The role of capacity constraints in Convolutional Neural Networks for learning random versus natural data

Christian Tsvetkov[a,*], Gaurav Malhotra[a], Benjamin D. Evans[a,b], Jeffrey S. Bowers[a]

[a]*School of Psychological Science, University of Bristol, 12a Priory Road, Bristol BS8 1TU, UK*
[b]*Department of Informatics, School of Engineering and Informatics, University of Sussex,Falmer, Brighton, BN1 9RH, UK.*

Abstract

Convolutional neural networks (CNNs) are often described as promising models of human vision, yet they show many differences from human abilities. We focus on a superhuman capacity of top-performing CNNs, namely, their ability to learn very large datasets of random patterns. We verify that human learning on such tasks is extremely limited, even with few stimuli. We argue that the performance difference is due to CNNs' overcapacity and introduce biologically inspired mechanisms to constrain it, while retaining the good test set generalisation to structured images as characteristic of CNNs. We investigate the efficacy of adding noise to hidden units' activations, restricting early convolutional layers with an information bottleneck, and using a bounded activation function. Internal noise was the most potent intervention and the only one which, by itself, could reduce random data performance in the tested models to chance levels. We also investigated whether networks with biologically inspired capacity constraints show

*Corresponding author

Email addresses: christian.tsvetkov@bristol.ac.uk (Christian Tsvetkov), gaurav.malhotra@bristol.ac.uk (Gaurav Malhotra), b.d.evans@sussex.ac.uk (Benjamin D. Evans), jeffrey.bowers@bristol.ac.uk (Jeffrey S. Bowers)

improved generalisation to out-of-distribution stimuli, however little benefit was observed. Our results suggest that constraining networks with biologically motivated mechanisms paves the way for closer correspondence between network and human performance, but the few manipulations we have tested are only a small step towards that goal.

Keywords: Deep Neural Networks, capacity, biological constraints, internal noise, information bottleneck

---

1. Introduction

Convolutional Neural Networks (CNNs) are extraordinarily successful in a diverse range of complex visual tasks such as object recognition, localisation, image segmentation and captioning (Ren et al., 2017; Chen et al., 2018; Tan & Le, 2021). These successes have inspired research exploring the extent to which CNNs and humans process visual information similarly. For example, previous research has shown that CNNs provide the best account of the representational geometry of object similarity in primate inferotemporal cortex (Khaligh-Razavi & Kriegeskorte, 2014). CNNs also outcompete alternatives in predicting neural firing patterns in various areas of the ventral visual stream by matching hidden layer activity to neural recordings (Yamins et al., 2014; Schrimpf et al., 2018). Others studies have focused on behavioural comparisons, showing some agreement between humans and CNNs in object similarity judgements and shape recognition (Kubilius et al., 2016; Peterson et al., 2017). Such results have generated enthusiasm for the claim that CNNs are the leading model of the visual system (Kubilius et al., 2019; Kriegeskorte, 2015).

However, for all these successes, there remain large gaps between human be-

haviour and CNN performance. Most often, this disconnect is observed in the form of mistakes of CNNs that are uncharacteristic of human performance. Perhaps most strikingly, CNNs confidently misclassify adversarial images that would not affect human performance (Dujmović et al., 2020). It is also the case that CNNs trained to recognise objects using one dataset often generalise poorly to images from different datasets even if they are created using the same data collection protocol (Recht et al., 2019), and are much more sensitive to image degradation (e.g., various forms of input noise) than human observers (Geirhos et al., 2018). These findings relate to the fact that CNNs often rely on qualitatively different features when classifying images compared with humans. For example, CNNs will often learn to classify objects on the basis of texture (Geirhos et al., 2019) rather than shape — a much more salient categorisation cue for humans (Biederman & Ju, 1988; Landau et al., 1998). Networks will even learn to classify objects on the basis of single pixels that are diagnostic of object category (Malhotra et al., 2020). This example highlights how CNNs often accomplish their objective using "shortcuts", learning features which do not match human experience and intuition (Geirhos et al., 2020).

We consider one instance of a second, often overlooked kind of divergence between humans and CNNs — cases in which network performance far surpasses human abilities. Though such high performance is desirable from an engineering perspective, such phenomena pose a challenge for the view that CNNs operate in a similar manner as to humans. One such striking example we focus on is that CNNs that perform well on object classification with datasets of labeled natural images can also learn to classify large datasets of images with random pixel intensities (which to a human observer look like images of TV-static), or natural image

datasets with arbitrary mappings between inputs and category labels (Zhang et al., 2017). Indeed, the models even learned to classify ~ 1 million such random inputs into 1,000 categories, conditions that parallel the `ImageNet` dataset (Deng et al., 2009). These random datasets required only a small multiple of additional presentations (< 5) compared to the unmodified `ImageNet` dataset (Zhang et al., 2017) in order to achieve similar levels of success, with the random pixel images being slightly easier than the shuffled labels. It seems unlikely that humans could do nearly so well under these conditions, especially in the case of the random pixel images.

In this article, we assess the disconnect between CNNs and human performance on these datasets and explore what conditions can make models perform more like humans. First, in order to confirm whether humans are indeed incapable of learning to categorise random datasets like those described, we designed a behavioural experiment that parallels the simulation study conducted by Zhang et al. (2017). Consistent with our expectations, participants performed near chance-level at classifying a small number of random pixel images. They were much better at learning the shuffled labels condition, but still required a substantial amount of training to map a small number of examples into two categories. Both results highlight how infeasible it is for humans to match CNNs in classifying ~ 1 million images into 1,000 categories.

Next, we investigated why human behaviour diverges so strikingly from CNN performance under these conditions. One possibility is that what is learnable to humans is influenced by our previous experience of the environment. On this view, by learning the regularities of the natural world we also restrict the kinds of information which can further be learned. In contrast, CNNs traditionally begin

training as a ‘blank-slate’, using randomly initialised weights, so their learning is unconstrained by any previous knowledge of the structure in the world. On this view, the disconnect in performance between the model and humans does not necessarily reflect any fundamental differences in the architecture of the CNNs and human visual cortex but instead suggests experience with the right data, or more appropriate training, would alleviate the mismatch. We tested whether CNNs pre-trained on large datasets of naturalistic images are impeded in subsequently learning random data. We found that pre-trained networks were in fact faster to learn the random datasets than untrained networks.

Another possibility is that the discrepancy in performance on these tasks is a consequence of fundamental differences in the processing capacity of CNNs and the human visual system. That is, unlike neurons in brains, units in CNNs compute with arbitrary precision, and CNNs are not limited by various anatomical constraints that characterise human vision (e.g., the optic nerve that serves as a severe bottleneck in communicating information from the retinas to rest of the brain). To explore this idea we looked at three biologically relevant mechanisms which can act as capacity constraints — internal noise, information bottlenecks and activation functions. We manipulated CNNs by including combinations of these constraints and examined their impact on classification accuracy for both natural and random images. Interestingly, we found that adding internal noise to hidden layer activations resulted in models that were still able to identify natural images but were poor at learning to classify random pixel images, whereas the other constraints had much less of an impact on performance.

Finally, we explored whether training models with biologically motivated capacity constraints improved their correspondence with human performance in

other aspects. In particular, we explored whether models with a limited capacity would show superior generalisation performance to test images that undergo distortions, such as adding salt-and-pepper or uniform noise. Our reasoning was that CNNs that learn to classify natural images without over-fitting the training data (e.g. through limited resources) may improve generalisation to out-of-distribution test images, as demonstrated in previous work (Evans et al., 2022). However, with minor exceptions, the constrained networks still failed to show improvements in generalisation accuracy compared to baseline models. Together, our findings highlight the importance of adding biological constraints to CNNs in order to match human limitations in classifying random data, but also show that additional innovations are required to improve the generalisation capacities of CNNs.

2. How well can humans learn random categories?

In previous work, Zhang et al. (2017) trained CNNs on two datasets. One dataset contained images with pixel intensities that were randomly sampled so that each image was akin to a randomly generated vector. These images were then randomly assigned to one of multiple categories. We call this the 'random pixel condition'. The other dataset consisted of naturalistic images from either the `CIFAR-10` or `ImageNet` dataset, which were randomly assigned to one of the categories at the start of the experiment. Accordingly, images that were members of the same class in the original dataset (e.g., two different images of a plane) could be assigned to different categories, whereas images that belonged to different categories in the original dataset (e.g., an image of a plane and an image of a horse) could be assigned to the same class. We call this condition the 'shuffled labels' condition. Zhang et al. (2017) showed that CNNs with different architectures can

learn to categorise images from both datasets, even when these datasets contained more than a million images.[1] The authors also found that CNNs learned to categorise the random pixel dataset more quickly (i.e., fewer training epochs were needed to converge to peak performance) than the shuffled label dataset.

In order to assess how humans learn random data we measured the ability of humans to perform a task similar to that which the networks in Zhang et al. (2017) had been trained on, but limited the training set to far fewer images on the assumption that humans cannot learn to classify over 1 million random images (due to human capacity limitations). This procedure involved categorising either random pixel images or images with shuffled labels. In both conditions the categories lacked any category-defining features.

