Edges to Shapes to Concepts: Adversarial Augmentation for Robust Vision

Aditay Tripathi*, Rishubh Singh, Anirban Chakraborty, Pradeep Shenoy

*Work done at Google Research India.

Texture bias in vision models

(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

(c) Texture-shape cue conflict

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

Image credits: Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. (ICLR 2019)

ELeaS enhance shape sensitivity in vision models

$\lambda *$

$+ (1-\lambda) *$

$=$

Distinguishing the augmented image entails differentiating the **relevant edges** representing the overall object shape from the **irrelevant edges** derived from the shuffled image.

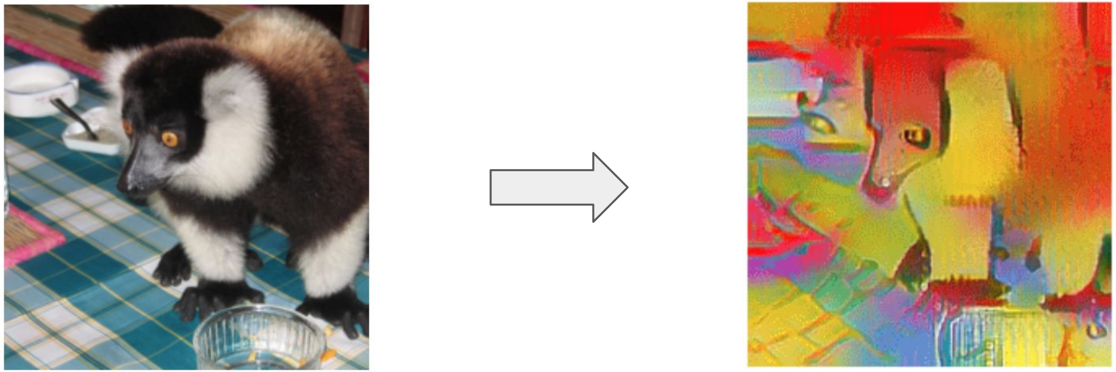
Deep networks v.s. Human behaviour

- Deep networks prioritize "local" features over global features, differing from human behavior.
- Image datasets like Imagenet may not accurately reflect cognitive concepts and real-world knowledge.
- Inductive biases are necessary in under-determined learning problems to guide the learning process.

Jacob, Georgin, et al. "Qualitative similarities and differences in visual object representations between brains and deep networks." *Nature communications* 12.1 (2021): 1872.

Related work

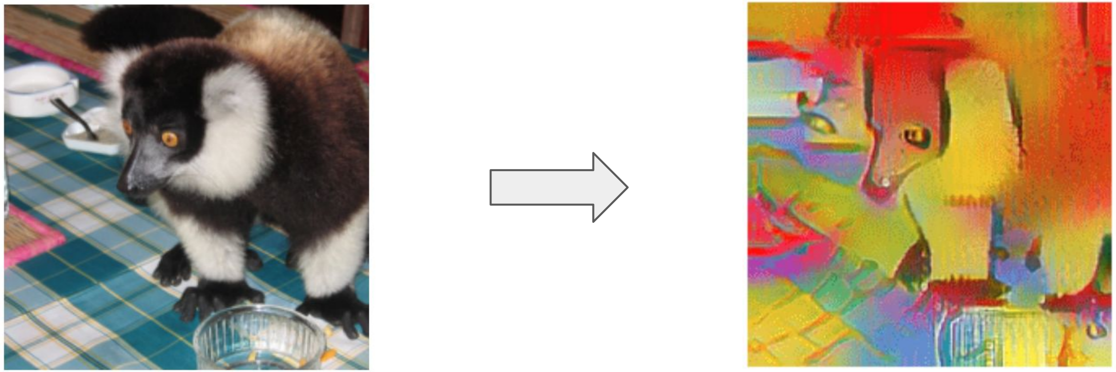
- Geirhos et al. proposed a data augmentation method that replaced an image's texture with a painting's texture through stylization.
- Later work expanded on this approach by replacing textures from other objects, not just paintings.



Geirhos, Robert et al. “ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.” *ArXiv* abs/1811.12231 (2018): n. pag.

Related work

- These approaches discourage relying too heavily on textural features in the learned model.
- However, they do not explicitly encourage or incentivize shape recognition.



