

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

Supplementary Material

A. Experimental Settings

A.1. Model Configuration

Our experiments cover four CNNs (ResNet-50/101, VGG-19, MobileNetV3) and three transformer-based models (ViT-B/32, Swin Transformer, CLIP-ViT). For brevity, we refer to ResNet-50, ResNet-101, VGG-19, and MobileNetV3 as R50, R101, V19, and M3, respectively. Table C details the network layers analyzed for identifying concept circuits. We utilize model architectures and their corresponding pretrained weights from Torchvision on ImageNet, except for the CUB-200-2011 dataset, where we employ the anonthors/cub200-resnet50 pretrained weights available on the Hugging Face Hub [2]. In transformer-based models, the feed-forward network (FFN) is known to serve as a memory mechanism, owing to its structural resemblance to the attention mechanism. Moreover, previous studies [12, 39] indicate that it captures a range of high-level, human-interpretable concepts. Based on this, we analyze the activations from the first linear layer of the FFN in each Transformer block.

A.2. Root Node Selection

The choice of root nodes determines the conceptual focus of the Granular Concept Circuits (GCCs). We define root nodes as neurons that rank among the top 1% in activation across the entire dataset. However, this threshold can be adjusted for broader exploration. For instance, selecting the top 10% highly activated neurons increases the number of discovered circuits, allowing for a wider variety of conceptual representations. While this approach enhances diversity, it introduces a trade-off: if a root node has lower activation levels in the given query, the retrieved circuit may be less relevant to the query itself. Therefore, the selection of root nodes should be carefully tuned to balance concept diversity and query relevance.

A.3. Parameter Selections.

Selections for τ_{NS} . To evaluate connectivity between two nodes, we introduce two scores, one of which is the Neuron Sensitivity Score (S_{NS}). This score quantifies the influence of a neuron (source node) in the current layer by measuring changes in the next layer when the source node is zero-masked. If a particular node in the next layer experiences an exceptionally large decrease, we consider it strongly associated with the source node.

To identify such significant changes, we recommend us-

Model	Layer Names
R50	layer1.0, ..., layer1.2, layer2.0, ..., layer2.3, layer3.0, ..., layer3.5, layer4.0, ..., layer4.2
R101	layer1.0, ..., layer1.2, layer2.0, ..., layer2.3, layer3.0, ..., layer3.22, layer4.0, ..., layer4.2
V19	1, 3, 6, 8, 11, 13, 15, 18, 20, 22, 25, 27, 29, 32, 34, 36
M3	0.0, 1.0, 1.1, 2.0, 2.1, 2.2, 3.0, 3.1, 3.2, 3.3, 4.0, 4.1, 5.0, 5.1, 5.2, 6.0
ViT-B	encoder.layers.0, ..., encoder.layers.11
CLIP-ViT	visual.transformer.resblocks.0, ..., resblocks.11
Swin-T	features.SwinTransformerBlock.0, ..., SwinTransformerBlock.11

Table C. Network layers used for constructing Granular Concept Circuits.

ing the Peak Over Threshold (POT) method, a statistical approach from Extreme Value Theory (EVT) commonly used in anomaly detection. POT models the tail behavior of a distribution by focusing on values exceeding a high threshold, without requiring assumptions about the underlying distribution. As shown in Figure 2, our score distribution is typically right-skewed, allowing us to identify sufficiently large score values as positive anomalies. Specifically, POT captures extreme events by selecting all observations above a predefined threshold, with exceedances modeled using the Generalized Pareto Distribution (GPD). Choosing an appropriate threshold is crucial—setting it too low includes non-extreme values, while setting it too high limits data availability. We examine the effect of the POT threshold on constructing GCCs by testing thresholds at 95, 90, 80, 70, and 60 percentiles. A looser threshold (e.g., 60) captures more nodes, while stricter thresholds may focus on the most significant ones. We quantitatively compare logit changes

when attenuating GCCs with different thresholds, using the same settings as Table 2.

