Hiding a plane with a pixel: examining shape-bias in CNNs and the benefit of building in biological constraints

Gaurav Malhotra*, Benjamin Evans, Jeffrey Bowers

School of Psychological Science
University of Bristol
Bristol, BS8 1TU, UK

Abstract

When deep convolutional neural networks (CNNs) are trained "end-to-end" on raw data, some of the feature detectors they develop in early layers resemble the representations found in early visual cortex. This result has been used to draw parallels between deep learning systems and human visual perception. In this study, we show that when CNNs are trained end-to-end they learn to classify images based on whatever feature is predictive of a category within the dataset. This can lead to bizarre results where CNNs learn idiosyncratic features such as high-frequency noise-like masks. In the extreme case, our results demonstrate image categorisation on the basis of a single pixel. Such features are extremely unlikely to play any role in human object recognition, where experiments have repeatedly shown a strong preference for shape. Through a series of empirical studies with standard high-performance CNNs, we show that these networks do not develop a *shape-bias* merely through regularisation methods or more ecologically plausible training regimes. These results raise doubts over the assumption that simply learning end-to-end in standard CNNs leads to the emergence of similar representations to the human visual system. In the second part of the paper, we show that CNNs are less reliant on these idiosyncratic features when we forgo end-to-end learning and introduce hard-wired Gabor filters designed to mimic early visual processing in V1.

*Corresponding author
Email address: gaurav.malhotra@bristol.ac.uk (Gaurav Malhotra)

Preprint submitted to Vision Research *March 5, 2020*

1. Introduction

Image recognition in traditional computer vision models proceeds in two stages. In the first stage, images are mapped onto a set of hand-crafted features. In the second stage, these features are mapped onto output categories. Consequently, the success of the image recognition algorithm strongly depends on identifying an appropriate set of features. Part of the appeal of deep learning models, such as convolutional neural networks (CNNs), has been in removing the first stage and letting the algorithm itself discover useful features. In this setting, image recognition proceeds "end-to-end", with raw pixels at one end and output categories at the other end. This method has been highly successful and indeed outperforms most traditional models of image recognition.

What is even more interesting from a neuroscience perspective is that when one trains these networks on images, the features learnt in the early layers seem to resemble features such as Gabor filters (Yosinski et al., 2014) which effectively extract edges from objects and are also found in early visual cortex (Petkov & Kruizinga, 1997). This gives credence to the belief that deep convolutional networks are capturing some fundamental aspects of human visual perception (Rajalingham et al., 2018). However, a closer inspection reveals that, in addition to features that resemble those found in the visual cortex, early layers also contain a number of features unlike those observed in the cortex (see Figure 1).

In this study, we examined (a) whether standard CNNs indeed perform image recognition in a fundamentally similar manner to human visual perception, and (b) whether image recognition performed by CNNs can be brought closer to humans by replacing end-to-end learning with learning that starts from a feature space similar to that found in human visual cortex.

We investigate these questions by focusing on a fundamental property of human image recognition, namely, it is largely a function of analyzing shape (Biederman, 1987; Hummel, 2013). A wealth of data from psychological experiments show that the shape of an object plays a privileged role in object recognition

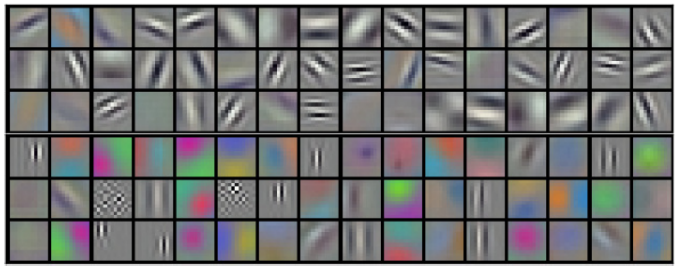


Figure 1: Example of 96 convolutional kernels learnt by the first convolutional layer from AlexNet, a high-performance convolutional neural network. Each kernel is of size $11 \times 11 \times 3$. Learning is performed on images of size $224 \times 224 \times 3$. Note that, in addition to filters that resemble Gabor filters, a number of other feature detectors also emerge from end-to-end learning. Figure taken from Krizhevsky et al. (2012).

compared to other diagnostic features such as size, colour, luminance or texture (Mapelli & Behrmann, 1997; Biederman & Ju, 1988). Experiments have also revealed that shape is extracted early (Leek et al., 2016) and automatically (Baker & Kellman, 2018) during human visual perception. Furthermore, experiments from developmental psychology show that this privileged status of shape starts early in life and becomes stronger with age (Landau et al., 1988). Note, these studies not only show that the visual system extracts shape during recognition, they also show that the human visual system prefers shape over other diagnostic features (e.g. color, texture, etc.) while performing recognition. In other words, it has a *shape-bias*.

What is still unsettled, however, is 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 et al. (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. On the one hand, 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. On the other hand, a shape-bias might be 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 et al., 2013). Indeed, 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. On this second view, CNNs have an *innate* shape-bias.

Some recent studies have suggested that CNNs rely on learning shape in order to categorise objects (Kubilius et al., 2016; Jozwik et al., 2017) and that a shape-bias is learned as a consequence of training on a particular dataset. For example, Ritter et al. (2017) observed that when an `Inception` model (Szegedy et al., 2016) was pre-trained on `ImageNet`, the representations in hidden layers were more similar for two (novel) objects that overlapped in shape than for two objects that overlapped in colour. Critically, they attributed this shape-bias to the statistical properties of the dataset itself. In another recent study, Feinman & Lake (2018) show that standard CNNs can show a shape-bias, just like children studied by Landau et al. (1988), when they are trained in an end-to-end manner on a controlled dataset, constructed in such a manner that the category name correlated with shape more than colour or texture.

