Accademia Musicale Studio Musica

Michele Della Ventura, *editor*

2020

Proceedings of the
International Conference on
**New Music Concepts
Inspired Education and
New Computer Science Generation**

Vol. 7

STUDIO
MUSICA

Accademia Musicale Studio Musica

International Conference on New Music Concepts
Inspired Education and
New Computer Science Generation

Proceeding Book
Vol. 7

Accademia Musicale Studio Musica
Michele Della Ventura
Editor

Preface

This volume of proceedings from the conference provides an opportunity for readers to engage with a selection of refereed papers that were presented during the International Conference on New Music Concepts, Inspired Education and New Computer Science Generation. The reader will sample here reports of research on topics ranging from a diverse set of disciplines, including mathematical models in music, computer science, learning and conceptual change; teaching strategies, e-learning and innovative learning, neuroscience, engineering and machine learning.

This conference intended to provide a platform for those researchers in music, education, computer science and educational technology to share experiences of effectively applying cutting-edge technologies to learning and to further spark brightening prospects. It is hoped that the findings of each work presented at the conference have enlightened relevant researchers or education practitioners to create more effective learning environments.

This year we received 57 papers from 19 countries worldwide. After a rigorous review process, 24 paper were accepted for presentation or poster display at the conference, yelling an acceptance rate of 42%. All the submissions were reviewed on the basis of their significance, novelty, technical quality, and practical impact.

The Conferece featured three keynote speakers: Prof. **Giuditta Alessandrini** (Università degli Studi Roma TRE, Italy), Prof. **Renee Timmers** (The University of Sheffield, UK) and Prof. **Axel Roebel** (IRCAM Paris, France).

I would like to thank the Organizing Committee for their efforts and time spent to ensure the success of the conference. I would also like to express my gratitude to the program Committee members for their timely and helpful reviews. Last but not least, I would like to thank all the authors for they contribution in maintaining a high-quality conference and I hope in your continued support in playing a significant role in the Innovative Technologies and Learning community in the future.

March 2020   Michele Della Ventura

Conference Chair

Michele Della Ventura, Accademia Musicale Studio Musica, Treviso, Italy

Keynote Speakers
Giuditta Alessandrini, Università degli Studi Roma TRE, Italy
Renee Timmers, The University of Sheffield, UK
Axel Roebel, IRCAM Paris, France

International Scientific Committee
Patricia Alessandrini, Goldsmiths, University of London, UK
Jioanne Armitage, University of Leeds, UK
Suzanne Aspden, Faculty of Music, University of Oxford, UK
Jean-Julien Aucouturier, IRCAM, Paris, France
Per Bloland, Miami University, Ohio, USA
Jeffrey Boehm, Bath Spa University, UK
David Carabias Galindo, University of Segovia, Spain
Marko Ciciliani, University for Music and Performing Arts Vienna, Austria
Sally Jo Cunningham,University of Waikato, New Zealand
Ching-Hua Chuan, University of North Florida, U.S.A.
Darryl N. Davis, University of Hull, UK
Marlo De Lara, University of Leeds, UK
Elga Dorner, Central European University, Budapest, Hungary
Simon Emmerson, De Montfort University, Leicester, UK
Travis Garrison, University of Central Missouri, USA
Inés María Monreal Guerrero, University of Valladolid, Spain
Duncan Williams, University of Plymouth, UK
Andrew Hankinson, Bodleian Libraries, University of Oxford, UK
Joseph Hyde, Bath SPA University, UK
Wladyslaw Homenda, Warsaw University of Technology, Poland
Orestis Karamanlis, Bournemouth University, UK
Alexandros Kontogeorgakopoulos, Cardiff Metropolitan University, UK
Steven Jan, University of Huddersfield, UK
Tae Hong Park, New York University Steinhardt, USA
Rudolf Rabenstein, University Erlangen-Nuremberg, Erlangen, Germany
Silvia Rosani, Goldsmiths, University of London, UK
Robert Rowe, New York University, USA
Nikos Stavropoulos, Leeds Beckett University, UK
Jacob David Sudol, Florida International University, U.S.A.
Eva Zangerle, University of Innsbruck, Austria

Contents

Learning Tools, Leraning Technologies, Learning Practices

Culture and Music

Poster presentation

New Music Concepts

Conditional Modelling of Musical Bars with Convolutional Variational Autoencoder

A. Oudad, H. Saito

Department of Computer Science, Keio University
adam@nak.ics.keio.ac.jp, hxs@nak.ics.keio.ac.jp

Abstract. Generational machine learning has been applied to various fields and is currently under exploration for music. Automatic music composition is receiving attention recently thanks to advance in distributed modelling using neural networks. Yet, musical models lack the ability to effectively capture harmonic and rhythmic features of music. We propose a model using variational autoencoder for encoding musical bars of ragtime piano music. Ragtime piano music is composed of left hand and right hand with different roles, thus we use righthand musical information to condition our variational autoencoder when reconstructing left-hand musical bars. We show our model is able to reconstruct and interpolate musical bars, thus providing useful musical bar embeddings for music generation.

