Musical Metacreation: Papers from the 2012 AIIDE Workshop
AAAI Technical Report WS-12-16

Improvising Musical Structure with Hierarchical Neural Nets

Benjamin D. Smith
Case Western Reserve University
Cleveland Institute of Art
Cleveland, Ohio

Guy E. Garnett
Illinois Informatics Institute
University of Illinois at Urbana-Champaign
Urbana, Illinois

Abstract

Neural networks and recurrent neural networks have been employed to learn, generalize, and generate musical examples and pieces. Yet, these models typically suffer from an inability to characterize and reproduce the long-term dependencies of musical structure, resulting in products that seem to wander aimlessly. We describe and examine three novel hierarchical models that explicitly operate on multiple structural levels. A three layer model is presented, then a weighting policy is added with two different methods of control attempting to maximize global network learning. While the results do not have sufficient structure beyond the phrase or section level, they do evince autonomous generation of recognizable medium-level structures.

Introduction

Creativity, and more specifically musical creativity, is difficult to quantify and subject to differences in experience, culture, goals, and other varied individual contexts. We therefore have focused our attention on novelty as a slightly easier path to understanding machine creativity. Furthermore, we take inspiration from cognitive models that imply a strong hierarchical structure underlies diverse brain activities such as pattern matching and associative memory. We have created and implemented a simple model of musical novelty that incorporates both of these features, such that complex notions of novelty emerge in higher hierarchical levels derived from simpler notions at the lower hierarchical levels. The lower levels are in turn influenced by feedback from higher levels, choosing local details that attempt to maximize, in some sense, the novelty at all levels rather than merely local levels.

Neural networks (NN), and particularly recurrent neural networks (RNN) (Eck and Schmidhuber 2002), while capable of learning note-to-note dependencies from given musical examples and generating new sequences as a result, suffer from a host of problems relating to the "vanishing gradients" problem (Hochreiter et al. 2001) in RNN. Conventional gradient learning methods error flow either quickly vanishes or explodes, resulting in an inability to robustly store past information about inputs. In music, long-term dependencies and relationships are definitive of formal structure, style, and phrasing. Without these long-term dependencies the music can appear aimless and incoherent (Mozer 1999; Smith and Garnett 2012).

We address this deficiency by creating a hierarchical structure of NNs in which each layer evaluates the output of the next lower layer, ideally providing long-term structure for the generated output. Two different models are proposed and examined, one using a directed acyclic graph structure and another adding a feedforward amplification control, with two variations, to push the system out of local minima. Following Schmidhuber (2009), we employ an intrinsically motivated prediction model allowing it to operate within a context of learning-by-doing, rather than by imitating preexisting exemplars, revealing the inherent characteristics of the system.

Model

Given the complexity and difficulty of defining creativity unambiguously, we start with notions of novelty in an immediate context. For a given observer (i.e. human performer or listener) sequences of musical events are often compared using a culturally and individually formed value function to determine the desirability of sequences. This is what we understand as musical preference, resulting in individual affinities for certain kinds of music or artists over others. Given appropriate experience, humans appear to have the ability to judge music, yet the theoretical details of this process of evaluation remain obscure and highly debatable. However, we posit that such a *musical preference* function exists, for the sake of the following formulation.

If such a function exists, at any given point in time within a musical piece it should be possible to identify the next most valuable event, or to predict a relative aesthetic value for all conceivable immediately subsequent events. To keep the choices tractable we consider an event to be a note onset and thus our problem becomes a matter of evaluating all possible subsequent notes, at any time point when a prediction is desired (say, every beat). Further, for the sake of constrained evaluation, limiting to four octaves of the western, equal-tempered, 12-tone scale we can consider 48 possibilities at each time step.

If we further assume that the state at any point in time

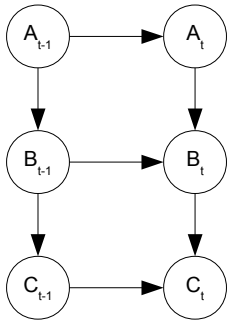

Figure 1: Single time slice of the DAG corresponding to the hierarchical intrinsically motivated generative model.

incorporates the relevant state from previous time steps, we have a simple Markov chain with the value function acting as a transition from state at time $t-1$ to the state at time t (the top level of fig. 1). If A_{t-1} is the previous musical context, or state, and our undefined aesthetic valuation function (which relates to the concept of *novelty*) serves as a transition function, predicting A_t, the next state our synthetic listener should find interesting is found by:

$$P(A_t|A_{t-1}) = novelty(A_t|A_{1:t}) \quad (1)$$

If we had such a function and a comprehensive memory of previous states ($A_{1:t}$, including previous pieces), iterating this function recursively could lead to a generative system capable of composing or improvising novel, *interesting* pieces. However, a system operating on a single level (such as the top layer of fig. 1) would require an extensive formulation of the context state in order to generate meaningful sequences with long-term relationships and dependencies. Additionally, fully predicting aesthetic preferences would seemingly involve the comprehensive auditory history of the individual in question, yet a simpler model could prove useful, creating pieces within a limited musical style or context to evaluate the application of the models in question.