Other studies have argued against a learned shape-bias when networks are trained on standard datasets such as `ImageNet`. For example, Geirhos et al. (2018) and Baker et al. (2018) manipulated the texture and shape of images independently and showed that standard CNNs trained end-to-end on `ImageNet` are biased towards using local features, such as texture, compared to the object's shape. However, in line with the results of Feinman & Lake (2018), Geirhos et al. (2018) also showed that CNNs develop a shape-bias when the training set is manipulated to make shape the most diagnostic feature.

As far as we are aware, however, no one in the machine learning community has argued that CNNs have (or should have) an innate shape-bias. That is, a bias

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

Figure 2: Images taken from `CIFAR-10` dataset and scaled up to 224x224 pixels. (a) Salt-and-pepper noise-like mask; (b) Uniform additive noise mask; (c) A single diagnostic pixel is inserted in the image (a dotted red circle is inserted here to illustrate the location of the pixel).

to identify objects on the basis of their shape when both shape and non-shape features are each highly diagnostic of category membership. In order to tease apart whether any shape-bias is learned or innate in standard CNNs, we modified the standard `CIFAR-10` dataset to simultaneously contain shape and non-shape features. We tried several types of non-shape features, such as noise-like 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 2). We carried out a sequence of experiments, where we manipulated the architecture of CNNs used, the learning algorithm, regularisation method and the type of learning regime used to train the CNNs. Our hypothesis was that, if CNNs have an innate shape-bias due to their architectural properties, they would rely more on shape compared to non-shape features. Furthermore, in order to determine whether we could induce an innate shape-bias we modified the architecture of our CNNs to include more constraints from the human visual system.

To preview our results, we found that standard CNNs trained on this modified `CIFAR-10` dataset learnt to depend on non-shape features that are diagnostic of object categories and often failed to learn (or retain) anything about shape under these conditions. These results suggest that 'vanilla' CNNs do not have an innate shape-bias even though they share some architectural properties of biological visual systems and discover some features resembling those found in their early layers. (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 innate biological constraints 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 although this change comes at a cost to the network's overall performance, it 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 relevant parameters for the Gabor filters, showing that a network using a bank of Gabor filters was, in general, less likely to rely upon idiosyncratic features present within the dataset. We argue that including Gabor filters as the first convolutional layer of CNNs makes them more similar to biological visual systems, becoming less sensitive to non-spatial details of images that can be predictive of object category.

2. Methods

We modified the `CIFAR-10` dataset (which contains 10 classes with $6,000$ 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 independent 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 each pixel value is then independently modified by adding a value sampled from the Uniform distribution. The width of this distribution was $[\mu - w, \mu + w]$ to this image, where $\mu \in [-50, 50]$ was the mean that depended on the category of the image and $2w$ was the width of the Uniform distribution which was set to 8 for images of all categories. The *single pixel* mask was created by replacing one pixel in each 224×224 image with a new pixel value. The location and colour of this pixel was category correlated: the location of the pixel, (x, y), was sampled from a 2D Gaussian distribution with a mean that depended on the category and a standard deviation that remained constant across categories. Similarly, each of the red, green and blue values of the pixel colour, (c_r, c_g, c_b), were drawn from a Gaussian distribution with a mean that depended on the category and a variance that remained constant across categories. If any value in a sampled set of (x, y, c_r, c_g, c_b) values fell out of their respective range, that value was re-sampled. Some example images are shown in Figure A.9.

We used a method similar to Geirhos et al. (2017) to preprocess images from the `CIFAR-10` dataset where each 32×32 pixel image was upscaled to 224×224 pixels using Lanczos resampling. For the single-pixel mask, we used 3-channel RGB images (or greyscale for Gabor-filter model) 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 $[0, 255]$.

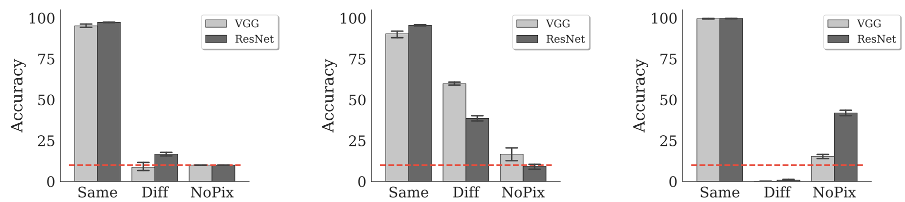
We trained the model on these modified sets of images and tested it under three conditions. During the 'Same' condition, the test set was modified in exactly the same manner as the training images, i.e., masks for each category were generated by using the same parameters as those used during training. In contrast, during the 'Diff' condition, the parameters of the noise masks for each category were swapped with another category. 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, where the mask was entirely absent during testing, 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 the 'Same' and 'NoPix' conditions should quantify the relative extent to which the network relied on shape and non-shape features.

Simulations were carried out using either a 16-layer `VGG` network (Simonyan & Zisserman, 2014) or a 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. When the networks were pre-trained, we replaced the fully-connected layer(s) of the `VGG`/`Resnet` pre-trained model such that the last fully-connected layer had 10 output units (corresponding to the 10 categories of CIFAR-10). 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 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 (or $1e-4$ throughout with the Gabor-filter model). 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.

3. Results

3.1. Experiments 1–3

We conducted three experiments, one for each type of noise mask described above. The results are shown in Figure 3. During all three experiments, we observed that both networks classify images with a nearly perfect accuracy during the 'Same' noise condition. When noise masks were swapped ('Diff' condition), this accuracy dropped; when the masks were completely removed



(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 shown in Figure 2. Training images contain (a) salt-and-pepper noise, or (b) additive uniform noise, or (c) just one diagnostic pixel. Each experiment shows test performance under three conditions – 'Same': the noise-like mask has the same properties for testing and training images of each category; 'Diff': the properties of the mask during testing are swapped with another category from training; 'NoPix': No mask is applied. The dashed (red) line indicates chance performance and error bars show 95% confidence intervals. Light and dark gray bars show accuracies on `VGG-16` and `ResNet-101` respectively.

