curr2vib: Modality Embedding Translation for Broken-Rotor Bar Detection

Amirhossein Berenji[0000−0003−3720−3015], Zahra Taghiyarrenani[0000−0002−1759−8593], and Sławomir Nowaczyk[0000−0002−7796−5201]

Center for Applied Intelligence Systems Research, Halmstad University, Sweden
{amirhossein.berenji}@hh.se

Abstract. Recently and due to the advances in sensor technology and Internet-of-Things, the operation of machinery can be monitored, using a higher number of sources and modalities. In this study, we demonstrate that Multi-Modal Translation is capable of transferring knowledge from a modality with higher level of applicability (more usefulness to solve an specific task) but lower level of accessibility (how easy and affordable it is to collect information from this modality) to another one with higher level of accessibility but lower level of applicability. Unlike the fusion of multiple modalities which requires all of the modalities to be available during the deployment stage, our proposed method depends only on the more accessible one; which results in the reduction of the costs regarding instrumentation equipment. The presented case study demonstrates that by the employment of the proposed method we are capable of replacing five acceleration sensors with three current sensors, while the classification accuracy is also increased by more than 1%.

Keywords: Induction Motor · Broken Rotor Bar · Fault Diagnosis · Predictive Maintenance · Contrastive pre-training · Multi-Modal Latent Translation

1 Introduction

Induction motors, mainly due to their affordable operational and maintenance costs alongside their reliability, are the most frequently used type of motors for industrial use cases [21]. The significance of their use in comparison to other equipment can be better understood by their share in energy consumption; they are estimated to consume up to 68% of the total energy in industrial sector, worldwide [2]. Therefore, optimizing the uptime of induction motors is of vital importance. Various faults can be expected to occur over the lifetime of this type of machinery. In particular, Broken Rotor Bar (BRB) problem − which is a partial crack, or a complete breakage, of the rotor bar − is categorized as one of the major faults of rotors [9]. Such an occurrence brings up different consequences, from increased power consumption [14] to unbalanced current in remaining rotor bars [9]. BRB can be detected by monitoring and analyzing a wide range of physical properties, with motor current and machinery vibrations considered to be among the most effective ones [7].

In recent years, enabled by the developments in the field of Internet-of-Things (IoT), we have witnessed an exponential growth in the amount of information that is being collected [18]. It has transformed the predictive maintenance (PdM) field, since the IoT is now the tool to collect, process and distribute large amounts of streaming data. The growth in the available information is not limited to the volume of data, but it also includes the variety of information being collected, in terms of different sources and sensor types [18, 13].

On the one hand, employing more modalities to solve any given problem is likely to improve the performance due to the inherent increase in the amount of available information. However, it is not always cost efficient, as the multi-source data is likely to include notable level of redundancy, potentially making the investment into additional equipment questionable. It has been shown that fusion of the data from different sources is not always helpful and extraction of high level features from key sources is often more important [18]. Moreover, multiple modalities are likely to vary from both accessibility (how easy it is to collect an arbitrary modality) and applicability (how useful this modality is to implement the in-hand task) point of view; therefore it can be logical to transfer knowledge from more applicable modality to more accessible modality to optimize the accessibility-applicability trade-off.

The contribution of this paper is an extension of our previous study [23], where we have compared vibrations against phase currents for BRB detection, and demonstrated that the former offer higher level of classification accuracy. Unfortunately, due to higher price and stricter requirements of proper sensors installation, vibrations is a less accessible modality in production environments. Building on these results, in this paper we demonstrate the possibility of employing modality embedding translation techniques to transfer knowledge from source (vibrations) to target (currents) modality in fault diagnosis case studies. We establish the effectiveness of this approach by showing that transferring the knowledge from vibrations to currents leads to increase in BRB detection accuracy.

Remaining of this paper is organized as follow: in Section 2, a number of previous studies preserving similarities to the present study are discussed. Afterwards, in Section 3, we introduce the proposed methods used in this study in details. Consecutively, in Section 4, experimental setups to evaluate the effectiveness of the proposed method is reported. Finally, yet importantly in Section 5 results from 4 is discussed and conclusions of this study is provided.

2 Related Works

2.1 Intelligent BRB Detection

Application of intelligent methods for detection and severity assessment of BRB problem have been studied in depth. For example, in [3], Empirical Mode Decomposition combined with an Adaptive Linear Network, alongside Feed Forward

Neural Network are employed to diagnose various types of defects in motor (including the BRB problem) based on motor current signal. Similarly, in [20], Wavelet Packet Decomposition is used to extract highly abstract set of features from stator current signals. The extracted feature set is next provided to a Multi-Layered Perceptron to classify the number of broken rotor bars in the induction motor. Besides stator current, machinery vibrations is also a great source of information for intelligent BRB detection. In [17], Sparse Representation is utilized to extract features from vibrations signals and these features are then used to evaluate the machinery health state, from BRB problem point of view. Likewise, in [19] the feature set extracted by Wavelet Discrete Transform is employed alongside K-Nearest Neighbors to not only detect complete BRB problem, but also to classify the severity of partial BRB. The methodology presented in that study is applicable to different levels of loading condition. Moreover, they had also considered the noise robustness of the proposed method.

