



An End to End Model for Automatic Music Generation: Combining Deep Raw and Symbolic Audio Networks

Rachel Manzelli, Vijay Thakkar, Ali Siahkamari, and Brian Kulis
Department of Electrical and Computer Engineering
Boston University
Boston, MA 02215 USA
fmanzelli, thakkarv, siaa, bkulisg@bu.edu

Ano de publicação: 2018

UNIRIO

Programa de Pós-Graduação em Informática

Disciplina: IA para Geração de Música

Professor: Jean-Pierre Briot

Aluna: Carina da Silva Chear

Proposta do artigo:



Uma abordagem para combinar dois tipos prevalentes de modelos de deep learning para geração de música: modelo de áudio bruto e modelo simbólico. É descrito um modelo, ainda em construção que treina um modelo de áudio bruto baseado na recentemente proposta arquitetura WaveNet, mas que incorpora as notas da composição gerados pelo modelo simbólico como um input secundário para a rede.

Os autores acreditam que combinar essas duas abordagens trará novas possibilidades para a geração de música e abrirá as portas para uma série de novas ferramentas de geração de música.

Ferramentas de geração de música podem ser amplamente classificadas em dois tipos:

Modelos Simbólicos e Modelos de Áudio Bruto

Modelos Simbólicos:

São a maioria dos sistemas de deep learning para geração de música;

Baseados nas melodias e notas de dados de treinamento existentes, estes modelos produzem novas melodias principalmente no formato MIDI (Musical Instrument Digital Interface), que é um formato e protocolo padrão para música eletrônica ou ainda representação de piano roll, representação de texto (notação ABC), representação de acordes e partituras.

Um cenário típico para produzir música nesse modelo é treinar e gerar no mesmo tipo de representação. Por exemplo, treinar sobre um dataset de arquivos MIDI e gerar novas melodias em MIDI a partir do modelo aprendido.

Modelos Simbólicos:

Uma abordagem popular nessa área é utilizar RNN(recurrent neural networks) e em particular, redes LSTM(Long Short Term Memory).

Alguns modelos usando RNN incluem o Blues Melody Generation System(improvisação de Blues), DeepBach, Concert, Celtic Melody Generation, modelos LTM biaxial e vários modelos que combinam RNNs com máquinas Boltzmann restritas(modelos probabilísticos (ou geradores) não supervisionados, baseados em energia).

Vantagens dos Modelos Simbólicos:

Redes LSTM são capazes de capturar estrutura melódica de média-escala de uma música razoavelmente bem.

São rápidas em gerar novas melodias

Desvantagens dos Modelos Simbólicos:

Perdem riqueza expressiva e nuances da música executada

Precisam ser executadas por um humano ou um sintetizador

Modelos de Áudio Bruto:



Operam diretamente na forma de onda do áudio, tanto no treinamento quanto na geração de música.

Esforços iniciais para gerar áudio bruto envolveram modelos usados primariamente para geração de imagem. Inicialmente geravam áudio bruto com ruído e desestruturado.

Modelos de Áudio Bruto:



Vantagens:

São capazes de produzir conteúdo musical mais rico e com mais nuances, que podem capturar emoções, humor, etc. Em trabalhos iniciais, produziu música de piano com som realista.

Também são capazes de produzir sons inéditos, como a combinação de múltiplos instrumentos, bem como a geração de vocais.

Modelos de Áudio Bruto:



Desvantagens:

Falham em capturar efeitos de médio e longo prazo na música; a maior parte da música gerada soa improvisada, sem estrutura clara.

As formas de onda são tipicamente amostradas em 16 kHz, resultando em um modelo computacionalmente caro, que requer muitas previsões por segundo de áudio.

WaveNet:



Em um nível alto, o WaveNet é uma arquitetura de aprendizagem profunda que opera diretamente em uma forma de onda de áudio bruta, projetado principalmente para aplicações de fala e foi utilizado no Google Assistant.

