What Musical Knowledge Does Self-Attention Learn?

Mikaela Keller, Gabriel Loiseau, Louis Bigo

What Musical Knowledge Does Self-Attention Learn ?

Mikaela Keller[1,2] **Gabriel Loiseau**[2] **Louis Bigo**[2]
[1] Inria
[2] Univ. Lille, CNRS, Centrale Lille
UMR 9189 CRIStAL, F-59000 Lille, France
{mikaela.keller,louis.bigo}@univ-lille.fr

Abstract

Since their conception for NLP tasks in 2017, Transformer neural networks have been increasingly used with compelling results for a variety of symbolic MIR tasks including music analysis, classification and generation. Although the concept of self-attention between words in text can intuitively be transposed as a relation between musical objects such as notes or chords in a score, it remains relatively unknown what kind of musical relations precisely tend to be captured by self attention mechanisms when applied to musical data. Moreover, the principle of self-attention has been elaborated in NLP to help model the "meaning" of a sentence while in the musical domain this concept appears to be more subjective. In this explorative work, we open the music transformer black box looking to identify which aspects of music are actually learnt by the self-attention mechanism. We apply this approach to two MIR probing tasks : composer classification and cadence identification.

1 Introduction

The Transformer (Vaswani et al., 2017) is a neural network architecture based on the self-attention mechanism that was designed for sequence prediction tasks (machine translation, syntactic parsing, etc.) in NLP. Subsequently, the self-attention principle has also been applied with success to improve MIR tasks including harmony analysis (Chen and Su, 2021) and generation with long-term coherence as demonstrated with Music Transformer (Huang et al., 2018b). The Music Transformer model has then inspired various researches including the generation of pop music (Huang and Yang, 2020) and guitar tablature (Chen et al., 2020).

Despite its increasing use in MIR tasks, the nature of the musical knowledge learned by Transformers is rarely studied. (Huang et al., 2018a) proposes a tool to visualise self-attention weights associated to a musical extract but without any systematic analysis. Inspired by NLP literature(Conneau et al., 2018; Coenen et al., 2019; Tenney et al., 2019; Manning et al., 2020) our work aims at opening the Music Transformer *black box* in order to extract its abstract representation of musical sequences and submit those representations to two selected MIR "probing" tasks : composer classification and cadence detection.

The self-attention mechanism is encoded within a transformer through matrices of coefficients, produced by *attention heads*, which are distributed in the subsequent layers of the network. Given a sequence of tokens $x_1, \ldots, x_T$ an attention head produces an attention matrix $A = (a_{ij})_{1 \leq i,j \leq T}$ where a_{ij} encodes "the attention that token x_i gives to token x_j" or the weight that x_j is going to play in in the next layer representation of x_i. The goal of our study[1] consists in identifying the musical knowledge that is encoded within these matrices in a trained Transformer. For this purpose we designed two "probing" datasets of musical sequences labeled with informations that were not explicitly available to the Transformer during training. The first dataset is labeled by the composer of the sequence. In the second dataset the sequences are characterized as containing a cadence (musical phrase ending) or not.

In the following we show, that a simple linear classifier fed with isolated attention matrices is able to discriminate between two composers when their styles are different enough. In contrast, an analogous experiment shows that marks of structural phenomena such as cadences appear more challenging to detect in attention matrices.

In the second part of our study, we examine attention values in order to gain insights into the

[1]Code avaliable at https://github.com/Music-NLP/MusicalSelfAttention

classification results. Our observations reveal various orientations (past or future) of attention spans among composers, as well as prominent attention values on theoretic cadence preparation points.

2 Attention Based Sequence Representation

In this work, the Music Transformer is used as a representation tool, to compute self-attention relations for any arbitrary musical sequence.

The MAESTRO dataset is used in this study to train the Music Transformer. This dataset gathers 1276 piano performances of pieces composed by 54 major composers of different styles, including Bach, Mozart, Beethoven or Debussy. In order to be compatible with the Transformer input format, the MAESTRO dataset is converted into sequences of tokens following the syntax proposed by (Huang et al., 2018b). This token representation includes `NOTE ON`, `NOTE OFF`, `TIME SHIFT`, and `VELOCITY` types. In this study, we trained[2] a Music Transformer neural network on this corpus as explained in (Huang et al., 2018b).