Proposed augmentation: ELeaS

- ELeaS (Edge Learning for Shape sensitivity), aims to enhance shape sensitivity in vision models.
- The two images are combined using a randomly sampled mixing weight.

$$i_s = \lambda * t + (1 - \lambda) * s$$

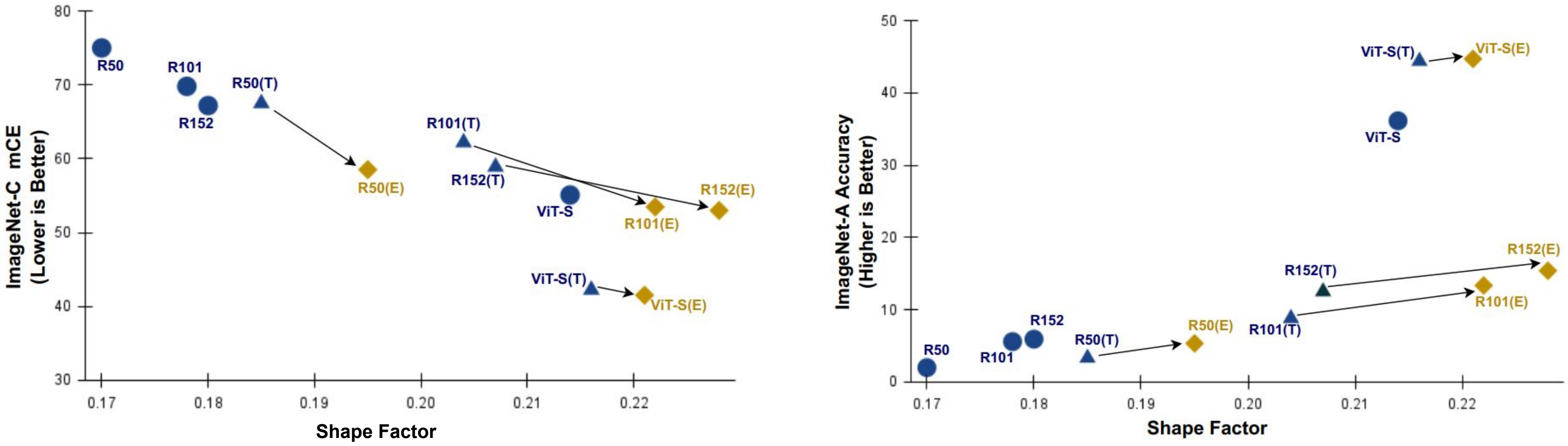
Im 1 | Im 1-Edgemap (***s***) | Im 2 | Im 2-Shuffled (***t***) | ELeaS ($\boldsymbol{i_s}$)

Proposed training strategy

- Each minibatch consists of a combination of natural images from **set I** and augmented images from **set B**.
- The training process minimizes the cross-entropy loss on both natural image samples and the augmentations.
- To control the induced shape sensitivity, a weighted mixture of cross-entropy loss is computed on the two image sets.

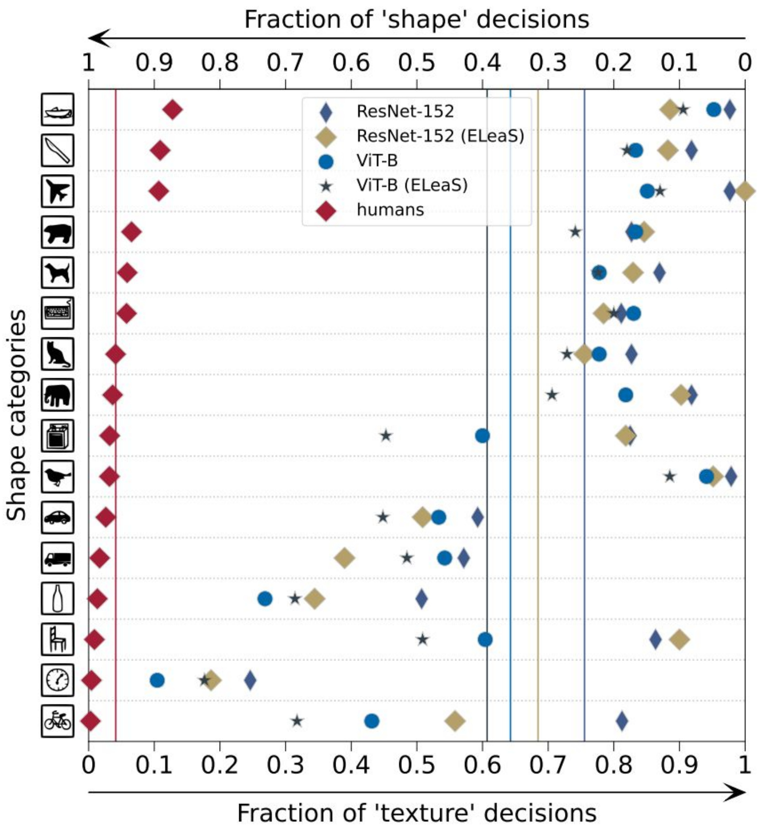
$$L(I, B, y_I, y_B) = \eta * CE(I, y_I) + (1 - \eta) * CE(B, y_B)$$

