Granular Concept Circuits: Toward a Fine-Grained Circuit Discovery for Concept Representations

Dahee Kwon[1*] Sehyun Lee[1*] Jeasik Choi[1,2]

[1]KAIST (Korea Advanced Institute of Science and Technology) [2]INEEJI

[*]Equal contribution.

ICCV HONOLULU HAWAII OCT 19-23, 2025

KAIST AI Kim Jaechul Graduate School

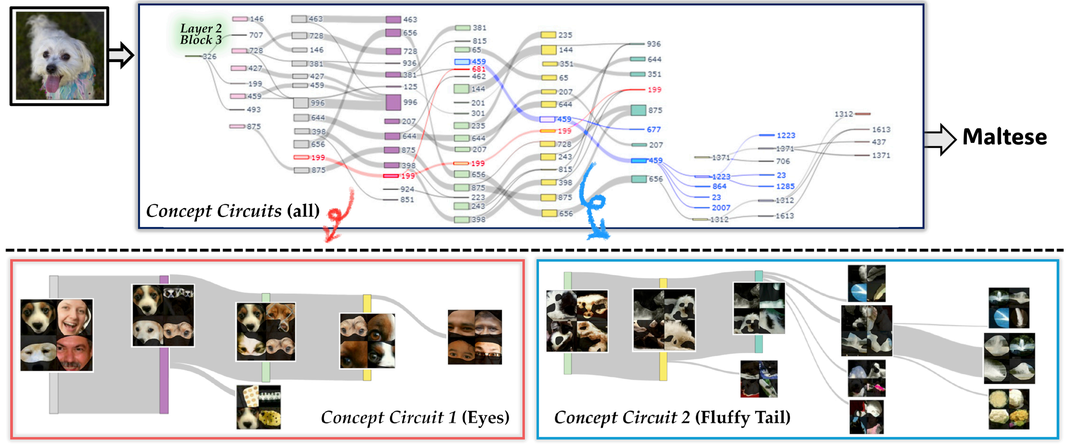
Project Page

Code

Introduction

The Problem Pinpointing visual concepts in DNNs is hard due todistributed neural representations. Existing methods often miss fine-grained concepts, focusing instead on single neurons or single, class-level circuits.

Our Solution We introduce Granular Concept Circuit (GCC), an effective circuit discovery method that automatically identifies multiple, concept-specific circuits for a given input query.



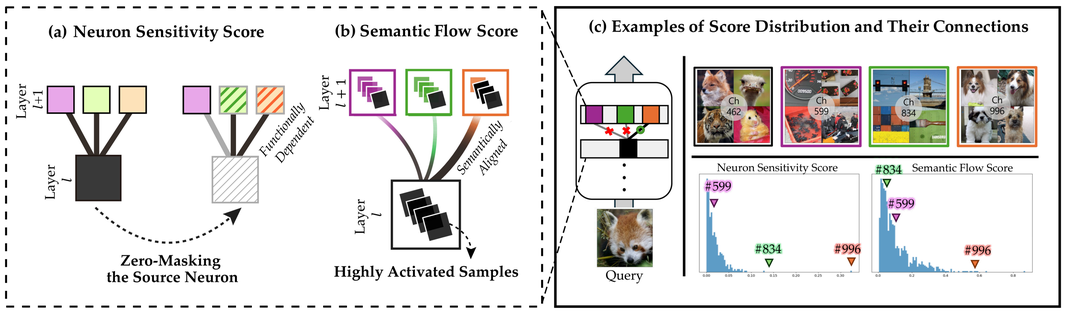
Methodology Details

Neuron Sensitivity Score (S_{NS}) quantifies functional dependency; measures the change in the target's activation when the source neuron is zero-masked.

$$\tilde{S}_{NS,c} = \max\left(0, f^{l+1}(a_c^l) - f^{l+1}(\hat{a}_c^l)\right) \Rightarrow S_{NS} = \frac{\tilde{S}_{NS}}{\sum \tilde{S}_{NS}}$$

