2022 IEEE International Conference on Prognostics and Health Management (ICPHM) | 978-1-6654-6615-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICPHM53196.2022.9815705

Gradient harmonized loss: Improving the performance of intelligent diagnosis models in large imbalance scenarios

Zhijun Ren
Key Laboratory of Education Ministry for Modern Design & Rotor-Bearing System
Xi'an Jiaotong University
Xi'an, China
renzhijun@stu.xjtu.edu.cn

Wenjun Su*
Key Laboratory of Education Ministry for Modern Design & Rotor-Bearing System
Xi'an Jiaotong University
Xi'an, China
wenjunsu@mail.xjtu.edu.cn

Tantao Lin
Key Laboratory of Education Ministry for Modern Design & Rotor-Bearing System
Xi'an Jiaotong University
Xi'an, China
lintantao@stu.xjtu.edu.cn

Rui Zhang
Key Laboratory of Education Ministry for Modern Design & Rotor-Bearing System
Xi'an Jiaotong University
Xi'an, China
cici666@stu.xjtu.edu.cn

Yongsheng Zhu*
Key Laboratory of Education Ministry for Modern Design & Rotor-Bearing System
Xi'an Jiaotong University
Xi'an, China
yszhu@mail.xjtu.edu.cn

Ke Yan
Key Laboratory of Education Ministry for Modern Design & Rotor-Bearing System
Xi'an Jiaotong University
Xi'an, China
yanke@mail.xjtu.edu.cn

Jun Hong
Key Laboratory of Education Ministry for Modern Design & Rotor-Bearing System
Xi'an Jiaotong University
Xi'an, China
jhong@mail.xjtu.edu.cn

***Abstract*—The natural distribution of monitoring data is imbalanced, which has a negative impact on the training of intelligent diagnosis models. Although researchers have proposed data-level and algorithm-level methods to solve this problem, these methods are only applicable to small imbalance scenarios. In order to correct the anomalies of model training under large imbalance scenarios, this paper proposes a gradient harmonized loss that coordinates the gradients of each class to prevent the majority class in the imbalanced data from dominating the training. The coordination of gradients is based on the similarity of the sample gradients, and the compression of similar gradients is achieved by defining different penalty rules for each class. Taking into account the computational efficiency and the training difficulty, the proposed method is further optimized in terms of gradient dimensionality reduction and parameter simplification respectively. The proposed method was verified using two sample sets with different imbalance ratios and compared with traditional methods. The results showed that the proposed method greatly improved the performance of the DCNN model in large imbalance scenarios.**

Keywords—rotating machinery, intelligent fault diagnosis, imbalanced data, large imbalance scenarios, gradient harmonization

I. INTRODUCTION

Rotating machinery is an indispensable part in manufacturing, transportation, energy, and other fields, where it contributes significantly to improving efficiency and reducing economic costs [1]. Once equipment breaks down, it will affect the operation efficiency, bring economic losses and, more seriously, endanger the safety of personnel [2]. Therefore, timely and reliable fault diagnosis is essential for mechanical equipment [3].

In the field of fault diagnosis, scholars have proposed numerous methods to improve the accuracy of fault identification. These methods can be further classified into signal processing-based methods and artificial intelligence-based methods [4]. In general, methods based on signal processing require a great deal of expert experience and finding fault features from massive and complex data is time-consuming and labor-intensive [5]. As a result, these methods are extremely constrained in practice. In contrast, because artificial intelligence-based methods can automatically identify fault status, they are effective for handling massive data [6].

According to the different feature extraction processes, artificial intelligence-based methods can be further divided into methods based on artificial features and methods based on adaptive features [6]. The former still relies on expert experience for the extraction of fault features, and these methods have insufficient generalization performance with complex data [7]. With the development of deep learning techniques, researchers have introduced deep neural networks to extract features. Since deep learning models can extract features adaptively based on data characteristics, these methods excel in terms of diagnostic difficulty and generalization ability [8]. Therefore, methods based on deep learning are currently the most popular in the field of fault diagnosis.

