The Seventh Asian-Pacific Symposium on Structural Reliability and Its Applications (APSSRA2020)
October 4–7 2020, Tokyo, Japan
T. Takada, I. Yoshida & T. Itoi (editors)

Interpretation of Deep Neural Network for Damage Pattern Classification Using Phase Plane

T. Kumagai[1], M. Kohiyama[2] and T. Yamashita[3]

[1]*Graduate Student, Graduate School of Science and Technology, Keio University. Email: kumataku_726@keio.jp*
[2]*Professor, Department of System Design Engineering, Faculty of Science and Technology, Keio University, Email: kohiyama@sd.keio.ac.jp*
[3]*Chief Researcher, National Research Institute for Earth Science and Disaster Resilience, Email: tyamashi@bosai.go.jp.*

Abstract: The author's research group has proposed a damage pattern classification method using a deep neural network (DNN) for monitoring the structural health of a building. We aimed to make the DNN for damage pattern classification, more accountable using the "interpretation" and "explanation" methods. We proposed generating input data that maximize the classification probability of each damage pattern, called a prototype, and drew a trajectory on a two-dimensional phase plane of appropriately selected state variables with color information on the degree of influence on the classification. We applied the proposed method to DNNs for damage classification of a wooden structure and a steel frame. The acceleration–velocity plane provided useful information about the mechanical characteristics of the DNN used in the damage pattern classification, and we thus demonstrated the effectiveness of the proposed method.

Keywords: structural health monitoring, deep neural network, interpretation.

1. Introduction

While deep neural networks (DNN) have been used in various fields with high discrimination accuracy, there exists a problem in terms of explanation for practical use. The contents of the DNN are kept in a black box, and the basis for discrimination is unclear. Specifically in fields related to human life, such as autonomous driving and structural health monitoring, there is a particular need for explanation. To address this problem, a number of studies have been conducted on the explanation and interpretation of DNNs in recent years (e.g., Ribeiro et al. 2016, Lin et al. 2017, Guidotti et al. 2018, Ribeiro et al. 2018, Samek et al. 2019, Lundberg and Lee 2019, Molnar 2020). There are two typical methods: one is an "interpretation" method that maps an abstract concept, for example, predicted class, into a domain that humans can understand, such as images and text. The other is an "explanation" method that scores the contribution of each feature to the determination of the DNN, and specifies the features regarded as important in the determination.

The purpose of this study is to propose a method to visually interpret and explain the DNN that classifies the damage pattern of a building, based on acceleration measurement records.

2. Damage pattern classification system

This study uses the damage pattern classification framework proposed by Yamashita et al. (Yamashita et al. 2018). First, a numerical analysis model of the target building is created, and an earthquake response analysis is performed assuming multiple damage patterns, then an earthquake response database is created. Using this, a DNN of damage pattern classifications is constructed by machine learning. After the system is installed in a real building, the response data observed during an earthquake are input to the DNN to determine the damage pattern.

In this study, we use a DNN composed of a total of five layers: an input layer, three hidden layers, and an output layer. The detailed structure of the DNN is described in Appendix. Acceleration time history data are used as input data, and the time length of one sample of input data is determined so that one or more force–displacement history loops can be drawn even when the target building is damaged.

3. Proposed method for interpreting and explaining the DNN

For the learned DNN, we propose a method that expresses the interpretation by activation maximization (AM) (Montavon et al. 2018) and the explanation by layer-wise relevance propagation (LRP) (Bach et al. 2015) in a two-dimensional graph of state variables (hereinafter referred to as phase plane) that can be mechanically interpreted.

