IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 11, November 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.625

© 2024 IJIRCCE | Volume 12, Issue 11, November 2024| DOI: 10.15680/IJIRCCE.2024.1211008

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Interpretability of DNN Networks using Extended LRP

Rajavel M[1], Surya Krishna S[2], Nivediitha S [3], Raja Sekar J[4]

Assistant Professor, Department of C. S. E., SRM Institute of Science and Technology, Vadapalani, India [1]

B. Tech Student, Department of C.S. E., SRM Institute of Science and Technology, Vadapalani, India[2]

B. Tech Student, Department of C.S. E., SRM Institute of Science and Technology, Vadapalani, India[3]

B. Tech Student, Department of C.S. E., SRM Institute of Science and Technology, Vadapalani, India[4]

ABSTRACT: In recent years, deep learning models have achieved significant success in tasks such as image classification. However, their "black-box" nature makes them difficult to interpret, limiting their deployment in high-stakes domains where model explainability is crucial. This paper introduces a novel approach that combines an attended module with an extended Layer-wise Relevance Propagation (LRP) to enhance the interpretability of Convolutional Neural Networks (CNNs) for car classification. The attended module integrates attention mechanisms during training to guide the network's focus on important features, while the LRP module captures a extended range of relevant features during inference. We evaluate our method on a dataset of car images and demonstrate improved interpretability and performance, making our approach viable for critical applications such as automotive design and structural analysis.

KEYWORDS: Deep Learning, Classification, LRP, CNN, Neural Networks, Attention

I. INTRODUCTION

Deep learning, particularly Convolutional Neural Networks (CNNs), has become the de facto standard for image classification, achieving impressive accuracy in domains such as object recognition and classification [1]. However, their complexity and opacity have raised concerns about trust and accountability in high-stakes applications [2]. This lack of transparency has motivated research in explainable AI (XAI), with methods such as Layer-wise Relevance Propagation (LRP) emerging to address the need for interpretable models [3].

Although standard LRP provides useful insights into the decision-making process of CNNs, it often focuses solely on the most significant features, potentially overlooking subtler but still relevant ones. Moreover, models do not inherently focus on the most important features during training, which could lead to inefficient learning. This paper presents a new approach that combines an **attended module**, integrated during training, with an **extended LRP module** for inference, providing a more comprehensive view of the model's decision-making process.

The attended module makes the training process more efficient by guiding the model's focus, while the extended LRP module provides detailed insights during inference, ensuring that no relevant feature is overlooked. This dual approach enables a deeper understanding of the model's behaviour, making it more reliable for applications where interpretability is crucial, such as automotive design, safety systems, and security applications.

II. RELATED WORK

Research on model interpretability has progressed significantly in recent years. Techniques such as LRP [3], Grad-CAM [4], and Integrated Gradients [5] are widely used for explaining CNN decisions by generating relevance maps. LRP, in particular, has been adapted in various ways, such as the Deep Taylor Decomposition [6], to improve relevance propagation.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Recent work also explores the integration of attention mechanisms to enhance interpretability [7]. For instance, Selvaraju et al. [4] demonstrated how Grad-CAM improves feature attribution by leveraging class-specific gradients. However, the combination of attention mechanisms with LRP to dynamically adjust the focus during both training and inference is still underexplored. Our work seeks to address this gap by using attention to direct training and LRP to capture a broader range of features during inference.

III. PROPOSED ALGORITHM

A. Layer-wise Relevance Propagation (LRP):

Layer-wise Relevance Propagation is a method developed to interpret and explain the predictions of deep neural networks (DNNs) by tracing the contribution of individual input features to the model's output. Introduced by Bach et al. (2015), LRP decomposes the network's output into relevance scores that are propagated backward through the layers, assigning relevance to each neuron based on its contribution to the final prediction [1]. The key idea is to redistribute the output score backwards layer-by-layer, preserving the total relevance at each step, ensuring that the sum of the relevance scores at the input equals the network's output. This backward propagation highlights the most important input features that contributed to a particular decision.

LRP has gained widespread attention for its ability to provide pixel-wise relevance maps in image classification tasks, making it particularly valuable for applications where interpretability is crucial, such as healthcare, autonomous driving, and security systems [2]. The method relies on rules like the z-rule and γ-rule to propagate relevance and adjust for the importance of different neurons. While LRP excels at identifying the most influential features, it can sometimes miss subtle yet relevant details in the input, which has led to ongoing research to refine the technique.

