An ablation approach to identifying functionally specific units from deep neural networks trained on object recognition

Abstract

Areas of the visual cortex respond to broad categories of objects (e.g. animate vs. inanimate objects). Ground truth for the specificity of function in these regions is established by lesioning the area in question. We seek to provide evidence of the representation of similar broad object categories in deep neural networks (DNNs) pre-trained on ImageNet. Prior neural network ablation studies have shown that single-unit and unit-pair ablations can negatively affect the overall performance of a model. However, selecting multiple units that interact as a functionally-specific group is challenging. To do so, we borrow an approach from the analysis of functional magnetic resonance imaging data in which we cluster units based on the similarity of their functional activation. We begin by clustering DNN units based on similarity in their average responses, then ablate each cluster and measure the specificity of the resulting performance deficit. These cluster ablation studies identified groups of units with performance deficits specific to the animate and inanimate objects with a differential effect of up to 66% in the test data set. Clusters of inanimate-specific units have similar activation profiles and are more prominent in later layers, similar to findings from neuroscience. Clusters of animate-specific units are also more common in later layers but show more distributed patterns of activation. This result provides evidence of groups of units in DNNs that are functionally selective for some broad image categories. We note that a similar approach could also be used to improve computational performance, through targeted pruning, and to remove bias from pre-trained artificial neural networks.

1 Introduction

Human visual cortex contains areas with specific functional roles that have been confirmed by lesion studies [1-5]. Studies of functional specificity have helped neuroscience to better understand biological solutions to vision. Finding similar functionally-specific groups of units in DNNs would provide a foundation for better interpretability and a mechanistic explanation of function. Here, we use a novel DNN ablation approach to identify groups of functionally specific units analogous to those seen in the brain. We first cluster groups of DNN units by the average activity produced in response to groups of image categories from ImageNet. Second, a mask layer is produced for each cluster that gates outputs from each unit in the target cluster. We then measure the resulting categorization performance deficit produced by the ablated network. If ablating a group of units produces a recognition deficit in one category but not others, that group of units can be said to be functionally specific for that category.

Approaches to identifying functionally specific units in DNNs can be divided into two families: model interpretability and ablation studies. Prominent methods of model interpretability such as layer wise relevance propagation (LRP) are used to build saliency maps in which backpropagation is employed to identify DNN units relevant to model predictions by assigning a score to each unit
with a layer [6]. Input perturbation methods, such as LIME, modify the input image to see how these changes affect categorization. Observing which image changes affect categorization provides a method to better understand the mechanism by which the DNN categorizes images [7]. Further, methods like TCAV or testing with concept activation vectors constructs support vector machines that represent high level features or concepts (e.g. gender, race, color) [8]. These interpretability methods, LRP, TCAV and others [9-12], show evidence of distinct representations for different categories within a layer. We extend these methods to move beyond modulating input or studying relevance within a layer to consider functionally specific representations across DNN layers for different image categories. We also confirm functional specificity by ablating units to causally establish their role in object identification.

Ablation studies of neural networks have established the necessity of a set of units by measuring performance differences with and without a targeted set of units [13]. Recent research on neural network ablations provide evidence for function-specific single units and unit pairs that can negatively impact the performance of the model when removed [14]. In some cases, single-unit ablations were even shown to improve the performance of the model for a particular image class. Here, we seek to ablate groups of neurons with similar activation profiles to identify functionally specific representations shared across broad categories of images. While it is possible to ablate single-units from simple feed forward neural networks with fewer units, identifying groups of units likely to have similar function, especially for larger networks like MobileNetV2 with millions of units, is challenging. We propose a novel ablation approach based on functional grouping commonly used in the analysis of functional magnetic resonance imaging data (fMRI). Clustering techniques in fMRI are often used to identify voxels (brain units) with similar functional activation profiles. For example, researchers have used gaussian mixture models to cluster parts of the brain with similar activation profiles both within a single study and to identify parts of the brain with similar activation profiles across studies [15]. Here, we apply the idea of clustering in unit response space to identify DNN units that are functionally specific to broad categories of objects.

