PAPER

▼ IMAGENET-TRAINED CNNS ARE BIASED TOWARDS TEXTURE; INCREASING SHAPE BIAS IMPROVES ACCURACY AND ROBUSTNESS (2019)

코끼리의 Texture를 가진 고양이는 CNN을 사용한 Computer의 Vision 상에서는 코끼리이고, 사람의 Human Vision으로 봤을 땐 고양이이다.

(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

Style Transfer 한 마지막 사진 → Indian Elephant로 분류함.

현재의 CNN(Convolution Neural Network)는 이미지를 분류하는데 있어 이미지의 shape가 아닌 texture에 biased 되어 있다.

classification을 수행하는 데 있어 shape 보다는 texture의 정보를 많이 사용한다.

하지만 Human vision은 그렇지 않다. Human Vision은 Classification에 특화되어 있다. 다중의 사물에 대해 정말 빠르게 Classification 작업을 수행한다. 사람들은 물체를 볼 때 texture 보다 shape 정보를 사용한다.

CNN 네트워크의 Object Recognition을 수행하는 데 집중해야할 것은 Shape? / Texture? 두가지 가설이 존재한다.

가설 1. Shape Hypothesis

CNN 네트워크는 low-level feature 부터 높은 level의 feature로 순차적으로 object를 인식한다.

Ex) Edges → wheels/windows → Car

CNN 네트워크를 처음 설계할 때부터 Human Vision을 포방하여 만들었기 때문에 CNN 네트워크가 texture 보다 shape를 더 특정하여 학습하도록 해야한다.

가설 2. Texture Hypothesis

CNN 네트워크는 texture가 살아있기만 한다면, shape가 없어지거나 무너져도 안정적으로 Image를 Classification 할 수 있다.

반대로, texture가 무너진다면, shape가 살아있어도 좋은 결과를 뽑아낼 수 없다.

하지만 Image를 통한 Object Recognition이라는 목표를 달성하는 데 있어서 Texture 정보만 사용하여도 충분하다.

두 가지 상충되는 가설을 실험적으로 어떤 가설이 올바른 가설인지 소개 및 해결방안을 제시하는 논문이다.

Using PSYCHOPHYSICAL 실험을 통해 (정신물리학적 실험)

실험자(사람) 들은 주어진 정사각형의 그림을 본 후 16개의 Label 중 선택해야 한다.

200ms	**200ms**	**1500ms**
Stimulus image	Pink noise mask	Select label

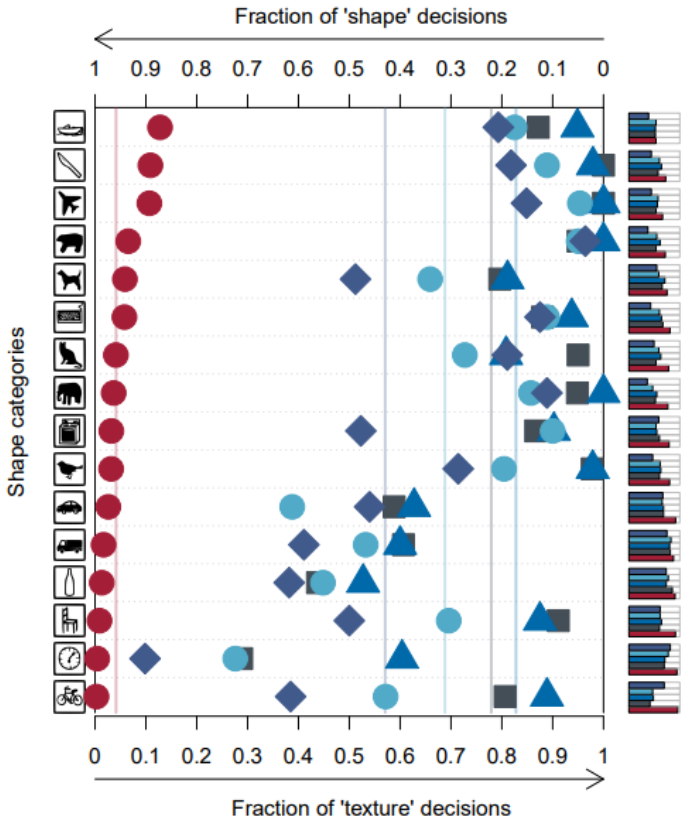
Human Vision의 feedback 시스템을 최소화하고, CNN과 공정하게 실험하기 위해 중간에 200ms의 Pink noise를 추가하였다.

Figure 3: Visualisation of Stylized-ImageNet (SIN), created by applying AdaIN style transfer to ImageNet images. Left: randomly selected ImageNet image of class `ring-tailed lemur`. Right: ten examples of images with content/shape of left image and style/texture from different paintings. After applying AdaIN style transfer, local texture cues are no longer highly predictive of the target class, while the global shape tends to be retained. Note that within SIN, every source image is stylized only once.

