ISA Transactions 97 (2020) 269–281

Contents lists available at ScienceDirect

ISA Transactions

journal homepage: www.elsevier.com/locate/isatrans

Practice article

Deep transfer network with joint distribution adaptation: A new intelligent fault diagnosis framework for industry application

Te Han [a,b], Chao Liu [a,c,*], Wenguang Yang [a,b], Dongxiang Jiang [a,b]

[a] Department of Energy and Power Engineering, Tsinghua University, Beijing 100084, China
[b] State Key Laboratory of Control and Simulation of Power System and Generation Equipment, Tsinghua University, Beijing 100084, China
[c] Key Laboratory for Thermal Science and Power Engineering of Ministry of Education, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:
Received 30 January 2019
Received in revised form 2 August 2019
Accepted 4 August 2019
Available online 12 August 2019

Keywords:
Transfer learning
Domain adaptation
Joint distribution adaptation
Intelligent fault diagnosis
Convolutional neural networks

ABSTRACT

In recent years, an increasing popularity of deep learning model for intelligent condition monitoring and diagnosis as well as prognostics used for mechanical systems and structures has been observed. In the previous studies, however, a major assumption accepted by default, is that the training and testing data are taking from same feature distribution. Unfortunately, this assumption is mostly invalid in real application, resulting in a certain lack of applicability for the traditional diagnosis approaches. Inspired by the idea of transfer learning that leverages the knowledge learnt from rich labeled data in source domain to facilitate diagnosing a new but similar target task, a new intelligent fault diagnosis framework, i.e., deep transfer network (DTN), which generalizes deep learning model to domain adaptation scenario, is proposed in this paper. By extending the marginal distribution adaptation (MDA) to joint distribution adaptation (JDA), the proposed framework can exploit the discrimination structures associated with the labeled data in source domain to adapt the conditional distribution of unlabeled target data, and thus guarantee a more accurate distribution matching. Extensive empirical evaluations on three fault datasets validate the applicability and practicability of DTN, while achieving many state-of-the-art transfer results in terms of diverse operating conditions, fault severities and fault types.

1. Introduction

In modern industry, machines and equipment are developing towards the direction of high-precision, high-efficiency, more automatic and more complicated, making the breakdown or even accidents more frequent. Intelligent monitoring and fault diagnosis systems, in a broad sense, have always been key to attaining the enhancement of security and reliability of industry equipment [1]. Over the past decade, various attempts have been made to design efficient algorithms or new ways for achieving superior diagnostic performance. These studies usually merge advanced signal processing algorithms and machine learning techniques to process machine data and make diagnostic decisions intelligently, leading to impressive results in many diagnosis cases [2–6].

Marvelous success on diverse intelligent fault diagnosis frameworks have been reported over the past decade [7–14]. However, two latent problems in these works may restrict extensive and flexible industry applications. (1) Most of designed methods or algorithms are validated based on a assumption: the training data and testing data follow a similar distribution. Take bearing fault diagnosis as an example. Lei et al. [1] utilized ensemble empirical mode decomposition (EEMD) and statistical parameters to extract features, and wavelet neural network (WNN) to intelligently classify and diagnose bearing health conditions. Verstraete et al. [10] designed a deep feature learning method using time–frequency images and convolutional neural networks (CNN) for bearing fault diagnosis. Feng et al. [11] presented a local connection network constructed by stacked auto-encoder (SAE) to extract shift-invariant features from bearing fault signals. Numerous other works can be found in related reviews [15,16]. In these works, the monitored signal is generally divide into many segments, i.e., samples. These samples are randomly partitioned into the training data and testing data. In this manner, the designed methods or algorithms are actually validated in the same data distribution. These reported works contribute to the development of more effective diagnosis methods utilizing expert knowledge or adaptively feature learning, while ignore the fact of distribution discrepancy. Due to the multiple loading conditions, working environments and fault severities for bearing, the distributions between training data and testing data are different in real situations. The diagnostic model is generally learned with the training data of limited conditions, and the

* Corresponding author at: Department of Energy and Power Engineering, Tsinghua University, Beijing 100084, China.
E-mail address: cliu5@tsinghua.edu.cn (C. Liu).

https://doi.org/10.1016/j.isatra.2019.08.012
0019-0578/

generalization error cannot be large enough to guarantee the success on the testing data for diverse application domains. (2) The success of intelligent fault diagnosis methods relies on the supervised training of labeled data. A large amount of training data is often required so that the hierarchical features can be fully learned and a stronger generalization ability can be achieved by the deep networks with millions of parameters. However, in real problems, especially for those unseen conditions, collecting sufficient typically labeled samples is usually an expensive or even impossible task. Considering the previous problem, it is even impossible to relearn the diagnostic model from sufficient fault samples for new diagnosis tasks.

Consequently, there is, in particular, a need to develop a framework that can solve the problem of distribution discrepancy between training data and testing data. And the obtained useful information from historical training tasks can be further borrowed to assist the diagnosis of new but similar testing tasks, instead of reconstructing and re-training a new diagnosis model from scratch. This could make the diagnosis systems more practical and flexible to be deployed in a variety of applications. Transfer learning provides a new tool for these problems and has proven its wide applicability spanning through various fields [17–21]. Different from traditional machine learning procedure, transfer learning framework focuses on enhancing the model performance and reduces the quantity of required sample in target domain by leveraging the transferable features or knowledge from source domain [22]. It is definitely a promising way for tackling the aforementioned challenges. Motivated by this, a new intelligent fault diagnosis framework, i.e., deep transfer network (DTN), is proposed in this work. A base network, CNN, is first learned on sufficient training data (source domain). Then, a joint distribution adaptation (JDA) scheme is devised to reduce the discrepancy and learn the shared features representation between training data and real-time monitored testing data (target domain) in real deployment scenario. Finally, the adapted model can be competent in the target tasks. On the basis of absorbing and drawing upon informed research, DTN utilizes the widely reported deep learning-based fault diagnosis framework, and further realize the goal of cross-domain fault diagnosis, making the diagnosis framework be more fit to practical engineering.

The major contributions of this work can be summarized as: (1) A new intelligent fault diagnosis framework, exploiting the idea of transfer learning, is proposed for more practical application scenarios. The DTN with JDA method can utilize the unlabeled target data to realize domain adaption, which conforms better with the real situation. (2) Compared with traditional maximum mean discrepancy (MMD) minimization in marginal distribution, the superiority of JDA has been demonstrated in the field of fault diagnosis. The experimental results show the DTN with JDA is able to provide a more accurate matching of feature distribution between domains. (3) Three benchmark datasets, i.e., wind turbine dataset, bearing dataset and gearbox dataset, are used for extensive empirical evaluations. Three novel transfer scenarios in mechanical fault diagnosis, namely, various operating conditions, diverse fault severity levels and different fault types, are considered. The presented framework achieves superior diagnosis performance in comparison with the state-of-the-art algorithms including supervised and domain adaption algorithms.

The remainder of this work includes. In Section 2, the literatures about intelligent fault diagnosis are reviewed to further show the advantages and limitations of existing studies. In Section 3, the preliminaries are described. In Section 4, the proposed intelligent fault diagnosis framework, DTN with JDA, is presented. The comparison methods and implementation details are also explained. The experiments, results and discussion are given in Section 5 and Section 6, respectively. Finally, the conclusions are drawn in Section 7.

2. Related works

To date, according to the procedure of diagnosis framework, most of previous studies can be divided into two stages. In the first stage, the diagnosis framework mainly includes three steps (1) data collection, (2) feature extraction and selection, and (3) fault classification (Fig. 1(a)) [7,23,24]. In this framework, a massive efforts have been devoted to manual feature extraction and selection. This process benefits from the extensive domain expertise captured by diagnosis specialist, but inevitably requires a large expenditure of labor and time. Besides, the designed features always aim at special application object, and thus have limited adaptability when facing new diagnosis issues or changing the physical characteristics of the original systems [25]. Moreover, the final decision-making resorts to pattern recognition methods, and the diagnostic performance is often sensitive to model parameters, such as the penalty factor and kernel function parameter in support vector machine (SVM), indicating the additional parameter optimization procedure need to be executed [26]. To tackle these issues, an adaptive feature learning based diagnostic framework with deep learning technology is emerging in the second stage [12,27–29]. With the aid of multilayer nonlinear modeling scheme, this framework provides an end-to-end learning procedure from input signals to output diagnosis labels, as shown in Fig. 1(b). The training process, in which the error estimated by the upper classification layer is back-propagated to update the parameters of lower feature descriptor layers, further guarantees the co-adaptation of the whole network. In the past few years, the deep learning models, including SAE [11,30], deep belief networks (DBN) [31] and, in particular, CNN [10,13,14,32–35], have gained much popularity and success in mechanical fault diagnosis issues, showing an extraordinary feature learning and fitting capacity. These studies significantly facilitate the application of artificial intelligence in fault diagnosis. However, as mentioned above, most of developed methods are still categorized into traditional machine learning without considering the problem of domain discrepancy. For real industrial diagnosis tasks,typically fault data is often limited, and the training data are generally from the experimental environment or the historical database of associated equipment. Due to the complex application scenarios of mechanical equipment, the real-time testing data may follow the different feature distribution. Consequently, the researches on cross-domain fault diagnosis have a significant practical sense.

