# Adding biological constraints to deep neural networks reduces their capacity to learn unstructured data

**Christian Tsvetkov, Gaurav Malhotra, Benjamin D. Evans, Jeffrey S. Bowers**
{christian.tsvetkov, gaurav.malhotra, benjamin.evans, jeffrey.bowers}@bristol.ac.uk
School of Psychological Science, University of Bristol, 12a Priory Road, Bristol BS8 1TU, UK

## Abstract

Deep neural networks (DNNs) are becoming increasingly popular as a model of the human visual system. However, they show behaviours that are uncharacteristic of humans, including the ability to learn arbitrary data, such as images with pixel values drawn randomly from a Gaussian distribution. We investigated whether this behaviour is due to the learning and memory capacity of DNNs being too high for the training task. We reduced the capacity of DNNs by incorporating biologically motivated constraints – an information bottleneck, internal noise and sigmoid activations – in order to diminish the learning of arbitrary data, without significantly degrading performance on natural images. Internal noise reliably produced the desired behaviour, while a bottleneck had limited impact. Combining all three constraints yielded an even greater reduction in learning capacity. Furthermore, we tested whether these constraints contribute to a network's ability to generalize by helping it develop more robust internal representations. However, none of the methods could consistently improve generalization.

**Keywords:** deep learning; biological details; memorization; generalization; internal noise; bottleneck

## Introduction

Not only do deep neural networks (DNNs) perform impressively across a range of visual tasks (He, Zhang, Ren, & Sun, 2015; Schroff, Kalenichenko, & Philbin, 2015), they are also becoming increasingly popular among vision researchers as models of the primate visual system (Kriegeskorte, 2015). This is due to the inspiration these networks draw from the architecture of the primate brain, combined with multiple studies showing that deep neural networks trained to classify images achieve high scores in predicting neural activity of primates exposed to the same stimuli (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014). However, DNNs also display some behaviours which are both undesirable in real-world applications, and also highly uncharacteristic of human behavior. Examples include adversarial images (Szegedy et al., 2013) – artificially created stimuli that networks misclassify in ways that humans never would.

Another example of strange behaviour in neural networks is their ability to learn any arbitrary data. Zhang, Bengio, Hardt, Recht, and Vinyals (2017) trained DNN models on a standard classification task, but used either synthetic, 'random' images, with pixel values randomly drawn from a Gaussian distribution, or natural images, where each example is arbitrarily paired with a random output label. The authors point out that in both conditions there are no class-specific patterns, yet the networks learn the data perfectly. They conclude that the networks must be memorizing every single example in order to achieve perfect accuracy. This memorization of unstructured data is in stark contrast with how humans and primates learn to categorize objects. Learning to classify a data set of 50,000 random 'noisy' images is out of reach for human observers. The human visual system is specialized at detecting structural regularities in the environment (Witkin & Tenenbaum, 1983), whereas noise images represent the complete opposite – unstructured data, with no correlations between pixels or examples. While the DNNs in Zhang et al. (2017) do take longer to learn arbitrary data compared to natural stimuli, the amount of extra training required is a single-digit scaling factor away from the number of steps required to learn a naturalistic data set, such as CIFAR10 (Krizhevsky, Nair, & Hinton, 2009).

We explore whether the ability of DNNs to learn arbitrary data sets is due to their overly large capacity relative to the tasks they're trained on. For a model to be seriously considered for investigating biological vision it should be able to perform the tasks humans and other animals can do, while at the same time failing when animals can't succeed either. The aim of this investigation is to bring deep neural networks' and humans' behavior closer together by reducing the capacity of DNNs to learn arbitrary data, without hindering categorization performance on natural images. To accomplish this, we introduce three biologically motivated constraints, which are either inspired by the primate brain, or use mechanisms analogous to observed biological phenomena.

First, we consider the idea of an information bottleneck (Tishby, Pereira, & Bialek, 2000) – the possibility that, by reducing the amount of information passed from some layer in the neural network to the next, (for example by narrowing the channel capacity), only essential information would be conveyed and any irrelevant details would be discarded. When regular patterns shared by all category members can be discovered, this could lead to developing more robust representations. Information bottlenecks could, in fact, play a pivotal role in reducing stimulus complexity to allow efficient processing in the perceptual system (Essen, Olshausen, Anderson, & Gallant, 1991). On the other hand, if no common pattern can be detected, like in the case of random pixel images, then a bottleneck might have a larger impact on network performance. Lindsey, Ocko, Ganguli, and Deny (2019) use the theory of information bottlenecks to show that reducing the

number of convolutional filters in the early layers of a convolutional neural network can create representations similar to the receptive fields found in the human retina and primary visual cortex.