However, most of the current methods are based on the unrealistic assumption that the data from each status is balanced [9]. In fact, equipment serves in a normal state for a long time and generates much more normal data than faulty data [10]. Imbalanced data has been demonstrated to be harmful to the training of deep learning models [11]. In order to solve the problem of model training with imbalanced data, researchers

This research was supported by the National Key R&D Program of China [grant number 2019YFB2004302]

have proposed data-level and algorithm-level methods. The data-level method balances the data for each status by increasing the data for the minority class or decreasing the data for the majority class to reduce the impact of imbalance on training [12, 13]. By weighting for each class, the algorithm-level method constrains the contribution of each class to the training to prevent the majority class from dominating the training [14, 15]. However, existing methods only focus on small imbalance situations. The data-level method cannot modify the sample set excessively because that would affect the original distribution of the data. The definition of weights in algorithm-level methods is difficult, and weights based on the number of samples can have very large or very small values in large imbalance scenarios, affecting the convergence of the model. Therefore, it is essential to develop imbalanced learning methods that are adapted to large imbalance scenarios.

Inspired by the studied in [16] and [17], the negative impact of imbalanced data on the deep learning model is caused by the imbalanced gradients. Therefore, this paper constructs a gradient harmonized loss based on the gradient relationship between the training samples to coordinate the gradients of each class in the training process and to correct the dominance of the majority class on the model training.

The rest of this paper is organized as follows. Section II describes the proposed gradient harmonized loss. In Section III, the application of the proposed method in the class imbalance problem is presented. The conclusion of this paper is summarized in Section IV.

II. Gradient Harmonized Loss

A. The calculation of gradients

Generally, the calculation process of the hidden layers and the output layer in a feedforward neural network can be described as

$$\boldsymbol{a}^{(l)} = f_l\left(\boldsymbol{z}^{(l)}\right) = f_l\left(\boldsymbol{W}^{(l)}\boldsymbol{a}^{(l-1)} + \boldsymbol{b}^{(l)}\right), \tag{1}$$

where $\boldsymbol{a}^{(l)} \in \Re^{M_l}$ is the output of the layer l. $\boldsymbol{W}^{(l)} \in \Re^{M_l \times M_{l-1}}$ and $\boldsymbol{b}^{(l)} \in \Re^{M_l}$ are the weight matrix and the bias vector from layer $l-1$ to layer l respectively. $f_l(.)$ represents the activation function of the layer l.

Given the loss function $L\left(\boldsymbol{y}, \overline{\boldsymbol{y}}\right)$, the gradient of the weight vector $\boldsymbol{w}_i^{(l)} \in \Re^{M_{l-1}}$ ($\boldsymbol{W}^{(l)} = \left[\boldsymbol{w}_1^{(l)}; ...; \boldsymbol{w}_i^{(l)}; ...; \boldsymbol{w}_{M_l}^{(l)}\right]$) in layer l is

$$\nabla \boldsymbol{w}_i^{(l)} = \frac{\partial L}{\partial \boldsymbol{w}_i^{(l)}} = \frac{\partial \boldsymbol{z}^{(l)}}{\partial \boldsymbol{w}_i^{(l)}} \cdot \frac{\partial \boldsymbol{a}^{(l)}}{\partial \boldsymbol{z}^{(l)}} \cdot \frac{\partial L\left(\boldsymbol{y}, \overline{\boldsymbol{y}}\right)}{\partial \boldsymbol{a}^{(l)}}, \tag{2}$$

where $\boldsymbol{y}$ and $\overline{\boldsymbol{y}}$ are the real label and the predicted label of the sample respectively.