In recent years, the broad application prospect of transfer learning has been viewed in different research areas. Several comprehensive surveys were made to review the present development of transfer learning [36,37]. In the surveys, transfer learning technology is categorized into many branches, such as inductive transfer learning, transductive transfer learning and unsupervised transfer learning. Domain adaptation, as a transductive transfer learning, fits the situation where the source domain labels are available, the target domain labels are unavailable. As this situation is normally seen in practical problems for various fields, the domain adaptation has also received wide attention. Some algorithms concerned, such as transfer joint matching (TJM) [38], transfer component analysis (TCA) [39], joint distribution adaptation (JDA) [40] and deep transfer learning [41–43] have been designed gradually for image classification. In intelligent fault diagnosis, there are only a few works considering the application of transfer learning to strengthen the applicability and flexibility of diagnosis framework for diverse domain tasks, as shown in Fig. 1(c). In [44], the authors developed a SAE based domain adaptation method for bearing diagnosis across diverse operating conditions, where a MMD term is utilized to measure the domain discrepancy. In [45], the authors proposed a domain

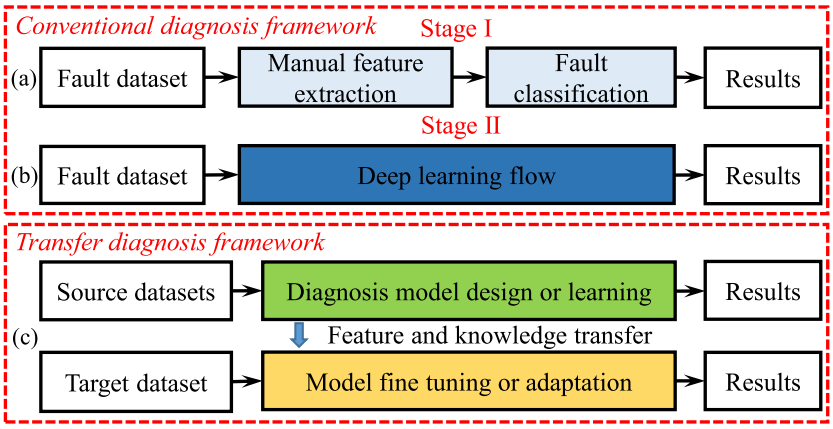

Fig. 1. Intelligent fault diagnosis framework. (a) Stage I, (b) Stage II and (c) New one.

adaptation method, which employed a MMD term to evaluate the discrepancy of normal category between source and target domains, and retained the sophisticated fault features with a weight regularization term. These studies have preliminarily explored the effectiveness of transfer learning in the field of intelligent fault diagnosis, but further works are needed to improve this framework in the following two aspects. (1) The transfer scenario should be extended to more challenging diagnosis tasks, such as the diverse fault severity levels and diverse fault types. (2) The previous studies only adapted the marginal distribution without considering the conditional distribution, leading to the neglect of the discrimination structures in rich labeled source data. Jointly reducing the discrepancy in both marginal distribution and conditional distribution may hold the potential to achieve superior transfer performance.

3. Preliminaries

3.1. Convolutional neural network

CNN, as a type of most effective deep learning models, has been widely used in image processing, computer vision and speech recognition. Typically, a CNN is composed of three types of layers, which are convolutional layers, pooling layers and fully-connected layers. The first step of CNN is to convolve the input signal with a set of filter kernels (1D for time-series signal and 2D for image). All the feature activations by convolution operation at different locations constitute the feature map. A nonlinear activation function, generally rectified linear unit (ReLU), is applied on the sum of feature maps. The operation of convolutional layer can be expressed as:

$$c_n^r = ReLU(\sum_m v_m^{r-1} * w_n^r + b_n^r) \tag{1}$$

where c_n^r is the nth output of convolutional layer r, n represents the number of filter in layer r, w_n^r and b_n^r are the nth filter and bias of layer r respectively, v_m^{r-1} is the mth output from previous layer $r-1$, $*$ denotes the convolution operation. The obtained feature map is then processed with a pooling layer by taking the mean or maximum feature activation over disjoint regions. By cascading the combination of convolutional layer and pooling layer, a multi-layer structure is built for feature description. Finally, the fully-connected layers, just like the layers in multi-layer neural network, are employed for classification. Given the training set $\{X_j\}_j$, the learning process of a CNN with K convolutional layers, including the parameters of filters $\{\boldsymbol{W}_i\}_{i=1}^K$, the biases $\{b_i\}_{i=1}^K$ and classification layers U, can be defined as an optimization task:

$$\min_{\{W_i\}_{i=1}^K, \{b_i\}_{i=1}^K} \sum_j \ell(h(\boldsymbol{X}_j), f(\{\boldsymbol{W}_i\}_{i=1}^K, \{b_i\}_{i=1}^K, \boldsymbol{U})) \tag{2}$$

where ℓ means the loss function to calculate the cost between true label $h(X)$ and predicted label by CNN model $f(\boldsymbol{X}, \{\boldsymbol{W}_i\}_{i=1}^K, \{b_i\}_{i=1}^K, \boldsymbol{U})$.

3.2. Transfer learning

For completeness, the definitions of transfer learning are first presented.

Definition 1 (*Domain*)**.** A domain $\mathcal{D}$ is composed of two components: a feature space X and a marginal probability distribution $P(X)$, where $X = \{x_1, \ldots, x_n\} \in \mathcal{X}$ is a particular training dataset, i.e., $\mathcal{D} = \{\mathcal{X}, P(X)\}$.

Definition 2 (*Task*)**.** A task $\mathcal{T}$ consists of two parts, a label space $\mathcal{Y}$ and a predictive function $f(X)$, which can be learned from the instance set X, i.e., $\mathcal{T} = \{\mathcal{Y}, f(X)\}$. Also, $f(X) = Q(Y|X)$ is the conditional probability distribution.

Definition 3 (*Transfer Learning*)**.** Given a source domain $\mathcal{D}_s$ with a learning task $\mathcal{T}_s$ and a target domain $\mathcal{D}_t$ with a learning task $\mathcal{T}_t$, transfer learning aims to facilitate the learning process of target predictive function $f_t(X)$ in $\mathcal{D}_t$ by using the related information or knowledge in $\mathcal{D}_s$ and $\mathcal{T}_s$, where $\mathcal{D}_s \neq \mathcal{D}_t$, or $\mathcal{T}_s \neq \mathcal{T}_t$. When the $\mathcal{D}_s = \mathcal{D}_t$ and $\mathcal{T}_s = \mathcal{T}_t$, it will be categorized into traditional machine learning task.

Two remarks should be emphasized here. The condition $\mathcal{D}_s \neq \mathcal{D}_t$ means $\mathcal{X}_s \neq \mathcal{X}_t \bigvee P_s(X_s) \neq P_t(X_t)$. And the condition of $\mathcal{T}_s \neq \mathcal{T}_t$ implies $Y_s \neq Y_t \bigvee Q_s(Y_s|X_s) \neq Q_t(Y_t|X_t)$.

3.3. Maximum mean discrepancy

MMD is an index to measure the discrepancy of two distributions. Given two dataset X_s, X_t, $P_s(X_s) \neq P_t(X_t)$ and a nonlinear mapping function ϕ in a reproducing Kernel Hilbert space $\mathcal{H}$ (RKHS), the formulation of MMD can be defined as:

$$MMD_{\mathcal{H}}(X_s, X_t) = \| \frac{1}{n_s} \sum_{i=1}^{n_s} \phi(x_i^s) - \frac{1}{n_t} \sum_{i=1}^{n_t} \phi(x_i^t) \|_{\mathcal{H}}^2 \tag{3}$$

In (3), we can find that the empirical estimation of the discrepancy for two distributions is considered as the distance between the two data distributions in RKHS. A value near zero for MMD means the two distributions are matched. In transfer learning, MMD is generally used to construct the regularization term for the constraint in feature learning, making the learned feature distributions more similar between different domains.

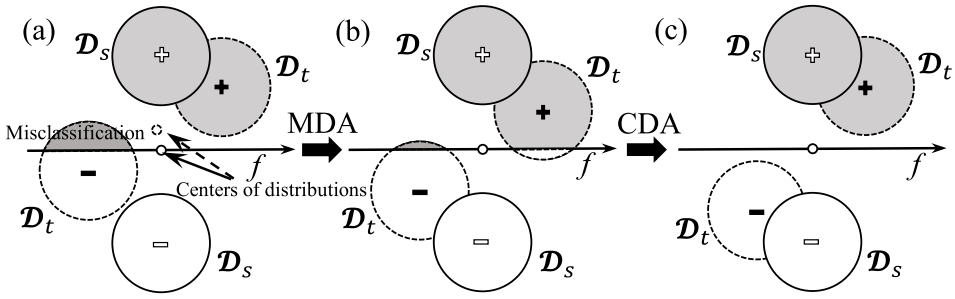

Fig. 2. An illustration of MDA and CDA, f: discriminative hyperplane, D_s: feature distribution in source domain, D_t: feature distribution in target domain.

4. Deep transfer network with joint distribution adaptation

4.1. Joint distribution adaptation

Generally, the probability distributions of diverse domains may exhibit significant difference not only in marginal distribution, which represents the cluster center of feature distributions, but also in conditional distribution for large amount of practical applications. From Fig. 2(a) to (b), it is clear the distributions for source and target domains are different. The direct use of trained discriminative hyperplane in source domain will lead to the extensive misclassification in target domain. The marginal distribution adaptation (MDA) contributes to improving transfer performance by aligning the two distribution centers. However, only adapting the marginal distributions is insufficient, since the discriminative hyperplanes may be different for diverse domain tasks. The conditional distribution adaptation (CDA), which aims to match the discriminative structures between labeled source data and unlabeled target data, is also indispensable and highly effective. An intuitive description of this consideration is illustrated from Fig. 2(b) to (c). Hence, in this part, we are dedicated to presenting a simple mathematical formulation of JDA, and further providing a specific deep transfer framework.

Problem formulation (joint distribution adaptation) In a fault diagnosis task, given a labeled source dataset $X_s = \{x_i^s, y_i^s\}_{i=1}^{n_s}$ and a unlabeled target dataset $X_t = \{x_i^t\}_{i=1}^{n_t}$, $\mathcal{X}_s = \mathcal{X}_t$, $\mathcal{Y}_s = \mathcal{Y}_t$, $P_s(X_s) \neq P_t(X_t)$, $Q_s(Y_s|X_s) \neq Q_t(Y_t|X_t)$. The weak form of transfer learning with domain adaptation is to learn a feature transform that simultaneously minimizes the discrepancy between marginal distribution and conditional distribution [39], i.e.,

$$\min D(P_s(\phi(X_s)), P_t(\phi(X_t))) \tag{4}$$

$$\textit{and} \quad \min D(Q_s(Y_s|\phi(X_s)), Q_t(Y_t|\phi(X_t))) \tag{5}$$

where D is the function to evaluate the domain discrepancy.

(1) MDA: The objective function of (4) is to minimize the distance between the two data distributions in RKHS, where we can apply MMD (3) to tackle it. The formula is described as:

$$MMD_{\mathcal{H}}^2(P_s, P_t) = \| \frac{1}{n_s} \sum_{i=1}^{n_s} \phi(x_i^s) - \frac{1}{n_t} \sum_{i=1}^{n_t} \phi(x_i^t) \|_{\mathcal{H}}^2 \tag{6}$$

where $\phi : X \to H$ is the nonlinear mapping function in RKHS.