Foi apresentado pela Deep Mind(empresa britânica, com o foco em pesquisas e desenvolvimento de máquinas de inteligência artificial) em 2016.

Tenta prever a próxima amostra de áudio.

A WaveNet modela diretamente a forma de onda bruta do sinal de áudio, uma amostra por vez. Além de produzir uma fala com um som mais natural, o uso de formas de onda cruas significa que a WaveNet pode modelar qualquer tipo de áudio, incluindo música.

Abordagem do artigo:



O trabalho propõe treinar uma rede LSTM biaxial para criar melodias simbólicas, e então tratar essas melodias como condicionamento local de um modelo WaveNet. Assim, o modelo LSTM permite criar estruturas melódicas de longo prazo na música, enquanto o componente WaveNet interpreta e expande a estrutura melódica gerada.

O trabalho ainda está em andamento mas são discutidos alguns resultados experimentais iniciais. Primeiro é discutida a sintonia do modelo WaveNet não condicionado para produzir música clássica para piano. Uma vez que este modelo é sintonizado, as discussões se estendem para o caso condicionado.

Geração de Música Não-condicionada com WaveNet



Como foi projetado para sintetizar fala, o modelo WaveNet foi ajustado para gerar música com base em contribuições de piano solo (cerca de 50 minutos de Chopin Nocturnes, do conjunto de dados YouTube-8M) bem como 350 músicas de vários gêneros de música eletrônica, obtido de No Copyright Sounds7.

Após treinar o modelo com música eletrônica multi-track, as gerações obtidas são barulhentas e geralmente contêm apenas informações de um único instrumento, como a linha de baixo ou a bateria. Com dados de treinamento complexos e diversos desse tipo, o WaveNet não generaliza bem. A mistura de letras e muitos tipos diferentes de sons na música não era consistente sobre os dados de treinamento, o que levou a gerações pobres.

Geração de Música Não-condicionada com WaveNet



No caso de treinamento em instrumentos individuais, como violoncelo e piano, os modelos WaveNet são capazes de produzir gerações musicais complexas sem perder a qualidade instrumental à medida que o número de amostras geradas aumenta

A rede é capaz de aprender dependências de curto alcance, incluindo ação de martelo e acordes simples. As gerações parecem manter uma emoção consistente por toda parte. No entanto, eles não são estruturados e não contem dependências temporais de longa prazo.

.

Condicionamento com Áudio Bruto



Uma vez que a melodia simbólica aprendida é obtida, ela é utilizada como uma segunda série temporal dentro do modelo de áudio bruto.

De acordo com o uso do condicionamento local pela WaveNet, é possível introduzir uma segunda série temporal (a partir do modelo LSTM para obter melodia de longo prazo).

No modelo WaveNet modificado, essas timeseries são as timeseries de incorporação de MIDI sob upsampling(upsampling é um método feito com o objetivo de aumentar a resolução da fonte de áudio, de maneira a minimizar ou até eliminar as potenciais limitações impostas pela amostragem (medida em kHz) e o número de bits usado para armazenar a informação digital (bit depth))

Avaliação



Geralmente as músicas geradas só podem ser avaliadas por ouvidos humanos.

Os autores planejam usar técnicas de avaliação de crowdsourcing como o Amazon Mechanical Turk5(plataforma em nuvem que permite outsourcing de tarefas para outros usuários, usando um sistema de pagamento embutido para compensar a participação dos usuários) para comparar as saídas geradas com outros sistemas e assim classificarem quais sons parecem mais naturais.

Por enquanto, os resultados foram avaliados em termos de complexidade e estrutura das saídas musicais obtidas até agora. Exemplos de resultados gerados estão disponíveis em <http://people.bu.edu/bkulis/projects/music/index.html>.