Here, numerical integration is used to calculate the velocity and displacement from the accelerograms. In a "prototype," which will be explained in the following paragraph, the initial conditions of velocity and displacement are not determined, so these optimizations are performed such that the trajectory of motion draws loops around the origin in a phase plane.

$$\text{Minimize } J_v(v_0) = \max_t d(t; v_0) - \min_t d(t; v_0) \qquad (1)$$

$$\text{Minimize } J_d(d_0) = \frac{1}{t_{\max}} \int_0^{t_{\max}} d(t; d_0)\, \mathrm{d}t \qquad (2)$$

AM maximizes the following adjusted log likelihood J_{AM} to generate a representative input $\hat{\boldsymbol{x}}$ of a class c, which is called a "prototype":

$$\text{Maximize } J_{\mathrm{AM}} = \log p(c|\hat{\boldsymbol{x}}) - \lambda \|\hat{\boldsymbol{x}} - \tilde{\boldsymbol{x}}\|^2 \qquad (3)$$

where $p(c|\hat{\boldsymbol{x}})$ is the likelihood function, λ is the weight factor, and $\tilde{\boldsymbol{x}}$ is the average of the inputs. In LRP, the degree of influence on the prediction of a feature is defined as a relevance score, and the score is backpropagated from output to input in accordance with the rule that the sum of these scores is conserved between layers, and highlighted for a specific input.

In this study, we explain the basis of DNN judgement by drawing the prototype on a phase plane of appropriately selected two state variables that can be mechanically interpreted.

4. Validation using experimental data of a wooden building

First, we verified the proposed method using the data of a wooden building. The target building (Fig. 1) is the wooden house of Specimen 4 in Project: Experiments for Verification of the Design Methods for Three-Story Wooden Houses by Post and Beam Construction. We obtained the experimental data from the Archives of Shaking table Experimentation dataBase and Information (ASEBI), which is disseminated by the Hyogo Earthquake Engineering Research Center, National Research Institute for Earth Science and Disaster Resilience (NIED) (NIED 2020). The house conforms to Seismic Grade 1 of a Housing Performance Indication System in the Housing Quality Assurance Act, which satisfies the minimum requirements of the Building Standard Law of Japan. During the shaking table tests, uniaxial shaking was applied to the shorter axis direction of the house. The ground motion specified in the Building Standard Law for ground type 2 (hereafter referred to as BSL) was input to the house with a duration of 20 s and amplification factors of 112.5% and 150% were used.

Figure 1. Specimen of wooden house (NIED 2020).

Based on this experiment, we created a three-dimensional frame model, and acceleration time history data were obtained by performing a time history response analysis on the model (Kohiyama et al. 2020). The simulated acceleration data were used as the input data for the neural network. With respect to the damage patterns, the initial state of the above experiment is defined as no damage (D1), and the states after excited by BSL with amplification factor of 112.5% and 150% input wave are defined as medium damage (D2) and severe damage (D3), respectively. As shown in Fig. 2, the depicted waveform of acceleration is measured on the table, and the DNN is learned with three damage patterns.

In this study, a time history response analysis was performed using *Wallstat ver. 4.1.3*, which is a numerical analysis program that uses a three-dimensional frame model (Nakagawa 2010, Nakagawa et al. 2010). In Fig. 3, the appearance of the model is illustrated. The yellow and orange parts indicate the degree of damage reaching the yield point load and maximum load, respectively.

