FACADE: A Framework for Adversarial Circuit Anomaly Detection and Evaluation

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

Abstract

We present FACADE, a novel probabilistic and geometric framework designed for unsupervised mechanistic anomaly detection in deep neural networks. Its primary goal is advancing the understanding and mitigation of adversarial attacks. FACADE aims to generate probabilistic distributions over circuits, which provide critical insights to their contribution to changes in the manifold properties of pseudo-classes, or high-dimensional modes in activation space, yielding a powerful tool for uncovering and combating adversarial attacks. Our approach seeks to improve model robustness, enhance scalable model oversight, and demonstrates promising applications in real-world deployment settings.

1. Introduction

In recent years, the field of machine learning has witnessed significant advancements propelled by improvements in learning algorithms and increased access to computational resources. These advancements have led to the development of larger and more capable models, offering remarkable performance on various tasks. However, as models grow in size and complexity, their interpretability diminishes (Gao & Guan, 2023). The sheer scale of modern deep learning models renders traditional methods of interpretation, such as feature importance or attribution, inadequate. The opacity of these models hinders our ability to understand the reasoning behind their predictions and opens the door to potential adversarial attacks.

Moreover, complex models possess numerous axes along which adversarial attacks can be targeted, making them more vulnerable. Adversarial attacks exploit small, carefully crafted perturbations to inputs that can mislead models into making incorrect predictions (Chakraborty et al., 2021). With the increase in model complexity, the space of potential adversarial perturbations expands exponentially, making detection and mitigation increasingly difficult.

Furthermore, as AI capabilities continue to grow, and as the freedoms afforded to such models continue to expand, the implications of a hijacked model pose substantial risks to society, for example through damage to critical systems or infrastructure. Robust mechanisms are needed to detect and prevent the misuse of models, especially as they become more powerful and potentially capable of autonomously evolving their behavior. An unsupervised framework for detecting anomalous behavior in models can serve as a crucial component in safeguarding against such risks.

In this paper, we propose mechanistic anomaly detection via probabilistic models for circuit mechanisms within models as a scalable method for model oversight. Our novel circuit-based framework aims to elucidate complex mechanistic pathways relevant to robustness and does so in an unsupervised fashion without any priors as to the nature of an adversarial attack. Our method develops probabilistic models operating in the geometry of neural activation space that facilitates the detection of deviations from the expected behavior, thereby enabling the identification of anomalous model outputs or adversarial attacks[1].

2. Mechanistic Anomaly Detection

2.1. Activations

Previously, a detailed analysis of flows in activation space proved computationally intensive and opaque. Insofar as neural networks apply a series of nonlinear geometric transformations to high-dimensional data manifolds, the propagation of data points through these transformations is computationally irreducible and largely uninterpretable directly (Cohen et al., 2020). However, understanding the propagation of data points in the high-dimensional activation space has profound implications for the reliability and security of our models, and this understanding may hold the key to investigating adversarial robustness.

[*]Equal contribution [1]Computer Science, Stanford University. Correspondence to: Rylan Schaeffer <rschaef@cs.stanford.edu>.

2nd AdvML Frontiers workshop at 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.
Copyright 2023 by the author(s).

[1]To understand how our proposal relates to emerging directions in adversarial machine learning, see Carranza et al. (2023).

Adversarial examples have been demonstrated to exploit the complex, and often poorly understood, geometry of the decision boundaries within this high-dimensional space (Gebhart et al., 2019). Thus, developing methods capable of elucidating these decision boundary structures and understanding the geometry of high-dimensional modes in activation space would lead to improved adversarial robustness. Prior literature and preliminary experiments demonstrate that neural networks in activation space learn pseudo-classes: intermediate groupings of features learned by the model that resemble high-density modes (Gebhart et al., 2019). An understanding of the distribution, composition, and shape of pseudo-classes within a network offers a lens into the mechanistic behavior of the model.

2.2. Circuit Mechanisms

As defined by Wang et al. (2022), a circuit is a subgraph of a neural network's overall computational graph. Small circuits have recently been investigated for their role in interpretability, in particular identifying circuits corresponding to certain visually meaningful properties of an image, e.g., orientation, curve-detection, color-detection circuits (Olah et al., 2020). Circuits have been demonstrated as a valuable intermediate between single-neuron and whole-model holistic interpretability, as they are well-conserved across mechanistically similar models and provide valuable insights into model behavior for a wide variety of architectures and datasets (Elhage et al., 2021).

