Detection of Rail Bottom Damage Defects Based on Recurrent Neural Network

Fengguang Zhou, Qinjun Zhao, Yuan Xu, Qinhua Xu, and Tao Shen(✉)

School of Electrical Engineering, University of Jinan, Jinan 250022, China
cse_st@ujn.edu.cn

Abstract. Today, the development of railway network has covered most of the country. This not only brings convenience to people, but also undoubtedly brings a lot of dangerous rail damages which will threaten the running safety of trains. One of the best ways to ensure the safety of railway transportation is to detect the possible damaged position in the rail before accidents. Under the circumstance, this paper studies the detection and identification of rail damages, establishes a set of rail bottom damage detection system based on Recurrent Neural Network, the detection accuracy is about 90%. Then, an adaptive moment estimation optimizer is added to the network model to realize the dynamic adjustment of the learning rate, and the training accuracy is also improved to over 95%. This system can quickly and accurately find out the damage at the bottom of the rail, which not only reduces the labor intensity of workers, but also improves the safety factor of railway transportation.

Keywords: Rail damage detection · Ultrasonic detect · Recurrent Neural Network · Adaptive moment estimation optimizer

1 Introduction

For a long time, railway transportation is the main lifeline of national economy. With the development of railway, the operation of a series of trains, high-speed trains and bullet trains has brought new vitality to modern transportation. However, with the increase of transportation process and service time, various kinds of damages will occur on the surface and inside of rail. Under the influence of these damages, coupled with the temperature stress caused by various extreme weather, the rail will suddenly break [1]. There is no doubt that this serious rail break will bring irreversible harm to the economy and personal safety of passengers. Therefore, it's of great significance to find a rapid detection and identification method for rail damage. Fortunately, with the development of electronic technology and computer technology, as well as the upgrading of nondestructive testing technology, it provides the possibility for rail internal inspection [2].

According to the investigation and analysis of NASA, there are about 70 kinds of nondestructive methods, which can be divided into six categories. However, in practical application, there are five common methods: ultrasonic testing, radiographic testing,

S. Wang et al. (Eds.): IoTCare 2021, LNICST 415, pp. 451–460, 2022.
https://doi.org/10.1007/978-3-030-94182-6_36

magnetic particle testing, penetrant testing and eddy current testing. In today's railway rail flaw detection field, ultrasonic flaw detection is more popular than other methods [3]. Compared with other conventional flaw detection methods, this method is suitable for internal defect detection, and has the advantages of wide detection range, high sensitivity, high efficiency, simple operation and low cost.

Railway is one of the earliest non-destructive testing departments in China. Non-destructive testing is a comprehensive applied science and technology, which detects macroscopic and microscopic defects by physical means without changing or affecting the performance of the tested object. Railway is one of the earliest non-destructive testing departments in China. Since the introduction of resonance ultrasonic flaw detector made in Switzerland in 1950s, the electronic components of rail flaw detector have experienced three development stages: tube-transistor-integrated circuit. The signal processing mode of rail flaw detector changes from analog to digital, and the display mode of rail flaw detection develops from A-type display to B-type display.

In China, the work of flaw detection workers in relevant departments is mainly to manually find the damaged parts from the flaw detection maps generated by ultrasonic flaw detection vehicles. However, with the continuous construction of new railways in China, the workload of flaw detection personnel will continue to increase. They need to look for damage from B-type display images with a long mileage on the playback software, which also includes a large number of repetitive B-type display images without value. Workers doing this kind of mechanical and tedious work for a long time will greatly reduce the correct rate of playback detection. Therefore, how to quickly identify and classify the rail flaw detection map is very important.

