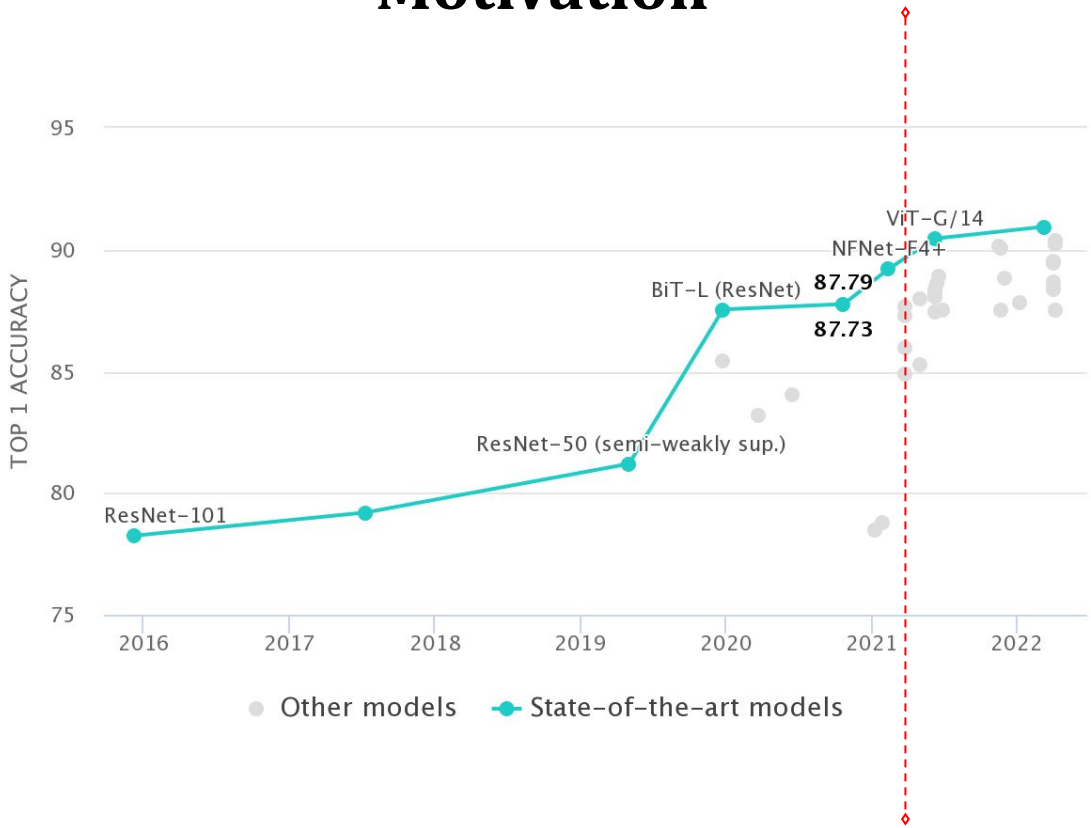
Inductive Biases in Vision Classifiers

Abhay Puri | Axel Bogos | Pulkit Madan | Jizhou Wang

Motivation

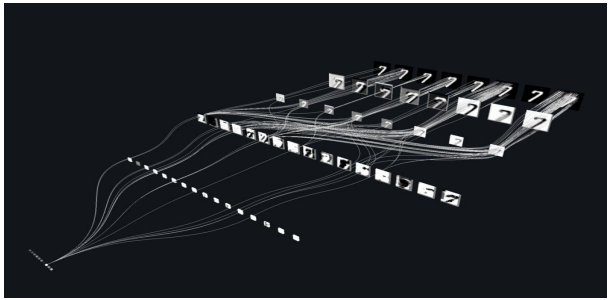

95
90
85
80
75
TOP 1 ACCURACY
2016
2017
2018
2019
2020
2021
2022
ResNet-101
ResNet-50 (semi-weakly sup.)
BiT-L (ResNet)
87.79
87.73
NFNet-F4+
ViT-G/14
Other models
State-of-the-art models

Background

Convolution Neural Networks

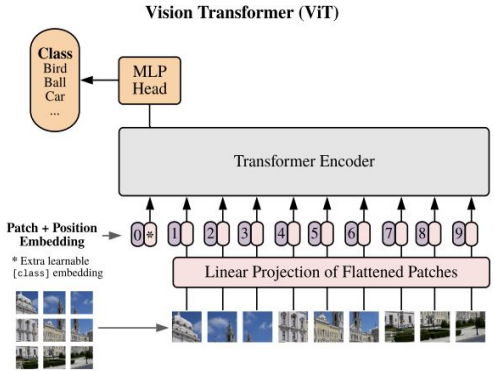
- Learning increasingly complex representations of objects.

