International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Music generation using Bidirectional Recurrent Neural Nets

Syeda Sarah Azmi[1], Shreekara C S[2], Shwetha Baliga[3]

[1]Student, Department of Electronics and Communication Engineering, RV College of Engineering, Bengaluru -59, Karnataka, India, syedasarahazmi.ec16@rvce.edu.in, 9742854662

[2]Student, Department of Electronics and Communication Engineering, RV College of Engineering, Bengaluru -59, Karnataka, India, shreekaracs.ec16@rvce.edu.in, 9482481103

[3]Assistant Professor, Department of Electronics and Communication Engineering, RV College of Engineering, Bengaluru -59, Karnataka, India, shwethaprabhun@rvce.edu.in, 8892303813

Abstract - *The advancement in neural network and deep learning is enabling the use of these technologies in several art and other fields, and produces outcome which are similar to humans. This paper proposes a method to generate music, different than the music samples it uses as the dataset using bidirectional recurrent neural networks (RNN) with Long Short Term Memory (LSTM) cells, efficiently, by understanding the complex relation between the different notes, pitches, timbre values, that constitutes music. Experiments with the Pokémon midi files showed that we can achieve a high performance in music generation using bidirectional recurrent neural networks with Long Short Term Memory (LSTM) cells. Further, the performance of the model with the variation in the number of epochs used is analyzed*

Key Words: **Terms-music, bi-directional recurrent neural networks, long short term memory, Machine Learning, Harmony**.

1. INTRODUCTION

Deep Learning classifiers and generators have already been used in several applications, over a wide spectrum of domains. A domain in which the use of deep learning algorithms is becoming popular is in the generation of music for different applications and software. This involves training some kind of a generator model on a large set of "music" data and coaxing it to produce similar kind of music with some variation. This is an interesting application of generators, as music as a dataset is much more complicated than images or text for direct application of generation- with each track composing of different notes, pitches, timbre values and so on. Therefore, analysis of conversion of music into some format which can easily be understood by models is a very important step which can make or break our "music generator".

After the advent of the computer, there is a powerful technology about data organization and construction [1]. This paper focuses on application of Bidirectional RNNs (with Long Short Term (LSTM) cells) and Markov chains for music generation. These models have been trained on MIDI files which contain the music data. The data has been pre-processed in different ways depending on the models and then used for training the models.

The structure of this paper is outlined as follows; in Section2, this paper discusses about different types of recurrent neural network and also the data representation of the input. In Section 3, discussion about dataset description and the pre-processing of the dataset is done. Section 4 discusses the architecture of bidirectional RNN and stacked bidirectional RNN along with the network architecture of RNNs. In Section 5, results are discussed and its detailed analysis is done and Section 6 concludes the paper.

2.0 BACKGROUND

A brief background of the related work is explained in the section that follows.

2.1 RELATED WORK

Music has been produced by computers in different ways throughout the years starting from a more rule or grammar oriented method to probabilistic methods as Markov Model. Currently, music production using artificial intelligence and neural network techniques is in much discussion and research.

The recurrent neural networks has applications in many fields, such as, music composition, predicting air pollution, Language Modelling and Generation Text, Machine Translation, Speech Recognition, Generation of image descriptions, Video Tagging, Text Summarization, Call Center Analysis, Face Detection. The neural network approach used for music generation should have the least error, to produce music with high accuracy. In 2018, T.Liu et al. [2] used recurrent neural networks based on LSTM for predicting Geomagnetic Field, and compared the performance between, Fully connected Neural network, recurrent neural network and LSTM Recurrent neural network and observed that average error and maximum error was reduced in the case of recurrent neural network with LSTM. In 2018, Y.Tsai et al [3] used recurrent neural network with LSTM to forecast the PM2.5 concentrations of various station of Taiwan, and

compared the performance between Artificial Neural Network and recurrent neural network with LSTM, saw a reduction in root mean square error in the case of LSTM.

In 2017, T.Hori et al. [4] performed music harmony acknowledgment from sound information utilizing distinctive bidirectional encoder-decoder LSTMs. A superior and precise music was perceived by the Conditional stacked bidirectional LSTM organize. In 2017, Yu Wang [5] developed a dynamic system identification system and observed that LSTM with multiple models were better in performance and speed to the conventional RNN and LSTM. D. Lee et al. [6] manufactured a connectionist fleeting characterization on an extremely huge dataset. And the acoustic model was designed for clean speech and noisy speech, and found that number of hidden layers are very less in the case of LSTM, compared to HMM.

2.2 DATASET REPRESENTATION

Music generation using bidirectional recurrent neural network with LSTM uses dataset which are in midi format. In 2017 D. M. Dhanalakshmy et al [7] proposed a system through which one can obtain data about enumerable parameters related to music such as the length of the track in terms of time as well as the pitch. Further, it also produced the resulting output in string format. This approach produced a very accurate midi file from a music sheet. The sample representation of MIDI file format is as shown in Fig -1 below.

```
'0, 0, Header, 1, 1, 1024\n',
'1, 0, Start_track\n',
'1, 0, Title_t, ""\n',
'1, 0, Pitch_bend_c, 0, 8192\n',
'1, 0, Note_on_c, 0, 60, 90\n',
'1, 512, Note_on_c, 0, 60, 90\n',
'1, 257536, End_track\n',
```

Fig -1: Sample Representation of a MIDI file.

