Explaining Scene-selective Visual Areas Using Task-specific Deep Neural Network Representations

Kshitij Dwivedi[1] **(kshitijdwivedi93@gmail.com), Michael F. Bonner**[2] **(mfbonner@jhu.edu),**
Gemma Roig[1] **(gemmar@mit.edu)**
[1]Information Systems Technology and Design, Singapore University Technology and Design, Singapore
[2]Department of Cognitive Science, Johns Hopkins University, Baltimore, MD

Abstract

Deep neural networks (DNNs) are currently the models that account for higher variance of the responses from the human visual cortex. In this work, we aim to explore the power of DNNs as a tool to gain insights into functions of visual brain areas. Particulary, we focus on scene selective visual areas. We use a set of DNNs trained to perform different visual tasks, comprising 2D, 3D and semantic aspects of scene perception, to explain fMRI responses in early visual cortex (EVC) and scene selective visual areas (OPA, PPA). We find that EVC representation is more similar to early layers of all DNNs and deeper layers of 2D-task DNNs. OPA representation is more similar to deeper layers of 3D DNNs, whereas PPA representation to deeper layers of semantic DNNs. We extend our study to performing searchlight analysis using such task specific DNN representations to generate task-specificity maps of visual cortex, and visualize their overlap with existing ROI parcels. Our findings suggest that DNNs trained on a diverse set of visual task can be used to gain insights into functions of visual cortex. Our approach has the potential to be applied beyond visual areas.

Keywords: Deep neural networks; fMRI; PPA; OPA.

Introduction

Deep neural networks (DNNs) are currently state-of-the-art models for explaining cortical responses in the visual cortex (D. L. K. Yamins et al., 2014; Cichy, Khosla, Pantazis, & Oliva, 2017; Tacchetti, Isik, & Poggio, 2016; Khaligh-Razavi & Kriegeskorte, 2014; Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; D. L. Yamins & DiCarlo, 2016) . DNNs trained on object classification task have been shown to explain human and monkey cortical responses in the inferior temporal cortex (IT) area, which is known to play a role in object recognition. Further, it has been revealed that unsupervised models are unable to explain the IT responses as well as the supervised models (Khaligh-Razavi & Kriegeskorte, 2014). This emphasizes the importance of using a model which has been optimized on a related task to the brain region under study. In this work, inspired by (Khaligh-Razavi & Kriegeskorte, 2014), we investigate for the first time if we can reveal functions of brain regions using DNNs trained on different aspects of visual perception. Our preliminary results can be found in (Dwivedi & Roig, 2018). To achieve this, we use 20 DNNs trained on 2D, 3D and semantic tasks from Taskonomy (Zamir et al., 2018) dataset and fMRI dataset collected on indoor scene images from (Bonner & Epstein, 2017) (Fig. 1a). We rank which type of DNNs explain the responses of ROIs better than others with Representational Similarity Analysis (Fig. 1b). To extend our investigations beyond the selected ROIs we perform searchlight analysis (Fig. 1c) to generate a whole brain task specificity plot using the task-specific DNNs.

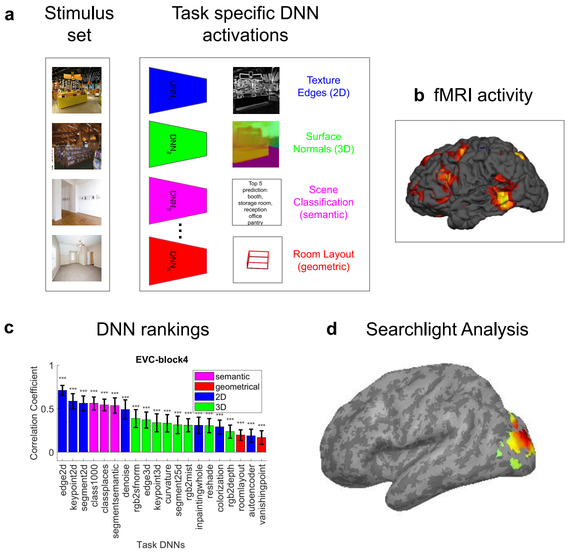
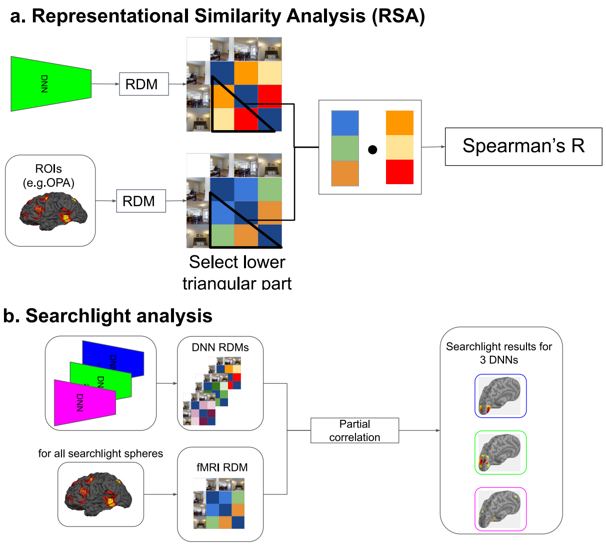


