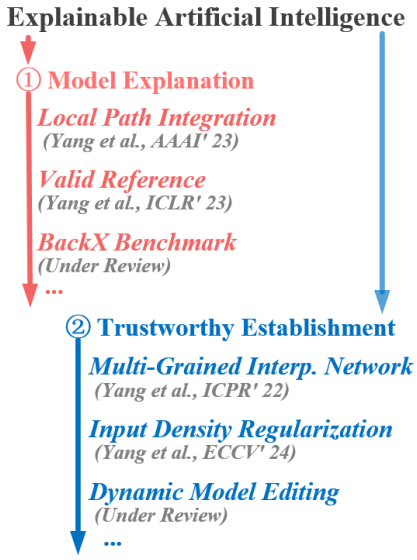
Peiyu Yang | Research Statement

Despite the great success of deep learning models across diverse domains, their inherent opacity results in a notable deficiency in the explainability of their decision-making process. This lack of transparency poses a significant barrier to deploying deep models in high-stakes applications.

Addressing this challenge, my research work focuses on explainable artificial intelligence (XAI), organizing around two main themes: ① **model explanations** and ② **establishing model trustworthiness**. For **model explanations**, I emphasize axiomatic attribution explaining methods and high-fidelity benchmarking criteria to assess explaining tools. Rather than prioritizing human interpretability, my work focus on ensuring the faithfulness of explanations to the model's decision making process. To **establish model trustworthiness**, I explore three complementary approaches: constructing self-explainable architectures, developing robust training strategies, and designing model editing methods for established models. In summary, these efforts improve transparency and reliability, promoting greater trust in deep learning systems. Below, I summarize my key contributions along these two themes and outline directions for future research.



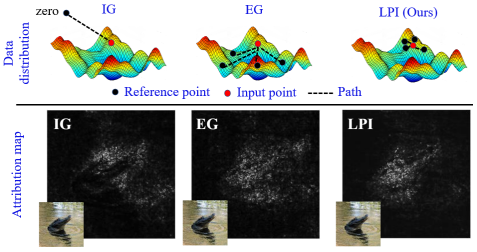
Research overview.

Research Progress

1 Model Explanation

Understanding the decision-making process of deep learning models is essential for their practical deployment. To enhance model transparency, I have developed novel attribution methods that provide additional reliability guarantees for model explanations [1, 2]. Moreover, I introduced a benchmark for explainable AI tools designed to conduct high-fidelity evaluations of explaining tools [3].

1) Local Path Integration for Model Prediction Attribution. Path attribution explaining methods often fail to satisfy the weak dependence explanatory axiom, which compromises their reliability in model explanation. To address this issue, we identified that path-based attribution can satisfy weak dependence by selecting a reference that invokes the same piecewise linear component. Building on this insight, we devise a process to identify the local distribution for the input. We then propose a Local Path Integration scheme aimed at stochastically integrating the model gradients over the references sampled from the local distribution, thereby ensuring reliability in path attributions. This work strengthens the theoretical foundations of the path-based attribution method, improving its reliability in model explanations.



Comparison of the proposed Local Path Integration (LPI) method [1] with existing methods.

2) Re-Calibrating Feature Attributions for Model Interpretation. The issue of reference choice ambiguity in path feature attribution methods presents a significant open challenge. To tackle this, we introduce a unified approach to re-calibrate a reliable reference for all integral-based attribution methods. Our method involves devising a technique to efficiently integrate over non-linear paths, enabling meaningful interpretations using actual attribution scores rather than absolute values. By incorporating our scheme, we achieve notable performance gains across a range of integral-based attribution methods, improving both local and global evaluation metrics. This recalibration method systematically eliminates reference selection ambiguity, offering a generalizable solution for improving the reliability of all path attribution methods.

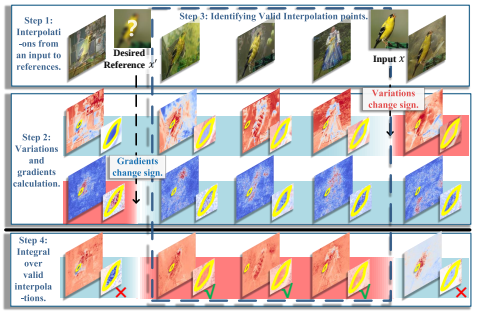


Illustration of recalibrating feature attributions using valid references [2].

