The contrasting roles of shape in human vision and convolutional neural networks

Gaurav Malhotra & Jeffrey Bowers
Department of Psychological Sciences
University of Bristol
Bristol, BS8 1TU, UK
{gaurav.malhotra,j.bowers}@bristol.ac.uk

Abstract

Convolutional neural networks (CNNs) were inspired by human vision and, in some settings, achieve a performance comparable to human object recognition. This has lead to the speculation that both systems use similar mechanisms to perform recognition. In this study, we conducted a series of simulations that indicate that there is a fundamental difference between human vision and vanilla CNNs: while object recognition in humans relies on analysing shape, these CNNs do not have such a *shape-bias*. We teased apart the type of features selected by the model by modifying the CIFAR-10 dataset so that, in addition to containing objects with shape, the images concurrently contained non-shape features, such as a noise-like mask. When trained on these modified set of images, the model did not show any bias towards selecting shapes as features. Instead it relied on whichever feature allowed it to perform the best prediction – even when this feature was a noise-like mask or a single predictive pixel amongst 50176 pixels.

Introduction

Object recognition in humans is largely a function of analyzing shape (Biederman, 1987; Hummel, 2013). A wealth of data from psychological experiments show that shape plays a privileged role in object recognition compared to other diagnostic features such as size, colour, luminance or texture. For example, Biederman and Ju (1988) showed that error rates and reaction times are virtually identical in a recognition task when full coloured photographs of objects are replaced by their line drawings even when colour was a diagnostic feature. This indicates that shape-based representations mediate recognition. Similarly, Mapelli and Behrmann (1997) found that, for patients with an object recognition deficit (visual agnosia), surface colour played minimal role in aiding object recognition unless the shape of the object was ambiguous, indicating that shape is instrumental to recognition, whereas surface characteristics such as colour and texture play only a secondary role. More recently, Baker and Kellman (2018) have shown that participants extract shape information automatically from arrays of dot patterns within the first 100ms of stimulus onset, even for tasks where extracting this information may be detrimental to performance on a task. Experiments from developmental psychology show that this privileged status of shape starts early in life and becomes stronger with age. For example, Landau, Smith, and Jones (1988) found that 2-3-year-old children as well as adults weight shape more heavily than size or texture when generalising the name of a learnt object to novel instances. They also found that the weight placed on shape increases in strength and generality from early childhood to adulthood.

By contrast, it is unclear whether shape plays a privileged role in how convolutional neural networks (CNNs) categorise objects. It is often claimed that CNNs learn representations of objects that are similar to the representations that monkeys and humans use when identifying objects (Rajalingham et al., 2018), and that CNNs largely rely on learning shape representations in order to categorise objects (Kubilius, Bracci, & de Beeck, 2016; Jozwik, Kriegeskorte, Storrs, & Mur, 2017). On the other hand, there are a growing number of studies that show that CNNs often categorise images on the basis on non-shape attributes of images. This is demonstrated by the existence of adversarial images that are confidently classified as a familiar category despite the lack of any shape information in the input (Nguyen, Yosinski, & Clune, 2015), adversarial images that contain the correct shape but altered colours that are confidently misclassified (e.g., categorizing an image of an airplane as a dog when only the colour of the plane has been manipulated), and large reductions in performance when trained coloured images are converted to greyscale (Geirhos et al., 2017) or the colours are inverted (Hosseini, Xiao, Jaiswal, & Poovendran, 2017). In addition, there are demonstrations that CNNs can easily learn to categorise random patterns of pixels that have no shape (Zhang, Bengio, Hardt, Recht, & Vinyals, 2016). All of these findings suggest that shape may not play a privileged role in how some well-known and high-performance CNNs perform object categorisation.

