Assessing Quality of Music Generated by the Transformer Models

github.com/gregwinther/folk_transformer

Sebastian G. Winther-Larsen
Center for Computing in Science Education, Department of Physics, 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 attention-based transformer model, 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 [1]. While originally used primarly for natural language processing (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 a minute of coherent 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 particularly promising candidates. 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,

built on the classical piano music dataset MAESTRO. These works are somewhat diversified in different music genres. In the pop music transformer [6] pop piano music is generated by the transformer. The paper shows that Transformers can do even better for music modeling, when the way a musical score is converted into the data fed to a Transformer model, is improved. This is performed by imposing a metrical structure in the input data, so that Transformers can be more easily aware of the beat-bar-phrase hierarchical structure in music. The new data representation maintains the flexibility of local tempo changes, and provides hurdles to control the rhythmic and harmonic structure of music. With this approach, this work claims to generate pop piano music with better rhytmic structure than the music transformer [2].

Another paper describes the Jazz transformer [4] which is in the other end of the spectrum related to complexity. Here the the Transformer-XL architecture is utilized to model lead sheets of jazz music. Moreover, the model endeavors to incorporate structural events present in the Weimar Jazz Database (WJazzD) for inducing structures in the generated music. Even though the training loss values are low, the esults are not impressive. Listening tests shows a clear gap between the ratings of the generated and real compositions. The work analyses the missing parts and presents a prediction system which in an analytical manner shed light on why machine-generated music to date still falls short of the artwork of humanity. This includes analyzing the statistics of the pitch class, grooving, and chord progression, assessing the structureness of the music with the help of the fitness scape plot, and evaluating the model's understanding of Jazz music.

In a system called LakhNES Donahue *et al.*, generate multi-instrumental music with the transformer [13]. Their success of music generation with the piano score generation is partially explained by the large volumes of symbolic data readily available for that domain. They leverage the recently-introduced NES-MDB dataset of four-instrument scores from an early video game sound synthesis chip[1]. They found this data to be well-suited to training with the Transformer architecture. The model was further improved with a pre-training technique to leverage the information in a large collection of heterogeneous music, namely the Lakh MIDI dataset. By performing transfer learning on the NES-MDB dataset, both the qualitative and quantitative performance from the target dataset was significantly improved.

Gan *et al.* use the transformer architecture to generate music, but with another approach. In a system called Foley Music they synthesize music from a silent video about people playing instruments [14]. A relationship between body key-points and MIDI recordings is established. Music generation is then formulated as a motion-to-MIDI translation problem, represented with a graph transformer framework that predict MIDI from motion. By testing the generator on different music performances the results is proven to outperform several existing systems in music generation.

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 high performance music model.

[1]The Nintendo Entertainment System (NES)

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 Eck & Schmidhuber [15], Yang & Lerch [16] and Wu & Yang [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 MAESTRO 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 [15]. 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 [17], 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 [16] and [4] 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.

A. Datasets

A brief summary of each of the datasets we have used in this study can be found in Table I.

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, virtuoso pianists perform on Yamaha Disklaviers which, in addition to being concert-quality acoustic grand pianos, utilize integrated high-precision MIDI capture and playback.

Since the **MAESTRO** dataset contains MIDI recordings from competions, the pieces are from a select set for each year. This means that many of the pieces are the same, but may

Table I
DATA SET DESCRIPTION

	MAESTRO v2	Americana
No. of songs	1282	5711
Total time [hours]	201	329
Mean length [min]	9.4	3.45

includes much variation within each performer's interpretation of the piece.

The **Americana** dataset is constructed from musical scores by Benjamin Robert Tubb[2]. These scores are in the public domain and composed between the ealry 1800s and 1922. The genres range from blues, ragtime, naval songs, hyms, minstres songs and spirituals.

B. Qantitative evaluation

For quantative evaluation we will be using the objective evaluation toolbox mgeval [16]. The toolbox lets us extract absolute metrics from MIDI files which lets us inspect the properties of both the dataset used for training and the generated dataset. The features extracted for absolute measures are divided into pitch-based features Pitch count (PC), Pitch class histogram (PCH), Pitch class transition matrix (PCTM), pitch range (PR) and Average pitch interval (PI) and rhythm-based features, Note count (NC), Average inter-onset-interval (IOI), Note length histogram (NLH) and Note length transition matrix (NLTM). These metrics can then be used to acquire the relative metrics between datasets with the use of exhaustive cross validation to acquire the distance between each sample of same set (intra-dataset) and another set (inter-dataset).

C. Evidence-Based Design for Assessment and Evaluation

As a rigorous and well-proven approach do construct a framework for assessing music composed by artificial intelligence models, we propose adapting the methodology of Evidence-Centered Design [18, 19] By working with Evidence-Centered Design (ECD), we engage in a modern approach to assessment design, for assessing complex knowledge and practices. ECD is originally applied to the contruction of psychometric learning assessment tools. Through to completion, it would take several years to construct such a tool, something that is well outside the scope of this study. However, we prospose to begin with the first step within ECD - **Domain Analysis**. This involves exploratory interviews of experts in the field in order to contruct a thematically organized and prioritized list of knowledge and practices to assess. Scecific to our study, we find it necessary to talk to professional musicians and composers in order to uncover what actually makes a good composition.

IV. RESULTS

AT THE TIME OF WRITING TRAINING OF MODELS HAVE NOT BEEN COMPLETED. THE TRAINING PROCESS HAS BEEN DEMANDING AND HAVE TAKEN LONGER TIME THAN EXPECTED. RESULT IS EXPECTED DURING THE WEEKEND. THE INITIAL RESULTS ARE PROMISING.

The plan - focus the comming days:

- Generate musicals samples with the two models. Distribute in questionaire for comparative assessment in A/B-style.
- Interview musical "experts", most likely in a focus group to uncover necessary components of a good composition. Would include playing generated music for participants.
- Apply quantitative analysis tools to generated music.
- Complet the paper with results and discussions.

V. DISCUSSION

VI. CONCLUSIONS AND FURTHER DEVELOPMENT

[2]These were scraped from a webpage that has since been taken down. Consequently, we are unable to provide a proper reference

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. Donahue, C., Mao, H. H., Li, Y. E., Cottrell, G. W. & McAuley, J. LakhNES: Improving multi-instrumental music generation with cross-domain pre-training. *arXiv preprint arXiv:1907.04868* (2019).
14. Gan, C., Huang, D., Chen, P., Tenenbaum, J. B. & Torralba, A. *Foley Music: Learning to Generate Music from Videos* 2020. arXiv: 2007.10984 [cs.CV].
15. 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.
16. Yang, L.-C. & Lerch, A. On the evaluation of generative models in music. *Neural Computing and Applications* **32,** 4773–4784 (2020).
17. Shin, H. *et al.* Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging* **35,** 1285–1298 (2016).
18. Mislevy, R. J., Steinberg, L. S. & Almond, R. G. Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives* **1,** 3–62 (2003).
19. Mislevy, R. J., Haertel, G., Riconscente, M., Rutstein, D. W. & Ziker, C. in *Assessing model-based reasoning using evidence-centered design* 19–24 (Springer, 2017).