Proceedings of 2018 ISFA
2018 International Symposium on Flexible Automation
Kanazawa, Japan, 15 - 19 July, 2018

(ISFA2018-L101)

MACHINE LEARNING ENABLED GEARBOX FAULT PATTERN RECOGNITION

Pei Cao
Department of Mechanical Engineering
University of Connecticut
Storrs, CT 06269, USA
pei.cao@uconn.edu

J. Tang
Department of Mechanical Engineering
University of Connecticut
Storrs, CT 06269, USA
Jiong.tang@uconn.edu

ABSTRACT

Fault pattern recognition in complex mechanical systems such as gearbox has always been a great challenge. The performance of a classic fault pattern recognition approach heavily depends on domain expertise and the classifier applied. This paper proposes a deep convolutional neural network-based transfer learning approach that not only entertains adaptive feature extractions, but also requires only a small set of training data. The proposed transfer learning architecture essentially consists of two sequentially connected pieces; first is of a pre-trained deep neural network that serves to extract features automatically, the second piece is a neural network aimed for classification which is to be trained using data collected from gearbox experiment. The proposed approach performs gear fault pattern recognition using raw accelerometer data. The achieved accuracy indicates that the proposed approach is not only sensitive and robust in performance, but also has the potential to be applied to other pattern recognition practices.

INTRODUCTION

Health monitoring is critical to modern machinery systems and has been motivating research aiming at fault pattern recognition techniques. Gearbox, as one of the most common components used in those systems, is prone to fault conditions and failures, because of the severe working condition with high mechanical loading and long operational time. Vibration signals are most widely used to infer the health condition of gear system, because they contain rich information and can be easily measured with off-the-shelf, low-cost sensors. However, the practice of fault pattern recognition using vibration signals is quite demanding. Generally, a manually selected feature extraction technique is first applied to vibration signals measured from a gear system that characterizes the fault-related features. Subsequently, a classifier is trained and applied to new signals to recognize fault occurrence in terms of type and severity. There have been extensive and diverse attempts in identifying useful features from gear vibration signals, which fall into three main categories: time-domain analysis (Zhou et al, 2008; Parey and Pachori, 2012), frequency domain-analysis (Fakhfakh et al, 2005; Li et al, 2015; Wen et al, 2015) and time-frequency analysis (Tang et al, 2010; Chaari et al, 2012; Yan et al, 2014; Zhang and Tang, 2018). Time-domain statistical approaches can capture the changes in amplitude and phase modulation caused by faults. In comparison, spectrum analysis can extract the features more efficiently such that distributed faults with clear sidebands can be detected. To deal with noise and at the same time utilize the transient components in vibration signals, many recent research efforts have place their focus on joint time-frequency domain analysis utilizing, such as Wigner-Ville distribution, short time Fourier transform, and various wavelet transforms. The time-frequency distribution in theory may lead to rich information regarding the time and frequency-related events in signals.

Although manual feature extraction methods have seen great successes, their effectiveness is hinged upon the specific features adopted in the diagnostic analysis. It is worth emphasizing that the choices of features as well as the often-applied signal processing techniques are generally based on domain. For example, while wavelet transforms have been popular, it is evident from large amount of literature that there does not seem to be a consensus on what kind of wavelet to be used for gear fault pattern recognition. This should not come as a surprise. On one hand, gear faults occur primarily at microstructure or even material level but their effects can only be observed indirectly at a system level; consequently, there exists a many-to-many relationship between actual faults and the observable quantifies (i.e., features) for a given gear system (Lu et al, 2012). On the other hand, different gear systems have different designs which lead to distinct dynamic characteristics. As such, the result on features manually selected and, to a large extent, the methodology employed to extract these features for one gear system design may not be easily extrapolated to a different gear system design.

Fundamentally, condition monitoring and fault diagnosis of gear systems belongs to the general field of data mining. The advancements in related algorithms and computational power have led to the wide spread of machining learning techniques to various applications. Most recently, deep neural network-based methods are progressively being investigated in gear fault pattern recognition in an automated manner with minimal tuning. For example, Zhang et al (2015) developed a deep learning network for degradation pattern classification and demonstrated the efficacy using turbofan engine dataset. Li et al (2016) proposed a deep random forest fusion technique for gearbox fault diagnosis. Weimer et al (2016) examined the usage of deep convolutional neural network for industrial inspection and demonstrated excellent defect detection results. Ince et al (2017) developed a fast motor condition monitoring system using a 1-D convolutional neural network with good classification accuracy. Abdeljaber et al (2017) performed real-time damage detection using convolutional neural network and showcased satisfactory efficiency.