Fortunately, with the development of artificial intelligence and other related technologies, more and more scholars have applied this technology in all walks of life. In article [4], the author introduces the feature parameter extraction of vehicle audio signal and the recognition algorithm of recurrent neural network. In order to optimize the traditional recurrent neural network model, the feature layer is added to the input layer to improve the model structure. The results of processing the audio data of four vehicle types show that the model can effectively identify different vehicle types, and the recognition accuracy rate exceeds 80%, which can meet the basic recognition requirements. In the article [5], Wang K studied the transmission characteristics of rail damage acoustic emission signals in wheel-rail coupling, and proposed a vehicle-mounted rail damage detection method based on convolutional neural network. Moreover, based on the improved convolutional neural network algorithm of multiple acoustic emission events probability, the classification rules of rail damage stage were established, and the constructed damage feature library was classified and trained, and the recognition rate reached the expected effect. Sun C applied the deep convolution neural network to the damage identification of B-type display images, designed the damage identification network architecture, and trained and fine-tuned the architecture, which made the accuracy of the model superior to the existing system of rail inspection vehicles, reached the index requirements of manual analysis, and improved the detection accuracy [6]. Wang Y combined adaptive moment estimation (Adam) algorithm optimizer with convolutional neural network, obtained better classification effect by adding learning rate correction factor, and at the same time reduced iterative oscillation and insufficient classification

stability of image samples [7]. Paper [8] proposes the first use of Adam algorithm to fast and stably converge large-scale tap coefficients of polynomial nonlinear equalizer. Different from serial least-mean square adaptive algorithm, Adam algorithm is a parallel processing algorithm, which can obtain globally optimal tap coefficients without being trapped in locally optimal tap coefficients. Owing to parallel processing and global optimization, Adam algorithm has much better performance on resisting the timing error. In conclusion, Adam algorithm shows great potential for converging the tap coefficients of PNLE in PAM8-based optical interconnects.

In this paper, the generation principle of B-type display image is analyzed first, and then a set of intelligent detection system for rail bottom damage of flaw detection data based on Recurrent Neural Network is designed, which can automatically eliminate undamaged data and identify damaged parts. Finally, on this basis, adaptive moment estimation (Adam) optimizer is added, which can get better training effect detection accuracy. After a series of training, the accuracy of the system can easily reach and exceed the manual accuracy, and the time required is greatly reduced.

2 Generation Process of Flaw Detection Data

2.1 Rail Flaw Detection Equipment

B-type display image data used in this paper are all from GCT-8C flaw detector produced by Xingtai ultrasonic testing equipment co., ltd., Hebei province, and the detection speed are about 10 km/h. This flaw detector is a small hand-push flaw detection tool with nine ultrasonic probes, which can cover all parts of a rail. As shown in Fig. 1, among the nine probes, six 70° probes are responsible for the data of the rail head, which is the most complicated. Two 37° probes are responsible for collecting the data of rail waist and rail bottom, and the last 0° probe is mainly used to calibrate the horizontal position of the flaw detector, so as to ensure that the other probes correspond to the relevant parts of the rail.

