LSTM Based Music Generation

International Conference on Machine Learning and Data Engineering
(iCMLDE2019), 02-04 December, 2019, Taipei, Taiwan

Falak Shah, Twisha Naik, Nisarg Vyas
InFoCusp Innovations Private Limited

Contents

- Introduction - ML based music generation approaches
- Music data terminology
- Note sequence based approach
- LSTM
- Magenta based models
- Contributions of our approach
- Results
- Conclusion

Introduction - ML based music generation approaches

- Learning a long-term structure such as melodies, is one of the difficult problems which is hard for machines.
- Basic machine learning systems have trouble generating a longer melody that follows a chord progression, or follows a multi-bar song structure of verses and choruses.
- The Deep Learning models like LSTMs are greatly useful for such tasks.
- Two major approaches to music generation - directly from wav files or generated based on sequence of notes- using music in form of midi or musicxml

Introduction (Past work in the domain)

- GAN like generative models for polyphonic music generation
- MidiNet which used CNN for music generation
- Wavenet architecture for raw audio waveforms
- An unsupervised approach called audio word2vec to learn inherent note representations
- Based on the magenta library - we propose variants to its basic melody RNN which uses stacked LSTMs to learn a sequence of notes
- But before moving to that, some music terminology is essential

Music data terminology

- **Notes** - The pitch and duration of a sound, and its representation in musical notation. Sa, Re, Ga, Ma, Pa, Dha, Ni (C to B) in different octaves

  Example: C4, F5, E7

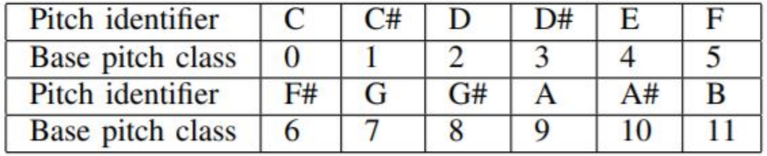
- **Melodies** - A sequence of notes
- **Chords** - Combination of 3 or more notes played simultaneously and they act in a supporting role forming the song’s harmony.

- **Bar -** It is a segment of time corresponding to specific number of beats. Typically, a piece consists of several bars of the same length, and in modern musical notation the number of beats in each bar is specified at the beginning of the score by the time signature.
- **Time signature -** A number denoted in form numerator over denominator - numerator showing number of beats in a bar and denominator showing what kind of notes constitute one beat.

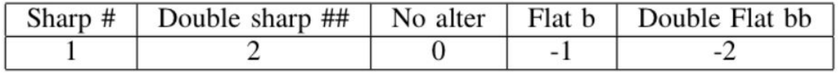
Bar and time signature have a major role when it comes to quantisation.

Music data terminology

- Digital representation of music: Music XML/ MIDI
- Note Sequence:



Pitch identifier	C	C#	D	D#	E	F
Base pitch class	0	1	2	3	4	5
Pitch identifier	F#	G	G#	A	A#	B
Base pitch class	6	7	8	9	10	11



Sharp #	Double sharp ##	No alter	Flat b	Double Flat bb
1	2	0	-1	-2

pitch class = (base pitch class + alter)%12 --- (1)

midi pitch = (12 + pitch class) + octave ∗ 12 --- (2)

Example: Note C#4 can be converted to midi pitch as follows:

From table 1 and 2 it can be seen that, base pitch class of C = 0 and alter of # = 1. The note itself suggests, octave = 4 Using equation 1, pitch class of C# = (0 + 1)%12 = 1

Using equation 2, midi pitch = (12 + 1) + 4*12 = 61

Data Preprocessing Pipeline

1. Note to pitch mapping

   Shown in the previous slide

2. Restricting note pitches within a range

   Since most notes in actual musical sequences fall in range 48-84, we restrict to that range

3. Time Splitting

   Divide song into chunks if there are multiple time splits

Data Preprocessing Pipeline (Continued)

4. Quantization

   Quantize since we want notes at fixed intervals. Shift start/ end times to nearest quanta

5. Melody Extraction

   Introduces two extra note events: No note or note off

6. Encoding

   Mapping all events to integers and convert to convert it into one hot vectors - we propose different variants which are described later in the slides

