A DEEP LEARNING MODEL FEATURE INTERPRETABILITY ANALYSIS METHOD FOR POWER SYSTEM TRANSIENT STABILITY ASSESSMENT

Yibo ZHOU[1,*], Liang ZHANG[2]

The mechanism underlying power system transient stability is complex. Deep learning models offer an effective solution for capturing complex mapping relationships, making them widely employed in transient stability assessment. However, the deep learning models face challenges in ensuring the effectiveness of feature extraction due to the lack of domain knowledge support. This limitation hampers improvements in evaluation accuracy. Furthermore, the inability to comprehend the acquired knowledge of the model raises concerns about trusting evaluation results, especially in security-sensitive scenarios. To address these issues, this article proposes a method for analyzing the interpretability of deep learning model features in power system transient stability assessment. Firstly, we construct a CNN model specifically designed for transient stability assessment. Then, we introduce a global interpretation method known as maximizing activation (AM) to obtain a comprehensive interpretation of typical stable modes within the model's injection space. Finally, the Class Activation Map (Grad-CAM) is utilized to identify dominant features in the injection space, providing guidance for the online application of transient stability assessment. The case studies show that this method can make operators easily understand the transient stability assessment knowledge learned by neural networks and improve the accuracy of transient stability assessment under the security region.

Keywords: transient stability assessment; deep learning model, feature interpretability analysis; stability pattern recognition.

1. Introduction

The machine learning (ML) method can realize end-to-end learning without manual feature extraction, such as deep neural network (DNN). This dramatically simplifies the dependence on expert knowledge and feature engineering. Therefore, DL has become one of the current common methods for solving complex classification, and regression problems [1].

[1] Key Laboratory of Modern Power System Simulation and Control & Renewable Energy Technology, Ministry of Education (Northeast Electric Power University), Jilin, Jilin Province, China, e-mail: zhouyiboaa@126.com;

[2] Key Laboratory of Modern Power System Simulation and Control & Renewable Energy Technology, Ministry of Education (Northeast Electric Power University), Jilin, Jilin Province, China, e-mail: longdongzhif-eng@163.com

With the increase in the level of the deep neural network, the gradually abstracted features make it difficult for humans to understand. Therefore, DNN is usually regarded as a "black box", and the model's performance judges the availability requirements. Sometimes, models that are difficult to be explained by human knowledge show significant differences in some scenarios. At this time, no apparent reason can be provided [2]. Therefore, this seriously restricts the widespread application of the DNN model in engineering. It is essential to study the interpretability of the DNN model.

There are mainly two ways to improve the interpretability of deep learning models at home and abroad: 1) combining human knowledge and feature visualization methods, explaining the key features of model extraction and recognition; 2) The attention mechanism is introduced to get the attention of the model to some features, to judge which input features are the dominant features. Reference [3] got the gradient of the network output relative to the input through the back-propagation algorithm. Then, saliency maps corresponding to the input are constructed to highlight important parts of the input samples. The deconvolution network reconstructs the feature map with the same dimension as the input sample [4]. This identifies pixels and regions that are significantly activated in the input picture. Reference [5] searched for the input mode of the bounded norm for the model and activated the selected remote unit to the maximum extent. The visualization of the calculation content of the team is realized in the input space. It helps people understand the special meaning of different neurons. The class activation map (CAM) method replaces the complete connection layer in CNN with the global average pooling layer [6]. By projecting the weight of the output layer to the convolution feature map, the core image region related to the label is identified. Then, the input samples' important regions with class discriminability are located. Reference [7] proposed a gradient weighted class activation map method (Grad-CAM). It can be used for any CNN model, and the calculation amount becomes smaller. The above methods provide a new way to improve the interpretability of DNN.

As we all know, power system transient stability assessment is a typical complex correlation map problem. It is difficult to build a map relationship between input features and system transient stability. Much work has been done on this issue. Reference [8] uses the generator active output, load active power, key line active power and other steady-state feature as inputs to build a CNN model. It realizes the stability evaluation of high accuracy, low misjudgment and low leakage. Reference [9] used trajectory analysis to build instability indicators, and used CNN to construct a composite neural network. Furthermore, the map relationship between the steady-state information of the system and the generator stability metrics is quantitatively described. This shows that DNN can effectively nonlinearly correlate the system's high-dimensional feature with the stability. However, when applied to

the DNN model of the power system TSA, the input samples are non-picture structures. This leads to the problem that the samples could be easier to understand intuitively. Reference [10] quantized the impact of input feature DT-based transient voltage stability model assessment results by SHAP index. To some extent, it explains the evaluation results of the model for each sample. Reference [11] obtained the input feature heat map using the Guided Grad-CAM algorithm and 1D-CNN model. It is found that node voltage has a more significant influence on TSA results than line power flow. These research results can provide new ideas for revealing the mechanism of transient stability.