('NoPix' condition), the categorisation accuracy was nearly at chance. For both the *salt-and-pepper* and *single pixel* experiments, performance in the 'Diff' condition was either at or below chance. Recall that the 'Diff' condition swaps the masks between categories. Therefore, a below chance performance reflects that the network is entirely relying on the mask to make category predictions, systematically predicting a different category to the original image category in `CIFAR-10`. These results are confirmed by the 'NoPix' condition: when the mask information is removed, the network struggles to make a prediction based on information within an image, with performance dropping to near-chance levels.

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 using some other features of the image beside the noise-like mask. However, even in this case, there was a significant drop in performance compared to the 'Same' condition.

The *additive noise* experiment showed an intriguing behaviour: when the noise-like mask was completely removed ('NoPix' condition) the model performed *worse* than when the images contained a mask from a different category ('Diff'

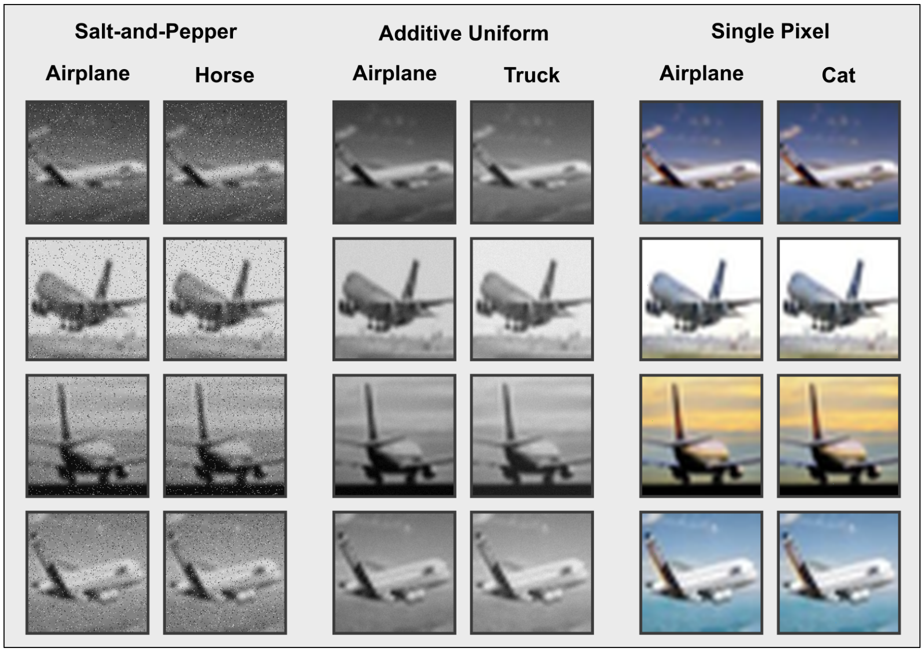


Figure 4: Four images from the `CIFAR-10` test-set that have been modified by adding a noise-like mask. Each image contains a different mask. However, all images in a column contain a mask with shared statistical properties. For example, all images in the first column contain salt-and-pepper masks drawn from the same distribution (see Methods) while images in the second column draw masks from a different distribution. Consequently the network classifies each image in the first column as an 'Airplane', while it classifies each image in the second column as a 'Horse'. Similarly, the two columns in the middle contain images with additive uniform noise masks drawn from two different distributions while the two columns on the right contain images with a single predictive pixel (nearly invisible to the naked eye).

condition). In other words, removing the mask made 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 appears to rely on the presence of the noise-like mask to make an inference.

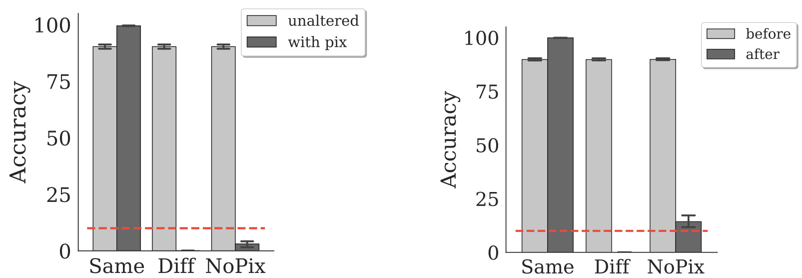
Furthermore, we obtained the same pattern of results irrespective of the type of regularisation used (we tried several well-known regularisation methods including *Batch Normalization*, *Weight Decay* and *Dropout*). These results clearly indicate that the model learnt to rely on features of the noise-like 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 preferred to classify based on this feature over other shape-related features present in each image. Figure 4 shows four example images that have been modified in the manner described above and are classified differently based on the mask superimposed on these images. Note that it is difficult for humans to distinguish the various salt-and-pepper and uniform noise masks that the CNNs use to make these image classifications.

The above results were obtained for networks that were pre-trained on `ImageNet`. Since these images are in the format 224×224 pixels, we upscaled all CIFAR-10 images to this size. A very similar pattern of results is obtained if the images are left unscaled (though in this case the networks had to be trained from scratch on the modified dataset). In fact, the upscaled images constitute a much stronger test as the network needs to learn a single predictive pixel amongst 50,176 pixels (224×224) instead of amongst 1,024 pixels (32×32). Results for conducting the above experiments on unscaled images of size 32×32 are shown in Appendix Appendix B.

3.2. Experiments 4 & 5

One possible reason why humans prefer to rely on shape-related features to categorise objects while standard CNNs do not, is that humans are guided by past experience when performing new categorisation tasks. So when a human sees an object with superimposed noise, they rely on shape-based information, paying less attention to non-shape related features such as the masks in the above images. We conducted two further experiments to test whether networks similarly generalise from concurrent and past experience. Both these experiments were conducted on the *single pixel* mask as this seems to be the most striking finding and we get the clearest pattern of results with this case.

