IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 12, 2020, accepted April 23, 2020, date of publication April 27, 2020, date of current version May 13, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2990572

Normalized Recurrent Dynamic Adaption Network: A New Framework With Dynamic Alignment for Intelligent Fault Diagnosis

CHENGDONG ZHENG, XIAOJING WANG, YIFAN HAO, KE WANG, AND XIN XIONG
School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China

Corresponding author: Xiaojing Wang (wangxjshu@163.com)

This work was supported by the National Natural Science Foundation of China under Grant 51705302.

ABSTRACT In the field of intelligent fault diagnosis, distribution divergence always exists between the training and testing sets (which could be considered as a source domain with known labels and a target domain without labels), which will lead to a significant degradation in the diagnosis performance of deep network. Generally, this problem is solved by transfer learning. Specifically, adapt the marginal distribution or jointly align the marginal and conditional distributions of two domains so that the classifier trained by labeled source data merely can correctly classify target data. However, when aligning the marginal and conditional distributions simultaneously, people usually gives them the equal weight while it is not in accordance with the general situations. In this paper, we propose a new framework called normalized recurrent dynamic adaption network (NRDAN) for intelligent fault diagnosis which not only adapts the marginal and conditional distributions of two domains simultaneously but also estimates the relative importance of two distributions dynamically and quantitatively. This framework adopts long short-term memory (LSTM) as the base network combined with layer normalization (LN) and mainly consists of a feature extractor, a dynamic adaption module, and a classifier. Finally, extensive experiments including transfer tasks between not only various operating conditions but also different machines are conducted to comprehensively evaluate the proposed method.

INDEX TERMS Intelligent fault diagnosis, deep learning, transfer learning, dynamic adaption, long short-term memory.

I. INTRODUCTION

Machinery and equipment are developing towards automation and intelligence in modern industry. It is expected that the health condition of machines can be monitored and the types of mechanical faults can be diagnosed effectively in order to reduce the economic loss and guarantee the workers' safety. Intelligent fault diagnosis frameworks utilizing deep learning technique have been applied in the field of fault diagnosis gradually and shows promising performance compared with those traditional machine learning based methods [1]–[3]. Deep network can extract features from raw data automatically instead of manually as shallow network, which indicates that the deep learning based fault diagnosis method can avoid the shortcoming of handcrafted features and the loss of primitive information [4], [5], and is suitable for end-to-end diagnosis. Besides, deep learning can contribute to a higher recognition accuracy for the fault diagnosis due to its stronger nonlinear expression ability [6], [7].

With these advantages, however, these deep model based intelligent fault diagnosis systems need to satisfy some requirements to achieve excellent diagnosis performance. First, a great amount of labeled data are required to train the deep network in order to fully learn representative features and obtain a strong generalization ability. Second, the training data and testing data should follow the same data distribution. Unfortunately, it is extremely difficult to meet these requirements in many engineering applications due to the following reasons. On the one hand, it is not only dangerous but also costly to acquire a large number of fault data samples directly from the monitoring machine since conducting fault experiments on the monitoring machine may lead to

The associate editor coordinating the review of this manuscript and approving it for publication was Szidónia Lefkovits.

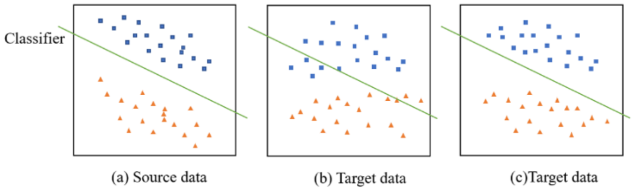


FIGURE 1. The distribution of target data before and after transfer learning.

catastrophic accident and take a lot of time [8]. On the other hand, the data samples acquired from the identical machine under different operating conditions may differ greatly in data distribution, which suggests that the diagnosis system may behave badly when the operating condition varies. Specially, when sufficient fault data are inconvenient to be acquired from the monitoring machine, it is expected to effectively monitor the target machine with the help of data samples obtained from other related but different machines. Nevertheless, the distribution discrepancy between the data from either the monitoring machine or other related machines can be exceedingly considerable because they are structurally different. This fact may result in the failure of the intelligent fault diagnosis system due to the deep model's poor generalization ability.

In such cases, transfer learning, i.e. transferring the knowledge learned from source domain into the new but related domain [9], [10], would be helpful to address these issues. As shown in Figure 1, relying on transfer learning, the domain-invariant features of the source and target domain data can be extracted by the deep model and provided to the classifier trained by labeled source data [11], [12]. Therefore, when there are new diagnostic tasks, there is no need to rebuild the network or train the classifier from scratch. This is especially suitable for the case that the labeled data in target domain is not sufficient to retrain an excellent model. With the help of transfer learning, the deep network based intelligent fault diagnosis framework can have a stronger generalization ability, and the quantity of samples required for new diagnosis tasks will also be reduced.

In real-world applications, the distribution discrepancy between the source and target domains, which decides the distribution alignment method to be applied, is variable for different target domains, as shown in Figure 2. To date, several state-of-the-art transfer learning methods have been applied to the field of intelligent fault diagnosis in succession. For example, [13] pays attention to the alignment of marginal distribution while [14] jointly aligns the marginal and conditional distributions of two domains and gives them the equal weight. However, they ignore the facts that it is not enough to perform the marginal distribution adaption merely, and the marginal and conditional distributions do not contribute equally to the domain divergence in many cases.

Therefore, it is necessary to develop a new framework to tackle the aforementioned problems. In this paper, we propose a new framework called normalized recurrent dynamic adaption network (NRDAN) for intelligent fault diagnosis which not only adapt the marginal and conditional distributions of two domains simultaneously but also evaluate the relative importance of two distributions dynamically and quantitatively. In this framework, long short-term memory (LSTM) is adopted as the base network for its advantages in time-series data processing and layer normalization, a simple yet powerful training trick with no requirements for batch size, is incorporated into the base network. The combination of LSTM and layer normalization makes it suitable for both end-to-end and online diagnosis, which better fits in with the necessity of real-world applications. Additionally, a dynamic adaption module which can dynamically and quantitatively adapt both marginal and conditional distributions is appended to the base network.

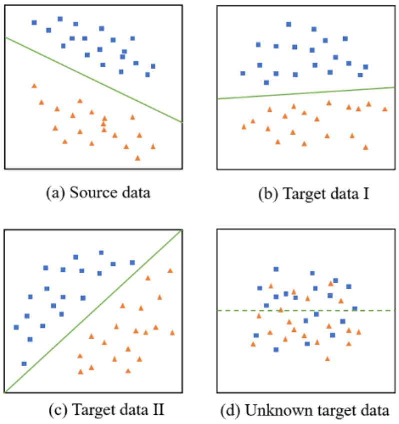


FIGURE 2. Different distributions for target data.

The main contributions of this work are summarized as follows:

1) We propose a novel diagnosis framework based on LSTM which could dynamically adjust the relative importance of marginal and conditional distributions in the transfer learning process.
2) Extensive experiments, which contains transfer tasks between not only various operating conditions but also different machines, are conducted to validate the effectiveness of the proposed framework and compare its performance with other state-of-the-art methods.
3) We further explore the reason of superiority of the proposed framework by providing the performance of NRDAN with diverse balance factors and the varying trend of balance factor with respect to the number of training iterations.
4) We incorporate layer normalization into the base network and comprehensively study the effect of the location where it is joined on the diagnosis performance.

The rest of this paper is structured as follows. Related work is reviewed in Section II. In Section III, some previous knowledge closely related to the proposed method is introduced. Section IV details the proposed framework. Extensive experiments and analysis are given in Section V. Conclusions of this paper are drawn in Section VI.

