Modeling Category-Selective Cortical Regions with Topographic Variational Autoencoders

Anonymous Author(s)
Affiliation
Address
email

Abstract

Category-selectivity in the brain describes the observation that certain spatially localized areas of the cerebral cortex tend to respond robustly and selectively to stimuli from specific limited categories. One of the most well known examples of category-selectivity is the Fusiform Face Area (FFA), an area of the inferior temporal cortex in primates which responds preferentially to images of faces when compared with objects or other generic stimuli. In this work, we leverage the newly introduced Topographic Variational Autoencoder to model of the emergence of such localized category-selectivity in an unsupervised manner. Experimentally, we demonstrate our model yields spatially dense neural clusters selective to faces, bodies, and places through visualized maps of Cohen's d metric. We compare our model with related supervised approaches, namely the TDANN, and discuss both theoretical and empirical similarities. Finally, we show preliminary results suggesting that our model yields a nested spatial hierarchy of increasingly abstract categories, analogous to observations from the human ventral temporal cortex.

1 Introduction

Category-selectivity is observed throughout the cerebral cortex. At a high level it describes the observation that certain localized regions of the cerebral cortex have been measured to respond preferentially to specific stimuli when compared with a set of alternative control images. It has been measured across a diversity of species [26, 51, 39], directly through fMRI and neural recordings [44], and more indirectly through observational studies of patients with localized cortical damage [38]. Examples of category-selective areas in the visual stream include the Fuisform Face Area (FFA) [26], the Parahippocampal Place Area (PPA) [14, 39], and the Extrastriate Body Area (EBA) [42] which respond selectively to faces, places, and bodies respectively. However, the extent of category-selectivity does not stop at such basic categories. Instead, selective maps have been observed for both more abstract 'superordinate' categories, such as animacy versus inanimacy [19, 31], as well as for more fine-grained 'subordinate' categories such as human-faces versus animal-faces [18]. These maps are seen to be superimposed on one-another such that the same cortical region expresses selectivity simultaneously to animate objects and human-faces, while other spatially disjoint regions are simultaneously selective to inanimacy and 'places' (images of scenes). Such overlapping maps have been interpreted by some researchers as nested hierarchies of increasingly abstract categories, potentially serving to increase the speed and efficiency of classification [16].

In interpreting these observations, one may naturally wonder as to the origins of such localized functional specialization. To date, researchers have demonstrated evidence for two potential explanations. Specifically, anatomical constraints such as the arrangement, layout, and properties of different cell bodies can be observed to vary slightly in different regions of the cortex in loose alignment with

Submitted to 3rd Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM 2021) of the Neural Information Processing Systems (NeurIPS) conference.

category selectivity [53, 9, 49]. Similarly, the principle of 'wiring length minimization' [32, 15] can be placed in this category, positing that evolutionary pressure has encouraged the brain to reduce the cumulative length of neural connections in order to reduce the costs associated with the volume, building, maintenance, and use of such connections. Computational models which attempt to integrate such wiring length constraints [34, 57] have recently have been observed to yield localized category selectivity such as 'face patches' similar to those of macaque monkeys. A second hypothesis for the emergence of category specialization which has recently gained increasing empirical support derives its explanatory power from information theory. Empirical studies have discovered that sufficiently deep convolutional neural networks naturally learn distinct and largely separate sets of features for certain domains such as faces and objects. Specifically, the work of Dobs et al. [13], showed that feature maps in the later layers of deep convolutional neural networks can be effectively segregated into object and face features such that lesioning one set of feature maps does not significantly impact performance of the network on classification of the other data domain. Such experiments suggest that the specialization of neurons may simply be an optimal code for representing the natural statistics of the underlying data when given a sufficiently powerful feature extractor.

Pursuant to these ideas, this work proposes that a single underlying information theoretic principle, namely the principle of redundancy reduction [6], may account for localized category selectivity while simultaneously serving as a principled unsupervised learning algorithm. Simply, the principle of redundancy reduction states that an optimal coding scheme is one which minimizes the transmission of redundant information. Applied to neural systems, this describes the ideal network as one which has statistically maximally independant activations – yielding a form of specialization. This idea served as the impetus for computational frameworks such as Sparse Coding [40] and Independant Component Analysis (ICA) [7, 12, 21, 22]. Interestingly, however, further work showed that features learned by linear ICA models were not entirely independant, but indeed contained correlation of higher order statistics. In response, researchers proposed a more efficient code could be achieved by modeling these residual dependencies with a hierarchical topographic extension to ICA [23, 25], separating out the higher order 'variance generating' variables, and combining them locally to form topographically organized latent variables. Such a framework shares a striking resemblance to models of divisive normalization [36, 5], but inversely formulated as a generative model. Ultimately, the features learned by such models were reminiscent of pinwheel structures observed in V1, encouraging multiple comparisons with topographic organization in the biological visual system [24, 25, 37].