Similar to the referenced studies, in this study we employ frequency domain signals of both vibrations three-phase currents to diagnose a squirrel cage induction motor, according to BRB problem.

2.2 Contrastive Representation Learning

When it comes to supervised learning of deep classification networks, cross-entropy loss is the most frequently used loss function [10]; alternatively, we may consider extraction of a feature set with optimum separability of classes as the objective of a learning process. A set of strategies known as Contrastive Representation Learning (CRL) are concerned with the construction of feature space, where different classes are sufficiently separable. CRL can be defined as learning by comparing the data [12]. Taking advantage of CRL strategies, one can be able to unlock higher level of classification accuracy, when compared with conventional baselines. For example, in [23], one step CRL-based pre-training turned out to be noticeably more effective for BRB classification. Moreover, the application of CRL-based pre-training is not limited to only classification tasks; in [15] contrastive pre-training is employed to learn de-noised sequence representations in both language and language-vision domains, based on self-supervised approaches. Similarly in [26], contrastive pre-training is utilized for event extraction in an unsupervised manner.

In our previous work [23], we showed that the application of CRL-based pre-training is an effective approach to overcome loading variation problem; therefore, in the presented work we also use this technique.

2.3 Multi-Source Fault Diagnosis

With the advances in IoT and sensors technology, information from more diverse sources is available. This has resulted in the application of Multi-Modal, or Multi-Source, techniques to PdM use cases. For instance, in [25], the traditional fusion of Multi-Source information is replaced by considering multiple sensors as different channels of the input fed to a Convolutional Neural Network.

This network is used to diagnose bearing faults, given time-domain vibration signals collected from three different locations. Moreover, in [1] a novel Hybrid Deep Neural Network is used to firstly extract two sets of features, temporal and spatial, respectively using Long Short Term Memory (LSTM) and Convolutional Neural Network (CNN) branches; subsequently, a fully-connected network is employed to fusion these two sets of features. The proposed architecture is used for remaining useful life estimation problem. Although fusion is beneficial in most cases, however, it is not always the best approach to take; mainly because of redundancy in multi-source datasets, or the added noise that comes from additional sensors. Therefore, a set of techniques is concerned with the maximization of the similarity over the representations derived from different modalities, or sources. As an example, in [16] a Deep Coupling Autoencoder is used to derive a joint representation from vibration and acoustic emission signals to capture the correlation between these two different modalities. The referenced methodology is shown to provide superior performance in comparison with traditional approaches, in bearing and gearbox fault diagnosis case studies.

Multi-Modal Translation, defined as the task to transfer or translate knowledge from a source modality to a target one [22], enables one to learn a mapping from a source modality to a target one. Multi-Modal Translation includes variety of applications, such as Image Captioning [8] (generation of a textual representation from an image) and Multi-Modal Speech synthesis [22] (generating audio given its textual representation). It is worth mentioning that Multi-Modal translation where the target modality is high-dimensional can get extremely challenging; one way to respond to this challenge is translating to a low-dimensional representation of the target modality containing higher level of semantic information in comparison with the input belonging to the source modality [27]. Taking this approach also saves the need to re-learn the latent space representation from its reconstructed version; making the implementation of consequent tasks, such as classification, easier.

3 Method

We propose a method that is based on Hybrid Classification, i.e., utilizing contrastive pre-training to derive the low-dimensional representation of the target modality. That embedding is then reconstructed, using a Pseudo-Autoencoder for Modality Embedding Translation, directly from the source modality. The whole process of extraction of low-dimensional representation and implementation of the modality embedding translation is demonstrated in Figure 1. Finally, in Section 3.3, we present the Centered Kernel Alignment which we use to highlight the similarity of representations learned by different networks.

