"Faithful to What?" On the Limits of Fidelity-Based Explanations

Jackson Eshbaugh
eshbaugj@lafayette.edu
Lafayette College Department of Computer Science

Motivation

- **Common Strategy:** approximate a neural network with a surrogate model, since neural networks are difficult to interpret.
- **Implicit Assumption:** high surrogate fidelity implies that a surrogate captures the NN's predictive structure.
  - *If a surrogate closely matches a model's behavior, it should preserve the structure responsible for its decisions.*
- **Result:** high surrogate fidelity does not guarantee access to predictive structure.

Methodology

- **Setup:** compare a linear baseline, a neural network (f), and a linear surrogate (g).
- **Training:** fit g to mimic f.
- **Metric:** measure fidelity with the **linearity score** $\lambda(f) \coloneqq R^2(f, g)$, which captures how linearly decodable a neural network's behavior is.
- **Interpretation:** high $\lambda(f)$ signals high linear decodability and low $\lambda(f)$ signals significant deviation from linearity.

Note: *$\lambda(f)$ describes the linearity of the learned function $f(x)$ relative to input x, not the linearity of the original data.*

Results

- Testing whether surrogate fidelity remains informative under distribution shift (California Housing dataset).
  - Training: middle 80% of the income distribution
  - Evaluation: IID within the middle 80% & OOD on the 10% lowest and highest income homes.

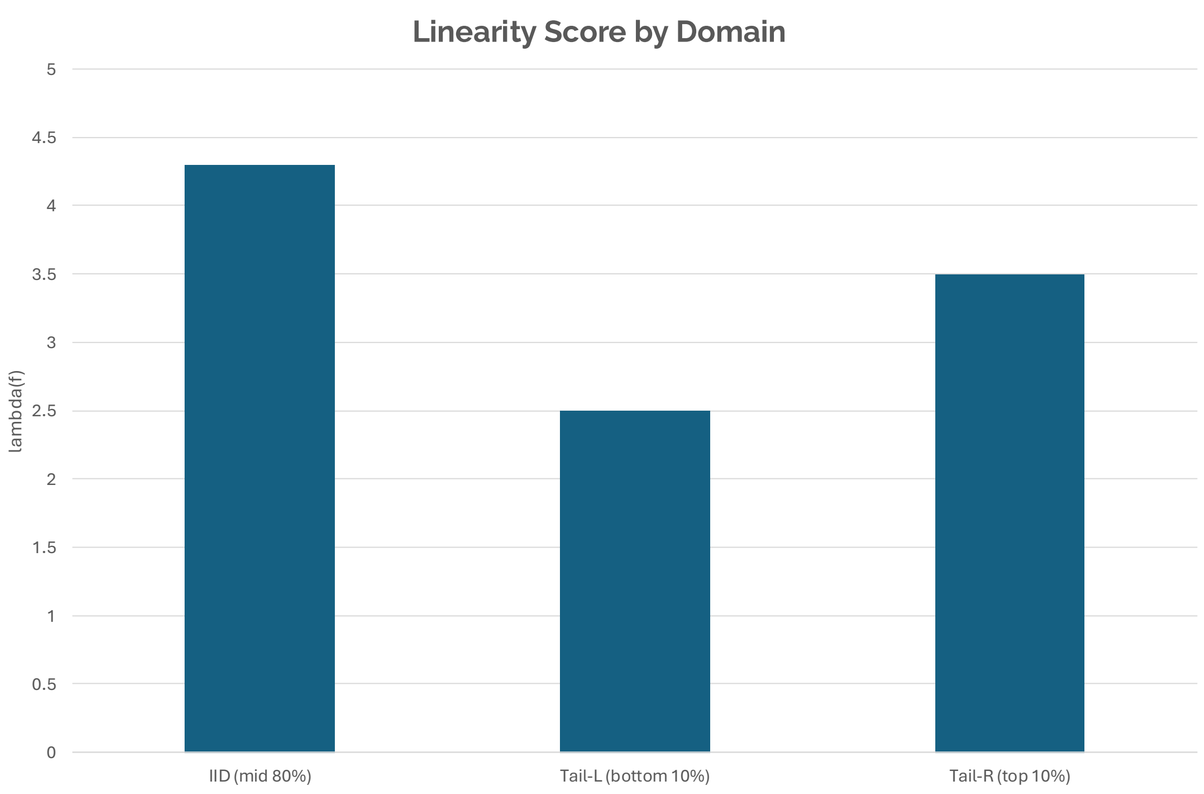
Domain	$\lambda(f)$	$RMSE(f)$	$RMSE(g)$	ΔRMSE
Mid 80%	**0.651**	0.519	0.718	**0.199**
Tail-L *(bottom 10%)*	0.289	0.548	0.676	0.128
Tail-R *(top 10%)*	**0.675**	0.927	0.841	**-0.086**

