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

Matthew Kowal[1,3] Richard P. Wildes[1,2] Konstantinos G. Derpanis[1,2,3]
[1]York University, [2]Samsung AI Centre Toronto, [3]Vector Institute
yorkucvil.github.io/VCC

Introduction. We present a new methodology for interpreting vision models, *the Visual Concept Connectome (VCC)*, which discovers human interpretable concepts and their interlayer connections in a fully unsupervised manner. Our approach simultaneously reveals fine-grained concepts at a layer, connection weightings across all layers and is amendable to global analysis of network structure (*e.g.* branching pattern of hierarchical concept assemblies). Quantitative and qualitative empirical results show the effectiveness of VCCs in the domain of object recognition.

Previous work has focused on interpreting models via feature attribution, which measures the contribution of individual inputs to a model's output [1, 10, 14]; so, explanations are of single pixels and may be difficult to interpret. Other work generates images that maximize activation of a model's features [3, 8, 13]. Like feature attribution, these approaches are qualitative and place most of the burden on the user to determine the concepts revealed. Concept-based interpretability, which identifies human-interpretable abstractions in a model's latent space [4–7], yield quantitative contributions of a concept to the model's output and explanations on the class-level. These approaches are easier to understand and validate, but they have not been used to explore *interlayer* relationships.

As it stands, no approaches can quantify the interlayer affect of a given distributed concept at one layer, l, to another concept at a different layer, l' (rather than the model *output*). Even though it is well established that deep networks learn to build concepts hierarchically as information flows through the network [9, 13], understanding the hierarchical representations has been under-researched. Indeed, little is known about the characteristics of this concept hierarchy for today's models. Questions abound: *How many concepts exist in a network? What are the connections and weights between concepts? Does the model architecture impact the hierarchical structure of concept abstractions?*

In response to these questions, we take inspiration from the biological notion of a *connectome* [11], defined as "a comprehensive structural description of the network of elements and connections forming a brain." Analogously, we present the *Visual Concept Connectome* (VCC), a comprehensive structural description of a deep pretrained network in terms of human-interpretable concepts and their relationships that form the internal representation maintained by the model. VCCs work in the *open-world* setting, *i.e.* the concepts and interlayer connections are discovered without the need for a predefined concept dictionary. We have implemented a futotally automated procedure for generating VCCs for arbitrary deep networks.

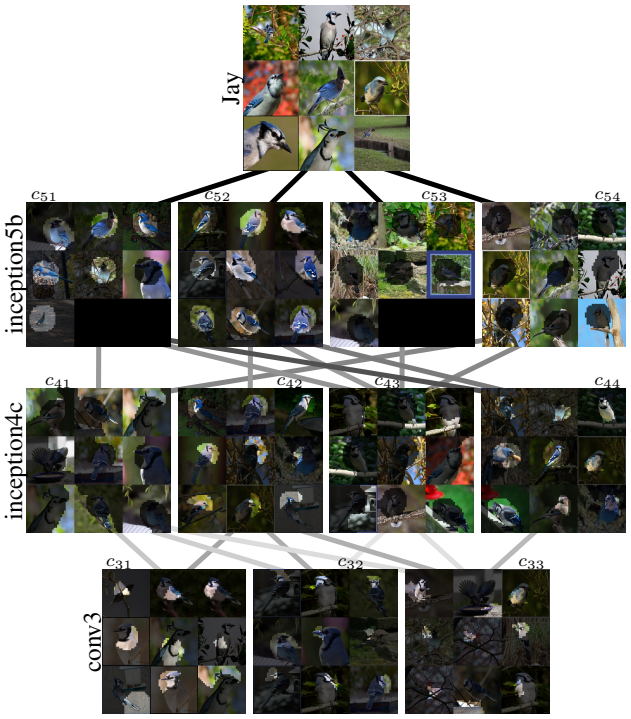

Figure 1. A VCC for three selected layers of a GoogLeNet model [12] targeting recognition of class "Jay". At each layer the visual concepts learned by a deep model for a given class are revealed as are the learned interlayer concept connections. For each concept, up to nine exemplars are shown as unmasked regions in a 3×3 image. Darker lines denote stronger connection weights.