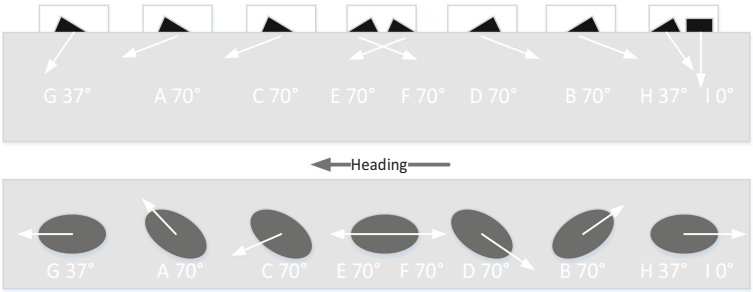

Fig. 1. Probe distribution diagram of flaw detector

2.2 Generation Principle of B-type Display Image

People call the propagation process of sound source vibration in medium (such as air) sound wave, and the mechanical wave with frequency higher than 20 kHz is called ultrasonic wave. In the field of flaw detection, ultrasonic has many advantages, such as good directivity, strong penetrating power, high sensitivity, small energy loss, etc. Moreover, ultrasonic will reflect and refract at heterogeneous interfaces, especially at gas-solid interfaces. Therefore, when the ultrasonic wave propagates to the interface between metal and defect, it will be partially or totally reflected. Each probe of the ultrasonic flaw detector will periodically transmit ultrasonic pulses and receive various reflected echoes at the same time. After the probe receives the reflected echo, it will be amplified and compared with the preset voltage threshold. If the sound pressure of the echo exceeds the current threshold, it means that the echo signal contains damage information, so it is necessary to draw a B-type display image in the output map. In the output result map, there are nine signal channels corresponding to nine ultrasonic probes of monorail flaw detector. The height of the drawing area from channel 1 to channel 6 is 48 pixels, corresponding to 48 sampling points. Starting from the first bit of the first data, when bit is 0, a point is drawn at the corresponding height. The drawing principle of the other three channels is the same.

2.3 Pretreatment of Flaw Detection Output Data

Railway transportation can be said to be an important part of the national strategic planning, so the rail flaw detection data obtained by flaw detection workers are not the original appearance of the data, but hexadecimal numbers after a series of encryption means. Before learning and training these data, it is necessary to convert the hexadecimal number to format, so as to obtain the original appearance of the data.

For B-type display images stored in hexadecimal numbers, each piece of data has special significance. Because this paper studies the detection of rail bottom damage, it takes the data of random section of rail bottom flaw detection as an example. Suppose the data obtained is 00 00 00 03 08 40 00 E3 FF, in which the first four bytes of data 00 00 00 03 are the current number of pulses, 08 is the current number of image channels, the next 40 00 is a mask for encrypting information, E3 is real data and the whole data ends with FF. To sum up, the flaw detection data of the rail bottom can be converted into 0xff, 0xff, 0xff, 0xff, 0xff, 0xff, 0xff, 0xff and 0xe3.

In order to indicate the position of data in sampling points, the drawing rules will use a mask of 1–2 bytes, that is, 8–16 bits, to indicate its position according to the different sampling points of different channels. For the bottom part of 7–8 channels, the number of sampling points is between 8 and 16 bytes, so a 2-byte (16-bit) mask is needed to indicate the source position of the data.

3 Construction of Recurrent Neural Network

3.1 Introduction of Recurrent Neural Network

Recurrent Neural Network (RNN) is an important branch of deep learning, because it can cyclically and recursively process historical data and model historical memory, and is suitable for processing information that is closely related in time and space series. It can learn complex mapping relationship from vector to vector, and it has self-connection inside, which can be used to process sequence data. Besides its biggest feature is that the results of each layer can be used as the input values of the next layer, so that its calculation results have the ability to remember the previous results and keep the data dependency.

At every moment, the output of RNN will be combined with the current input and the model state. As we can see from the following Fig. 2, on the left is a single recurrent neural network, which is composed of input vector x, hidden layer state s and output vector h, and other W, U and V are all weight parameters. Among them, the output of hidden layer has two functions, one is to feed back to itself circularly, and the other is to propagate to neurons in the next moment in sequence. Therefore, the main characteristic of RNN is that the hidden layer state s_t can memorize the sequence data, and the relationship between sequence information can be captured through the hidden layer, so that the sequence with any length can be processed with limited parameters.