3.1 Hybrid Classification

Siamese neural networks are one way to implement a contrastive pre-training. As the name suggests, a Siamese network is made of two exactly identical networks;

not only using the same architecture, but also sharing parameters. During its training process, the network is fed with positive pairs (both instances belong to same class) and negative pairs (instances belong to different classes). It is trained to aggregate all the observations sharing a class in the same region of the latent feature space it reconstructs (embedding); and simultaneously, to project observations from different classes to separate regions. Different options are available to train a Siamese network, including Contrastive Loss, defined as:

$$ContrastiveLoss = (1 - Y)D_w^2 + (Y)\frac{1}{2}(\max\{0, m - D_w\})^2, \tag{1}$$

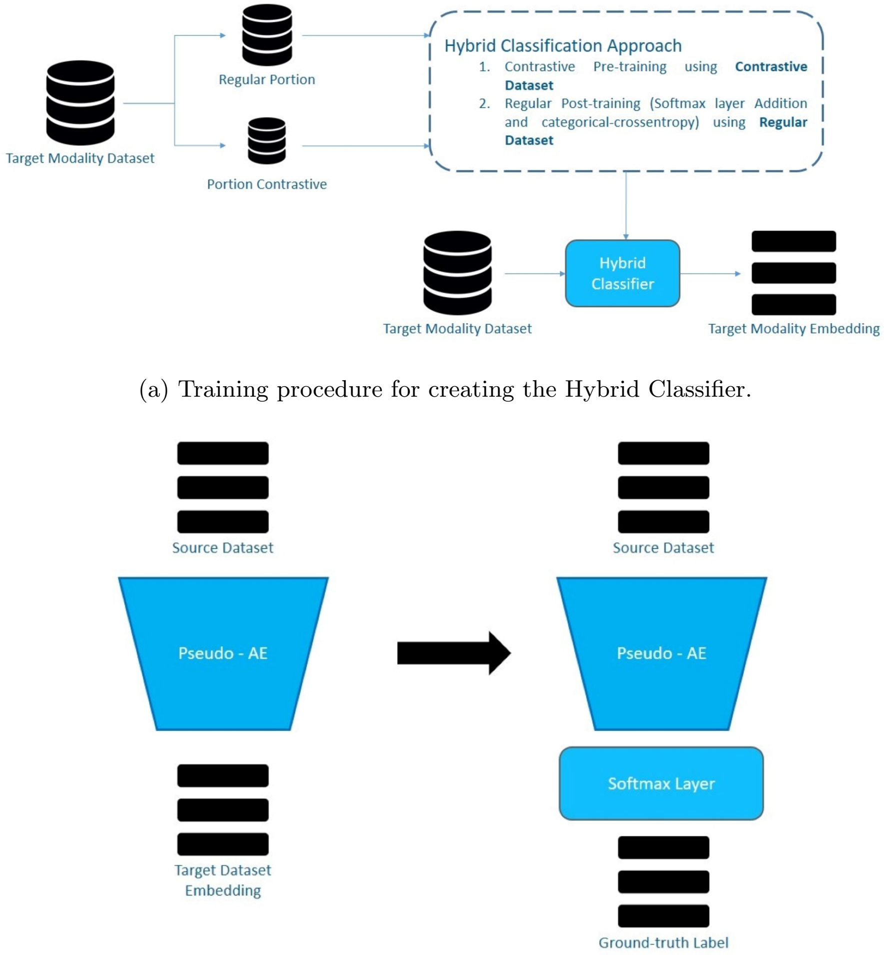
where Y is the label of a given pair, either 0 (for negative pairs) or 1 (for positive pairs), D_w is the similarity of the embedding of the observations in a pair and the m is the margin used to set a base value for the desired distance between negative pairs.

As mentioned earlier, access to a low-dimensional representation of the target modality is essential for the modality embedding translation task. In our previous work, we demonstrated that the application of contrastive pre-training is capable of improving the classification accuracy [23]. We use the same approach here, by first training a hybrid classifier and then re-using the low-dimensional representation created this way to train a Pseudo-AE network. Training the hybrid classifier involves two steps; first, we train a feature extractor network using contrastive learning approaches; second, a softmax layer is added to the feature extractor and the whole network is trained as a classifier. It is noteworthy that we divide the training dataset into two distinct portions, Contrastive and Regular. They are used during pre-training and actual training, respectively. The process of training the hybrid classifier and extracting the representation (embedding) is demonstrated in the Figure 1a.

3.2 Pseudo-AE for Modality Embedding Translation

Modality embedding translation is implementable using different methods, including an Autoencoder-like network, pseudo-AE for short. Such a network can be used to learn a mapping from source modality to a lower dimensional representation of the target modality. Autoencoders are networks capable of reconstructing a given input at its output, with the constraint of learning a lower dimensional representation of the input in its bottle-neck. Similarly, a pseudo-AE can be defined as a network capable of reconstructing an arbitrary but somehow related representation from a given input. Taking such an approach, we are able to reconstruct a representation, originally extracted from target modality, using only the source modality. A pseudo-AE can be trained using a similarity maximizing loss function, such as a Mean Squared Error.

