



Progress in Interpretability Research of Convolutional Neural Networks

Wei Zhang^{1,2}, Lizhi Cai^{1,2}, Mingang Chen^{2(✉)}, and Naiqi Wang^{1,2}

¹ School of Information Science and Engineer,
East China University of Science and Technology, Shanghai, China
zweing1995@163.com, slytherinwnq@163.com,
clz@ssc.stn.sh.cn

² Laboratory of Computer Software Testing & Evaluating, Shanghai
Development Center of Computer Software Technology, Shanghai, China
{clz, cmg}@ssc.stn.sh.cn

Abstract. Convolutional neural networks have made unprecedented breakthroughs in various tasks of computer vision. Due to its complex nonlinear model structure and the high latitude and complexity of data distribution, it has been criticized as an unexplained “black box”. Therefore, explaining the neural network model and uncovering the veil of the neural network have become the focus of attention. This paper starts with the term “interpretability”, summarizes the results of the interpretability of convolutional neural networks in the past three years (2016–2018), and analyses them with interpretable methods. Firstly, the concept of “interpretability” is introduced. Then the existing research achievements are classified and compared from four aspects, data characteristics and rule processing, model internal spatial analysis, interpretation and prediction, and model interpretation. Finally pointed out the possible research directions.

Keywords: Convolutional neural networks · “black box” · Interpretability

1 Introduction

The concept of artificial neuron was first proposed in the 1940s. After decades of research, Yann LeCun designed and trained LeNet-5 model — classic CNN structure, in 1998, giving the basic component and framework structure of Convolutional Neural Network. Later, neural networks such as AlexNet, VGGNet, GoogLeNet, ResNet, DenseNet appeared, and they became deep and complex. These convolutional neural networks have achieved unprecedented breakthroughs in various tasks of computer vision, such as image classification, semantic segmentation, target detection and visual problem answering.

Although these neural networks have been successful in various scenarios, the entire network lacks intuitive and understandable components, making the results of the network model difficult to interpret. In particular, the application of neural network in the fields of medicine, financial markets, criminal justice, etc., interpretability is an extremely important standard for model evaluation, and has become the most worrying “black box”.

Therefore, how to understand the interpretability of neural networks better is a common concern of academia and industry.

2 An Overview of “Interpretability”

As for interpretability, there is no strict mathematical symbol definition and no general (non-formula) literal definition. However, with the development of artificial intelligence, it is particularly important to study the interpretability of models. In general, it is far from enough to obtain simple prediction results for models from low-cost general fields (such as commodity recommendation) to high-cost key fields (such as finance and medical treatment). People began to pay attention to how the model made predictions.

The PhD student Leilani Gilpin from MIT’s Computer Science and Artificial Intelligence Lab (CSAIL) has published a paper [1] that analyzes “interpretability” and several related semantic approximate terms, classifies the current machine learning model interpretability methods, and puts forward the evaluation of interpretability methods. Gilpin informally defines “interpretability” as understanding what the model does or has done. This paper discusses the difference between “explanation” and “interpretability”. In a word, the model with interpretability can be interpreted by default, but not vice versa. The proposed interpretative understanding is divided into three types: (i) was proposed some explanation, while the key to this explanation does not represent a model will make the decision making process, but can provide a certain degree of reason to make a choice; (ii) was the representation of data in the network; (iii) was the establishment of a network model that generates interpretation.

In the 2nd ICML 2017 Workshop on Human Interpretability in Machine Learning (WHI), Google brain senior research scientist, Been Kim [2] reported on the interpretability study of machine learning and provided a preliminary understanding of the “interpretability” study of the AI model. This is a tutorial report that shows what is interpretability, why interpretability, and what we can do on interpretability. She said, Interpretation is the process of giving explanations To Humans. Comparing the AI model with traditional software shows that the AI model also needs security, debugging, principle support, iterative optimization and fairness and legality. Been Kim divided the third question into three aspects: pre-modeling, modeling time, and post-modeling. For example, consider data distribution before modeling, consider feature functions in modeling process, and consider hidden layer information in model completion.