According to the chain rule, $\partial L\left(\boldsymbol{y}, \overline{\boldsymbol{y}}\right) / \partial \boldsymbol{a}^{(l)}$ can be described as

$$\frac{\partial L\left(\boldsymbol{y}, \overline{\boldsymbol{y}}\right)}{\partial \boldsymbol{a}^{(l)}} = \frac{\partial \boldsymbol{z}^{(l+1)}}{\partial \boldsymbol{a}^{(l)}} \cdot \frac{\partial \boldsymbol{a}^{(l+1)}}{\partial \boldsymbol{z}^{(l+1)}} \cdot \frac{\partial L\left(\boldsymbol{y}, \overline{\boldsymbol{y}}\right)}{\partial \boldsymbol{a}^{(l+1)}}. \tag{3}$$

Thus, $\nabla \boldsymbol{w}_i^{(l)}$ can be further described as

$$\nabla \boldsymbol{w}_i^{(l)} = \frac{\partial \boldsymbol{z}^{(l)}}{\partial \boldsymbol{w}_i^{(l)}} \cdot \frac{\partial \boldsymbol{a}^{(l)}}{\partial \boldsymbol{z}^{(l)}} \cdot \frac{\partial \boldsymbol{z}^{(l+1)}}{\partial \boldsymbol{a}^{(l)}} \cdot \ldots \cdot \frac{\partial \overline{\boldsymbol{y}}}{\partial \boldsymbol{z}^{(L)}} \cdot \frac{\partial L\left(\boldsymbol{y}, \overline{\boldsymbol{y}}\right)}{\partial \overline{\boldsymbol{y}}}. \tag{4}$$

If the gradients of each layer are harmonized, the computational burden will be increased significantly, especially for deep neural networks. According to (4), the gradient of each layer is related to $\partial L\left(\boldsymbol{y}, \overline{\boldsymbol{y}}\right) / \partial \overline{\boldsymbol{y}}$, so we can harmonize $\partial L\left(\boldsymbol{y}, \overline{\boldsymbol{y}}\right) / \partial \overline{\boldsymbol{y}}$ to complete the approximate harmonization of the network gradients. Li et al. [18] used the same approximation when harmonizing the size of the gradients, which was proved to be feasible.

B. The Indicator for gradient harmonization

Under the condition of data imbalance, the reason for the abnormally large gradient in the majority class is the large similarity between the gradients. As shown in Fig. 1, two gradients with the same modulus form three synthetic gradients at three different similarity degrees, and we can find that the modulus of the synthetic gradients varies significantly ($\nabla T_3 > \nabla T_2 > \nabla T_1$). In fact, in Fig.1(c), either ∇T_{31} or ∇T_{32} can represent the optimization direction of the two samples, however, adding the two gradients together may result in over-optimization of the two samples. Therefore, the gradients with large similarity need to be constrained to construct the ideal optimization direction and step. In this study, the gradients with large similarity were compressed to prevent the formation of larger gradients, rather than just discarding some gradients. Discarding some gradients may cause deviations from the ideal gradient, while compressing the gradient ensures that the direction of the synthetic gradient remains unchanged.

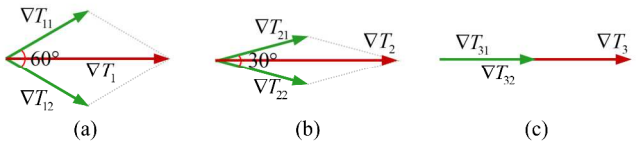

Fig. 1. A comparison between synthetic gradients of two gradient vectors with different spatial relations. (a) The angle between the two gradients is 60°. (b) The angle between the two gradients is 30°. (c) The angle between the two gradients is 0°.

Given the gradient of sample i

$$\boldsymbol{\delta}_i = \frac{\partial L\left(\boldsymbol{y}_i, \overline{\boldsymbol{y}_i}\right)}{\partial \overline{\boldsymbol{y}_i}}, \tag{5}$$

the gradient similarity of sample i with sample j in the same class can be described as

$$GS_{ij} = \arccos\left(\frac{\boldsymbol{\delta}_i \boldsymbol{\delta}_j^T}{\|\boldsymbol{\delta}_i\|_2 \|\boldsymbol{\delta}_j\|_2}\right), \tag{6}$$

where $\boldsymbol{\delta}_j^T$ represents the transposition of $\boldsymbol{\delta}_j$. $\|\cdot\|_2$ represents the modulus of the vector.