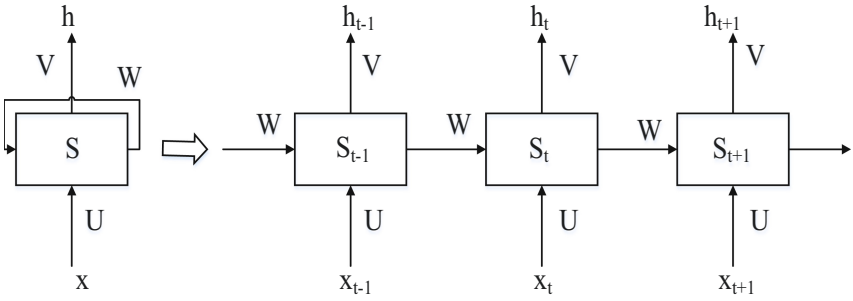

Fig. 2. Structure model of Recurrent Neural Network

3.2 Input Data Processing

The purpose of this paper is to detect the existence of rail bottom damage defects, so it is necessary to extract the features of this part of data before it can be used as the input of the system. In the preparation work at the early stage of the experiment, the flaw detection data at the bottom of the rail are found out accurately by manual marking, which includes both damage data and normal data. In this experiment, 4014 pieces of data are shared, and each piece of data was 15, so the system input was a 4014*15-dimensional matrix. In the data collection stage, different flaw detectors are used by different workers, and their flaw detection speed may be different, which leads to slightly different characteristics

of the same injury data. Therefore, in the experiment, the activation function Sigmoid is used to normalize the original input matrix, that is, to normalize all the data to [0,1], which can reduce the system operation and the system error. The specific expression of this function is as shown in Eq. 1.

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

3.3 Specific Introduction of Training Model

There are three layers in this model, including input layer, hidden layer and output layer. The characteristic of RNN is that the propagation between layers includes forward propagation and backward propagation. RNN can be regarded as the result that the same neural network structure is replicated many times in time series. This replicated structure is called loop body, and how to design loop body is also the key to solve practical problems of this network model. The classification of this paper is mainly binary, so Sigmoid function is used in forward propagation, and cross entropy cost function is defined to represent the average error of all samples.

$$C = -\frac{1}{m}\sum_{x}\left[ylna + (1-y)ln(1-a)\right] \tag{2}$$

In which C represents cost function, x represents samples, y represents actual value,a represents output value of input matrix calculated by activation function, and n represents total number of samples. When the error is bigger, the gradient is bigger, and the adjustment speed of parameters is wider, so the final training speed is faster. In the back propagation of RNN, it is assumed that the loss function is cross entropy. In all the forward propagation, the inverse calculation is carried out, and the gradient is calculated. The gradient descent algorithm is used to update the weight.

3.4 Adaptive Moment Estimation

Adaptive moment estimation (Adam) is a popular optimization method to estimate large-scale parameters in neural networks. As we all know, when calculating on the whole data set, as long as the learning rate is low enough, we can always get non-negative progress on the loss function. When using this method, it is necessary to set an initial learning rate first, and then introduce a second-order momentum based on this learning rate. This variable can effectively reduce the convergence speed of the system, which can effectively prevent the system from exceeding the optimal solution in the calculation process.

4 Training and Testing

As explained above, the flaw detection data of rail bottom are distributed in channel 7 and channel 8, and the example of damage is shown in Fig. 3, in which the marked with black box represents the bottom defect, and the unmarked one represents the normal echo or the weld echo signal under another flaw detection condition.

