

Bearing faulty prediction based on knowledge distillation

Chung-Wen HUNG, Zheng-Jie LIAO, Chun-Liang LIU

National Yunlin University of Science and Technology, Taiwan
123 University Road, Section 3, Douliou, Yunlin 64002, Taiwan, R.O.C

Email: wenhung@yuntech.edu.tw, M11212039@yuntech.edu.tw, clliu@yuntech.edu.tw

Abstract

This paper employs knowledge distillation to train teacher and student models using different motor bearing vibration datasets. The signal is transformed from the time domain to the frequency domain using Fast Fourier Transform (FFT), and a Convolutional Neural Network (CNN) model is used to recognize the bearing conditions. The teacher model is a deeper model trained with a larger dataset, while the student model is a shallower model trained with less data. The student model is guided by the soft labels provided by the teacher model. The results demonstrate that knowledge distillation improves the student model's recognition performance and enables knowledge transfer, allowing the student model to achieve good recognition accuracy even with limited training data.

Keywords: Bearing fault detection · Knowledge distillation · CNN

1. Introduction

Bearings are crucial to the operation of rotating motors, such as induction motors. Therefore, the condition of motor bearings significantly impacts the operational state of machinery [1]. In recent years, advancements in artificial intelligence have facilitated the widespread application of deep learning, with techniques based on deep neural networks proven effective for diagnostics using vibration signals [2], [3]. However, existing fault diagnosis networks are large in scale and contain numerous parameters, making them difficult to deploy on embedded computing platforms. This study utilizes model compression techniques to reduce the model size while maintaining a certain level of accuracy, with the goal of enabling deployment on edge devices.

2. Materials and Methods

2.1. Artificial Neural Networks

Artificial Neural Networks (ANN) are the foundational structure of deep learning, comprising an Input Layer, Hidden Layers, and an Output Layer. Each hidden layer consists of a different number of neuron nodes, which connect with one another through each node's multiplication by various weights and addition of a bias term. The calculation is shown in Equation (1), where F represents the chosen activation function, y_n is the output of the current layer (and the input to the next layer), x_i represents the node of the current layer, M is the number of nodes in the previous layer, $w_{j,i}$ is the weight connecting the nodes, and $bias$ is the offset term. Before passing to the next layer, each neuron undergoes a linear or nonlinear transformation through an activation function. This transformation allows neural networks to learn and simulate data patterns and decision boundaries of varying complexity, enhancing the model's flexibility.

$$y_n(n) = \sum_{i=0}^M F * (x_i(n) \times w_{j,i}(n) + bias_n) \quad (1)$$

2.2. Convolution Neural Network

Convolutional Neural Networks (CNNs) are the most widely used method for feature extraction. The structure of a one-dimensional CNN consists of a Convolution Layer, a Pooling Layer, and a Fully Connected Layer. The Fully Connected Layer, typically organized as an artificial neural network, performs classification after computing the weights. CNNs are a frequently used type of neural network in deep learning, particularly for tasks that require efficient feature extraction and classification.

2.3. Knowledge Distillation

Knowledge Distillation (KD) [4] is a model compression technique designed to train a simpler student model to replicate the behavior of a more complex teacher model, achieving comparable or even superior recognition performance. The principle of knowledge distillation is shown in Equation (1), where z_i represents the output logits for the i -th class, and a temperature variable T is introduced. A higher T results in a smoother probability distribution from the softmax output, increasing the entropy of the distribution, which helps the model focus more on negative labels. This can enhance the generalization ability of the student model. During high-temperature distillation, as shown in Equation (2), the objective function consists of a weighted combination of the distillation loss \mathcal{L}_{soft} and the student loss \mathcal{L}_{hard} is introduced to avoid fully trusting the teacher model, effectively reducing the propagation of errors to the student model. The weights α and β represent the respective proportions, with their sum equal to 1.

$$q_i = \frac{\exp(\frac{z_i}{T})}{\sum_j \exp(\frac{z_j}{T})} \quad (1)$$

$$\mathcal{L} = \alpha \mathcal{L}_{soft} + \beta \mathcal{L}_{hard} \quad (2)$$

3. Data and Results

3.1. Case Western Reserve University dataset

In this study, we utilized the Case Western Reserve University (CWRU) ball bearing vibration dataset [5], which is commonly employed for fault diagnosis in induction motor bearings. The vibration testing platform, shown in Fig. 1, consists of components arranged from left to right: motor, torque sensor, and dynamometer. Two accelerometers were installed on the motor's drive end and fan end, sampling at 48 kHz and 12 kHz, respectively. For this study, we selected the 48 kHz drive-end bearing data to evaluate the proposed classification method.

