Generalisation in humans and deep neural networks

Robert Geirhos\textsuperscript{1,3,5} Carlos R. Medina Temme\textsuperscript{1}\* Jonas Rauber\textsuperscript{2,3}\*
Heiko H. Schütt\textsuperscript{1,4,5} Matthias Bethge\textsuperscript{2,6,7}\* Felix A. Wichmann\textsuperscript{1,2,6,8}\*

\textsuperscript{1}Neural Information Processing Group, University of Tübingen
\textsuperscript{2}Centre for Integrative Neuroscience, University of Tübingen
\textsuperscript{3}International Max Planck Research School for Intelligent Systems
\textsuperscript{4}Graduate School of Neural and Behavioural Sciences, University of Tübingen
\textsuperscript{5}Department of Psychology, University of Potsdam
\textsuperscript{6}Bernstein Center for Computational Neuroscience Tübingen
\textsuperscript{7}Max Planck Institute for Biological Cybernetics
\textsuperscript{8}Max Planck Institute for Intelligent Systems
\*Joint first / joint senior authors
\textsuperscript{5}To whom correspondence should be addressed: robert.geirhos@bethgelab.org

Abstract

We compare the robustness of humans and current convolutional deep neural networks (DNNs) on object recognition under twelve different types of image degradations. First, using three well known DNNs (ResNet-152, VGG-19, GoogLeNet) we find the human visual system to be more robust to nearly all of the tested image manipulations, and we observe progressively diverging classification error-patterns between humans and DNNs when the signal gets weaker. Secondly, we show that DNNs trained directly on distorted images consistently surpass human performance on the exact distortion types they were trained on, yet they display extremely poor generalisation abilities when tested on other distortion types. For example, training on salt-and-pepper noise does not imply robustness on uniform white noise and vice versa. Thus, changes in the noise distribution between training and testing constitutes a crucial challenge to deep learning vision systems that can be systematically addressed in a lifelong machine learning approach. Our new dataset consisting of 83K carefully measured human psychophysical trials provide a useful reference for lifelong robustness against image degradations set by the human visual system.

1 Introduction

1.1 Deep neural networks as models of human object recognition

The visual recognition of objects by humans in everyday life is rapid and seemingly effortless, as well as largely independent of viewpoint and object orientation \cite{1}. The rapid and primarily foveal recognition during a single fixation has been termed \textit{core object recognition} (see \cite{2} for a review). We know, for example, that it is possible to reliably identify objects in the central visual field within a single fixation in less than 200 ms when viewing “standard” images \cite{2,4}. Based on the rapidness of object recognition, core object recognition is often thought to be achieved with mainly feedforward processing although feedback connections are ubiquitous in the primate brain. Object recognition
in the primate brain is believed to be realised by the ventral visual pathway, a hierarchical structure consisting of areas V1-V2-V4-IT, with information from the retina reaching the cortex in V1 (e.g. [5]).

Until a few years ago, animate visual systems were the only ones known to be capable of broad-ranging visual object recognition. This has changed, however, with the advent of brain-inspired deep neural networks (DNNs) which, after having been trained on millions of labeled images, achieve human-level performance when classifying objects in images of natural scenes [6]. DNNs are now employed on a variety of tasks and set the new state-of-the-art, sometimes even surpassing human performance on tasks which only a few years ago were thought to be beyond an algorithmic solution for decades to come [7,8]. Since DNNs and humans achieve similar accuracy, a number of studies have started investigating similarities and differences between DNNs and human vision [9–24]. On the one hand, the network units are an enormous simplification given the sophisticated nature and diversity of neurons in the brain [25]. On the other hand, often the strength of a model lies not in replicating the original system but rather in its ability to capture the important aspects while abstracting from details of the implementation (e.g. [26,27]).