In this work, we leverage the recently introduced Topographic Variational Autoencoder [28, 27], a modern instantiation of such a topographic generative model, and demonstrate that it is capable of modeling localized category selectivity as well as higher order abstract organization, guided by a single unsupervised learning principle. We quantitatively validate category selectivity through visualization of Cohen's d effect size metric for different image classes, showing selective clusters for faces, bodies, and places. We compare our model with the supervised wiring cost proxy of Lee et al. [34] and demonstrate qualitatively similar results with an unsupervised learning rule. Finally, we show preliminary results indicating that our model contains a nested spatial hierarchy of increasingly abstract categories, similar to those observed in the human ventral temporal cortex [18, 16].

2 Background

Topographic Generative Models Bayesian modeling, the theoretical framework underlying probabilistic generative models, has been proposed in multiple studies as a potential model of human learning [50, 1]. Abstractly, the goal of a probabilistic generative model is to accurately capture the true data generating process. A typical framework for generative models includes defining a joint distribution over observations $\mathbf{X}$ and unobserved latent variables $\mathbf{T}$, assuming that the joint factorizes into the product of a conditional 'generative' distribution and a prior: $p_{\mathbf{X},\mathbf{T}}(\mathbf{x},\mathbf{t}) = p_{\mathbf{X}|\mathbf{T}}(\mathbf{x}|\mathbf{t})p_{\mathbf{T}}(\mathbf{t})$. In standard generative models such as ICA or Variational Autoencoders (VAEs) [46, 29], it is common to define a prior over latent variables such that all variables are independant. Topographic generative models differ from this by instead having a more complex correlation structure defined by the spatial distance between variables in a pre-defined topographic layout. In Topographic ICA (TICA) [23], such a local-correlation structure was shown to be efficiently achievable through a 2-layer hierarchical generative model. Specifically, at the highest layer, a set of 'variance generating' variables $\mathbf{V}$ are independently sampled and subsequently summed in local neighborhoods to determine the variance of lower level topographic variables $\mathbf{T}$. Formally, for $\mathbf{V} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the variances of $\mathbf{T}$ are given

as $\sigma = \phi(\mathbf{WV})$, for an appropriate point-wise non-linearity ϕ, and a locally connected matrix $\mathbf{W}$ (such as a convolution matrix). Then the vector $\mathbf{T} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ can be seen to have correlations of variance across elements T_i & T_j if these elements share a subset of their variance generating variables V. Interestingly, the proposed wiring length proxy of prior work [34] turns out to be based on the same underlying statistical property as the topographic generative models described in this work, namely local correlation. This suggests that these ideas may not be mutually exclusive, and hints at a potential fundamental connection between wiring length minimization and a generative modeling perspective of the brain.

The Topographic VAE Inspired by linear topographic generative models such as TICA, the Topographic Variational Autoencoder (TVAE) was recently introduced to train deep *nonlinear* latent variable models with topographic structure. The model places a Topographic Product of Student's-T prior [54, 41] over the latent variables, and achieves efficient training through a hierarchical construction identical to that of TICA. Formally, the model parameterizes the conditional generative distribution with a powerful function approximator $p_\theta(\mathbf{x}|g_\theta(\mathbf{t}))$, and trains the paramters of this model through the use of two approximate posteriors $q_\phi(\mathbf{z}|\mathbf{x})$ and $q_\gamma(\mathbf{u}|\mathbf{x})$ which are combined to construct the topographic $\mathbf{t}$ variable. Explicitly:

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}\big(\mathbf{z}; \mu_\phi(\mathbf{x}), \sigma_\phi(\mathbf{x})\mathbf{I}\big) \qquad p_\theta(\mathbf{x}|g_\theta(\mathbf{t})) = p_\theta(\mathbf{x}|g_\theta(\mathbf{z}, \mathbf{u})) \tag{1}$$

$$q_\gamma(\mathbf{u}|\mathbf{x}) = \mathcal{N}\big(\mathbf{u}; \mu_\gamma(\mathbf{x}), \sigma_\gamma(\mathbf{x})\mathbf{I}\big) \qquad \mathbf{t} = \frac{\mathbf{z} - \mu}{\sqrt{\mathbf{Wu}}} \tag{2}$$

The parameters θ, ϕ, γ and μ are then optimized to maximize the likelihood of the data through the Evidence Lower Bound (ELBO):

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})q_\gamma(\mathbf{u}|\mathbf{x})}\left(\log p_\theta(\mathbf{x}|g_\theta(\mathbf{t})) - D_{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p_\mathbf{Z}(\mathbf{z})] - D_{KL}[q_\gamma(\mathbf{u}|\mathbf{x})||p_\mathbf{U}(\mathbf{u})]\right) \tag{3}$$