In the face of a model with nonlinear solid fitting ability, it is essential to understand its internal working principle and turn it into a "gray box" or even a "white box". Enhancing the interpretability of the DNN classification model can be divided into the following two parts: one is to strengthen the understanding of input samples, and the other is to understand the extracted features of the model. The recognition and enhancement of DNN model features can enhance the knowledge of the weight of each layer of the model and improve trust in the model. It can also help engineers and technicians improve the model's structure and parameters and promote the popularization of deep network model in practical projects.

Under the background that deep neural network has achieved excellent results, aiming at the problem of poor interpretability of the model, this article adopts a DNN feature recognition method. Taking the CNN model as an example, we use the Grad-CAM algorithm to recognize and locate the features of the CNN model. It directed CNN to find the specific location of key data in the sample. Then, the activation maximization (AM) algorithm is used to construct sample features to explain the model. At the same time, a new input sample construction method is proposed. The training samples of CNN are constructed in the form of geographical wiring diagram, which enhances the readability of sample data and facilitates the visualization of features. Finally, the effectiveness and feasibility of this method are verified in the IEEE-39 system.

2. Feature Interpretability Analysis Method

The so-called interpretability refers to the ability to display in a human-understandable way. The core is how to understand the relationship between the feature of input samples and the output. Visualizing the components extracted from CNN models has become an effective tool to reveal the differences between CNN and human recognition. The feature information extracted from the trained CNN model implies the convolution kernel weight. The weights are difficult to understand, while the original sample features are understandable. Therefore, the components extracted by CNN from samples can be reconstructed under the actual sample dimensions using appropriate algorithms, and the regions where the main

features exist can be pointed out. This section will introduce two interpretable methods: feature recognition and feature enhancement algorithms.

2.1 Grad-CAM

The samples are input into a CNN model, and the time for the model to make judgments is very short and the accuracy meets the expectation. The operator wants to know what information CNN "sees" in the input sample during this process before making a corresponding judgment. To solve this problem, the Grad-CAM method can explain better.

Grad-CAM comes from improving CAM. CAM changes the fully connected layer to global average pooling (GAP) layer. Then CNN can provide a feature map for corresponding sample categories. The map k weight to the corresponding category is w_k^c in Fig. 1. By taking out the graph corresponding to the category, and then weighting and summing its corresponding feature graph, the final output is a class activation map. CAM can find features related to output classes. The larger the value of each point in map, the higher the attention of model to corresponding area. Finally, the location information of the features concerned by the model can be obtained.

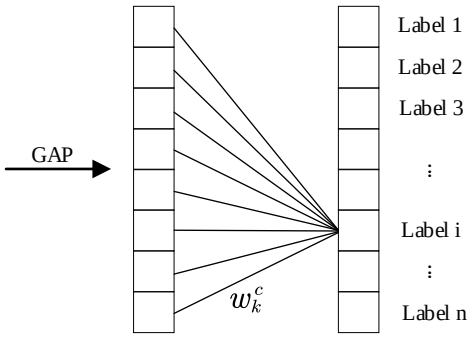


Fig. 1. CAM Schematic diagram

CAM needs to change the model structure and retrain it. This is time-consuming and laborious in practice. Therefore, Grad-CAM improved CAM. The result calculated is identical to that of the original CAM, as shown in (1).

$$w_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{1}$$

$$M^c = \sum_k \mathrm{Re}lu\left(w_k^c * A^k\right) \tag{2}$$

Where Z is the number of pixels; y^c is the score of category c; A^k is the k-th feature map; A_{ij}^k is the value in the (i,j) position; M^c is the calculated CAM of the category c. At the same time, the score y_i after softmax is related to all categories.

If only the gradient is calculated, it is assigned to the corresponding position after taking the absolute value. The result is also called a saliency map. It can indicate which small changes of data in the input sample can have a more significant impact on the CNN output score, that is, which data CNN is sensitive to.