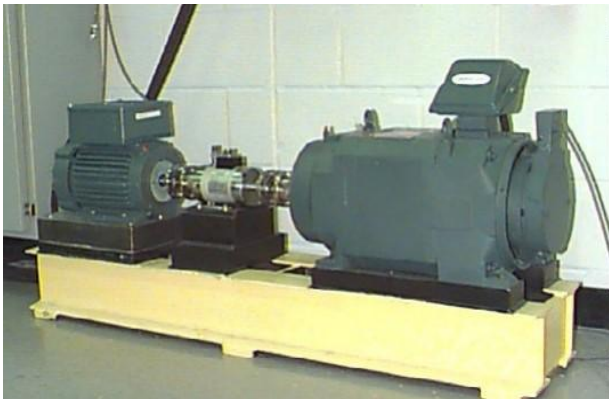


Fig. 1 CWRU platform

The drive-end bearing has four distinct conditions: inner race fault, outer race fault, ball fault, and normal condition. Fault diameters are categorized into three sizes: 0.007, 0.014, and 0.021 inches. The motor was tested under four different load levels: 0 hp, 1 hp, 2 hp, and 3 hp, which correspond to rotational speeds of 1797 rpm, 1772 rpm, 1750 rpm, and 1730 rpm, respectively. Table 1 presents the various fault conditions utilized in this study, along with the labels assigned to inner race faults, outer race faults, and normal conditions.

Table 1 CWRU fault conditions

| Condition | Fault size(in.) | label |
|-----------|-----------------|-------|
| normal | 0 | 0 |
| Inner | 0.007 | 1 |
| Inner | 0.014 | 2 |
| Inner | 0.021 | 3 |
| Outer | 0.007 | 4 |
| Outer | 0.014 | 5 |
| Outer | 0.021 | 6 |

3.2. Triaxial bearing vibration dataset

The triaxial bearing vibration dataset (TBVD) [6] contains three-dimensional (X, Y, and Z axis) vibration data for induction motor bearing faults. The testing platform, shown in Fig. 2, consists of a three-phase induction motor and an AC generator with a variable load. An accelerometer is mounted on the housing near the drive-end bearing of the three-phase induction motor, sampling at a rate of 10 kHz. The bearing conditions include inner race faults, outer race faults, and normal conditions, with fault severity levels of 0.7, 0.9, 1.1, 1.3, 1.5, and 1.7 mm. Vibration data were collected under load levels of 100 W, 200 W, and 300 W. Table 2 provides details of the various faults in the dataset and the labels assigned to the selected inner race faults, outer race faults, and normal conditions.



Fig. 2 TBVD platform

Table 2 TBVD fault conditions

| Condition | Fault size (mm) | label |
|-----------|-----------------|-------|
| normal | 0 | 0 |
| Inner | 0.7 | 1 |
| Inner | 1.3 | 2 |
| Inner | 1.7 | 3 |
| Outer | 0.7 | 4 |
| Outer | 1.3 | 5 |
| Outer | 1.7 | 6 |

In this study, the CWRU vibration dataset was used as training data for the teacher model. The input data for the model involved downsampling the 48 kHz vibration signal to 10 kHz using linear interpolation. The data was then segmented into segments of 2048 points each, followed by a Fast Fourier Transform (FFT). This process yielded a total of 1122 samples. The waveforms for each condition are shown in Fig. 3, demonstrating distinct frequency distributions under various fault conditions.

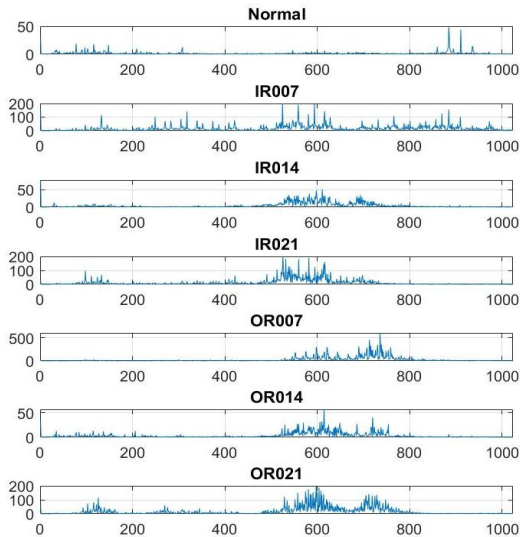


Fig. 3 CWRU data waveform

For output classification, faults of the same severity across three load levels were grouped into a single class, resulting in a total of seven conditions. Detailed model training parameters are listed in Table 3. The training and testing data were divided in a 7:3 ratio. Accuracy (ACC) was used as the evaluation metric, calculated as shown in Equation (1), where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. The confusion matrix for the teacher model on the test set is shown in Fig. 4, with an ACC of 0.997, which indicates excellent classification performance.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Table 3 Teacher model parameter

| | |
|--------------------------------|-------------------|
| CNN Filters | 22 |
| CNN Stride | 8 |
| CNN Kernel | 10 |
| CNN Activate Function | Swish |
| ANN Hidden layer | 64-64-64-64-64-64 |
| Hidden layer Activate Function | Sigmoid |
| Classes | 7 |
| Params | 39,667 |
| Learning Rate | 0.0001 |
| Epoch | 200 |

| True \ Predicted | True | | | | | | |
|------------------|------|----|----|----|----|----|----|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 0 | 53 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 49 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 40 | 0 | 0 | 1 | 0 |
| 3 | 0 | 0 | 0 | 58 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 45 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 50 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 0 | 41 |

