Texture vs shape bias in computer vision networks

Aksenenko Veronika 211

How CNNs recognise objects?

Shape hypothesis - CNNs combine low-level features to complex shapes until the object can be classified

Texture hypothesis - object textures are more important than global object shapes for CNN object recognition

Cue conflict

(c) Texture-shape cue conflict

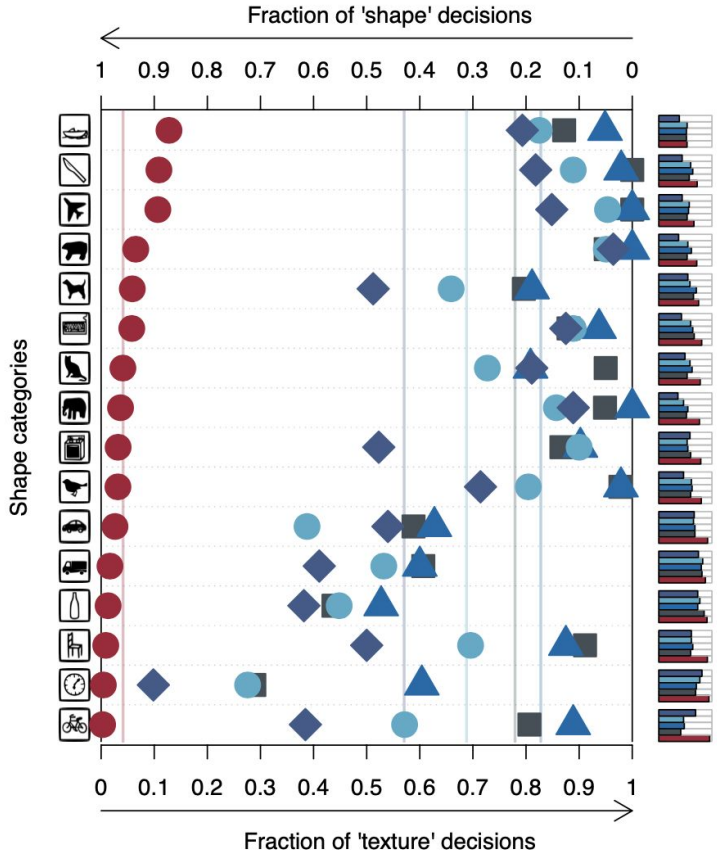
63.9%	**Indian elephant**
26.4%	indri
9.6%	black swan

(a) Texture image

81.4%	**Indian elephant**
10.3%	indri
8.2%	black swan

(b) Content image

71.1%	**tabby cat**
17.3%	grey fox
3.3%	Siamese cat



Human observers (red circles)

AlexNet (purple diamonds)

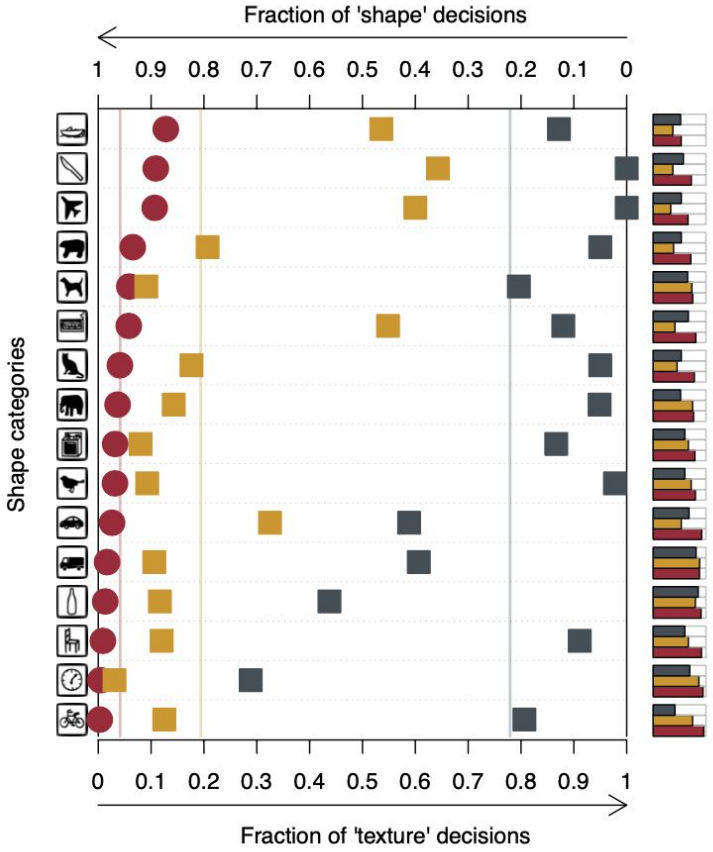
VGG-16 (blue triangles)

GoogLeNet (turquoise circles)

ResNet-50 (grey squares).

