Evaluating Neuron Explanations: A Unified Framework with Sanity Checks

Tuomas Oikarinen, Ge Yan, Tsui-Wei (Lily) Weng

GitHub: https://github.com/Trustworthy-ML-Lab/Neuron_Eval, **Paper:** https://arxiv.org/abs/2506.05774

ICML International Conference On Machine Learning

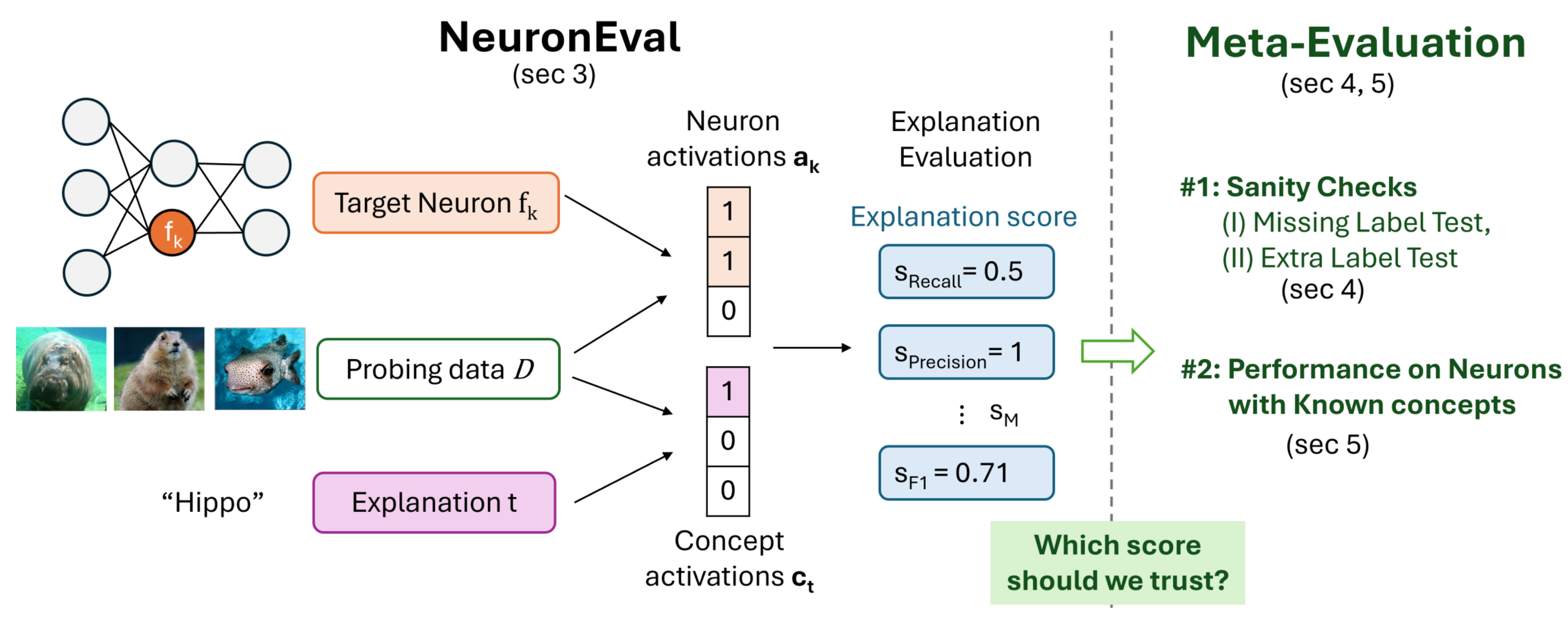
UC San Diego

Trustworthy ML Lab @ UCSD - lilywenglab.github.io

Motivation: Lacking reliable evaluation of the faithfulness of neuron explanations in mechanistic interpretability

- Generating **text explanations** for individual neurons or feature vectors of a neural network is important for mechanistic interpretability.
- For these explanations to be useful, we must understand how **reliable** and **faithful** they are.
- Currently, different papers use very different evaluation methods, often with little justification, making comparison hard.

