DOI: 10.23977/acss.2026.100105
ISSN 2371-8838 Vol. 10 Num. 1

A Survey of Deep Learning Interpretability Methods: Current Status and Challenges

Suyang Wu

*School of Electronic and Information Engineering, University of Science and Technology Liaoning,
Anshan, China*
1615079253@qq.com

Keywords: Deep Learning; Interpretability; Black Box Problem; Causal Inference; High-Risk Domains

Abstract: Deep learning models have demonstrated excellent performance in numerous fields such as image recognition, natural language processing, and medical diagnosis. However, due to their complex network structures and nonlinear mapping mechanisms, they exhibit significant "black box" problems, which restrict their reliable application in high-risk domains. This paper systematically combs through the core value and development history of deep learning interpretability research, classifies existing interpretability methods into three major categories: feature visualization-based methods, model decomposition-based methods, and causal inference-based methods, deeply analyzes the core principles, applicable scenarios, advantages and disadvantages of each type of method, focuses on discussing the application requirements of interpretability in high-risk fields such as medical care and finance, and finally looks forward to potential breakthrough directions such as the integration of causal and statistical interpretation frameworks in the future. The research aims to provide a comprehensive overview of the current status and directional guidance for deep learning interpretability research, and help promote the credible development of deep learning models.

1. Introduction

1.1. Research Background

With the improvement of computing power and the growth of data volume, deep learning models have achieved breakthrough progress in many fields. From image recognition and speech recognition to natural language generation, their application scenarios have continuously expanded, and even extended to high-risk domains with extremely high requirements for reliability and safety, such as medical diagnosis, financial risk control, and autonomous driving. However, deep learning models, especially deep neural networks, are usually regarded as "black boxes"—their decision-making process relies on complex nonlinear interactions of massive parameters, which are difficult for humans to intuitively understand and explain. When models have decision-making biases in high-risk domains, the root cause of the bias cannot be traced, which may lead to serious safety accidents and social risks. For example, in medical diagnosis, if a deep learning-based lesion detection model gives an incorrect diagnosis result, it will directly threaten the patient's life and health; in financial risk control, unreasonable credit approval decisions of the model may cause economic losses to financial institutions and users. Therefore, solving the "black box" problem of deep learning and carrying out interpretability research have become the key premise for promoting the safe and reliable application of deep learning models.

1.2. Core Value

Deep learning interpretability research has multiple core values, covering model optimization, risk control, trust building and other dimensions. Firstly, at the model optimization level, interpretability can help researchers and developers understand the key basis for model decisions, identify biases and defects in the model's feature learning process, provide targeted guidance for model structure improvement and parameter tuning, and thereby enhance the generalization ability and robustness of the model. Secondly, at the risk control level, by explaining the model's decision-making logic, potential decision-making risks of the model in special scenarios can be identified in advance, reducing the safety hazards of applying the model in high-risk domains and meeting the regulatory requirements of relevant industries. Thirdly, at the trust building level, clear decision explanations can enhance users' trust in deep learning models and promote their application in practical scenarios, which is also the core embodiment of interpretability as a trust bridge between users and models. In addition, interpretability research can also promote the development of artificial intelligence ethics and fairness. By analyzing the correlation between model decisions and input features, it can avoid models making discriminatory decisions due to biases in the learning data.

1.3. Development History

The development history of deep learning interpretability research is closely related to the evolution of deep learning models, which can be roughly divided into three stages. The first stage is the early exploration stage (before 2010). At this stage, deep learning models have not yet been widely popularized, and interpretability research mainly focuses on simple neural networks. The research methods are relatively basic, mostly focusing on post-hoc statistical analysis of model output results. There are few relevant research results, and no systematic research system has been formed. The second stage is the rapid development stage (2010-2020). With the successful application of deep convolutional neural networks, recurrent neural networks and other models in computer vision and natural language processing fields, the "black box" problem has become increasingly prominent, and interpretability research has received widespread attention. A large number of interpretability methods based on feature visualization and gradient analysis have emerged in this stage, such as Gradient-Weighted Class Activation Mapping (Grad-CAM) and feature visualization of deep convolutional networks. The research focus is on revealing the key features relied on by model decisions, forming a preliminary research framework. The third stage is the in-depth expansion stage (since 2020). Interpretability research has extended from a single post-hoc explanation to pre-interpretability model design and in-process decision-making tracking. At the same time, it integrates cross-disciplinary theories such as causal inference and symbolic logic, making the research direction more diversified, focusing on the reliability, generalization and causality of explanations to meet the strict requirements of high-risk domains.

