Title of your thesis

Your Name

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Your Department

Your Advisor, Chair
First Committee
Second Committee
Third Committee
Last Committee

December 4, 2020
Blacksburg, Virginia

Keywords: Some Keywords, Subject matter, etc.

Title of your thesis

Your Name

ABSTRACT

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Dedicated to Virginia Tech.

Acknowledgments

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Contents

List of Figures **vii**

List of Tables **viii**

1 Introduction **1**

2 Background **2**

2.1 Expressive Performance Generation 2

2.2 Data 4

2.3 Transformers 4

2.4 Evaluation 5

3 Related Work **7**

3.1 Existing EPG models 7

3.2 Datasets 8

4 Experiments **10**

4.1 Model and Experiments 10

4.2 Evaluation 11

5 Results **13**

5.1 Quantitative 13

5.2 Qualitative 13

6 Discussion **14**

Appendices **15**

Appendix A Appendices I **16**

A.1 A1 16

A.2 A2 16

Bibliography **17**

List of Figures

List of Tables

Chapter 1

Introduction

- Introduce the idea of musical research with computers. Talk about the illiac suite [1] and Music Information Retrieval.
- Significance of machine learning on the field
- Introduce idea of expressive musical performance. Brief conversation about the different performance components (articulation, dynamics, timing).
- Using Transformer architecture which hasn't been done in the field.
- Richard: report results

Chapter 2

Background

Provide additional context to the problem of expressive musical performance (EMP) and the model domain we are applying (Transformers). Make sure to talk about the intricacies behind the musical side of the problem, given that it is generally not as well known in computer science, ML, and AI. Introduce the idea of music information retrieval (MIR) research, and how EMP fit's into this research. Provide sufficient detail at a high level detailing exactly what EMP is and why it is an interesting problem, specifically for machine learning. Cover what type of data is required for the problem.

Cover the Transformer and why it is worth it to apply this model to the problem domain.

2.1 Expressive Performance Generation

Define expressive performance generation (EPG) at a technical level (data features). Give background into how it fits into MIR research.

- Define a Score and Performance
  - Talk about differences between score and performance at a higher level.
  - Score includes symbolic representation of music and includes pitch, tempo (sometimes), timing, dynamics, and phrasing.

  - Performance is an interpretation of a score. Includes the note pitches, tempo, timing, deviation, articulation, and dynamics.
  - EPG is the task of creating a model which takes in a score (usually in the form of MusicXML) and outputs a performance (usually MIDI).

Richard

Add reference to data format section

  - Score to Performance Alignment. Necessary to the notes of a performance with their corresponding position in a score. Because performances are so varied, this is a non-trivial problem.
  - Papers
    * Basis Mixer [2]
    * Computational Models for Expressiveness [3]
  - Richard: Add reference to later section which talks about feature engineering in detail
  - Richard: Create (find) figures for score and performance

- Explain how expressive performance generation fits into music generation research
  - Generation as subset of MIR research
  - Different components of generation. Composition, performance, and synthesis.
  - Richard: Create graph showing (or referencing other graphs) of where performance generation fits into music generation as a whole
  - Papers
    * This time with feeling [4]
    * Deep learning for music generation survey [5]

2.2 Data

A brief section about the data used for the problem. Introduce MusicXML and MIDI

- MusicXML
  - A text based representation of a musical score.
  - Created as a way to standardize score data among different notation software.
  - Useful for EMP research because of the standardized format.
  - Contains all relevant information about the score and it's related features.

Richard
Add reference to feature section

- MIDI
  - Event based protocol for digital representation of musical instruments.
  - Used in a variety of ways, most commonly known for it's use in DAW software to represent easily editable tracks for music production.
  - Can be synthesized in many different ways.
  - Contains all of the needed information to represent a musical performace. .

Richard
Reference feature section

2.3 Transformers

Provide context to why transformers are important and the problems they've solved in nlp.

- Intuition behind transformers and why they are so powerful in sequence modeling
- Attention is all you need paper [6]
  - State of the art in translation tasks

  - New architecture for sequence modeling using only attention. No recurrent network

- BERT [7]
  - Transformer Encoder only
  - Self-supervised learning and pre-training. Includes having a simple multi-layer perceptron at the end to make it useful

- Music Transformer [8]
  - Builds off of This Time with Feeling[4] paper. Both composition and performance generation at the same time
  - Implements full transformer architecture
  - Achieves better results than LSTM

- Question: Can a transformer model be applied to only performance generation with an encoder only architecture to achieve better results than current state of the art models?. Intuition says yes given the results from Music Transformer.

