Explainable AI - Understanding and Interpreting Computer Vision Models

Gunti Lahari
IIT Hyderabad
ai22btech11008@iith.ac.in

J Hima Chandh
IIT Hyderabad
ai22btech11009@iith.ac.in

S Divija
IIT Hyderabad
ai22btech11026@iith.ac.in

K Anuraga Chandan
IIT Hyderabad
ai22btech11011@iith.ac.in

***Abstract*—Explainable Artificial Intelligence (XAI) plays a critical role in making deep learning models more interpretable and trustworthy, particularly in high-stakes domains like autonomous driving and healthcare. However, existing XAI methods often struggle with faithfulness and plausibility, limiting their effectiveness in object detection and image classification tasks. This study reviews two key papers addressing these challenges: Human Attention-Guided Explainable AI for Computer Vision, which integrates human attention to improve XAI methods, and Model-Agnostic Explainable AI for Object Detection in Image Data, which focuses on creating interpretable explanations that work across different object detection models. By analyzing their methodologies, contributions, and limitations, this study highlights how human attention data and model-agnostic techniques can enhance explainability. A clear understanding of AI decisions, biases, and vulnerabilities is essential for trust, enabling better model refinement and performance optimization in real-world applications..**

***Keywords*—XAI (Explainable AI), Grad-CAM Grad-CAM++, Human Attention Deep Learning , saliency map Computer vision, Black-box testing Object detection, Hierarchical masking**

I. Introduction

Explainable AI (XAI) refers to techniques and methods that help interpret and understand the decision-making process of AI models, particularly deep learning models. In computer vision, deep neural networks, such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), have demonstrated impressive performance but are often seen as ”black boxes” due to their complex internal representations. XAI aims to make these models more transparent, interpretable, and trustworthy by providing explanations for their predictions.

Explainable AI (XAI) ensures trust, transparency, and accountability in computer vision by helping users understand model decisions. This is crucial in high-stakes fields like healthcare, finance, and autonomous systems, where unexplained errors can have serious consequences. XAI aids in debugging and bias detection by revealing hidden biases and model misinterpretations, improving robustness. It also supports regulatory compliance, as laws like GDPR mandate explainability in sensitive applications. Additionally, XAI enhances human-AI collaboration, enabling users to interpret predictions and make better decisions. By fostering trust and reliability, XAI makes AI-driven vision systems more ethical and effective.

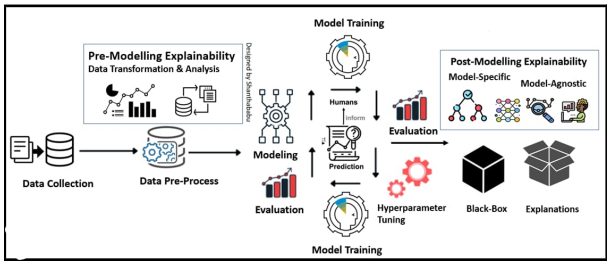

Fig. 1. A basic XAI Framework

Here are different types of methods evolved over time:

1) *Saliency-Based Methods* (*Gradient-Based XAI*): Saliency-based methods highlight important image regions that influence model predictions. Saliency maps compute gradients to identify key pixels, while deconvolution networks backpropagate activations to reconstruct relevant features. Guided Backpropagation enhances this by suppressing irrelevant gradients for clearer visualizations, improving interpretability.
2) *Class Activation Mapping* (*CAM*)*and its Variants*: CAM-based methods generate heatmaps to highlight image regions most responsible for a model's classification decision. Class Activation Mapping (CAM) uses global average pooling to assign weights to feature maps, but it only works for CNNs without fully connected layers. Grad-CAM extends this to a wider range of architectures by using gradients to weight feature maps. Guided Grad-CAM combines Grad-CAM with Guided Backpropagation for high-resolution, class-discriminative visualizations. Smooth Grad-CAM++ further improves Grad-CAM by reducing noise and enhancing feature importance, making explanations clearer and more reliable
3) *Perturbation-Based Methods* Perturbation-based methods explain AI decisions by modifying input data and analyzing how predictions change. LIME approximates complex models with simpler ones by perturbing inputs and observing prediction shifts. DLIME extends this for deep learning models, enhancing local explanations. SHAP uses Shapley values to fairly assign feature importance, ensuring consistent explanations. RISE generates class-specific heatmaps by

occluding image regions and measuring their impact on predictions, while dRISE improves this for deep learning, providing more precise class-discriminative insights.

