Misc ML Paper Notes

William Yang

July 13, 2025

Contents

1 Multimodal Automated Interpretability 1
1.1 Intro 1

1 Multimodal Automated Interpretability

1.1 Intro

MAIA is an interpretability agent that can predict and explain the behavior of neural networks in computer vision models. Specifically, it does better in reducing sensitivity to spurious features and identify inputs that are misclassified.

Neuron description paradigm -

-