Long Short Term Memory (LSTM) Networks

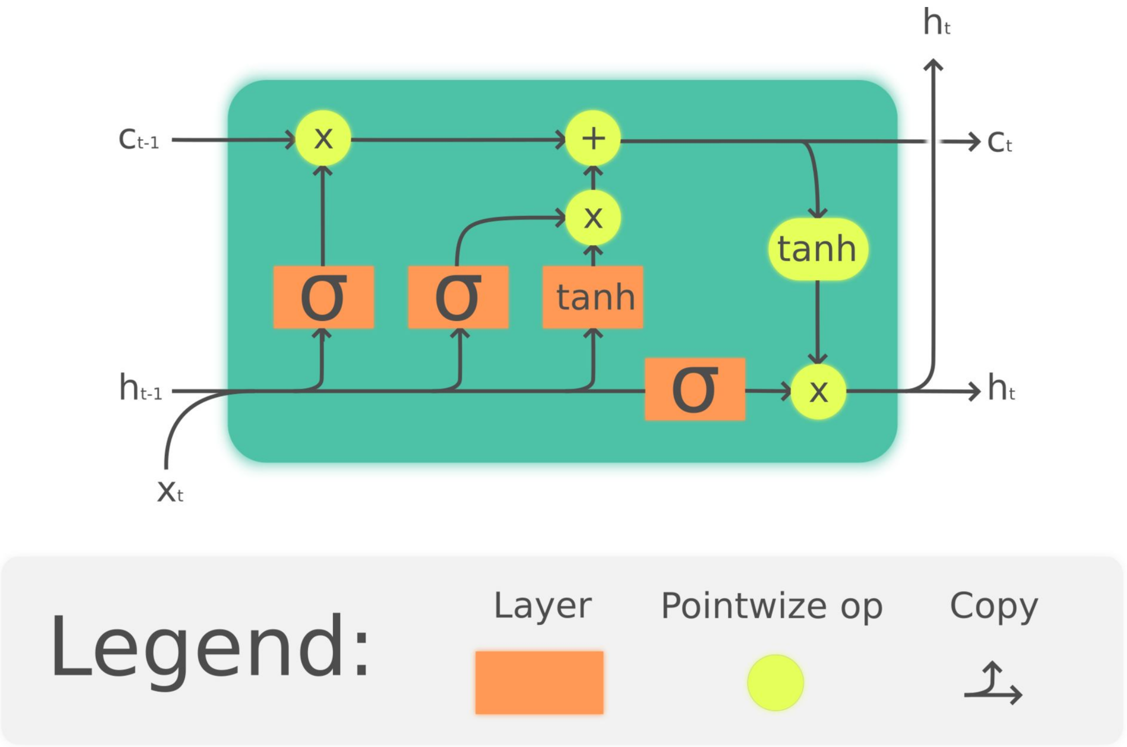
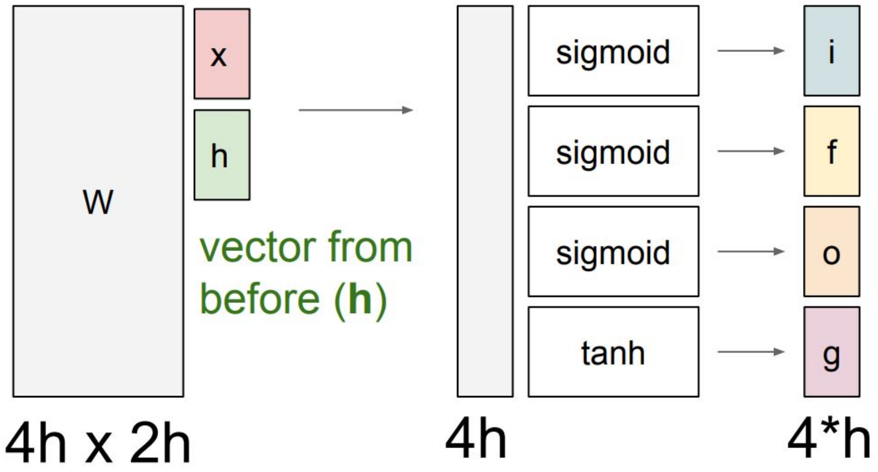


Image Courtesy: Wikipedia

LSTM as a structure has ability to retain information till needed, update if necessary or delete if no longer required from the memory.