II. RELATED WORK

Transfer learning becomes an increasingly popular topic in the area of fault diagnosis recently and abundant efforts have been made to develop transfer learning based fault diagnosis framework. Xie *et al.* [15] presented a fault diagnosis method combining transfer component analysis (TCA) and support vector machine (SVM) to investigate gearbox diagnosis under various operating conditions. Lu *et al.* [16] established a deep neural network (DNN) model utilizing transfer learning to extract general features, which are then input into the SVM classifier trained by labeled source data and normal category data in the target domain. Wen *et al.* [13] proposed a fault diagnosis method based on sparse auto-encoder (SAE) and incorporated maximum mean discrepancy (MMD) term into the network to reduce distribution discrepancy between the source and target domains. Guo *et al.* [17] constructed a one-dimension convolutional neural network (CNN) utilizing domain adaption and adversarial learning to reduce the domain shift and studied the performance of the proposed method by conducting experiments on bearing datasets obtained from different machines. Li *et al.* [18] proposed a 2-stage deep general neural networks based fault diagnosis method, which utilizes multi-kernel MMDs and can provide reliable diagnosis results when testing data in fault conditions are not available for training. Some researchers also developed diagnosis frameworks to realize multi-layer distribution adaption in order to efficiently extract more transferable features [19], [20]. Additionally, some attempts have been made to reduce the marginal and conditional divergences simultaneously and the corresponding diagnosis systems have been validated to outperform those based on marginal distribution adaption method [14], [21].

The transfer learning based diagnosis methods mentioned above can be roughly divided into two categories: (1) marginal distribution adaption, which merely aligns the marginal distribution in the last hidden layer or multiple hidden layers; and (2) joint distribution adaption, which adapts the marginal and conditional distributions jointly. Unfortunately, they do not realize the fact that the marginal and conditional distributions are not equally important to the domain shift, while the method NRDAN proposed in this paper can address the problem by dynamically and quantitatively estimating the relative importance of each distribution.

III. PRELIMINARIES

A. PROBLEM DESCRIPTION

In transfer learning, normally, there is a source domain $D_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ with n_s labeled samples and a target domain $D_t = \{(x_i^t)\}_{i=1}^{n_t}$ with n_t unlabeled samples, where $X = \{x_i\}_{i=1}^{n}$ represents the feature space and $Y = \{y_i\}_{i=1}^{n}$ is the corresponding label space, respectively. Unlike the traditional deep learning scenario that the data distribution of the training and testing datasets are almost same or extremely similar, in this paper, we suppose that a more general case exists in the transfer tasks. Specifically, the marginal distribution $P(X)$ and conditional distribution $Q(Y|X)$ of the two aforementioned domains are different from each other, i.e. $P(X^s) \neq P(X^t)$and $Q(Y^s|X^s) \neq Q(Y^t|X^t)$. The purpose of transfer learning is to align the distributions of two domains and enable the network to learn more general features so that the classifier trained by labeled source data cannot discriminate whether a data sample comes from the source domain or target domain. As a result, the classifier can achieve satisfying recognition effect on data samples from both source and target domains.

B. MAXIMUM MEAN DISCREPANCY

Maximum mean discrepancy (MMD) [22] is widely adopted as a distribution distance metric in transfer learning [11], [23], [24] due to its non-parametric characteristics and satisfying effectiveness. This paper adopts multi-kernel MMD (MK-MMD) [25] for better performance. For domain adaption problems discussed in this paper, the distribution discrepancy between the source domain and target domain can be measured as the squared distance between the kernel embeddings in a reproducing kernel Hilbert space (RKHS), i.e.:

$$\begin{aligned} D(D_s, D_t) &= \left\| E\left[\varphi\left(X^s\right)\right] - E\left[\varphi\left(X^t\right)\right] \right\|_{\mathcal{H}}^2 \\ &= \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \varphi\left(x_i^s\right) - \frac{1}{n_t} \sum_{j=1}^{n_t} \varphi\left(x_j^t\right) \right\|_{\mathcal{H}}^2 \\ &= \frac{1}{n_s^2} \sum_{i=1}^{n_s} \sum_{j=1}^{n_s} k_\varphi\left(x_i^s, x_j^s\right) \\ &\quad + \frac{1}{n_t^2} \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} k_\varphi\left(x_i^t, x_j^t\right) \\ &\quad - \frac{2}{n_s n_t} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} k_\varphi(x_i^s, x_j^t) \end{aligned} \tag{1}$$

where $k_\varphi(\cdot)$ denotes the kernel function.

C. MARGINAL DISTRIBUTION ADAPTION

Marginal distribution adaption (MDA), which was firstly realized on deep neural networks to implement transfer learning by Tzeng *et. al.* [23] in 2014, has been applied to the field of intelligent fault diagnosis and achieved encouraging performance [17], [19], [26], [27]. MDA mainly relies on aligning the marginal distributions of two domains to conduct transfer learning and the corresponding formula can be calculated as:

$$D_P(D_s, D_t) = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \varphi\left(x_i^s\right) - \frac{1}{n_t} \sum_{j=1}^{n_t} \varphi\left(x_j^t\right) \right\|_{\mathcal{H}}^2 \tag{2}$$

D. JOINT DISTRIBUTION ADAPTION

Recently some researchers have applied joint distribution adaption (JDA) to the field of transfer learning [14], [21]. JDA aligns the marginal and conditional distributions simultaneously and has been shown to outperform MDA in most cases. The MMD term of conditional distribution adaption (CDA) can be defined as:

$$D_Q\left(D_s, D_t\right) = \left\| E[\varphi\left(Y^s \,|\, X^s\right)] - E[\varphi\left(Y^t \,|\, X^t\right)] \right\|_{\mathcal{H}}^2 \tag{3}$$

Note that it is not feasible to evaluate the conditional distribution because of the absence of ground-truth labels for target data. According to the sufficient statistics when sample sizes are large, the class conditional distribution $Q(X|Y)$ can be used to approximate $Q(Y|X)$ because $Q(X|Y)$ and $Q(Y|X)$ can be quite involved [28], [29]. Supposing that each domain contains a total of C categories, then the corresponding MMD term of conditional distribution adaption (CDA) can be described as:

$$D_Q\left(D_s^c, D_t^c\right) = \left\| \frac{1}{n_s^c} \sum\nolimits_{x_i^s \in D_s^c} \varphi(x_i^s) - \frac{1}{n_t^c} \sum\nolimits_{x_j^t \in D_t^c} \varphi(x_j^t) \right\|_{\mathcal{H}}^2 \tag{4}$$

where $c \in \{1, \ldots, \mathcal{C}\}$ is the class indicator, $n_s^c = \left|D_s^c\right|$ and $n_t^c = \left|D_t^c\right|$ denote the number of samples belonging to class c from source and target domains, respectively. $D_s^c = \left\{x_i^s | x_i^s \in D_s \wedge y\left(x_i^s\right) = c\right\}$ and $D_t^c = \left\{x_j^t | x_j^t \in D_t \wedge \hat{y}\left(x_j^t\right) = c\right\}$, containing the samples whose class labels are exactly c, are the subset of D_s and D_t, respectively. In the above formula, y(·) denotes the true labels of data samples from source domain. It is worth noting that the true labels for target data are not available in unsupervised domain adaption and hence replaced by predicting labels $\hat{y}(\cdot)$. Although the pseudo labels of target data predicted by the classifier are rather unreliable at the initial iterations, they will be updated as the training process of the network and thus become more accurate.

By integrating the marginal and conditional distribution distances, the MMD term for JDA can be represented as:

$$D\left(D_s, D_t\right) = D_P\left(D_s, D_t\right) + \sum\nolimits_{c=1}^{C} D_Q\left(D_s^c, D_t^c\right) \tag{5}$$

where the first term is the marginal distribution distance between the source and target domains while the last term denotes the sum of conditional distribution distance for each category.

E. LONG SHORT-TERM MEMORY

Recurrent neural network (RNN) [30], [31] has been widely applied in varieties of fields from machine translation [32] and language modeling [33] to speech recognition [34] and recommendation systems [35] due to its powerful ability of sequential data processing. Different from other types of neural networks such as convolutional neural network, the information in RNN propagates between not only two connected layers but also two adjacent time steps simultaneously. This distinctive characteristic of RNN leads to great advantages in time-series data processing.

However, the basic structure of RNN is rarely used in actual situations because it is difficult to train. As a gated variant of the original RNN, long short-term memory (LSTM) successfully relaxes the exploding and vanishing gradient problems which the original RNN suffers [36], and is adopted as the structure of the proposed method in this paper. The architecture of LSTM memory block with a single cell is exhibited in Figure 3. It can be seen that the LSTM and the standard RNN are similar in overall structure, except that the hidden neurons in the hidden layer are replaced by memory blocks. Additionally, input gate, forget gate and output gate are introduced into the memory block to make sure that the memory blocks can store information over long periods of time.

