Examining Shape and Texture Representations in Human Brain Using Deep Convolutional Neural Networks

Shreya Gupta
Symbolic Systems/ Stanford University
shreyagupta@stanford.edu

Abstract

Imagenet-trained deep convolutional neural networks (dCNNs) are among the best models of the human visual system and of human object recognition behavior. However, recent findings suggest that these dCNN models diverge from human behavior in very noteworthy ways, such as their preference for texture over shape information and their susceptibility to high spatial frequency perturbations. We hypothesize that the robust performance of human observers on such visual discrimination tasks must be supported by some cortical substrates, although recent evidence suggests that these representations might not be found along the ventral visual stream. Here, we make use of whole-brain neuroimaging data from the Natural Scenes Dataset to finetune dCNN models by maximizing predictivity of various cortical regions along the ventral and dorsal visual streams. We then propose to evaluate the performance of these cortically-constrained dCNN models on a texture-shape cue-conflict task. We seek to identify which cortically-constrained dCNN models would be most shape-biased and therefore best predictive of human behavior in a texture-shape discrimination task. We hypothesize that cortical regions downstream of the ventral temporal cortex might contain the best representations for matching human behavior, although it is also possible that a dCNN model jointly fit to multiple brain areas (e.g. V1 and VTC) might also provide the best model of human behavior. These results provide a framework for understanding the the robust perceptual abilities of human observers and help us to identify the cortical substrates underlying robust object recognition.

1 Introduction

Human visual recognition mechanism has been a topic of research interest since a long time. We are able to detect objects in images irrespective of angle and light variations, with a noteworthy accuracy and speed. Attempts have hence been made to understand our neural mechanisms. A landmark feat was achieved when Yamins et al. (2014) were able to predict IT neural data using a class of deep convolutional neural network (dCNNs) without explicitly training the model on the neural data. This inspired a chain of studies which tried to model human visual system using artificial neural networks. Rajalingham et al. (2018) were able to predict human object recognition behavior using pre-trained dCNN models and training multiclass logistic regression to output the class label for the images shown to humans and monkeys.

Figure 1: Classification of a standard ResNet-50 of a) texture-source image of an elephant skin b) shape-source image of a cat and c) cue-conflict image of a cat superimposed on an elephant skin. Image and values taken from (Geirhos et al., 2018)

While CNNs have done a remarkable job at predicting neural responses, they are still susceptible to adversarial attacks. Szegedy et al. (2013) showed by adding minute noises to the image pixels the output of the deep learning model changed from a dog to red wine. These noises were tactfully placed and were unseen to the naked eye yet fooled the model. This problem was deepened when Geirhos et al. (2018) demonstrated the diverging behavior of CNNs from human behavior. While humans classified the image in Figure 1 c) as a cat based on its shape, the ResNet model confidently labelled it an Indian Elephant based on its texture.

Jagadeesh and Gardner (2022) found the cause of the divergence in dCNNs from human behavior. They showed that the ventral visual stream were also generating texture-like representations using an oddity identification task. They found that the primary visual cortex was able to identify the odd image in a triplets of objects when they had different shape and texture (for example, image of a tiger and two different images of elephants) but was not able to distinguish between the triplets when they had only shape variation (for example, distinguishing tiger from a set of scrambled tigers). This raises an important question: if the ventral visual stream is generating texture-like representations, how are we as humans able to classify images based on their shapes?

Section 2 formalises our goals for this project, Section 3 enlists the methodology we adopt to bridge this gap and Section 4 delineates the dataset and its relevant properties. Section 5 describes the proposed experiments to test our hypothesis. Section 6 and 7 describe what results we hypothesize receiving and the interpretation of each of the possible outcomes along with some future work possible in this direction.

2 Problem Statement

The main goal of this study is to create dCNN-based models of different brain areas using NSD Data (Allen et al., 2022) and make an attempt to understand how the brain areas interplay during object detection. More concretely, we want to understand which sections of the brain are biased towards the shape of the incoming image and which areas pay more importance to the texture of the image while performing object detection. We want to use this information to hypothesize what might happen in our object detection mechanisms.

3 Methods

For the purposes of this study, we make use of the Natural Scenes Dataset (NSD) (Allen et al., 2022) to get blood-oxygen level dependent (BOLD) activity of different brain areas when human subjects were shown various images. We then construct a model of each brain area by fine-tuning a pre-trained deep convolutional neural network (dCNN) to learn mappings from the dCNN to the voxel responses for the images (Figure 2). Finally we examine the representational geometry of each of our models of cortical regions to understand how shape- or texture-biased they are to understand the underlying neural mechanisms behind object detection (Figure 3).

