ISA Transactions 93 (2019) 341–353

Contents lists available at ScienceDirect

ISA Transactions

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

Practice article

# Learning transferable features in deep convolutional neural networks for diagnosing unseen machine conditions

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

## HIGHLIGHTS

- A deep transfer learning framework for diagnosing unseen faults in target applications.
- Exploring the feature transferability in disparate levels of the pre-trained model.
- Enables to build and train a large network efficiently with limited data for new tasks.

## ARTICLE INFO

*Article history:*
Received 5 October 2018
Received in revised form 17 March 2019
Accepted 18 March 2019
Available online 25 March 2019

*Keywords:*
Intelligent fault diagnosis
Transfer learning
Convolutional neural networks
Fine tuning
Mechanical systems

## ABSTRACT

Recent years have witnessed increasing popularity and development of deep learning spanning through various fields. Deep networks, and in particular convolutional neural network (CNN) have also achieved many state-of-the-art competition results in the intelligent fault diagnosis of mechanical systems. However, most of the existing studies have been performed with the assumption that the same distribution holds for both the training data and the test data, which is not in accord with situations in real diagnosis tasks. To tackle this problem, a transfer learning framework based on pre-trained CNN, which leverages the knowledge learned from the training data to facilitate diagnosing a new but similar task, is presented in this work. First, the CNN is trained on large datasets to learn the hierarchical features from the raw data. Then, the architecture and weights of the pre-trained CNN are transferred to new tasks with proper fine-tuning instead of training a network from scratch. To adapt the pre-trained CNN in a specific case, three transfer learning strategies are discussed and compared to investigate the applicability as well as the significance of feature transferability from the different levels of a deep structure. The case studies show that the proposed framework can transfer the features of the pre-trained CNN to boost the diagnosis performance on unseen machine conditions in terms of diverse working conditions and fault types.



## 1. Introduction

The condition-based maintenance (CBM) strategy has received great attention in modern industries, with its prominent advantages of availability, sustainability, safety, and reliability. Unlike breakdown maintenance and time-based maintenance, CBM aims to schedule the maintenance time and planning depending on the actual running state of the equipment. As the process of making a decision on machine conditions is based on appropriate analyses of the monitoring data, diagnostics are the core blocks of CBM [1]. Essentially, machine diagnostics can be formulated as a pattern classification and recognition problem [2].

Little over a decade ago, artificial intelligence techniques brought about the idea that the machine health conditions can be efficiently assessed by a well-trained classifier instead of a diagnosis specialist, which is commonly known as intelligent fault diagnosis [3,4]. Backed by an automatic operating procedure, high-efficiency data processing capability as well as relatively satisfactory diagnosis accuracy, this strategy led to a series of works dealing either with feature extraction and selection, or with the design and optimization of classification algorithms, in the fields of diagnostics for mechanical equipment [5–7]. With reference to the recent literature, most of the research about intelligent fault diagnosis can be categorized into two groups: manual feature extraction methods and deep learning methods (as shown in Fig. 1(a)). The general steps, including the data preprocessing, feature extraction and selection and pattern recognition, are needed in the manual feature extraction methods. The influential

* 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.2019.03.017
0019-0578/

algorithms in data mining cover noise filtering, dimensionality reduction, instance reduction, imbalanced data preprocessing [8–10], have been widely studied in fault diagnosis problem. However, for manual feature extraction methods, relevant studies have revealed two inherent weaknesses. (1) The proper designs of feature extraction and selection approaches largely rely on prior knowledge about the objects of analysis. (2) The prevailing pattern recognition algorithms such as artificial neural network (ANN) and support vector machine (SVM) for automatically identifying the machine health condition, are shallow learning architectures, indicating that they are unable to approximate the highly complex non-linear functions [11–13]. Tackling some of the weaknesses of the traditional framework, the second deep learning methods have enjoyed increasing popularity for a wide range of problems, such as image processing, computer vision, video processing, natural language processing, etc., in recent years [14]. One of the great advantages of this framework is that feature learning is automatic. Based on the labeled data and back-propagation, the deep model methods can initiatively capture the essential features for diagnosing machine conditions, rather than depending on the classification capacity of handcrafted features or pattern recognition algorithms. The deep neural network, especially convolutional neural networks (CNNs), has also led to many state-of-the-art competitive results in the fault diagnosis of diverse mechanical objects, such as spindle bearings, planetary gearboxes and rotor systems [15–24].

Although the marvelous success of deep learning in a variety of diagnosis applications has often been reported, the topic is still largely open. First, these works are mostly under the assumption of the same distribution between the training data and the test data. The diagnostic ability will evidently degenerate when the training and testing data have different feature distribution. Second, larger amounts of data are often required to train the deep learning models, while typically labeled fault samples are generally scarce in actual industrial tasks. Consequently, it would be highly desirable to develop methods that can leverage knowledge from pre-existing tasks (source domain) to facilitate model training and diagnosis in a unseen machine diagnostic problem (target domain) that is similar but not same as the existing task (namely transfer learning), as shown in Fig. 1(b). This would allow the knowledge to be adequately utilized across different tasks, so that new diagnosis applications can be done more flexible, and more importantly, to improve the learning ability with limited number of samples.

Currently, transfer learning based algorithms and frameworks have been widely studied in different research areas, such as image classification [25,26], text classification [27], acoustic event recognition [28] and biometrics [29]. According to the latest surveys in this community [30,31], transfer learning techniques can be mainly categorized into three classes: inductive transfer learning, transductive transfer learning and unsupervised transfer learning. In this work, we focus on the inductive transfer learning for fault diagnosis tasks where sufficiently labeled data in the source domain are available for model pre-training and some labeled data in the target domain are employed to induce a transferred model for use in target tasks. Combined with the rise of deep learning, the signal characteristics can be adaptively and hierarchical compiled as the hierarchical weights of the deep network. Instead of training a deep neural network from scratch, the network parameters can be easily recaptured in other new tasks for feature and knowledge transfer. In the field of object recognition, Oquab et al. [32] designed a deep CNN based transfer learning method for the reuse of the convolutional layers learned on large-scale annotated datasets (ImageNet) to compute the mid-level image representation for target tasks. To reduce the effects of different image statistics (types of objects, viewpoints) across domains, new adaptation layers (two fully connected layers) were added to replace the output layers of pre-trained networks, and trained with a limited amount of target data. The competition results on two image datasets show the high feature transferability of ImageNet-trained CNN for object and action recognition tasks. Also utilizing CNN, the paper by Yosinki et al. [33] further investigated how well the features of each layer transfer from the source domain to the target domain in image classification problems. This research reveals two issues, i.e., optimization difficulties when separating certain layers from the whole network without considering the fragile co-adapted features on successive layers, and the specialization of higher layer features to the source domain, which may increase the difficulties in bridging the gaps between the source and target tasks.

However, studies about transfer learning with the pre-trained deep network in fault diagnosis cases are few. Zhang et al. trained a shallow ANN with enough source data from the bearing data center of Case Western Reserve University (CWRU) for bearing fault diagnosis tasks, and then they transferred the parameters and modified the structure into new but similar tasks with a small amount of target data under different working conditions [34]. The other scenarios, such as diverse fault types across domains, will also lead to a large domain discrepancy. Investigations of transfer fault diagnosis under more scenarios and fault datasets are crucial and necessary. Besides, the transferability of the features in the deep structure is not well discussed with mechanical fault data, and the proper transfer strategy with respect to fault data quality is not clear. Motived by this, this work presents three strategies, namely fine-tuning of the fully connected layers, fine-tuning of the whole pre-trained model, and fine-tuning of the feature descriptor layers, to explore feature transferability in diagnosis applications. With two diagnosis datasets (i.e. an open-access gearbox fault dataset [35], and our single-stage cylindrical straight gearbox fault dataset), the performance of the presented transfer learning strategies is investigated with the assumption that limited samples are available in the target task, where two scenarios i.e., diverse working conditions and diverse fault types, are considered. The main contributions of this work are summarized as follows:

(1) A novel deep transfer learning framework is presented on the basis of the pre-trained CNN, consisting of three strategies: (i) transferring the feature descriptor (convolutional layers) and fine tuning the classifier (fully-connected layers), (ii) transferring and fine-tuning both the feature descriptor and the classifier, and (iii) transferring the classifier and fine-tuning the feature descriptor.

(2) Through exploring the feature transferability in disparate levels of the pre-trained model, the questions of 'how to transfer' and 'what to transfer' [36] are answered for actual industrial diagnosis applications, where approaches that can handle new but similar scenarios (target tasks) with no need for large amounts of new labeled data and intense expert knowledge, are highly appreciated.

(3) The presented framework is successfully applied in two different scenarios, (i) feature transfer between varying working conditions to properly identify the fault, and (ii) feature transfer between diverse fault types to correctly isolate the new fault by its location.

The remaining part of this paper is organized as follows. In Section 2, the research backgrounds are briefly described. In Section 3, we present the proposed transfer learning strategies in the mechanical diagnosis application. Section 4 presents two case studies, and the base systems. The results and a thorough discussion are made in Sections 5 and 6, respectively. Finally, the conclusions are drawn in Section 7.

