Navigational affordance cortical responses explained by scene-parsing model

Kshitij Dwivedi[0000−0001−6442−7140] and Gemma Roig[0000−0002−6439−8076]

Singapore University of Technology and Design
kshitij_dwivedi@mymail.sutd.edu.sg, gemma_roig@sutd.edu.sg

Abstract. Deep Neural Networks (DNNs) are the leading models for explaining the population responses of neurons in the visual cortex. Recent studies show that responses of some task-specific brain regions can also be explained by a DNN trained for classification. In this work, we propose that responses of task-specific brain regions are better explained by DNNs trained on a similar task. We first show that responses of scene selective visual areas like parahippocampal place area (PPA) and Occipital Place Area (OPA) are better explained by a DNN trained for scene classification than one trained for object classification. Next, we consider a particular case of OPA which has been shown to encode navigational affordances. We argue that a scene parsing task, which predicts the class of each pixel in the scene is more related to navigational affordances than scene classification. Our results show that the responses in OPA are better explained by the scene parsing model than the scene classification model.

Keywords: Deep neural networks, Representational Similarity Analysis, Occipital Place Area, Neural encoding

1 Introduction

In recent works, DNNs have been shown to explain the responses of the human visual cortex. In several recent works [16, 12, 15, 8, 3, 17, 7], it has also been demonstrated that responses in visual cortex during perception and neural network activations of different layers of a DNN are highly correlated. Areas from higher visual cortex have been shown to be more correlated with the deeper layers [17, 7] and areas of lower visual cortex have been shown to be highly correlated with the initial layers of the DNN [7].

DNNs have also been used to explain responses of brain areas associated with specific visual tasks. In a recent work by Bonner and Epstein [2], they explore the possibility of explaining the navigational affordances with the functional Magnetic Resonance Imaging (fMRI) activation patterns in the OPA. They show that fMRI responses in OPA are associated with the navigational affordances of the scenes. In a subsequent work [1], they explore if layers of a DNN trained for scene classification can serve as a computational model of navigational affordance related responses in the OPA.

In this work, we investigate if the responses of a brain area performing specific visual tasks are explained better by a computational model performing a similar task rather than a generic classification model. In scene parsing task, the aim is to predict the class labels for all the locations in the image. The output of the scene parsing task can label the free space available for navigation and the obstacles present in the scene. Thus, we argue that a scene parsing model will explain the spatial scene property like navigational affordances, and hence, the OPA responses better than a scene classification model. We investigate this in the following steps:

1. We investigate if a scene classification model better explains the responses in scene-selective areas PPA [6] and OPA [5] better than an object classification model.
2. We investigate if a model trained on a potential task similar to navigational affordance such as scene parsing explains OPA responses and behavioral model for navigational affordances better than a scene classification model.
3. We perform a detailed comparative analysis of the specific class labels with the OPA and PPA responses to gain more insights into the functionality of these areas.

The results from all the experiments above suggest that due to task similarity, scene parsing model explains cortical responses to the navigational affordance in scenes better than a classification model. Our results reinforce the use of models trained to perform similar tasks for explaining responses of task-specific areas in the visual cortex.

2 Methods

In the first section, we describe RSA [9] which is a standard method to compare the correlation of computational and behavioral models with human brain activity. In the second section, we briefly describe the dataset we used in this work and then in the following sections we provide the details of the DNN models used for analysis.

2.1 Representation 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). In the case of comparison with DNNs, we compute the correlation of RDMs of the brain responses with the RDM of layer activations of the DNNs.

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 fMRI responses while

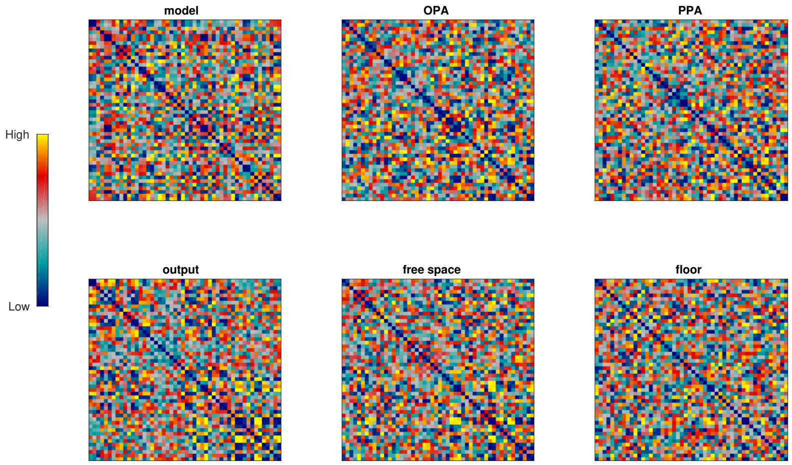


Fig. 1: Top row: RDM of the behavioral model for navigational affordance followed by RDMs of brain responses in OPA and PPA. Bottom row: RDMs of final layer, free space labels, and floor mask output of a scene parsing DNN

