Music Composition using Recurrent Neural Networks

Nipun Agarwala, Yuki Inoue, and Axel Sly

Dataset + Preprocessing

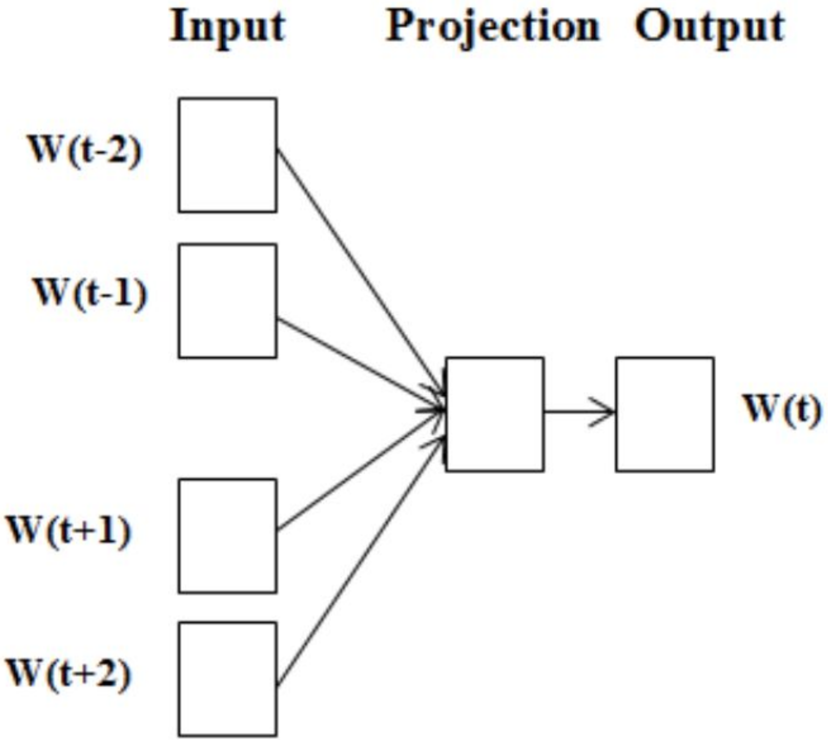
- Encoding: Text format - ABC notation
- Augmentation: Each song transposed to 4 random keys


```
X: 1
T: A Cup Of Tea
Z: dafydd
S: https://thesession.org/tunes/3038#setting3038
R: reel
M: 4/4
L: 1/8
K: Amix
|:eA (3AAA g2 fg|eA (3AAA BGGf|
eA (3AAA g2 fg|1afge d2 gf:|2afge d2 cd||
|:eaag efgf|eaag edBd|eaag efge|afge dgfg:|
```

```
X:1
T:ACupOfTea
R:reel
M:4/4
L:1/8
K:Amix
Q:1/4=100
|:eA(3AAAg2fg|eA(3AAABG
Gf|eA(3AAAg2fg|1afged2gf:
|2afged2cd|||:eaagefgf|eaa
gedBd|eaagefge|afgedgfg:|
```

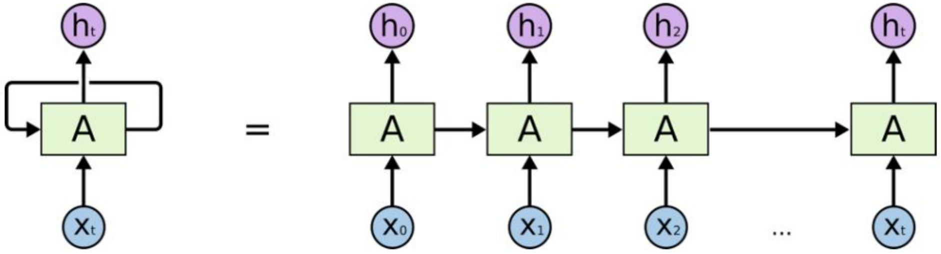
Methodology + Results: CBOW

- Implemented as a baseline for other models
- Context window: previous n characters, instead of traditional balanced context window
- 20% accuracy: overfit, not enough expressive capacity
- Nothing musical produced



