Topographic Vision Transformers

Yash Shah, Daniel L. K. Yamins
Stanford University

Abstract

Functional organization in the form of topographic maps is a hallmark of many cortical systems and is believed to arise from biophysical efficiency, such as the minimization of neuronal wiring length. Recently, Margalit et al. (2024) developed the TDANN as a topographic convolutional neural network (CNN) that recapitulated gross ventral stream topography while minimizing feedforward wiring length. However, standard CNNs lack mechanisms for within-layer long-range interactions that are well identified in the primate visual cortex. Here we leverage a vision transformer (ViT), which learns to behave locally like CNNs through training and possesses long-range interactions via self-attention, to learn topographic properties. We find that a topographic ViT reproduces key topographic motifs, maintains high object categorization performance, and shows reduced inter- and intra-layer wiring length. We thus introduce a new class of topographic models that can express hypotheses about the roles of local vs. long-range cortical interactions in the brain.

Keywords: topography; deep neural networks; vision transformers; ventral visual cortex; long-range connections

Motivation

Topographic maps—which portray the arrangement of neurons in characteristic spatial patterns—have been ubiquitously found in various cortical systems such as visual (Hubel & Wiesel, 1962), auditory (Humphries, Liebenthal, & Binder, 2010), parietal (Harvey, Klein, Petridou, & Dumoulin, 2013), precentral (Wong, Kwan, MacKay, & Murphy, 1978), and medial entorhinal (Obenhaus et al., 2022).

Recently, Margalit et al. (2024) built the TDANN—a topographic convolutional neural network (CCN)—by imposing a proxy for biophysical efficiency on the task-optimization learning objective and found that the model predicted topographic maps in early and higher ventral visual cortex. CNNs have established themselves as mechanistic models of biological vision, in part due to the hard-coded hierarchy, underlying biologically-plausible computation, and extensive vetting on both neural and behavioral data (Lindsay, 2021; Yamins et al., 2014). Subsequently, relatively sparse but growing attention has been paid to vision transformers (ViTs) (Vaswani et al., 2017) as models of the ventral visual cortex, which not only perform computations that are believed to be biological, perhaps resembling neuron-astrocyte interactions in the brain (Kozachkov, Kastanenka, & Krotov, 2023), but also effectively predict neural and behavioral visual data (Conwell, Prince, Kay, Alvarez, & Konkle, 2024). A lesser known fact about ViTs is that they learn to behave like CNNs through training by establishing a local-to-global hierarchy in effective receptive field patterns (Huang, Kotar, Lee, Cao, & Yamins, 2024). Additionally, they implement what can be thought of as within-layer long-range lateral connections-like interactions via self-attention. Long-range horizontal connections have been widely identified in primate visual cortex (Liang et al., 2017), but seem to be lacking in standard CNNs. ViTs, thus, present themselves as good candidates for understanding the potential contribution of both local and long-range interactions within the framework of building a topographic model, as well as for studying the effects thereby on those interactions.

Results

We build a topographic ViT by training a ViT base-16 on both a task and spatial objectives. The task objective is self-supervised contrastive learning via the MoCo v3 (Chen, Xie, & He, 2021) training objective. The spatial objective encourages nearby pairs of model units from its attention layers, which are placed on a simulated cortical sheet, to have more correlated responses than distant pairs (following Margalit et al. (2024)).

Reproducing ventral stream topography

The topographic ViT recapitulates V1 preference maps for orientation, spatial frequency, and color, exhibiting pinwheel-like iso-orientation domains and punctate color blobs (Figure 1(a)). Quantitative analyses of the difference in preference as a function of pairwise cortical distance, circular variance, and preferred orientation produce curves closely matching those observed in macaque V1. Although the model slightly underperforms the TDANN on the map smoothness metric and has a larger cardinality index, it achieves a higher pinwheel density and a greater proportion of strongly orientation-selective units, aligning more closely with macaque data.

Similarly, the topographic ViT produces category-selective patches in its most VTC-like layer (Figure 1(b)). Selectivity maps are slightly less smooth, if not as smooth, compared to human VTC data on various domain categories from the fLoc stimuli set (Stigliani, Weiner, & Grill-Spector, 2015). While the model produces a higher average patch count than the TDANN, the average patch surface area across categories approaches measurements from human VTC.

Furthermore, the topographic ViT effectively predicts macaque electrophysiological responses under a linear mapping, without any compensation of the hierarchical alignment between model layers and visual areas (Figure 1(c)).