Results. Figure 1 shows VCC for three selected layers of GoogLeNet [12] targeting the class "Jay"; examination reveals how hierarchical concepts are assembled in this

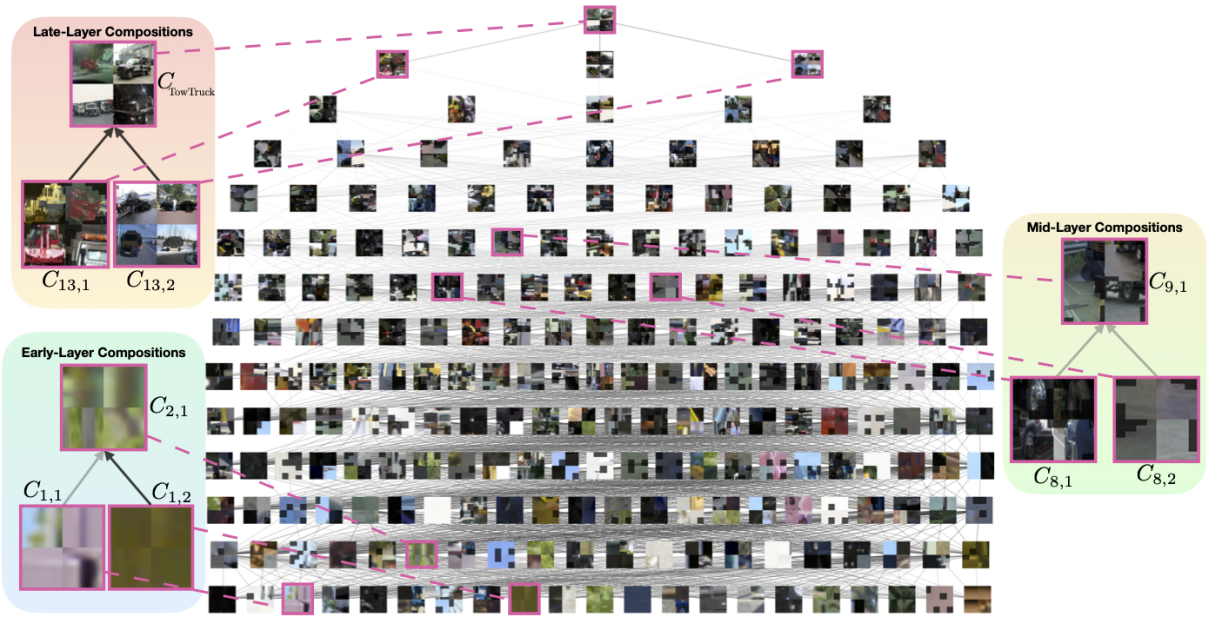

Figure 2. Shown is a VCC for every convolutional layer of a VGG16 model [12] trained on ImageNet [2] targeting recognition of class "Tow Truck". A closer visualization of VCC subgraphs reveals interesting compositions occurring at different levels of abstraction corresponding at different depths of the model. At early layers (bottom left), we observe oriented patterns ($C_{1,1}$) and brown color ($C_{1,2}$) composing the concept of green and brown orientation ($C_{2,1}$). Middle layers (right) show the concept of 'wheel on the road' ($C_{9,1}$) being composed of wheels ($C_{8,1}$) and regions of asphalt ($C_{8,2}$). The final layer concepts (top left) show that both foreground objects, *e.g.* tow trucks ($C_{13,1}$), and background regions, *e.g.* road, trees, humans, or car being towed ($C_{13,2}$), concepts highly influence the final category ($C_{TowTruck}$).