A more tractable solution is to create a hierarchical model that groups sets of events in reasonable chunks (say, 3-5 events) and abstracts each layer into higher levels of structure. In musical terms this is analogous to the relationship of notes to motives to sub-phrases to phrases, etc. This formal abstraction can stack up to the level of a complete piece and beyond. Each level applies a similar transition function, asking, for example: after motive A, which motive will be optimally aesthetically interesting?

Figure 1 depicts such a model commencing at the note level at the bottom, C, and abstracting progressively with each higher level. The arrows indicate the influence of each layer's previous state on the next, given our novelty function, and the influence of the higher levels of abstraction on the note level, C. The directed acyclic graph (DAG) depicted in fig. 1 admits to the following factorization:

$$P(A_{1:T}, B_{1:T}, C_{1:T}) = P(A_1)P(B_1|A_1)P(C_1|B_1) \quad (2)$$
$$\times \prod_{t=2}^{T} P(A_t|A_{t-1})P(B_t|B_{t-1}, A_t)P(C_t|C_{t-1}, B_t)$$

Data Representation

Based on (Gjerdingen 1990; Smith and Garnett 2011; 2012) we take the operative context state to be a feature vector that uses spatial encoding (Davis and Bowers 2006) at the lowest level to create a short-term memory of note events. To keep the problem manageable only pitch (or a rest) is encoded, producing a vector comparable to that employed in (Smith and Garnett 2012), comprising the last five pitches, pitch classes, intervals, interval classes, and register information. Additional musical aspects, such as durations, tempo, and dynamics are ignored for the moment in order to examine the computational model in its barest form.

This feature vector is analyzed by an Adaptive Resonance Theory (ART)(Carpenter, Grossberg, and Rosen 1991) neural net in order to produce a classification which is then encoded into a feature vector for the next higher layer to analyze (as employed by Gjerdingen (1990)). Because the ART identifies new categories and classifications as new patterns are perceived the state space for the upper layers is unbounded, and typically expands over time (with a maximum rate of one category per input, but typically much slower, around one category per 5-10 inputs, decreasing over time). In our simple test cases we typically generate on the order of two-thousand events (notes or rests) resulting in several hundred category identifications.

As described above, the transition from the current state to the next state is done in a predictive fashion, testing all possible proceeding inputs and measuring the value of the transition. From Schmidhuber (2009), the function used to calculate this value attempts to maximize the intrinsic interest of the algorithm as measured by the amount of change in the ART neural nets. Schmidhuber (2009) proposes models to measure intrinsic interest and we employ a variation that contrasts the expansion of the ART nets with the amount of change in the weights of the nodes. We calculate the intrinsic reward that any layer generates based on a given transition by (β and γ are constants set to limit the result $[0, 1]$):

$$\frac{log(\beta - (|\Delta A_t - \gamma| + \gamma))}{s} \quad (3)$$

When new categories are identified they cause the ART to define new nodes, increasing the amount of data space (s) required to store the ART, and signifying a move towards chaos. The amount of residual change in the weights of the nodes (ΔA_t) similarly reflects the boredom-chaos continuum, where purely repetitious inputs produce no adaptation in the network ($\Delta A_t \approx 0$) and novel inputs produce significant changes ($\Delta A_t \approx 1$). The constant γ specifies the maximal point within the boredom-chaos range (nominally $r \approx 0.5$). Combining these two measures of change causes the algorithm to prefer inputs that cause adaptation in the node's weights without causing a new category to be created.

Algorithmically, the graph is resolved in a top-down, comprehensive fashion (which will be replaced with a Monte Carlo approach as more degrees of freedom are introduced, with further research). By testing each possible subsequent state at the each level a set of weights are calculated employing the novelty function, above. Here the values

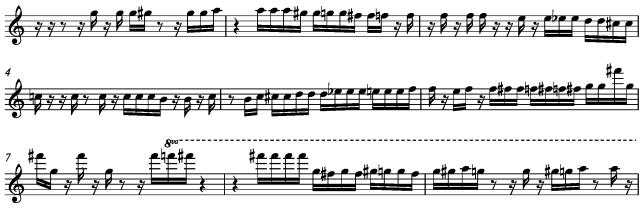

Figure 2: Musical fragment generated by computational model. See discussion section, below.

predicted by each pair of neighboring layers are summed:

$$P(B_t|B_{t-1}, A_t) = novelty(B_{t-1})\alpha + (1-\alpha)A_t \quad (4)$$

$$P(C_t|C_{t-1}, B_t) = novelty(C_{t-1})\alpha + (1-\alpha)B_t \quad (5)$$

Where $\alpha = 1/3$.

