Using pre-trained neural networks to validate the impact of distinct clusters of neurons for object recognition task

Abhijit Suresh, R. McKell Carter

Institute of Cognitive Science, Department of Psychology and Neuroscience, Department of Computer Science, University of Colorado Boulder

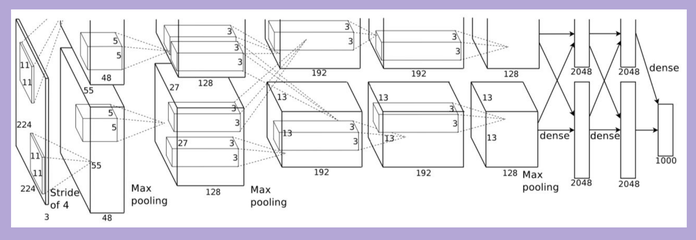
How are functions structured in an artificial neural network?

Motivation

- The task of identifying the selective importance of a group of neurons in the human brain requires lesion which is not possible.
- Leveraging results from measures of functional importance in artificial neural networks like TCAV and LRP, we seek to extend these solutions to the network as a whole for super-categories of images..

Data

- We begin with Alexnet, a convolutional neural network trained to classify a given image into one of the 1000 categories.



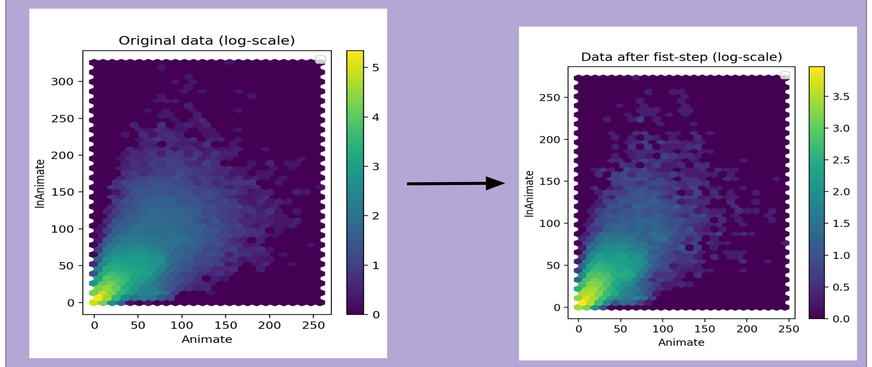
- We then work to characterize Alexnet functions for two super-categories of images.

Animate

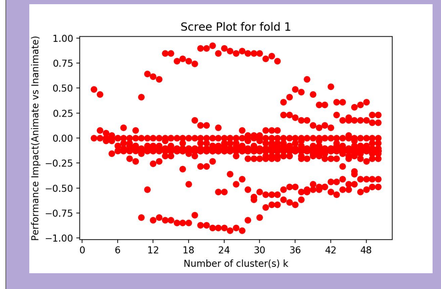
Inanimate

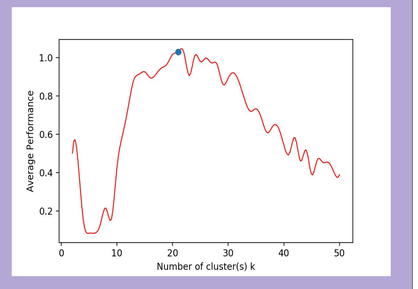
Approach: lesion units with similar activation for a targeted super-category

- Artificial lesioning is the process of removing a specific cluster of neurons from participating in the classification process.
- Alexnet has 658,272 neurons. As the first step, we use k-Means to perform a 10-fold data reduction yielding 65827 summary data points.

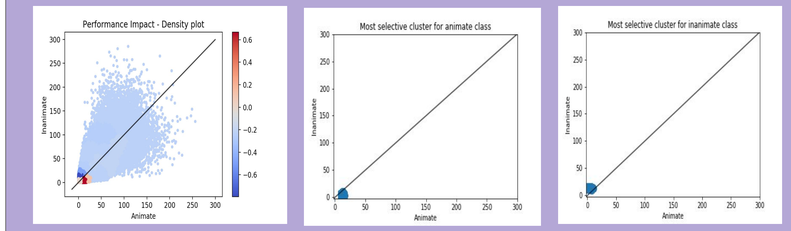


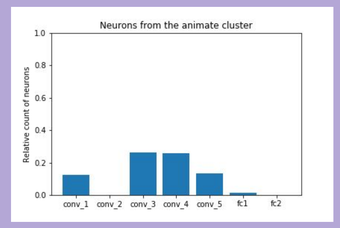
- Using the results obtained from the first step, we proceed to cluster further using algorithms such as mixture models, ward clustering and HDBSCAN.
- We perform 4-fold cross validation on the training set to identify the number of clusters (k) for which it is possible to identify a cluster that can be lesioned with the largest impact on only one of the two super-categories. The optimal k occurs where the performance distance between clusters is maximal.

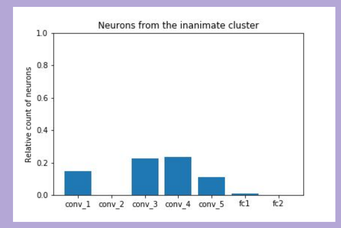


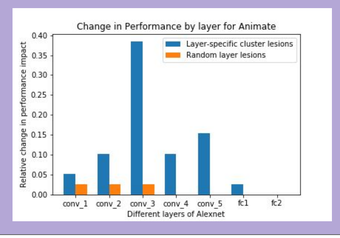


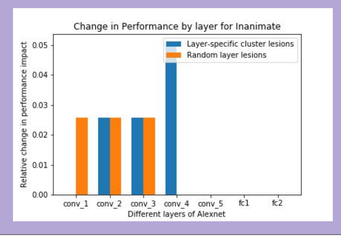
Results - Measuring performance deficit











Discussion

1. Using this method, we are able to identify clusters of neurons in a pre-trained visual object recognition network (AlexNet) whose 'lesioning' creates substantial deficits in performance (sensitivity reductions of 0.6-0.8) specific to the targeted object class.
2. We identified that the neurons belonging to the middle layers lesioned with the proposed method have more impact on the classification of animate images.
3. The absence of layer-specific impacts in the identification of inanimate objects leads to two surprising conclusions.
   a. Some functional selectivity in a network is not possible to identify using techniques like TCAV.
   b. It is possible to isolate units specific to functions spread across an ANN like AlexNet.