In Experiment 4, we divided the training set into two subsets. The first subset ('with pix') contained three randomly chosen categories from `CIFAR-10` and, as described above, contained a category-correlated pixel in all images of these categories. The second subset ('unaltered') contained the remaining seven categories from `CIFAR-10` which were 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 concurrently. 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 categorise images in the 'with pix' subset. All other details of the experiment remain the same as in Experiment 1.



(a) Experiment 4 (b) Experiment 5

Figure 5: Accuracy 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.

The results from this experiment are shown in Figure 5a. 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 5 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, 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, introducing 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 into each image.

We observed that, 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 90% during the 'before' phase to 0% during the 'after' phase in the 'Diff' condition (Figure 5b). 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) but it remains to be seen whether this can overcome some of these problems.

3.3. Experiment 6

It could be argued that the diagnostic non-shape features that we inserted provide a very strong diagnostic signal. For example, in the single-pixel condition, each image contains the pixel in roughly the same location. Since it is unclear to what extent large datasets such as `ImageNet` or `CIFAR-10` contain such idiosyncratic (but reliable) features, we decided to examine how the behaviour of the network changes when only a subset of images contain a diagnostic non-shape feature. We again restricted this experiment to the case of a single diagnostic pixel as this was the most striking finding in the above experiments. We also restricted testing to the `VGG-16` network, as very similar results were found for

`VGG-16` and `ResNet-101` above. The location and colour of this pixel were fixed across all images of a category, but we introduced stochasticity in the presence of this pixel within a training image. Figure 6 shows the change in accuracy for the 'NoPix' condition with a decrease in the probability with which a pixel is present in a training image. We specifically focus on the 'NoPix' condition as the accuracy on this condition is inversely correlated with how much the network relies on this pixel to predict the output category.

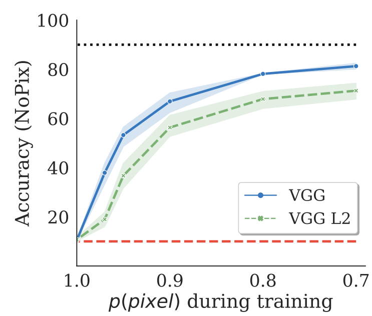


Figure 6: Accuracy of the model on images containing no mask, as a function of the fraction of training images containing a diagnostic pixel. The solid (blue) and dashed (green) lines plot this relation for a network trained without and with weight-decay, respectively. The dashed (red) line at the bottom shows chance performance. The dotted (black) line at the top shows performance of a network trained on images without any noise mask.

It is clear from this figure that the network continues to rely on this informative pixel, even when it is not present in all the images. For example, the network's performance drops from around 90% when it is trained on the unmodified `CIFAR-10` dataset to around 70% when it is trained on a modified dataset that contained the pixel in 90% of the images. As we decreased the proportion of images containing the pixel, the performance increased, but still did not achieve the performance of the unmodified `CIFAR-10` when only 70% of images contained such a pixel. The increase in performance with decrease in the proportion of images containing the diagnostic pixel is consistent with the hypothesis that the learning algorithm selects the feature based on the predictive power of the feature; as the single pixel becomes less predictive, the network starts relying on other features to choose the output category. Lastly, we also observed that L2 regularisation made the performance of the network worse on the original images when a diagnostic pixel was inserted on a fraction of the images. While L2 regularisation should help the network learn a more general solution, in this case it led to the opposite effect.

4. 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-like masks that are diagnostic of output categories of the 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.

There is good reason to believe that filtering an image through a bank of Gabor filters will reduce high-frequency noise present within these images. Convolving an image with a Gabor kernel filters the image based on the shape of the kernel. Thus, much like simple cells, Gabor kernels act like oriented edge or bar detectors for particular spatial frequencies, filtering noisy information outside their bandwidth.

4.1. 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 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, θ, four phases, ψ, and two aspect ratios, γ, (defining the ellipticity of the filter) while the wavelength, λ, and bandwidth, b, were systematically varied. The corresponding values are given in Table 1. Additionally, the kernels were set to be 31×31 pixels, with an odd number chosen in order to centre the kernels on each image pixel. We chose a fairly large size for the Gabor filters (note this is distinct from the spatial scale, σ) to allow the Gaussian envelope to decay to near-zero at the edges and thus avoid any truncation artefacts when computing the convolutions. The filters were plotted to visually confirm that they had largely decayed to zero near the borders of the frame, avoiding boundary effects (see Figure C.11).

Table 1: Parameters used for constructing sets of Gabor filters.

Parameter	**Symbol**	**Values**
Orientation	θ	$\{0, \frac{\pi}{8}, \frac{\pi}{4}, \frac{3\pi}{8}, \frac{\pi}{2}, \frac{5\pi}{8}, \frac{3\pi}{4}, \frac{7\pi}{8}\}$ *radians*
Phase shift	ψ	$\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$ *radians*
Aspect ratio	γ	$\{0.5, 1\}$
Wavelength	λ	varied: $3, 4, 5, 6, 7, 8$ *pixels/cycle*
Spatial bandwidth	b	varied: $1, 1.4, 1.8$ *octaves*

As with the previous experiments, CIFAR-10 images were manipulated by adding one of the following types of noise: Salt and Pepper, Additive or Single pixel but remained in their original size of 32×32 pixels. All images were converted to greyscale and fed into the modified network under the same training and test conditions described previously.