Maintaining high object categorization performance

A key limitation of the TDANN was the reduction in ImageNet object categorization performance under a spatial constraint. In contrast, not only does a task-only ViT outperform the TDANN by approximately 25% on object categorization, the

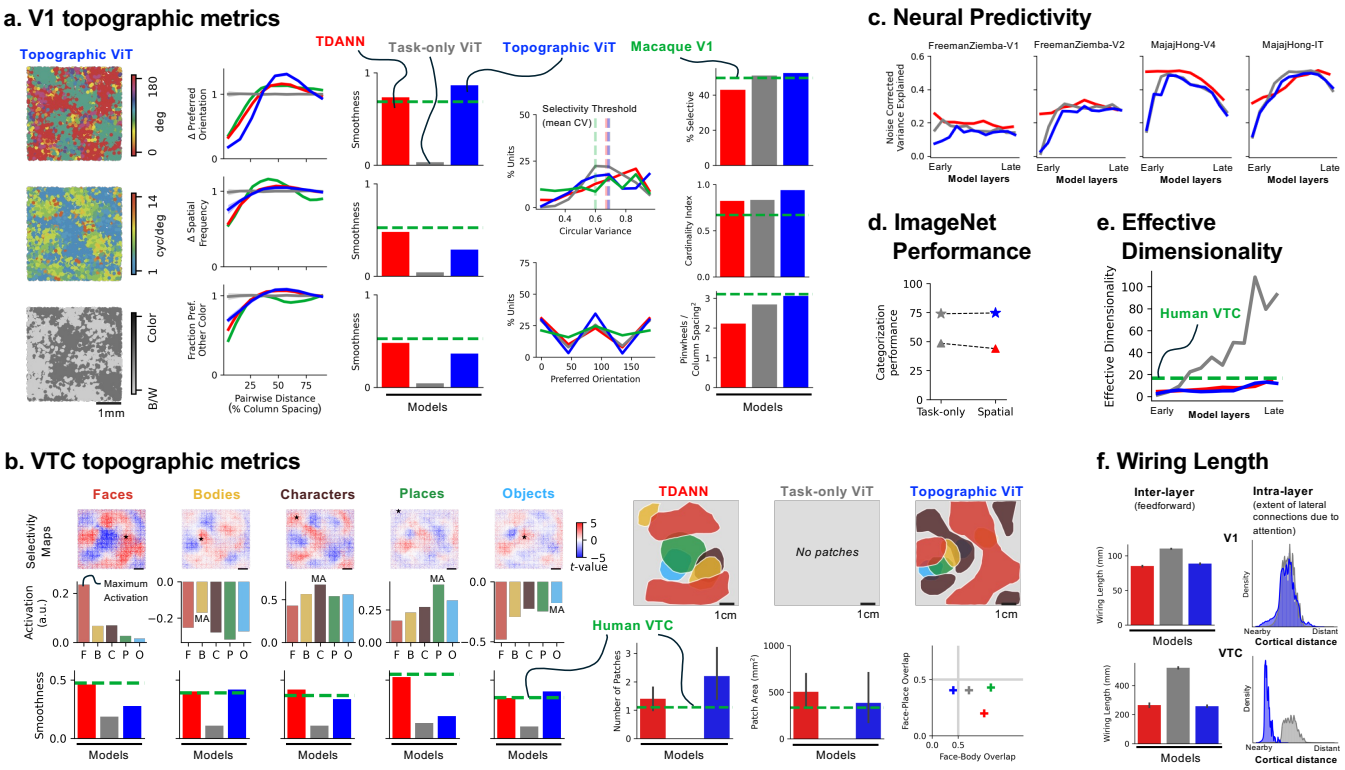

Figure 1: **The topographic ViT reproduces V1- and VTC-like topography, does not incur a performance drop on object categorization, and minimizes both inter- and intra-layer wiring length.** **(a)** Moving from left to right across columns, in a breadth-first order: orientation, spatial frequency, and color preference maps for a random inset on the simulated cortical sheet; pairwise difference in preference as a function of pairwise cortical distance; map smoothness; distribution of circular variance (vertical lines represent thresholds for strong selectivity as determined by the mean circular variance (Ringach et al., 2002)); distribution of preferred orientations; percentage of strongly orientation selective units; proportion of units preferring cardinal directions over obliques; and density of pinwheel-like discontinuities. **(b)** Selectivity for various categories from the fLoc stimuli set and responses to these categories for a unit highlighted with a black star; map smoothness for each category; category-selective patches; average number of patches across categories; average surface area of patches; and overlap between units that are face- and body-selective vs. face- and place-selective. **(c)** Variance explained under model units to neural data mapping via linear regression. **(d)** ImageNet validation set object categorization performance. **(e)** Estimated effective dimensionality. **(f)** Feedforward and within-layer wiring length. Within-layer "wiring" length is computed by looking at the attention-weighted distance on the simulated cortical sheet between each pair of units. For more details on the metrics, please refer to Margalit et al. (2024).

topographic ViT does not incur a performance drop due to the imposition of a spatial constraint (Figure 1(d)).