2. Classification and Analysis of Deep Learning Interpretability Methods

According to the core ideas and technical paths of interpretation, existing deep learning interpretability methods can be divided into three major categories: feature visualization-based methods, model decomposition-based methods, and causal inference-based methods. Each type of method approaches the "black box" problem from different angles, with unique principle mechanisms and applicable scenarios, as well as their own limitations. This section will conduct a detailed combing and comparative analysis of various methods.

2.1. Feature Visualization-Based Methods

The core idea of feature visualization-based methods is to convert the abstract features learned by each layer of the deep learning model into visually understandable information for humans. By analyzing the visualization results, the key input features relied on by model decisions are clarified. Deep learning models, especially deep convolutional neural networks, usually learn basic features such as edges and textures in the lower layers, and more complex semantic features in the higher layers. Such methods convert the feature maps of each layer of the network into images through techniques such as backpropagation and activation maximization, enabling researchers to intuitively observe the process of the model extracting and processing input information, and then infer the basis of the model's decisions. For example, the activation maximization method optimizes the input image to maximize the activation of a specific neuron in a certain layer of the network, and the optimized image is the feature pattern learned by that neuron. The recently proposed Feature CAM method further improves the interpretation effect of feature visualization methods by combining perturbation and activation techniques to generate saliency maps that are more interpretable by humans than the traditional Grad-CAM++[3]. In terms of applicable scenarios, such methods are mainly suitable for deep learning models in the field of computer vision, such as image classification, object detection, and semantic segmentation. In these scenarios, the input data is images, and the feature visualization results can be directly combined with human visual cognition, facilitating researchers to understand the model's decision-making logic; in addition, they can also be used for simple natural language processing tasks, such as text classification, by visualizing the distribution of word embedding features to analyze the model's learning of different semantic features. Such methods have the advantages of being intuitive and easy to understand, and simple to implement. The visualization results can be directly associated with human visual cognition, reducing the understanding threshold of interpretation. Moreover, most of them are model-agnostic post-hoc interpretation methods, which do not need to modify the model structure and have strong versatility. However, they also have significant limitations: first, the interpretation is highly subjective, and different researchers may have different interpretations of the visualized features, making it difficult to form a unified interpretation standard; second, the interpretation ability for deep networks is limited. With the increase of network layers, the semantic clarity of the feature visualization results gradually decreases, making it difficult to accurately correspond to the specific content in the input image; third, it cannot reveal the interaction between features, and it is difficult to fully reflect the model's decision-making mechanism.

2.2. Model Decomposition-Based Methods

Model decomposition-based methods realize the interpretation of model decisions by decomposing and analyzing the network structure, parameter distribution or decision-making process of deep learning models, and converting complex "black box" models into a series of simple and interpretable sub-models or rules. Their core logic is that the decision-making logic of deep learning models is contained in their network structure and parameters, and the function and role of each part in the decision-making process can be clarified by decomposing the components of the model. According to different decomposition methods, it can be further divided into three categories: model simplification, rule extraction and gradient analysis. Model simplification reduces model complexity by pruning the network structure, quantifying parameters and other methods; rule extraction converts the decision-making logic into human-understandable logical rules such as "if-then" by analyzing the mapping relationship between model parameters and input-output; gradient analysis measures the influence of each input feature on the decision result by calculating the gradient of the model output to the input feature. In terms of application scope, such methods can be applied to multiple fields such as image recognition, natural language processing, and recommendation systems. Among them, model simplification and rule extraction are more suitable for models with relatively regular structures and moderate parameter scales, and have significant advantages in scenarios that require clear decision rules, such as credit approval models in financial risk control; gradient analysis methods are suitable for various deep learning models, especially widely used in scenarios that need to quickly locate key input features, such as the localization of lesion areas in medical image diagnosis. The advantages of such methods are that the interpretation is highly objective, the results are directly derived from the model's structure and parameters, which can avoid the subjectivity of human interpretation, and the clear decision rules generated by rule extraction methods are also convenient for regulatory auditing. However, they also have many shortcomings: first, there is a trade-off between "performance and interpretability". Simplifying the model or extracting rules may lead to a decrease in performance; second, gradient-based methods are sensitive to minor perturbations of input data, which affects the stability and reliability of interpretation; third, for complex deep networks, decomposition is difficult, which may lead to one-sided interpretation.