Contrary to expectations, the degree of logit drop remains largely consistent across different POT thresholds. We hypothesize that this is because only a small portion of the network is actively engaged with a given query, while the rest is redundant. Since looser POT thresholds do not significantly improve performance but may introduce redundant circuits and increase computational cost, we recommend using a POT threshold between 90 and 95. While the optimal parameter may vary depending on the dataset or model, our empirical results suggest that thresholds in this range provide sufficiently strong performance. Note that the threshold for S_{NS} can also be determined using other well-established statistical methods, such as Interquartile Range-based outlier detection [35].

POT	95	90	80	70	60	AVG
Original	17.17	16.77	17.28	16.71	16.52	-
Random	15.66	15.15	15.61	14.81	14.54	(▼1.74)
Ours	6.41	6.60	10.42	6.92		(▼5.66)
Ours ^C	16.12	15.41	16.12	15.15	15.43	(▼1.24)

Table D. Logit change comparison using threshold values estimated by POT in ResNet50.

Selections for τ_{SF} . The Semantic Flow Score (S_{SF}) ensures that information from the source node is meaningfully shared with the connected node. Unlike the Neuron Sensitivity Score (S_{NS}), which directly measures connectivity strength, the Semantic Flow Score helps filter out spuriously aligned nodes among those with high S_{NS} values. To achieve this, we introduce a threshold (τ_{SF}) to control the required level of semantic alignment. A stricter τ_{SF} ensures that only genuinely related nodes are retained, while a looser τ_{SF} may weaken the constraint. This balance is particularly important in complex models, which often exhibit polysemantic characteristics, where an overly high threshold could inadvertently remove valid connections. To address this trade-off, we set τ_{SF} as the average value across all nodes in the target layer, ensuring that only nodes with above-average semantic similarity to the source node are retained. This approach maintains a necessary level of information sharing while allowing the threshold to adapt flexibly based on model and layer characteristics.

A.4. Visualization strategy

To reveal the specific visual features that contribute to a neuron’s activation and enable a precise analysis of learned representations, we apply cropping and masking techniques based on the activation map of a given sample and neuron.

The activation map is upsampled to match the input query image size, and Gaussian blur is applied to suppress noise and improve spatial coherence. The processed activation map is then thresholded to generate a binary mask, isolating highly activated regions. The largest connected region is identified using a bounding box, which is subsequently used to crop and standardize the visualization. Finally, less activated areas are darkened by overlaying the binary mask with lower transparency onto the original image, improving interpretability.

B. Quantitative Results on Transformers

In line with the quantitative results on convolution-based models (Table 2), we evaluate performance degradation in transformer-based models by ablating the discovered circuits. Unlike the previous setting where we measured logit drops, we use accuracy drop as the evaluation metric here, as the transformer architecture includes layer normalization, which can suppress logit differences while still affecting the final decision. The results are presented in Table E. For CLIP-ViT, classification is based on the text embedding with the highest cosine similarity to the image embedding. The results show that ablating our identified circuits leads to a substantial accuracy drop of 33.85%, whereas randomly ablating the same number of nodes results in only a 0.81% decrease. This indicates that our method can be effectively extended to transformer-based models.

	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 E. Impact of circuit ablation on ViT and its variants.

C. User-Study

To evaluate the interpretability and effectiveness of Granular Concept Circuits (GCC), we conducted a user study with 33 participants. The objective was to assess whether GCC visualizations offer meaningful insights into model behavior and enhance human understanding of circuit representations. Figure 1 to Figure L illustrate the questions presented in the user study, while the corresponding responses are summarized in Figure 5.

User study of Concept Circuit


In this section, you are asked to access the relationships between images (or image sets).

Not shared

* Indicates required question

Q1-1. Select all the cases where you think the images share common features with the query. A set of 4 images represents a common concept shared by the given images.