Reducing intrinsic dimensionality and wiring length

Next, we analyze the effects of spatial constraints on learned model features. We find that the topographic ViT shows reduced intrinsic dimensionality of population responses across all layers, mirroring the TDANN (Figure 1(e)).

Spatial constraints also minimize inter-layer (feedforward) wiring length in the model (Figure 1(f) left). Additionally, because ViTs incorporate long-range lateral connections-like interactions via self-attention, we can measure within-layer "wiring" length based on attention-weighted cortical distances between pairwise unit interactions. Intriguingly, we observe that long-range interactions are heavily penalized, likely reflecting the high "wiring" cost they incur (Figure 1(f) right). For both inter- and intra-layer interactions, wiring length is penalized more in higher than early model layers.

Discussion

In this paper, we introduce a topographic ViT by incorporating a spatial constraint into a self-supervised task objective, aiming to assess whether the model can capture the gross topography of ventral visual cortex—and, if so, to investigate the roles of local and long-range interactions under this constraint. We find that the topographic ViT reproduces key aspects of both V1- and VTC-like topography, maintains high categorization performance, and exhibits reduced inter- and intra-layer wiring length. Further investigation is needed to clarify the contribution of long-range interactions in early model layers, where such connections are not as strongly penalized, and to understand their role in shaping the emergence of topography.

Acknowledgements

This work was supported by the following awards: To D.L.K.Y.: Simons Foundation grant 543061, National Science Foundation CAREER grant 1844724, National Science Foundation Grant NCS-FR 2123963, Office of Naval Research grant S5122, ONR MURI 00010802, ONR MURI S5847, and ONR MURI 1141386 - 493027. We also thank the Stanford HAI, Stanford Data Sciences and the Marlowe team, and the Google TPU Research Cloud team for computing support.

References

Chen, X., Xie, S., & He, K. (2021). An empirical study of training self-supervised vision transformers. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 9640–9649).

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2024). A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature communications*, *15*(1), 9383.

Harvey, B. M., Klein, B. P., Petridou, N., & Dumoulin, S. O. (2013). Topographic representation of numerosity in the human parietal cortex. *Science*, *341*(6150), 1123–1126.

Huang, F., Kotar, K., Lee, W., Cao, R., & Yamins, D. (2024). Are vits as global as we think?-assessing model locality for brain-model mapping. *Cognitive Computational Neuroscience*.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, *160*(1), 106.

Humphries, C., Liebenthal, E., & Binder, J. R. (2010). Tonotopic organization of human auditory cortex. *Neuroimage*, *50*(3), 1202–1211.

Kozachkov, L., Kastanenka, K. V., & Krotov, D. (2023). Building transformers from neurons and astrocytes. *Proceedings of the National Academy of Sciences*, *120*(34), e2219150120.

Liang, H., Gong, X., Chen, M., Yan, Y., Li, W., & Gilbert, C. D. (2017). Interactions between feedback and lateral connections in the primary visual cortex. *Proceedings of the National Academy of Sciences*, *114*(32), 8637–8642.

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of cognitive neuroscience*, *33*(10), 2017–2031.

Margalit, E., Lee, H., Finzi, D., DiCarlo, J. J., Grill-Spector, K., & Yamins, D. L. (2024). A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron*, *112*(14), 2435–2451.

Obenhaus, H. A., Zong, W., Jacobsen, R. I., Rose, T., Donato, F., Chen, L., ... Moser, E. I. (2022). Functional network topography of the medial entorhinal cortex. *Proceedings of the National Academy of Sciences*, *119*(7), e2121655119.

Ringach, D. L., Shapley, R. M., & Hawken, M. J. (2002). Orientation selectivity in macaque v1: diversity and laminar dependence. *Journal of neuroscience*, *22*(13), 5639–5651.

Stigliani, A., Weiner, K. S., & Grill-Spector, K. (2015). Temporal processing capacity in high-level visual cortex is domain specific. *Journal of Neuroscience*, *35*(36), 12412–12424.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wong, Y., Kwan, H., MacKay, W., & Murphy, J. (1978). Spatial organization of precentral cortex in awake primates. i. somatosensory inputs. *Journal of neurophysiology*, *41*(5), 1107–1119.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619–8624.