Dr. Zachary C. Lipton, of the University of California, San Diego, and assistant professor of computer science at Carnegie Mellon University, shared a report on “The Mythos of Model Interpretability” [3] on the ACM Queue and discussed the interpretability of the supervised machine learning model. Lipton said that people have realized the importance of interpretability for a model, especially in key areas such as medicine, criminal justice systems, and financial markets. He believes that the results of the interpretable analysis of the deep model from the current academia can be seen that people generally agree with the term “interpretability”, but there is absence of a definition. In other words, the meaning of “interpretability” is unclear, so that there are various papers that claim to be interpretable after optimizing a model or building a model. Such an article may interpret the model based on different starting points,

leading to such a vague situation. Lipton divided the work of “interpretability” into two categories by analyzing the need for interpretability research. The first relates to transparency, i.e., how does the model work? The second consists of post-hoc explanations, i.e., what else can the model tell me? Finally, in order to standardize the “interpretability” study, he proposed that the interpretability study of the model should achieve one of the above two as a specific goal.

3 Convolutional Neural Networks “Interpretability” Research

Combining the above researches, in this paper, the “interpretability” of the convolutional neural network model (hereinafter referred to as the model) is summarized into four aspects (As shown in Fig. 1):

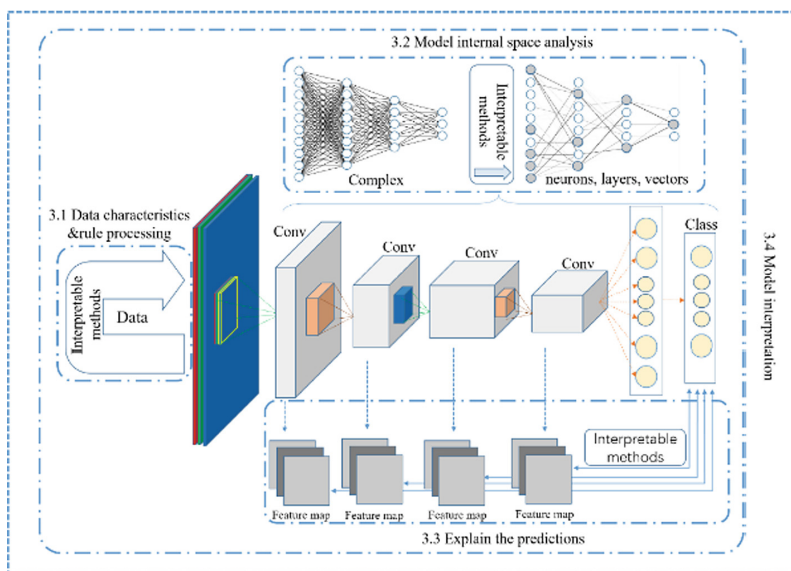


Fig. 1. Interpretable research structure diagram

- Data characteristics and rule processing. Initial exploration of model data or adding some known rules to the model (see Sect. 3.1 for details).
- Model internal space analysis. By analyzing the internal components of the model, such as unit level, hierarchical analysis (see Sect. 3.2 for details).
- Explain the predictions. Focus on the analysis of the results of the model, that is, post-hoc explanations (see Sect. 3.3 for details).
- Model interpretation. Based on the entire model, such as model simulation, construction of interpretable systems (see Sect. 3.4 for details).

3.1 Data Characteristics and Rule Processing

Data Characteristic. [4] has proposed feature selection can help to build better models with finer data. Removing the unrelated and redundant attributes can reduce the complexity of the model, so that the model can be understood and explained. When understanding a model, the first starting point is the characteristics of the original data. [5, 6] analyzed the influence of the original data features on the model interpretability from medical image data and material microstructure image data. [5] compared the results of the original image scaling $\times 1/2$, $\times 1/4$ and $\times 1/8$ post-resolution pairs, indicating that the difference in the details of the two images provides an explanation for the prediction. However, the experimental comparison can only confirm that the detail features in the image can increase the prediction accuracy, but it is difficult to explain the influence of the details on the model decision. [6] used CNN to extract micro-texture features on Titanium, Steel, and Powder dataset images, and discussed the generalization and classification features between datasets when convolutional neural networks are used for microscopic image classification.

The selection of data features as a specific method of Model interpretation [7]. Based on the maximization of mutual information between selected features and response variables, a function model based on learning method is established to extract the feature subset with the largest amount of information in each given example. Then, an importance score is assigned to a given instance prediction result for each feature, allowing the relative importance of each feature to vary from instance to instance.