Deep neural network is undoubtedly a powerful tool in pattern recognition and data mining. As an end-to-end hierarchical system, it inherently blends the two essential elements in condition monitoring, feature extraction and classification, into a single adaptive learning frame. It should be noted that the amount of training data required for satisfactory results depends on many aspects of the specific problem being tackled, such as the correctness of training samples, the number of classes, and the degree of separation between each class. In most machinery diagnosis investigations, the lack of labeled training samples is a common issue. To improve the performance given limited training data, some recent studies have attempted to combine data processing and data augmentation techniques, e.g., discrete wavelet transform (Saravanan and Ramachandran, 2010), antialiasing/decimation filter (Ince et al, 2017), and wavelet packet transform (Li et al, 2016), with neural networks for fault diagnosis. Nevertheless, the data processing techniques employed, subjected to selection based on domain expertise, may hurt the objective nature of neural networks and to some extent undermines the usage of such tools.

In this research, we present a deep neural network-based transfer learning approach utilizing limited time domain data for gearbox fault pattern recognition. This approach inherits the non-biased nature of neural network-type methods that avoids the manual selection of features. Meanwhile, the issue of limited data is overcome by formulating a new architecture with two parts. Massive image data (1.2 million) from ImageNet are used first to train the original deep neural network model. Then the parameters of the original neural network are partially transferred to construct the first part of the proposed architecture. The second part of the neural net is further trained using experimentally generated gear fault data. With this new architecture, highly accurate gear fault pattern recognition can be achieved using limited time-domain data without any subjective data processing techniques to assist feature extraction.

APPROACH FORMULATION

The proposed approach is built upon deep neural network transfer-learning. In this section, we start from the fundamental formulations of deep convolutional neural network and transfer learning, followed by specific architecture developed for gear fault pattern recognition.

Deep Convolutional Neural Network

Convolutional Neural Network (CNN) is a class of biologically inspired neural network featuring one or multiple convolutional layers that simulate human visual system (LeCun et al, 1990). In recent years, due to the enhancement in computational power and the dramatic increase in the amount of data available in various applications, CNN-based methods have shown significant improvements in performance and thus have become the most popular class of approaches for pattern recognition tasks such as image classification (Krizhevsky et al, 2012), natural language processing (Kim, 2016), recommending systems (Van den Oord, et al, 2013) and fault detection (Ince et al, 2016). CNN learns how to extract and recognize characteristics of the target task by combining and stacking convolutional layers, pooling layers and fully connected layers in its architecture. Figure 1 illustrates a simple CNN with an input layer to accept input images, a convolution layer to extract features, a ReLU layer to augment features through non-linear transformation, a max pooling layer to reduce data size, and a fully connected layer combined with a Softmax layer to classify the input to defined labels. The parameters are trained through a training dataset and updated using back propagation algorithm to reflect the features of the task that may not be recognized otherwise. The basic mechanism of layers in CNN is outlined as follows.

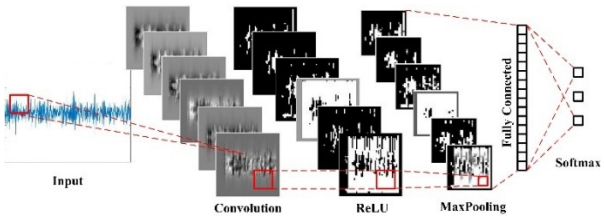


Figure 1 An example convolutional neural network.

Convolutional layer. Each feature map in the convolution layer shown in Figure 1 is generated by a convolution filter. Generally, the input and convolution filters are tensors of size $m \times n$ and $p \times q \times K$ (K is the number of filter used), respectively. Stride (i.e., step size of the filter sliding over input) is set to 1 and padding (i.e., the number of rows and columns to insert around the original input) is set to 0. The convolution operation can be expressed as,

$$y_{d_1,d_2,k} = \sum_{i=0}^{p}\sum_{j=0}^{q} x_{d_1 i, d_2 j} \times f_{i,j,k} \quad (1)$$

Where y, x and f denote the element in feature map, input and convolutional filter, respectively. $f_{i,j,k}$ represents the element on the i-th column and j-th row for filter k. $y_{d_1,d_2,k}$ is the element on the d_1-th column and d_2-th row of feature map k. And $x_{d_1 i, d_2 j}$ refers to the input element on the i-th column and

j-th row of the stride window specified by d_1 and d_2. Equation (1) gives a concise representation of the convolution operation when the input is 2-demensional, and stride and padding are 1 and 0. Higher dimension convolution operations can be conducted in a similar manner. In CNN, multiple convolution filters are used in a convolutional layer, each acquiring a feature piece in its own perspective from the input image specified by the filter parameters. Regardless of what and where a feature appears in the input, the convolution layer will try to characterize it from various perspectives that have been tuned automatically by the training dataset.

ReLU layer. In CNN, ReLU (rectified linear units) layers are commonly used after convolution layers. In most cases, the relationship between the input and output is not linear. While the convolution operation is linear, the ReLU layer is designed to take non-linear relationship into account, as shown in the equation below,

$$\bar{y} = \max(0, y) \tag{2}$$

The ReLU operation is applied to each feature map and returns an activation map. The depth of the ReLU layer equals to that of the convolution layer.