- Inductive biases:
  - Local receptive field
  - Translation equivariance
  - Shift invariance



Vision Transformers

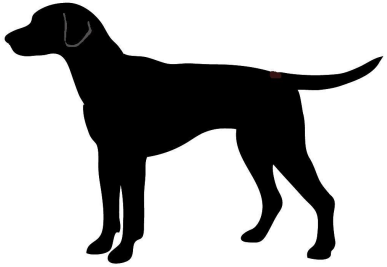
- Patches + Positional Embedding
- No inductive bias towards a local spatial structure, or translation invariance.
- Learn allocation of attention.



Introduction

Comparing *how* these architectures perform classification.

- Analyse how they models the global and local features in the images.
  - Global Representation: Shape
  - Local Representation: Texture



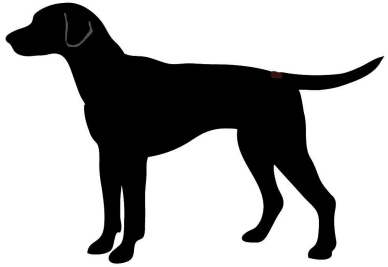
- Two ways to model this behaviour:
  - Error consistency on standard datasets.
  - Testing on specially designed datasets.

Introduction

Comparing *how* these architectures perform classification.

- Analyse how they models the global and local features in the images.
  - Global Representation: Shape
  - Local Representation: Texture

- Two ways to model this behaviour:
  - Error consistency on standard datasets.
  - **Testing on specially designed datasets.**



Introduction

Comparing *how* these architectures perform classification.

(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

Geirhos et al, 2019

Figure - Classification of a standard ResNet-50 of (a) a texture image (elephant skin: only texture cues); (b) a normal image of a cat (with both shape and texture cues), and (c) an image with a texture-shape cue conflict, generated by style transfer between the first two images.

Introduction

Geirhos

- ResNets vs Humans inductive biases
- Dataset: Imagenet
- Training: Transfer Learning

Our work

- CNNs vs ViTs inductive biases
- Custom Dataset
- Training: Fine-tuning
- Application on medical task.

Stylized ImageNet (Geirhos, 2019)

- Built by applying AdaIN (Huang, 2017) style-transfer to ImageNet
- Maps ImageNet classes to 16 overarching classes such as cat, dog, car etc..
- Source of style can be **in-distribution** relative to the content, as in (a) or **out-of-distribution** relative to the content, as in (d).
- Prevents a CNN from “solving” IN solely by texture cues.

(a) Texture image

(b) Content image

(c) Texture-shape cue conflict

(d) Out-of-distribution cue conflict

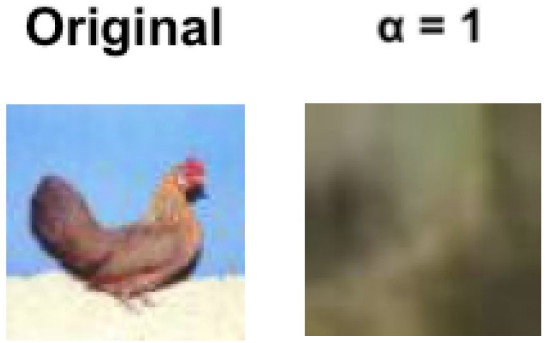
Dataset

- Stylized ImageNet (Geirhos et al., 2019): drop-in replacement for ImageNet, but nearly just as big!
- Tiny ImageNet (Le et al, 2020): much smaller and portable, but not stylized…

Custom Stylized Tiny(ish) ImageNet!

A Few Challenges…

- Style-transfer and low resolution do not mix well
- Tiny ImageNet classes ≠ Stylized ImageNet classes



Dataset

Sample ImageNet Images with OOD Stylization

Truck
n03345487

Elephant
n02504458

Cat
n02124075

Boat
n04612504

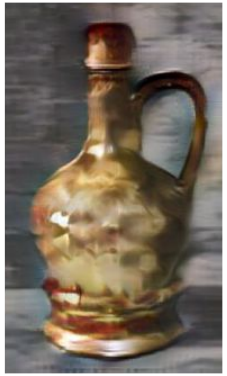
Out-of-distribution stylization
(Kaggle Painter's by Number dataset → IN style transfer)

