Rolling Bearing Fault Diagnosis Based on Single Gated Unite Recurrent Neural Networks

Wenwen Tan¹, Yuansheng Sun², Dawei Qiu³*, Yapeng An¹ and Ping Ren¹
¹ College of Information Science & Technology, Beijing University of Chemical Technology, Beijing 100029, China
² Institute of High Energy Physics, Chinese Academy of Science, Beijing 100049, China
³ Beijing Key Laboratory of Mobile Computing and Pervasive Device (Institute of Computing Technology, Chinese Academy of Sciences), Beijing 100190, China

corresponding author: Dawei Qiu
tanwenwen90@163.com, syuans1989@126.com, qiudawei@ict.ac.cn, 850016457@qq.com, 2018200773@mail.buct.edu.cn

Abstract: In order to meet the real-time requirement of bearing fault diagnosis, a simplified strategy of the traditional long short term memory (LSTM) neural network structure is proposed. We designed a new single gated unite (SGU) recurrent neural network. In view of the non-stationary and non-linear characteristics of bearing fault vibration data, we use wavelet packet decomposition to extract the features as input signals of bi-directional single gated unite (Bi-SGU) to complete the diagnosis of 10 types of bearing data. Simulation results show that the proposed method can ensure the accuracy of bearing fault diagnosis, reduce the number of network parameters by 36%, and improve time efficiency by 41%.

1. Introduction
Rolling bearings are important components of rotating machinery. Bearing fault is one of the main causes of rotating machinery fault [1-3]. Affected by multiple factors, rolling bearings are prone to failure, causing problems, such as delays in production, increased repair costs, and even casualties. Therefore, it is necessary to carry out fault diagnosis research on rolling bearings.

Technologies used for bearing fault detection can be divided into model-based methods and data-driven methods, or subdivided into methods based on physical models, statistical models, artificial intelligence and hybrid technologies. Among the artificial intelligence methods, the data-driven artificial neural network method has attracted the attention of many researchers due to its good data learning ability and generalized reasoning ability. In [4], researchers used empirical mode decomposition (EMD) to extract energy in different frequency bands of bearing vibration signals as the input of artificial neural networks for bearing fault diagnosis. Paper [5] used the energy moment eigenmode function to extract feature vectors, then fed into BP neural network. Paper [6] compared the performance of three artificial neural networks in bearing fault diagnosis including single-layer perceptron (MLP), radial basis function (RBF) and probabilistic neural network (PNN). Reference [7] used convolutional neural network (CNN) to directly process raw bearing vibration data. Reference [8] used deep convolutional neural networks for bearing fault diagnosis. Reference [9] used a convolution-gated recurrent network on a small sample
data set, achieving an accuracy of 97%. The disadvantage of above methods is cannot use the time attribute of the bearing data which is closely related to bearing fault diagnosis. The recurrent neural network (RNN) has a good ability to process sequence data, which can solve this problem. Reference [10] made a full comparison of the performance of BP neural network and traditional RNN in fault diagnosis. The results show that RNN has an improvement in algorithm convergence speed, accuracy and algorithm stability than BP neural network. However, one of the problems of training the traditional RNN is data vanishing gradient. As a special type of RNN, long short term memory network (LSTM) [11] not only has the ability to process sequence data, but also can alleviate the problem of vanishing gradient existing in traditional RNN. Literature [12] used LSTM for aero engine fault diagnosis and remaining life prediction, achieving good results.

Although the standard LSTM has good performance, the network needs more run time relatively, which does not meet the real-time requirement of bearing fault diagnosis. The reason is that LSTM has three “gates”, each “gate” needs to learn its own parameters. The problem above also causes the high carbon emissions of computers, making artificial intelligence been questioned in terms of environmental protection. In view of the above problems, we design an SGU structure which constitutes Bi-SGU for bearing fault diagnosis. Under the premise of ensuring the accuracy of fault diagnosis, the number of network parameters is reduced by 36% and the running time is reduced by 41%.