3.0 DATASET DESCRIPTION AND PREPROCESSING

As stated earlier, music data is complex and difficult to represent as it composes of different attributes. Single instrumental music or track is much simpler to process, especially Piano instrumental tracks. Also using MIDI files helps in pre-processing as MIDI files contain metadata which can be used to convert the tones into some other format suitable to our model. Hence, we chose a corpus of piano music to train our models.

The dataset consists of a set of 307 piano tracks from Pokémon games, all in the form of MIDI files of length 1 minute approximately. The tracks neither have pauses nor vocals, which makes it easier for the model to train from it. The tracks used for training are stored in a folder named "Pokémon MIDIs".

For pre-processing and extracting information from the music files, we used a software tool called Music21. Music21 is a set of tools which can be incorporated in python to do in-depth analysis and generation of musical tones and ragas. It builds on pre-existing frameworks. The pre-processing and representation of data vary with the model.

4.0 MODEL AND ARCHITECTURE

4.1 BIDIRECTIONAL LSTM

Bidirectional repetitive neural systems (BiRNN) are extremely simply assembling two free RNNs. The info grouping is taken care of in typical time request for one system, and in invert time request for another. The yields of the two systems are typically linked at each time step, however there are different choices, for example summation.

The structure as shown in Fig -2 permits the systems to have both in reverse and forward data about the arrangement at each time step.

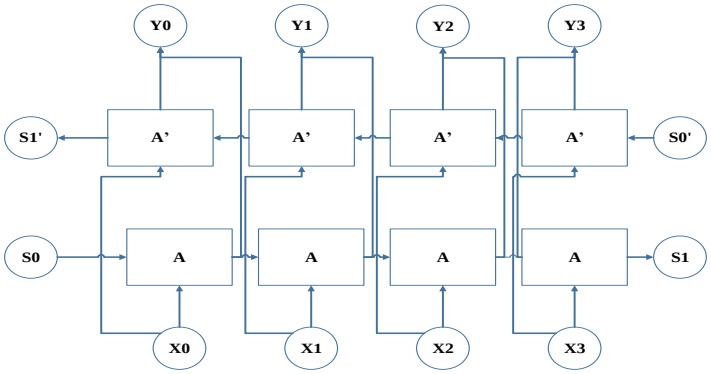

Fig -2: Bi-directional Recurrent Neural Network with LSTM

Analysis of MIDI files shows that the dataset can viably be treated as content information, with each note deciphered as a "word". This analogy makes it evident that LSTM cells can be effectively used on the processed music notes, along with temporal sequence to generate new "words", a.k.a notes in our case. Therefore, the data was preprocessed and represented as a numerical sequence of dimensions (num_inputs, 100, 1). The idea while training was that given a sequence of 'n' notes in sequence, the model should predict the 'n+1'th note efficiently. So a label set is used in which one hot encoding is used for the 497 classes of notes. Also since

Tensor flow is being used at the backend, the input and output need to be converted as a set of tensors, each with the correct input and output dimensions. These are then used for training the network.

4.2 NETWORK ARCHITECTURE

The Bidirectional LSTM has 1 input layer, 3 hidden layers (Bidirectional LSTM cells) and an output layer. Since Tensorflow is being used in implementation, the input and output layers have not been explicitly defined and are implemented using tensors and matrix operations.

Input layer: Implemented using a sequence of tensors, each of shape (batch size, input dim). Input dim has been used as 1 since each input is a single note and a batch size of 100.

Hidden layer 1: Layer consisting of 512 hidden units, where each unit consists of a forward and backward Bidirectional LSTM cells with forget-bias=0.1.

Hidden layer 2: Layer consisting of 512 hidden units, where each unit consists of a forward and backward Bidirectional LSTM cells with forget-bias=0.1. On this dropout with drop probability of 0.2 has been added.

Hidden layer 3: Layer consisting of 512 hidden units, where each unit consists of a forward and backward Bidirectional LSTM cells with forget-bias=0.1.

Output layer: The output is a single output of dimensions (1,497) which is a one hot vector for the output note class. This is actualized by network duplication of Weight framework with the yield of the past layer and including the inclination grid.

4.3 STACKED BI-DIRECTIONAL LAYERS

In the network, as shown in Fig -3, stacked architecture has been used for the bidirectional hidden layers, where-in each layer's output is passed as input for the next hidden layer. This provides a cumulative effect on the processing as compared to alternately stacked hidden layers. The initial input is given only to the first hidden layer and the output is taken from the last hidden layer has been used to compute final output.