2.3. Causal Inference-Based Methods

Causal inference-based methods break through the statistical correlation-based interpretation framework of traditional interpretability methods, and realize the causal interpretation of model decisions by establishing the causal relationship between input features and model outputs. Such methods hold that the decision-making of deep learning models not only relies on the statistical correlation between features and outputs, but also should be based on the causal relationship between them. Only by clarifying the causal impact of input features on output results can reliable interpretation be achieved. Its core is to identify causal features in input features and exclude confounding features by introducing causal inference tools such as causal graphs and intervention experiments, and then reveal the causal logic of model decisions. For example, by intervening in the value of a certain input feature and observing the change of the model's output result, it is judged whether there is a causal relationship between the feature and the output result[2]. Researchers have constructed a Structural Causal Model (SCM) as an abstraction of specific aspects of convolutional neural networks, and proposed a method for quantitatively ranking convolutional layer filters based on counterfactual importance. This method has been verified on mainstream convolutional neural network architectures such as LeNet5, VGG19 and ResNet32. Another study uses causal generative learning as a tool to explain image classifiers. By changing causal attribute values through counterfactual reasoning, calculating Shapley values and contrastive explanations, it identifies the pixels that have the most significant impact on the classifier's decisions. The generated counterfactual explanations are more interpretable than existing tools[4]. In terms of applicable scenarios, such methods are suitable for scenarios that require high reliability and causality of interpretation, and have important application value especially in high-risk fields such as medical care, finance, and justice. They can clarify the causal relationship between lesion features and disease diagnosis, identify causal factors affecting credit risk, and also be used to solve the fairness problem of models. The biggest advantage of such methods is that the interpretation has high reliability and generalization, which can effectively exclude the influence of statistical confounding factors and provide reliable guidance for model robustness optimization. However, they also have obvious limitations: first, the computational complexity is high, requiring high computing power and data volume, making it difficult to apply to deep learning models with large parameter scales; second, it is difficult to identify causal relationships. In high-dimensional input scenarios, the causal relationships between features are complex, which may lead to inaccurate inference results; third, the method has poor versatility. The causal relationships in different fields are significantly different, making it difficult to establish a unified causal interpretation framework.

3. Application Requirements of Deep Learning Interpretability in High-Risk Domains

The particularity of high-risk domains such as medical care and finance determines that their requirements for deep learning interpretability are much higher than those of ordinary domains. They not only require clear and accurate interpretation results, but also need to meet multiple requirements such as regulatory compliance, auditability, and traceability. This section will combine the specific application scenarios of medical and financial fields to analyze their core requirements for deep learning interpretability.

3.1. Application Requirements in the Medical Field

The medical field is directly related to human life and health, and the interpretability of deep learning models is the core premise for their clinical application. The specific requirements include the following three aspects. First, the accuracy and reliability of interpretation. The interpretation results of the model must accurately reflect the real basis for its decisions, and the basis must comply with medical common sense and clinical practice. For example, in the lesion detection model, the interpretation results must accurately point out the specific lesion areas and features that lead the model to make a positive diagnosis, and the areas and features must be supported by medical theories to avoid misdiagnosis by doctors due to incorrect interpretation. The recently proposed causally informed interpretable early prediction model identifies potential causal relationships for prediction through causal discovery, which can not only provide clear causal path references for clinical diagnosis, but also show better generalization among different patient groups[5]. Second, the balance between interpretability and professionalism. The interpretation results must be understandable to non-professional patients to enhance their trust in the diagnosis results, and also meet the clinical analysis needs of professional doctors. This requires interpretability methods to generate multi-level interpretation results, including both easy-to-understand natural language explanations and professional medical feature analysis reports. Third, the traceability and compliance of interpretation. The medical field is subject to strict supervision. The decision-making process and interpretation results of deep learning models must have complete traceability, and be able to record key information such as model input data, parameter changes, and decision-making basis to meet the requirements of medical accident liability identification and regulatory auditing. In addition, the interpretation results must also comply with relevant regulations on the protection of medical data privacy to avoid leakage of patients' sensitive information.