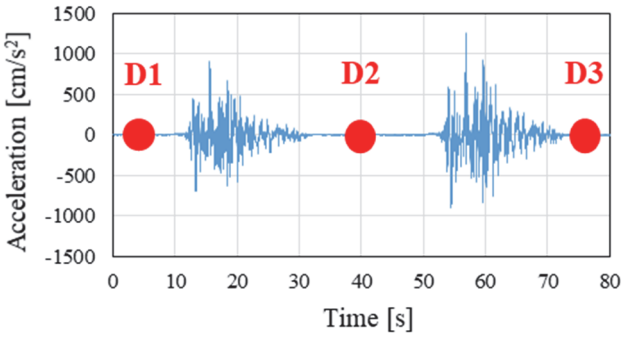
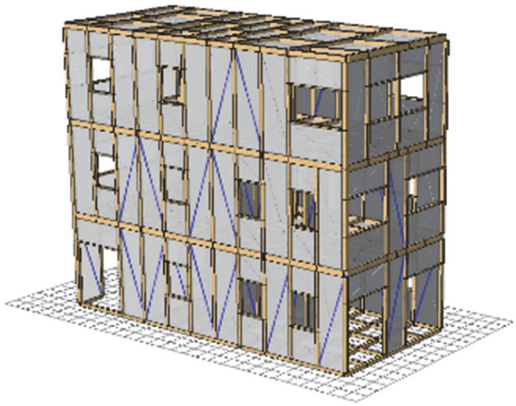
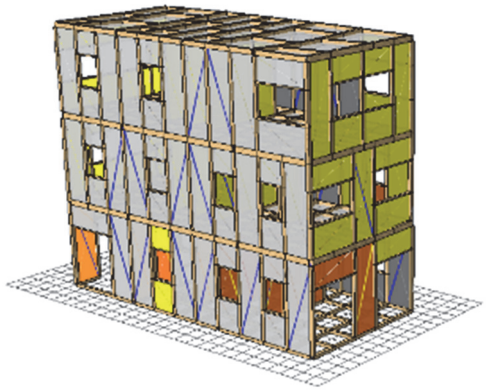


Figure 2. Relationship between damage patterns and input acceleration.



(a) D1



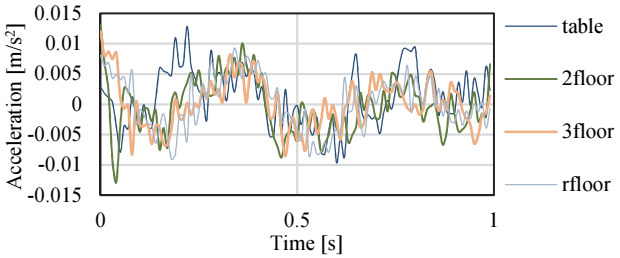
(b) D3

Figure 3. Appearance of analysis model and damage parts.

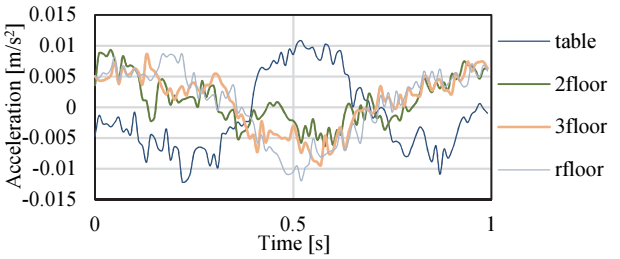
The prototypes of D1 and D3 generated by AM and shown in time history waveforms are depicted in Fig. 4. It can be confirmed that the natural period becomes longer as the damage progresses.

The relationship between the damping force and velocity is sometimes expressed in a force–velocity relationship graph with force on the vertical axis and velocity on the horizontal axis. Therefore, in this study, we substituted force with the acceleration obtained by dividing the force by a constant mass and drew a prototype in the

acceleration–velocity plane to obtain information on the damping force, as shown in Fig. 5.

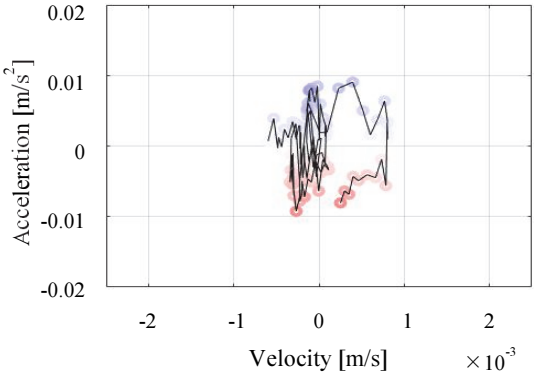


(a) D1

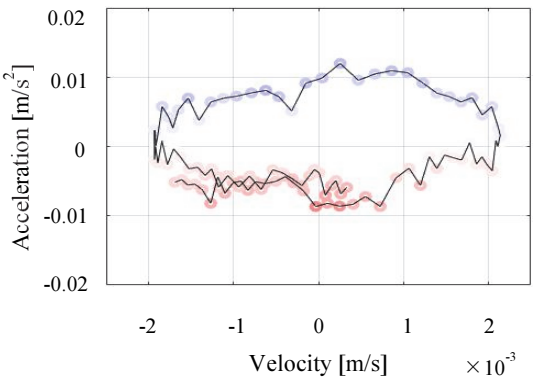


(b) D3

Figure 4. Acceleration time history of the prototypes.