The training process of Grad-CAM needs to be included. In addition, the results of the Grad-CAM output are guided by labels. For the same sample, different labels can make CNN view the original sample at various locations, while the saliency map does not have this feature.

Grad-CAM can show which part of the input sample CNN sees before making the corresponding prediction. It can check whether CNN pays attention to the correct area in the input sample. This method has achieved remarkable results in the field of image recognition.

2.2 Activation Maximization

It is challenging to analyze the interpretability of CNN by reducing the dimension of high-dimensional information in the convolution kernel. Therefore, a new method is used to visualize features, namely activation maximization. It can visualize the optimal input of each layer of neurons. The optimal input is the input sample that can make the activation value of the model output larger. "The activation value should be as high as possible" can be interpreted as "the most likely" for the layer. As a result, the optimal input could reveal which features the chosen neuron might have understood. The idea of AM is very intuitive. For a trained network, the optimal input can display the CNN extracted features in the dimension of input samples. This feature display is not achieved by dimensionality reduction of high-dimensional features, which avoids the dimensionality reduction process of high-dimensional features. Instead, a "most expected" input sample of CNN is gradually generated through iterative training. In this input sample, features must be extracted by the network, and the objective function:

$$x^* = \arg\max_x a_{i,l}(\theta, x) \tag{3}$$

The training process of AM:

(1) Create an initial input sample of random numbers and feed it into CNN, and forward propagation be done.

(2) Utilize backpropagation to determine the gradient $\frac{\partial a_{i,l}}{\partial x}$ of the active value relative to the input.

(3) Update input:

$$x^{(n+1)} = x^{(n)} + \eta \cdot \frac{\partial a_{i,l}\left(\theta, x^{(n)}\right)}{\partial x^{(n)}} \tag{4}$$

Where η is the learning rate.

(4) Repeat steps (1), (2) and (3) until there is no noise data in the input or the maximum amount of iterations is completed.

According to the iteration process of AM, we can infer that while presenting the characteristics, AM also somewhat enlarges them. This is so that the iterative process' objective of maximizing the activation value can be achieved, and some feature in the original sample makes the activation size of the feature map smaller than the maximum activation value. A higher activation value indicates that the sample's features are far more noticeable.

Generally, a convolutional neural network with deep structure has convolution, pooling, full connection layer and output layer. The full connection layer contains information related to all categories, which is difficult for humans to understand and visualize. For different CNN output results, we want to know what kind of samples CNN "most expects" to input. Therefore, this method is applied to the output layer of CNN (before softmax), which may give a reasonable explanation to the classification results of convolutional neural networks.

Then, the iterative training's objective function changes when the output layer is used:

$$\arg\max_{x} S_c(x) - \lambda \| x\|_2^2 \tag{5}$$

Where $S_c(x)$ is the score of category c; To guarantee that the final result is as similar to the original sample as feasible without becoming overly abstract and challenging to understand, regularization parameter λ is used. The reason for taking the score before softmax is that the maximum score after softmax may be achieved by minimizing the score of other categories, so focus on $S_c(x)$ to verify that category c is the only optimization target for all efforts and have nothing to do with other categories. Regularization parameters are introduced to govern the output, thereby making the final output more natural because of results acquired by using this method to deepen CNN will be more abstract and challenging to comprehend.

In addition, an initial image can be selected to replace the sample initialized by noise data initially set by the algorithm. In this way, the initial sample can be used as a guide to add features learned by the CNN model. If the initial image is completely irrelevant to the content learned by CNN, the original image will be given new features. This process is often used as style transfer in the field of image recognition, that is, to convert the image from the original style to another style, while ensuring that the main content of the image does not change.

In actuality, for the convolutional neural network-based transient stability assessment model, AM is used to produce a sample that CNN considers the most "stable" and the most "unstable". And depending on the colors shown in the sample pictures, we can find the feature of the system trend. This can provide great help for the power system TSA.

3. Sample Construction Method for Improving Model Interpretability

As we all know, the construction of input information matrix has a significant impact on the model's performance. For power system transient stability analysis, the power flow before failure reflects the power operating point and offers extensive data on transient stability. Therefore, for the CNN model adopted in this article, the power flow data is selected to construct the CNN input samples.