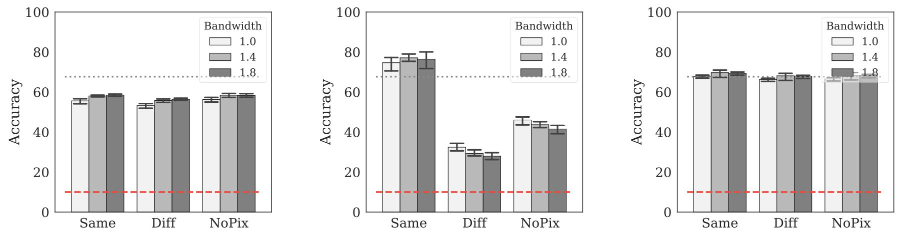
4.2. Results

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 is shown for direct comparison to earlier results in Figure 7 for $\lambda = 5$ and $b = \{1, 1.4, 1.8\}$. The trends in network performance for each test condition are plotted against λ in Figure 8. The performance was found to be largely insensitive to variations in b for this range but the full trends are included in Figures C.12 and C.13.

It is evident from the largely flat performance profiles across the test conditions in Figure 7 that the network is no longer reliant upon the noise-like 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.

Figure 8 shows that although performance gradually declines with increasing λ (as the filters represent decreasing spatial frequency information), the effect of



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

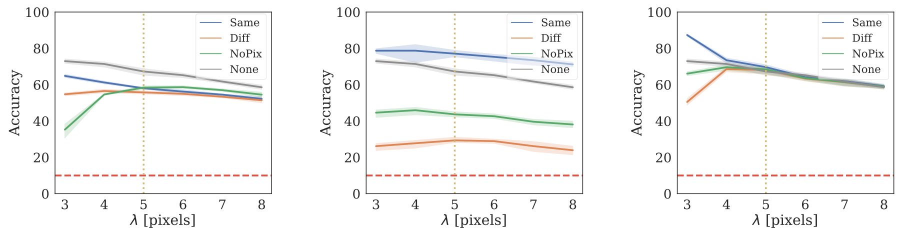
Figure 7: Accuracy on test images under the three types of noise-like masks. The shading of the bars indicates the three filter bandwidths tested. The dotted (grey) line indicates performance on the standard `CIFAR-10` images, the dashed (red) line indicates chance performance and error bars show the 95% confidence intervals. In all cases, the wavelength of the sinusoid component was fixed at $\lambda = 5$.

the noise-like masks has been eliminated by 4 or 5 pixels/cycle (demonstrated by the convergence of performance curves in Figures 8a and 8c) and is robust throughout a wide range of the parameter space. The additive noise condition still affects the network performance but to a lesser extent than the CNNs that were trained end-to-end, with performance well above chance throughout the parameter range under all conditions.

5. Discussion & Conclusions

In a series of simulations we found that standard CNNs do not show a shape-bias when trained on images that include both shape and non-shape features diagnostic of object category. That is, standard CNNs do not have an innate shape-bias. Instead, the models learnt to categorise objects on the basis of non-shape features that were 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 popular datasets contain various biases due to the different conditions and motivations for their construction (Torralba & Efros, 2011). As such, biases like the ones we engineered may well be present in these



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

Figure 8: Accuracy on test images under the three types of noise-like masks plotted against varying wavelength, λ. In addition to the standard noise conditions, 'None' indicates the original images (no noise mask) were used for training and testing to provide a performance baseline. The shaded bands around each line represent the 95% confidence intervals, the horizontal (red) dashed line represents chance performance and the vertical (yellow) dotted line represents the point in parameter space corresponding to Figure 7. In all cases, the median bandwidth was used, $b = 1.4$ octaves, with very similar trends exhibited at the other bandwidths tested (see Figure C.12).

datasets, which standard networks may be picking up on. This hypothesis is in line with a recent study conducted by Jo & Bengio (2017) who observed that standard CNNs have a tendency to learn the surface statistical properties of images as opposed to high-level abstractions. Indeed, this adds to a body of evidence showing that standard CNNs trained on ImageNet categorize images on the basis of texture rather than shape (Geirhos et al., 2018).

This tendency for learning surface statistical properties may help explain the vulnerability of CNNs to adversarial attacks. It is well known that CNNs show several 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). Ilyas et al. (2019) have recently argued that many adversarial attacks can be attributed to learning "*non-robust features*" present within datasets – that is, features that are predictive of an image category in a dataset but highly sensitive to small perturbations of the image and hence incomprehensible to human beings. In contrast, a high-level feature, such as shape, is robust to small deformations and the human preference for relying on shape makes them less vulnerable to small, high-frequency changes within images.

To be clear, our results do *not* show that CNNs cannot rely on shape if it is the only or primary diagnostic feature. Indeed, if the most diagnostic feature in our dataset was shape (rather than the noise-like masks), then we expect CNNs would learn to rely on shape, consistent with the work by Feinman & Lake (2018). However the hypothesis we set out to test is not whether networks can learn to identify objects on the basis of shape, but rather, whether CNNs have an innate shape-bias – that is, whether or not CNNs *prefer* to rely on shape in the presence of other diagnostic features. Our results show that this is not the case.

We also found that pre-processing images through a bank of Gabor filters and mapping them to a more biologically plausible feature space can make CNNs less sensitive to some types of non-shape diagnostic signals. Of course, we do not want to suggest that preprocessing images in this manner ensures that CNNs rely on shape to perform classification, or start exhibiting a shape-bias. Clearly, if one designed a predictive feature with a spatial extent that can pass through the bank of Gabor filters, the network would end up using it to perform categorisation, instead of relying on the object's shape. What we show here is that if one replaces end-to-end learning with learning that takes as its input a biologically plausible feature space, namely a bank of Gabor filters, it makes the network more robust to a range of idiosyncratic non-shape features. We chose the parameters of these Gabor filters based on neurophysiological data and found that these results hold, not just for particular values of parameters but for an entire range of parameters. So the crucial element does not seem to be learning the correct values of these parameters but having the correct form of filters.