Eventually, the gradient similarity vector formed by sample i and other samples in the same class is

$$\boldsymbol{GS}_i = \left[GS_{i1},...,GS_{ij},...,GS_{iN}\right], \tag{7}$$

where N represents the number of samples in the same class.

Based on the gradient similarity vector, the gradient density function of sample i is formulated as

$$GD_i = \sum_{j=1}^{N} \varphi\left(GS_{ij}, GS_T\right), \tag{8}$$

where

$$\varphi(x,y) = \begin{cases} 1 & \text{if } 0 \le x \le y \\ 0 & \text{otherwisw} \end{cases}, \tag{9}$$

and GS_T represents the gradient similarity threshold.

Now we define the gradient harmonizing function as

$$\bar{\boldsymbol{\delta}}_i = \frac{\boldsymbol{\delta}_i}{GD_i}. \tag{10}$$

According to (10), we can embed the gradient density GD_i into classification loss, and then the gradient harmonized loss (GHL) is

$$L_{GH}\left(\boldsymbol{y}_i^c, \overline{\boldsymbol{y}_i^c}\right) = \frac{1}{GD_i^c} L\left(\boldsymbol{y}_i^c, \overline{\boldsymbol{y}_i^c}\right), \tag{11}$$

where c is the class label.

C. The definition of the gradient similarity threshold GS_T

According to (8) and (9), the gradient similarity threshold GS_T determines the region size for calculating the gradient similarity. In imbalanced data, if all classes share the same threshold, the synthetic gradient from the majority class still dominates the training, and if different thresholds are defined for each class, the tuning difficulty for thresholds increases geometrically with the number of classes. Therefore, in this study, an adaptive threshold definition method is proposed, which is described as

$$GS_{\mathrm{T1}}^c = \alpha\left(\mathrm{sigmoid}\left(\left(\frac{N^c}{\sum_{c=1}^{C} N^c} - \frac{1}{C}\right) * 5C\right) - 0.5\right), \tag{12}$$

where α is a hyperparameter used to define the threshold for each class. C is the number of classes.

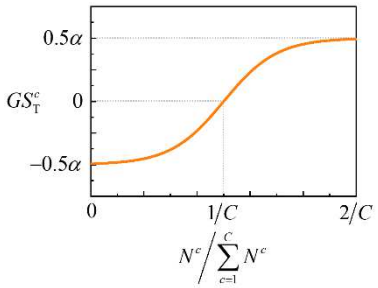

Fig. 2. The relationship between the class distribution and the corresponding class threshold.

The relationship between the class distribution and the corresponding class threshold is shown in Fig. 2. For classes with a small number of samples, the corresponding threshold GS_{T1}^c is small, otherwise, the threshold GS_{T1}^c is large. Moreover, the thresholds can change adaptively according to the sample distribution, which makes the threshold definition for imbalanced data with a continuous distribution more efficient.

According to (12), once the training samples are fixed, the thresholds for each class are then determined and remain constant throughout the training process. However, a constant threshold can have a negative impact on the convergence of the model training, since the gradient of the samples is small in the late training period, which means that smaller weights will further compress the sample gradients, resulting in slow or no convergence of the samples. Therefore, the F1-score was introduced as an annealing factor that varies dynamically with training. The final gradient similarity threshold is defined as

$$GS_{\mathrm{T}}^c = \alpha\gamma\left(\mathrm{sigmoid}\left(\left(\frac{N^c}{\sum_{c=1}^{C} N^c} - \frac{1}{C}\right) * 5C\right) - 0.5\right), \tag{13}$$

where γ represents the F1-score.

According to (13), as the samples gradually converge, the gradient similarity threshold decreases, which means that the penalty for similar gradients decreases, facilitating the convergence of the model in the later training period.

D. The optimizations of the method

Equation (6) only calculates the similarity of two samples, however, in practice it is necessary to calculate the similarity between all samples in the same class. The similarity matrix of samples in the same class can be described as

$$\boldsymbol{GS} = \arccos\left(\frac{\boldsymbol{\delta}\boldsymbol{\delta}^T}{\|\boldsymbol{\delta}\|_{2,1} \|\boldsymbol{\delta}\|_{2,1}^T}\right), \tag{14}$$

where $\boldsymbol{\delta} = \left[\boldsymbol{\delta}_1;...;\boldsymbol{\delta}_i;...;\boldsymbol{\delta}_{N_c}\right] \in \Re^{N_c \times C}$.

