ATLANTIS PRESS

Advances in Social Science, Education and Humanities Research (ASSEHR), volume 300

2018 International Workshop on Education Reform and Social Sciences (ERSS 2018)

Automatic Composition of Music by LSTM

Yifan Zhao [1, a], Lixun Liu [1, *, b], Yifan Huang [2, c], Nadan Fang [1, d]

[1]Zhuhai College of Jilin University, Zhuhai City, Guangdong Province, 519000, China;

[2]Jilin University, Changchun City, Jilin Province, 130000, China.

[a]z870040648@outlook.com,[*,b]emtrix@163.com,[c]240558283@qq.com,[d]642166607@qq.com

Abstract. LSTM is used in different fields, such as Machine Translation, Time Series Prediction and so on. Machine Translation with LSTM uses Embedding vector for word and computes loss of two kinds of languages. And decreasing the loss of the Model and fixing it, which can translate one language to another one. Music and its lyrics are governed by rules which are declarative and non-deterministic. In this paper, compositing will be addressed with LSTM which can automatic analysis the rules of the music scores and lyrics. Automatic Composing can be looked the same as the Machine Translation. The Music score is a kind of language and the lyrics is another kind of language, which can use the Translation Model to fix it, and then to implement composing tasks.

Keywords: Automatic Composing, LSTM, Embedding Matrix of word, Translation Model.

1. Introduction

Deep Neural Networks (DNNs) are popular in recent years. DNNs are powerful in Computer Vision and Image Identification because they can do billions of parallel computing in a second [1-5]. Different from other DNNs that are to detect a picture in a second with the convolution, which cares less about the influence of the time series. Recurrent Neural Networks (RNNs) play a key role in solving the general sequences problem. In general, RNNs are also an encoding-base model and can learn in sequences to sequences model. Similar methods including: LSTM, GRU, TBCNN, etc.[5][6][7].

LSTM is a sort of encoding-base model and performs well in sequences to sequences model, especially in translation. One LSTM encodes input sequences as a fixed-length vector. Then another LSTM reads the vector representation and decodes it into an output sequence. This model achieved great results in many difficult sequence prediction problems and quickly became a popular model. However, decoding is limited to fixed-length vector for the reason that input sequence is encoded into a fixed-length vector.[5][8].

To solve the problem of fixed-length vector, researchers proposed the Inner-Attention. The core idea of the Inner-Attention is to break the limitation of the traditional structure of encoding-base model which relies on a fixed-length vector. Inner-Attention is implemented by using the LSTM's inner-output to train a model which can learn from the inner-output selectively and correlate it with the model output. Using Inner-Attention makes it easy to understand how the input information affects the output and helps researchers to understand the endogenous of the model.[5] But, there is large amount of computational complexity in Inner-Attention.

Music is the art or science of combining vocal or instrumental sounds with a view to beauty or coherence of form and expression of emotion. [4] In view of universality, Music score is composed of different sequences of pitches, which is closely related to waves of different frequencies. Different kinds of notes with different neighboring notes in different second convey numerous meaning. With numerous meaning, people can make a lot of lyrics.

Analyzing Music is a difficult thing. Traditional process is using transform such as FFT to extract features of the music. But it's not suited for all different music style, and not easy to find the Features' rule. Even though researchers use the Answer Set Programming (ASP) to solve the problem and achieve the Automatic Composition, this model still has some problem, for example the difficult global structure problem.[3].

In this paper, we propose an architecture is Inner-Attention with Sliding Window to solve the above problems. We show that the Bi-LSTM with Slide-Attention can achieve composition. In fact,

Automatic Composing can be treated as the Machine Translation. The Music score can be regarded as a kind of language, and the lyrics can be regarded as another kind of language. It can achieve composing comparable to human composing. Furthermore, qualitative analysis show that the model we made fixed perfectly in the Dataset we built.

2. Methodology

In this chapter, we describe Data Set and Slide-Attention. There are two LSTMs in the model. There is Bi-LSTM comes before the Slide-Attention called Pre-LSTM and use $a^{\langle t \rangle} = [\vec{a}^{\langle t \rangle}; \overleftarrow{a}^{\langle t \rangle}]$ to represent the concatenation of two direction activations of Pre-LSTM. And another LSTM which after the Slide-Attention called Post-LSTM. The Pre-LSTM goes T_x time steps. The Post-LSTM goes T_y time steps.

2.1 Data Set

Music Score is difficult to pick-up, because most of the Music Score is written in the PDF. Therefore, there is a great to pick-up the Major Melody from MIDI. MIDI is composed of 128 notes, which is represent of -1 musical scale to 9 musical scale.[9] It is given in the Table 1. We also collect the lyrics in contact of MIDI. Both of Score and Lyrics in some popular music has the repeating part. So, we cut them into two to reduce the amount of data.