3.1 Construction of input sample matrix based on node connection relationship

Table 1

Feature Variables in Power Flow Information

Variable types	Electrical parameters
Generator	active power output P_G, reactive power output Q_G
Load	active power P_{Load},reactive power Q_{Load}
Line/Transformer	Head end power P_{Line}、Q_{Line} voltage angle difference at two ends of the branch. $\Delta\theta = \theta_i - \theta_j$
Node	node voltage amplitude U_m, voltage phase-angle θ_i

For a power system with *N* nodes, the sample matrix can be constructed from the power flow information in the form of an admittance matrix. The row and column labels of the matrix correspond to the node number one by one. Information about nodes is represented by the matrix's primary diagonal elements, while the upper and lower triangular elements represent branch information. Different electrical parameters can be respectively constructed into admittance like matrices and stacked into three-dimensional matrices. The feature variables included in the power flow information are shown in Table 1.

The power flow information mainly includes active power, reactive power and voltage, so the dimension of sample matrix *F* is *N* * *N* * 3. The specific construction method is shown below.

1. The active power injected by each system bus is represented by the matrix's first layer and the active power transmitted by branches.

$$F(i,i,1) = P_{Gi} - P_{\text{Load } i} \quad i = 1,2,\ldots,N \tag{6}$$

$$F(i,j,1) = P_{\text{Line } (i,j)} \quad i,j = 1,2,\ldots,N \quad i \neq j \tag{7}$$

Where $P_{Line(i,j)}$ represents the active power at the head end of the branch connecting node *i* and *j*, and the node with a small number is the head end.

2. The second layer of the matrix represents the reactive power injected by each node of the system and the reactive power transmitted by branches.

$$F(i,i,2) = Q_{Gi} - Q_{\text{Loadi } i} \quad i = 1,2,\ldots,N \tag{8}$$

$$F(i,j,2) = Q_{\text{Line } (i,j)} \quad i,j = 1,2,\ldots,N \quad i \neq j \tag{9}$$

Where $Q_{Line(i,j)}$ represents the reactive power at the head end of the branch connecting node *i* and *j*, and the node with a small number is the head end.

3. The third layer of the matrix represents the bus voltage amplitude and the voltage angle difference at two ends of the branch.

$$F(i,i,3) = U_{mi} \quad i = 1,2,\ldots,N \tag{10}$$

$$F(i,j,3) = \Delta\theta_{ij} \quad i,j = 1,2,\ldots,N \quad i \neq j \tag{11}$$

The input sample data generated by the above method contains almost all the power flow information. The network topology is hidden in the data, and the data is complete and the physical meaning is clear. It is also clearly separable in computer vision, which facilitates CNN to select data and extract features. However, we will find that more data in the samples constructed in this way are concentrated near the main diagonal, and the data is relatively dense. This makes it more difficult for people to understand the sample and analyze the interpretability of the CNN model. Therefore, we need a convenient sample construction method for CNN interpretability analysis.

3.2 Sample matrix construction method based on geographical wiring diagram

A sample building technique based on the geographical wiring diagram is suggested in order to use CNN's image recognition capabilities and to simplify the interpretability study. An RGB image serves as a representation of each power flow sample (3D matrix), in which circles are used to represent node and branches. The power flow information is entered with in diagram based on such a system wiring design. The pixel value of each channel is [0, 255]. C_N is the pixel value of each channel of the node, and C_L is the pixel value of the branch. Since the picture pixel contains *R*, *G*, and *B* channels, the following method is used to convert the power flow data to the pixel value.

1. The active power that node *i* injects and branch *j* transmits is known as the *R* channel.

$$C_{Ni} = P_{Gi} - P_{\text{Load } i} \quad C_{Lj} = P_{\text{Line } j} \tag{12}$$

2. The reactive power that node *i* injects and branch *j* transmits is known as the *G* channel.

$$C_{Ni} = Q_{Gi} - Q_{\text{Load } i} \quad C_{Lj} = Q_{\text{Line } j} \tag{13}$$

3. The voltage amplitude at node *i* and the phase angle difference between two ends of branch *j* are both represented by the *B* channel.

$$C_{Ni} = U_{mi} \quad C_{Lj} = \Delta\theta_j \tag{14}$$

4. Normalize the above converted values to [0,255].

$$C' = \frac{255}{C_{\max} - C_{\min}} \times \left(C - C_{\min}\right) \tag{15}$$

Where C' is the normalized pixel value, C is the non normalized pixel value, and $C_{\max}$ and $C_{\min}$ are the maximum and minimum values of the same category variable (C_N or C_L) respectively.