2 Methods

To ensure we identify convergent representations that are not idiosyncrasies of a single category, we group eight image classes into two broad (and untrained) categories: animate and inanimate objects. Animate and inanimate objects also has the benefit of being two broad categories that for which previous representational similarity between the brain and Alexnet has been demonstrated [16]. The images used in this study was taken from the ImageNet validation dataset which contains 50,000 images across 1000 image categories [17]. We chose 8 image categories that could be grouped into animate and inanimate categories (4 different image categories randomly sampled within each category). We carefully removed images from the inanimate subset that contained animate objects (i.e. a person using a pen). The final number of images in the animate and inanimate subset was 194 in each. We report results from MobileNetV2 (alpha=0.35) [18] here but have completed similar analyses on Alexnet [19], VGG16 [20], and Resnet101 [21].

Clustering DNN units can be challenging since contemporary object-recognition DNNs can contain millions of units. Past approaches have used single unit embedding techniques to identify groups of neurons with similar activation specificity [22]. Here, we assume that unit activation profiles can be grouped using a mixture model in which each cluster is thought of as a function that supports object recognition. Recent ablation work shows that neither activation specificity [23] nor activity levels [22] are predictive of functional selectivity for single units. We therefore test the performance deficit that results from ablating each cluster of units regardless of activation level or specificity. The number of processes in the mixture model \((k)\) is a free parameter estimated by four-fold cross-validation on an ablation ‘training’ set. The \(k\) parameter is potentially constrained one or both of two factors: 1. The number of separable processes for the image categories used. 2. The expected number of units in a functional cluster. In order to make clustering computationally feasible for a large number of units, we employ a two-step clustering approach [24]. In the first step, we cluster activation values of units averaged across animate and inanimate categories objects using batch \(k\)-means [25]. We gather the clusters from the \(k\)-means step and cluster further using gaussian mixture models. Functional specificity of a cluster of units is measured by the deficit in performance post ablation of that particular cluster.
We introduce a novel method of ablation that allows the ablation of individual units in convolutional networks. Ablations of single units in DNNs are typically performed by manually setting their incoming weights and biases to zeros [14][22]. Similarly, ablations in convolution neural networks is performed by setting the weights and biases of a selected kernel to zero [14] and thereby preventing information flow into those neurons. Here, we use lambda layers to support ablation of selected units across different DNN layers (Figure 1). The global lambda layer consists of weights that define a binary ablation mask. The lambda layer can either ablate (zero the output for) or permit the unit to participate in object recognition. The units are ablated by multiplying the activation output ($O_{i,j}$) of a DNN unit with 0.

$$O_{i,j} = \begin{cases} O_{i,j} & \text{if the unit is permitted} \\ 0 & \text{if the unit is ablated} \end{cases}$$

Where $i$ denotes the index of the lambda layer and $j$ denotes the index of the unit in the corresponding layer. We included lambda units after each unit with a trainable parameter.

Performance impact of a cluster is computed as the difference between the top-5 testing accuracy of the pre-ablated model with the cluster-ablated top-5 testing accuracy (the top-5 predictions on the individual object class labels). A cluster of units would be described as having an animate-specific function if its ablation led to a significant drop in performance for animate test images while not affecting the performance on the inanimate test images. Similarly, ablating a cluster of units with inanimate-specific function would produce a significant drop in performance for inanimate images while not affecting the performance of animate images. Here, we define this relationship as the maximum potential performance impact ($MPPI$) for a cluster($c$):

$$MPPI_{c} = d(animate) - d(inanimate)$$

Where $d(category)$ refers to the performance deficit caused by the ablation on the particular image super category.