Max pooling layer. Max pooling down-samples a sub-region of the activation map to its maximum value,

$$\hat{y} = \max_{L_1 \le i \le U_1,\, L_2 \le i \le U_2} \bar{y}_{i,j} \tag{3}$$

where $L_1 \le i \le U_1$ and $L_2 \le j \le U_2$ define the sub-region. The max pooling layer not only makes the network less sensitive to location changes of a feature but also reduces the size of parameters, thus alleviating computational burden.

Transfer Learning

The performance of a convolutional neural network can be improved by upscaling the CNN equipped. The scale of a CNN concurs with the scale of the training dataset. Naturally, the deeper the CNN, the more parameters need to be trained, which requires a substantial amount of valid training samples. Nevertheless, in the application of gear fault pattern recognition, the training data is always not as sufficient as that of other tasks such as natural image classification. Transfer learning, on the other hand, can achieve prominent performance commensurate with large scale CNNs using only a small set of training date. By partially deploying a pre-trained neural network, transfer learning provides a possible solution to improve the performance of a novel task with small training dataset. Classic transfer learning approaches transfer (copy) the first n layers of a well-trained network to the target network of layer $m > n$. Initially, the last $(m-n)$ layers of the target network are left untrained. They are trained subsequently using the training data from the novel task.

Transfer learning becomes possible and promising because, as has been discovered by recent studies, the layers at the convolutional stages (convolution layers, ReLU layers and pooling layers) of the convolutional neural network trained on large dataset indeed extract general features of inputs, while the layers of fully connected stages (fully connected layers, softmax layers, classification layers) are more specific to task (Zeiler and Fergus, 2013; Sermanet et al, 2014). Therefore, the n layers transferred to the new task can be regarded as a well-trained feature extraction tool and the last few layers serve as a classifier to be trained. Even with substantial training data, initializing with transferred parameters can improve the performance in general (Yosinski et al, 2014). In this research, transfer learning is implemented to gearbox fault pattern recognition. The CNN is well-trained in terms of pulling characteristics from images. As illustrated in Figure 5, the parameters in the convolution stage, i.e., the parameters used in the convolution filter, the ReLU operator and the max pooling operator are transferred to the fault pattern recognition task. The parameters used in the fully connected layer and the softmax layers are trained subsequently using a small amount of training data generated from gear fault experiments.

Proposed Architecture

In this sub-section we present the details of the proposed architecture. The deep CNN adopted in this study as base architecture is originally proposed by Krizhevsky et al (2012), which is essentially composed of 5 convolution stages and 3 fully connected stages (Figure 2). This base architecture showed its extraordinary performance in Large Scale Visual Recognition Challenge 2012 (ILSVRC2012), and has since been repurposed for other learning tasks (Shie et al, 2015). To the best of our knowledge, the architecture has yet to be used for fault pattern recognition using time domain inputs. In the base architecture, the parameters are trained using approximately 1.2 million human/software labeled 3D true-color nature images from ImageNet. The trained parameters in the first 7 stages are well-polished in the sense of characterizing high-level abstractions of the input image and thus have the potential to be used for other tasks with image inputs.

In gear fault pattern recognition, vibration signals can be sampled using accelerometers as gear rotates. Such vibration signals can then be represented by 3D grey-scale/true-color images as. Although the vibration images may look different from the images used to train the original CNN, useful features can be extracted in a similar manner if the CNN adopted is able to identify high-level abstractions. Therefore, as a deep convolution neural network, the first 7 stages of the base architecture can be transferred to facilitate gear fault pattern recognition. The first 7 stages indeed serve as a general well-trained tool for automatic feature extraction. The more stages and layers used, the higher level of features can be obtained. The final stage is left to be trained as a classifier using the experiment data specific to the pattern recognition task. As specified in Table 1, a total number of 24 layers are used in the proposed architecture; the parameters and specifications used in the first 21 layers are transferred from the base architecture.

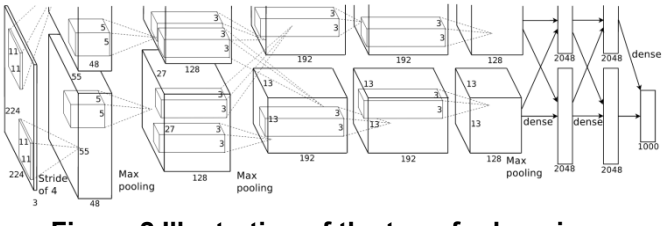


Figure 2 Illustration of the transfer learning architecture (adopted from (Krizhevsky et al, 2012)).

In this study, the loss function used is the cross-entropy function given as follows,

$$E(\boldsymbol{\theta}) = -\hat{\mathbf{L}} \ln\left(CNN(\mathbf{X},\boldsymbol{\theta})\right) + \gamma \left\|\boldsymbol{\theta}\right\|_2 = -\hat{\mathbf{L}} \ln \mathbf{L} + \gamma \left\|\boldsymbol{\theta}\right\|_2 \quad (4)$$

where $\left\|\boldsymbol{\theta}\right\|_2$ is a l_2 normalization term to prevent the network from over-fitting. Equation (4) quantifies the difference between correct output labels and predicted labels. And the loss is then back propagated to update the parameters using the stochastic gradient descent method (Sutskever et al, 2013) given as,

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \alpha \nabla E(\boldsymbol{\theta}_i) + \beta(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}) \quad (5)$$

where α is the learning rate, i is the number of iteration, and β stands for the contribution of previous gradient step.

