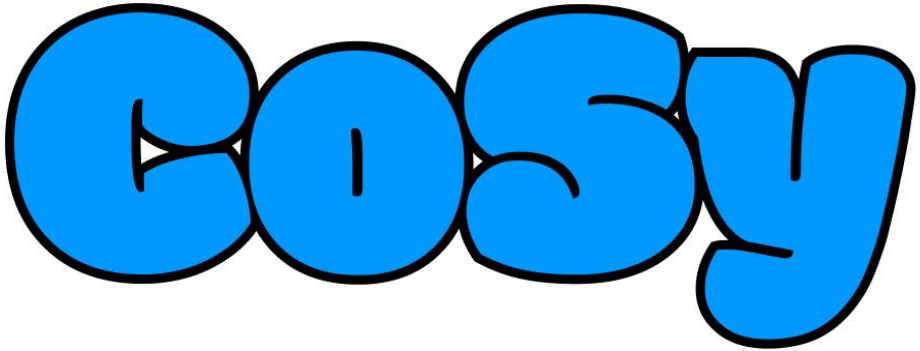
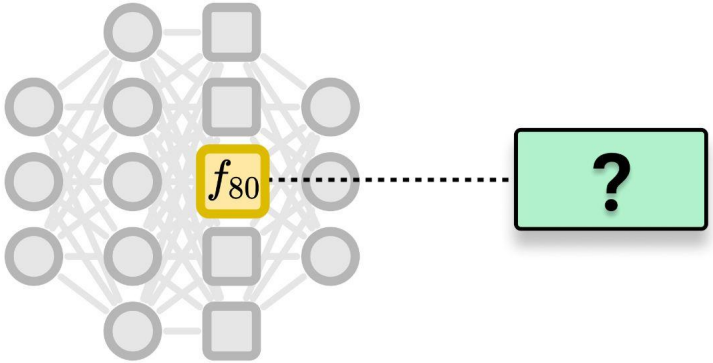


NEURAL INFORMATION PROCESSING SYSTEMS

CoSy: Evaluating Textual Explanations of Neurons

Laura Kopf, Philine Lou Bommer, Anna Hedström, Sebastian Lapuschkin, Marina M.-C. Höhne, Kirill Bykov

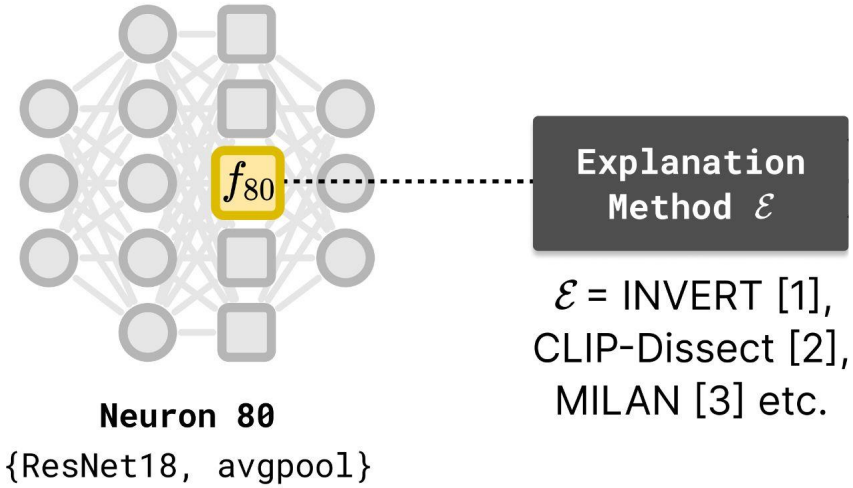
What do Neurons Detect?