Dynamic Model

A single layer, one-note look ahead model based on Schmidhuber (2009) appears to lack longer formal structures (Smith and Garnett 2012), yet this hierarchical model (see fig. 1) shows hints of long-term memory and dependencies at longer time scales. However, the fixed nature of the layer's relationship (equations 4 and 5, above) maximizes local reward gain, denying note choices that would be deemed locally boring, at the expense of longer-term gain. At any point in time only a limited number of state transitions are available for a given layer (due to the short-term memory nature of the feature encoding). As an example consider the following sequences: AAA, AAB, ABC, and CBA where each character represents a distinct musical event (or note). If the current state is the first sequence (AAA) direct transitions (accomplished by appending individual characters to the sequence) are only possible to AAA and AAB. Yet, the upper layer (in any 2 layer relationship) may weight all of the accessible transitions at 0, giving significant weight to inaccessible transitions, such as a transition from AAA to CBA which would result in two intermediate sequences (AAC and ACB). This can be seen in fig. 3, showing reward generated at each layer over time, where each layer has periods of 0 reward as the system crosses over undesirable sequences to get to more interesting areas. The most extreme cases see the system stuck repeating a single note (a local maximum), where the only viable next step is a repetition of the current state. Simultaneously, the generated intrinsic reward for all layers drops to 0.

To counter this effect we consider the influence of the middle layer on the lower layer as a policy that is controlled by a yet higher layer (see fig. 4, providing a single connection, X, between the lowest layer, C, and a tertiary layer, A). This model is closer in spirit to recent work by Hinton (2006), using deep networks and restricted Bolzman machines.

When the reward resulting from the transitions of A decline, we assume that this is a the result of becoming stuck in an area of non-novel states which B has been unable to move out of. The solution is to give the lowest layer, C, more freedom to move to novel areas in these situations, which will

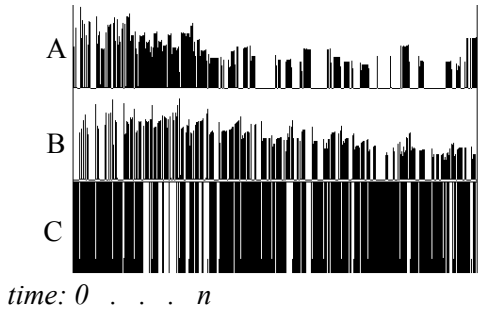

Figure 3: Reward generated at each level over time.

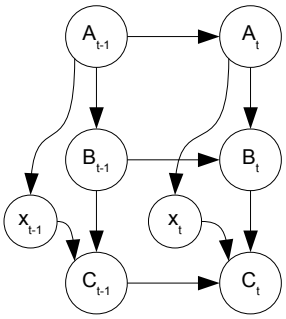

Figure 4: Single time slice of the DAG with the inter-level policy node.

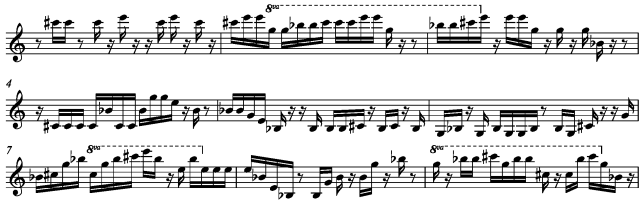

Figure 5: Musical fragment generated by model with policy nodes.

give A and B more opportunity for reward after a few time steps. Conversely, when A is receiving more reward, B is given precedence in determining future outputs.

Feedforward Policy

Finally, we consider an alternative model with the same goal. At time t the sum of the weights for a given layer is calculated and this sum is used to determine its weighting policy $\alpha = X_t$ for the next time step (see fig. 6). Thus if the higher layer, in any two-layer configuration, is proving too restrictive (due to isolation in local minima), the lower layer will be given more priority and freedom to pull the system to more novel areas at the next time-step.

$$X_t = \frac{\sum P(C_t^i|C_{t-1}^i, B_t^i)}{i} \quad (6)$$

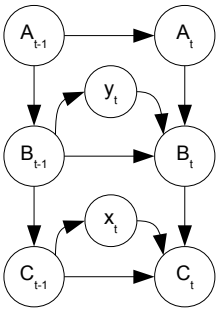

Figure 6: Single time slice of the DAG with feedforward policy nodes.