However, some recent studies have argued that convolutional neural networks can show a shape-bias. Ritter, Barrett, Santoro, and Botvinick (2017) took an Inception model, a high-performance CNN (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016), and presented novel objects to the model that had been pre-trained to recognise the categories from ImageNet dataset. They found that the representations in hidden layers were more similar for two (novel) objects that overlapped in shape than for two objects that overlapped in colour. They interpret this proximity in hidden layer representations between objects of same shape as a shape-bias. In another study, Feinman and Lake (2018) trained a CNN on a controlled dataset containing synthetic images that differed on three dimensions: shape, colour and texture. They found that when this dataset was constructed in such a manner that the

(a) Salt-and-pepper noise (b) Additive noise (c) Single diagnostic pixel

Figure 1: Hidden in plane sight. Images taken from CIFAR-10 dataset and scaled up to 224x224 pixels. (a) Image is converted to greyscale and we add a salt-and-pepper noise-like mask to each training image; (b) Image is converted to greyscale and we add uniform additive noise mask to each training image; (c) A single diagnostic pixel is inserted in the image (dotted red circle is inserted here to illustrate the location of the pixel).

category name correlated with shape more than colour or texture, the network had a higher propensity for classifying novel objects based on shape rather than colour or texture. In other words, the network learns to reflect the feature bias of the training set; when the biased feature is a shape, the network shows shape-bias.

Both these studies assume that shape-bias is a property of the environment itself. Feinman and Lake (2018) explicitly make shape more diagnostic than any other feature in the dataset, while Ritter et al. (2017) assume that this is implicitly the case. However, it is not not clear that shape is necessarily the most diagnostic feature in the environment of biological systems and it is also unclear whether deep neural networks would develop an inductive bias for shape when this is not the most diagnostic feature. Our goal in this study was to test the stronger claim that CNNs show a shape-bias even when there is no such bias in the dataset. Within the psychological literature it is still unsettled whether our visual system identifies objects on the basis of shape because we learn through experience that shape is the most reliable cue to object identification or because there are innate inductive biases that make shape a privileged cue from the beginning (for discussion see Elman, 2008; Xu, Dewar, & Perfors, 2009).

It is certainly possible that CNNs have an inductive bias to rely on shape given that the depth of the architecture and pooling operations enables them to combine features of the stimuli in a hierarchical manner where lower layers represent high-frequency features while higher layers represent more abstract features, such as the shape, which are invariant to local changes of input (Bengio, Courville, & Vincent, 2013). If shape emerges due to this hierarchical composition of features, it is possible that it is preferred to other features (such as colour or texture) that do not lend themselves to such a hierarchical composition. Henceforth we use the term shape-bias to refer to the hypothesis that the visual system has an innate inductive bias to rely on shape cues to identify objects rather than the view that the visual system learns to identify objects on the basis of whatever visual cues are most strongly associated with object category.

Here we systematically explore the impact of non-shape features in the categorisation performance of convolutional neural networks on CIFAR-10 images. We introduced non-shape features to images by adding informative noise-like masks to the training set. We tried several types of masks and an extreme version where the non-shape feature consisted of just a single pixel with a location correlated to the image category (see Figure 1). We show that vanilla CNNs, that perform object classification on CIFAR-10 to near human level, nevertheless learn and depend on non-shape features that are highly diagnostic of object categories and often fails to learn anything about shape under these conditions. These results did not depend on the type of network architecture used, the learning algorithm or regularisation method indicating that this was a property of a broad class of CNNs rather than the particular setup chosen by us. This highlights that, even though they mimic the hierarchical architectural and learning processes of biological vision, the vanilla architectures and algorithms for learning in CNNs simply pick up whatever statistical structure is most relevant to learning the training set, with shape playing no special role. To dispel any confusions at the outset, we would like to emphasise that this does *not* imply that CNNs do not encode shape information under any circumstance, but that shape does not seem to be weighted more than other diagnostic features, even when these features are noise-like masks or the luminance of a single pixel.

Experiments

We modified the CIFAR-10 dataset (which contains 10 classes with 6000 images per class, see `https://www.cs.toronto.edu/~kriz/cifar.html`) so that each image contained not only features that pertain to the shape (e.g. object outlines) but also features without any shape information. As non-shape features we used noise-like masks that were combined with the original image. Two different types

of masks were used: the *salt-and-pepper noise mask* turned a certain proportion of image pixels to either black or white, while a *additive uniform noise mask* added a value sampled from a uniform distribution to each pixel of an image. We also tested an extreme form of the salt-and-pepper noise mask where only one pixel was turned to a particular colour. In this case the location and colour of the pixel were different for different categories but correlated for images within a category. Masks were independently sampled for each category but were either fixed for all images in a category (in which case the mask predicted the category) or sampled from a distribution with category-dependent parameters (in which case these parameters predicted the category). So these modified images concurrently contained features that were related to shape and features without shape information.