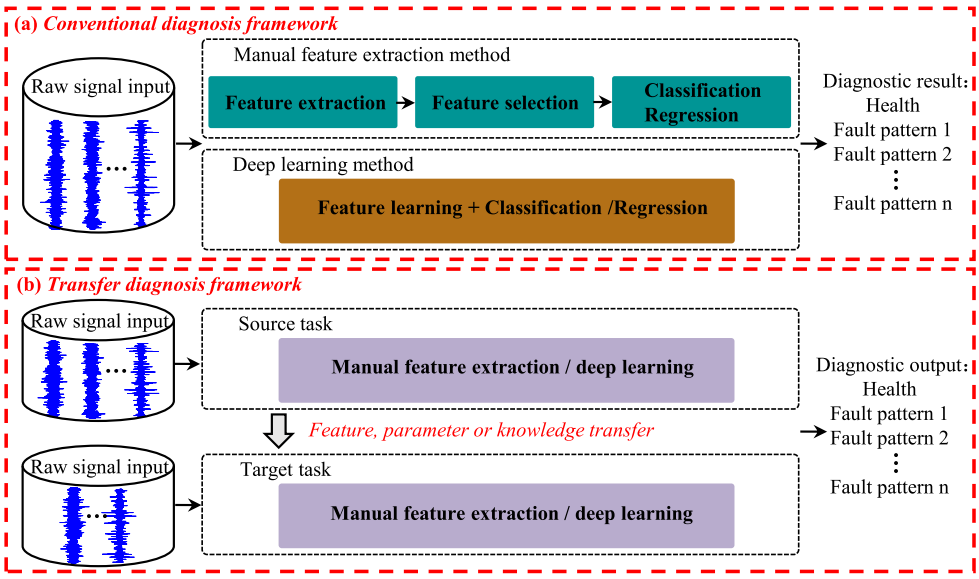


**Fig. 1.** Intelligent fault diagnosis framework. (a) Conventional diagnosis framework and (b) transfer diagnosis framework.

## 2. Related works

### 2.1. End-to-end diagnosis framework via CNN

The main advantage of deep networks is the automatic feature learning ability, which is capable of adaptively capturing and extracting fault-sensitive features from the raw signal, providing an end-to-end framework for mechanical fault diagnosis. This work focuses on a type of widely used deep network, i.e., CNN. A typical CNN generally consists of a multiple combinations of one or more convolutional layers and pooling layers, followed by one or more fully connected layers.

**Feature extraction:** The convolutional layer convolves the small filter kernels with the larger input signal to obtain a different feature activation value at each location, named as a feature map. The feature maps for all of the filters along the depth of this layer are further processed by the activation unit, and they are composed as the inputs of higher layers. The operation in a convolutional layer is summarized as:

$$o_j^l = ReLU(\sum_i o_i^{l-1} * k_{ij}^l + b_j^l) \tag{1}$$

where $o_j^l$ is the $j$th output at the $l$th layer, $o_i^{l-1}$ denotes the $i$th input at the $(l-1)$th layer, $k_{ij}^l$ is the convolutional kernel between the $i$th input and the $j$th output, $b_j^l$ is the basis, $*$ represents the convolution operation and ReLU denotes the nonlinear activation function, i.e., a rectified linear unit.

After convolution, the obtained feature activation may retain a high dimension, and this will be computationally challenging when directly applying them for classification tasks. Hence, a pooling layer, which is also called sub-sampling layer, often follows each convolutional layer to aggregate the statistics of the convolved features at various locations. A max-pooling process is described as:

$$o_j^{l+1}(i) = \max_{(i-1)W+1 \le t \le iW} o_j^l(t) \tag{2}$$

where $o_j^l(t)$ is the feature map of the $j$th neuron at the $l$th layer, $W$ is the width of a local region for pooling, and $o_j^{l+1}(i)$ denotes the output after pooling operation.

**Classification:** After a series of convolutions and pooling, several fully connected layers, mainly including one or more hidden layers and a classification layer, are added to classify the abstract high-level features. Since each neuron in one layer connects to all of the neurons in another layer, these types of layers are identical to the ones in traditional multilayer neural network. A softmax regression is generally appended to the classification layer. The output of the softmax regression is defined as:

$$\begin{aligned} h_\theta(x^{(i)}) &= \begin{bmatrix} p(y^{(i)}=1|x^{(i)};\theta) \\ p(y^{(i)}=2|x^{(i)};\theta) \\ \cdots \\ p(y^{(i)}=c|x^{(i)};\theta) \end{bmatrix} \\ &= \frac{1}{\sum_{j=1}^k \exp(\theta_j^T x^{(i)})} \begin{bmatrix} \exp(\theta_1^T x^{(i)}) \\ \exp(\theta_2^T x^{(i)}) \\ \cdots \\ \exp(\theta_k^T x^{(i)}) \end{bmatrix} \end{aligned} \tag{3}$$

where the predicted label $y^{(i)} \in \{1, 2, \ldots, c\}$, $c$ is the number of categories and $\theta$ is the collection of the parameters of model. With the predefined loss function, the predicted error can be quantitatively evaluated and back propagated to optimize the parameters of network.

In the diagnostic application of CNN, Verstraete et al. [20] explored different time–frequency analysis methods to form the original time–frequency images, which were then fed into CNN for bearing condition classification and diagnosis. Liu et al. [21] developed a dislocated layer to intercept signal segments and to generate a 2D matrix for feature learning in CNN. Ding et al. [22] presented a feature learning method using multilayers of CNN from wavelet packet energy images for bearing fault diagnosis. In [23], the authors utilized 1D-CNN to fuse the multi-sensor signal into data-levels for gearbox condition classification, and they achieved a superior performance than manually selected features. In [24], the CNN was shown to be able to automatically learn features with wide kernels in the first convolutional layer from the raw 1D mechanical vibration signal. Although encouraging results have been achieved, and the research on intelligent fault diagnosis is facilitated by the deep algorithms, these works fail to fully consider the domain difference problem. For instance, the CWRU-bearing fault dataset has been widely used and investigated in many studies, and the training set and testing set are drawn from the same working load or fault diameter. Due to the complex mechanical structure, the changeable working conditions and the varying degrees of ambient noise, the machine

signals generally present various manifestations and keep domain independent. The model trained with the source-domain data may perform poorly in the target domain. Consequently, most of the recent deep methods for intelligent fault diagnosis have their limitations and more general methods based on the transfer learning framework are needed.

### 2.2. Transfer learning

In transfer learning, a domain $\mathcal{D}$ is composed of two parts: a feature space $\mathcal{X}$, and a marginal probability distribution $P(X)$, where $X = \{x_1, \ldots, x_n\} \in \mathcal{X}$. Taking the fault diagnosis as an example, each machine condition contains many samples, $X$ is a particular training set, $x_i$ means the $i$th feature vector or instance corresponding to a certain machine condition, and $\mathcal{X}$ is the whole feature space of all vectors.

For a specific domain, $\mathcal{D} = \{\mathcal{X}, P(X)\}$, a task $\mathcal{T}$ consists of two parts, a label space $\mathcal{Y}$ and a predictive function $f(\cdot)$, which can be learned from the instance set $\{x_i, y_i\}$, where $x_i \in X$ and $y_i \in \mathcal{Y}$. Referring to the fault diagnosis, $\mathcal{Y}$ is a set of the labels of each machine condition, $y_i$ takes on a value such as 0,1, …to represent different condition categories, and $f(x)$ is the learned model that can predict the corresponding label of a new machine sample.

Now, we have both defined the domain $\mathcal{D} = \{\mathcal{X}, P(X)\}$ and task $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$. A source domain data is denoted as $\mathcal{D}_s = \{(x_1^s, y_1^s), \ldots, (x_{n_s}^s, y_{n_s}^s)\}$, where $x_i^s \in \mathcal{X}_s$ is a data sample of $\mathcal{D}_s$ and $y_i^s \in \mathcal{Y}_s$ is the corresponding category label of $x_i^s$. Similarly, a target domain dataset is denoted as $\mathcal{D}_t = \{(x_1^t, y_1^t), \ldots, (x_{n_t}^t, y_{n_t}^t)\}$, where $x_i^t \in \mathcal{X}_t$ is a data sample of $\mathcal{D}_t$ and $y_i^t \in \mathcal{Y}_t$ is the corresponding category label of $x_i^t$. Generally, $0 \le n_t \ll n_s$. Then, a formal definition of transfer learning can be presented. 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 $f_t(\cdot)$ 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 $\mathcal{D}_s = \mathcal{D}_t$ and $\mathcal{T}_s = \mathcal{T}_t$, the learning process can be identified with the traditional machine learning problem.

