Adding biological constraints to CNNs makes image classification more human-like and robust

Gaurav Malhotra
School of Psychological Science
University of Bristol
Bristol, BS8 1TU, UK
gaurav.malhotra@bristol.ac.uk

Benjamin D. Evans
School of Psychological Science
University of Bristol
Bristol, BS8 1TU, UK
benjamin.evans@bristol.ac.uk

Jeffrey S. Bowers
School of Psychological Science
University of Bristol
Bristol, BS8 1TU, UK
j.bowers@bristol.ac.uk

Abstract

In this study, we show that when standard convolutional neural networks (CNNs) are trained end-to-end on datasets containing low-level and spatially high-frequency features, they are susceptible to learning these potentially idiosyncratic features if they are predictive of the output class. Such features are extremely unlikely to play a major role in human object recognition, where instead a strong preference for shape is observed. Through a series of empirical studies, we show that standard CNNs cannot overcome this reliance on non-shape features merely by making training more ecologically plausible or using standard regularisation methods. However, we show that these problems can be ameliorated by forgoing end-to-end learning and processing images initially with Gabor filters, in a manner that more closely resembles biological vision.

Introduction

A fundamental property of human image recognition is that it is largely a function of analyzing shape (Biederman, 1987). A wealth of data from psychological experiments show that the global shape of an object plays a privileged role in object recognition compared to other diagnostic features such as size, colour, luminance or texture (Biederman & Ju, 1988; Landau, Smith, & Jones, 1988). In other words, it has a *shape-bias*. However, it is still unsettled whether we learn to have a shape-bias through experience or there are innate inductive biases that make shape a privileged cue (see Elman, 2008 and Xu, Dewar, & Perfors, 2009).

Similarly there are two possible reasons why CNNs trained in an end-to-end manner may develop an inductive bias to rely on shape. Some recent studies have argued that shape may be the most diagnostic feature in a trained dataset and this causes the CNN to learn to rely on shape to perform categorisation – i.e., CNNs can have a *learned* shape-bias (Ritter, Barrett, Santoro, & Botvinick, 2017; Feinman & Lake, 2018). On the other hand, a shape-bias might be *innate* and the product of the architecture of the CNN itself. For instance, the multiple layers and pooling operations enable a CNN to combine features of the stimuli in a hierarchical manner, and this might result in lower layers representing high-frequency features and higher layers representing more abstract features, such as shape (Bengio, Courville, & Vincent, 2013).

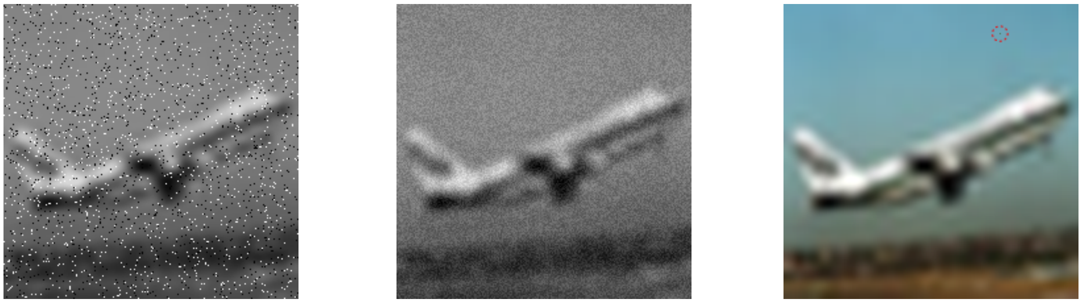
Our goal in this study was to tease apart whether any shape-bias is learned or innate in standard CNNs. To do this, we trained some standard CNNs on a dataset that modified the standard `CIFAR-10` dataset to simultaneously contain shape-based and non-shape features (Figure 1). We found that standard CNNs trained on this modified dataset learn to depend on non-shape features that are diagnostic of object categories and often fail to learn anything about shape under these conditions. These results suggest that that vanilla CNNs do not have an innate shape-bias. (Note 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).

We hypothesised that the lack of innate shape-bias in standard CNNs reflects a lack of an innate biological constraint in how they model human vision. To test this hypothesis we replaced the first convolutional layer of a standard CNN with a bank of unmodifiable Gabor filters designed to mimic simple cells in V1 cortex. We found that doing this, comes at a cost to the network's overall performance but made the CNN far less reliant on non-shape features, such as noise-like masks or single diagnostic pixels. We also found that these results were robust across a range of neurophysiologically plausible parameters for the Gabor filters showing that it's not the particular value of the parameters but the process of filtering itself that makes classification robust to non-shape features present within the dataset.

