The Notorious Difficulty of Comparing Human and Machine Perception

Judy Borowski*[1], Christina M. Funke*[1], Karolina Stosio[1, 3],
Wieland Brendel† [1-3], Thomas S. A. Wallis† ‡ [1], Matthias Bethge† [1-3]
* joint first authors, † joint senior authors, firstname.lastname@bethgelab.org
[1]University of Tübingen, Maria-von-Linden-Str. 6, Germany; [2]Werner Reichardt Centre for Integrative Neuroscience; [3]Bernstein Center for Computational Neuroscience. ‡Current affiliation: Amazon Development Center Germany GmbH. The author's contribution to this work was prior to his employment at Amazon.

Abstract

With the rise of machines to human-level performance in complex recognition tasks, a growing amount of work is directed towards comparing information processing in humans and machines. These works have the potential to deepen our understanding of the inner mechanisms of human perception and to improve machine learning. Drawing robust conclusions from comparison studies, however, turns out to be difficult. Here, we present three case studies to highlight common shortcomings that can easily lead to fragile conclusions. These pitfalls include sub-optimal training procedures or architectures that lead to premature claims regarding gaps between human and machine performance, unequal testing procedures that lead to different decision behaviours, and finally human-centred interpretation of results.

Addressing these shortcomings alters the conclusions of previous studies. We show that neural networks can, in fact, solve same-different tasks, that they do experience a "recognition gap" on minimally recognisable to maximally unrecognisable images, and finally, that despite their ability to solve closed-contour tasks, neural networks use different strategies than humans. To counter these three pitfalls, we provide guidelines on how to compare humans and machines in visual inference tasks.

Keywords: neural networks; deep learning; human vision

Introduction

How do biological brains infer environmental states from sensory data? This has been a central question in neuroscience and psychology for over 100 years. In vision, early theorists (Alhazen, 1083; Helmholtz, 1925; Howard, 1996) proposed that perception was a process of unconscious inference: given ambiguous sensory data, we usually perceive a stable and coherent world. While these models have had some notable successes in explaining aspects of visual perception, until recently they were unable to perform most of the natural tasks we effortlessly perform daily.

Advances in computer vision, particularly with respect to convolutional neural networks, have created systems that can now perform these tasks at human levels. For example, deep neural networks can perform complex tasks such as object recognition (Krizhevsky, Sutskever, & Hinton, 2012), saliency prediction (Kümmerer, Theis, & Bethge, 2014), depth estimation (Eigen & Fergus, 2015) and semantic segmentation (Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2017). These successes invite the enticing possibility that we may learn from one system by studying the other (Hassabis, Kumaran, Summerfield, & Botvinick, 2017; Jozwik, Kriegeskorte, Cichy, & Mur, 2019; Kar, Kubilius, Schmidt, Issa, & DiCarlo, 2019; Kriegeskorte & Douglas, 2018; Lake, Ullman, Tenenbaum, & Gershman, 2017).

We show in this paper that such comparisons can be fraught with difficulties. We present three case studies that demonstrate different pitfalls in comparing human and machine performance.

Synthetic Visual Reasoning Test — Sub-optimal Architectures Can Lead to Fragile Conclusions

In order to compare human and machine performance in learning relationships between visual shapes, Fleuret et al. (2011) created the Synthetic Visual Reasoning Test (SVRT) consisting of 23 problems (Fig. 1A). They showed that humans need only few examples to understand the underlying concepts. Stabinger, Rodríguez-Sánchez, and Piater (2016) as well as Kim et al. (2018) assessed the performance of deep convolutional architectures on these problems. They found that some visual reasoning tasks were more difficult than others. Specifically, their models performed well on tasks that concern the spatial arrangement of objects — so-called spatial tasks — however, they struggled to learn tasks that involve the comparison of shapes — so-called same-different tasks (Fig. 1B). Based on these findings, Kim et al. (2018) claimed that same-different relationships would be difficult to learn in feed-forward convolutional architectures. They suggested that such comparisons might require feedback as found in recurrent systems.