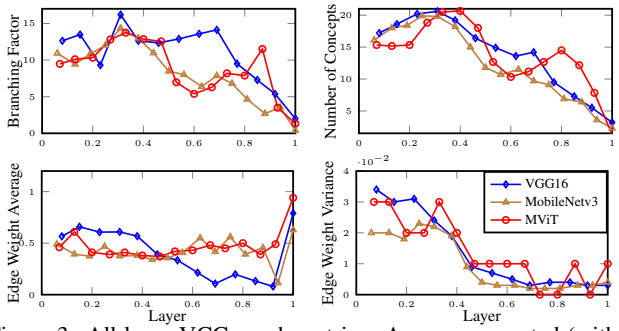

Figure 3. All layer VCC graph metrics. Averages reported (with normalized layers) over 10 VCCs for 10 random ImageNet classes.

network: The inception5b 'bird' concepts (c_{51}, c_{52}) form as selective weighting of inception4c concepts background (c_{41}), bird part (c_{42}, c_{44}) and tree branch (c_{43}), while the inception5b background concepts (c_{53}, c_{54}) form differently from weighting of solely inception4c background part concepts (*e.g.* tree branch (c_{43}) and green leaves (c_{41})). Notably, the network separates subspecies of Jay in the final layer (*e.g.* Blue Jay (c_{52}) and other types (c_{51})). Further, the concepts found in inception4c are composed from varying combinations of colors and parts found in conv3 (*e.g.* various bird parts (c_{31}, c_{33}) contribute to the bird concepts at inception4c). In the end, both scene and object contribute with strong weights to the final category.

An all layer VGG-16 VCC visualization is presented in Fig. 2. As discussed in the caption, hierarchical concept assemblies again are revealed, with both target object as well as its background contributing to the final classification. While both the few layer and all layer visualizations reveal the concept representations of the models under analysis, they afford different levels of granularity: Few layer visualization provides a concise summary with a focus on specific layers, while all layer provides very detailed study.

VCCs are not just a visualization tool, as they also support quantitative analysis of a network's concept structure. Figure 3 shows quantitative analyses on all layer VCCs for three diverse models: a standard CNN (VGG-16), an efficient model (MobileNetv3) and a transformer (MViT). For all models, concept composition is nonlinear across layers, with branching factor ranging 5-15 and converging to approximately two near the last layers. The peak number of concepts, around 20, is at 30-40% of network depth. In contrast, each model displays unique edge weight characteristics: VGG16's average weights decrease in later layers, MobileNetv3's drop greatly before the final layer and MViT maintains consistent values. These results indicate that penultimate concepts differ between ConvNets vs. transformers. While all models generally show decreased variance with layer depth (*i.e.* greater concept diversity in earlier layers), transformers show a marked variance increase in the final layer, indicating greater compositionality. These differences lend insight into the generally stronger classification performance of transformers vs. ConvNets: Higher edge weight variance suggests stronger ability to weight concepts selectively based on categorization importance.

References

[1] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[3] Christoph Feichtenhofer, Axel Pinz, Richard P Wildes, and Andrew Zisserman. Deep insights into convolutional networks for video recognition. *International Journal of Computer Vision*, 128:420–437, 2020. 1

[4] Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. CRAFT: Concept recursive activation factorization for explainability. In *Conference on Computer Vision and Pattern Recognition*, pages 2711–2721, 2023. 1

[5] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, 2019.

[6] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, and Fernanda Viegas. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning*, pages 2668–2677, 2018.

[7] Matthew Kowal, Achal Dave, Rares Ambrus, Adrien Gaidon, Konstantinos G Derpanis, and Pavel Tokmakov. Understanding video transformers via universal concept discovery. *arXiv preprint arXiv:2401.10831*, 2024. 1

[8] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. https://distill.pub/2017/feature-visualization. 1

[9] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 2018. https://distill.pub/2018/building-blocks. 1

[10] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *International Conference on Computer Vision*, pages 618–626, 2017. 1

[11] S. Seung. *Connectome: How the Brain's Wiring Makes Us Who We Are*. Houghton Mifflin Harcourt, 2012. 1

[12] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. 1, 2

[13] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833, 2014. 1

[14] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 1