Rule Processing. Traditional machine learning is generally considered to be more suitable for interpretation with rules, and Boolean rules are one of the simplest interpretable classification models [8]. For the depth model, some optimized rule-based methods are equally applicable.

Rule-Based Extraction. When a known model is built according to a priori rule, it is theoretically easy to understand the model. On the contrary, the decision process of the model can be studied by extracting rules from the model. Rule extraction can be divided into (i) decomposition-based methods; (ii) model-agnostic methods (for machine learning models, not discussed here). The former, for example, the DeepRED [9] algorithm that is able to extract rules from deep neural networks. The basis of this method is the CRED [10] that contain both continuous (real-valued) and discrete literals. This decomposition algorithm used the decision tree to describe the behavior of the hidden layer elements of the NN. DeepRED extracted intermediate rules for each layer of the DNN through the CRED algorithm, then merged the previously generated rules and generated behaviors(rules) describing the DNN through input data.

Embedding of Prior Rules. A priori rules is embedded in the NN to explaining the model. [11] proposed a rule embedded neural network (ReNN) to cope with the shortcomings of neural network. ReNN breaks down the “black box” of ANN into two parts: local-based reasoning (local patterns learned from data sets) and global-based reasoning (a priori knowledge of human long-term accumulation) (As shown in Fig. 2). Through the local inference mode and rule analysis of the ReNN, the entire network is better interpretable.

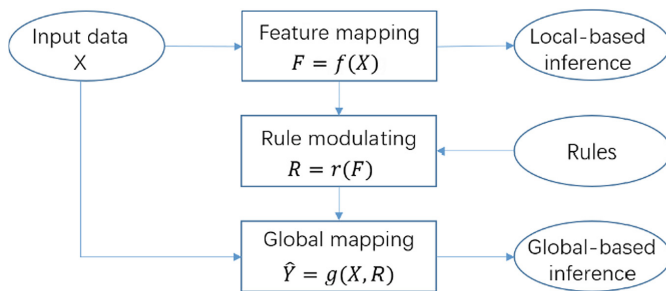


Fig. 2. Computational graph of ReNN(adapted from [7]).

3.2 Model Internal Space Analysis

The classical CNN structure given by Yann LeCun usually consists of input layer, hidden layer and output layer. Understanding the neural network from such a model structure requires an analysis of the roles of its components. It is divided into neural unit, neural network layer, hierarchical neuron combinations and perturbation-based models interpret four aspects of analysis.

Component Analysis Based on Neural Unit. On the analysis of neural network neurons, on the one hand, under what conditions a single neuron is activated, on the other hand, when a neuron is activated, it expresses information. The method displays the sensing area of the activated neuron by maximizing the activated input image and highlights a particular portion of the neuron image used to activate the convolutional layer through the deconvolution network.

[12, 13] both adopted the activation maximization method to analyze the information contained in the neuron, that is, to find the optimal stimulus of each unit by performing gradient descent to maximize the activation of the unit, mainly calculating the input sample when the activation of the i th neuron in the j th hidden layer is maximized. The downside is that the complex input distribution will fail. The latter optimized it and proposed that a single neuron can detect multiple characteristics (color, size and direction) on the original basis, while the existing maximum activation method only considers one of them. Therefore, the algorithm is proposed to synthesize the multi-aspect information that each neuron can express into the sample activation image through the activation method, which can more fully understand the function of each neuron.

The other [14], which adds a deconvolution operation (convolution operation is carried out on the filter with both horizontal and vertical directions reversed) on each convolutional layer of the classification CNN to visualize the image region activated by each neuron. In this paper, 9 images with the highest activation value are shown after convolution of each feature image. It can be seen that each feature map is “interested” in different images.

[15] evaluated the consistency between individual neural units and the quantified interpretability of visual semantic concepts (color, material, structure, parts, objects and scenes). And indicated that neural units is assigned different identifiable labels.

Hierarchical-Based Component Analysis. The formula $y = h(\omega \cdot x + b)$ for each layer of the neural network is the transformation of the input vector x to the output vector y . Where, $\omega \cdot x$ represents lifting dimension, scaling and rotation, b represents translation, and the function $h(x)$ represents distortion, namely the transformation of linear and non-linear matrix space is completed. The graphical explanation can be seen here [16, 17]. From the perspective of classification neural network, a hyperplane is found in the space after the linear transformation of the original space through the nodes in each layer and the nonlinear transformation of the activation function. This is explained by a operation from each layer of the neural network.

