Shape Matters: Investigating Shape Bias in MLPs

Tristan Gollmart Daniele Roncaglioni Lisa Walz Matthias Wüest

Abstract

In this work we explore whether MLPs are naturally biased towards using shape or texture as the main feature for image classification decisions. CNNs are prone to being texture biased, especially when trained using strong augmentations, while Vision Transformers are more shape biased. Our findings reveal that MLPs heavily favor extracting shape cues rather than texture. The predominance of the shape bias is almost as high as that of humans, and substantially higher than that of vision transformers and of course CNNs. This indicates that MLPs process information in a fundamentally different way, with high variance in the importance of input pixels, and a propensity for identifying cohesive shape areas as a whole rather than patterns of edges or texture.

1. Introduction

A strong feature of the human visual system is its ability to robustly generalize to previously unseen distributions. Needless to say, such generalization abilities are desirable not only for humans, but also for real-world machine vision systems. (Geirhos et al., 2018b) compared the robustness of human vision with convolutional neural networks (CNNs) on object recognition, revealing CNNs' superiority on trained distortions but poor generalization to new ones. In a similar setup, (Geirhos et al., 2018a) explored texture and shape biases in humans and CNNs, revealing: (1) while humans rely on shape, ImageNet-trained CNNs exhibit a strong texture bias, (2) the texture bias can be reduced by training on specific datasets, and (3) these new networks demonstrate exceptional robustness to various image distortions. This suggests that introducing a shape bias into deep neural networks (DNNs) may enhance overall distortion robustness. (Hermann & Kornblith, 2019) suggested that CNN's texture bias results from training data augmentations, proposing alternatives that improve shape bias. (Geirhos et al., 2021) evaluated more recent machine learning (ML) developments and found that top models are now outperforming humans on out-of-distribution datasets.

In a different line of research, (Bachmann et al., 2023) explored the multilayer perceptron (MLP) as a viable architecture for vision tasks, motivated by their mathematical simplicity, inference efficiency, and low inductive bias. The results suggest that the lack of inductive bias can be compensated by scale, with large-scale pre-trained MLPs achieving competitive accuracies comparable to CNNs trained from scratch on different datasets.

The ability of such MLPs to generalize is largely unexplored. In this work, we make the following contributions:

- We experimentally confirm the shape bias in MLPs.
- We observe the robustness of the shape bias to shape destroying augmentations.
- We characterize the shape bias: MLPs construe shape more as solid, colored silhouettes rather than seeing it as a collection of semantically meaningfully positioned edges.
- We compare the magnitude of the shape bias to that of humans and other architectures.
- We visualize, explore and discuss the weight dynamics of MLPs.

2. Models and Methods

2.1. Experiments on Imagenette

We have investigated two Inverted Bottleneck MLPs (B-MLP) from (Bachmann et al., 2023): *B-6/Wi-512* and *B-12/Wi-1024*.
As a dataset we used Imagenette (IN10 or IN when clear from context) (Howard), which is a 10-class subset of Imagenet1k images in native resolution. We have created a stylized version of this dataset (SIN10 or SIN, see Fig 1) as explained in (Geirhos et al., 2018a).

We trained our models from scratch on IN and SIN (64x64) avoiding data augmentations that a priori seem to have a high chance of disrupting shape or texture information such as aggressive cropping or color changes. We then evaluated their performance on the other dataset (IN to SIN, SIN to IN) in order to use the generalization behaviour to gain insights about which features MLPs learn and how they use these to make predictions. Moreover we trained the models on higher resolution versions of the images (160x160) to determine whether the more readily available textural information will tilt MLPs towards focusing on texture.
The core insights of these experiments reveal themselves

through relative performances and directional changes so reaching high absolute accuracies is not a necessity. To nonetheless observe MLPs at higher accuracies we also investigated the impact of using more disruptive data augmentations, and we explored the effect of pre-training on a very large corpus of images (Imagenet21k) in conjunction with more aggressive augmentations by fine-tuning models by (Bachmann et al., 2023).

