A Framework to Interpret Deep Learning-Based Health Management System with Human Interactions

Namkyoung Lee[1], Michael H. Azarian[1], and Michael G. Pecht[1]

[1] *Center for Advanced Life Cycle Engineering (CALCE), University of Maryland, College Park, MD 20742, USA*

nklee@umd.edu
mazarian@umd.edu
pecht@umd.edu

ABSTRACT

Deep learning has shown good performance in detecting a product's faults and estimating the remaining useful life of a product. However, it is hard to interpret deep learning-based health management systems because deep learning is often regarded as a black box. In order to make a maintenance decision based on the result of the management system, humans need to know how it gave the outcome. This study aims to develop a framework that utilizes human interactions during system development to understand the internal process of deep learning. The study will demonstrate the framework on bearing datasets.

1. PROBLEM STATEMENT

Deep learning is a type of machine learning algorithms that utilizes multiple layers of neural layers to classify or fit the trend of data. Compared to conventional machine learning algorithms such as a support vector machine and a decision tree, deep learning can learn features by itself and has more flexibility to fit data using many parameters. Because of these characteristics, deep learning performed better when solving complex problems.

Since detecting faults and predicting remaining useful life (RUL) of a product can be complex problems, deep learning has been used for diagnostics and prognostics for many products. Although deep learning has shown promising results, reasoning the results was difficult because deep learning is considered as a black box. However, in order to adopt deep learning-based diagnostics and prognostics in practice, the logic inside deep learning should be transparent so that humans can understand the outcome of the management system.

In order to reason deep learning, several methods have been addressed. One approach was to visualize self-learned features of a neural network (Simonyan, Vedaldi, and Zisserman, 2013; Dosovitskiy, Springenberg, and Brox, 2015). Since a neural network gives a result on the basis of the features, understanding the features can help comprehend the logic of the network.

Another approach to interpret deep learning was conducting sensitivity analysis on a deep learning model (Julian, Olden, and Donald, 2002). The authors investigated the weights that connect neurons inside the model. Through the analysis, humans can understand which inputs influence the output of the model most.

Although the aforementioned approaches showed the feasibility of understanding deep learning, a few steps still remain when the input of a deep learning model also needs interpretation, which is common in the data for diagnostics and prognostics. For example, vibrations collected nearby gears need processing to interpret the health state of the gear. Likewise, several steps that require human interactions are needed during the development phase to understand the model.

2. EXPECTED CONTRIBUTIONS

This study will provide a framework that provides methods to interpret a deep learning-based health management system with human involvement. Although the main purpose of this framework is to understand deep learning for diagnostics and prognostics, the framework also aids in validating and improving the performance of a deep learning model as a result. To be specific, the framework will provide three methods that will help understand the behavior of a deep learning model as shown in Figure 1.

The diagram in Figure 1. depicts the process of developing a deep learning model. Dashed lines in the diagram indicate the suggested methods that have human interactions.

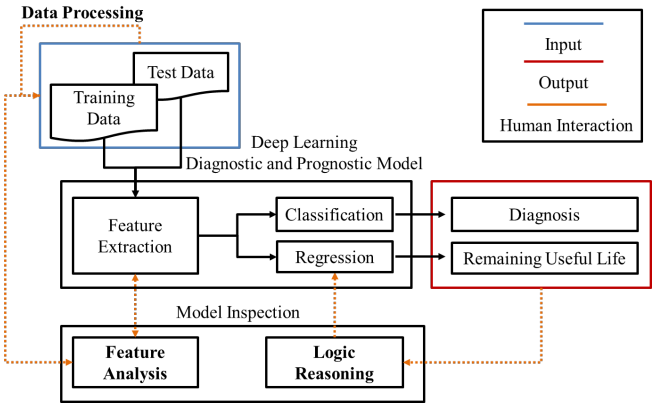

Figure 1. Human interactions on developing a deep learning diagnostic and prognostic model to improve the interpretability of the model.

First, this study will present the method to process raw input to better understand the meaning of the input. Especially, the method focused on processing spectral data to find features of interests using domain knowledge. This process can guide a deep learning model to learn certain features that have already known as good features for diagnostics and prognostics.

Second, the framework will visualize self-learned features by using an autoencoder as a deep learning architecture. Autoencoders can reconstruct inputs with self-learned features. The method will use this function to infer the characteristics that are extracted by features. Using this method, humans can inspect the self-learned features to determine whether these features are useful for diagnostics or prognostics.

Lastly, this study will provide a method to extract and simplify the logic inside a trained deep learning model. To extract the logic, sensitivity analysis will be conducted on the model. After the analysis, the logic will be transferred to a decision tree to make this logic transparent. This method enables humans to understand the behavior of a trained deep learning model. Therefore, humans can validate the logic of the model and get insights for diagnostics and prognostics.