3) XAI Benchmark for High Fidelity Evaluation. Assessing the faithfulness of attribution methods remains challenging due to the absence of accurate ground truth for attributing model predictions. To develop a reliable benchmark, we first

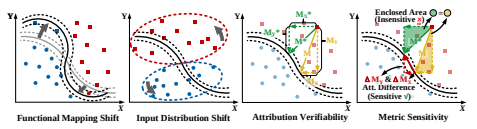
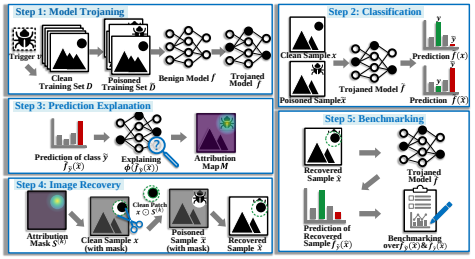


Illustration of fidelity criteria for explainable XAI benchmark [3].

identify a set of fidelity criteria that reliable benchmarks for attribution methods are expected to fulfill. Next, we leverage backdoor neural Trojans to design precise ground truth for attribution evaluations, introducing a Backdoor-based eXplainable AI benchmark (BackX). We theoretically establish the superiority of BackX in adhering to the desired fidelity criteria over the existing benchmarks. With extensive analysis, we also identify the setup of different attribution methods for fair benchmarking. This setup is ultimately employed for a comprehensive comparison of existing methods using our BackX benchmark. Our analysis also provides guidance for defending against backdoor attacks with the help of attribution methods. This work advances fidelity evaluation in XAI and provides new insights into the distinct characteristics of attribution methods.

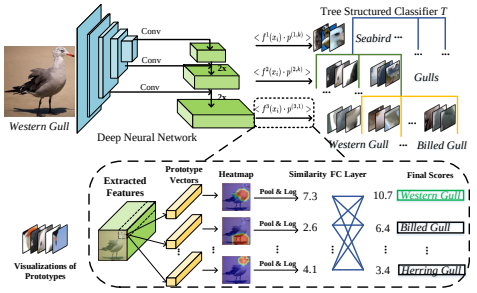


The pipeline of BackX benchmark [3].

2 Model Trustworthy Establishment

In addition to explaining deep models, it is equally crucial to establish trust in their decision-making processes. In pursuit of model trustworthiness, we established a multi-grained self-explainable network [4]. Next, we developed a regularization technique to regulate the model's reliance on non-robust features [5]. Furthermore, we proposed a dynamic model editing framework for correcting unreliable model behaviors.

1) Multi-Grained Interpretable Network. To ensure enough detail in explaining the model's predictions, we developed a tree-structured classifier capable of hierarchically classifying input images. This hierarchical approach allows prototypical features to be learned at different granularities. Consequently, the model is equipped to provide the model's decision-making pathway with varying levels of granularity in explanations, facilitating self-interpretable deep neural networks. Experimental results demonstrate that our model achieves competitive prediction accuracy and simultaneously produces high-quality explanations of its decisions. The proposed architecture enhances model transparency by offering multi-grained and human-aligned explanations, making deep learning more interpretable and actionable in critical applications.



Architecture of the proposed multi-grained interpretable network [4].

2) Robust Regularization Against Model Reliance on Non-Robust Features. To establish a model with general robustness, we begin by leveraging attributions to identify both robust and non-robust features. We then establish the correlation between the model's reliance on non-robust features and the smoothness of data marginal density. Motivated by the distinctive characteristics of features, we propose an efficient technique for regularizing the gradients of data density, which also addresses the numerical instability inherent in the underlying optimization problem. Extensive experiments validate the effectiveness of the proposed method, providing clear evidence of its capability to address the feature leakage problem and mitigate spurious correlations. This work provides insights into understanding feature robustness in deep learning and contributes to achieving general robustness across various domains.

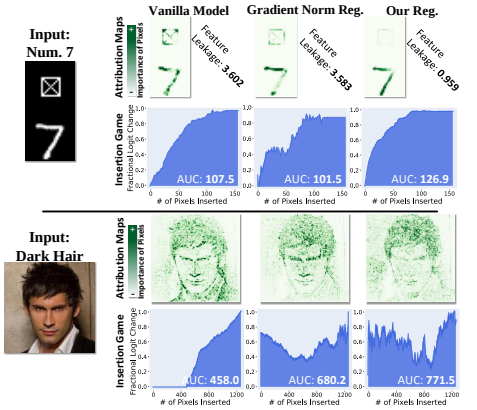
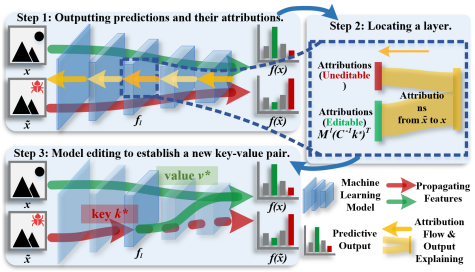


Illustration of the robustness of our regularization [5] on BlockMNIST and CelebA-Hair.

