First Steps in Using Topographic Deep Artificial Neural Network Models to Generate Hypotheses about Not-yet-detected Functional Neural Aggregates in the Ventral Stream

Kamila M. Jozwik (kmjozwik@mit.edu)
Department of Psychology
University of Cambridge, Downing Site, Craik-Marshall Building, CB2 3AR United Kingdom
McGovern Institute for Brain Research and Center for Brains, Minds, and Machines
Massachusetts Institute of Technology, Cambridge, MA 02139 United States

Hyodong Lee (hyo@mit.edu)
McGovern Institute for Brain Research
Massachusetts Institute of Technology, Cambridge, MA 02139 United States

Nancy Kanwisher (ngk@mit.edu)
Department of Brain and Cognitive Sciences
McGovern Institute for Brain Research
Center for Brains, Minds, and Machines
Massachusetts Institute of Technology, Cambridge, MA 02139 United States

James J. DiCarlo (dicarlo@mit.edu)
Department of Brain and Cognitive Sciences
McGovern Institute for Brain Research
Center for Brains, Minds, and Machines and MIT Quest for Intelligence
Massachusetts Institute of Technology, Cambridge, MA 02139 United States

Abstract

Although several types of spatially-aggregated neural functional selectivities have been reported in the inferior temporal (IT) cortex of humans and monkeys, such as face, place, and body selectivities, broad swaths of IT have yet to be similarly characterized. Here, we present the first steps of using Topographic Deep Artificial Neural Networks (TDANNs) as hypothesis generators of not-yet-detected spatially-aggregated IT functional selectivities. To isolate the shared selectivities across a population of TDANNs, we applied hyperalignment to the IT layer of ten TDANNs. We then analyzed the shared underlying functional representations to identify eleven predicted neuronal functional selectivity clusters. After mapping these clusters back to the spatial IT maps in each TDANN, we find that face-selective units – which spatially aggregate in TDANNs – are strongly loaded on one of these functional clusters. On visual inspection, the other functional clusters appear to be selective for scenes, animal bodies, and mid-level object properties. Topographic ANNs, when analyzed in this manner, could be used to predict novel spatially-aggregated selectivities shared by all brains and to predict the spatial relationships between those functional aggregates. Both types of predictions could then be tested via targeted fMRI experiments.

Keywords: object vision; inferior temporal cortex; topography; selectivity; dimensionality; deep artificial neural networks

Introduction

Humans and monkeys easily recognize objects, thanks to the neural computations conducted along the ventral visual pathway, which culminates in the inferior temporal (IT) cortex. Primate and human IT each have spatially-aggregated neural functional selectivities: nearby neurons tend to have similar response properties, and clustering of neural selectivity for some object categories is particularly strong (e.g., faces and bodies, with novel selectivities being proposed (Bao, She, McGill, & Tsao, 2020)). While specific deep ANNs are now widely used to model the image-evoked functional properties of ventral stream neurons (Yamins, Hong, & Cadieu, 2014; Schrimpf et al., 2018), only recently have newer ANN architectures been able to attempt to explain the spatial organization of those neurons (aka functional topography) (Lee et al., 2020; Blauch, Behrmann, & Plaut, 2022; Doshi & Konkle, 2022). In principle, these types of models should allow prediction of not-yet-detected functional selectivities that could then be assessed via targeted experiments. However, a key challenge and opportunity is that these models suggest that no two brains are exactly identical – either functionally or spatially. We developed an approach to meet that challenge: 1) hyperalignment to extract the underlying shared functional neural representational space, 2) Principal Component Analysis (PCA) and functional clustering to identify functional sub-types of neurons, and 3) re-projection of these shared functional sub-types back into spatial coordianates for inspection and prediction of experiments that depend on spatial aggregration of neural functional sub-types (e.g., fMRI). Here, we apply this approach to one topographic ANN architectural family – Topographic Deep Artificial Neural Networks (TDANNs), we demonstrate successful recovery of face-preferring spatial aggregates, and we illustrate other functional subtypes predicted by this family of topographic models.