3 Methods

Evaluation Following prior computational work [34, 57] and fMRI studies [4], we use Cohen's d metric [11, 48], a measure of standardized difference of two means, as our selectivity metric. Given the means $\bar{m}_1$ & $\bar{m}_2$ and standard deviations σ_1 & σ_2 of two sets of data, the d metric is given as:

$$d = \frac{\bar{m}_1 - \bar{m}_2}{\sqrt{\frac{1}{2}\left(\sigma_1^2 + \sigma_2^2\right)}} \tag{4}$$

This value is unitless and can be seen as expressing the difference between two means in terms of units of 'pooled variability'. In this work, the mean $\bar{m}_1$ corresponds to the mean activation of a single neuron computed across an entire dataset of class-specific target images (e.g. faces), while $\bar{m}_2$ is the mean activation of the same neuron across a dataset of control images which do not contain this class.

Datasets The dataset used for training both the TDANN and TVAE is a composition of the ImageNet [47] and Labeled Faces in the Wild (LFW) datasets [20], following Lee et al. [34]. The TDANN was trained to classify the 1000 distinct image classes from ImageNet, plus one generic face class encompassing all of LFW. The TVAE used no such class labels and was trained entirely unsupervised. To measure the category selectivity of the models, the primary test face dataset used in Figures 1, 2, & 4 was a ∼25,000 image subset of VGGface2 [8]. The control 'object' dataset for Figures 1 & 4 was composed of roughly half of the images (25,000) from the validation set of ImageNet. To measure selectivity to body parts and places in Figure 2, we created a 'body' dataset composed of headless body images [55] and hands [3], and used the Place365 dataset [59] for places. In Figure 2, the 'control' set used to compute selectivity for each class was defined to be the compliment of the test set, i.e. all other datasets besides the target category of interest.

Models All models are trained on top of features extracted by the convolutional layers of a pre-trained Alexnet model [33]. The Alexnet architecture was chosen to match the setup from Lee et al. [34] and Zhang et al. [57], and has further been shown to have remarkable similarities to hierarchical processing in the human visual stream [56, 10, 17]. For the TVAE, we randomly initialize and train a single linear layer encoder and decoder with 4096 output neurons, arranged in a 64x64 grid with circular boundary conditions to avoid edge effects. For the TDANN, we randomly initialize and train all three fully connected layers of Alexnet, imposing the spatial correlation loss over both 'FC6' and 'FC7' as in [34]. See Section A.1 for all hyperparameters and training details.

4 Experiments

In the following, we explore the category-selectivity of top-level neurons trained with the Topographic VAE framework on realistic images. We observe that neurons do indeed become category-selective, and that selective neurons tend to group together to form localized category-selective regions for a variety of domains including faces, bodies, and places. We compare these results with a non-topographic baseline (pre-trained Alexnet), and a re-implementation of the TDANN [34], observing qualitatively similar results. Additionally, following Zhang et al. [57], we plot selectivity maps to more abstract concepts (such as animacy and real-world size), and observe that such maps overlap in an intuitive manner, suggesting the existence of a nested spatial hierarchy of categories.

Localized Category-Selectivity In Figure 1, we plot the continuous value of Cohen's d metric for all neurons as arranged in a 2-d grid. The baseline (left) shows the first fully connected layer (FC6) of a pre-trained Alexnet architecture. As expected, the neurons of this model have no defined spatial organization and thus result in a random selectivity map. We note the existence of class-selective neurons is not guaranteed, but their appearance here is in-line with observations from prior work [34, 45, 2]. Secondly, we compare our TVAE model (middle) with our re-implementation of the TDANN [34] (right). We observe that both models demonstrate the emergence of face-selective clusters of comparable size and density. We see that the TVAE framework appears to yield smoother topographic maps, perhaps due to the unified objective function and unsupervised learning rule when compared with the competing supervised classification loss and wiring cost regularization of the TDANN. To validate the robustness of these category selective regions, in Figure 5 of the appendix we plot these same selectivity maps for four different test face datasets, and observe that the face clusters have roughly the same size and location despite the varying statistics across datasets.