2.1. Methods

We recruited 41 participants using the online platform Prolific (`https://www.prolific.co`) and randomly assigned each participant to either the random pixel condition or shuffled labels condition. Participants in both conditions completed a categorisation task in which they were asked to learn how to classify 20 images. In each condition, these 20 images were randomly selected from a larger dataset and assigned to one of two categories (10 images per category). Images used in the shuffled labels condition were taken from the Novel Objects and Unusual Names (NOUN) version 2 dataset (Horst & Hout, 2016) and randomly assigned to one of the two categories. This dataset comprises of photographs of inanimate objects

[1] Learning random datasets was characterised as memorisation (Zhang et al., 2017; Arpit et al., 2017), but we note that, while CNNs trained on these datasets cannot generalise to entirely new examples, they do generalise to perturbed exemplars of studied items. See Appendix B.

which have been judged as being novel using an inter-rater naming consensus (Horst & Hout, 2016) (see Figure 1, dataset II). The novelty of these images ensures that images are unfamiliar to humans. The objects depicted in the images did not share defining features with each other, making this a challenging task to learn (although CNNs can learn an even more challenging task where images with similar features were assigned to different categories).

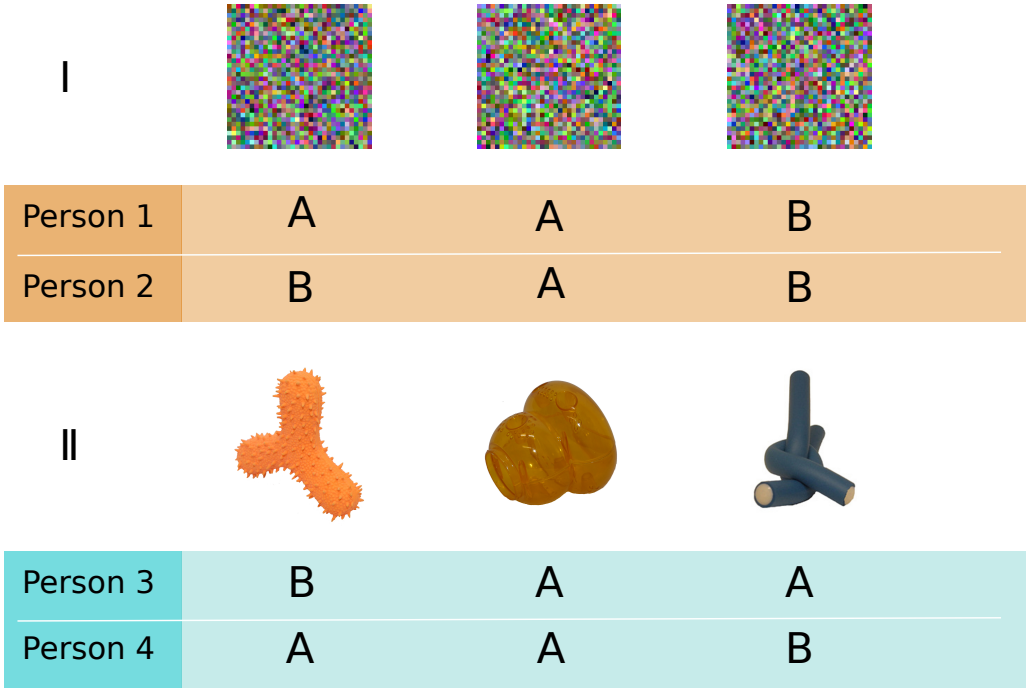


Figure 1: Examples of (I) random pixel images and (II) novel object images shown in the 'shuffled labels' condition in the behavioural experiment. Stimuli were randomly assigned to one of two categories, A or B. The stimulus selection and category assignment was different for every participant — e.g., the labels given to Participant 1 were not the same as those given to participant 2.

Images in the Random pixel condition were generated by drawing samples from Gaussian distributions for each colour channel, with the mean and stan-

dard deviation of each matching the values derived from the `CIFAR-10` dataset (Krizhevsky et al., 2009). The generated images were sized 32 by 32 pixels to match the size of `CIFAR-10` images used in the simulations of Zhang et al. (2017). To overcome any confounding effects of visual acuity of participants, these images were then upscaled (without smoothing) to 224 by 224 pixels. Some examples of these images are shown in Figure 1, dataset I.

Participants were instructed to indicate the categories by pressing either '`x`' or '`z`' on their keyboard. On each trial, a fixation cross was presented on the screen for 250ms. Afterwards, each stimulus was displayed for 300ms. All stimuli were presented on a neutral grey background. Trials were separated by a presentation of an empty neutral grey screen for 500ms. Responses within the first 150ms of presentations were considered too quick and a warning message was displayed to participants to pay more attention to the stimulus before responding. These trials were not considered for analysis. Similarly, we took responses longer than 3000ms as evidence that participants were not attending and these trials were also excluded.

Participants first completed a 'training' phase, receiving feedback about their response for each trial. The training phase consisted of 300 trials for the shuffled labels condition and 500 trials for the random pixel images condition. The larger number of trials for the random pixel images condition was motivated by a pilot study that showed participants struggled to learn this task and we wanted to ensure that they received sufficient training. Breaks of up to 5 minutes were included midway through, and upon completion of the training trials. Participants then completed a 'test' phase, in which they no longer received any feedback, with the number of trials equal to 20% of training trials (60 and 100 trials, respectively).

2.2. Results and Discussion

We computed the average categorisation accuracy of participants for the second half of the training phase and for the test phase (Figure 2). In the test phase of the random pixel condition, participants' accuracy (M = 0.529, SD = 0.11) was not statistically different from chance (One-sample t-test, two-tailed, $t(20) = 1.229$, $p = 0.233$). In contrast, most subjects learned to categorise the images in the shuffled labels condition (M = 0.895, SD = 0.157). The difference in accuracy between the two groups was statistically significant (two-tailed, $t(40) = 8.639$, $p < 0.001$, Cohen's d = 2.699, 95%CI=[1.825; 3.572]).

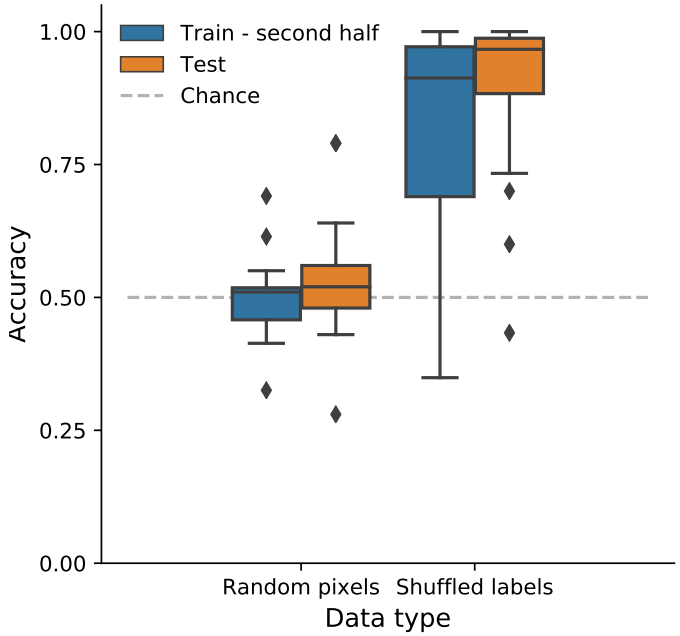


Figure 2: Humans do not improve much over chance levels when classifying random pixel images. Boxes show average accuracy for the second half of the training phase (blue) and the test phase (orange)

The results of the behavioural experiments match our expectations. Participants were near or at chance performance in categorising the 20 random pixel images into two categories, even after 500 learning trials. Three individuals who succeeded in achieving above-chance performance were invited to answer a post-experiment debrief question about what strategy they used to complete the task. Out of three participants, two replied, specifying they tried to distinguish stimuli by focusing only on certain patches within the images (such as the top-left quadrant) and paying attention to the local properties of that region (e.g. "This one had more red overall"), rather than paying attention to individual pixels. Although these heuristics allowed some participants to achieve an above chance performance in a 2-way categorisation task involving a small dataset of 20 images, they would become much more difficult to execute when increasing the number of examples or categories. Scaling the strategy to a 1,000-way classification of ~ 1 million images is infeasible.

In contrast to the random pixel condition, almost all participants were able to learn to classify the 20 novel objects into two categories in the shuffled labels condition with over 80% accuracy in the test phase. It is unlikely humans can rival the capacity of a network to learn to classify ~ 1 million images into 1,000 random categories, but the results showed another disconnect between humans and neural networks — humans find it more difficult to learn to classify random pixel images than novel objects with shuffled labels, whereas Zhang et al. (2017) observed that models did not show any preference for learning the structured data (shuffled labels) compared to random pixels. Indeed, CNNs took *longer* to learn to categorise stimuli in the shuffled labels condition compared with the random pixel condition. Together, these results highlight a fundamental misalignment in

information processing between the models and the human visual system.