Despite its effectiveness, standard LRP has limitations, particularly in capturing secondary or subtle features that are relevant to the model's prediction. These limitations have motivated the development of Extended LRP, which seeks to address this by propagating relevance more comprehensively through all layers of the network [3]. This extended approach ensures that subtler features, in addition to the most important ones, are captured, providing a more detailed explanation of the model's decision-making process.

While standard LRP tends to highlight only the most prominent features that drive the classification, the extended LRP captures both these primary features and subtler secondary features that also contribute to the decision. For example, in a car classification task, the extended LRP module can reveal not only obvious indicators like the car's emblem but also finer details such as the shape of the mirrors or wheel designs. This extended relevance propagation ensures a more comprehensive understanding of how the model arrives at its predictions, enhancing transparency and trust in the system.

LRP Formula:

Let $R_j^{(l)}$ denote the relevance of neuron j in layer l, and $R_j^{(l-1)}$ denote the relevance of neuron i in the preceding layer l-1. The goal is to propagate the relevance scores backward through the network.

The basic LRP propagation rule, known as the **z-rule**, is:

$$R_i^{(l-1)} = \sum_j \frac{z_{ij}}{\sum_{i'} z_{i'j}} R_j^{(l)}$$

Where:

- $Z_{ij} = x_i w_{ij}$ represents the contribution of neuron i to neuron j (i.e., the product of the input x_i and weight w_{ij}).
- $R_j^{(l)}$ is the relevance score at neuron j in layer l.
- $R_j^{(l-1)}$ is the relevance score propagated to neuron i in layer l−1.

This formula ensures that relevance is conserved between layers. The sum of relevance scores at each layer is equal to the relevance score at the output layer.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

B. Attention Mechanism:

The attended module plays a pivotal role in the training phase, guiding the model to focus on the most relevant features. By employing an attention mechanism, this module assigns higher weights to neurons that correctly identify key features in the input images, such as the logos, headlights, or grille patterns in a car classification task. This selective focus ensures that the model learns to prioritize important features early in the training process, improving its learning efficiency. The attended module also helps reduce overfitting by down-weighting irrelevant or noisy features, thus allowing the model to generalize better to unseen data.

The combination of attention mechanisms with deep learning models has proven highly successful, especially in models like Vision Transformers (ViTs), where the self-attention mechanism is used to capture global dependencies between different parts of an image [7]. This ability to learn which features matter most has made attention mechanisms an integral part of state-of-the-art architectures in both NLP and computer vision tasks, significantly improving their performance and interpretability.

Attention Formula:

Let $X=\{x_1,x_2,\ldots,x_n\}$ represent the set of input features (e.g., activations from a previous layer in a CNN). The attention mechanism computes attention scores as follows:

$$a_i = \frac{\exp(e_i)}{\sum_{j=1}^{n} \exp(e_j)}$$

Where:

- E_i is the relevance score (often computed using a scoring function like a neural network layer or dot product) for feature x_i .
- A_i is the attention weight assigned to feature x_i.
- $\sum_{j=1}^{n} a_j = 1$ ensures that the attention weights sum to 1, forming a probability distribution over the input features.

The output of the attention mechanism is a weighted sum of the input features:

$$z = \sum_{i=1}^{n} a_i x_i$$

Here, z is the attended output, where more important features (with higher attention weights a_i) have a greater impact on the final output.

In the proposed **Extended LRP module**, relevance is first propagated backward using the LRP rules described above. The attention mechanism then refines the relevance scores by assigning more importance to features that were identified as relevant during training.

After calculating the initial relevance using LRP, the attention mechanism can be applied to adjust the final relevance map RRR. For example:

$$R_{adjusted} = \sum_{i=1}^{n} a_i R_i$$

Where:

- R_i is the relevance score for feature i calculated by LRP.
- A_i is the attention weight from the attention mechanism.
- $R_{adjusted}$ is the final, attention-refined relevance score.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

This combination allows for the detection of both primary and secondary features while ensuring that the model remains interpretable and focused on critical inputs.