(2) CDA: The conditional distribution in (5) is intractable in the absence of classification ground truth. We rewrite it into the following form:

$$\min D\left(\frac{Q_S(\phi(X_s)|Y_s) \cdot P_s(Y_s)}{P_s(\phi(X_s))}, \frac{Q_t(\phi(X_t)|Y_t) \cdot P_t(Y_t)}{P_t(\phi(X_t))}\right) \tag{7}$$

For our problem, we have $P(Y_s) = P(Y_t)$, as the labels of the source domain and the target domain are assumed with the same distribution. If the marginal distribution for (4) holds, the optimization problem in (7) becomes

$$\min D(Q_s(\phi(X_s)|Y_s), Q_t(\phi(X_t)|Y_t)) \tag{8}$$

The above objective function is noted as CDA. This step is essential for an accurate and robust distribution adaptation. However, it is still intractable as Y_t is unknown. Some previous studies proposed a circuitous way by exploiting the pseudo labels for target data to handle the CDA in unsupervised domain adaptation [40,46]. With the aid of the pre-trained models on labeled source data, pseudo labels for target data can be preliminarily supplied. Supposing a total of C categories and the category $c \in \{1, \ldots, C\}$, the distance index, MMD, can be defined to measure the mismatch of conditional distributions $Q_s(x_s|y_s = c)$ and $Q_t(x_t|y_t = c)$ of c category,

$$MMD_{\mathcal{H}}^2(Q_s^{(c)}, Q_t^{(c)}) = \| \frac{1}{n_s^{(c)}} \sum_{x_i^s \in \mathcal{D}_s^{(c)}} \phi(x_i^s) - \frac{1}{n_t^{(c)}} \sum_{x_j^t \in \mathcal{D}_t^{(c)}} \phi(x_j^t) \|_{\mathcal{H}}^2 \tag{9}$$

where $D_s^{(c)} = \{x_i : x_i \in \mathcal{D}_s \wedge y(x_i) = c\}$, $y(x_i)$ is the true label, and $n_s^{(c)} = |D_s^{(c)}|$, $D_t^{(c)} = \{x_j : x_j \in D_t \wedge \widehat{y}(x_j) = c\}$, $\widehat{y}(x_j)$ is the pseudo label and $n_t^{(c)} = |D_t^{(c)}|$.

It should be noted that, although there are probably many mistakes in the initial pseudo labels, one can iteratively update the pseudo labels in the stage of model optimization to obtain the optimal prediction accuracy under the current learning conditions.

(3) JDA: By integrating marginal MMD and conditional MMD, a regularization term of JDA can be written as:

$$D_{\mathcal{H}}(J_s, J_t) = MMD_{\mathcal{H}}^2(P_s, P_t) + \sum_{c=1}^{C} MMD_{\mathcal{H}}^2(Q_s^{(c)}, Q_t^{(c)}) \tag{10}$$

where J_s and J_t is the joint probability distribution of $\mathcal{D}_s$ and $\mathcal{D}_t$, respectively. Minimizing the (10) can guarantee the match both in marginal distribution and marginal distribution with sufficient statistics.

4.2. Deep transfer network

Having introduced the regularization term of JDA, we now turn to the establishment of DTN, attempting to realize the goal of domain adaptation under deep learning framework. CNN is utilized as the basic model in this work.

Generally, we can train a CNN model on the sufficient source data from scratch with the optimization task defined in (2). The cross-entropy ℓ_{ce} between estimated probability distribution and true label is served as the loss function. When applying the pre-trained CNN model to domain adaptation, a new objective function is redefined by integrating the ℓ_{ce} and regularization term of JDA, rewritten as:

$$\mathcal{L}(\Theta) = \ell_{ce} + \lambda D_{\mathcal{H}}(J_s, J_t) \tag{11}$$

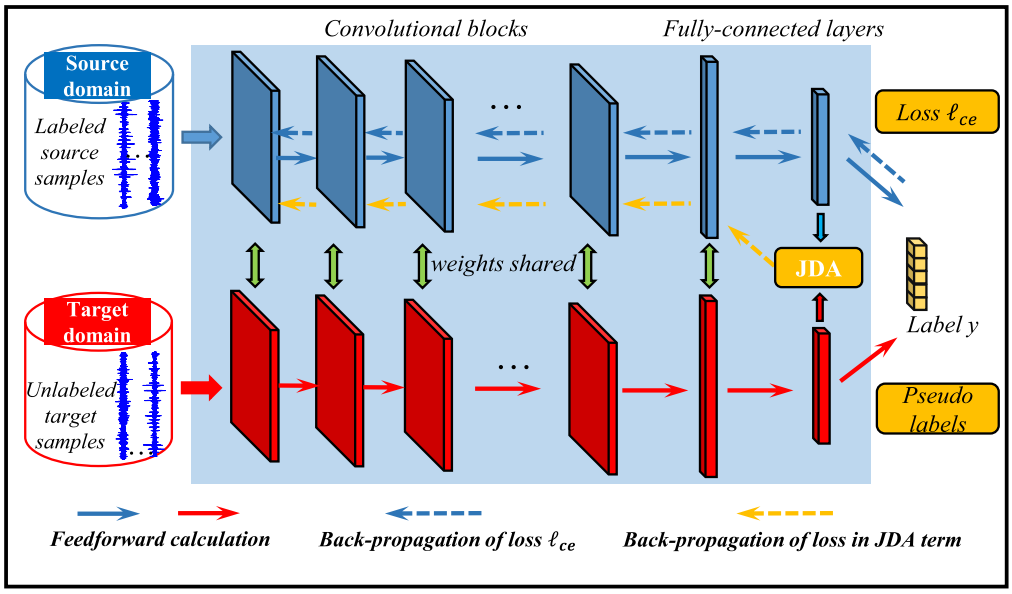

Fig. 3. The architecture of DTN for unsupervised domain adaptation.

where $\theta = \{W^i, b^i\}_{i=1}^{l}$ is the parameter collection of a CNN with l layers and λ is non-negative regularization parameter. It should be emphasized that the mapping function ϕ in RKHS $\mathcal{H}$ is the nonlinear feature transform learned by deep models herein. For CNNs, the features always change from general to specific with the increase of layer depth. The upper layers tend to represent more abstract features, which may result in a larger domain discrepancy [47]. Consequently, we deploy the regularization term on the last hidden fully-connected layer, namely the layer in front of discrimination layer, that is, $\phi(x) = h_{l-1}(x)$, where $h_{l-1}(\cdot)$ is the feature map by the nonlinear feature transform of the first $(l-1)$ layers. The JDA regularization term employed in conjunction with deep models can generate the mapping function ϕ by adaptively learning from data, and avoid to manually set the parameterized kernel function.

The architecture of proposed DTN with JDA is illustrated in Fig. 3. A domain-shared CNN is utilized to extract signal characteristics for both source data and target data. That is, the structure and weights of convolutional blocks and fully-connected layers keep consistent in source and target domains. By executing the forward pass, the two terms in (11) can be calculated, namely, the traditional cross-entropy loss ℓ_{ce} and the regularization term of JDA. Then, the backpropagation algorithm and mini-batch stochastic gradient descent (SGD) are utilized for network optimization. On the one hand, by optimizing the loss ℓ_{ce}, the model is animated to capture the discriminant structure from the labeled source data. On the other hand, by optimizing the regularization term of JDA, the model can further reduce the discrepancy of feature distributions between domains and learn domain-invariant feature representation so that the learnt discriminant structure in source domain can also be applied to target data.

The gradient of objective function with respect to network parameters is

$$\nabla_{\theta^l} = \frac{\partial \ell_{ce}}{\partial \Theta^l} + \lambda(\nabla D_{\mathcal{H}}(J_s, J_t))^T\left(\frac{\partial \phi(x)}{\partial \Theta^l}\right) \tag{12}$$

where $\frac{\partial \phi(x)}{\partial \Theta^l}$ are the partial derivatives of the output of $(l-1)$th layer with network parameters. The detailed formulations of $\nabla D^2_{\mathcal{H}}(J_s, J_t)$ are described as:

$$\nabla D_{\mathcal{H}}(J_s, J_t) = \nabla MMD^2_{\mathcal{H}}(P_s, P_t) + \sum_{c=1}^{C} \nabla MMD^2_{\mathcal{H}}(Q_s^{(c)}, Q_t^{(c)}) \tag{13}$$

Algorithm 1 Training Procedure of DTN with JDA

Input: Given the dataset $D_s = \{x_i^s, y_i^s\}_{i=1}^{n_s}$in source domain, unlabeled dataset $D_t = \{x_i^t\}_{i=1}^{n_t}$ in target domain, the architecture of deep neural network, the trade-off parameters λ.
Output: Transferred network and predicted labels for target samples
1: **begin**
2: Train a base deep network on the source dataset D_s
3: Predict the pseudo labels $\widehat{Y}_0 = \{y_i^t\}_{i=1}^{n_t}$ for target samples with base network
4: **repeat**
5: $j = j + 1$
6: Compute the regularization term of JDA according to (10)
7: Network optimization with respect to (11)
8: Update the pseudo labels $\widehat{Y}_j$ with optimized network
9: **until** convergence or $\widehat{Y}_j = \widehat{Y}_{j-1}$
10: Check the diagnosis performance of transferred network on other target samples.

$$\nabla MMD^2_{\mathcal{H}}(P_s, P_t) = \begin{cases} \frac{2}{n_s}\left(\frac{1}{n_s}\sum_{i=1}^{n_s}\phi(x_i^s) - \frac{1}{n_t}\sum_{j=1}^{n_t}\phi(x_j^t)\right), & x \in \mathcal{D}_s \\ \frac{2}{n_t}\left(\frac{1}{n_t}\sum_{j=1}^{n_t}\phi(x_j^t) - \frac{1}{n_s}\sum_{i=1}^{n_s}\phi(x_i^s)\right), & x \in \mathcal{D}_t \end{cases} \tag{14}$$

$$\text{and} \quad \nabla MMD^2_{\mathcal{H}}(Q_s^{(c)}, Q_t^{(c)}) = \begin{cases} \frac{2}{n_s^{(c)}}\left(\frac{1}{n_s^{(c)}}\sum_{x_i^s \in \mathcal{D}_s^{(c)}}\phi(x_i^s) - \frac{1}{n_t^{(c)}}\sum_{x_j^t \in \mathcal{D}_t^{(c)}}\phi(x_j^t)\right), & x \in \mathcal{D}_s \\ \frac{2}{n_t^{(c)}}\left(\frac{1}{n_t^{(c)}}\sum_{x_j^t \in \mathcal{D}_s^{(c)}}\phi(x_j^t) - \frac{1}{n_s^{(c)}}\sum_{x_i^s \in \mathcal{D}_s^{(c)}}\phi(x_i^s)\right), & x \in \mathcal{D}_t \end{cases} \tag{15}$$

4.3. Training strategy

The training procedure of this framework mainly consists of two parts: (1) the pre-training on labeled source data and (2)