Keywords. Deep Learning, Algorithmic Music Composition, Variational Inference, Music Information Processing.

1 Introduction

Automatic music composition, or algorithmic music generation, refers to the field of Artificial Intelligence that aims to generate music with computers. Most notable approaches include statistical approaches using hidden markov model [1] with relative success. As far as we incorporate enough carefully engineered-features, such as chord progressions, such models are able to produce musical improvisations. Results are often heavily constrained by chord transitions, sometimes abrupt.

The advances of distributed approach using neural networks have made huge progress possible on various fields related to generation using machines. Distributed models learn features via deep neural architectures, such as Long Short Term Memory (LSTM) networks [2] which can learn long-term dependencies, or Convolutional Neural Networks (ConvNet) [3] which efficiently process images and audio spectrograms.

Efficient learning frameworks have been developed around distributed approaches, such as Generative Adversarial Networks (GAN) [4] and Variational Autoencoders (VAE) [5][6]. GAN, given huge amount of data, are usually able to outperform other

generative models, by carefully controlling adversarial learning and avoiding mode collapse. Yet VAE have been preferred on relatively small datasets (thousands to hundreds of thousands of samples), as it is the case for molecule sequence modelling [7].

In this paper, we provide a way to model music using pianoroll representation. We also propose some metrics, incorporated to our loss function during training to assess the ability of the model to learn meaningful representation of music. To assess the reconstruction ability of our model, we use ragtime piano music, a musical genre for piano with separate left-hand and right-hand respective roles. Its structure has been defined by Scott Joplin in 1908 [8]. Fig. 1 shows an excerpt from The Entertainer, by Scott Joplin.

Fig. 1. Example of ragtime music from The Entertainer, by Scott Joplin (1902).

2   Data processing

The dataset of ragtime piano music is a collection of 109 files encoded in MIDI format. Files were scraped from a ragtime piano music collection on the internet (https://trachtman.org/). MIDI files are processed to separate left- and right-hand parts and obtain pianoroll representation in two different MIDI files. In ragtime music, left and right hands play very distinct roles in different range of notes. Higher notes are usually reached by right hand while lower notes are reached by right hand. We assume the first track to be the one dedicated to right hand, and the second one to be dedicated to left hand. Then, each MIDI file is exported to pianoroll using pretty_midi python library, with a sampling frequency computed with the following formula (1).

$$f = \frac{q}{rt_1} \tag{1}$$

where r is the resolution of the MIDI file (which is a number in ticks per quarter notes) and t_1 the time occurrence of the first tick. Finally, q is a parameter we can set to decide how many timesteps are equivalent to the duration of one quarter note. With $q = 4$, we match a timestep to the duration of a 32th note. The resulting pianorolls are cut into smaller 2D matrices of 32 timesteps, corresponding to one measure in the original score. We get a dataset of 1271328 pianoroll representations of measures, labelled with right or left hand, that is, their role in the original scores.

In the following, we refer to a sample from the dataset as $x = (x^r, x^l)$

where x^r is the righthand pianoroll and x^l is the lefthand pianoroll. For timestep j and note pitch i, the pianoroll has binary activation $x_{ij} = \left(x_{ij}^r, x_{ij}^l\right)$ with 0 for note off, and 1 for a note on.

3   Model

We use a variational autoencode to encode musical bars into a fixed-sized vector representation. The model is evaluated on reconstruction of musical bars for left hand, and conditioned using a dense layer connection on corresponding right hand's musical bar.

To encode an input left-hand pianoroll, a convolutional block is applied two times. This block consists of of a convolutional layer of 256 filters and kernel of size 4x4, a batch normalization layer, and ReLU activation. Then a fully-connected layer maps the 256 feature maps to an intermediate embedding of dimension 128. The latent space dimension is set to 32. We condition the output of the encoder with the righthand pianoroll corresponding to the input lefthand pianoroll. We keep the harmonic features of the righthand pianoroll and concatenate its features to the intermediate representation of the lefthand pianoroll. The harmonic features are computed using the *Harmony* operation detailed in next section. The decoder network uses a deconvolutional block, similar to the encoder. The overall network architecture is described by Fig. 2. Since we aim to predict binary target values, the neural network is initialized using a uniform distribution. The neural networks parameters are initialized with a uniform distribution, and RMSprop is used for optimization.
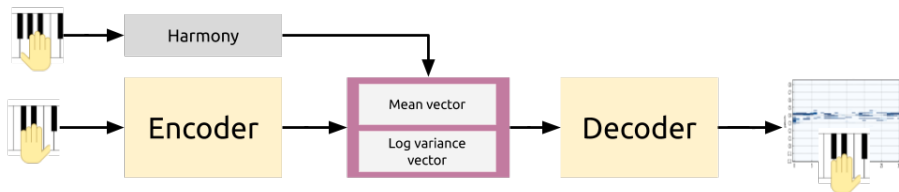
Fig. 2. Model architecture for left-hand musical bar reconstruction, conditioned on righthand musical bar. Harmony block is defined in next section