One of the most remarkable properties of the human visual system is its ability to generalise robustly. Humans generalise across a wide variety of changes in the input distribution, such as across different illumination conditions and weather types. For instance, human object recognition is largely unimpaired even if there are rain drops or snow flakes in front of an object. While humans are certainly exposed to a large number of such changes during their preceding lifetime (i.e., at “training time”, as we would say for DNNs), there seems to be something very generic about the way the human visual system is able to generalise that is not limited to the same distribution one was exposed to previously. Otherwise we would not be able to make sense of a scene if there was some sort of “new”, previously unseen noise. Even if one never had a shower of confetti before, one is still able to effortlessly recognise objects at a carnival parade. Naturally, such generic, robust mechanisms are not only desirable for animate visual systems but also for solving virtually any visual task that goes beyond a well-confined setting where one knows the exact test distribution already at training time. Deep learning for autonomous driving may be one prominent example: one would like to achieve robust classification performance in the presence of confetti, despite not having had any confetti exposure during training time. Thus, from a machine learning perspective, general noise robustness can be used as a highly relevant example of lifelong machine learning [28] requiring generalisation that does not rely on the standard assumption of independent, identically distributed (i.i.d.) samples at test time.

### 1.2 Comparing generalisation abilities

Generalisation in DNNs usually works surprisingly well: First of all, DNNs are able to learn sufficiently general features on the training distribution to achieve a high accuracy on the i.i.d. test distribution despite having sufficient capacity to completely memorise the training data [29], and
considerable effort has been devoted to understand this phenomenon (e.g. [30–32]). Secondly, features learned on one task often transfer to only loosely related tasks, such as from classification to saliency prediction [33], emotion recognition [34], medical imaging [35] and a large number of other transfer learning tasks [36]. However, transfer learning still requires a substantial amount of training before it works on the new task. Here, we focus on a third setting that adopts the lifelong machine learning point of view of generalisation [37]: How well can a visual learning system cope with a new image degradation after it has learned to cope with a certain set of image distortions before. As a measure of object recognition robustness we can test the ability of a classifier or visual system to tolerate changes in the input distribution up to a certain degree, i.e., to achieve high recognition performance despite being evaluated on a test distribution that differs to some degree from the training distribution (testing under realistic, non-i.i.d. conditions). Using this approach we measure how well DNNs and human observers cope with parametric image manipulations that gradually distort the original image.

First, we assess how top-performing DNNs that are trained on ImageNet, GoogLeNet [38], VGG-19 [39] and ResNet-152 [40], compare against human observers when tested on twelve different distortions such as additive noise or phase noise (see Figure 2 for an overview)—in other words, how well do they generalise towards previously unseen distortions? In a second set of experiments, we train networks directly on distorted images to see how well they can in general cope with noisy input, and how much training on distortions as a form of data augmentation helps in dealing with other distortions. Psychophysical investigations of human behaviour on object recognition tasks, measuring accuracies depending on image colour (grayscale vs. colour), image contrast and the amount of additive visual noise have been powerful means of exploring the human visual system, revealing much about the internal computations and mechanisms at work [42–48]. As a consequence, similar experiments might yield equally interesting insights into the functioning of DNNs, especially as a comparison to high-quality measurements of human behaviour. In particular, human data for our experiments were obtained using a controlled lab environment (instead of e.g. Amazon Mechanical Turk without sufficient control about presentation times, display calibration, viewing angles, and sustained attention of participants). Our carefully measured behavioural datasets—twelve experiments encompassing a total number of 82,880 psychophysical trials—as well as materials and code are available online at [https://github.com/rgeirhos/generalisation-humans-DNNs](https://github.com/rgeirhos/generalisation-humans-DNNs).

2 Methods

We here report the core elements of employed paradigm, procedure, image manipulations, observers and DNNs; this is aimed at giving the reader just enough information to understand experiments and results. For in-depth explanations we kindly refer to the comprehensive supplementary material, which seeks to provide exhaustive and reproducible experimental details.