We trained the model on these modified sets of images and tested it under three conditions. During the 'Same' condition, the test set was modified in exactly the same manner – i.e., either images in each category were generated by using the same mask as that for the training images of that category (when the mask was fixed) or they were generated by using the same parameters as the parameters used to generate noise masks for training images of that category (when the mask was variable). In contrast, during the 'Diff' condition, the noise masks (or their parameters) for each category were swapped with another category. So, for example, a noise mask that was used in the 'DOG' category during training was inserted into images in the 'CAT' category during testing. The premise here was that if the model based it's decisions on shape-related features, then it would ignore the noise mask and the performance during 'Same' and 'Diff' condition should be similar. On the other hand, if the model relied on properties of the (non-shape) mask, then it's performance would be worse in the 'Diff' condition compared to the 'Same' condition. Finally, we used a third, 'NoPix', condition to estimate the extent to which the network relied on features of the noise mask. In this condition, we presented the network with a version of the image without any mask, with the premise that the difference between the performance in 'Same' and 'NoPix' condition should quantify the relative extent to which the network relied on shape-based and non-shape features. We ran all of the simulations using the well-known VGG-16 network (Simonyan & Zisserman, 2014) and checked that our main results replicate for a deeper network, ResNet-101 (He, Zhang, Ren, & Sun, 2016). To give the model the best chance to recognise shape-based features, all simulations were carried out on CNNs that had previously been trained on ImageNet categories and replaced only the fully-connected layers to perform the new classification task. We then turned the learning rate to a small value and trained these networks on the new classification task.

Methods

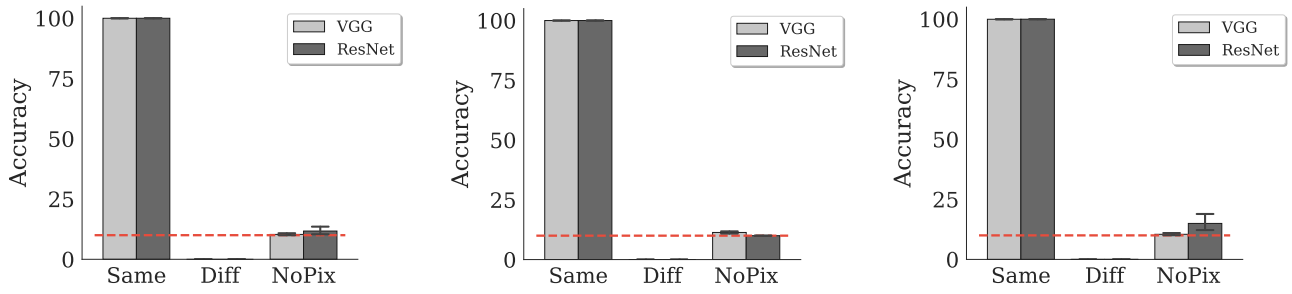
We used a method similar to Geirhos et al. (2017) to transform images from the CIFAR-10 dataset. All transformations were performed using the Pillow fork of the Python Imaging Library (`https://pillow.readthedocs.io`). Each 32x32 pixel image was rescaled to 224x224 pixels using the `PIL.Image.LANCZOS` method. For the single-pixel mask, we used 3-channel RGB images while for the salt-and-pepper and additive noise mask, we transformed images to greyscale. When images were transformed to greyscale, their contrast was adjusted to 80% by scaling the value of each pixel using the formula: $0.8 \times v + \frac{1-0.8}{2} \times 128$, where v was the original value of the pixel in the range $[0, 255]$.

The salt-and-pepper mask was created by taking the transformed greyscale image and setting each pixel to either black or white with a probability p. When the mask was fixed for a category (Experiment 1–3 below), all images had the exact same set of pixels that were turned either black or white and the p was set to 0.05. When the mask varied from image to image within a category (Experiment 4 below), the pixels were sampled independently for each image and the probability p was fixed for each category but varied between categories in the range $[0.03, 0.06]$.

