The LSTM Neural Network Based on Memristor

Ziqi Chu 1, Hui Xu 2 and Haijun Liu*

1College of Electronic Science, National University of Defense Technology, Changsha, Hunan, 410073, China
2College of Electronic Science, National University of Defense Technology, Changsha, Hunan, 410073, China
*Corresponding author’s e-mail: liuhaijun@nudt.edu.cn

Abstract. The recurrent neural network adds the concept of time series on the basis of the traditional multi-layer feedforward neural networks, provides the memory function, and makes the network show good modeling ability on time-series data. Therefore, this paper proposes a LSTM (Long Short-Term Memory) neural network based on memristor. It establishes a discrete weighted LSTM network model by simplifying the traditional recurrent neural network, and uses memristor arrays on the premise of ensuring recognition performance. We realize the function of weight matrix to improve the structure of LSTM neural network, and finally carry out simulation research on the proposed neural network. And due to the volatility and yield of memristors, this paper also demonstrates and analyzes the impact of these two characteristics on network performance, and the performance level of the LSTM neural network based on memristor is verified under the existing preparation level. Experiments on the TIMIT speech database show that the proposed neural network in this paper has good accuracy and its speech recognition performance is superior.

1. Introduction

In the field of deep learning, recurrent neural network (RNN) is one of the widely studied deep network models with information feedback. Compared with feedforward neural networks, recurrent neural networks incorporate the concept of time series, which provides learning and memory functions for neural networks, maintains long-term dependence on time series, and has good modeling capabilities for time-series scenarios. However, because the recurrent neural networks use the back propagation algorithm to adjust the model parameters, when there are pretty many hidden layers, the gradient explosion or gradient disappearance may occur, which will affect the prediction accuracy of the model. To this end, Hochreiter and Schmidhuber proposed the Long Short-Term Memory Recurrent Neural Network (LSTM) algorithm [1]. Compared to RNN, LSTM adds a forget gate, an input gate to control the state of the cell, and an output gate to control the output of the cell in the neuron, which suppresses the gradient disappearance to a certain extent and maintains the information dependency for a longer period. The LSTM model can be used to process and predict events with long intervals and delays in the time series. At present, the LSTM neural network has been widely applied to speech recognition, machine translation, text classification, computer vision and other fields[2],[3],[4].

The implementation of LSTM neural network is mainly based on CMOS devices, resulting in the waste of space and excessive power consumption. While the appearance of memristor provides a good solution. Compared with the traditional neural networks based on CMOS devices, the memristor based
neural network has the characteristics of small size, low power consumption and high integration \[5\], and has multi-resistance characteristics, which is very suitable for the weight matrix circuit device in the LSTM neural network.

Memristors have achieved extensive research results in CNN, SNN and other aspects. For example, the convolutional neural network chip based on memristors improves the performance by 7.7 times under the premise that the area is basically equivalent \[6\]. However, the current research on recurrent networks is carried out on the basis of continuous weights, and the lack of recurrent neural network models based on discrete weights is difficult to achieve through memristors. Therefore, it is pretty important to consider how to properly simplify the neural network structure under the condition that the network recognition performance is not degraded, so that it is significant to implement the memristor array. At the same time, the memristor array also has non-ideal characteristics such as volatility and yield. It’s obvious that these problems need to be considered when analyzing the performance of the artificial neural networks based on memristors. In this paper, the LSTM neural network model based on memristor is constructed first, and then the effect of the yield and volatility of the memristors on the performance of the proposed neural network is analyzed.

2. The LSTM neural network Based on Memristor

2.1. LSTM neural network based on memristor

The key of LSTM neural network is to set cell state to process time series information. At the same time, LSTM neurons design three kinds of gated units to remove or add elements to cell state. And the LSTM neural network are mainly composed of three gated units, such as the forget gate, input gate and output gate. The unique gated units are used to learn and memorize the sequence data to maintain long-distance time series information dependence and achieve high-precision prediction.

As shown in Figure 1, the LSTM neuron has the input gate, output gate, and forget gate. The input gate mainly processes input data. The forget gate determines the current neuron's retention of historical information. The output gate represents the output result of the neuron. Suppose the input sequence is \((x_1,x_2,\ldots,x_t)\), then at time \(t\), the calculation formula of each parameter of LSTM neuron is as follows:

\[
i_t = S(W_i *[h_{t-1},x_t] + b_i) \tag{1}
\]

\[
f_t = S(W_f *[h_{t-1},x_t] + b_f) \tag{2}
\]

\[
o_t = S(W_o *[c_t,h_{t-1},x_t] + b_o) \tag{3}
\]

Figure 1. The structure of a memristor-based LSTM cell.
where $i_t, f_t,$ and $o_t$ are the input of the input gate, the forget gate, and the output gate at time $t$ respectively; $x_t$ denotes the input to the LSTM neuron at time $t$, $h_{t-1}$ indicates the output state of the hidden layer at time $t-1$; $b_i, b_f,$ and $b_o$ are their corresponding offset vectors; then $h_t$ represents the output of the hidden layer at time $t$; $S$ is the Sigmoid function; $W_i, W_f$ and $W_o$ are respectively the weight matrix of the input gate, the forget gate and the output gate of the neuron, $W_c$ is the weight between the input and the cell unit.