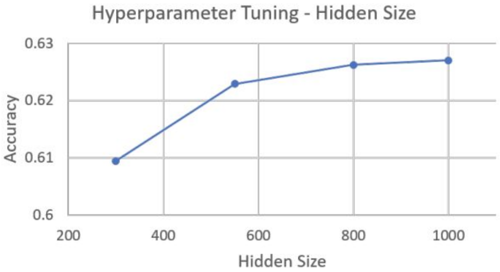
Methodology + Results: Character RNN

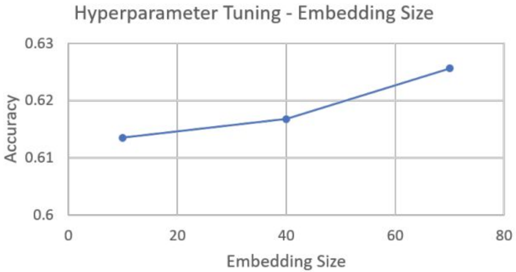
- One hot vectors of characters converted to character embeddings, passed through the Char-RNN, passed through a shared weight matrix and a softmax layer to get probabilities
- Cell types:
  - RNN: 39.5%
  - GRU: 47.5%
  - LSTM: 51.7% - final resulting model: 59.5%
- Output was passable, but not able to predict presence of bar lines

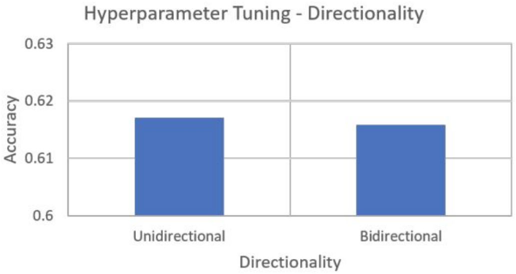


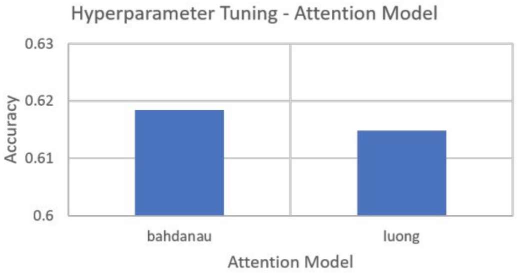
Methodology + Results: Seq2Seq

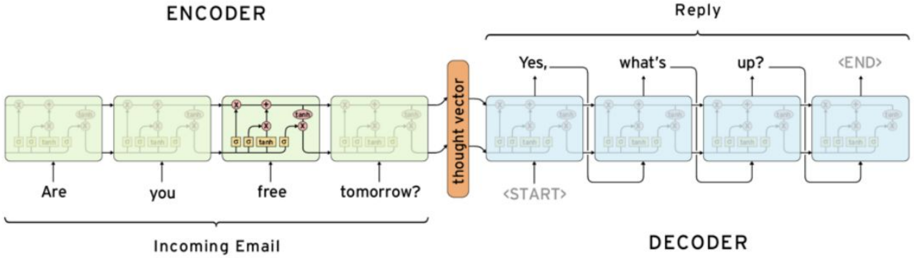
- Encoding and decoding: Character RNN
- 65.5% accuracy









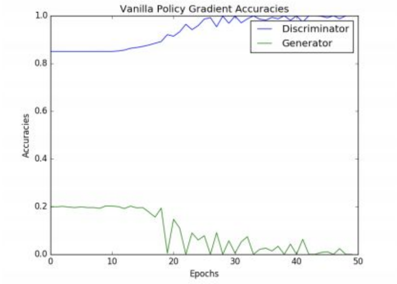


Methodology + Results: GAN

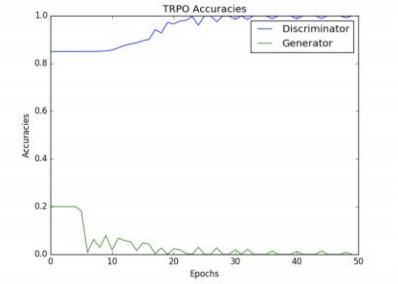
- Generator: Character RNN
- Embeddings passed into a 5 layer CNN for classification
- Output of the CNN to backpropagate policy gradients
- Trust Region Policy Optimization (TRPO):

$$\sum_{i=1}^{n} \frac{\pi}{\pi_{old}} r_i + KL(\pi_{old}||\pi)$$

π: newly predicted distribution
π_old: old distribution
r_i: reward for the ith sample



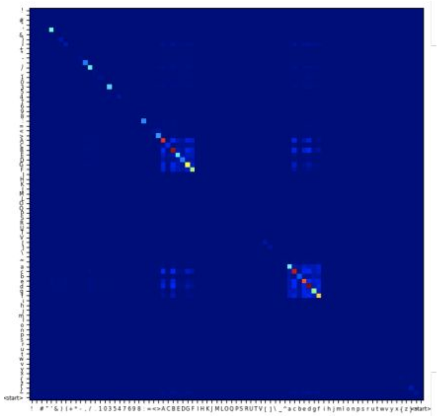
(a) GAN accuracies for simple Policy Gradients