Fig. 4 confusion matrix

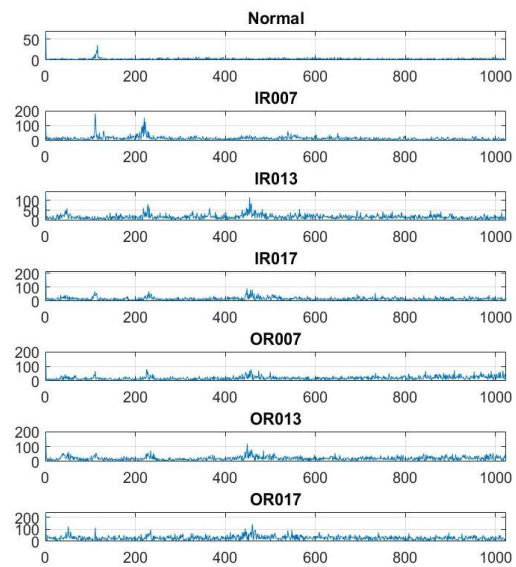


Fig. 5 TBVD data waveform

For the student model, the triaxial bearing vibration dataset was used. Since the teacher model only takes single-axis vibration data as input, it is essential to align the data dimensions between the teacher and student models to enable soft-label supervision from the teacher model. Therefore, the X-axis data from the triaxial dataset was selected as the input for the student model. The 10 kHz vibration data was segmented into 2048-point segments and transformed using FFT, resulting in 455 samples. The waveforms for each condition are shown in Fig5, revealing distinct frequency distributions under different fault states.

The data was divided into training and testing sets in an 8:2 ratio, and a 5-fold cross-validation method was applied to the dataset, with ACC averaged over the five folds. Detailed training parameters for the student model are shown in Table 4. The structure of the student model has fewer parameters compared to the teacher model.

Table 4 Student model parameter

| | |
|--------------------------------|-------------------|
| CNN Filters | 22 |
| CNN Stride | 8 |
| CNN Kernel | 10 |
| CNN Activate Function | Swish |
| ANN Hidden layer | 32-64-32-64-32-64 |
| Hidden layer Activate Function | Sigmoid |
| Classes | 7 |
| Params | 22,291 |
| Learning Rate | 0.0001 |
| Epoch | 1000 |

The training results are summarized in Table 5. When the student model was trained independently, it achieved an accuracy of 0.824. When the student model reused the CNN layers from the teacher model and trained only the subsequent ANN classification layers, accuracy increased to 0.839. By performing knowledge distillation with soft labels provided by the teacher model, the student model's recognition accuracy reached 0.892. This represents a 6.8% improvement over training of the student model independently and only a 2% difference from the performance of the teacher model. Moreover, the student model achieved comparable recognition performance while requiring only 56% of the parameters of the teacher model.

Table 5 model accuracy

| Model | ACC |
|---|-------|
| Teacher | 0.997 |
| Triaxial data with the teacher model architecture | 0.912 |
| Independent training of the student model | 0.824 |
| KD for the student model's ANN layers | 0.839 |
| KD student model | 0.892 |

4. Conclusion

This study utilizes knowledge distillation to facilitate the student model's learning from the teacher model's expertise. Under similar operating conditions but different working states, experimental results show that knowledge distillation can effectively reduce the model size to 56% of the teacher model. Furthermore, it allows for knowledge transfer, resulting in only a 2% difference in accuracy between the student model and the teacher model.

5. References

1. Xie, J., et al.: A novel bearing fault classification method based on XGBoost: the fusion of Deep Learning-based features and empirical features. *IEEE Trans. Instrum. Meas.* 70, 1–9 (2021)
2. Yang, R., Zhang, Z., Chen, Y.: Analysis of vibration signals for a ball bearing-rotor system with raceway local defects and rotor eccentricity. *Mech. Mach. Theor.* 169, 104594 (2022).
3. Li, X., et al.: A vibration fault signal identification method via sest. *Electronics* 11(9), 1300 (2022).
4. Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." *arXiv preprint arXiv:1503.02531* (2015).
5. Brjapon: 'CWRU Bearing Dataset'. <https://www.kaggle.com/datasets/brjapon/cwru-bearing-datasets>. (2023)
6. Kumar, D., et al.: Triaxial bearing vibration dataset of induction motor under varying load conditions. *Data Brief* 42, 108315 (2022).

Authors Introduction

Dr. Chung-Wen HUNG



He received the Ph.D. degrees in Electrical Engineering from National Taiwan University in 2006. Currently he is a Professor in National Yunlin University of Science & Technology. His research interests include the IoT, IIoT, and AI application.

Mr. Zheng-Jie LIAO



He received the B.S. degrees and now he is studying for the M.S. degree in electrical engineering from National Yunlin University of Science and Technology.

Dr. Chun-Liang LIU



He received the Ph.D. degrees in Electrical Engineering from National Taiwan University of Science and Technology in 2014. Currently he is a Professor in National Yunlin University of Science & Technology.