In our previous work [23], we showed that vibrations offer a significantly higher classification performance, compared to current. On the other hand, the collection of multi-point vibrations from an induction motor is far more challenging compared with three-phase currents; in most cases, it requires invasive


Target Modality Dataset
Regular Portion
Portion Contrastive
Hybrid Classification Approach
1. Contrastive Pre-training using Contrastive Dataset
2. Regular Post-training (Softmax layer Addition and categorical-crossentropy) using Regular Dataset
Target Modality Dataset
Hybrid Classifier
Target Modality Embedding
Source Dataset
Pseudo - AE
Target Dataset Embedding
Source Dataset
Pseudo - AE
Softmax Layer
Ground-truth Label

(a) Training procedure for creating the Hybrid Classifier.

(b) Pseudo-AE Training and Post-training Procedure

Fig. 1: Visual Demonstration of the Proposed Method

measurements which do not suit practical online monitoring use cases. Moreover, current sensors are likely to be more affordable in comparison with their vibration counterparts. Last but not least, in the case study presented here, by taking advantage of the modality embedding translation technique, we would be able to decrease the number of required sensors from 5 (number of vibrometers) to 3 (number of current sensors), resulting in a more affordable technical infrastructures for data collection and storage.

To perform the modality embedding translation, we assume that we have access to synchronously measured signals from both modalities. Moreover, we also assume that we have access to the corresponding superior (task-specific) embeddings of the target modality (vibrations), for every observation of the source modality. In section 3.1, we have explained the procedure used to extract such superior embedding.

Using the Pseudo-AE network, we are able to learn a mapping from three phase currents FFT spectra towards the latent space of vibrations embedding. Having access to such a mapping, we will be able to reconstruct the corresponding vibration embedding, given an arbitrary observation in the three-phase current spectra. Once the mapping is learned, we are adding a softmax layer on top of the Pseudo-AE network and post-training it – utilizing Categorical Cross-Entropy loss function. This way, the network can be used for induction motor health state diagnosis, from BRB point of view, based only on the currents input data. Having the currents to vibrations embedding mapping learned sufficiently well, we are able to improve the performance of current-only dependent BRB detection classifier beyond what is possible by learning directly from raw data.

3.3 CKA for Representation Similarity Comparison

The effectiveness of a network in the fulfillment of a modality translation task, can be done by comparing the representations learned by the network at each layer. In the modality embedding translation task, an ideal translator should have representations similar to the ones from the source modality network in the early layers, while the final layers should be more similar to those of a network trained on the target modality. This way, we can make sure that a mapping from the source modality to the desired subspace of the target modality is learned well.

A number of techniques from a field known as Representational Similarity can be used to capture and quantify the similarity between two arbitrary embeddings. Among various proposed metrics, all possess different advantages and disadvantages; however, Centered Kernel Alignment (CKA) is considered as the current state of the art [4]. CKA not only enables measuring similarity between representations derived by different layers of the same network, but is also capable of quantifying the similarity between representations at different layers of different networks [11].

CKA mainly relies on the idea that the similarity of two sets of representations can be measured by calculating the similarity between every pair of examples in each set separably and comparing the similarity structures. Consider X and Y as two matrices including representations derived from n examples. Dot product can be used to evaluate the level of similarity between the representations, as demonstrated in the Equation 2:

$$\langle vec(XX^T), vec(YY^T)\rangle = tr(XX^TYY^T) = \|Y^TX\|_F^2 \tag{2}$$

Assuming that X and Y are centered, it implies Equation 3:

$$\frac{1}{(n-1)^2} tr(XX^TYY^T) = \|cov(X^T, Y^T)\|_F^2 \tag{3}$$

By employment of the Hilbert-Schmidt Independence Criterion [6], Equations 2 and 3 can be generalized to inner products from kernel Hilbert spaces; moreover, squared Forbenius norm of the cross-covariance matrix turns into the squared Hilbert-Schmidt norm of the cross-covariance operator [11]. Considering $K_{i_j} = k(x_i, y_j)$ and $L_{i_j} = l(x_i, y_j)$ where k and l are two kernels, empirical estimator of HSIC can be defined as:

$$HSIC(K, L) = \frac{1}{(n-1)^2} tr(KHLH), \tag{4}$$

where H is the centering matrix $H_n = I_n - \frac{1}{n}11^T$. A normalization step can make it invariant to isotropic scaling $S(X, Y) \neq S(\alpha X, \beta Y)$ for all $\alpha, \beta \in \mathbb{R}^+$. Normalized HSIC is known as Centered Kernel Alignment:

$$CKA(K, L) = \frac{HSIC(K, L)}{\sqrt{HSIC(K, K)HSIC(L, L)}}) \tag{5}$$

In this work, we employ CKA to compare the representations derived by the vibration embedding modality translator network, given corresponding current observation (curr2vib for short). This way, we would be able to investigate the goodness of the mapping learned by modality latent space translator in transforming input from the source modality (currents frequency spectra) to the latent space originally derived from the target modality (vibrations frequency spectra).