Another biologically inspired intervention we implement is the concept of internal noise, also referred to as 'neural variability'. Computations in the brain are not fault-proof and it is a long-standing observation that presentation of the same stimulus can often result in different neural response patterns (Stein, Gossen, & Jones, 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 by small variation in neurotransmitter release, which can have a significant net effect on the behavior of post-synaptic cells (Stein et al., 2005). Adding internal Gaussian noise to the activation values of hidden units in a neural network is also a known regularization technique in machine learning.

Finally, we consider the role of the activation function on a neural network's capacity. Typically, modern deep learning models use rectified linear (ReLu) units, the activity of which is thresholded at zero for negative values and increases linearly for positive ones. We contrast this with the Sigmoid activation function, which is often encountered in older, connectionist models of cognitive processes, as well as many neurophysiological models of neural population dynamics (Wilson & Cowan, 1972). While there is an ongoing debate about the biological plausibility of Sigmoid versus ReLu activation functions (see Glorot, Bordes, & Bengio, 2011), we focus on their relative capacities for representing information. The activation of rectified 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 $0-1$ and therefore provide a natural constraint on the representational capacity of these units.

Our results suggest that the considered mechanisms modulate the learning of unstructured data by lowering the networks' capacity to memorize. In order to gain further insight into how the constraints affect the internal states of the network, we investigated whether models with constrained capacity learn more robust, invariant representations. We evaluated generalization performance on a modified data set, created using image manipulations which have not been observed during training. Geirhos et al. (2018) have previously shown that humans exhibit a far greater robustness to such manipulations compared to deep convolutional neural networks. This task adds another challenge for the constrained deep neural networks to match human behaviour.

# Methods

## Models and training procedure

All experiments were conducted using two model architectures, `small-inception` and `small-alexnet`, adapted from Zhang et al. (2017)[1]. Both are convolutional neural networks, with a similar number of trainable parameters, differing primarily in the implementation of inception modules and a greatly increased network depth in `small-inception`.

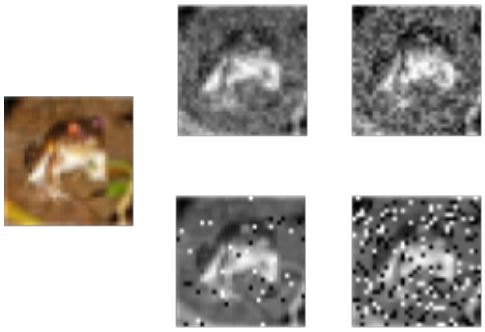


Figure 1: Examples of manipulated images from Experiment 2. Top: uniform noise; Bottom: salt and pepper noise

We trained models with a varying amount of internal Gaussian noise, with or without a bottleneck, and using either ReLU or sigmoid hidden unit activation functions. All experiments use four realizations with different random seeds for each combination of these hyperparameters to improve robustness and generalization of the results.

All networks were implemented and trained using the `tensorflow.keras` library for python[2]. Models were trained for 100 epochs using a batch size of 128. The optimizer used for training was Stochastic Gradient Descent (SGD), with an initial learning rate of 0.1 for `small-inception` and 0.01 for `small-alexnet` and scheduled learning rate decay rate of 0.05 every epoch.

## Materials

The networks were trained for image classification using either the CIFAR10 (Krizhevsky et al., 2009) image data set, or on a random-image data set. CIFAR10 consists of 10 object categories, with 5,000 training and 1,000 test examples per category, for a total of 50,000 stimuli. Unless otherwise stated, we use a matching number of classes and examples of random data in all experiments. The random-image data was created by sampling pixel values from a Gaussian distribution with a mean and standard deviation matching those of CIFAR10, calculated independently for each channel.