Figure 1: **Overview of our approach:** We use **a)** activations of DNN trained on 2D, 3D, semantic and geometric tasks, and **b)** fMRI responses, both on stimuli set from (Bonner & Epstein, 2017) to find **c)** rankings of DNNs for each brain ROI using RSA. **d)** We highlight task specificity in visual cortex using searchlight.

Our results from ROI analysis suggest that EVC responses are better explained by 2D-tasks trained DNNs, OPA responses by 3D DNNs, and PPA responses by DNNs trained on semantic tasks. While in this work, we only focus on scene images and scene-selective visual areas, our approach is general and can be extended to other stimuli types and other brain areas, as suggested by our searchlight results.

Methods

In this section, after we detail how we select task-specific activations from the DNNs, we describe the fMRI data and the analysis performed.



a. Representational Similarity Analysis (RSA)

b. Searchlight analysis

Figure 2: **Methods used: a)** representational similarity analysis for DNNs' rankings, and **b)** Searchlight Analysis to visualize task-specificity of the whole brain.

Task-specific DNNs' Representations

We use task-specific DNNs trained on Taskonomy dataset (Zamir et al., 2018)[1]. Taskonomy dataset is a large-scale image dataset containing 4 million images with annotations and pretrained DNN models available for 26 vision related tasks. The tasks included in this dataset cover most common computer vision tasks related to 2D (*e.g.* edge detection), 3D (*e.g.* depth estimation), and semantics (*e.g.* scene and object classification). The DNN models trained on different tasks from Taskonomy dataset share a common encoder architecture and have a task-specific decoder, which varies according to the output structure of each task. The encoder is a fully convolutional ResNet-50(He, Zhang, Ren, & Sun, 2016) consisting of 4 residual blocks each containing multiple convolutional layers, without any pooling layer. For selecting task-specific representation we select the activation from the last layer of the final encoder's block (block-4).

fMRI data

We use the fMRI data from (Bonner & Epstein, 2017). The stimuli images used for analysis consist of 50 indoor environments images. The subjects' fMRI responses were obtained while they performed a category-recognition task (bathroom or not). For RSA analysis we use the precomputed subject averaged RDMs of the EVC, PPA and OPA. For more details about the fMRI data please refer to (Bonner & Epstein, 2017)

Representational Similarity Analysis (RSA)

RSA is used to compare the information encoded in brain responses with a computational or behavioral model by computing the correlation of the corresponding representation dissimilarity matrices (RDMs).

Representation Dissimilarity Matrix (RDM). The RDM for a dataset is constructed by computing dissimilarities of all possible pairs of stimulus images. For fMRI data, the RDMs are computed by comparing the pairwise fMRI responses, while for DNNs the RDMs are computed by comparing the pairwise layer activations for each image pair in the dataset. The dissimilarity metric used in this work is $1-\rho$, in which ρ is the Pearsons correlation coefficient, as illustrated in Fig. 2a.

Statistical Analysis. We use RSA toolbox (Nili et al., 2014) to compute RDM correlations, as well as their corresponding p-values and standard deviation using stimulus-label randomization test. For determining which RDM better explains the neural RDMs, we perform a bootstrap test. The number of bootstrap iterations for all the analysis was set to 5000.

Searchlight Analysis.

We perform a searchlight analysis to visualize the overlap of the DNNs' searchlight results with predefined ROIs, and to reveal which other regions have similar representations similar to the task-specific DNNs selected. We also generate a color coded task-specificity map to visualize which brain area show higher correlation with which task type (Fig. 2b).

Visualizing Overlap with Existing ROIs. We know from the ROI analyses which DNNs are the best fit for each ROI. We next used exploratory searchlight analyses to visualize the fit of these models throughout the rest of the brain. To do this, we selected the DNNs that showed highest correlation with EVC, PPA, and OPA. We then performed a searchlight analysis using partial correlation of one DNN RDM with RDMs of each searchlight sphere while selecting the other 2 DNN RDMs as the control variables. Using this analysis we aimed to visualize the cortical regions that have representations specific to 2D, 3D and semantic DNNs.