(a) D1



(b) D3

Figure 5. Prototypes in the acceleration–velocity plane.

In Fig. 5, the absolute acceleration of the top layer of the target structure was used as the acceleration, and the velocity of the top layer relative to the layer just below the top was used as the velocity. From the figure, the difference in the damping characteristics can be confirmed from the width of the acceleration–velocity trajectory loops. Note that the red and blue colors in the figure represent positive and negative contributions to the prediction, respectively, which were obtained by LRP.

5. Validation using steel frame data

Next, we verified the proposed method using the results of the shaking table experiment of a steel frame performed at NIED. Four braces were attached to the second layer of the experimental specimen, and we emulated multiple damage patterns by fastening and relaxing the brace turnbuckles. Details of the target structure are described in Yamashita (2016), and the specimen is a 1:3 scaled model with a story height of 1.157 m and first natural period of 0.21 s under the condition that all braces are removed. The original scale structure was supposed to be a four-story steel structure with a height of 14 m and a slab of 6 × 12 m. Uniaxial shaking table tests were conducted using a large-scale shaking table at NIED, Tsukuba. One hundred different simulated ground motions with small to large amplitudes and short to long durations were input to the specimen for each damage pattern.

The measured acceleration records were divided into samples with lengths of 1 s, which were used as the input data for the neural network (Kohiyama et al. 2020). For the purpose of confirming the basis of the DNN judgement other than the natural period, the DNN is learned with four damage patterns, D9–D12, which are shown in Fig. 6.

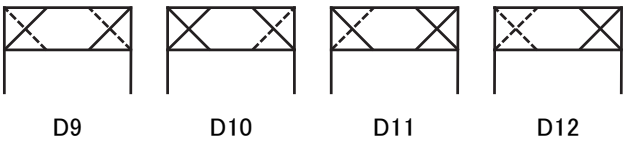
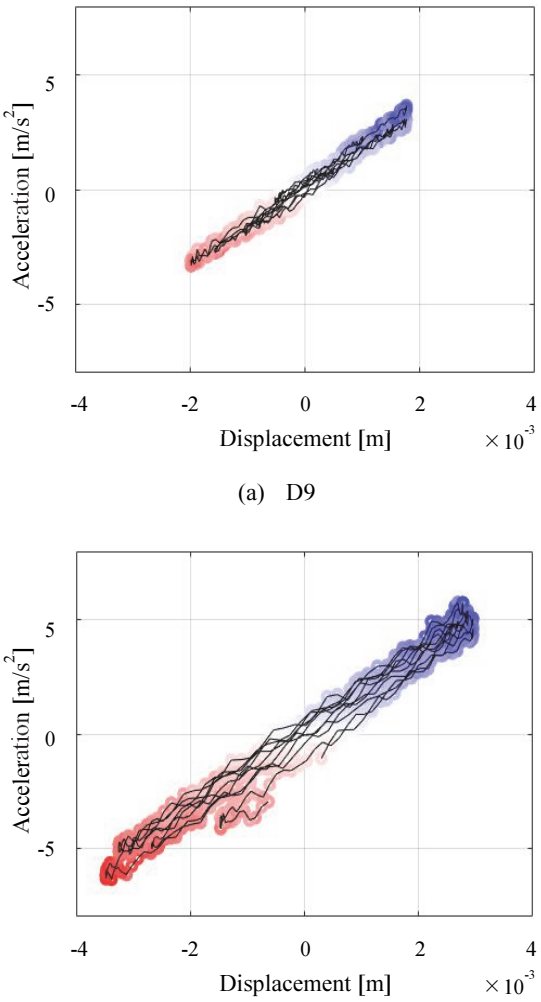


Figure 6. Damage patterns (broken lines represent the 2 mm stretch of the brace).