## Capacity manipulations

Internal noise was implemented using Gaussian noise regularization from the `tensorflow.keras` module. Internal noise was added to the output of the activation function of every convolutional or fully-connected layer in each neural network, except for the output layer. The standard deviation of the noise was varied from 0 to 1.2, with an increment of 0.05 for networks using ReLU activation units, and between

[1]Please consult source for further details about the architectures
[2]TensorFlow 2.0 for Python 3.6

0 and 0.2, incremented by 0.02 for sigmoid networks. The lower values of noise for the sigmoid networks is due to the upper bound of the range of the function being 1.

Following Lindsey et al. (2019), we incorporated an information bottleneck by reducing the number of convolutional filters in the first convolutional layer of each network (immediately following the input) as much as possible without impairing performance on the CIFAR10 data set ($96 \rightarrow 2$ for `small-inception` and $200 \rightarrow 8$ for `small-alexnet`). While this manipulation reduces the total number of trainable parameters only minimally, 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 layer 1 and a prevalence of Gabor-like filters in subsequent layers. We hypothesize such changes would be detrimental for unstructured data, which may require processing more lower-level, idiosyncratic features, possibly on the level of individual pixels.

### Generalization tests

We applied several types of image distortion to either the training or test set, derived from Geirhos et al. (2018)[3]. Specifically, we focus on uniform noise and salt-and-pepper noise. Although there are other available manipulations in this problem set, we consider these two the most challenging problems, as they are the ones for which the deep neural networks tested by Geirhos et al. (2018) exhibit the largest drop in performance, and also those which diverge most from human results. The variance of the uniform noise and the probability value of the salt-and-pepper noise were varied systematically between 0 and 0.5, with an increment of 0.05.

## Results

### Experiment 1

In the first experiment, we trained each network using a bottleneck, internal noise, or a combination of both, on either CIFAR10 or the random-image data and assess their categorization performance. Figure 2 summarizes the training accuracy for each model. Only data from the `small-inception` models are displayed, as the results were comparable for `small-alexnet` architectures. Test accuracy is also included for the internal noise condition.

The bottleneck models do not show any diminished learning capacity for random data – they fit the random data with perfect accuracy (see light versus dark bars in Figure 2 at noise level 0.0). On the other hand, the noise manipulation was effective. As we increased the variance of the internal noise, the training accuracy of the random image model started decreasing, showing that the network failed to learn unstructured data. Crucially, the performance for structured (CIFAR10) data showed only a minor drop for values of internal noise which reduced learning on unstructured data to chance levels (Figure 2 - Left). Finaly, changing the activation function of the networks to sigmoid does not appear, by itself, to diminish the networks' ability to learn random data (Figure 2 - Bottom).

We also examined whether the network's capacity for learning unstructured information can be further limited by combining a bottleneck with internal noise. In order to do this, we considered a network with a bottleneck as well as a 'moderate' ($\sigma = 0.2$ for ReLU, $\sigma = 0.04$ for sigmoid) or a 'high' ($\sigma = 0.4$; $\sigma = 0.1$ ) level of internal noise. The results are shown in the (light versus dark) bar plots on the right of Figure 2. We observed that a combination of the two constraints is indeed more effective at diminishing random data learning compared to the noise condition alone, especially at a 'moderate' level of internal noise.

### Experiment 2

One interpretation of the above findings is that, as intended, injecting internal noise reduced the memory capacity of the networks such that they were unable to learn the random images. Another possibility, however, is that the internal noise selectively impairs the ability of the network to learn noise-like patterns, and that the capacity of the networks for other types of inputs, structured or not, remains unaffected.

Therefore, to investigate the effect of injecting internal noise further, we trained multiple models with a large internal noise (greater than the threshold value of 0.6, see Figure 2) to categorize random images or CIFAR10 stimuli, while varying the number of total classes (from 10 to 2), as well as the number of images seen in each class (from 1000 to 10). If internal noise works by reducing the learning capacity rather than impairing the learning of unstructured information, we should observe improvement in training accuracy for both the CIFAR10 and noise-like patterns when there are fewer categories and examples per category.