Query



(a)

(b)


(c)

(d)

☒ (a)
 ☐ (b)
 ☐ (c)
 ☒ (d)

Q1-2. Select all the cases where you think the images share common features with the query. A set of 4 images represents a common concept shared by the given images.

Query



(a)

(b)


(c)

(d)

☐ (a)
 ☐ (b)
 ☐ (c)
 ☐ (d)

Q2-1. Select the case that you think is most related to the source feature image on the left. If we were to draw a connection from the left image, which case would be the most natural next image?

source feature



(a)


(b)

(c)

☐ (a)
 ☐ (b)
 ☐ (c)

Q2-2. Select the case that you think is most related to the source feature image on the left. If we were to draw a connection from the left image, which case would be the most natural next image?

source feature



(a)

(b)

(c)

☐ (a)
 ☐ (b)
 ☐ (c)


User study of Concept Circuit

In this section, a query image is provided along with its corresponding circuit. The circuit is a subsequence that represents the features of the given query image. Each element of a circuit consists of an image set, where each image set represents a specific attribute or pattern. An image set consists of 4 images that share common features.

Q3-1. Do you think the circuit below effectively represents the feature for this part?

Please take a look at the highlighted part of the query image. Does the given circuit effectively explain the features represented by the highlighted part? (The relationships between the concepts shared by the 4 images are represented in the circuit.)

Query Image



1

2

3

4

5

Disagree


☐
☐
☐
☐
☐

Agree

Q3-1. Do you think the circuit below effectively represents the feature for this part?

Please take a look at the highlighted part of the query image. Does the given circuit effectively explain the features represented by the highlighted part? (The relationships between the concepts shared by the 4 images are represented in the circuit.)

Query Image



1

2

3

4

5

Disagree


☐
☐
☐
☐
☐

Agree

Q3-1. Do you think the circuit below effectively represents the feature for this part?

Please take a look at the highlighted part of the query image. Does the given circuit effectively explain the features represented by the highlighted part? (The relationships between the concepts shared by the 4 images are represented in the circuit.)

Query Image



1

2

3

4

5

Disagree

☐
☐
☐
☐
☐

Agree

Figure I. User study questions (Part 1). The questions are presented in a sequential order from top-left to bottom-left, followed by top-right to bottom-right. Participants were instructed to read each question carefully and select their responses based on the given options.

Q3-1. Do you think the circuit below effectively represents the feature for this part?

Please take a look at the highlighted part of the query image. Does the given circuit effectively explain the features represented by the highlighted part?
(The relationships between the concepts shared by the 4 images are represented in the circuit.)

Query Image

1 2 3 4 5

Disagree ☐ ☐ ☐ ☐ ☐ Agree

Q3-2. Do you think the each circuit shown above sufficiently represent the diverse concepts present in the query?

1 2 3 4 5

Disagree ☐ ☐ ☐ ☐ ☐ Agree

User-study of Concept Circuit

In this section, a query image is provided along with its corresponding circuit. The circuit is a **subsequence that represents the features of the given query image**. Each element of a circuit consists of an image set, where each image set represents a specific attribute or pattern. An image set consists of 4 images that share common features.

Q4-1. Do you think the circuit below effectively represents the feature for this part?

Please take a look at the highlighted part of the query image. Does the given circuit effectively explain the features represented by the highlighted part?
(The relationships between the concepts shared by the 4 images are represented in the circuit.)

Query Image

1 2 3 4 5

Disagree ☐ ☐ ☐ ☐ ☐ Agree

Q4-1. Do you think the circuit below effectively represents the feature for this part?

Please take a look at the highlighted part of the query image. Does the given circuit effectively explain the features represented by the highlighted part?
(The relationships between the concepts shared by the 4 images are represented in the circuit.)


Query Image

1 2 3 4 5

Disagree ☐ ☐ ☐ ☐ ☐ Agree

Figure J. User study questions (Part 2). Continuation of the user study evaluation, following the same structure as Figure I.