3) Dynamic Model Editing for Correcting Model Unreliable Behavior. The performance of neural network models deteriorates due to their unreliable behavior on corrupted input samples and spurious data features. Owing to their opaque nature, rectifying models to address this problem often necessitates arduous data cleaning and model retraining, resulting in huge computational and manual overhead. We proposed to employ model editing techniques to efficiently correct the model's unreliable behavior. Moreover, we introduce an effective attribution-based layer localization method, which identifies a primary layer responsible for the unreliable model behavior. Furthermore, we propose a dynamic model editing capable of dynamically adjusting the target editing layer. Additionally, our approach achieves notable performance and efficiency in correcting model unreliabilities with the requirement of only a single sample, indicating its potential for widespread applicability.



Workflow of dynamic model editing.

Future Research Directions

Moving forward, I aim to deepen my research on model interpretability and trustworthiness. With the growing impact of large language models (LLMs), understanding and mitigating their uncertainty is crucial for ensuring their safe deployment. In addition, I plan to develop techniques for diagnosing and correcting unreliable behaviors in deep models through explainability-driven probing methods.

1) Uncertainty Estimation in Large Language Models (LLMs). Despite their remarkable success, LLMs frequently generate hallucinated or unreliable outputs, raising concerns about their robustness and trustworthiness in real-world applications. Accurately quantifying uncertainty in LLM outputs is essential for improving their reliability, yet existing Bayesian neural network approaches are infeasible due to the sheer scale and inaccessibility of LLM parameters. To address this challenge, I plan to focus on two key aspects of uncertainty estimation in LLMs:

- **Uncertainty Decomposition.** The complexity of LLM outputs makes uncertainty estimation nontrivial, as different sources of uncertainty (e.g., epistemic and aleatoric) may contribute differently depending on the task. I aim to develop a structured uncertainty decomposition framework that disentangles these uncertainty components, providing interpretable and actionable insights into LLM confidence. This decomposition will enable more precise uncertainty quantification, ultimately improving reliability in high-stakes applications.
- **Revisiting Uncertainty Taxonomies for LLMs.** Traditional uncertainty categorization may not fully capture the nuances of LLM-generated content, particularly in open-ended generative tasks. I intend to investigate how uncertainty manifests differently in LLMs compared to traditional deep learning models, reassessing existing taxonomies and developing new formulations better suited to large-scale generative AI. This will provide a foundation for future research on uncertainty-aware LLMs.

3) Probing Unreliable Behaviors in Deep Models. Beyond LLMs, deep models used in fields such as vision and structured data analysis continue to exhibit unreliable behaviors that are difficult to diagnose. While existing explainability tools provide post hoc insights, they often fail to proactively uncover failure cases. I plan to develop an active probing framework that systematically analyzes model behavior, allowing for more targeted identification of failure modes. This framework will be integrated with model editing techniques to dynamically adjust model responses, enhancing robustness without requiring full retraining.

References

[1] Peiyu Yang, Naveed Akhtar, Zeyi Wen, and Ajmal Mian. Local path integration for attribution. In *AAAI Conference on Artificial Intelligence, AAAI*, 2023.

[2] Peiyu Yang, Naveed Akhtar, Zeyi Wen, Mubarak Shah, and Ajmal Mian. Re-calibrating feature attributions for model interpretation. In *International Conference on Learning Representations, ICLR*, 2023.

[3] Peiyu Yang, Naveed Akhtar, Jiantong Jiang, and Ajmal Mian. Backdoor-based explainable AI benchmark for high fidelity evaluation of attribution methods. *arXiv preprint arXiv:2405.02344*, 2024.

[4] Peiyu Yang, Zeyi Wen, and Ajmal Mian. Multi-grained interpretable network for image recognition. In *International Conference on Pattern Recognition, ICPR*, 2022.

[5] Peiyu Yang, Naveed Akhtar, Mubarak Shah, and Ajmal Mian. Regulating model reliance on non-robust features by smoothing input marginal density. In *European Conference on Computer Vision, ECCV*, 2024.