3.2. Application Requirements in the Financial Field

The financial field involves a large number of capital transactions and risk control. The requirements for deep learning interpretability focus on compliance, fairness and risk controllability, including the following three aspects. First, the compliance and auditability of interpretation.

Financial institutions are subject to strict supervision by regulatory authorities such as the China Banking and Insurance Regulatory Commission and the China Securities Regulatory Commission. The decision-making logic of deep learning models (such as credit approval models and risk assessment models) must comply with relevant financial regulations and regulatory requirements. The interpretation results must be clear and clear, and understandable and auditable by regulatory authorities. For example, in the credit approval model, the specific reasons for rejecting a credit application must be clearly explained, and the reasons must comply with relevant regulations such as anti-discrimination and fair credit. Second, the fairness and non-bias of interpretation. The decision results of financial models must not have discrimination based on sensitive features such as gender, race, and region. Interpretability methods must be able to identify potential biased features in model decisions and ensure that the interpretation results can reflect the fairness of model decisions. For example, researchers or financial institutions can verify whether the credit approval model makes decisions solely based on reasonable features such as the applicant's credit status and repayment ability, and is not influenced by sensitive features (e.g., gender, race, region) by explaining the model's decision-making basis. Third, the risk early warning of interpretation. The financial market has a high degree of uncertainty. The interpretability of deep learning models must help financial institutions identify potential decision-making risks of the model in advance and adjust the model strategy in a timely manner. For example, financial institutions can identify the decision-making flaws of the risk assessment model in extreme market environments by explaining the model's decision-making logic, thereby avoiding financial risks arising from model failures.

4. Future Outlook of Deep Learning Interpretability Research

Although significant progress has been made in deep learning interpretability research, there are still many challenges, such as the reliability, generalization, and causality of interpretation have not been completely solved. In the future, interpretability research will develop towards integrating multi-disciplinary theories, constructing unified interpretation frameworks, and improving the practical value of interpretation. The specific breakthrough directions include the following aspects.

4.1. Integration of Causal and Statistical Interpretation Frameworks

Existing statistical correlation-based interpretability methods lack consideration of causal relationships, resulting in insufficient reliability of interpretation results; while causal inference-based methods have problems of high computational complexity and poor versatility. In the future, the integrated causal and statistical interpretation framework will become an important breakthrough direction in interpretability research. This framework will combine the efficiency of statistical learning with the reliability of causal inference, quickly identify the correlation patterns between input features and output results through statistical methods, then verify the causal relationships in the correlation patterns through causal inference, exclude the influence of confounding factors, and finally generate interpretation results with both efficiency and reliability. As Das and Rad emphasized in their research, constructing an abstract framework of deep learning models with the help of causal inference ideas can support arbitrary causal interventions and queries, thereby deeply understanding the internal working mechanism of the model, which provides an important idea for the integrated causal and statistical interpretation framework. Koch et al.'s research also pointed out that the near-nonparametric nature of deep learning enables it to estimate smooth response surfaces and capture heterogeneous treatment effects of individual units, providing technical support for integrated causal and statistical interpretation methods[6]. For example, deep learning models can be used to learn the statistical correlation between input features and outputs, and then causal graph models can be used to verify the causal relationships of correlated features, clarify the causal hierarchical relationships between features, and realize in-depth causal interpretation of model decisions. This integrated framework can effectively balance the efficiency and reliability of interpretation, and improve the application value of interpretability methods in complex practical scenarios.

4.2. Design and Optimization of Pre-Interpretable Models

The current mainstream interpretability methods are mostly post-hoc explanations, that is, interpreting the model's decision-making process after the model training is completed. There are problems such as disconnection between interpretation and the model's decision-making process, and possible inaccuracy of interpretation results. In the future, the design and optimization of pre-interpretable models will become a research hotspot, which is consistent with the research idea of integrating interpretability constraints into the modeling process. By introducing interpretability constraints in the model design stage, the model can not only learn task-related features during the training process, but also actively learn interpretable decision-making logic, solving the "black box" problem from the root. For example, a deep learning model with a modular structure can be designed, where each module corresponds to a specific function and decision-making logic, and the interpretability of the model's decision-making process can be realized by clarifying the interaction between modules; interpretability regular terms can also be introduced into the loss function to guide the model to learn human-understandable feature patterns. Pre-interpretable models can avoid the limitations of post-hoc explanations, improve the accuracy and reliability of interpretation, and are more suitable for application requirements in high-risk domains.

