2021 KNS Autumn Meeting

Enhancing the Explainability of AI Models in NPPs with Layer-wise Relevance Propagation

Seung Geun Kim, Seunghyoung Ryu, Hyeonmin Kim, Kyungho Jin, Jaehyun Cho
Korea Atomic Energy Research Institute

Thursday, October 21nd

Contents

I. Introduction

I. Introduction

- Artificial intelligence (AI) technology has been rapidly advanced.
- The nuclear field is also following this trend, and there are efforts to apply AI technology for many purposes.
  - Event/accident diagnosis
  - Component prognostics and health monitoring (PHM)
  - Simulation acceleration
  - Automated/autonomous operation

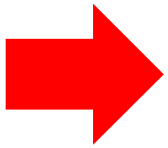
AI-based Operation support systems (OSSs)

- However, deep neural network (DNN) has a critical drawback that it has low 'explainability'.
- Its inherent 'black-box' characteristic makes difficult to know the internal logic or output deducing process of the model.
- Importance of the model's explainability is more emphasized when the model is related to,
  - Safety-critical systems.
  - Problems including moral/legal issues.

I. Introduction

- The issue on explainability should be thoroughly considered when applying the AI models to the nuclear power plant (NPP).
- However, many studies on developing AI-based OSSs for NPPs did not considered the explainability issue.

- To enhance the explainability, various explainable AI (XAI) methods have been developed.
- Each method has its own advantages and disadvantages.
- Recent studies on XAI methods have revealed that the layer-wise relevance propagation (LRP) shows outstanding performance compared to the other methods.

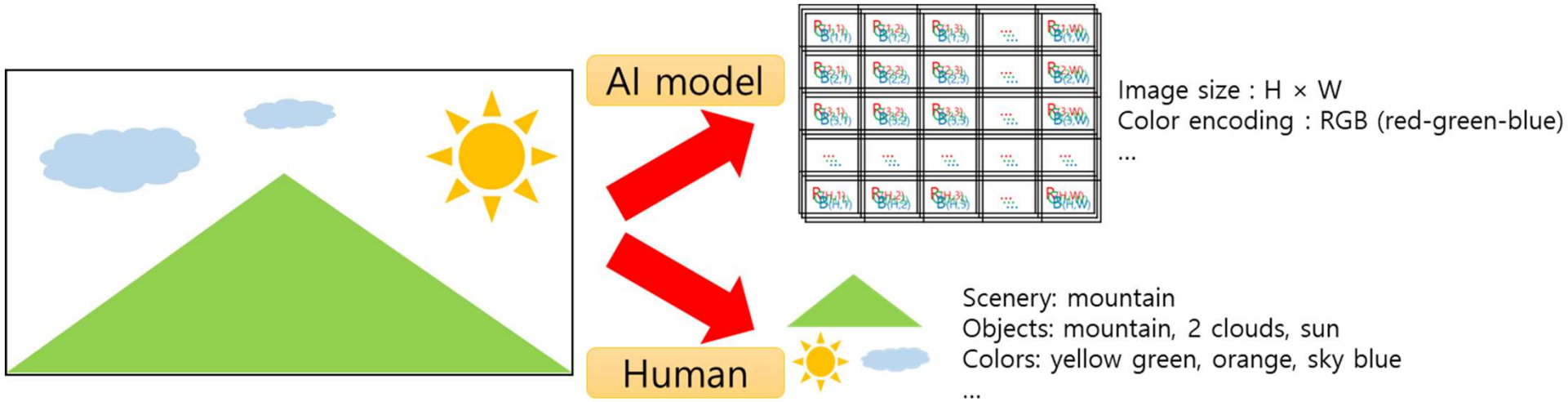


LRP is applied to check whether the application of XAI method could provide better explanations on the AI models in NPP OSSs.

II. Preliminaries

II. Preliminaries

- Explainable artificial intelligence (XAI)
  - Data recognition schemes of AI model (or computer) and human are different.
  - This discrepancy makes difficult to intuitively understand many AI models for human.
  - XAI is a set of methods and technology that make given AI model easily understood by human.



II. Preliminaries

- There are many kinds of XAI methods.
  - Each XAI method has its own advantages and limitations.
  - XAI methods can be categorized with various criteria.
  - Yu Zhang et al. proposed taxonomies for classifying XAI methods.