for DNNs the RDMs are computed by comparing the layer activations for each image pair in the dataset. In this work, we consider OPA and PPA RDMs for comparison as these areas have been hypothesized to represent scene affordances [2] and scene layout [6] respectively. We also compare the DNN RDMs with a behavior Navigational Affordance Map (NAM) [1] that represents navigational affordances in a scene. The top row in Fig. 1 shows RDMs of NAM, OPA, PPA obtained from the dataset and bottom row shows the RDMs of the final layer output, combined activations of categories corresponding to free space, and activation of the floor of the scene parsing DNN.

The dissimilarity metric used in this work is $1 - \rho$ where ρ is the Pearson's correlation coefficient. Although in previous work [1], where a scene classification DNN was compared with the navigational affordance the dissimilarity metric used was the Euclidean distance, we observed that with $1-\rho$ as the dissimilarity metric, the correlation was higher. Hence, in this work for all the analysis $1 - \rho$ is used as the dissimilarity metric to compute RDMs of layer activations. We did not use PCA on layer activations as done in [1] since the spatial information in the case of convolutional layer outputs is lost by performing PCA.

Statistical analysis We use RSA toolbox [13] to compute RDM correlations and corresponding p-values and standard deviation using bootstrap similar to [1]. For determining which RDM better explains the behavioral or neural RDMs, we perform a two-sided statistical comparison. The p-values are estimated as the proportion of bootstrap samples further in the tails than 0. The number of bootstrap iterations for all the analysis was set to 5000.

2.2 Navigational Affordance Dataset and Model

The stimuli images used for analysis consisted of 50 images of indoor environments. The subject's fMRI responses were obtained while they performed a category-recognition task (bathroom or not). In this work, we directly use the precomputed RDMs of the navigational affordance map (NAM), PPA and OPA provided by Bonner and Epstein [1]. Recall that RDMs are constructed by computing dissimilarities of all possible pairs in the dataset as explained in section 2.1.

To obtain NAM, first, an independent group of subjects was asked to indicate the paths in each image starting from the bottom using a computer mouse. The probabilistic maps of paths for each image were created followed by histogram construction of navigational probability in one-degree angular bins radiating from the bottom center of the image. This histogram represents a probabilistic map of potential navigation routes from the viewer's perspective. For further details of the navigational affordance model or dataset, we refer the reader to [2, 1].

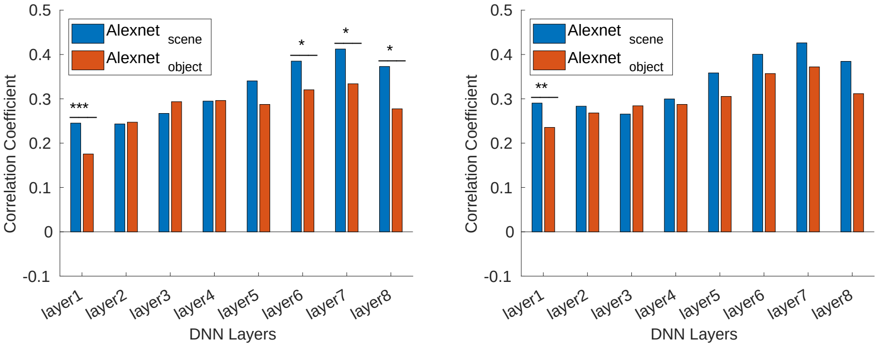
2.3 Deep Neural Network Models to explain brain responses

In this section, we describe the architecture of the DNN models used in the analysis.

Object classification model We used Alexnet [10] which we refer as Alexnet$_{\text{object}}$, trained on Imagenet [4] dataset (an object classification dataset) as the object classification model. The Alexnet model [10] consists of 5 convolutional layers each followed by a pooling layer and 3 fully connected layers after the last pooling layer.