4 Experiments

Three different experiments are carried out in this study. This section starts with introducing the dataset and the pre-processing procedure. Next, in Section 4.2, we present results of training hybrid classifiers directly on the raw data of different modalities. This is followed, in Section 4.3, with the demonstration of improvements provided by training the Pseudo-AE model. Last but not least, in Section 4.4, we employ CKA to compare the similarity of representations derived from different networks and modalities.

4.1 Dataset and Pre-processing Procedure

Data is the essential ingredient of every data-driven study and ours is not an exception. We took advantage of the experimental dataset for detecting and diagnosing rotor broken bar in a three-phase induction motor [24] to carry out our case study. This dataset provides us with both electrical (phase voltages and currents) and mechanical (multi point vibrations) signals. Five different states from BRB problem point of view (from zero to four broken rotor bars), over eight different levels of mechanical torque as loading conditions are available in this dataset. In this study, we consider four distinct levels of mechanical torque to consider the various loading condition, corresponding to 12.5%, 50%, 62.5% and 100% of nominal load. The classification problem we take into account in this study is to predict the number of broken rotor bars (from zero to four ones), over a balanced training and testing dataset, from both loading conditions and number of broken rotor bars.

The original time-domain signals are split into shorter ones, using 1024 and 6667 points-long windows for vibrations and currents signals, respectively. Moreover, Fast Fourier Transform (FFT) is employed to map time domain signals to frequency domain signals, resulting in 512 and 3333 points long vibrations and currents signals, in frequency domains. The 5 point vibration signals collected from different location and three phase currents are then concatenated horizontally to form 2560 and 9999 points long signals for vibrations and currents modalities.

The whole dataset is randomly split into training (75%) and testing (25%) sets. In addition to that, min-max scaling is used to normalize the feature space.The fact that by the application of min-max scaling every frequency components in frequency spectrum is regarded as an individual feature, makes this scaling strategy an optimal choice for the problem in hand.

4.2 Hybrid Classification by Contrastive Pre-training

As mentioned in Section 3.1, we employ contrastive pre-training to train a hybrid vibrations classifier. This classification network is used to extract a 64-dimensional representation of the original vibrations input (2560 long space); we believe it is a reasonable size to compress the original 2560-dimensional space. The referenced low dimensional representations are derived from the last layer of the classification network, excluding the softmax layer (since this layer is expected to contain the feature set with the highest level of abstraction). This latent subspace would be later used to learn a mapping from currents to the vibrations embedding latent subspace. This process is discussed in more detail in Section 4.3.

To train the hybrid vibration classification network, we start with splitting the training vibrations dataset into regular and contrastive portions, with a ratio of 25% contrastive to regular. Afterwards, 10 pairs are created per observation in the contrastive portion of the training dataset, consisting of five positive and five negative ones. These pairs are used to conduct a contrastive pre-training process for the feature extractor of the hybrid vibration classification network. The feature extractor utilizes a multi-layered perceptron architecture with 2560-1280-640-580-512-256-128-64 neurons per layer. All the layers use hyperbolic tangent as the activation function. During the contrastive pre-training Contrastive Loss is used as the loss function, number of epochs is 100 and learning rate is 0.00001. It is worth mentioning that, the choice of learning rate and epoch, not only for this specific experiment but also for all the experiments carried out in this paper, is done to 1) keep training process properly smooth by using relatively low learning rate and 2) achieving the best possible model parameters by the employment of surpass number of epochs. Having the pre-training process finished, a softmax layer is added to the feature extractor to form a classification network and the remaining 75% portion of the training data is used to post-train the classification network. During the post-training process Categorical Cross-entropy – as the most frequent choice of loss function in multi-class classification problems– is used as the loss function, number of epochs is 400 and learning rate is 0.000001. Having the whole network post-trained, the latent space required to conduct the modality embedding translation process is now extractable. This can be done by extracting the representations available in the last layer of the classification network before softmax layer, corresponding to all the vibrations observations available in the training dataset.

