The Tangled Web they Weave: Exploring Neural Networks with Directed Topology

Michael Bleher, Institute for Mathematics, Heidelberg University.

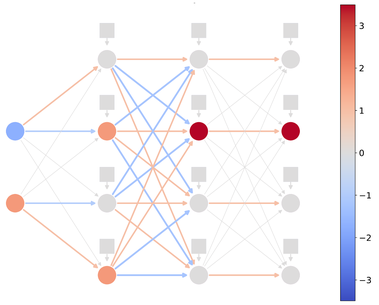
Artificial Neural Networks

Feedforward Multi-Layer Perceptrons

$$f : \mathbb{R}^n \to \mathbb{R}^{k_1} \to \ldots \to \mathbb{R}^{k_L} \to \mathbb{R}^m$$

$$a^{(0)} = x \in \mathbb{R}^n$$
$$a^{(\ell)} = \sigma(W^{(\ell)} a^{(\ell-1)} + b^{(\ell)}) \in \mathbb{R}^{k_\ell}$$

σ activation function, $W^{(\ell)}$ weight matrix, $b^{(\ell)}$ bias



MLPs are structural causal models
$\rightsquigarrow$ causal reasoning (Pearl [1]).

Internal Neuron Activations

$$a(x) := (a^{(1)}, \ldots, a^{(L)}) \in M \subset \mathbb{R}^N$$

neural manifold

Mechanistic Interpretability

Neural networks learn to
- recognize **features**, and
- perform **internal computations**.

$\longrightarrow$ Features must be represented in the neuron activation $a(x) \in M$.

Feature Representations

- **Distributed**: Activation patterns span multiple neurons
- **Superposition**: Feature-specific patterns overlap
- **Polysemanticity**: Individual neurons respond to multiple unrelated concepts

Task *Disentangle superpositions of distributed features on polysemantic neurons.*

Why Directed Topology?

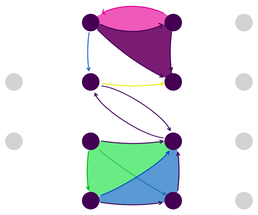
State of the art investigates *coordinate axes* of the feature space (e.g. ICA, NMF, SAE[2]), or detects logical circuits (e.g. CMA[3], DAS[4], ACDC[5]).

Why (directed) topology?

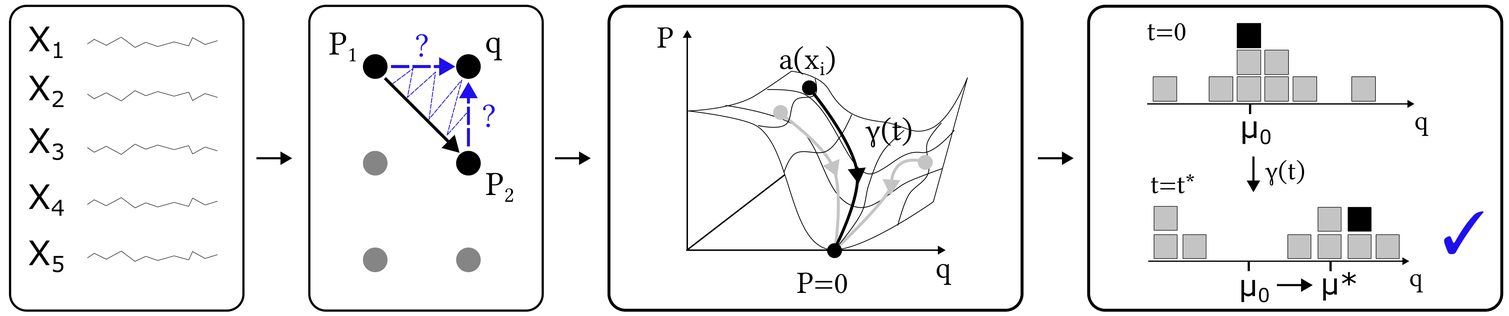
"What fires together, wires together"
- distributed features $\rightarrow$ 'semantic neighbourhoods'
- polysemantic neurons $\rightarrow$ intersections
- superposition $\rightarrow$ simplicial complexes
- causal structure $\rightarrow$ directionality

Idea

(Filtered) **directed simplicial complex** as feature representations and interactions.