3. Improving correspondence between networks and human behaviour

The experiment in Section 2 established two types of differences between humans and CNNs, namely, a) humans struggle to learn even a small set of unstructured data (random pixels condition), while CNNs are much better at learning this type of data, and b) humans are better at learning structured data that is randomly assigned to categories (shuffled labels condition) while CNNs are equally capable of learning both naturalistic images with shuffled labels and unnatural images with random pixels (Zhang et al., 2017).

Here we explore whether different training environments and architectural modifications can make CNN performance more qualitatively similar to the human results. First, in an attempt to model the effect of human experience — participants spend a lifetime classifying naturalistic images prior to taking part in the experiment — we investigated whether pre-training CNNs on datasets of naturalistic images limits the learnability of the unstructured datasets. Second, we introduced various biologically inspired constraints to CNNs designed to reduce network capacity to learn unstructured data, namely, the introduction of processing bottlenecks, internal noise, and bounded activation functions, as detailed below.

3.1. The role of experience

The divergence in learning and memorisation abilities between humans and CNNs might be explained by the differences in their visual experience. Living in a world of statistical regularities and predictable structures, human beings learn and adapt to these reliable features. One consequence of this could be becoming

less adept at dealing with information that lacks structure. For example, experts in a given domain like chess are remarkably better than novices in remembering and reconstructing chessboard configurations from games, but perform on an equal footing with less experienced players when reconstructing random configurations of pieces (Chase & Simon, 1973). Indeed, the effects of early visual experience on development and behaviour can be profound. For example, limiting the visual experience of kittens exclusively to either horizontal or vertical bars makes the animals insensitive to the oriented stimuli that they were not previously exposed to (Blakemore & Cooper, 1970). We examine whether a comparable argument can be extended to CNNs. Does experience with the natural world, operationalised as pre-training a network on a dataset of natural image, set limits on what kind of data can be further learned by that network? Would this experience cause learning unstructured data, such as random pixel images, to be more difficult?

3.1.1. Methods

We trained `ResNet-50`, a standard CNN architecture (He et al., 2016), to classify either the shuffled label `CIFAR-10` dataset or random pixels images. In order to assess the impact of past experience with structured data, we compared a baseline condition in which the weights were randomly initialised (using uniform Xavier initialisation (Glorot & Bengio, 2010)) to CNNs with the weights learned by pre-training the network on the `ImageNet` dataset (Deng et al., 2009), a large database of annotated naturalistic images. For the pre-trained networks we included two training conditions. In a transfer learning condition (Yosinski et al., 2014), we 'froze' the weights of all convolutional and pooling layers (meaning that these weights were not updated during further training) and re-initialised the weights of the fully connected layers. In a 'fine-tuning' condition we allowed all

the weights of the network to be modified during training, similar to the baseline condition. We used three different seeds for every condition. The networks were trained until convergence. The model architecture was modified by inserting one hidden layer with 1024 units immediately prior to the output layer, adding ~ 2 million parameters to the network. This modification was necessary because initial tests indicated that without it, when all convolutional layers in the network were frozen, none of the datasets could converge to sufficiently high accuracy.

Images from `CIFAR-10` were upsampled with smoothing (Burt & Adelson, 1983) from 32×32 pixels to 224×244 pixels to match the input size of the pre-trained networks. Random pixel images were created in resolution 32×32 pixels, then resized without smoothing to 224×224, forming a grid of 32 squares of size 7×7. All images were represented in RGB format. The size of the random pixel dataset matched the size of the `CIFAR-10` training data at 50,000 images. The datasets were the same in each iteration for all models.

3.1.2. Results and Discussion

We assessed the accuracy of the networks on the unmodified data, random pixel and shuffled labels datasets in the three training conditions (Figure 3). Strikingly, pre-training the networks on `ImageNet` did not prevent them from learning the random datasets. Both random pixels images and shuffled labels data were eventually learned with perfect accuracy across all conditions.[2] While pre-training

[2]Note that, unlike the findings of Zhang et al. (2017), the networks trained on the shuffled labels dataset converged faster than those trained on random pixel images. We found that this relationship was not a robust one, and varied based on e.g., the choice of architecture, or even the learning rate.

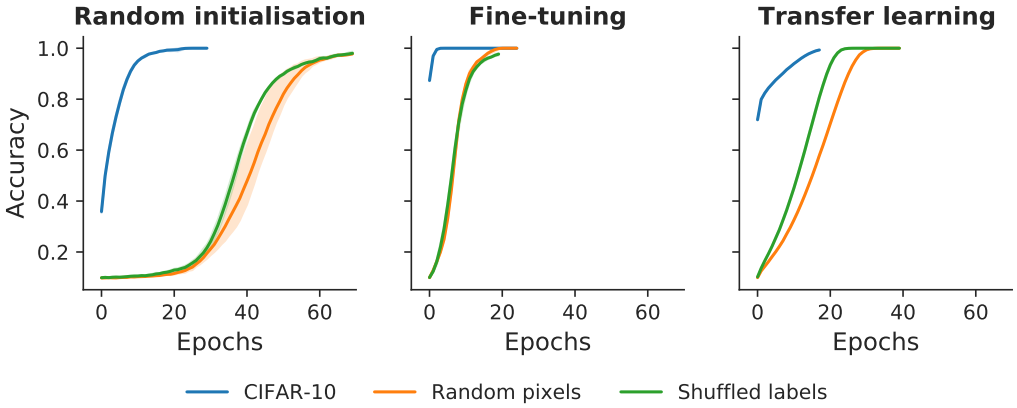


Figure 3: Pre-training on `ImageNet` does not prevent learning random data. Training accuracy is shown for `ResNet-50` networks trained to classify either unmodified `CIFAR-10` (blue), the same dataset with shuffled labels (green), or Random pixel images (orange). The models were either trained from a random initialisation (**Left**), or were pre-trained on `ImageNet`, fine-tuning the entire network (**Right**), or transfer learning: freezing the convolutional weights and training only the classifier (**Center**).

accelerated learning on the structured (`CIFAR-10`) dataset, it also made learning faster on the unstructured datasets, for both the fine-tuning and transfer learning conditions. Crucially, we did not observe an interaction between pre-training and the type of data being learned. That is, we did not find pre-training to be preferentially beneficial to the speed of learning the `CIFAR-10` dataset compared to the random pixels dataset.

These results do not corroborate the hypothesis that experience with natural images restricts the network's ability to learn random data. Indeed, the networks pre-trained with `ImageNet` learned the random dataset faster than those with randomly initialised weights.

3.2. The role of biological constraints

Next, we investigated whether introducing biologically-inspired constraints would limit the capacity of CNNs to learn random data. We experimented with three such constraints (see Figure 4).

First, we introduced **internal noise** — or variability — in the activations of units in the CNN. There are various sources of internal noise in the brain. For example, it is a long-standing observation that presentation of the same stimulus can often result in different neural response patterns (Stein et al., 2005). This variability can be observed at multiple spatial and temporal scales, caused by environmental factors or specific cell properties. The largest source of neuronal noise is synaptic, produced stochasticity in neurotransmitter release, which can have a substantial net effect on the behaviour of post-synaptic cells (Stein et al., 2005). Adding internal noise to the activation values of hidden units in a neural network reduces the precision of the signal, and in turn, may reduce the capacity of the network to encode large numbers of random categories. If the precision of network computations is more critical for the random datasets due to the networks having to rely on fine-grained idiosyncrasies in the data, then internal noise should start to drown this delicate signal and selectively limit CNNs' ability to classify random data.

Another biologically inspired concept we explored is an **information bottleneck** (Tishby et al., 1999) in a hidden layer. Information bottlenecks could, in principle, play a pivotal role in reducing representational complexity to allow efficient processing in the perceptual system (Essen et al., 1991) by discarding irrelevant details and passing on the most reliable and robust features of the inputs (Evans et al., 2022). We hypothesised that if category members share some

regular features — as is the case with `CIFAR-10` categories — an information bottleneck should encourage these features to be discovered, leading to robust representations. On the other hand, if no common features for each category can be detected, then a bottleneck is expected to have a large impact on network performance.

Lastly, we considered the role of the **activation function** on a neural network's capacity. Typically, modern deep learning models use activation functions which permit a large range of activations. For example, the rectified linear (ReLU) units lead to output activations that are bounded at the lower end but unbounded at the upper end. In contrast, the Sigmoid activation function, which is often encountered in connectionist models of cognitive processes and neurophysiological models of neural population dynamics (Wilson & Cowan, 1972), bounds output activation at both the lower as well as the upper end. Although there is lack of consensus about the biological plausibility of Sigmoid versus ReLu activation functions (see Glorot et al., 2011), we focus on the effect of their range on regulating the capacity for representing information. The activation of ReLU units can grow as large as needed, and their representational capacity is only limited by the degree of precision imposed by the numeric data. Units using the Sigmoid activation function, on the other hand, have their output values limited within the range $(0, 1)$. When paired with the limited precision caused by floating-point operations in computers, they therefore provide a natural constraint on the representational capacity of these units.