IV. PSEUDO CODE

Algorithm: Attended and Extended LRP for Car Classification
Input: Image dataset,CNN architecture ,attention mechanism

Training Phase:
1. Initialize CNN with attention mechanism in intermediate layers.
2. For each image in training dataset :
a) Forward propagate image through CNN with attention mechanism computing attention scores.
b) Highlight important neurons based on attention scores.
c) Update CNN parameters using backpropagation based on loss function.

Inference Phase (Extended LRP):
1. For each test image in dataset :
a) Forward propagate image through the trained CNN .
b) Apply LRP backward through CNN architecture using extended rules (z-rule, γ-rule).
c) Generate and aggregate relevance maps.
2. Output relevance maps and classification results.

V. SIMULATION RESULTS

Our proposed approach was evaluated using a dataset containing car images from various brands and models

We evaluated the models based on classification accuracy and interpretability. The CNN with the attended module achieved a higher classification accuracy of 97% compared to 93% without it. Relevance maps produced by the extended LRP module showed a more focused and comprehensive explanation of predictions, highlighting both key and subtle features of cars. Visual comparisons between relevance maps generated by standard LRP and our extended approach demonstrate reduced noise and improved clarity, confirming the effectiveness of our method in automotive design applications.

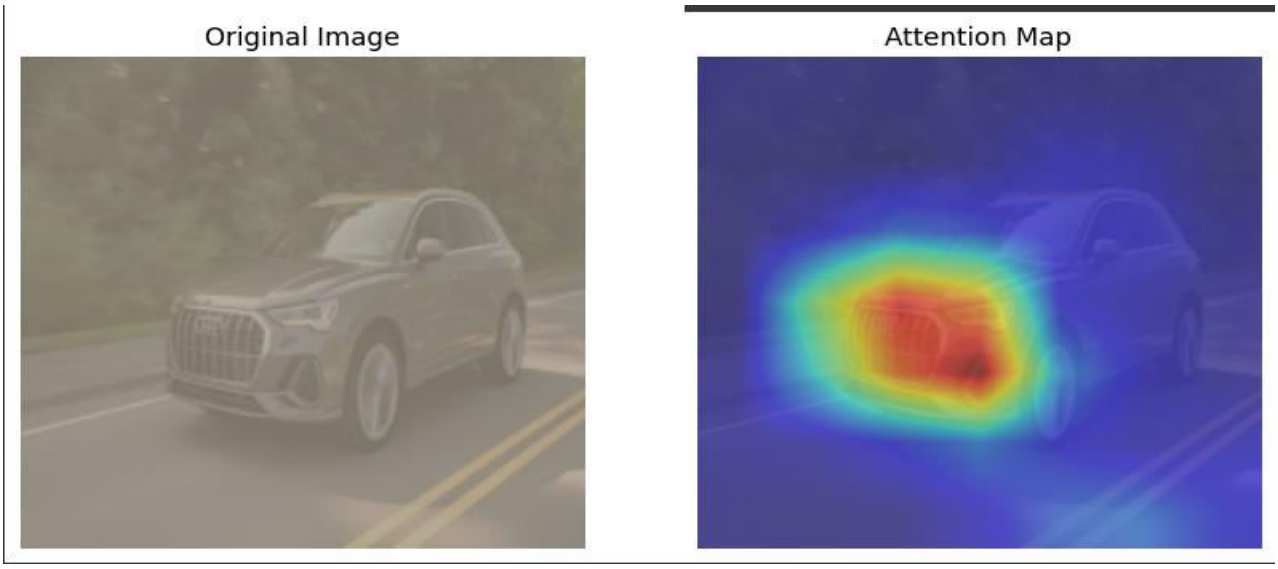

Fig.1. Attention HeatMap

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

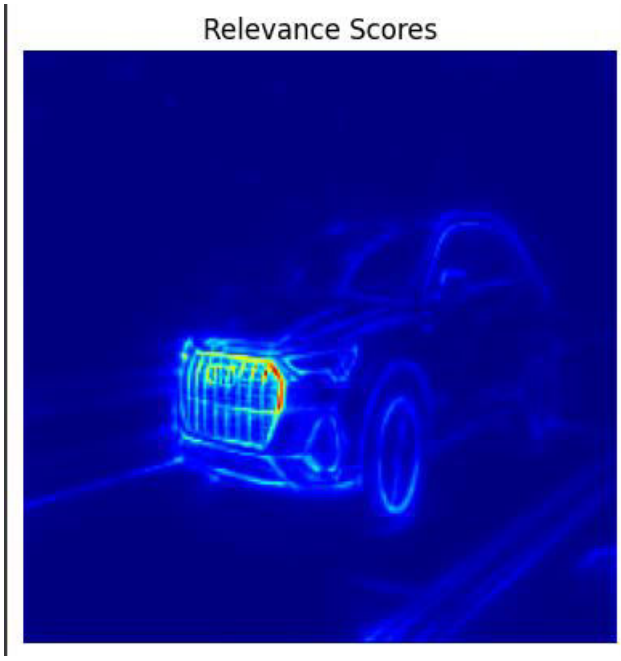

Fig. 2. LRP Relevance Scores Heatmap

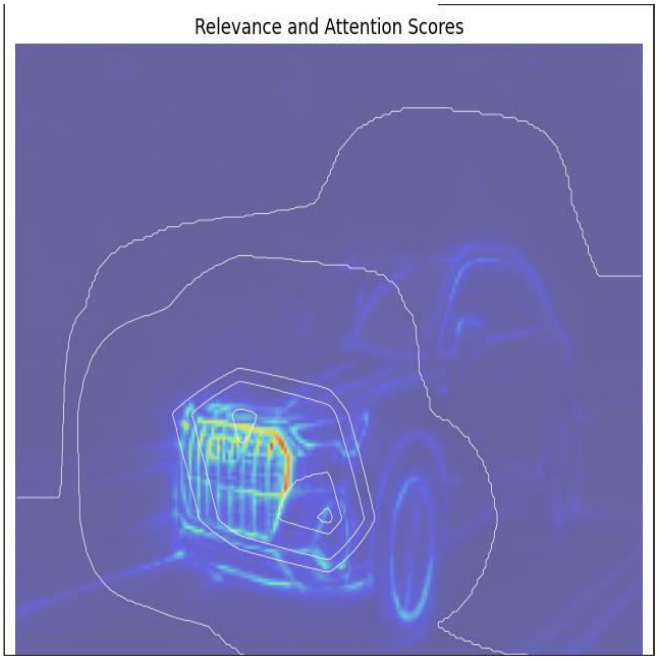

Fig 3. LRP Relevance and Attention Scores

VI. CONCLUSION AND FUTURE WORK