The additive uniform noise mask was created by taking the transformed greyscale image and adding a value sampled from the uniform distribution $[-w, w]$ to this image, where $2w$ was the width of the uniform distribution and was set to 8. When the noise mask was fixed, this sampling was done only once per category and the same mask was added to each image. When the mask was variable, it was sampled independently for each image from a distribution $[\mu - w, \mu + w]$, where μ was the mean that depended on the category and varied in the range $[-50, 50]$.

The single pixel mask was created by choosing a random location, (x, y), (sampled from a uniform distribution on the interval $[0, 224]$) on the image and changing the colour of the pixel to a value c (sampled from a uniform distribution on the interval $[0, 255]$). When the mask was fixed for each category, (x, y, c) remained constant for all images in a category, but varied between categories. When the mask was variable, each of x, y and c were sampled independently for each image from a Gaussian distribution with a constant variance and a mean that depended on the category of the image. If any value in a sampled set of (x, y, c) values fell out of their respective range, that value was re-sampled.

Simulations were carried out using either a 16-layer VGG network (Simonyan & Zisserman, 2014) or 101-layer ResNet network provided by the `torchvision` package of `Pytorch`. These networks were either trained from the scratch on the modified dataset or were first pre-trained on ImageNet and then trained on the modified dataset. When the networks were pre-trained, we replaced the fully-connected layers of the VGG/Resnet pre-trained model with three/one fully-connected layer(s) with 10 units (for 10 categories) on the output layer. Since the results remain qualitatively the same, we report the results for the networks pre-trained on ImageNet. We tried a number of different optimization algorithms, including RMSProp, SGD and Adam (Kingma & Ba, 2014). Results again remained qualitatively the same. We



(a) Salt-and-pepper noise (b) Additive noise (c) Single diagnostic pixel

Figure 2: Accuracy on test images under the three types of noise-like masks shown in Figure 1. 'Same': the noise-like mask has same properties for test and training images of each category; 'Diff': the properties of the mask during test are swapped with another category from training; 'NoPix': No mask is inserted. The dashed (red) line indicates chance performance and error bars show 95% confidence interval. Light and dark gray bars show accuracies on `VGG-16` and `ResNet-101`.

started with a learning rate of $1e-3$ when training the network from scratch and used a learning rate of $1e-5$ when fine-tuning a pre-trained network. In all cases, we used cross-entropy as the loss function. The input to both types of networks was a 3-channel RGB image. For greyscale images, all three channels were set to the same value.

Experiment 1

In the first experiment, all images in a category had the exact same noise mask. For salt-and-pepper mask, this meant that noise masks were sampled independently for each category, but the same set of pixels in each image were modified for all images in a category. Similarly, for the additive uniform noise mask, the same mask was added to each image in a category. For the single pixel noise, the location and colour of the added pixel were independently sampled for each category, but kept constant for all images in a category.

The results of the first experiment are shown in Figure 2. We obtain the same pattern of results for all three cases: when noise mask in the test images matches the noise mask in training images, the model classifies images nearly perfectly; when noise masks are swapped, the accuracy drops to zero; when the mask is completely removed, the categorisation accuracy is at chance. Furthermore, we get the same pattern of results on both VGG and ResNet networks and irrespective of the type of regularisation used (we tried several well-known regularisation methods including *Batch Normalization*, *Weight Decay* or *Dropout*). These results clearly indicate that the model learns to completely rely on features of the noise-like mask, rather than any shape-related information present in the images. Even in the extreme case, where only one pixel amongst 50176 was diagnostic of the category, the model prefers to classify based on this feature over other shape-related features present in each image.

Experiment 2 & 3

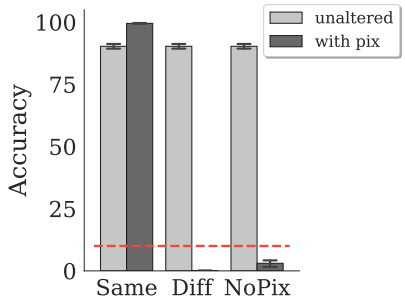
One possible reason why humans prefer to rely on shape-related features to categorise objects while CNNs do not is that humans are guided by past experience and bring this past knowledge to new categorisation tasks. So when a human sees an object with superimposed noise, they generalise from past experience and look for shape-based information, paying less attention to non-shape related features such as the noise-like mask in above images. We conducted two further experiments to test whether networks similarly generalise from concurrent and past experience.