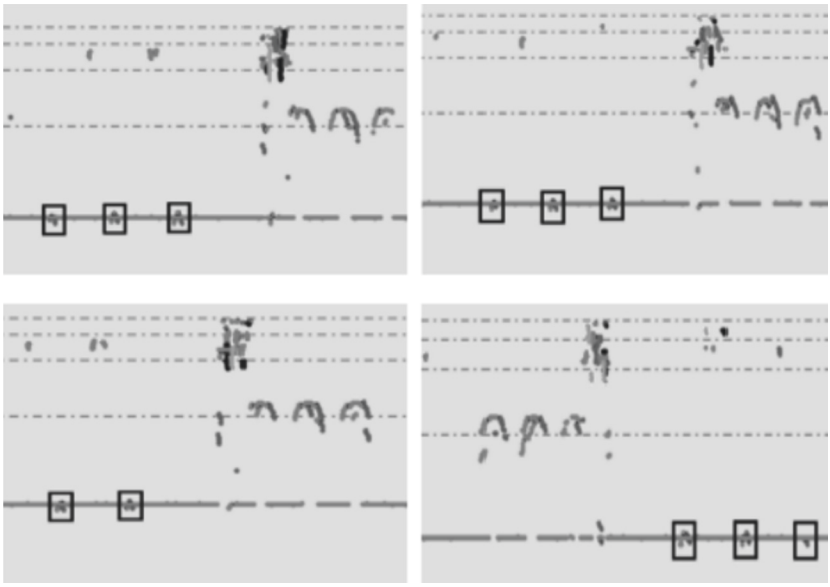

Fig. 3. Case of rail bottom damage

All the data used in this paper are from each line within the jurisdiction of Jinan Railway Bureau in Shandong Province in recent years, and the amount of data is sufficient. In the early stage of the experiment, the flaw detection data on the whole line were sliced, and the images containing the echoes at the bottom of the rail were extracted, and then marked and integrated. About 90% of the data set is selected as the training set, and the rest is used as the test set. The experimental results are shown in Fig. 4 (Table 1).

It can be seen from the waveform of the cost function of the experimental results that when the training times are the same, different learning rates will lead to great differences in training results. When the learning rate is set to 0.0001, the initial value of the cost function is relatively small, but the convergence speed in the later period is slow, that is, the average variance of the sample training results converges slowly and remains a large value at the end of training, which also leads to the low accuracy. However, when the learning rate is increased to 0.00018, the convergence speed is increased compared with that before, and the average variance of the sample training results converges to a smaller value at the end of training, which corresponds to an increase in the training accuracy.

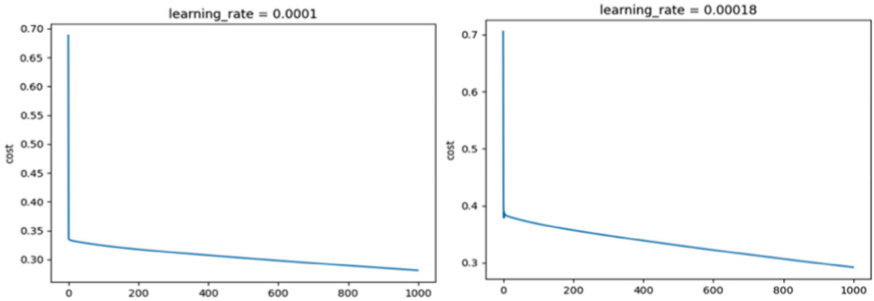

Fig. 4. Cost function image under different learning rate

Table 1. Training accuracy record

Training times	Learning rate	Training accuracy
100000	0.0001	85.85%
500000	0.0001	95.79%
100000	0.00018	91.56%
500000	0.00018	98.11%

When analyzing the training results of a single picture, we can find that the value of cost function decreases with the increase of training times, which means that the average variance of the whole sample training is getting smaller and smaller, that is, the accuracy of the network model is getting higher and higher, and finally reaches 98.11% after 500,000 training.

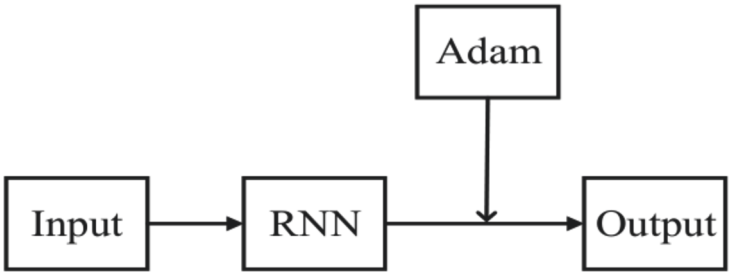

Fig. 5. System with Adam optimizer

However, when combined with adaptive moment estimation optimizer shown in Fig. 5, the system can automatically adjust the learning rate according to the training situation, so that better training results can be obtained. Adam is a variant of gradient descent algorithm, and it dynamically adjusts the learning rate of each parameter by using the first-order moment estimation and the second-order moment estimation of

gradient. However, the learning rate of each iteration parameter has a certain range, so the learning rate (step size) will not become large because of the large gradient, and the value of the parameter is relatively stable. When the system function is close to the optimal solution, the step size of system convergence is reduced, which can prevent the system from crossing the optimal solution in the calculation process and causing local divergence. Figure 6 shows the accuracy of the system after adding Adam optimizer. It can be seen from the figure that the system can get better training effect with less training times.