4   Incorporating rhythmic and harmonic components

We define operations which extract rhythmic and harmonic information from pianorolls. Then, we define a novel loss function with these operations, as well as a conidtional *harmony* block which inform the neural network of the righthand pianoroll when reconstructing the left-hand pianoroll, as it is shown in Fig. 2. The *Harmony* operation aims at extracting notes played in the pianoroll.

$$\text{Harmony}(x, p) = \sum_{t,k} x_{t,12k+p}$$

For the implementation, we use a dilated convolution over pitches, with a dilation factor of 12. The *Harmony* operation retrieves the number of timesteps during which a given note is played in the pianoroll. We also define a rhythmic vector, which detects note on and note off events in the pianoroll. For doing so, we use a 2D convolutional kernel with ReLU activation $k_{ON} = [-1,1]$ for note on, and $k_{OFF} = [-1,1]$ for note off. We then sum over all pitches to obtain a time vector of note events.

$$\text{NoteON}(x, t) = \sum_{p} \text{ReLU}\big(x_{t+1,p} - x_{t,p}\big)$$

$$\text{NoteOFF}(x, t) = \sum_{p} \text{ReLU}\big(x_{t,p} - x_{t+1,p}\big)$$

We thus obtain three feature vectors giving rhythmic and harmonic information of a pianoroll. The loss function which is used as objective of learning is defined as

$$L(x, \hat{x}) = \sum_{t,p} \big(x_{t,p} \log(\widehat{x_{t,p}}) + (1 - x_{t,p}) \log(1 - \widehat{x_{t,p}})\big)$$

$$+ \frac{1}{32} \sum_{t} \big(\text{NoteON}(x, t) - \text{NoteON}(\hat{x}, t)\big)^2$$

$$+ \frac{1}{32} \sum_{t} \big(\text{NoteOFF}(x, t) - \text{NoteOFF}(\hat{x}, t)\big)^2$$

$$+ \frac{1}{12} \sum_{p} \big(\text{Harmony}(x, p) - \text{Harmony}(\hat{x}, p)\big)^2$$

The above loss function adds a binary cross-entropy component between input and output pianorolls, respectively x and $\hat{x}$, for minimizing reconstruction error, with mean squared root error components on rhythmic and harmonic vectors of input and output.

5   Results

Experiments were conducted on Nvidia GTX 1080 Ti. The model is trained over 1000 epochs. Training takes about 10 hours to complete. Reconstruction results are shown in Fig. 3 and 4. Results in Fig. 3 show that the system was able to learn to reconstruct from its internal representation of musical bars. Reconstruction is noisy but at listening, the sound is harmonically and rhythmically consistent with the original score. In Fig. 4, we evaluated the ability of the system to interpolate between two pianorolls. On the left and right are two different pianorolls from the dataset. In the center is a concatenation of the

first half of the pianoroll on the left and the second half of the pianoroll on the right.

The system effectively learns to reconstruct chords and melodies without explicitly designed to do so. Outputs also suggest that the system captures some insight on syncopation from the original pianoroll, as it is able to organize output notes with approximately the same duration as the input pianoroll. The main problem displayed is that the pianorolls have a lot of notes which sometimes disregard principles of harmony. In Fig. 5, we compare the impact of adding the righthand pianoroll conditioning for left-hand pianoroll reconstruction. Thanks to this conditioning, the loss is able to converge faster.