2.1 Paradigm, procedure & 16-class-ImageNet

For this study, we developed an experimental paradigm aimed at comparing human observers and DNNs as fair as possible by using a forced-choice image categorisation task. Achieving a fair psychophysical comparison comes with a number of challenges: First of all, many high-performing DNNs are trained on the ILSRVR 2012 database [50] with 1,000 fine-grained categories (e.g., over a hundred different dog breeds). If humans are asked to name objects, however, they most naturally categorise them into so-called entry-level categories (e.g., dog rather than German shepherd). We thus developed a mapping from 16 entry-level categories such as dog, car or chair to their corresponding ImageNet categories using the WordNet hierarchy [51]. We term this dataset “16-class-ImageNet” since it groups a subset of ImageNet classes into 16 entry-level categories (airplane, bicycle, boat, car, chair, dog, keyboard, oven, bear, bird, bottle, cat, clock, elephant, knife, truck). In every experiment, then, an image was presented on a computer screen and observers had to choose the correct category by clicking on one of these 16 categories. For pre-trained DNNs, the sum of all softmax values mapping to

---

1 Still, DNNs usually need orders of magnitude more training data in comparison to humans, as explored by the literature on one-shot or few-shot learning (see e.g. [23] for an overview).
2 We have reported a subset of these experiments on arXiv in an earlier version of this paper [41].
3 This is the same paradigm as reported in [49].
a certain entry-level category was computed. The entry-level category with the highest sum was then taken as the network’s decision. A second challenge is the fact that standard DNNs only use feedforward computations at inference time, while recurrent connections are ubiquitous in the human brain [52, 53]. In order to prevent this discrepancy from playing a major confounding role in our experimental comparison, presentation time for human observers was limited to 200 ms. An image was immediately followed by a 200 ms presentation of a noise mask with $1/f$ spectrum, known to minimise, as much as psychophysically possible, feedback influence in the brain.

2.2 Observers & pre-trained deep neural networks

Data from human observers were compared against classification performance of three pre-trained DNNs: VGG-19 [39], GoogLeNet [38] and ResNet-152 [40]. For each of the twelve experiments that were conducted, either five or six observers participated (with the exception of the colour experiment, for which only three observers participated since similar experiments had already been performed by a number of studies [48, 55, 56]). Observers reported normal or corrected-to-normal vision.

2.3 Image manipulations

A total of twelve experiments were performed in a well-controlled psychophysical lab setting. In every experiment, a (possibly parametric) distortion was applied to a large number of images, such that the signal strength ranged from ‘no distortion / full signal’ to ‘distorted / weak(er) signal’. We then measured how classification accuracy changed as a function of signal strength. Three of the employed image manipulations were dichotomous (colour vs. greyscale, true vs. opponent colour, original vs. equalised power spectrum); one manipulation had four different levels (0, 90, 180 and 270 degrees of rotation); one had seven levels (0, 30, ..., 180 degrees of phase noise) and the other distortions had eight different levels. Those manipulations were: uniform noise, controlled by the ‘width’ parameter indicating the bounds of pixel-wise additive uniform noise; low-pass filtering and high-pass filtering (with different standard deviations of a Gaussian filter); contrast reduction (contrast levels from 100% to 1%) as well as three different manipulations from the eidolon toolbox [57]. The three eidolon experiments correspond to different versions of a parametric image manipulation, with the ‘reach’ parameter controlling the strength of the distortion. Additionally, for experiments with training on distortions, we also evaluated performance on stimuli with salt-and-pepper noise (controlled by parameter $p$ indicating probability of setting a pixel to either black or white; $p \in [0, 10, 20, 35, 50, 65, 80, 95\%]$). More information about the different image manipulations is provided in the supplementary material (Section Image preprocessing and distortions), where we also show example images across all manipulations and distortion levels (Figures 10, 11, 12, 13, 14). For a brief overview, Figure 2 depicts one exemplary manipulation per distortion. Overall, the manipulations we used were chosen to reflect a large variety of possible distortions.

