Convolution vs. Attention for Image Classification

Abhay Puri
abhay.puri@umontreal.ca

Jizhou Wang
jizhou.wang@umontreal.ca

Axel Bogos
axel.bogos@umontreal.ca

Pulkit Madan
pulkit.madan@umontreal.ca

Code Repository

Abstract

Texture and shape are two key components of image recognition, both for humans and within the computer vision field. However, it has been shown that convolutional neural networks are biased towards classifying images through local texture cues rather than a global shape representation (Geirhos et al., 2019). In this work, we aim to expand the evaluation of inductive biases via performance on Stylized-ImageNet (Geirhos et al., 2019) to attention-based models such as Vision Transformers ((Dosovitskiy et al., 2021) and ConvNeXt (Liu et al., 2022). We also introduce a smaller, subsampled version of Stylized-ImageNet in the spirit of Tiny-ImageNet (Le & Yang, 2015) for easier and more iterative experimenting of various models inductive biases. Despite fine-tuning on a much smaller dataset than Geirhos et al., we obtain similar gains in shape-bias representations for attention-based models and better performance than ImageNet pretrained models on an external medical task where both global shape and local texture cues are important to the classification task (Codella et al., 2017).

1 Introduction

Image recognition in computer vision is widely regarded as one of the core tasks of deep learning, and indeed one that has sparked the current wave of interest in deep learning with the introduction of Alexnet (Krizhevsky et al., 2012), one of the first convolutional neural network (CNN) to reach state-of-the-art performance on ImageNet (Deng et al., 2009). Despite the success of CNNs, there has been a recent trend with new architectures inspired from natural language processing which utilizes attention modules described as Vision Transformers (ViT) (Dosovitskiy et al., 2021).

These two architectures have been shown to have different inductive biases when trained to classify conflicting images based on their shape vs. texture (Geirhos et al., 2019). Our experiments have shown that training on a more diverse out of distribution (OOD) stylized dataset allows models regardless of architecture to learn a more global shape representation. This, however, does reduce the accuracy for the task which the architecture was first pretrained on but allows for better global representation and performance when generalizing towards other tasks, such as melanoma skin cancer classification where both local and global representations are important. (Codella et al., 2017)

We will conduct a brief review of the relevant literature, introduce a smaller version of the Stylized-ImageNet dataset brought forth by (Geirhos et al., 2019), briefly expand on the experimental setup before finally discussing the performance and impact on the inductive biases for each model fine-tuned on Stylized-ImageNet and the performance of these models on the melanoma skin cancer task.

2 Literature review

2.1 Background and Previous Work

Convolution Neural Networks (CNNs) have shown remarkable accuracy on benchmark datasets such as ImageNet. Commonly, CNNs are thought to recognize objects by building complex representations in the later layers by utilizing simpler information such as edges, corner, blobs in the initial layers. (Geirhos et al., 2019) go beyond the metric of accuracy and look into *how* the CNNs are doing the task of classification. Through their experiments on Stylized ImageNet (SIN), it has been shown that CNNs cling on to local information of texture and don't pay (as much) attention to the global aspect (shape) of the object in the image. This behaviour is also in contrast to how humans classify the objects as they show that we are biased towards shape. The authors point out that this behaviour of CNNs might not be particularly because of how CNNs function but because the local information (texture) is enough to achieve a strong performance on ImageNet and the network doesn't *need to* perform the harder task of looking at the global context (Occam's razor). Through experiments on the model *BagNet* the authors show that SIN can't be solved using just local features. As a solution i.e., to reduce the dependence on local features (texture) and increase the bias towards global features (shape), authors utilize the technique of data augmentation where they train ResNet-50 (pre-trained ImageNet) on the SIN dataset and then fine-tune it on ImageNet. Experiments show that the resulting network has an increased the bias towards global features (shape), is more robust to various types of noise, improves classification accuracy on ImageNet, and object detection accuracy on the Pascal VOC mAP dataset.