Methods

We modified the `CIFAR-10` dataset so that each image contained not only features that pertain to the shape (e.g. object outlines) but also features without shape information. As non-shape features we used three types of noise-like masks that were combined with the original image. 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. This 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 $[\mu - w, \mu + w]$ to this image, where μ was the mean that depended on the category and varied in the range $[-50, 50]$ and $2w$ was the width of the uniform distribution and was set to 8. 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 $[0, 255]$). 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.

We trained the model on these modified sets of images and



(a) Salt-and-pepper noise

(b) Additive noise

(c) Single diagnostic pixel

Figure 1: A modified dataset consisting of images taken from CIFAR-10 dataset (scaled up to 224x224 pixels) as well as a noise-like mask. (a) Image + salt-and-pepper noise-like mask; (b) Image + uniform additive noise mask; (c) Image + a single diagnostic pixel (a dotted red circle is inserted here to illustrate the location of the pixel).

tested it under three conditions. During the 'Same' condition, the test set was modified in exactly the same manner as the training images. In contrast, during the 'Diff' condition, the parameters of the noise masks for each category were swapped with another category, creating a cue conflict. Finally, during a 'NoPix' 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.

Simulations were carried out using either a 16-layer VGG network or 101-layer ResNet network provided by the torchvision package of Pytorch and Keras with TensorFlow. These networks were either trained from scratch on the modified dataset or were first pre-trained on ImageNet and then trained on the modified dataset. Since the results remain qualitatively the same, we report the results for the networks pre-trained on ImageNet.

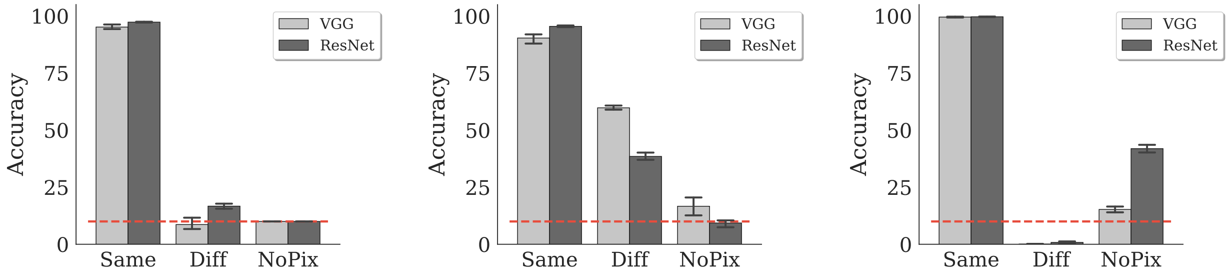
Results

We conducted three experiments, one for each type of noise mask described above. The results are shown in Figure 2. During all three experiments, we observed that both networks classify images with a nearly perfect accuracy during the 'Same' noise condition. When noise masks are swapped ('Diff' condition), this accuracy drops; when the mask is completely removed ('NoPix' condition), the categorisation accuracy is nearly at chance. For both the *salt-and-pepper* and *single pixel* experiments, performance in the 'Diff' condition is either at chance or below chance indicating that the trained network learns to categorise based on the noise-like mask in both experiments. During the *single pixel* experiment, accuracy in the 'NoPix' condition was somewhat better for ResNet-101 than VGG-16 indicating that, in this case, the network may be picking on some other features of the image beside the noise-like mask. However, even in this case, there is a significant drop in performance compared to 'Same' condition. The *additive noise* experiment showed an intriguing behaviour: 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.

In further experiments, we also examined whether the type of training had any effect on these results. In one experiment, we modified the dataset such that only some of the categories contained a diagnostic pixel. In another experiment, we trained the network on an unmodified CIFAR-10 training set before subsequently training on the modified training set containing a diagnostic pixel. In both experiments, results remained qualitatively the same as above – i.e., if the network learnt anything about shape, it wasn't able to generalise this knowledge across categories or across time. Furthermore, we obtained the same pattern of results irrespective of the type of regularisation used (we tried several well-known regularisation methods including *Batch Normalization*, *Weight Decay* and *Dropout*) or type of optimisation algorithm (SGD / RMSProp / Adam). These results clearly indicate that the model learns to 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 50,176 was diagnostic of the category, the model prefers to classify based on this feature over other shape-related features present in each image.

