Developing a Measure Image and Applying It to Deep Learning

Lucas Francesco Piccioni Costa
Universidade Tecnológica Federal do Paraná
Brazil
luccos@alunos.utfpr.edu.br

André Yoshio Caram Ogoshi
Universidade Estadual do Centro-Oeste
Brazil
a.yoshio@hotmail.com

Marcella Scoczynski Ribeiro Martins
Universidade Tecnológica Federal do Paraná
Brazil
marcella@utfpr.edu.br

Hugo Valadares Siqueira
Universidade Tecnológica Federal do Paraná
Brazil
hugosiqueira@utfpr.edu.br

Abstract

The use of intelligent systems linked to musical tasks such as automatic composition, classification, and Music Information Retrieval has increasingly shown itself to be a promising field of study, not only from a computational, but also from a musical point of view. This paper aims to develop an innovative method capable of producing a coded image that contains all the information of a musical measure, generating a structure that can be used in several computational applications involving machine learning, especially deep learning and convolutional neural networks (CNNs). To illustrate the usefulness of this method, the measure image is applied to a CNN to solve the problem of automatic musical harmonization. This brief application achieves better results than those known in the literature, demonstrating the method's effectiveness.

Introduction

The art of music undergoes constant changes, sometimes appropriating science to achieve its transformations. Music and computing can be considered indivisible through the generation of new sounds and even acting in composition (Webster, 2002).

But this does not mean that computers can understand music: the human element is fundamental in this process. Specifically, when the subject is automatic musical computation, it is necessary to encode musical information, often wholly represented by a musical score. It is natural that in the encoding process, some information is lost, which can negatively affect the performance of automated systems.

The most common ways to perform this encoding process involve numeric vectors (Franklin, 2006; Laden & Keefe, 1989; Mozer, 1994; Todd, 1989). Other authors use coding

approaches with image representations, such as Velarde et al. (2016) and Modrzejewski et al. (2019).

This paper aims to present a standardized way to elaborate musical visual representations focusing on intelligent systems applications. The analysis space is limited to one musical measure at a time. First, the measure has its rhythmic information standardized. Then, an image capable of containing all melodic and rhythmic information of that measure is built, regardless of the time signature and tempo. To illustrate the method's usefulness, we apply the Measure Image (MI) in the automatic harmonization task. For a melodic input, defining the best chord to harmonize it is necessary. We obtain better results than those found in the literature, using a CNN with simple architecture.

1 Standardization of Rhythmic Information

The length of a musical measure is governed by a time signature, which indicates the number and type of rhythmic figure that fills it. There are numerous possible time signatures, which causes a problem, as this way, songs with different time signatures will have different lengths, highlighting the need for standardization.

By default, the duration s of the semibreve is equal to 1, and the other durations are fractions of it. The fraction is directly related to the denominator number of the time signature TS. The numerator, in turn, indicates the total of that figure type that will complete a measure.

Knowing this, one can define the durations vector of the k rhythmic figures of a measure as in Equation 1:

$$\dot{\mathbf{d}} = \left[\dot{d}_1, \dot{d}_2, \ldots, \dot{d}_k\right]^T \mid \dot{d}_i = \frac{s}{d_i},\ d_i \geq 1,\ \sum_{i=1}^{k} \dot{d}_i = s.TS. \tag{1}$$

To standardize measure durations independent of TS, we want to define the normalized durations vector. To do so, we perform the scalar multiplication of TS^{-1} by $\dot{\mathbf{d}}$ forming the vector $\dot{\mathbf{d}}^{(\dot{\mathrm{N}})}$, that is, $TS^{-1}.\dot{\mathbf{d}} = \dot{\mathbf{d}}^{(\dot{\mathrm{N}})}$, which has properties according to Equation 2:

$$\dot{\mathbf{d}}^{(\dot{\mathrm{N}})} = \left[\dot{d}_1^{(\dot{\mathrm{N}})}, \dot{d}_2^{(\dot{\mathrm{N}})}, \ldots, \dot{d}_k^{(\dot{\mathrm{N}})}\right]^T \mid \dot{d}_i^{(\dot{\mathrm{N}})} = \frac{s}{\dot{d}_i^{(\dot{\mathrm{N}})}},\ d_i^{(\mathrm{N})} = d_i.TS,\ \sum_{i=1}^{k} \dot{d}_i^{(\dot{\mathrm{N}})} = s. \tag{2}$$

In possession of this information, it is possible to perform rhythmic standardization. The next step is to encode this information into a MI.