How to induce texture bias?

- Texture recognition is easier than shape recognition?
- Stylized ImageNet



ResNet-50 on Stylized-ImageNet (orange squares)

ResNet-50 on ImageNet (grey squares)

Human data (red circles)

Some more results

architecture	IN→IN	IN→SIN	SIN→SIN	SIN→IN
ResNet-50	92.9	16.4	79.0	82.6
BagNet-33 (mod. ResNet-50)	86.4	4.2	48.9	53.0
BagNet-17 (mod. ResNet-50)	80.3	2.5	29.3	32.6
BagNet-9 (mod. ResNet-50)	70.0	1.4	10.0	10.9

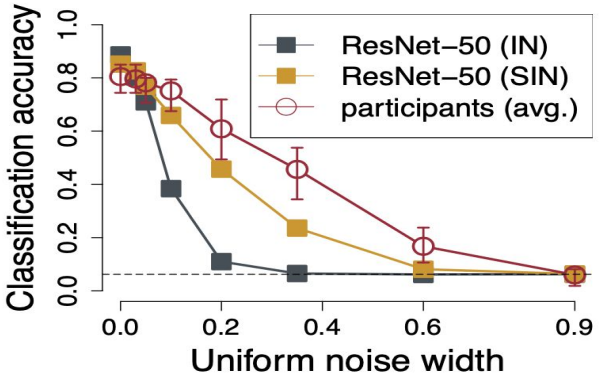
- SIN is a much harder task than IN since textures are no longer predictive
- ImageNet features generalise poorly to SIN
- SIN data set we propose does actually remove local texture cues, forcing a network to integrate long-range spatial information

Are there any benefits of shape bias?

+ accuracy

+ robustness against distortions

			object recognition		object detection	
name	training	fine-tuning	top-1 IN accuracy (%)	top-5 IN accuracy (%)	Pascal VOC mAP50 (%)	MS COCO mAP50 (%)
vanilla ResNet	IN	-	76.13	92.86	70.7	52.3
	SIN	-	60.18	82.62	70.6	51.9
	SIN+IN	-	74.59	92.14	74.0	53.8
Shape-ResNet	SIN+IN	IN	**76.72**	**93.28**	**75.1**	**55.2**



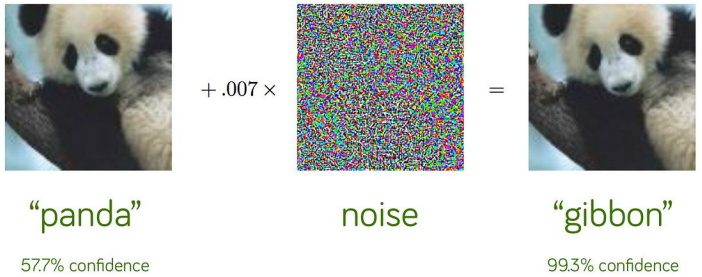
(a) Uniform noise

Results

- CNNs mostly recognise texture rather than shapes
- We can develop shape bias with suitable dataset
- Shape bias is beneficial

Why texture bias could cause problems?

- Vulnerable to adversarial attacks

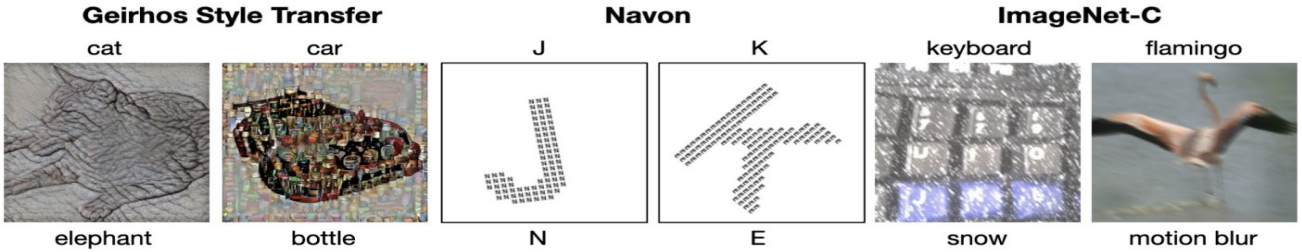


- Makes it difficult for models to learn human-relevant vision task
- CNNs model of choice for modelling primate visual cortex
- Sensible to compare human and machine vision