More interestingly shown by (Hermann et al., 2019), texture bias in CNNs is not inherently due to their architecture or training objective but due to the target task and the data it encounters. By applying a more naturalistic data augmentations such as colour distortion, noise and blur, the models are able to learn a more global shape representation. It is comparable to humans, who have encountered more data with different augmentations to arrive at a more 'shape-like' global representation.

2.1.1 Vision Transformers

Dosovitskiy et al. (2021) based their experiments on transformer scaling successes in NLP. They experimented with applying a standard Transformer directly to images, with the fewest possible modifications. This was achieved by splitting an image into patches and providing the sequence of linear embeddings of these patches as an input to a transformer. Image patches are treated the same way as tokens (words) in an NLP application. The model is trained on image classification in supervised fashion. The author highlighted that transformers lack some of the inductive biases inherent to CNNs, such as translation equivariance and locality. This is one of the reasons why they do not generalize well when trained on insufficient amounts of data.

2.1.2 Human Vision Comparisons

Building on (Geirhos et al., 2019),(Tuli et al., 2021) explores the relationship between the shape and texture biases of convolutional nets—in this instance ResNets (He et al., 2015) and ViT (Dosovitskiy et al., 2021) compared to human perception. Notably, they establish that attention-based methods are more biased towards classification *via* shape rather than texture, which in turns also is more similar to human vision biases. Furthermore, the errors made by ViT are more consistent; to measure this, they introduce 2 metrics: Cohen's kappa and the Jenson-Shannon (JS) distance; both of which are founded on a joint probability distribution of misclassifications. These conclusions are obtained under the premise than both models are fine-tuned on an augmented or stylized dataset.

2.2 Concurrent Work and Possible Improvements

In recent years, hybrid architectures (or at least introducing hybrid design paradigms) have been gaining popularity—two such architectures are discussed here. Swin Transformers Liu et al. (2021), based on ResNet-50, the authors iteratively update design choices to nudge the convolutional architecture towards similar architectural features; for example, increasing kernel size to emulate a global receptive field. While some fundamental inductive biases of prototypical ConvNets make their way into the ConvNeXt (Liu et al., 2022), we wish to inquire about the shape and texture biases of

ConvNeXt since, as it has been previously mentioned, attention-based and ConvNets vary greatly in this regard.

Another hybrid architecture that tackles the inductive biases by combining both **Co**nvolutional and Self-**At**tention **Net**works with relative attention called **CoAtNet** (Dai et al., 2021). Convolutions are desirable for their translation equivariance and self-attention for their input-adaptive weighting and global receptive field. Depending on the context of the classification problem, these stacked convolutional self-attention mechanisms are able to focus on the inductive biases relevant for the task at hand, which is especially important for generalization.

Foundation models such as **Florence** (Yuan et al., 2021) and **CLIP** (Radford et al., 2021) are models trained on broad data that can be generalized to a wide range of downstream tasks could provide an insight on inductive biases of shape/texture. Most of these models utilizes a cross-modal shared representation space with visual-language components that adapts to solving vision tasks in Space-Time-Modality space. They have been similarly pre-trained on a contrastive text-image learning task and can be fine-tuned to classify shape/texture as our downstream task.

3 Methods

3.1 ImageNet Stylization & Data set setup

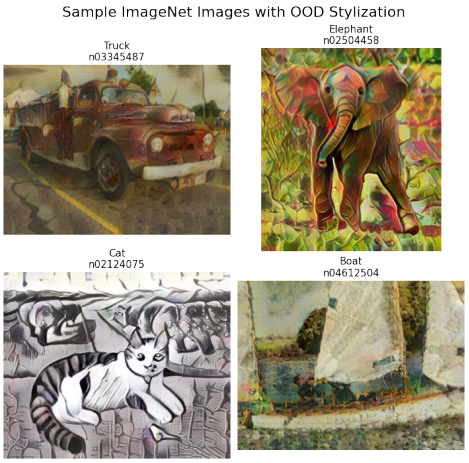
Due to computation and time limitations, we will be comparing our experiments for shape texture bias on a subset of Stylized ImageNet (SIN) that was introduced by Geirhos et al.. Stylized Image-Net consists of the original Image Net (Deng et al., 2009) upon which a style transfer is applied; this style transfer is an implementation of AdaIN style transfer approach introduced by Huang & Belongie (2017).