Table 1 Specifications of the proposed architecture

Stage	Layer	Name	Specifications
1 (transferred)	1	Convolution	11*11*96
	2	ReLU	N/A
	3	Normalization	5 channels/element
	4	Max pooling	3*3
2 (transferred)	5	Convolution	5*5*256
	6	ReLU	N/A
	7	Normalization	5 channels/element
	8	Max pooling	3*3
3 (transferred)	9	Convolution	3*3*384
	10	ReLU	N/A
4 (transferred)	11	Convolution	3*3*384
	12	ReLU	N/A
5 (transferred)	13	Convolution	3*3*256
	14	ReLU	N/A
	15	Max pooling	3*3
6 (transferred)	16	Fully connected	4096
	17	ReLU	N/A
	18	Dropout	50%
7 (transferred)	19	Fully connected	4096
	20	ReLU	N/A
	21	Dropout	50%
8 (to be trained)	22	Fully connected	9
	23	Softmax	N/A
	24	Classification	Cross entropy

EXPERIMENTAL STUDIES

Experimental Setup

Many types of mechanical faults and failures can occur to gears in a gearbox. Vibration signals collected from such a system are usually used to reveal information about its operating condition. In this study, experimental cases are carried out on a two-stage gearbox with replaceable gears as shown in Fig. 3. The speed of the gear is controlled by a motor. The torque is supplied by a magnetic brake which can be adjusted by changing its input voltage. A 32-tooth pinion and an 80-tooth gear are installed on the first stage input shaft. The second stage consists of a 48-tooth pinion and 64-tooth gear. The input shaft speed is measured by a tachometer and gear vibration signals are measured by an accelerometer. The signals are recorded through a dSPACE system (DS1006 processor board, dSPACE Inc., Wixom, MI) with sampling frequency of 20 KHz. As illustrated in Fig. 4, nine different gear faults are introduced to the pinion on the input shaft including health, missing tooth, root crack, spalling and chipping tip with 5 different levels of severity. Dynamic responses of a system involving gear mechanism are angle-periodic. The gearbox system is usually treated as time-periodic system while the rotational speed is assumed to be constant. This assumption is generally not accurate because of load disturbances, geometric tolerances, and motor control errors, etc (Zhang and Tang, 2018). In this study, the originally time-domain vibration signals are converted from time-even to angle-even with evenly angular increment.

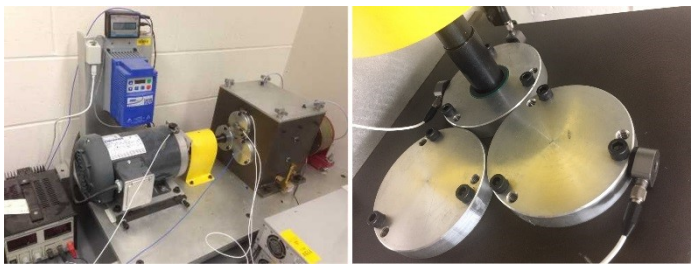
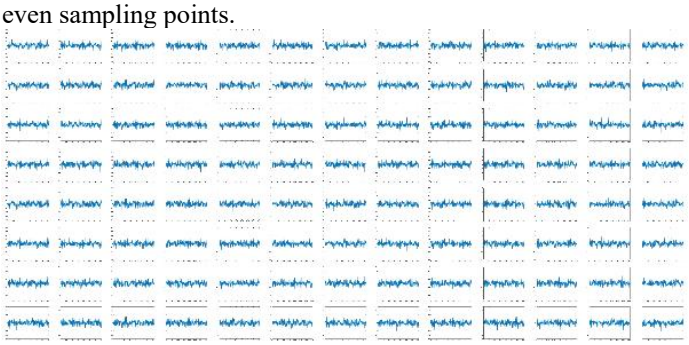


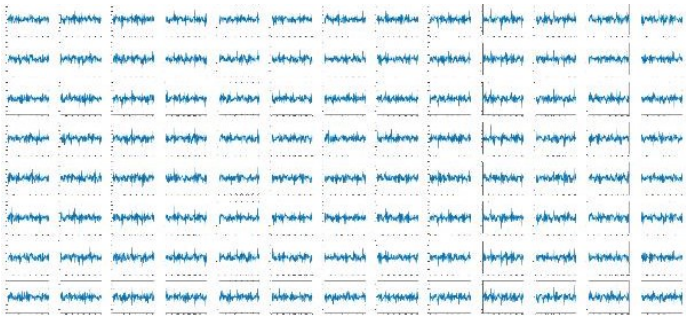
Figure 3 Gearbox for experimental study.

