Is Robustness to Transformations driven by Invariant Neural Representations?

Syed Suleman Abbas Zaidi[1,2,*], Xavier Boix[1,3,*], Neeraj Prasad[1], Sharon Gilad-Gutnick[1], Shlomit Ben-Ami[1], Pawan Sinha[1]

[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology (USA), [2]Department of Electrical and Computer Engineering, Technical University of Munich (Germany), [3]Center of Brains, Minds and Machines, Massachusetts Institute of Technology (USA)

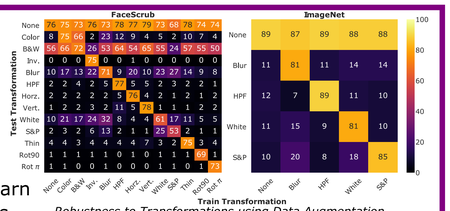
https://arxiv.org/abs/2007.00112

Overview

Premise

Deep Convolutional Neural Networks (DCNNs) can achieve remarkable robustness to transformations using data augmentation. However, how this robustness is achieved is largely unknown.

A hypothesis from the field of Neuroscience is that DCNNs learn neural representations which are invariant to transformations.



Robustness to Transformations using Data Augmentation

Contribution

We study the role of invariant neural representations on two fronts: 1) We test DCNNs for Across-Category Robustness, a phenomenon which is a direct consequence of invariance. 2) We define a metric to directly measure invariance on a single-neuron level.

We found that invariance emerges as the network sees more categories with transformations and the strength of invariance depends on the type of transformation.

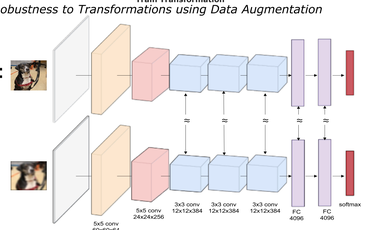


Illustration of Invariant Neural Representations

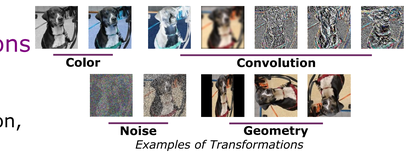
Methods

Network Architectures, Datasets and Transformations

Datasets: FaceScrub and ImageNet

Architectures: AlexNet (FaceScrub) and ResNet (ImageNet)

Transformations: Grayscale, Color Rotation, Contrast Inversion, Low-Pass, High-Pass, Horizontal and Vertical Filtering, White Noise, Salt and Pepper Noise, Thinning, 90° and 180° Rotation.



Examples of Transformations

Paradigm to Evaluate the Across-Category Robustness

Robustness: Ability of a DCNN to correctly classify transformed images measured by test accuracy.

Seen-Transformed Categories: Categories that the network has seen both with and without transformation during training.

Unseen-Transformed Categories: Categories that the network has only seen without transformation.

Within-Category Robustness: Test accuracy of a DCNN on seen-transformed categories.

Across-Category Robustness: Test accuracy of a DCNN on unseen-transformed categories.

Paradigm: We evaluate the Across-Category Robustness in DCNNs with increasing number of seen-transformed categories.

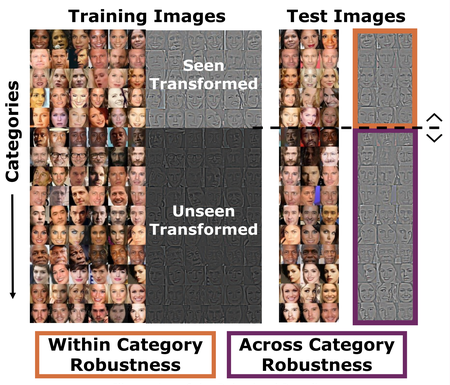
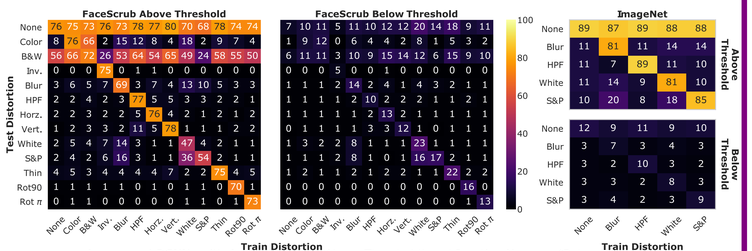


Illustration of the Behavioral Paradigm.

Invariance to Transformations at the Individual Neural Level

Active Neurons:
A neuron is active for an image if its activation is above a certain threshold such that any sub-threshold activity could be silenced without a significant drop in the DNN's performance.



Accuracy of DCNNs with Active Neurons (Above Threshold) and Inactive Neurons (Below Threshold)

Invariance of an Active Neuron:
If $f_k(x)$ is the activation of neuron k for image x and T(x) is the transformed version of image x, then the invariance measure $h_k(x)$ of neuron k given image x is calculated as:

$$h_k(x) = 1 - \left| \frac{f_k(T(x)) - f_k(x)}{f_k(T(x)) + f_k(x)} \right|$$

The value $h_k(x)$ is high if the activation of neuron k for x and T(x) are close to each other. The Invariance coefficient I_k of a neuron is calculated by averaging over all images for which neuron k is active.