---

*But see e.g. [54] for a critical assessment of this argument.*
Figure 3: Classification accuracy and response distribution entropy for GoogLeNet, VGG-19 and ResNet-152 as well as for human observers. ‘Entropy’ indicates the Shannon entropy of the response/decision distribution (16 classes). It here is a measure of bias towards certain categories: using a test dataset that is balanced with respect to the number of images per category, responding equally frequently with all 16 categories elicits the maximum possible entropy of four bits. If a network or observer responds prefers some categories over others, entropy decreases (down to zero bits in the extreme case of responding with one particular category all the time, irrespective of the ground truth category). Human ‘error bars’ indicate the full range of results across participants. Image manipulations are explained in Section 2.3 and visualised in Figures 10, 11, 12, 13 and 14.
2.4 Training on distortions

Beyond evaluating standard pre-trained DNNs on distortions (results reported in Figure 3), we also trained networks directly on distortions (Figure 4). These networks were trained on 16-class-ImageNet, a subset of the standard ImageNet dataset as described in Section 2.1. This reduced the size of the unperturbed training set to approximately one fifth. To correct for the highly imbalanced number of samples per class, we weighted each sample in the loss function with a weight proportional to one over the number of samples of the corresponding class. All networks trained in these experiments had a ResNet-like architecture that differed from a standard ResNet-50 only in the number of output neurons that we reduced from 1000 to 16 to match the 16 entry-level classes of the dataset. Weights were initialised with a truncated normal distribution with zero mean and a standard deviation of $\sqrt{\frac{1}{n}}$ where $n$ is the number of output neurons in a layer. While training from scratch, we performed on-the-fly data augmentation using different combinations of the image manipulations. When training a network on multiple types of image manipulations (models B1 to B9 as well as C1 and C2 of Figure 4), the type of manipulation (including unperturbed, i.e. standard colour images if applicable) was drawn uniformly and we only applied one manipulation at a time (i.e., the network never saw a single image perturbed with multiple image manipulations simultaneously, except that some image manipulations did include other manipulations per construction: uniform noise, for example, was always added after conversion to greyscale and contrast reduction to 30%). For a given image manipulation, the amount of perturbation was drawn uniformly from the levels used during test time (cf. Figure 3). The remaining aspects of the training followed standard training procedures for training a ResNet on ImageNet: we used SGD with a momentum of 0.997, a batch size of 64, and an initial learning rate of 0.025. The learning rate was multiplied with 0.1 after 30, 60, 80 and 90 epochs (when training for 100 epochs) or 60, 120, 160 and 180 epochs (when training for 200 epochs). Training was done using TensorFlow 1.6.0 [58]. In the training experiments, all manipulations with more than two levels were included except for the eidolon stimuli, since the generation of those stimuli is computationally too slow for ImageNet training. For comparison purposes, we additionally included colour vs. greyscale as well as salt-and-pepper noise (for which there is no human data, but informal comparisons between uniform noise and salt-and-pepper noise strongly suggest that human performance will be similar, see Figure 1c).

3 Generalisation of humans and pre-trained DNNs towards distortions

In order to assess generalisation performance when the signal gets weaker, we tested twelve different ways of degrading images. These images at various levels of signal strength were then shown to both human observers in a lab and to pre-trained DNNs (ResNet-152, GoogLeNet and VGG-19) for classification. The results of this comparison are visualised in Figure 5. While human and DNN performance was similar for comparatively minor colour-related distortions such as conversion to greyscale or opponent colours, we find human observers to be more robust for all of the other distortions: by a small margin for low contrast, power equalisation and phase noise images and by a larger margin for uniform noise, low-pass, high-pass, rotation and all three eidolon experiments. Furthermore, there are strong differences in the error patterns as measured by the response distribution entropy (indicating biases towards certain categories). Human participants’ responses were distributed more or less equally amongst the 16 classes, whereas all three DNNs show increasing biases towards certain categories when the signal gets weaker. These biases are not completely explained by the prior class probabilities, and deviate from distortion to distortion. For instance, ResNet-152 almost solely predicts class bottle for images with strong uniform noise (irrespective of the ground truth category)5 and classes dog or bird for images distorted by phase noise. One might think of simple tricks to reduce the discrepancy between the response distribution entropy of DNNs and humans. One possible way would be increasing the softmax temperature parameter and assuming that model decisions are sampled from the softmax distribution rather than taking the argmax. However, increasing the response DNN distribution entropy in this way dramatically decreases classification accuracy and thus comes with a trade-off (cf. Figure 8 in the supplementary material).