Reasons of texture bias

- Datasets
- Augmentation
- Objective
- Architecture

Datasets

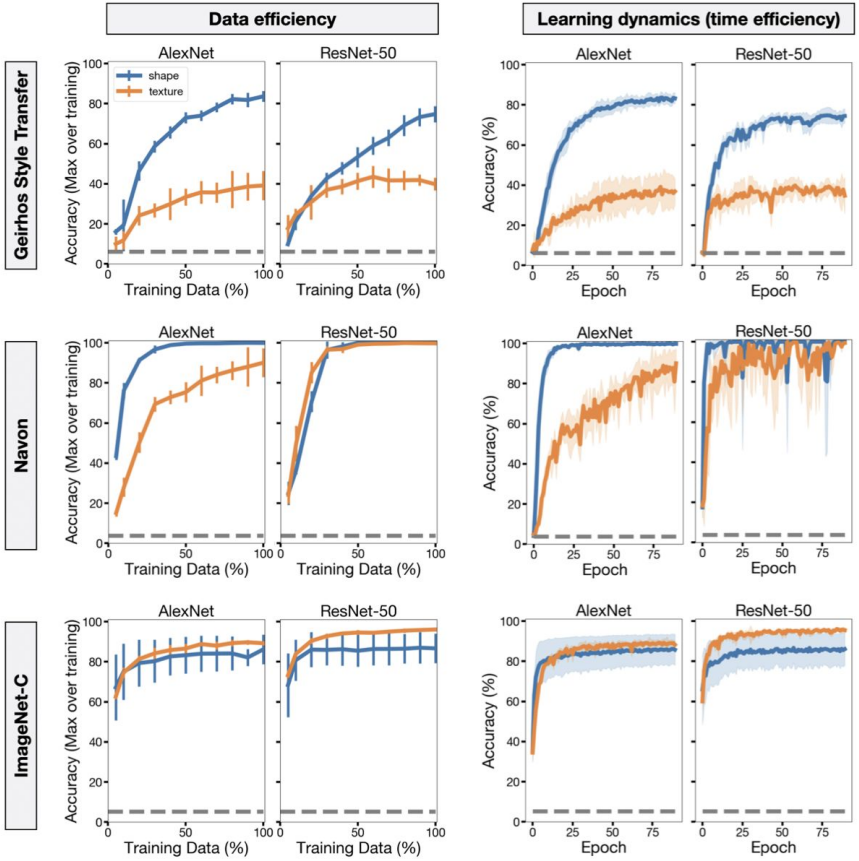


Shape bias - the percentage of the time it classified images from the GST dataset according to shape, provided it classified either shape or texture correctly

We call a model *shape-biased* if its shape bias is > 50%, and *texture-biased* if it is < 50%

Shape match and *texture match* indicate - the percentage of the time the model chose the image's correct shape or texture label, respectively, as opposed to choosing some other label

Datasets


Data efficiency
Learning dynamics (time efficiency)
Geirhos Style Transfer
Navon
ImageNet-C
AlexNet
ResNet-50
shape
texture
Accuracy (Max over training)
Accuracy (%)
Training Data (%)
Epoch

Role of augmentation

- Random-crop data augmentation increases texture bias
- Appearance-modifying data augmentation reduces texture bias

Model	Shape Bias		Shape Match		Texture Match		ImageNet Top-1 Acc.	
	Random	Center	Random	Center	Random	Center	Random	Center
AlexNet	28.2%	**37.5%**	16.4%	**19.3%**	**41.8%**	32.1%	**56.4%**	50.7%
VGG16	11.2%	**15.8%**	7.6%	**10.7%**	**60.1%**	57.1%	**71.8%**	62.5%
ResNet-50	19.5%	**28.4%**	11.7%	**16.3%**	**48.4%**	41.1%	**76.6%**	70.7%
Inception-ResNet v2	23.1%	**27.9%**	15.1%	**19.8%**	50.2%	**51.2%**	**80.3%**	77.3%