To delineate our methodology further, we obtain the image data for a subject. The image data consist of the images that were shown to the subject. Then, for a pre-specified brain area (for example, primary visual cortex, ventral temporal cortex or posterior parietal cortex) that we want to inspect, we obtain the voxel responses against those images. Voxel responses consist of the BOLD values of the voxels in the pre-specified brain area when an image is shown. The methodology to obtain these voxel responses is elucidated in Section 4.

After obtaining the voxel responses and the image data, we load a pre-trained dCNN and use a linear regression to learn the mapping from the dCNN to voxel values. When loading the dCNN we truncate it to a pre-specified layer (for example, the fourth pooling layer) that yields the best mapping (i.e. best performance) from the dCNN to voxel responses. To learn this mapping we use the features generated by the truncated dCNN, F and learn the weights, W of the linear regression layer in equation 1.

$$F \cdot W = V \tag{1}$$

Here V represents the voxel responses of the brain area we are trying to fit our model on. This is represented in Figure 2

Once we have learned the mapping function in equation 2, we have a neural model of each voxel. cnn in equation 2 refers to dCNN features, w and b represent weight and bias values learnt in the linear mapping process. For a given brain area, we can now examine how this model of a cortical region would perform in various behavioral metrics. One such metric is a texture vs shape bias task as elucidated in (Geirhos et al., 2018).

$$f(cnn, w, b) = \text{voxel responses} \tag{2}$$

This task helps us test whether the brain area's final prediction is biased based on the given image's shape or its texture. We use the texture-source image (for example elephant skin) and shape-source image (for example cat) from (Geirhos et al., 2018) and generate the cue-conflict image (the cat with elephant skin in our example) using style transfer.

To find the inclination of the brain region, we show the cortical model a cue-conflict image and compute the response vector using the dCNN

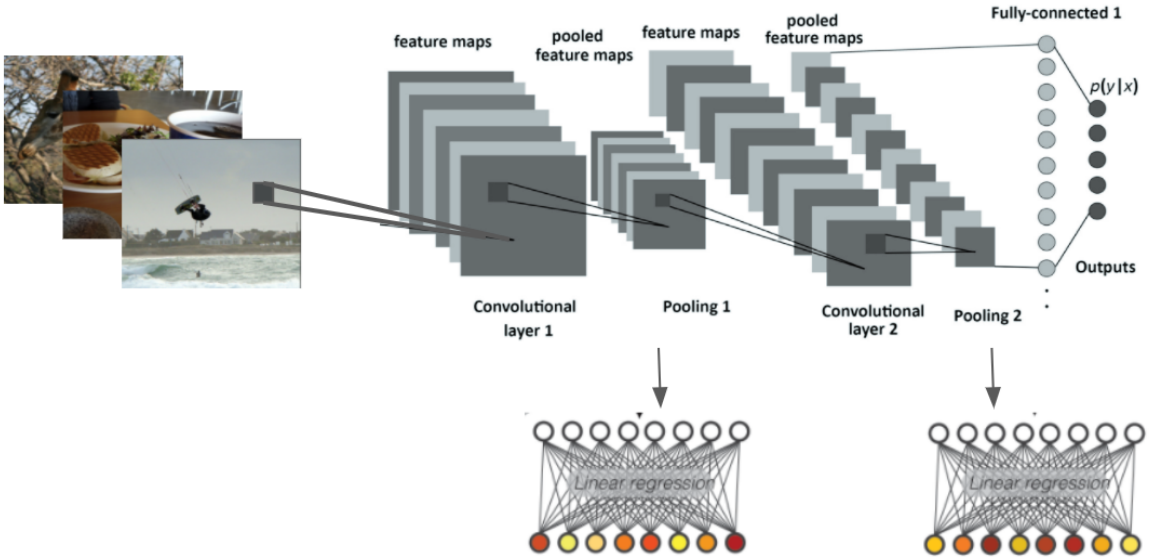

Figure 2: Deep convolutional neural network model of different parts of the primary visual cortex. Imagenet-trained VGG-19 features from different pool layers (pool1 and pool2 here) are fit to predict voxels belonging to different brain areas.

model succeeded by the linear regression model. This response vector consists of the predicted responses of each voxel in that brain area. We then show the model the texture-source image (for example, elephant skin in this case) and the shape-source image (for example, cat in this case) and for each of those, compute a response vector following the process described previously. Finally, we compute the similarity between the response vector for the cue-conflict image (cat shape on an elephant skin) and the response vector for texture-source image and shape-source image respectively. This is demonstrated in Figure 3.

4 Dataset

In this paper we use the recently released Natural Scenes Dataset (NSD) by Allen et al. (2022) . It is a large-scale fMRI dataset conducted at ultra-high-field (7T) strength. The dataset consists of fMRI responses of 8 healthy adult human subjects when they were shown a combined number of 73,000 colored natural scenes over 40 sessions consisting of 750 trials. Each subject was shown 10,000 images, of which 9,000 were unique and 1,000 were common across all images. Each image was shown thrice to a subject and while viewing the images, they were engaged in a recognition task.