실험에 사용된 Image 들은 [Grayscale, Silhouette, Edges, Texture] 을 변경하여 만든 데이터셋이다. (Stylized-ImageNet)

Manupulate 한 이미지를 바탕으로 Human Vision과 CNN의 Shape bias/Texture bias에 대해 실험적으로 Classification을 수행했다.

Figure 4: Classification results for human observers (red circles) and ImageNet-trained networks AlexNet (purple diamonds), VGG-16 (blue triangles), GoogLeNet (turquoise circles) and ResNet-50 (grey squares). Shape vs. texture biases for stimuli with cue conflict (sorted by human shape bias). Within the responses that corresponded to either the correct texture or correct shape category, the fractions of texture and shape decisions are depicted in the main plot (averages visualised by vertical lines). On the right side, small barplots display the proportion of correct decisions (either texture or shape correctly recognised) as a fraction of all trials. Similar results for ResNet-152, DenseNet-121 and Squeezenet1_1 are reported in the Appendix, Figure 13.



Human Observer는 Shape decision에 bias 되어있고, CNN 모델들은 대부분 texture decision에 bias 되어 있다.

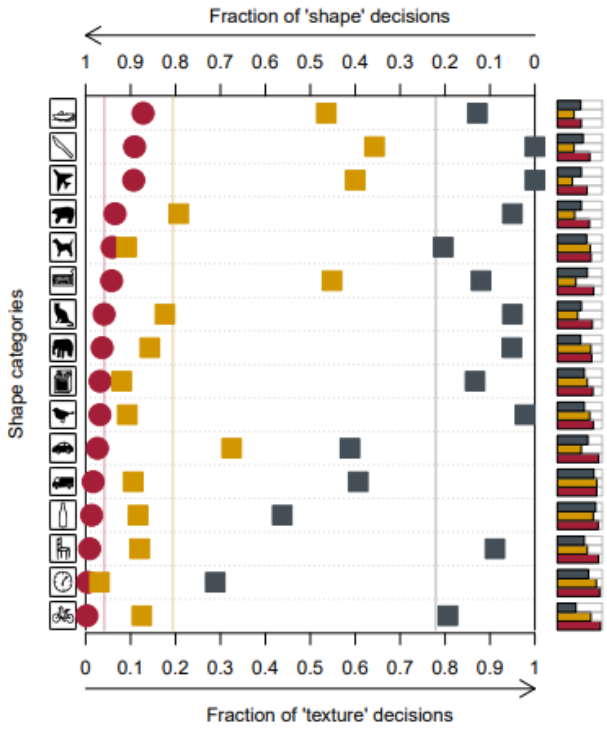
대부분의 경우 CNN과 실험자 모두 주어진 이미지를 분류하는데 성공했다.

하지만 **Silhouette** 이미지, 윤곽선을 검은 색으로 채운 이미지 분류는 사람의 정확도가 훨신 높았다.

→ 사람의 Vision이 texture 정보가 거의 없는 이미지에 더 잘 대처한다.

→ 원본 이미지로 학습된 CNN 모델은 Domain Shifts (원본 → 스케치) 를 통한 학습된 적 없는 이미지에 대해 전혀 대처하지 못한다.

Figure 5: Shape vs. texture biases for stimuli with a texture-shape cue conflict after training ResNet-50 on Stylized-ImageNet (orange squares) and on ImageNet (grey squares). Plotting conventions and human data (red circles) for comparison are identical to Figure 4. Similar results for other networks are reported in the Appendix, Figure 11.



빨간색 원 - 사람 , 노란색 네모 - SIN으로 학습한 ResNet-50, 파란색 네모 IN으로 학습한 ResNet-50

이를 해결하기 위해 SIN (Stylized-ImageNet) 데이터세트를 추가하여 학습하였다.

SIN(Stylized-ImageNet)이란 훈련데이터 Object 의 texture 정보를 무작위로 선택하여 만든 새로운 데이터셋이다.

SIN(Stylized-ImageNet)과 IN(ImageNet)을 같이 학습한 ResNet-50 아키텍처를 Shape-ResNet이라고 부른다.

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**

훈련 데이터에 SIN을 추가하면 Object Detection 성능이 **70.7**→**75.1**로 높이 상승한다.

Object Detection의 경우 Texture bias 보다 Shape bias에 더 집중해야 한다.

SIN을 사용하면 모델 견고성이 향상된다고 볼 수 있다.

Human Vision의 Object Detection을 좀 더 체계적이고 사실적으로 모사하고 싶다면, SIN을 도입하고 Shape bias를 강조하는 방법을 찾은 것 처럼, Human Vision이 Object를 인식하는 방식을 더 심도있게 이해하고 그에 맞는 모델을 만들어야 한다.