FACADE
A Framework for Adversarial Circuit Anomaly Detection and Evaluation
AdvML ICML 2023 Workshop ▪ Blue Sky

Dhruv Pai* [1] Andres Carranza* [1] Rylan Schaeffer* [1] Arnuv Tandon* [1] Sanmi Koyejo [1]

[1]Computer Science, Stanford University

Summary

- We propose FACADE: a geometric & probabilistic framework for unsupervised mechanistic anomaly detection in deep neural networks, geared towards adversarial attack mitigation.
- FACADE elucidates circuit contributions to the properties of high-dimensional activation modes, aiding in adversarial attack identification.
- This framework should bolster model robustness, facilitate scalable model oversight, and demonstrate potential for real-world applications.

Introduction

- The growth of machine learning has resulted in powerful yet less interpretable models, increasing susceptibility to adversarial attacks.
- Complex models present detection challenges, & misuse can risk societal harm.
- We introduce an unsupervised framework for anomaly detection in these models.
- Leveraging circuit-based analysis and probabilistic models, our method aids in identifying deviations and potential adversarial attacks.

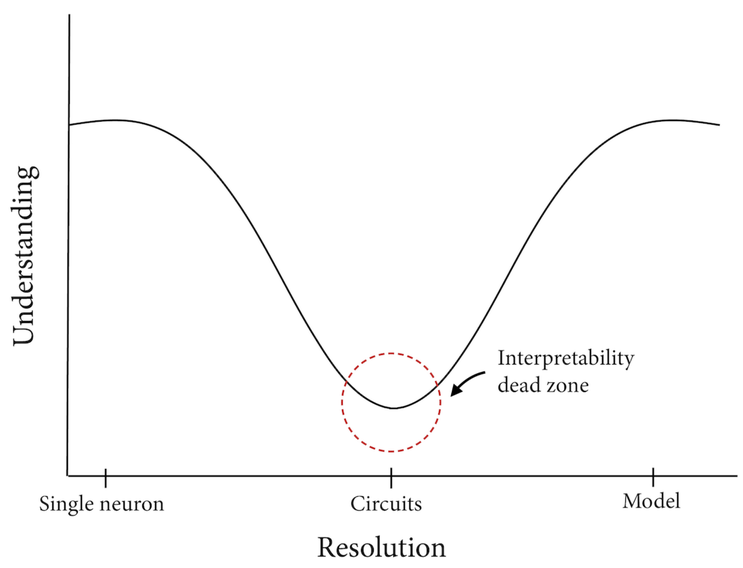

Figure 1. Interpretability Dead Zone

Activation Space

- Neural network activations are models in high-dimensional activation space, providing mechanistic behavior insights upon evaluation.
- Insights into data propagation within this high-dimensional space are crucial for enhancing model reliability and security.
- Adversarial examples manipulate decision boundaries in this space.
- Techniques that illuminate these boundaries and activation space geometry can bolster adversarial robustness.
- Analyzing activations at scale is computationally intensive and can lack transparency.

Circuit Mechanisms

- Circuits, subgraphs of a neural network's computational graph, enhance model interpretability by highlighting meaningful properties.
- They provide an effective bridge between single-neuron and whole-model interpretability across different architectures and datasets.
- Traditional circuit interpretability, focusing on specific adversarial or visual features, requires impractical prior knowledge of adversarial attacks in real-world scenarios.
- An unsupervised approach to circuit interpretability could unveil mechanistic anomalies, improving model invariance, efficiency, and scalability.

References

[1] U. Cohen, S. Chung, D. D. Lee, and H. Sompolinsky.
Separability and geometry of object manifolds in deep neural networks, 2020.

[2] A. Conmy, A. N. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso.
Towards automated circuit discovery for mechanistic interpretability, 2023.

[3] T. Gebhart, P. Schrater, and A. Hylton.
Characterizing the shape of activation space in deep neural networks, 2019.

[4] K. Wang, A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt.
Interpretability in the wild: a circuit for indirect object identification in gpt-2 small, 2022.