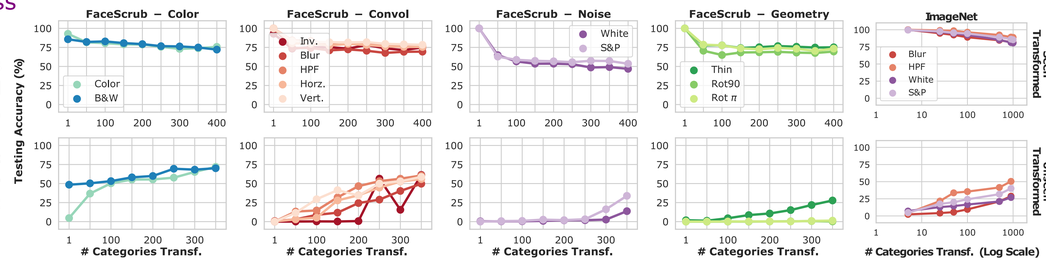
Results

Results on Across-Category Robustness

Across-Category Robustness Improves as the Number of Seen-transformed Categories is Increased.

Color, Convolutional and Noise Transformations lead to Higher Across-Category Robustness than Geometric Transformations.

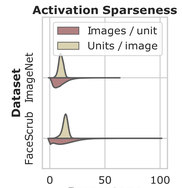
Across-Category Robustness depends on the Visual Homogeneity of the Object Categories.



Within-category robustness (top) and Across-category robustness (bottom) in FaceScrub and ImageNet, against number of seen-transformed categories in the training set

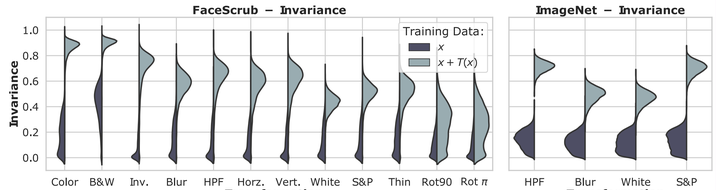
Results on Invariance to Transformations at the Individual Neural Level

Less than 20% of the Images Activate a Neuron.
Neurons are only active for a few images because they are tuned to specific features.
Only a few neurons are active for each image.



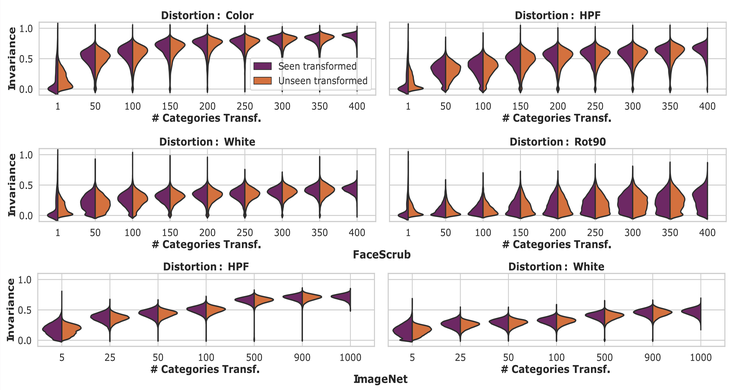
Distribution of Active Neurons

Invariance due to Data Augmentation depends on the Transformation Type
Data Augmentation gives rise to invariant neural representations. Strength of invariance is high for color and convolutional transformations where as low in geometric ones.



Violin plots illustrating the distribution of the invariance coefficient of neurons in the penultimate layer, for FaceScrub and ImageNet, respectively

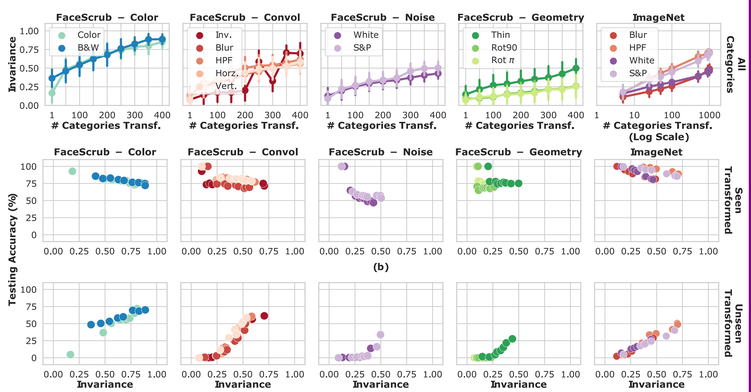
Invariance Increases when the Number of Seen-transformed Categories Increases.
The invariance measure remarkably increases when the number of seen-transformed categories is increased. Invariance is identical for seen and unseen-transformed categories. Invariance is category independent.



A few violin plots of the invariance coefficient for FaceScrub and ImageNet against number of seen-transformed categories

Invariance Explains the Across-Category Robustness.

The trends of invariance follow that of across-category robustness. Invariance is not needed to achieve within-category robustness. Extent of across-category robustness is correlated with the amount of invariance for all transformations



(Row 1) Average invariance for different number of seen-transformed categories, transformations and datasets.
(Row 2&3) Relationship between invariance and within and across-category robustness

Conclusion

Increasing seen-transformed categories increases robustness to unseen-transformed categories and makes the neural representations more invariant to transformations.

Invariance drives across-category robustness.

Invariance emerges at the individual neural level, signifying the importance of individual neurons in DCNN analysis.

Future Work

Understanding why invariant representations emerge for some transformations.

What are the neural mechanisms for robustness in the absence of invariance (e.g. for geometric transformations).

Acknowledgements

We are grateful to Tomaso Poggio for his insightful advice & warm encouragement. This work is supported by the Center for Brains, Minds & Machines (funded by NSF STC award CCF-1231216), Fujitsu Laboratories Ltd. (Contract No. 40008401 & 40008819), the MIT-Sensetime Alliance on AI & the R01EY020517 grant from National Eye Institute (NIH).