Augmentation	Shape Bias	Shape Match	Texture Match	ImageNet Top-1 Acc.
Baseline	19.5%	11.7%	48.4%	76.6%
Rotate 90°, 180°, 270°	19.4%	10.8%	45.1%	75.7%
Cutout	**21.4%**	12.3%	45.2%	76.9%
Sobel filtering	**24.8%**	12.8%	38.9%	71.2%
Gaussian blur	**25.2%**	14.1%	41.7%	75.8%
Color distort.	**25.8%**	15.3%	44.2%	76.9%
Gaussian noise	**30.7%**	17.2%	38.8%	75.6%

Effect of training objective

Rotation classification

Guess rotation

Exemplar

Triplet Loss

BigBiGAN

Generator - converts latent codes into images

Encoder - converts images to latent codes

SimCLR

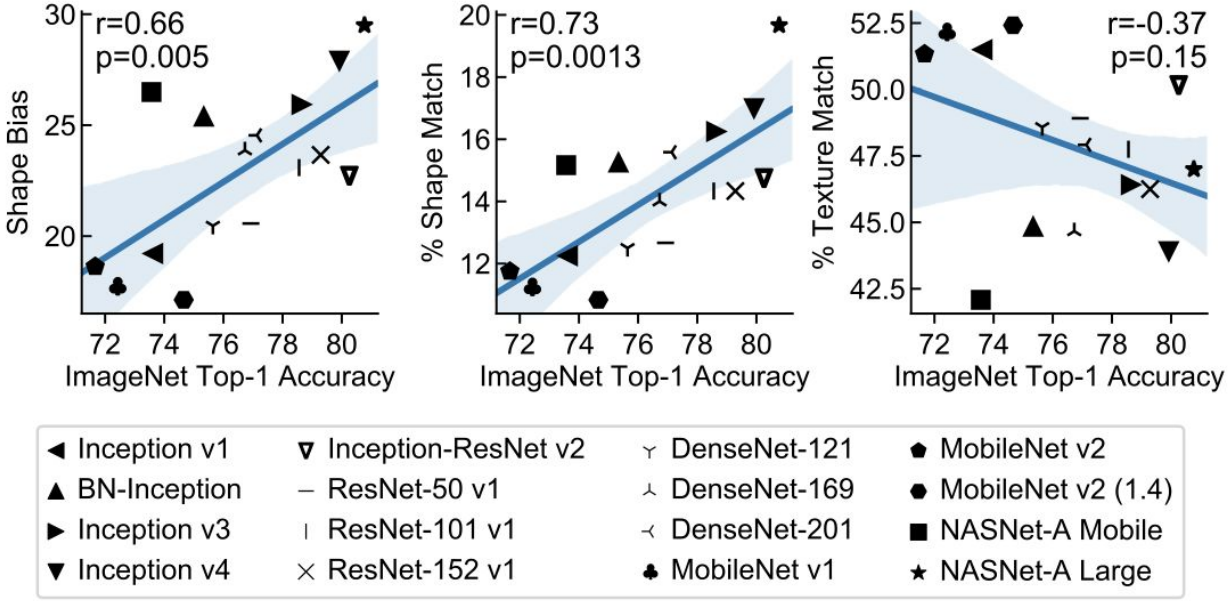
Learns representations by maximizing agreement between differently augmented views of the same data example

Effect of training objective

- Rotation model had significantly lower texture bias than supervised models
- Shape bias is higher for AlexNet than for ResNet-50