Scene classification models We used the same model as above (Alexnet) but trained on Places [18] dataset (a scene classification dataset) as the scene classification model (referred as Alexnet$_{\text{scene}}$). For comparison with scene parsing model we choose VGG16 [14] trained on Places as the scene classification model (VGG$_{\text{scene-class}}$). The reason behind the different choice of scene classification models was that we were unable to find a pretrained scene parsing model with similar architecture as Alexnet. The VGG16 model contains 13 convolutional layers with 5 pooling layer after a convolutional block of either 2 or 3 convolutional layers and 3 fully connected (FC) layers after the last convolutional layer.

Scene parsing models We use fully convolutional modification of VGG16 [11] trained on the scene parsing dataset ADE20k [19], [20] as the scene parsing model ($\text{VGG}_{\text{scene-parse}}$). In $\text{VGG}_{\text{scene-parse}}$, the FC layers are replaced by convolutional layers to predict pixel-wise spatial mask. We use pyramid scene parsing network ($\text{PSP}_{\text{scene-parse}}$) for performing analysis of class specific masks as $\text{PSP}_{\text{scene-parse}}$ outperforms $\text{VGG}_{\text{scene-parse}}$ on scene parsing task and hence the class masks are more accurate and suitable for this particular analysis. The $\text{PSP}_{\text{scene-parse}}$ model introduces a pyramid pooling module that fuses features of four different scales to obtain superior performance on scene parsing task.



(a) PPA

(b) OPA

Fig. 2: RSA of (a) PPA with layers of DNN trained on scene and object classification.(b) OPA with layers of DNN trained on scene and object classification. The asterisk at the top indicates the significance of difference (*p <0.05, **p <0.01, ***p <0.001)

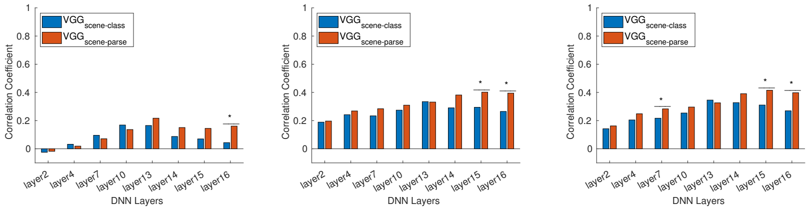
3 Results

Here, we first report the correlation results of the scene-selective areas (OPA and PPA) with an object classification model ($\text{Alexnet}_{\text{object}}$) and a scene classification model ($\text{Alexnet}_{\text{scene}}$). Then, we report the correlation results of the NAM, OPA, and PPA with a scene parsing model ($\text{VGG}_{\text{scene-parse}}$) and a scene classification model ($\text{VGG}_{\text{scene-class}}$). Finally, we investigate category specific activations of the scene-parsing model and compare the correlations of NAM, OPA, and PPA with relevant categories.

3.1 Scene vs. Object classification

We compare the correlation of all pooling and fully connected layer outputs of $\text{Alexnet}_{\text{scene}}$ and $\text{Alexnet}_{\text{object}}$ with scene-selective brain areas (OPA and PPA).

From the comparison result with PPA, we observe that for all the layers except pool3 the Alexnet$_{\text{scene}}$ show a higher correlation (Fig. 2(a)). A similar trend is observed by comparing with OPA(Fig. 2(b)). The results support our hypothesis that a model trained on a related type of images better represents the brain activity.



(a) NAM (b) OPA (c) PPA

Fig. 3: RSA of (a) NAM with layers of DNN trained on scene classification and parsing,(b) OPA with layers of DNN trained on scene classification and parsing, and, (c) PPA with layers of DNN trained on scene classification and parsing. The asterisk at the top indicates the significance of difference (*p <0.05, **p <0.01, ***p <0.001)

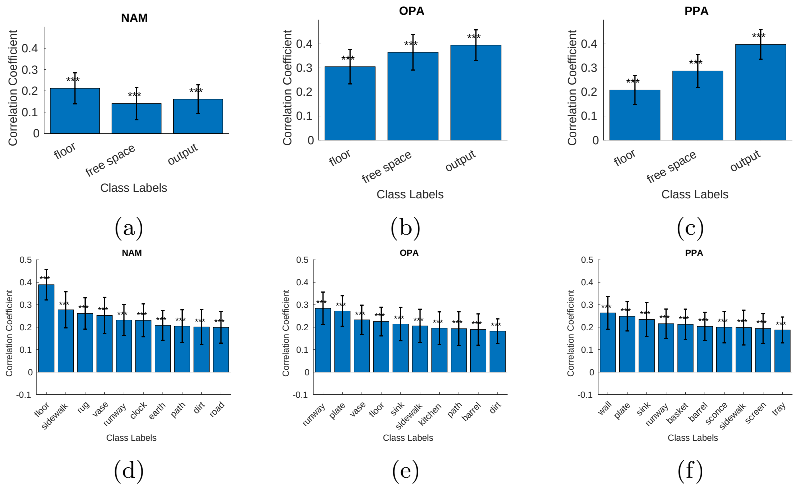
3.2 Scene-parsing vs. Scene-classification