To demonstrate the framework, the study will provide a case study that will guide people who want to understand and verify their deep learning-based diagnostic or prognostic model.

3. RESEARCH PLAN

In order to develop and demonstrate the framework, four tasks should be completed as follows:

1. Devise a framework that allows human involvement while developing a deep learning model for diagnostics and prognostics
2. Select a product and experimental datasets to apply the framework
3. Review literature to acquire domain knowledge about failure mechanisms of the targeted product, which will be used while applying the framework
4. Apply the framework on the targeted product and develop a deep learning-based prognostic model.

The research will be started by developing a framework that was introduced in the previous section. In order to prove the framework, a case study will be conducted using available prognostic and health management data. The data will be selected based on two constraints. First, the data should have enough experimental data to train a deep learning model. Second, the tested product should have already been researched to obtain domain knowledge about the product. After the data selection process, literature will be reviewed to acquire knowledge. Finally, the framework will be applied to the targeted product by developing a prognostic model using deep learning.

3.1. Work Performed

The framework to understand deep learning was already developed and explained in the previous section. To demonstrate the framework, a rolling-element bearing was selected as a target product because many experimental data on the bearing are available in public. Among them, bearing datasets used for PHM 2012 prognostic challenge (Nectoux et al., 2012) were used as a benchmark.

In general, failures of the bearings have been monitored through spectral analysis (Graney and Starry, 2011). The most commonly used precursors in spectral analysis were ball pass frequency outer race and ball pass frequency inner race. In addition, statistical features gathered from time domain, frequency domain, and time-frequency domain were also used as precursors in common (Xia et al., 2012).

These precursors were calculated from training data of the bearing datasets and the changes in the precursors were monitored. Among the tested precursors, Fourier transform results from raw vibration data showed good degradation trends as shown in Figure 2.

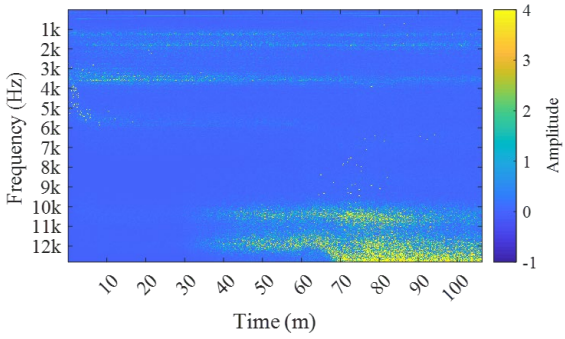

Figure 2. The results of Fourier transform from bearing2_2 dataset over time.

The graph in Figure 2. is a scalogram of horizontal vibrations collected from bearing2_2 dataset. The x-axis and y-axis represent the operation time of the bearing and frequency bands each. The color of the graph represents the amplitude of Fourier transform results. As the bearing degrades, the amplitudes of the high-frequency bands monotonically increased. Since the results showed good degradations trends, the results were used as input of a deep learning model.

To extract and visualize features from the input, a stacked autoencoder was trained. The size of the autoencoder was configured to get the root mean squared errors between inputs and outputs less than 10^{-3}. As a result, the trained autoencoder had double layers as shown in Figure 3.

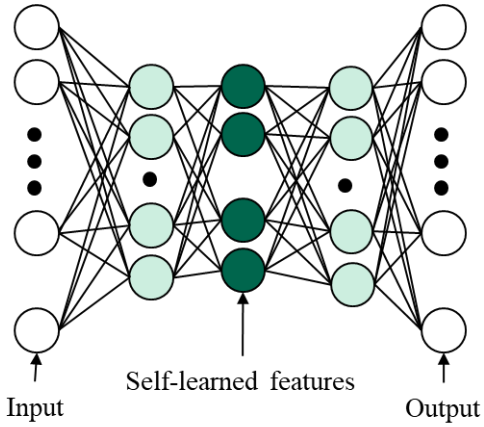

Figure 3. An architecture of a stacked autoencoder that was utilized to extract self-learned features from spectra data.

The autoencoder had five neural layers that have 176, 45, 8, 45, 176 neurons each. The autoencoder encodes a vector that has the size of 176 to 8 features, and then decodes the features to a 176 size vector again. To reason the characteristics of the 8 features, all bearing datasets were encoded and decoded and changes in input were compared with the changes of each feature. Through the analysis, it was confirmed that the features extracted by the autoencoder represent a bundle of frequency bands that are highly correlated.