To simplify the interpretation process, reducing the number of variables can be beneficial for stable labels. Hence, a single fault is selected as the fault set, with the category, location, and time of each sample remaining unchanged. In this scenario, the stability label is determined based on whether the system transient is stable or not. In this way, each sample is drawn as an RGB image. With the help of this sample creation technique, features can be visualized and their interpretability can be examined after the information in the system topology has been fully displayed. A topological diagram illustrating the relative positions of the network nodes is depicted in Fig. 2.

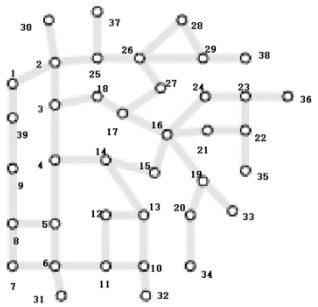


Fig. 2. System node label comparison

4. Case Studies

The IEEE-39 system is used to evaluate the performance of the proposed model. Simulation experiments are performed in PSASP. The generators are the 2th order model and the loads are the constant impedance model. The fault is set to a three-phase short circuit, assumed to occur on bus 18 and cleared after lasting 0.1s. All the loads are set between 80% and 120% of the original load levels, respectively and the power of generations is also scaled in the same proportion. The total

simulation time is 5s. The system's topology does not change before and after the short circuit. A total of 11421 samples are obtained from simulation, of which 7845 are stable and 3576 are unstable. The training set and testing set are randomly divided according to the ratio of 4:1. The deep neural network model is built based on the PyTorch 2.0 framework. The structure of the model, as well as the detailed parameters of the convolutional layers and pooling layers, are depicted in Fig. 3.

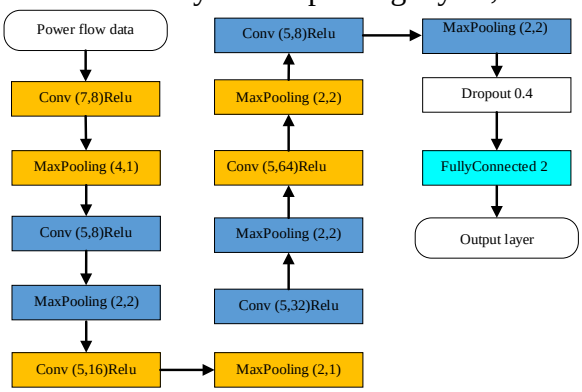
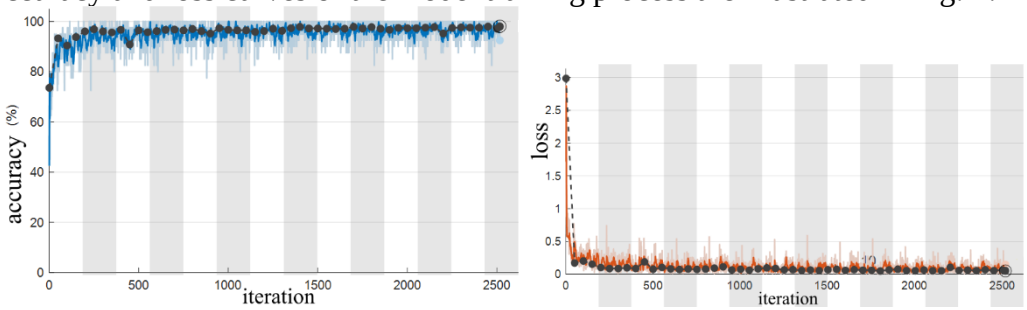


Fig. 3. Transient stability evaluation model based on convolutional neural networks

The training set is utilized for the iterative training of the CNN model, while the testing set is employed to evaluate and verify the model's performance. The accuracy and loss curves of the model training process are illustrated in Fig. 4.



(a) accuracy (b) loss

Fig. 4. Accuracy and loss curves of model training process

To evaluate the transient stability assessment performance of the deep learning model constructed in this case study, the metrics of accuracy and miss rate, as indicated by equation (16) and (17) respectively, are employed.

$$I_{ACC}\% = \frac{T_P + T_N}{T_P + F_N + F_P + T_N} \times 100\% \quad (16)$$

$$I_{MAR}\% = \frac{F_P}{F_P + T_N} \times 100\% \quad (17)$$

In the equation, I_{ACC} represents the accuracy metric, and I_{MAR} represents the miss alarm rate metric. T_P represents the number of true positive samples correctly classified by the model; T_N represents the number of true negative samples correctly classified by the model; F_N represents the number of false negative samples, where positive samples are misclassified as negative by the model; F_P represents the number of false positive samples, where negative samples are misclassified as positive by the model.