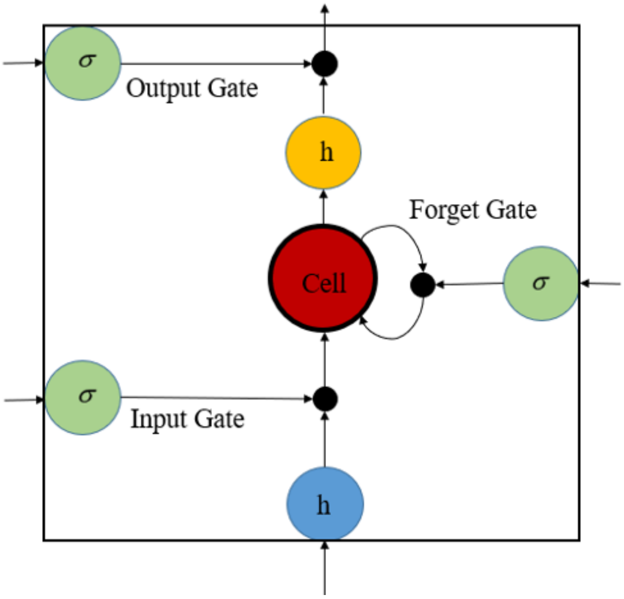


FIGURE 3. **LSTM memory block with a single cell.**

The equations for basic LSTM adopted in this paper are given as follows:

$$\begin{pmatrix} f_t \\ i_t \\ o_t \\ g_t \end{pmatrix} = W_h h_{t-1} + W_x x_t + b \tag{6}$$

$$c_t = \sigma\left(f_t\right) \odot c_{t-1} + \sigma\left(i_t\right) \times \tanh(g_t) \tag{7}$$

$$h_t = \sigma\left(o_t\right) \odot \tanh(c_t) \tag{8}$$

where g_t is the current cell input, i_t, f_t, and o_t are the output values of the input gate, the forget gate, and the output gate at current time, respectively. The state of cell is denoted c_t and the cell output is represented as h_t. The gate activation function is sigmoid and represented as $\sigma(\cdot)$, so that the output values of these gates are between 0 and 1. W_x is the weight matrix connecting the input layer and hidden layer at current time t. W_h denotes the weight matrix of the hidden layer between the current time and the previous time. b is the corresponding bias value.

F. BATCH NORMALIZATION

It is not easy to train deep neural networks partially because of the phenomenon that the distribution of each layer's inputs changes during training process. Ioffe and Szegedy [37] proposed batch normalization (BN) to address this problem called internal covariate shift by normalizing each dimension of layer inputs over a mini-batch and then scaling and shifting the normalized values. It is especially worth noting that BN performs differently for training and inference. Specifically, once the network has been trained, use the population, rather than mini-batch, statistics to predict, and the means and variances are fixed during reference.

Several attempts have been made to apply batch normalization to recurrent neural networks [38], [39], however, the experimental results indicate that BN is not suitable for RNNs because of their distinctive structures.

G. LAYER NORMALIZATION

Being the same with batch normalization (BN), layer normalization (LN) is initially proposed to reduce the training time of deep neural networks by promoting the corresponding convergence processes. Unlike BN whose effect is considerably reliant on the size of mini-batch, LN has no requirements for the quantity of training samples since it computes the mean and variance on each sample independently. This characteristic makes it more convenient to apply to the neural network because LN performs the same operation at either training or inference stage. It has been confirmed that LN works well when implemented with fully connected layers, and is particularly beneficial for recurrent neural networks [40]. Similarly, in order to describe conveniently, LN is defined as a function with two adaptive parameters, i.e. gains α and biases β:

$$LN\left(\mathbf{z};\alpha,\beta\right)=\frac{(\mathbf{z}-mean)}{std}\odot\alpha+\beta \tag{9}$$

$$mean=\frac{1}{D}\sum\nolimits_{i=1}^{D}z_i \tag{10}$$

$$std=\sqrt{\frac{1}{D}\sum\nolimits_{i=1}^{D}(z_i-mean)^2} \tag{11}$$

where z_i is the i^{th} element and D is dimension of the vector $\mathbf{z}$, respectively.

After incorporating LN, the aforementioned equations of LSTM are modified as follows:

$$c_t=\sigma\left(f_t\right)\odot c_{t-1}+\sigma\left(i_t\right)\odot\tanh(g_t) \tag{12}$$

$$h_t=\sigma\left(o_t\right)\odot\tanh(LN(c_t;\alpha_3,\beta_3)) \tag{13}$$

$$\begin{pmatrix} f_t\\ i_t\\ o_t\\ g_t\end{pmatrix}=LN\left(W_hh_{t-1};\alpha_1,\beta_1\right)+LN\left(W_xx_t;\alpha_2,\beta_2\right)+b \tag{14}$$

where α_i, β_i are the scale and shift parameters, respectively.

IV. NORMALIZED RECURRENT DYNAMIC ADAPTION NETWORK

A. DYNAMIC ADAPTION

Despite being superior to the MDA method, the JDA method is not robust enough to deal with practical applications since it treats the marginal and conditional distributions with equal weight while it is not true in many cases. Therefore, in this paper, dynamic adaption which can dynamically adjust the relative importance of each distribution is introduced to tackle the problem. According to [10], [41], we adopt $\mathcal{A}$-distance as the basic measure of cross-domain discrepancy to evaluate the relative importance of two distributions. Concretely, the proxy $\mathcal{A}$-distance is defined as:

$$d_{\mathcal{A}}(D_s,D_t)=2(1-2\epsilon) \tag{15}$$

where ϵ represents the error of a linear classifier discriminating the source and target features generated by the feature extractor (i.e. a binary problem). Then the $\mathcal{A}$-distance for marginal distribution can be computed directly according to the above formula and written as:

$$d_m=d_{\mathcal{A}}(D_s,D_t) \tag{16}$$

As for the $\mathcal{A}$-distance of conditional distribution, we refer to the method stated in part D of Section III and thus the $\mathcal{A}$-distance in each class can be calculated as:

$$d_c=d_{\mathcal{A}}(D_s^c,D_t^c) \tag{17}$$

where D_s^c and D_t^c represent the features belonging to class c in source domain and target domain, respectively. Hence the conditional $\mathcal{A}$-distance for all categories can be obtained as $\sum_c^C d_c$. Finally, the balance factor μ weighing the relative importance of marginal and conditional distributions can be estimated as:

$$\mu=1-\frac{d_m}{d_m+\sum_c^C d_c} \tag{18}$$

where the denominator in the above equation can be considered as the whole discrepancy between domains thus the balance factor μ denotes the weight of conditional distribution. The larger balance factor indicates the conditional distribution alignment is more dominant and the feature of two domains is relatively similar.

Based on the equations of joint distribution and balance factor, the dynamic adaption can be formally defined as follow:

$$D\left(D_s,D_t\right)=(1-\mu)D_P\left(D_{s,}D_t\right)+\mu\sum\nolimits_{c=1}^{C}D_Q\left(D_s^c,D_t^c\right) \tag{19}$$

where $\mu\in[0,1]$, $D_P(D_{s,}D_t)$ is the MMD term of marginal distribution and $D_Q\left(D_s^c,D_t^c\right)$ represents the MMD term of conditional distribution for class c.

