Pre-Training on High-Quality Natural Image Data Reduces DCNN Texture Bias

Niklas Müller (n.muller@uva.nl)
Psychology Research Institute, University of Amsterdam, The Netherlands

Iris I. A. Groen[*] (i.i.a.groen@uva.nl)
Informatics Institute, University of Amsterdam, The Netherlands
Psychology Research Institute, University of Amsterdam, The Netherlands

H. Steven Scholte[*] (h.s.scholte@uva.nl)
Psychology Research Institute, University of Amsterdam, The Netherlands

[*] Shared senior author

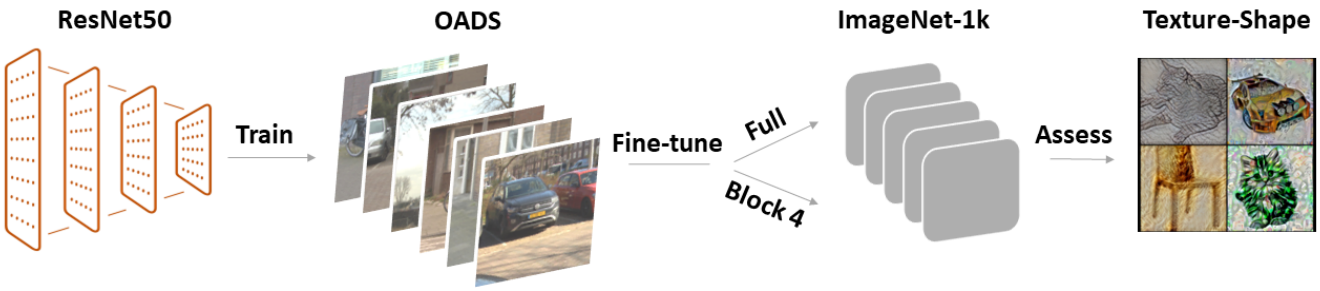

Figure 1: Experimental paradigm of pre-training (OADS), fine-tuning (ImageNet) and behavioural assessment (texture-shape).

Abstract

Deep Convolutional Neural Networks (DCNNs) perform increasingly well on visual tasks like object recognition while also closely aligning with human brain activity. However, model behaviour also differs from human behaviour in important ways. One prominent example of this difference is that DCNNs trained on ImageNet exhibit a texture bias, while humans are consistently biased towards object shape. Previous work suggests DCNN shape bias can be increased by training on purposely designed stimuli (e.g. stylized images). Here, we present an alternative method that reduces texture bias: pre-training on high-resolution natural images that more closely approximate human visual experience. Our training pipeline needs no data augmentation but solely relies on visual features that occur in everyday scenes. Our method and dataset provide an opportunity to build DCNNs that operate on high-resolution images and may aid in closing the gap between human visual processing and DCNNs.

Keywords: Natural human vision; Deep convolutional neural networks; Open Amsterdam Data Set; Texture bias

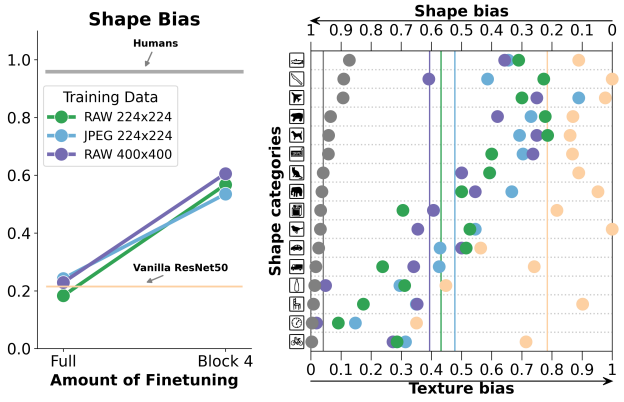

Figure 2: Shape bias increases when pre-training on OADS. Left: average shape bias of models finetuned across all layers, or only on block 4. Colours indicate data quality of images used during pre-training. Gray horizontal bar denotes human shape bias. Orange bar denotes shape bias of a standard ImageNet trained ResNet50. Right: shape bias separated by image class for models finetuned on Block 4. Vertical lines show average across all classes.

Introduction

Previous studies show that DCNNs rely strongly on texture-like information to classify objects, while humans rely more on object shape **(Geirhos et al., 2018)**. Data augmentation and training on stylized datasets that exhibit more shape-like structures increases DCNN shape bias compared to DCNNs trained only on ImageNet **(Geirhos et al., 2021)**. While this partially closes the behavioural gap between DCNNs and humans, the engineering nature of these solutions weakens the claim that DCNNs are accurate models of human vision.

Human visual processing starts at the retina, which is capable of processing visual inputs at several times higher resolution than ImageNet images used for standard DCNN training. High-resolution retinal and precortical processing give rise to neural tuning to contrast edges **(Schiller & Tehovnik, 2015)**, likely facilitating cortical coding of object shape. Consistent with this idea, we here show that pre-training DCNNs on a high resolution natural image dataset that more closely matches the information density that the human eye provides to the visual cortex increases DCNN shape bias, thereby making network behaviour match human behaviour more closely.

Our contributions are twofold: first, we introduce the Open Amsterdam Data Set (OADS) consisting of high-resolution, uncompressed images of natural outdoor scenes captured in the streets of Amsterdam. Each image contains several object annotations, allowing it to be used for object classification and localization tasks. Due to the high-resolution of the scenes, objects have rich information of local features and closely resemble human visual experience. Second, we show that pre-training on OADS reduces texture bias in DCNNs. Our pre-training pipeline does not require engineered data augmentation and therefore enables a direct comparison of both behaviour and activity patterns between humans and DCNNs.