As for the expression of each layer of the neural network, [18] proposed that each layer of the classification neural network recognized the distribution of each category in the two data sets of ImageNet-CNN and Places-CNN, as well as the detection of an object by a single neuron. [19] illustrated the transferability of neural networks, quantifies the comparison between the universality and specificity of each layer of deep convolutional networks, and two factors affecting its portability are found: fragile coadaptation of middle layers and specialization of higher layers.

Component Analysis Based on Hierarchical Unit Combination. Instead of studying individual neurons or the concept of layers in a neural network, exploring linear combinations of hierarchical units brings new perspectives.

On the basis of theory, [20, 21] proposed different concepts. The former mapped semantic concepts to vectors based on the corresponding filter response. Analysis model internal filter proof: (i) In most cases, need more than one filter to code a concept; (ii) Not a single filter specific to a concept; (iii) For single filter activation, filter embedding can better represent the meaning of the representation and its relationship with other concepts. The latter proposed two counter-intuitive properties of deep neural network. [21] found that it is the space, rather than the individual units, that contains the semantic information in the high layers of neural networks. This provides a new point of view with the general understanding of neural networks. At the same time, this paper also proposed the existence of adversarial example in neural networks.

Subsequently, [22] introduced the concept activation vector(CAV), and represented that the model interpretation is formally expressed as a set of model state space vector Em and a set of unknown human understandable concepts Eh . The model interpretation can be expressed as a mapping relationship $g:Em \rightarrow Eh$. As a way of conversion between Em and Eh , a set of human understandable input data examples are defined as concepts. The relative importance of concepts to classification is quantitatively analyzed to explain the neural network. [23] proposed each neuron in DNN is interpreted as an activation vector whose value is the scalar output generated by it on the input data. By collecting two groups of neurons and then outputting the alignment feature, it can be seen that the potential representations acquired by the two networks have similar characteristics. The advantages of this method are: one is to compare the representations learned by the two neural networks, and the other is to explain the representations of DNN hidden layer learning.

Internal Space Analysis Based on Perturbation. The perturbation in the neural network is not to delete or modify the model structure, but to input the processed test samples and then observe the prediction results of the neural network. The specific

processing methods include occlusion experiment, noise study and adversarial sample study. [14] studied which region of the image has the largest effect, the experiment used a gray square to cover different positions of objects and then monitored the output of the classifier. The results are consistent with the results of human cognitive knowledge, that is, the key position in line with human knowledge will have a greater impact. [24] studied the effects of noises on the interpretation of neural networks. The deep Taylor decomposition is used to show the interpretation results of different interpretation rules in response to noise. [25] found out which part of the image has the greatest influence on its output score when disturbed, so as to understand the search position of the algorithm. [26] proposed a new scoring formula on the basis of antagonistic samples and characteristic scores. Based on the adversarial example, seeking the minimum data perturbation of model input can identify the important input characteristics and the minimum allowable data perturbation by looking for the maximum data perturbation that does not change the output. Among them, occlusion experiment is the most consistent with human cognition, but it has a strong artificial purpose. Some noise studies have achieved good results. Although the results of adversarial sample experiment are eye-catching, there are some deficiencies for human understanding.

[27] proposed LIME(Local Interpretable Model-agnostic Explanation), a novel explanation technique that explains the predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. Observing the predictive behavior of the model by perturbing the input samples (actually a sampling method), and then assigning weights based on the distances of the perturbed data points from the original data, based on which they learn an interpretable model and prediction results. The essence of perturbation is that these around disturbances must be understandable by humans.

3.3 Explain to Predict

This part focus on the results of the model, such as analyzing the reason why an image is classified into a certain category from a CNN, which is explained by the two aspects of feature and visualization. And, feature interpretation is divided into feature importance and feature text interpretation.

Feature Importance. The interpretation of feature importance is to evaluate the features concerned by the model, and then measure the importance of the features with scores. Compared with the perturbation-based method above, it is easier because each perturbation requires a forward propagation of the network, which is computationally inefficient.