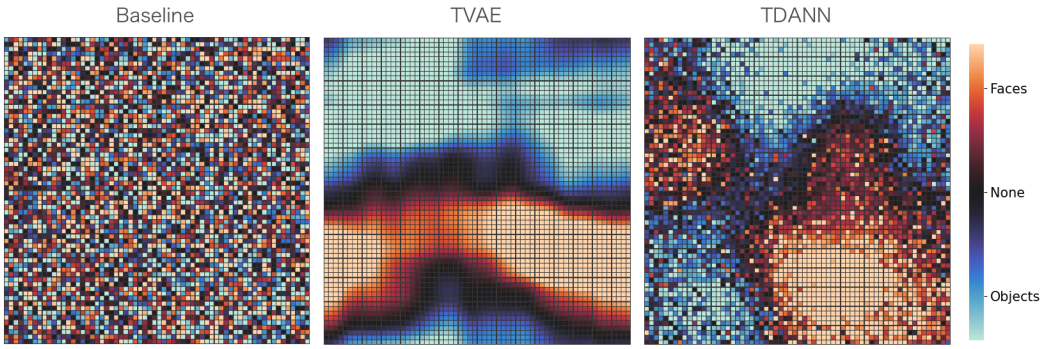

Figure 1: Face vs. Object selectivity for a non-topographic baseline, Topographic VAE, and TDANN. We see the TVAE has an emergent face cluster qualitatively similar to that of the TDANN.

Face, Body & Place Clusters Next, in Figure 2, we plot the simultaneous selectivity of neurons in our TVAE model to multiple classes including faces, bodies, and places. To create a map of multi-class selectivity, we follow prior work and threshold the d metric at 0.85, considered a 'strong effect' [48] and computed to be to be equivalent to a threshold of 0.65 for noisy neural recordings in monkeys [34]. In the plot we observe an overlap of neurons with selectivity to faces and bodies, as seen in prior computational work [34] and fMRI studies [43, 52].

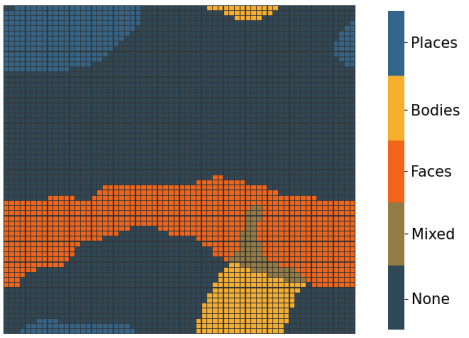

Figure 2: TVAE category-selectivity, $d \geq 0.85$

Locally Distributed Activations To understand better how exactly individual images are represented by the TVAE, we present the activation maps corresponding to a single image from an array of classes in Figure 3. We see the representation of each image is still distributed, but most strongly activates in the associated category-selective region.

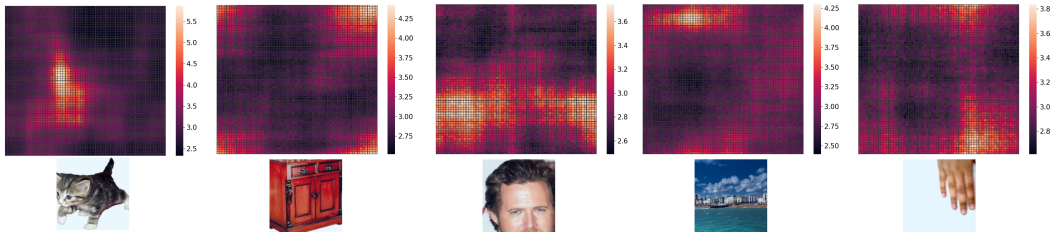

Figure 3: Activations for single images. Left to Right: Animate, Inanimate, Faces, Places, & Hands

Nested Spatial Hierarchy of Categories Following Zhang et al. [57] we additionally measure the selectivity maps of our TVAE model with respect to more abstract categories such as animacy and real-world object size, obtaining such datasets from the Konkle lab database [31, 30]. Specifically, Figure 4 shows Cohen's d maps for animate versus inaminate objects (top), and for big versus small objects (middle), overlayed on the face versus object map (bottom). At the largest scale, we observe an intuitive overlap of spatial maps, specifically inanimate objects, large objects, and the place cluster from Figure 2 all overlap in the top left and right corners of the map. We additionally highlight the maximum activating neurons for three separate input images. We see the image of a red dresser activates a region which is simultaneously selective to places, large, and inanimate objects, echoing the nested spatial hierarchies thought by Grill-Spector & Weiner [16] to exist in the brain.

5 Discussion

In this work we demonstrate the ability of topographic generative models, namely Topographic Variational Autoencoders, to model the emergence of category-selective cortical areas as well as more abstract spatial category hierarchies. We see the model agrees qualitatively with prior work and observations from neuroscience while being founded on a single information theoretic principle.

We note that this study is inherently preliminary and is limited by both the small size of the models used, as well as the feature extraction by a pre-trained convolutional model. It is possible that class-level features and even hierarchical organization are already partially present in some form in the 9216-dimensional feature vectors used as input, and thus it is unclear how much feature extraction the TVAE model is itself learning. Nevertheless, we highlight that this limitation is shared across all existing prior work [34, 57], and that there is nothing fundamentally limiting the TVAE framework from extending to train full deep convolutional networks end-to-end. This is in contrast to the existing related methods which either require a supplementary learning algorithm to guide feature extraction [34], or do not scale to high dimensional inputs [57].