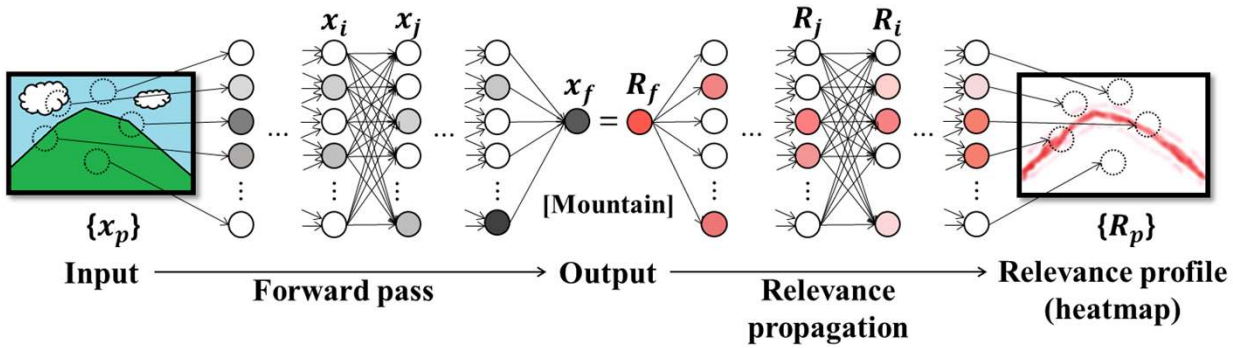
Dimension	Category	Description
Passive vs. Active Approaches	Passive	Post-hoc explain trained neural network
	Active	Actively change the network architecture or training process for better interpretability
Type of Explanations (to explain a prediction/class by…)	Examples	Provide example(s) which may be considered similar or as prototype(s)
	Attribution	Assign credit (or blame) to the input features (e.g. feature importance, saliency masks)
	Hidden semantics	Make sense of certain hidden neurons/layers
	Rules	Extract logic rules (e.g. decision trees, rule sets and other rule formats)
Local vs. Global Interpretability	Local	Explain network's predictions on individual samples (e.g. a saliency mask for a input image)
	Semi-local	In between, for example, explain a group of similar inputs together
	Global	Explain the network as a whole (e.g. a set of rules/a decision tree)

II. Preliminaries

- XAI can be utilized for,
  - Enhancement of the model performance
  - Verification of the model
  - Learning new insights from the model
  - Solving ethical/legal problems of the model
  - …

II. Preliminaries

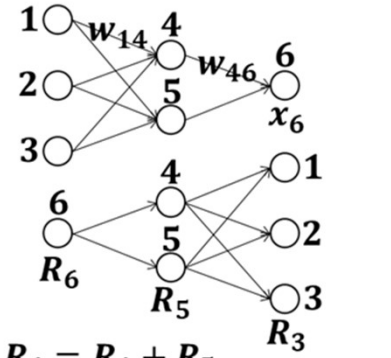
- Layer-wise relevance propagation (LRP)
  - LRP is one of the XAI methods that can be applied for DNN-based models.
  - Investigates the contribution of input's each element to the model's output, referred as 'relevance'.
  - Based on the two concepts: relevance conservation law and relevance propagation rules



  - Advantages of LRP
    - Tends to deduce clearer results.
    - Tends to deduce more informative results.
    - High performance.
  - Limitations of LRP
    - Requires more computation resources and human efforts.
    - Most of studies have focused on classification problems with image and natural language data.

II. Preliminaries

- Relevance conservation law
  - Law that the sum of every nodes' relevance values within one layer is always same to that of another layers.
  - Similar to the energy conservation or mass conservation law in physics.
- Relevance propagation rules
  - Various formula that are applied for relevance calculation.
  - Every rules should follow relevance conservation law.
  - User should determine which rule to be applied for each layer.