DeepLIFT (Deep Learning Important FeaTures) [28], an algorithm for recursive predictive interpretation of depth models that assigns importance scores to inputs for a given output. The difference is that DeepLIFT uses backpropagation to calculate the scores, so they can be efficiently obtained by a single reverse network propagation. LRP (Layer-Wise Relevance Propagation) [29] achieved pixel-level decomposition, using a single pixel to evaluate the impact of sample images in a kernel classifier and neural network, and visualize it. This method is equivalent to performing a DeepLIF operation, activating all input reference values (DeepLIFT will set a reference value for

each input) to zero. [30] proposed the use of Shapley value to quantify the importance of characteristics of a given input, and proposed a sample-based method and “kernel SHAP” to approximate Shapley value. The commonality of the above methods is to use a local additional model to approximate the local model, and its inadequacy is also localized.

The difficulty with the importance of features is that it is difficult to evaluate them with experience. To compensate for this shortcoming, [31] proposed an integrated gradient approach. And two basic axioms that the attribution method should satisfy—sensitivity and implementation invariance. Integrated gradient is a new attribute method guided by these axioms. This method does not require any network tools, and can be easily calculated by a few calls to gradient operations.

For the deficiency of localization, [7] proposed the L2X (Learning to Explain) method, which learns the feature selection function different from the local approximation method of the previous function in the global scope, and takes the instantiated feature selection as the method of model interpretation. In particular, the importance score of each feature of an instance is given to indicate which features are the key for the model to predict on this instance.

Feature Text Interpretation. [32] focused on the description and interpretation of the recognition features, for example, when the neural network identifies a bird, it will give “this is a bird, because its beak is recognized” instead of “filter ith is activated at the highest level in the model”. Such an explanation would be more useful to non-professionals with no knowledge of modern computer vision. The paper proposes that such interpretations must meet two criteria: they must be class sensitive and accurately describe specific image instances (Table 1).

Table 1. Feature importance methods comparison

Method	Train	Efficiency	Locality	Model-agnostic
LIME [27]	No	Low	Yes	Yes
DeepLIFT [28]	No	High	Yes	No
LRP [29]	No	High	Yes	No
SHAP [30]	No	Low	Yes	Yes
Integrated gradient [31]	No	High	No	Yes
L2X [7]	Yes	High	No	Yes

Summary of the properties of different methods. “Training” indicates whether a method requires training on an unlabeled data set. “Efficiency” qualitatively evaluates the computational time during single interpretation. “Locality” indicates whether a method is locally additive. “Model-agnostic” indicates whether a method is generic to black-box models (adapted from [7]).

[33] interpreted the output results of the model by generating counterfactual explanations of text types afterwards. The counterfactual interpretation here refers to a description of a characteristic fact that distinguishes Category A from Category B, for example, “This is not a scarlet tanager because it has no black wings.”

Visualization. Further research into the interpretability of neural networks has shown that Deep Visualization is a good way to understand neural networks, and leads to the direction of Deep Visualization. These methods are mainly composite images, which can be divided into two directions: gradient-based method and network-based activation method.

Gradient-Based Approach. The Gradient Explanation of the input x is $E_{grad}(x) = \frac{\partial S}{\partial x}$. The gradient quantifies the extent to which a change in each input will change the predicted value $S(x)$ in its neighborhood. By iterating the gradient of the objective function and updating the random input x , the original image can be reconstructed in reverse [34] or the image that maximizes the score for a certain category can be realized [35].

[36, 37] proposed DeConvNet and [38] proposed Guided Backpropagation (GBP) based on the gradient method where negative gradient entries are set to zero while back-propagating through a ReLU unit to generate a clearer visualization.

[31] combined the axioms of previous research to guide a new approach, called Integrated Gradients (IG). With summing over scaled versions of the input solves gradient saturation. IG for an input x is defined as $E_{IG}(x) = (x - \bar{x}) \times \int_0^1 \frac{\partial S(\bar{x} + \alpha(x - \bar{x}))}{\partial x} d\alpha$, where \bar{x} is a “baseline input” that represents the absence of a feature in the original input x .