As noted, this robustness to perturbations across the three test manipulations comes at the cost of a decrease in overall performance, e.g. dropping from the standard result of around 95% accuracy (with the unmodified `CIFAR-10` dataset) to around 70% when Gabor filters are included in `VGG16` (see 'None' for $\lambda \geq 4$

in Figure 8). This decrease in performance may be partly due to discarded colour information and the restriction to individual wavelengths and bandwidths (rather than a full range) for the sake of systematic evaluation. However, the Gabor kernels themselves filter out an additional source of information, namely *unstructured*, spatially high-frequency features, further lowering performance. From a machine learning perspective the reduction in accuracy is a problem. However, from a psychological perspective the resultant flat performance profile gained by these convolutional constraints suggests that the excellent performance of existing CNNs relies on extracting such high-frequency features that humans ignore (or are insensitive to). Accordingly, we argue that this accuracy drop demonstrates the fragility and biological implausibility of solutions found by end-to-end trained models, rather than an inadequacy of adding the Gabor filters as a front-end to CNNs.

In this study, we imposed a biological constraint by replacing end-to-end learning with a biologically motivated feature space. Another possible approach is to preserve end-to-end learning while changing the architecture of the CNN in such a way that a similar feature space of Gabor filters is learned. Recently, Lindsey et al. (2019) have shown that imposing such architectural constraints, such as a retinal "bottleneck", can lead to the emergence of antagonistic centre-surround fields found in retinal ganglion cells, followed by Gabor-like receptive fields. It remains to be seen whether such a constraint could be used to overcome vulnerabilities of standard CNNs to non-shape features present within datasets. However, even if this approach proves to be successful, it is important to note that neurophysiological research shows that oriented receptive fields in V1 are innate rather than learnt through experience (Chapman & Stryker, 1993; Wiesel & Hubel, 1974).

Rather than learning Gabor filters end-to-end in response to image datasets, from a biological perspective, the more appropriate question might be to explain how these filters develop in response to evolutionary pressures. From an engineering perspective the challenge now is to advance this new direction, closing the performance gap while retaining the robustness.

References

Baker, N., & Kellman, P. J. (2018). Abstract shape representation in human visual perception. *Journal of Experimental Psychology: General*, *147*, 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*, e1006613.

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

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

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

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

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

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*, (pp. 32–46).

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, .

Jo, J., & Bengio, Y. (2017). Measuring the tendency of cnns to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*, .

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*, 1233–1258. URL: http://jn.physiology.org/cgi/content/abstract/58/6/1233.

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., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, *114*, 3521–3526.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

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*, e1004896.

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

Leek, E. C., Roberts, M., Oliver, Z. J., Cristino, F., & Pegna, A. J. (2016). Early differential sensitivity of evoked-potentials to local and global shape during the perception of three-dimensional objects. *Neuropsychologia*, *89*, 495–509.

Lindsey, J., Ocko, S. A., Ganguli, S., & Deny, S. (2019). A Unified Theory of Early Visual Representations from Retina to Cortex through Anatomically Constrained Deep CNNs. *arXiv e-prints*, (p. arXiv:1901.00945). arXiv:1901.00945.

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

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

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*, 83–96. doi:10.1007/s004220050323.

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*, (pp. 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).

Wiesel, T. N., & Hubel, D. H. (1974). Ordered arrangement of orientation columns in monkeys lacking visual experience. *Journal of comparative neurology*, *158*, 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*, (pp. 263–284).

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems* (pp. 3320–3328).

Appendix A. Example Images

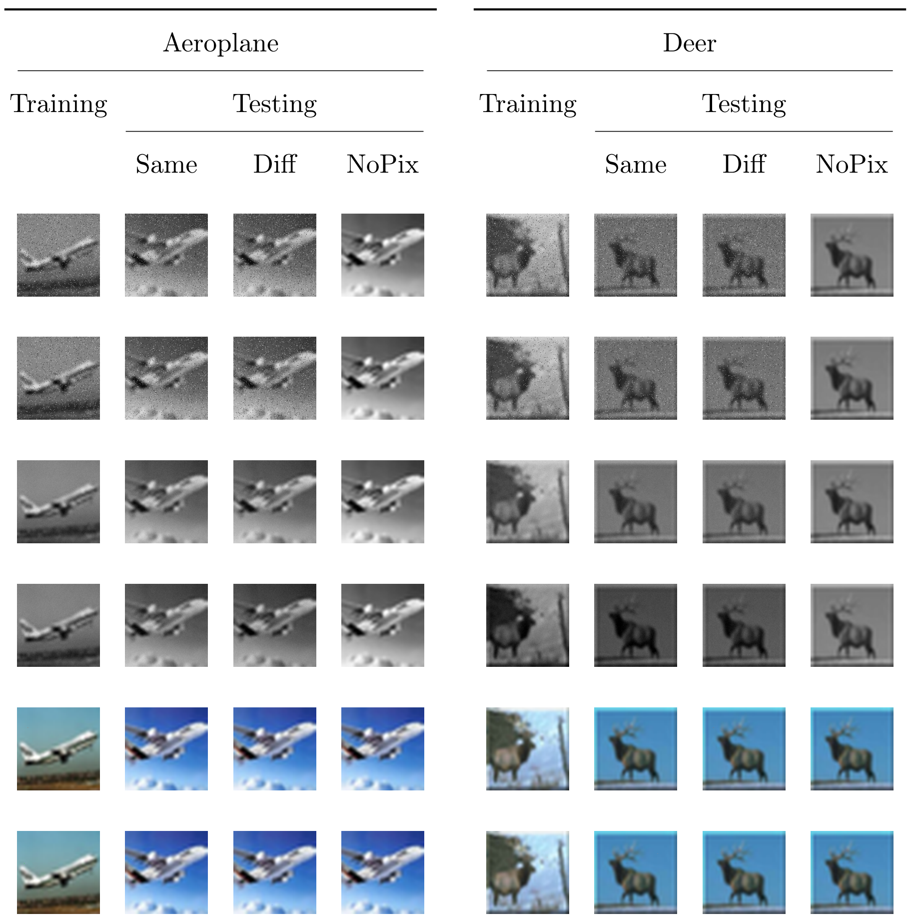
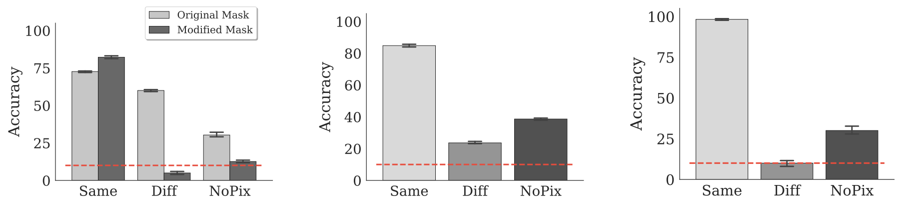