Neuron 80

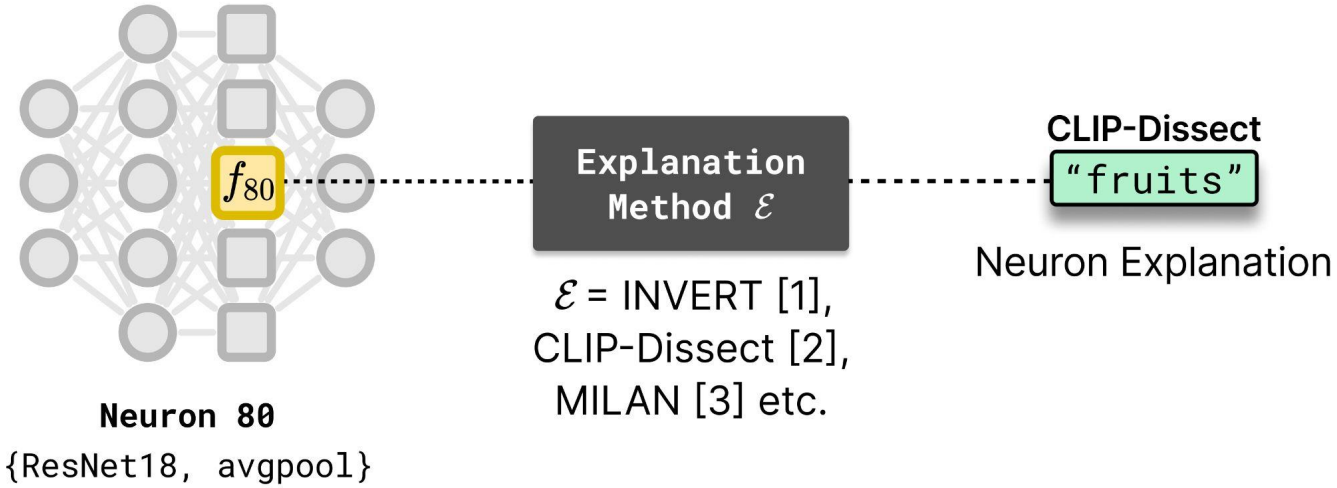
{ResNet18, avgpool}

Neuron Description Methods



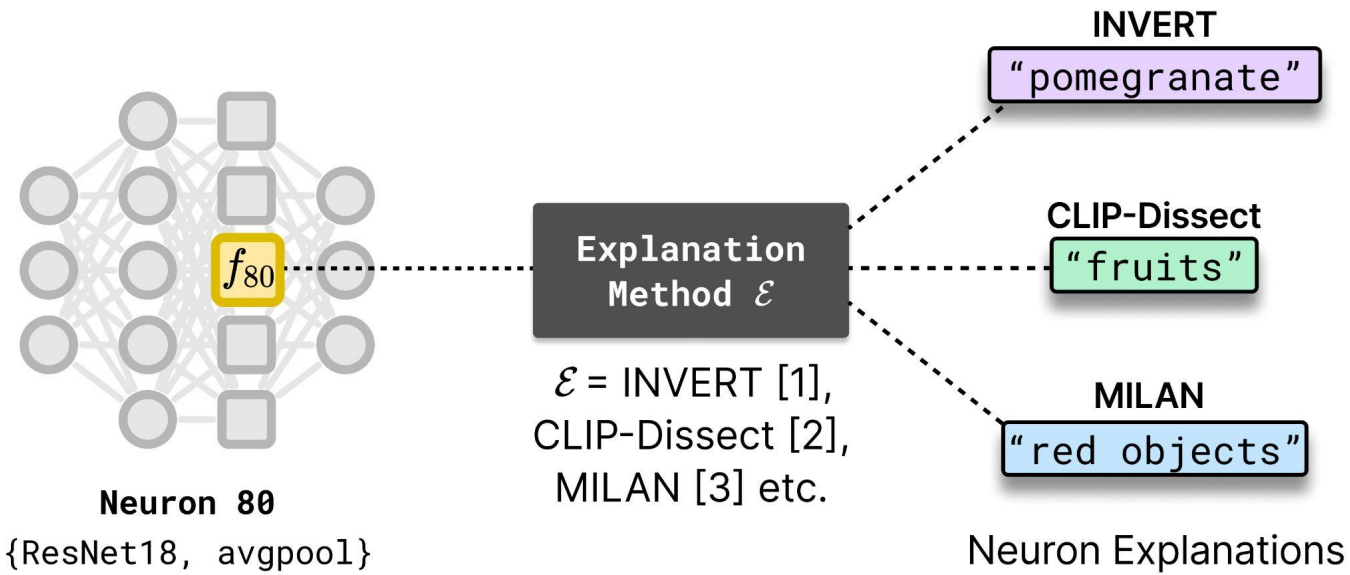
[1] Bykov, Kirill, et al. "Labeling Neural Representations with Inverse Recognition." *Conference on Neural Information Processing Systems*. 2023.
[2] Oikarinen, Tuomas, et al. "CLIP-Dissect: Automatic Description of Neuron Representations in Deep Vision Networks." *International Conference on Learning Representations*. 2022.
[3] Hernandez, Evan, et al. "Natural language descriptions of deep visual features." *International Conference on Learning Representations*. 2021.