Table 2

Performance Comparison of Different Models

Model	I_{ACC}% (entire dataset)	I_{ACC}% (test set)	I_{MAR}%
SVM	95.03%	94.88%	4.78%
DT	96.50%	96.91%	2.61%
BP	96.32%	95.46%	1.30%
CNN	97.76%	97.08%	0.61%

Table 2 presents the performance of the convolutional neural network in transient stability assessment. Compared to other machine learning models with default parameters, the CNN achieves a classification accuracy of 97.08% on the testing set, demonstrating high performance in transient stability assessment. This indicates that the CNN can effectively identify and extract spatial correlations between variables represented in the form of power flow diagrams through two-dimensional spatial convolutions.

4.1 Results and analysis of crucial feature identification

Taking an unstable sample in the testing set as an example, we use Grad-CAM and CAM to generate the thermodynamic diagram, as shown in Fig. 5. It can be seen that there is little difference between the two, which indicates the effectiveness of Grad-CAM.

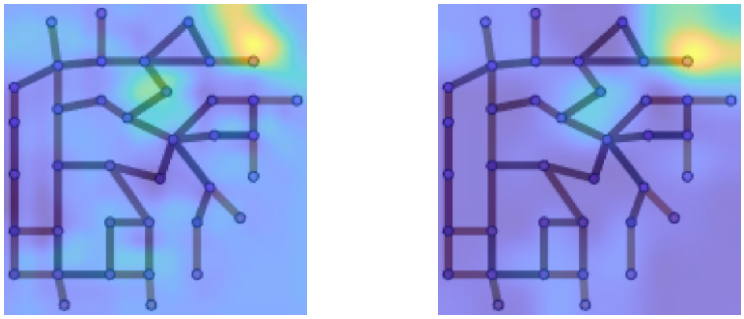


Fig. 5. The CAM (left) and Grad-CAM (right) results

We found that the CNN model will first and focus on the 38 node generator. This shows that the generator data has a very powerful feature, and CNN can conduct transient stability assessment based on its significant feature when facing instability samples.

As the 38 node is PV bus, the voltage is a fixed value, the active power is controlled by the system generation load proportionally, and the reactive power and voltage phase angle are calculated by the power flow. Therefore, we plot the reactive power of 38 node generators in the data set according to their transient stability labels in Fig. 6. The abscissa represents the reactive power output by the generator. At the same time, the ordinate has no physical meaning, just to evenly distribute the plotted points in the plane.

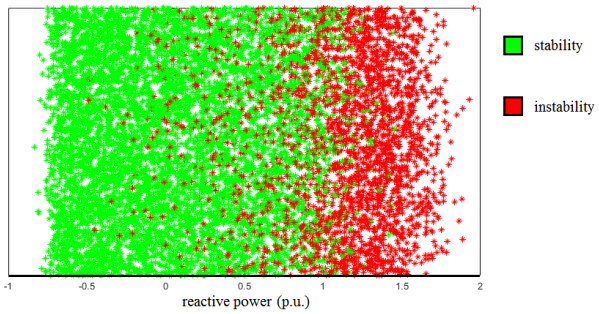


Fig. 6. Corresponding diagram of reactive power of 38 node generator and stability label

It can be seen from Fig. 6 that the reactive power of 38 node generator has excellent classification effect. When the reactive power is more significant than 1.2 p.u, the sample may be transient instability. When it is less than 1.2 p.u, CNN needs to consider other feature variables for stability judgment. This is consistent with the information obtained in the comparison chart in the appendix.

It can also be seen from Fig. 5 that when CNN evaluates the stable samples, although most of the strong activation areas of the samples are still 38 node, CNN also focuses on 9, 14, 15, 21 nodes, and 31, 32 node generators. These regions are just weak activation regions in the sample activation graph. This shows that the data concerned in the CNN stability judgment is multivariate, and the information contained in the weak activation region of the feature map can also play a key role in the process of CNN output results. In other words, compared with areas without valid data, because some regions in the activation map are strongly activated and some regions are weakly activated, CNN can make prediction judgments based on this.