[35, 38] demonstrated that the gradient could be used for extracting a saliency map of an image. However, they also tend to be noisy, covering many irrelevant pixels and missing many relevant ones. SmoothGrad [39] achieved the denoising effect by adding noise to the image, then sampling the similar image, and average the sensitivity map of the sampled image. Take an input x and average the resulting sensitivity maps E , $E_{SG}(x) = \frac{1}{N} \sum_{i=1}^N E(x + N(0, \sigma^2))$, where $N(0, \sigma^2)$ represents Gaussian noise with standard deviation σ .

Above classifier-dependent saliency maps can be utilized to analyze the inner workings of a specific network. [40] proposed a saliency map extraction method that does not rely on a classifier, which can find the portion of the image that any classifier can use.

Methods Based on Network Activation. Activation maximization is the search for an image that maximizes the activation of a specific neuron (also known as a “unit,” “feature,” or “feature detector”) to reveal the neural response what it has learned (the features it has detected) in DNN. This technique can be performed for output neurons, such as neurons that classify image types [35], or for each hidden neuron in DNN [12, 41, 42], to explain the representation of neuron activation during prediction [34, 43].

Another set of visual activation methods not only focus on single neuron activation, but also take into account the global information of the image.

[31, 39, 44, 45] Integrated Gradient, SmoothGrad, CAM, GradCAM, each method show that the correlation between highly activated region (the area where neurons are highly activated) and highly sensitive region (the area where changes have the greatest influence on the output).

These methods provide useful insights into deep neural networks, but they also have some shortcomings. Based on gradient method, artifacts caused by discontinuity of gradient in the process of back propagation; Based on the network activation method, when the filter response is displayed in the deeper sensory field, the enlarged activation diagram may lose the details obviously. [28, 46] proposed methods to alleviate the problem of introducing artifacts. There is no good solution to the problem of missing details in the enlargement of the activation diagram.

Semantic-Based Feature Representation. [47] proposed the Network Dissection framework, a method for accurately calculating the receptive field regions of neural activation in feature maps. The Network Dissection effectively partitions the input image into multiple parts with various semantic definitions (accurate estimates of receptive fields) that match six semantic concepts (such as scenes, targets, parts, materials, textures, and colors). The semantics directly represent the meaning of the features to improve the interpretability of neurons. [20] proposed the Net2Vec framework, in which semantic concepts are mapped to vector embeddings based on corresponding filter responses. Through this method, the article can better describe the semantics of the filter and its relationship with other semantics. However, the common shortcoming of both is that the interpretation of network components (neurons, filters) is limited by semantic concept annotations, and the annotation of new concepts is costly.

For the deficiencies of the above methods, [48, 49] proposed an unsupervised method, that is, without the annotation concept part. [48] presented a graphical explanatory diagram that reveals the hidden semantic features in pre-trained CNNs. In the explanatory graph, each node represents a part pattern, and each edge encodes co-activation relationships and spatial relationships between patterns. [49] proposed a decision tree for coding potential decision patterns stored in a fully connected layer. The decision tree quantitatively interprets the logic of each CNN prediction, that is, given an input image, the decision tree tells people which object parts activate which filters for the prediction and how much they contribute to the prediction score. The decision tree can be used to explain the basic principles of each CNN prediction at the semantic level, which object parts are used by CNN for prediction.

3.4 Model Interpretation

Explain the model from the perspective of the entire model. The main methods are: simulation model and establishment of an interpretable model system.

Model processing. One is to simulate the model by constructing a simple human-understandable model to simulate the decision function of the depth model, so that the results of the simple model are close to the original model results to achieve the purpose of interpretation. [50] proposed Model Compression method to simulate a shallow network training shallow network, and obtain a single-layer neural network model. This new shallow model can achieve the same effect as the depth model. [51] also uses the method of compressing the model, but it is trimmed according to the filter importance index in the CNN model to achieve the effect of compression. The filter importance index is defined as the classification accuracy reduction (CAR) of the network after pruning that filter.

The other is through model decomposition, which is usually using decision trees that are well interpretable in machine learning as a tool. Both the DeepRED [9] and CRED [10] algorithms in the Rule Processing Section decompose the DNN model into decision tree models to obtain interpretable rules.