Neuron Description Methods



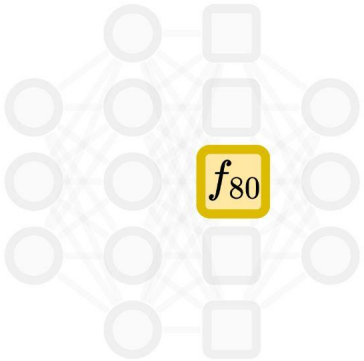
[1] Bykov, Kirill, et al. "Labeling Neural Representations with Inverse Recognition." *Conference on Neural Information Processing Systems*. 2023.
[2] Oikarinen, Tuomas, et al. "CLIP-Dissect: Automatic Description of Neuron Representations in Deep Vision Networks." *International Conference on Learning Representations*. 2022.
[3] Hernandez, Evan, et al. "Natural language descriptions of deep visual features." *International Conference on Learning Representations*. 2021.

Neuron Description Methods



[1] Bykov, Kirill, et al. "Labeling Neural Representations with Inverse Recognition." *Conference on Neural Information Processing Systems*. 2023.
[2] Oikarinen, Tuomas, et al. "CLIP-Dissect: Automatic Description of Neuron Representations in Deep Vision Networks." *International Conference on Learning Representations*. 2022.
[3] Hernandez, Evan, et al. "Natural language descriptions of deep visual features." *International Conference on Learning Representations*. 2021.

Neuron Description Methods



Neuron 80

{ResNet18, avgpool}

What is a good neuron explanation?

INVERT

"pomegranate"

CLIP-Dissect

"fruits"

MILAN

"red objects"

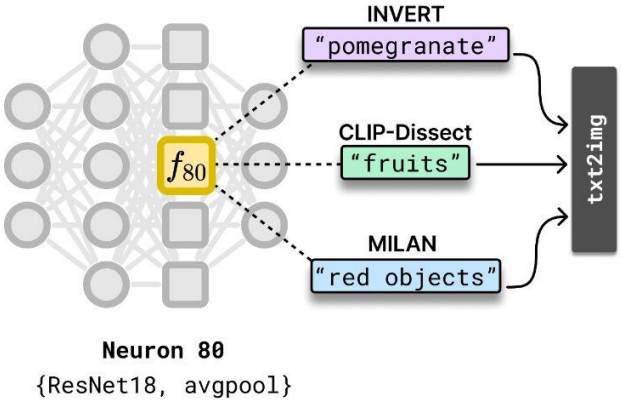
Neuron Explanations

[1] Bykov, Kirill, et al. "Labeling Neural Representations with Inverse Recognition." *Conference on Neural Information Processing Systems*. 2023.
[2] Oikarinen, Tuomas, et al. "CLIP-Dissect: Automatic Description of Neuron Representations in Deep Vision Networks." *International Conference on Learning Representations*. 2022.
[3] Hernandez, Evan, et al. "Natural language descriptions of deep visual features." *International Conference on Learning Representations*. 2021.

CoSy evaluates Textual Explanations of Neurons

Problem: Evaluate Textual Explanations

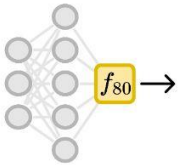
What is a good explanation for the neuron?



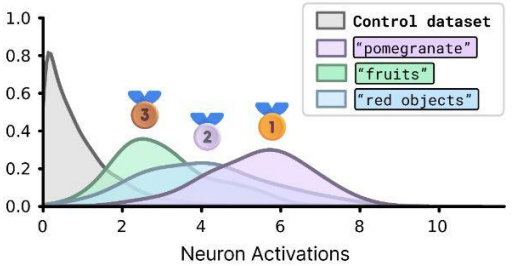
Solution: Estimate Quality with CoSy Framework