Task-specificity Map For this analysis, we select 14 DNNs from Taskonomy dataset. We next perform a searchlight analysis using partial correlation of RDM of deeper layer (block 4) of each DNN with fMRI RDMs of each searchlight sphere while selecting the RDM of early layer of DNN as the control variable. The assumption behind using RDMs of early layer as control variable is that early layers have EVC-like representation and therefore by partialling out its effect, only the task-specific signal of each RDM is left behind. After obtaining the searchlight results for each RDM we find which type of task shows the maximum correlation at each searchlight sphere and color code it according to task type.

[1] pretrained models downloaded from https://github.com/StanfordVL/taskonomy/tree/master/taskbank

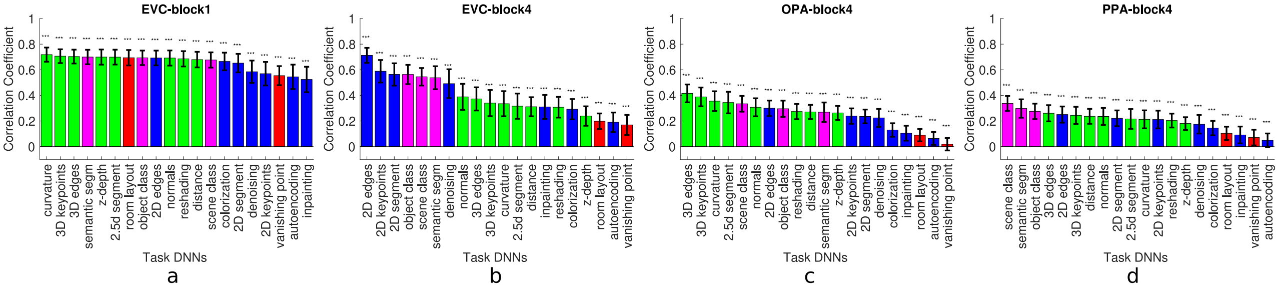


Figure 3: RSA of Brain ROIs with task-specific DNNs.

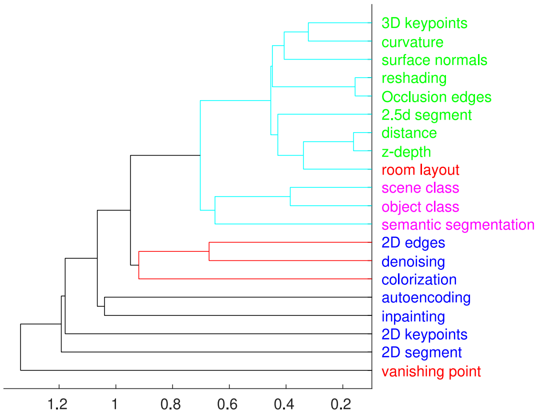


Figure 4: Hierarchical clustering on DNN RDMs.

Variance Partitioning Analysis. Variance partitioning method is used to determine the unique and shared contribution of individual models when considered in conjunction with the other models. We describe the analysis by considering the case of OPA predicted by 3D, 2D, and semantic DNNs. First, the off-diagonal elements of the OPA RDM are assigned as the dependent variable (predictand). Then, the off-diagonal elements of 3 DNN RDMs are selected as the independent variable. Then, we perform seven multiple regression analysis: one with all three independent variables as predictors, three with three different possible combinations of two independent variables as predictors, and three with individual independent variables as the predictors. Then, by comparing the explained variance (r^2) of a model used alone with the explained variance when it was used with other models, the amount of unique and shared variance between different predictors can be inferred. For the other variance partitioning analysis, the predictors and predictands were modified accordingly, and we followed the same steps.

Results

We first report the result of DNN rankings for EVC, OPA and PPA. We next report the results of searchlight analysis.

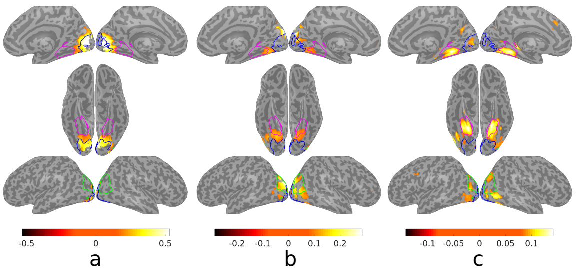


Figure 5: Searchlight analysis of task-specific DNNs.

DNN Rankings for ROIs

In Fig. 3a, and b, we show the correlation of EVC RDMs with block 1 and block 4 RDMs of 20 DNNs, respectively. We observe that EVC shows high correlation with early layers of all DNNs and deeper layers of 2D DNNs. In Fig. 3c and d, we show the correlation of OPA and PPA with block 4 RDMs of 20 DNNs. We observe that OPA shows high correlation with deeper layers of 3D DNNs, and PPA shows high correlation with those of semantic DNNs. The results suggest that EVC is biased towards 2D scene representations, the best fit for OPA are 3D scene representations and the preference for PPA are semantic representations of the indoor scenes.