In the force–displacement plane, the relationship between the restoring force as well as the damping force and displacement appears in a loading history loop. Therefore, we again propose replacing the force by acceleration, and the trajectory of motion was drawn on the acceleration–displacement plane to obtain information on stiffness, plastic deformation, and damping.

In Figs. 7 and 8, the prototypes of damage patterns D9 and D12 drawn on the acceleration–velocity plane and the acceleration–displacement plane, respectively, are shown. When the regression line for the prototype trajectory drawn on the acceleration–displacement plane is determined, the slope of the line corresponds to the stiffness. From Fig. 7, it can be confirmed that the stiffnesses of D9 and D12 are similar, which indicates that their natural periods are also similar.



(a) D9

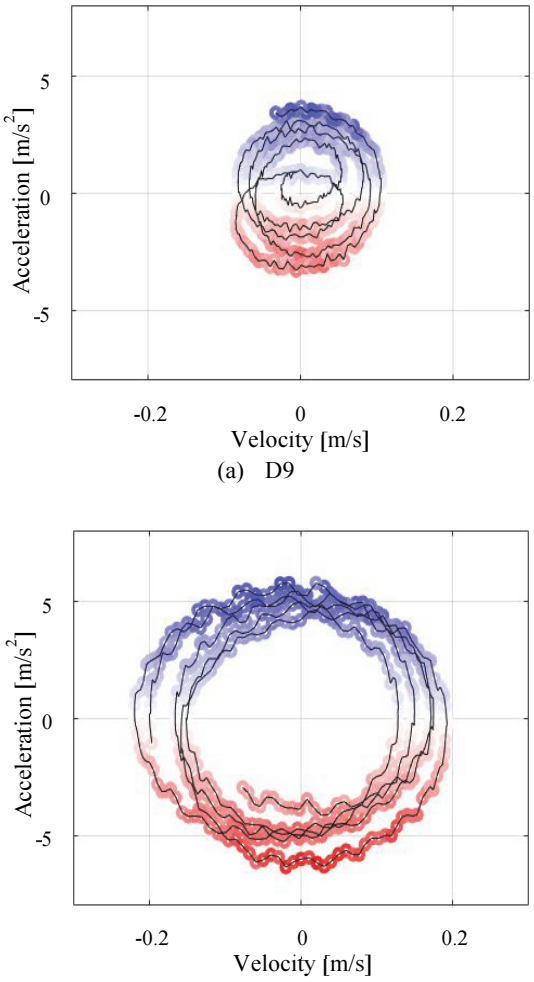
(b) D12

Figure 7. Prototypes on the acceleration–displacement plane.

Because two braces in the upper right direction are effective in D9, the damping force applied to the top layer in the lower-left direction is larger than that in the lower-right direction. Therefore, it can be inferred that an asymmetric vibration characteristic is realized. As can be observed in Fig. 7 (a), the displacement in the positive direction is smaller than that in the negative direction. Thus, a correspondence between the asymmetric vibration characteristics and damage patterns can be confirmed. In addition, it can be confirmed from Fig. 8 that there is a difference in the dispersions of the acceleration peaks in D9 and D12.

6. Conclusion

We proposed a method to visually explain the characteristics of a trained DNN for damage pattern classification based on acceleration records by graphical representation in a two-dimensional phase plane of appropriately selected two state variables employing the AM interpretation method and LRP explanation method.



(a) D9

(b) D12

Figure 8. Prototypes on the acceleration–velocity plane.

We demonstrated the validity of the proposed method through applications to DNN for damage pattern classification of a wooden house and steel frame. The damping characteristics and asymmetric vibration characteristics could be observed in the acceleration–velocity plane and we confirmed the usefulness of the proposed method.

References

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R. and Samek, W. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE*, 10(7), e0130140.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D. 2018. A survey of methods for explaining black box models, ACM Computing Surveys, 51(5), Article 93: 1–42.

