QR Code

DO NOT REMOVE

SMARTSAT COOPERATIVE RESEARCH CENTRE

University of South Australia

Green Triangle Forest Industries Hub

Improve interpretability of Information Bottlenecks for Attribution with Layer-wise Relevance Propagation

Xiongren Chen, Jiuyong Li, Jixue Liu, Stefan Peters, Lin Liu, Thuc Duy Le, Anthony Walsh

Introduction

Deep learning models, especially in computer vision, have achieved remarkable accuracy. However, their decision-making remains enigmatic. Researchers have developed visualization techniques, like attribution maps, to discern which input features most influence a model's decision. These maps assign importance to each input feature, based on its impact on the model's outcome. Yet, they lacked human-perceptual clarity. Our proposed Relevance-IBA combines IBA [1] with the LRP's [2] backpropagation method, enhancing attribution map clarity and aligning better with human intuition. We advocate for a segmentation-oriented evaluation, emphasizing interpretability that resonates with human perception.

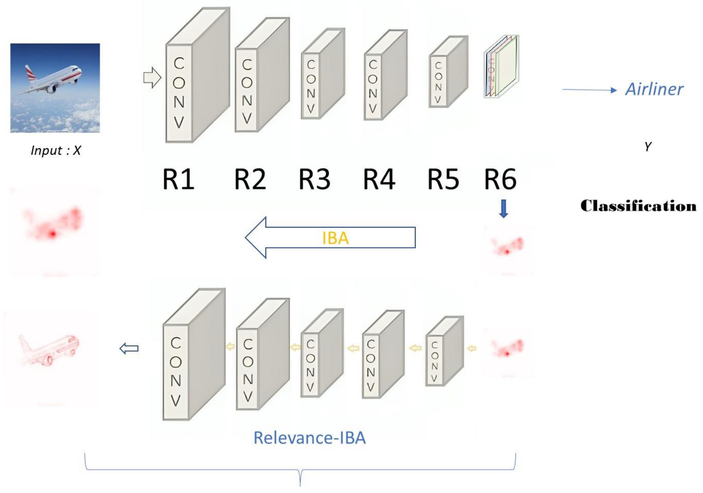
Aims

1. To propose the Relevance-IBA method, which combines IBA with the LRP's backpropagation method, aiming to enhance the clarity of attribution maps and make them more aligned with human intuition.

2. To advocate for a segmentation-oriented evaluation technique that emphasizes interpretability techniques' ability to highlight clear object boundaries and intricate visual details, aligning more closely with human perception.

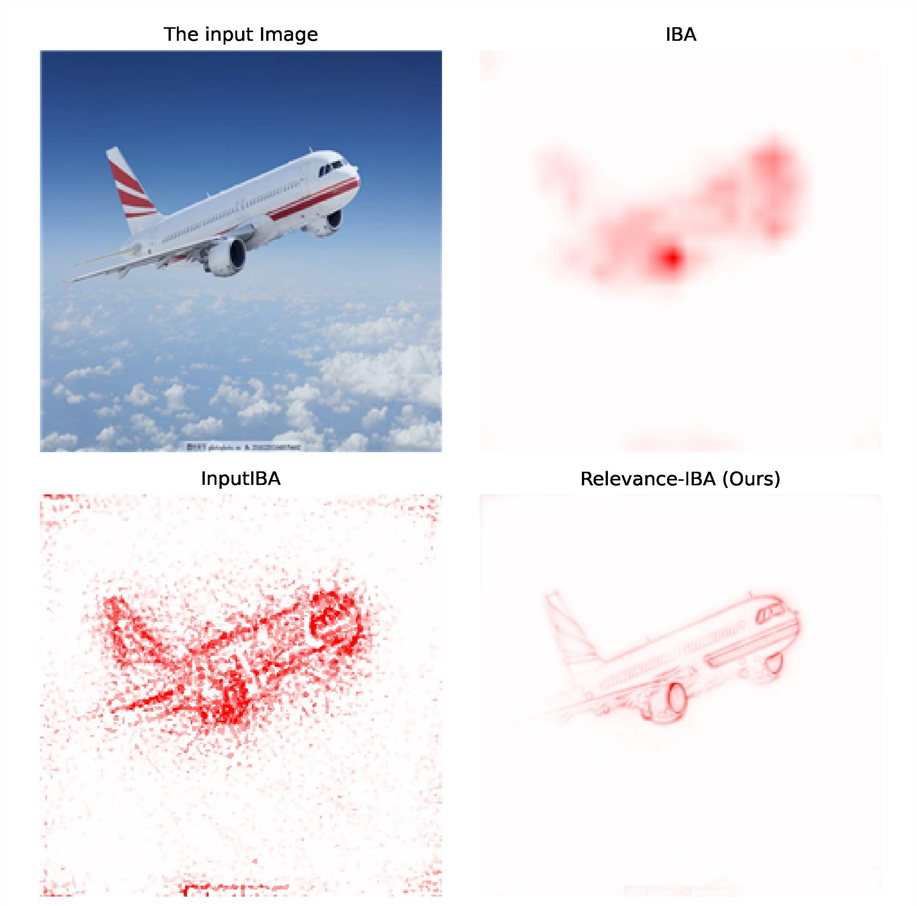
Methods

We introduce Relevance-IBA as an advancement over the traditional IBA. While IBA directly transfers the importance score to the input layer by leveraging the spatial invariance of the CNN and bypassing intermediate layers, Relevance-IBA offers a more nuanced propagation method. It transfers the importance score to the input layer in a step-by-step manner, following a sequence from the deepest layer to the front layer and then to the Input.

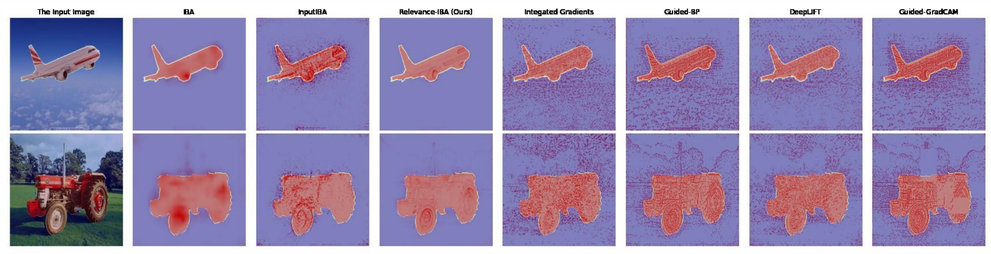


We introduce Relevance-IBA to employ a more advanced propagation method, transferring the importance score to the input layer step by step. For instance, the importance score flows in the sequence: R6 → R5 → R4 → R3 → R2 → R1 → Input.

Results



Relevance-IBA is capable of generating pixel-wise attribution maps, marking a substantial advancement over IBA and InputIBA. This enhancement leads to greater interpretability that aligns with human perception by accurately highlighting the contours and intricate details of the predicted object.



A demonstration of pixel importance allocation by different methods. Our method predominantly concentrates important pixels within the object segments, indicating its efficacy and precision.

References

[1] K. Schulz, L. Sixt, F. Tombari, and T. Landgraf, "Restricting the flow: Information bottlenecks for attribution," in International Conference on Learning Representations, 2020. [Online]. Available: https://openreview.net/forum?id=S1xWh1rYwB

[2] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. M¨uller, and W. Samek, "On pixel-wise explanations for non-linear classifier deci- sions by layer-wise relevance propagation," PloS one, vol. 10, no. 7, p. e0130140, 2015.

Australian Government
Department of Industry, Science, Energy and Resources
AusIndustry Cooperative Research Centres Program