To validate the fitness of the features for prognostics, the sensitivity analysis was conducted. Through this process, the most responsive features to predict the remaining useful life of bearings were the amplitude of high frequencies above 10 kHz for both horizontal and vertical vibrations. Since natural frequencies of bearing's components can generate frequencies above 2 kHz, the fitness of the features was verified.

3.2. Remaining Work

Among the three methods that were proposed in the framework, developing the method to extract logic from a deep learning model is in process. In particular, the approach to transform neural networks to decision trees needs to be developed. Due to the high complexity of a deep learning model, two problems need to be solved.

Transforming the entire structure of a trained model is not feasible and even if it is possible, understanding the logic is challenging. Therefore, the model should be simplified without losing much information.

Another issue is the limitation of sensitivity analysis on understanding the model. Since the result of the analysis gives only the expected responses of the model, inferred logic based on the result is not deterministic. In addition, the analysis has drawbacks in interpreting interactions between inputs. As the number of input increases, the computational power to interpret the interactions increases exponentially.

4. CONCLUSION

Interpretability of deep learning models is a big obstacle to embrace a deep learning-based health management system. To understand deep learning, this study presents a framework that includes human interactions while training a deep learning model. The framework will provide three methods that can be used to validate and improve a deep learning model. Among them, two methods were already developed and demonstrated on a developing prognostic model for rolling element bearings. However, the usage of the framework will not be limited to bearings because the methodologies provide general approaches to understand deep learning.

ACKNOWLEDGMENT

The authors also thank the Center for Advanced Life Cycle Engineering (CALCE) at the University of Maryland and its over 150 supporting companies.

REFERENCES

Dosovitskiy, A., Tobias Springenberg, J., and Brox, T. (2015). Learning to generate chairs with convolutional neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1538-1546.

Graney, B., and Starry, K. (2012), Rolling element bearing analysis. Materials Evaluation vol. 70.1, pp. 78–85.

Nectoux, P., Gouriveau, R., Medjaher, K., Ramasso, E., Chebel-Morello, B., Zerhouni, N., & Varnier, C. (2012). Pronostia: An experimental platform for bearings accelerated degradation tests. IEEE International Conference on Prognostics and Health Management, pp. 1–8.

Olden, J. D., and Jackson, D. A. (2002). Illuminating the "black box": a randomization approach for understanding variable contributions. Artificial neural networks. Ecological modeling, vol. 154(1-2), pp. 135-150.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualizing image

classification models and saliency maps. arXiv preprint arXiv:1312.6034.

Xia, Z., Xia, S., Wan, L., and Cai, S. (2012) Spectral Regression based Fault Feature Extraction for Bearing Accelerometer Sensor Signals. Sensors, vol. 12(10), pp. 13 694–13 719.

BIOGRAPHIES

Namkyoung Lee received the B.S. degree in electronics from Hanyang University, Seoul, Korea in 2015 and the M.S. degree in computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea in 2017. He is currently working toward Ph.D. degree in mechanical engineering from the University of Maryland, College Park, MD, USA. His research interests include data-driven prognostics and healthy management of mechanical systems, and diagnostics and prognostics of car components.

Michael H. Azarian received the B.S.E. degree in chemical engineering from Princeton University, Princeton, NJ, USA, and the M.E. and Ph.D. degrees in materials science and engineering from Carnegie Mellon University, Pittsburgh, PA, USA. He spent over 13 years in industry. He is a Research Scientist with the Center for Advanced Life Cycle Engineering, University of Maryland, College Park, MD, USA. His research interests include the analysis, detection, prediction, and prevention of failures in electronic and electromechanical products. He holds five U.S. patents.

Dr. Azarian is chair of the SAE G-19A Test Laboratory Standards Development Committee which is responsible for the AS6171 family of standards on detection of counterfeit electrical, electronic, and electromechanical parts. He also co-chairs the working group responsible for the IEEE 1624 standard on organizational reliability capability of suppliers of electronic products.

Michael G. Pecht received the B.S. degree in acoustics, M.S. degrees in electrical engineering and engineering mechanics, and Ph.D. degree in engineering mechanics from the University of Wisconsin at Madison, Madison, WI, USA, in 1976, 1978, 1979, and 1982, respectively. He is the Founder of the Center for Advanced Life Cycle Engineering, University of Maryland, College Park, MD, USA, where he is also a Chair Professor. He has been leading a research team in the area of prognostics.

Dr. Pecht is a Professional Engineer and a Fellow of the American Society of Mechanical Engineers. He was the recipient of the IEEE Undergraduate Teaching Award and the International Microelectronics Assembly and Packaging Society William D. Ashman Memorial Achievement Award for his contributions in electronics reliability analysis. He served as the Chief Editor of IEEE Transactions on Reliability for eight years and an Associate Editor for IEEE TRANSACTIONS on Components and Packaging Technologies.