Table 1. The architecture of MIDI

ID	0	1	2	…	126	127
Musical Scale	-1	-1	-1	…	9	9
Pitch	C	C#	D	…	F#	G

2.1.1 Notes Embeddings in Scores

So, the core idea here is that we want both of Notes that have their similar neighbors, to be similar in this new virtual representation. For Notes, Different style of Music has its own rule, such as different kinds of musical chord.[10] So, every Note have the most probable neighbors in different case.

Basically, the way usually use these Notes is push one-hot vector through Neural Network which have been trained by the 5 neighbors Notes.[11] Then, there is a weight matrix W multiply by the one-hot matrix X to get the vector which represent Notes or Texts.

$$\ll Notes \gg = WX = \sum w_i x_i \tag{1}$$

Finally, we plot the distribution in 2-D which is reduced dimension by PCA. We get some similar neighbors of Notes, for example, ID 71 which is B is most similar with ID 52 which is E etc. It's shown in the figure 1.

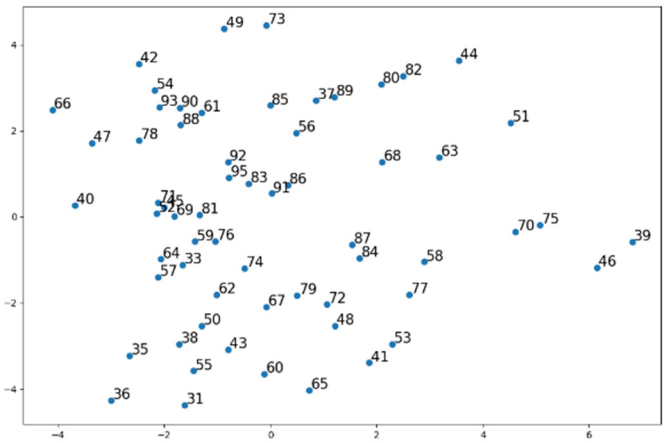

Figure 1. The Distribution of MIDI

2.1.2 Characters Embeddings in Lyrics

Because every Chinese character have its own meaning, especially in Lyrics and Poem. So, based on Chinese feature, we use single Chinese characters instead of words as elements to train embedding matrix.

On the one hand, using single Chinese characters as training elements will reduce the amount of data and make every character to express its own meaning. Chinese common-used Characters only have 3,500. It's much less the common-used word. On the other hand, it will bring some new combinations which do not have in Chinese words. But somehow, in Chinese new words are very common, especially in the internet. Because the parameter of embeddings depends on the amount of input, this way takes the input from thousands of words to nearly 3,500, which can reduce the most of parameter.

2.2 Slide-attention

Inner-Attention is powerful to help us to understand what happen in side of LSTM. The post-attention LSTM passes the output activation $s^{\langle t \rangle}$ and the hidden cell state $c^{\langle t \rangle}$. At time t, LSTM only takes $s^{\langle t \rangle}$ and $c^{\langle t \rangle}$ as input.

Traditionally, at step t, given all the hidden states of the Pre-LSTM which is $\left[a^{\langle 1 \rangle}, a^{\langle 2 \rangle}, \ldots, a^{\langle T_x \rangle}\right]$ and the previous hidden state of the Post-LSTM which is $s^{\langle t-1 \rangle}$. We will compute the attention weight and output the context vector.

$$context^{\langle t \rangle} = \sum_{t'=0}^{T_x} \alpha^{\langle t,t' \rangle} a^{\langle t' \rangle} \tag{2}$$

But, in this way, a simple popular music has more than 500 input, it means that $T_x \geq 500$, it takes too much parameter to compute. Because lyrics just communicate with some music notes but not all the music notes. So, we propose the Slide-Attention architecture. The Slide-Attention is using the Slide-Window to select the several hidden state such as $\left[a^{\langle t \rangle}, a^{\langle t+1 \rangle}, \ldots, a^{\langle t+n \rangle}\right]$ instead of using the all the hidden state and concating the previous hidden state become $s^{\langle t-1 \rangle}$. And then we will compute the attention weight and output the context vector.

$$context^{\langle t \rangle} = \sum_{t'=t}^{t+n} \alpha^{\langle t,t' \rangle} a^{\langle t' \rangle} \tag{3}$$

In the equation, n is the window's length which is depend on T_y. For example, the Pre-LSTM is T_x and the length of Post-LSTM is T_y, so that the length of the Slide-Windows $n = T_x - T_y + 1$. This way changes the input of the Inner-Attention for all the hidden state into several hidden state which just communicate the Post-LSTM output.