It has 4 gates to do that as shown in the adjacent figure and as described in the images in the next slides.

LSTM Structure



$$\boxed{\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}}$$

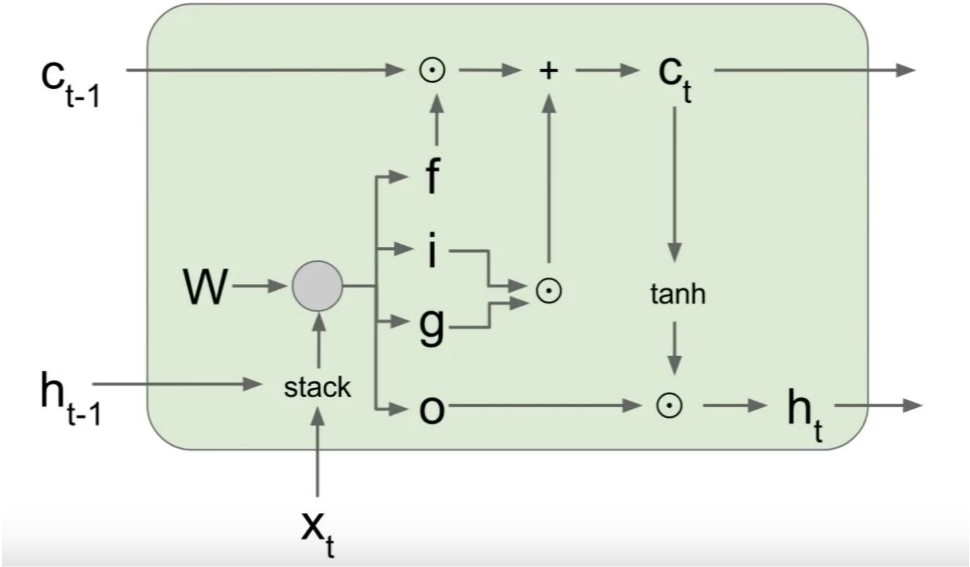
$$c_t = f \odot c_{t-1} + i \odot g$$

$$h_t = o \odot \tanh(c_t)$$

Image Courtesy: Justin Johnson lecture notes CS231n

f: Forget gate, Whether to erase cell state
i: Input gate, Whether to write to cell
g: Gate gate, How much to write to cell state
o: Output gate, How much to reveal to the cell state

LSTM Structure Continued



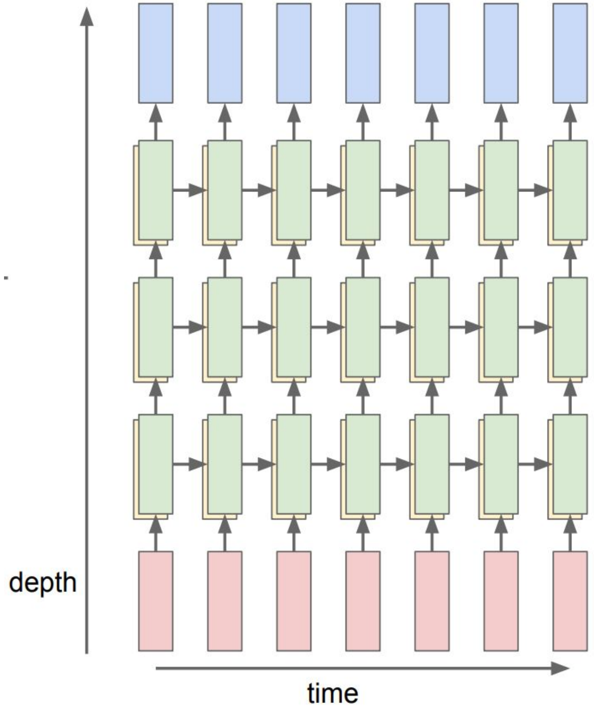
$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}$$

$$c_t = f \odot c_{t-1} + i \odot g$$

$$h_t = o \odot \tanh(c_t)$$

Image Courtesy: Justin Johnson lecture notes for CS231n

Stacked RNN structure



When we have a stacked structure, for the higher layers, we will have outputs from the previous layer RNN as an input.