These results are in line with previous findings reporting human-like processing of chromatic information in DNNs [19] but strong decreases in DNN recognition accuracy for image degradations like

5 A category-level analysis of decision biases for the uniform noise experiment is provided in the supplementary material, Figure 9.
4 Training DNNs directly on distorted images

We trained one network per distortion directly and from scratch on (potentially manipulated) 16-class-ImageNet images. The results of this training are visualised in Figure 4 (models A1 to A9). We find that these specialised networks consistently outperformed human observers, by a large margin, on the image manipulation they were trained on (as indicated by strong network performance on the diagonal). This is a strong indication that currently employed architectures (such as ResNet-50) and training methods (standard optimiser and training procedure) are sufficient to ‘solve’ distortions under i.i.d. train/test conditions. We were able to not only close the human-DNN performance gap that was observed by [13] (who fine-tuned networks on distortions, reporting improved but not human-level DNN performance) but to surpass human performance in this respect. While the human visual system has been exposed to a number of distortions during evolution and lifetime, we clearly had no exposure whatsoever to many of the exact image manipulations that we tested here. Thus, our human data show that a high level of generalisation is, in principle, possible. There may be many different reasons for the discrepancy between human and DNN generalisation performance that we find: Are there limitations in terms of the currently used network architectures (as hypothesised by [60]), which may be inferior to the human brain’s intricate computations? Is it a problem of the training data (as suggested by e.g. [61]), or are today’s training methods / optimisers not sufficient to solve robust and general object recognition? In order to shed light on the dissimilarities we found, we performed a second batch of experiments by training networks directly on distorted images.
cells of models A1 to A9. Overall, we find that training on a certain distortion slightly improves performance on other distortions in a few instances, but is detrimental in other cases (when compared to a vanilla ResNet-50 trained on colour images, model A1 in the figure). Performance on salt-and-pepper noise as well as uniform noise was close to chance level for all networks, even for a network trained directly on the respectively other noise model. This may be surprising given that these two types of noise do not seem very different to a human eye (as indicated in Figure 1c). Hence, training a network on one distortion does not generally lead to improvements on other distortions.

Since training on a single distortion alone does not seem to be sufficient to evoke robust generalisation performance in DNNs, we also trained the same architecture (ResNet-50) on two additional settings. Models B1 to B9 in Figure 4 show performance for training on one particular distortion in combination with uniform noise (training consisted of 50% images from each manipulation). Uniform noise was chosen since it seemed to be one of the hardest distortions for all networks, and hence they might benefit from including this particular distortion in the training data. Furthermore, we trained models C1 and C2 on all but one distortion (either uniform or salt-and-pepper noise was left out).

We find that object recognition performance of models B1 to B9 is improved compared to models A1 to A9, both on the distortions they were actually trained on (diagonal entries with red rectangles in Figure 4) as well as on a few of the distortions that were not part of the training data. However, this improvement may be largely due to the fact that models B1 to B9 were trained on 200 epochs instead of 100 epochs as for models A1 to A9, since the accuracy of model B9 (trained & tested on uniform noise, 200 epochs) also shows an improvement towards model A9 (trained & tested on uniform noise, 100 epochs). Hence, in the presence of heavy distortions, training longer may go a long way but incorporating other distortions in the training does not seem to be generally beneficial to model performance. Furthermore, we find that it is possible even for a single model to reach high accuracies on all of the eight distortions it was trained on (models C1 and C2), however for both left-out uniform and salt-and-pepper noise, object recognition accuracy stayed around 11 to 14%, which is by far closer to chance level (approx. 6%) than to the accuracy reached by a specialised network trained on this exact distortion (above 70%, which serves as a lower bound on the achievable performance).