3.2.1. Methods

We used two model architectures, `small-inception` and `small-alexnet`, adapted from Zhang et al. (2017, p. 12). These models were not pre-trained as in

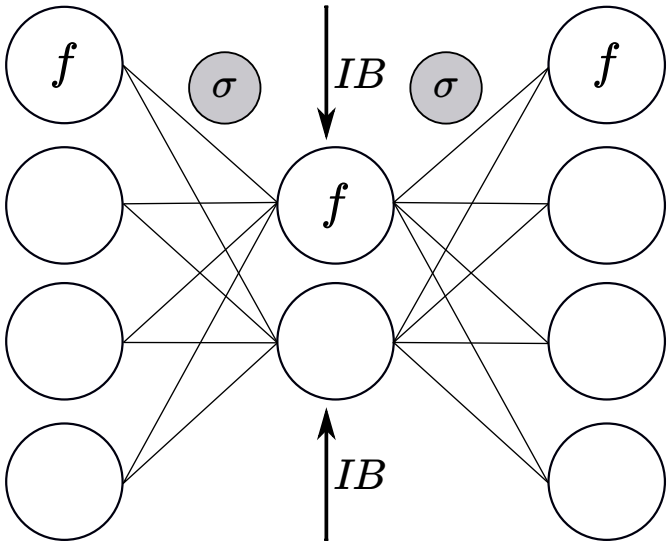


Figure 4: Illustration of biologically-motivated capacity constraints. Gaussian noise (σ) was added to the output of activation functions of hidden units. Hidden units had Sigmoid instead of ReLU activation function (f). The information bottleneck (IB) was introduced by reducing the number of convolutional filters in a layer, modifying the architecture of the network.

Section 3.1, but started with randomly initialised weights. While the investigation into the role of experience demanded working with models pre-trained on a large natural image dataset, the architectures selected for this study are scaled down versions of popular models, reducing computational demands. This reduction in trainable parameters permits a wider and more thorough exploration of relevant conditions and hyperparameters.

We implemented internal noise by adding random values drawn from a Gaussian distribution ($\mu = 0, \sigma = [0 : x]$) to the output of the activation function of every hidden unit in the neural network.[3] We varied the standard deviation of internal noise (σ) from $0 - 1.2$, with an increment of 0.05 for networks using ReLU activation units, and from $0 - 0.2$, in 0.02 steps for Sigmoid networks.

Inspired by Lindsey et al. (2019), we implemented an information bottleneck by reducing the number of convolutional filters in the first convolutional layer. This procedure aims to model the reduction from photoreceptors to ganglion cells in the retina. We experimented with several values for the number of filters in the bottleneck and settled on the smallest number possible without catastrophically impairing performance on the `CIFAR-10` dataset. As a result, we reduced the number of filters in `small-inception` from $96 \rightarrow 2$ and from $200 \rightarrow 8$ for `small-alexnet`. Although this manipulation only reduces the total number of trainable parameters by a small amount, it has been demonstrated to qualitatively alter the types of filters learned in early convolutional layers, resulting in the development of center-surround receptive fields in the first layer and a prevalence of Gabor-like filters in subsequent layers, similar to what is observed in the human

[3] Implemented as a regularisation layer from `tensorflow.keras` module.

visual system (Lindsey et al., 2019).

We generated model instances for all four combinations of the activation (ReLU or Sigmoid) and bottleneck (yes or no) constraints. Each network instance was trained to classify the three types of datasets described in Section 3.1: (a) the unmodified `CIFAR-10` dataset, (b) a version of the `CIFAR-10` dataset with shuffled labels (c) random pixel images. All networks were trained for 100 epochs, with a batch size of 128 (39,000 steps). We used learning rates which lead to the fastest convergence on all datasets (determined empirically; $lr = 0.1$ for `small-inception` and $lr = 0.01$ for `small-alexnet`). A step decay algorithm was used to decrease learning rate by 5% after each epoch. Internal noise was added to the output of the activation function of each hidden unit in every convolutional or fully-connected layer of the CNNs. All results were averaged over four different random seeds for each network instance, with new random pixel images and a shuffled labels permutation for every seed. The same hyperparameters were used for all datasets and seeds.

3.2.2. Results and Discussion

Results from all simulations are summarised in Figure 5 (top row), which shows accuracy on the training set after 100 training epochs for all three datasets for the `small-inception` architecture. Each panel plots the accuracy after 100 epochs as a function of the internal noise. Performance on the `CIFAR-10`, random pixel and shuffled labels datasets are shown using blue, orange and green lines. Bold lines show performance without a bottleneck while dashed lines show performance using a bottleneck. [4]

[4] Results for `small-alexnet` are qualitatively similar — see Appendix A.

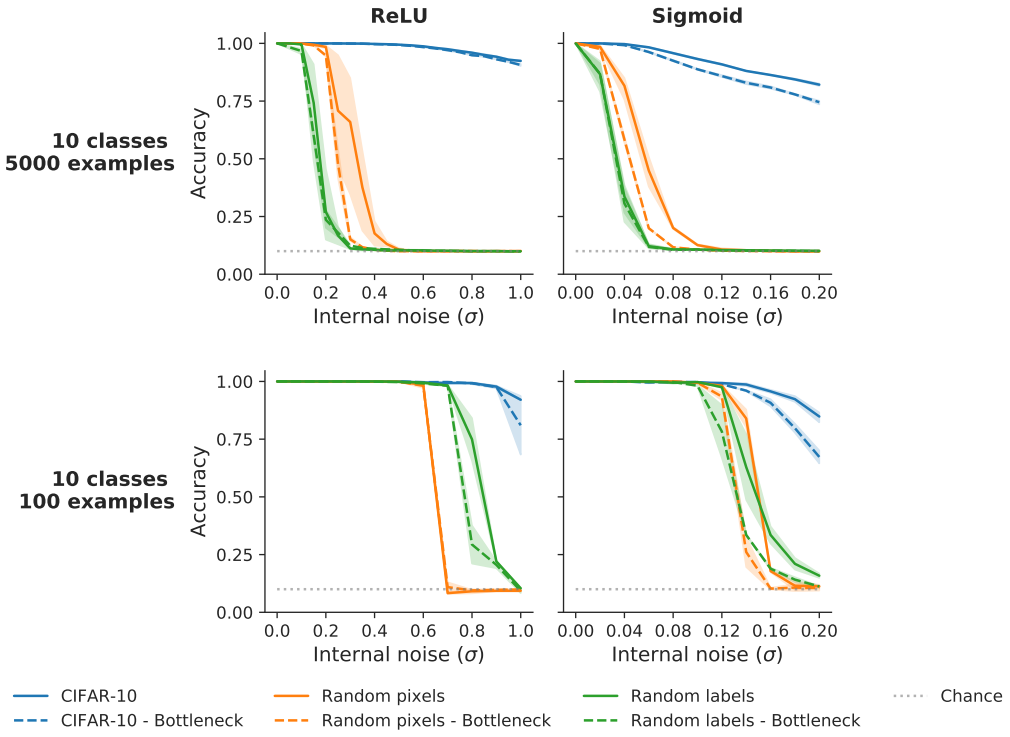


Figure 5: **Top**: Training with biological constraints severely reduces CNNs' ability to learn random datasets. These figures represent the results described in Section 3.2. **Left**: ReLU activation function. **Right**: Sigmoid activation function. The graphs show the impact of the capacity manipulations on training accuracy (y-axis) for the three types of datasets (blue: unmodified `CIFAR-10`; green: modified `CIFAR-10` with shuffled labels; orange: dataset of random pixels images). The internal noise levels used are plotted on the x-axis. Solid lines indicate model accuracy for models without an information bottleneck; dashed lines indicate models with a bottleneck. **Bottom**: CNNs trained on smaller datasets require greater values of internal noise for random dataset learning to decrease. The figures represent the results described in Section 3.3

We observed that internal noise had a large, selective effect on the ability of the network to learn the two unstructured datasets. That is, within the explored range of values, noise had little impact on the training set classification accuracy for

CIFAR-10 images (although the effect was more noticeable for small-alexnet networks). On the other hand, increasing noise values progressively hindered performance on the random pixel and shuffled labels conditions, with training accuracy quickly dropping to chance. Comparing networks using ReLU hidden units with those using Sigmoid units, it appears changing the activation function, by itself, did not impact the ability of the networks to learn to classify the random datasets. However, the activation function interacted with the networks' tolerance to internal noise, with ReLU unit networks requiring ~ 5 times more noise compared with Sigmoid units to produce a similar decline in accuracy for the random pixels and shuffled labels conditions (see different scales on x-axes). Nevertheless, the same pattern of results was obtained with the two activation functions, with a selective effect of increased noise on the two unstructured datasets but little impact on classifying CIFAR-10 images.