2.2. Interpretability methods

We used a number of complementary interpretability methods to gain a better understanding of how MLPs process and classify images in the context of texture and shape.

Pixel-level attention: To understand how much attention the MLP is paying to each input pixel, we compute the L2 norm of the gradient of the output logits with respect to each input pixel. We then average this quantity over multiple inputs to find the aggregated importance given to each pixel.

Activation maximization: To discern the network's specific output detections, we employed activation maximization, a technique generating images that maximize the activation of a chosen output. We adhered to an implementation by (stanford.edu). The method is based on (Simonyan & Zisserman, 2014) and (Yosinski et al., 2015). Parameters were qualitatively tuned to produce visually meaningful results.

Attribution maps: Attribution maps score input pixels by "relevance" or "contribution" (Ancona et al., 2019). We employed "Gradient" (Simonyan & Zisserman, 2014) to show the impact of perturbing input pixels on the network output and "Gradient * Input" (Shrikumar et al., 2017) to depict the contribution of individual input pixels to the final output. "SmoothGrad" (Smilkov et al., 2017) was added for noise reduction. We relied on implementations in "Captum" (Kokhlikyan et al., 2020).

2.3. Texture vs shape bias and distortion robustness benchmarks

To contextualize MLP texture vs. shape bias and distortion robustness within the existing literature, we utilized the "model-vs-human" toolbox (Geirhos et al., 2021). This tool enables cue conflict and distortion robustness experiments akin to (Geirhos et al., 2018a) and (Geirhos et al., 2018b) on ImageNet-trained models. The toolbox allows for comparisons with human observers and models from (Geirhos et al., 2018b), (Geirhos et al., 2018a), and (Geirhos et al., 2021). We assessed B-12/Wi-1024+DA by (Bachmann et al., 2023) against human observers and three baselines (ResNet-50, ResNet-50 SIN, and CLIP ViT-B/32) that include low and high shape biases, as well as recent ML developments.

3. Results

3.1. MLPs are more shape than texture focused

Overall, the generalization behavior of B-MLPs suggests they learn to primarily extract shape cues. This finding is supported by several observations.

The first indicator is that the drop in performance we observe for models trained on IN when applied to SIN is much smaller for MLPs than what was observed by (Geirhos et al., 2018a) for CNNs. Models can in principle use both texture and shape cues to classify on IN, while only shape cues are available for SIN. Models focusing on shape should comparatively generalize quite well and only a small performance drop would be observed. As the CNNs tested in (Geirhos et al., 2018a) predominantly focus on texture, they perform poorly when tested on SIN. (Geirhos et al., 2018a) finds that for CNNs the performance drops by about 80%, while we find the relative drop in performance going from IN to SIN to be about 30% (for both pre-trained and from scratch trained B-MLPs) (see Table 3.1).
Conversely, when training on SIN, a model is forced to focus on shape attributes, and thus should generalize well also to IN. We find that performance increases when a model trained on SIN is applied to IN, as expected, although the magnitude of the jump in performance for MLPs suggests that shape information can be much more readily extracted from IN than SIN. Indeed the stylising can severely affect the shapes, which seems to be the case upon human inspection (see Fig 1). The jump from SIN to IN seems to be more modest for CNNs ((Geirhos et al., 2018a)).

Model	IN/IN	IN/SIN	SIN/SIN	SIN/IN
B6-W512	58.7%	41.9%	43.3%	50.7%
B6-W512(IN21K)	84.9%	64.4%	71.3%	81.4%
B12-W1024	58.9%	42.0%	44.4%	52.3%

Table 1. Top-1 accuracy comparisons of MLPs trained from scratch or fine-tuned from baselines trained on IN21k by (Bachmann et al., 2023). "/" indicates: train or fine-tuning data / test data.