2 Measure Image Construction

A musical measure contains a lot of information, particularly the duration and pitch of each note when observing the melodic context. So that none of this information is lost, we propose building an image capable of containing it.

A three-dimensional matrix $x \times y \times z$ is considered per image, where x and y are coordinates of discrete and finite values called pixels, and dimension z is a composition of RGBA dimensions (Adler et al., 2003; Gonzalez & Woods, 2010).

The visual piano roll approach is applied to compose an image with melodic information, a form of representation that uses horizontal lines to define the duration of notes, and each line

represents a pitch (FL Studio, 2021). Figure 1 demonstrates the construction and meaning of each component of the MI.

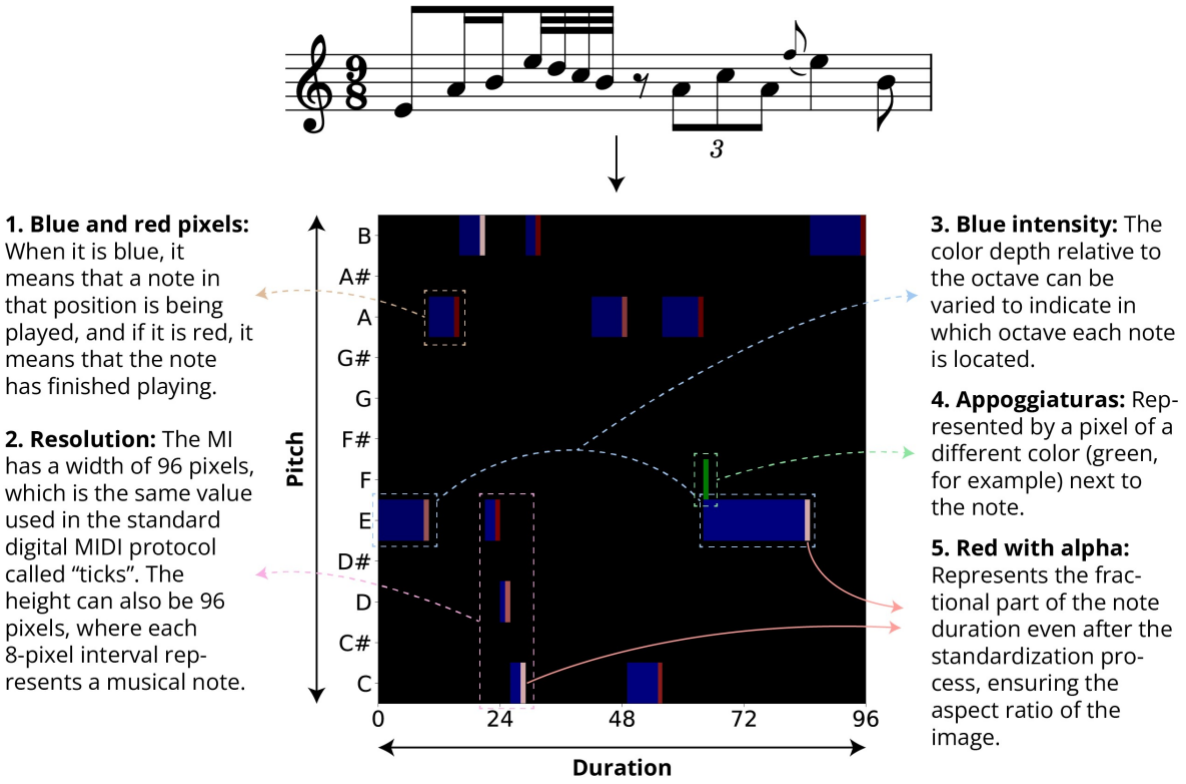

Figure 1: MI originated from a measure with all significant parts explained: (1) how to represent a note duration and pitch, (2) the resolution and size of the image, (3) how to represent different octaves, (4) appoggiaturas, and (5) the use of the alpha channel to store fractional duration information.