Our Contribution: A unified evaluation framework + 2 meta evaluation tests to identify reliable evaluation metrics



Contribution #1: A unified evaluation framework for neuron explanations - *NeuronEval*

Our proposed unified framework: NeuronEval

- We unify 20 diverse evaluations from existing work under the same mathematical framework
- Our framework works for any neuron explanations and directions in activations space, including (1) *features in Sparse Autoencoder (SAE)*, (2) *Linear probing*, (3) *Steering vectors*, (4) *Concept bottleneck models*, (5) *TCAV*, and (6) *individual neurons*.
- The evaluations mostly differ on:
  (a) How concept labels $\mathbf{c_t}$ are sourced
  (b) Which metric is used to compare similarity between $\mathbf{a_k}$ and $\mathbf{c_t}$

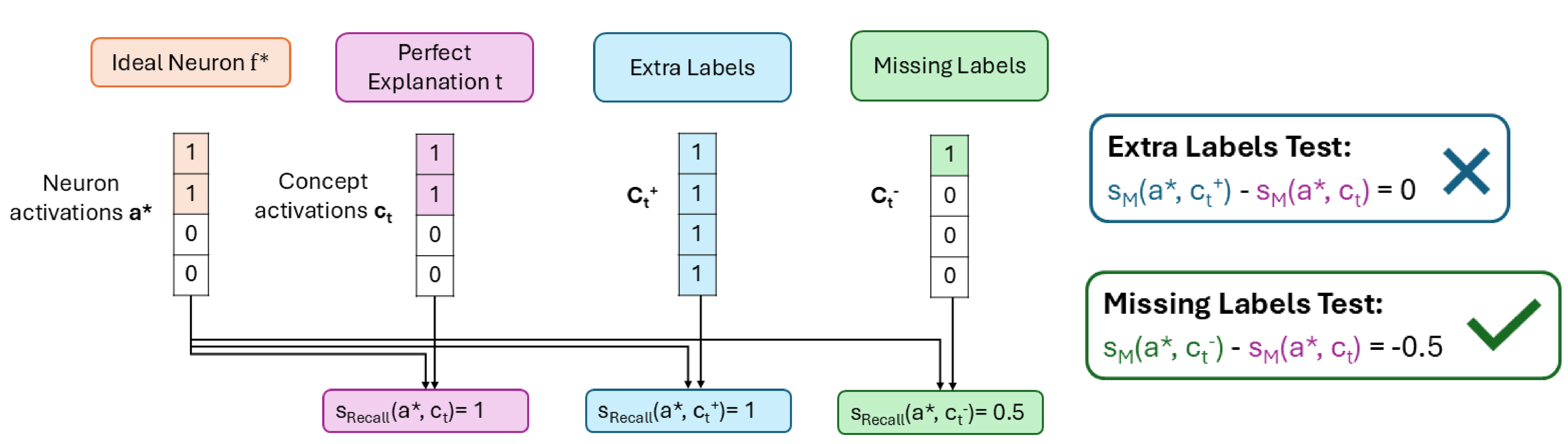
***NeuronEval*: A General Framework to Evaluate Neuron Explanations**

Metric M	Study	Concept Source c_t	Granularity	Domain
Recall	(Zhou et al., 2015)	Crowdsourced	Whole Input	Vision
~Recall	(Bau et al., 2017; Oikarinen & Weng, 2023) (Oikarinen et al., 2023; Bai et al., 2025)	Crowdsourced	Whole Input	Vision
Precision	(Srinivas et al., 2025)	Generative	Whole Input	Vision
F1-score	(Huang et al., 2023)	Generative + Model	Whole Input	Language
	(Gurnee et al., 2023)	Labeled data	Per-token	Language
IoU	(Bau et al., 2017; Mu & Andreas, 2020) (La Rosa et al., 2024)	Labeled data	Per-pixel	Vision
Accuracy	(Koh et al., 2020)	Labeled data	Whole Input	Vision
~AUC	(Zimmermann et al., 2023)	Crowdsourced	Whole Input	Vision
Inverse AUC	(Bykov et al., 2023)	Labeled data	Whole Input	Vision
	(Kopf et al., 2024)	Generative	Whole Input	Vision
Correlation(T&R)	(Bills et al., 2023)	Model	Per-token	Language
Correlation	(Oikarinen & Weng, 2024)	Model	Whole Input	Vision
Spearman Correlation(T&R)	(Bricken et al., 2023; Templeton et al., 2024)	Model	Per-token	Language
~WPMI	(Oikarinen & Weng, 2023)	Model	Whole Input	Vision
MAD	(Kopf et al., 2024)	Generative	Whole Input	Vision
~MAD	(Shaham et al., 2024)	Generative	Whole Input	Vision
	(Singh et al., 2023)	Generative	Whole Input	Language