We also observed that the bottleneck manipulation was not sufficient to prevent learning of unstructured data on it's own, although its effect interacted with the degree of internal noise and the activation function. That is, for networks with a bottleneck, a slightly lower internal noise value was needed to prevent the network from learning unstructured data.

3.3. Does noise affect capacity?

The findings in Section 3.2 suggest that adding internal noise to the activations of hidden units was the most effective method for reducing CNNs' ability to learn random datasets. The next step was to work out *why* it was effective. Our hypothesis was that this is because noise decreases capacity. If this is the case, then even after inserting noise the network should be able to learn the random pixel and shuffled labels datasets, as long as they do not exceed the now lowered

capacity. We manipulated this demand on the network by decreasing the number of exemplars in each category.

We trained the same model architectures used in Section 3.2 on identical tasks — classifying `CIFAR-10` images with either unmodified or shuffled labels, as well as random pixel images. However, we reduced the size of the datasets so that each class contains fewer examples than the full dataset. Lowering the size of the dataset lowers the difficulty of the problem the network has to solve because the model has to fit fewer data points. We hypothesised that if adding internal noise affects capacity, then by reducing the dataset size, and therefore task difficulty, the level of internal noise necessary to suppress random data learning would increase. For example, if CNNs with internal noise σ_{full} were no longer be able to learn a dataset of 50,000 random pixel images, they could still learn a smaller dataset of, e.g., 1,000 examples. That is, a larger noise value, σ_{part} ($\sigma_{part} > \sigma_{full}$) would be required to prevent the CNNs from learning the smaller dataset.

3.3.1. Method

We recreated the simulations in Section 3.2, using the same network architectures, but decreased the number of classes and examples per class in each dataset. For the unmodified `CIFAR-10` and the shuffled labels conditions, a stratified sampling procedure was used to ensure an equal number of examples were sampled from each class. For the random pixels condition, an appropriate number of images were generated and randomly assigned to each class. The tested datasets consisted of 10 classes, with 100 examples per class.

Training the neural networks on the smaller dataset with the same batch size and number of training epochs used when training on the full dataset would result in overall fewer weight updates. Therefore, we scaled the number of train-

ing epochs, while keeping the batch size the same, so that networks learning the smaller dataset would have the same number of training steps as those trained on the full dataset.

3.3.2. Results and Discussion

The pattern of results corroborates our hypothesis, with greater values of internal noise needed to prevent CNNs from learning the unstructured datasets. This is easily seen when compared with results from training on the datasets of 50,000 images used in Simulation 3.2 on the top row of (Figure 5).[5]

There was another interesting, unanticipated effect of training the models with smaller datasets, namely, the relative difficulty of learning random pixels versus shuffled labels swapped when trained on 100 compared to 5,000 images per category in the ReLU condition. However, we do not consider this finding to be robust, as the results were less clear in the Sigmoid condition. We also found that the relative difficulty of the two random conditions varied in the pre-training simulationss in Section 3.1, influenced by various factors such as hyperparameter choice. We do not have a good explanation for this effect, but a successful model of human vision should not only explain the very limited capacity of humans to learn unstructured data, but also the much greater difficulty humans experience in learning the random pixel stimuli. Although our introduction of biological constraints is a step in the right direction, we have not identified a condition that allows to disassociate the ability to learn these two forms of random data within the existing CNN framework.

[5]See Appendix A, Figure A.1 for comparable results for `small-alexnet` models.

4. Do capacity constraints improve generalisation?

In the simulations in Sections 3.2 and 3.3 we have shown how adding biologically motivated constraints makes CNNs more similar to humans in that the neural networks' capacity to learn random data is reduced without a notable disruption to their ability to classify natural images. We now consider whether learning with these constraints has consequences for other areas where correspondence between CNN and human performance is lacking, namely, out-of-distribution (*o.o.d.*) generalisation. That is, CNNs struggle to generalise to images that differ in their statistics from images that the network was trained on (Geirhos et al., 2018; Sinz et al., 2019). Here, we look at the performance of CNNs when presented with images that have been manipulated with image distortions not observed during training. This problem is challenging for CNNs, with performance of models degrading rapidly after even small amounts of image distortions or other perturbations (Geirhos et al., 2019). Human performance also deteriorates under these conditions, but more smoothly and to a much lesser extent (Geirhos et al., 2019). We tested whether training CNNs with resource capacities limited by biologically-inspired constraints on naturalistic data will force the networks to learn more general representations that support improved *o.o.d.* generalisation.

4.1. Methods

To test *o.o.d.* generalisation, we followed the method used by Geirhos et al. (2019) to modify the `CIFAR-10` test set. We created two test sets, each of which converted images from the `CIFAR-10` test set to grayscale and added a type of noise that was not seen during training. In the first set, we added uniform (zero-centred) noise to each image — that is, we modified each pixel by a random value

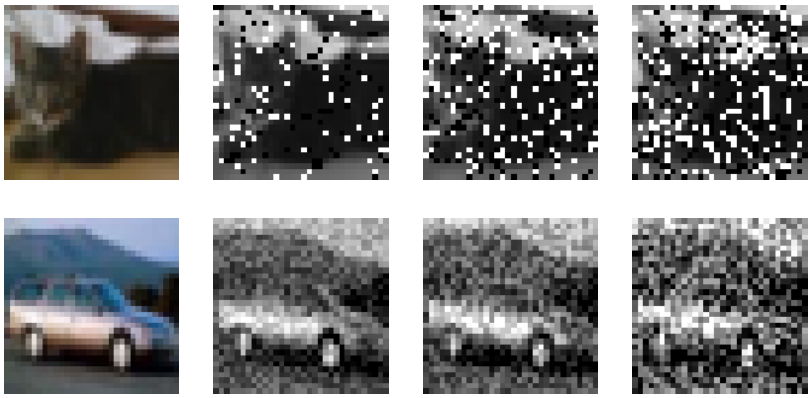


Figure 6: Example CIFAR-10 images with different amounts of salt and pepper noise (**top**) or uniform noise (**bottom**).

drawn from the distribution $[-a, +a]$, in which we varied the value a in the range $[0, 0.5]$ in increments of 0.05. In the second dataset, we applied salt-and-pepper noise to each image, where each pixel of an image can be switched to black or white with a defined probability, p, in the range $[0, 0.5]$, varied in increments of 0.05 (5%). Examples of both types of image degradation are shown in Figure 6.

We trained the same set of CNN architectures used in Section 3.2 to classify the unmodified CIFAR-10 training set and tested each model instance on the two *o.o.d.* generalisation test sets. We trained the networks with every combination of constraints — using either ReLU or Sigmoid unit activation, with or without a bottleneck, while varying the level of internal noise. The results were averaged over four random seeds for each network.

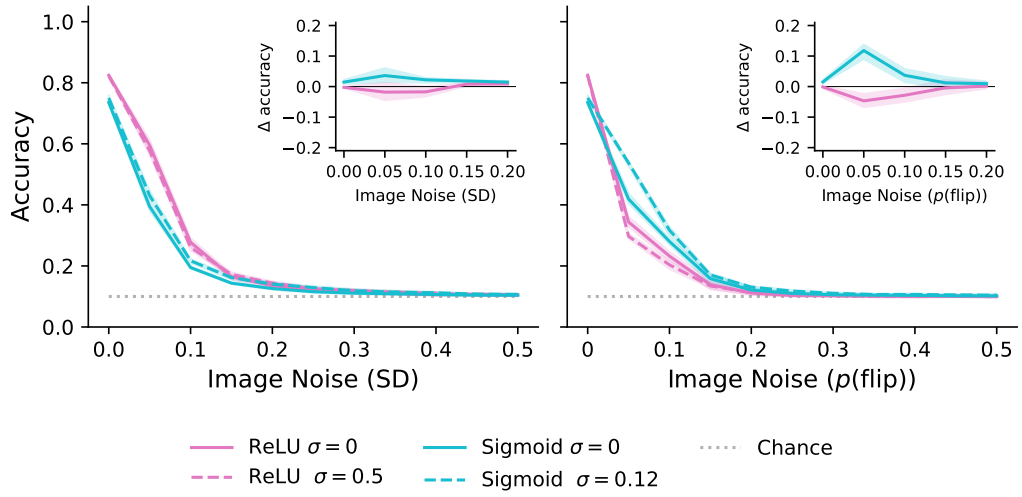
4.2. Results and Discussion

Figure 7 shows the change in generalisation accuracy for `small-inception` networks, trained with a particular value of internal noise, relative to a baseline network trained without activation noise.[6] The inset shows the change in accuracy from the baseline model. Consistent with previous findings (Geirhos et al., 2018), the networks are not much affected by the very low levels of image degradation, but increasing it even slightly further rapidly lowers the accuracy of all models to chance performance (10%). In contrast, human performance decreases much more gradually with increased degradation (Geirhos et al., 2018).