3 Applying the Measure Image to Deep Learning

Aiming at applying MI as an input to an automatic harmonization system, we wanted to test a CNN architecture to obtain results using the *CSV Leadsheet Database* (Lim et al., 2017). The database was first standardized in terms of melody (all songs were transposed to the key of C major), harmony (only major and minor triads were considered), and rhythm (applying Equation 2 to the duration of each note). A simple model based on AlexNet (Krizhevsky et al., 2012) was used, as illustrated in Figure 2.

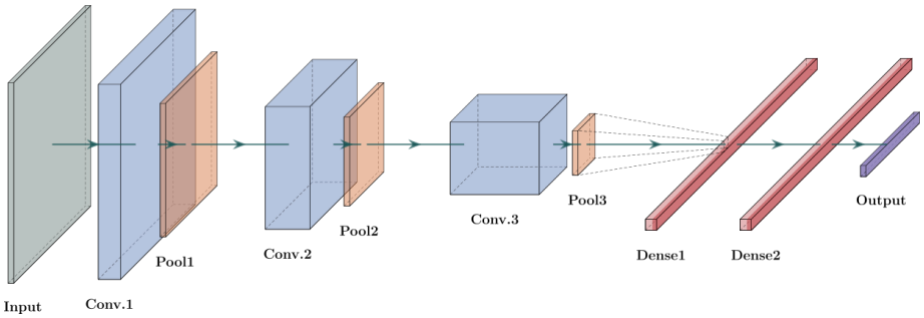

Figure 2: Illustrative structure of the CNN architecture applied during the experiment.

In this model, three layers of convolution (Conv.) and subsampling (Pool) applying the max function, i.e., max-pooling, are interspersed with an increasing number of filters (32, 64, and 128, respectively). The sequence places three fully connected layers (Dense) at the end, where Dense1 and Dense2 have 128 neurons each and 24 neurons at the Output. Each convolution and fully connected layer uses the ELU activation function, excepting the output layer, which performs softmax activation. Dropout and batch normalization techniques were applied.

Only 30% of the database was used, being divided 60% for training, 20% validation, and 20% testing. The model was run 30 times to be cross-validated, resulting in average and best accuracy (ACC) of 50.88% and 52.34%, respectively, and average and best Cohen's Kappa (κ) of 38.31% and 40.37%, respectively.

Based on the subjective levels of κ, the average result can be defined as reasonable, with the best value being just above the lower limit to be considered a moderate or still adequate result (Artstein & Poesio, 2008).

The ACC values allow us to identify an improvement compared to the results in the literature, especially when compared to the work of Lim & Lee (2017), and against a random guess, being equal to 4.17% for the 24 chord classes considered. Analyzing the resulting normalized confusion matrix illustrated in Figure 3 can better explain these statements.

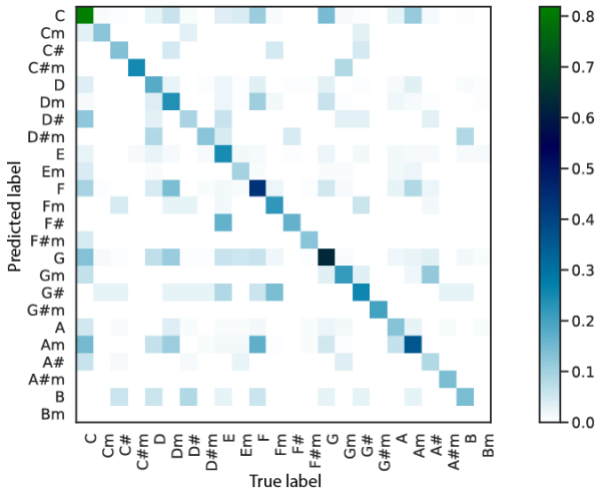

Figure 3: Confusion matrix of the CNN model generated with the results of its best execution.