Interpretable model system. Building an interpretable model, [52] proposed a method to modify traditional CNN into an interpretable CNN to clarify the knowledge representation in high conv-layers of CNNs. In an interpretable CNN, each filter in a high conv-layer represents a particular object portion. And it automatically assigns each filter in a high conv-layer to the object portion during the learning process. The explicit knowledge representations in CNN can help people understand the logic within CNN.

[53] proposed the learning of qualitatively interpretable models for object detection based on the R-CNN. This method utilize a top-down hierarchical and compositional grammar model embedded in a directed acyclic AND-OR Graph (AOG) to explore and unfold the space of latent part configurations of RoIs. Then proposed an AOG Parsing operator to substitute the RoI Pooling operator widely used in R-CNN. In detection, a bounding box is interpreted by the best parse tree derived from the AOG on-the-fly, which is treated as the extractive rationale generated for interpreting detection.

4 Summary

Being able to understand a “black box” model is the most important issue related to model security, model optimization, and model generalization, especially in medical, financial and other engineering applications. Therefore, model interpretability has been the focus of research in recent years. This paper summarizes the related work based on the interpretability of the CNN, such as the meaning of “interpretability”, and the classification of interpretable methods. Then, we find that the current model interpretability research is divergent, and there is no unified main line. They are basically based on their own previous studies and turned to interpret the results of these studies. Therefore, future studies on interpretability can focus on the following points:

(1) *Conceptual definition of “interpretability”*

At present, there is no unified definition of “interpretability” in the academia. This is not appropriate for the development of follow-up research. It is necessary to formulate a brief explanation for “interpretability”.

(2) *Visual interpretation is the focus of interpretable studies*

Of the 53 references cited in this paper, 21 involved “observing” and understanding models from the perspective of visual interpretation. This is not a denial of other work, but it seems to be a trend, because the graphical interpretation gives the most direct understanding.

(3) *Establishing an interpretable system is the goal*

At present, there are not many achievements in the research on the construction of an interpretable system, but with the emphasis on the concept of “interpretability”, people need a complete interpretative system to meet the needs of interpretation. Such a

system does not merely provide a local interpretation, but an integrated end-to-end interpretation system.

Acknowledgment. This work was funded by Science and Technology Commission of Shanghai Municipality Program (No. 17411952800, No. 18441904500, 18DZ1113400) and Science and Technology Department of Hainan Province (No. ZDYF2018022).

References

1. Gilpin, L.H., et al.: Explaining Explanations: An Overview of Interpretability of Machine Learning (2018)
2. WHI: 2017 Homepage. <https://sites.google.com/view/whi2017/home>
3. Lipton, Z.C.: The Mythos of Model Interpretability. *Communications of the ACM* (2016)
4. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**(6), 1157–1182 (2003)
5. Geras, K.J., et al.: High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks (2017)
6. Ling, J., et al.: Building data-driven models with microstructural images: generalization and interpretability. *Mat. Discov.* **10**, 19–28 (2018). S235292451730042X
7. Chen, J., et al.: Learning to Explain: An Information-Theoretic Perspective on Model Interpretation (2018)
8. Freitas, Alex A.: Comprehensible classification models: a position paper. *ACM Sigkdd Explor. Newsl.* **15**(1), 1–10 (2014)
9. Zilke, J.R.: DeepRED – Rule Extraction from Deep Neural Networks (2016)
10. Sato, M., Tsukimoto, H.: Rule extraction from neural networks via decision tree induction. In: *International Joint Conference on Neural Networks* (2001)
11. Wang, H.: ReNN: Rule-embedded Neural Networks (2018)
12. Erhan, D., et al.: Visualizing higher-layer features of a deep network. *University of Montreal*, 1341.3(1) (2009)
13. Nguyen, A., Jason Y., Jeff C.: Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks (2016). arXiv preprint [arXiv:1602.03616](https://arxiv.org/abs/1602.03616)
14. Zeiler, M.D., Fergus, R.: Visualizing and Understanding Convolutional Networks (2013)
15. Bolei, Z., et al.: Interpreting Deep Visual Representations via Network Dissection. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1 (2018)
16. <http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/>
17. <https://cs.stanford.edu/people/karpathy/convnetjs/demo/classify2d.html>
18. Zhou, B., et al.: Object detectors emerge in deep scene CNNs (2014). arXiv preprint [arXiv:1412.6856](https://arxiv.org/abs/1412.6856)
19. Yosinski, J., et al.: How transferable are features in deep neural networks?. *Eprint Arxiv* 27, 3320–3328 (2014)
20. Fong, R., Vedaldi, A.: Net2Vec: Quantifying and Explaining how Concepts are Encoded by Filters in Deep Neural Networks (2018)
21. Szegedy, C., et al.: Intriguing properties of neural networks. *Computer Science* (2013)
22. Kim, B., et al.: Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV) (2017)
23. Raghu, M., et al.: SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability (2017)