How DNN representations are clustered? To observe if the DNN representations are clustered to corresponding task type we perform hierarchical clustering similar to (Dwivedi & Roig, 2019) on block 4 RDMs of each task. As shown in Fig. 4, the tasks are indeed clustered into different clusters of 2D,3D and semantic tasks.

Visualizing the ROI overlap with DNN searchlight

The top 95th-percentile correlation for the searchlight results are displayed in Fig. 5. As expected, we observe that there is high overlap between the ROIs and searchlight result of DNNs that showed the highest correlation with corresponding ROI. We observe that beyond the above mentioned ROIs, other visual areas are also highlighted as showing 2D, 3D, and semantic functionality.

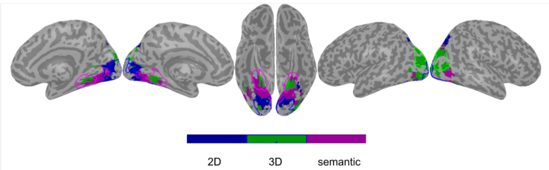


Figure 6: **Task specificity map.** We use blue for 2D, green for 3D, and magenta for semantic tasks.

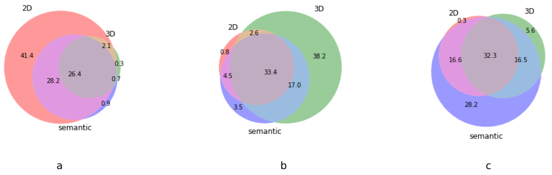


Figure 7: **Variance partitioning results** for **a)** EVC, **b)** OPA, and **c)** PPA.

Task-specificity map

We visualize the color coded map obtained using Task-specificity analysis in Fig. 6. We observe that 2D representations are formed first in the early visual cortex hierarchy, followed by semantic and 3D in deeper visual cortex.

Variance Partitioning Results

We combined RSA with variance partitioning (Nimon & Oswald, 2013) analysis to investigate how uniquely does the most correlated 2D, 3D, and semantic trained DNNs explain the responses of EVC, OPA and PPA, respectively. In variance partitioning approach, we can divide the unique and shared variance contributed by all of its predictors. As shown in Fig. 7a, 2D DNN explains 41.4% of EVC variance uniquely, 3D DNN explains 38.25% of OPA variance uniquely and semantic DNN explains 28.17% of PPA variance uniquely.

Discussion

In this work, we sought to find functions of the different areas in visual cortex using DNNs optimized to perform tasks on different aspects of visual perception. We demonstrated using DNNs trained on different aspects of scene perception that representation of EVC is similar to 2D DNNs, OPA is similar to 3D DNNs and and PPA to semantic DNNs. Our results suggest that OPA encodes information about 3D structure of the scene while PPA encodes semantic information about the scene. Our results are consistent with recent neuroimaging findings (Henriksson, Mur, & Kriegeskorte, 2019). In our work we were able to achieve such findings using a purely computational approach. We believe our method based on using task-specific trained DNNs opens new horizon for investigating functions of visual cortex and beyond.

Acknowledgments

K.D. was funded by SUTD Presidents Graduate Fellowship.

References

Bonner, M. F., & Epstein, R. A. (2017). Coding of navigational affordances in the human visual system. *Proceedings of the National Academy of Sciences*, *114*(18), 4793–4798.

Cichy, R. M., Khosla, A., Pantazis, D., & Oliva, A. (2017). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, *153*, 346–358.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(June), 1–13.

Dwivedi, K., & Roig, G. (2018). Navigational affordance cortical responses explained by scene-parsing model. In *Proc. of european conference on computer vision workshops.*

Dwivedi, K., & Roig, G. (2019). Representation similarity analysis for efficient task taxonomy & transfer learning. In *Proc. of IEEE computer vision and pattern recognition.*

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Henriksson, L., Mur, M., & Kriegeskorte, N. (2019). Rapid invariant encoding of scene layout in human opa. *Neuron*.

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, *10*(11).

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS computational biology*, *10*(4), e1003553.

Nimon, K. F., & Oswald, F. L. (2013). Understanding the results of multiple linear regression: Beyond standardized regression coefficients. *Organizational Research Methods*, *16*(4), 650–674.

Tacchetti, A., Isik, L., & Poggio, T. (2016). Invariant recognition drives neural representations of action sequences. , 1–23.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356.

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*, *111*(23), 8619–8624.

Zamir, A. R., Sax, A., Shen, W. B., Guibas, L. J., Malik, J., & Savarese, S. (2018). Taskonomy: Disentangling task transfer learning. In *Proc. of IEEE conference on computer vision and pattern recognition.*