According to (14), $\boldsymbol{GS}$ has a time complexity of

$O\left(CN_c^2+CN_c+3N_c^2\right)$. When the number of samples and classes is large, the computation of the gradient similarity matrix will be time-consuming. So, we improved the calculation process for the gradient similarity matrix.

When the loss function is mean square error, the gradient $\boldsymbol{\delta}_i$ can be described as

$$\begin{aligned}\boldsymbol{\delta}_i&=\left[\delta_{i1};...;\delta_{ij};...;\delta_{iC}\right]\\&=-\frac{2}{K}\left[y_{i1}-\bar{y}_{i1};...;y_{ij}-\bar{y}_{ij};...;y_{iC}-\bar{y}_{iC}\right],\end{aligned}\tag{15}$$

where $\boldsymbol{y}_i=\left[y_{i1};...;y_{ij};...;y_{iC}\right]$ and $\bar{\boldsymbol{y}}_i=\left[\bar{y}_{i1};...;\bar{y}_{ij};...;\bar{y}_{iC}\right]$.

When the loss function is root mean square error, the gradient $\boldsymbol{\delta}_i$ can be described as

$$\begin{aligned}\boldsymbol{\delta}_i&=\left[\delta_{i1};...;\delta_{ij};...;\delta_{iC}\right]\\&=-\frac{\left[y_{i1}-\bar{y}_{i1};...;y_{ij}-\bar{y}_{ij};...;y_{iC}-\bar{y}_{iC}\right]}{\sqrt{C\sum_{j=1}^{C}\left(y_{ij}-\bar{y}_{ij}\right)^2}},\end{aligned}\tag{16}$$

These two loss functions are often used to train deep learning models. According to (15) and (16), we can find that $\sum_{j=1}^{C}\delta_{ij}=0$, which means the gradients of all samples lie on a C-dimensional hyperplane. Therefore, we can construct a two-dimensional coordinate system on this hyperplane to reduce the dimensionality of high-dimensional data. Given two original gradients $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ as well as a coordinate transformation matrix $\boldsymbol{A}$, the gradients after reducing the dimensionality can be described as

$$\begin{cases}\bar{\boldsymbol{\delta}}_1=\boldsymbol{\delta}_1\boldsymbol{A}\\\bar{\boldsymbol{\delta}}_2=\boldsymbol{\delta}_2\boldsymbol{A}\end{cases}.\tag{17}$$

In order to ensure that the gradient similarity remains the same before and after transforming, namely $GS\left(\boldsymbol{\delta}_i,\boldsymbol{\delta}_j\right)=GS\left(\bar{\boldsymbol{\delta}}_i,\bar{\boldsymbol{\delta}}_j\right)$, the transformation shown in (17) must be an orthogonal transformation, namely $\boldsymbol{A}\boldsymbol{A}^T=\boldsymbol{E}$. Therefore, $\boldsymbol{A}$ is an orthogonal matrix that can be calculated from two linearly independent vectors based on Gram-Schmidt orthogonalization.

When the loss function is mean square error or root mean square error, $\boldsymbol{GS}$ has a time complexity of $O\left(5N_c^2+2N_c+2CN_c\right)$ after reducing the dimensionality of gradients. Therefore, the time complexity is reduced to almost $5/(C+3)$ of the original. Taking the subsequent sample set as an example, when the number of the majority samples is 3200 and the number of classes is 12, the time complexity of $\boldsymbol{GS}$ reduced to 33% of the original after reducing the dimensionality of gradients.