B. NORMALIZED RECURRENT DYNAMIC ADAPTION NETWORK

As shown in Figure 4, the architecture of normalized recurrent dynamic adaption network (NRDAN) mainly consists of three parts: a feature extractor, a feature classifier, and a dynamic adaption module. The feature extractor is composed of 12 layers, including 4 LSTM layers, 2 fully connected layers and 6 LN layers followed by each hidden layer. The number of neurons contained in each hidden layer is [200, 200, 200, 200, 100, 50], sequentially. In the training stage, raw vibration signals from source and target domains are fed into the feature extractor simultaneously to obtain general features, which will be input into the classifier and dynamic

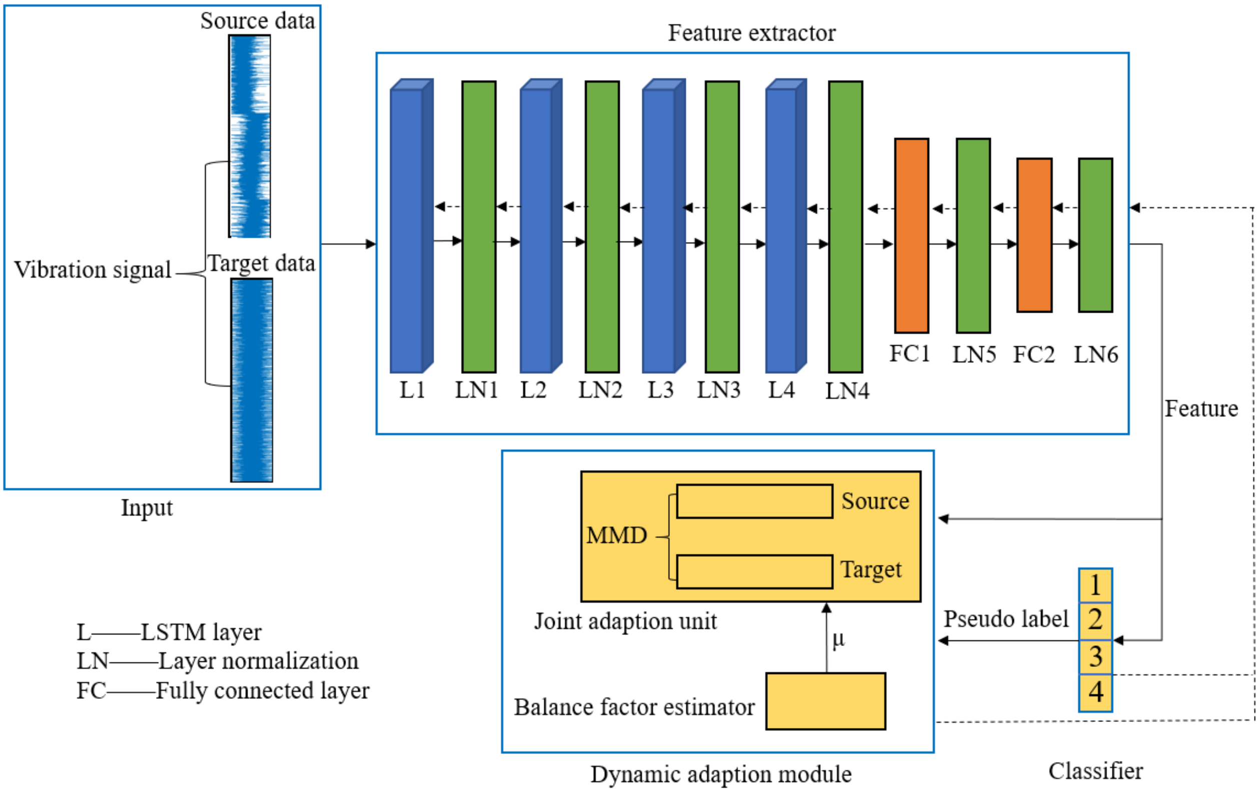


FIGURE 4. Structure illustration of the framework. The recurrent structures of LSTM layers are omitted for simplicity.

adaption module. The dynamic adaption module receives not only the features of source and target data generated by the feature extractor but also the pseudo labels of target data predicted by the classifier and true labels of source data in order to jointly adapt the marginal and conditional distributions between two domains. Specifically, the balance factor estimator uses a linear classifier (e.g. SVM) to classify the features from source and target domains and then obtains the classification error of domain to calculate the balance factor μ, which will be input into the joint adaption unit to contribute to the calculation of joint discrepancy.

```
Algorithm 1 Training Process of NRDAN
Input: Labeled source data ($x^s$,$y^s$ ), unlabeled target data $x^t$ regularization parameter $\lambda$
Output: Transferable features and predicted labels
1: repeat
2:  Sample a mini-batch data from both the source and target domains
3:  Feed the mini-batch data into the network and obtain the features and labels
4:  Calculate the joint discrepancy between domains and classification loss for source data
5:  Update the trainable parameters $\Theta$
6:  After an epoch, update the balance factor $\mu$
7: untilConvergence
```

Eventually, the objective function of this framework can be composed of classification loss $\mathcal{L}_c$ obtained by the classifier and joint discrepancy $D(D_s, D_t)$ obtained by the dynamic adaption module:

$$\mathcal{L}(\Theta) = \mathcal{L}_c + \lambda D(D_s, D_t) \tag{20}$$

where $\Theta = \{W, b, \alpha, \beta\}$ is a collection of trainable parameters including the weight W and bias b in the hidden layers, and the scale parameter α and shift parameter β in the LN layers. λ is a nonnegative hyperparameter determining the weight of regularization term and set to 0.25 in this paper.

By minimizing the above objective function, the trainable parameters will be updated. With the continuous optimization of the network, more transferable features can be obtained which leads to better classification effect on the target data. Note that in order to estimate the distribution divergence comprehensively and obtain a relatively stable value, the balance factor is updated after each epoch rather than each mini-batch. The training process is summarized in Algorithm 1.

V. EXPERIMENTS AND ANALYSIS

A. DATA DESCRIPTION

In this section, in order to evaluate the proposed framework against other state-of-the-art transfer learning methods, extensive experiments are conducted on several bearing datasets including CWRU [42], IMS [43], and XJTU-SY [44].

TABLE 1. Primary information of datasets utilized in the transfer tasks.

Name	Source	Condition	Speed (rpm)	Load	Sample length	Training samples	Testing samples	Total
A	CWRU	IF,OF,BF,NC	1797	0 HP	1200	480	480	960
B	CWRU	IF,OF,BF,NC	1750	2 HP	1200	480	480	960
C	IMS	IF,OF,BF,NC	2000	6000 lbs	1200	480	480	960
D	XJTU-SY	IF,OF,CF,NC	2250	11 KN	1200	480	480	960

1) CWRU BEARING DATASET

Case Western Reserve University (CWRU) bearing dataset was collected from a test rig primarily consisting of a motor, a torque transducer and a dynamometer. Single point faults were introduced in the inner race, outer race and ball of the test bearings separately which support the motor shaft. Each fault condition contains several different fault diameters representing different degrees of fault severity. Vibration signals were acquired from both these fault conditions and the normal condition using accelerometers attached to the housing with magnetic bases and placed at the 12 o'clock position. The dataset A and B, which are the subsets of CWRU dataset and differ in operating condition (i.e. motor speed and motor load), contain four health conditions (i.e. inner race fault (IF), outer race fault (OF), ball fault (BF) and normal condition (NC)) and 960 data samples, respectively.

2) IMS BEARING DATASET

Intelligent Maintenance System (IMS) bearing dataset was generated by conducting test-to-failure experiments on these test bearings mounted on a shaft. A radial load generated by a spring mechanism was applied to the bearing housing and the rotating speed was kept stable at 2000 RPM by an AC motor. High sensitivity quartz ICP accelerometers were installed on the bearing housing to collect vibration data of these test-to-failure bearings. At the end of the test-to-failure experiments, all failures, i.e. inner race failure, ball failure and outer race failure, occurred in these bearings after exceeding their designed life time. A dataset named C which contains normal condition and the above three fault conditions is constructed based on IMS bearing dataset.

3) XJTU-SY BEARING DATASET

XJTU-SY bearing dataset was provided by Xi'an Jiaotong University and the Changxing Sumyoung Technology. Run-to-failure experiments under three different operating conditions were conducted to observe the whole degradation processes of tested bearings which were installed on the test platform to support shaft. Rotating speed of the shaft can be adjusted by a motor speed controller and radial force applying to the housing of tested bearings is controlled by the hydraulic loading system. Two accelerometers were mounted on the horizontal axis and vertical axis of the housing respectively to collect the vibration signals of testing bearings. At the end of these run-to-failure experiments, varieties of failures occurred on the tested bearings including inner race fault (IF), cage fracture (CF), outer race fault (OF), etc. A subset of XJTU-SY named D is established to prepare for the transfer experiments between machines.