A biologically plausible feature space

In this section, we tested the hypothesis that adding a biological constraint may make the network less reliant on the noise mask types that are diagnostic of output categories of stimuli. To do so, we replaced the first convolutional layer of VGG-16 with unmodifiable Gabor filters, rather than allow the model to form its own feature space end-to-end. Gabor filters have been shown to be a good model of the simple cell receptive fields found in the early visual cortex of cats (Jones & Palmer, 1987) and primates (Petkov & Kruizinga, 1997) and are regarded as the standard model of simple cells amongst neuroscientists.



(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. Training images contain (a) salt-and-pepper noise, or (b) additive uniform noise, or (c) just one diagnostic pixel. Test performance is shown under three conditions – ‘Same’: the mask has same parameters during testing and training within each category; ‘Diff’: the parameters of masks used during testing are swapped with another category; ‘NoPix’: No mask is inserted. The dashed (red) line indicates chance performance and error bars show 95% confidence intervals.

Methods

The Gabor function is an oriented sinusoidal grating convolved with a Gaussian envelope:

$$g_{\lambda,\theta,\phi,\sigma,\gamma}(x,y)=\exp\left(-\frac{x_\theta^2+\gamma^2 y_\theta^2}{2\sigma^2}\right)\exp\left(i\left(\frac{2\pi x_\theta}{\lambda}+\phi\right)\right) \tag{1}$$

with the following definitions:

$$x_\theta = x\cos\theta + y\sin\theta \qquad y_\theta = -x\sin\theta + y\cos\theta \tag{2}$$

where x and y specify the position of a light impulse in the visual field (Petkov & Kruizinga, 1997).

Rather than specify the width of the Gaussian component in pixels, it is more natural to set the bandwidth, b, which describes the number of cycles of the sinusoid within the Gaussian envelope. The standard deviation of the Gaussian factor, σ, is therefore set indirectly through b, and λ:

$$\sigma = \frac{\lambda}{\pi}\sqrt{\frac{\ln 2}{2}}\cdot\frac{2^b+1}{2^b-1} \tag{3}$$

Throughout each simulation where Gabor filters were used, the first convolutional layer of VGG-16 was replaced with a fixed bank of Gabor filters (each 31×31 pixels) designed to model the early primate visual cortex and match the number of output channels (64) defined in the original CNN. Each such bank had eight orientations, $\theta \in \left\{0, \frac{\pi}{8}, \frac{\pi}{4}, \frac{3\pi}{8}, \frac{\pi}{2}, \frac{5\pi}{8}, \frac{3\pi}{4}, \frac{7\pi}{8}\right\}$ radians, four phases, $\psi \in \left\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\right\}$ radians, and two aspect ratios, $\gamma \in \left\{\frac{1}{2}, 1\right\}$, (defining the ellipticity of the filter) while the wavelength, λ, and bandwidth, b, were systematically varied.

As with the previous experiments, `CIFAR-10` images were mainipulated by adding one of the following types of noise: *salt-and-pepper*, *additive uniform* or *single pixel* but remained in their original size of 32×32 pixels. All images were converted to greyscale according to the ITU BT.601 conversion formula ($Y = 0.299\cdot R + 0.587\cdot G + 0.114\cdot B$) and presented under the same training and test conditions described previously.

Results

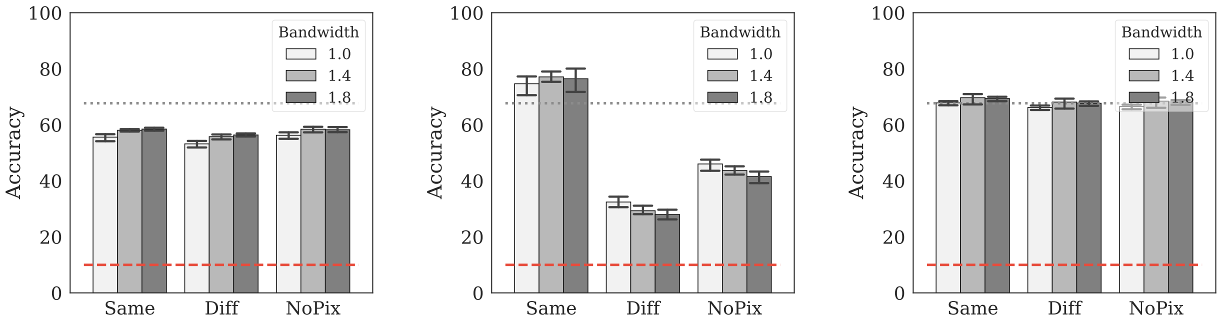
To test the hypothesis that the reliance of the network on the noise masks was due to high spatial frequency information contained in these images, we systematically varied the two key parameters of the Gabor filters most pertinent to this idea: λ and b. The wavelength of the sinusoidal component, λ was varied in the range $[3..8]$ pixels/cycle while the bandwidth of the Gaussian component, b, was chosen from $\{1.0, 1.4, 1.8\}$ octaves in accordance with measurements from macaque visual cortex (Petkov & Kruizinga, 1997), with σ automatically calculated for each combination of parameters according to Equation 3. For each experimental condition, five realisations were run with different randomised initial conditions.