Directed Simplicial Complexes from Neuron Activations



Fix Context → **Candidate Simplex** → **Coherent Counterfactual Ablation** → **Check for Collective Response**

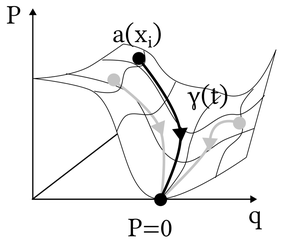
Coherent Counterfactual Ablation

- *Coherent*: keep structural equations intact (stay on neural manifold M).
- *Counterfactual*: 'what if $a|_P$ was different?'
- *Ablation*: different as in $a|_P \to 0$.

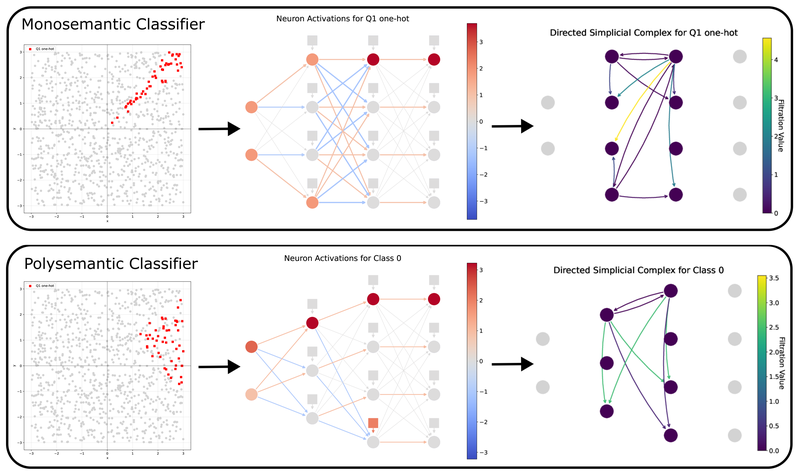
Construction

1. Pick $x_i \in X$ with activation $a(x_i) \in M \subset \mathbb{R}^N$
2. Find $v \in T_{a(x_i)}M$ s.t. $a(x_i)|_P$ decreases.
3. Integrate to path $\gamma(t)$ (flow equation $\dot{\gamma}(t) = v$, $\gamma(0) = a(x_i)$)

The pullback $\gamma(t)|_q$ quantifies the causal influence of the neurons P on the probe neuron q at time t.



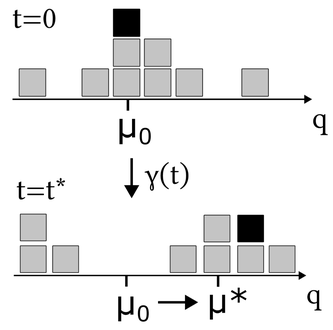
Experiments: Monosemantic vs Polysemantic Quadrant Classifiers



Collective Response

For fixed probe neuron q

1. Compute change of activations
$$\Delta q_i(t) = (\gamma_i(t) - \gamma_i(0))|_q$$
2. Perform Wilcoxon signed-rank test
   - Null Hypothesis: median$\{\Delta q_i(t)\} = 0$
   - Reject if $p < \alpha = 0.01$
3. filtration $t^* = \min\{t : p(t) < \alpha\}$
   (earlier $t^* \implies$ more immediate influence)



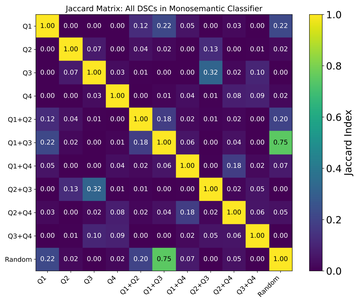
Conclusions & Future Work

Take-away

- *feature-specific*,
- *compositional*, and
- *higher-order* 'causal units'

Next steps

- (directed) topological invariants
- dictionary learning
- algebra of feature compositions?



References

[1] Judea Pearl et al. "Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference". In: *Synthese-Dordrecht* 104 (1995).

[2] Nelson Elhage et al. "Toy Models of Superposition". In: *Transformer Circuits Thread* (2022).

[3] Jesse Vig et al. "Investigating Gender Bias in Language Models Using Causal Mediation Analysis". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 12388–12401.

[4] Atticus Geiger et al. "Finding Alignments Between Interpretable Causal Variables and Distributed Neural Representations". In: *Conference on Causal Learning and Reasoning* (2024).

[5] Arthur Conmy et al. "Towards Automated Circuit Discovery for Mechanistic Interpretability". In: *Advances in Neural Information Processing Systems* 36 (2023).