Enhancing the Explainability of AI Models in Nuclear Power Plants with Layer-wise Relevance Propagation

Seung Geun Kim [a*], Seunghyoung Ryu [a], Hyeonmin Kim [b], Kyungho Jin [b], Jaehyun Cho [b]
[a]Applied Artificial Intelligence Laboratory/[b]Risk Assessment and Management Research Team, Korea Atomic Energy Research Institute, 111, Daedeok-daero 989beon-gil, Yuseong-gu, Daejeon, South Korea, 34057
[*]Corresponding author: sgkim92@kaeri.re.kr

1. Introduction

With the rapid advance of artificial intelligence (AI) technology, there have been enormous number of applications across the various fields. Nuclear field is also following this trend, and there have been many studies that utilize AI models for solving problems such as event diagnosis and automated/autonomous operation.

However, deep neural network (DNN) which takes the biggest portion of recent AI technology applications has a limitation that it is not transparent and has low explainability. In the case of DNN-based model, it is difficult to know the internal logic of the model or how the model deduces outputs from given input. Due to this limitation, there is a hesitation on practical application of DNN-based models to the safety-critical fields and the fields related to moral/legal issues, although their performances are acceptable.

To overcome the limitation of low explainability, a number of explainable AI (XAI) methods have been proposed. XAI methods can provide detail explanations such as model's internal logics and relations between inputs and outputs. However, although the explainability issue is crucial for safety-critical nuclear field, there is a lack of studies that dealing with XAI.

In this study, as an effort to enhance the explainability and the practicality of AI models in nuclear field, layer-wise relevance propagation (LRP) [1] was investigated, which is one of the XAI methods that has shown better performance in many applications compared to other XAI methods.

The rest of paper is organized as follows. In chapter 2, brief explanations on XAI and LRP are provided. In chapter 3, experiments for feasibility check are described and chapter 4 concludes the paper.

2. Preliminaries

2.1 Explainable Artificial Intelligence

Explainable artificial intelligence (XAI) is a technology that makes human to easily understand the AI model. Most of AI models are different from human in terms of data processing and problem solving methods. For example, AI model recognizes images with pixel-wise RGB values while human does not. XAI is proposed to alleviate the difficulty of understanding the AI model's internal processes or the reason why certain output is deduced.

XAI can be applied for various AI methods. However, as not only DNN's popularity but also their inherent explainability issue, most of recent XAI related studies are focusing on explaining DNN-based models.

There are various kinds of XAI methods, and they have their own range of application, advantages and limitations. These methods can be classified according to their type of applications, type of explanations and range of explanations. Table I is showing the taxonomies for classifying XAI methods, suggested by Yu Zhang et al. [2].

Table I: Taxonomy for classifying XAI methods

Dimension	Category	Description
Type of applications	Passive	Post-hoc explain
	Active	Actively change the network architecture or training processes
Type of explanations	Examples	Provide examples
	Attributions	Assign credit to the input features
	Hidden semantics	Make sense of hidden neurons/layers
	Rules	Extract logic rules
Range of explanations	Local	Explain on individual samples
	Semi-local	In between the local and global
	Global	Explain the network as a whole

In this paper, among various XAI methods, LRP that belongs to the passive/attributions/local category is selected. Brief descriptions on LRP are provided in next sub-chapter.

2.2 Layer-wise Relevance Propagation

Layer-wise relevance propagation (LRP) [1] is an XAI method that can be applied for DNN-based models. LRP explains the model by calculating 'relevance', which implies the level of contribution of specific part of the input for deducing the output. Since relevance calculation processes are mathematically well-defined, LRP tends to deduce clearer result compared to other XAI methods that belongs to the same category. Moreover, LRP tends to deduce more informative results as the relevance itself has same meaning to the 'importance' of input's corresponding part. Figure 1 is a schematic of LRP.

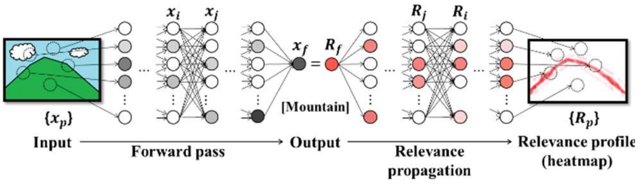

Fig. 1. Schematic of LRP

LRP is based on two fundamental concepts. First concept is the relevance conservation law, which means that the sum of the relevance is always conserved at the each layer within the model. Second concept is the relevance propagation rules, which are formula for relevance calculation through the layers. Relevance propagation rules include various formula that follows relevance conservation law, and the user should define which formula to be applied. Figure 2 is a summary of two core concepts of LRP.