In the LSTM network, the concatenated vector is multiplied by the weight matrix and obtained outputs go through activation functions to form gate values. Memristors with variable resistance and capability of remembering its resistance state are ideal devices for weight matrix hardware. Therefore, memristor-based LSTM network realizes the function of weight matrix through memristor crossbar array[7], [8]. Memristor has two states $R_{on}$ and $R_{off}$ that can be controlled by applied voltage amplitude and pulse duration, we use this property to realize the representation of positive and negative numbers in the weight matrix. To be specific, it can be represented as a difference of two memristor conductances[9].

Figure 2 shows the structure of LSTM neural network with 3 hidden layers. The deep LSTM neural network is based on the LSTM neural unit, and thus constructs the network model with multiple hidden layers, and continuously removes redundant information in the data set through the forget gates to maintain long-distance dependencies, so it has stronger predictive performance and better generalization ability.

2.2. The memristor-based LSTM network voice recognition processing flow

In this paper we firstly use MFCC(Mel frequency cepstrum coefficient)[10] to process the feature of speech samples. The input speech information will be preprocessed into a matrix, so that it can act as the input vector of the LSTM network.
Figure 3. The flow of samples processing

Figure 4 shows the architecture of LSTM neural network based on memristor. The proposed neural network architecture is mainly divided into two parts. The first part is the basic LSTM neural network. The input vector $x_t$ of the network at each time is 13 dimensions, the output vector $h_t$ is 10 dimensions, and the number of network parameters is: $4 \times ((n + m) \times m + m)$, where $n=13$ is embedding dim, and $m=10$ is hidden size of the LSTM network. The second part is a softmax function layer, which classifies the output vectors and judges the recognition results.

Figure 4. The structure of a speech recognition LSTM network
The LSTM neural network based on memristor firstly obtains the mathematical LSTM neural network with continuous weight through the training process of LSTM neurons, then uses memristor to train the weight matrix of the network in data set, so as to determine the value of the weight matrix, and then discretizes the value of the weight matrix, so as to realize the modeling process of the performance of LSTM neural network based on memristor.

3. Experimental results and analysis
Firstly, this paper designs an LSTM neural network which can recognize and classify the speech information of digit 0-9. The samples are selected from the TIMIT speech database. The length of speech samples are cut to 1s, and the frame length is 25ms, frame shift is 10ms. MFCC is used to process the feature of the sample. The feature value of each frame is extracted, and a vector sequence of 1 * 13 with length of 100 is obtained. [input_shape=(100,13)].

3.1 Effect of the resistance states
In order to realize the simulation of LSTM neural network based on memristor, we first select the appropriate number of memristor’s resistance states, set the number of resistance states as \( a \), and the maximum weight is \( w_{\text{max}} \) and minimum weight is \( w_{\text{min}} \). The weight range is divided into \( a \) intervals, then when the weight of the network is in the \( n \)th interval, the weight size can be shown in (6).

\[
\frac{(2n + 1)(w_{\text{max}} - w_{\text{min}})}{2a} + w_{\text{min}} \quad (6)
\]

Take the number of resistance states \( a \) as 2, 4, 8, 16 respectively, and conduct 20000 cycles on the network, and the simulation results are as follows:

| the number of \( a \) | 2   | 4   | 8   | 16  |
|-----------------------|-----|-----|-----|-----|
| Recognition rate      | 31.68% | 80.27% | 98.07% | 99.13% |

Table 1 shows that the recognition rates of the LSTM network under different number of resistance states. It can be seen that when the number of resistance states of the memristor is 2 or 4, the recognition accuracy of the network is poor; when the resistance state of the memristor is more than 8, the recognition performance of the network is improved significantly, and when the resistance state is 16, it is not significantly improved compared with 8. Therefore, when the resistance state of the memristor is 8, the non ideal characteristics of the memristor are simulated.