In summary, aiming to improve the performance of rolling bearing fault diagnosis, this paper proposes a novel structure of simplified Bi-LSTM neural network. The main contributions of the paper are as follows:

• We propose a simplified LSTM neural network named SGU (single gated unite recurrent neural network). SGU only uses one gate for fault diagnosis to reduce parameters and to save run time.
• We propose to use Bi-SGU for fault diagnosis to make full use of information at every time step.
• A rolling bearing fault diagnosis architecture of Bi-SGU recurrent neural network combined with wavelet packet decomposition is proposed.

2. Bearing fault diagnosis
The bearing fault diagnosis processing in this paper is shown in figure 1. After the vibration data is pre-processed, using the wavelet packet decomposition to extract the energy of time-frequency feature, which is standardized as the input of a neural network for fault diagnosis.

![Figure 1. Framework of rolling bearing fault diagnosis](image-url)
2.1. Feature extraction

The original rolling bearing vibration data is shown in figure 2. Compared with normal data, the peak value of the fault vibration data increased obviously, the inner race and the outer race fault data have a certain periodicity. Since time-domain vibration signal cannot provides comprehensive information for fault diagnosis, it is necessary to consider frequency-domain information. In addition, the vibration signal caused by local bearing failure is non-stationary, and its statistical characteristics are a function of time. The Fourier spectrum analysis method is not suitable for the analysis of non-stationary signals, so a suitable method needs to be selected for processing. Wavelet packet decomposition is a theory derived from wavelet transform. It has good time-frequency localization ability and is suitable for processing transient signals [13]. Literature [14] used the combination of wavelet packet and AR spectrum to extract the bearing signal features to improve the diagnosis accuracy. Therefore, this paper uses wavelet packet decomposition as the feature extraction method. Taking one-dimensional signal S as an example, the three-level wavelet packet decomposition is shown in figure 3. In this paper, Daubechies 10 and Symlets 8 wavelets are used to perform five-layer wavelet packet decomposition on rolling bearing vibration data. Energy is calculated from wavelet coefficients and normalized as the input feature. Each sample has 32 features. The energy formula is shown in (1). First six features are shown in figure 4.

\[ E(j) = \left( \sum_{j=0}^{N} c^2 \right)^{1/2} \]  

(1)

In the formula, \( N = 2n \), \( n \) is the number of layers of wavelet packet decomposition, and \( j = 0,1,2 \ldots \), \( N \), \( c \) are coefficients of wavelet packet decomposition.
2.2. Network Architecture

A traditional standard LSTM structure with peephole is shown in figure 5. Where, $x_t$ is the input at time $t$, $h_{t-1}$ is the output at time $t-1$, $x_t$ and $h_{t-1}$ decide the value of $c_t$ which is the candidate memory at time $t$. LSTM contains three "gate" structures including "forget gate", "input gate" and "output gate". "Gate" is a method to select information. The values of the three "gates" depend on $x_t$, $h_{t-1}$ and $c_{t-1}$. The activation function of "gate" is sigmoid. The output value of "gate" is between 0 and 1. The standard LSTM formulas are shown in (2) ~ (6):

$$c_t = \tanh(W_{ce} \cdot x_t + W_{ch} \cdot h_{t-1} + b_c)$$  \hspace{1cm} (2)
$$f_t = \text{sigmoid}(W_{cf} \cdot x_t + W_{fh} \cdot h_{t-1} + p_f * c_{t-1} + b_f)$$  \hspace{1cm} (3)
$$i_t = \text{sigmoid}(W_{ci} \cdot x_t + W_{hi} \cdot h_{t-1} + p_i * c_{t-1} + b_i)$$  \hspace{1cm} (4)
$$o_t = \text{sigmoid}(W_{co} \cdot x_t + W_{ho} \cdot h_{t-1} + p_o * c_t + b_o)$$  \hspace{1cm} (5)
$$c_t = \tilde{c_t} * i_t + c_{t-1} * f_t$$  \hspace{1cm} (6)
$$h_t = o_t * \tanh(c_t)$$  \hspace{1cm} (7)