Kohiyama, M., Oka, K. and Yamashita, T. 2020. Detection method of unlearned pattern using support vector machine in damage classification based on deep neural network. *Structural Control Health Monitoring*. DOI: 10.1002/stc.2552

Lin, Y., Nie, Z. and Ma, H. 2017. Structural damage detection with automatic feature-extraction through deep learning. *Computer-Aided Civil and Infrastructure Engineering*, 32: 1025–1046.

Lundberg, S. M. and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, Edited by Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R. Red Hook, New York, pp. 4765–4774.

Molnar, C. 2020. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. https://christophm.github.io/ (accessed on May 28, 2020).

Montavon, G., Samek, W. and Müller, K.-R. 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73: 1–15.

Nakagawa, T. 2010. Development of analysis method for collapsing behavior of wooden post-and-beam houses during earthquake. *Building Research Data*, 128: 1–90. (in Japanese)

Nakagawa, T., Ohta, M., Tsuchimoto, T. and Kawai, N. 2010. Collapsing process simulations of timber structures under dynamic loading III: numerical simulations of the real size wooden houses. *Journal of Wood Science*, 56(4): 284-292.

National Research Institute for Earth Science and Disaster Resilience. *E-Defense archives ASEBI: Experiment on verification of design method of three-story wooden frame construction method*. https://www.edgrid.jp/datas (accessed on May 28, 2020).

Ribeiro, M. T., Singh, S. and Guestrin, C. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, San Diego, California, June 12–17, 2016. ACL, Stroudsburg, Pennsylvania, pp. 97–101.

Ribeiro, M. T., Singh, S. and Guestrin, C. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, February 2–7, 2018. AAAI, Palo Alto, California, pp. 1527–1535.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R. 2019. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Springer, Berlin.

Yamashita, T. 2016. Preliminary analysis for shake table test of small-scaled steel frame using detailed FEM. *Summaries Tech Papers Annual Meeting of Architectural Institute of Japan*. 2016; B-3:701–702. (in Japanese)

Yamashita, T., Kohiyama, M., Watanabe, K., Matsuzaki, K. and Mori, Y. 2018. Estimation of damaged brace members installed in a steal frame using deep neural network. *Proceedings of the Conference on Computational Engineering and Science*, 23: G-03-01 1–6. (in Japanese)

Appendix

The DNN used in this study consists of an input layer, three hidden layers, and an output layer as shown in Fig. 9. The number of nodes in the input layer is the product of the number of sensors, sampling frequency, measurement length, and number of acceleration direction components; for example, 4 × 100 Hz × 1 s × 1 = 400 in Section 4 and 3 × 500 Hz × 1 s × 1 = 1500 in Section 5. The number of nodes in the output layer is the number of damage patterns, i.e., three in Section 4 and four in Section 5. The number of nodes in the hidden layers is set so that it will be halved when proceeding to the next layer. As the activation function, ReLU (rectified linear unit) is used in the hidden layers and the softmax function in the output layer. Cross-entropy is used as a loss function for calculating and updating the training parameters. In addition, Adam (adaptive moment estimation) is used as a parameter updating method in training the DNN.

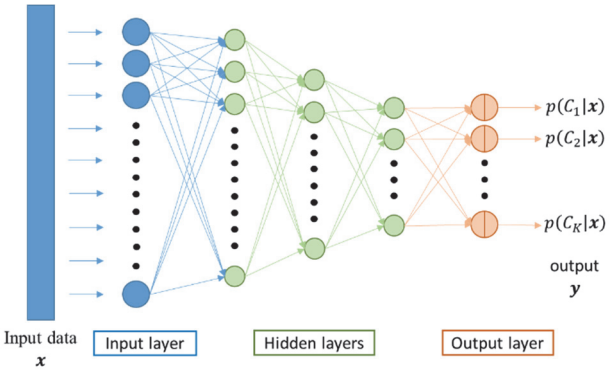


Figure 9. Structure of the DNN used in the study.