However, for our experiment two different AdaIN style transfers were applied to the ImageNet subsample. We will refer to an *out-of-distribution style-transfer* as a style-transfer from a dataset that is not derived from ImageNet to ImageNet, and an *in-distribution style-transfer* as a style-transfer from a dataset that is derived from ImageNet to a dataset that is also derived from ImageNet. As recommended by Geirhos et al., applying an out-of-distribution style-transfer to the training data, while maintaining in-distribution style-transfer allows for the proper evaluation of the inductive bias of models (instead of a proxy measurement of model capacity for learning a noisy image representation). Hence, we have applied a style-transfer from the painter by numbers dataset (which is a subset of WikiArt (Saleh & Elgammal, 2015), available here) and a style-transfer from ImageNet to the validation and test data. Samples of out-of-distribution style-transfer and in-distribution style-transfer can be found in figure 1. Henceforth, we will refer to performance metrics on the in-distribution style-transfer test data as "on SIN" and will differentiate between "pre-trained models" as models pre-trained on standard ImageNet and "fine-tuned models" as models pre-trained on ImageNet and fine-tuned on out-of-distribution stylized train data.

Our final dataset consists of 207 stylized ImageNet classes, each containing 500 training images, 50 validation images and 50 test images for a total of 124,200 images. Each of these files are named according to two classes: the original class of the image, and the class after which it was stylized; henceforth, we will refer to these labels as the *shape label* and the *texture label* respectively. Each of these 207 classes is mapped to one of 16 *Human-Level* classes as introduced by Geirhos et al.. It is on these 16 classes that will ultimately be measured the shape and texture biases of each model. Henceforth, we will refer to this dataset as *Stylized ImageNet Subset*.

3.2 Experimental Setup

Experiments will largely follow two axes; firstly, an **evaluation** of the shape bias of each model will be conducted; secondly, an inquiry into possible training methods or architectural changes to **reduce** the bias of various models will be conducted. Finally, a comparative evaluation on the melanoma dataset (Codella et al., 2017) where shape and texture are both inherently important to the classification task.



(a) Style Transfer from painter by numbers to IN samples

(b) Style Transfer from IN to IN samples

Figure 1: Out-of-Distribution and In-Distribution Style-Transfer Examples

3.2.1 Bias Evaluation

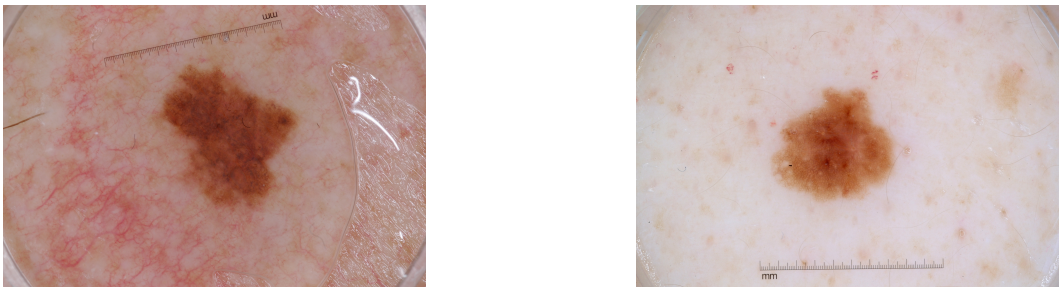
To first evaluate the underlying shape bias of each model of interest (e.g., ResNet-50, Vit-B-32, ConvNext and CoAtNet), the pre-trained on ImageNet version of each model will be trained on *stylized tiny ImageNet* with the loss computed with respect to the shape label. The accuracy of each model on ImageNet and stylized ImageNet (SIN) will be recorded. Then, the bias of each model will be compared through the shape decision ratio (SDR). The SDR is simply defined as the ratio of shape label predictions conditioned on the prediction of the model being either the shape or texture label without cue conflict.