Methods

The OADS dataset contains 5330 images with a resolution of 5496x3672 pixels. Each image contains multiple annotations consisting of an object bounding box and a class label, yielding a total of 64152 annotations across 16 object classes, including 'Van', 'Bench', 'Lamppost', 'Bike' and others. Images were collected at outdoor locations in Amsterdam and were all

taken from a human observer perspective.

Here, we use pre-training on the OADS as an alternative approach to data augmentation. We compared publicly available ImageNet trained ResNet50 models **(He, Zhang, Ren, & Sun, 2016)** to equivalent networks pretrained on the OADS and finetuned on ImageNet. We systematically varied model training conditions in three aspects. First, to investigate if image quality affects DCNNs texture bias, we manipulated the OADS image data used for pre-training to either match ImageNet resolution, or exhibit JPEG compression, or combinations of the two. Further, we varied the ratio of data quality that is kept during compression. Third, we varied the number of parameters that are allowed to be trained during fine-tuning.

Uncompressed OADS images were cropped using the object annotation boxes. Crops were resized to 400x400 pixels (high resolution) and directly used as input to the networks or first JPEG compressed (compression quality=40/60/90%) and then resized to 224x224 pixels (low resolution) to approximate ImageNet image quality. To disentangle the effect of image resolution versus JPEG compression, we included two control conditions, one where crops were JPEG compressed but maintained higher resolution and one where crops were downsampled to the lower resolution but not compressed.

Models were trained for 30 epochs on the 16 OADS object class categories and then finetuned on ImageNet-1k object classification for 15 epochs using PyTorch **(Paszke et al., 2019)**. We varied the amount of influence of the pre-training during evaluation by fine-tuning two versions of each model. The first version only fine-tunes the last residual block (block 4 out of 4) while the second allows all network weights to adapt.

Object classification test accuracy was used as a measure of transferability as a function of the number of fine-tuned parameters (*Full* vs *Block 4*). Behavioural similarity between humans and all (pre-trained) models was assessed using the *model-vs-human* benchmark and more specifically, images from the cue-conflict dataset **(Geirhos et al., 2021)**. These images exhibit the texture of an image from one class while the shape of an image from a distinct class is superimposed onto the texture (**Fig. 1**). All cue-conflict images are part of ImageNet. Texture bias was assessed as the number of texture decisions over the sum of texture and shape decisions.

For comparison, we include a model that was pretrained on ImageNet-1k and finetuned on a stylized version of ImageNet (here referred to as "SIN ResNet50") **(Geirhos et al., 2018)**.

Results

During pre-training, all models achieved a training accuracy of $> 99\%$ and a test accuracy of $> 95\%$. During fine-tuning on ImageNet-1k all models achieved a training accuracy of $> 98\%$. Test (generalization) accuracy is shown in **Table 1**.

All OADS pre-trained models show a clear increase in shape bias compared to the Vanilla ResNet50 when only Block 4 of the model was finetuned (**Fig. 2, left**). Importantly, this gain in shape bias disappears when fine-tuning the entire network. This suggests that features learned in lower-level layers by training on OADS contribute to shape bias.

Table 1: OADS-RAW refers to models pretrained on uncompressed images, OADS-JPEG to models pretrained on JPEG compressed images. Percentages refer to the ratio of data quality preserved during compression. Numbers refer to the image resolution used during training on OADS. Longer fine-tuning (15-30 epochs) did not further increase test accuracy.

		Accuracy		**Shape Bias**	
Model		Full	Block 4	Full	Block 4
Vanilla ResNet50		0.76		0.21	
SIN ResNet50		0.60		0.81	
OADS	RAW 224x224	0.59	0.37	0.18	0.57
	RAW 400x400	0.57	0.30	0.23	0.60
	JPEG 40% 224x224	0.61	0.33	0.24	0.55
	JPEG 60% 224x224	0.61	0.34	0.27	0.53
	JPEG 90% 224x224	0.58	0.33	0.21	0.52

Human shape bias is more prominent for certain classes than for others; similarly, OADS pre-trained networks show a stronger increase in shape bias to certain classes than others, generally following the same trend as humans (**Fig. 2, right**).

We find no systematic effect of JPEG compression on the shape bias. Similar to SIN ResNet50, classification performance of OADS-trained networks is lower than the Vanilla ResNet50. This suggests the presence of an inherent trade-off between ImageNet accuracy and shape bias.

Discussion

Pre-training on high-resolution, uncompressed naturalistic images increases DCNN shape bias. Systematic comparison of pre-trained models shows that the more a model is exposed to ImageNet (fine-tuning of more parameters) or its features (compression or lower resolution), the more texture bias increases. This suggests that texture bias need not be inherent to DCNNs but might rather depend on the type and quality, and the extent of control thereof, of the training data.

Our approach relies on the level of local detail in the training images and does not require data augmentation or training on a purposely designed dataset. We posit that shape bias increases because OADS more closely mimicks the information density of inputs to the human eye. An important next step is a separate texture vs. shape bias benchmark that does not require fine-tuning on ImageNet. Moreover, the uncompressed format of OADS offers the possibility to develop custom pre-processing pipelines (e.g., biologically inspired compression or filters) to more closely match primate visual processing.

Conclusion

We contribute a new high-resolution natural image dataset for DCNN training and provide evidence that pre-training on these images - thereby more closely mimicking human visual experience - offers an alternative to engineering pipelines or training on artificial stimuli to increase DCNN similarity with humans.

Acknowledgments

This work is supported by the Interdisciplinary PhD Programme of University of Amsterdam Data Science Center.

References

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2021). Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, *34*, 23885–23899.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc.

Schiller, P. H., & Tehovnik, E. J. (2015). *Vision and the visual system*. Oxford University Press, USA.