An illustrative example of the familiar performance bar chart for $\lambda = 5$ and $b = \{1, 1.4, 1.8\}$ is given in Figure 3 (for direct comparison to earlier results in Figure 2), showing performance to be largely insensitive to variations in b for this range.

It is evident from the largely flat performance profiles across the test conditions in Figure 3 that the network is no longer reliant upon the noise masks for correctly classifying the `CIFAR-10` images (albeit with some lingering difficulty with additive noise). In all cases, performance on the ‘Diff’ condition is greater than zero and performance on the ‘NoPix’ condition is greater than chance (10%). This trend can also be seen to hold across a biologically relevant range of variation in bandwidth.

Discussion & Conclusions

In a series of simulations we found that standard CNNs do not show an innate shape-bias when a stronger non-shape signal is present within the training dataset. Instead, the models learnt to categorise objects on the basis of non-shape features strongly correlated with the output class, even when the features were as small as a single pixel. Of course, we engineered our dataset to contain diagnostic non-shape features, but it is well-known that datasets contain various biases due to the different conditions and motivations for their construction (Torralba & Efros, 2011). So biases like the one we engineered



(a) Salt-and-pepper noise

(b) Additive noise

(c) Single diagnostic pixel

Figure 3: Accuracy on test images under the three types of noise-like masks. In all cases, the wavelength of the sinusoid was fixed, $\lambda = 5$. The shading of the bars indicates the filter bandwidth. The dotted (grey) line indicates performance on the standard CIFAR-10 images, the dashed (red) line indicates chance performance and error bars show the 95% confidence intervals.

may well be present in these datasets. 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, Yosinski, & Clune, 2015) and being overly sensitive to colour, noise or even single pixels in images (Su, Vargas, & Kouichi, 2017).

One reason why humans may be less vulnerable to relying on non-shape statistical regularities may be that the architecture of our visual system imposes innate constraints on the type of information used for classification. One such constraint is that primary visual cortex is organized into hyper-columns composed of simple cells that identify edges of various orientations. Our hand-coding of the Gabor filters is well motivated by the fact that neurophysiology has shown that the ordered arrangement of simple cells are not learnt in response to the statistical structure of the world, but are innately specified and emerge in visual cortex in animals who have no visual experience (Chapman & Stryker, 1993; Wiesel & Hubel, 1974).

In line with this view, we observed that vulnerabilities to these non-shape features can be ameliorated when we replaced end-to-end learning by learning on a bank of Gabor filters that are the standard model of simple cells (Jones & Palmer, 1987). Moreover, we chose the parameters of these Gabor filters based on neurophysiological data and found that these results hold, not just for particular parameter values but for an entire range. In conclusion, the crucial element does not seem to be learning the correct values of these parameters but having the correct form of filters.

References

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

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

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

Chapman, B., & Stryker, M. P. (1993). Development of orientation selectivity in ferret visual cortex and effects of deprivation. *Journal of Neuroscience*, *13*(12), 5251–5262.

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*.

Jones, J. P., & Palmer, L. A. (1987). An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, *58*(6).

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

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).

Petkov, N., & Kruizinga, P. (1997). Computational models of visual neurons specialised in the detection of periodic and aperiodic oriented visual stimuli: bar and grating cells. *Biological Cybernetics*, *76*(2), 83–96. doi: 10.1007/s004220050323

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*.

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

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

Wiesel, T. N., & Hubel, D. H. (1974). Ordered arrangement of orientation columns in monkeys lacking visual experience. *Journal of comparative neurology*, *158*(3), 307–318.

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.