Modelling Interval Relations in Neural Music Language Models

Radha Kopparti [1] and **Tillman Weyde** [1]

Abstract

In this study, we explore the use of modelling of pitch intervals and interval relations in pitch with neural networks. Intervals and their relations are essential features of music, but in neural networks, the trend is to use raw data as input and not to model any higher level aspects of the music. We propose to use Relation Based Patterns (RBP) to integrate intervals (early and mid fusion) and interval relations (late fusion) into the network structure. We observe significant improvements in pitch prediction for the Essen Folk Song Collection for RBP over standard networks, and for mixed over unsigned and signed interval representation.

1. Introduction

Pitch relations in the form of melodic intervals are an essential feature of music, especially the structural understanding of melody. Popular deep neural networks are typically used with raw data and network structures are designed to be generic and not specialised for a specific task. However, small datasets, as common in symbolic music, and abstract relations like repetitions or intervals are conditional where neural networks fail to work well (Lake & Baroni, 2018; Marcus, 2018)

In this work we extend the approach of designing networks structures that facilitate the learning of abstract relations as in (Weyde & Kopparti, 2018) and integrating them with standard neural networks, in order to benefit from the pattern recognition capabilities of neural networks and knowledge based modelling.

The modelling of intervals in sequence prediction has been done in earlier music prediction models since the 1990s in (Conklin & Witten, 1995; Dubnov, 1998; Pearce & Wiggins, 2004; 2012; Paiement et al., 2009; Paiement & Yves Grandvalet, 2009). Related approaches have also been used in cognitive modelling music learning (Saffran, 1999; Rohrmeier & Rebuschat, 2012). In these models pitch intervals are defined between adjacent notes. There are more general approaches, called combinatorial metrics by (Polansky, 1996), but these have to our knowledge not been used in pitch prediction models.

Connectionist models have been modelling per-note features (Cherla et al., 2013; 2015) or more recently have focused on relations between subsequences rather than notes within sequences (Lattner et al., 2016; 2018). Skip-grams do model non-adjacent relations as in (Sears et al., 2017; Herremans & Chuan, 2017), but they haven't been used to model the relationships between pitch intervals.

In this work, we use Relation Based Patterns (RBP) (Weyde & Kopparti, 2018) for modelling abstract interval relations patterns within the neural network structure. In this work, we used monophonic folk melodies (Schaffrath & Huron, 1995) to test the RBP models performance with different input representations.

2. Interval Modelling

In RBP, DR units act as repetition detectors by comparing every pair of input values using the absolute of the difference of the inputs in one-hot encoding (Weyde & Kopparti, 2018; Kopparti & Weyde, 2018). However, integer encoding as MIDI pitch values is more efficient and encodes the structure of higher and lower pitches. With integer encoding of pitch, the DR units represent interval size. We now propose D (difference) units, with activation $f(x, y) = y - x$. We integrate D units into LSTM networks in Early, Mid and Late Fusion settings (Weyde & Kopparti, 2018). Early and Mid Fusion involves intervals within the context as D(R) units concatenated with the input (early) or first hidden layer (mid fusion). In one-hot encoding, there is a D(R) unit for every pitch value for every pair of notes. In integer encoding, there is one D(R) unit per pair of notes.

Late fusion involves mapping from intervals within the context to intervals between context notes and predicted note, in parallel to a standard network. In this case, the output contains for every note in the input context, a probability

[1]Research Centre for Machine Learning, Department of Computer Science, City University of London, UK . Correspondence to: Radha Kopparti <radha.kopparti@city.ac.uk>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

distribution over the possible intervals between each context note and the predicted note. These distributions are added, normalised, mapped back to the output pitch space and averaged with the normal neural network prediction using a trainable weighting. See (Weyde & Kopparti, 2018) for a more detailed description.

3. Results

We compare performance of both forms of encoding and note the overall performance of the LSTM model with and without RBP below. We also tested standard RNN and GRU networks, but do not report the results as they performed similarly but consistently worse than LSTM networks.

The range of context lengths is [2,3,4,5,6,7,8,9], following the idea that human short term memory stores up to 9 items of information (Miller, 1956). The number of hidden units and number of epochs is set to 20 and 30 after performing a grid search over [10,20,30,50] for each parameter. We use 2 hidden layers in all the neural networks, as this gave the best results out of [1,2,3]. We used ADAM optimisation with a learning rate of 0.01. The loss function and evaluation metric is cross entropy C between the original distribution p and predicted distribution q, defined as

$$C(p, q) = -\Sigma_{x \in S}\, p(x) \log q(x) \qquad (1)$$

where S is the set of possible events, i.e. pitches.

Table 1 and 2 give the overall performance of the models for different context lengths using one-hot and integer encoding with D units in RBP Early, Mid and Late fusion.

Context Length	Without RBP	With RBP		
		Early Fus	Mid Fus	Late Fus
n= 2	2.9213	2.8534	2.8523	2.8056
n= 3	2.8932	2.8456	2.8413	2.8023
n= 4	2.8906	2.8502	2.8478	2.7959
n= 5	2.8712	2.8478	2.8432	2.7922
n= 6	2.8676	2.8353	2.8321	2.7862
n= 7	2.8612	2.8324	2.8236	2.7812
n= 8	2.8527	2.8255	2.8224	2.7621
n= 9	2.8514	2.8124	2.8105	2.7539

Table 1. Average cross entropy with different RBP variants for various context lengths *n* using one-hot encoding.

The overall performance of the one-hot encoding with D units is worse than with integer encoding for all the RBP variants and worse than previous results with DR units, that we have not reported here for space reasons. Late fusion performs better than early and mid in all cases. All differences are significant with $p < .05$ in a Wilcoxon signed rank test over the context length. We also see that models with greater context lengths perform better.