3.2.2 Bias Reduction

Once a baseline has been established for the shape and texture bias of each model, methods to reduce the bias of each model may be considered. These experiments belong either to training paradigm changes or architectural changes. One such training experiment will be to reframe the training on stylized tiny ImageNet as a multiclass objective. To do so, target labels will be one-hot encoded and the loss function will be modified to reflect that predicting whether the shape or texture labels is a correct prediction. Once again, the change in bias of each model will be represented as the SDR.

3.2.3 Frozen vs. Unfrozen Finetuning

2 methods of fine-tuning the models pretrained on ImageNet have been considered. First, we considered freezing all layers except for the final fully connected classification layer of each model, leaving the core of the pre-trained weights intact in our fine-tuning phase. The alternative would be to allow our fine-tuning training phase to backpropagate through the full network and hence modify the core pretrained weights. For the majority of the discussion here, our results refer to the first fine-tuning method, consisting of freezing all but the final layer weights. While unfrozen experiments have been conducted and will be expanded upon in future works, we limit the core of our discussion to frozen pretraining for the following reasons. First, our number of fine-tuning samples is several orders of magnitudes lower than the number of ImageNet examples used for the pretraining, yielding a very low training examples to model parameters ratio for our fine-tuning. Secondly, we wish to investigate the preliminary impression that shape-biased representations are indeed learned by pre-trained models on ImageNet, but it is the final classification layer that favours local-texture features rather than global shape ones. By improving the shape-bias of each model solely by fine-tuning the classification layer, our results seem to agree with the latter hypothesis.



(a) Malignant (b) Benign

Figure 2: Melanoma Dataset Samples.

3.2.4 Melanoma Dataset & Setup

A study by (Kalouche, 2016), which tried to classify skin cancer using different vision classification models has shown that CNNs were able to accurately predict melanoma compared to non-CNN methods. In our case, we will try to show that the dataset in which the deep learning models are pretrained on can also have an impact when fine-tuning to an OOD task such as melanoma classification.

The dataset we will be using is from the 2017 Internation Skin Imaging Collaboration (ISIC) Challenge (Codella et al., 2017). It contains 2000 train, 150 validation and 600 test images. A sample of benign and malignant tumour is shown in figure 2. The ratio of malignant to benign tumours is roughly 1:4.35, therefore weights are applied during training to rebalance the losses. The model train and validation pipeline are the same as the SIN dataset.

4 Results and Analysis

4.1 Experiment Tracking

The evaluation of the representation learning capabilities of ResNet, Vision Transformer (ViT), and ConvNeXt was tracked using the Weights&Biases dashboard. Weights&Biases is a central cloud dashboard that keeps track of hyperparameters, system metrics, and predictions of deep learning models. Everything about our models such as performances, architecture and parameters used, system information (e.g., number of CPUs/GPUs used), running time are tracked. Some of our experimental results are shown here.

4.2 Overview of Results on SIN

Overall, observing table 1, we see that fine-tuning on SIN improves significantly the combined shape and texture accuracy of each model (with the biggest jump in performance being for ViT16 with a 0.404 percentage point gain in %-correct-both on SIN compared to the non-finetuned model). Comparing the distinct performance on shape and texture, we see that the significant gain in %-correct-shape by fine-tuning is not accompanied by a comparable drop in performance in %-correct-tex, indicating that we are indeed learning a richer representation instead of unilaterally changing the focus of the classification layer. Lastly, we also see a significant increase in the shape bias of all models.

Observing figure 3, we see that every after fine-tuning on SIN every model has significantly increased its shape bias, thus pushing each model towards learning a more global representation. Increasing the shape-bias has been shown by (Geirhos et al., 2019) to increase the robustness and performance of trained models on the original task compared to the equivalent model with its original inductive biases. This is also comparable to the human observer experiment from (Geirhos et al., 2019) where a high bias towards shape is also represented. The results are also presented in an interactive plot here.

