Are Topographic Deep Convolutional Neural Networks Better Models of the Ventral Visual Stream?

Kamila Maria Jozwik (kmjozwik@mit.edu)
University of Cambridge and McGovern Institute for Brain Research, Center for Brains, Minds and Machines at Massachusetts Institute of Technology, 43 Vassar St
Cambridge, MA 02139 United States

Hyodong Lee (hyo@mit.edu)
Department of Electrical Engineering and Computer Science at MIT

Nancy Kanwisher (ngk@mit.edu)
McGovern Institute for Brain Research and Department of Brain and Cognitive Sciences at MIT
and
James J. DiCarlo (dicarlo@mit.edu)
McGovern Institute for Brain Research and Department of Brain and Cognitive Sciences at MIT

Abstract:

Neural computations along the ventral visual stream, -- which culminates in the inferior temporal (IT) cortex -- enable humans and monkeys to recognize objects quickly. Primate IT is organized topographically: nearby neurons have similar response properties. Yet the best models of the ventral visual stream - deep artificial neural networks (ANNs) – have "IT" layers that lack topography. We built Topographic Deep ANNs (TDANNs) by incorporating a proxy wiring cost alongside the standard ImageNet categorization cost in the two "IT-like" layers of AlexNet (Lee et al., 2018), by specifying that "neurons" that have similar response properties should be physically close to each other. This cost both induced topographic structure and altered tuning characteristics of model IT neurons. We presented 2560 naturalistic images to monkeys and to ANNs. We found that, relative to the base (nontopographic) model, the "neurons" in the "IT" layer of some of the TDANN models matched actual IT neurons slightly better, and the dimensionality of the TDANN "IT" neural population was much closer to that of the measured monkey IT neural population. We also found that, while TDANNs did not show a statistically significant better match to human object discrimination behavior, detailed analysis suggests a trend in that direction. Taken together, TDANNs may better capture properties of IT cortex and wiring costs might be the cause of topographic organization in primate IT.

Keywords: object vision; IT; ANN; topography; dimensionality

Introduction

Humans and monkeys recognize objects with ease, thanks to the neural computations conducted in the ventral visual pathway, which culminates in the inferior temporal (IT) cortex. Primate IT has a topographical organization: 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). In recent years, deep artificial neural networks (ANNs) have revolutionized computer vision, and have also been shown to be the best models of the ventral stream in that they account for monkey V4 and IT responses far better than other models (Yamins et al., 2014). Yet ANNs lack topography in "IT" layers, which limits their suitability as models of the ventral stream.

To determine whether topography might be important to the functioning of the ventral stream, we built Topographic Deep ANNs (TDANNs) by incorporating a proxy wiring cost alongside the standard ImageNet categorization cost in AlexNet (Lee et al., 2018). The proxy wiring cost is implemented in the two penultimate layers of TDANNs by specifying that "neurons" placed in proximity on an artificial tissue map should have similar response properties.

We used four variants of TDANNs with increasing proxy wiring costs to see how topography affects a model's functional fidelity with primate IT (i.e., the model's ability to predict median single IT site response measures), its representational dimensionality (measured by participation ratio) relative to IT, and its predictivity of human behavior.

Methods

Topographic Deep ANNs

The architecture of TDANNs was based on AlexNet (Krizhevsky et al, 2012): five convolutional layers and two fully-connected layers (fc6 and fc7). The topographic constraint was applied to "IT" layers fc6 and fc7, as these layers showed the highest predictivity of IT representations. In addition to training the models to classify 1.2 million images into 1000 categories using the ImageNet dataset, we added a wiring cost

constraint, as follows. We assigned a random position for each of the model units in "IT" layers (fc6 and fc7) on a two-dimensional artificial tissue map before training, simulating cortical maps in monkey IT (Figure 1). The size of the tissue map was 10mm x 10mm, which corresponds to the processing of 8° of visual angle at the center of gaze. We derived the cost function as a local correlation rule from monkey IT neural recordings, where the pairwise response correlation of neurons are high for nearby pairs and decrease as a function of their cortical distance. Along with image classification task, our models were trained to satisfy this local correlation rule. We tuned the strength of the wiring cost constraint (relative to the image classification cost) by manipulating a 'loss weight' parameter. We used loss weight 0 (base model), and increasingly stronger loss 10, 20, 30, and 40 and trained each model three times with different initialization conditions. All results presented here are based on three randomly-initialized models for each parameter setting.