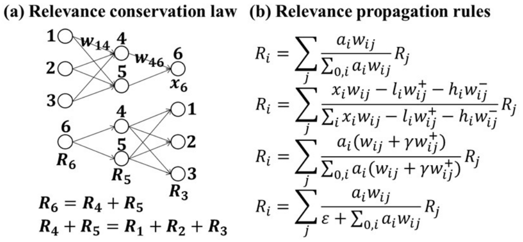
(a) Relevance conservation law
(b) Relevance propagation rules
1
2
3
w_{14}
4
5
w_{46}
6
x_6
6
R_6
4
5
R_5
1
2
3
R_3
$R_6 = R_4 + R_5$
$R_4 + R_5 = R_1 + R_2 + R_3$
$R_i = \sum_j \frac{a_i w_{ij}}{\sum_{0,i} a_i w_{ij}} R_j$
$R_i = \sum_j \frac{x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-}{\sum_i x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-} R_j$
$R_i = \sum_j \frac{a_i (w_{ij} + \gamma w_{ij}^+)}{\sum_{0,i} a_i (w_{ij} + \gamma w_{ij}^+)} R_j$
$R_i = \sum_j \frac{a_i w_{ij}}{\varepsilon + \sum_{0,i} a_i w_{ij}} R_j$

Fig. 2. Examples of (a) relevance conservation law and (b) relevance propagation rules

Although LRP has shown better performance compared to many other XAI methods, there are limitations that it is difficult to apply, and requires high computation resource. Additionally, as a limitation not only for LRP but also for many XAI methods, there is a problem that most of recent studies are focusing on classification problems with image and natural language data [3, 4]. AI models applied to the nuclear field are mostly based on time-series data, but related studies are scarce. LRP can be applied for every kind of data and every DNN architectures in theory. However, for the practical application and meaningful results, it is needed to conduct additional studies.

3. Experiments

In this study, to check the applicability of LRP, it was applied for accident diagnosis problem which is a classification problem based on time-series data. The experiments can be summarized into three steps including data acquisition and processing, model development and training, and application of LRP.

3.1 Data Acquisition and Processing

It is needed to develop a target model first in order to apply LRP. For the experiments, since accident diagnosis is one of the most popular problem for applying AI model, a simple accident diagnosis model was developed for feasibility study.

To acquire the data for model training and testing, MARS (Multidimensional analysis of reactor safety) code developed by KAERI [5] was used. Considered accident scenarios were SGTR (Steam generator tube rupture) with break sizes A, 2A, 4A, and MSLB (Main steam line break) with break sizes A, 2A. Control of several components were included to broaden the range of data distribution. For 900 seconds from the emergency reactor trip, 19 kinds of instrumentation signals were obtained. List of obtained instrumentation signals are shown in Table II.

Totally 104,615 datasets were obtained including 62,775 SGTR datasets and 41,840 MSLB datasets. Among them, 84,615 datasets consist of 50,775 SGTR datasets and 33,840 MSLB datasets were used for model training and rest of 20,000 datasets consist of 12,000 SGTR datasets and 8,000 MSLB datasets were used for model testing.

After the simulation, min-max scaling was applied to make the measurement values between zero to one.

Table II: List of obtained instrumentation signals and their units

Instrumentation signals	Units
Reactor power	%
SG* 1/2 level difference	%
RCP* 1 on/off	-
RCP 2 on/off	-
RCP 3 on/off	-
RCP 4 on/off	-
SG 1 level (wide range)	%
SG 2 level (wide range)	%
SG 1 level (narrow range)	%
SG 2 level (narrow range)	%
FW* flow to SG 1	L/sec
FW flow to SG 2	L/sec
AFW* flow to SG 1	L/sec
AFW flow to SG 2	L/sec
PZR* pressure	kg/cm^{2}A
PZR level	&
RCS* subcooling margin	ºC
SG 1 pressure	kg/cm^{2}A
SG 2 pressure	kg/cm^{2}A

*SG: steam generator, RCP: reactor coolant pump, FW: feedwater, AFW: auxiliary feedwater, PZR: pressurizer, RCS: reactor cooling system

3.2 Model Development and Training

For the experiments, a simple accident diagnosis model with five fully-connected layers was developed. As activation functions, rectified linear unit (ReLU) activation function was applied for all layers except the last layer with log-softmax activation function. The model receives a vector with length 17,100 that includes 19 kinds of instrumentation signals for 900 seconds, and deduces a vector with length 5 that includes probabilities for each accident category. Figure 3 is a schematic of the developed accident diagnosis model.