Sample ImageNet Images with In-Distribution Stylization

Original: Car
Style: Elephant

Original: Keyboard
Style: Bird

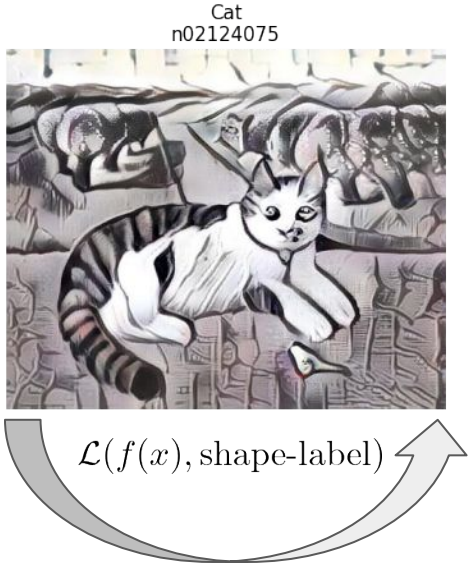
Original: Bottle
Style: Dog



Original: Bottle
Style: Elephant

In-Distribution (IN → IN Style Transfer)

Training (OOD Stylization)



- Learn a global representation with **out-of-distribution** stylization.
- Loss is with respect to the shape label; here “cat”.

Evaluation (In-Distribution Stylization)

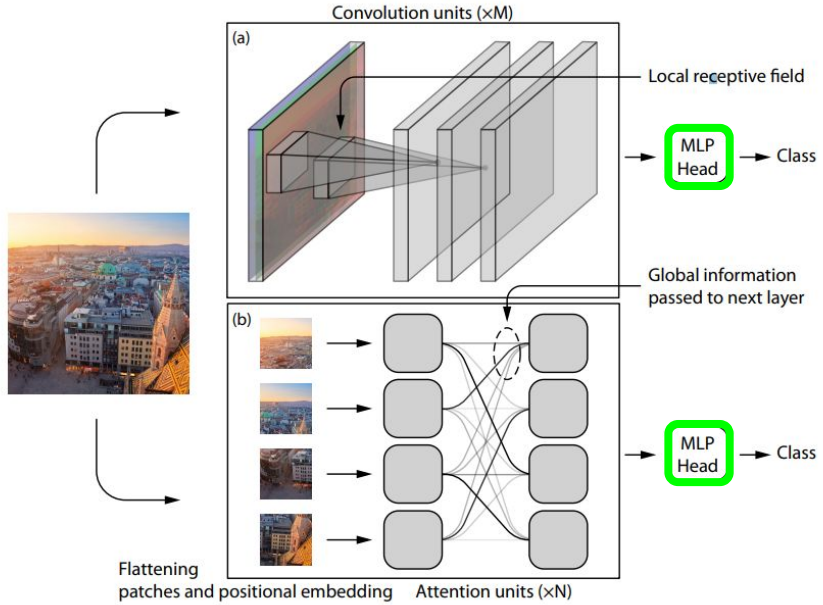
Shape label: **Car**
Texture label: **Elephant**

“Correct” predictions:
texture **OR** shape labels

$$\text{shape-bias} = \frac{\sum \#\text{ of correct shape preds}}{\sum \#\text{ of total correct preds}}$$

Results

- Using models pre-trained on ImageNet
- Fine-tuning *only* the MLP/classification heads.



Results

Pretrained (Top-1 accuracy) on SIN

Model	Pre-trained (%)
ResNet50	0.8
ConvNeXt	0.4
ViT-16	0.5
ViT-32	0.3

Results

Pretrained vs Fine-tuned (Top-1 accuracy)

Model	Pre-trained (%)	Fine-Tuned (%)
ResNet50	0.8	48.1
ConvNeXt	0.4	65.1
ViT-16	0.5	64.6
ViT-32	0.3	52.6

Results

BagNets

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

Top-5 accuracy

Geirhos et al, 2019

% correct shape

Model	Pre-trained (%)	Fine-tuned (%)
ResNet50	35.7	86.0
ConvNeXt	34.7	90.9
ViT-16	36.8	**93.4**
ViT-32	32.9	89.0

