Turning Polysemantic Neurons Into Pure Features by Identifying Relevant Circuits

PURE
PURIFYING REPRESENTATIONS

Fraunhofer HHI
TECHNISCHE UNIVERSITÄT BERLIN
BIFOLD

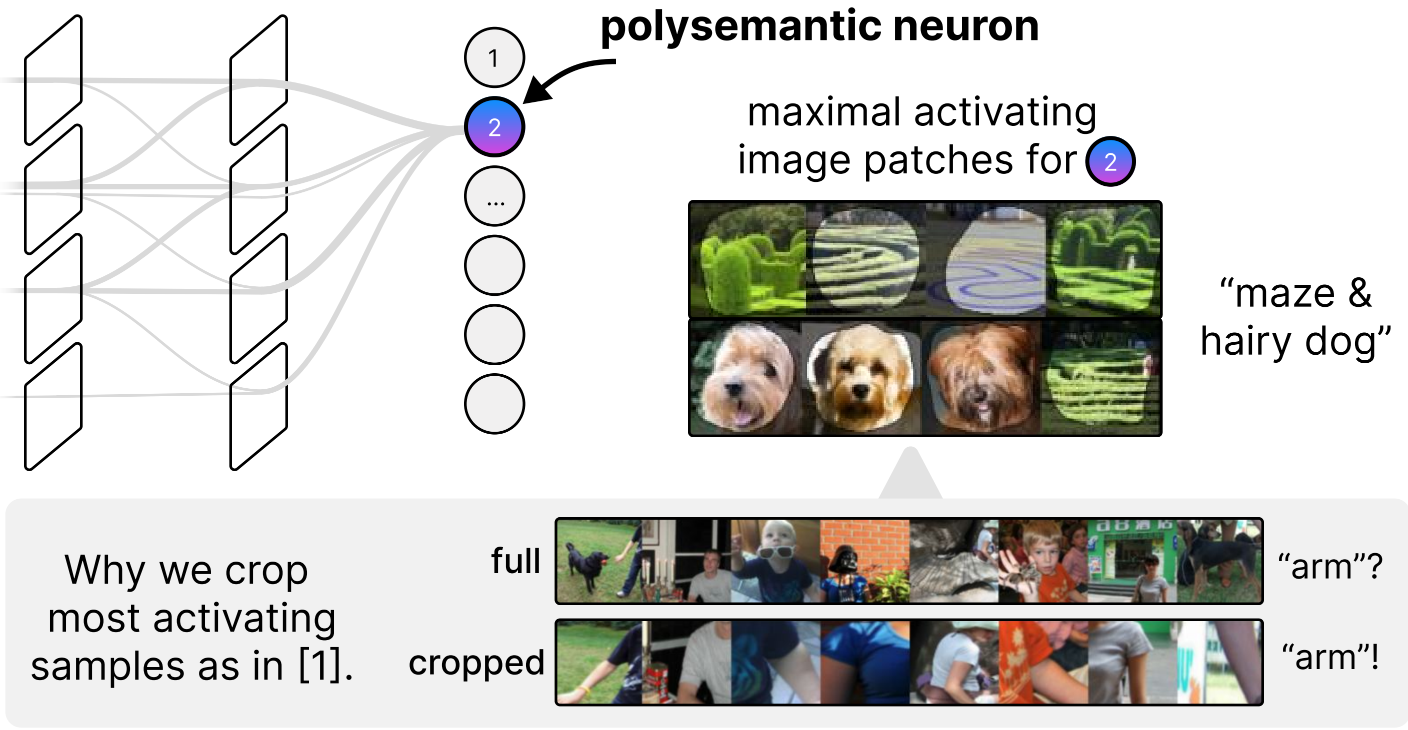
Maximilian Dreyer, Erblina Purelku, Johanna Vielhaben, Wojciech Samek, Sebastian Lapuschkin

CVPR
JUNE 17-21, 2024
SEATTLE, WA

What do neurons encode?