$$R_6 = R_4 + R_5$$

$$R_4 + R_5 = R_1 + R_2 + R_3$$

Relevance conservation law

$$R_i = \sum_j \frac{a_i w_{ij}}{\sum_{0,i} a_i w_{ij}} R_j$$

$$R_i = \sum_j \frac{x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-}{\sum_i x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-} R_j$$

$$R_i = \sum_j \frac{a_i (w_{ij} + \gamma w_{ij}^+)}{\sum_{0,i} a_i (w_{ij} + \gamma w_{ij}^+)} R_j$$

$$R_i = \sum_j \frac{a_i w_{ij}}{\varepsilon + \sum_{0,i} a_i w_{ij}} R_j$$

Relevance propagation rules

III. Experiments

III. Experiments

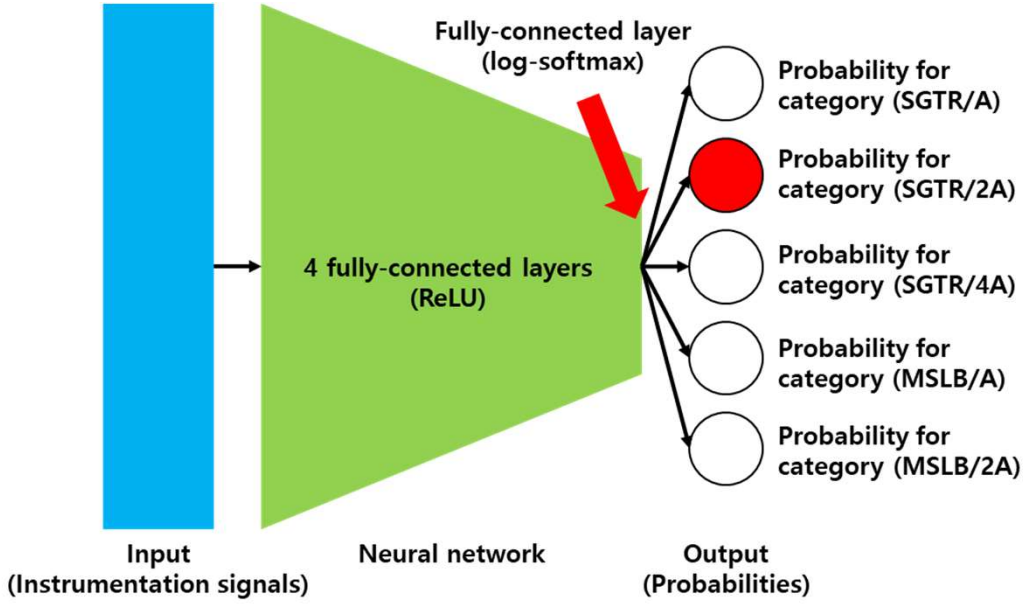
- Data preparation
  - Simulation based on MARS* code is conducted (reference plant: OPR1000).
  - SGTR* (with break sizes A/2A/4A) and MSLB* (with break sizes A/2A) accident cases are simulated.
  - Control of several components are included.
  - 19 kinds of instrumentation signals are obtained for 900 seconds from the emergency reactor trip.
  - Min-max scaling is applied to adjust the ranges of every signals from zero to one.
  - Totally 104,615 data sets are acquired (training: 50,775 SGTR, 33,840 MSLB / testing: 12,000 SGTR, 8,000 MSLB).

Instrumentation signals	Units
Reactor power	%
Steam generator 1 / 2 level difference	%
Reactor coolant pump 1 on/off	-
Reactor coolant pump 2 on/off	-
Reactor coolant pump 3 on/off	-
Reactor coolant pump 4 on/off	-
Steam generator 1 level (wide range)	%
Steam generator 2 level (wide range)	%
Steam generator 1 level (narrow range)	%
Steam generator 2 level (narrow range)	%
Feedwater flow to steam generator 1	L/sec
Feedwater flow to steam generator 2	L/sec
Aux. feedwater flow to steam generator 1	L/sec
Aux. feedwater flow to steam generator 2	L/sec
Pressurizer pressure	kg/㎠A
Pressurizer level	%
Reactor cooling system subcooling margin	℃
Steam generator 1 pressure	kg/㎠A
Steam generator 2 pressure	kg/㎠A

*MARS: Multidimensional analysis of reactor safety
*SGTR: steam generator tube rupture
*MSLB: main steam line break

III. Experiments

- Model development and training
  - Simple accident diagnosis model is developed.
  - 4 FC* layers with ReLU* activation function + 1 FC layer with log-softmax activation function
  - Input: instrumentation signals (size: 17100 = 19 variables × 900 sec/variable)
  - Output: probabilities for every accident categories (size: 5)
  - Optimized model has shown 92.8% mean accuracy.