2.4 Evaluation

- Evaluation is particularly difficult for a problem like EPG because there is no ”correct” interpretation of a score. However, there is at least a vaguely understood relationship between a score marking and how a performaner should use that marking within the context of a performance. For example, if a crescendo marking is used in a score, the performer should at the very least increase the volume of the performance relative to the current volume of the piece. The amount which the volume should increase or the rate at which it increases are not clearly defined, but the fact of the increase of volume itself is. This is the fundamental intuition behind the motivation to build computational models for expressive performance. Nonetheless, it still remains a difficult job to evaluate a given EPG model because of the ambiguity of what is ”correct” or not.

- Evaluation methods used so far in EPG models are broken into two categories, quantitative and qualitative.
- Quantitative:
  - This follows standard techniques for experimentation of evaluation of ML models in general. It usually involves calculating a numerical value for a models inference on a separate test data set that was not used for model training or model selection. . Common metrics for regression like problems are mean squared error (MSE) and the pearson correlation coefficient (R2).

    Richard
    Find reference for ML training and evaluation
  - Due to the nature of EPG model evaluation mentioned above, it is not clear that ”better” quantiative metric score for a given model over another indicates that the performance of the model is superior. .

    Richard
    Find section in Garcon survey that references this point
- Qualitative
  - Qualitative evaluation methods involve gathering human feedback by playing performances of a given models performance to an audience and getting ratings or judgement of the model according to a predefined questionnare or survey method. The nature of these evaluation methods is not consistent in the current literature and remains a challenge for the field to solve in the future. .

    Richard
    Find section in
  - Richard: Conduct more research for reference on current methods for qualitative evaluation

Chapter 3

Related Work

3.1 Existing EPG models

- KTH system [9]. A rule-based system for expressive performance. Rules are selected through a empirical process based on human feedback.

- YQX. A Bayesian network that models timing, dynamics, and articulation [10]. Won the 2008 RenCon contest.

- Basis Function Models [2]
    - Linear Basis Functions. Uses Least Squares regression and Bayesian models with about the same performance
    - Non-Linear Basis Functions. Uses standard feed-forward network. FFNN perform better than Linear models. Also uses an RNN.

- Giraldo and Ramirez use several different ML algorithms, including Decision Trees, k-NN, SVM's, and FFNN to build an expressive performance generation system for improvisational Jazz guitar [11].

- Moulieras and Pachet use a Maximum Entropy model to infer the underlying distribution of expressive performance and build a generation system trained from a mix of popular music. Their expressive model outperforms base models in listening tests [12].

Richard

This needs more exploration. Lot of possibilities for future work

- Jeong builds two versions of virtuosoNet, one using a recurrent hierarchical attention network (HAN) [13], and another using a recurrent graph network [14]. These models are built using a dataset order of magnitudes larger than other datasets and attempt to model the expressive performance feature of the pedal, which no other model does. The code for the models is also open source so it was chosen as the starting place for this work.

  Richard: Add more papers and expand upon the existing research a bit more. Isn't completely necessary but will be good for my overall understanding

3.2 Datasets

- Talk about the fundamental limitations of gathering data for this problem, especially in relation to other fields [9]. Because of this, the lack of high-quality data is limited.

- The dataset used for the virtuosoNet [14] [13] will be the dataset used for the experiments. At the time the experiment started it was the largest publicly available dataset applicable to the EPG systems, and was chosen for use. A recent publication [15] builds off of the dataset used for the virtuosoNet with more sophisticated alignment and some extensions to the size (dataset is named ASAP). ASAP would be more appropriate for future use.

- One of the necessary data processing tasks for EMP is the alignment between the score and performance of a given piece. Because there is always an inherent interpretation of a composition by a performer, there is no clear mapping between any given score and performance. It remains necessary to have some sort of alignment process to match each note in the performance with its related position in the score.

  Richard: Reference this in the introduction

Richard: This needs more research. Find relevant papers to cite, as well as show a diagram that makes it clear why alignment is necessary

.

Chapter 4

Experiments

4.1 Model and Experiments

- Due to the open-source nature of virtuosoNet project and its attempt to build a more cohesive EPG model by introducing the pedal as an expressive feature and training on a much larger dataset, we built off of this model.

- Because of the significant advances in other sequence modeling domains (such as NLP) and the indication of increased performance of another related task with the Music Transformer [8], the main question we want to answer is whether we can see similar increases in model performance by applying a Transformer ANN architecture to the problem domain.

- We will experiment with a transformer encoder only architecture similar to BERT. The problem includes a 1-1 to mapping between every note in the score and a related note in a performance. This is different than seq-2-seq modeling problem such as neural machine translation which maps a sequence of one length to another sequence of a different length, which is what the full Transformer architecture was intended for. The Transformer Encoder can be seen as as a large encoder that learns the best representation for a given feature set. The model we'll build will use a simple FFNN that accepts the output of the transformer encoder to decode this representation and give the final feature set which is then used to create a performance. This is similar to the BERT architecture and it's intended application.