In Experiment 2, we divided the training set into two subsets. The first subset ('with pix') contained three randomly chosen categories from CIFAR-10 and, like above, contained a category-correlated pixel in all images of these categories. The second subset ('unaltered') contained the remaining seven categories from CIFAR-10 and was left unaltered – i.e. we did not add the category-correlated pixel to images of this subset. We trained a VGG-16 network on all ten categories at the same time. We were interested in finding out whether the network generalised from one subset to another and started using the features used to categorise images in the 'unaltered' subset to images of the 'with pix' subset. All other details of the experiment remain same as Experiment 1.

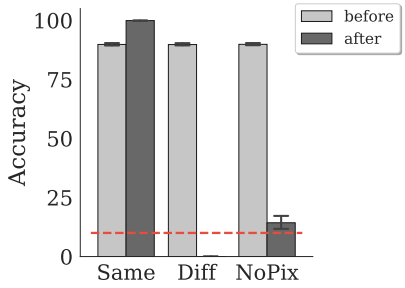
The results from this experiment are shown in Figure 3a. The model learnt to predict the images in the 'unaltered' subset with nearly 90% accuracy. However the performance on the 'with pix' subset still completely depended on the location and colour of the added pixel: accuracy was nearly 100% when test images contained the pixel in the same location, but dropped below chance when this pixel was removed. Thus, the network did not seem to generalise the features (concurrently) learnt in the 'unaltered' categories to the categories containing the diagnostic pixel.

In Experiment 3 we tested what happens when the network is first trained on images that did not contain such a pixel (a 'before' phase) followed by a second ('after') phase in which such a pixel was inserted in the training set. In the first phase, a we trained a VGG-16 network on an unaltered CIFAR-10 training set. Once the network had learnt this task, we trained it on the modified set of images in a second phase, introduc-

ing a predictive pixel in each category. So all that changes between the 'before' and 'after' phases is the insertion of a single category-correlated pixel to each image.



(a) Generalising between subsets



(b) Generalising from one time to another

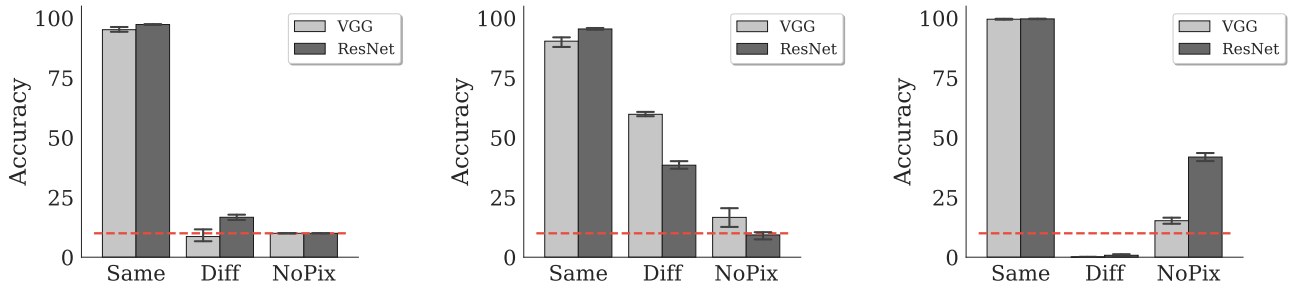
Figure 3: Lack of generalisation. Accuracy under Same, Diff and NoPix conditions for (a) two subsets: an 'unaltered' subset where no noise-like mask was inserted in training images and a 'with pix' subset where a single diagnostic pixel was inserted, and (b) for two phases: a 'before' phase, where a pre-trained VGG network was trained on images without any noise masks and tested on the three conditions, and an 'after' phase, where the model from before phase was then trained on images with a single diagnostic pixel.