Fig. 3. Charles Johnson's Dill Pickles rag extract (top) and reconstruction by the system (bottom).
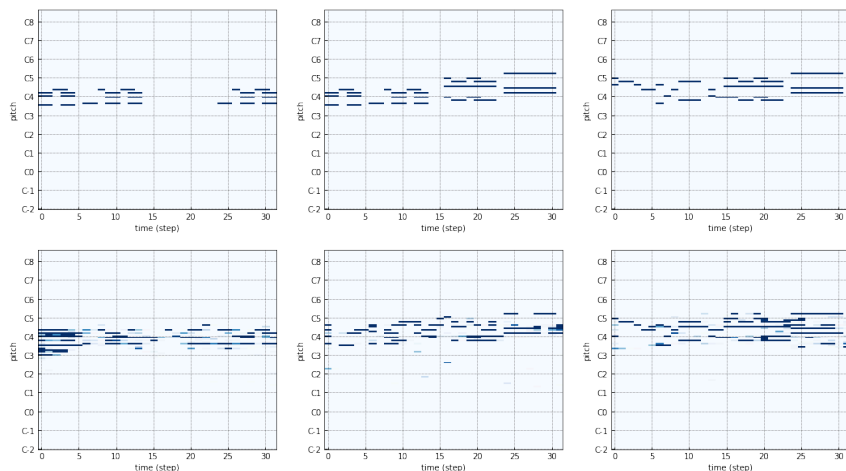

Fig. 4. Pianoroll half-interpolation. Top-left and top-right pianorolls are real samples from the dataset. Top-center pianoroll is a concatenation of first half and second half of respectively top-left and top-right pianorolls. Under each pianoroll is the reconstruction by the system.
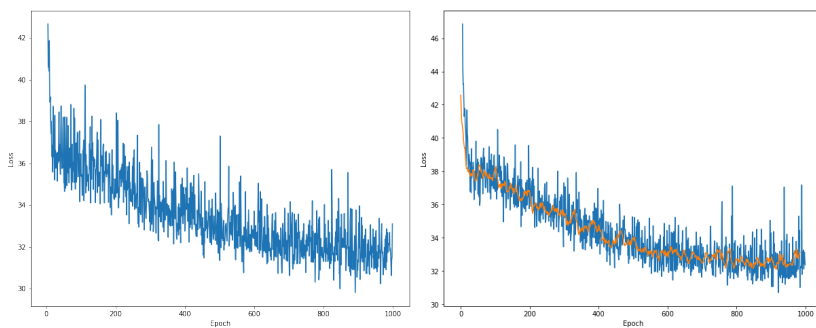
Fig. 5. Reconstruction loss. Left graph shows reconstruction loss of standard variational autoencoder on left-hand pianoroll. Right graph shows loss of variational autoencoder using harmonic and rhythmic information of righthand and left-hand pianorolls. In red is a moving average over loss.

6   Conclusion

In this paper, we show that provided minimal constraints on pianorolls, music representations can be learnt using a Variational Autoencoder. Adding additional constraints on the loss function help in keeping rhythmic and harmonic information in the output pianoroll, as well as speeding up convergence.

State-of-the art in Music generation is evolving and by exploring various music generation techniques, we can reach human-level composition, that is, generated music which sounds plausible for a human discriminator. The objective of this research was to assess how well the Variational Autoencoder approach is able to learn a specific piano genre, that is Ragtime piano music. We believe this work to be a good basis for music generation, by sampling vectors from the latent space, and predicting next pianoroll segment, using, for example, recurrent neural networks. Further work will involve improving the reconstruction quality of the network. We believe quality can be improved by learning on bigger datasets and fine-tuning to our ragtime piano dataset.

References

[1]   J.Gillick, T. Kevin, R. Keller, "Machine Learning of Jazz Grammars", Cambridge : MIT Press, 9 2010, Comput. Music J., Vol. 34, pp. 56-66. ISSN: 0148-9267.

[2]   E.Douglas, J. Schmidhuber, "Learning the Long-Term Structure of the Blues", ICANN, 2002.

[3]   A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio", 2016.

[4]   H. Dong, W. Hsiao, L. Yang, Y. Yang, "MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompani-

ment", Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018. pp. 34-41.

[5]   G. Brunner, A. Konrad, Y. Wang, R. Wattenhofer, "MIDI-VAE: Modeling Dynamics and Instrumentation of Music with Applications to Style Transfer", Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, pp. 747-754, September 23-27, 2018.

[6]   A. Roberts, J. Engel, C. Raffel, C. Hawthorne, D. Eck, "A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music", Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, pp. 4361-4370, July 10-15, 2018.

[7]   R. Gomez, J. Wei, D. Duvenaud, J. Hernandez-Lobato, "Automatic chemical design using a data-driven continuous representation of molecules", Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018.

[8]   S. Joplin, "School of ragtime: 6 Exercises for Piano", 1908.

This book presents a collection of selected papers that present the current variety of all aspect of the research at a high level, in the fields of music, education and computer science. The book meets the growing demand of practitioners, researchers, scientists, educators and students for a comprehensive introduction to key topics in these fields. The volume focuses on easy-to-understand examples and a guide to additional literature.

Michele Della Ventura, editor

New Music Concepts, Inspired Education, Computer Science

Revised Selected Papers