The part of the data that was relevant to us was extracting the fMRI responses of different brain areas. For the purpose of this study, we used the single-trial fMRI values which are estimated by fitting hemodynamic response functions individually for each voxel, denoising voxel responses by regressing out nuisance variables, and using ridge regression to regularize beta estimates. This process yields an accurate estimate of the response of each voxel on each trial, which they then averaged across repetitions of the same image to derive a trial-averaged response of each voxel to each image.

We also use the voxel locations of different brain areas from the primary visual cortex, ventral temporal cortex or posterior parietal cortex etc. Finally, we map the voxel locations and their responses with the images that were shown to the subjects. Since each image is shown thrice to a subject, we average the fMRI response across the three trials to obtain a single fMRI response for each voxel for that image.

5 Experiments

We use the NSD data to obtain the voxel responses against each image shown to a subject. There is a raft of possible experiments that can be conducted with this type of data. Even though there is no major difference between the kind of fMRI response we choose (from considering the brain as a big cube or flattening the surface), we use the native surface fMRI response for its simplicity. More details about it are given in (Allen et al., 2022).

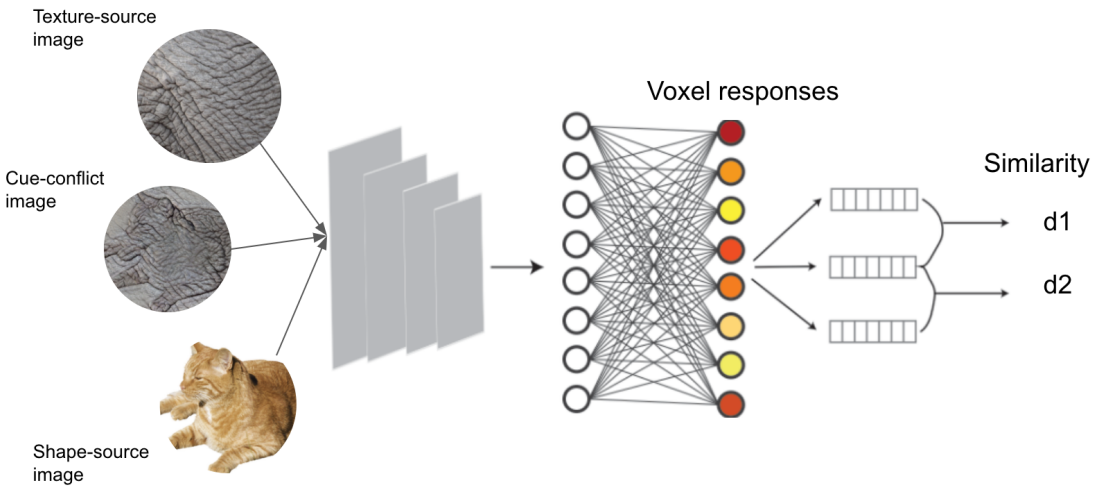

Figure 3: Model architecture while evaluating the shape and texture bias of a pre-defined brain area. The the three images (texture-source, shape-source and cue-conflict) are passed through a pre-trained dCNN and the linear regression layer to generate voxel responses. Thereafter similarity of the cue conflict voxel response is calculated with the voxel response of the texture source and shape-source images respectively and a judgment made.

Second set of experiments surrounds the choice of brain regions we want to consider. We will consider early visual cortical ROIs, defined retinotopically (e.g. V1, V2, V3, hV4). We will also examine high-level visual cortical ROIs, defined with a functional localizer (E.g. FFA, PPA, OTS, etc). We will also examine parietal lobe ROIs, defined with an anatomical atlas (From Wang atlas, we can look at IPS0, IPS1, IPS2, IPS3, IPS4, IPS5 and SPL1). For each of these areas, we can obtain their voxel locations and responses from the NSD data.

Post choosing the fMRI responses of the brain areas, we can experiment with the choice of pre-trained deep convolutional neural networks (dCNN). For the purposes of our study we are currently using a VGG19 (Simonyan and Zisserman, 2014), but we can also replace it with other dCNN models, like ResNet50 (He et al., 2016) or visual transformer (Dosovitskiy et al., 2020). The purpose of this dCNN is to generate features upto a pre-defined layer. The layer we truncate to is also a hyper-parameter of the study. We hypothesize the early layers in dCNN will generate better representations of high level visual cortex layers (like V1 and V2) whereas the later layers in dCNN have more potential to be better neural models of late visual cortex layers. For example, in (Cadena et al., 2019) pooling layer 1 was found to be more synchronous to V1 whereas output pool layer 4 was found more synchronous to IT in (Yamins et al., 2014). Even though their objectives were marginally tangential, it gives a rough idea of how high level features in dCNN correspond to early visual cortex layers for visual tasks.