In future work, we intend to explore hierarchical extensions of the TVAE, modeling topographic organization of features at multiple levels of the visual processing pipeline while simultaneously training directly on raw pixel inputs. Such a model would validate the idea of end-to-end unsupervised category-selectivity while simultaneously providing a learned decoder from latent space to image space, opening new avenues for experimentation.

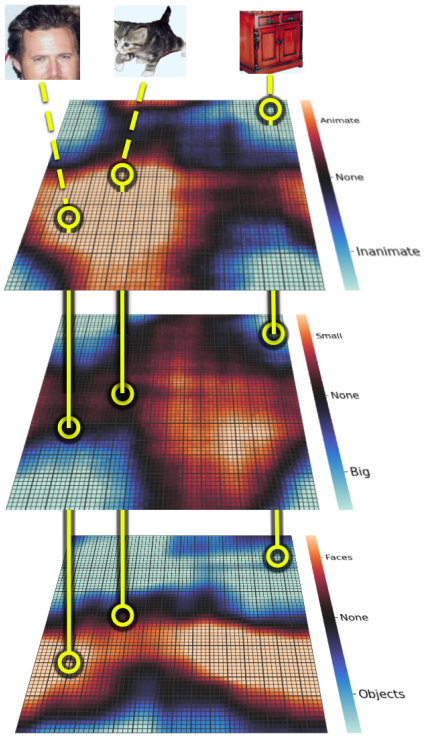

Figure 4: Selectivity maps for abstract categories: Animate vs. Inanimate (top), Small vs. Big (middle), and Faces vs. Objects (bottom). We highlight the maximum activating neurons for the individual images from Figure 3 across all maps, demonstrating their place in the proposed nested spatial hierarchy.

References

[1] Theory-based bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10(7):309–318, 2006. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2006.05.009. URL https://www.sciencedirect.com/science/article/pii/S1364661306001343. Special issue: Probabilistic models of cognition.

[2] Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nature Communications*, 10(1):4934, 2019. doi: 10.1038/s41467-019-12623-6. URL https://doi.org/10.1038/s41467-019-12623-6.

[3] Mahmoud Afifi. 11k hands: gender recognition and biometric identification using a large dataset of hand images. *Multimedia Tools and Applications*, 2019. doi: 10.1007/s11042-019-7424-8. URL https://doi.org/10.1007/s11042-019-7424-8.

[4] Paul L. Aparicio, Elias B. Issa, and James J. DiCarlo. Neurophysiological organization of the middle face patch in macaque inferior temporal cortex. *Journal of Neuroscience*, 36 (50):12729–12745, 2016. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.0237-16.2016. URL https://www.jneurosci.org/content/36/50/12729.

[5] Johannes Ballé, Valero Laparra, and Eero Simoncelli. Density modeling of images using a generalized normalization transformation. 11 2015.

[6] Horace B Barlow et al. Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1(01), 1961.

[7] Anthony J. Bell and Terrence J. Sejnowski. An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, 7(6):1129–1159, 11 1995. ISSN 0899-7667. doi: 10.1162/neco.1995.7.6.1129. URL https://doi.org/10.1162/neco.1995.7.6.1129.

[8] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 67–74. IEEE, 2018.

[9] Julian Caspers, Nicola Palomero-Gallagher, Svenja Caspers, Axel Schleicher, Katrin Amunts, and Karl Zilles. Receptor architecture of visual areas in the face and word-form recognition region of the posterior fusiform gyrus. *Brain structure & function*, 220, 10 2013. doi: 10.1007/s00429-013-0646-z.

[10] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(1):27755, 2016. doi: 10.1038/srep27755. URL https://doi.org/10.1038/srep27755.

[11] Jack Cohen. *Statistical Power Analysis for the behavioral sciences*. L. Erlbaum Associates, 1988.

[12] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3): 287–314, 1994.

[13] Katharina Dobs, Julio Martinez, Alexander J.E. Kell, and Nancy Kanwisher. Brain-like functional specialization emerges spontaneously in deep neural networks. *bioRxiv*, 2021. doi: 10.1101/2021.07.05.451192. URL https://www.biorxiv.org/content/early/2021/07/06/2021.07.05.451192.

[14] Russell A. Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392:598–601, 1998.

[15] David C. Van Essen. A tension-based theory of morphogenesis and compact wiring in the central nervous system. *Nature*, 385(6614):313–318, 1997. doi: 10.1038/385313a0. URL https://doi.org/10.1038/385313a0.

[16] Kalanit Grill-Spector and Kevin S. Weiner. The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8):536–548, 2014. doi: 10.1038/nrn3747. URL https://doi.org/10.1038/nrn3747.

[17] Umut Güçlü and Marcel A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35 (27):10005–10014, 2015. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.5023-14.2015. URL https://www.jneurosci.org/content/35/27/10005.