In addition, we also found that CNN pays more attention to bus parameters than line data. On the one hand, there is redundancy between various data in such a large sample. On the other hand, as the key component of the system, the parameters of generator and load are reflected on the bus.

4.2 Results and analysis of maximizing activation

The optimal input of the CNN model generated by AM algorithm is shown in Fig. 7 and Fig. 8. The diagram is superimposed with the original system wiring diagram to facilitate the search for specific feature information. In the dimensions of the first input samples, we can see the universal traits CNN extracted and emphasized.

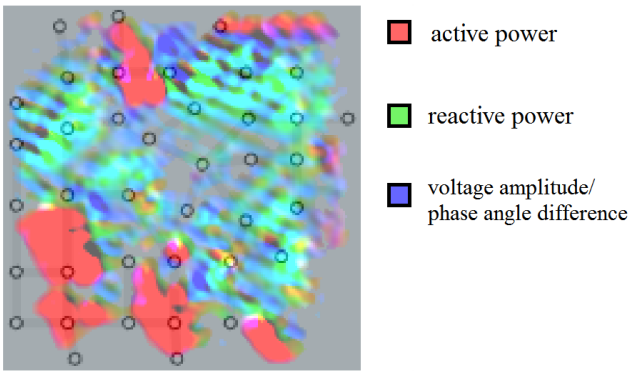


Fig. 7. Preferred input of instability label

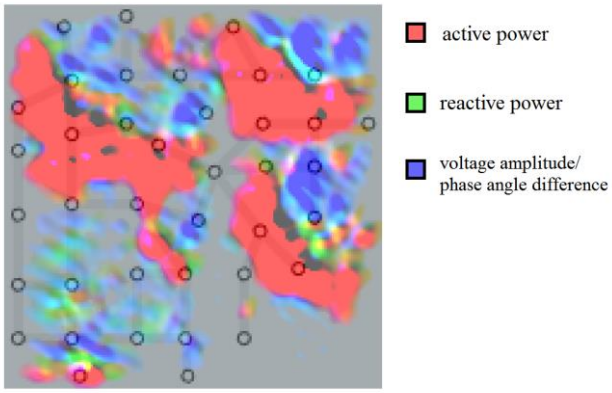


Fig. 8. Preferred input of stability label

Here, the output of AM should be described as follows: 1) The color represents the category of electrical parameters, that is, red represents active power, green represents reactive power, and blue represents voltage amplitude/phase angle difference. 2) The color depth represents the numerical value of electrical parameters.

In the training process, the weights of each parameter of the CNN model are fixed values, and the AM output results contain typical features extracted by CNN. As the input samples are drawn based on geographical wiring diagram, the features reflected by AM generation results are associated with their specific locations. CNN believes that the optimal inputs of stability and instability are significantly different,

with obvious differences in categories and regions. There are the following differences:

1. For the preferred input of the instability label, 30, 31, 32 and 35 nodes turn yellow, the color of the pure load node is very light, and that of 3 and 18 nodes are light. This indicates that CNN believes that the reactive power output of generators at 30, 31, 32 and 35 nodes in the instability sample are too large, and the overall load level of the system is too large;

2. For the preferred input of the stability label, the color near the 6 node is blue. Except for the load centers represented by 16, 21, and 24 nodes, the red color of other pure load nodes are darker. This indicates that the stability features extracted by CNN are that the voltage of the 6 nods is slightly higher and the active power demand of the load outside the load center is smaller.

It can be seen from the above comparison differences that when the active output of the system balancer is large and the reactive load rate of the branch is too high, the voltage of each node will be reduced, which is easy to lose stability. It shows that there must be some relationship between power angle instability and voltage. If the active output of individual generator is large, the energy injected into the system after failure will be significant and the acceleration area will be large, which is not conducive to the system's stability. When the system voltage is slightly high, the power flow to the fault point is small, and the energy injected into the system is small, so instability is not easy to occur.

The actual correspondence between the system-related feature variables and the stable labels is shown in Fig. 9 and Fig. 10. This also proves that CNN has a robust feature extraction capability and AM has a feature amplification effect.