Models	%-correct-both	%-correct-shape	%-correct-tex	Top-1-acc-shape	Top-5-acc-shape	Top-1-acc-tex	Top-5-acc-tex	Shape-Bias	No. of Params
ResNet-50	0.528	0.357	0.332	0.008	0.030	0.005	0.022	0.534	424143
ResNet-50 (finetuned)	0.899	0.860	**0.358**	0.481	0.736	**0.008**	**0.035**	0.932	424143
ConvNeXt	0.531	0.347	0.351	0.004	0.021	0.004	0.022	0.493	159183
ConvNeXt (finetuned)	0.930	0.909	0.346	**0.651**	0.849	0.006	0.030	0.965	159183
ViT-16	0.540	0.368	0.341	0.005	0.023	0.004	0.023	0.535	56828367
ViT-16 (finetuned)	**0.944**	**0.934**	0.337	0.646	**0.869**	0.005	0.025	**0.982**	56828367
ViT-32	0.513	0.329	0.338	0.003	0.018	0.005	0.024	0.487	56828367
ViT-32 (finetuned)	0.910	0.890	0.336	0.526	0.778	0.005	0.025	0.966	56828367

Table 1: Comparison of models across shape, texture and combined metrics. The best performance for each column (metric) is bolded.

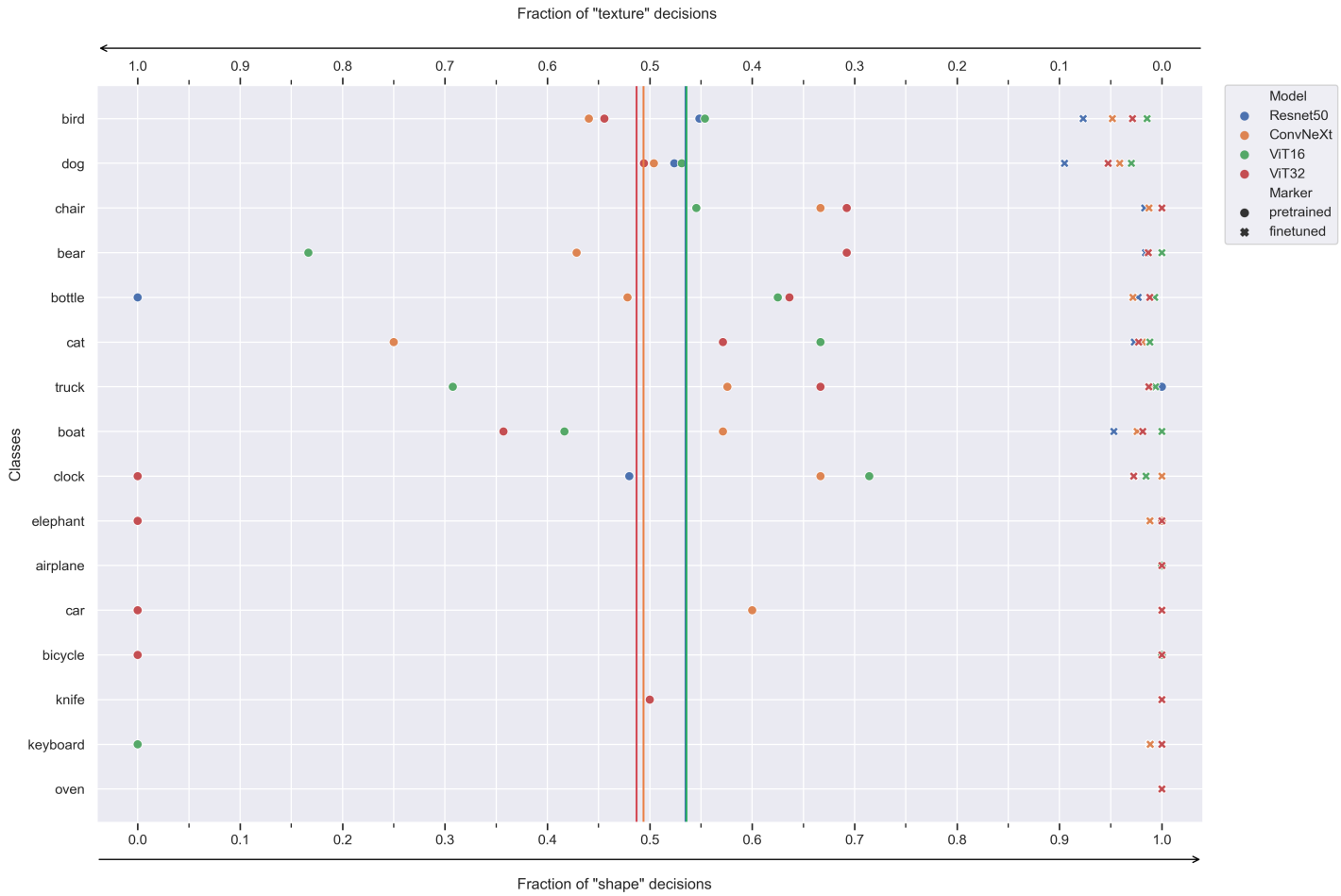


Figure 3: Overview of the per-class shape-bias of both pretrained and fine-tuned models. Vertical lines are the average shape-bias of non-fine-tuned models.

4.3 ResNet

ResNet50 trained on ImageNet (IN) achieves a top-1 accuracy of 76.13% when tested on IN. However, when tested on Stylized-ImageNet (SIN) obtained after applying in-distribution style-transfer, it achieves a suboptimal top-1 accuracy of only 0.8%. From 1a we can observe that after applying for out-of-distribution style transfer, we retain the global shape but the local texture cues are no longer highly predictive of the target class and thus, we can hypothesize that convolutional neural networks are looking at local features of texture to *solve* IN. Now, to evaluate if the SIN can be solved using local features, Geirhos et al. conducted experiments using *BagNets*(Brendel & Bethge, 2019) which utilize the ResNet-50 architecture but with smaller receptive fields of a size such as (9 x 9) pixels and reported that although *BagNets* can achieve an accuracy of 70% on IN, it only achieves an accuracy

of 1.4% on SIN. Following the results of experiments conducted by us and Geirhos et al. we can conclude that:

- ResNets are using local features to solve IN and thus are biased towards texture.
- SIN can't be solved using local features when there are local conflicts.