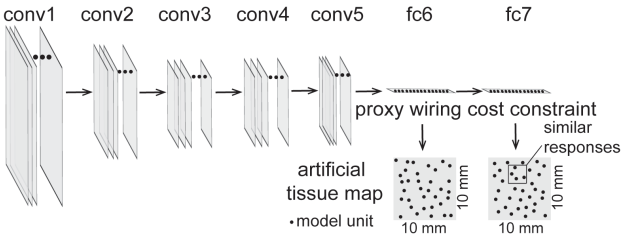

Figure 1. Topographic Deep ANNs (TDANNs) architecture. The proxy wiring cost is implemented in the two penultimate layers of TDANNs by specifying that "neurons" placed in proximity on an artificial tissue map should have correlated response properties. The fall-off of response correlation vs distance varied in a range based on monkey IT neural recordings. During training, weights were tuned to maximize object classification performance while adhering to the local correlation function.

Monkey Recordings

Neural responses to 2560 stimuli were recorded from 168 IT neurons (Majaj et al., 2015). The images were from eight object categories (animals, boats, cars, chairs, faces, fruits, planes, tables). An image was generated by pasting an object on a naturalistic background with random position, pose, and size of an object. Recordings were acquired from two monkeys, each implanted with two Utah arrays in IT. Images were shown for 100ms at 8° visual angle. Neural firing rate was averaged in the window between 70 ms and 170 ms. We measured the neuronal predictivity of ANN units using cross-validated partial least squares regression (PLS) with 25 principal components to map ANN units to each neural site. The neuronal predictivity was expressed as Pearson correlation between the ANN predictions and measured neural responses.

Dimensionality Estimates

We used the participation ratio as our dimensionality measure (Gao et al., 2017). We randomly subsampled the number of units from ANNs that we had neurons in our monkey recordings (N = 168), multiple times, and added Poisson noise comparable to the noise in monkey recordings to ANN units before computing the participation ratio. We obtained similar results with four other unit subsampling methods.

Behavioral Measure

Human behavioral data were acquired for 240 images from 24 categories (Rajalingham et al., 2018). A sample image was presented for 100 ms followed by two choices: an object from the initial image with potential variations in position, pose, and size, and another object from one of the remaining 23 categories. Behavioral performance was represented in a 240 image x 24 category matrix showing accuracy for each combination of the sample image and discrimination category. Similarly, we trained a classifier on the base (nontopographic) model's and TDANN's activations to mimic the task performed by humans, and created corresponding 240x24 matrices for the models. We then compared each model to human behavior by correlating these matrices (Pearson correlation).

Results

Neural Predictivity of TDANNs

We used four variants of TDANNs with increasing proxy wiring costs to ask how topography affects a model's functional fidelity with primate IT, i.e. the model's ability to explain and predict IT spiking response measures. We found that the functional fidelity of TDANNs ("IT" layer, fc7) was slightly higher than the non-topographic base model (Figure 2). Neural predictivity was the highest for loss 20 and loss 30. However, neural predictivity of the TDANN started dropping at loss 40, suggesting that optimal proxy wiring costs exists that maximizes neural predictivity.

Dimensionality of TDANNs

The number of linearly independent coding dimensions (aka "dimensionality") is an important characterization of each ventral stream area, yet it is still unclear if conventional deep ANNs match the brain's dimensionality. We expected that the proxy wiring cost constraint might reduce the dimensionality, however, it could be that the dimensionality is reduced too much or

not enough relative to monkey IT. We found that the base model dimensionality (measured by participation ratio) is 3.71 higher than measured monkey IT dimensionality. In contrast, TDANNs were closer to empirical estimates of the (subsampled) dimensionality of IT (Figure 3). To compare the dimensionality of monkey IT and the models, we subsampled the same number of features from models as we have neurons in monkey recordings (N = 168), and added Poisson noise comparable to the noise in monkey recordings. We do not claim that we have estimated the full dimensionality of IT (as we have a very limited image set and a limited neural sample). We simply claim that the same dimensionality analyses applied to IT and to the models showed that TDANNs were more similar to IT than the base model.