Step 1:
Generate Synthetic Data

Step 2:
Collect Neuron Activations



Step 3:
Score Explanations



Scoring Functions

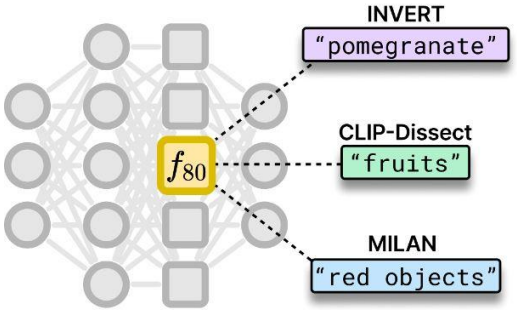
$$\Psi_{\text{AUC}}(\mathbb{A}_0, \mathbb{A}_1) = \frac{\sum_{a \in \mathbb{A}_0} \sum_{b \in \mathbb{A}_1} \mathbf{1}[a < b]}{|\mathbb{A}_0| \cdot |\mathbb{A}_1|}$$

$$\Psi_{\text{MAD}}(\mathbb{A}_0, \mathbb{A}_1) = \frac{\frac{1}{m}\sum_{b \in \mathbb{A}_1} b - \frac{1}{n}\sum_{a \in \mathbb{A}_0} a}{\sqrt{\frac{1}{n-1}\sum_{a \in \mathbb{A}_0}(a - \tilde{a})^2}}$$

CoSy evaluates Textual Explanations of Neurons

Problem: Evaluate Textual Explanations

What is a good explanation for the neuron?