(b) GAN accuracies for TRPO Gradient updates

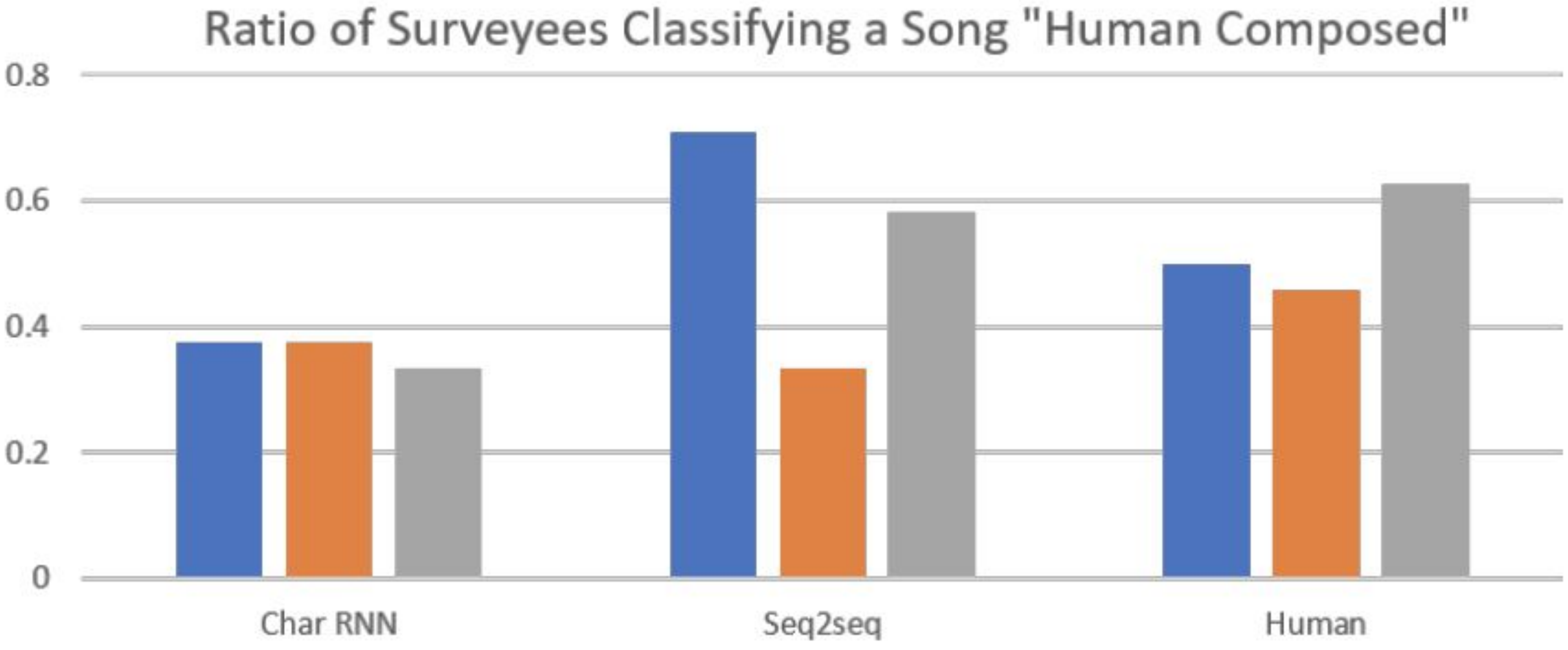
Char-RNN Produced Music

Char-RNN



Seq2Seq Produced Music

Survey



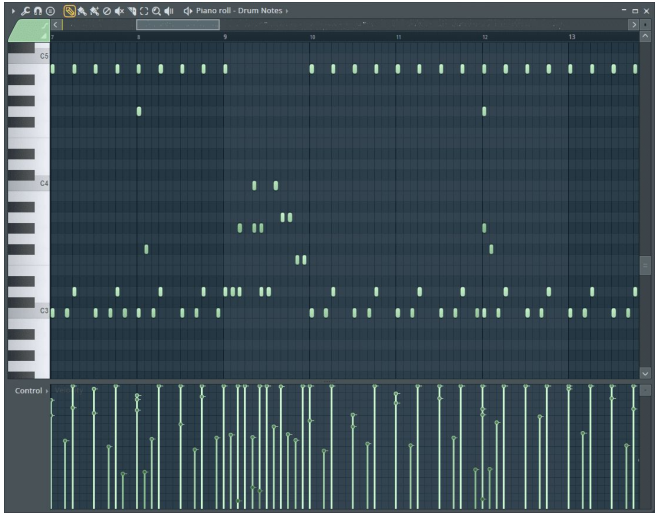
Ratio of Surveyees Classifying a Song "Human Composed"
0.8
0.6
0.4
0.2
0
Char RNN
Seq2seq
Human