Figure 4 Nine pinions with different health conditions (five levels of severity for chipping tip).

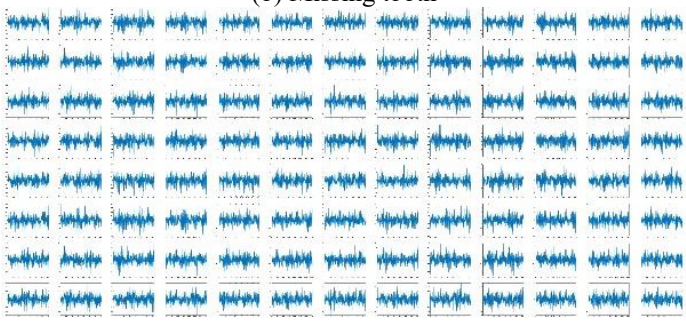
For each gear condition, 104 signals are generated using the gearbox system. For each signal, 3,600 angle-even samples are recorded during 4 gear revolutions first for the case study. Figure 5 shows all 104 signals of each type of gear condition where the vertical axis is the acceleration of the gear (rad/s^2) and the horizontal axis corresponds to the 3,600 angel-even sampling points.



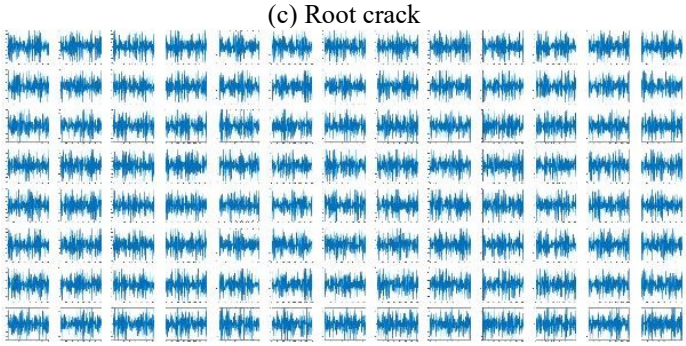
(a) Healthy



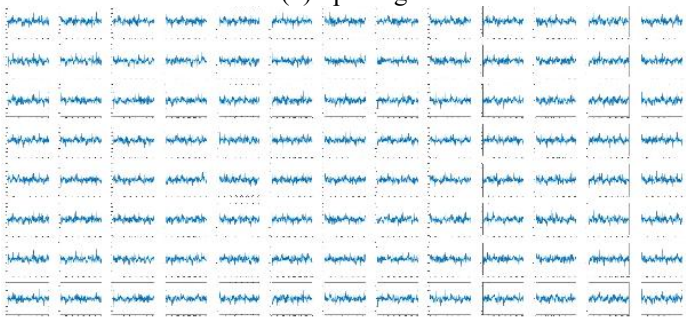
(b) Missing tooth



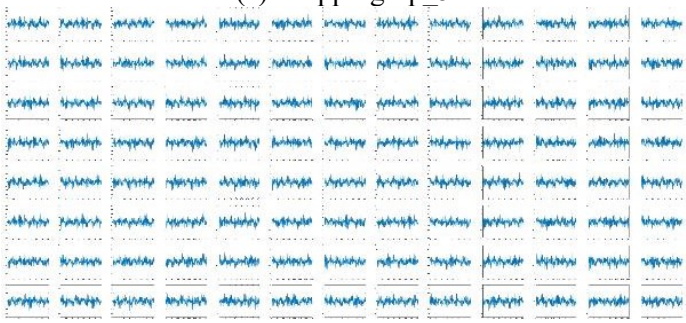
(c) Root crack



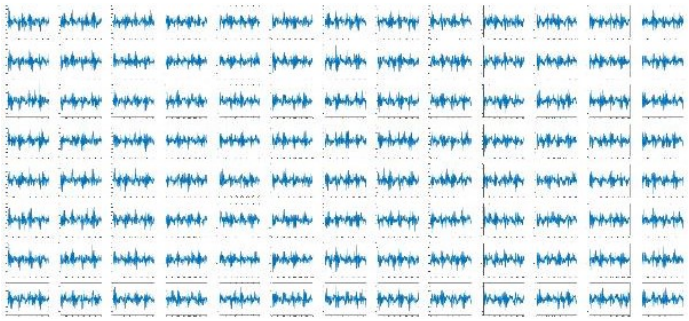
(d) Spalling



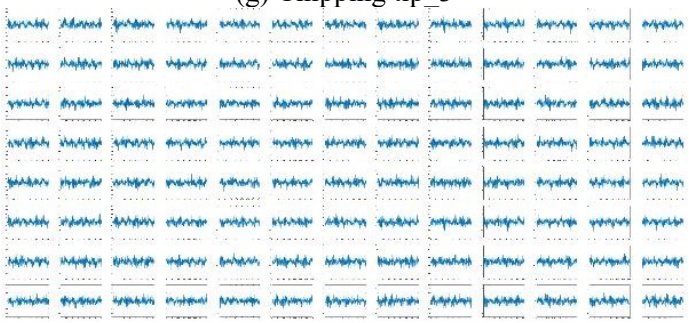
(e) Chipping tip_5



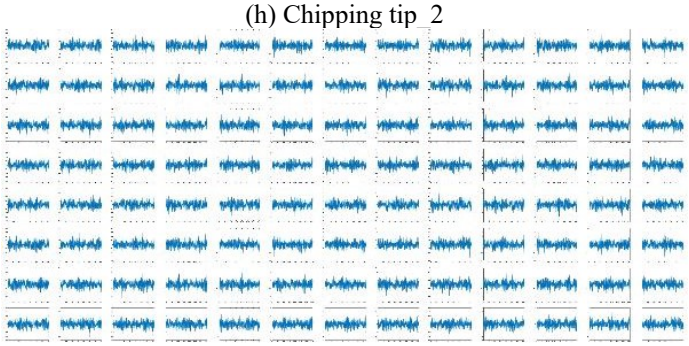
(f) Chipping tip_4



(g) Chipping tip_3



(h) Chipping tip_2



(i) Chipping tip_1 (most severe)

Figure 5 Vibration signal samples of different gear health conditions.

Case Study - 3,600 sampling points with varying training data size

104 vibration signals are generated for each gear condition. In the case studies, a portion of the signals are randomly selected as training data while the rest serves as validation data. To demonstrate the performance of the proposed approach towards various data sizes, the size of the training dataset ranges from 80% (83 training data per condition, 83*9 data in total) to 2% (2 training data per condition, 2*9 data in total) of all the 104 signals for each health condition.

Table 2 Classification results (3,600 sampling points)

Method / Training data	Transfer learning Accuracy (%)	Local CNN Accuracy (%)	AFS-SVM Accuracy (%)
80% (83 per condition)	Mean: 100	Mean: 97.57	Mean: 87.48
60% (62 per condition)	Mean: 100	Mean: 80.74	Mean: 87.72
40% (42 per condition)	Mean: 100	Mean: 76.63	Mean: 86.67
20% (21 per condition)	Mean: 99.92	Mean: 69.69	Mean: 86.24

10% (10 per condition)	Mean: 99.41	Mean: 55.82	Mean: 83.83
5% (5 per condition)	Mean: 94.90	Mean: 44.11	Mean: 79.89
2% (2 per condition)	Mean: 72.22	Mean: 27.99	Mean: 62.44

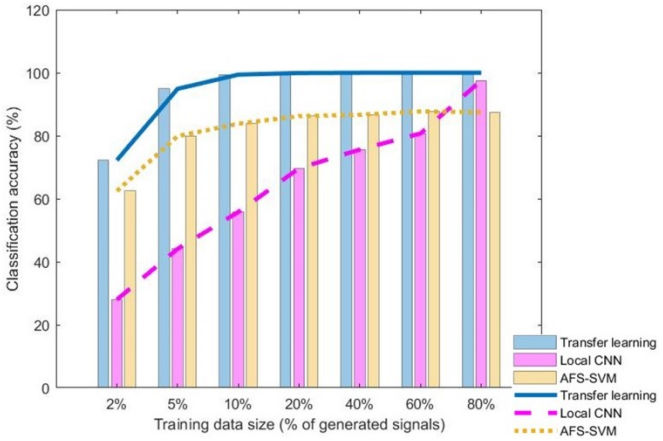