In the recent literature, Pan and Yang made a comprehensive survey on the current progress of transfer learning, and they present a detailed categorization of transfer learning techniques [30]. Weiss et al. further discussed different transfer learning scenarios using many real-world applications [31]. The main transfer learning approaches can broadly be categorized into domain adaptation (DA) methods, which adapt the model learned on the source-domain labeled data to a specific target domain with unlabeled target data, and inductive approaches, which utilize a few labeled target datasets to fine-tune the pre-trained models that were trained on a large source dataset. For DA methods, either maximum mean discrepancy (MMD) minimization or adversarial training are studied to align the source-domain distributions to the target domain distribution in many fields. The mechanical fault diagnosis cases with DA methods have also been reported recently. Li et al. developed a multi-layer domain adaptation method with the multi-kernel metric for rolling bearing fault diagnosis under diverse working conditions [37]. Guo et al. integrated the binary cross entropy loss-based domain recognition errors maximization and MMD-based distribution distance minimization for domain adaptation between different machine fault datasets [38]. Zhang et al. developed adversarial adaptive 1D convolutional neural networks with an adversarial training framework for bearing fault diagnosis [39]. In some fault diagnosis applications, it is often the case that there are only a few typical fault samples rather than plenty of unlabeled samples in the target domain, especially in the state that more and more manufacturers are collecting typical operating data during factory tests, which probably include several fault scenarios aided by development of fault injection techniques. In this scenario, the reuse of pre-trained models with different fine-tuning strategies is the site of this investigation. To the best of our knowledge, research in inductive transfer learning approaches are limited in the field of machine fault diagnosis. Specifically, the applicability of diverse transfer learning strategies is not well discussed and compared as well as the significance of feature transferability from different levels of a deep structure (e.g., different layers of a CNN). Motived by existing work utilizing the pre-trained CNN, we employed the pre-trained CNN as a starting point, and we explored effective transfer learning strategies for solving practical diagnosis challenges.

## 3. A transfer learning framework for machine fault diagnosis

Essentially, a deep CNN model can be divided into a total of two parts. The initial convolutional and max-pooling blocks are used to learn the signal feature, and to appropriately represent the input signal, which can be treated as a feature descriptor (purple region in the left part in Fig. 2). Additionally, the subsequent fully-connected layers are trained to make a decision for a supervised classification problem, which can be seen as a classifier (buff region in the right section in Fig. 2). The transfer learning based framework for machine fault diagnosis is illustrated in Fig. 2. First, sufficient fault data in the source domain are applied to train a base network, which is also referred as a pre-trained CNN. The intrinsic features for different signal patterns can be learned and compiled as the weights of the network. Given a specific network architecture and large-scale source samples $\{x_i^s, y_i^s\}_{i=1}^{n_s}$, the optimization objective in the offline pre-training step can be formulated as:

$$arg \min_{\boldsymbol{\theta}} \sum_i \ell(y_i, f(x_i, \boldsymbol{\theta})) \tag{4}$$

where $\ell$ denotes the loss function to estimate the cost between the true label $y_i$ and the real predicted label by the model $f(x_i, \boldsymbol{\theta})$, $\boldsymbol{\theta} = \{\boldsymbol{\theta}^{(l)}\}_{l=1}^{L}$ is the collection of network parameters of the feature descriptor and classifier, respectively, and $\boldsymbol{\theta}^{(l)}$ denotes the weights and the basis of layer $l$. The back-propagation (BP) and stochastic gradient descent (SGD) are used to optimize (4):

$$\boldsymbol{\theta} = \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} E[\ell(\boldsymbol{\theta})] \tag{5}$$

where $\alpha$ is the learning rate.

Then, for a new but similar diagnosis task, the parameters of the pre-trained CNN are copied to a target network. When the labels are different between the source and target domain, a general method is to remove the classification layer, and to randomly initialize the parameters of the layer, as shown in Fig. 2. In order to learn the transferable features from the pre-trained CNN, a fine tuning step, which back-propagates the errors of the target samples $\{x_i^t, y_i^t\}_{i=1}^{n_t}$, will be executed to adapt the network to the target domain. In a real diagnosis problem, collecting and labeling enough target fault samples is a labor-intensive and costly task. Therefore, we are assuming that the samples of the target domain are significantly less than those of the source domain. Also, this brings in the risk of overfitting with limited training samples during the fine-tuning step. To overcome this problem, reducing the number of parameters to be fine-tuned is usually adopted by fixing the transferred parameters of a number of layers (these layers are also known as the frozen layers), while training the other layers toward the target task. Supposing that layer $l$ needs to be frozen, the original optimization of SGD in layer $l$ is presented in (6):

$$\boldsymbol{\theta}^{(l)} = \boldsymbol{\theta}^{(l)} - \alpha^{(l)} \frac{\partial E(\ell(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}^{(l)}} \tag{6}$$

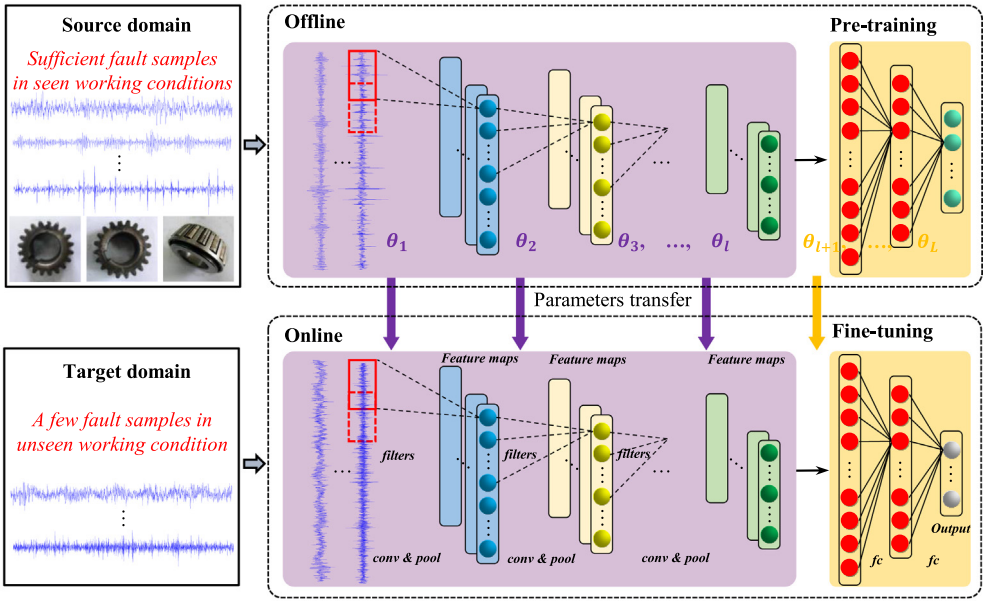


**Fig. 2.** The framework of transfer learning using a pre-trained CNN.

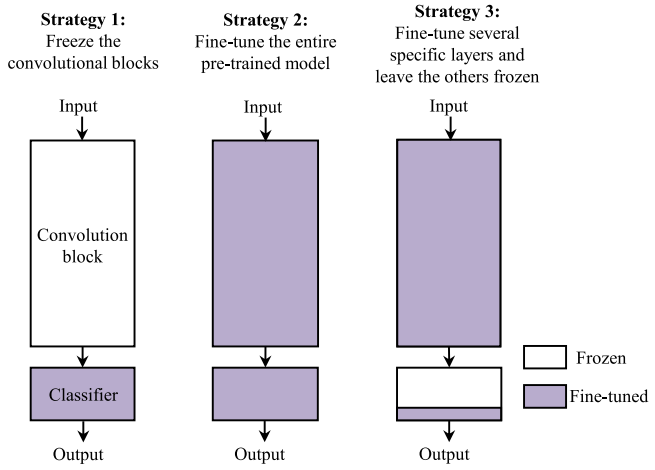


**Fig. 3.** Schematic explanation of three transfer learning strategies.

and the learning rate $\alpha^{(l)}$ in layer $l$ is set to zero, i.e., $\alpha^{(l)} = 0$. In this way, the parameters in layer $l$ can remain unchanged during optimization, and are namely frozen layers.

Since how to proceed on using the pre-trained CNN differs radically in different fields, it is significant to explore which layers need to be fine-tuned, and whose features are transferable in machine fault diagnosis problems. For this reason, we mainly present three different transfer strategies, and we further discuss the merits and demerits of each transfer strategy in two case studies. The detailed descriptions of the three schemes are given as follows, and the schematic presentation is illustrated in Fig. 3:

Strategy 1: Following the practices of image recognition, the pre-trained CNN can be treated as a feature extraction mechanism, which means that we can directly transfer the feature descriptor and frozen weights in the target domain. Only the classifier, which refers to specialized features, needs to be customized or adjusted with target data. We can remove the classifier of the pre-trained CNN and totally train a new one from scratch, or we can transfer the weights of the pre-trained classifier and adapt it to target tasks via fine tuning.

Strategy 2: A supervised fine tuning process is implemented after the unsupervised pretraining to improve the performance in most of the deep models. Similarly, a proper way for transfer learning is to retain the architecture and weights of the pre-trained CNN, and to transfer the model to the target domain. Then, we can retrain the whole network with the initial weights of the pre-trained CNN via fine-tuning.

Strategy 3: In image recognition, the CNN can be trained on many standard and large database [40], such as ImageNet, Pascal Voc and COCO. This process can facilitate the feature descriptor of pre-trained CNN in capturing some universal features such as curves and edges, that are also relevant to other image recognition tasks. However, the inside mechanism of feature learning in deep CNN for the mechanical signal is still not clear, and the front convolutional blocks may not refer to the general features. Moreover, since different mechanical structure, working conditions or fault types all can generate different signal patterns, it may be difficult to guarantee for certain that the learned feature representation with a specific source dataset can be certainly applicable to new machine fault data. Consequently, this strategy focuses on fine-tuning the feature descriptor part and the top classification layer.

