical piano music to a more unsystematic, irregular challenging music genres. (sebastian describes how and why). Americana is used to exemplify the benchmark principles.

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 (ref) model and the Reformer (ref) as a particularly promising candidate. 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

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. 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).