Datasets de Treino:



Foi utilizado o algoritmo Melodia para gerar dados de treinamento. Detectando automaticamente o tom de uma melodia, o algoritmo é capaz de juntar dados simbólicos com uma entrada de áudio bruta. No entanto, o algoritmo detecta apenas a melodia principal de uma música.

Experimentos também foram realizados usando o Saarland Music Data, que contém 50 partes de música de piano ocidental com sequências MIDI perfeitamente alinhadas. Devido aos recursos limitados e foco do trabalho do artigo se concentrar exclusivamente no violoncelo, os resultados extensivos do treinamento neste conjunto de dados não foram gerados.

Datasets de Treino:



Atualmente é explorado o uso do banco de dados MusicNet para treinamento, pois esses dados apresentam tanto áudio bruto quanto anotações melódicas.

Outros metadados também estão incluídos, como o compositor da peça, o instrumento com o qual a composição é tocada e a posição de cada nota na estrutura métrica da composição. A música é separada por gênero; há mais de 900 minutos de piano individual, o que provou ser muito útil em treinamento em apenas um instrumento.

Geração de áudio modelo condicionado



Em um espectrograma de frequência de pico (figura), foi observado que, para um determinado tempo, tanto entrada como geração de áudio bruta têm poder similar em suas respectivas frequências; por exemplo a cor no espectrograma se intensifica em intervalos de tempo similares ao longo da música. Isso indica que o modelo está prevendo amostras no tempo com o ritmo do áudio bruto fornecido.

As alturas relativas entre as frequências de pico são semelhantes para cada peça de música, indicando que a frequência da saída está mudando de forma semelhante em toda a peça. Isso indica que a saída de áudio bruto segue de perto a dinâmica do sinal de condicionamento. A diferença na altura vertical total dos espectrogramas é devida ao fato de que a entrada simbólica foi interpretada com um instrumento diferente da saída de áudio bruta e foi reproduzida em uma frequência geral diferente.