A Study on LSTM Networks for Polyphonic Music Modeling

Adrien Ycart and Emmanouil Benetos

Dataset + Model

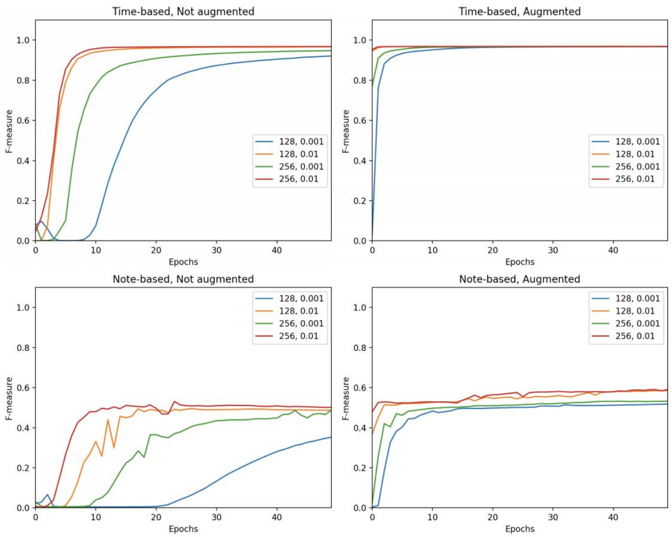
- Piano-roll representation (88xT binary matrix) of MIDI data
  - Augmentation: by 12x, transposed to all keys
- Time-based and note-based timesteps
- LSTM with 88 inputs, one single hidden layer, and 88 outputs
- Output sent through a sigmoid and thresholded



Time Step

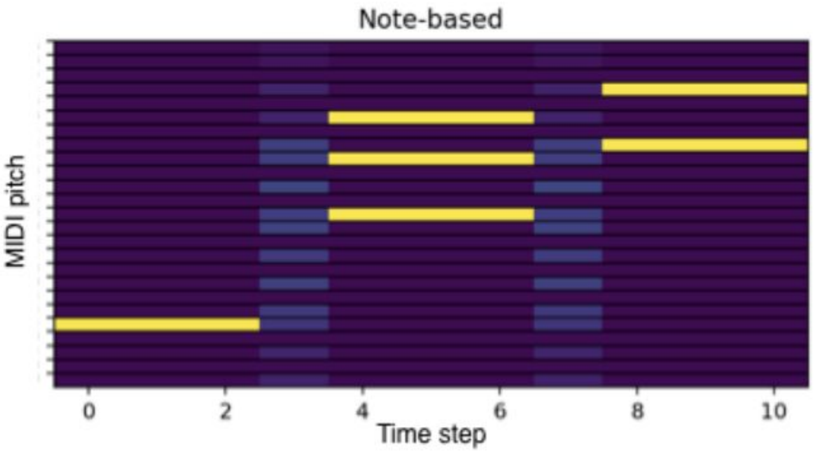
Time-based

- Fixed: 10 ms
- Predictive accuracy higher



Note-based

- Variable: 16th note
- Learned rhythmic structure



Audio Transcription

	F-Measure	Precision	Recall
Full audio, raw_piano			
Baseline	0.455	0.960	0.299
128, 0.001	0.458	0.938	0.303
256, 0.001	0.458	0.941	0.303
128, 0.01	0.460	0.959	0.303
256, 0.01	**0.460**	**0.961**	**0.303**
Right hand in C, raw_post, Synth			
Baseline	**0.670**	0.898	**0.535**
128, 0.001	0.556	0.955	0.393
256, 0.001	0.607	**0.966**	0.442
128, 0.01	0.522	0.834	0.380
256, 0.01	0.527	0.877	0.377
Full note-based, raw_piano			
Baseline	**0.526**	**0.963**	**0.361**
128, 0.001	0.434	0.624	0.332
256, 0.001	0.440	0.651	0.332
128, 0.01	0.478	0.852	0.332
256, 0.01	0.481	0.875	0.332