[18] James Haxby, Maria Gobbini, Maura Furey, Alumit Ishai, Jennifer Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science (New York, N.Y.)*, 293:2425–30, 10 2001. doi: 10.1126/science.1063736.

[19] James Haxby, Jyothi Swaroop Guntupalli, Andrew Connolly, Yaroslav Halchenko, Bryan Conroy, Maria Gobbini, Michael Hanke, and Peter Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72:404–16, 10 2011. doi: 10.1016/j.neuron.2011.08.026.

[20] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[21] Aapo Hyvärinen. Independent component analysis in the presence of gaussian noise by maximizing joint likelihood. *Neurocomputing*, 22(1-3):49–67, 1998.

[22] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[23] Aapo Hyvärinen, Patrik O Hoyer, and Mika Inki. Topographic independent component analysis. *Neural computation*, 13(7):1527–1558, 2001.

[24] Aapo Hyvärinen, Jarmo Hurri, and Patrick O Hoyer. *Natural image statistics: A probabilistic approach to early computational vision.*, volume 39. Springer Science & Business Media, 2009.

[25] Aapo Hyvärinen and Patrik O. Hoyer. A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research*, 41(18): 2413–2423, 2001. ISSN 0042-6989. doi: https://doi.org/10.1016/S0042-6989(01)00114-6. URL `https://www.sciencedirect.com/science/article/pii/S0042698901001146`.

[26] Nancy Kanwisher, Josh McDermott, and Marvin M. Chun. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11): 4302–4311, 1997. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.17-11-04302.1997. URL `https://www.jneurosci.org/content/17/11/4302`.

[27] T. Anderson Keller and Max Welling. Predictive coding with topographic variational autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1086–1091, October 2021.

[28] T. Anderson Keller and Max Welling. Topographic vaes learn equivariant capsules, 2021.

[29] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[30] Talia Konkle and Alfonso Caramazza. Tripartite organization of the ventral stream by animacy and object size. *Journal of Neuroscience*, 33(25):10235–10242, 2013. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.0983-13.2013. URL `https://www.jneurosci.org/content/33/25/10235`.

[31] Talia Konkle and Aude Oliva. A real-world size organization of object responses in occipitotemporal cortex. *Neuron*, 74:1114–24, 06 2012. doi: 10.1016/j.neuron.2012.04.036.

[32] Alexei A Koulakov and Dmitri B Chklovskii. Orientation preference patterns in mammalian visual cortex: a wire length minimization approach. *Neuron*, 29(2):519–527, 2001.

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL `https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf`.

[34] Hyodong Lee, Eshed Margalit, Kamila M. Jozwik, Michael A. Cohen, Nancy Kanwisher, Daniel L. K. Yamins, and James J. DiCarlo. Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. *bioRxiv*, 2020. doi: 10.1101/2020.07.09.185116. URL `https://www.biorxiv.org/content/early/2020/07/10/2020.07.09.185116`.

[35] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[36] Siwei Lyu and Eero P. Simoncelli. Nonlinear image representation using divisive normalization. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008. doi: 10.1109/CVPR.2008.4587821.

[37] Libo Ma and Liqing Zhang. Overcomplete topographic independent component analysis. *Neurocomputing*, 71(10-12):2217–2223, 2008.

[38] Valentina Moro, Cosimo Urgesi, Simone Pernigo, Paola Lanteri, Mariella Pazzaglia, and Salvatore Aglioti. The neural basis of body form and body action agnosia. *Neuron*, 60:235–46, 11 2008. doi: 10.1016/j.neuron.2008.09.022.

[39] Shahin Nasr, Ning Liu, Kathryn J. Devaney, Xiaomin Yue, Reza Rajimehr, Leslie G. Ungerleider, and Roger B. H. Tootell. Scene-selective cortical regions in human and nonhuman primates. *Journal of Neuroscience*, 31(39):13771–13785, 2011. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.2792-11.2011. URL `https://www.jneurosci.org/content/31/39/13771`.

[40] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision research*, 37(23):3311–3325, 1997.

[41] Simon Osindero, Max Welling, and Geoffrey E. Hinton. Topographic Product Models Applied to Natural Scene Statistics. *Neural Computation*, 18(2):381–414, 02 2006. ISSN 0899-7667. doi: 10.1162/089976606775093936. URL `https://doi.org/10.1162/089976606775093936`.

[42] Marius V. Peelen and Paul E. Downing. Selectivity for the human body in the fusiform gyrus. *Journal of Neurophysiology*, 93(1):603–608, 2005. doi: 10.1152/jn.00513.2004. URL `https://doi.org/10.1152/jn.00513.2004`. PMID: 15295012.