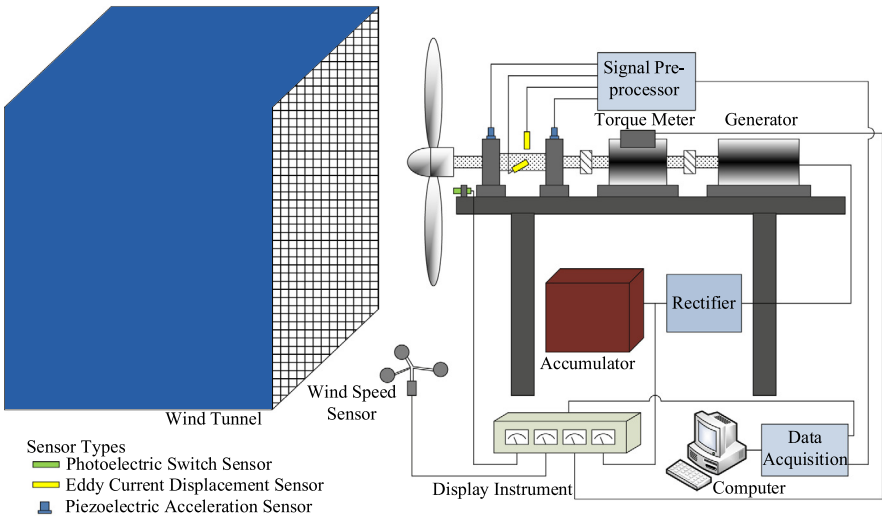

Fig. 4. Illustration of wind turbine experimental platform.

the network adaptation in target domain with the input of both labeled source data and unlabeled target data. It should be noted that the dataset is generally divided into small batches, which are fed into the network for training. A desirable batch size should be as large as possible to cover the variance of the whole dataset, whereas a too large batch size will also increase the calculation burden. It is a trade-off between transfer performance and computational effectiveness. Besides, the same amount of samples from source and target domain are used for network adaptation. When the data sizes are different across domains, the re-sampling can be applied in the smaller dataset to keep the same sample size in the source and target domains. The whole adaptation steps of DTN with JDA are listed in Algorithm 1.

5. Experiments

In this section, experiments on three mechanical fault datasets are conducted to demonstrate the efficiency, superiority as well as practical value of proposed transfer framework. Mechanical equipment may appear diverse failure mode during the long-time operation. Different faults may present different characteristics. The studies of intelligent fault diagnosis focus on classifying the signal samples from different health conditions and make diagnostic decisions automatically. In the three datasets, the frequently occurred faults in mechanical systems are artificially introduced to machines so as to simulate diverse health conditions. The vibration signal under diverse machine conditions are collected. The performance of proposed method and comparative methods can be further tested in these fault datasets.

5.1. Data description

(1) Wind Turbine Fault Dataset: The first dataset is from our wind turbine experimental platform, whose schematic diagram is illustrated in Fig. 4. This dataset contains ten machine conditions, which are health, front bearing pedestal loosening (FB), back bearing pedestal loosening (BB), rolling element fault of front bearing (RF), inner-ring fault of front bearing (IF), outer-ring fault of front bearing (OF), misalignment in horizontal direction (MH), misalignment in vertical direction (MV), variation in airfoil of blades (VB) and yaw fault (YF) respectively (corresponding labels 0–9). All these faults can basically simulate the typical failure modes from wind wheel to drive chain of a real wind turbine.

To create working conditions close to reality, we change the power of axial flow fan in the wind tunnel to generate varying loading conditions (i.e., varying wind speeds). The experiments are performed under six different wind speeds ranging from 5.8 m/s to 11.5 m/s (loads 0–5). And the corresponding speeds of wind wheel range from 255 rpm to 300 rpm. The raw vibration data is collected by accelerometers. The sampling frequency is 20 kHz. The time-domain waveforms of diverse machine conditions under load 5 are presented in Fig. 5. When the machine is health (condition 0), it is clear the vibration amplitude maintains in low level and the signal components related to rotating frequency is dominated. When the faults are introduced to machines (conditions 1 to 9), obvious impulse characteristics appear, especially for bearing-related faults (conditions 3 to 5). And the signal components are more complex.

For clarity, the denotation of A→B is utilized to represent the transfer task from source dataset A to target dataset B. In the wind turbine fault dataset, we aim to explore the transfer ability of proposed framework across diverse operating conditions. Consequently, six transfer tasks are designed for empirical evaluation (listed in Table 1). For instance, A→B: the source dataset A contains the samples of ten machine conditions under load 0–2, while the target dataset B is composed of the samples under load 3–5. In Table 1, the unlabeled target samples are utilized for domain adaptation. No information of label can be used in this process. After domain adaptation, another set of testing target samples with labels are used to evaluate the performance of transferred diagnosis model.

(2) Bearing Fault Dataset: The bearing fault dataset is an open-access dataset from Case Western Reserve University [48]. Four different bearing conditions: health, outer ring fault (OF), rolling element fault (RF) and inner ring fault (IF) (corresponding labels 0–3), are considered in this dataset. The experiments are performed under four motor speeds (1797, 1772, 1750 and 1730 rpm) at a sampling frequency of 12 kHz. For each kind of fault, single point faults with different severity levels are introduced to test bearing respectively. In most existing studies, the samples with the same fault type but different severity levels are treated as distinct categories. Indeed, the signal characteristic of certain fault type always varies with the severity level. Therefore, we aim to investigate the performance of proposed transfer framework across diverse fault severities in this dataset.

For simplicity, we select two fault severity levels with the fault diameters (FD) of 0.18 mm and 0.53 mm to construct transfer tasks: G→H, H→G. The dataset G is composed of the samples of four bearing conditions under four motor speeds and the fault diameter of OF, RF and IF cases is 0.18 mm. The dataset H is

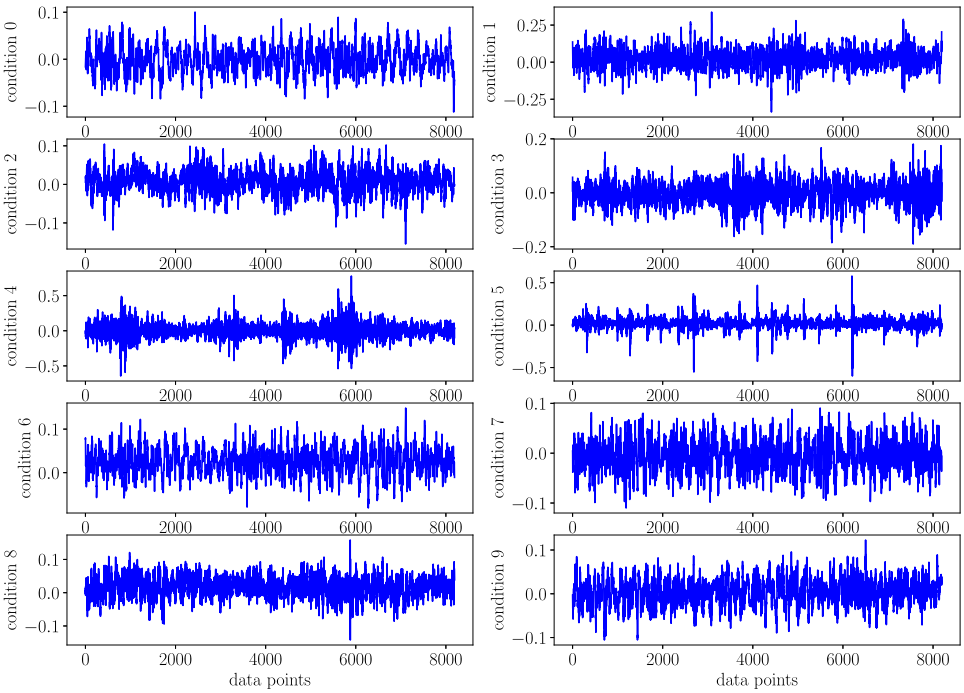

Fig. 5. Time waveforms of collected vibration data.

Table 1
Designed transfer tasks across diverse operating conditions.

Transfer tasks	Source domain	Target domain	Unlabeled target samples	Testing target sample	Machine conditions
A→B	Load 0–2	Load 3–5	24000	4000	
B→A	Load 3–5	Load 0–2	24000	4000	10
C→D	Load 2	Load 3–5	24000	4000	conditions
D→C	Load 3–5	Load 2	12000	4000	(labels
E→F	Load 2	Load 5	12000	4000	0–9)
F→E	Load 5	Load 2	12000	4000	

formed by the health samples and fault samples with 0.53 mm fault diameter.

(3) Gearbox Fault Dataset: The gearbox fault dataset collected from our single-stage cylindrical straight gearbox test rig (as shown in Fig. 6) is analyzed in the scenario where the domain discrepancy between specific fault types are expected to be bridged by transfer learning. Sometimes, it may be more practical to confirm the location of failure instead of specific types. Considering the example of gearbox, identifying the fault location, such as gear fault or bearing fault, is beneficial for monitoring and maintenance. That said, certain types of fault occurred in one component, such as bearing inner race fault or outer race fault, can be defined as one category. Besides, it may be impossible to obtain the fault data of various fault types and train a diagnosis model with high accuracy for a complex mechanical system. Consequently, the transfer performance across similar but diverse fault types is of great practical significance. In the experiments, we introduced two types of faults, i.e., gear root crack (RC) and tooth surface spalling (TS), to high-speed cylindrical gearing, and another two types of faults, i.e., outer race fault (OR) and roller fault (RO), to high-speed conical bearing. The vibration data is collected with a sampling frequency of 20 kHz.

We state three conditions of gearbox, including health, gear fault and bearing fault in this dataset (corresponding labels 0–2), and design two transfer tasks: I→J, J→I. The dataset I contains the samples of health, bearing OR and gear RC. The dataset J is formed by the samples of health, bearing RO and gear TS (see Table 2).

5.2. Comparison studies

(1) Comparison methods: The proposed framework will be compared with several state-of-the-art methods in the field of intelligent fault diagnosis: (1) SVM [26]; (2) Random forest (RF) [26]; (3) Empirical mode decomposition analysis (EMD) [1]; (4) CNN [13,33]; (5) TJM [38]; (6) TCA [39]; (7) JDA [40]; (8) DTN with MDA and (9) DTN with JDA (this work). These baseline methods can be categorized into two subsettings: the standard diagnosis methods (1)–(4) and the transfer learning based techniques (5)–(9).

In (1)–(2), the popular statistical features, such as root mean square and kurtosis, are extracted from raw data in time and frequency domains to form the input of the classifiers [7,26,49] . In (3), EMD is applied to decompose the raw signal into a sequence of intrinsic mode functions (IMF). The energy distribution of first five IMFs is calculated as the input features for classifier. In (4), using the deep learning flow, CNN. In the transfer learning based techniques (5)–(9), TJM, TCA and JDA are the shallow transfer learning methods, and thus we also extract the statistical features from raw data, then conduct the unsupervised domain adaptation, finally make diagnosis results with classifier. In deep learning flow, a comparison between the proposed method and DTN with MDA method by removing the CDA term in objective function is made. The pre-trained base network resorts to the optimal CNN model in source domain, that is, the trained model in (4).