For the networks with ReLU activation, the accuracy of the models with internal noise was $5 - 10\%$ lower than the performance of the baseline models, for both additive uniform and salt-and-pepper degradation. That is, training with noise offered no advantage in generalisation to these images. On the other hand, the CNNs using sigmoid activations and internal noise did show improvement over the baseline model for both types of degradation (see Insets in Figure 7). The size of the effect was small for uniform noise ($< 5\%$) and somewhat larger for salt-and-pepper noise ($\sim 10\%$). The magnitude of the effect was also enhanced by the addition of a bottleneck (Appendix C, Figure C.2). However, this advantage was observed only for low levels of image degradation and quickly disappeared as the performance of all models deteriorated. Furthermore, the advantage was either not observed, or considerably smaller for `small-alexnet` networks (see Appendix A, Figure A.2).

The constraints appear to have slightly improved generalisation abilities for

[6] See Appendix C for results for models with information bottleneck.



[t]

Figure 7: `small-inception` networks with internal noise have better generalisation for low levels of input degradation for sigmoid (teal), but not ReLU (pink) activation function. **Left**: Uniform noise. **Right**: Salt-and-pepper noise. Solid lines represent baseline models trained without constraints. Dashed lines represent networks trained with one specific noise level (0.5 for ReLU and 0.12 for sigmoid). Insets highlight deviation of models with internal noise from the baseline model (distance from the midline). So values above the x-axis show a better performance compared to baseline while values below show a worse performance.

some combinations of conditions, but not for others, and only for some low levels of noise. As the advantage is further contingent on factors such as the model architecture, we conclude that the results do not support our prediction that training with constraints will overall improve robustness to image distortions. Clearly, additional constraints are needed to explain the human capacity to generalise to *o.o.d.* datasets. All models, both with and without constraints, quickly deteriorated as the amount of image degradation increased, at a similar pace. Only the networks with Sigmoid activation showed an improvement in accuracy, but only for low amounts of input noise. Furthermore, it is not clear why the effect is larger for salt-and-pepper noise compared with uniform noise. Finally, it should be taken into account that in absolute terms the accuracy of ReLU networks for uniform input noise was consistently higher than that of Sigmoid networks. For salt-and-pepper noise, the reverse was true — CNNs with sigmoid activations performed better than networks with ReLU units.

5. General Discussion

If CNNs are to be considered models of human vision they should not only succeed in classifying images under conditions in which humans succeed, but also fail under conditions in which human fail. Here we carried out a series of experiments that explore a striking example of CNNs outperforming humans, namely, in their ability to classify random data, in the form of images that look like TV static (random pixel images) or naturalistic images that are randomly assigned to categories (shuffled labels), as reported by Zhang et al. (2017). Our key contribution is to introduce biological constraints to CNNs that allow them perform well on unmodified naturalistic image data but fail on random data. These findings

suggest that these constraints, or some version of them, should be implemented in CNNs when modelling human vision in order to provide better alignment between their performance characteristics.

To summarise, we report five main findings. First, in two behavioural experiments we confirm that humans perform far worse when learning random data compared to standard CNNs (Figure 2). In addition, we observed that humans are able to learn (a small set of) images with randomly assigned labels, but are mostly unable to learn images of random pixels. This behaviour contrasted with CNNs, which learn random pixels eitehr more easily Zhang et al. (2017), or as in our simulations, learn the two types of random data at similar rates, with random pixels sightly easier than shuffled labels in some conditions and slightly harder in others.

Second, we found that exposing CNNs to structured data (pre-training on `ImageNet`) did not prevent them from learning random data. Surprisingly, pre-training actually made it easier for CNNs to learn to classify the random datasets (Figure 3), suggesting that this superhuman capacity reflects their architectures rather than their background training history.

Third, we introduced three biologically inspired modifications to CNNs designed to reduce their resource capacities. The addition of activation noise designed to model neural noise was the most important factor in selectively reducing performance on random data, but the introduction of a bottleneck (modelling the optic nerve) and a sigmoidal activation function (corresponding to the bounded firing rate of neurons) contributed to more human-like performance as well (Figure 5, **Top**).

Fourth, we show that these biological constraints worked by reducing network

capacity rather than through some alternative mechanism. That is, we showed that CNNs with these constraints could still learn noise patterns, but far fewer of them (Figure 5, **Bottom**). This is the pattern of results to be expected if the manipulation interferes with capacity.

Finally, we did not find that the pre-training and the biological constraints helped much to improve better generalisation to *o.o.d.* data. Indeed, most of the networks did not improve on, or fared worse than baseline models with no constraints when tasked to classify test set images perturbed with different types of noise (Figure 7). Constraints did appear to improve performance slightly in CNNs with sigmoid activations, but benefits were mostly restricted to low levels of one particular type of input noise.

What do our results say about the correspondence between CNNs and human vision? The first point to note is that our findings highlight the importance of adding biological constraints that reduce CNN resource capacity when attempting to model human vision. Although the extraordinary success of state-of-the-art CNNs in image classification is often taken as evidence that CNNs are promising models of human vision, these models are clearly exploiting resource capacities that far exceed those available to networks in the human visual system, allowing models to learn random data almost as well as structured data. Biologically plausible CNNs need to succeed in classifying images without the ability to learn random data and our own findings highlight three constraints that appear to be relevant to this.

However, clearly the pre-training and the biological constraints are not enough to reconcile the CNN and human results. First, the networks do not reproduce the pattern of relative difficulty of the random datasets that humans exhibit. Learning

to classify random pixel images is much more difficult for humans than classifying structured objects in arbitrary categories. This pattern of results was not consistently observed with CNNs. And more importantly, the networks with capacity constraints, that were selectively good at classifying structured data, were not much better at *o.o.d.* generalisation. Clearly a range of additional constraints and perhaps new models are required to capture this key feature of human object recognition.

Arpit et al. (2017) have also explored how to reduce the memorisation of random data in CNNs, although they did not consider the relevance of their findings to human vision. Specifically, they assessed the ability of several commonly used regularisation techniques to reduce memorisation of random data. Interestingly, some of the findings of Arpit et al. (2017) differ from our own results. Most relevant to our work, they concluded that the introduction of internal Gaussian noise was ineffective as it reduced performance on both structured and unstructured data. By contrast, our results show that it is possible to find levels of noise that selective impair performance on unstructured data. It is not clear why we reached different conclusions, but it may be explained by differences in the implementation of models. Firstly, we applied internal noise both during training and testing of the networks. This is not conventional practice when using regularisers, but bears closer correspondence to the biological brain. Secondly, Arpit et al. (2017) do not describe the methodological details of their procedure, such as at which points within a network the noise is added. This lack of procedural details prevents further comparisons between the results of the two studies.

It is interesting to note that a similar pattern of learning unstructured data is observed in the domain of natural language processing as well. That is, while

deep networks show impressive successes in learning natural languages, unlike humans, they also find it easy to learn impossible languages (Mitchell & Bowers, 2020). Both findings highlight missing inductive biases in standard CNNs that result in networks with superhuman learning capacities in some domains. A key challenge in creating neural network models with greater correspondence to human performance is to introduce the relevant constraint and biases so that models learn what they should learn (under realistic training conditions), and at the same time do not learn what humans cannot.

In summary, we have found that adding three basic biological constraints to CNNs leads to them exhibiting more human-like performance through impeding their ability to classify random data, while preserving their ability to classify structured data. However, models trained with the examined constraints are not better at *o.o.d.* generalisation. Developing models that succeed where humans succeed, but also fail where humans fail, is an important challenge for future work in developing better models of human vision.

References

Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., & Lacoste-Julien, S. (2017). A closer look at memorization in deep networks. In D. Precup, & Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning* (pp. 233–242). PMLR volume 70 of *Proceedings of Machine Learning Research*. URL: `https://proceedings.mlr.press/v70/arpit17a.html`.

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

//www.sciencedirect.com/science/article/pii/0010028588900242. doi:https://doi.org/10.1016/0010-0285(88)90024-2.

Blakemore, C., & Cooper, G. F. (1970). Development of the brain depends on the visual environment. *Nature*, *228*, 477–478. URL: http://www.nature.com/articles/228477a0. doi:10.1038/228477a0.

Burt, P., & Adelson, E. (1983). The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, *31*, 532–540. doi:10.1109/TCOM.1983.1095851.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*, 55–81. URL: https://www.sciencedirect.com/science/article/pii/0010028573900042. doi:https://doi.org/10.1016/0010-0285(73)90004-2.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*, 834–848. doi:10.1109/TPAMI.2017.2699184.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Dujmović, M., Malhotra, G., & Bowers, J. S. (2020). What do adversarial images tell us about human vision? *eLife*, *9*. URL: https://doi.org/10.7554/elife.55978. doi:10.7554/elife.55978.

Essen, D. V., Olshausen, B. A., Anderson, C. H., & Gallant, J. T. (1991). Pattern

recognition, attention, and information bottlenecks in the primate visual system. In B. P. Mathur, & C. Koch (Eds.), *Visual Information Processing: From Neurons to Chips* (pp. 17 – 28). International Society for Optics and Photonics SPIE volume 1473. doi:10.1117/12.45537.