Context Length	Without RBP	With RBP		
		Early Fus	Mid Fus	Late Fus
n= 2	2.8512	2.7862	2.7623	2.7259
n= 3	2.8503	2.7812	2.7689	2.7214
n= 4	2.8467	2.7734	2.7632	2.7209
n= 5	2.8231	2.7423	2.7402	2.7062
n= 6	2.8123	2.7384	2.7362	2.6927
n= 7	2.7867	2.6925	2.6903	2.6767
n= 8	2.7834	2.6916	2.6843	2.6621
n= 9	2.7657	2.6826	2.6732	2.6527

Table 2. Average cross entropy with different variants of RBP various context lengths *n* with integer encoding.

We also evaluated the model with a concatenation of DR and D units (unsigned and signed interval representation). The cross entropy results of D and DR units and their combination in late fusion are given in table 3. The combined D/DR units perform best in terms of cross entropy (significantly) and in prediction accuracy, where the performance is 29%, 33% and 35% for $n = 9$ respectively.

Context Length	RBP Late Fusion		
	D units	DR units	DR and D units
n= 2	2.7259	2.6232	2.5984
n= 3	2.7214	2.6254	2.5925
n= 4	2.7209	2.6232	2.5864
n= 5	2.7062	2.6065	2.5802
n= 6	2.6927	2.5878	2.5724
n= 7	2.6767	2.5957	2.5714
n= 8	2.6621	2.5868	2.5654
n= 9	2.6527	2.5927	2.5658

Table 3. Average cross entropy with D units, DR units and DR and D units combined for RBP in Late Fusion for various context lengths *n*.

4. Conclusions

Integration of interval representations into neural music language models improves pitch prediction. We find that integer encoding of pitch is more effective and efficient than one-hot encoding. Signed and unsigned interval representations with DR and D units are effective, at most when combined. Late fusion, which models the relations between intervals within the context and with the predicted note is consistently more effective than early and mid fusion that uses interval information only for input features that are used directly to predict pitch. The approach of modelling intervals in the network structure is overall successful and a motivation for designing networks for other tasks, such as modelling musical rhythm and dynamics or linguistic tasks like word prediction based on embeddings.

References

Cherla, S., Weyde, T., Garcez, A., and Pearce, M. A distributed model for multiple viewpoint melodic prediction. *International Society for Music Information Retrieval Conference*, pp. 15–20, 2013.

Cherla, S., Weyde, T., Garcez, A., and Tran, S. N. Hybrid long-and short-term models of folk melodies. *International Society for Music Information Retrieval Conference*, pp. 584–590, 2015.

Conklin, D. and Witten, I. H. Multiple viewpoint systems for music prediction. *Journal of New Music Research*, (24.1):51–74, 1995.

Dubnov, S., A. G. . E.-Y. R. Universal classification applied to musical sequences. *In Proceedings of the 1998 International Computer Music Conference, San Francisco*, pp. 332–340, 1998.

Herremans, D. and Chuan, C.-H. Modeling musical context with word2vec. *arXiv preprint arXiv:1706.09088*, 2017.

Kopparti, R. and Weyde, T. Evaluating repetition based melody prediction over different context lengths. *ICML Joint Workshop on Music and Machine Learning*, 2018.

Lake, B. and Baroni, M. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, pp. 2879–2888, 2018.

Lattner, S., Grachten, M., and Widmer, G. Imposing higher-level structure in polyphonic music generation using convolutional restricted boltzmann machines and constraints. *arXiv preprint arXiv:1612.04742*, 2016.

Lattner, S., Grachten, M., and Widmer, G. A predictive model for music based on learned interval representations. In Gómez, E., Hu, X., Humphrey, E., and Benetos, E. (eds.), *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, pp. 26–33, 2018. ISBN 978-2-9540351-2-3. URL http://ismir2018.ircam.fr/doc/pdfs/179_Paper.pdf.

Marcus, G. F. Deep learning: a critical appraisal. *arXiv:1801.00631*, 2018.

Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

Paiement, J.-F. and Yves Grandvalet, a. S. B. Predictive models for music. *Connection Science*, (21.2-3):253–272, 2009.

Paiement, J.-F., Bengio, S., and Eck, D. Probabilistic models for melodic prediction. *Artificial Intelligence*, (173.14): 1266–1274, 2009.

Pearce, M. and Wiggins, G. Improved methods for statistical modelling of monophonic music. *Journal of New Music Research*, 33(4):367–385, 2004.

Pearce, M. T. and Wiggins, G. A. Auditory expectation: The information dynamics of music perception and cognition. *Topics in cognitive science*, (4.4):625–652, 2012.

Polansky, L. Morphological metrics. *Journal of New Music Research*, 25(4):289–368, 1996.

Rohrmeier, M. and Rebuschat, P. Implicit learning and acquisition of music. *Topics in Cognitive Science*, (4.4): 525–553, 2012.

Saffran, J. R. Statistical learning of tone sequences by human infants and adult. *Cognition*, (70.1):27–52, 1999.

Schaffrath, H. and Huron, D. *The Essen Folksong Collection in the Humdrum Kern Format*, 1995.

Sears, D. R. W., Arzt, A., Frostel, H., Sonnleitner, R., and Widmer, G. Modeling harmony with skip-grams. In Cunningham, S. J., Duan, Z., Hu, X., and Turnbull, D. (eds.), *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, pp. 332–338, 2017. ISBN 978-981-11-5179-8. URL https://ismir2017.smcnus.org/wp-content/uploads/2017/10/18_Paper.pdf.

Weyde, T. and Kopparti, R. Modeling identity rules with neural networks. *https://arxiv.org/abs/1812.02616*, 2018.