The primary information for the four datasets employed in the subsequent transfer tasks is summarized in Table 1. Note that, dataset D generated from XJTU-SY bearing dataset contains cage fracture which is different from ball fault contained in the other datasets.

B. EXPERIMENTAL SETUP

First of all, in order to validate the effect of dynamic adaption, we compare it with other related transfer learning methods on the same base network: 1) Deep marginal distribution adaption network (DMDAN, a deep transfer network with marginal distribution adaption); 2) Deep joint distribution adaption network (DJDAN, a deep transfer network with joint distribution adaption); 3) Normalized recurrent dynamic adaption network without layer normalization (NRDAN_LN, the proposed method which dynamically adapts two distributions). It should be noted that all the methods are performed on the same base network without incorporating LN layer. The training epoch is set to 2000, where the Adam optimizer is adopted for the first 1000 epochs and the gradient descent (GD) optimizer is used for the last 1000 epochs so that the network can be trained rapidly and a convergence result can be obtained. The initial learning rate of Adam optimizer and GD optimizer is set to 0.001 and 0.01, respectively. Additionally, the learning rate for GD optimizer is adjusted using the formula $\mu = \frac{\mu_0}{(1+10*p)^{0.75}}$, where μ_0= 0.01, p is the training progress linearly changing from 0 to 1 [46]. The transfer tasks are represented by letters and arrows for simplicity. For example, transfer task A→B denotes that the network is trained with the data of training sets from both source domain A and target domain B, and then tested with the data of testing set from target domain B. It is worth noting that the labels of training data from target domain are not available in the diagnosis experiments. Each diagnosis task is repeated ten times to

TABLE 2. Comparison of diagnosis performance (%) for three related transfer learning methods. The deep network is performed with marginal distribution, joint distribution, and dynamic adaption, respectively.

Method	A→B	A→C	A→D	B→A	B→C	B→D	C→A	C→B	C→D	D→A	D→B	D→C	Average
DMDAN	83.25	67.08	70.66	91.46	66.46	63.54	74.29	61.63	82.5	91.2	71.3	75.33	74.89
DJDAN	87.02	72.81	70.86	92.64	70.48	73.54	79.4	63.79	91.58	93.33	77.42	81.31	79.52
NRDAN_LN	**88.96**	**72.89**	**75.13**	**93.71**	**74.26**	**73.31**	**82.17**	**65.63**	**93.01**	**92.88**	**77.29**	**82.54**	**80.98**

TABLE 3. Diagnosis results (%) of various fault diagnosis methods.

Method	A→B	A→C	A→D	B→A	B→C	B→D	C→A	C→B	C→D	D→A	D→B	D→C	Average
Source only	80.62	35.62	26.88	63.96	33.12	23.96	43.96	37.71	32.92	24.79	24.17	25.42	37.76
TCA	51.88	25	27.08	58.54	25.83	23.33	20.83	24.38	32.71	31.04	25.83	17.92	30.36
DCTLN	92.5	98.12	87.5	98.54	97.08	93.75	92.29	72.29	87.61	90.42	77.92	94.16	90.18
DTN with JDA	95.41	92.71	95.42	91.46	94.79	96.87	89.38	76.87	93.95	88.54	82.5	97.91	91.32
NRDAN	**95.76**	**98.44**	**96.98**	**98.4**	**98.12**	**98.12**	**95.21**	**80.42**	**95.73**	**97.99**	**87.22**	**99.75**	**95.18**
Labeled target	95.05	98.5	99.71	94.79	98.5	99.71	94.79	95.05	99.71	94.79	95.05	98.5	97.01

obtain the average accuracy as the final result. The diagnosis results for all of the tasks are shown in Table 2.

Secondly, we demonstrate the superiority of the proposed framework by comparison with other state-of-the-art fault diagnosis frameworks: 1) Source only (a deep network without transfer learning which is trained with source data only); 2) Transfer component analysis (TCA, a traditional transfer learning approach) [45]; 3) Deep convolutional transfer learning network (DCTLN) [17]; 4) Deep transfer network with joint distribution adaption (DTN with JDA) [14]; 5) Normalized recurrent dynamic adaption network (NRDAN, the proposed framework); 6) Labeled target (a deep network without transfer learning which is trained with labeled target data only). Note that the diagnosis methods *source only* and *labeled target* are implemented on the same base network with NRDAN and the later method *labeled target* is performed without LN. During the diagnosis experiments, the training iteration of the methods *source only*, NRDAN, and *labeled target* is set to 300. Other learning strategies are keeping the same with the requirement mentioned above. Except for the methods 1) and 6), other methods are trained with the labeled training data from the source domain and unlabeled training data from the target domain. By contrast, the networks of *source only* and *labeled target* are trained with labeled training data from the source domain and target domain, respectively. All of the diagnosis methods are tested on the testing sets of target domains. The corresponding testing results are listed in Table 3.

C. RESULTS AND ANALYSIS

From the diagnosis results shown in Table 2, we can obtain some observations. Firstly, according to Figure 5 and Table 2, the proposed method NRDAN_LN outperforms the most related method DJDAN in most cases and achieves a relatively higher average accuracy of all the diagnosis tasks. This fact clearly verifies the superiority of dynamic adaption. Secondly, the diagnosis accuracy of each task depends on the distribution divergence between domains involved in the transfer task. For instance, domain A and B are generated by the identical machine under different operating conditions thus the overall distributions of domain A and B are rather similar while domain B and C are acquired from different machines so that the domain discrepancy between B and C is extremely considerable. Therefore, the diagnosis accuracy of task A→B outweighs that of task C→B. Finally, the diagnosis performance may vary greatly with the change of transfer direction even though the transfer method and domains are consistent. For example, the diagnosis accuracy of task B→C substantially outweighs that of task C→B.

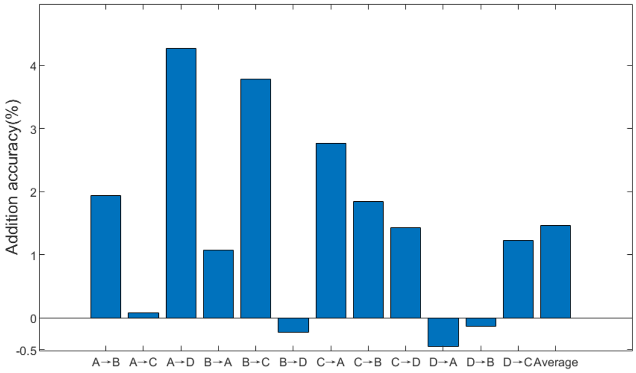


FIGURE 5. The addition accuracy (%) of NRDAN_LN compared with DJDAN. For instance, the diagnosis accuracy of NRDAN_LN on task A→D is 4.27% more than that of DJDAN.