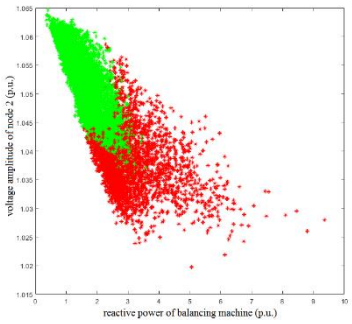


Fig. 9. Mapping of 2 node voltage amplitude, balancing machine reactive power and stability labels (green indicates stability, red indicates instability)

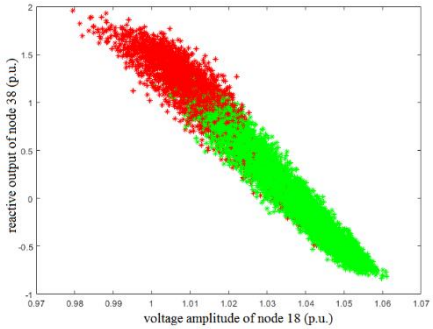


Fig. 10. 18 node voltage amplitude, 38 node reactive output and the corresponding diagram of the stability label (green indicates stability, red indicates instability)

5. Conclusion

In this article, we utilize the Grad-CAM and AM algorithms to analyze the feature visualization and interpretability of the model. We aim to uncover the preferred input for CNN and explore the relationship between strong and weak activation in the activation graph and the model output.

The proposed method in this article is applied to the IEEE-39 system, and the interpretability analysis yields the following conclusions:

1. Upon observing the thermal diagram generated by Grad-CAM, we discover that CNN primarily focuses on the reactive output of 38 node generators when evaluating instability samples. However, in cases where this feature is not prominent, the model shifts its attention to data from nodes 17 and 27.

2. When CNN assesses stability, it examines both the regions with strong and weak activations, which serve as the basis for evaluation. The model's accurate predictions are possible due to the presence of strong activation in one area and weak activation in another, as depicted in the activation map.

3. Through the analysis of the optimal input obtained by AM, we ascertain that the node information of the system plays a more critical role in CNN's transient stability assessment compared to the branch information. Notably, the active and reactive power output of the balancing machine, the reactive power output of certain generator nodes, the voltage amplitude of node 2, and the load level near the fault point exhibit significant classification effects. This demonstrates that CNN effectively captures the nonlinear mapping relationship between power flow and transient stability.

Acknowledgment

This paper was supported in part by the National Natural Science Foundation of China (Key Project Number: 51877034).

REFERENCES

[1]. Li Bingzhen, Liu Ke, Gu Jiaojiao, et al. Review of the researches on convolutional neural networks. Computer Era, 2017, 40(06):1229-1251.
[2]. Ji Shouling, Li Jinfeng, Du Tianyu, et al. Surveyon Techniques, Applications and Security of Machine Learning Interpretabilit. Journal of Computer Research and Development, 2019, 56(10): 2071-2096.
[3]. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv: 1312. 6034, 2013.
[4]. Zeiler M D, Fergus R., Visualizing and understanding convolutional networks. European conference on computer vision. Springer, Cham, 2014: 818-833.
[5]. Couteaux V, Nempont O, Pizaine G, et al. Towards interpretability of segmentation networks by analyzing Deep Dreams. Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support. Springer, Cham, 2019: 56-63.
[6]. Zhou Bolei, Khosla A, Lapedrize A, et al. Learning deep features for discriminative localization. Proc of the 28th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ; IEEE, 2016: 2921-2929
[7]. Selvaraju Ramprasaath R. et al., Grad-cam: Visual explanations from deep networks via gradient based localization. Proceedings of the IEEE international conference on computer vision. 2017.
[8]. Tian Fang, Zhou Xiaoxin, Shi Dongyu, et al. Power System Transient Stability Assessment Based on Comprehensive Convolutional Neural Network Model and Steady-state Features. Proceedings of the CSEE, 2019, 39(14): 4025-4032.
[9]. An J, Yu J, Li Z, et al. A Data-driven Method for Transient Stability Margin Prediction Based on Security Region. Journal of Modern Power Systems and Clean Energy, 2020, 8(6):1060-1069.
[10]. Zhou Ting, Yang Jun, Zhan Xiangpeng, et al. A data-driven method for transient voltage stability assess-ment and its interpretability analysis. Automation of Electric Power Systems, 2021, 45(11): 4416-4425.
[11]. Zhao Kai, Shi Libao. Transient stability assessment of power system based on improved one-dimensional convolutional neural network. Power System Technology: 1-14.
[12] Qin Z, Yu F, Liu C, et al. How convolutional neural network see the world - A survey of convolutional neural network visualization methods.2018.