FACADE Approach

We present here the FACADE approach to unsupervised adversarial circuit detection:

- Identify high-dimensional modes of input data propagating through networks, e.g., using probabilistic Dirichlet Process Mixture models for unsupervised clustering in intermediate activation space for a given density threshold λ
- Elucidate circuits responsible for pseudoclass formation and propagation through causal discovery and Automatic Circuit DisCovery (ACDC)
- Determine manifold and kernel density properties of pseudoclass propagation through circuits and in relation to final classes through mean-field theoretic approximation
- Generate a distribution over circuits as they contribute to changes in manifold properties of pseudoclasses as they propagate through the network, e.g., effective reduction in radius or dimension
- Repeat the above algorithm for a sweep of λ values allows for circuit distribution evaluation across a variety of features and mechanistic pathways.

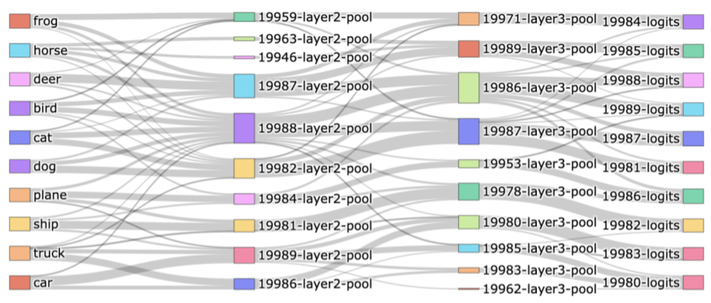

Figure 2. Pseudoclass Propagation Identified through FACADE

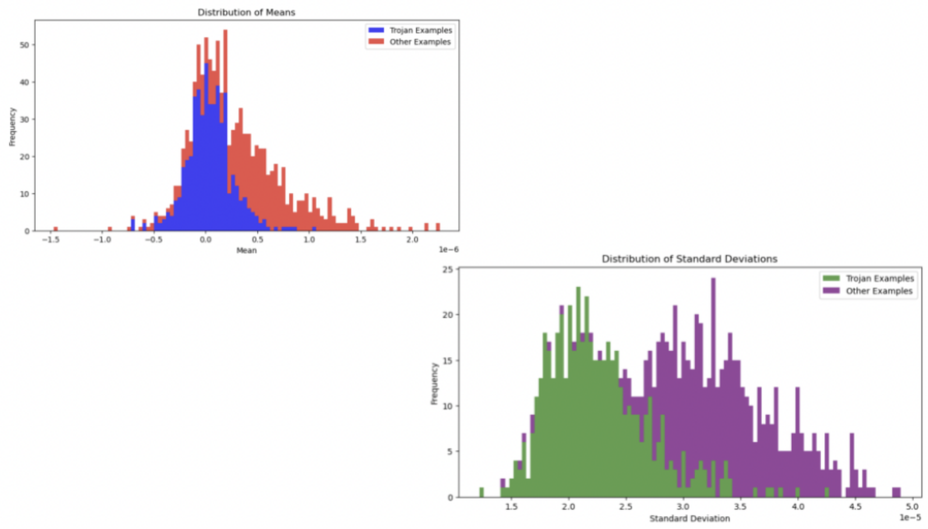

Figure 3. TRAK distributions

Blue Sky Impact

- Analyzing anomalous circuits or using FACADE to prune circuits could significantly improve adversarial robustness.
- Adversarial circuits, seen as probabilistic outliers in geometric transformations, can be highlighted on FACADE distributions and reverse-engineered to enhance robustness through specific weight tuning.
- FACADE requires sufficient training examples to capture meaningful activation flows without supervision.
- At test time, FACADE distributions with probabilistic thresholding can autonomously identify and prevent mechanistic anomalies and adversarial attacks.

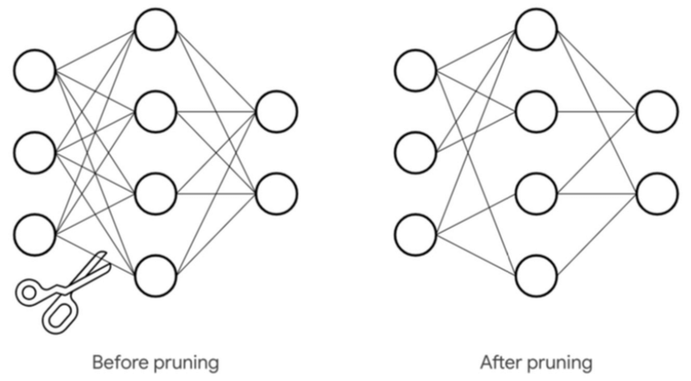

Figure 4. FACADE Circuit Pruning

dhruvpai@stanford.edu