We observed that (Figure 3b), instead of relying on past experience with these images, the model learnt to completely rely on the predictive pixel to perform categorisation – accuracy dropped from nearly 100% to 0% between 'Same' and 'Diff' conditions. Crucially, the model completely forgot about how to perform categorisation when the predictive pixel was removed – accuracy was close to chance in the 'NoPix' condition during the 'after' phase. Thus learning about the diagnostic feature seemed to be accompanied by unlearning previously learnt representations. This, catastrophic forgetting, is a well-known problem in neural networks (McCloskey & Cohen, 1989) and contrasts with how humans transfer their knowledge from one task to another. Some recent solutions to catastrophic learning in neural networks have been suggested, such as Elastic Weight Consolidation (Kirkpatrick et al., 2017) and it remains to be seen whether this can overcome some of these problems.

Experiment 4

The non-shape features used in the experiments above have all been completely invariant from one image to another within a category. It can be argued that these features are selected by the model over other shape-based features because they provide a very strong predictive signal. It is possible that if these features contained larger variance, the model would be more likely to rely on shape-based features while performing categorisation. In the next experiment, we introduced variability in the non-shape features by sampling the noise-like mask independently from a distribution for each training and test image within a category. In order to make these noise-like masks diagnostic of an image's category, a parameter of this distribution correlated with an image's category. For the salt-and-pepper noise, this meant that the probability, p, of changing a pixel to black or white was different for each category. Thus, the parameter, p, became diagnostic of the category. However, the masks now varied from image to image and were independently sampled with the (category-dependent) probability, p. Similarly, for the additive uniform noise, masks could vary from one image to other within a category but the mean of the distribution depended on each category (see Methods above for details). For the single diagnostic pixel, the inserted pixel could vary in location and colour from one image to the other, but were generated from a Gaussian distribution with a mean determined by the category of the image and a fixed standard deviation. We ran these simulations on both VGG-16 and Resnet-101 and aside from the way in which the dataset was generated, all other details remain same as Experiment 1.

The results of introducing a variable noise mask are shown in Figure 4. Introducing variability in the location and colour of the single diagnostic pixel brought very little change to the VGG model's behaviour (compare Figure 4c with Figure 2c). Performance in the NoPix condition was somewhat better for ResNet, however the pattern of result remained the same – performance dropped substantially from the Same to NoPix condition. Similarly, introducing variability in the salt-and-pepper masks lead to only a minor change in behaviour of the model, with accuracy in 'Diff' condition dropping to chance, rather than 0%. The most intriguing change in behaviour occurred when variability was introduced to the additive uniform noise mask (Figure 4b). While the VGG and ResNet networks differed quantitatively in these results, the pattern of results remained the same: when the noise mask was completely removed (NoPix condition) the model performed *worse* than when the images contained a noise mask from a different category (Diff condition). In other words, removing the mask makes the image less informative for the model, not only compared to images with the correct category-correlated (Same) mask, but also compared to images with the incorrect (Diff) mask – the model seems to rely on the presence of the noise-like mask to make an inference.



(a) Salt-and-pepper noise (b) Additive noise (c) Single diagnostic pixel

Figure 4: Accuracy on test images when the noise mask varies between images of a category. Training images contain (a) salt-and-pepper noise, or (b) additive uniform noise, or (c) just one diagnostic pixel. The dashed (red) line indicates chance performance. See Figure 2 for a description of the 'Same', 'Diff' and 'NoPix' conditions.

Related Work

Su, Vargas, and Kouichi (2017) demonstrated that CNNs trained on CIFAR-10 and ImageNet can be fooled by introducing a single adversarial pixel, with error rates of 68% and 41%, respectively. Unlike our approach the model was trained with uncorrupted images and the authors systematically searched for an adversarial pixel that lead to any sort of error (so-called non-targeted attack). So, in contrast to our goal, the goal of their study was not to explore whether CNNs systematically learn non-spatial information. However, the findings are in line with ours – the CNNs trained by them do not seem to be categorising based on shape. Rather, it must be that there was, by chance, some pixel value that was highly correlated with a given output category and the model picked up on this idiosyncratic correspondence. As a consequence, when this pixel was added to another category the model was fooled.

Two recent studies – Geirhos et al. (2018) and Baker, Lu, Erlikhman, and Kellman (2018) – manipulate the texture and shape of images independently and show that CNNs trained on ImageNet are biased towards picking up texture compared to shape. These results are again in line with our results and show that CNNs will make inferences on whichever feature is most predictive in the training set. Indeed, when Geirhos et al. (2018) make the texture less diagnostic of category, the model seems to use non-texture features for performing classification. Our findings go beyond past work by highlighting the extent to which CNNs categorize objects on the basis of non-shape features even when it is given concurrent or prior training without such non-shape features. Indeed, a single diagnostic pixel can override all the shape information present in the training images.