In addition to the time complexity, the difficulty of hyperparameter optimization needs attention. Although the proposed method introduces only one hyperparameter α, there is a coupling between the effects of α and the hyperparameter *batch size* on the model training. In fact, the coupling of multiple parameters increases the difficulty of model optimization, which seems to receive little attention. Therefore, we propose using the gradient density GD_i to simulate the role of batch size in model training. Since the gradient weighting factor $1/GD_i$ has a range of [0,1], when $1/GD_i$ is close to 0, the contribution of the gradient to the training is extremely small. Therefore, we can eliminate such samples to simulate the random selection of samples, which is the main purpose of *batch size*. In this way, only the α needs to be adjusted in hyperparameter tuning, which reduces the difficulty of model training.

III. Application of GHL in Intelligent Fault Diagnosis

A. Datasets description

In this study, the bearing dataset was used to demonstrate the validity of the proposed method. The dataset is published by the Bearing Data Center at Case Western Reserve University [19], as shown in Table I. In the dataset, there are 12 bearing health statuses in total. The bearings in each fault status operated under the same conditions, with a speed of around 1700 r/min and a load of 2 hp. The bearing in the normal status ran at 1700 r/min with loads of 0, 1, and 2 hp. In the experiment, the accelerometer was installed on the driving end of the motor at 12 o'clock, and the sampling frequency is 12 kHz.

TABLE I. The Bearing Dataset

Health status [a]	Fault size (mm)	Health status	Fault size (mm)
NC	-	RF3	0.53
OF1	0.18	RF4	0.71
OF2	0.36	IF1	0.18
OF3	0.53	IF2	0.36
RF1	0.18	IF3	0.53
RF2	0.36	IF4	0.71

[a] NC represents the normal condition; OF represents the outer race fault; RF represents the roller fault; and IF represents the inner race fault.

B. Sample sets description

Based on the above dataset, 2 different sample sets were constructed. The number of training samples in each sample set is shown in Table II. The number of testing samples in two bearing sample sets is 108. The dimension of samples is 1024. In each sample set, the training samples and the testing samples are split according to the time series, namely the data at the earlier time is the training sample and the data at the later time is the testing sample [11].

C. Model description

In this study, a deep convolutional neural network (DCNN) was constructed to verify the validity of the proposed method. The DCNN model contains seven layers: the input layer, the first convolution layer that contains 16 convolutional kernels with size (1×49), the first max pooling layer that contains a pooling kernel with size (1×4), the second convolution layer

that contains 16 convolutional kernels with size (1×21), the second max pooling layer that contains a pooling kernel with size (1×4), the fully connected layer whose dimensions is 100, and the output layer. In subsequent sections, the original model is called DCNN, and the model with GHL is called DCNNG.

TABLE II. THE SAMPLE SET GENERATED FROM THE BEARING DATASET

Health status	The number of samples	
	Set 1	Set 2
NC	3200	3200
OF1	64	16
OF2	64	16
OF3	64	16
RF1	64	16
RF2	64	16
RF3	64	16
RF4	64	16
IF1	64	16
IF2	64	16
IF3	64	16
IF4	64	16

D. Diagnostic results

In order to demonstrate the effectiveness of the method proposed in this paper, a popular imbalanced learning method was also introduced. Jia et al. [20] introduced class costs in deep convolutional neural networks to constrain the losses of different classes (DCNNC), which can prevent the gradient of the majority class from dominating the training. The diagnostic results of the original DCNN model, the DCNNC model, and the DCNNG model proposed in this paper are shown in Table III.

TABLE III. THE DIAGNOSTIC RESULTS

Sample set	Method	Test accuracy (%)
Set 1	DCNN	88.83±5.78
	DCNNC	88.45±5.46
	DCNNG	95.28±2.51
Set 2	DCNN	70.36±8.59
	DCNNC	76.21±7.07
	DCNNG	81.84±3.98

As can be seen from Table III, the DCNNG achieves an average accuracy of 95.28% in Set 1 and 81.84% in Set 2, which are the best one among the three methods. The original DCNN obtains low accuracy in two sample sets because the majority class in the imbalanced sample set dominates the training of the model, resulting in difficult convergence of the minority class and low accuracy of the minority class. In Set 1, the average accuracy of DCNNC is lower than that of the original DCNN, while in Set 2, the average accuracy of DCNNC is higher than that of the original DCNN. The unstable performance improvements from DCNNC are due to the large imbalance rate resulting in larger weights for the minority classes, which may weaken the training of the majority classes.

