Visual Concept Connectome (VCC): Open World Concept Discovery and their Interlayer Connections in Deep Models

Matthew Kowal[1,2] Richard P. Wildes[1,3] Konstantinos G. Derpanis[1,2,3]

[1]York University, [2]Vector Institute, [3]Samsung AI Centre Toronto

Highlight CVPR JUNE 17-21, 2024 SEATTLE, WA

Project Page: yorkucvil.github.io/VCC

Motivation and Research Questions

The goal of this work is to discover human interpretable concepts and their interlayer connections in a deep network. We present the Visual Concept Connectome (VCC): a novel approach for simultaneously discovering concepts at every layer and their connection weightings across any layers. This paper aims to answer the following questions with VCCs:

1. What patterns of concepts and connections arise in models trained for different tasks?
2. Does model architecture impact the hierarchical structure of concept abstractions?
3. Do models implicitly cluster super-classes into sub-classes?

Method

i) Top down clustering into feature space image segments

ii) Layer-wise concept discovery

iii) Interlayer concept connectivity (ITCAV)

Results

Top: An all layer VCC for the VGG16 model targeting the class Tow Truck. At early layers, oriented pattens and brown color compose green+brown orientation. Middle layers show wheels on road being composed of wheels and asphalt regions. In the final layer, both foreground objects and background regions highly influence the final class.

Bottom: A three layer VCC for the GoogLeNet model targeting the class Jay. The inception5b bird concepts form as selective weightings of inception4c concepts background, bird part, and tree branch, while the inception5b background concepts form differently from green leaves and tree branches. Notably, the network separates subspecies of Jay in the final layer as Blue Jay (c_{52}) and other Jay species (c_{51}).

We compute standard metrics for analyzing tree-like data structures averaged over 50 ImageNet classes. We find: (i) The number of concepts decrease through layers for all models. (ii) Transformers have higher edge variance at the final layers (i.e., concepts are more compositional). (iii) Robust and SSL models rely less on earlier concepts (likely due to transformation robustness).

We apply VCCs to debug model failures in a compositional nature using the L2 distance between the concepts in both class VCCs. The *Tricycle* is incorrectly classified as a *Rickshaw* due to the collapsible hood and background concepts in layers two through four.

We validate the three components of VCCs: (a) our segments follow the receptive field closer than baselines. (b) Removing concepts impacts performance more than random directions. (c,d) ITCAV connection strength of a concept is predictive of the target logit value when classifying concept segments.

Conclusion

1) We introduced the VCC, a method for discovering human interpretable concepts and their interlayer connections in a fully unsupervised way.
2) ConvNets were found to rely heavily on final layer concepts for prediction, while transformers were found to have more diverse connection strengths in the final layer and a more complex assembly of later concepts from earlier ones.
3) We applied VCCs to model debugging and showed how the hierarchical concept representation can indicate how mistakes arise.