3.2 Effect of the variation
Due to the limitation of fabrication technology, the resistance of memristor synapses fluctuates to some extent. In LSTM, when the resistance value of the memristor used to simulate artificial synapse fluctuates in different degrees, the calculation formula of the resistance value is as follows:

\[
R = A \times \left( 1 - \frac{B}{2} \right), \quad A \times \left( 1 + \frac{B}{2} \right) >
\]

Among them, \( A \) is the resistance value of the memristome synaptic device in the ideal status, \( B \) is the variation, and the resistance value of the memristome synaptic device obeys the normal distribution in this range. When the maximum value of the memristor synapse device is taken, the fluctuation of the resistance causes little change in the conductivity, that is to say, it has little influence on the current flowing through the synapse, which can be ignored. Therefore, this paper only considers the influence of the fluctuation of the minimum resistance on the recognition rate of the LSTM.
At this time, we can set the synapse to be undamaged, that is, the yield is 100%. Set different volatility to test the recognition rate of LSTM, and the experimental results are shown in the table 2:

| Variation | 5%    | 10%   | 15%   | 20%   | 25%   |
|-----------|-------|-------|-------|-------|-------|
| Recognition rate | 97.89% | 94.53% | 89.28% | 81.34% | 70.37% |

It can be concluded from the table 2 that the recognition rate of the LSTM network based on memristor decreases gradually with the increase of the volatility of the memristor resistance value. When the volatility of the resistance value reaches 15%, the recognition rate of the network drops below 90%.

### 3.3 Effect of the yield

In practice, the memristor synaptic devices in LSTM may be partially damaged, which can be divided into three situations: 1) Memristor in high resistance state; 2) Memristor in low resistance state; 3) Memristor in opposite state. Assuming that there is no fluctuation in the resistance value of the memristor at this time, it is defined in this topic that when the memristor is damaged, the resistance value is always in the high resistance state, and different yield is set to test the recognition rate of the pulse neural network. When the fluctuation rate is 0, the results are as follows:

| Yield  | 100%   | 99%    | 98%    | 97%    | 96%    |
|--------|--------|--------|--------|--------|--------|
| Recognition rate | 98.07% | 97.15% | 95.86% | 94.36% | 91.13% |

It can be seen from table 3 that the recognition rate of the network decreases with the decrease of yield. When the yield of the memristor array is 96%, the recognition rate of the network decreases to about 90%.

### 4. Conclusion

In this paper, by simplifying the mathematical LSTM network model, the LSTM neural network model based on memristor with discrete weights is established on the premise of ensuring the recognition performance. And the weight matrix of the simplified LSTM neural network is mapped to the memristor array, so that the function of the weight matrix can be realized through the memristor array. In the experiment on TIMIT speech database, we analyze the influence of the number of memristor resistance states on the recognition accuracy of neural network, and discuss the influence of the memristor’s volatility and yield on the recognition rate. The final experimental results verify the accuracy and performance of the proposed LSTM neural network based on memristor.

### Acknowledgments

This research was supported in part by the National Key Research and Development Plan of MOST of China under Grant No. 2019YFB 2205100, the National Natural Science Foundation of China (No. 61471377, 61804181, 61604177 and 61704191) and NUDT Science Support Program (No. JC-15-04-02).

### References

[1] Wu J, Long J, Liu M. Evolving RBF neural networks for rainfall prediction using hybrid particle swarm optimization and genetic algorithm[J]. Neurocomputing, 2015, 148: 136-142.
[2] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, “Convolutional lstm network: A machine learning approach for precipitation nowcasting,” in Advances in neural information processing systems, 2015, pp. 802 – 810.

[3] A. Graves, N. Jaitly, and A.-r. Mohamed, “Hybrid speech recognition with deep bidirectional lstm,” in Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. IEEE, 2013, pp. 273 – 278.

[4] D. Soutner and L. Muller, “Application of lstm neural networks in language modelling,” in International Conference on Text, Speech and Dialogue. Springer, 2013, pp. 105 – 112.

[5] Podani, J. (1994) Multivariate Data Analysis in Ecology and Systematics. SPB Publishing, The Hague.

[6] Hu Fei, You Zhiqiang, Liu Peng, Kuang Jishun. Design of Convolutional Neural Network Circuit Based on Memory Cross Array [J]. Computer Research and Development,2018,55(05):1097-1107.

[7] K. Smagulova, O. Krestinskaya, and A. P. James, “A memristor-based long short term memory circuit,” Analog Integrated Circuits and Signal Processing, vol. 95, no. 3, pp. 467 – 472, 2018.

[8] A. Irmanova and A. P. James, “Multi-level memristive memory with resistive networks,” in Postgraduate Research in Microelectronics and Electronics (PrimeAsia), 2017 IEEE Asia Pacific Conference on. IEEE, 2017, pp. 69 – 72.

[9] R. Hasan, T. M. Taha, and C. Yakopcic, “On-chip training of memristor crossbar based multilayer neural networks,” Microelectronics Journal, vol. 66, pp. 31 – 40, 2017.

[10] Davis, S. and Mermelstein, P. (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing, 28, 357-366.