E. Improvement mechanism

In order to visualize the improvement mechanism of the proposed method, the variation trend of the average gradient similarity within each class of the DCNN model and the DCNNG model is counted, as shown in Fig. 3. Bases on (7), the average gradient similarity within each class can be described as

$$\overline{GS} = \sum_{i=1}^{N}\sum_{j=1}^{N} GS_{ij}. \tag{18}$$

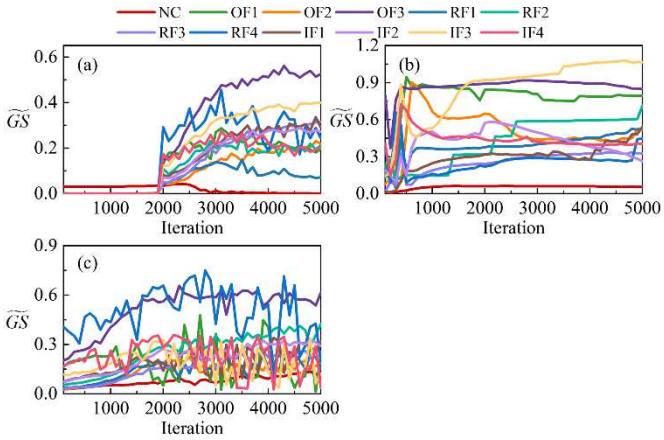

Fig. 3. The variation trend of the average gradient similarity within each class. (a) The model is DCNN and the sample set is Set 1. (b) The model is DCNNG and the sample set is Set 1. (c) The model is DCNN and the sample set is balanced, where the number of samples in each class is 16.

We can see that under the balanced sample set, the average similarity of the gradients within each class shows a fluctuating trend, while under the imbalanced sample set, the majority class dominates training at the beginning of iteration and the average similarity of the gradients within each class remains constant, which is detrimental to the convergence of the model and leads to a low diagnostic accuracy of the model. In contrast, after gradient coordination, the majority class no longer dominates the initial iterations of the model and the average similarity of the gradients within each class shows a fluctuating trend.

IV. CONCLUSIONS

In this paper, a gradient harmonized loss is proposed for solving the imbalanced learning problem in large imbalance scenarios. By compressing gradients with large similarities, abnormally large gradients are avoided, and by applying different compression rules to different classes, the domination of the majority class over the training is avoided. In the compression rule, the definition of the threshold can change adaptively with the sample distribution and the training process. In addition, a novel dimensionality reduction strategy and a hyperparameter simplification strategy were designed to reduce the computational burden as well as the training difficulty, respectively. Two sample sets with different imbalance rates were used to verify the effectiveness of the method. The results showed that the method greatly improved the performance of the DCNN model in large imbalance scenarios.

The definition of the gradient similarity threshold makes the method proposed in this paper inapplicable to balanced datasets. Therefore, in future work, it will be necessary to construct a method for defining the threshold for both balanced and

imbalanced situations.

ACKNOWLEDGMENT

Thanks to Prof. Zhu, Prof. Yan, and Prof. Hong for their guidance in the content and writing of this study. Mr. Lin and Mr. Zhang provided great help to the experiment of this research.

REFERENCES

[1] L. C. Brito, G. A. Susto, J. N. Brito, and M. A. V. Duarte, "An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery," Mech. Syst. Signal Proc., vol. 163, 108105, Jan. 2022.

[2] Y. Guan, Z. Meng, D. Meng, J. Meng, and F. Meng, "2MNet: Multi-sensor and multi-scale model toward accurate fault diagnosis of rolling bearing," Reliab. Eng. Syst. Saf., vol. 216, 108017, Dec. 2021.

