IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, June 2025
Impact Factor: 7.67

MAIA – A Multimodal Automated Interpretability Agent for Explainable AI

Vedant Jawalekar[1], Vijay Hatte[2], Naman Shetty[3], Nikita Khawase[4], Anil Walke[5]
Students, Artificial Intelligence & Data Science[1,2,3]
Head of Department - Artificial Intelligence & Data Science[4]
Assistant Professor - Artificial Intelligence & Data Science[5]
ISBM College of Engineering, Pune, India

Abstract: *Explainable Artificial Intelligence (XAI) has emerged as a critical field to demystify the decision-making process of complex deep learning models. This paper introduces MAIA – a Multimodal Automated Interpretability Agent – developed as a web-based platform to enhance interpretability, fairness analysis, and transparency of AI models. MAIA offers an integrated set of modules for neuron visualization, bias detection, feature attribution, and natural language summarization. Designed for educators, researchers, and developers, MAIA leverages advanced techniques such as Grad-CAM, Integrated Gradients, and Pegasus-XSum to interpret and present AI decisions in an understandable way. This paper details the architecture, implementation, and real-world use cases demonstrating MAIA's capabilities as a complete XAI toolkit.*

Keywords: Explainable AI, Model Interpretability, Grad-CAM, Pegasus-XSum, Bias Detection

I. INTRODUCTION

Despite their performance, modern deep learning models often function as black boxes, leaving users uncertain about the reasoning behind predictions. This lack of interpretability hampers adoption in sensitive domains such as healthcare and education. MAIA, a Multimodal Automated Interpretability Agent, addresses this challenge by enabling users to engage with deep learning models through visual and textual insights. It is particularly useful for students, educators, and developers seeking to understand and trust AI systems.

II. CORE MODULES AND SYSTEM DESIGN

A. Neuron Visualizer

This module offers a graphical interface to observe neuron activations across convolutional layers. Using Grad-CAM, it highlights the spatial regions of an image that contribute most to a model's output, helping users visually trace the reasoning path of the network.

B. Bias Detection Engine

By analysing the model's responses to inputs with varying demographic features, this component identifies inconsistencies and flags potential biases. It helps assess fairness in model behaviour across different user groups.

C. Feature Attribution Tool

Built with the Integrated Gradients technique, this module assigns scores to input pixels or regions based on their influence on the output. A corresponding bar chart helps visualize these contributions, offering insight into feature importance.

D. Summarization Interface

To aid non-technical users, MAIA incorporates a text summarization engine powered by the Pegasus-XSum model. It converts dense model outputs into simple, readable summaries, enhancing accessibility and understanding.

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-27520

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, June 2025
Impact Factor: 7.67

III. TECHNOLOGICAL FOUNDATION

- **Frontend:** React.js, styled with Tailwind CSS
- **Backend:** Django framework with REST API support
- **AI Models:** ResNet18 (for image processing), Pegasus-XSum (for summarization)
- **Libraries**: PyTorch, Captum (for interpretability), Hugging Face Transformers, Matplotlib
- **Custom APIs:** Handle image uploads, heatmap rendering, bias logic, feature scoring, and summary generation

IV. APPLICATIONS AND USE CASES

MAIA is built for a diverse audience:

- **Students** exploring deep learning architectures
- **Educators** demonstrating model behavior in classrooms
- **Researchers** studying ethical AI systems
- **Developers** performing sanity checks on model decisions before deployment

Its intuitive interface and explainable outputs make it particularly suitable for academic and instructional environments.

V. RESULTS AND EVALUATION

When tested on classification tasks, MAIA successfully identified key visual cues used by models, detected subtle prediction biases, and produced faithful text summaries. Heatmaps generated for different layers illustrated model focus areas, while feature attribution confirmed their consistency. The summarization component also improved user understanding during qualitative evaluations.

VI. CONCLUSION

MAIA presents a unified approach to AI interpretability by combining multiple techniques in a single interface. It empowers users to scrutinize and trust model behaviour through visualisation, fairness checks, and plain-language summaries. With future integration of more complex data types and model architectures, MAIA aims to serve as a standard tool in the XAI landscape.

ACKNOWLEDGMENT

I express sincere gratitude to my faculty and peers who guided the development of MAIA. I also acknowledge the open-source communities behind PyTorch, Captum, and Hugging Face, whose tools were instrumental in building this platform.

REFERENCES

[1] R. R. Selvaraju et al., “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” ICCV, 2017.
[2] M. Sundararajan et al., “Axiomatic Attribution for Deep Networks,” ICML, 2017.
[3] J. Zhang et al., “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization,” ICML, 2020.
[4] M. T. Ribeiro et al., “Why Should I Trust You?: Explaining the Predictions of Any Classifier,” KDD, 2016.