Figure 6 Classification results of the three methods when training data size varies.

Table 2 shows the classification results where the mean accuracy is the average of 5 training attempts. The classification accuracy is the ratio of the correctly classified validation data to the total validation dataset. As illustrated in Fig. 6, the proposed transfer learning approach has the best classification accuracy for all types of data size. Even if only 5 vibration signals per condition are selected for training, the proposed approach can achieve an extraordinary 94.90% classification accuracy, which further escalates to 99%-100% when 10% and more training data are used. On the other hand, while the performance of AFS-SVM reaches the plateau (showing only minimal increments) after 20% date is used for training, the classification accuracy of local CNN gradually increases with data size from 27.99% to 97.57% and surpasses AFS-SVM eventually when 80% data is used for training, indicating the significance of the size of training data to properly train a neural network. Although the data size greatly affects the performance of a CNN in general senses, the proposed transfer learning architecture still exhibits high classification accuracy because only one fully connected stage needs to be trained locally, which notably lowers the standard of the data required by a CNN in terms of achieving satisfactory outcome. Figure 7 shows the convergent histories (mini-batch accuracy) of the proposed approach and local CNN when 5% data is used for training. As can be seen from the comparisons, transfer learning gradually converges in accuracy while local CNN inclines to 'random walk' due to insufficient data. Compared with AFS-SVM, the proposed approach not only excels in performance, but also requires no pre-processing effort, angel-frequency analysis in this case, which makes the proposed approach more unbiased in feature extraction and readily applicable to other pattern recognition practices. The proposed approach also shows satisfactory outcomes it the regard of robustness. As demonstrated in Fig. 8, it has the smallest variance in all cases, while the performance of the under-trained local CNN oscillates the most.