(2) Implementation details: For (1)–(4), we use the labeled source data to train the model, which will be applied to diagnose

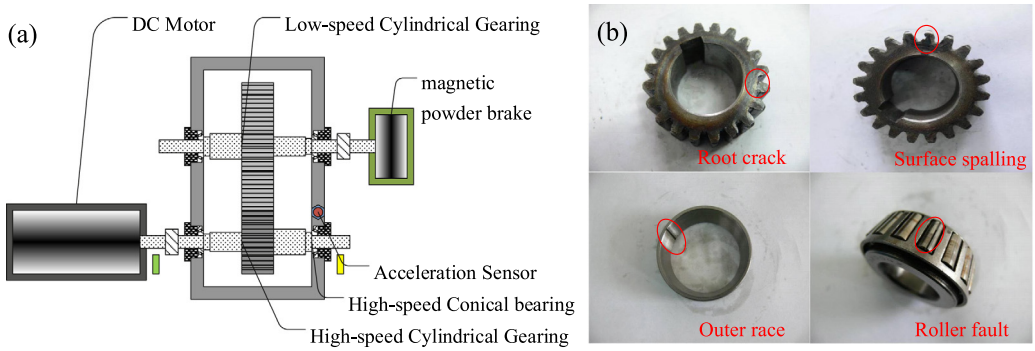

Fig. 6. The single-stage cylindrical straight gearbox test rig: (a) Schematic diagram of gearbox test rig; (b) The damaged components.

Table 2
Designed transfer tasks across diverse fault severities and types.

Transfer tasks	Source domain	Target domain	Unlabeled target samples	Testing target sample	Machine conditions
G→H	FD 0.18	FD 0.53	12000	4000	4 conditions
H→G	FD 0.53	FD 0.18	12000	4000	(labels 0–3)
I→J	H, OR, RC	H, RO, TS	12000	4000	3 conditions
J→I	H, RO, TS	H, OR, RC	12000	4000	(labels 0–2)

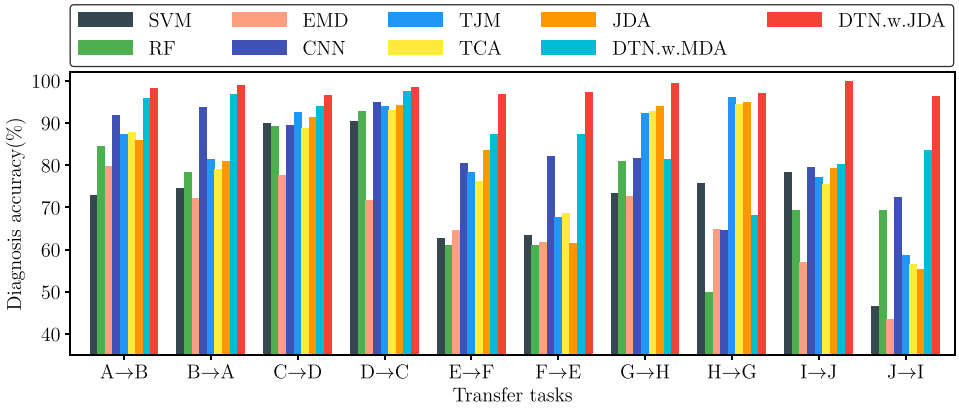

Fig. 7. Comparison of the diagnosis accuracy of diverse methods on ten transfer tasks.

Table 3
Model structure of used CNN.

No.	Layer type	Kernel number	Kernel size	Activation
1	Convolution1	16	128 × 1	ReLU
2	Max-pooling1	–	2 × 1	ReLU
3	Convolution2	32	64 × 1	ReLU
4	Max-pooling2	–	2 × 1	ReLU
5	Convolution3	64	16 × 1	ReLU
6	Max-pooling3	–	2 × 1	ReLU
7	Convolution4	128	3 × 1	ReLU
8	Max-pooling4	–	2 × 1	ReLU
9	Convolution5	256	2 × 1	ReLU
10	Max-pooling5	–	2 × 1	ReLU
11	Fully-connected	1	128	ReLU
12	Fully-connected	1	64	ReLU
13	Output	1	n	Softmax

unlabeled target data. For (5)–(7), TJM, TCA and JDA simultaneously process the labeled source data and unlabeled target data for dimension reduction. The classification model is then trained with low-dimensional features in source domain, and deployed on target domain. Herein, both the SVM and RF are adopted to achieve an optimal diagnosis performance as the final results. The RBF is adopted as the kernel in SVM. The model parameters, i.e., penalty factor and kernel function parameter, are both decided by grid search in the sense of 5-fold cross validation. In RF, the performance is substantially robust to the two structure parameters in RF, i.e., number of trees n_{try} and the number of random feature subset m_{try}. Hence, the n_{try} and m_{try} are empirically set to 500 and $\lfloor\sqrt{m}\rfloor$ respectively, where m is the dimension of input vector [26]. Referring to literatures of TJM, TCA and JDA [38–40], the adaptation regularization parameter λ is set by searching $\lambda \in \{1e^{-2}, 1e^{-1}, 1, 10, 20, 50, 100\}$. The kernel type is selected from RBF and linear, and the dimension after adaptation is optimized with the strategy of trial and error. In DTN methods, the adaptation regularization parameter λ is set by searching $\lambda \in \{1e^{-4}, 1e^{-3}, 1e^{-2}, 5e^{-2}, 1e^{-1}, 5e^{-1}, 1\}$. The learning rate of SGD is set to 0.01.

In CNN and DTN, the architecture and parameters setting are listed in Table 3. In this work, the structure of 5 convolutional and max-pooling layers is used. Considering the overfitting problem, the L2-norm regularization term are introduced for the network parameters, whose weight decay is set to 1e-4. The batch size is chosen from 16 to 128, the SGD is used as the optimizer with a learning rate of 0.01.

6. Results and discussion

6.1. Results

The diagnosis results of ten tasks are shown in Tables 4–6 and Figs. 7–9, respectively. Each result is an average of 20

Table 4
Diagnosis accuracy (%) on ten transfer tasks with different methods.

Methods	A→B	B→A	C→D	D→C	E→F	F→E	G→H	H→G	I→J	J→I	Avg
SVM	72.8 ± 1.9	74.5 ± 1.5	89.8 ± 0.8	90.4 ± 1.2	62.6 ± 1.0	63.4 ± 1.5	73.4 ± 1.3	75.7 ± 1.3	78.2 ± 0.9	46.6 ± 0.6	72.7 ± 1.2
RF	84.4 ± 0.9	78.3 ± 2.4	89.1 ± 0.3	92.7 ± 1.1	60.9 ± 0.7	61.0 ± 0.7	80.8 ± 1.1	49.9 ± 1.3	69.4 ± 0.9	69.2 ± 7.8	73.6 ± 1.7
EMD	79.8 ± 0.9	72.2 ± 0.9	77.7 ± 4.2	71.7 ± 6.7	64.5 ± 6.2	61.8 ± 1.2	72.5 ± 4.4	64.8 ± 10.7	57.0 ± 6.9	43.4 ± 2.8	66.5 ± 4.5
CNN	91.8 ± 0.3	93.8 ± 0.4	89.4 ± 3.1	94.9 ± 0.2	80.4 ± 4.0	82.1 ± 4.6	81.7 ± 5.3	64.5 ± 10.0	79.5 ± 1.5	72.3 ± 1.3	83.0 ± 3.1
TJM	87.4 ± 1.9	81.4 ± 2.4	92.5 ± 1.3	93.9 ± 0.4	78.3 ± 5.4	67.6 ± 1.3	92.2 ± 5.8	96.0 ± 6.9	77.1 ± 2.2	58.7 ± 5.2	82.5 ± 3.3
TCA	87.8 ± 1.7	79.0 ± 2.8	88.7 ± 0.5	92.9 ± 0.5	76.1 ± 4.0	68.6 ± 1.2	92.8 ± 7.2	94.4 ± 8.5	75.5 ± 4.7	56.4 ± 6.9	81.2 ± 3.8
JDA	86.0 ± 2.2	81.0 ± 2.6	91.3 ± 1.6	94.1 ± 1.3	83.6 ± 2.8	61.4 ± 3.4	93.9 ± 10.9	94.8 ± 11.0	79.2 ± 1.2	55.4 ± 5.4	82.1 ± 4.2
DTN.w.MDA	95.9 ± 2.9	96.9 ± 0.4	94.0 ± 1.1	97.4 ± 0.4	87.3 ± 1.2	87.4 ± 1.7	81.3 ± 5.0	68.2 ± 7.9	80.1 ± 1.0	83.6 ± 1.8	87.2 ± 2.3
DTN.w.JDA	98.3 ± 0.2	98.9 ± 0.5	96.6 ± 0.2	98.5 ± 0.2	96.8 ± 0.5	97.3 ± 0.2	99.3 ± 0.4	97.1 ± 8.7	99.9 ± 0.1	96.3 ± 0.5	97.9 ± 1.2

Table 5
Missing alarm rate (%) on ten transfer tasks with different methods.

Methods	A→B	B→A	C→D	D→C	E→F	F→E	G→H	H→G	I→J	J→I	Avg
SVM	18.4 ± 2.2	17.7 ± 1.1	8.0 ± 0.6	7.5 ± 0.8	29.9 ± 0.9	38.3 ± 2.2	27.0 ± 1.1	14.2 ± 2.2	15.4 ± 0.8	52.5 ± 0.4	22.9 ± 1.2
RF	12.4 ± 0.8	11.6 ± 1.7	8.5 ± 0.2	5.0 ± 0.6	40.0 ± 1.0	39.5 ± 1.5	12.7 ± 0.9	62.6 ± 0.4	24.3 ± 6.2	28.6 ± 11.7	24.5 ± 2.5
EMD	19.5 ± 0.9	27.2 ± 0.9	21.3 ± 3.7	29.0 ± 7.0	35.6 ± 7.2	37.4 ± 1.3	27.8 ± 5.0	35.4 ± 13.2	44.6 ± 6.6	58.4 ± 2.9	33.6 ± 4.9
CNN	8.2 ± 0.3	6.1 ± 0.4	10.6 ± 3.0	5.0 ± 0.2	19.7 ± 4.2	17.9 ± 4.7	18.3 ± 5.3	35.3 ± 9.9	21.0 ± 2.0	27.1 ± 1.6	16.9 ± 3.2
TJM	9.8 ± 1.0	13.4 ± 2.0	6.4 ± 1.1	5.4 ± 0.4	16.3 ± 3.4	29.6 ± 5.4	6.0 ± 4.2	2.5 ± 3.8	17.0 ± 2.7	25.7 ± 4.7	13.2 ± 2.9
TCA	9.9 ± 1.2	15.6 ± 2.5	10.4 ± 0.9	6.3 ± 0.5	19.6 ± 1.4	25.3 ± 5.0	5.3 ± 4.4	3.4 ± 4.7	16.1 ± 2.7	31.0 ± 11.2	14.3 ± 3.5
JDA	11.9 ± 1.5	13.9 ± 2.1	7.1 ± 1.5	5.4 ± 1.1	14.1 ± 2.2	46.6 ± 2.5	8.3 ± 15.7	7.7 ± 16.2	14.9 ± 1.0	35.2 ± 11.1	16.5 ± 5.5
DTN.w.MDA	4.2 ± 0.3	3.2 ± 0.5	6.0 ± 1.0	2.6 ± 0.4	12.8 ± 1.2	12.7 ± 1.8	18.9 ± 5.0	32.2 ± 7.5	19.8 ± 1.0	16.4 ± 1.8	12.9 ± 2.1
DTN.w.JDA	1.7 ± 0.2	1.1 ± 0.5	3.4 ± 0.2	1.6 ± 0.2	3.2 ± 0.5	2.6 ± 0.5	0.8 ± 0.4	2.8 ± 8.5	0.1 ± 0.1	3.7 ± 0.5	2.1 ± 1.2

Table 6
False alarm rate(%) on ten transfer tasks with different methods.