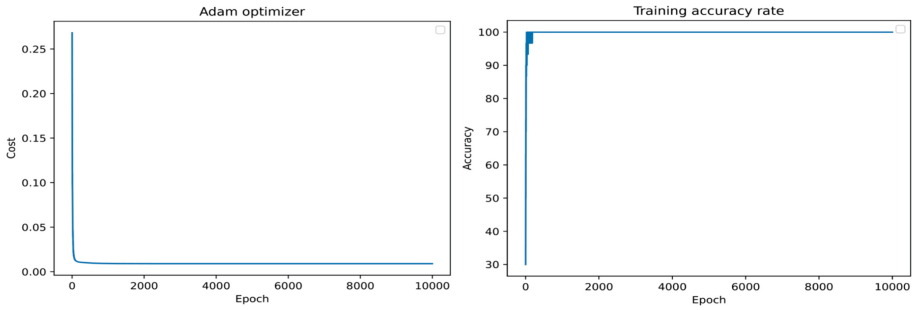

Fig. 6. Training results with Adam optimizer

Table 2. Test sample of rail bottom damage

Sample number	Whether it is defect	Expected output	Actual output
1	Yes	1	1
2	Yes	1	1
3	Yes	1	1
4	Yes	1	1
5	Yes	1	1
6	Yes	1	1
7	Yes	1	1
8	No	0	1
9	No	0	0
10	No	0	0

After the whole training, to ensure the reliability of the system designed in this paper, some brand-new data fragments were intercepted from the flaw detection data of previous years, which included both injury data and normal echo data. After a round of testing, 10 test results were randomly selected and counted in Table 2. From the table,

it can be seen that the accuracy of the test basically accords with the results obtained in the training process.

5 Conclusions

Firstly, this paper analyzes the present situation of railway development in China, and gives the damage caused by rail damage and the causes of these damages. Then, combined with some current research status, an automatic detection system of rail bottom damage based on Recurrent Neural Network is designed. What's more, combines with Adam optimizer, a better detection effect is obtained. After training more than 4,000 samples, the accuracy of the system can reach the accuracy in the actual working process. At the same time, the detection speed far exceeds the manual detection speed.

In the future research, if you want a better and more accurate classification effect, you can use SoftMax classifier to achieve multi-level classification of injuries. You can also use the same method to make a larger data set, so that after the system is fully trained, check the detection effect of the system.

References

1. Tian, G., Gao, B., Gao, Y., Wang, P., et al.: Overview of inspection and monitoring technology for railway rail defects. Chin. J. Sci. Instrum. **37**(08), 1763–1780 (2016)
2. Zhang, H., Song, Y., et al.: Summary of nondestructive testing and evaluation technology for rail defects. Chin. J. Sci. Instrum. **40**(02), 11–25 (2019)
3. Li, W.: Study on ultrasonic detection of rail damage and location. North University of China (2013)
4. He, J.: Vehicle recognition method based on recurrent neural network. W. Commun. Technol. **10**(44), 160–162 (2020)
5. Wang, K.: Research on Vehicle Identification Method of Rail Damage Based on Convolutional Neural Network. Harbin Institute of Technology (2017)
6. Sun, C., Liu, J.: Intelligent identification method of rail damage based on deep learning. China Rail. Sci. **39**(05), 51–57 (2018)
7. Wang, Y., Xu, P.: Classification of CNN electron microscope medical images based on improved Adam optimizer. J. Xi'an Univ. Posts Telecommun. **24**(05), 26–33 (2019)
8. Zhou, J., et al.: Adaptive moment estimation for polynomial nonlinear equalizer in PAM8-based optical interconnects. Opt. Express **27**(22), 32210–32216 (2019)