Generally, two factors, i.e., the dataset size and similarity, guide the selection of transfer strategy. When the target dataset is large and similar to the pre-trained CNN, of course, we can fine-tune the entire model, or we can train some layers and leave others frozen. When the target dataset is large but different from the source data, strategy 2, fine-tuning the entire model, is a good choice. On the contrary, for a small but similar target dataset, it will be better to choose strategy 1, which just needs to remove the last classification layer, run the frozen convolutional blocks as a fixed feature extractor, and then use the resulting features to train a new classifier. In the last case, the target dataset is both small and dissimilar to the source dataset. Such a scenario may happen fairly frequently in fault diagnosis tasks. When fine-tuning the whole model, the risk of overfitting increases. If only fine-tuning the shallow end of the pre-trained CNN, the adapted model may fail to learn anything useful. Therefore, retraining the front convolutional blocks and the classification layer may be efficient. A detailed analysis and discussion about the diverse transfer strategies will be presented based on two case studies of mechanical fault diagnosis.

**Table 1**
Descriptions of detailed fault patterns.

| Label | Gear | | | | Bearing | | | | | | Shaft | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 32T | 96T | 48T | 80T | IS:IS | ID:IS | OS:IS | IS:OS | ID:OS | OS:OS | Input | Output |
| 1 | G | G | G | G | G | G | G | G | G | G | G | G |
| 2 | C | G | E | G | G | G | G | G | G | G | G | G |
| 3 | G | G | E | G | G | G | G | G | G | G | G | G |
| 4 | G | G | E | Br | B | G | G | G | G | G | G | G |
| 5 | C | G | E | Br | In | B | O | G | G | G | G | G |
| 6 | G | G | G | Br | In | B | O | G | G | G | Im | G |
| 7 | G | G | G | G | In | G | G | G | G | G | G | Ks |
| 8 | G | G | G | G | G | B | O | G | G | G | Im | G |

IS = input shaft; :IS = input side; ID = idler shaft; OS = output shaft; :OS = output side. G: good; C: chipped; E: eccentric; Br: broken; B: ball; In: inner race; O: outer race; Im: imbalance; Ks: keyway sheared.

## 4. Experimental descriptions

### 4.1. Diagnosis case 1

The gearbox fault dataset from the 2009 challenge data of Prognostics and Health Management (PHM) society is used as the first diagnosis case. This fault dataset is representative of generic industrial gearbox fault data. The details of the gearbox structure, the positions of the apparatuses that are used to collect the data, and an overview of the gearbox are presented in Fig. 4. The experiments were performed on two sets of gears, i.e. spur gears and helical gears, at 30, 35, 40, 45 and 50 Hz shaft speeds, under high and low loadings. As listed in Table 1, eight types of conditions were processed on the gearbox. There were, in total, three shafts, i.e., the input shaft (IS), the idler shaft (ID) and the output shaft (OS). The bearings are installed on the two sides, i.e., input side (:IS) and output side (:OS). In bearing faults, for instance, the symbol of IS:IS means that the input shaft bearing is on the input side. In this study, the second column of the vibration signal (output end) of the spur gearbox under high load was selected for discussion. The sampling frequency was 66.67 kHz, and each sample contained 6144 data points.

In most of the existing studies on the dataset, the training data and testing data are from the same mechanical condition and working condition; that is, the training data and testing data follow the same distribution. To analog the real diagnosis problem, this work focuses on the application of transfer learning to solve the domain discrepancy. We partitioned the five shaft speeds into two domains, whose scenario settings are given in Table 2. The samples under first three shaft speeds (30 Hz, 35 Hz, 40 Hz) are used as source data, while the ones under the 45 Hz and 50 Hz shaft speeds are served as target data.

### 4.2. Diagnosis case 2

The second fault dataset is from our single-stage cylindrical straight gearbox. The schematic diagram of the gearbox test rig and the damaged components are presented in Fig. 5. Since the high-speed stage is more easily subject to failure, most of the faults were introduced to the high-speed cylindrical gearing and conical bearings. Two types of faults, i.e., the tooth broken (TB) and the root crack (RC), were processed on the high-speed gearing, and the other two kinds of faults, i.e., the outer race fault (OF) and the roller fault (RF), were created in the high-speed bearings and low-speed bearings, respectively. The vibration signal of the gearbox was collected from the casing near the high-speed stage. The sampling frequency was 20 kHz, and each sample contains 4096 data points. The experiments were conducted under three working speeds loads (900 rpm, 1200 rpm and 1500 rpm) with a load of 4 N m.

In general, the deep CNN can show excellent diagnosis performance with the help of supervised learning using the samples collected under the known failure types. Unfortunately, it may be hard to guarantee that the fault types occurring are well within the existing scope in practical diagnosis applications. Thus, the diagnosis accuracy was mostly unsatisfactory, because of the different distributions between the training set and the test set. Considering this situation, we defined four gearbox conditions, which were health, gear fault, high-speed bearing fault and low-speed bearing fault in this case study. Each fault condition contained two specific fault types. One fault type was used in the source domain, while the other one served as the target domain. The detailed scenario setting is presented in Table 3. The performances of the three transfer learning strategies between the different fault types will be discussed in this case.

### 4.3. Baseline system

On the basis of deep learning framework, all the experiments employ an end-to-end diagnosis procedure without additional data preprocessing or feature extraction. Considering the impact of architecture settings of CNNs, three 1D-CNNs with different depths were designed, to make a credible study. The detailed architectures and parameters of three CNNs are shown in Fig. 6. Following the design in [24], the size of the first convolutional kernel was set to a large value (64 and 128), with the purpose of restraining the high-frequency noise. The number of kernels increased with the depth of the layers for achieving hierarchical and abstract feature representations. The dropout layers were added to avoid overfitting. Other details were stated as: the activation function was ReLU, the batch size was selected from 16 to 64, the optimizer was the SGD with a learning rate of 0.01, and the experiments were implemented in the PyTorch framework.

## 5. Results

This section conducts case studies to verify effectiveness and efficiency of the investigated techniques mainly from two aspects: the comparison between traditional framework and transfer learning framework considering domain discrepancy problem, and the performance investigation of diverse transfer strategies.

### 5.1. Results in traditional framework

Firstly, the diagnosis experiments using traditional deep learning framework was performed to test the generalization ability of the pre-trained CNN. Sufficient data in the source domain (5000 samples in the PHM 2009 dataset, and 15 000 samples in the gearbox dataset), which followed a uniform distribution among different machine conditions, were employed to train and validate a CNN from scratch. Another 1000 testing samples from the source domain and target domain respectively, were selected for verification. In addition to the above three 1D-CNNs [24], another two 2D-CNN methods in [20,21] were also tested. In [20], the 1D vibration signals were converted to time–frequency images by short-time Fourier transform (STFT), and then used to train a 2D-CNN. Similarly, in [21], 2D matrixes were generated from the vibration signals by a dislocated layer, this method is referred to as DTS-CNN. The architectural settings of the CNNs are the same as the ones in the original literature. The detailed diagnosis results are shown in Tables 4 and 5. It is clear that high accuracies were achieved when the testing samples were from the source domain, while the performance apparently drops when diagnosing the samples in the target domain. For instance, the detailed diagnosis results of CNN3, namely the CNN with 5 convolutional layers for the two domains in the PHM 2009 dataset are shown in

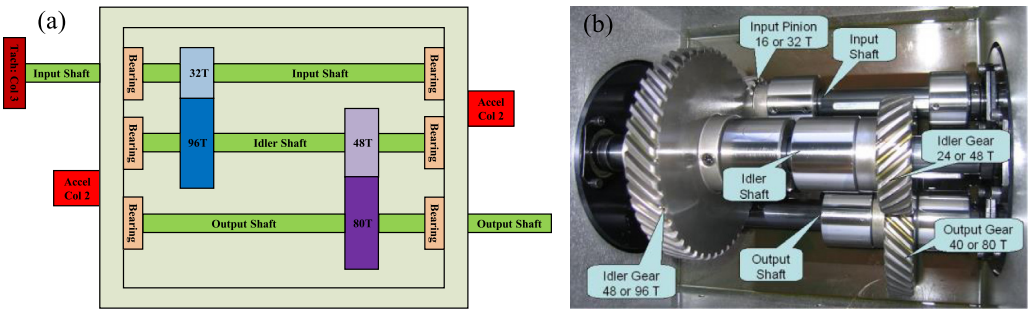


**Fig. 4.** The gearbox in the 2009 Challenge Data of the Prognostics and Health Management (PHM) society: (a) Schematic diagram; (b) Overview of the gearbox.