Evans, B. D., Malhotra, G., & Bowers, J. S. (2022). Biological convolutions improve dnn robustness to noise and generalisation. *Neural Networks*, *148*, 96–110. URL: https://www.sciencedirect.com/science/article/pii/S0893608021004780. doi:https://doi.org/10.1016/j.neunet.2021.12.005.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*, 665–673. URL: https://doi.org/10.1038/s42256-020-00257-z. doi:10.1038/s42256-020-00257-z.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=Bygh9j09KX.

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc. volume 31. URL: https://proceedings.neurips.cc/paper/2018/file/0937fb5864ed06ffb59ae5f9b5ed67a9-Paper.pdf.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Y. W. Teh, & M. Titterington (Eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (pp. 249–256). Chia Laguna Resort, Sardinia, Italy: PMLR volume 9 of *Proceedings of Machine Learning Research*. URL: https://proceedings.mlr.press/v9/glorot10a.html.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In G. Gordon, D. Dunson, & M. Dudík (Eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (pp. 315–323). Fort Lauderdale, FL, USA: PMLR volume 15 of *Proceedings of Machine Learning Research*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). doi:10.1109/CVPR.2016.90.

Horst, J. S., & Hout, M. C. (2016). The novel object and unusual name (noun) database: A collection of novel images for use in experimental research. *Behavior Research Methods*, *48*, 1393–1409. URL: https://link.springer.com/article/10.3758/s13428-015-0647-3. doi:10.3758/s13428-015-0647-3.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLOS Computational Biology*, *10*, 1–29. doi:10.1371/journal.pcbi.1003915.

Kriegeskorte, N. (2015). Deep neural networks: A new frame-

work for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446. URL: https://doi.org/10.1146/annurev-vision-082114-035447. doi:10.1146/annurev-vision-082114-035447. PMID: 28532370.

Krizhevsky, A., Nair, V., & Hinton, G. (2009). Cifar-10 (canadian institute for advanced research), . URL: http://www.cs.toronto.edu/~kriz/cifar.html.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLOS Computational Biology*, *12*, e1004896. URL: https://dx.plos.org/10.1371/journal.pcbi.1004896. doi:10.1371/journal.pcbi.1004896.

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., Issa, E., Bashivan, P., Prescott-Roy, J., Schmidt, K., Nayebi, A., Bear, D., Yamins, D. L., & DiCarlo, J. J. (2019). Brain-like object recognition with high-performing shallow recurrent anns. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (pp. 12805–12816). Curran Associates, Inc. volume 32. URL: https://proceedings.neurips.cc/paper/2019/file/7813d1590d28a7dd372ad54b5d29d033-Paper.pdf.

Landau, B., Smith, L., & Jones, S. (1998). Object shape, object function, and object name. *Journal of Memory and Language*, *38*, 1–27. URL: https://www.sciencedirect.com/science/article/pii/S0749596X9792533X. doi:https://doi.org/10.1006/jmla.1997.2533.

Lindsey, J., Ocko, S. A., Ganguli, S., & Deny, S. (2019). The effects of neural resource constraints on early visual representations. In *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=S1xq3oR5tQ.

Malhotra, G., Evans, B. D., & Bowers, J. S. (2020). Hiding a plane with a pixel: examining shape-bias in cnns and the benefit of building in biological constraints. *Vision Research*, *174*, 57 – 68. URL: http://www.sciencedirect.com/science/article/pii/S0042698920300742. doi:https://doi.org/10.1016/j.visres.2020.04.013.

Mitchell, J., & Bowers, J. (2020). Priorless recurrent networks learn curiously. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 5147–5158). Barcelona, Spain (Online): International Committee on Computational Linguistics. URL: https://aclanthology.org/2020.coling-main.451. doi:10.18653/v1/2020.coling-main.451.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2017). Adapting deep network features to capture psychological representations: An abridged report. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17* (pp. 4934–4938). URL: https://doi.org/10.24963/ijcai.2017/697. doi:10.24963/ijcai.2017/697.

Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2019). Do ImageNet classifiers generalize to ImageNet? In K. Chaudhuri, & R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning* (pp. 5389–5400). PMLR volume 97 of *Proceedings of Machine Learning Research*. URL: https://proceedings.mlr.press/v97/recht19a.html.

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster r-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*, 1137–1149. URL: https://doi.org/10.1109/tpami.2016.2577031. doi:10.1109/tpami.2016.2577031.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*, . URL: https://www.biorxiv.org/content/10.1101/407007v2. doi:10.1101/407007.

Sinz, F. H., Pitkow, X., Reimer, J., Bethge, M., & Tolias, A. S. (2019). Engineering a less artificial intelligence. *Neuron*, *103*, 967–979. URL: https://www.sciencedirect.com/science/article/pii/S0896627319307408. doi:https://doi.org/10.1016/j.neuron.2019.08.034.

Stein, R. B., Gossen, E. R., & Jones, K. E. (2005). Neuronal variability: noise or part of the signal? *Nature Reviews Neuroscience*, *6*, 389–397. doi:10.1038/nrn1668.

Tan, M., & Le, Q. (2021). Efficientnetv2: Smaller models and faster training. In M. Meila, & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning* (pp. 10096–10106). PMLR volume 139 of *Proceedings of Machine Learning Research*. URL: https://proceedings.mlr.press/v139/tan21a.html.

Tishby, N., Pereira, F. C., & Bialek, W. (1999). The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communica-*

tion, Control and Computing (pp. 368–377). URL: https://arxiv.org/abs/physics/0004057.

Wilson, H. R., & Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical Journal*, *12*, 1–24. doi:10.1016/s0006-3495(72)86068-5.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*, 8619–8624. URL: https://www.pnas.org/content/111/23/8619.full.pdf. doi:10.1073/pnas.1403112111.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc. volume 27. URL: https://proceedings.neurips.cc/paper/2014/file/375c71349b295fbe2dcdca9206f20a06-Paper.pdf.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. URL: https://openreview.net/forum?id=Sy8gdB9xx.

Acknowledgements

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).

Appendix A.

This section contains figures displaying results for several simulations described in the Section 3 for networks with the `small-alexnet` model architecture. Results from Figure A.1 correspond to those for `small-inception` from Figure 5. Figure A.2 corresponds to Figure 7.

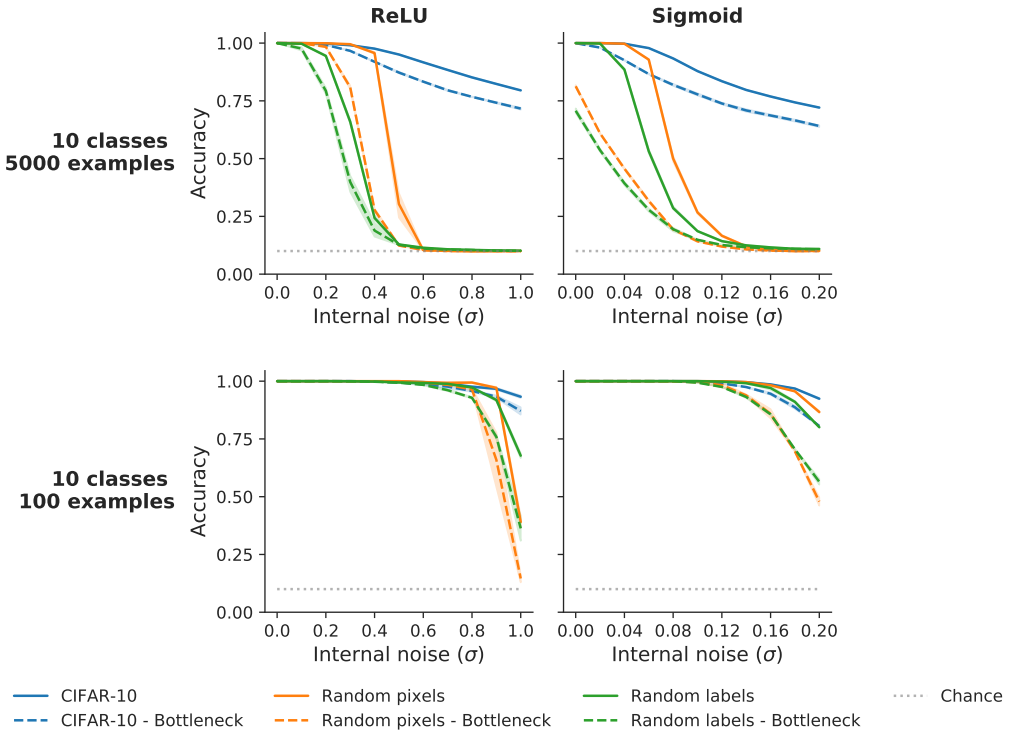


Figure A.1: **Top**: Accuracy for `small-alexnet` models with biological constraints. **Left**: ReLU activation function. **Right**: Sigmoid activation function. The graphs show the impact of the capacity manipulations on training accuracy for the three types of datasets (blue: unmodified `CIFAR-10`; green: modified `CIFAR-10` with shuffled labels; orange: dataset of random pixels images). The internal noise levels used are plotted on the x-axis. Solid lines indicate model accuracy for models without an information bottleneck; dashed lines indicate models with a bottleneck. **Bottom**: Training set accuracy for models trained with biologically inspired constraints. Networks trained on `CIFAR-10` images, with either unmodified (**blue**) or shuffled (**green**) labels, and random pixel images (**orange**). The CNNs architecture either included an information bottleneck (**dashed lines**), or did not (**solid lines**). Hidden units used either ReLU (**left**) or sigmoid (**right**) activation functions. Networks trained with a limited number of examples per class (10 classes with 100 examples).