The new equation will be:

$$\boxed{h_t^l = \tanh W^l \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}}$$

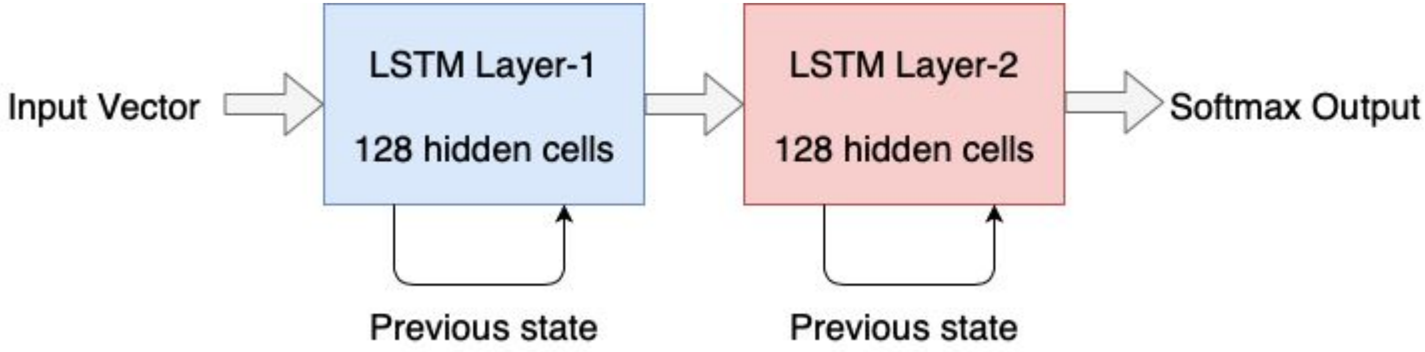
$h \in \mathbb{R}^n$ $W^l \ [n \times 2n]$

Magenta based models

Basic Model - note sequence based approach

- Model Architecture:

  Stacked LSTMs: Two layers of LSTMs with 128 hidden units each



At prediction time

- Initially, a primer sequence is given which sets the cell states of LSTM and which will help in predicting the next notes.

- Input - Previous note in one-hot encoded form and hidden state
- Output - Current note

Proposed Encoding methods

1. Neighbours distribution:

Provide additional information to the model about the target. The model is trying to learn the output note (i.e the next note) distribution given a particular input node.

2. Consecutive pitch difference:

The central idea behind this model is that any melody is defined by the difference of consecutive notes and not particular notes.