Shape sensitivity v.s. Robustness



Islam, Md Amirul, et al. "Shape or texture: Understanding discriminative features in cnns." *arXiv preprint arXiv:2101.11604* (2021).

Improved shape-sensitivity



Experimental results

Model	Method	IN-A(↑)	IN-R(↑)	IN-C(↓)	IN-Sketch(↑)	IN-1K(↑)
Resnet50	Vanilla	2.0	36.2	75.0	23.5	76.4
	TSD [21]	3.3	40.8	67.5	28.3	76.9
	ELEAS	**5.4**	**41.7**	**58.5**	**29.7**	**77.1**
Resnet101	Vanilla	5.6	39.3	69.8	27.1	78.0
	TSD [21]	8.8	44.3	62.2	32.3	**78.8**
	ELEAS	**13.4**	**44.4**	**53.5**	**32.4**	78.6
Resnet152	Vanilla	5.9	41.3	67.2	28.4	78.6
	TSD [21]	12.5	45.5	58.9	33.3	79.7
	ELEAS	**15.4**	**45.7**	**53.0**	**34.7**	79.0
ViT-S	Vanilla	16.6	36.1	55.1	33.2	74.6
	Vanilla FT	27.6	43.8	44.3	34.7	80.6
	TSD [21]	27.4	44.4	42.2	32.4	76.4
	ELEAS	**28.5**	**45.0**	**41.5**	**35.3**	**81.1**

Experimental results

Model	Method	IN-A(↑)	IN-R(↑)	IN-C(↓)	IN-Sketch(↑)	IN-1K(↑)
Resnet50	Vanilla	2.0	36.2	75.0	23.5	76.4
	TSD [21]	3.3	40.8	67.5	28.3	76.9
	ELEAS	**5.4** +2.1	**41.7**	**58.5** -9.0	**29.7**	**77.1**
Resnet101	Vanilla	5.6	39.3	69.8	27.1	78.0
	TSD [21]	8.8	44.3	62.2	32.3	**78.8**
	ELEAS	**13.4** +4.6	**44.4**	**53.5** -8.7	**32.4**	78.6
Resnet152	Vanilla	5.9	41.3	67.2	28.4	78.6
	TSD [21]	12.5	45.5	58.9	33.3	79.7
	ELEAS	**15.4** +2.9	**45.7**	**53.0** -5.9	**34.7**	79.0
ViT-S	Vanilla	16.6	36.1	55.1	33.2	74.6
	Vanilla FT	27.6	43.8	44.3	34.7	80.6
	TSD [21]	27.4	44.4	42.2	32.4	76.4
	ELEAS	**28.5** +0.9	**45.0**	**41.5** -0.7	**35.3**	**81.1**

Segmentation and Detection performance

- Only backbone model is changed.
- The models are evaluated on the COCO-Val2017 dataset.

Model	Object Detection			Instance Segmentation		
	mAP	AP@0.50	AP@0.75	mAP	AP@0.50	AP@0.75
Vanilla	39.87	60.21	43.33	36.35	57.39	38.79
TSD [21]	37.82	58.98	41.29	33.87	55.41	35.85
ELEAS	**41.65** +1.78	**61.83**	**45.43**	**37.63** +1.28	**58.99**	**40.33**

Improved shape-sensitivity leads to improved Object detection and segmentation performance for free.

Conclusion

ELeaS training leads to improved shape sensitivity.

Improved shape sensitivity leads to improved model robustness.

Enhanced shape sensitivity improves segmentation and detection performance.