Methods and Results

Topographic Deep ANNs (TDANNs)

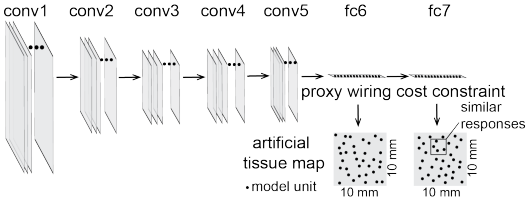

Figure 1: Architecture of Topographic Deep Artificial Neural Networks.

The architecture of TDANNs (Lee et al., 2020) is based on AlexNet (Krizhevsky, Sutskever, & Hinton, 2012). The topographic constraint was applied to "IT" layers fc6 and fc7, as these layers showed the highest predictivity of IT representations at the single unit level. TDANNs are constructed by assigning a random position for each of the model units in 10mm x 10mm IT layers on a two-dimensional artificial tissue map before training, simulating cortical maps in monkey cIT and aIT (Figure 1). TDANNs implement a pair-wise correlation rule within each IT layer: conceptually the rule aims to enforce response correlations to be high for pairs of nearby neurons, and to gradually decrease as a function of cortical distance between the pair. The hyperparameters of this rule (spatial fall-off) are determined from prior primate IT studies. To generate an individual TDANN model (which we take as a model of an individual subject's ventral stream), we initialize the entire network with random weights, and then optimize ("train") to try to satisfy both this local correlation rule in both IT layers and the image classification task using ImageNet Large Scale Visual Recognition Challenge database (Russakovsky et al., 2015). Here we built a population of ten TDANN models using this method (each from a different random starting initialization). We extracted TDANN activations for fc6 based on over 26,000 grayscale images from the THINGS database (Hebart et al., 2019) enriched with 1,700 face images from Labelled Faces in the Wild (Huang, Mattar, Berg, & Learned-Miller, 2008) given that the THINGS database contains few explicit categories that have faces in them.

Approach for defining spatially-aggregated neural functional selectivities in TDANNs: hyperalignment, PCA, and clustering

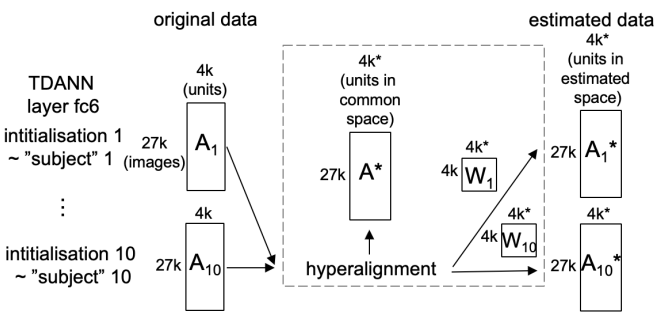

Figure 2: Schematic of hyperalignment of different initialisation conditions of ten TDANNs (see text).

We consider the variation in this stochastic TDANN model generation process (above) to be conceptually analogous to variation across different subjects (Mehrer, Spoerer, Kriegeskorte, & Kietzmann, 2020). All results presented here are based on a population of ten TDANN models (i.e., ten TDANN subjects).