Methods	A→B	B→A	C→D	D→C	E→F	F→E	G→H	H→G	I→J	J→I	Avg
SVM	22.1 ± 1.9	25.8 ± 1.5	11.8 ± 0.7	11.1 ± 1.3	36.1 ± 0.9	36.6 ± 1.3	26.8 ± 1.0	24.3 ± 0.3	22.9 ± 1.1	53.4 ± 0.5	27.1 ± 1.1
RF	15.8 ± 0.9	21.4 ± 2.4	11.2 ± 0.4	8.7 ± 1.1	39.2 ± 0.7	38.9 ± 0.7	19.3 ± 1.1	49.3 ± 1.3	32.8 ± 0.6	31.6 ± 7.6	26.8 ± 1.7
EMD	20.1 ± 0.9	27.6 ± 0.9	21.9 ± 3.8	27.9 ± 6.6	34.8 ± 6.6	37.0 ± 2.1	27.0 ± 4.0	35.3 ± 11.2	44.0 ± 6.7	57.7 ± 3.6	33.3 ± 4.6
CNN	7.3 ± 0.3	5.4 ± 0.3	9.5 ± 2.6	4.0 ± 0.2	17.2 ± 3.8	16.6 ± 3.5	17.5 ± 5.3	48.1 ± 9.7	12.7 ± 0.7	18.1 ± 1.6	15.6 ± 2.8
TJM	12.6 ± 1.8	18.4 ± 2.4	7.6 ± 1.4	6.7 ± 0.3	20.6 ± 5.0	32.5 ± 1.5	7.7 ± 5.9	3.9 ± 6.6	22.8 ± 1.9	41.8 ± 5.0	17.5 ± 3.2
TCA	12.3 ± 1.7	21.2 ± 2.8	11.6 ± 0.6	7.4 ± 0.6	23.3 ± 3.4	30.5 ± 1.2	7.3 ± 7.2	5.6 ± 8.5	25.1 ± 4.4	43.6 ± 7.4	18.8 ± 3.8
JDA	14.1 ± 2.2	19.2 ± 2.5	8.8 ± 1.5	6.3 ± 1.4	15.7 ± 2.7	43.8 ± 5.0	5.9 ± 10.7	5.0 ± 10.5	21.1 ± 1.1	44.6 ± 4.5	18.5 ± 4.2
DTN.w.MDA	3.8 ± 0.2	2.7 ± 0.3	5.5 ± 0.9	2.4 ± 0.2	10.8 ± 0.9	11.5 ± 1.0	18.4 ± 5.2	39.6 ± 2.8	19.7 ± 1.0	12.5 ± 0.9	12.7 ± 1.3
DTN.w.JDA	1.6 ± 0.2	1.1 ± 0.4	2.3 ± 1.0	1.4 ± 0.2	2.9 ± 0.3	2.5 ± 0.3	0.7 ± 0.3	4.4 ± 13.0	0.1 ± 0.1	3.3 ± 0.4	2.0 ± 1.6

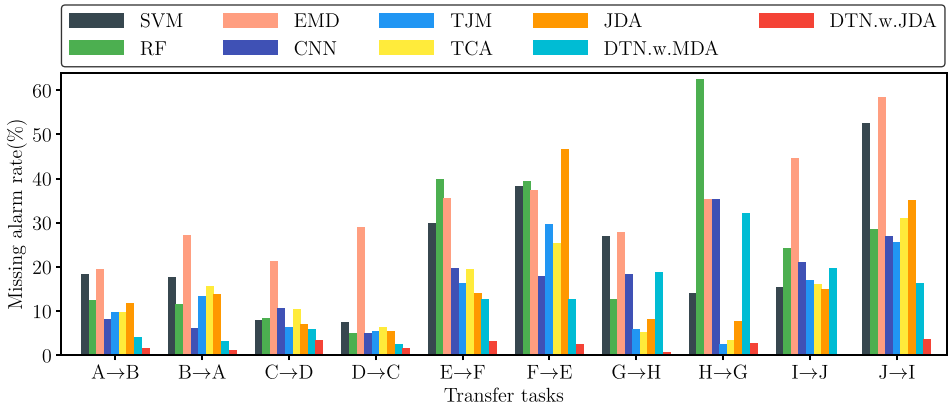

Fig. 8. Comparison of the MAR of diverse methods on ten transfer tasks.

random tests, where the training set and testing set are randomly splitted. To comprehensively show the capabilities of proposed method, three performance indices, i.e., average diagnosis accuracy, missing alarm rate (MAR), false alarm rate (FAR) [50], are reported. Several encouraging observations are firstly noted. (1) The DTN with JDA method in this work significantly outperforms the other methods. The stable average accuracies and low root-mean-square error under different transfer scenarios (over 96% for all tasks) validate the effective and robust domain adaptation ability of proposed method. The better performances of DTN with JDA can also be found for MAR and FAR (much lower than comparative methods). (2) The diagnosis performance in standard methods (the first four) is much improved with domain adaptation for most cases. Especially, the average accuracy of DTN with JDA is 97.9%, and makes a 14.9% transfer improvement, comparing with the baseline CNN, 83.0%. (3) The deep learning methods always present a superior performance to the shallow methods no matter in standard diagnosis framework or transfer learning framework, conforming its extraordinary feature learning and representation capacity as well as a stronger feature transferability. (4) By jointly adapting the marginal distribution and conditional distribution, the DTN with JDA in this work significantly promotes the adaptation ability of previous DTN

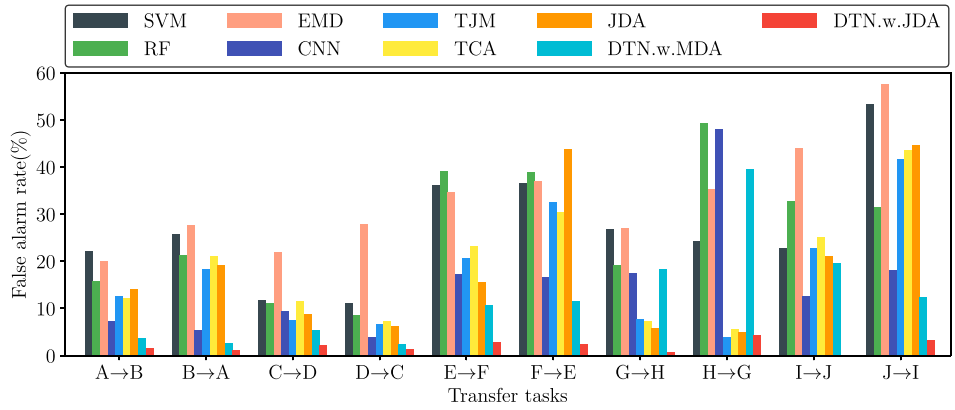

Fig. 9. Comparison of the FAR of diverse methods on ten transfer tasks.

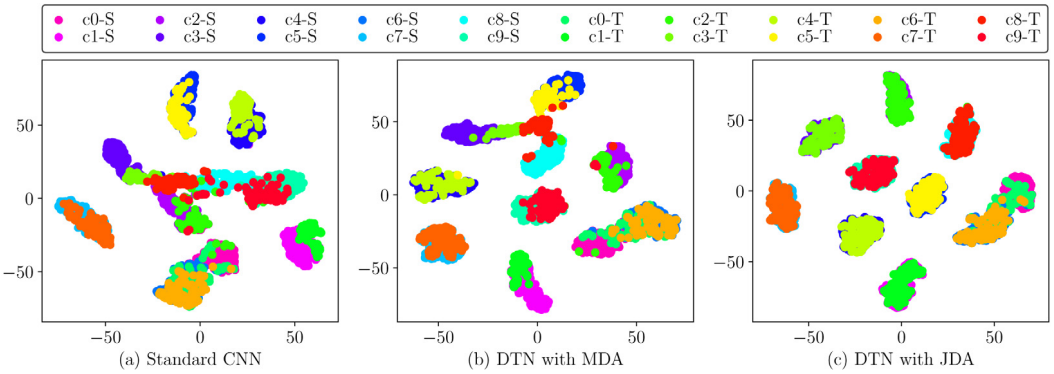

Fig. 10. Network visualization in task E→F: t-SNE is applied on the feature representation of last hidden fully-connected layer for both the source data and target data. There are total 10 categories in wind turbine dataset (corresponding labels 0–9). S represents the samples in source domain and T means the target domain. For instance, the c5-T corresponds to the samples of category 5 (inner-ring fault of bearing, as introduced above) in target domain.

Table 7
Computation complexity for diverse deep methods with wind turbine dataset in task A→B.

Methods	Time (s/epoch)	Memory (MB)
CNN	4.52	1016.6
DTN.w.MDA	7.12	1548.0
DTN.w.JDA	33.78	1548.5

with MDA, especially under the transfer scenarios of diverse fault severity levels and diverse fault types.

To show the real-time practicality of the proposed framework, the computation complexity of diverse methods in task A→B is compared in Table 7. Generally, the deep learning methods require higher computation complexity but achieve better performance than shallow methods, and thus we only listed the results of three deep methods here. Since the DTN with JDA calculates more intermediate variables in CDA, it needs more computing time and memory than standard CNN and DTN with MDA. This work focuses on the investigation of effectiveness in DTN. The training process is implemented in batch manner. Another online learning manner can train the deep network from sequential data flow. The transform of transfer diagnosis framework from batch learning to online learning will largely reduce the real-time computing time and memory [51].