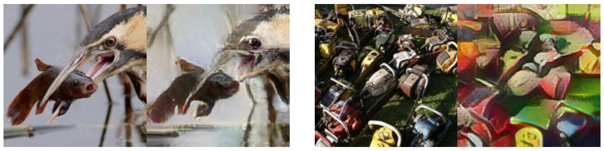


Figure 1. Left: shapes are preserved and easily recognizable. Right: shapes changed too much and are barely recognizable.

Secondly, increasing the training image sizes from 64x64 to their native resolution of 160x160 seems to not really improve the learning capabilities of the model, as the deviations in accuracy of the higher resolution models fall

within -0.5% to +0.1%. In principle it would be reasonable to expect that if MLPs had an easy time learning texture cues, then the much increased level of detail and increased model capacity should translate into higher accuracies.
Not only do MLPs naturally tend towards learning shape, it seems like it would be actively hard to tilt them towards learning texture. For CNNs (Hermann & Kornblith, 2019) finds that aggressive random crops tilt the model towards learning texture. This makes sense since aggressive cropping destroys the global shape of the object, and the network cannot learn to use that as a reliable feature. Adding aggressive random crops to our MLP training surprisingly manages to increase accuracy of both IN-IN and IN-SIN by approximately 2% and 1% respectively. The former is expected as the crops add variety to the data and the network can use texture to classify, but the latter is more surprising as the increase in performance must come from something other than texture. What we have discussed so far is also consistent with what we observe when fine-tuning models by (Bachmann et al., 2023) through linear probing. The transferability of the learning closely mimics our findings in the case where we train from scratch, even though in absolute terms we attain much higher accuracies thanks to the extensive pre-training. It is important to note that the pre-training was done with the use of aggressive cropping, so also here we see how cropping does not really bias MLPs towards texture. Probably aggressive crops push the MLPs towards learning more local, small scale shape features.

3.2. MLPs compared to other architectures

With the “model-vs-human” toolbox, we were able to largely reproduce texture vs shape bias and distortion robustness of both human observers and the three baseline models reported in (Geirhos et al., 2018b), (Geirhos et al., 2018a) and (Geirhos et al., 2021). The subsequent evaluations of B-12/Wi-1024+DA revealed several findings.

First, the shape bias of B-12/Wi-1024+DA (81.8%) is still below human levels (95.9%) (see Fig 2). However, its value surpasses not only ResNet-50 (21.4%), as expected from the experiments on the Imagenette dataset, but also CLIP ViT-B/32 (57.3%) and any other model evaluated in (Geirhos et al., 2021). We hypothesize that parameter sharing of ViTs makes such models rely more on texture than MLPs. Also, as seen in the Imagenette experiments (Table 1), there is hope that the shape bias of MLPs can be further increased if trained on a suitable data set.

Second, we see that having a high shape bias does not guarantee high general distortion robustness (see Fig 3). Besides a generally lower accuracy at no distortion, B-12/Wi-1024+DA shows a particularly high reliance on colors (see panels (a), (b), (c)), low frequency cues (see (e), (i)) and orientation (see (d)). We hypothesize that it is rather the training dataset size and selected augmentations that make a model robust, and not the shape bias itself. In future work, it might be insightful to study the distortion robustness of MLPs with additional data augmentations, such as stylization, rotations or color augmentations.

Third and most interestingly, we see that ResNet-50 SIN, which exhibits a shape bias (81.4%) very similar to that of B-12/Wi-1024+DA, is much more reliant on high frequency information in the input images. This suggests that shape can be seen as either a collection of edges, or as cohesively colored silhouettes and helps explain why random crops increase MLPs’ accuracies without tilting them to texture. Indeed, our interpretability methods support this finding. The activation maximisation plots in Fig 4 show that images maximising given output classes have a significant amount of long range patterns. This can also be seen in the weight maps of Fig 6 where there is some long range heterogeneity in the importance of input pixels. And Fig 5 reveals that the MLP assigns importance to object pixels in a more long range and less precise manner than the CNN.