Figure 3 illustrates results for `small-alexnet` based models trained with a large internal noise. It can be seen from this Figure that training accuracy improves from chance levels when the network was trained on a dataset with 5000 examples/category and 10 categories (see Figure 2) to above chance levels when the number of examples or categories decreases. Furthermore, decreasing the number of exemplars or classes steadily improves performance. We also noted that this also led to an improvement in performance for the structured (CIFAR10) dataset, but there was less variability in the model accuracy of CIFAR10-trained models, which perform very well throughout the conditions. Thus, these results lend support to the hypothesis that injecting internal noise decreases the learning capacity of networks rather than prevent them from learning unstructured information.

### Experiment 3

Our results from Experiment 1 show that the internal noise models do not affect natural and random images in the same way. Further, Experiment 2 suggest that the manipulations

[3]Code adapted from https://github.com/rgeirhos/generalisation-humans-DNNs

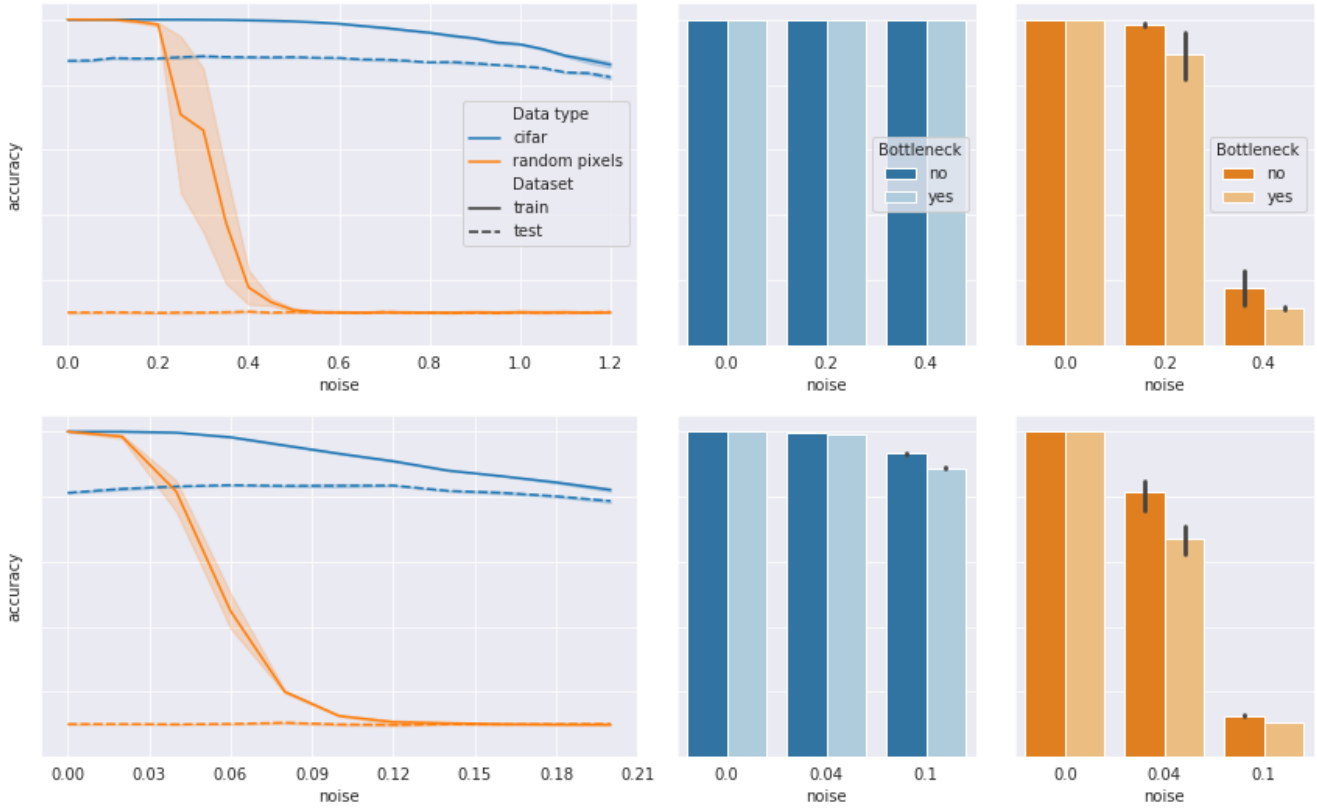



Figure 2: Accuracy for `small-inception` models with biological constraints. **Top**: ReLU activation function. **Bottom**: Sigmoid activation function. The x-axis represents the variance of internal Gaussian noise. The origin represents models without internal noise. On the left, lines visualize the training (solid) and test (dashed) data accuracy of models with internal noise on the CIFAR10 data set or a random image data set. On the right, bars show performance for models with (light bars) or without (dark bars) a bottleneck for various values of internal noise.