The Transformer architecture trained for this study includes 6 layers, each composed of 4 attention heads. Given an input sequence of T elements an attention head produces a square real-valued attention matrix $X = (x_{ij})$ of dimension $T \times T$. The value x_{ij} is usually interpreted as the attention that the elements at position i has for the element at position j. Once the transformer is trained, it has the ability to systematically abstract any musical sequence of size T by a set of $6 \times 4 = 24$ attention matrices of size $T \times T$. Through probing tasks NLP literature (Tenney et al., 2019; Manning et al., 2020) has reported that lower attention heads seem to attend to lower level abstractions, such as syntactic parsing, while deeper layers attend to higher level abstract such as coreference resolution. Assuming that some of this knowledge is transferable to the musical domain we have chosen to focus on the deeper layer of the network for representing the sequences in our MIR inspired probing tasks. We have chosen to collapse the 4 attention matrices produced by the last layer into an average matrice, and to use these $T \times T$ coefficients as the input to the classification tasks that we define in the next section[3]. Figure 1 illustrates this pipeline.

3 Agnostic Probing Tasks

In this section, we describe two probing tasks that aim at highlighting the musical knowledge encoded in attention values computed by the Music Transformer. The first task is a composer classification and the second one is cadence detection. Both tasks are formulated as supervised binary classification performed on the attention matrices described in section 2.

3.1 Composer Identification

We evaluate the ability of learned attention representations to model musical style through a composer identification task.

We used a subset of the MAESTRO dataset that contains unique composer performances to create several binary classification tasks `composer1` vs `composer2`. To better highlight the ability of attention values to capture stylistic information, we deliberately selected composers that are known to be close in term of style, such as Haydn and Mozart, and far apart, such as Bach and Chopin.

For each couple, a set of training musical sequences of fixed size are abstractly represented as attention matrices (see Section 2). The training sets are balanced and contain 2648 sequences from each of the composers. The corresponding abstract representations are then given as input to a logistic regression classifier with $l2$-regularization that is trained to assign composer authorship to any input attention matrix. The experiment is repeated 5 times, sampling various training sets for every couple of composers and for various sizes of sequences. Figure 2 displays the average performance of the classifiers over a separate and fixed test set[4] of 426*2 sequences. A random classifier is here expected to have a 50% accuracy.

Low standard deviations, illustrated by vertical lines on each experiment, show that given a couple of composers the accuracy is quite stable with respect to the various training sets. Figure 2 also shows that the accuracy generally tends to increase with the size of the sequences (which was not obvious since when increasing the size of the sequence we increase quadratically the search space number of dimensions without increasing

[2] Using the implementation in https://github.com/jason9693/MusicTransformer-tensorflow2.0

[3] Although probing tasks are often performed on other outputs of the transformer, limiting the transformation of attention values facilitates their musical interpretation in this work.

[4] We used MAESTRO train/test split to insure that a same piece could not appear both in the train and the test set

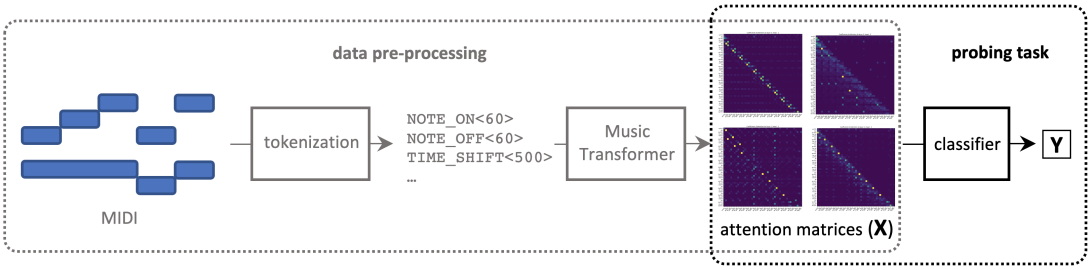

Figure 1: Pipeline used for the two probing tasks. The left part illustrates the systematic representation of a midi sequence into a set of self-attention values thanks to the Music Transformer. The right part illustrates how a probing task is formulated as a classification problem on attention values.