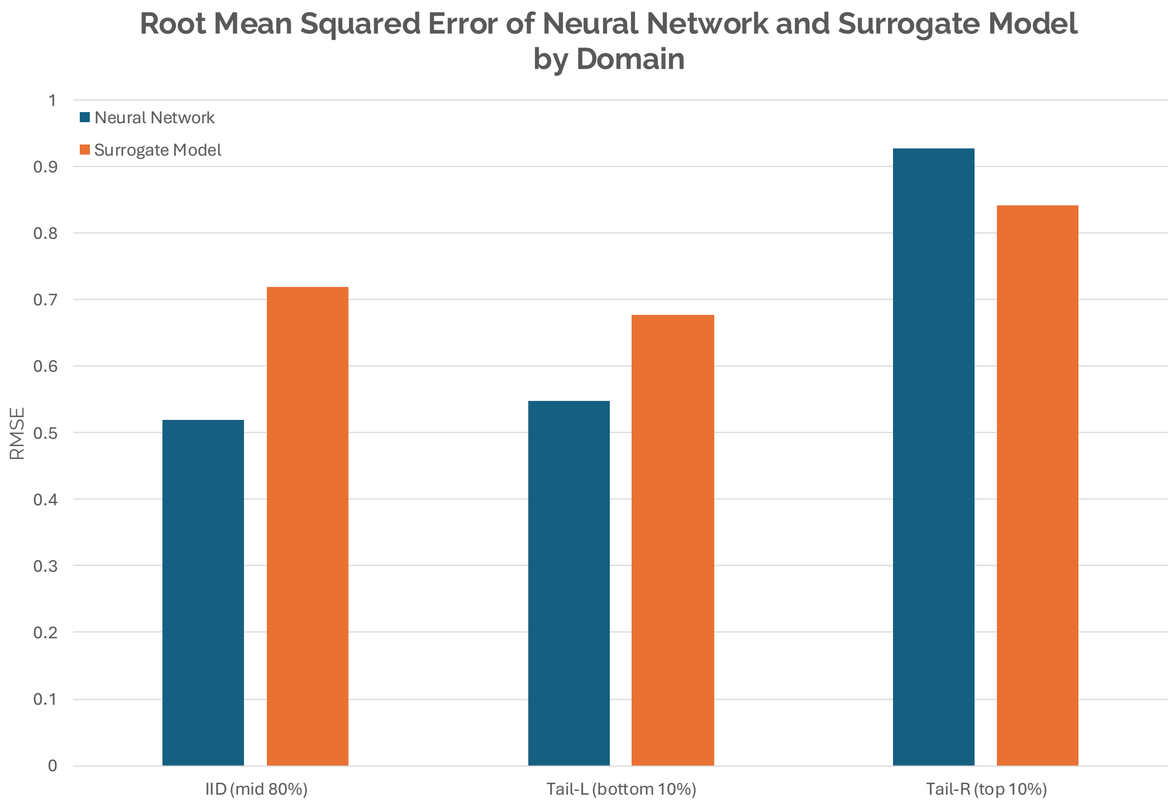
$$\Delta RMSE = RMSE(g) - RMSE(f)$$

- Surrogates perform substantially worse than the network in the IID region and better in the Tail-R (high income) region **despite nearly identical $\lambda(f)$ values**.
  - Nearly identical $\lambda(f)$, **completely different predictive behavior.**
- High-fidelity explanations may fail to reflect (or even invert) comparative predictive performance under shift.

Implications

- There is a fundamental distinction between surrogate fidelity and model behavior.
- Fidelity-based explainations are informative about a model's behavior, but not necessarily about the structure that governs predictive performance.
- Surrogates may faithfully explain a model's behavior while failing to capture the task-relevant structure responsible for the model's predictive advantage.





Acknowledgements

I thank Professors Jorge Silveyra, Jefferey Pfaffmann, and Sofia Serrano for their guidance throughout the editing and submission process. I also thank the Lafayette College Department of Computer Science for their support.

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, November 2018.

[2] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144, San Francisco California USA, August 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939778.

LAFAYETTE

2nd Workshop on Scientific Methods for Understanding Deep Learning
Sunday, April 26, 2026 · Rio de Janeiro, Brazil

ICLR