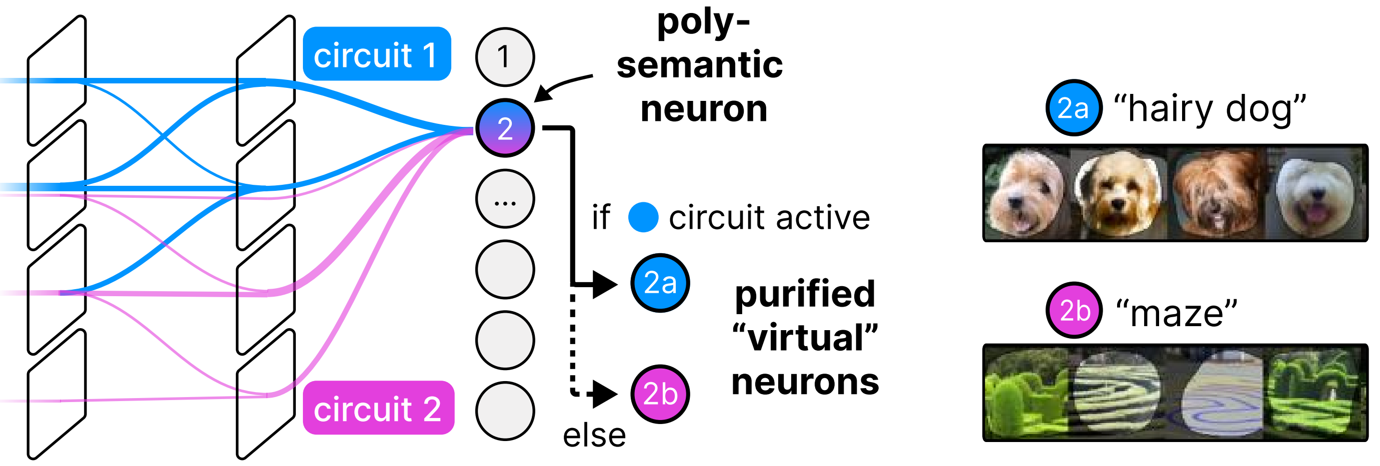
Studying neurons can be difficult due to polysemanticity, redundancies, etc.

with PURE, we tackle **polysemanticity:**



Idea

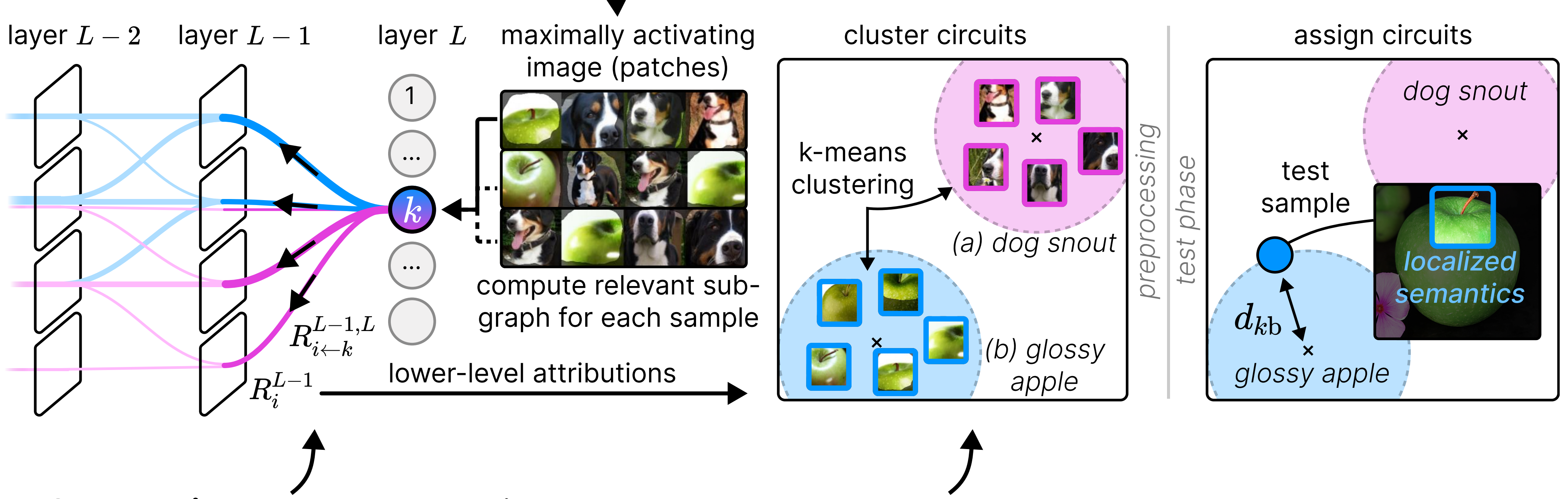
Each pure feature corresponds to a specific sub-graph.