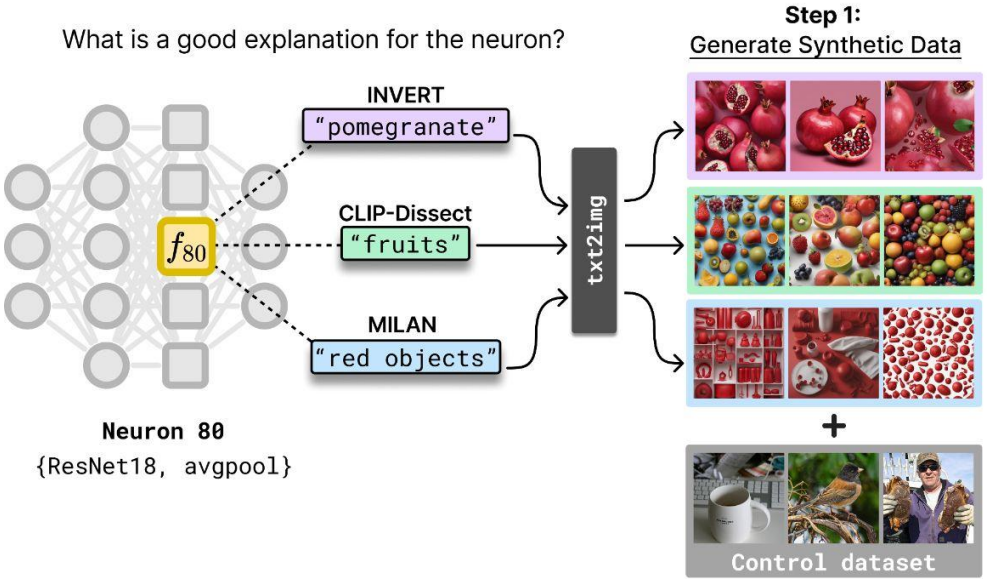


Neuron 80
{ResNet18, avgpool}

CoSy evaluates Textual Explanations of Neurons

Problem: Evaluate Textual Explanations

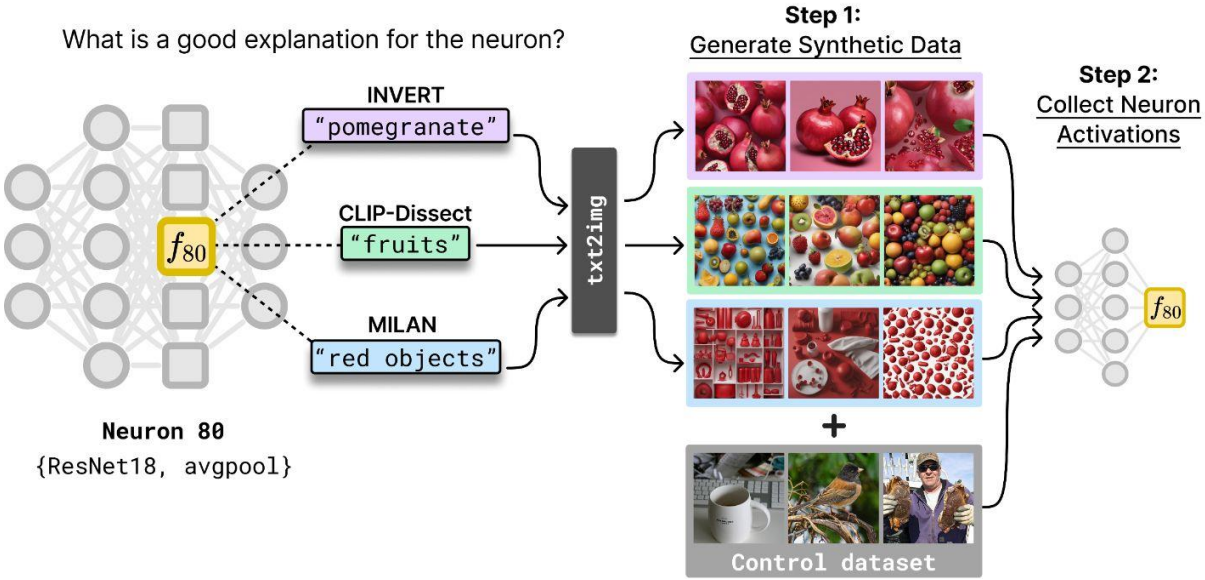
Solution: Estimate Quality with CoSy Framework



CoSy evaluates Textual Explanations of Neurons

Problem: Evaluate Textual Explanations

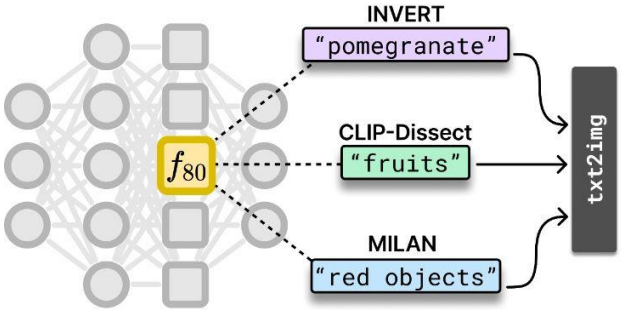
Solution: Estimate Quality with CoSy Framework



CoSy evaluates Textual Explanations of Neurons

Problem: Evaluate Textual Explanations

What is a good explanation for the neuron?

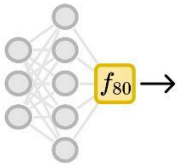


Neuron 80
{ResNet18, avgpool}

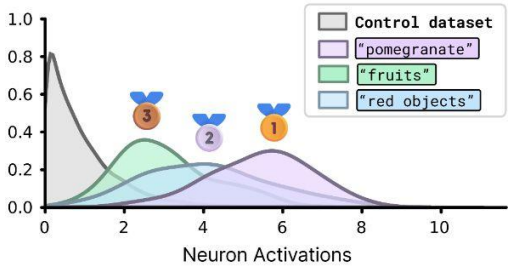
Solution: Estimate Quality with CoSy Framework

Step 1:
Generate Synthetic Data

Step 2:
Collect Neuron Activations



Step 3:
Score Explanations



Scoring Functions

$$\Psi_{\text{AUC}}(\mathbb{A}_0, \mathbb{A}_1) = \frac{\sum_{a \in \mathbb{A}_0} \sum_{b \in \mathbb{A}_1} \mathbf{1}[a < b]}{|\mathbb{A}_0| \cdot |\mathbb{A}_1|}$$