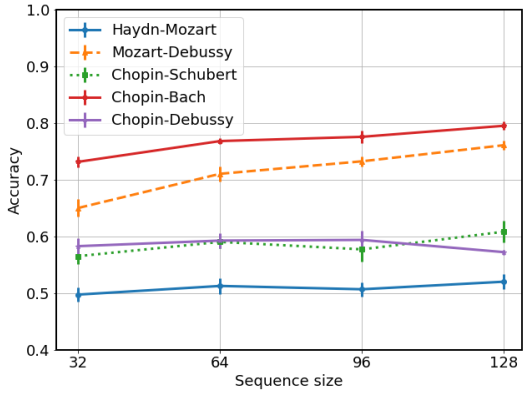

Figure 2: Mean accuracy of composer identification on attention matrices computed from sequences of various lengths.

the number of examples). The difficulty of the classification task of a pair of composers certainly relates to how they differ in style. Interestingly, by using birth date gaps as rough proxy for style differences, the accuracies appears to match the difficulty of the tasks[5].

3.2 Cadence Detection

Cadences are structural breaks widely used in the classical repertoire to emphasize the end of a musical phrase. Cadence are often associated with a closure feeling that resolves a tension region (Blombach, 1987). This concept therefore appears as a promising candidate to validate the principle of self-attention in music as the short past that precedes a cadence is supposed to be organized in close relation with the upcoming cadence. This short past is sometime referred to as the *preparation* of the cadence.

The present task consists in evaluating how much the attention values encode the presence of a cadence. Our hypothesis is that cadential points and preparation points should have important mutual attention one for each other if they appear concomitantly within the training set. Attention matrices are computed as explained in section 2 through a Transformer which is trained on the MAESTRO corpus. Given the pieces of music present in the MAESTRO dataset, it can reasonably be hypothesized that cadences, that are typical of the classical era, are sufficiently represented in the training set to be modeled by the Transformer. Similarly to the composer identification task, a set of attention matrices, that represent musical sequences with and without cadences, are used to train a logistic regression classifier. For this purpose, we use a dataset of 24 fugues from J.-S. Bach with cadence annotation (Giraud et al., 2015). A set of 3864 sequences of 64 tokens is sampled from the fugue dataset, a third of which include a cadence[6] while the remaining do not include any cadence. We use a leave-one-piece-out strategy to evaluate the performance of the cadence classification and compare it to a random classification on each fold of the cross-validation. The micro-averaged F1 score of the cadence classifiers is 0.458 as compared to 0.315 for the random classifier. This results seems to suggest that attention values learned by the Transformer do encode some information about the notion of cadence.

3.3 Discussion

Cadences belong to high level elements of tonal musical language. Despite their unified closure meaning, they can be realized through a large vari-

[5] Birth date gaps (in years) : Chopin-Bach: 125, Debussy-Mozart: 106, Debussy-Chopin: 52, Mozart-Haydn: 24 and Chopin-Schubert: 13

[6] A same cadence can appear several time in our dataset but at different positions and necessarily in the 2nd half of the sequence in order to favor the inclusion of the *preparation* of the cadence within the sequence.

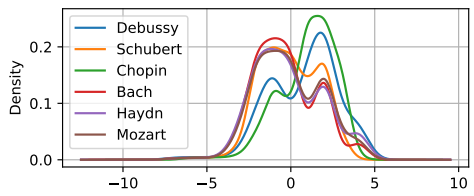

Figure 3: Distribution of the attention spans. The horizontal axis shows the length of attention spans (in tokens).

ety of musical surfaces, which makes their modeling particularly complex (Bigo et al., 2018). Musical style, on the other hand, can refer to lower level relationships between musical objects, like pitch intervals. It is therefore interesting to observe that our attention based classification approach give results better than chance both on style modelling and on cadence detection.

4 Musical Interpretation of Self-Attention Relations

In this section, we provide a few exploratory analysis to gain musical insights on the data that was given in input to the probing classifiers.

4.1 How Do Transformers Learn About Composers ?

As explained in section 2, the composer discrimination probing task was performed using an average attention matrix A_x computed from each sequence x. We averaged A_x for each composer over a subset of 1000 sequences used for training the linear classifiers. The sequence are of fixed size ($T = 64$). The result is a matrix $M = (m_{ij})_{1 \leq i,j \leq T}$ where m_{ij} is the average attention that the i^{th} token gives to the j^{th} token in the sequences of a given composer. We consider that a token at position i "looks at" a token in position j, *ie* it has an attention span of at least $i - j$, if the coefficient m_{ij} is greater than a certain threshold. In Figure 3 we report the distribution of attention spans for a threshold of 0.04 ($\approx 7\% - 10\%$ of coefficients) for several composers.