Figure A.9: Examples of images used for training and testing. The columns show the condition under which the image was used and the rows show the type of noise-like mask applied. These masks are, respectively, (row 1) salt-and-pepper noise with a fixed mask, (row 2) salt-and-pepper noise with a variable mask, (row 3) additive uniform noise with a fixed mask, (row 4) additive uniform noise with a variable mask, (row 5) single diagnostic pixel, with fixed location and colour and (row 6) single diagnostic pixel with variable location and colour.

Appendix B. Results for 32×32 images



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

Figure B.10: Accuracy of `VGG-16` convolutional neural network on test images of size 32×32 under (a) salt-and-pepper, (b) additive uniform, and (c) single pixel noise-like masks. The 'Same', 'Diff' and 'NoPix' conditions are the same as in Figure 3. we modified the `VGG-16` network from the original (Simonyan & Zisserman, 2014) network so that the first layer consists of three channels each of size 32×32. Instead of using a network that is pre-trained on ImageNet (which contains images in the 224×224 format), we trained the network from scratch on the modified datasets containing 32×32 images. Light gray bars in (a) show noise-like masks generated in the same manner as for the 224×224 images above. Since different categories differ in the rate of the salt-and-pepper noise (see Methods above), this method of generating noise leads to a much weaker diagnostic signal for 32×32 pixel images. When the strength of this diagnostic signal is increased, the same pattern of results reappears (dark gray bars). For (b) & (c) the amount and type of noise remains as used for the 224×224 pixels images and described in the Methods section above.

Appendix C. Gabor filters

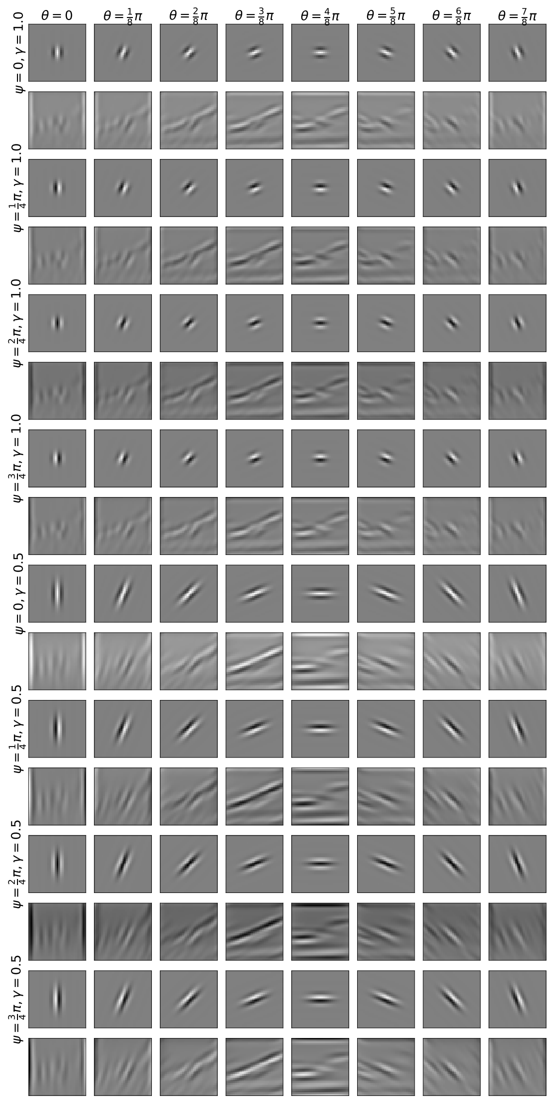


Figure C.11: Illustrative set of Gabor filters used in the first convolutional layer of the network with $\lambda = 5$ and $b = 1.4$. Orientation varies from 0 to $\frac{7}{8}\pi$ across each row, while down each column *psi* varies from 0 to $\frac{3}{4}\pi$ and γ varies from 1 to 0.5. The Gabor kernels are displayed on odd rows while the results of their convolution with an example image from the training set (Figure 2) are shown on even rows.