Taken together, these findings indicate that data augmentation with distortions alone may be insufficient to overcome the generalisation problem that we find. It may be necessary to move from asking “why are DNNs generalising so well (under i.i.d. settings)?” to “why are DNNs generalising so poorly (under non-i.i.d. settings)?”. It is up to future investigations to determine how DNNs that are currently being handled as computational models of human object recognition can solve this challenge. At the exciting interface between cognitive science / visual perception and deep learning, inspiration and ideas may come from both fields: While the computer vision sub-area of domain adaptation (see [63] for a review) is working on robust machine inference in spite of shifts in the input distribution, the human vision community is accumulating strong evidence for the benefits of local gain control mechanisms. These normalisation processes seem to be crucial for many aspects of robust animal and human vision [46], are predictive for human vision data [21, 64] and have proven useful in the context of computer vision [65, 66]. It could be an interesting avenue for future research to determine whether there is a connection between neural normalisation processes and DNN generalisation performance.

5 Conclusion

We conducted a behavioural comparison of human and DNN object recognition robustness against twelve different image distortions. In comparison to human observers, we find the classification performance of three well-known DNNs trained on ImageNet—ResNet-152, GoogLeNet and VGG-19—to decline rapidly with decreasing signal-to-noise ratio under image distortions. Additionally, we find progressively diverging patterns of classification errors between humans and DNNs with weaker signals. Our results, based on 82,880 psychophysical trials under well-controlled lab conditions, demonstrate that there are still marked differences in the way humans and current DNNs process...
object information. These differences, in our setting, cannot be overcome by training on distorted images (i.e., data augmentation): While DNNs cope perfectly well with the exact distortion they were trained on, they still show a strong generalisation failure towards previously unseen distortions. Since the space of possible distortions is literally unlimited (both theoretically and in real-world applications), it is not feasible to train on all of them. DNNs have a generalisation problem when it comes to settings that go beyond the usual (yet often unrealistic) i.i.d. assumption. We believe that solving this generalisation problem will be crucial both for robust machine inference and towards better models of human object recognition, and we envision that our findings as well as our carefully measured and freely available behavioural data may provide a new useful benchmark for improving DNN robustness and a motivation for neuroscientists to identify mechanisms in the brain that may be responsible for this remarkable robustness.

**Author contributions**

The initial project idea of comparing humans against DNNs was developed by F.A.W. and R.G. All authors jointly contributed towards designing the study and interpreting the data. R.G. and C.R.M.T. developed the image manipulations and acquired the behavioural data with input from H.H.S. and F.A.W.; J.R. trained networks on distortions; experimental data and networks were evaluated by C.R.M.T., R.G. and J.R. with input from H.H.S, M.B. and F.A.W.; R.G. and C.R.M.T. worked on making our work reproducible (data, code and materials openly accessible; writing supplementary material); R.G. wrote the paper with significant input from all other authors.

**Acknowledgments**

This work has been funded, in part, by the German Federal Ministry of Education and Research (BMBF) through the Bernstein Computational Neuroscience Program Tübingen (FKZ: 01GQ1002) as well as the German Research Foundation (DFG; Sachbeihilfe Wi 2103/4-1 and SFB 1233 on “Robust Vision”). The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting R.G. and J.R.; J.R. acknowledges support by the Bosch Forschungsstiftung (Stifterverband, T113/30057/17); M.B. acknowledges support by the Centre for Integrative Neuroscience Tübingen (EXC 307) and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DoI/IBC) contract number D16PC00003.

We would like to thank David Janssen for his invaluable contributions in shaping the early stage of this project. Furthermore, we are very grateful to Tom Wallis for providing the MATLAB source code of one of his experiments, and for allowing us to use and modify it; Silke Gramer for administrative and Uli Wannek for technical support, as well as Britta Lewke for the method of creating response icons and Patricia Rubisch for help with testing human observers. Moreover, we would like to thank Nikolaus Kriegeskorte, Jakob Macke and Tom Wallis for helpful feedback, and three anonymous reviewers for constructive suggestions.

**References**


[https://github.com/rgeirhos/generalisation-humans-DNNs](https://github.com/rgeirhos/generalisation-humans-DNNs)


