Quantifying the quality of shape and texture representations in deep neural networks

Fenil R. Doshi (fenil_doshi@fas.harvard.edu), Talia Konkle, George A Alvarez

Department of Psychology, Harvard University

VISION LAB

Kempner INSTITUTE

Introduction

Humans can perceive objects by their shape, despite variations texture.

DNNs are emerging as de-facto models of human perception, but are known to over-rely on texture

We propose two metrics to measure shape in DNNs :

1. Corrected Shape Bias
2. Configural Shape Index

Beyond Shape Bias

While the Shape-Bias metric has been very useful, it has several broader issues

1. Shape-bias requires output classification

Can we provide a measure based on activations that can be probed at each layer, and in self-supervised models, without fine-tuning?

2. Shape is operationalized *against* Texture.

Models can be correct by shape or by texture but not both, but it's possible to represent both shape and texture well: Can we measure shape and texture representations *independently*?

3. What qualities define strong shape representations?

Going forward, we would like establish clearer desiderata for strong shape representations: To begin, we propose that a strong shape-representation ought to have ***high tolerance to shape-preserving affine transformations*** (e.g., changes in position, orientation and scale), and **low tolerance (high sensitivity) to *shape-destructive transformations*** (e.g., scrambling parts of an object/scene).

Metric 1: Accuracy-Scaled Shape Bias

Standard Shape Bias (Geirhos et al., 2018)

Cue-Conflict Stimuli

Example: Task: classify object category

$$\text{Standard Shape Bias} = \frac{\text{total \# correct shape decisions}}{\text{total \# correct decisions (total correct shape + total correct texture)}}$$

Problem:

Standard Shape Bias doesn't account for accuracy

	Model A	Model B
Total Trials	1200	1200
# Correct Shape Decisions	**1**	**300**
# Correct Texture Decisions	0	300
Standard Shape Bias	**1.0**	**0.5**
Accuracy-Scaled Shape Bias	**0.028**	**0.35**

- The objectively poor Model A (1 correct response total), has a higher standard shape-bias score than the stronger Model B.
- Standard Shape Bias scores for Untrained-Resnet50 (0.52) and SSL-Resnet50 (0.217) !?

Proposed Adjustment:

Scale the shape bias by overall shape-accuracy so that the score reflects *bias* and *accuracy.*

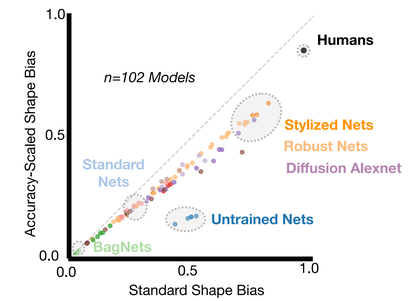
Accuracy-Scaled Shape Bias

$$\textit{standard shape-bias} \times \textit{overall shape-accuracy}$$

$$\sqrt{\frac{\text{\# Correct Shape Decisions}}{\text{Total \# Correct (shape + texture)}}} \times \sqrt{\frac{\text{\# Correct Shape Decisions}}{\text{Total Number of Trials}}}$$

* *square root keeps the score on a 0-1 scale*

Standard Shape Bias vs. Accuracy-Scaled Shape Bias

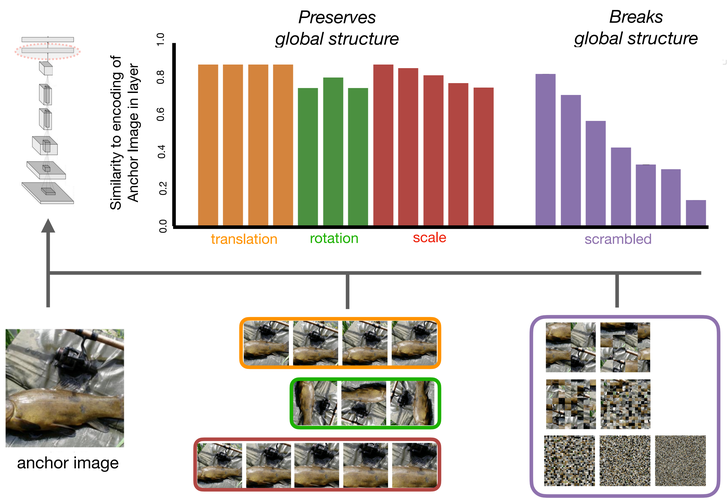


- **Untrained models scores are now corrected**
- **Overall order across models is maintained (r=0.94)**

Metric 2: Configural Shape Index

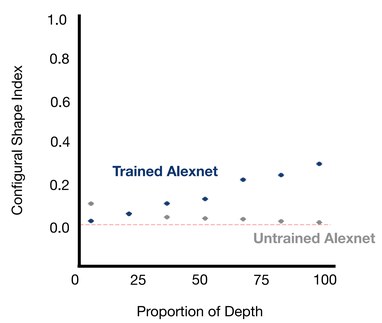
Key Idea:

The encoding of an anchor image should be similar to itself over translation, rotation, and scale; and different from itself when scrambled.

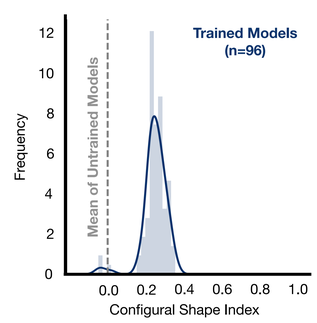


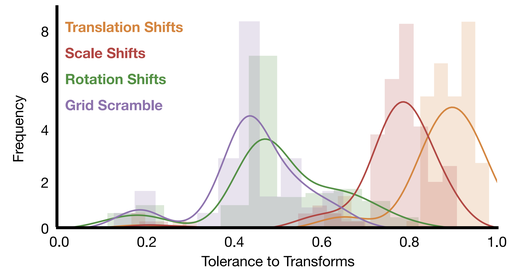
$$\textit{Configural Shape Index} = \textit{Tolerance (Translation, Rotation, Scale)} - \textit{Tolerance (Scrambling)}$$

Configural Shape Index increases across layer hierarchy in trained Alexnet



Across models, penultimate layers all have similar, relatively low, configurable shape information



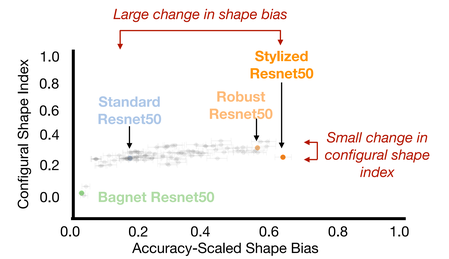


Across models, tolerance to translation and scale variation is good.

However, rotated images are as similar as scrambled images, highlighting the main areas for improvement

Metric Comparison

Training strategies that increase shape bias do not increase configural shape index



Conclusions

- If you're measuring shape bias, consider reporting **Accuracy-Scaled Shape-Bias.**
- For more graded measures, consider **Configural Shape Index** - stricter metric for measuring the quality of shape representations
- Instead of conceiving of this as one axis; think of it as two separate capacities

Shape ⟷ Texture (Vs.)

Texture Sensitivity ↑ / Shape Sensitivity →