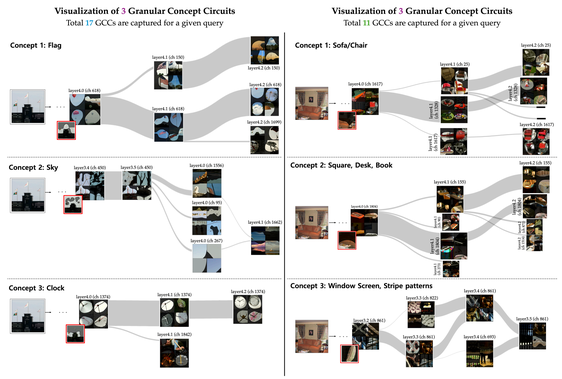
Semantic Flow Score (S_{SF}) ensures semantic alignment; measures the overlap between the set of highly activated samples for the source and target neurons.

$$S_{SF} = \frac{|\mathcal{S}_{\text{src}} \cap \mathcal{S}_{\text{tgt}}|}{|\mathcal{S}_{\text{src}}|}$$

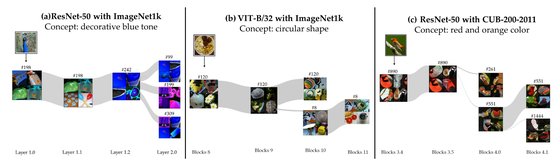


Qualitative Results

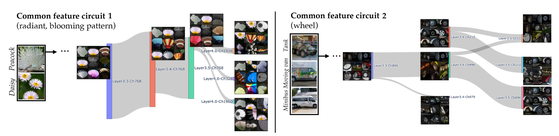
Granular Circuits: Disentangling Multiple Concepts per Query



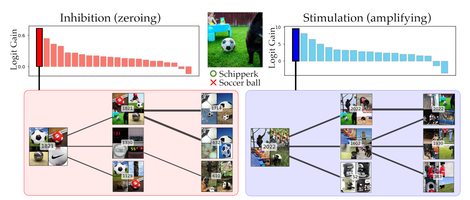
Versatility: GCCs Across Models and Datasets



Finding Common Concepts Across Classes



Auditing Errors via Circuit Manipulation The figure audits the model's misclassification of a 'schipperke'image as 'soccer ball'by illustrating the circuit's influence on the true class logit when its neurons are inhibited or stimulated.



Quantitative Results

Average Logit Drop After Neuron Ablation

	R50	R101	V19	M3	Avg
Original	17.17	17.46	20.94	17.34	-
Random	15.66	13.80	19.03	15.01	(▽ 2.35)
Ours	**6.41**	**6.18**	**12.93**	**12.95**	(▽ **8.60**)
Ours^C	16.12	14.58	19.93	15.88	(▽ 1.74)

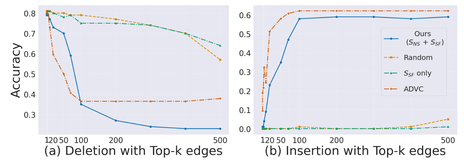
Table 1. Comparison on average logit drop after ablating neurons with random selection (**Random**), found circuits (**Ours**), and non-relevant neurons (**Ours**^{C}).

1. **Faithfulness**: Ablating neurons within discovered GCCs (**Ours**) causes a massive logit drop (▽ 8.60), confirming they are critical components.
2. **Completeness**: Ablating an equal number of neurons outside the discovered circuits (**Ours**^{C}) causes a minimal drop (▽ 1.74), showing the circuits capture almost all necessary signal.

	ViT	Swin-T	CLIP-ViT	Avg
Original	76.61	81.92	62.19	73.57
Random	77.19	81.34	59.75	72.76 (▽ 0.81)
Ours	**58.47**	**36.92**	**23.78**	**39.72** (▽ **33.85**)

Table 2. Impact of circuit ablation on ViT and its variants.

Insertion and Deletion Game validates the quality and fidelity of GCC's dual-score connections by progressively deleting or inserting ranked edges.



User-study Responses Responses from over 31 participants demonstrate the discovered circuits are perceptually meaningful and interpretable to humans.