[3] Z. Ren, Y. Zhu, Y. Fu, C. Fu, K. Yan, and J. Yan, "Fault diagnosis with imbalanced data based on auto-encoder," in 11th International Conference on Prognostics and System Health Management (PHM-2020 Jinan), Jinan, China, Oct. 23-25, 2020.

[4] Y. Ding, J. Zhuang, P. Ding, and M. Jia, "Self-supervised pretraining via contrast learning for intelligent incipient fault detection of bearings," Reliab. Eng. Syst. Saf., vol. 218, 108126, Oct. 2021.

[5] Z. Ren, Y. Zhu, K. Yan, K. Chen, K. Kang, Y. Yue et al., "A novel model with the ability of few-shot learning and quick updating for intelligent fault diagnosis," Mech. Syst. Signal Proc., vol. 138, 106608, Apr. 2020.

[6] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi, "Applications of machine learning to machine fault diagnosis: A review and roadmap," Mech. Syst. Signal Proc., vol. 138, 106587, Apr. 2020.

[7] K. Zhang, J. Chen, S. He, E. Xu, F. Li, and Z. Zhou, "Differentiable neural architecture search augmented with pruning and multi-objective optimization for time-efficient intelligent fault diagnosis of machinery," Mech. Syst. Signal Proc., vol. 158, 107773, Sep. 2021.

[8] D. Gao, Y. Zhu, W. Kang, H. Fu, K. Yan, and Z. Ren, "Weak fault detection with a two-stage key frequency focusing model," ISA Trans., Jun. 2021.

[9] J. Wu, Z. Zhao, C. Sun, R. Yan, and X. Chen, "Learning from Class-imbalanced Data with a Model-Agnostic Framework for Machine Intelligent Diagnosis," Reliab. Eng. Syst. Saf., vol. 216, 107934, Dec. 2021.

[10] S. Xing, Y. Lei, B. Yang, and N. Lu, "Adaptive knowledge transfer by continual weighted updating of filter kernels for few-shot fault diagnosis of machines," IEEE Trans. Ind. Electron., vol. 69, no. 2, pp. 1968-1976, Feb. 2021.

[11] Z. Zhao, T. Li, J. Wu, C. Sun, S. Wang, S. Yan et al., "Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study," ISA Trans., vol. 107, pp. 224-255, Dec. 2020.

[12] W. Zhang, X. Li, X. Jia, H. Ma, Z. Luo, and X. Li, "Machinery fault diagnosis with imbalanced data using deep generative adversarial networks," Measurement, vol. 152, 107377, Feb. 2020.

[13] A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," J. Artif. Intell. Res., vol. 61, pp. 863-905, Apr. 2018.

[14] C. Zhang, K. C. Tan, H. Li, and G. S. Hong, "A cost-sensitive deep belief network for imbalanced classification," IEEE Trans. Neural Netw. Learn. Syst., vol. 30, no. 1, pp. 109-122, Jan. 2019.

[15] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 8, pp. 3573-3587, Aug. 2018.

[16] R. Anand, K. G. Mehrotra, C. K. Mohan, and S. Ranka, "An improved algorithm for neural-network classification of imbalanced training sets," IEEE Trans. Neural Netw. Learn. Syst., vol. 4, no. 6, pp. 962-969, Nov. 1993.

[17] Z. Ren, Y. Zhu, W. Kang, H. Fu, Q. Niu, D. Gao et al., "Adaptive cost-sensitive learning: Improving the convergence of intelligent diagnosis models under imbalanced data," Knowledge-Based Syst., vol. 241, Apr. 2022.

[18] B. Y. Li, Y. Liu, and X. G. Wang, "Gradient harmonized single-stage detector," in 33rd AAAI Conference on Artificial Intelligence (AAAI 2019), Honolulu, USA, Jan. 27-Feb. 01, 2019.

[19] "Seeded Fault Test Data." Bearing Data Center of Case Western Reserve University. http://csegroups.case.edu/bearingdatacenter/home (accessed Oct. 20, 2021).

[20] F. Jia, Y. Lei, N. Lu, and S. Xing, "Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization," Mech. Syst. Signal Proc., vol. 110, pp. 349-367, Sep. 2018.