It is easy to notice the present diagonal line, indicative of the good classification performance of the system. Similarly, there is a greater density of correct classification for the chords of C major, F major, and G major.

The success of this simple model compared to classical techniques in the literature demonstrates the promise of the results obtained by the application of MI using CNNs.

Conclusion

A new way of representing and encoding musical measures was developed, intended for application in intelligent systems. It was put to the test by serving as the input to an automatic harmonizing system that obtained reasonable results considering its simple construction and comparing its results with existing approaches and random baselines.

In the future, we intend to explore more possible applications for MI, such as classification and automatic composition tasks, in addition to the development of a generic application that can be used in development environments for quick and easy use.

Acknowledgments

The authors thank the Brazilian agencies Coordination for the Improvement of Higher Education Personnel (CAPES) - Financing Code 001, Brazilian National Council for Scientific and Technological Development (CNPq), processes number 40558/2018-5, 315298/2020-0, and Araucaria Foundation, process number 51497, and Federal University of Technology - Parana (UTFPR) for their financial support.

References

Adler, M., Boutell, T., Bowler, J., Brunschen, C., Costello, A. M., Crocker, L. D., Dilger, A., Fromme, O., Gailly, J., Herborth, C., Jakulin, A., Kettler, N., Lane, T., Lehmann, A., Lilley, C., Martindale, D., Mortensen, O., Pickens, K. S., Poole, R. P., … Wohl, J. (2003). *Portable Network Graphics (PNG) Specification*. https://www.w3.org/TR/2003/REC-PNG-20031110/

Artstein, R., & Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34(4), 555–596. https://doi.org/10.1162/coli.07-034-R2

FL Studio. (2021). *Piano roll.* https://www.image-line.com/fl-studio-learning/fl-studio-online-manual/html/pianoroll.htm

Franklin, J. A. (2006). Recurrent Neural Networks for Music Computation. *INFORMS Journal on Computing* 18(3), 321–338. https://doi.org/10.1287/ijoc.1050.0131

Gonzalez, R. C., & Woods, R. C. (2010). *Processamento digital de imagens* (3rd ed.). Pearson Prentice Hall.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger

(Eds.), *Advances in Neural Information Processing Systems* 25. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Laden, B., & Keefe, D. H. (1989). The Representation of Pitch in a Neural Net Model of Chord Classification. *Computer Music Journal* 13(4), 12–26. https://doi.org/10.2307/3679550

Lim, H., & Lee, K. (2017). Chord Generation from Symbolic Melody Using BLSTM Networks. *Proceedings of the 18th International Society for Music Information Retrieval Conference*, 621–627. https://doi.org/10.5281/zenodo.1417327

Lim, H., Rhyu, S., & Lee, K. (2017). *CSV Leadsheet Database*. Music and Audio Research Group. http://marg.snu.ac.kr/chord_generation/

Modrzejewski, M., Dorobek, M., & Rokita, P. (2019). Application of Deep Neural Networks to Music Composition Based on MIDI Datasets and Graphical Representation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), LNAI* 11508, 143–152.

https://doi.org/10.1007/978-3-030-20912-4_14

Mozer, M. C. (1994). Neural Network Music Composition by Prediction: Exploring the Benefits of Psychoacoustic Constraints and Multi-scale Processing. *Connection Science* 6(2–3), 247–280. https://doi.org/10.1080/09540099408915726

Todd, P. M. (1989). A Connectionist Approach to Algorithmic Composition. *Computer Music Journal* 13(4), 27. https://doi.org/10.2307/3679551

Velarde, G., Weyde, T., Chacón, C. C., Meredith, D., & Grachten, M. (2016). Composer recognition based on 2D-filtered piano-rolls. *Proceedings of the 17th International Society for Music Information Retrieval Conference (New York City, August 7–11, 2016),* 115–121.

Webster, P. (2002). Historical Perspectives on Technology and Music. *Music Educators Journal* 89(1), 38–43. https://doi.org/10.2307/3399883