First, to isolate the shared functional selectivities across a population of TDANN models, we applied hyperalignment method (Haxby et al., 2011), which was initially developed to study brains, to the IT layer (cIT) of ten TDANN models thereby projecting all initialization conditions to a common space (Figure 2). The common space is initially defined as the first TDANN "subject" (reference), but is subsequently refined by iteratively combining the common space with the projected TDANN subject. The projection matrices W are Procrustean transformations (shift, scaling, and rotation). Conceptually, hyperalignment is intended to separate a functional representation that is shared across TDANN subjects from functional properties that are idiosyncratic to individual TDANN subjects. The hyperalignment was successful in reducing the mean squared error to the reference TDANN subject by 65.3% (std = 0.8%) as compared to non-hyperaligned TDANN subjects. We interpret this result as 65.3% of the functional variance is (on average) shared across TDANN subjects, whereas the remaining variance is idiosyncratic. Second, we then analyzed the underlying shared functional representation. We performed cross-validated PCA on our estimate of this functional representation, and identified 16 PCs (using the participation ratio (Gao et al., 2017)) that explained 55.2% of the shared variance. We performed K-means clustering on the neurons projected in the 16-PC functional space specifying N clusters = 2,...,20 and found that N clusters = 11 gave the highest silhouette score. Third, we mapped these functional clusters back to the spatial IT maps in each TDANN model/subject.

Using TDANNs to generate hypotheses about not-yet-detected functional neural clusters

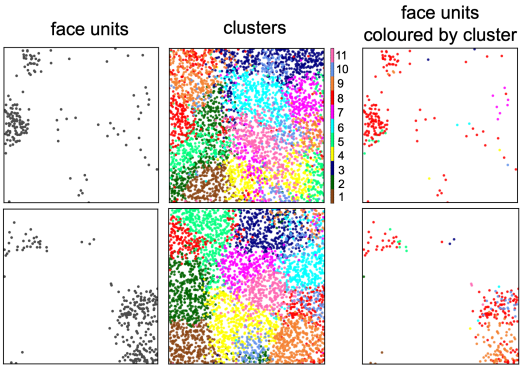

Figure 3: Face-selective units and clusters for two example TDANN initialisations.

First, as a positive control, we tested whether we can identify spatially-aggregated face-selective units in TDANNs using two different face stimulus sets (Downing, Chan, Peelen, Dodds, & Kanwisher, 2005; Stigliani, Weiner, & Grill-Spector, 2015). If our approach allows us to find known brain selectivities like face selectivity, we can use it to generate hypotheses for not-yet-detected spatially-aggregated IT functional selectivities. Spatially-aggregated face selective units (identified by d-prime) that overlapped between the two independent stim-

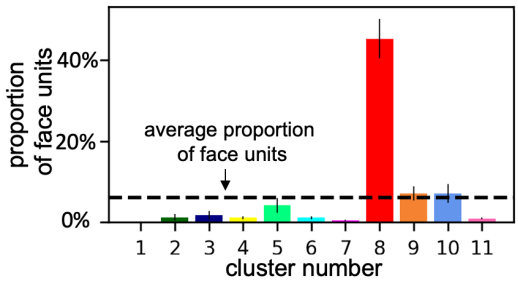

Figure 4: Proportion of face units in each cluster (mean across TDANN initialisations, error bars depict standard error of the mean across initialisations). Dashed line represents average proportion of face units.

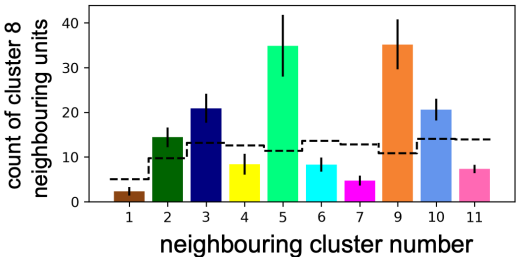

Figure 5: Neighbouring clusters of cluster 8. For each unit of cluster 8, we find the unit closest to it, and if it belongs to another cluster we add it to that neighbouring cluster's count (mean across TDANN initialisations, error bars depict standard error of the mean across initialisations). Dashed line is the average of all such plots across clusters.