To test if we can reduce the texture bias in CNNs and see if they can learn a more global representation of shape to solve SIN, we fine-tune ResNet50 pretrained on IN on SIN. The fine-tuning is performed by training only the classification layers and keeping the rest of the layers of the network *frozen*. We experiment with multiple training regimes (different learning rates, optimizers) and achieve the best top-1 accuracy of 48.1%.

With increase in top-1 accuracy of 1.4% to 48.1%, we can say that the texture bias in CNNs is not inherently due to their architecture but due to the target task. As IN is solvable using local features, CNNs performed the easier task of looking at local cues of texture *(Occam's Razor)*. For SIN, CNNs are able to learn *some* global representation as we see a big increase in the top-1 accuracy the metric *% correct-shape* but with the best-case accuracy still below 50%, CNNs leave a lot to be desired when it comes to learning global cues such as shape. By looking at the figure 4 and examining the misclassified sample on the right, we can see the CNNs are still clinging to the local cues as a dog with a bird's texture, is a bird to the CNN.

Figure 4: Prediction of ResNet50 Under Texture and Shape Conflict

4.4 ViT

As shown in table 1, pre-trained ViT16 and ViT32 on ImageNet respectively achieve 0.5% and 0.3% top-1 accuracy rate on SIN, hence showing that despite favouring a more global learned representation through the attention mechanism, ViT models are not able to directly solve SIN. However; after fine-tuning the out-of-distribution stylized dataset, ViT16 and ViT32 achieve top-1 accuracy rates on SIN of 64.6% and 52.6%. While it is initially surprising that ViT32 did not prove better at capturing long-range shape representations than ViT16 given its larger kernel size, we hypothesize the following possible explanation to be further explored time-permitting in the final deliverable: the smaller patch size of ViT 16 may allow it to better capture and single-out key shape features, such as wheels in trucks and cars, where those shapes occupy relatively small portions of images without introducing too much supplementary stylization noise.
We also note that despite a very high number of parameters, fine-tuning on roughly 100k stylized images proved significant in increasing the shape-bias and by extension the bias towards learning a global representation.