6.2. Network visualization

In order to give a clear and intuitive understanding of proposed framework, t-distributed stochastic neighbor embedding (t-SNE), is utilized for network visualization. For comparison, the visualization results of standard CNN (that is, the pre-trained base network for further domain adaptation), DTN with MDA and DTN with JDA in three transfer tasks are presented in Figs. 10–12, respectively.

Task E→F is to realize the domain adaptation across diverse operating conditions. First, as shown in Fig. 10(a), most of the 10 categories of source samples are well separated with the standard CNN, while the feature distributions of same category between source and target domains are not aligned well. And even worse, a large overlapping areas can be inspected among the target samples of certain categories, such as 2, 3 and 8. These observations suggest the domain discrepancy exists not only in marginal distribution, but also in conditional distribution, which may result in the degraded diagnosis results in conventional framework. In Fig. 10(b) and (c), under the transfer learning framework, we can find the obvious improvement of distribution adaptation. In particular, the same category between domains is aligned very well by DTN with JDA, and a consonant and legible discriminant structure can be observed for both source and target categories.

Task G→H is to adapt the distribution across diverse fault severity levels. In Fig. 11(a), the standard CNN assembles the distributions of OF and IF in target domain, and the source OF and target IF are mixed, explaining the unsatisfactory accuracy in Table 4. As a contrast, in Fig. 11(c), the distribution of same category between source and target domains are well matched with JDA. Interestingly, in Fig. 11(b), we can observe that MDA relocates the target OF and IF away from the corresponding distribution in source domain. Naturally, marginal distribution only reflects the cluster structure for the feature distribution of all categories, and MDA aims to explicitly reduce the distance

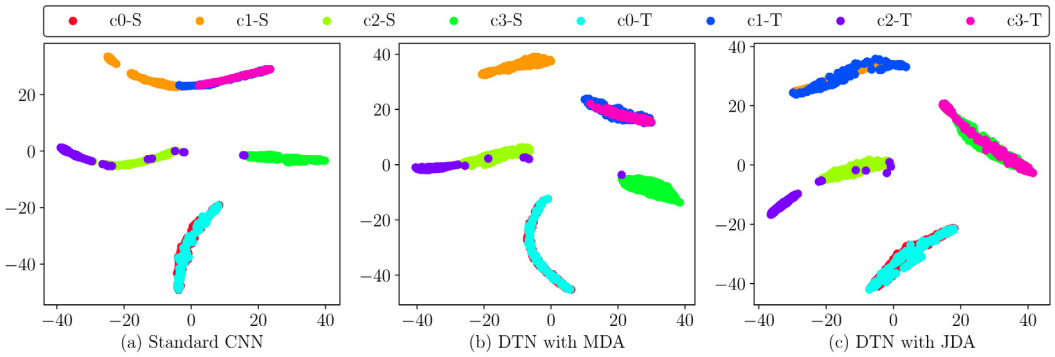

Fig. 11. Network visualization in task G→H: there are total 4 categories in bearing dataset (corresponding labels 0–3).

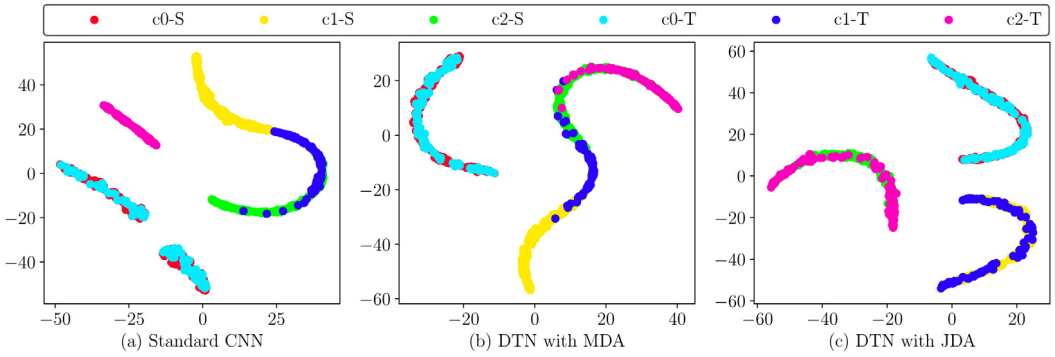

Fig. 12. Network visualization in task I→J: there are total 3 categories in gearbox dataset (corresponding labels 0–2).

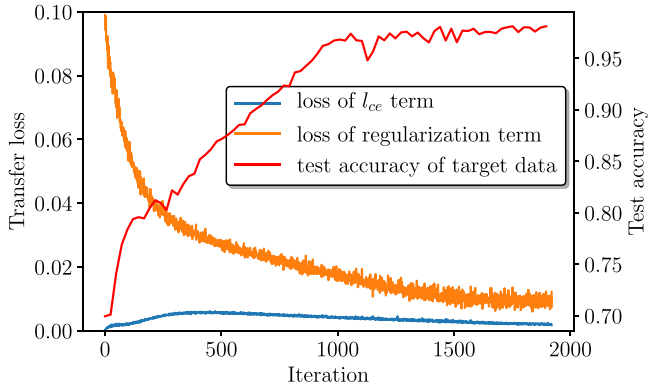

Fig. 13. The transfer loss curves and test accuracy via DTN with JDA.

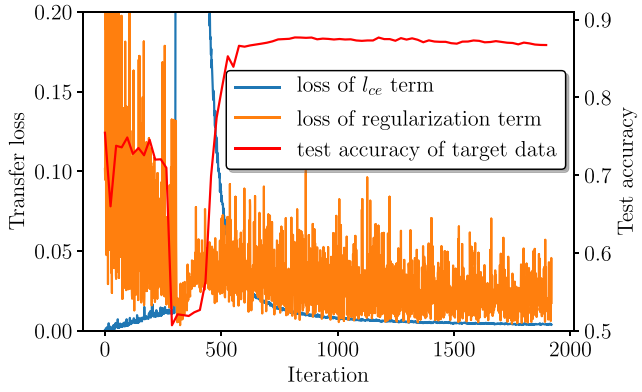

Fig. 14. The transfer loss curves and test accuracy via DTN with MDA.

between the cluster centers of different domains. When the conditional distributions are same across domains, MDA helps to correct the overall shift of feature space. However, in the field of fault diagnosis, the difference in conditional distributions may be prevalent. Consequently, unlike single MDA, the JDA which simultaneously adapts the marginal distribution and conditional distribution is promising in these cases. As shown in Fig. 12, similar results can be found in transfer task I→J across diverse fault types. Both the transfer accuracy and network visualization show that JDA supersedes the performance of MDA.

6.3. Convergence analysis

Since the additional regularization term is appended to objective function for the transfer training, the convergence analysis is necessary to illustrate the transfer ability. The transfer loss curves and test accuracy curve for DTN with JDA and DTN with MDA in task G→H are plot in Figs. 13 and 14, respectively. Here, we separately display the ℓ_{ce} term and regularization term for ease of observation. At the beginning, the losses of regularization term for the two methods are both around 0.1, and the ones of ℓ_{ce} term are almost negligible. From Fig. 13, the loss of JDA regularization term converges to a certain degree after a series of iterations, accompanied by the continued increase of test accuracy of target data. However, from Fig. 14, the loss of MDA regularization term finally fluctuates in a high level, and the test accuracy is confined around 87%. Besides, it is clear to observe the loss of ℓ_{ce} term presents an abrupt increase after around 300 iterations. Essentially, the ℓ_{ce} term and regularization term in objective function try to reduce domain discrepancy while preserving the original discriminant structure in source domain. One possible reason for the jump is that the gradient direction of the parameter optimization for regularization term conflicts with that of ℓ_{ce} term, causing a significant spike in transfer loss and test accuracy.

The analysis reveals that the use of JDA regularization term is capable of facilitating the network training and guaranteeing a stronger feature transferability.

6.4. Discussion

In the traditional intelligent fault diagnosis framework, whether for shallow methods or the deep learning methods, the diagnosis performance varies a lot on different tasks. For instance, in the first six tasks on wind turbine dataset, RF get the best performance in task C→D and D→C (89.1% and 92.7%), while degraded results in tasks E→F and F→E (60.9% and 61.0%). It is reasonable because the operating conditions between C and D are closer than those between E and F, and thus the data for C and D shares a more similar feature space, leading to a higher diagnosis accuracy. This phenomenon actually reveals the inherent drawback in conventional diagnosis framework, that the feature distribution discrepancy between source domain and target domain is neglected. The success much relies on the similarity between source and target distributions, whereas a large discrepancy across domains is common and inevitable in practical diagnosis applications. The proposed transfer diagnosis framework provides an effective measure for resolving the problem mentioned above, and DTN with JDA achieves the desirable performance both in diagnosis indices and feature visualization.

In domain adaptation methods, the DTN achieves superior performance than shallow domain adaptation methods, such as TJM, TCA and JDA. The shallow methods require manual feature extraction, which may suffer from the interference of redundant and irrelevant features. And more importantly, this process is not flexible and not able to meet the need of adaptivity. DTN establishes the domain adaptation in deep learning flow, and is capable of adaptively learning intrinsic fault characteristics. It is also worth noting that the complexity of the domain adaptation process always changes with the scenarios. In the easy transfer tasks, e.g. C→D and D→C, all these transfer learning based techniques get the relatively satisfactory results. However, in several hard tasks, e.g. E→F and J→I, where the source and target data could be substantially dissimilar, the performance drop in the comparative transfer methods, such as DTN with MDA, convincingly illustrates that the difficulties of domain adaptation will accordingly increase. The comprehensive assessments under diverse transfer scenarios further demonstrate the pivotal role of JDA in DTN.

This work proposes a novel diagnosis framework for considering the deep feature learning and cross-domain feature distribution alignment simultaneously. It may overcome the shortcomings in existing studies and have a certain significance in the practical diagnosis application. Although the effectiveness of proposed DTN with JDA has been demonstrated from the aspects of diagnosis indices, feature visualization and loss convergence in ten experimental tasks, it still has limits in assumed conditions, where the faults occur both in source and target domains, and the fault labels are also the same. However, the monitoring data of industrial process is mostly under health conditions, and the occurred fault types may differ from the known ones in source domain. As a result, these factors introduce additional difficulties into the application of transfer diagnosis framework. The integration of data cleaning and selection techniques into this framework has an important significance.