ViTs vs ResNets

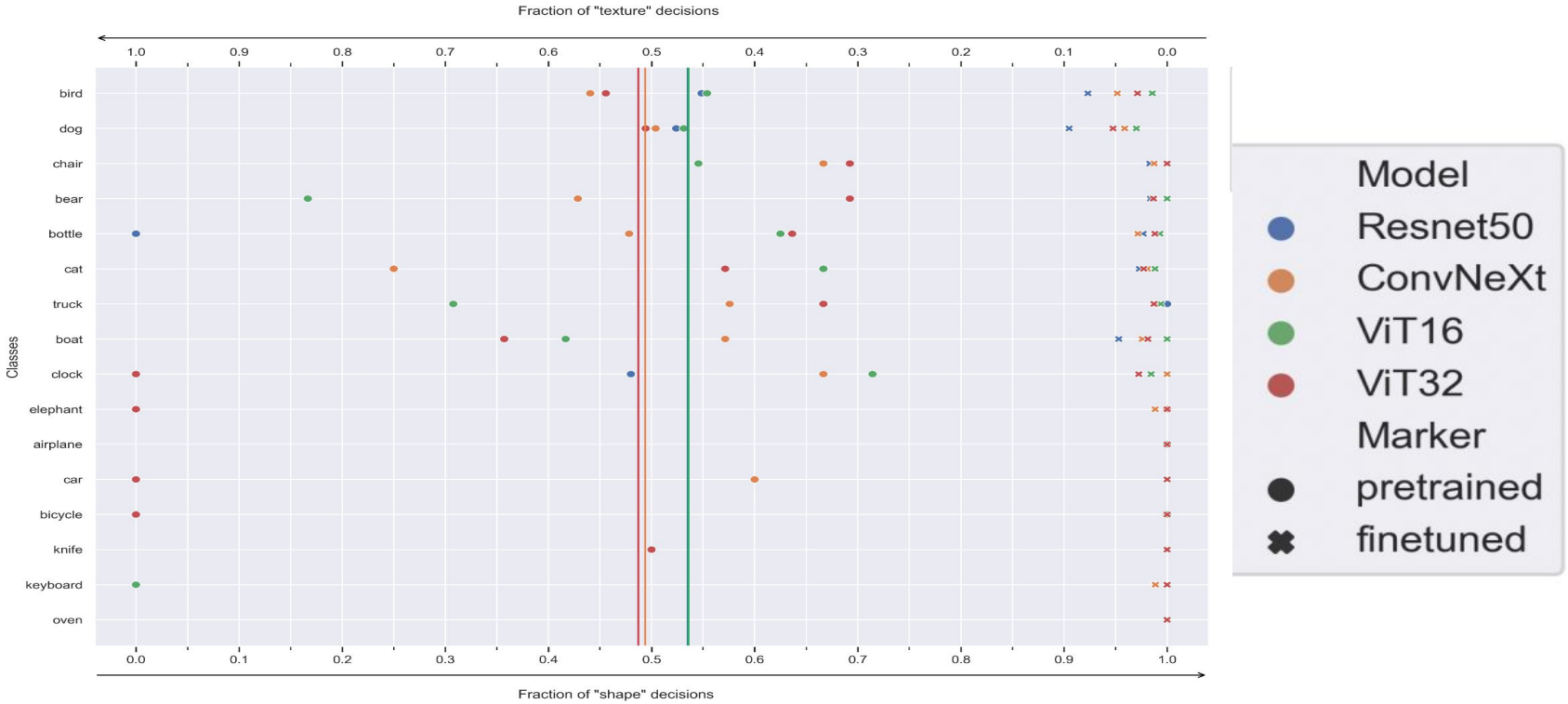
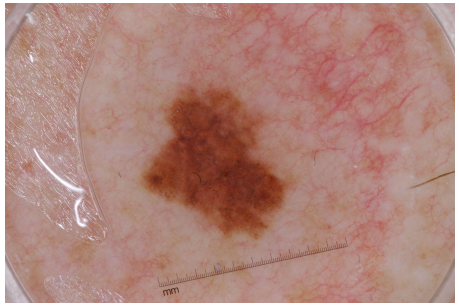


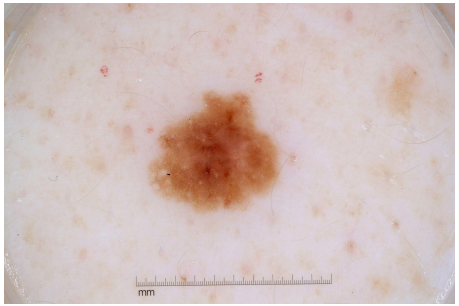
Figure - Overview of the per-class shape-bias of both pre-trained and fine-tuned models.

Melanoma Classification

Malignant



Benign



- 2017 International Skin Imaging Collaboration (ISIC) Challenge Dataset
- Most serious type of skin cancer
- Malignant to Benign Ratio (Train): 1:4.35
- Same training pipeline as SIN

Results After Fine-tuning

Method	Accuracy
ResNet-50 (IN)	80.5
ResNet-50 (IN-SIN)	**82.0**
ViT32 (IN)	78.2
ViT32 (IN-SIN)	**80.5**

Conclusion

- Training on SIN with OOD stylization leads to a more global representation. Models can be pushed towards a “Shape” representation.

- Bias is dependent on target task and is not inherent due to architecture but due the type of data it encounters.

- Classification tasks leads to better performance when the model is pre-trained on stylized representations.

Future Work

- More rigorous experiments with different model architectures (CoAtNet, CLIP), data augmentations (Color distortion, noise, blur)
- Error Consistency and Model Biases
- Experiments on other datasets

Questions?