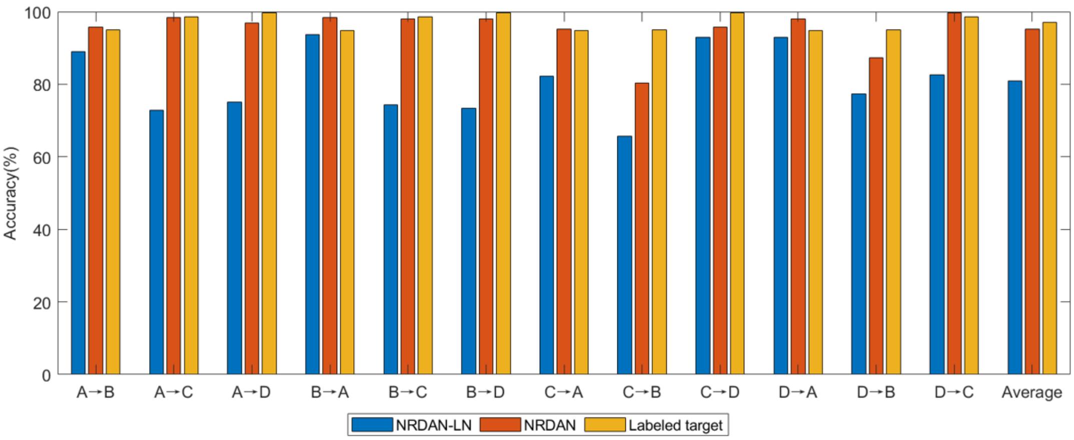
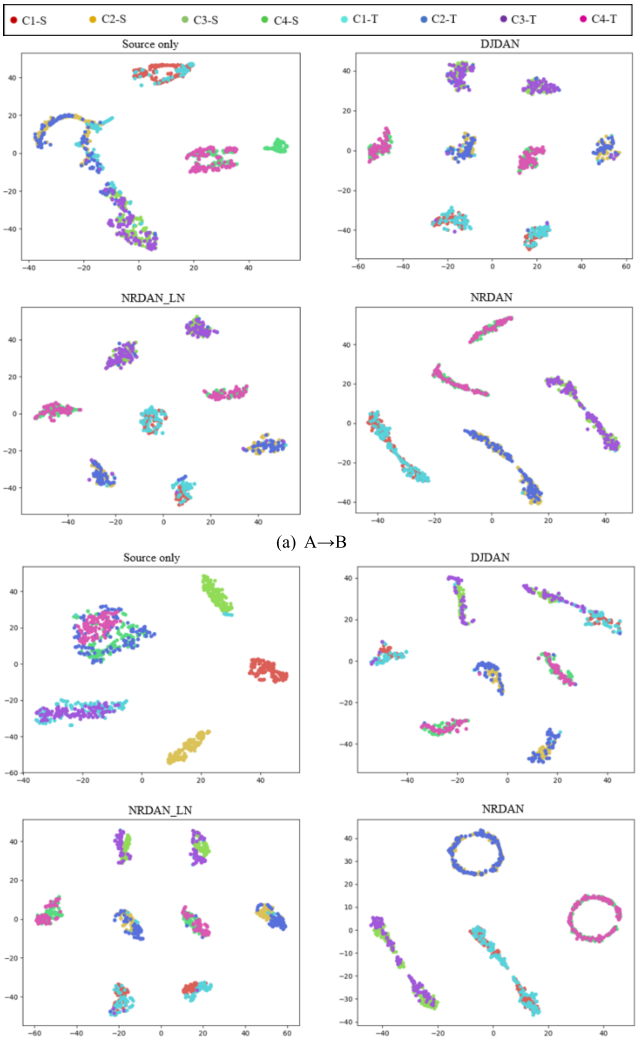


FIGURE 6. Comparison of the diagnosis performance for NRDAN_LN, NRDAN, and labeled target.

According to the results shown in Table 3, we can make the following observations. First, compared with other related transfer learning based diagnosis frameworks, the proposed NRDAN achieves a significantly higher average of diagnosis accuracy (more than 95%) on all transfer tasks, which confirms that the NRDAN outperforms other state-of-the-art diagnosis methods. Second, according to Figure 6, the diagnosis performance of NRDAN is always close and sometimes even superior to that of the method *labeled target*, which acts as an upper bound in the experiments. The approaching average accuracy of NRDAN and *labeled target* for all transfer tasks further validates the effectiveness of our proposed framework. Third, in comparison with the method *source only*, other deep transfer networks extremely improve the diagnosis accuracy on all transfer tasks, which demonstrates the necessity of transfer learning when distribution divergence exists between the training and testing sets. Finally, as shown in Figure 6, we can find that LN is a simple yet powerful trick since the NRDAN makes a considerable transfer improvement by comparison with NRDAN_LN.

In order to provide the visualization of distribution discrepancy between domains intuitively and exhibit the effect of transfer learning vividly, t-distributed stochastic neighbor embedding (t-SNE) is utilized to map the features automatically extracted by deep network from source and target domains into a two-dimension space. According to Figure 7, the method *source only* can basically classify samples of each category without transfer learning in task A→B, but the distributions of features from source and target domains are not aligned well. By contrast, the method *source only* can effectively separate the four categories of source domain in task D→C while it is incapable of discriminating the features from target domain. The degraded performance for *source only* can be explained that the distribution divergence between domain D and C is so considerable that the classifier trained by source data only cannot classify the features from target domain. After implementing transfer learning, the features learned by DJDAN and NRDAN_LN are correctly classified



(a) A→B

(b) D→C

FIGURE 7. t-SNE visualization of features extracted by feature extractor. The letter in the legend indicates whether the features are from source or target domain and the number denotes the category of features. For example, C1-S suggests that the features belonging to the first category come from source domain.

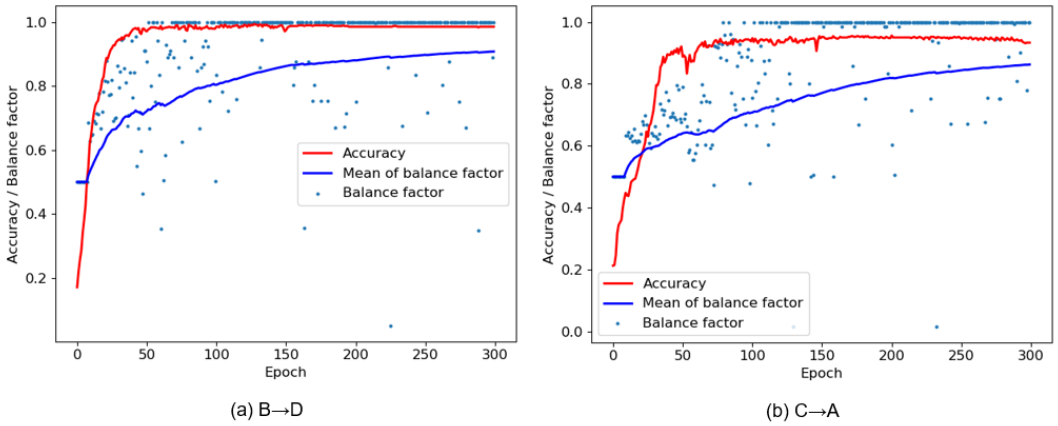


FIGURE 8. Varying trend of balance factor μ with respect to the number of training iterations.

in most cases and the distributions of source and target features are aligned very well both in task A→B and task D→C, which evidently demonstrates the effect of transfer learning. However, the features of the identical category are always separated into two distinct parts. Apparently, the internal shift of a class needs to be further reduced. With the combination of transfer learning and LN, the features learned by NRDAN are perfectly aligned with the sharply decreased shift intra class and increased distance between classes. There is no doubt that the distribution alignment has been improved a lot compared with the case of without LN.

D. THE REASON OF SUPERIORITY OF DYNAMIC ADAPTION IN COMPARISON WITH JOINT ADAPTION

In this part, we will explore the reason why dynamic adaption outperforms joint adaption which adapts marginal and conditional distributions simultaneously and is closest to the proposed method.

Figure 8 provides the varying trend of balance factor with respect to the number of training iterations for task B→D and C→A, where the scattering point represents balance factor, the blue solid line is the mean of balance factors, and the red solid line denotes the accuracy curve of NRDAN. Note that the balance factor is initialized as 0.5 at the beginning of training. It can be seen that the balance factor is changing throughout the training process and increasing gradually with the number of iterations, which can be explained that the distribution discrepancy between domains is reduced little by little in the training process thus the conditional distribution is more and more dominant. Therefore, the diagnosis accuracy and balance factor have the similar trends.

In addition, as shown in Figure 9, the balance factor μ is fixed in the experiments and it is easy to find that the performance of NRDAN on a certain transfer task varies with the change of balance factor and the optimal value of balance factor for each transfer task is different from each other due to distinct distribution divergence between domains for different tasks. For example, when the balance factor is 0.1, NRDAN achieves the best accuracy for task D→C while it is true for task D→B when the balance factor is equal to 0.7.