[43] Mark A. Pinsk, Kevin DeSimone, Tirin Moore, Charles G. Gross, and Sabine Kastner. Representations of faces and body parts in macaque temporal cortex: A functional mri study. *Proceedings of the National Academy of Sciences*, 102(19):6996–7001, 2005. ISSN 0027-8424. doi: 10.1073/pnas.0502605102. URL `https://www.pnas.org/content/102/19/6996`.

[44] Mark A. Pinsk, Kevin DeSimone, Tirin Moore, Charles G. Gross, and Sabine Kastner. Representations of faces and body parts in macaque temporal cortex: A functional mri study. *Proceedings of the National Academy of Sciences*, 102(19):6996–7001, 2005. ISSN 0027-8424. doi: 10.1073/pnas.0502605102. URL `https://www.pnas.org/content/102/19/6996`.

[45] Rajani Raman and Haruo Hosoya. Convolutional neural networks explain tuning properties of anterior, but not middle, face-processing areas in macaque inferotemporal cortex. *Communications Biology*, 3(1):221, 2020. doi: 10.1038/s42003-020-0945-x. URL `https://doi.org/10.1038/s42003-020-0945-x`.

[46] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *ArXiv*, abs/1401.4082, 2014.

[47] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[48] Shlomo S. Sawilowsky. New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, 8(2):597–599, Nov 2009. doi: 10.22237/jmasm/1257035100. URL `http://dx.doi.org/10.22237/jmasm/1257035100`.

[49] Zeynep M. Saygin, David E. Osher, Kami Koldewyn, Gretchen O Reynolds, John D. E. Gabrieli, and Rebecca Saxe. Anatomical connectivity patterns predict face-selectivity in the fusiform gyrus. *Nature neuroscience*, 15:321 – 327, 2012.

[50] Joshua B. Tenenbaum. Bayesian modeling of human concept learning. In *NIPS*, 1998.

[51] Doris Y. Tsao, Winrich A. Freiwald, Roger B. H. Tootell, and Margaret S. Livingstone. A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670–674, 2006.

[52] Kevin Weiner and Kalanit Grill-Spector. Neural representations of faces and limbs neighbor in human high-level visual cortex: Evidence for a new organization principle. *Psychological research*, 77, 12 2011. doi: 10.1007/s00426-011-0392-x.

[53] Kevin S. Weiner, Golijeh Golarai, Julian Caspers, Miguel R. Chuapoco, Hartmut Mohlberg, Karl Zilles, Katrin Amunts, and Kalanit Grill-Spector. The mid-fusiform sulcus: A landmark identifying both cytoarchitectonic and functional divisions of human ventral temporal cortex. *NeuroImage*, 84:453–465, 2014.

[54] Max Welling, Simon Osindero, and Geoffrey E Hinton. Learning sparse topographic representations with products of student-t distributions. In *Advances in neural information processing systems*, pp. 1383–1390, 2003.

[55] william Clemons. Human body identification and verification dataset. In *Mendeley Data*, December 2018.

[56] Daniel L K Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016. doi: 10.1038/nn.4244. URL `https://doi.org/10.1038/nn.4244`.

[57] Yiyuan Zhang, Ke Zhou, Pinglei Bao, and Jia Liu. Principles governing the topological organization of object selectivities in ventral temporal cortex. *bioRxiv*, 2021. doi: 10.1101/2021.09.15.460220. URL `https://www.biorxiv.org/content/early/2021/09/17/2021.09.15.460220`.

[58] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. *CoRR*, abs/1702.08423, 2017. URL `http://arxiv.org/abs/1702.08423`.

[59] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

A Experimental details

All code for running the experiments in this paper can be found at the following repository:
`https://github.com/AnonSVHRM2021/FFA_Modeling_with_TVAE`

A.1 Training details

Dataset Preprocessing In order to eliminate variability between different datasets, all images were first reshaped to 256×256. A random percentage of the image area (between 8% to 100%) and a random aspect ratio (between $\frac{3}{4}$ and $\frac{4}{3}$) were then chosen, and each image was then cropped according to these values. Finally, the crops were resized to the final shape of 224×224. All images were then normalized by the mean $[0.48300076, 0.45126104, 0.3998704]$ and standard deviation $[0.26990137, 0.26078254, 0.27288908]$.

TDANN Hyperparameters The TDANN model was trained with stochastic gradient descent, a learning rate of 1×10^{-3}, standard momentum of 0.9, and a batch size of 128 for 10 epochs. Explicitly, the loss function was given by a sum of the classification cross entropy loss, the spatial correlation losses for both layers FC6 and FC7, and weight decay of 5×10^{-4}. A fixed weight of $10 \times \frac{1}{4096^2}$ was multiplied by the spatial correlation loss before backpropagating as this was found necessary to qualitatively match the results from Lee et al. [34]. Contrary to the original TDANN work, we did not randomly initialize the locations of the neurons, and instead spaced them evenly on a grid of the same size. We found the spatial correlation loss to still function equally well in this setting, and detail our implementation in Section A.2 below.