3. Experimental Results

While training we can see the speed of every turn of the Slide-Attention is faster than the traditional Inner-Attention. Figure 2 shows that before the epoch 10, there are no difference between two architectures, and after the epoch 10, the Slide-Attention looks a lot faster than the traditional Inner-Attention. And we plot the loss, and we also can see that convergence rate is faster than the traditional Inner-Attention. Therefore, using several hidden states instead of using all hidden states can optimize the traditional model.

Since the problem has a fixed output, different step of the output has different softmax units to generate the different Chinese characters of the output. We visualize the attention values in the network. On the generate plot we can observer the values of the attention weight for each character of the predicted output.

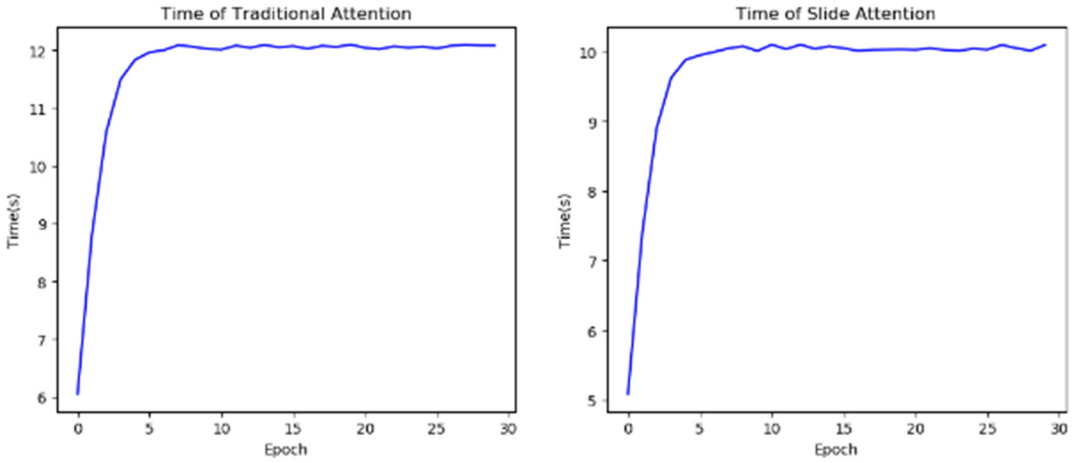

Figure 2. Compare Traditional Attention with Slide Attention

4. Conclusion

In this paper, we propose a new architecture with Slide-Window, which looks like 1-D convolution but not just 1-D convolution, and cencats the previous hidden state. From the above viewpoint, we verify that this architecture can improve the computing speed in the Attention. And in the Automatic Composition of Music Application, we will observer that most of central music note help predict the words, and chord music note hasn't much impact on the predicting the words. It looks the same as our music rule, and this may the new way to achieve automatic composition.

Acknowledgments

(1)Premier Key-Discipline Enhancement Scheme Supported by Guangdong Government Funds ,2016GDYSZDXK036.

(2) Innovation and Entrepreneurship Training Program for College Students in Guangdong Province, Music Generator Based on Machine Learning, 201813684026S.

(3) Advantage Disciplines in Zhuhai City,2015YXXK02.

References

[1]. Graves, Alex. Long Short-Term Memory. Supervised Sequence Labelling with Recurrent Neural Networks. Springer Berlin Heidelberg, 2012:1735-1780.

[2]. Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.

[3]. Boenn, Georg, et al. "Automatic Composition of Melodic and Harmonic Music by Answer Set Programming." International Conference on Logic Programming Springer-Verlag, 2008:160-174.

[4]. Dictionary, Oxford English. "Oxford english dictionary." Retrieved May 30 (2008): 2008.

[5]. Bahdanau, Dzmitry, K. Cho, and Y. Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate." Computer Science (2014).

[6]. Sutskever, Ilya, O. Vinyals, and Q. V. Le. "Sequence to Sequence Learning with Neural Networks." 4(2014):3104-3112.

[7]. Liu, Yang, et al. "Learning Natural Language Inference using Bidirectional LSTM model and Inner-Attention." (2016).

[8]. Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078 (2014).

[9]. Selfridge-Field, Eleanor. Beyond MIDI: the handbook of musical codes. MIT Press, 1997.

[10]. Music. "Alfred Over the Rainbow from The Wizard of Oz Vocal, Piano/Chord Sheet Music." Virtual Sheet Music Inc.

[11]. Bengio, Yoshua, et al. "Neural Probabilistic Language Models." Journal of Machine Learning Research 3.6(2003):1137-1155.