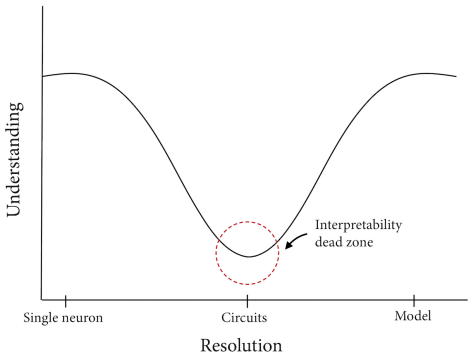

Figure 1. The intermediate scale of interpretability is simultaneously the most poorly understood and the most complex, yet could offer the greatest insights into MAD.

Problematically, circuit interpretability has focused overwhelmingly on uncovering circuits responsible for specific adversarially or visually meaningful features specified *a posteriori* (Olah et al., 2020; Wang et al., 2022; Conmy et al., 2023). Such an approach requires a specification of the features used in an adversarial attack before circuit interpretability is applied, which is rarely the case in a real-world deployment setting (Carlini et al., 2019). Supervised circuit interpretability also extrapolates poorly across mechanistically distinct models. We therefore motivate an unsupervised circuit interpretability approach with the promise of revealing novel insights into mechanistic anomalies, while improving model invariance, computational efficiency, and scalable model oversight.

2.3. Blue Sky MAD Approach

To identify adversarial examples, and their corresponding circuits through mechanistic anomaly detection, we propose a novel probabilistic, geometric framework for creating unsupervised distributions over circuits in a deep neural network titled FACADE: A Framework for Adversarial Circuit Anomaly Detection and Evaluation. Specifically, we envision a four-step approach, where the hyperparameter λ can be interpreted as setting the resolution of circuit analysis.

1. Utilize the probabilistic Dirichlet Process Mixture model (Blei & Jordan, 2006; Kulis & Jordan, 2011) for unsupervised clustering (DP-Means) to identify "pseudoclass" modes in intermediate activation space for a given density threshold λ (Dinari & Freifeld, 2022)
2. Elucidate circuits responsible for pseudoclass formation and propagation through causal discovery and Automatic Circuit DisCovery (ACDC) (Nauta et al., 2019) (Conmy et al., 2023)
3. Determine manifold and kernel density properties of pseudoclass propagation through circuits and in relation to final classes through mean-field theoretic approximation (Cohen et al., 2020)
4. 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

Repeating the above algorithm for a sweep of λ values allows for circuit distribution evaluation across a variety of features and mechanistic pathways. By analyzing anomalous circuits in the distribution or employing FACADE to prune circuits, we envision significant gains in adversarial robustness. Adversarial circuits, identified as probabilistic outliers in geometric transformations, would stand out on FACADE distributions and could easily be reverse-engineered to derive how adversarially-susceptible pseudoclasses can be made more robust by surgical tuning of weights. It is worth noting that FACADE relies on sufficiently many training examples to capture meaningful activation flows in an unsupervised fashion. However, if this condition is met, at test-time FACADE distributions with a simple probabilistic thresholding approach can identify and prevent mechanistic anomalies and adversarial attacks autonomously.

References

Blei, D. M. and Jordan, M. I. Variational inference for dirichlet process mixtures. 2006.

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., and Kurakin, A. On evaluating adversarial robustness, 2019.

Carranza, A., Pai, D., Tandon, A., Schaeffer, R., and Koyejo, S. Deceptive alignment monitoring, 2023.

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., and Mukhopadhyay, D. A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, 6(1):25–45, 2021. doi: https://doi.org/10.1049/cit2.12028. URL https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/cit2.12028.

Cohen, U., Chung, S., Lee, D. D., and Sompolinsky, H. Separability and geometry of object manifolds in deep neural networks. *Nature communications*, 11(1):746, 2020.

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

Dinari, O. and Freifeld, O. Revisiting dp-means: fast scalable algorithms via parallelism and delayed cluster creation. In *Uncertainty in Artificial Intelligence*, pp. 579–588. PMLR, 2022.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Gao, L. and Guan, L. Interpretability of machine learning: Recent advances and future prospects. *IEEE MultiMedia*, pp. 1–12, 2023. doi: 10.1109/MMUL.2023.3272513.

Gebhart, T., Schrater, P., and Hylton, A. Characterizing the shape of activation space in deep neural networks. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 1537–1542. IEEE, 2019.

Kulis, B. and Jordan, M. I. Revisiting k-means: New algorithms via bayesian nonparametrics. *arXiv preprint arXiv:1111.0352*, 2011.

Nauta, M., Bucur, D., and Seifert, C. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):312–340, 2019.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.