3. Pitch difference w.r.t. start note:

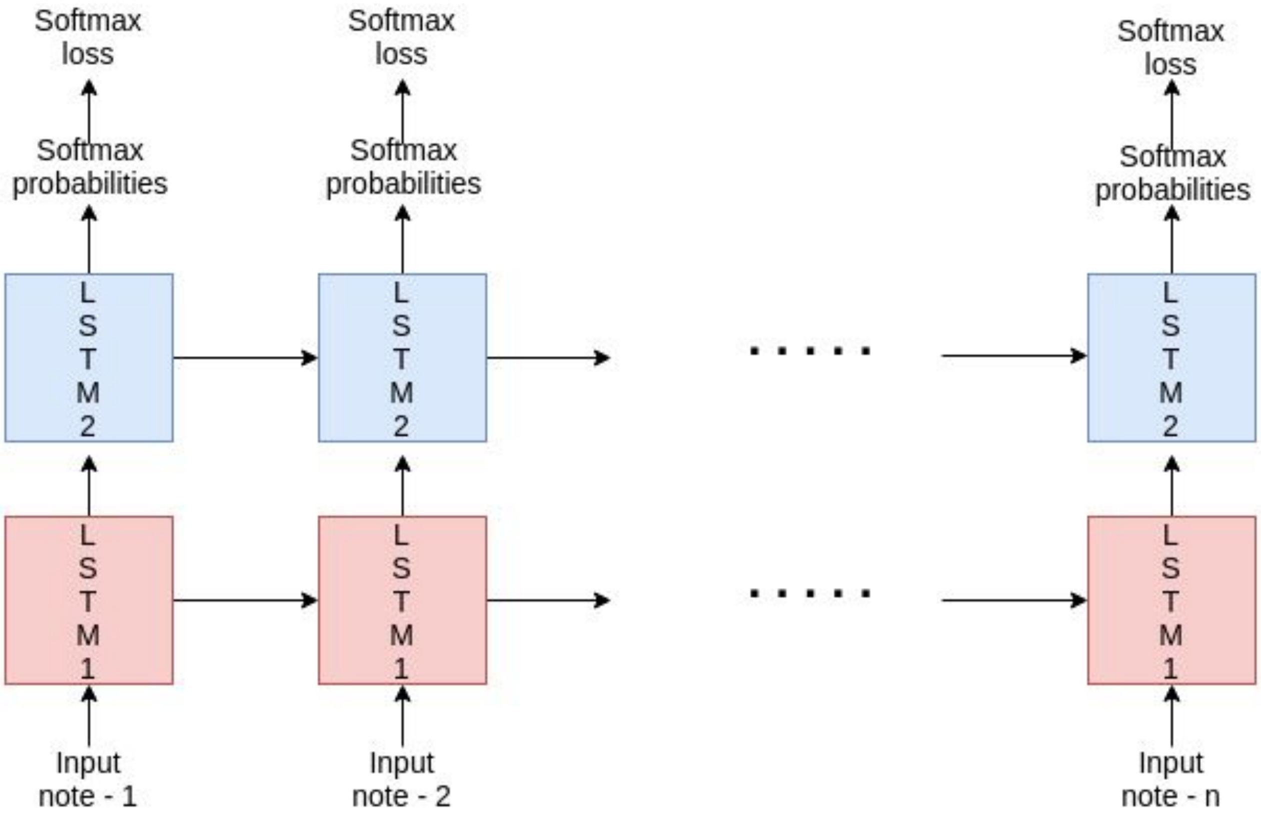
Different note sequences with differences being same from the start note sound very similar. Based on this observation, we can reduce the variations in the input data by encoding the notes