decrease the networks' capacity. The disproportionate impact could be an indicator that, even though CIFAR10 categorization performance does not diminish, there are changes to the internal representations the models are learning. We hypothesize that the constrained capacity of the models could be affecting what convolutional filters are learned, possibly by focusing on more robust, invariant features. To investigate, we test the generalization performance of the networks trained with internal noise, with and without a bottleneck, on images from the test set modified by out-of-domain manipulations (examples in Figure 1).

It can be argued that a more direct way of investigating changes to the models' internal representations would be to visualize the features learned by neurons in each channel of the convolutional layers by various techniques (Yosinski, Clune, Nguyen, Fuchs, & Lipson, 2015). We acknowledge that such methods can be very useful for understanding neural networks. However, they are not suited to answering questions about the qualitative differences between representations. Knowing the details of differences in two models' learned representations does not necessarily tell us which model, if any, would exhibit better generalization. Our implicit measure of robustness is better equipped to answer such questions.

All models in this experiment were trained on the CIFAR10 data set. The networks with ReLU activation do not appear to benefit from the noise training in terms of their generalization performance (Figure 4 - Top). Overall accuracy is similar across all tested ReLU models, but the baseline network, trained without any internal noise, performs the best. The sigmoid networks, on the other hand, show a slight improvement in accuracy on the modified test images in the salt-and-pepper noise condition (Figure 4 - Bottom). The models with a moderate amount of internal noise are better at classifying the degraded images at low levels of input noise. However, the effect is not maintained for higher amounts of input noise. This advantage is furthered in the internal noise and bottleneck condition (Figure 4 b) - Bottom), with internal noise models performing better than baseline, for both uniform and salt-and-pepper.

## Discussion

The proposed capacity constraints produced results with mixed efficacy with respect to the goal of the project. Adding

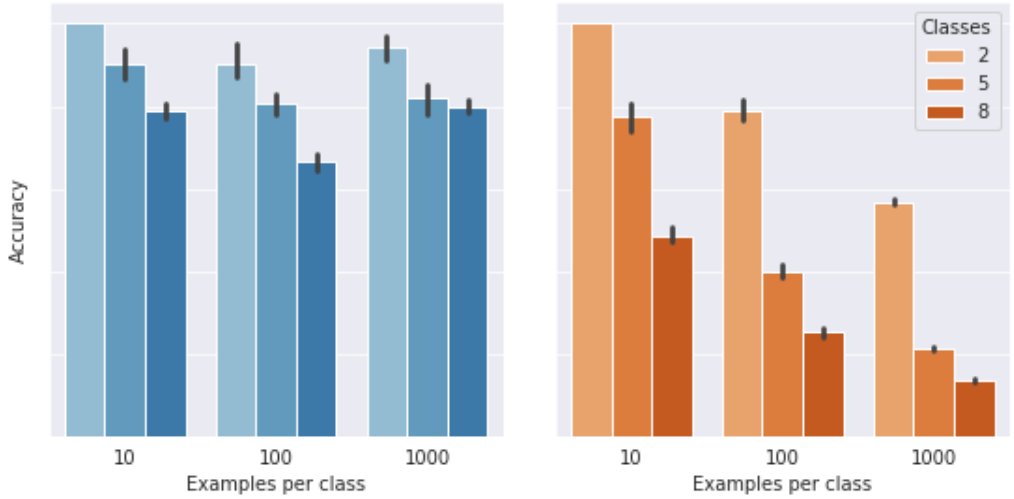



Figure 3: Training accuracy of models using different numbers of classes and examples per class, using internal noise of $\sigma = 0.8$ (higher than threshold from Experiment 1). Darker bars represent a greater number of classes. Left: CIFAR10 accuracy. Right: Random image accuracy.