4) *Relevance Propagation and Attribution Methods*
These methods trace important pixels by propagating relevance backward. LRP assigns pixel-wise scores, Deep Taylor Decomposition refines this using second-order approximations, and Integrated Gradients averages gradients from a baseline to the input for better interpretability.

Despite advancements, making AI models interpretable remains challenging. There is often a trade-off between accuracy and interpretability, requiring methods that balance both. Different XAI techniques may produce conflicting explanations, highlighting the need for standardization. Real-world deployment demands adaptation beyond research settings, especially in critical fields like healthcare and finance. Additionally, ensuring explanations align with human reasoning and domain expertise is essential for improving trust and usability.

In this project, we aim to replicate and analyze explainability methods from two key research papers: HAG-XAI and BODEM. By implementing these techniques, we will evaluate their effectiveness in generating interpretable explanations for deep learning models in computer vision. Our goal is to compare their performance, understand their strengths and limitations, and explore potential improvements for better explainability.

II. Problem Statement

Understanding the reasoning behind predictions made by deep learning models is essential, especially in image classification and object detection. While Explainable AI (XAI) techniques attempt to address this need, existing methods such as Saliency Maps, Class Activation Mapping (CAM), Grad-CAM, LIME, SHAP, and Layer-wise Relevance Propagation (LRP) suffer from various limitations:

- Gradient-based methods (e.g., CAM, Grad-CAM, LRP) require internal model structures, making them unsuitable for black-box settings.
- Saliency Maps often fail to precisely localize the most critical regions influencing a model's decision.
- Perturbation-based methods (e.g., LIME, SHAP) can be computationally expensive and produce inconsistent explanations.
- Existing black-box explanation methods like RISE struggle with noise and misattribution of irrelevant pixels, leading to misleading explanations.

To address these challenges, this research focuses on two explainability approaches designed for deep learning-based image classification and object detection models:

1) Human Attention-Guided XAI (HAG-XAI)
   - Enhances explainability by integrating human attention priors into saliency maps.
   - Utilizes learnable Gaussian smoothing and activation weighting to refine explanation quality.
   - Employs loss functions such as Pearson Correlation Coefficient (PCC) and RMSE loss to align explanations with human perception.
2) Black-box Object Detection Explanation by Masking (BODEM)
   - Introduces a model-agnostic XAI approach that does not require access to gradients, class probabilities, or objectness scores.
   - Implements hierarchical structured masking to improve stability and precision in saliency maps.
   - Reduces noisy attributions observed in traditional black-box methods like RISE, improving reliability.

Key Research Objectives

- Compare BODEM with existing XAI methods (e.g., Grad-CAM, LIME, human-guided attention models) across various datasets
- Assess interpretability, localization accuracy, computational efficiency, and human alignment of different explainability techniques
- Investigate vulnerabilities in explainability approaches, including adversarial sensitivity and unintended biases.
- Improve model robustness and fairness by refining XAI methods that mitigate misleading explanations.

By systematically evaluating human-guided and model-agnostic XAI methods, this research aims to enhance transparency, improve decision interpretability, and contribute to the development of more explainable and resilient AI systems.

III. Literature Review

In this we include explanation from two papers [1] and [2] respectively.

A. *Human Attention-Guided Explainable Artificial Intelligence for Computer Vision Models[1]*

We had already mentioned about saliency maps above. These Saliency-based explanations are evaluated in two ways:

- *Faithfulness:* How well the highlighted features truly influence the AI's decision. This is tested by removing or adding highlighted parts and checking the model's performance change.
- *Plausibility:* How understandable the explanation is to humans, often measured through human judgments.

These limitations make current XAI methods unsuitable for safety-critical systems like autonomous driving and medical diagnosis, highlighting the need for better explainability in object detection models.

Using Humsn Attention to Improve XAI.. Human attention is increasingly used in AI to enhance model explanations. In tasks like Visual Question Answering (VQA), it helps AI focus on relevant details, improving predictions and interpretability. While common in some AI fields, its use in object detection is still rare. Recently, human attention has also been used to assess the credibility of AI explanations by comparing them to human focus patterns. This approach helps

ensure that AI-generated explanations are more intuitive and understandable.