Similarly, a hybrid classifier utilizing currents as the input is implemented, also using hyperbolic tangent as the activation function and following 9999-7500-6000-4500-3000-1500-750-500-250-50 architecture. For the contrastive pre-training of this network, four pairs are created per each observation in the contrastive portion. Moreover, the choices of hyperparameters such as loss function, number of epochs and learning rate for both contrastive pre-training and categorical cross-entropy post-training are kept the same as the ones used for hybrid vibrations classifier. To account for the randomness, experiments are conducted 5 times and mean of the classification accuracy is used as the metric to evaluate the performance, as it is the most frequent metric to evalueate the performance of a classifier in balanced classification problems. Results regarding the classification performance of hybrid classifiers on the testing dataset are shown in the first two rows of Table 1. As it is clearly observable, both modalities are offering +90% accuracy in classification of the BRB detection problem. Additionally, vibrations are offering significantly higher performance in comparison with current.

Table 1: Average (AVG) and Standard Deviation (STD) of classification accuracy of each network.

Network	**AVG**	**STD**
Currents	0.9096	0.0070
Vibrations	0.9769	0.0033
curr2vib	0.9204	0.0041

4.3 Modality Embedding Translation using Pseudo-AE

Different approaches are available for Modality Embedding Translation; in this study, we employ a Pseudo-AE network, utilizing a Multi-Layered Perceptron with the architecture of 9999-6000-3000-750-250-150-50-64. In the proposed architecture, the last layer before the output is kept to a lower-dimensional compared with the output to preserve the constraint of learning the lowest dimensional representation in the middle layers of network. Moreover, in all the neurons of this network, hyperbolic tangent is used as the activation function. Besides, Logarithmic Mean Squared Error, 100 and 0.0000001 are used as the loss function, number of epochs and learning rate during the Seq2Seq reconstruction training of the Pseudo-AE network. Once the mapping from current to vibrations embedding is learned, we need a post-training process to make a classification network out of the Pseudo-AE network. This is done by addition of a softmax layer to the Pseudo-AE network and employment of Categorical Cross-entropy as the loss function of the whole classification network. Categorical Cross-Entropy is chosen, as it is the most frequent option to use for multi-class classification tasks. Moreover, 0.0000001 and 2000 are employed as the learning rate and number of epochs for the implementation of the post-training process. It is worth mentioning that the whole training dataset is employed to learn the mapping from current to vibrations embedding, however, similar to the Section 3.1 only 75% of the training dataset is used during the post-training process.

In the final row of Table 1, the performance of the proposed method (curr2vib) is presented. When compared with the performance of current-based hybrid classifier, we managed to increase the classification performance by more than 1% due to taking advantage of the modality embedding translation technique and the vibrations embedding. Moreover, lower STD of the curr2vib classifier in comparison with hybrid current classifier demonstrates the higher level of stability of this approach.

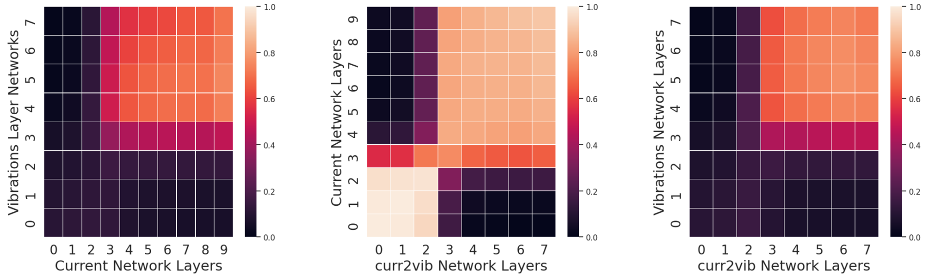
4.4 Using CKA to Evaluate the Effectiveness of Pseudo-AE to Translate Modality Embedding

Comparison of the representations learned by neural networks at different layers can be used to quantify the similarities between the set of features learned at each layer. In particular, in this study we employ CKA – current state of the art tool to investigate the similarities between representations learned by different networks at different layers – to evaluate the effectiveness of our proposed method in the extraction of features similar to the target modality, given the source modality as the input. Representations extracted for these comparisons are derived from observations included in the testing dataset. Moreover, we employed the implementations[1] provided by authors of [11].

Using the heatmaps present in Figure 2, we are able to compare the similarity of the representations learned by different models, pairwise. The color which has

[1] https://cka-similarity.github.io/

filled the cells of these heatmaps is an indicator of the similarity scores, measured using CKA technique. In the Figure 2a, the similarity between representations extracted from Vibrations Classifier and Currents Classifier is demonstrated. As it is expected, representations at the initial layers are not similar, since the two networks are fed with information belonging to different modalities as input. Moreover, significant increase in the similarities of the representations is observed among those extracted from fourth and further layers; clearly, in these layers both networks are able to extract related, highly-abstract feature sets. Besides that, in Figure 2b, representations extracted from Currents Classifier, and curr2vib Classifier are compared. Unlike the previous figure, in this figure, a noticeable level of similarity is found between first three layers of the networks; this makes intuitive sense, as they are fed with identical inputs. Moreover, we experience significant reduction in the similarity from the fourth layer, showing that the features learned by two networks in these layers differ, which is the reason for the gap between these two networks in the classification of BRB problem.