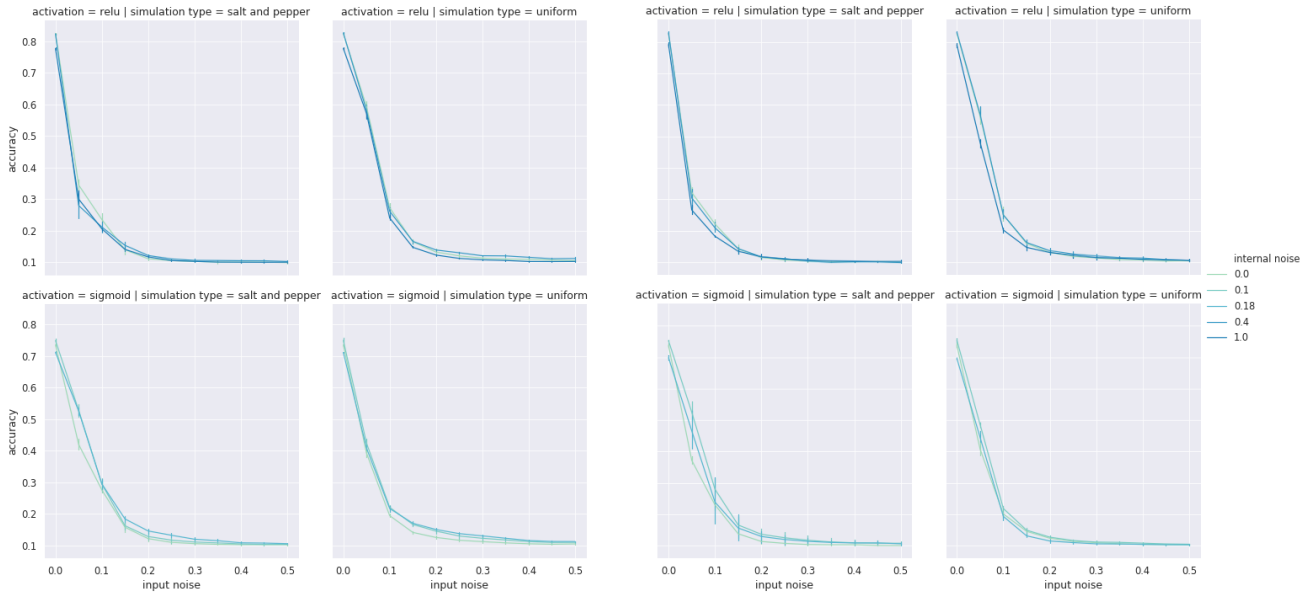
internal noise to hidden layer activation values achieved the desired effect of reducing the ability of the models to learn random images, while retaining a nearly unchanged training and generalization performance on natural images. On the other hand, the bottleneck intervention turned out to not be effective as a way of constraining the neural networks' learning capacity by itself, in any of the conditions. Nevertheless, the combined constraints condition showed that it did have a net effect on learning capacity. One explanation could be that the implementation of the information bottleneck was not adequate for the task. It is difficult to compute just how severe the constraint is as the unit activations are not only unbounded in the case of ReLU networks, but also have a large precision, since they are represented as 32-bit floating point numbers. This makes estimating the channel capacity of the bottleneck in bits difficult, since it is uncertain what level of detail the models require for discrimination (does a difference of $10^{-5}$ matter for classification?) Future work could narrow down the possible values by introducing precision-limiting measures such as rounding the activation values or using a binary activation function.

Arplt et al. (2017) have conducted a similar experiment, using several kinds of model regularization – techniques that reduced overfitting a model to the training data – in order to reduce 'memorization' of random stimuli without impacting test performance on a natural image data set. Interestingly, our results from Experiment 1 are at odds some of the data reported in that study. Specifically, Arplt et al. (2017) also use internal Gaussian noise regularization, but conclude it is not effective in decreasing memorization. There are some important differences between the two studies. First, while we focus on the random images task from Zhang et al. (2017), Arplt et al. (2017) chose to work with the random label condition instead. Secondly – and critically – the values for the variance of the Gaussian noise they explore are too low to noticably decrease memorization capacity. That behavior becomes most prominent at values around twice the maximum range considered by Arplt et al. (2017). Finally, they report a reduction in test data accuracy on the CIFAR10 dataset, which we do not observe in our experiments. Further efforts will be needed to establish why the results of the two studies diverge and whether they would hold in the context of a direct replication. Preliminary findings support the robustness of the data.