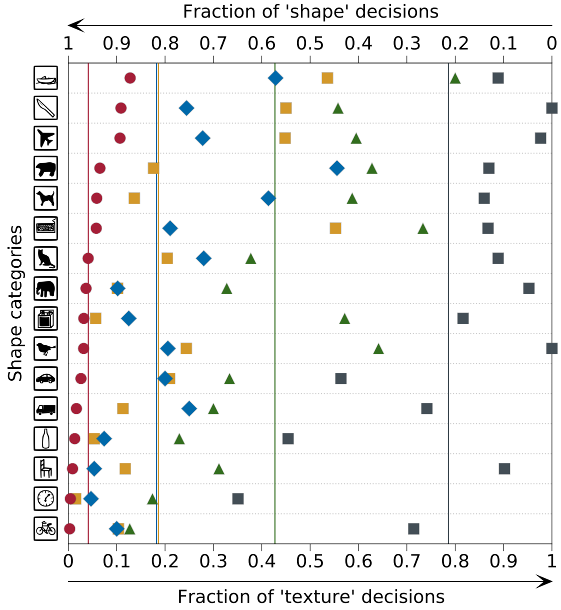
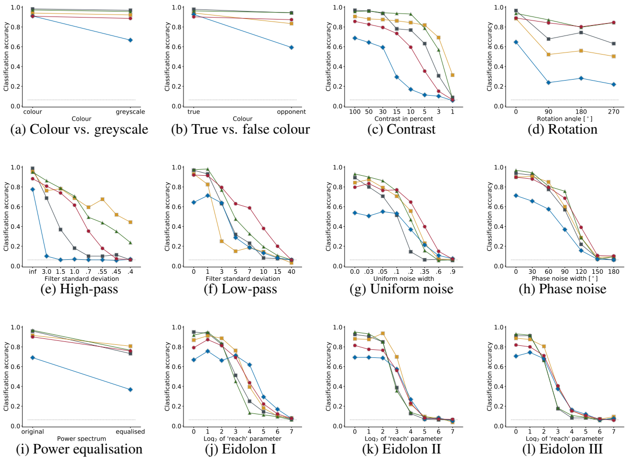


Figure 2. Shape vs. texture biases for stimuli with cue conflict for human observers (◯), ResNet-50 (□), ResNet-50 SIN (□), CLIP ViT-B/32 (△) and B-12/Wi-1024+DA (◇). Averages over all categories visualized by vertical lines.

3.3. Why MLPs behave differently to CNNs

To enhance our understanding of the underlying mechanics of MLPs, we follow the pixel-level attention methodology outlined in section 2.2. Throughout all our experiments involving MLPs, as well as those conducted by (Bachmann



(a) Colour vs. greyscale (b) True vs. false colour (c) Contrast (d) Rotation

(e) High-pass (f) Low-pass (g) Uniform noise (h) Phase noise

(i) Power equalisation (j) Eidolon I (k) Eidolon II (l) Eidolon III

Figure 3. Robustness towards parametric distortions for ResNet-50, ResNet-50 SIN, CLIP ViT-B/32 and B-12/Wi-1024+DA, as well as for human observers.

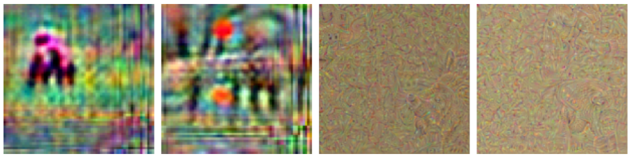


Figure 4. Activation maximization results of the ImageNet classifiers B-12/Wi-1024+DA (left) and ResNet-50 (right) for the target outputs “brown bear” and “african elephant”.