(a) Hybrid Vibrations Classifier and Hybrid Currents Classifier

(b) Hybrid Current Classifier and curr2vib Classifier

(c) Hybrid Vibrations Classifier and curr2vib Classifier

Fig. 2: Plots of CKA values of pairwise comparisons of the three networks.

Last but not least, in Figure 2c, representations from Hybrid Vibrations Classifier and curr2vib Classifier are compared. Again, as in Figure 2a, the initial layers are not similar as the inputs belong to different modalities. Moreover, significant increase in the similarities is noticeable from the fourth layers to the end of the networks; this increase happens in the similar region where the Figure 2b experienced the drop in the similarity, demonstrating that the transformation of the representations available in the fourth layer to the rest of the network extracted by the curr2vib Classifier is making them more similar to the ones extracted by the Hybrid Vibrations Classifier. Being more similar to the representations extracted from Hybrid Vibrations Classifier, rather than the ones extracted from Hybrid Currents Classifier, can be considered as the reason behind the improvement in the classification performance.

5 Discussion and Conclusion

Comparison of the similarity of the representations learned by source-only based classifier (currents classifier), target-only based classifier (vibrations classifier) and Multi-Modal Embedding Translation classifier (curr2vib classifier), showed that the proposed method is capable of learning, using only the weaker source modality, representations similar to those coming from the stronger target modality. Therefore, this approach exploits some of the principles underlying Knowledge Distillation; a set of techniques and approaches to transfer what a superior model (teacher), or ensemble of them, has learned, to an inferior one (student) [5]. Knowledge Distillation is mainly concerned with improving the performance of a model with the help of another model. According to the above definition, vibration classifier is the teacher model and the curr2vib is the student model; Moreover, as the teacher model in this study is kept non-trainable during the knowledge transfer process, curr2vib utilizes an offline distillation scheme.

This study applies Modality Embedding Translation – as a Multi-Modal approach – to transfer knowledge from source modality (with high classification performance but expensive to collect) to the target one (cheaper, but with lower performance). As shown in the case study investigated, employment of such strategy is capable of improving the performance, when compared against conventional approaches learning on raw data in target modality separately. Although both modalities are required during the training process, in the deployment stage only the target modality is needed; therefore this approach is considerably more affordable in comparison with sensor fusion. Using the proposed strategy, we are able to replace expensive instrumentation pieces of equipment with more affordable ones while the performance is kept within acceptable range. One limitation is that the implementation of the proposed method requires having access to synchronously measured signals from both modalities, which can be hard to provide. Although measuring signals from both modalities simultaneously tends to reduce the data collection time, however, it is not always cost efficient to record both modalities at the same time, as it would require data acquisition equipment with higher capacities. The future works on this topic can be directed towards development of strategies to eliminate this constraint.

6 Acknowledgments

This work was partially supported by Vinnova and by CHIST-ERA grant CHIST-ERA-19-XAI-012 funded by Swedish Research Council.

References

1. Al-Dulaimi, A., Zabihi, S., Asif, A., Mohammadi, A.: A multimodal and hybrid deep neural network model for remaining useful life estimation. Computers in Industry **108**, 186–196 (2019)