**Table 2**
Scenario setting of transfer learning in diagnosis case 1.

| Domain types | Source domain | Target domain |
|---|---|---|
| Description | Eight gearbox conditions: G, C, E, Br, B, In, O, Im and Ks, (Corresponding labels 0–7). | |
| Working conditions | Shaft speeds: 1800 rpm, 2100 rpm and 2400 rpm | Shaft speeds: 2700 rpm and 3000 rpm |

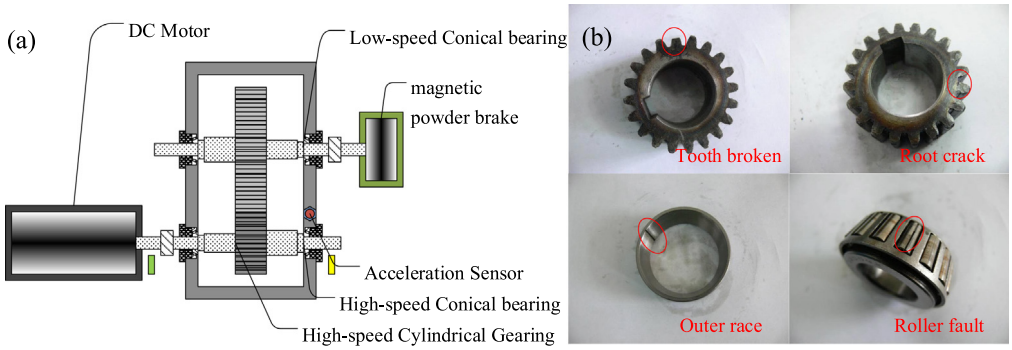


**Fig. 5.** The single-stage cylindrical straight gearbox test rig: (a) Schematic diagram of gearbox test rig; (b) The damaged components.

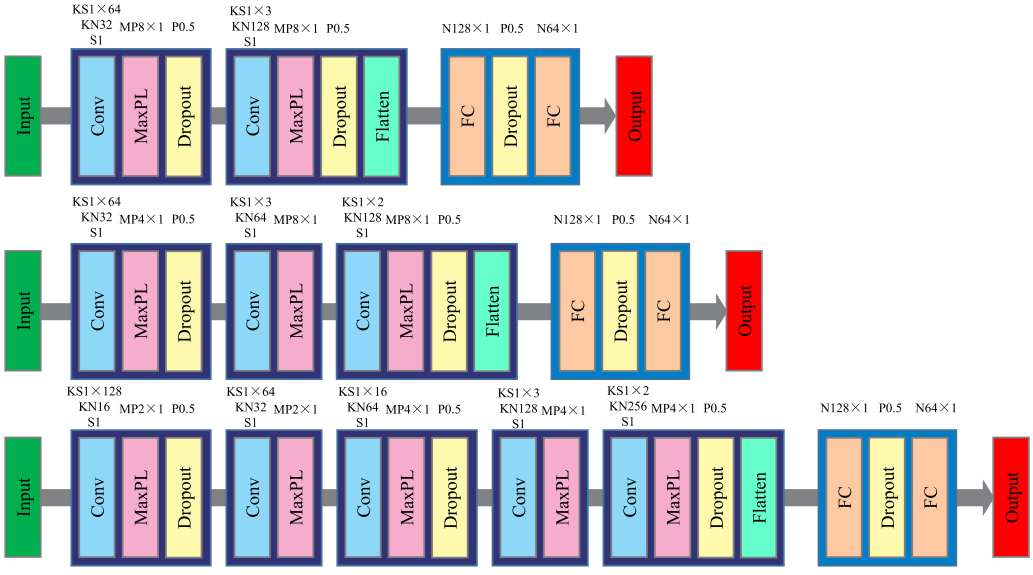


**Fig. 6.** The designed architectures of the three CNNs, CNN1 with 2 convolutional layers, CNN2 with 3 convolutional layers and CNN3 with 5 convolutional layers. Conv: convolutional layer, MaxPL: max-pooling layer, FC: fully-connected layer. In Conv, KS, KN and S represent kernel size, number of kernels and the stride size, respectively. In MaxPL, MP is the pooling size. In Dropout, P donates the percentage. Additionally in FC, N donates the neuro size.

Fig. 7. These results reveal that the pre-trained CNN may not be directly applicable to tasks that are unseen during model training, since a domain discrepancy generally exists between the source and target tasks.

### 5.2. Results in transfer learning framework

After showing the necessity of transfer learning, we executed the above mentioned transfer learning flow on the pre-trained CNN. Since the size of the dataset plays an important role in

**Table 3**
Scenario setting of transfer learning in diagnosis case 2.

| Domain types | Source domain | Target domain |
|---|---|---|
| Description | Four gearbox conditions: health, high-speed gear fault, high-speed bearing fault and low-speed bearing fault (Corresponding labels 0–3). | |
| Fault conditions | Health, TB in high-speed gearing, OF in high-speed bearing and OF in low-speed bearing | Health, RC in high-speed gearing, RF in high-speed bearing and RF in low-speed bearing |

**Table 4**
Testing accuracies (%) using pre-trained CNN for PHM 2009 dataset.

| Domain types | No. of samples (training/testing) | CNN1 | CNN2 | CNN3 [24] | STFT-CNN [20] | DTS-CNN [21] |
|---|---|---|---|---|---|---|
| Source domain | 5000/1000 | 98.4 | 98.6 | 98.3 | 99.2 | 96.3 |
| Target domain | −/1000 | 54.7 | 67.1 | 63.1 | 75.0 | 44.4 |

**Table 5**
Testing accuracies (%) using pre-trained CNN for gearbox fault dataset.

| Domain types | No. of samples (training/testing) | CNN1 | CNN2 | CNN3 | STFT-CNN | DTS-CNN |
|---|---|---|---|---|---|---|
| Source domain | 15 000/1000 | 99.8 | 100 | 100.0 | 99.9 | 98.6 |
| Target domain | −/1000 | 77.1 | 53.7 | 48.4 | 28.5 | 45.8 |

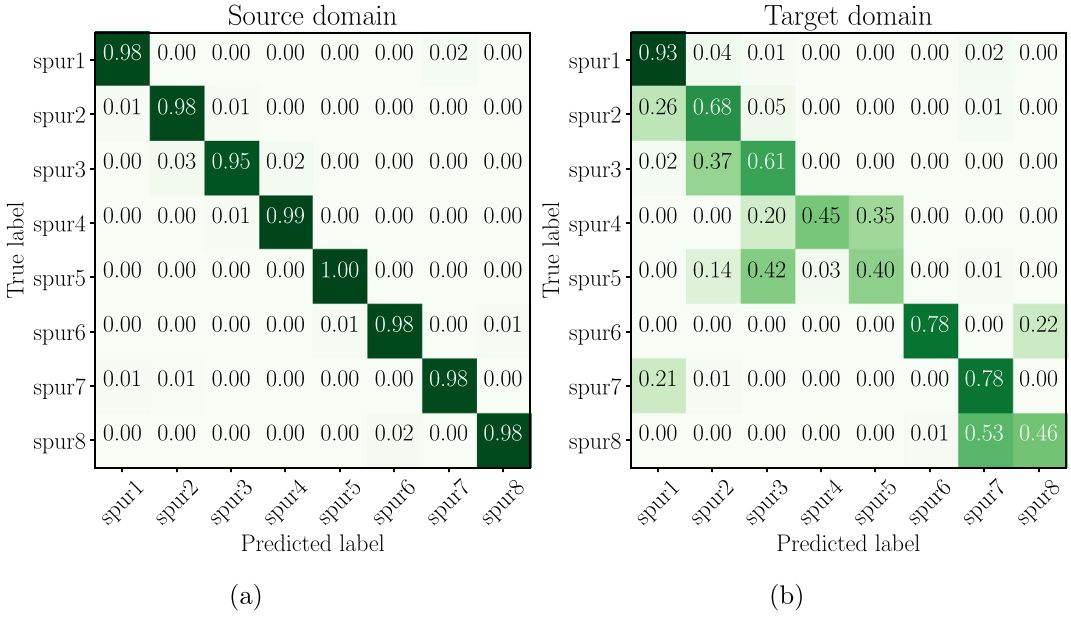


**Fig. 7.** Diagnosis results of the pre-trained CNN with 5 convolutional layers for two domains in PHM 2009 dataset.

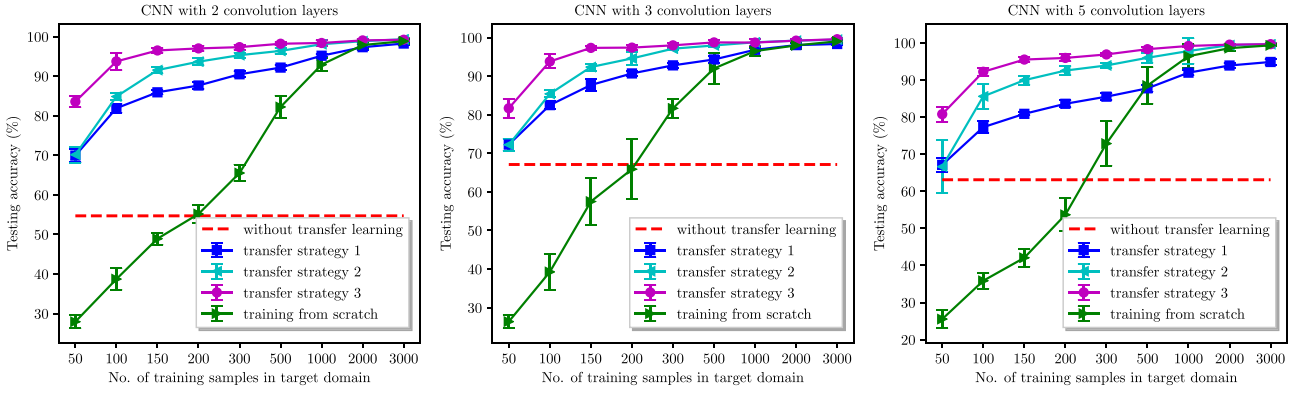


**Fig. 8.** Results comparison of diverse strategies in target domain with 3 CNNs in PHM 2009 dataset.

achieving better results for transfer learning, different numbers of target samples were used to fine-tune the pre-trained CNN. For comparison, the same amount of target samples were also used to train the network from scratch. Figs. 8 and 9 give the testing results of the diverse strategies with the 3 network structures in two datasets. It was found that the diagnosis performances were significantly improved by transfer learning in all the trials, compared with the base accuracy (red dotted line) as well as the