Richard: Come up with a more detailed explanation of this modeling choice. Also create a visual diagram that explains the transformer encoder with the simple regression model sitting on top of it

- Because we are using the same dataset used to train virtuosoNet, we will directly compare the performance a Transformer model to the existing virtuosoNet models using the same quantitative metric, MSE.

  Richard: Come up with specific model experiments and comparison in a table. Table doesn't have to have results but needs the general outline that will be used in the final paper

4.2 Evaluation

- Quantitative: Because we are using the same dataset used to train virtuosoNet, we will directly compare the performance a Transformer model to the existing virtuosoNet models using the same quantitative metric, MSE.

  Richard: Come up with specific model experiments and comparison in a table. The table doesn't have to have results but needs the general outline that will be used in the final paper

- Due to time and resource constraints, no sophisticated qualitative evaluation was conducted for the models. However, a personal evaluation was used during the entire model development process.

  Richard: Talk about method used for personal analysis

-

Chapter 5

Results

5.1 Quantitative

Add table with results of experiments along with explanations.

5.2 Qualitative

Give personal qualitative report.

Chapter 6

Discussion

Richard: The following are some interesting discussion ideas that have come up so far. There is no telling if these will be in the final paper or not after conducting more experiments.

- Transformer performs worse according the quantitative metrics. This could be because it doesn't build in a specific hierarchical layer that is specific the problem. It is a much more generic model. There is a lot of room for exploration into experimenting with different architectures based on the Transformer to better fit the problem domain.

  Richard: Add more discussion based on more results

- Transformer appears to be a more dynamic model than the recurrent virtuosoNet model that makes more ”mistakes”. Does this mean that it is more ”human”.

Richard

Add discussion of uncanny valley

- Pedal in performance is messy. Could be because of problems in the feature and modeling, or could just be because it is a difficult problem to model.

-

  Richard: Discussion on qualitative results

Appendices

Appendix A

Appendices I

A.1 A1

A.2 A2

Bibliography

[1] O. Sandred, M. Laurson, and M. Kuuskankare, "Revisiting the illiac suite–a rule-based approach to stochastic processes," *Sonic Ideas/Ideas Sonicas*, vol. 2, pp. 42–46, 2009.

[2] C. Eduardo, *Computational modeling of expressive music performance with linear and non-linear basis function models.* PhD thesis, JOHANNES KEPLER UNIVERSITY LINZ, 2018.

[3] C. E. Cancino-Chacón, M. Grachten, W. Goebl, and G. Widmer, "Computational models of expressive music performance: A comprehensive and critical review," *Frontiers in Digital Humanities*, vol. 5, p. 25, 2018.

[4] S. Oore, I. Simon, S. Dieleman, D. Eck, and K. Simonyan, "This time with feeling: Learning expressive musical performance," *Neural Computing and Applications*, vol. 32, no. 4, pp. 955–967, 2020.

[5] S. Ji, J. Luo, and X. Yang, "A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions," *arXiv preprint arXiv:2011.06801*, 2020.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, pp. 5998–6008, 2017.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[8] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, I. Simon, C. Hawthorne, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck, “Music transformer,” *arXiv preprint arXiv:1809.04281*, 2018.

[9] A. Friberg, R. Bresin, and J. Sundberg, “Overview of the kth rule system for musical performance,” *Advances in Cognitive Psychology*, vol. 2, no. 2-3, pp. 145–161, 2006.

[10] G. Widmer, S. Flossmann, and M. Grachten, “Yqx plays chopin,” *AI magazine*, vol. 30, no. 3, pp. 35–35, 2009.

[11] S. Giraldo and R. Ramirez, “A machine learning approach to ornamentation modeling and synthesis in jazz guitar,” *Journal of Mathematics and Music*, vol. 10, no. 2, pp. 107–126, 2016.

[12] S. Moulieras and F. Pachet, “Maximum entropy models for generation of expressive music,” *arXiv preprint arXiv:1610.03606*, 2016.

[13] D. Jeong, T. Kwon, Y. Kim, K. Lee, and J. Nam, “Virtuosonet: A hierarchical rnn-based system for modeling expressive piano performance.,” in *ISMIR*, pp. 908–915, 2019.

[14] D. Jeong, T. Kwon, Y. Kim, and J. Nam, “Graph neural network for music score data and modeling expressive piano performance,” in *International Conference on Machine Learning*, pp. 3060–3070, 2019.

[15] F. Foscarin, A. Mcleod, P. Rigaux, F. Jacquemard, and M. Sakai, “Asap: a dataset of aligned scores and performances for piano transcription,” in *ISMIR 2020-21st International Society for Music Information Retrieval*, 2020.