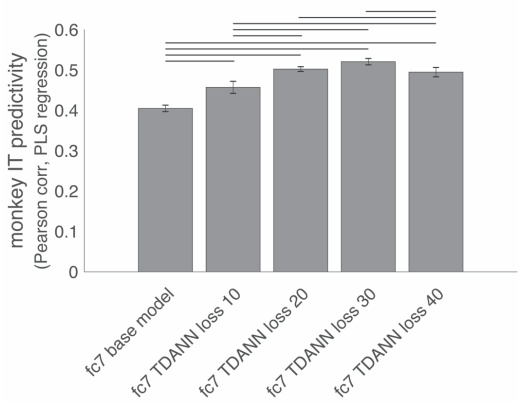

Figure 2. Effect of the proxy wiring cost constraint on neural IT predictivity. Median raw IT site response predictivity using PLS regression with cross-validation (Pearson correlation) averaged over 3 randomly initialized models for each model class ("IT" layer, fc7). IT predictivity of TDANNs was significantly better than the non-topographic base model. Error bars represent standard deviation across 3 randomly initialized models. Significant differences between model IT predictivity are indicated by horizontal lines (two-sample t-test, $p<0.05$).

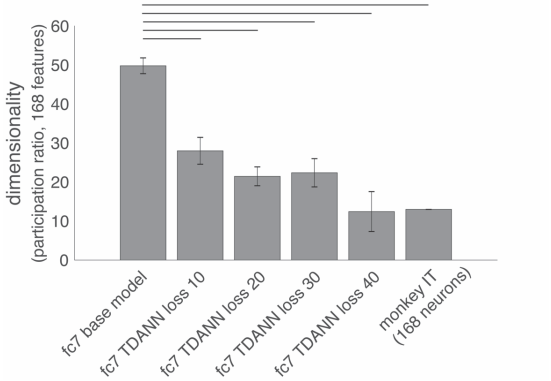

Figure 3. Effect of the proxy wiring cost constraint on dimensionality. Dimensionality was calculated for a set of the same 2560 images using participation ratio (estimated after subsampling of features and added Poisson noise) in "IT" layer (fc7) in the base model and TDANNs with varying amount of topographical loss. The dimensionality of TDANNs is comparable to measured monkey IT dimensionality. Error bars represent standard deviation across 3 randomly initialized models. Significant differences between the dimensionality of the models, and models' dimensionality and monkey IT are indicated by horizontal lines (two-sample t-test, $p<0.05$).

Behavioral predictivity of TDANNs

A good model of the brain should also be able to predict human and monkey object discrimination behavior. As humans and monkeys have very similar behavioral patterns on the task we evaluated (Rajalingham et al., 2018), we only looked at human behavioral data. Human behavioral data was acquired for 240 images, where subjects indicated what object was presented on an image selecting one of the two choices (presented after the sample image). To mimic the task performed by humans, we trained a classifier on the base model and TDANNs activations and compared behavioral matrices of humans and ANNs. TDANNs with loss 10 and 20 were able to predict human behavior at a similar level to the base model (Figure 4A). However, larger proxy wiring cost started hurting behavioral predictivity, pointing again that there is an optimal amount of proxy wiring constraint. In our case, that seems to be loss 20, as it improved neuronal predictivity, brought dimensionality closer to that observed in IT, and did not affect behavioral predictivity. It could be that our behavioral measure is not sensitive enough to detect differences between the base model and TDANN loss 20, so to increase that sensitivity we need to calculate a behavioral score using only the subset of stimuli for which the base model and TDANNs give the most different category predictions. We selected the stimuli with the largest absolute difference (in either direction) between the base model and the TDANN topo loss 20 (residuals) for category predictions that were consistently different across 3 randomly initialized models (Figure 4B, N = 18) and recomputed behavioral scores for these stimuli. We stress that we looked for predictions that are different between the base model and the TDANN, not the predictions where the TDANN was better. The TDANN loss 20 seemed to predict the responses to this subset of stimuli (Figure 4C) slightly better than the base model.