For computing the correlation with OPA, PPA, and NAM, we use the outputs of 5 pooling layer and 3 fully connected layers of VGG$_{\text{scene-class}}$ and 5 pooling layers and convolutionalized version of 3 fully connected layers of VGG$_{\text{scene-parse}}$.

In general, from Fig. 3 we observe that deeper layers of VGG$_{\text{scene-parse}}$ model have higher correlation values with the behavioral model and brain responses than the earlier layers. Further, for all three cases, we observe that the difference in correlation values of VGG$_{\text{scene-parse}}$ and VGG$_{\text{scene-class}}$ is more significant in the deeper layers with higher correlation values for VGG$_{\text{scene-parse}}$ layers.

One explanation for these results that supports our hypothesis is that the deeper layers of the DNNs are more task-relevant while earlier layers perform generic feature processing. A related possible explanation might be that since VGG$_{\text{scene-parse}}$ is a fully convolutional model, it's last three layers are convolutional while in VGG$_{\text{scene-class}}$ the last 3 layers are fully connected. This suggests that convolutional layers may better represent a spatial scene property such as navigational affordance. The convolutional layer output has information about the spatial structure of the scene in explicit form while fully connected layers lose the spatial information, and therefore this might be another possible reason for the high difference in the correlation values. This again supports our hypothesis of task-related models being more correlated as compared to a generic model.

The results also suggest that spatial information is preserved in the higher brain areas such as PPA and OPA and the models with the fully connected layers may not represent these areas better than fully convolutional models. The results show navigational affordance related model VGG$_{\text{scene-parse}}$ shows a higher correlation in most of the layers with NAM, PPA, and OPA.



(a) (b) (c)

(d) (e) (f)

Fig. 4: Top: RSA of final layer output of VGG$_{\text{scene-parse}}$, free space, floor labels with (a) NAM,(b) OPA, and (c) PPA . Bottom: Top-10 correlated classes with (d) NAM,(e) OPA, and (f) PPA. Error bars represent bootstrap ±1 s.e.m. (*p <0.05, **p <0.01, ***p <0.001).

3.3 Floor and free space labels

To investigate further what information is present in the OPA activity, we first separate out labels from the ADE20k dataset which correspond to free space. We found 13 such labels (road, floor, sidewalk, etc.) that correspond to free spaces. Since the images in the dataset were from indoor scenes, we considered one more case with only floor label. The output of VGG$_{\text{scene-parse}}$ consists of 151 channels in which 150 channels correspond to a class in the ADE20k dataset, and one channel corresponds to the background. Therefore, we investigated if the output of channels corresponding to free space classes such as roads and floor might have a higher correlation with the NAM and OPA.

Using RSA, we compute the correlation of the final layer of the scene parsing with NAM and OPA and compare it with the output containing only free space labels (13 channels) and floor labels (1 channel). From the results of the comparison, shown in Fig. 4 (top row), we observe that although for RSA analysis with NAM the output with only floor label shows the highest correlation this is not the case with the OPA. The results suggest that OPA might encode information more than just the floor labels which are highly representative of the navigational affordance in the images considered.

Top correlated classes To gain further insights about the information encoded in OPA and PPA we computed the correlation of each class activation from the DNN with place areas OPA and PPA and also with the NAM. From the class activation maps with high correlation values, we may gain some insights about what is encoded in the place areas. For this analysis, we choose a highly accurate scene-parsing model $\text{PSP}_{\text{scene-parse}}$ which generates more accurate masks than $\text{VGG}_{\text{scene-parse}}$ model used in the previous analysis. We take the activations of the last output layers which has 150 channels corresponding to each class in the ADE20k dataset and then compute RDMs corresponding to each channel output for RSA analysis.

For the NAM (Fig. 4(d)), as expected the floor class has the highest correlation. The next few classes that showed the highest correlation values were also indicative of free space such as rug, sidewalk, runway, etc. Surprisingly, the objects such as vase and clock also showed high correlation. This might be because vase and clock are typically placed on floor and wall, respectively.