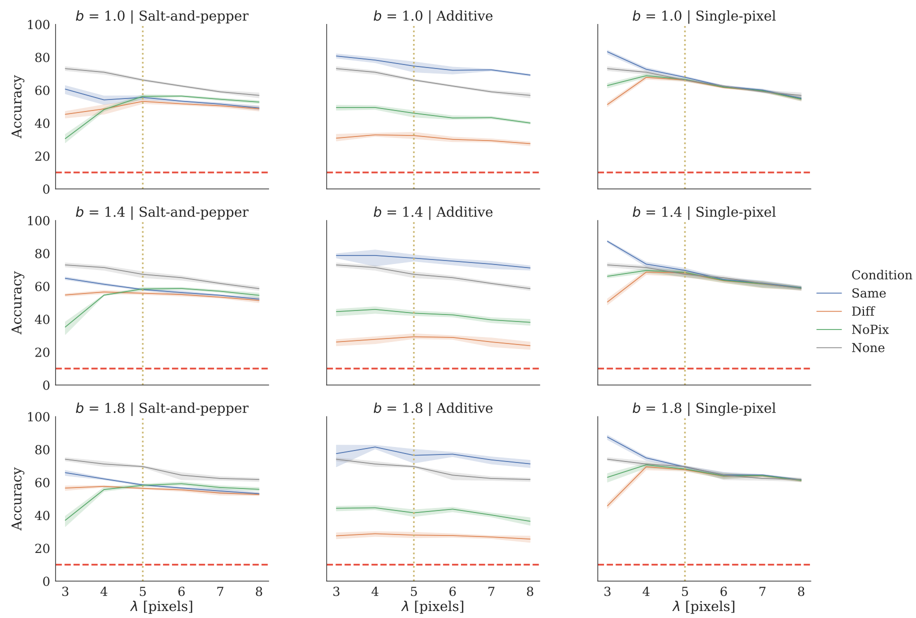


Figure C.12: Accuracy on test images under the three types of noise-like masks plotted against varying wavelength λ for each noise mask (columns) and three bandwidths, b (rows). In addition to the standard noise conditions, 'None' indicates the original images (no noise mask) were used for training and testing to provide a performance baseline. The shaded bands around each line represent the 95% confidence intervals, the horizontal (red) dashed line represents chance performance and the vertical (yellow) dotted line represents the point in parameter space corresponding to Figure 7. The middle row ($b = 1.4$) corresponds exactly to Figure 8 but is reproduced here for direct comparison to the performance curves obtained at other bandwidths.

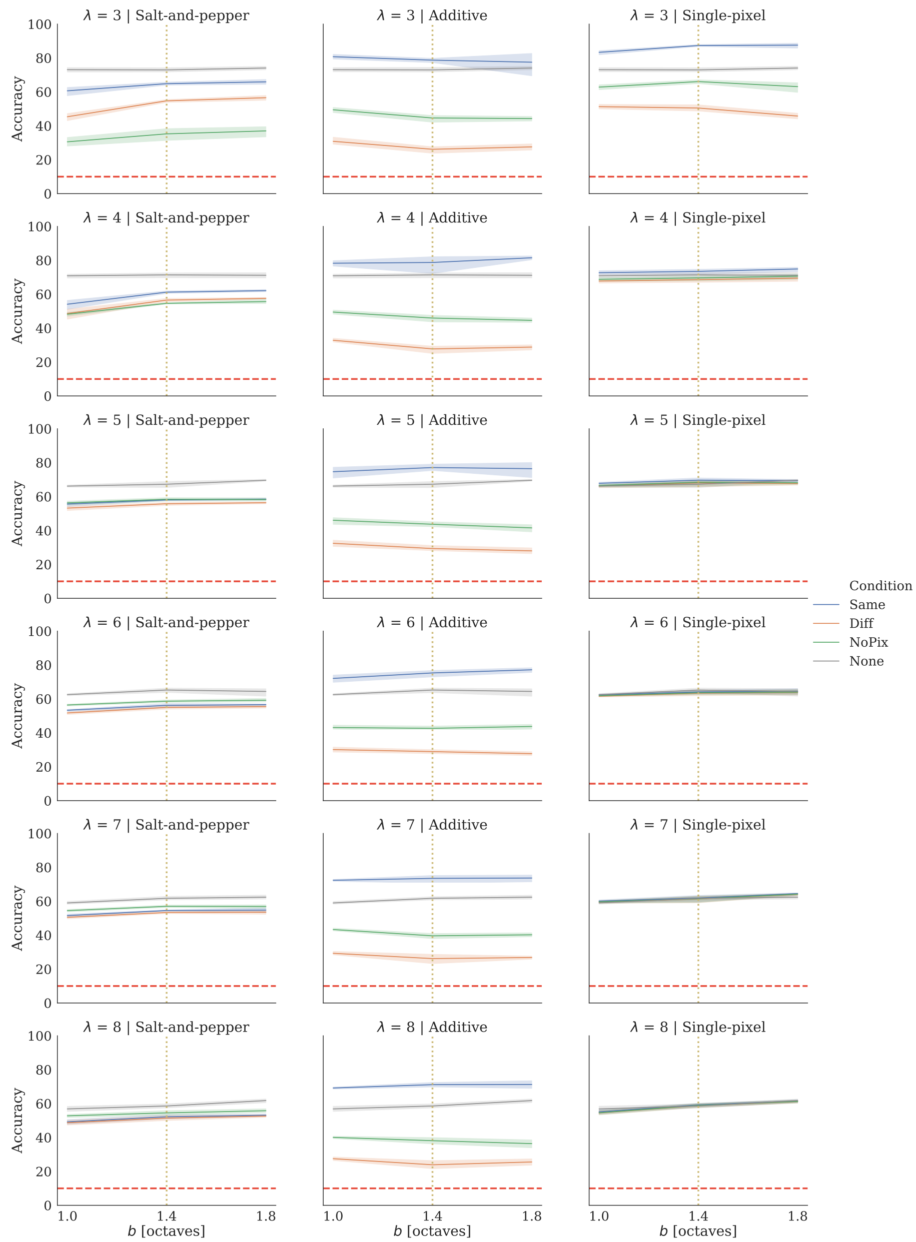


Figure C.13: Accuracy on test images under the three types of noise-like masks plotted against varying bandwidth, b for each mask (columns) and six wavelengths, λ (rows). In addition to the standard noise conditions, 'None' indicates the original images (no mask) were used for training and testing to provide a performance baseline. The shaded bands around each line represent the 95% confidence intervals, the horizontal (red) dashed line represents chance performance and the vertical (yellow) dotted line represents the point in parameter space corresponding to Figure 8 ($b = 1.4$). These are the same data used in Figure C.12 but transposed in order to explicitly see the performance trends with varying bandwidth.