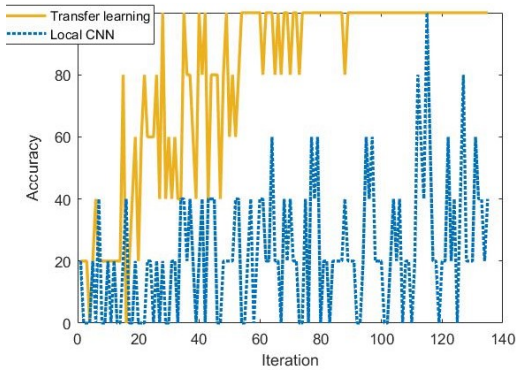


Figure 7 Convergent histories of transfer learning and local CNN for 5% training data.

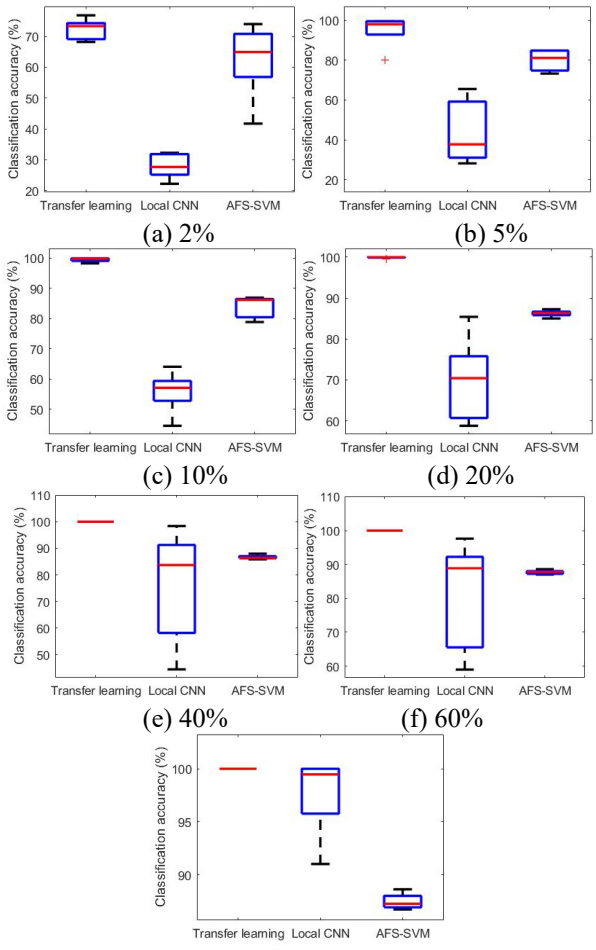


Figure 8 Box plots of classification results of the three methods when training data size varies.

The transferred stages of the proposed architecture attend to extract the high-level abstract features of the input that cannot be recognized otherwise, even if the input is different from that of the previous task. Figure 9 gives an example of such procedure by showing the feature maps generated in each convolution layer by the proposed architecture when it is used to classify a gearbox vibration signal. It is seen that the abstraction level of the input image continuously escalates from

the 1st feature map to the 5th feature map. In general, the number of convolution stages equipped is correlated with the level of abstraction the features can be represented in the CNN. As demonstrated in this case study, the base architecture is indeed transferable towards gear fault pattern recognition tasks and the proposed approach performs well with raw image signal inputs, which indicates the transferred layers constructed in this study are generally applicable to represent useful features of an input image in high-level abstraction.

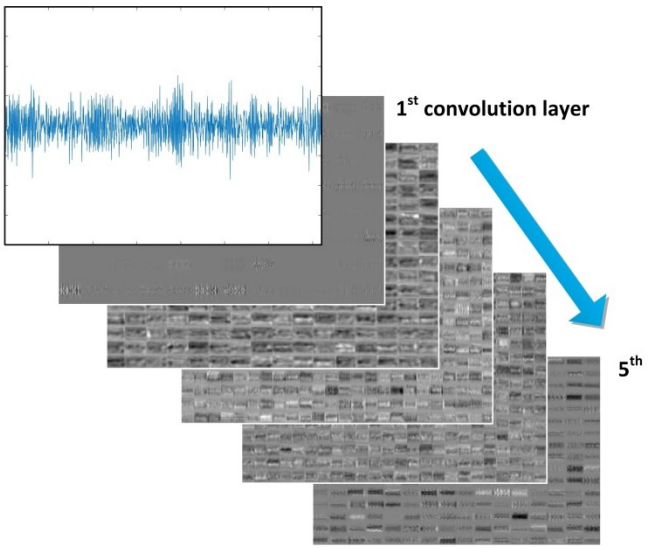


Figure 9 Feature maps extracted by 5 convolution layers of the proposed transfer learning approach.