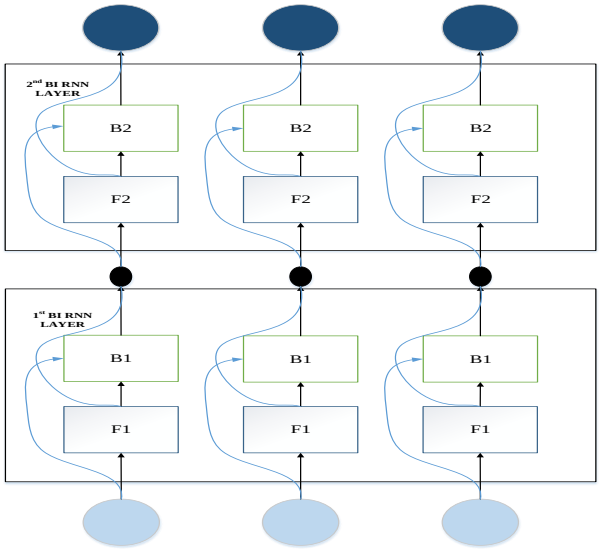

Fig -3: Stacked BiRNN Layers

5.0 RESULTS AND ANALYSIS

After training the Bidirectional model, it was tested by feeding a seed sequence of 100 notes. The model then generated the next 500 notes, which was converted into a track and stored as a MIDI file. The track generated was 2 min long and had similar note-sequence as the dataset.

The training loss was observed over 30 epochs as shown in the Chart -1. It is observed that using 15 epochs would be optimum in terms of training loss and the time for training. Table -1 shows the variation in accuracy with the number of epochs used. Though the accuracy and training loss for 30 epochs have better values but the training time for it is twice of the training time required for the training of 15 epochs this would affect the efficiency when the data set is considerably large. Hence, 15 epochs is selected as the optimum number.

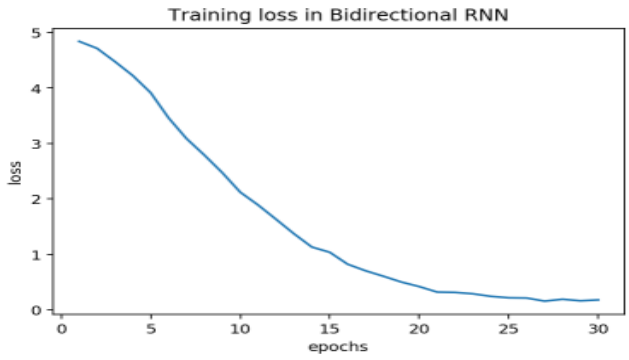

Chart -1: shows the graphical representation of variation of training loss in Bi-directional RNN versus the number.

Table -1: shows the accuracy of the Bi-directional RNN model for varied number of epochs.

No. of epochs	Accuracy
10	68%
15	89%
30	99%

6. CONCLUSIONS AND FUTURE SCOPE

In this paper, bi-directional RNNs using LSTM are introduced with the goal to generate music that is unique and differs from that of the dataset used. The model took input from both the positive and negative directions which allowed it to have information at each step in a bi-directional manner. The MIDI files could effectively be used as text data thus enabling the use of LSTM cells. The Bi-directional LSTM network has three hidden layers with the second layer having a drop out to prevent overfitting. The stacked nature of the Bi-directional hidden layers provided a cumulative effect on preprocessing. It was observed that training with 15 epochs is most optimum in terms of time and training loss. New music was generated after training by inputting a single note to the model.

The future work for this would be using a Long Short-Term Memory (LSTM) along with a Recursive Generative Adversarial Network (RNN-GAN) to generate music but this is more computationally expensive and would require a high powered Graphical Processing Unit (GPU).

REFERENCES

[1] G. Qian, "A Music Retrieval Approach Based on Hidden Markov Model," *2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, Qiqihar, China, 2019, pp. 721-725.

[2] T. Liu, T. Wu, M. Wang, M. Fu, J. Kang and H. Zhang, "Recurrent Neural Networks based on LSTM for Predicting Geomagnetic Field," *2018 IEEE International Conference on Aerospace Electronics and Remote Sensing Technology (ICARES)*, Bali, 2018, pp. 1-5.

[3] Y. Tsai, Y. Zeng and Y. Chang, "Air Pollution Forecasting Using RNN with LSTM," *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, Athens, 2018, pp. 1074-1079.

[4] T. Hori, K. Nakamura and S. Sagayama, "Music chord recognition from audio data using bidirectional encoder-decoder LSTMs," *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Kuala Lumpur, 2017, pp. 1312-1315.

[5] Yu Wang, "A new concept using LSTM Neural Networks for dynamic system identification," *2017 American Control Conference (ACC)*, Seattle, WA, 2017, pp. 5324-5329.

[6] D. Lee *et al*., "Long short-term memory recurrent neural network-based acoustic model using connectionist temporal classification on a large-scale training corpus," in *China Communications*, vol. 14, no. 9, pp. 23-31, Sept. 2017.

[7] D. M. Dhanalakshmy, H. P. Menon and V. Vinaya, "Musical notes to MIDI conversion," *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Udupi, 2017, pp. 799-804.