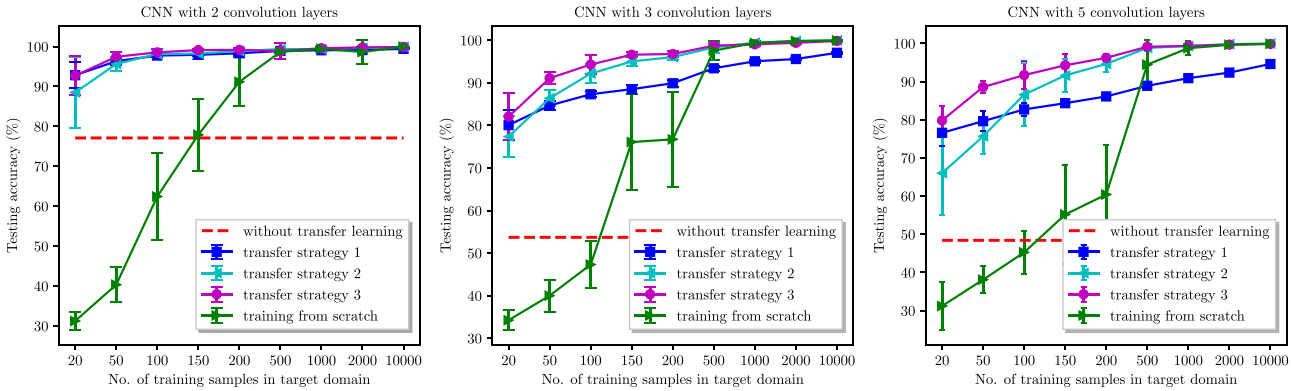


**Fig. 9.** Results comparison of diverse strategies in target domain with 3 CNNs in gearbox dataset.

results of traditional 2D-CNN methods in Tables 4 and 5. The results were reasonable, because more target data were used to fine tune the model, and the domain discrepancy problem was considered. It was worth mentioning that noticeable performance boosts were achieved in transfer learning framework only with a few target samples, such as 50 samples in PHM 2009 dataset and 20 samples in gearbox dataset. Meanwhile, the accuracies for training from scratch remained high with sufficient target samples, but they exhibited an evident drop as the number of samples decreased. In transfer learning flows, the strategy 3 attained a higher diagnostic accuracy than the other two methods with different sample sizes. Taking the results of CNN with 2 convolution layers in PHM 2009 dataset as an example, a satisfactory diagnostic rate (93.7%, 83.6%) could be still reached when using a small quantity of target samples (100, 50), suggesting that the feature representation layers are more easily adapted to the target domain with limited target supervision, while avoiding overfitting. For comparison, the other two transfer strategies could only obtain lower accuracies for most of the trials. The superior transfer capacity of strategy 3 could be confirmed according to extensive experiments, especially under limited target samples.

### 5.3. Network visualization

To gain a better understanding of the transferred network and to offer insights into the feature representation, the t-SNE technique was conducted to visualize the flatten layer in the transferred CNN. Taking the transferred network in PHM 2009 dataset as an example, the 2D feature maps are shown in Fig. 10. The corresponding confusion matrixes are shown in Fig. 11. It can be found that large areas were overlapping between diverse categories when the network was trained from scratch, indicating that limited high-quality training data was unable to efficiently train a network with millions of parameters.

In the transfer learning flow, the feature distributions of the different conditions were more separable. Especially, in the transferred network with strategy 3, the feature distributions of the same bearing condition were successfully clustered together, while those belonging to different categories were well separated (corresponding to a total 92.2% diagnostic rate), which means that the gaps across the source and target domains were effectively bridged by fine-tuning the feature descriptor part. For the transferred network with strategy 1, the weights of feature descriptor were frozen, that is, the front feature representation layers were directly copied from the pre-trained CNN without modifying. Due to the nonlinear mapping ability of the final fully-connected layers, the diagnosis accuracy reached up to 85.6% from 63.1% in Table 4 via fine tuning. However, the frozen feature descriptor was lack of the feature representation ability for the target data, which may restrict the further improvement of diagnosis performance. Some categories, such as spur 4 and spur 5, were seriously overlapped. For the transferred network with strategy 2, it was clear that the features also did not concentrate well, and some categories, such as spur 7 and spur 8, were slightly overlapping. All of these visualizations can clearly explain the diagnosis performances of each transfer strategy.

**Table 6**
Efficiency evaluation of diverse strategies with 3 CNNs in PHM 2009 dataset.

| Strategies | CNN1 | | CNN2 | | CNN3 | |
|---|---|---|---|---|---|---|
| | Epoch | Time (s) | Epoch | Time (s) | Epoch | Time (s) |
| Training from scratch | 341 | 138.2 | 239 | 117 | 317 | 414.4 |
| Transfer strategy 1 | 108 | 27.2 | 103 | 28.3 | 37 | 12.3 |
| Transfer strategy 2 | 116 | 46.8 | 53 | 26.4 | 82 | 106.5 |
| Transfer strategy 3 | 176 | 64.1 | 39 | 17.9 | 104 | 131.6 |

### 5.4. Network training efficiency

Generally, the initialization based on pre-training has been proved to be significant for accelerating convergence in deep architecture, and thus the transfer learning framework utilizing the pre-trained CNN could possess higher computational efficiency, compared with the traditional training from scratch. In this part, the efficiency comparison of transfer diagnosis techniques and traditional ones in the PHM 2009 fault dataset was given for the quantitative assessment. The sufficient target samples (3000) were used to train a network for scratch or fine tune the pre-trained CNN in source domain. The early stopping with patience 50 was employed as the training regulation. High diagnosis accuracies can be achieved by all these strategies, as shown in Fig. 8. The training epochs and time for diverse strategies were recorded. As listed in Table 6, the three transfer strategies are all clearly superior to the traditional training in terms of efficiency. Specially, the 12.3 s of transfer strategy 1 in CNN3 and the 17.9 s of transfer strategy 3 in CNN2 are far less than the original training times 414.4 s, 117 s respectively. The availability and flexibility of transfer learning framework can be further illustrated not only in target accuracy, but also the computational efficiency.

## 6. Discussion

For transfer strategy 1, as the convolution kernels in the pre-trained CNN are completely learned from the source data, thus, they are tailored to the source tasks. Keeping the weights of the convolution kernels frozen and only fine tuning the classifier could not essentially adapt the feature representation to