24. Kindermans, P.J., et al.: Investigating the influence of noise and distractors on the interpretation of neural networks (2016)
25. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: 2017 IEEE International Conference on Computer Vision (ICCV) IEEE Computer Society (2017)
26. Hara, S., et al.: Maximally Invariant Data Perturbation as Explanation (2018). arXiv preprint [arXiv:1806.07004](https://arxiv.org/abs/1806.07004)
27. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: Explaining the predictions of any classifier. In: 22nd ACM SIGKDD International Conference ACM (2016)
28. Shrikumar, A., Greenside, P., Kundaje, A.: Learning Important Features Through Propagating Activation Differences (2017)
29. Sebastian, B., et al.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE **10**(7), e0130140 (2015)
30. Lundberg, S., Lee, S.I.: A Unified Approach to Interpreting Model Predictions (2017)
31. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic Attribution for Deep Networks (2017)
32. Hendricks, L.A., et al.: Generating Visual Explanations. In: European Conference on Computer Vision (2016)
33. Hendricks, L.A., et al.: Generating Counterfactual Explanations with Natural Language (2018)
34. Mahendran, A., Vedaldi, A.: Understanding Deep Image Representations by Inverting Them (2014)
35. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. Computer Science (2013)
36. Zeiler, M.D., et al.: Deconvolutional networks. Computer Vision & Pattern Recognition (2010)
37. Mahendran, A., Vedaldi, A.: Salient deconvolutional networks. In: European Conference on Computer Vision (2016)
38. Springenberg, J.T., et al.: Striving for simplicity: the all convolutional net (2014). arXiv preprint [arXiv:1412.6806](https://arxiv.org/abs/1412.6806)
39. Smilkov, D., et al.: Smoothgrad: removing noise by adding noise (2017). arXiv preprint [arXiv:1706.03825](https://arxiv.org/abs/1706.03825)
40. Zolna, K., Krzysztow, J.G., Kyunghyun, C.: Classifier-agnostic saliency map extraction (2018). arXiv preprint [arXiv:1805.08249](https://arxiv.org/abs/1805.08249)
41. Le, Q.V., et al.: Building high-level features using large scale unsupervised learning (2011). arXiv preprint [arXiv:1112.6209](https://arxiv.org/abs/1112.6209)
42. Yosinski, J., et al.: Understanding neural networks through deep visualization (2015). arXiv preprint [arXiv:1506.06579](https://arxiv.org/abs/1506.06579)
43. Dosovitskiy, A., Thomas B.: Inverting visual representations with convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
44. Zhou, B., et al.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
45. Selvaraju, R.R., et al.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization (2016)
46. Oramas, J., Kaili W., Tinne, T.: Visual explanation by interpretation: improving visual feedback capabilities of deep neural networks (2017). arXiv preprint [arXiv:1712.06302](https://arxiv.org/abs/1712.06302)
47. Bau, D., et al.: Network dissection: quantifying interpretability of deep visual representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
48. Zhang, Q., et al.: Interpreting CNN Knowledge via an Explanatory Graph (2017)

49. Zhang, Q., et al.: Interpreting CNNs via Decision Trees (2018)
50. Ba, L.J., Caruana, R.: Do deep nets really need to be deep?. In: International Conference on Neural Information Processing Systems, MIT Press (2014)
51. Abbasi-Asl, R., Yu, B.: Interpreting Convolutional Neural Networks Through Compression (2017)
52. Zhang, Q., Wu, Y.N., Zhu, S.-C.: Interpretable Convolutional Neural Networks (2018)
53. Wu, T., et al.: Towards Interpretable R-CNN by Unfolding Latent Structures (2017). arXiv preprint [arXiv:1711.05226](https://arxiv.org/abs/1711.05226)