4.3. Research on Cross-Modal Interpretability Methods

With the development of deep learning, the application of cross-modal models (such as text-image fusion models and speech-text conversion models) has become increasingly widespread. However, existing interpretability methods are mostly designed for single-modal models, making it difficult to adapt to the complex decision-making process of cross-modal models. In the future, research on cross-modal interpretability methods will receive more attention. Such methods need to integrate feature information from different modal data, reveal the interaction mechanism and decision-making logic between cross-modal features, and generate unified and coherent interpretation results. For example, in the text-image fusion sentiment analysis model, the interpretability method must be able to clarify how text features and image features work together to affect the sentiment judgment results, and generate an interpretation report covering both text and image features. The research on cross-modal interpretability methods will expand the application scope of interpretability research and promote the credible development of cross-modal deep learning models.

4.4. Construction of Interpretability Evaluation System

At present, deep learning interpretability research lacks a unified evaluation system. The evaluation standards, indicators and data sets of different interpretability methods are quite different, making it difficult to objectively and fairly compare the interpretation effects of different methods. In the future, constructing a unified and comprehensive interpretability evaluation system will become the key to promoting the standardized development of interpretability research. This evaluation system should include quantitative evaluation indicators and qualitative evaluation indicators. Quantitative indicators can include the accuracy, stability, and generalization of interpretation, while qualitative indicators can include the interpretability and professionalism of interpretation; at the same time, a standardized evaluation data set should be established, covering the application scenarios of deep learning models in different fields and tasks, to provide a unified benchmark for the evaluation of interpretability methods. A unified evaluation system can standardize the development direction of interpretability research, promote communication and cooperation between different research teams, and accelerate the innovation and application of interpretability technologies.

5. Conclusion

The "black box" problem of deep learning has become a key bottleneck restricting its reliable application in high-risk domains, and interpretability research has important theoretical and practical value. This paper systematically combs through the core value and development history of deep learning interpretability research, classifies existing interpretability methods into three major categories: feature visualization-based methods, model decomposition-based methods, and causal inference-based methods, deeply analyzes the principles, applicable scenarios, advantages and disadvantages of various methods, discusses the application requirements of interpretability in high-risk fields such as medical care and finance, and looks forward to future breakthrough directions such as the integration of causal and statistical interpretation frameworks and the design of pre-interpretable models.

Each existing interpretability method has its own advantages and disadvantages. Feature visualization-based methods are intuitive and easy to understand but highly subjective; model decomposition-based methods are highly objective but have a trade-off between "performance and interpretability"; causal inference-based methods have high reliability but high computational complexity. The requirements for interpretability in high-risk domains focus on accuracy, reliability, compliance and traceability, which cannot be fully met by existing methods. In the future, by integrating multi-disciplinary theories, constructing unified interpretation frameworks, optimizing the design of pre-interpretable models, and establishing standardized evaluation systems, the research level of deep learning interpretability will be effectively improved, promoting deep learning models to develop in a more credible and reliable direction, and laying the foundation for their wide application in high-risk domains. As pointed out in relevant studies, the theoretical basis research and algorithm innovation of interpretability are necessary ways to open the black box of deep models, and also important supports for promoting the improvement of fairness and generalization of deep learning[1].

References

[1] Hong, Xiangyu, et al. "DePass: Unified Feature Attributing by Simple Decomposed Forward Pass." arXiv preprint arXiv:2510.18462 (2025).

[2] Narendra, Tanmayee, et al. "Explaining deep learning models using causal inference." arXiv preprint arXiv:1811.04376 (2018).

[3] Clement, Fricy, Ji Yang, and Irene Cheng. "Feature CAM: interpretable ai in image classification." arXiv preprint arXiv:2403.05658 (2024).

[4] Taylor-Melanson, Will, Zahra Sadeghi, and Stan Matwin. "Causal generative explainers using counterfactual inference: a case study on the Morpho-MNIST dataset." Pattern Analysis and Applications 27.3 (2024): 89.

[5] Cheng, Yuxiao, et al. "Causally-informed Deep Learning towards Explainable and Generalizable Outcomes Prediction in Critical Care." arXiv preprint arXiv:2502.02109 (2025).

[6] Koch, Bernard J., et al. "A Primer on Deep Learning for Causal Inference." Sociological Methods & Research 54.2 (2025): 397-447.