CONCLUSIONS

In this paper, a deep convolutional neural network-based transfer learning approach is developed for pattern recognition. The proposed approach not only entertains adaptive feature extractions, but also requires only a small set of training data compared to other locally trained convolution neural networks. The proposed transfer learning architecture consists of two parts; the first part is constructed with a piece of a pre-trained deep neural network that serves to extract the features automatically from the input, the second part is a fully connected stage for classification which needs to be trained using experiment data. Experiment studies have been conducted using pre-processing free raw accelerometer data. The performance of the proposed approach is highlighted by varying the size of training data. The classification accuracies outperform other methods such as locally trained convolution neural network and angle-frequency analysis-based support vector machine by as much as 50%. The achieved outcome indicates that the proposed approach is not only remarkable in gear fault pattern recognition, but also has the potential to be readily applicable to other fault diagnosis or pattern recognition practices.

ACKNOWLEDGMENT

This research is supported by the National Science Foundation under grant IIS – 1741171.

REFERENCES

Abdeljaber, O., Avci, O., Kiranyaz, S., Gabbouj, M. and Inman, D.J., 2017. Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks. Journal of Sound and Vibration, 388, pp.154-170.

T. Fakhfakh, F. Chaari, M. Haddar, Numerical and experimental analysis of a gear system with teeth defects, International Journal of Advanced Manufacturing Technology, 25 (2005) 542-550.

B. Tang, W. Liu, T. Song, Wind turbine fault diagnosis based on Morlet wavelet transformation and Wigner-Ville distribution, Renewable Energy, 35 (2010) 2862-2866.

Ince, T., Kiranyaz, S., Eren, L., Askar, M. and Gabbouj, M., 2016. Real-time motor fault detection by 1-D convolutional neural networks. IEEE Transactions on Industrial Electronics, 63(11), pp.7067-7075.

Kim, Y., 2014. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

Li, D.Z., Wang, W. and Ismail, F., 2015. An enhanced bispectrum technique with auxiliary frequency injection for induction motor health condition monitoring. IEEE Transactions on Instrumentation and Measurement, 64(10), pp.2679-2687.

Li, C., Sanchez, R.V., Zurita, G., Cerrada, M., Cabrera, D. and Vásquez, R.E., 2016. Gearbox fault diagnosis based on deep random forest fusion of acoustic and vibratory signals. Mechanical Systems and Signal Processing, 76, pp.283-293.

LeCun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E. and Jackel, L.D., 1990. Handwritten digit recognition with a back-propagation network. In Advances in neural information processing systems (pp. 396-404).

Lu, Y., Tang, J. and Luo, H., 2012. Wind turbine gearbox fault detection using multiple sensors with features level data fusion. Journal of Engineering for Gas Turbines and Power, 134(4), p.042501.

F.P.G. Márquez, A.M. Tobias, J.M.P. Pérez, M. Papaelias, Condition monitoring of wind turbines: Techniques and methods, Renewable Energy, 46 (2012) 169-178.

Parey, A. and Pachori, R.B., 2012. Variable cosine windowing of intrinsic mode functions: Application to gear fault diagnosis. Measurement, 45(3), pp.415-426.

Shie, C.K., Chuang, C.H., Chou, C.N., Wu, M.H. and Chang, E.Y., 2015, August. Transfer representation learning for medical image analysis. In Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE (pp. 711-714). IEEE.

Sutskever, I., Martens, J., Dahl, G. and Hinton, G., 2013, February. On the importance of initialization and momentum in deep learning. In International conference on machine learning (pp. 1139-1147).

Saravanan, N. and Ramachandran, K.I., 2010. Incipient gear box fault diagnosis using discrete wavelet transform (DWT) for feature extraction and classification using artificial

neural network (ANN). Expert Systems with Applications, 37(6), pp.4168-4181.

Van den Oord, A., Dieleman, S. and Schrauwen, B., 2013. Deep content-based music recommendation. In Advances in neural information processing systems (pp. 2643-2651).

W. Wen, Z. Fan, D. Karg, W. Cheng, Rolling element bearing fault diagnosis based on multiscale general fractal features, Shock and Vibration, 2015 (2015).

Weimer, D., Scholz-Reiter, B. and Shpitalni, M., 2016. Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection. CIRP Annals-Manufacturing Technology, 65(1), pp.417-420.

R. Yan, R.X. Gao, X. Chen, Wavelets for fault diagnosis of rotary machines: a review with applications, Signal Processing, 96 (2014) 1-15.

Zhang, C., Sun, J.H. and Tan, K.C., 2015, October. Deep belief networks ensemble with multi-objective optimization for failure diagnosis. In Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on (pp. 32-37). IEEE.

Zhang, S. and Tang, J., 2018. Integrating angle-frequency domain synchronous averaging technique with feature extraction for gear fault diagnosis. Mechanical Systems and Signal Processing, 99, pp.711-729.

Zhou, W., Habetler, T.G. and Harley, R.G., 2008. Bearing fault detection via stator current noise cancellation and statistical control. IEEE Transactions on Industrial Electronics, 55(12), pp.4260-4269.