

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

$$f : \mathbb{R}^n \rightarrow \mathbb{R}^{k_1} \rightarrow \dots \rightarrow \mathbb{R}^{k_L} \rightarrow \mathbb{R}^m$$

$$a^{(0)} = x \in \mathbb{R}^n$$

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

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

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

Internal Neuron Activations

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

neural manifold

Directed Simplicial Complexes from Neuron Activations

Fix Context: X_1, X_2, X_3, X_4, X_5

Candidate Simplex: P_1, P_2 , probe neuron q

Coherent Counterfactual Ablation: Path $\gamma(t)$ from $P=0$ to $a(x_i)$ on the manifold M .

Check for Collective Response: Activation change $\mu_0 \rightarrow \mu^*$ over time $t=0$ to $t=t^*$.

Mechanistic Interpretability

Neural networks learn to

- recognize **features**, and
- perform **internal computations**.

→ 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
- **Polysemy**: Individual neurons respond to multiple unrelated concepts

Task Disentangle superpositions of distributed features on polysemantic neurons.

Coherent Counterfactual Ablation

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

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 .

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: $\text{median}\{\Delta q_i(t)\} = 0$
 - Reject if $p < \alpha = 0.01$
3. filtration $t^* = \min\{t : p(t) < \alpha\}$
 (earlier t^* \Rightarrow more immediate influence)

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.

Experiments: Monosemantic vs Polysemantic Quadrant Classifiers

Monosemantic Classifier: Neuron Activations for Q1 one-hot, Directed Simplicial Complex for Q1 one-hot

Polysemantic Classifier: Neuron Activations for Class 0, Directed Simplicial Complex for Class 0

Jaccard Matrix: All DSCs in Monosemantic Classifier

	Q1	Q2	Q3	Q4	Q1+Q2	Q1+Q3	Q1+Q4	Q2+Q3	Q2+Q4	Q3+Q4	Random
Q1	1.00	0.00	0.00	0.00	0.12	0.22	0.05	0.00	0.03	0.00	0.22
Q2	0.00	1.00	0.07	0.00	0.04	0.02	0.00	0.13	0.00	0.01	0.02
Q3	0.00	0.07	1.00	0.03	0.01	0.00	0.00	0.32	0.02	0.10	0.00
Q4	0.00	0.00	0.03	1.00	0.00	0.01	0.04	0.01	0.08	0.09	0.02
Q1+Q2	0.12	0.04	0.01	0.00	1.00	0.18	0.02	0.01	0.02	0.00	0.20
Q1+Q3	0.22	0.02	0.00	0.01	0.18	1.00	0.06	0.00	0.04	0.00	0.75
Q1+Q4	0.05	0.00	0.00	0.04	0.02	0.06	1.00	0.00	0.18	0.02	0.07
Q2+Q3	0.00	0.13	0.32	0.01	0.01	0.00	0.00	1.00	0.02	0.05	0.00
Q2+Q4	0.03	0.00	0.02	0.08	0.02	0.04	0.18	0.02	1.00	0.06	0.05
Q3+Q4	0.00	0.01	0.10	0.20	0.00	0.00	0.02	0.05	0.06	1.00	0.00
Random	0.22	0.00	0.02	0.20	0.75	0.07	0.00	0.05	0.00	0.00	1.00

Legend: Jaccard Index (0.0 to 1.0)

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