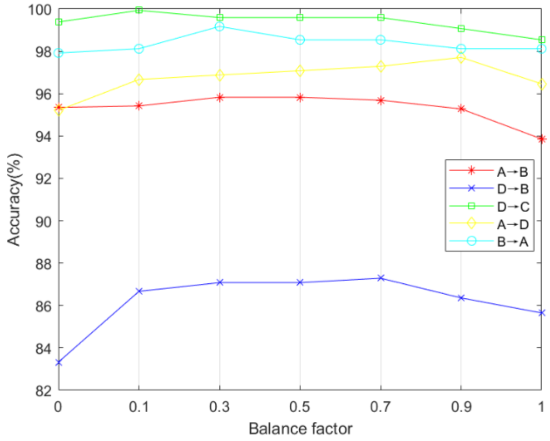


FIGURE 9. The performance of NRDAN with diverse balance factors.

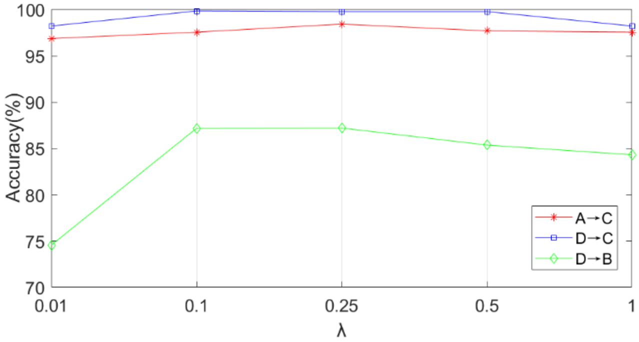


FIGURE 10. Parameter sensitivity analysis for regularization parameter λ.

To sum up, it is necessary and effective to apply dynamic adaption into transfer learning.

E. PARAMETER SENSITIVITY ANALYSIS

We investigate the effect of regularization parameter λ through experiments with a range of $\lambda \in \{0.01, 0.1, 0.25, 0.5, 1\}$. Figure 10 provides the diagnosis performance of NRDAN by varying λ on task A→C, D→B, and

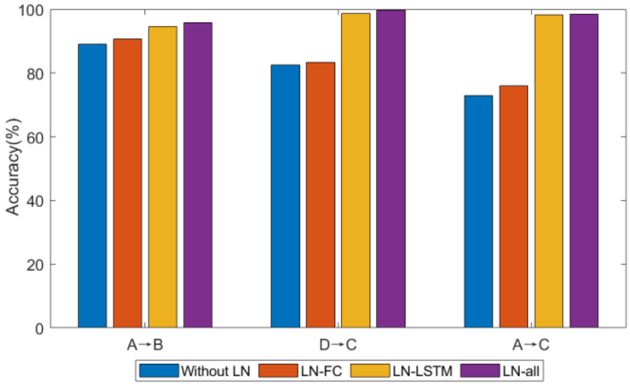


FIGURE 11. The performance of NRDAN with different LN operations.

TABLE 4. Investigation into the structure of LSTM. The numbers in the square bracket indicate the quantity of LSTM layers in the feature extractor and how many neurons each layer contains. For example, the case 1 represents that the number of LSTM layers in the feature extractor is four and each layer contains 100 neurons. Each case is tested with the same fully connected layers.

Case	Structure of LSTM	Accuracy(%)
Case1	[100,100,100,100]	96.67
Case2	**[200,200,200,200]**	**98.5**
Case3	[400,400,400,400]	97.08
Case4	[800,800,800,800]	74.79
Case5	[200,200]	96.87
Case6	[200,200,200]	98.38
Case7	[200,200,200,200,200]	97.08
Case8	[200,200,200,200,200,200]	75.21

D→C. These accuracy curves are bell-shaped, i.e. the diagnosis accuracy first increases and then decreases when λ increases gradually. The experimental results show that the regularization parameter between 0.1 and 0.5 is beneficial to realize satisfying transfer performance.

F. ABLATION STUDY FOR LN

We implement ablation study to investigate the effect of the position where LN layer incorporates on the transfer performance. As shown in Figure 11, experiments are conducted on task A→B, A→C, and D→C with four settings of LN, i.e. without LN, normalizing fully connected layers, normalizing LSTM layers, and normalizing all the layers of feature extractor. It can be observed that NRDAN performs better when implementing LN operations and achieves the best diagnosis performance when normalizing all the layers in the feature extractor. Normalizing LSTM layers can also contribute to a satisfying performance which is approaching the case of normalizing all the layers. By contrast, the effect of normalizing fully connected layers is just slightly better than the baseline without LN. Therefore, LN is particularly beneficial for LSTM layers.

In this paper, the reason for LN achieving significant improvement in performance can be explained as the following reasons. 1) First of all, as suggested in [40], LN is particularly beneficial for recurrent neural networks, which is also confirmed in our paper (The results show that the normalization of LSTM layers plays a major role). 2) It is worth noting that the method NRDAN_LN (NRDAN without LN) is performed without any other regularization (e.g. dropout and weight regularization). This is an important reason for LN showing great improvement in performance in the experiments. As we all know, the LSTM is easier to overfit in comparison with the CNN. 3) According to the Figure 7, the internal shift of each class is sharply reduced and the distance between classes increases a lot due to LN, which is helpful to align the data distribution and beneficial for domain adaption.

G. STRUCTURE OF THE FEATURE EXTRACTOR

In this part, we explore the influence of the structure of the feature extractor on the diagnosis accuracy. Because the LSTM layers in the feature extractor plays a major role in feature extraction, we primarily determine the structure of LSTM through experiments. The network is trained with data from the training set of dataset B and tested on the testing set of dataset B. As shown in Table 4, a total of eight cases are listed in the table and we adopt the diagnosis accuracy as the measurement to evaluate the structure of the feature extractor. According to the experiment results shown in Table 4, we finally adopt the parameters of case 2 in this paper.

VI. CONCLUSION

Generally, distribution divergence between domains is reduced by adapting the marginal distribution or jointly aligning the marginal and conditional distributions so that the classifier trained by labeled source data merely can correctly classify target data. This paper proposes a novel diagnosis framework named NRDAN, which could dynamically adjust the relative importance of marginal and conditional distributions in the transfer process to better fit in with real-world applications. NRDAN is based on LSTM and adopts LN to normalize the outputs of hidden layers. Extensive experiments, which contain transfer tasks between not only various operating conditions but also different machines, are conducted and the experimental results show NRDAN is effective and outperforms other state-of-the-art transfer learning methods. Finally, we further explore the reason of superiority of dynamic adaption. NRDAN is capable of dealing with more general cases and boosting the popularization of intelligent fault diagnosis in practical applications. Future work will pay attention to further evaluation on other types of fault datasets and applying NRDAN to real-world applications.

REFERENCES

[1] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep learning and its applications to machine health monitoring: A survey," *J. Latex Class Files*, vol. 14, no. 8, pp. 1–14, 2015.

[2] M. He and D. He, "Deep learning based approach for bearing fault diagnosis," *IEEE Trans. Ind. Appl.*, vol. 53, no. 3, pp. 3057–3065, May 2017.

[3] Z. Chen, S. Deng, X. Chen, C. Li, R.-V. Sanchez, and H. Qin, ''Deep neural networks-based rolling bearing fault diagnosis,'' *Microelectron. Rel.*, vol. 75, pp. 327–333, Aug. 2017.

[4] Y. LeCun, Y. Bengio, and G. Hinton, ''Deep learning,'' *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[5] M. Wang, H.-X. Li, X. Chen, and Y. Chen, ''Deep learning-based model reduction for distributed parameter systems,'' *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 46, no. 12, pp. 1664–1674, Dec. 2016.

[6] P. Tamilselvan and P. Wang, ''Failure diagnosis using deep belief learning based health state classification,'' *Rel. Eng. Syst. Saf.*, vol. 115, pp. 124–135, Jul. 2013.

[7] M. Gan, C. Wang, and C. Zhu, ''Construction of hierarchical diagnosis network based on deep learning and its application in the fault pattern recognition of rolling element bearings,'' *Mech. Syst. Signal Process.*, vols. 72–73, pp. 92–104, May 2016.

[8] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, ''Machinery health prognostics: A systematic review from data acquisition to RUL prediction,'' *Mech. Syst. Signal Process.*, vol. 104, pp. 799–834, May 2018.

[9] S. Jialin Pan and Q. Yang, ''A survey on transfer learning,'' *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[10] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, ''A theory of learning from different domains,'' *Mach. Learn.*, vol. 79, nos. 1–2, pp. 151–175, May 2010.