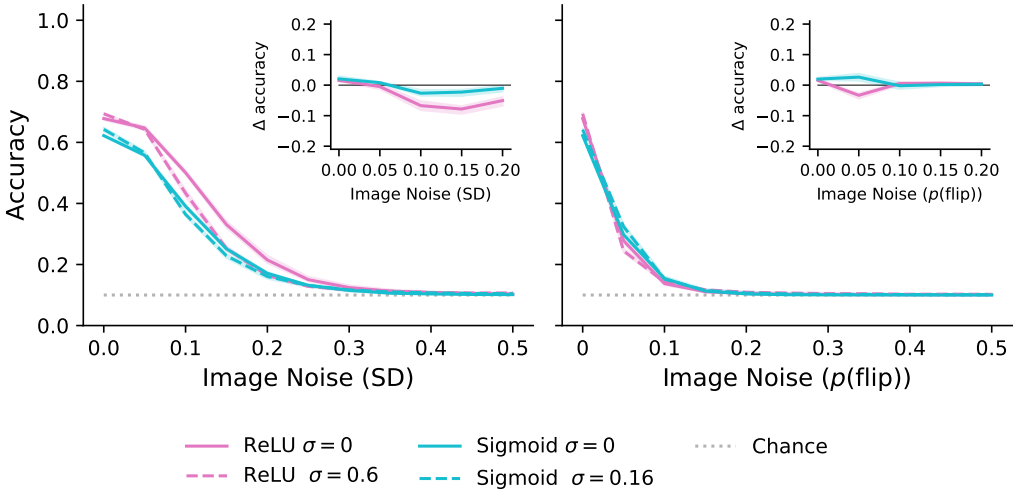


Figure A.2: `small-alexnet` models trained with internal noise (orange lines) do not generalise to novel distorted images better than a baseline model (blue lines). **Left**: Uniform noise. **Right**: Salt-and-pepper noise. Solid lines represent baseline models trained without constraints. Dashed lines represent networks trained with one specific noise level (0.6 for ReLU and 0.16 for sigmoid). Insets highlight deviation of models with internal noise from the baseline model (distance from the midline). Values above the x-axis show a better performance compared to baseline while values below show a worse performance.

Appendix B.

Neural networks that classify random data are often described as performing 'memorisation' (Zhang et al., 2017; Arpit et al., 2017). The basic idea is that networks are representing each specific input pattern rather than extracting common features amongst members of a given class. Indeed, that must be the case given that there are no common shared features of members of a given class, as with random pixel images. One plausible prediction is that CNNs that classify images on the basis of memorising all the training patterns will not only fail to generalise well to test stimuli, but will also generalise less effectively if the training data were to be perturbed in a way not experienced during training. Indeed, learning invariant patterns is often thought to support generalisation in CNNs.

To explore generalisation in CNNs trained to memorise random data we trained `small-alexnet` networks to classify random pixel images. Afterwards, taking a smaller sample of the training dataset, we changed the values of a randomly selected proportion of the pixels in each image. The new values were sampled from a uniform distribution $a \sim U(0., 1.)$. We then assess the accuracy of the networks on this modified training data. We find that, while network performance decreases as a larger proportion of the original pixel values are changed, their accuracy remains relatively high (Figure B.1). This effect is present in both ReLU (pink) and Sigmoid (teal) networks.This highlights that CNNs trained on random data can generalise in the sense that they are robust to perturbations in the data they were trained on.

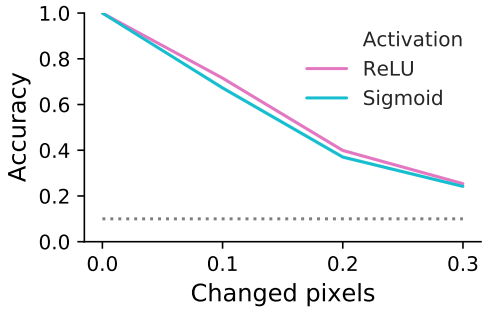


Figure B.1: Prediction accuracy of `small-alexnet` models with ReLU (pink) or Sigmoid (teal) activation on modified random pixel training data. The CNNs performance degrades gradually as a greater proportion of pixels is changed, not all at once.

Appendix C.

The figures below show the generalisation performance of CNNs trained with internal noise plus an information bottleneck, for ReLU and Sigmoid units. They correspond to the results depicted in Figures 7 and A.2.

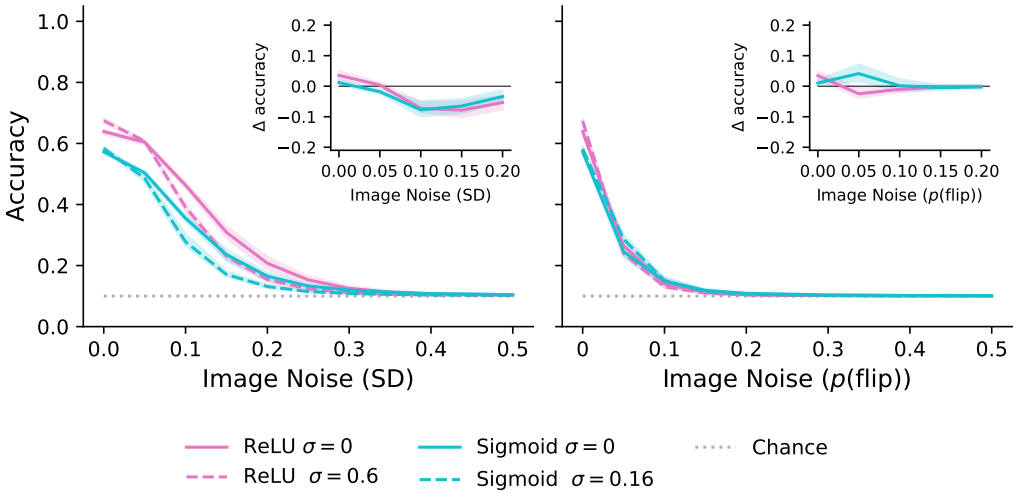


Figure C.1: `small-alexnet` models trained with internal noise and a bottleneck (dashed lines) do not generalise to novel distorted images better than a baseline model (solid lines). **Left**: both networks with ReLU (pink) and Sigmoid (teal) activations perform worse than the baseline on uniform input noise.**Right**: There is a slight improvement for sigmoid networks with constraints (teal), but not ReLU (pink) over the baseline performance for salt-and-pepper input noise.

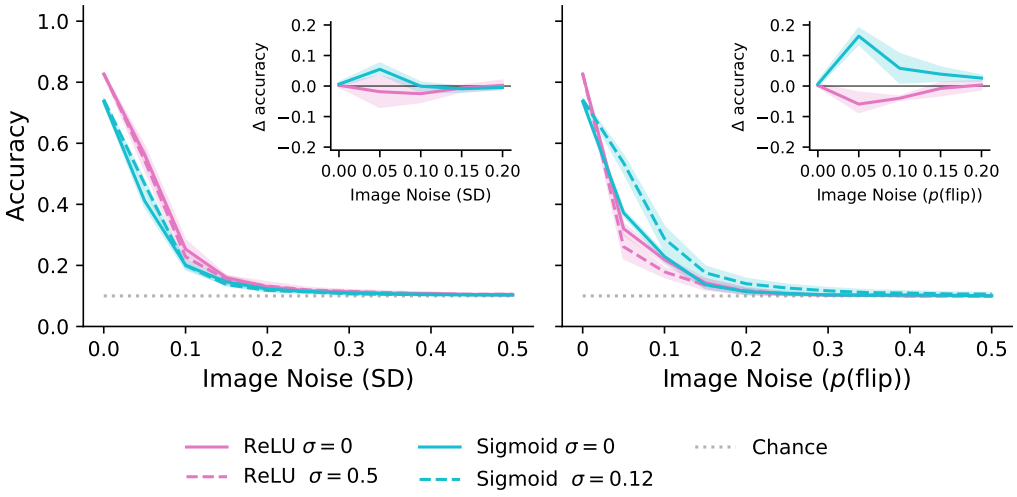


Figure C.2: `small-inception` models trained with internal noise and a bottleneck (dashed lines) showed improved generalisation to novel distorted images over the baseline models (solid lines) for networks with Sigmoid units (teal), but not ReLU (pink). The effect was larger for images with salt and pepper noise (**right**) than for images with uniform noise (**left**). The average deviation in accuracy from the baseline model is highlighted in the inset, with values above the mid-line indicating an improvement in performance.