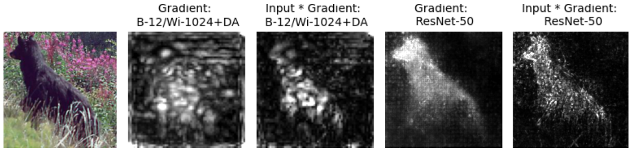


Figure 5. Attribution maps generated with “Gradient” and “Gradient * Input” for the ImageNet category “brown bear”.

et al., 2023), we observe distinct patterns indicating the varied significance of different input pixels. The importance of input pixels is very heterogeneous: most models demonstrate a focused attention on the image’s corners, borders, and center. To test that this visualization is indeed indicative of performance, we manipulate 16 pixels (or 0.4% of pixels) of the pretrained *B-12/Wi-1024* architecture. When choosing random pixels, the test performance drops by 0.05%, whereas selecting 4 pixels in each of the corners leads to a performance drop of about 45%. In contrast, CNNs assign homogeneous weights to all inputs by design due to weight sharing. To validate this, we conduct the weight analysis on ResNet-18, which revealed a uniform distribution of weights as expected. While MLPs have the possibility to learn convolutions, our trained MLPs seem to behave differently. This helps to explain why they do not immediately adopt the texture focus of CNNs.

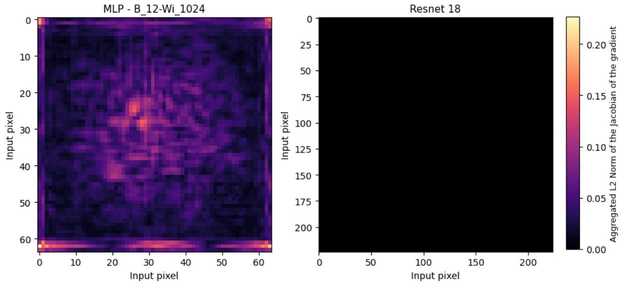


Figure 6. Sensitivity of predictions to different input pixels, comparing (Bachmann et al., 2023): *B-12/Wi-1024* to ResNet18

4. Discussion & Summary

Studying MLPs has significant relevance due to their reduced inductive bias compared to CNNs. The investigation of MLPs is not only relevant for Machine Learning but can also bring us closer to modeling the intricacies of human vision. Our study, which extends results involving CNNs and Vision Transformers to MLPs, has highlighted the substantial influence of architectural choices on texture and shape biases within neural networks. We showed that MLPs are generally biased towards using shape over texture as features to make predictions. This bias is stronger not only than in CNNs, but also than in the new Vision Transformer architectures, and approaches human-like bias level for shape.

However, an underdeveloped aspect is the impact of the training dataset. The notable divergence in performance between models trained on IN and IN21K (see Table 3.1) provides supporting evidence in this regard. Additionally, our findings show that MLPs do not inherently learn convolutional operations. Nevertheless, it is essential to acknowledge that this observed behavior might be influenced by the specific characteristics of the training data employed in our experiments. To gain deeper insights into the intricacies of human vision, aligning the data distribution more closely with the visual stimuli perceived by the human brain thus represents a promising path for future research.

A further aspect to look into is evaluating the performance of a hybrid model equipped with some inductive bias towards a Toeplitz matrix, but with the freedom of an MLP to learn the most useful features such as both low and high frequency patterns during training. This could be a generalized form of a CNN, in which weight sharing is enforced by the respective kernel used. Finally their propensity for identifying areas might prompt semantic segmentation as an interesting direction for future research.

References

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. Gradient-based attribution methods. *Explainable AI: Interpreting, explaining and visualizing deep learning*, pp. 169–191, 2019.

Bachmann, G., Anagnostidis, S., and Hofmann, T. Scaling mlps: A tale of inductive bias, 2023.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *CoRR*, abs/1811.12231, 2018a. URL http://arxiv.org/abs/1811.12231.

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., and Wichmann, F. A. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31, 2018b.

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

Hermann, K. L. and Kornblith, S. Exploring the origins and prevalence of texture bias in convolutional neural networks. *CoRR*, abs/1911.09071, 2019. URL http://arxiv.org/abs/1911.09071.

Howard, J. Imagenette. URL https://github.com/fastai/imagenette/.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMLR, 2017.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

stanford.edu. Cs231 assignment on network visualization (pytorch). URL https://shorturl.at/fhxHN.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., and Lipson, H. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.