[11] M. Long, Y. Cao, J. Wang, and M. I. Jordan, ''Learning transferable features with deep adaptation networks,'' in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 97–105.

[12] C. Persello and L. Bruzzone, ''Kernel-based domain-invariant feature selection in hyperspectral images for transfer learning,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 5, pp. 2615–2626, May 2016.

[13] L. Wen, L. Gao, and X. Li, ''A new deep transfer learning based on sparse auto-encoder for fault diagnosis,'' *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 49, no. 1, pp. 136–144, Jan. 2019.

[14] T. Han, C. Liu, W. Yang, and D. Jiang, ''Deep transfer network with joint distribution adaptation: A new intelligent fault diagnosis framework for industry application,'' *ISA Trans.*, vol. 97, pp. 269–281, Feb. 2020.

[15] J. Xie, L. Zhang, L. Duan, and J. Wang, ''On cross-domain feature fusion in gearbox fault diagnosis under various operating conditions based on transfer component analysis,'' in *Proc. IEEE Int. Conf. Prognostics Health Manage. (ICPHM)*, Jun. 2016, pp. 1–6.

[16] W. Lu, B. Liang, Y. Cheng, D. Meng, J. Yang, and T. Zhang, ''Deep model based domain adaptation for fault diagnosis,'' *IEEE Trans. Ind. Electron.*, vol. 64, no. 3, pp. 2296–2305, Mar. 2017.

[17] L. Guo, Y. Lei, S. Xing, T. Yan, and N. Li, ''Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data,'' *IEEE Trans. Ind. Electron.*, vol. 66, no. 9, pp. 7316–7325, Sep. 2019.

[18] X. Li, W. Zhang, and Q. Ding, ''Cross-domain fault diagnosis of rolling element bearings using deep generative neural networks,'' *IEEE Trans. Ind. Electron.*, vol. 66, no. 7, pp. 5525–5534, Jul. 2019.

[19] Z. An, S. Li, J. Wang, Y. Xin, and K. Xu, ''Generalization of deep neural network for bearing fault diagnosis under different working conditions using multiple kernel method,'' *Neurocomputing*, vol. 352, pp. 42–53, Aug. 2019.

[20] B. Yang, Y. Lei, F. Jia, and S. Xing, ''An intelligent fault diagnosis approach based on transfer learning from laboratory bearings to locomotive bearings,'' *Mech. Syst. Signal Process.*, vol. 122, pp. 692–706, May 2019.

[21] Z. Tong, W. Li, B. Zhang, F. Jiang, and G. Zhou, ''Bearing fault diagnosis under variable working conditions based on domain adaptation using feature transfer learning,'' *IEEE Access*, vol. 6, pp. 76187–76197, 2018.

[22] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, ''Analysis of representations for domain adaptation,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 137–144.

[23] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, ''Deep domain confusion: Maximizing for domain invariance,'' 2014, *arXiv:1412.3474*. [Online]. Available: http://arxiv.org/abs/1412.3474

[24] M. Long, J. Wang, and M. I. Jordan, ''Deep transfer learning with joint adaptation networks,'' in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 2208–2217.

[25] A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur, ''Optimal kernel choice for large-scale two-sample tests,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1205–1213.

[26] B. Zhang, W. Li, X.-L. Li, and S.-K. Ng, ''Intelligent fault diagnosis under varying working conditions based on domain adaptive convolutional neural networks,'' *IEEE Access*, vol. 6, pp. 66367–66384, 2018.

[27] X. Li, W. Zhang, and Q. Ding, ''A robust intelligent fault diagnosis method for rolling element bearings based on deep distance metric learning,'' *Neurocomputing*, vol. 310, pp. 77–95, Oct. 2018.

[28] J. Wang, Y. Chen, S. Hao, W. Feng, and Z. Shen, ''Balanced distribution adaptation for transfer learning,'' in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 1129–1134.

[29] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, ''Transfer feature learning with joint distribution adaptation,'' in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2200–2207.

[30] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, ''Learning representations by back-propagating error,'' *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[31] J. L. Elman, ''Finding structure in time,'' *Cognit. Sci.*, vol. 14, pp. 179–211, 1990.

[32] D. Bahdanau, K. Cho, and Y. Bengio, ''Neural machine translation by jointly learning to align and translate,'' 2014, *arXiv:1409.0473*. [Online]. Available: https://arxiv.org/abs/1409.0473

[33] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, ''Exploring the limits of language modeling,'' 2016, *arXiv:1602.02410*. [Online]. Available: http://arxiv.org/abs/1602.02410

[34] A. Graves, A.-R. Mohamed, and G. Hinton, ''Speech recognition with deep recurrent neural networks,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649.

[35] C.-Y. Wu, A. Ahmed, A. Beutel, A. J. Smola, and H. Jing, ''Recurrent recommender networks,'' in *Proc. 10th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2017, pp. 495–503.

[36] S. Hochreiter and J. J. Schmidhuber, ''Long short-term memory,'' *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[37] S. Ioffe and C. Szegedy, ''Batch normalization: Accelerating deep network training by reducing internal covariate shift,'' 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[38] C. Laurent, G. Pereyra, P. Brakel, Y. Zhang, and Y. Bengio, ''Batch normalized recurrent neural networks,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2657–2661.

[39] T. Cooijmans, N. Ballas, C. Laurent, Ç. Gäláehre, and A. Courville, ''Recurrent batch normalization,'' 2016, *arXiv:1603.09025*. [Online]. Available: http://arxiv.org/abs/1603.09025

[40] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, ''Layer normalization,'' 2016, *arXiv:1607.06450*. [Online]. Available: http://arxiv.org/abs/1607.06450

[41] Y. Mansour, M. Mohri, and A. Rostamizadeh, ''Domain adaptation: Learning bounds and algorithms,'' in *Proc. COLT*, 2009, pp. 1–16.

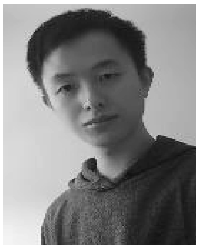
[42] (2000). *Case Western Reserve University Bearing Data Center Website*. [Online]. Available: http://csegroups.case.edu/bearingdatacenter/home

[43] H. Qiu, J. Lee, J. Lin, and G. Yu, ''Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics,'' *J. Sound Vib.*, vol. 289, nos. 4–5, pp. 1066–1090, Feb. 2006.

[44] B. Wang, Y. Lei, N. Li, and N. Li, ''A hybrid prognostics approach for estimating remaining useful life of rolling element bearings,'' *IEEE Trans. Rel.*, vol. 69, no. 1, pp. 401–412, Mar. 2020.

[45] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, ''Domain adaptation via transfer component analysis,'' *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[46] Y. Ganin, ''Domain-adversarial training of neural networks,'' *J. Mach. Learn. Res.*, vol. 17, no. 59, pp. 1–35, 2016.

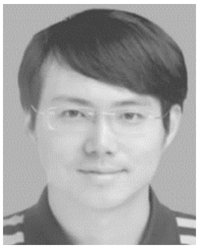


CHENGDONG ZHENG is currently pursuing the M.S. degree with Shanghai University, Shanghai, China. His research interests include deep learning, transfer learning, active control, and their application in the smart bearing.

XIAOJING WANG was born in Shanghai, China, in 1970. She received the Ph.D. degree in mechanical engineering from Shanghai University, Shanghai, China. She is currently a Professor of mechanical engineering with Shanghai University. Her research interests include vibration reduction and active control of bearing, smart bearing, and rotor dynamics.

YIFAN HAO is currently pursuing the M.S. degree with Shanghai University, Shanghai, China. Her research interests include vibration reduction and active control for journal bearing.

KE WANG is currently pursuing the M.S. degree with Shanghai University, Shanghai, China. His research interests include dynamics of sliding bearing and the smart vibration reduction of bearing.



XIN XIONG received the Ph.D. degree from Zhejiang University, Hangzhou, China, in 2012. He is currently an Assistant Professor of mechanical engineering with the School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China. His current research interests include fault diagnosis of mechanical systems, remaining useful life prediction of mechanical components, and rotor dynamics.