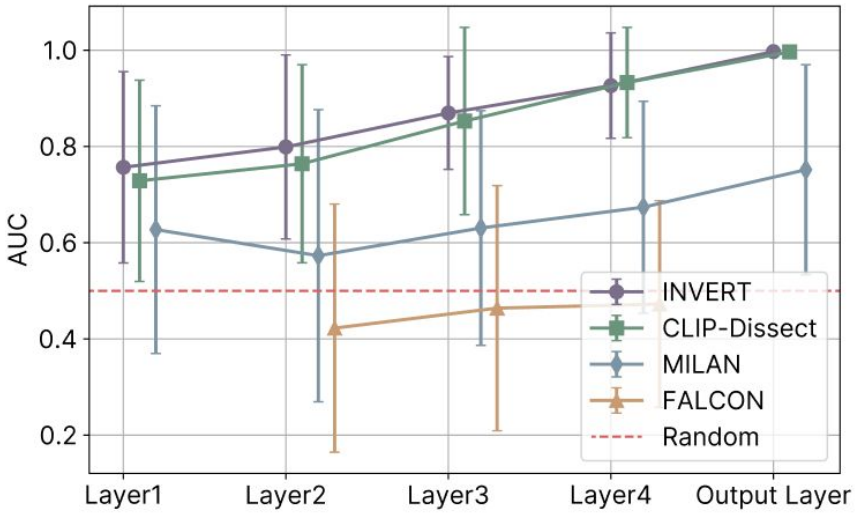
$$\Psi_{\text{MAD}}(\mathbb{A}_0, \mathbb{A}_1) = \frac{\frac{1}{m}\sum_{b \in \mathbb{A}_1} b - \frac{1}{n}\sum_{a \in \mathbb{A}_0} a}{\sqrt{\frac{1}{n-1}\sum_{a \in \mathbb{A}_0}(a - \bar{a})^2}}$$

CoSy gives Quantitative Performance Comparison

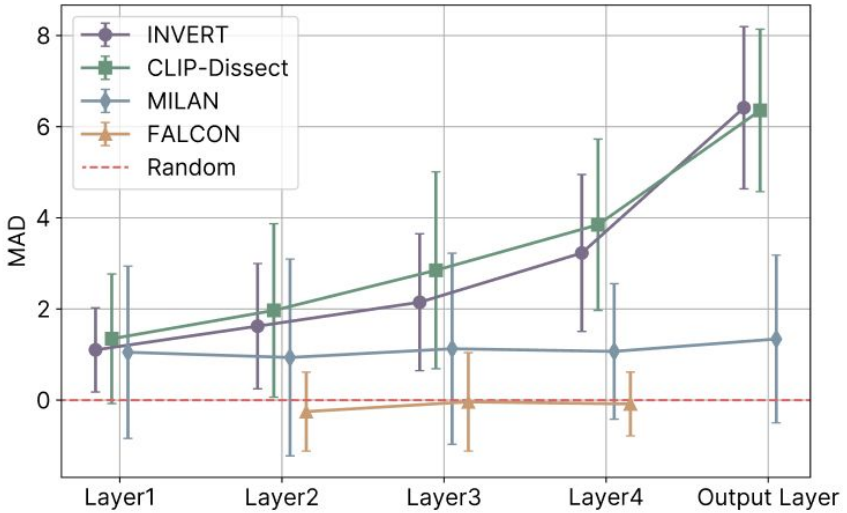
Dataset	Model	Layer	Method	AUC (↑)	MAD (↑)
ImageNet	ResNet18	Avgpool	MILAN	0.61±0.23	0.69±1.35
			CLIP-Dissect	**0.93±0.11**	**3.85±1.88**
			INVERT	**0.93±0.11**	3.23±1.72
	ResNet50	Avgpool	MILAN	0.44±0.23	-0.08±0.72
			CLIP-Dissect	0.95±0.08	**4.98±2.57**
			INVERT	**0.96±0.06**	4.62±2.26
	ViT-B/16	Features	MILAN	0.53±0.19	0.12±0.76
			CLIP-Dissect	0.78±0.19	1.29±1.01
			INVERT	**0.89±0.17**	**1.67±0.82**
	DINO ViT-S/8	Layer 11	MILAN	0.59±0.21	0.37±0.91
			CLIP-Dissect	**0.95±0.08**	**4.59±2.62**
			INVERT	0.73±0.27	2.70±3.48
Places365	DenseNet161	Features	MILAN	0.56±0.28	0.44±1.30
			CLIP-Dissect	0.82±0.21	**2.52±2.33**
			INVERT	**0.85±0.16**	2.21±1.95
	ResNet50	Avgpool	MILAN	0.65±0.28	1.11±1.67
			CLIP-Dissect	0.92±0.11	**3.73±2.39**
			INVERT	**0.94±0.08**	3.54±1.99

Now different neuron description methods can be ranked against each other.