### 3 Results

The number of clusters ($k$) was chosen (Figure 2a) as the value of $k$ for which the difference in MPPI between the animate selective and inanimate selective clusters was maximized (4-fold cross validation): $\delta MPPI = MPPI_{animate}(maximum impact animate cluster) - MPPI_{inanimate}(maximum impact inanimate cluster)$. Starting with $k=2$ (to $\approx 192$ in powers of 2), we clustered DNN units, ablated each cluster, and determined the resulting MPPI score. A cluster of units with MPPI score close to 1 indicates functional specificity for animate images while a score close to -1 denotes specificity for inanimate images. The final value of $k$ is chosen based on the difference between the clusters with the maximum (animate-specific) and minimum (inanimate-specific) MPPI score. On
average across the four-fold cross-validation, the MPPI difference was largest for MobileNetV2 at $k=256$ with an average number of units per cluster of 22,365 (Figure 2).

We next sought to identify the units in common across functionally selective clusters and folds of the cross-validation. For each fold from the cross-validation, we took the intersection of units present in the top-$m$ functionally selective clusters for animate and inanimate categories based on the MPPI score. $m$ was then chosen to be the number of clusters that produced a number of intersection units closest to the average number of units in each cluster $a_{k=256}$ (22,365). For animate-selective clusters, 23,926 common units (Figure 2b) were identified at $m=23$. For inanimate-selective clusters, 21,672 common units (Figure 2b) were found at $m=13$.

An ablation of the units common across the top-23 most selective animate clusters reduced top-5 performance on animate test images from 87.17% (non-ablated) to 0% (Figure 2c). This same animate-specific ablation had a smaller impact on performance on inanimate images (69.23% non-ablated to 35.89% animate-selective ablated). An ablation of the units common across the 13 inanimate-selective clusters, reduced performance on inanimate test images from 69.23% to 5.12% (Figure 2c). Ablating the inanimate-specific group of units had little impact on animate performance (87.17% non-ablated to 66.67% inanimate specific ablated).

To further confirm the functional selectivity of these units, we visualized which parts of the images were utilized during categorization by whole and ablated networks. Saliency maps are visual representations of the internal state of a neural network [26]. Here, we compare the saliency maps of DNNs pre and post ablation (Figure 3). We generated these maps using gradient back-propagation starting with the output unit and tracing it back to the input [26]. We look at samples from the test set which were correctly classified pre-ablation (indicated by green dotted outline) and misclassified post-lesion (indicated by red solid outline). We found that an ablation to the units with the maximally selective impact for the animate category leads to loss in salience across relevant pixels for animate test samples (Figure 3 top image row), likely contributing to misclassification. Ablation of the inanimate-specific units did not reduce salient pixels in the target area of the example image. For inanimate images the opposite was true. The inanimate-specific ablation altered salience for areas of the inanimate image but the animate-specific ablation had a much smaller effect on the inanimate image (Figure 3, bottom image row). It is interesting to note that salience on the inanimate images is commonly limited to the areas surrounding the object of interest. These results were consistent for the majority of images.

Selective units were not uniformly distributed across network layers. Both animate and inanimate specific units were found to be a higher proportion of the layer in the later layers of the network.
Figure 3: Salience maps of test samples for animate and inanimate ablations in MobileNetV2

Figure 4: Distribution of animate-specific (left) and inanimate-specific (right) units distributed across the 104 layers of MobileNetV2.

(Figure 4). These results support previous findings in which representations from later layers were found to be better able to distinguish image categories [27].

4 Conclusion

Using a novel ablation method, we were able to isolate groups of units that when ablated produced performance deficits with strong selectivity. The observed selectivity (62%), measured as the difference in test performance between animate and inanimate categories, compares favorably to single-unit and unit-pair performance differences produced during exhaustive ablations in simpler networks (e.g., 44.5% in [28]). These functionally selective ablations led to changes in image salience for the image type targeted. In accordance with previous research on representation similarity [16], the identified selective units made up a larger fraction of later layers. Our results show that clusters of functionally selective units with confirmed causal effects are identifiable in a large DNN. We note that a similar procedure could be used for other applications on DNNs such as targeted printing and model interpretability.
References