2. Beleiu, H.G., Maier, V., Pavel, S.G., Birou, I., Pică, C.S., Dărab, P.C.: Harmonics consequences on drive systems with induction motor. Applied Sciences **10**(4), 1528 (2020)
3. Camarena-Martinez, D., Valtierra-Rodriguez, M., Garcia-Perez, A., Osornio-Rios, R.A., Romero-Troncoso, R.d.J.: Empirical mode decomposition and neural networks on fpga for fault diagnosis in induction motors. The Scientific World Journal **2014** (2014)
4. Csiszárik, A., Kőrösi-Szabó, P., Matszangosz, Á., Papp, G., Varga, D.: Similarity and matching of neural network representations. Advances in Neural Information Processing Systems **34**, 5656–5668 (2021)
5. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision **129**(6), 1789–1819 (2021)
6. Gretton, A., Bousquet, O., Smola, A., Schölkopf, B.: Measuring statistical dependence with hilbert-schmidt norms. In: International conference on algorithmic learning theory. pp. 63–77. Springer (2005)
7. Gritli, Y., Di Tommaso, A., Filippetti, F., Miceli, R., Rossi, C., Chatti, A.: Investigation of motor current signature and vibration analysis for diagnosing rotor broken bars in double cage induction motors. In: International Symposium on Power Electronics Power Electronics, Electrical Drives, Automation and Motion. pp. 1360–1365. IEEE (2012)
8. Hossain, M.Z., Sohel, F., Shiratuddin, M.F., Laga, H.: A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR) **51**(6), 1–36 (2019)
9. Kanović, Ž., Matić, D., Jeličić, Z., Rapaić, M., Jakovljević, B., Kapetina, M.: Induction motor broken rotor bar detection using vibration analysis—a case study. In: 2013 9th IEEE international symposium on diagnostics for electric machines, power electronics and drives (SDEMPED). pp. 64–68. IEEE (2013)
10. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in Neural Information Processing Systems **33**, 18661–18673 (2020)
11. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: International Conference on Machine Learning. pp. 3519–3529. PMLR (2019)
12. Le-Khac, P.H., Healy, G., Smeaton, A.F.: Contrastive representation learning: A framework and review. IEEE Access **8**, 193907–193934 (2020). https://doi.org/10.1109/ACCESS.2020.3031549
13. Lei, Y., Yang, B., Jiang, X., Jia, F., Li, N., Nandi, A.K.: Applications of machine learning to machine fault diagnosis: A review and roadmap. Mechanical Systems and Signal Processing **138**, 106587 (2020)
14. Lizarraga-Morales, R.A., Rodriguez-Donate, C., Cabal-Yepez, E., Lopez-Ramirez, M., Ledesma-Carrillo, L.M., Ferrucho-Alvarez, E.R.: Novel fpga-based methodology for early broken rotor bar detection and classification through homogeneity estimation. IEEE Transactions on Instrumentation and Measurement **66**(7), 1760–1769 (2017)
15. Luo, F., Yang, P., Li, S., Ren, X., Sun, X.: Capt: contrastive pre-training for learning denoised sequence representations. arXiv preprint arXiv:2010.06351 (2020)
16. Ma, M., Sun, C., Chen, X.: Deep coupling autoencoder for fault diagnosis with multimodal sensory data. IEEE Transactions on Industrial Informatics **14**(3), 1137–1145 (2018)

17. Morales-Perez, C., Rangel-Magdaleno, J., Peregrina-Barreto, H., Amezquita-Sanchez, J.P., Valtierra-Rodriguez, M.: Incipient broken rotor bar detection in induction motors using vibration signals and the orthogonal matching pursuit algorithm. IEEE Transactions on Instrumentation and Measurement **67**(9), 2058–2068 (2018)
18. Ran, Y., Zhou, X., Lin, P., Wen, Y., Deng, R.: A survey of predictive maintenance: Systems, purposes and approaches. arXiv preprint arXiv:1912.07383 (2019)
19. Rangel-Magdaleno, J., Peregrina-Barreto, H., Ramirez-Cortes, J., Morales-Caporal, R., Cruz-Vega, I.: Vibration analysis of partially damaged rotor bar in induction motor under different load condition using dwt. Shock and Vibration **2016** (2016)
20. Sadeghian, A., Ye, Z., Wu, B.: Online detection of broken rotor bars in induction motors by wavelet packet decomposition and artificial neural networks. IEEE Transactions on Instrumentation and Measurement **58**(7), 2253–2263 (2009)
21. Spyropoulos, D., Mitronikas, E., Dermatas, E.: Broken rotor bar fault diagnosis in induction motors using a goertzel algorithm. In: 2018 XIII International Conference on Electrical Machines (ICEM). pp. 1782–1788. IEEE (2018)
22. Summaira, J., Li, X., Shoib, A.M., Abdul, J.: A review on methods and applications in multimodal deep learning. arXiv preprint arXiv:2202.09195 (2022)
23. Taghiyarrenani, Z., Berenji, A.: An analysis of vibrations and currents for broken rotor bar detection in three-phase induction motors. In: PHM Society European Conference. vol. 7, pp. 43–48 (2022)
24. Treml, A.E., Flauzino, R.A., Suetake, M., Maciejewski, N.A.R.: Experimental database for detecting and diagnosing rotor broken bar in a three-phase induction motor. IEEE DataPort (2020)
25. Wang, J., Wang, D., Wang, X.: Fault diagnosis of industrial robots based on multi-sensor information fusion and 1d convolutional neural network. In: 2020 39th Chinese Control Conference (CCC). pp. 3087–3091. IEEE (2020)
26. Wang, Z., Wang, X., Han, X., Lin, Y., Hou, L., Liu, Z., Li, P., Li, J., Zhou, J.: CLEVE: Contrastive Pre-training for Event Extraction. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 6283–6297. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-long.491, https://aclanthology.org/2021.acl-long.491
27. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. Advances in neural information processing systems **30** (2017)