Discussion

The proxy wiring cost added to the base model altered the tuning characteristics of the model IT neurons. We

found that, relative to the base model, the "neurons" in the "IT" layer of some of the TDANN models were a slightly closer match to actual IT neurons, and that the dimensionality of the TDANN "IT" neural population was closer to that of the measured monkey IT neural population. While these more brain-like ventral stream models did not show a statistically significantly better match to human object discrimination behavior, detailed analysis suggests a trend in that direction.

It would be interesting to explore the similarity of representational dimensions in TDANNs and in monkey IT using a larger stimulus set and larger neural sample. It would be also interesting to extend our approach to all layers of the base model by incorporating proxy wiring cost at all layers of the model and evaluate its ability to predict neural responses across the ventral stream.

In summary, our results suggest that TDANNs may better capture properties of IT cortex and the wiring costs might be the cause of topographic organization in primate IT. Broadly, these results also show that using brain observations not already in place in deep ANNs (here topography) can lead to improved models of the brain.

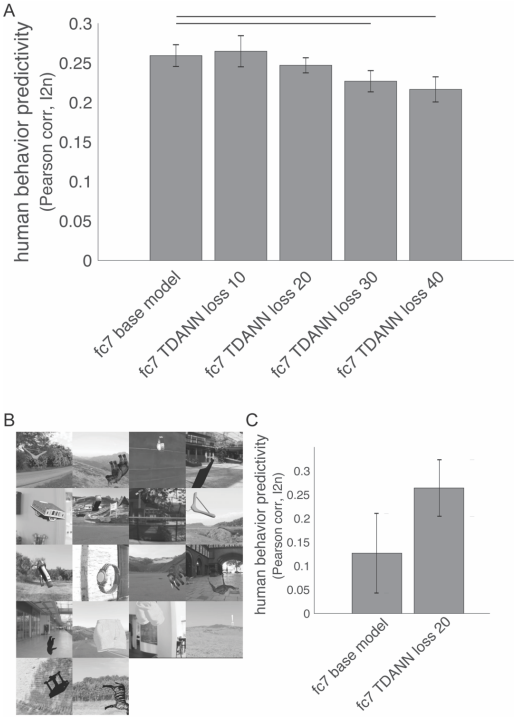

Figure 4. Effect of the proxy wiring cost constraint on behavioral predictivity. **A**. Humans indicated which object was presented in an image, selecting one of the two choices (presented after the sample image) for each of 240 sample images. We trained a classifier on the base model and TDANNs activations to mimic this task and compared behavioral matrices of humans and ANNs (Pearson correlation between behavioral and ANN matrices, image-by-image patterns, broken down by the object choice alternatives - I2n). **B**. Selected stimuli with the largest absolute difference between the base model and the TDANN loss 20 for category predictions (residuals, N = 18). **C**. Behavioral predictivity for selected stimuli. Error bars represent standard deviation across 3 randomly initialized models. Significant differences between model dimensionality are indicated by horizontal lines (two-sample t-test, $p<0.05$).

Acknowledgments

This work was funded by the Sir Henry Wellcome Postdoctoral Fellowship (206521/Z/17/Z) to KMJ, the National Institutes of Health grant (DP1HD091947) to NK, the Simons Foundation grants (SCGB [325500, 542965]) to JJD, and the Center for Brains, Minds and Machines (NSF STC award CCF-1231216).

References

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

Krizhevsky, A., Sutskever, I., Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*. pp. 1097–1105.

Lee, H., DiCarlo, J. J. (2018). Topographic Deep Artificial Neural Networks (TDANNs) predict face selectivity topography in primate inferior temporal (IT) cortex. *Conference on Cognitive Computational Neuroscience.*

Majaj, N. J., Hong, H., Solomon, E.A., and DiCarlo, J.J. (2015). Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. *Journal of Neuroscience*, 35(39):13402–13418

Rajalingham, R., Issa. E. B., Bashivan, P., Kar K., Schmidt, K., and DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, pages 0388–18.

Yamins, D. L. K., 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 of the United States of America*, 111, 8619–8624.