Since collecting human attention data is slow and difficult, researchers created a model called Human Saliency Imitator (HSI) that learns from human data to generate similar attention maps automatically. This helps evaluate AI explanations for image classification models.

However, HSI relies on deep learning, making it hard to interpret, and it doesn't work for object detection since it can't focus on specific objects. So far, there's no explainable version of HSI for object detection. this study explores whether human attention patterns can improve them.

In image classification (object recognition) and object detection (visual search), human attention—observed through eye movements—helps distinguish objects from others or efficiently locate targets. This suggests that human attention could guide AI models toward important features that are both accurate (faithful) and understandable (plausible).

They first tested two common gradient-based XAI methods (Grad-CAM and Grad-CAM++) and then developed FullGrad-CAM and FullGrad-CAM++ to explain object detection models. They compared the saliency maps from these methods with human attention maps. Next, they created HAG-XAI, a model trained on human attention data, which combines different explanatory features to make AI explanations more human-like. Finally, they tested whether HAG- XAI improved explanations across different tasks and datasets. Since its parameters are interpretable, it also helps understand what makes explanations better. Additionally, HAG-XAI could generate human-like attention maps for benchmarking object detection models.

1) Study1: COMPARISONS BETWEEN XAI SALIENCY MAPS AND HUMAN ATTENTION MAPS :

1) *Models and Dataset Used for Image Classification*
For image classification, the study used ResNet-50 and Xception models. The dataset was a subset of the ImageNet database, which originally contains 1,000 image classes.
A subset of 144 images from the testing set was selected, covering 18 classes:
   - 8 natural objects (e.g., ants, corn, lemon, etc.)
   - 10 artificial objects (e.g., broom, laptop, phone, etc.)

   Since the images had inconsistent resolutions, they were padded with white borders to a unified resolution of 520 × 400. The input layers of both models were modified to accept images of this resolution.

2) *Models and Dataset Used for Object Detection*
For object detection, the study used:
   - YOLO-v5s (one-stage model)
   - Faster R-CNN (two-stage model)

   Since XAI for object detection is crucial for automated driving safety, they used BDD-100K, a well-annotated driving image database. All images had a resolution of 1280 × 720.
   For training and testing:
   - Training: 69,400 images with five labels (car, truck, bus, person, rider)
   - Testing: 10,000 images (validation set)

   For further experiments, two independent test datasets (A B) were randomly selected from the validation set, each containing 160 images.
   To test generalization, they used a pretrained model on the MS-COCO database (which has 80 object classes).
3) *Human Attention Data for Image Classification*
The human attention data used in this study were obtained from Qi et al which included two tasks.
   - Classification (Cls) Task – Participants judged the class label of an image.
   - Explanation (Exp) Task – Participants explained why an image belonged to a certain class.

   Experimental Setup:
   - The experiment was conducted on a laptop with a $255\,\mathrm{mm} \times 195\,\mathrm{mm}$ screen.
   - Display resolution: 1024×768 pixels.
   - Each image was displayed at $9.68^\circ \times 12.32^\circ$ of visual angle (viewing distance: $60\,\mathrm{cm}$).
   - Eye fixation data were smoothed using a Gaussian kernel ($\sigma = 21$ pixels, $\sim 1^\circ$ visual angle) to generate human attention maps.

   Task Procedures:
   a) Classification Task
      i) A drift check was performed at the screen center.
      ii) A fixation cross appeared in the upper left corner.
      iii) After fixating for 250 ms, an image appeared on the right side.
      iv) The participant named the class label aloud, and the image disappeared once their response was detected.
   b) Explanation Task
      i) A similar procedure was followed, but a class label appeared where the fixation cross was.
      ii) A textbox below the label allowed participants to type explanations.
      iii) They were asked to explain as if to a young child.
      iv) The trial ended when they pressed Enter after typing their explanation.
4) *Human Attention Data for Object Detection*
To generate human attention maps for object detection, we collected human eye movement data during a vehicle detection task
   Experimental Setup:
   - Monitor**:** 15.6-inch screen (1920 × 1080 pixels).
   - Image resolution: 1024 × 576 pixels.
   - Viewing distance: 55 cm.

- Visual angle: 34.2° × 20.8°.