where, $W_{ce}, W_{ch}, W_{cf}, W_{fh}, W_{ci}, W_{hi}, W_{co}, W_{ho}$ are parameters that the network needs to learn. Sigmoid and tanh are non-linear activation functions. The formulas are (7) and (8), respectively:
Although the standard LSTM structure in figure 5 has good performance in many fields, problems in bearing fault diagnosis appear. Due to the three "gates" of LSTM need to learn parameters respectively, resulting in longer network running time relatively. The result does not meet the real-time requirement of bearing fault diagnosis. The above problem also causes high carbon emissions of computers, making artificial intelligence been questioned in terms of environmental protection. The reason is that LSTM was originally designed to handle the complex logic of natural language. But the logic of frequency domain energy feature is relatively simple. While the bearing fault diagnosis task needs higher real-time performance. Therefore, a corresponding simplified strategy is proposed, a new structure SGU with only one "gate" is designed as shown in figure 6, Formulas are shown in (10) ~ (13).

\[
sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{8}
\]

\[
tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{9}
\]

\[
\begin{align*}
\tilde{c} &= \tanh(W_c \cdot x_t + W_{hc} \cdot h_{t-1} + b_c) \\
g_t &= \text{sigmoid}(W_g \cdot x_t + W_{hg} \cdot h_{t-1} + p_g \cdot c_{t-1} + b_g) \\
c_t &= (\tilde{c} + c_{t-1}) \cdot g_t \\
h_t &= \tanh(c_t) \tag{13}
\end{align*}
\]

where, \( W_c, W_{hc}, W_g, W_{hg} \in \mathbb{R}^{N \times N}, p_g, b_c, b_g \in \mathbb{R}^N \) are the parameters that the network needs to learn.
Considering that SGU can only use current and previous information, we propose bi-directional SGU (Bi-SGU) for bearing fault diagnosis to make full use of information at every time step. The neural network architecture is shown in figure 7, including an input layer, a forward SGU layer, a backward SGU layer, and an output layer. Time-frequency energy features of the bearings obtained by wavelet packet decomposition are processed by the input layer, and then used as the input of the forward SGU and the backward SGU. The forward SGU layer and the backward SGU layer are composed of multiple SGU structures.

3. Experimental setup and analysis

3.1. Experimental data
The experimental data used in this paper is an open data set provided by the Bearing Center of Case Western Reserve University in the United States. The data acquisition platform is shown in figure 8. The tested bearings were equipped with EDM technology to arrange single-point faults, including inner race fault, outer race fault and ball fault. Each fault has three different fault size. Including 0.007 inches, 0.014 inches, 0.021 inches, as shown in table 1. An accelerometer with a bandwidth of up to 5000 Hz is installed on the end cover of the experimental motor, and the vibration data of the test bearing under different working conditions is collected by a recorder with a sample frequency of 48 kHz.

![Figure 8. Data collection platform](image)

| No. | Fault Diameter | Fault Location |
|-----|----------------|----------------|
| 1   | -              | Normal         |
| 2   | 0.007          | Inner Race     |
| 3   | 0.007          | Outer Race     |
| 4   | 0.007          | Ball           |
| 5   | 0.014          | Inner Race     |
| 6   | 0.014          | Outer Race     |
| 7   | 0.014          | Ball           |
| 8   | 0.021          | Inner Race     |
| 9   | 0.021          | Outer Race     |
| 10  | 0.021          | Ball           |

For the original bearing data, 2048 is used as a sample length to perform continuous and non-repeated truncation, obtaining 2420 samples totally. Using the method proposed in this paper to diagnose above 10 kinds of bearing data.