CoSy provides new Insights into Explanation Methods



a) AUC across ResNet18 Layers



b) MAD across ResNet18 Layers

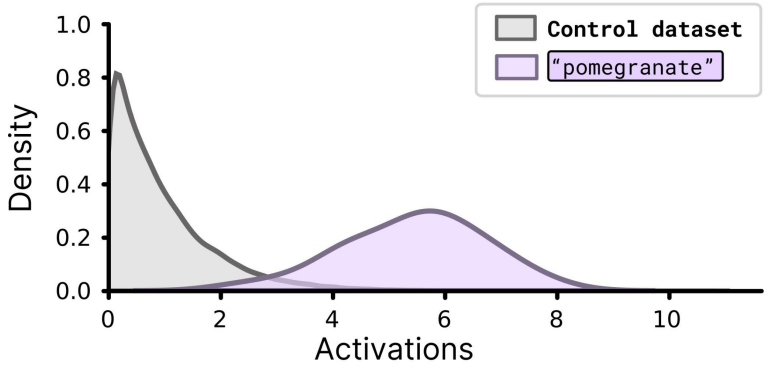
New insight: Explanation methods struggle to explain lower layer neurons.

Conclusion

Conclusion

CoSy

- **evaluates** textual **explanations of neurons**



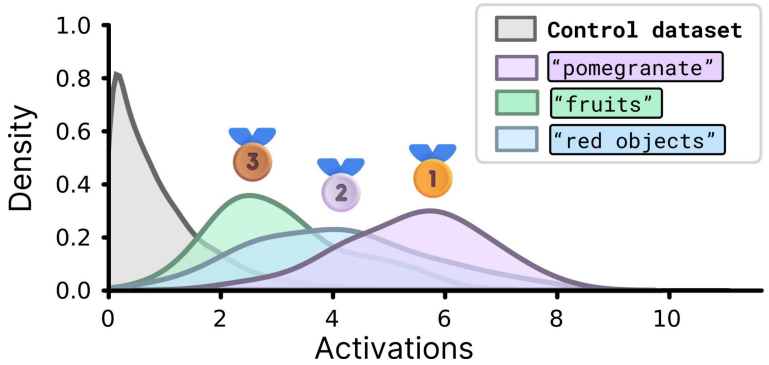
Ψ_{AUC}(`Control dataset`, `"pomegranate"`) = 0.99

Evaluation Score

Conclusion

CoSy

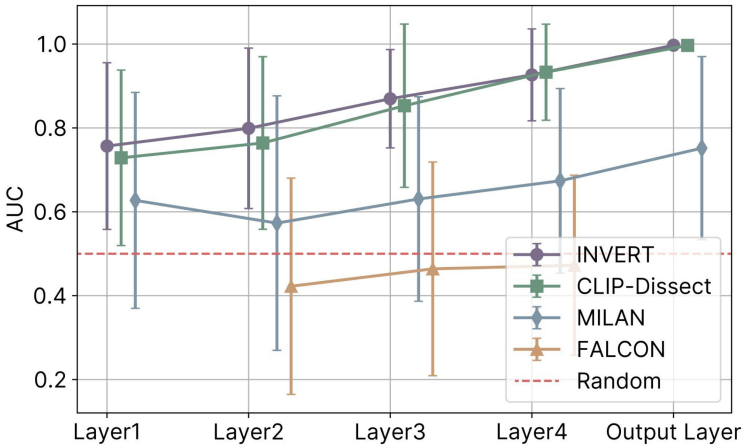
- **evaluates** textual **explanations of neurons**
- enables **comparison** of different textual **explanation methods**



Conclusion

CoSy

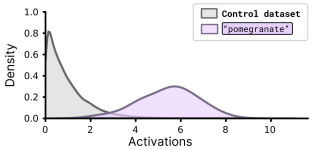
- **evaluates** textual **explanations of neurons**
- enables **comparison** of different textual **explanation methods**
- provides **new insights** into these **explanation methods**



Conclusion

CoSy

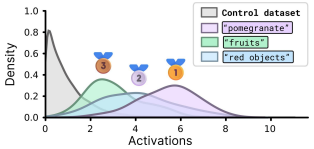
- **evaluates** textual **explanations of neurons**



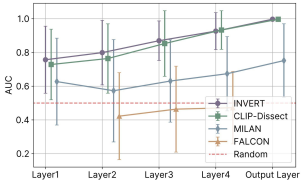
$\Psi_{\text{AUC}}(\texttt{Control dataset}, \texttt{"pomegranate"}) = 0.99$

Evaluation Score

- enables **comparison** of different textual **explanation methods**



- provides **new insights** into these **explanation methods**



Get CoSy !

Paper

https://arxiv.org/abs/2405.20331

Code

https://github.com/lkopf/cosy