Task Procedure:

a) A solid circle appeared at the screen center for a drift check.
b) A fixation cross was shown for 0.5 seconds.
c) A driving scene image was displayed at the screen center.
d) Participants searched for vehicle objects ('car', 'truck', 'bus').
e) They pressed the spacebar when they detected all targets.
f) The screen then turned blank, and participants clicked on detected target locations using a mouse.

For image classification, saliency maps were computed using the last convolutional layer before global average pooling.

For object detection,

- YOLOv5s: Used the last convolutional layer in the neck module, considering the multi-scale branch.
- Faster RCNN: Used the last convolutional layer of the backbone for global saliency.

Traditional gradient-based methods produced noisy maps, with Faster RCNN showing grid-like patterns due to 7×7 ROI pooling, while YOLOv5s had more focused saliency.

2) Methods: **XAI Methods**

For image classification Grad-CAM and Grad-CAM++ are used for explanation.The vanilla Grad-CAM highlights regions based on the importance of features with respect to a certain class score. Let M(.) is an is an object detection model(ex: yolo-v5) then $y_m = M(I)$ for input image I. where $m = 1, 2..., N_{obj}$ is the output classification probability of m^{th} detected object. The Grad-CAM map can be expressed as :

$$S_G = \sum_{m=1}^{N_{\text{obj}}} \mu \left(ReLU \left(\sum_{k=1}^{N_{\text{ch}}} \frac{1}{Z} \sum_{ij} \frac{\partial y^m}{\partial A_{ij}^k} A^k \right) \right), \quad (1)$$

where A_k(Activation map in k^{th} layer) $= \sum(w_i f_i)$ f_i is feature map of the last convolutional layer , is the max-min normalization function, N_{ch} is the number of channels in A_k. Z is a normalization term defined by a global average pooling operation

In GRAD-CAM++ modified gradient term as :

$$S_G^* = \sum_{m=1}^{N_{\text{obj}}} \mu \left(ReLU \left(\sum_{k=1}^{N_{\text{ch}}} \frac{1}{Z} \sum_{ij} \alpha_{ij}^{km} ReLU \left(\frac{\partial y^m}{\partial A_{ij}^k} \right) A^k \right) \right), \quad (2)$$

where α_{jm}^{ij} is a coefficient in (i, j) position for m^{th} detected object and A ReLU function is applied to the gradient term to retain the most important features with a positive gradient value.

FullGrad-CAM & FullGrad-CAM++ for Object Detection

In object detection models, gradient maps contain useful spatial information, but Grad-CAM and Grad-CAM++ ignore this due to global average pooling. This leads to saliency maps that highlight irrelevant areas, reducing correlation with detected targets. To address this remove global average pooling, preserving spatial details. Generate object-specific saliency maps, improving interpretability for object detection models.

The FullGrad-CAM is defined as

$$S_F = \sum_{m=1}^{N_{\text{obj}}} \mu \left(ReLU \left(\sum_{k=1}^{N_{\text{ch}}} \frac{\partial y^m}{\partial A^k} \odot A^k \right) \right), \quad (3)$$

Full Grad-CAM++ is defined as

$$S_F^* = \sum_{m=1}^{N_{\text{obj}}} \mu \left(ReLU \left(\sum_{k=1}^{N_{\text{ch}}} ReLU \left(\frac{\partial y^m}{\partial A^k} \odot A^k \right) \right) \right), \quad (4)$$

Traditional gradient-based methods produced noisy maps, with Faster RCNN showing grid-like patterns due to $7 * 7$ ROI pooling, while YOLOv5s had more focused saliency.

Evaluation metrics and FAG-XAI will be discussed in the next report

B. Model-agnostic Explainable Artificial Intelligence for Object Detection in Image Data (2)

Interpretable AI Approaches

1) XAI Methods:
White Box (Need model internal structure)
   - Interpretation requires AI expertise

   Black Box (Relies on only input and output of the model)
   - Explanations are often more practical
2) Techniques:
   a) Saliency Maps
      - Highlights the regions in an image that contribute more to the model's

      decision.
   b) Perturbation-based Methods
      - Systematically alters parts of an image to observe changes in the model's output.
      - Helps in identifying significant factors.
   c) Use of Interpretable Models
      - Decision trees or linear models, which, although less complex, offer greater transparency.
   d) Model-Agnostic Methods
      - LIME
      - SHAP
3) Challenges of LIME and SHAP:
   While powerful and popular for explaining ML model predictions, they face significant challenges when applied to object detection models.
   - Requires access to class probabilities or objectness scores for the prediction (not always available in black-box testing).