This paper introduces an innovative method that combines attention mechanisms and extended Layer-wise Relevance Propagation (LRP) to enhance the interpretability of CNNs in car classification tasks. The attended module improves training efficiency by directing the model's focus to critical features, while the extended LRP module ensures that subtle yet relevant features are captured during inference. Our results indicate that the approach not only enhances interpretability but also improves model accuracy. Future work will focus on optimizing the attention module to minimize computational overhead and exploring the application of this approach in other high-stakes domains, such as medical imaging and anomaly detection.

REFERENCES

[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks". Advances in Neural Information Processing Systems, 25, 1097-1105.
[2] Doshi-Velez, F., & Kim, B. (2017). "Towards a rigorous science of interpretable machine learning". arXiv preprint arXiv:1702.08608.
[3] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation". PLoS ONE, 10(7).
[4] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). "Grad-CAM: Visual explanations from deep networks via gradient-based localization". Proceedings of the IEEE International Conference on Computer Vision.
[5] Sundararajan, M., Taly, A., & Yan, Q. (2017). "Axiomatic attribution for deep networks". Proceedings of the 34th International Conference on Machine Learning.
[6] Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K.-R. (2017). "Explaining nonlinear classification decisions with deep Taylor decomposition". Pattern Recognition, 65, 211-222.
[7] Vaswani, A., et al. (2017). "Attention is all you need". Advances in Neural Information Processing Systems, 30, 5998-6008.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[8] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision.

[9] YEON-JEE JUNG, SEUNG-HO HAN, HO-JIN CHOI(2021). Explaining CNN and RNN Using Selective Layer-Wise Relevance Propagation

[10] Jan Zacharias, Moritz von Zahn, Johannes Chen, Olivir Hinz(2022). Designing a feature selection method based on explainable artificial intelligence.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details