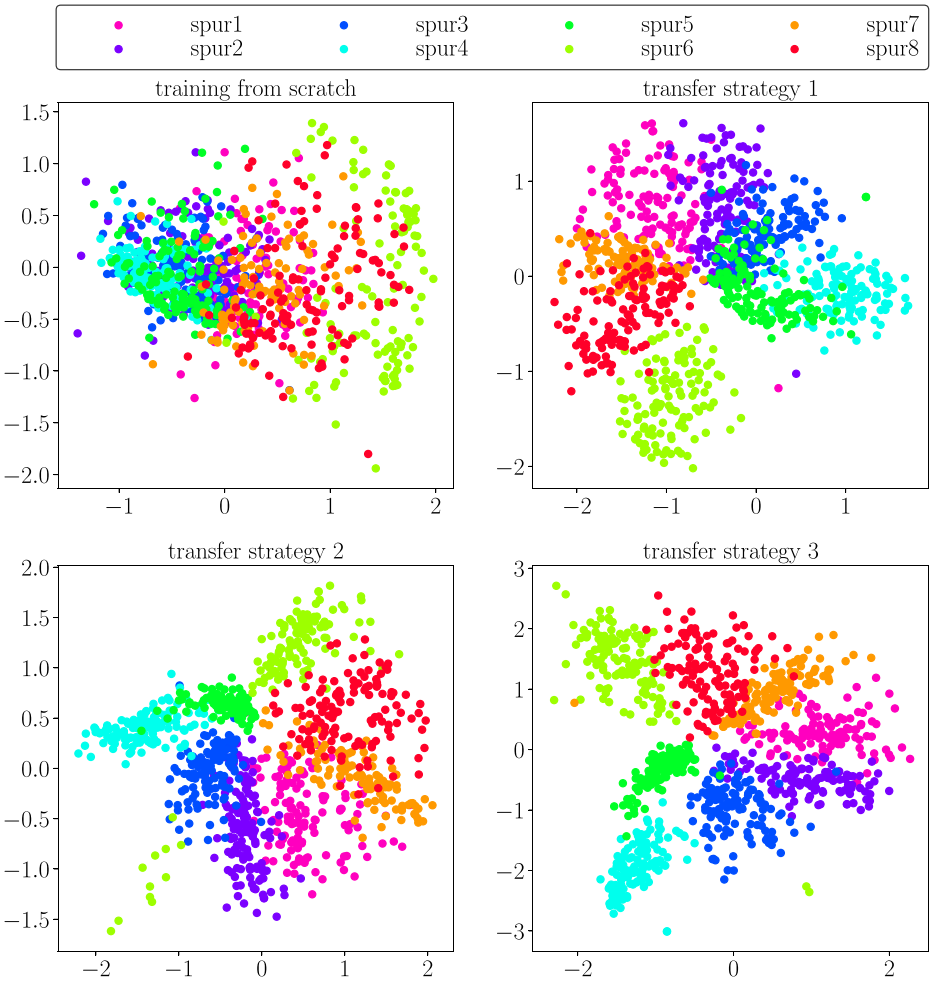


**Fig. 10.** Network visualization in PHM 2009 dataset with 100 target samples and the pre-trained CNN with 5 convolutional layers.

the target task when the signal characteristics exhibited distinct differences across the domains. Different working conditions or machine failure types generally result in different dynamic characteristics, and they further present diverse signal characteristics. On account of this, the strategy of only fine-tuning the classifier part may be inapplicable for learning transferable features with a large domain discrepancy for machine fault diagnosis tasks.

For the transfer strategy 2, in principle, fine tuning the whole pre-trained CNN can benefit from the co-adaptation of the feature descriptor and classifier, that is, the weights can interact with each other on successive layers, so that the pre-trained CNN will be thoroughly transferred and adapted to the target domain. Basically, if there is a large number of target samples, a new neural network can be directly trained from scratch, whether the lower level filters or the higher level classifiers are completely learned from the target data. In this transfer strategy, the pre-trained CNN gives the initial architecture and weights to facilitate the fine tuning, whose process is similar to the generally network training. Therefore, fine-tuning the whole pre-trained CNN can bring out the greatest potential for learning the transferable features among these transfer strategies, especially when the similarity between the source task and the target task is low. However, it is worth emphasizing that training whole networks with limited target samples may easily lead to overfitting. The experimental results of the two datasets also prove this fact. When providing sufficient samples, the joint fine-tuning strategy is able to do just as well as only fine-tuning the feature descriptor, whereas the case of using a small sample size exhibits a clear performance drop.

Contrary to strategy 1, the results of strategy 3 in both of the two datasets shows a particularly surprising effect: keeping the classifier frozen and fine tuning the feature descriptor results in a network with stronger feature transferability. More precisely, from Fig. 10, the features between the different categories in the target domain are well-separated, which means that this transfer strategy is capable of learning the inherent features, and giving a proper representation of the target samples. By fine-tuning the transferred feature descriptor, the weights of the convolution kernel have been fully adjusted to the target domain. Two remarks should be highlighted here about this transfer strategy. (1) As shown in Fig. 12, we can find that most of the parameters exist in the first fully-connected layer. Since the classifier is left frozen, only the weights of the feature descriptor need to be fine-tuned. Owing to the fewer learning parameters, it can still achieve satisfactory transfer results, even with a small training sample size. (2) As more fault categories cannot be added to the frozen classifier, another more flexible approach is to remove the classifier after fine-tuning. The front convolutional layers are used to extract features, and the popular classifiers, not limited but including SVM, which can perform well with high dimensional inputs and small sample problems, are used for decision-making.

When the base network has a deeper structure (more than 3 convolutional layers), it is important to quantify the generality versus the specificity of each layer so as to find the task-specific layers. By fine tuning these task-specific layers, the risk of overfitting can be lowered with only limited target samples, and the transferability and scalability of the pre-trained network will be

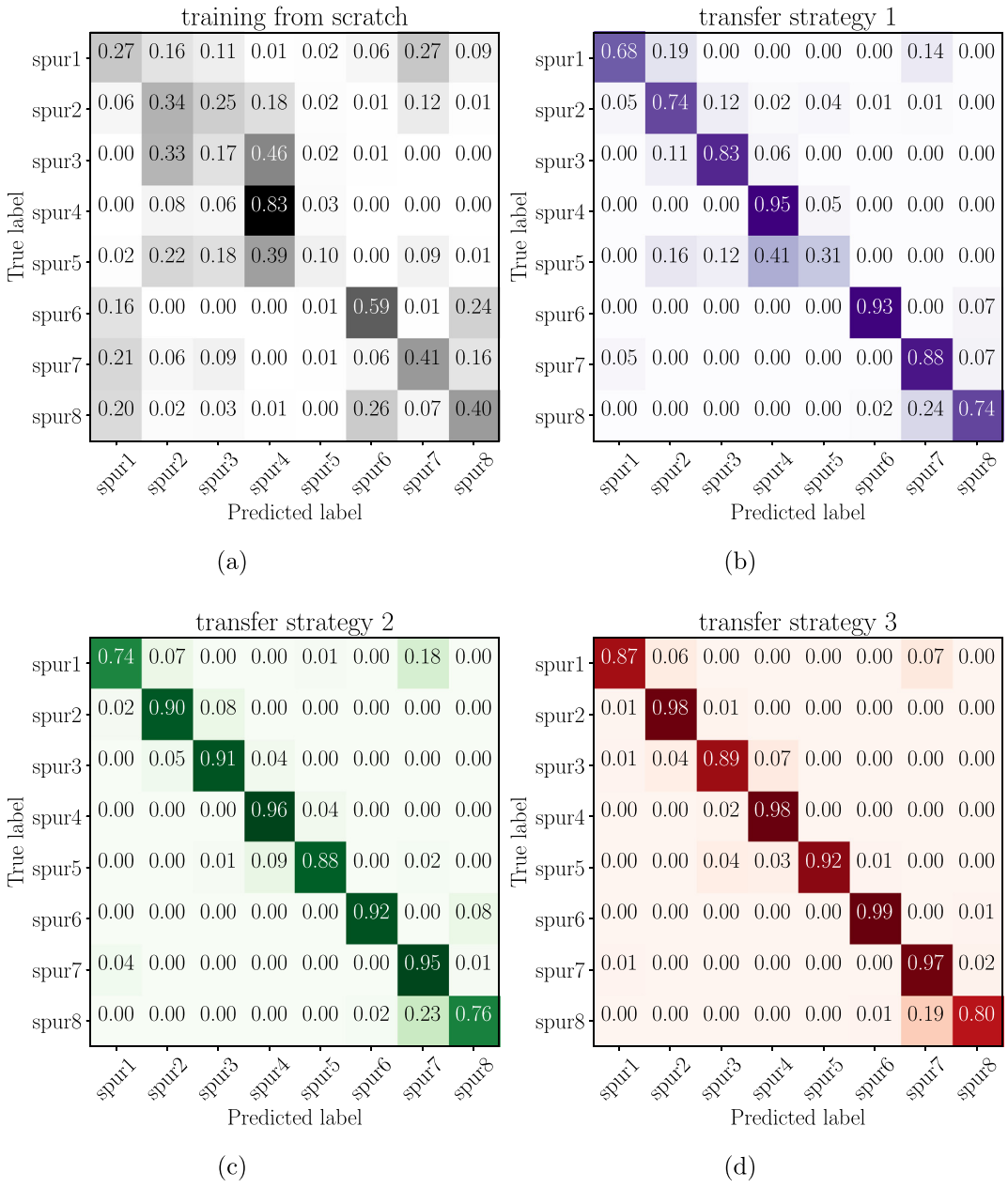


**Fig. 11.** The corresponding confusion matrixes for the four strategies.

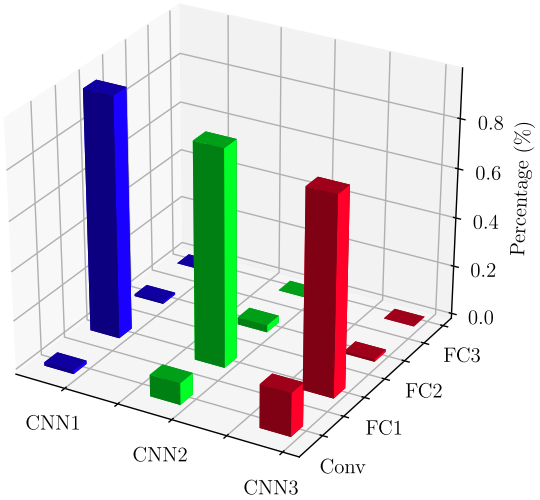


**Fig. 12.** The percentages of parameters for diverse parts in the CNNs.

further enhanced. In the above analysis, we explore the transfer of the whole feature descriptor of the base network to the target domain, and we achieve encouraging diagnosis results via fine tuning. Moreover, the diverse layers in the feature descriptor for the pre-trained CNN with 5 convolutional layers are frozen to investigate the degrees of generality or specificity. The results are given in Fig. 13. It is clear the performance is greatly improved from the blue line to the purple line, indicating that these layers are more specific, and that the features of the higher layers must be closely related to the dataset and task in the source domain. On the contrary, the slight improvement from the green line to red line suggests that the features in the first convolution layer may be general for many applications, rather than for a specific task. These results have revealed that the features in the deep neural network always present a transition from general to specific with the depth increasing, which provides a reference for how to transfer the specific convolutional layers in a deep structure.