information in terms of difference from the start note.

4. Bunch notes in input:

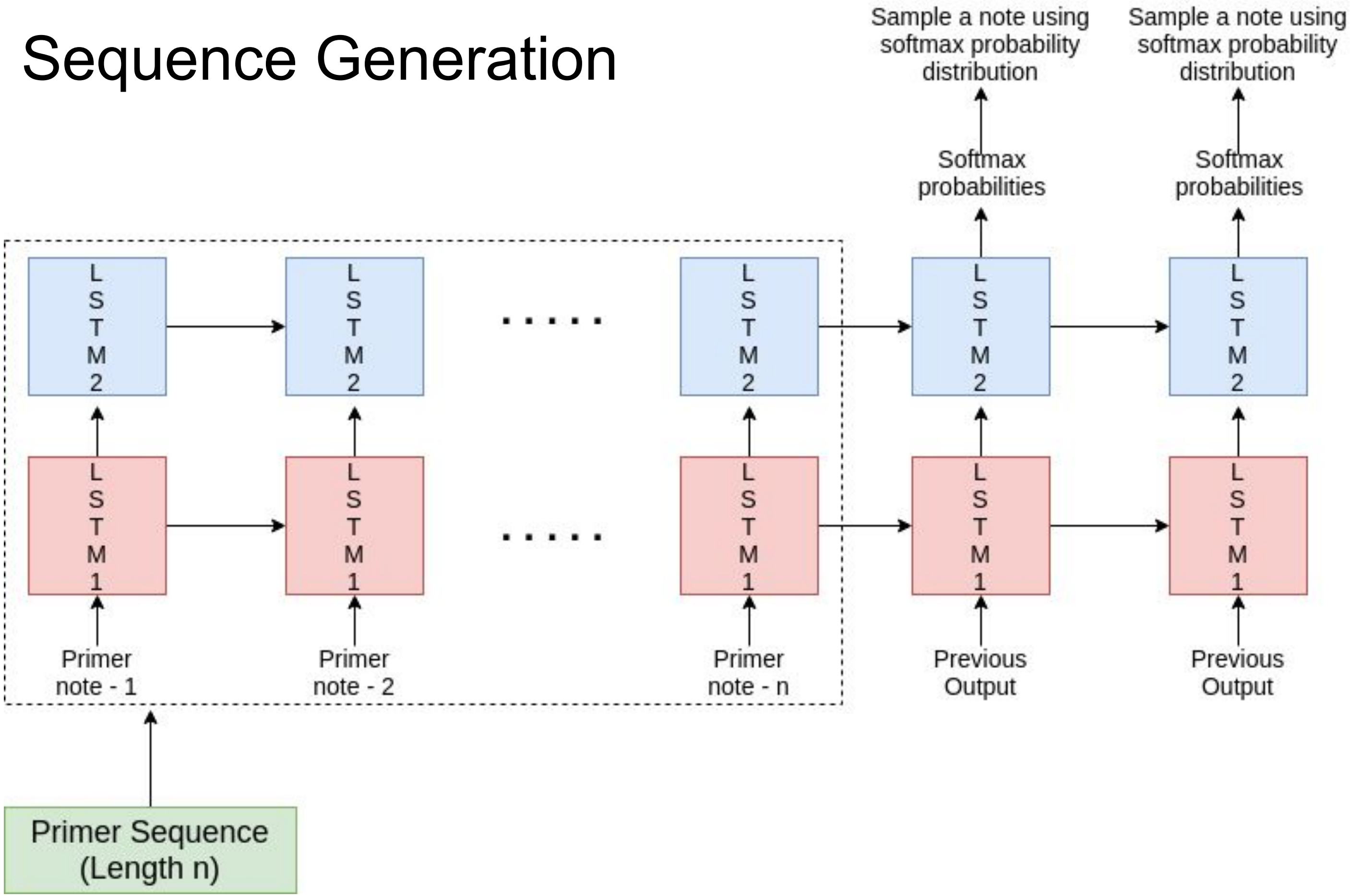
Instead of relying on the model’s hidden states for the information about previous notes, we explicitly provide a few of them in the model input.