- The random sampling approach used by these methods can lead to inconsistency in the saliency maps generated. - This may result in noisy explanations.

Explanation Methods for CV Applications:
Various explanation methods have been developed for Computer Vision (CV) applications. However, most of these methods focus on white-box explanations.

- Few address black-box explanations for Deep Learning (DL) models in image processing tasks.

1) RISE
   - Estimates a saliency map by probing the object classification model using randomized masking of the input image.
   - Calculates importance scores for pixels by measuring the difference in class probabilities before and after masking.
   - DRISE extends this pixel-wise image masking strategy to provide an attribution method for explaining object detection models.
2) Challenges of RISE:
   For these methods, there are challenges:
   - Requires probability scores over classes and an objectness score for every bounding box to estimate the saliency map.(This limits their utility when only bounding box coordinates are available.)
   - Random masking can result in equal importance being assigned to both relevant and irrelevant pixels.

Proposed Solution: BODEM
To address the above challenges, we propose BODEM.

- It is a model-agnostic method that can be applied to any object detection system, regardless of the underlying ML model.
- Importantly, it does not require access to class probabilities or objectness scores.
- This makes it suitable for black-box explanation scenarios where only bounding box predictions are available.

Stages in BODEM:
There are 3 stages:

1) Hierarchical random mask generation
2) Model Inquiry
3) Saliency Estimation

Problem Formulation:
Let

- I be the input image of dimensions WH.
- An object detector $f(I)$ outputs detections where:

$$f(I) \rightarrow \text{O= } \{\text{o}_1, o_2, o_3, ..., o_n\}$$

The set of objects detected by f is represented as follows:

- Each object o_n is represented by bounding box coordinates in 2D space:

$$o_n = (x_1, y_1, x_2, y_2) \quad (2)$$

- Our objective is to generate a saliency map SM_n for each detected object o_n.
- The saliency maps have the same dimensions as the input image.

Explanation
Our method assigns values that indicate the importance of each pixel in relation to the target object.

- Uses a random mask generation technique combined with a hierarchical masking algorithm.
- Generates a final saliency map with smoother salient areas.
- Incorporates controlled randomness to reduce noise.

Mask Generation
A common approach for providing black-box explanations in image processing models involves masking different parts of the input image.

- A primary issue is that random masks may simultaneously cover both relevant and irrelevant pixels.
- Since object detection models rely on masked inputs, irrelevant pixels may receive importance scores similar to relevant ones, resulting in noise in the saliency maps.
- To overcome this, we propose a masking technique that integrates both random and hierarchical masking strategies.
- The hierarchical approach ensures that the final saliency map highlights the most important regions with minimal noise.

Model Inquiry

- **Mediator Role:** The model inquiry module acts as a bridge between the explanation model and the object detection model, forwarding masked images and retrieving detected bounding boxes.
- **Model-Agnostic Nature:** It does not require knowledge of the detection model's architecture, loss function, or parameters, making it adaptable to any object detection system.
- **Black-Box Compatibility:** This method is ideal for black-box testing scenarios, as it can explain object detection models without needing internal model details.

Saliency Estimation
The primary function of saliency estimation is to determine the importance of pixels within the input image by assessing the difference between the original and new predictions. It then updates the saliency map SM_n, which represents the significance of various parts of the image to the object o_n. Initially, the saliency map SM_n is filled with zero values.

REFERENCES

[1] Guoyang Liu, Jindi Zhang , Antoni B. Chan,Janet H. Hsiao "Human Attention-Guided Explainable Artificial Intelligence for Computer Vision Models" arXiv:2305.03601v1 [cs.CV] 5 May 2023

[2] Milad Moradi, Ke Yan,David Colwell,Matthias Samwald,Rhona Asgari "Model-agnostic explainable artificial intelligence for object detection in image data"Tricentis GmbH, Leonard-Bernstein-Straße 10, 1220 Vienna, Austria

[3] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in Proceedings of the IEEE international conference on computer vision, Conference Proceedings, pp. 618–626.
[4] https://www.sciencedirect.com/org/science/article/pii/S1526149224003084