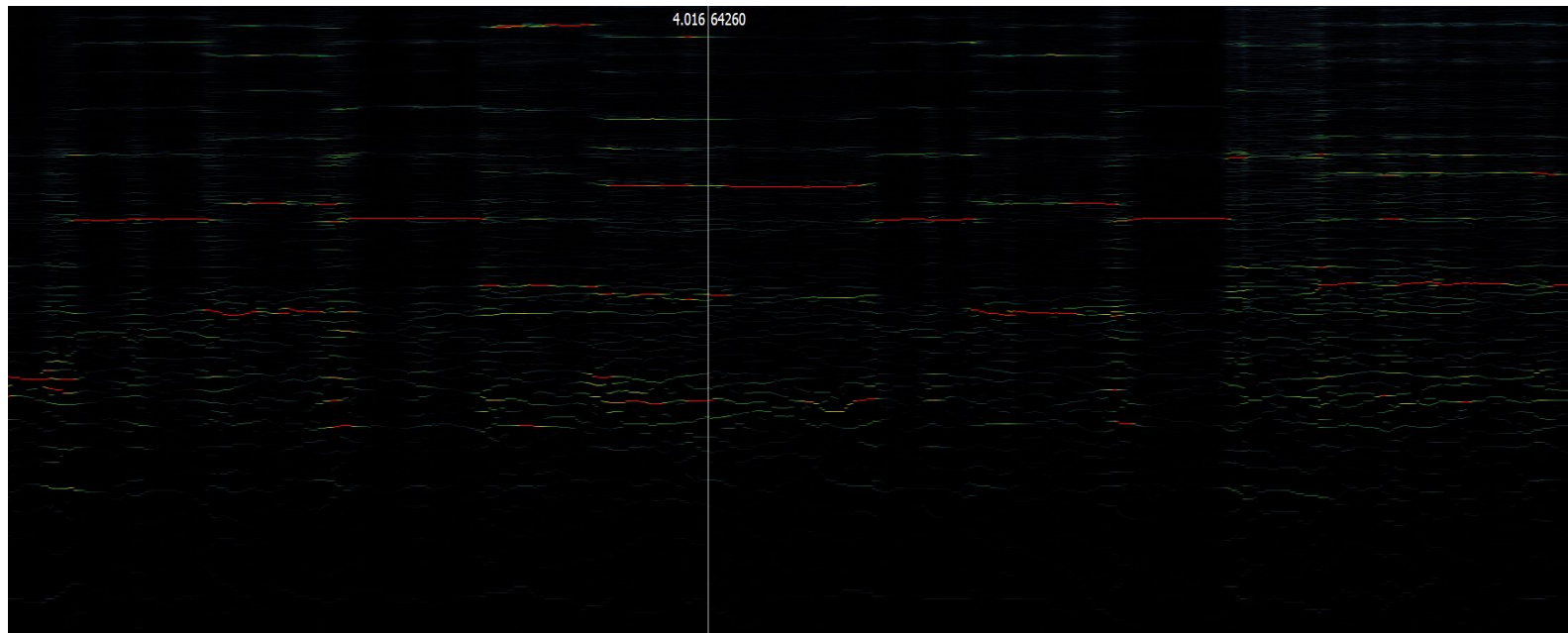
Geração de áudio modelo condicionado

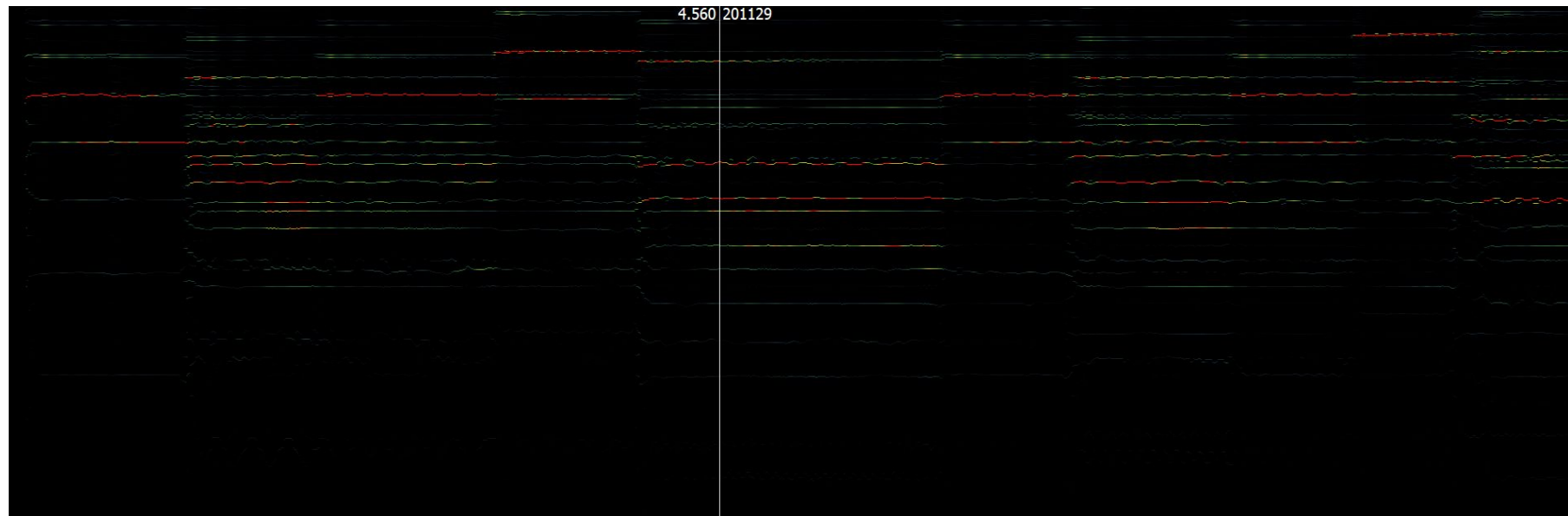


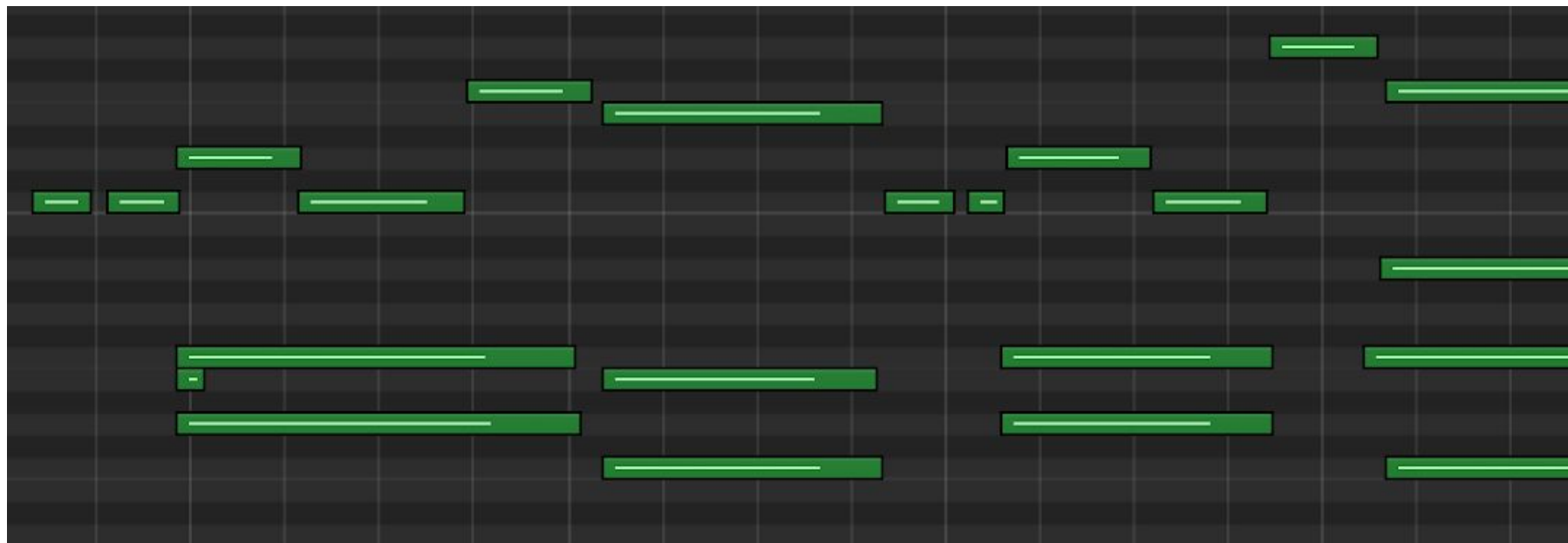
Também foi notado que, embora várias notas estejam sendo reproduzidas ao mesmo tempo, ao longo da entrada simbólica, o áudio bruto do modelo indica som nessas frequências em timesteps (mais suavemente, entretanto, uma vez que a cor não é tão intensa). Esta análise se generaliza para todas as peças geradas com o modelo proposto.

O trabalho proposto conseguiu transformar com sucesso um modelo não estruturado com pouca dependência de longo alcance em um modelo com gerações guiadas que exibem certas características.

Figura: Os espectrogramas de frequência de pico para a saída de áudio bruta de um modelo de violoncelo (primeiro) e seu sinal de condicionamento, que é uma parte da melodia de Feliz Aniversário tocada em sinos (meio). Apesar de as duas formas de onda parecerem diferentes, a figura mostra a estrutura semelhante de cada espectrograma à sequência de condicionamento (ultimo).









The audio files are posted at <https://ismir2018submission.wixsite.com/music/music>.



FIM