TVAE Hyperparameters The TVAE was trained with stochastic gradient descent, a learning rate of 1×10^{-5}, standard momentum of 0.9, and a batch size of 128 for 30 epochs. The global topology was set to a single 2D torus (i.e. a 2D grid with circular boundary conditions), and the local topology was set to sum of local regions of size 25×25, i.e. the kernel used to convolve over $\mathbf{u}$ was of size 25×25 and contained all 1's. The μ parameter was initialized to 40, and trained simultanously with the remainder of the model parameters.

A.2 Spatial Correlation Loss of TDANN

The exact form of the spatial correlation loss used for training the TDANN in this paper is given as:

$$\text{SpatialCorrelationLoss}(\mathbf{z}) = \sum_{i}^{n} \sum_{j \neq i}^{n} \left| C_{ij}(\mathbf{z}) - \frac{1}{D_{ij} + 1} \right| \tag{5}$$

where $\mathbf{z}$ an n-dimensional vector of activations, C is the normalized cross correlation matrix (e.g. a matrix of Pearson correlation coefficients), and D is a matrix containing the 'cortical distances' in millimeters between all pairs of neurons i and j. In this work, we defined all neurons to be equally spaced in a 2-D grid of 10mm $\times$ 10mm. This resulted in a horizontal and vertical spacing between neurons of 0.15625mm and a diagonal spacing of 0.22097087mm. Unlike the TVAE, the TDANN grid was not defined to have circular boundary conditions in order to match the original model.

B Additional Results

B.1 Robustness to Face Dataset Choice

To investigate the robustness of face selectivity across different face datasets, and ensure the observed clusters were not a dataset dependant phenomenon, selectivity maps for four different face datasets were computed for both models. Explicitly, the four datasets included: a 25,000 subset of VGGface2 [8], 10137 images from UTKface [58], 24684 images from CelebA [35], and the Labled Faces in the Wild [20] dataset upon which the model was trained. The resulting selectivity maps can be seen to be highly overlapping in Figure 5 below.

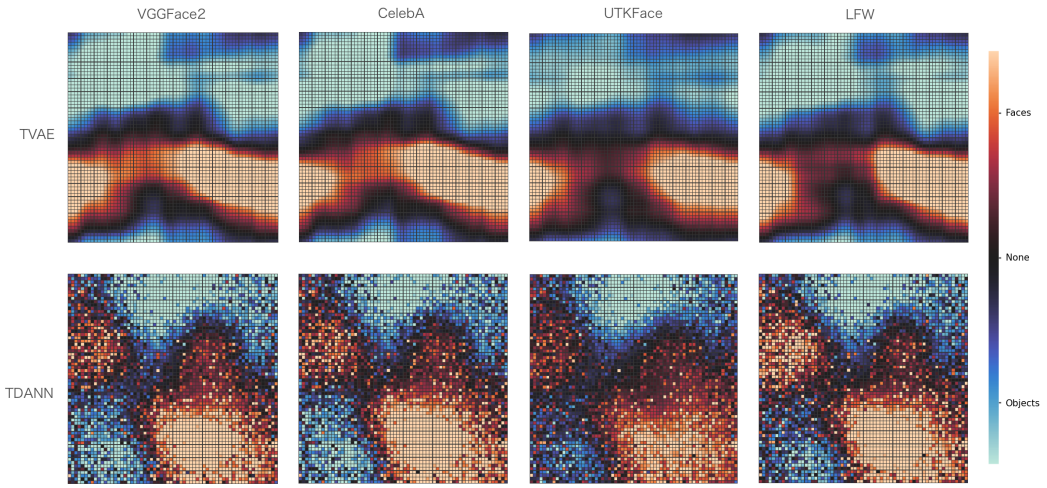

Figure 5: Face vs. Object selectivity maps for four different face datasets. We see that for both the TVAE and TDANN the relative locations and sizes of the face and object selective clusters are relatively stable despite the differences in the underlying test datasets used.

B.2 Impact of Topographic Prior on Data Likelihood

To ensure the TVAE was learning to be a reasonable generative model of the input (the 9216-dimensional feature vector from Alexnet), we measured the likelihood of a validation dataset under both our TVAE and a baseline non-topographic VAE of the same architecture. We found the baseline VAE to achieve roughly 3.4 bits per dimension (BPD) while the Topographic VAE achieved roughly 3.6 BPD in the same number of iterations. This appears consistent with the intuition and observation that the topographic prior does not prevent learning, but rather acts as a regularization term on the model. In future work we hope to quantify this regularization effect more precisely and determine in which situations it may be beneficial for generalization or computational efficiency.