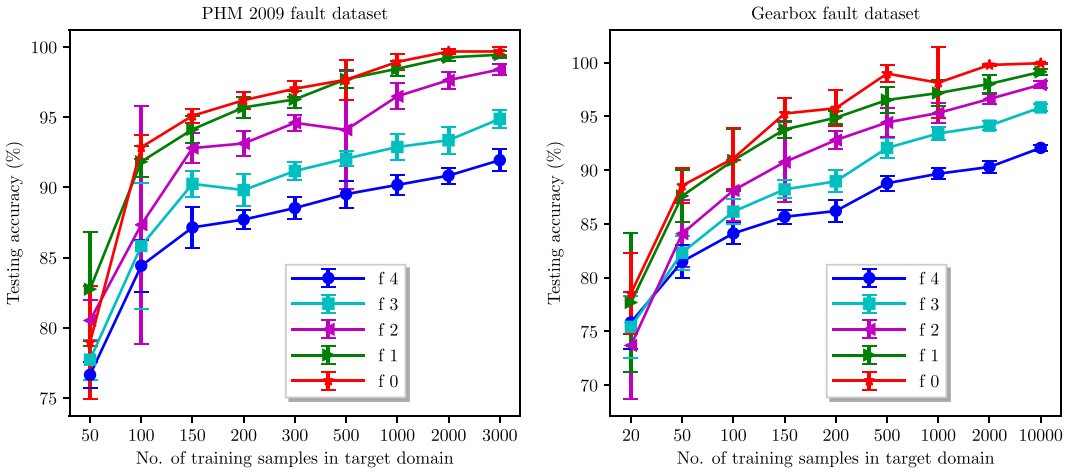


**Fig. 13.** The investigation of the generality versus the specificity of each layer in the CNN with convolutional layers: f-n means freezing the first n convolutional layers; for example, f-4 means freezing the first 4 convolutional layers and only fine-tuning the last convolutional layer.

## 7. Conclusions and future works

Transfer learning is a promising tool for improving the performance of target problems by exploiting knowledge from the previous tasks that are different but similar. Unlike traditional machine learning techniques on fault diagnosis, transfer learning focuses more on transferring existing knowledge or skills to novel tasks, to meet the needs for there being not enough data to train a model from scratch, especially if some of the faults rarely occur in reality. By exploring the transferable features of a well-developed model, it enables us to build and train a large network efficiently and accurately when high-quality training data is limited for new tasks. With this motivation, this work presents a novel transfer learning framework on the basis of a pre-trained CNN, where the feature transferability at different levels of the deep structure is discussed and compared in two case studies. The results show that the proposed framework can transfer the features of the pre-trained CNN to the target domain with different working conditions or fault types, and a high accuracy is achieved for both cases.

One of the fundamental requirements for transfer learning is the presence of pre-trained models that perform well on source tasks. In many fields, such as computer vision and natural language processing, the pre-trained models are usually shared while being trained to a relatively universal and stable state [40]. Aiming at the frequently occurring faults, such as bearing faults, gear faults and rotor-related faults, general deep learning architectures need to be developed based on many standard mechanical fault datasets, so as to achieve universal pre-trained models. Besides, in this work, the transfer experiments are all performed in the same machinery. The success of transfer learning between different objects or machines will make real-world applications more flexible. Thus, this will facilitate the application of transfer learning between experiments and industrial tasks. Consequently, further work is being carried out including (i) the establishment of a state-of-the-art and universal pre-trained model for mechanical fault diagnosis tasks, and (ii) the application of transfer learning between the experimental data and the industrial data.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 11572167). The authors would like to express their sincere gratitude to lab associate Mr. Shaohua Li for his contribution on the experiments. Special thanks should also be expressed to the editors and reviewers for their review work.

## Conflict of 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.

## References

[1] Lee J, Wu F, Zhao W, Ghaffari M, Liao L, Siegel D. Prognostics and health management design for rotary machinery systemsreviews, methodology and applications. Mech Syst Signal Process 2014;42:314–34.
[2] Jiang D, Liu C. Machine condition classification using deterioration feature extraction and anomaly determination. IEEE Trans Reliab 2011;60:41–8.
[3] Widodo A, Yang BS, Gu DS, Choi BK. Intelligent fault diagnosis system of induction motor based on transient current signal. Mechatronics 2009;19:680–9.
[4] Henriquez P, Alonso JB, Ferrer MA, Travieso CM. Review of automatic fault diagnosis systems using audio and vibration signals. IEEE Trans Syst Man Cybern Syst 2014;44:642–52.
[5] Han T, Jiang D, Sun Y, Wang N, Yang Y. Intelligent fault diagnosis method for rotating machinery via dictionary learning and sparse representation-based classification. Measurement 2018;118:181–93.
[6] Han T, Jiang D, Qi Z, Lei W, Kai Y. Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery. Trans Inst Meas Control 2018;40:2681–93.
[7] Liu C, Jiang D, Yang W. Global geometric similarity scheme for feature selection in fault diagnosis. Expert Syst Appl 2014;41:3585–95.
[8] Sabina B, Ivana B, Verheij ER, Raymond R, Sunil K, Macdonald IA, et al. Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. Anal Chem 2006;78:567–74.
[9] Cateni S, Colla V, Vannucci M, Vannocci M. A procedure for building reduced reliable training datasets from real-world data. Acta Press 2014.
[10] Garca S, Luengo J, Herrera F. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. Knowl-Based Syst 2016;98:1–29.
[11] 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.
[12] Cui L, Yu Z, Gong X, Kang C. Application of pattern recognition in gear faults based on the matching pursuit of a characteristic waveform. Measurement 2017;104:212–22.
[13] Shao H, Jiang H, Zhao H, Wang F. A novel deep autoencoder feature learning method for rotating machinery fault diagnosis. Mech Syst Signal Process 2017;95:187–204.
[14] Schmidhuber J. Deep learning in neural networks: An overview. Neural Netw 2015;61:85–117.
[15] 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.
[16] 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.

[17] 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.
[18] Wang S, Xiang J, Zhong Y, Zhou Y. Convolutional neural network-based hidden markov models for rolling element bearing fault identification. Knowl-Based Syst 2018;144:65–76.
[19] Jia F, 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.
[20] 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.
[21] 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;13:1310–20.
[22] Ding X, He Q. Energy-fluctuated multiscale feature learning with deep convnet for intelligent spindle bearing fault diagnosis. IEEE Trans Instrum Meas 2017;66:1926–35.
[23] Jing L, Wang T, Zhao M, Wang P. An adaptive multi-sensor data fusion method based on deep convolutional neural networks for fault diagnosis of planetary gearbox:. Sensors 2017;17:414.
[24] 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;17:425.
[25] 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.
[26] Zhao B, Huang B, Zhong Y. Transfer learning with fully pretrained deep convolution networks for land-use classification. IEEE Geosci Remote Sens Lett 2017;14:1436–40.
[27] Lu Z, Zhu Y, Pan SJ, Xiang EW, Wang Y, Yang Q. Source free transfer learning for text classification. In: AAAI conference on artificial intelligence. 2014.
[28] 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.
[29] Kandaswamy C, Monteiro JC, Silva LM, Cardoso JS. Multi-source deep transfer learning for cross-sensor biometrics. Neural Comput Appl 2016;28:1–15.
[30] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2010;22:1345–59.
[31] Weiss K, Khoshgoftaar TM, Wang DD. A survey of transfer learning. J Big Data 2016;3:9.
[32] 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.
[33] Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks?, Eprint Arxiv 27 (2014) 3320–3328.
[34] Zhang R, Tao H, Wu L, Guan Y. Transfer learning with neural networks for bearing fault diagnosis in changing working conditions. IEEE Access 2017;PP. 1–1.
[35] PHM GD. Phm data challenge 2009., https://www.phmsociety.org/competition/PHM/09.
[36] Wei Y, Zhang Y, Yang Q. Learning to transfer, Eprint Arxiv (2017).
[37] Xiang L, Wei Z, Qian D, Jian-Qiao S. Multi-layer domain adaptation method for rolling bearing fault diagnosis. Signal Process 2019;157:180–97.
[38] Liang G, Yaguo L, Saibo X, Tao Y, Naipeng L. Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data. IEEE Trans Ind Electron 2018;PP. 1–1.
[39] Zhang B, Li W, Hao J, Li XL, Zhang M. Adversarial adaptive 1-d convolutional neural networks for bearing fault diagnosis under varying working condition, Eprint Arxiv (2018).
[40] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. Vol. 1. 2012, p. 1097–105.