7. Conclusion

Intelligent fault diagnosis in real industrial applications is suffering the difficulty of model re-training as of the discrepancy between the source domain (where the model is learnt) and the target domain (where the model is applied). However, re-training the model is challenging and probably unrealistic as of the lack of sufficient labeled data in practical applications. To address this issue, this work presents a DTN to take advantage of a pre-trained network from the source domain and get the model transferred with unlabeled data from the target domain, where a novel domain adaptation approach, JDA, is presented. Through extensive experiments on three datasets, the results show that the DTN with JDA outperforms the state-of-the-art approaches. Compared with the shallow methods, i.e., SVM, RF, EMD, TJM, TCA and JDA, DTN with JDA achieves 25.2%, 24.3%, 31.4% 15.4%, 16.7%, 15.8% improvements on the average accuracy in ten diagnosis tasks. In deep learning framework, DTN also effectively increase the diagnosis accuracy from 83.0%, 87.2% to 97.9% in comparison with basic CNN and DTN with MDA. The network visualization further provides the interpretation of diagnosis results, and DTN with JDA is shown to obtain a more accurate feature distribution alignment across domains. Moreover, the DTN with JDA presents smooth convergence and avoids negative adaptation in comparison with MDA.

Using DTN with JDA, it is promising that the learnt diagnosis models from experimental or real datasets can be transferred to new but similar applications in a more efficient and accurate way, which could benefit kinds of industrial applications. Further work will pursue (i) quantitative assessment approaches of similarity and transferability between diverse domains, (ii) application on imbalanced distribution of machine conditions and (iii) hyper-parameter selection with intelligent optimization algorithms [52, 53].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No. 11572167 and 11802152). The authors would like to express their sincere gratitude to Mr. Shaohua Li for his contributions on the acquisition of experimental data.

References

[1] Lei Y, He Z, Zi Y. Eemd method and wnn for fault diagnosis of locomotive roller bearings. Expert Syst Appl 2011;38(6):7334–41.
[2] Jiang D, Liu C. Machine condition classification using deterioration feature extraction and anomaly determination. IEEE Trans Reliab 2011;60(1):41–8.
[3] Cui L, Huang J, Hao Z, Zhang F. Research on the meshing stiffness and vibration response of fault gears under an angle-changing crack based on the universal equation of gear profile. Mech Mach Theory 2016;105:554–67.
[4] Gong X, Qiao W. Current-based mechanical fault detection for direct-drive wind turbines via synchronous sampling and impulse detection. IEEE Trans Ind Electron 2015;62(3):1693–702.
[5] Yunusa-Kaltungo A, Sinha JK, Nembhard AD. A novel fault diagnosis technique for enhancing maintenance and reliability of rotating machines. Struct Health Monit 2015;14(6):231–62.
[6] Cui L, Huang J, Zhang F, Chu F. Hvsrms localization formula and localization law: Localization diagnosis of a ball bearing outer ring fault. Mech Syst Signal Process 2019;120:608–29.
[7] Shen Z, Chen X, Zhang X, He Z. A novel intelligent gear fault diagnosis model based on emd and multi-class tsvm. Measurement 2012;45(1):30–40.
[8] Li Y, Wang X, Liu Z, Liang X, Si S. The entropy algorithm and its variants in the fault diagnosis of rotating machinery: A review. IEEE Access 2018;6:66723–41.
[9] Li Y, Wang X, Si S, Huang S. Entropy based fault classification using the Case western reserve university data: A benchmark study. IEEE Trans Reliab 2019. DOI: 10.1109/TR.2019.2896240.

[10] Verstraete D, Ferrada A, Droguett EL, Meruane V, Modarres M. Deep learning enabled fault diagnosis using time-frequency image analysis of rolling element bearings. Shock Vib 2017;2017:1–17.
[11] Feng J, Lei Y, Guo L, Lin J, Xing S. A neural network constructed by deep learning technique and its application to intelligent fault diagnosis of machines. Neurocomputing 2018;272:619–28.
[12] Jia F, Lei Y, Lin J, Zhou X, Lu N. Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. Mech Syst Signal Process 2016;72–73:303–15.
[13] Han T, Liu C, Yang W, Jiang D. A novel adversarial learning framework in deep convolutional neural network for intelligent diagnosis of mechanical faults. Knowl-Based Syst 2019;165:474–87.
[14] Wen L, Li X, Gao L, Zhang Y. A new convolutional neural network based data-driven fault diagnosis method. IEEE Trans Ind Electron 2017;65(7):5990–8.
[15] Liu R, Yang B, Zio E, Chen X. Artificial intelligence for fault diagnosis of rotating machinery: A review. Mech Syst Signal Process 2018;108:33–47.
[16] Cerrada M, Sánchez RV, Li C, Pacheco F, Cabrera D, Oliveira JVD, Vásquez RE. A review on data-driven fault severity assessment in rolling bearings. Mech Syst Signal Process 2018;99:169–96.
[17] Oquab M, Bottou L, Laptev I, Sivic J. Learning and transferring mid-level image representations using convolutional neural networks. In: IEEE conference on computer vision and pattern recognition. 2014, p. 1717–24.
[18] Mun S, Shin M, Shon S, Kim W, Han DK, Ko H. DNN Transfer learning based non-linear feature extraction for acoustic event classification. IEICE Trans Inf Syst 2017;100(9).
[19] Qureshi AS, Khan A, Zameer A, Usman A. Wind power prediction using deep neural network based meta regression and transfer learning. Appl Soft Comput 2017;58:742–55.
[20] Khatami A, Babaie M, Tizhoosh HR, Khosravi A, Nguyen T, Nahavandi S. A sequential search-space shrinking using CNN transfer learning and a radon projection pool for medical image retrieval. Expert Syst Appl 2018;100:224–33.
[21] Han T, Liu C, Yang W, Jiang D. Learning transferable features in deep convolutional neural networks for diagnosing unseen machine conditions. ISA Trans 2019. http://dx.doi.org/10.1016/j.isatra.2019.03.017.
[22] Wei Y, Zhang Y, Yang Q. Learning to transfer, Eprint Arxiv (2017).
[23] Liu C, Jiang D, Yang W. Global geometric similarity scheme for feature selection in fault diagnosis. Expert Syst Appl 2014;41(8):3585–95.
[24] Zhao C, Feng Z, Wei X, Qin Y. Sparse classification based on dictionary learning for planet bearing fault identification. Expert Syst Appl 2018;108:233–45.
[25] Lei Y, Jia F, Lin J, Xing S, Ding SX. An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data. IEEE Trans Ind Electron 2016;63(5):3137–47.
[26] Han T, Jiang D, Zhao Q, Wang L, Yin K. Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery. Trans Inst Meas Control 2018;40(8):2681–93.
[27] Costilla-Reyes O, Scully P, Ozanyan KB. Deep neural networks for learning spatio-temporal features from tomography sensors. IEEE Trans Ind Electron 2018;65(1):645–53.
[28] Han T, Liu C, Yang W, Jiang D. An adaptive spatiotemporal feature learning approach for fault diagnosis in complex systems. Mech Syst Signal Process 2019;117:170–87.
[29] Jiao J, Zhao M, Lin J, Zhao J. A multivariate encoder information based convolutional neural network for intelligent fault diagnosis of planetary gearboxes. Knowl-Based Syst 2018.
[30] Lu C, Wang ZY, Qin WL, Ma J. Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification. Signal Process 2017;130(C):377–88.
[31] Shao H, Jiang H, Wang F, Wang Y. Rolling bearing fault diagnosis using adaptive deep belief network with dual-tree complex wavelet packet. ISA Trans 2017;187–201.
[32] Jing L, Zhao M, Li P, Xu X. A convolutional neural network based feature learning and fault diagnosis method for the condition monitoring of gearbox. Measurement 2017;111:1–10.
[33] Zhang W, Peng G, Li C, Chen Y, Zhang Z. A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. Sensors 2017;17(2):425.
[34] Liu R, Meng G, Yang B, Sun C, Chen X. Dislocated time series convolutional neural architecture: An intelligent fault diagnosis approach for electric machine. IEEE Trans Ind Inf 2017;13(3):1310–20.
[35] Sun W, Zhao R, Yan R, Shao S, Chen X. Convolutional discriminative feature learning for induction motor fault diagnosis. IEEE Trans Ind Inf 2017;13(3):1350–9.
[36] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2010;22(10):1345–59.
[37] Weiss K, Khoshgoftaar TM, Wang DD. A survey of transfer learning. J Big Data 2016;3(1):9.
[38] Long M, Wang J, Ding G, Sun J, Yu PS. Transfer joint matching for unsupervised domain adaptation. In: IEEE conference on computer vision and pattern recognition. 2014, p. 1410–7.
[39] Pan SJ, Tsang IW, Kwok JT, Yang Q. Domain adaptation via transfer component analysis. IEEE Trans Neural Netw 2011;22(2):199.
[40] Long M, Wang J, Ding G, Sun J, Yu PS. Transfer feature learning with joint distribution adaptation. In: IEEE international conference on computer vision. 2014, p. 2200–7.
[41] Long M, Cao Y, Wang J, Jordan MI. Learning transferable features with deep adaptation networks, Eprint Arxiv (2015) 97–105.
[42] Long M, Zhu H, Wang J, Jordan MI. Deep transfer learning with joint adaptation networks, Eprint Arxiv (2016).
[43] Ghifary M, Kleijn WB, Zhang M, Balduzzi D, Li W. Deep reconstruction-classification networks for unsupervised domain adaptation. In: European conference on computer vision. 2016, p. 597–613.
[44] Wen L, Gao L, Li X, Wen L, Gao L, Li X. A new deep transfer learning based on sparse auto-encoder for fault diagnosis. IEEE Trans Syst Man Cybern 2017;1–9.
[45] Lu W, Liang B, Cheng Y, Meng D, Yang J, Zhang T. Deep model based domain adaptation for fault diagnosis. IEEE Trans Ind Electron 2017;64(3):2296–305.
[46] Xu Zhang S-FCSW, Yu FX. Deep transfer network: Unsupervised domain adaptation, Eprint Arxiv, arXiv:1503.00591 (2015).
[47] Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks?, Eprint Arxiv 27 (2014) 3320–3328.
[48] Center BD. Case western reserve university bearing data, http://csegroups.case.edu/bearingdatacenter/pages/download-data-file, 2013.
[49] Rauber TW, Boldt FDA, Varejao FM. Heterogeneous feature models and feature selection applied to bearing fault diagnosis. IEEE Trans Ind Electron 2015;62(1):637–46.
[50] Xu J, Wang J, Izadi I, Chen T. Performance assessment and design for univariate alarm systems based on far, mar, and AAD. IEEE Trans Autom Sci Eng 2012;9(2):296–307.
[51] Wang X, Hou Z, Yu W, Jin Z. Online fast deep learning tracker based on deep sparse neural networks. In: International conference on image and graphics. Springer; 2017, p. 186–98.
[52] Patwal RS, Narang N, Garg H. A novel TVAC-PSO based mutation strategies algorithm for generation scheduling of pumped storage hydrothermal system incorporating solar units. Energy 2018;142:822–37.
[53] Garg H. A hybrid GSA-GA algorithm for constrained optimization problems. Inform Sci 2019;478:499–523.