4.5 ConvNeXt

As shown in table 1, pre-trained ConvNeXt achieves a top-1 accuracy of 0.4% on SIN, hence also showing that it is not able to solve directly solve SIN. Once fine-tuned on an out-of-distribution stylization, ConvNeXt achieves 65.1% accuracy on SIN. Along with ViT16, ConvNeXt achieves

Table 2: Summary of experiments evaluated on the melanoma dataset.

Model	Pretrained	Finetuned	Accuracy	F1-Score
ResNet-50	IN	No	0.540	0.59
ResNet-50	IN	Yes	0.805	0.72
ResNet-50	IN→SIN	Yes	**0.820**	**0.78**
ViT32	IN	No	0.680	0.68
ViT32	IN	Yes	0.782	0.71
ViT32	IN →SIN	Yes	**0.805**	**0.72**

the highest top-1 accuracy and shape-bias. Since ConvNeXt outperforms CNNs by a big margin, we can hypothesize that the addition of the attention module to the CNN architecture has allowed the ConvNeXt to look at a more global representation and thus achieve a higher score for accuracy and shape bias. Interestingly, ConvNeXt achieves higher accuracy than ViT models while having a higher texture bias. This means that the ConvNeXt can strike a balance between adhering to local and global representation which makes it particularly attractive for tasks where both the local and global representations are important, especially in a limited computing environment as ConvNeXt uses a magnitude less number of parameters are compared to ViT.

4.6 Melanoma Dataset Results

From our experiments on SIN, we are able to conclude that we are able to tune the texture-shape bias in different model architectures. Thus, we utilize these findings on the task of melonoma classification, where features of both shape and texture are important for accurately identifying the type of cancer (malignant or benign). After evaluating on ResNet-50 and ViT32, the results shown in table 2 concludes that pretraining the models (pretrained on IN) on SIN improves the classification accuracy of each respective models on the melanoma dataset by a few percentage points. This shows that the model has better generalization when pretrained a more diverse dataset before fine-tuning.

4.7 Problems Encountered & Challenges

For data exploration, the dataset balance for the 16 Human-Level classes with their shape and texture labels shown in figure 5 displays an imbalance which can introduce a bias during training without weight balancing. Although this may not be a big problem since we're training on the shape decision of 207 stylized classes which are balanced and this wouldn't affect the average result (over all 16 classes) for the shape-texture bias.

Certain models have different output heads for the final layer, we have been able to adapt them during training for the relevant datasets.

5 Conclusion

From our study, we conclude that the biases of shape and texture are not only influenced by the type of architecture that we are using but also by the target text at hand. ImageNet (IN) is solvable using just the local cues of texture and thus the models didn't have to perform the harder task of looking at the global representation of shape. However, SIN required the models to learn such representation and we notice the models were indeed able to pick up more global cues related to shape, and the extent of the learned shape representation (shape bias) is influenced by the choice of architecture. By training on SIN, we are able to *tune* the biases present in the architectures and we utilize the findings to increase the accuracy of classification of melanoma by simply finetunining the representations pretrained on ImageNet, before training it on the melonoma dataset.

5.1 Future Work & Further Discussions

In this work, we are performing fine-tuning of just the classification layer(s) of the network(s) and thus are essentially not learning new features but, aggregating the already learned features in a way that is able to better capture the global representation of shape, and have a better generalizing performance as we notice in the case of melanoma classification. In our future work we would like to

conduct an exploratory study into *how* information gathering evolves with training on SIN, which leads to learning of better global features like shape. It would also be interesting to compare how the final representation and information gathering of the networks fine-tuned on SIN compare with the models that utilize transfer learning (training the whole network instead of just classification layer).