When we know which sub-graph is active, we also know which feature is present.

PURE: Purifying Representations

1. Find **most activating samples** for a polysemantic neuron.

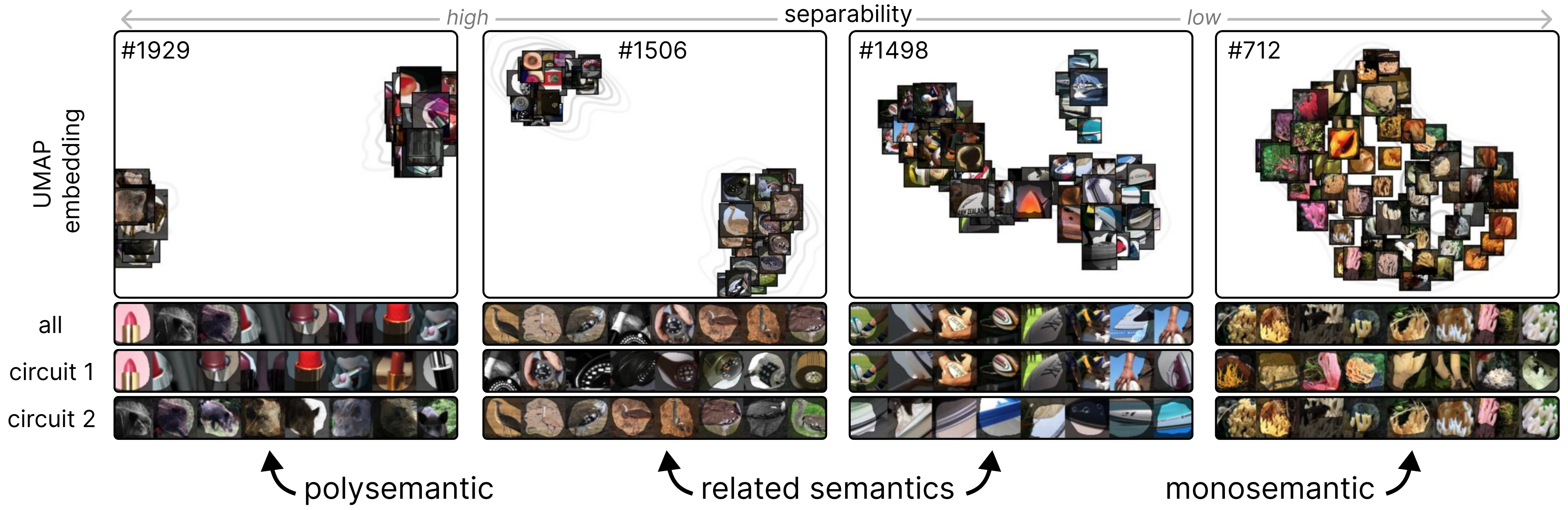


2. **Explain neuron** and attribute lower-level neurons.