3.2. Neural network settings
In this paper, 2420 samples are divided into training set, validation set and test set, with sizes of 1620, 400 and 400. The Batch size is 100. The number of neurons is 32 in the input layer, and neurons in the forward-backward SGU network is 100. There are 10 neurons in the output layer.
3.3. Model of fault diagnosis

In order to verify the effectiveness of the structure designed in this paper, experiments are performed using the same training set, validation set and test set. In order to eliminate the influence of random factors on the experiment, the test results in table 2 are the average of 10 experiments. Test set size is 400. From figure 9 and table 2, it can be found that the number of parameters and the test time are significantly reduced after using a single-layer SGU compared with use a single-layer LSTM. But the train loss of SGU and LSTM are about 0.5, and the test accuracy are about 99%. The result is not very satisfactory. After using Bi-LSTM and Bi-SGU, as the iteration increases, the train loss of the model decreases gradually. It can be found that when train iteration is about 150, the loss is close to zero, tends to be stable. Test accuracy has improved significantly.

![Figure 9. The train-loss.](image)

| Algorithm | Accuracy (%) | Parameters | Test Time (s) |
|-----------|--------------|------------|--------------|
| LSTM      | 98.5         | 81610      | 0.3446       |
| SGU       | 98.48        | 42510      | 0.2187       |
| Bi-LSTM   | 99.518       | 225610     | 0.7652       |
| Bi-SGU    | 99.58        | 144810     | 0.4514       |

The results show that Bi-SGU guarantees the accuracy of fault diagnosis, and the network parameters and model run time are reduced significantly compared to Bi-LSTM. Where, Bi-LSTM has 225610 parameters, run time is 0.7652 seconds, and Bi-SGU has 144810 parameters, run time is 0.4514 seconds. The number of parameters is reduced by 36%, and the network operation time is reduced by 41%.

3.4. Comparison with other algorithms

BPNN, SVM, CNN and the Bi-SGU method proposed in this paper are compared in table 3. The results show that the accuracy of Bi-SGU is 99.58%, which higher than other algorithms.

| Algorithm | Accuracy (%) | Algorithm | Accuracy (%) |
|-----------|--------------|-----------|--------------|
| BPNN      | 96.35        | CNN       | 98.81        |
| SVM       | 88.56        | Bi-SGU    | 99.58        |

3.5. The influence of Wavelet Packet Decomposition layer number

The function of wavelet packet decomposition is to extract the time-frequency features of bearing data. The feature is the key of the accuracy of rolling bearing fault diagnosis. If features are not capable of expressing the original data, the fault diagnosis cannot be accurate. The diagnosis accuracy varies with the number of wavelet packet decomposition layers. The higher the layers, the finer the extracted features, but the feature dimensions also increases with the number of wavelet decomposition layers, which will affect the accuracy and the time of fault diagnosis. So different layer numbers of wavelet packet decomposition are tested. As shown in table 4, the fault diagnosis accuracy gradually increases when the number increases from 3 to 5. When the number of decomposition layers is 6 to 9, the accuracy does not improve significantly. When the number of layer is 10, the accuracy decreases significantly. In
addition, the more layers of wavelet packet decomposition, the longer time the network need to run. After comprehensive consideration of time and accuracy, a 5-layer wavelet packet decomposition is chosen.

| Layer | accuracy (%) | parameters | test time (s) |
|-------|--------------|------------|---------------|
| 3     | 95.52        | 96810      | 0.2401        |
| 4     | 99.35        | 112810     | 0.2927        |
| 5     | 99.58        | 144810     | 0.4100        |
| 6     | 99.47        | 208810     | 0.4980        |
| 7     | 99.45        | 336810     | 0.8907        |
| 8     | 99.48        | 592810     | 1.6991        |
| 9     | 99.45        | 1104810    | 3.7595        |
| 10    | 98.75        | 2128810    | 8.5265        |

### 3.6. Fault diagnosis on variable load

The bearing load is changed in the actual bearing operating environment. To verify the robustness of the model, experiments are performed under three different load data, load1, load2, and load3. The diagnostic results are shown in table 5.

| load   | accuracy (%) | Load1 | Load2 | Load3 |
|--------|--------------|-------|-------|-------|
| Load1  | 99.48        | 98.70 | 98.36 |       |
| Load2  | 99.30        | 99.56 | 95.42 |       |
| Load3  | 98.80        | 97.30 | 99.58 |       |

It can be known from table 5 that the method proposed in this paper can still achieve an accuracy about 95% under a variable load, which shows that Bi-SGU model has a good generalization ability.

