Explainable AI – Neural Network Visualization Tool

Sarib Samdani &
Advait Vinod

Motivation

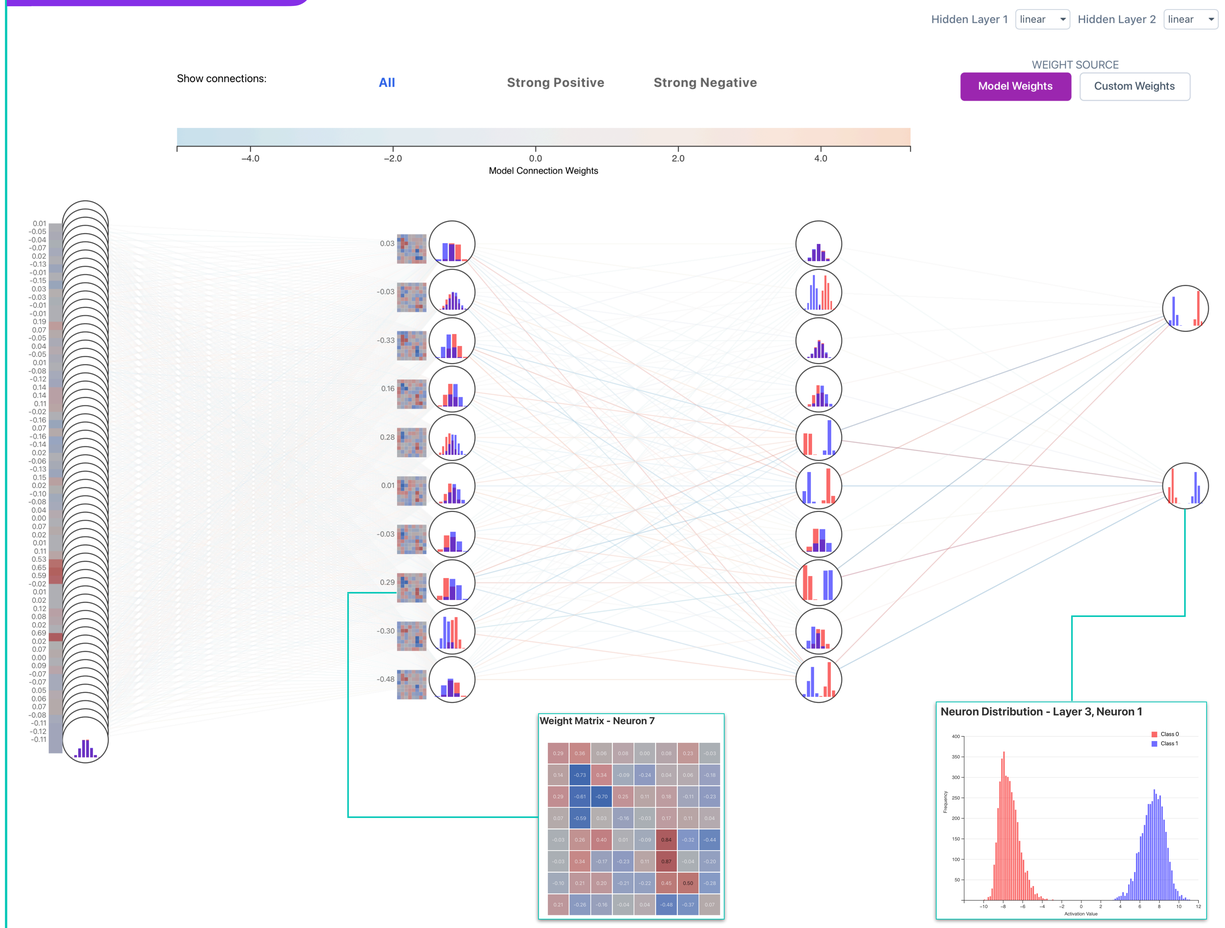
- Neural networks' inherent opacity poses fundamental challenges in critical domains (e.g. healthcare) where understanding decision-making processes is essential for safety and accountability
- Current black-box nature severely limits deployment, as it It conflicts strict EU laws
- Existing interpretability methods frequently produce unreliable or inconsistent explanations, undermining trust in model analysis
- Systematic visualization and verification of internal network dynamics represents a critical step toward bridging the interpretability gap in deep learning systems

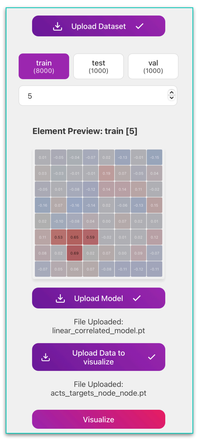
Background

- The paper XAI-TRIS [1] introduces a benchmark dataset for evaluating explainable AI methods in non-linear image classification tasks with known ground truth explanations, providing the foundation for our work and dataset we used to test this tool in a controlled environment
- Four classification scenarios are presented: linear, multiplicative, translations/rotations, and XOR, each using tetrominoes (only L or T shaped) overlaid on different background types to induce suppressor variables
- Experiments show that many popular XAI methods struggle to outperform random baselines and edge detection methods, particularly in scenarios with correlated backgrounds
- Results indicate that explanations can be inconsistent across equally-performing model architectures, highlighting potential risks of misinterpretation when deploying XAI in critical applications

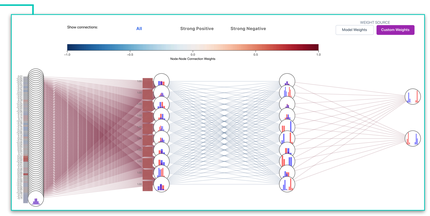
[1] Stefan Haufe, Benedict Clark & Rick Wilming: XAI-TRIS: non-linear image benchmarks to quantify false positive post-hoc attribution of feature importance

Method





- User has to upload their trained model and it's activations as PyTorch files
  - This provides all the necessary information to create the Neural Network, with every neuron containing its activations-histogram
  - Optionally the targets (defining the classes) and custom-defined node interaction values can be added, which are visualized instead of the weights. In the example the node interaction values are set to 1, -1 and 1
- Optionally the dataset can be uploaded
  - Visualizes pixel-to-neuron mappings of selected samples next to the input layer, alongside the learned weight patterns of the first hidden-layer neurons' reshaped as matrix



Technology

- Backend in Python
  - FastAPI with Uvicorn ASGI server
  - PyTorch integration for model loading and tensor operations with CPU support
  - Key functions: Validates input format; Extracts network structure, weights and activations; Manages file operations; Provides RESTful API endpoints with CORS
- Frontend in TypeScript/JavaScript
  - React, D3.js for activations visualisation as histogram for each neuron
  - HTTP polling for neural network visualization updates and state management
  - Dataset Visualization with custom color mapping and reshaping matrices for first hidden layer

Link to the Github Repository with source code and additional information: https://github.com/saribx/Explainable-AI-neural-network-visualisation-tool

Conclusion

- Developed a visualization tool that provides structured insights into neural network decision-making processes
- Enables methodical exploration of network behavior through activation and weight visualization, advancing model interpretability
- Offers integration with PyTorch models and datasets, supporting further XAI research
- Successfully tested on XAI-TRIS benchmark dataset, demonstrating the tool's utility in exploring model internal operations
- Open-source architecture enables immediate community contributions for expanding capabilities