Conclusions

In a series of simulations we found that some high-performance convolutional networks trained to categorise CIFAR-10 images that included noise-like masks diagnostic of the output categories often learned to categorise on the basis of these masks rather than features present within the CIFAR-10 images themselves. Indeed, the models often entirely relied on the masks, and performed at floor when the noise was removed from the images. This clearly highlights that, when a shape-bias is not present within the training dataset itself, these models do not show a shape-bias due to their own architectural or algorithmic properties.

In our experiments, we specifically engineered our dataset to contain invariant non-shape features. One might object that large datasets like ImageNet and CIFAR-10 don't contain such features so that the models trained on these datasets end up relying on shape to perform categorisation. But it is well-known that popular datasets contain various biases due to conditions under which the images were captured as well as the different motivations for construction of the datasets (Torralba & Efros, 2011). So biases like the one we engineered may well be present in these datasets and networks trained on these datasets may be picking on these features. This, in turn, implies that these networks may be relying on entirely different set of features and representations to perform classification than human beings or other animals.

If CNNs do indeed rely too heavily on non-shape features present within datasets, it could also be the source of various idiosyncratic behaviours such as being confounded by fooling images (Nguyen et al., 2015) or being overly sensitive to colour (Hosseini et al., 2017), noise (Geirhos et al., 2017) or even single pixels in images (Su et al., 2017). The alternative hypothesis that the human visual system learns to categorize objects on whatever statistical regularities are strongest in the input cannot be ruled out on the basis of our findings, but it would predict that humans would show a similar pattern of result to these models, such as picking up on single pixels or noise-like masks to categorise stimuli. In addition, this view also needs to explain why human beings are not susceptible to adversarial attacks such as the non-shape fooling images in the same manner as vanilla CNNs. We are currently carrying modelling and behavioural work to provide further insights into the computational benefits of inducing a shape-bias to CNNs and how these modified CNNs relate to human vision.

References

Baker, N., & Kellman, P. J. (2018). Abstract shape representation in human visual perception. *Journal of Experimental Psychology: General*, *147*(9), 1295.

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, *14*(12), e1006613.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological review*, *94*(2), 115.

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive psychology*, *20*(1), 38–64.

Elman, J. L. (2008). The shape bias: an important piece in a bigger puzzle. *Developmental science*, *11*(2), 219.

Feinman, R., & Lake, B. M. (2018). Learning inductive biases with simple neural networks. *arXiv preprint arXiv:1802.02745*.

Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969*.

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).

Hosseini, H., Xiao, B., Jaiswal, M., & Poovendran, R. (2017). On the limitation of convolutional neural networks in recognizing negative images. In *Machine learning and applications (icmla), 2017 16th ieee international conference on* (pp. 352–358).

Hummel, J. E. (2013). Object recognition. *Oxford handbook of cognitive psychology*, 32–46.

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in psychology*, *8*, 1726.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, *114*(13), 3521–3526.

Kubilius, J., Bracci, S., & de Beeck, H. P. O. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, *12*(4), e1004896.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive development*, *3*(3), 299–321.

Mapelli, D., & Behrmann, M. (1997). The role of color in object recognition: Evidence from visual agnosia. *Neurocase*, *3*(4), 237–247.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation* (Vol. 24, pp. 109–165). Elsevier.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 427–436).

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, 0388–18.

Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. *arXiv preprint arXiv:1706.08606*.

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

Su, J., Vargas, D. V., & Kouichi, S. (2017). One pixel attack for fooling deep neural networks. *arXiv preprint arXiv:1710.08864*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2818–2826).

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In *Ieee conference on computer vision and pattern recognition (cvpr)* (pp. 1521–1528).

Xu, F., Dewar, K., & Perfors, A. (2009). Induction, overhypotheses, and the shape bias: Some arguments and evidence for rational constructivism. *The origins of object knowledge*, 263–284.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.

Acknowledgement

"This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 741134)".