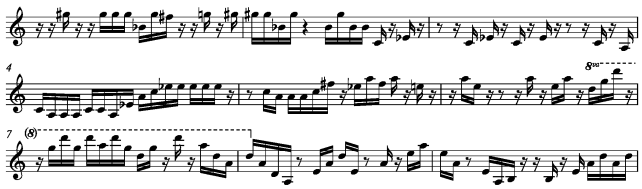

Figure 7: Musical fragment generated by model with feedforward policy nodes.

Discussion

These models attempt to address the problems of generating long-term musical structures employing NNs. However, even the best human improvisers required many hours of experience and training, and yet the models presented here are effectively beginners. The ability to introduce them to preexisting musical styles, or to generalize based on the system's own past preferences, has not been approached herein. The intrinsically motivated nature of the system indicates that these models will exhibit the same tendencies after training on musical examples, albeit with different surface textures and harmonic language.

Each of the musical figures exhibits characteristics of repetition and development, creating structure within the few bars depicted. In fig. 2 chromatic fragments, alternatively hesitant and continuous, move down and then up the scale until, arriving at the starting pitch of G, the texture breaks into major-7^{ths}. This instigates an ascension up two octaves where a new chromatic pattern commences, now without any immediate pitch repetition but on the same pitches as the opening sequence. Thus we see a statement interrupted followed by a return with variation.

Figure 5 employs different material, using a fully diminished arpeggio, and does not show any significant harmonic movement. However the same rhythmic pauses followed by short clusters of notes are in evidence. The strongest structural element seems to be register, as the fragment starts with an ascension up an octave, followed by a rapid descent down three octaves. After a couple of bars of centered around the low B-flat, two rapid climbs are seen delineated by a rapid descent in the middle. The last stave of the figure can be seen as a repetition with temporal diminution of the first two staves, creating a structure of acceleration at the phrase level.

The third example, fig. 7, appears to be setting up a longer time-scale development, as it continues to move through different harmonic material. Beginning with a short exploration of the pitches around G it quickly finds the diminished arpeggio which is used through the end of the second stave. However the appearance of the E-natural sets off a string of fourths and fifths which fill through the end of the figure. The continuation of this example sees the fifths replaced by a whole-tone scale which evolves into chromatics, briefly, before the augmented arpeggio is discovered. This leads to an alternation of major thirds with fourths and fifths before the run was halted.

Models one and two may be creating higher order structures, however a longer analysis of many scores would be required to make a conclusive argument. The rapacious nature of the third model, continually exploring new harmonic spaces, may be a direct result of the policy node formulation. Whenever the weights of the upper layer become too restrictive, i.e. possibly attempting to force the system into repeating material, the policy node denies enforcement, allowing the lowest layer to continue exploring freely.

Conclusions

Three variations of a hierarchical NN model were presented towards the generation of musical improvisations with higher-level dependencies and structure. All of the models successfully generated musical material which revealed pattern repetition and manipulation in non-obvious, yet intelligible fashion. However, evidence of long-term structure cannot be fully exposed in these short fragments. Longer examples for evaluation will be made available at http://ben.musicsmiths.us.

References

Carpenter, G.; Grossberg, S.; and Rosen, D. 1991. Fuzzy art: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural networks* 4(6):759–771.

Davis, C., and Bowers, J. 2006. Contrasting five different theories of letter position coding: Evidence from orthographic similarity effects. *Journal of Experimental Psychology: Human Perception and Performance* 32(3):535.

Eck, D., and Schmidhuber, J. 2002. A first look at music composition using lstm recurrent neural networks. *Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale*.

Gjerdingen, R. 1990. Categorization of musical patterns by self-organizing neuronlike networks. *Music Perception* 339–369.

Hinton, G.; Osindero, S.; and Teh, Y. 2006. A fast learning algorithm for deep belief nets. *Neural computation* 18(7):1527–1554.

Hochreiter, S.; Bengio, Y.; Frasconi, P.; and Schmidhuber, J. 2001. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.

Mozer, M. 1999. Neural network music composition by prediction: Exploring the benefits of psychoacoustic constraints

and multiscale processing. *Musical Networks: Parallel Distributed Perception and Performance* 227.

Schmidhuber, J. 2009. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. *Anticipatory Behavior in Adaptive Learning Systems* 48–76.

Smith, B., and Garnett, G. 2011. The self-supervising machine. In *Proc. New Interfaces for Musical Expression*.

Smith, B., and Garnett, G. 2012. Reinforcement learning and the creative, automated music improviser. In *Proc. EvoMUSART*.