Also, we fine-tune the networks (pretrained on IN) on SIN, by penalizing the missclassifcations of shape. Thus, during this process we suspect that the network might be losing some essential local representations of texture that it learned while training on IN, to minimize the loss on shape labels. Therefore, we would like to explore if fine-tuning with a joint dataset of IN + SIN we would lead to learning a representation that preserves the key local cues of texture while learning the global cues of shape.

Furthermore, we utilize a subset of IN to create our custom dataset: SIN. Granted enough compute, we would like to repeat our experiments on SIN formed using the complete ImageNet dataset.

In our choice of networks, we experiment with both traditional as well latest convolutional and attention based architectures. In our future work, we would also like to experiment with "fusion" models like CoAtNet (Dai et al., 2021) which has dedicated modules for capturing both local and global representations.

Another inquiry in regards to the training and data setup is to critique the human's biases towards shape. The dataset which humans were trained on generally have a more "shape" based label. This may be an indication that the experiment setup by (Geirhos et al., 2019) for the human evaluation of shape-bias is inherently biased towards shapes compared to textures. It would be interesting to further test the model by conducting training experiments based on textures to see if the shape-decision ratio (SDR) still tends towards shape. Otherwise conduct human experiments where they classify unseen shape classes in terms of evaluating SDR.

For the applications of this work, we would like to perform more robust experiments on the melonoma dataset as well as tasks in other domains to test for statistical significance of model representation generalization after training on stylized versions of the datasets.

6 Contributions of each team member

Jizhou Wang has given input to the project report and has worked on coding the model train/test pipeline, visualization of shape-bias and other evaluation metrics.
Pulkit Madan has given input to the project report, and contributed towards literature review, model training, and evaluation of models.
Abhay Puri has given input to the project report, contributed towards the literature review, set up the experiment tracking mechanisms and model evaluation.
Axel Bogos has given input to the project report, contributed towards the literature review, conducted the data engineering and data preparation, and model training.

References

Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet, 2019. URL `https://arxiv.org/abs/1904.00760`.

Noel C. F. Codella, David A. Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin K. Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (ISIC). *CoRR*, abs/1710.05006, 2017. URL `http://arxiv.org/abs/1710.05006`.

Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit,

and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Katherine L. Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks, 2019. URL https://arxiv.org/abs/1911.09071.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510, 2017.

Simon Kalouche. Vision-based classification of skin cancer using deep learning. 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pp. 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge. 2015.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature, 2015. URL https://arxiv.org/abs/1505.00855.

Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L. Griffiths. Are convolutional neural networks or transformers more like human vision?, 2021.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision, 2021.

A Appendix

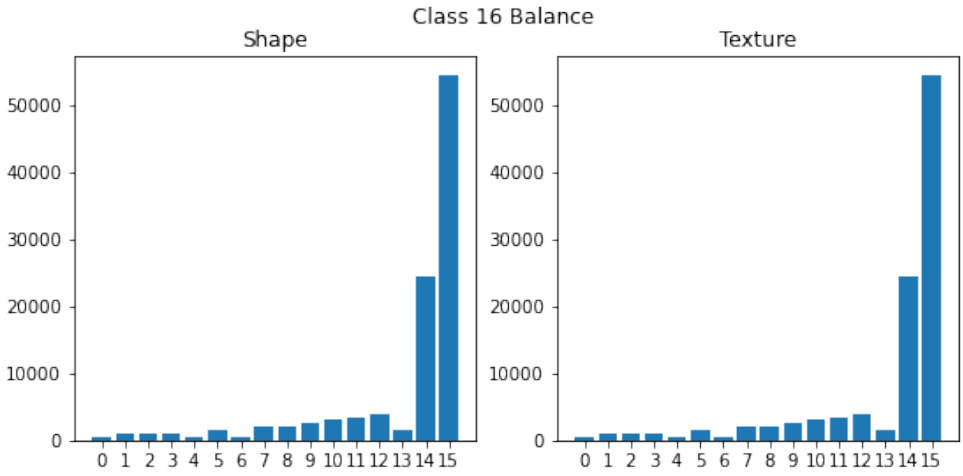


Figure 5: Dataset Balance of the 16 Human-Level Classes