A bunch of 4 notes was given as the input instead of just one note.

Model Training



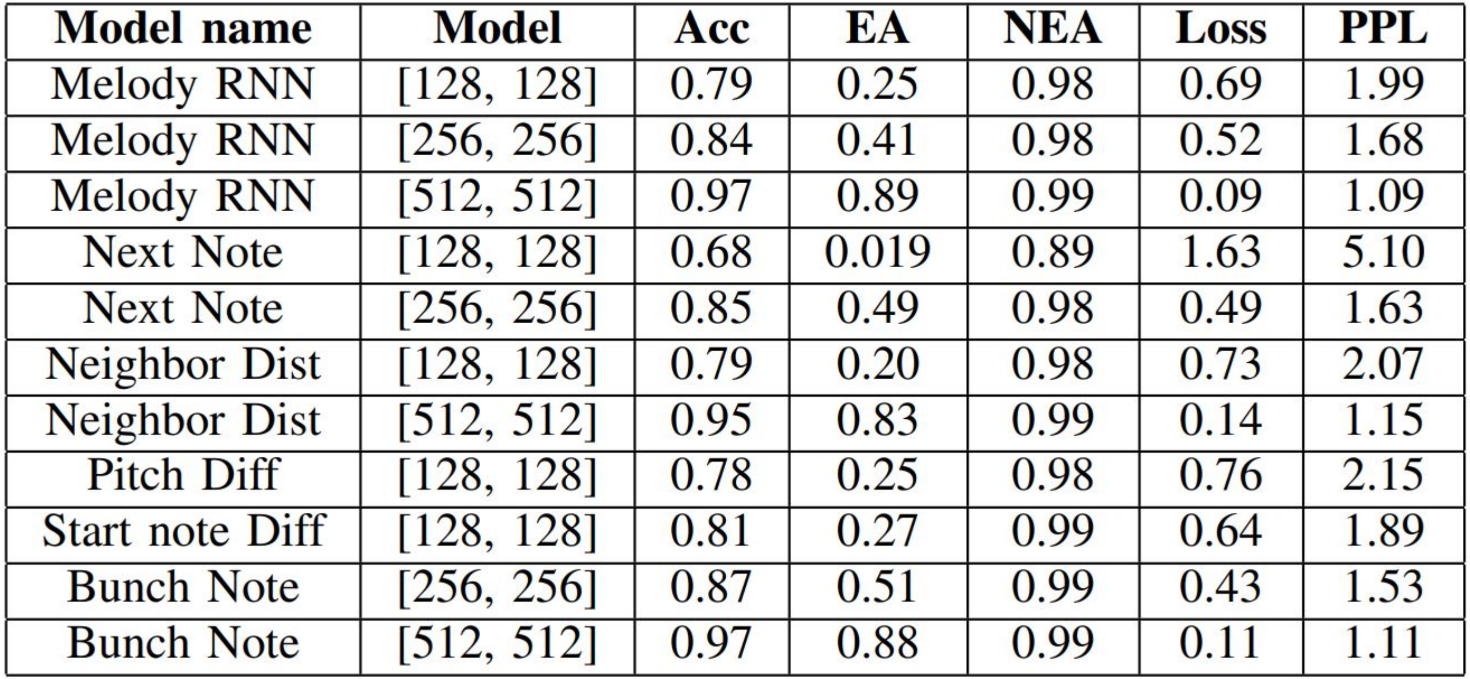
Sequence Generation



Our Contributions

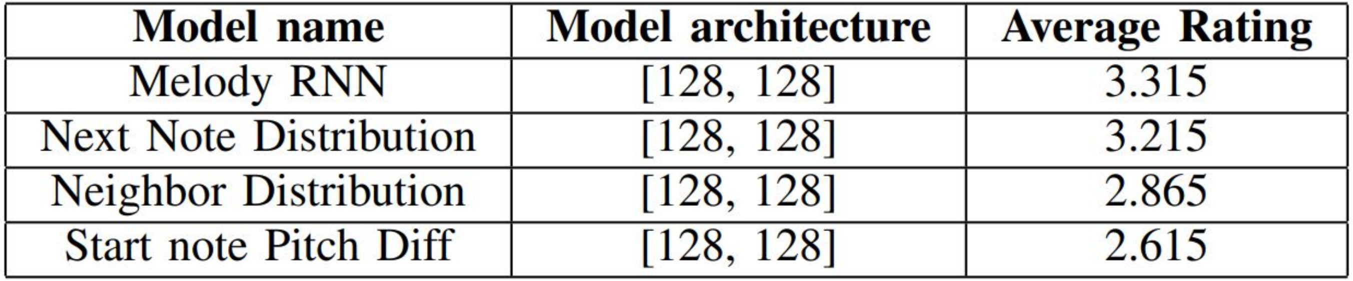
- Quantitative evaluation of magenta based models - so as to have a benchmark set
- Encoding variants introduced based on the similarity of structure in music domain
- Used these variants to generate melodies building further on magenta library

Quantitative Results



Model name	Model	Acc	EA	NEA	Loss	PPL
Melody RNN	[128, 128]	0.79	0.25	0.98	0.69	1.99
Melody RNN	[256, 256]	0.84	0.41	0.98	0.52	1.68
Melody RNN	[512, 512]	0.97	0.89	0.99	0.09	1.09
Next Note	[128, 128]	0.68	0.019	0.89	1.63	5.10
Next Note	[256, 256]	0.85	0.49	0.98	0.49	1.63
Neighbor Dist	[128, 128]	0.79	0.20	0.98	0.73	2.07
Neighbor Dist	[512, 512]	0.95	0.83	0.99	0.14	1.15
Pitch Diff	[128, 128]	0.78	0.25	0.98	0.76	2.15
Start note Diff	[128, 128]	0.81	0.27	0.99	0.64	1.89
Bunch Note	[256, 256]	0.87	0.51	0.99	0.43	1.53
Bunch Note	[512, 512]	0.97	0.88	0.99	0.11	1.11

Subjective evaluation results



Model name	Model architecture	Average Rating
Melody RNN	[128, 128]	3.315
Next Note Distribution	[128, 128]	3.215
Neighbor Distribution	[128, 128]	2.865
Start note Pitch Diff	[128, 128]	2.615

Music Samples

Neighbour distribution

🔊

Start note pitch difference

🔊

Melody RNN

🔊

Conclusion

- Presented baseline results on magenta models and variants of the models trained on different hyperparameters which can be helpful for quantitative analysis.
- Introduced variants of the encoding techniques based on
  a) Additional information which can be useful to LSTMs for predictions and
  b) The observation that the melody is relatively independent of the starting note, and the difference in pitches of notes from the start note preserves the characteristics of a song.
- These variants produce melodies which are comparable or better in perceptual quality compared to the baseline magenta models

Thank You