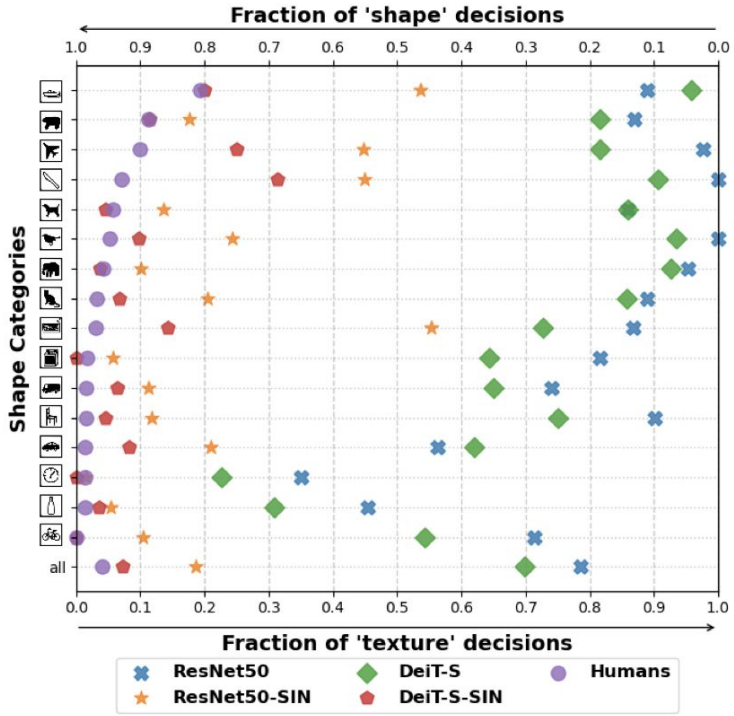
Objective	Shape Bias		Shape Match		Texture Match		ImageNet Top-1 Acc.	
	AlexNet	ResNet-50	AlexNet	ResNet-50	AlexNet	ResNet-50	AlexNet	ResNet-50
Supervised	29.8%	21.9%	17.5%	13.5%	41.2%	48.2%	57.0%	75.8%
Rotation	47.0%	32.3%	21.6%	14.2%	24.3%	29.8%	44.8%	44.4%
Exemplar	29.9%	14.4%	12.6%	7.5%	29.5%	44.7%	37.2%	41.8%
BigBiGAN	-	31.9%	-	17.7%	-	37.7%	-	55.4%
SimCLR	-	37.0%	-	17.3%	-	29.4%	-	69.2%
Supervised w/ SimCLR aug.	-	40.4%	-	23.1%	-	34.0%	-	76.3%

Effect of architecture



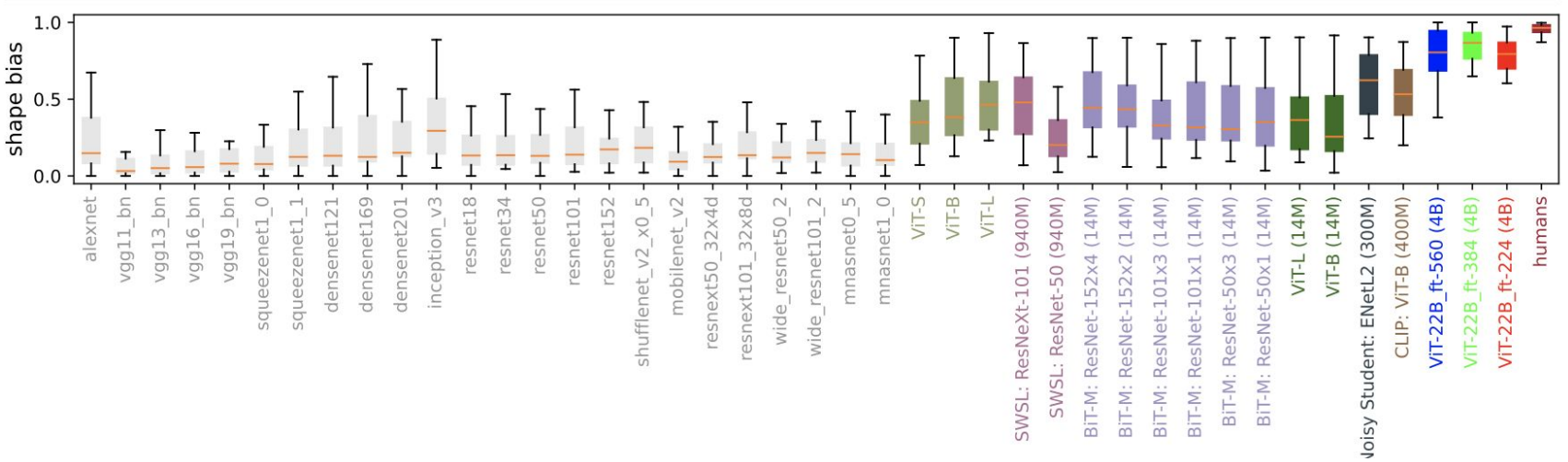
Vision transformers (ViT)

- ViTs are significantly less biased towards local textures, compared to CNNs
- ViTs demonstrate shape recognition capability comparable to that of human visual system



Vision transformers (ViT)

- ViT-22B fine-tuned on ImageNet (red, green, blue) have the highest shape bias recorded in a ML model to date



Results

- Shape bias is useful
- Shape bias depends on various factor
- ViTs perform best of all

Hyperparameters