For OPA (Fig. 4(e)), although 50 percent of the labels in the top-10 list included labels corresponding to free space, rest of the labels include objects like plate, vase, sink, kitchen, and barrel. One possible explanation for these classes is the experimental design in which the OPA responses were recorded. The subjects were asked to classify whether the room displayed is a bathroom or not. The objects such as sink, plate, and vase are highly indicative of the room type, and OPA responses may be related to the classification task. Therefore, the high correlation of OPA with these objects is explained by assuming that OPA is involved in the classification task. Further, knowing the scene category is also crucial for planning navigation. A related possible explanation is that the objects also suggest the spatial layout of the scene by indicating the presence of obstacles and therefore can be relevant for navigational affordances.

PPA, on the other hand, is hypothesized to represent the spatial layout of the scenes and is insensitive to the navigational affordance as shown in [1]. The results from this analysis (Fig. 4(f)) are consistent with [1] as the majority of the labels with high correlation are objects that are indicative of scene layout and category and only a few of the highly correlated classes correspond to free space.

4 Conclusion

In this work, we demonstrated that task-specific areas in the visual cortex are better explained by a model trained to perform a similar task. In particular, we first showed that responses of scene selective visual areas are better explained by a DNN trained on the similar type of the images. Next, we showed that OPA activity which has been hypothesized to be associated with the navigational affordances shows a higher correlation with task-relevant deeper layers of a scene parsing DNN than a scene classification DNN.

Our results also show that a DNN model trained for scene parsing task may provide more insights about the brain responses associated with navigational affordances (OPA brain area). With the scene parsing model, we were able to perform the detailed analysis with each class activation showing that OPA responses are also highly correlated with the objects that are indicative of the scene type. This suggests that OPA also plays a role in scene classification since knowing scene category is also crucial for planning navigation.

Acknowledgement

This work was funded by the MOE SUTD SRG grant (SRG ISTD 2017 131). Kshitij Dwivedi was also funded by SUTD Presidents Graduate Fellowship.

References

1. Bonner, M.F., Epstein, R.A.: Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. PLOS Computational Biology . https://doi.org/10.1371/journal.pcbi.1006111
2. Bonner, M.F., Epstein, R.A.: Coding of navigational affordances in the human visual system. Proceedings of the National Academy of Sciences **114**(18), 4793–4798 (2017)
3. Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., Oliva, A.: Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Scientific Reports **6**(June), 1–13 (2016). https://doi.org/10.1038/srep27755, http://dx.doi.org/10.1038/srep27755
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 248–255. IEEE (2009)
5. Dilks, D.D., Julian, J.B., Paunov, A.M., Kanwisher, N.: The occipital place area is causally and selectively involved in scene perception. Journal of Neuroscience **33**(4), 1331–1336 (2013)
6. Epstein, R., Harris, A., Stanley, D., Kanwisher, N.: The parahippocampal place area: Recognition, navigation, or encoding? Neuron **23**(1), 115–125 (1999)
7. Horikawa, T., Kamitani, Y.: Generic decoding of seen and imagined objects using hierarchical visual features. Nature communications **8**, 15037 (2017)
8. Khaligh-Razavi, S.M., Kriegeskorte, N.: Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. PLoS Computational Biology **10**(11) (2014). https://doi.org/10.1371/journal.pcbi.1003915
9. Kriegeskorte, N., Mur, M., Bandettini, P.A.: Representational similarity analysis-connecting the branches of systems neuroscience. Frontiers in systems neuroscience **2**, 4 (2008)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
11. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
12. Martin Cichy, R., Khosla, A., Pantazis, D., Oliva, A.: Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. NeuroImage **153**, 346–358 (2017). https://doi.org/10.1016/j.neuroimage.2016.03.063, http://dx.doi.org/10.1016/j.neuroimage.2016.03.063

13. Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N.: A toolbox for representational similarity analysis. PLoS computational biology **10**(4), e1003553 (2014)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
15. Tacchetti, A., Isik, L., Poggio, T.: Invariant recognition drives neural representations of action sequences pp. 1–23 (2016). https://doi.org/10.1371/journal.pcbi.1005859, http://arxiv.org/abs/1606.04698
16. Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., DiCarlo, J.J.: Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences **111**(23), 8619–8624 (2014). https://doi.org/10.1073/pnas.1403112111, http://www.pnas.org/cgi/doi/10.1073/pnas.1403112111
17. Yamins, D.L., DiCarlo, J.J.: Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience **19**(3), 356 (2016)
18. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence (2017)
19. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. arXiv preprint arXiv:1608.05442 (2016)
20. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 633–641 (2017)