### 4. Conclusion

Rolling bearing is an important part of the rotating system. Fault of rolling bearings will cause huge economic losses and increase security risks. Moreover, the working conditions of rolling bearings are harsh, the state of the bearings may change at any moment. Therefore, it is necessary to ensure the accuracy of bearing fault diagnosis and improve the time effect. In this paper, a novel SGU recurrent neural network structure is designed. The experiment proved that the structure can reduce the number of network parameters by 36% and the running time by 41% while ensure the fault diagnosis accuracy. This work provides a good idea for rolling bearing fault diagnosis. In the future, we will study how to improve the accuracy of bearing fault diagnosis, and continue to study how to further improve the time efficiency.

### References

[1] Jiang L, Shi T L and Xuan J P. Fault diagnosis of rolling bearings based on Marginal Fisher analysis[J]. Journal of Vibration and Control, 2014, 20(3): 396-418.

[2] Van H B, Qu Y Z and He D. Bearing fault diagnosis based on a new acoustic emission sensor technique[J]. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, 2015, 229(2):105-118.

[3] Xu J, Tong S G, Cong F Y and Zhang Y D. The application of time-frequency reconstruction and correlation matching for rolling bearing fault diagnosis[J]. Proceedings of the Institution of Mechanical Engineers, Part C. Journal of mechanical engineering science, 2015(17): 3291-3295.

[4] Yang Y, Yu D J and Cheng J S. A roller bearing fault diagnosis method based on EMD energy
entropy and ANN[J]. Journal of Sound and Vibration, 2006, 294(1): 269-277.

[5] Qin T L, Yang Y, Cheng H and Xue S. Rolling bearing fault diagnosis based on intrinsic mode function energy moment and BP neural network[J]. Journal of Vibration, Measurement & Diagnosis, 2008(03): 229-232, 300.

[6] Samanta B S, Al-Balushi K R and Al-Araimi S A. Artificial neural networks and genetic algorithm for bearing fault detection[J]. Soft Computing-A Fusion of Foundations, Methodologies and Applications, 2006, 10(3): 264-271.

[7] Lee D, Siu V, Cruz R A and Yetman C J. Convolutional neural net and bearing fault analysis[C]/Proceedings of the International Conference on Data Mining (DMIN). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2016: 194-200.

[8] Tian Y L and Liu X Y. A deep adaptive learning method for rolling bearing fault diagnosis using immunity[J]. Tsinghua Science and Technology, 2019, 24(6): 750-762.

[9] Yang P and Su Y C. Fault diagnosis of rolling bearing based on convolution gated recurrent network. Journal of Aerospace Power, 2019, 34(11): 2432-2439.

[10] Chen Q R. Comparison between two fault diagnosis methods based on neural network[J]. Proceedings of The Chinese Society for Electrical, 2005(16): 112-115.

[11] Hochreiter S and Schmidhuber J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[12] Yuan M, Wu Y T and Lin L. Fault diagnosis and remaining useful life estimation of aero engine using LSTM neural network[C]/ IEEE International Conference on Aircraft Utility Systems, 2016: 136-241.

[13] Zhang S, Chen J, She L H and Bao X R. Use of optimal wavelet packet decomposition for the long-term prediction of variable-bitrate video traffic[J]. IET Communications, 2010, 4(11): 1277-1287.

[14] He S Y, Liu Y and Tian X C. Research on bearing fault diagnosis based on wavelet packet-auto regressive model spectrum and deep learning[J]. Application Research of Computers, 2019, 36(06): 1758-1761, 1766.