After extracting the truncated pre-trained VGG, we have multiple ways to adapt it to learn the voxel responses. One of the ways which includes training a linear regression to learn the mapping from the feature vectors obtained from the truncated dCNN is explained in the methodology section. An alternate experiment is to fine-tune the entire dCNN to voxel responses. Since the data size is larger than the datasets used in (Yamins et al., 2014) and (Jagadeesh and Gardner, 2022), it is also worth experimenting training a dCNN from scratch.

We can generate these dCNN-based neural models for any or all the brain areas we want to consider and analyze. Thereafter, we use a Pearson Correlation (Benesty et al., 2009) to extract the similarity between the neural response generated by our model for the cue-conflict image with those generated by the texture-source and shape-sources images respectively.

6 Results

Our results will majorly be two folds. The first part will consist of the accuracy scores from how well our neural dCNN-models were able to fit the fMRI-

voxel responses. We expect that we will be able to fit neural responses well in most visual regions. We also anticipate, based on prior evidence (Yamins et al., 2014), that in early visual cortical regions, voxel responses will be best explained by responses from early layers of the VGG19 model, whereas in late visual cortical regions, voxel responses will be better explained by later layers of the model.

For the choice of adapting a dCNN for predicting voxel responses, we predict that fine-tuning the entire VGG might have the best performance, followed by a linear regression layer model. We anticipate that training an entire CNN from scratch is less likely to give promising results, and even if it learns to emit voxel responses, its ability to learn a visual task is questionable.

In the second set of results, we report the Pearson Correlation (or results from a similar similarity metric) between the neural response generated by our model for the cue-conflict image with those generated by the texture-source and shape-sources images respectively. A few different possible outcomes are possible. These are shown in Figure 4. It is possible that fitting to the high-level visual cortex makes the model more shape-biased, and shape-bias increases from early (e.g. V1) to late (e.g. VTC) ventral stream regions. This is reflected in the light blue colored line in Figure 4. The opposite outcome is also possible - early visual cortex contains more spatial and contour information, so maybe fitting to early regions might increase shape bias more than late ventral stream regions. This is represented by the dark blue line in Figure 4. The third scenario is possibly some joint readout of early (e.g. V1) and late (e.g. VTC) regions generate the most shape-biased representation, as suggested by (cite Jagadeesh Gardner 2021). The orange line represents this outcome. Another possible outcome is the possibility that dorsal stream regions (i.e. parietal lobe) actually contain shape representations, not ventral stream regions. Finally, it is also possible that shape-bias is not significantly prevalent through any of these regions (or a combination of these). It is represented by the green line in Figure 4.

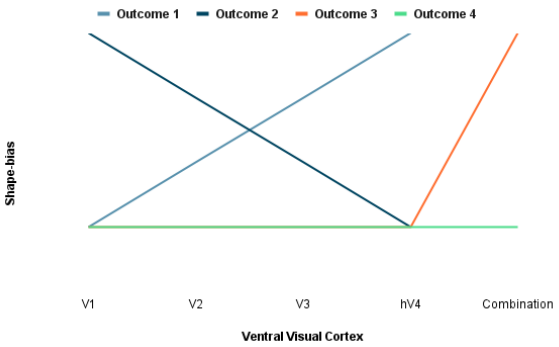

Figure 4: Possible Outcomes of shape-bias across different regions (V1 - hV4) in primary visual cortex. Combination reflects a potential joint readout of early and late Visual Temporal Cortex regions.

7 Discussion and Future Directions

The results and its analysis would be a step into understanding where the texture-bias in our object detection mechanism originates from and when and how the shape-bias overpowers this texture bias in the process. If texture bias originates from the higher level layers and increases from early to late ventral stream regions, this would indicate that the late regions have more contribution towards the overall prediction and vice versa. If, however, a combination of these generates the most shape-biased representation, that would indicate an interplay between the regions. If the final possible outcome is true, i.e. none of the regions or a reasonable combination of them yields a bias towards shape while detecting objects, that would indicate the possible interference of regions outside the ones NSD considers. The alternate explanation for such an outcome is also the inability of our model to capture this difference.

The future directions of this research is to explore and examine fMRI responses from the frontal lobe for a similar experiment. It is also worth exploring other metrics to measure the shape and texture bias of the cortical regions.

References

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. 2022. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. 2019. Deep convolutional models improve predictions of macaque v1

responses to natural images. *PLoS computational biology*, 15(4):e1006897.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. 2018. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Akshay V. Jagadeesh and Justin L. Gardner. 2022. Texture-like representation of objects in human visual cortex. *bioRxiv*.

Rishi Rajalingham, Elias B. Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J. DiCarlo. 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*, 38(33):7255–7269.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624.