A Neural Network for Predicting Musical Pitch

Srikanth Cherla, Artur Garcez, Tillman Weyde
{abfb145,a.garcez,t.e.weyde}@city.ac.uk
Music Informatics Research Group & Machine Learning Group
Department of Computer Science

CITY UNIVERSITY LONDON

Machine Learning and Music

Machine Learning aims to develop models that can generalize over data through examples. Models realized in this manner can be used to classify, and make predictions about new data presented to them. Machine Learning techniques have been successfully applied in the analysis of text, video, audio and various other sensor data. The increasing availability these days, of music in various digital formats, and online, has made it feasible to apply these techniques to extract information directly from audio and symbolic music data. The area of ***Music Information Retrieval*** (MIR) deals with such tasks, and applies this knowledge to automatic music recommendation, transcription, classification, generation, etc. In the present work, we are interested in the ***analysis of sequential patterns in music***, which influence our perception of musical style, the emotions we associate with music, and our notions of similarity in music. We are also interested in ***automatic music generation*** with these models.

Music, Language & Neural Networks

Both language and music have a well-defined temporal structure. Furthermore, tonal music can be expressed using sets of musical symbols, much in the same way as language with linguistic characters. Between the two, language has been more extensively studied in Computer Science due to its relatively early digitization, and its application to the analysis of content in documents. From the MIR point-of-view, this motivates ***the extension of computational models that have been successful at language modelling,*** combined with appropriate domain knowledge, ***to music*** [3]. This is challenging because music results from the interaction between multiple musical facets, namely timbre, rhythm, melody, and harmony, which allows for a large number of possible variations in how these are combined for its creation.

In the past decade, there has been an increasing interest in the use of ***neural networks*** for modelling language. It was found that through the distributed representations that these models learn, they are able to ***efficiently capture information*** in longer word-sequences, and ***learn language semantics*** in addition to only the syntax. One can also often explain their behaviour by appropriately interpreting the learned representations. These desirable features help them ***overcome*** some of the limitations pertaining to the ***curse of dimensionality*** of the more widely used Markov models [1]. We wish to employ these desirable features of neural networks in the analysis of musical pitch sequences.

Restricted Boltzmann Machines

The ***Restricted Boltzmann Machine*** (RBM) is an undirected, bipartite neural network consisting of a *visible* and a *hidden* layer. The two layers are fully connected to each other, but there exist no connections between any two hidden units, or any two visible units. In a typical application, the visible layer would represent data, and the hidden layer, the features learned from the data. Figure 1 illustrates the essential structure of the RBM.

Although this model was first introduced in 1986 [4], it received only limited attention as it had a computationally expensive learning algorithm. This changed in 2002, with the discovery of a tractable approximation of the original algorithm, known as ***Contrastive Divergence*** [2]. RBMs have since evolved into other variants, and been increasingly applied to a variety of machine learning tasks. In addition to the above listed advantages of neural networks, multiple RBMs can also be stacked to create a ***deep network*** which can learn representations of the data at multiple levels of abstraction.

Neural Probabilistic Pitch Prediction Model

Our model for pitch prediction is based on a ***discriminative variant of the RBM***, and models the conditional distribution $p(s_t|s_{(t-n+1)}^{(t-1)})$, where s_t is the t^{th} element in a sequence s_1^t of MIDI pitch values. The Markov assumption limits the temporal memory of the model. It is illustrated in Figure 1.The model is trained using Contrastive Divergence learning, on sequences of length n. Given a context of length $(n-1)$, it ***predicts a probability distribution*** over the possible continuation pitches of this context.

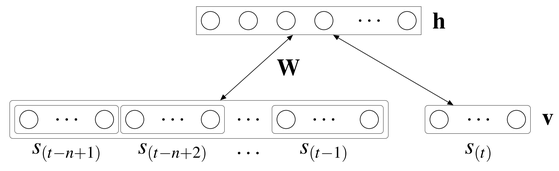

Figure 1: *The prediction model. Each set of nodes in the visible layer is the one-of-m representation of a pitch value. Those sets grouped together to the left make up the context $s_{(t-n+1)}^{(t-1)}$ of length $(n-1)$, and contain $n \times m$ nodes. The set of nodes to the far right correspond to the pitch s_t to be predicted.*

Evaluation & Results

We performed a ***two-fold evaluation***. The first measures the ***information content*** of the prediction model in terms of cross entropy (*cf.* Table 1). In the second evaluation (*cf.* Table 2), an ensemble of these models is put to test in a task to ***classify folk melodies according to origin***.

Model	RBM-2	RBM-3	RBM-4	RBM-5	RBM-6	RBM-7	RBM-8	RBM-9
C. Ent.	3.059	2.894	2.815	2.771	2.737	2.743	2.734	2.751

Table 1: *Variation in cross entropy of the prediction model with sequence length n. Its value progressively decreases until $n = 6$. On the other hand, that of an n-gram model was typically found to saturate at $n = 3$ and increase thereafter.*

The results show that in the first task, there is a progressive increase in predictive performance with context length. The model ***performs competitively to state-of-the-art n-gram models***, with scope for further improvement. It fairs reasonably well in the second task with a ***classification accuracy of*** 59.93%.

	Nova-Scotia	Alsace	Yugoslavia	Switzerland	Austria	Germany	China
Nova-Scotia	74.83%	5.96%	0.00%	2.65%	4.63%	4.63%	7.28%
Alsace	13.19%	29.67%	7.69%	10.99%	19.78%	16.48%	2.19%
Yugoslavia	10.08%	10.08%	42.02%	14.29%	14.29%	8.40%	0.84%
Switzerland	6.45%	17.20%	6.45%	35.48%	22.58%	7.53%	4.30%
Austria	10.58%	19.23%	4.81%	12.50%	41.35%	9.62%	1.92%
Germany	11.27%	11.74%	2.35%	9.86%	8.92%	53.99%	1.88%
China	6.75%	1.27%	0.00%	0.84%	3.80%	0.00%	87.34%

Table 2: *Confusion matrix for the classification task, illustrating the poor classification of melodies from geographically close regions of Europe.*

Conclusions & Future Work

The present work demonstrates the use of Discriminative RBMs to learn sequences of musical pitch, and their application to folk melody classification. While these initial results are encouraging, and comparable to the state-of-the-art in the area, there is room for further improvement through the ***use of improved variants of the basic RBM*** employed here. We would also like to consider other interesting applications of the models. Of immediate interest is the ***analysis of the features learned*** by the RBM in the hidden units.

Acknowledgements: Thanks to Dr. Marcus Pearce for valuable feedback at different stages of this work, and Son Tran for several useful discussions on RBMs and machine learning. Srikanth Cherla's work is funded by a Ph.D. studentship from City University London.

References

[1] Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A Neural Probabilistic Language Model. Journal of Machine Learning Research 3, 1137–1155 (2003)

[2] Hinton, G.E.: Training products of experts by minimizing contrastive divergence. Neural computation 14(8), 1771–1800 (2002)

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

[4] Smolensky, P.: Parallel distributed processing: explorations in the microstructure of cognition, vol. 1. chap. Information processing in dynamical systems: foundations of harmony theory, pp. 194–281. MIT Press (1986)