*FC: Fully-connected
*ReLU: Rectified linear unit

III. Experiments

- Application of LRP
  - Several data sets are selected randomly, and their outputs are deduced from the trained model.
  - LRP is applied to calculate relevance value for each element in the input.
  - Applied relevance propagation rules: LRP-zB rule (input layer), LRP-generic rule (other layers)

LRP-zB rule :

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

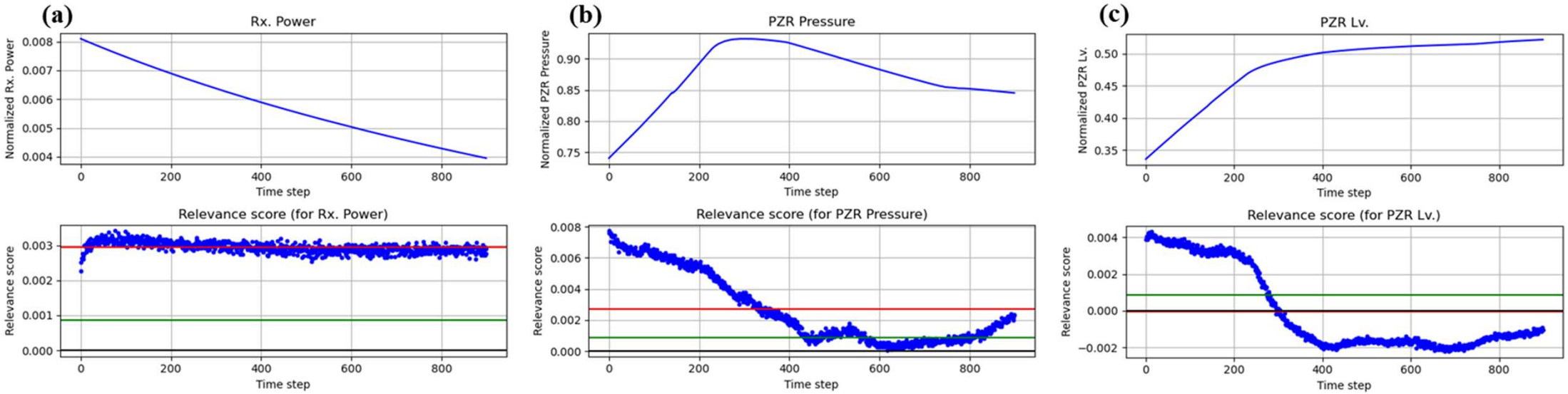
LRP-generic rule :

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

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

III. Experiments

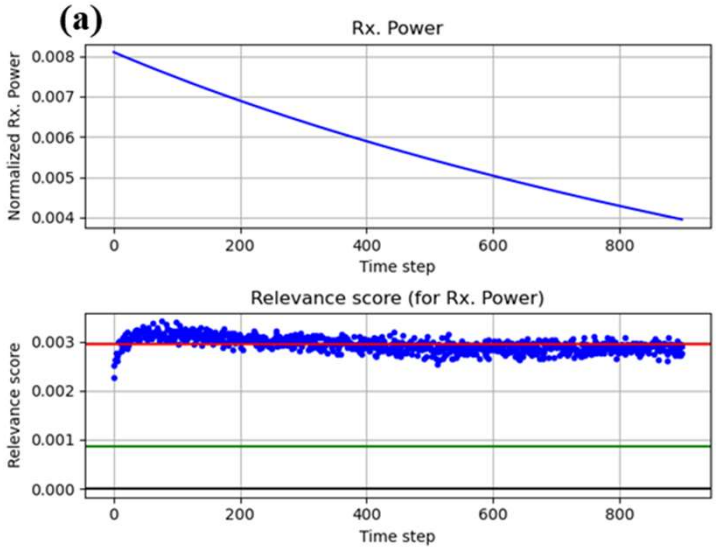
- Exemplary results (SGTR/2A case)



*Green line: mean relevance value over all variables and entire time steps
*Red line: mean relevance value for each variable over entire time steps

- (a) : example that the variable is more important than other variables and the relevance values are almost constant over time.
- (b) : example that the variable is more important than other variables, while the relevance values change over time.
- (c) : example that the variable is not much important, while the relevance values change over time.