Further work will be required to explain how and why the sigmoid models show some improvement to out-of-domain generalization to manipulated images, albeit modest, while networks with ReLU activations do not. The internal representations of both models could be analyzed and compared in other ways, for instance by visualizing the receptive fields of hidden units in different layers and inspecting the differences in the patterns the units have specialized to detect.

Our results establish that, at least in the context of the architectures and data sets studied, DNNs can endure significant constraints while maintaining their testing performance on natural images – highlighting that they otherwise operate well over the required capacity. In this context, it is interesting to consider examples of deep learning models which manage to match or surpass human performance (He et al., 2015; Schroff et al., 2015). Would these models be able to perform at the same human-like levels if their learning capacity is controlled to match other human constraints?

While in Experiment 3 we focused only on two forms of image manipulations to study the generalization behavior of constrained models, other focused manipulations could also offer insights into the kinds of features models are sensi-




(a) Without bottleneck

(b) With bottleneck

Figure 4: Generalization performance on out-of-domain image manipulations for combinations of constraints. Left: Uniform image noise. Right: Salt-and-pepper noise.

tive to. For example, it has been proposed that convolutional neural networks already focus too much on spatially high-frequency information such as texture (Geirhos et al., 2019). Using manipulations such as a low-pass filter could uncover whether models with capacity constraints are more prone to attune to spatially low-frequency information than unconstrained models.

## Conclusion

We demonstrate that biologically inspired mechanisms can be effective at reducing the capacity of deep neural networks. The resulting models are more consistent with human behaviour, as they are less capable of learning arbitrary inputs such as random noise images. Further work is necessary to determine how a reduced capacity influences internal representations. Results from Experiment 3 suggest that more severe constraints, such as combining internal noise, sigmoid activations and a bottleneck, show modest improvements in generalization to unobserved image manipulations. The mechanisms of this behaviour are yet to be elucidated.

## Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, Grant ID 741134.

## References

Arplt, D., Jastrzbskl, S., Bailas, N., Krueger, D., Bengio, E., Kanwal, M. S., ... Lacoste-Julien, S. (2017). A closer look at memorization in deep networks. In *34th international conference on machine learning, icml 2017* (Vol. 1, pp. 350–359). Retrieved from https://arxiv.org/pdf/1706.05394.pdf

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* (Vol. 1473, pp. 17 – 28). SPIE. doi: 10.1117/12.45537

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.* Retrieved from 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 31* (pp. 7538–7550). Curran Associates, Inc.

Glorot, X., Bordes, A., & Bengio, Y. (2011, 11–13 Apr). Deep sparse rectifier neural networks. In G. Gordon, D. Dunson, & M. Dudík (Eds.), *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (Vol. 15, pp. 315–323). Fort Lauderdale, FL, USA: PMLR.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep residual*

*learning for image recognition.*

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014, 11). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLOS Computational Biology*, *10*(11), 1-29. doi: 10.1371/journal.pcbi.1003915

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*(1), 417-446. (PMID: 28532370) doi: 10.1146/annurev-vision-082114-035447

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

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.* Retrieved from `https://openreview.net/forum?id=S1xq3oR5tQ`

Schroff, F., Kalenichenko, D., & Philbin, J. (2015, Jun). Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi: 10.1109/cvpr.2015.7298682

Stein, R. B., Gossen, E. R., & Jones, K. E. (2005). Neuronal variability: noise or part of the signal? *Nature Reviews Neuroscience*, *6*(5), 389–397. doi: 10.1038/nrn1668

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). *Intriguing properties of neural networks.*

Tishby, N., Pereira, F. C., & Bialek, W. (2000). *The information bottleneck method.*

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

Witkin, A. P., & Tenenbaum, J. M. (1983). On the role of structure in vision. In J. Beck, B. Hope, & A. Rosenfeld (Eds.), *Human and machine vision* (p. 481 - 543). Academic Press. doi: https://doi.org/10.1016/B978-0-12-084320-6.50022-0

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*(23), 8619–8624. doi: 10.1073/pnas.1403112111

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). *Understanding neural networks through deep visualization.*

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.* Retrieved from `https://openreview.net/forum?id=Sy8gdB9xx`