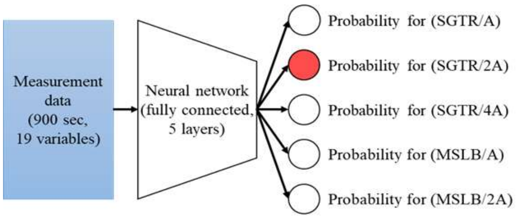
Measurement data (900 sec, 19 variables)
Neural network (fully connected, 5 layers)
Probability for (SGTR/A)
Probability for (SGTR/2A)
Probability for (SGTR/4A)
Probability for (MSLB/A)
Probability for (MSLB/2A)

Fig. 3. Schematic of the developed accident diagnosis model

Negative log-likelihood function was used as a loss function, and Adam optimizer [6] with default setting was used as an optimizer. The model was repeatedly trained with applying various hyperparameter sets. The model that shown best performance for testing data was applied at the next LRP application step. Selected model shows about 92.8% accuracy for entire data, and the accuracy for training data was slightly higher.

3.3 Application of LRP

After the development of accident diagnosis model, LRP was applied for the randomly selected datasets. For the relevance propagation rules, LRP-zB rule (equation 1) was applied for input layer and LRP-ε rule (equation 2) was applied for another layers. Equations for these two rules can be represented as follows.

$$R_i = \sum_j \frac{a_i w_{ij} - l_i w_{ij}^{+} - h_i w_{ij}^{-}}{\sum_i a_i w_{ij} - l_i w_{ij}^{+} - h_i w_{ij}^{-}} R_j \qquad (1)$$

$$R_i = \sum_j \frac{a_i \rho(w_{ij})}{\varepsilon + \sum_{0,i} a_i \rho(w_{ij})} R_j \qquad (2)$$

Where R_i is i-th node's relevance value, a_i is i-th node's value, l_i is maximum input value, h_i is minimum input value, w_{ij} is weight between node i and j, w_{ij}^{+} is max(0, w_{ij}), w_{ij}^{-} is min(0, w_{ij}), ε is stabilization factor, and ρ is activation function.

Several parts of the result are shown in Figure 4. The graphs at upper positions are about the instrumentation signals, and the graphs at lower positions are about the relevance values. The red line in the relevance graph shows the mean value of 900 relevance values for corresponding instrumentation signal. The green line in the relevance graph shows the mean value of total 17,100 relevance values. If the red line is above the green line, it implies that the signal has higher importance than many other signals. In contrast, if the red line is below the green line, it implies that the signal has lower importance than many other signals.

Cases (a), (b) and (c) in Figure 4 are showing representative patterns of LRP analysis results.