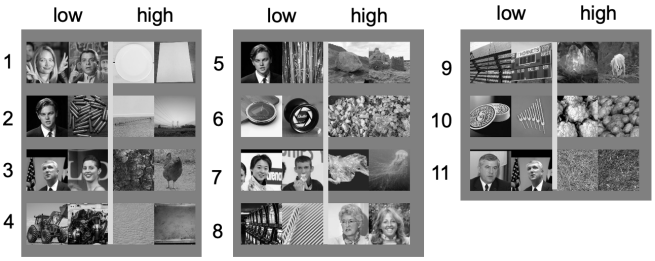

Figure 6: Images that have lowest and highest load on each cluster (mean across TDANN initialisations).

ulus sets were found in every TDANN initialisation. Subsequently, we investigated the correspondence between these face-selective units and 11 functional selectivity clusters revealed by our analysis. We mapped these functional clusters back to the spatial IT maps in each TDANN model (Figure 3). We found that face units are strongly loaded on cluster 8 (Figures 3 and 4). We also found that images that most highly load on cluster 8 contain mostly human faces (Figure 6). Clusters 5 and 9 were the clusters that were spatially nearby cluster 8 (Figure 5). Images that most highly load on these clusters contained rocks and animal bodies (Figure 6). On visual inspection, the other functional clusters appeared to be selective for mid-level object properties: 1 - white objects, 2 - scenes, 4 - sand texture, 6 - clutter, 10 - many small objects, 11 - grassy texture (Figure 6). Clusters 3 and 7 are harder to describe. We believe that TDANNs and other rapidly emerging topographic ANN architectures (Blauch et al., 2022; Doshi & Konkle, 2022) could be used to predict novel spatially-aggregated selectivities shared by all brains and to predict the spatial relationships between those functional aggregates. Both types of predictions could then be tested via targeted fMRI experiments.

Acknowledgements

This work was supported by the Wellcome Trust Grant [206521/Z/17/Z] to KMJ, the National Science Foundation Grant [2124136] to NK and JJD, the Simons Foundation Grant [542965] to JJD, and The Center for Brains, Minds and Machines [NSF-STC, CCF-1231216].

References

Bao, P., She, L., McGill, M., & Tsao, D. (2020). A map of object space in primate inferotemporal cortex. *Nature*, *583*, 103–108.

Blauch, N., Behrmann, M., & Plaut, D. (2022). A connectivity-constrained computational account of topographic organization in primate high-level visual cortex. *Proceedings of the National Academy of Sciences*, *119*, e2112566119.

Doshi, F., & Konkle, T. (2022). Visual object topographic motifs emerge from selforganization of a unified representational space. *bioRxiv*.

Downing, P., Chan, A., Peelen, M., Dodds, C., & Kanwisher, N. (2005). Domain specificity in visual cortex. *Cerebral cortex*, *16*, 1453–1461.

Gao, P., Trautmann, E., Yu, B., Santhanam, G., Ryu, S., Shenoy, K., & Ganguli, S. (2017). A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv*.

Haxby, J., Guntupalli, J., Connolly, A., Halchenko, Y., Conroy, B., Gobbini, M., ... Ramadge, P. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, *72*, 404-16.

Hebart, M., Dickter, A., Kidder, A., Kwok, W., Corriveau, A., Van Wicklin, C., & Baker, C. (2019). Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLoS ONE*, *14*, e0223792.

Huang, G., Mattar, M., Berg, T., & Learned-Miller, E. (2008). Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*.

Lee, H., Margalit, E., Jozwik, K., Cohen, M., Kanwisher, N., Yamins, D., & DiCarlo, J. (2020). Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. *bioRxiv*.

Mehrer, J., Spoerer, C., Kriegeskorte, N., & Kietzmann, T. (2020). Individual differences among deep neural network models. *Nature Communications*, *11*, 5725.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211–252.

Schrimpf, M., Hong, H., Issa, E., Kar, K., Prescott-Roy, J., Rajalingham, R., ... DiCarlo, J. (2018). Brain-score: which artificial neural network is most brain-like? *bioRxiv*.

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

Yamins, D., Hong, H., & Cadieu, S. E. S. D. D. J., C.F. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*, 8619–8624.