3. Cluster attributions and **find circuits**.

Qualitative Experiments

Apply PURE and sort neurons according to the effect of disentanglement.



Quantitative Experiments

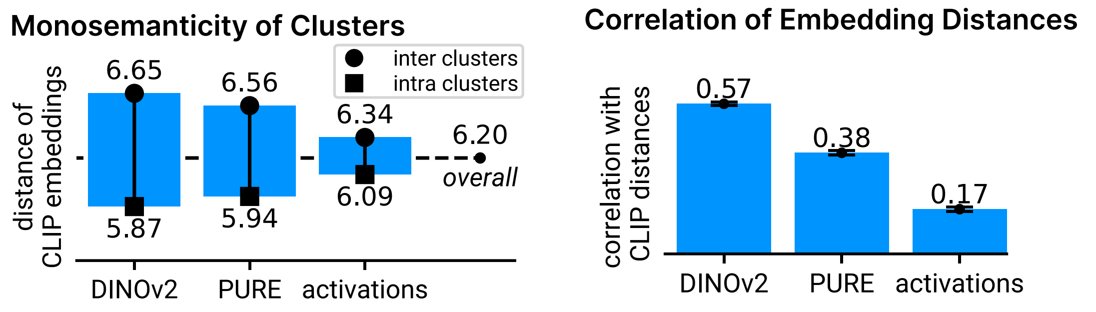
We use foundation model embeddings (e.g., CLIP [2] and DINOv2 [3]) to measure monosemanticity before and after purification of ResNet models.

Idea: embedding distances for maximally activating patches should decrease.

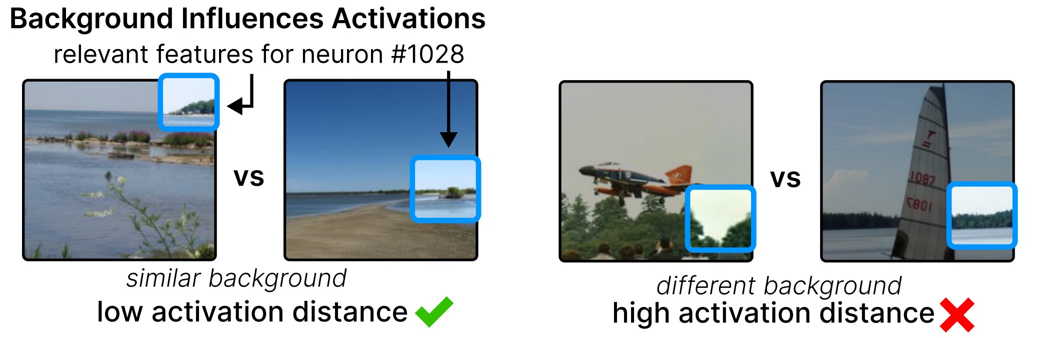
$$\mathbf{D}^k_{ij} = \sqrt{\left(\mathbf{e}^{\mathrm{CLIP}}_i - \mathbf{e}^{\mathrm{CLIP}}_j\right)^2}$$

with embedding $\mathbf{e}^{\mathrm{CLIP}}_i$ of max. act. image patch i of neuron k

PURE achieves better disentanglement compared to activation-based clustering.



PURE is more neuron-specific, as activations take into account all present features.



Outlook & Conclusion

→ Application to language, e.g., Large Language Models.
→ Studying the benefits of PURE for concept-based explanations, probing, and unlearning.
→ Performing an ablation study & user study for validation.

References

[1] Achtibat, Reduan, et al. **"From attribution maps to human-understandable explanations through concept relevance propagation."** Nature Machine Intelligence 5.9 (2023): 1006-1019.
[2] Radford, Alec, et al. **"Learning transferable visual models from natural language supervision."** International conference on machine learning. PMLR, 2021.
[3] Oquab, Maxime, et al. **"DINOv2: Learning Robust Visual Features without Supervision."** Transactions on Machine Learning Research (2023).

PAPER CODE