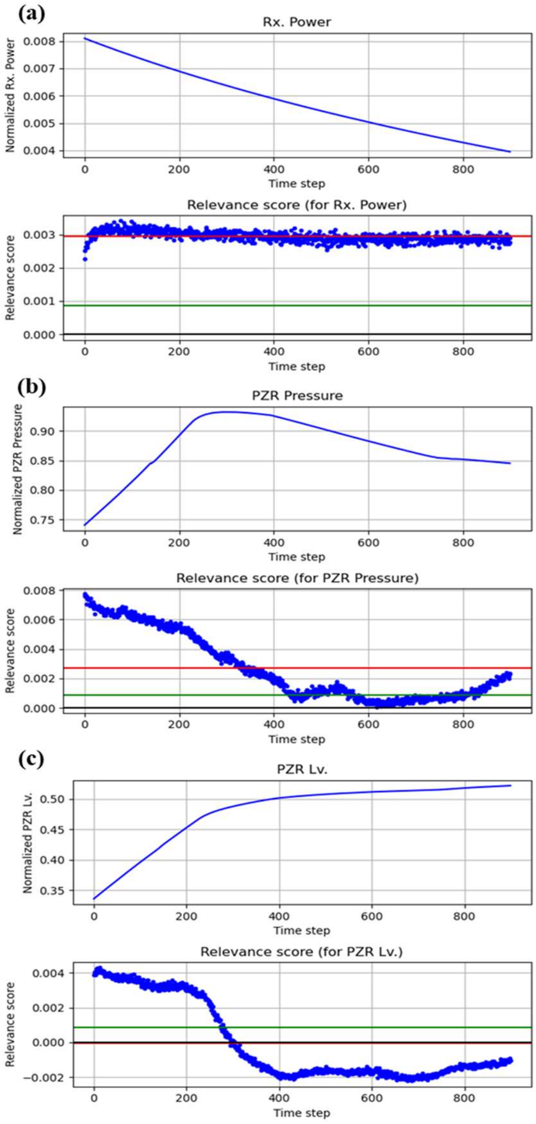

Fig. 4. Examples of LRP application result

Case (a) is the example about when the signal has more influence compared to other signals over all time steps. As relevance values are consistent over all time steps, correlation analysis could deduce similar results with LRP.

Case (b) is the example about when the signal has more influence compared to other signals, but not for every time steps. As shown in graph, relevance values at the front part are higher than relevance values at the back part. Relevance values at the back part are near the green line, implying that the corresponding elements have less influences on classification result.

Case (c) is the example about when the signal has less influence compared to other signals, but not for every time steps. Similar to the case (b), relevance values at the front part are higher than the relevance values at the back part. However, as relevance values at the back part have negative numbers and cancelling the effect of front part, overall influence of the signal is decreased. Negative relevance values imply that the corresponding elements have negative influences on classification result.

From the cases (b) and (c), it was revealed that the application of LRP could lead to the acquisition of additional information that cannot be found by the conventional methods.

For the image data, LRP application leads to highly intuitive results that high relevance values are deduced for the pixels around the object boundaries. However, although it was shown that LRP application could reveal the 'important parts' also for the time-series data, there is a limitation that the results are not much intuitive and may require domain knowledge for proper interpretation.

4. Conclusion

In this study, LRP which is one of the XAI method was adopted to nuclear field for enhancing the AI model's explainability. To check the applicability of LRP, experiments were conducted that apply LRP to the simple accident diagnosis model. Although it is difficult to quantitatively compare the degree of explainability, it was empirically found that the application of LRP could provide more information than conventional methods.

Through the application of LRP, it would be possible to easily find improvements such as improving the accuracy in the development stage of AI-based operation support systems. Even after the development, additional information given by LRP can be used for further improvements with proper expert supervisions. Application of LRP would be beneficial for not only in the terms of performance, but also for complementary cooperation between human operator and operation support system, since it enables the operator to conduct detailed review on AI model's output.

Current XAI studies are mostly focusing on classification problems with image or natural language data, and therefore an AI model for the accident diagnosis-which is a representative classification problem in nuclear field-was used in the experiments. However, since AI models in nuclear field are applied to various problems other than classification problem, it is needed to investigate on how to apply XAI and LRP for such problems. Moreover, it is needed to conduct studies on other XAI methods to find appropriate method for each problem.

ACKNOWLEDGMENT

This work was supported by the Ministry of Science, ICT and Future Planning of the Republic of Korea and the National Research Foundation of Korea (NRF-2020M2C9A1061638).

REFERENCES

[1] S. Bach et al., "On Pixel-wise Explanations for Non-linear Classifier Decisions by Layer-wise Relevance Propagation", PLos ONE 10(7): e0130140, 2015

[2] Y. Zhang et al., "A Survey on Neural Network Interpretability", arXiv preprint arXiv:2012.14261v2, 2021

[3] B. K. Iwana et al., "Explaining Convolutional Neural Networks using Softmax Gradient Layer-wise Relevance Propagation", 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 4176-4185, 2019

[4] H. Bharadhwaj, "Layer-wise Relevance Propagation for Explainable Deep Learning Based Speech Recognition", 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 168-174, 2018

[5] W. J. Lee et al., "Development of MARS for Multi-dimensional and Multi-purpose Thermal-hydraulic System Analysis", Japan-Korea Symposium on Nuclear Thermal Hydraulics and Safety, Oct 15-18, 2000, Fukuoka, Japan,

[6] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization", 3rd International Conference for Learning Representations, 2015, San Diego.