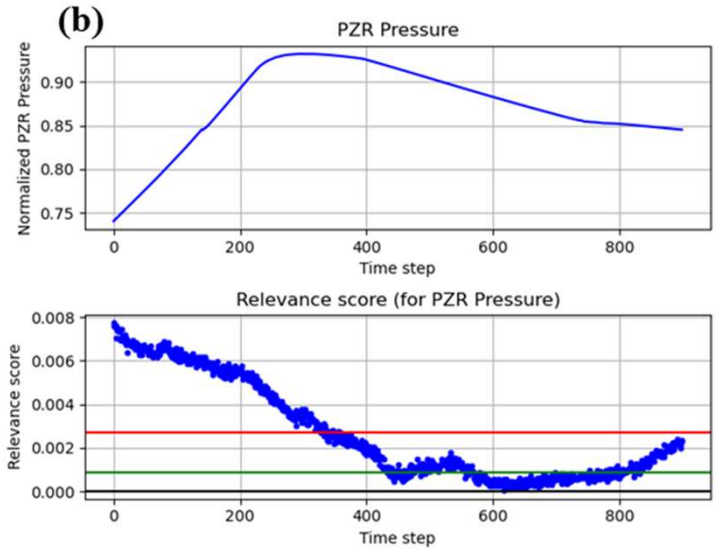
III. Experiments



*Green line: mean relevance value over all variables and entire time steps
*Red line: mean relevance value for each variable over entire time steps

- Since the relevance values are almost constant, similar results can be obtained by conventional correlation analysis.

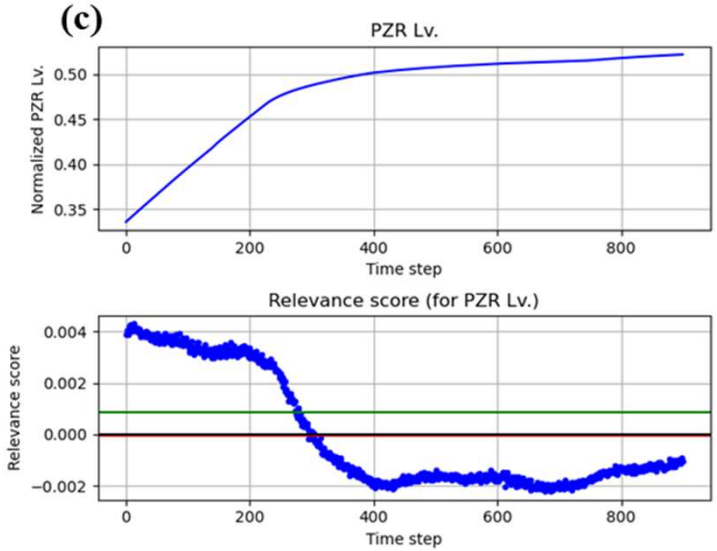
III. Experiments



*Green line: mean relevance value over all variables and entire time steps
*Red line: mean relevance value for each variable over entire time steps

- Relevance values are relatively higher at the front part compared to the back part.
- In the back part, relevance values are located near the green line, which means that corresponding part is not much important for deducing output.

III. Experiments



*Green line: mean relevance value over all variables and entire time steps
*Red line: mean relevance value for each variable over entire time steps

- Relevance values are relatively higher at the front part compared to the back part.
- This case is showing that although some variables' mean relevance value is low, still there could exist 'important parts'.

IV. Conclusion

IV. Conclusion

- LRP is introduced as an effort for adopting XAI method to the AI-based OSSs in NPPs.
- As a feasibility study, application of LRP is conducted for the simple accident diagnosis model.
- It is empirically revealed that the application of LRP to the AI-based OSSs could deduce informative results.
  - Application of LRP enables the precise element-wise relevance analysis, which is not available for conventional methods.
  - By inspecting these results, operators could make advanced decisions.

IV. Conclusion

- Expectations
  - Model for OSS can be further enhanced by precisely investigating and improving weak points.
  - Innovative knowledge can be deduced from the model if the model's mechanism is different from the human knowledge.
  - LRP may grants operators a second chance to reconsider the model's decision, when the model is suspected to deduce wrong or bad decision.
  - These enhancements could lead to the increment of public acceptance and reliability of AI-based OSSs.

- Future works
  - Further investigations on applying LRP to non-classification problems.
  - Studies on application of other kinds of XAI methods and comparing their performances.

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)

Thank you for your listening