Contribution #2: Meta Evaluations + Insights on reliable evaluation metrics

Meta-Evaluation 1: Sanity Checks

The idea is to measure whether a metric can differentiate between a **perfect explanation** and (i) Overly generic explanation - Extra Labels Test, or (ii) Overly specific explanation – Missing Labels Test



Meta-Evaluation #1	(I) Missing Labels Test		(II) Extra Labels Test		Pass
	Experimental	Theoretical	Experimental	Theoretical	
Recall	98.66%	100.00%	0.00%	0.00%	×
Precision	45.73%	0.00%	99.81%	100.00%	×
F1-score	93.68%	100.00%	99.82%	100.00%	✓
IoU	93.62%	100.00%	99.81%	100.00%	✓
Accuracy	23.79%	60.00%	70.37%	69.68%	×
Balanced Accuracy	98.65%	100.00%	53.67%	60.00%	×
Inverse Balanced Accuracy	64.18%	60.00%	99.50%	100.00%	×
AUC	94.96%	100.00%	59.18%	60.00%	×
Inverse AUC	52.81%	60.00%	99.99%	100.00%	×
Correlation	99.41%	100.00%	99.92%	100.00%	✓
Correlation (T&R)	87.83%	100.00%	60.26%	43.64%	×
Spearman Correlation	64.05%	67.20%	49.21%	44.08%	×
Spearman Correlation (T&R)	80.04%	100.00%	59.81%	19.68%	×
Cosine	99.45%	100.00%	99.26%	100.00%	✓
WPMI	95.89%	100.00%	58.84%	100.00%	×
MAD	59.81%	60.00%	99.34%	100.00%	×
AUPRC	95.61%	100.00%	99.46%	100.00%	✓
Inverse AUPRC	99.15 %	100.00%	95.58%	89.54%	×

- ➢ Most existing evaluation metrics **fail at least one** of our sanity tests.
  This includes popular metrics such as:
  - *Recall* (Only evaluating highly activating inputs)
  - *Correlation with Top-and-Random* sampling
  - Generative evaluations like *MAD* and *Inverse AUC*
- ➢ This is often caused by poor performance on unbalanced activations

Meta-Evaluation 2:

Comparing evaluation performance on neurons with known concept:

- Metrics that pass sanity checks perform the best

Method	Avg. AUPRC	Avg. Rank
Recall	0.6722	11.30
Precision	0.8039	7.90
F1-score/IoU	0.8140	6.70
Accuracy	0.7215	10.80
Balanced Accuracy	0.7979	7.30
Inverse Balanced Acc.	0.8087	7.10
AUC	0.7652	11.00
Inverse AUC	0.7569	10.90
Correlation	**0.8765**	**1.60**
Correlation(T&R)	0.6606	10.70
Spearman Correlation	0.0853	16.20
Spearman Correlation(T&R)	0.3418	15.40
Cosine	<u>0.8666</u>	<u>2.30</u>
WPMI	0.7999	7.30
MAD	0.6952	8.80
AUPRC	0.8406	3.90
Inverse AUPRC	0.6904	9.30

In conclusions, **we recommend using the following metrics** for *reliable* evaluation:

- **Correlation, AUPRC, F1-score & IoU**

Paper

Code