The figure shows that the learned attention span rarely exceeds five tokens in the past or in the future. Interestingly, the attention learned on early composers such as Bach, Haydn, Mozart, and Schubert seem to focus *towards* tokens in the short past. In contrast, Chopin and Debussy attention is turned *towards* tokens in the short future, which might be partly related to a stylistic rupture of the composers with the classical era. Confirming this hypothesis would require a deeper study.

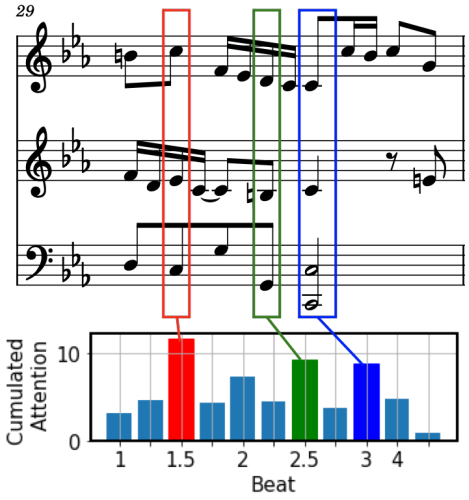

Figure 4: Cumulated attention on successive offsets of bar 29 of Fugue 2 of the *Well-Tempered Clavier* from Bach. A perfect authentic cadence is annotated on beat 3 (blue frame). Other points of prominent attention (red and green) correspond to important *preparation* points of the cadence.

4.2 How Do Transformers Learn About Cadences ?

In this experiment we observe the information within the attention matrix A_x of a sequence containing a cadence. The sequence can be divided into `TIME SHIFT` events that can be aligned with the beat pulse of the piece extract. Figure 4 shows the cumulated attention between `TIME SHIFT` events in regard with the sheet music.

5 Conclusions and Perspectives

We proposed in this work an original approach to improve our understanding of the musical knowledge that the self-attention mechanism can learn. In spite of instructive results, these experiments highlight the difficulty to interpret neural values within a multi layer model but also confirm the necessity to pursue our efforts in that quest of comprehension of music deep learning models.

Futur works include experimenting with other probing tasks, such as harmony and tonality analysis, in order to better understand how Transformer architectures learn these high level concepts. It could also be interesting to test those tasks on different layers of the network to see if there is a gradation in the information levels of abstraction.

Acknowledgments

The authors are grateful to the Algomus and Magnet teams and to Kamil Akesbi for fruitful discussions. This work is supported by a special interdisciplinary funding (AIT) from the CRIStAL laboratory.

References

Louis Bigo, Laurent Feisthauer, Mathieu Giraud, and Florence Levé. 2018. Relevance of musical features for cadence detection. In *International Society for Music Information Retrieval Conference (ISMIR 2018)*.

Ann Blombach. 1987. Phrase and cadence: A study of terminology and definition. *Journal of Music Theory Pedagogy*, 1(2):225–251.

Tsung-Ping Chen and Li Su. 2021. Attend to chords: Improving harmonic analysis of symbolic music using transformer-based models. *Transactions of the International Society for Music Information Retrieval*, 4(1).

Yu-Hua Chen, Yu-Hsiang Huang, Wen-Yi Hsiao, and Yi-Hsuan Yang. 2020. Automatic composition of guitar tabs by transformers and groove modeling. *arXiv preprint arXiv:2008.01431*.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of bert. *arXiv preprint arXiv:1906.02715*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Mathieu Giraud, Richard Groult, Emmanuel Leguy, and Florence Levé. 2015. Computational fugue analysis. *Computer Music Journal*, 39(2):77–96.

Anna Huang, Monica Dinculescu, Ashish Vaswani, and Douglas Eck. 2018a. Visualizing music self-attention.

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. 2018b. Music transformer. *arXiv preprint arXiv:1809.04281*.

Yu-Siang Huang and Yi-Hsuan Yang. 2020. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1180–1188.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.