Q4-2. Do you think the each circuit shown above sufficiently represent the diverse *
concepts present in the query?

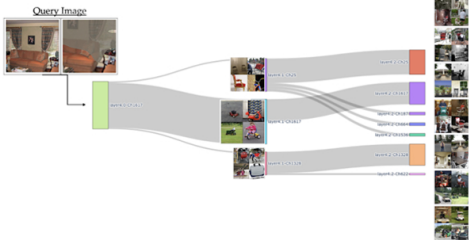


12345

Disagree
☐
☐
☐
☐
☐
Agree

Q5-1. Do you think the circuit below effectively represents the feature for this part? *

Please take a look at the highlighted part of the query image. Does the given circuit effectively explain the features represented by the highlighted part?
(The relationships between the concepts shared by the 4 images are represented in the circuit.)

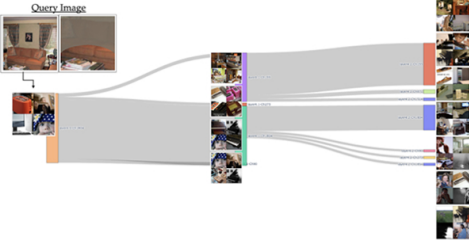


12345

Disagree
☐
☐
☐
☐
☐
Agree

Q5-1. Do you think the circuit below effectively represents the feature for this part? *

Please take a look at the highlighted part of the query image. Does the given circuit effectively explain the features represented by the highlighted part?
(The relationships between the concepts shared by the 4 images are represented in the circuit.)

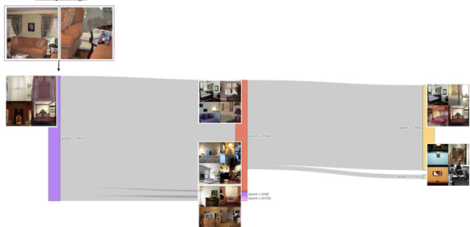


12345

Disagree
☐
☐
☐
☐
☐
Agree

Q5-1. Do you think the circuit below effectively represents the feature for this part? *


Please take a look at the highlighted part of the query image. Does the given circuit effectively explain the features represented by the highlighted part?
(The relationships between the concepts shared by the 4 images are represented in the circuit.)



12345

Disagree
☐
☐
☐
☐
☐
Agree

Q5-2. Do you think the each circuit shown above sufficiently represent the diverse *
concepts present in the query?




12345

Disagree
☐
☐
☐
☐
☐
Agree

User-study of Concept Circuit

In this section, you are asked to evaluate how well the concept circuit captures and represents the common features shared across the query image set.

Q6-1. Do you think the circuit below is related to the given query? *



12345

Disagree
☐
☐
☐
☐
☐
Agree

Figure K. User study questions (Part 3). Continuation of the user study evaluation, following the same structure as Figure J.

Q6-2. Do you think the circuit below is related to the given query? *

1 2 3 4 5

Disagree ☐ ☐ ☐ ☐ ☐ Agree

Q6-3. Do you think the circuit below is related to the given query? *

1 2 3 4 5

Disagree ☐ ☐ ☐ ☐ ☐ Agree

User-study of Concept Circuit

In this section, you are asked to compare the circuits based on two criteria:

- Representativeness** – How well the circuit captures and represents the concepts found in the query image.
- Diversity** – The extent to which the circuit includes a diverse range of concepts.

Which circuits represent the diverse concepts of the query image? *

1 2 3 4 5

Left circuit represents more diverse concepts. ☐ ☐ ☐ ☐ ☐ Right circuit represents more diverse concepts.

Which circuits represent the diverse concepts of the query image? *

1 2 3 4 5

Left circuit represents more diverse concepts ☐ ☐ ☐ ☐ ☐ Right circuit represents more diverse concepts ☐ ☐ ☐ ☐ ☐

Figure L. User study questions (Part 4). Final set of evaluation questions in the user study.