In our experiment, we trained a feed-forward neural network based on the ResNet50-architecture (He, Zhang, Ren, & Sun, 2016) on the SVRT problems and show that our models could indeed perform well on both types of tasks (Fig. 1B). Specifically, our models were able to reach accuracies above 90% for all tasks. Thus, while other network architectures may be more efficient at learning same-different tasks than feed-forward networks (Kim et al., 2018), our results show that these tasks are not an inherent limitation of the network architecture per se. These findings highlight that when comparing which tasks are easy or difficult for humans or machines, it is important to keep in mind that a low performance of the machine model might arise from sub-optimal model parameters or training procedures. Hence, inferring general statements

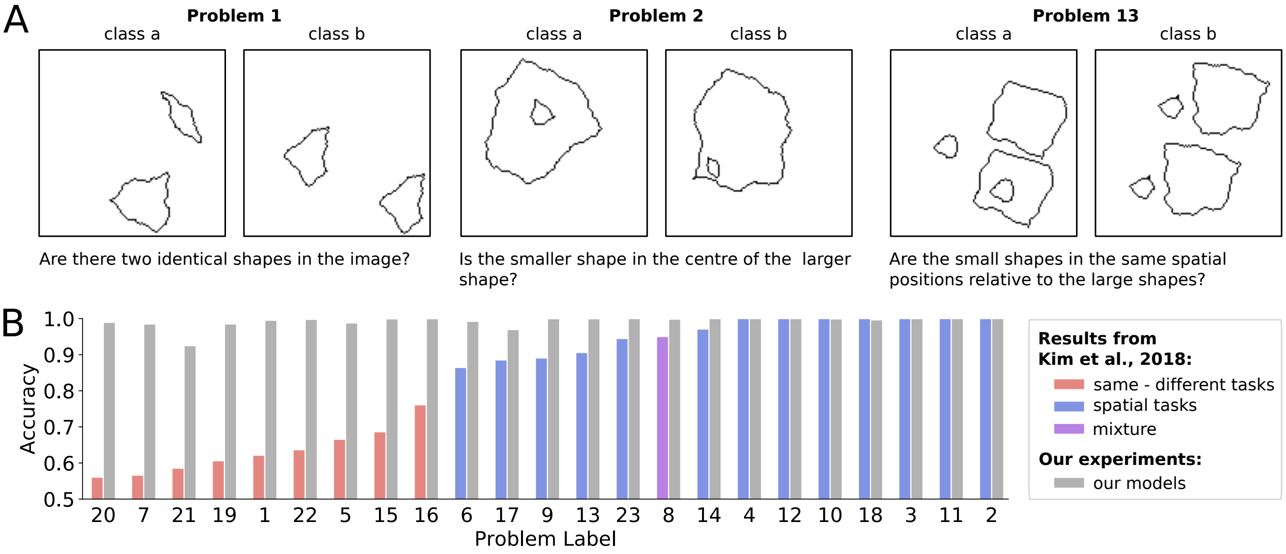

Figure 1: A: For three of the 23 SVRT problems, two example images representing the two opposing classes are shown. For each of the problems, the task is to find the rule that separates the images and to sort them accordingly. B: Kim et al. (2018) trained a DNN on each of the problems. They found that same-different tasks (red bars) are harder than spatial tasks (blue bars). Bars are re-plotted from Kim et al. (2018). Our ResNet50-based models (grey bars) reached high accuracies for all problems.

about failures of certain architectures necessitates a very thorough optimisation of all relevant parameters.

Recognition Gap — Humans and Machines Should Be Exposed to Equivalent Settings

Ullman et al. (2016) claimed the “human visual system uses features and processes that are not used by current models and that are critical for recognition.” This statement was based on experiments showing that humans’ ability to recognise image crops dropped sharply when the patch became too small or the resolution too low (Fig. 2 A, left column and Fig. 2 B, horizontal bar). In contrast, their machine algorithms did not show this ’recognition gap’ when tested on these human-selected stimuli (Fig. 2 B, right column).

Here, we investigated the same phenomenon in a very similar experimental design but with one crucial difference: instead of testing machines on human-selected patches, we tested them on machine-selected patches. In this way, we ensured that machines were tested in exactly the same way that humans were tested, i.e. humans and machines were each tested on patches they selected. We found that our VGG-based model (Simonyan & Zisserman, 2014) did experience a similarly strong recognition gap between minimally recognisable and maximally unrecognisable stimuli (Fig. 2B) - just like found for humans by Ullman et al. (2016).

We hereby illustrated that appropriately aligned testing conditions for both humans and machines are inevitable to compare perceptual phenomena between the two systems.

In the next section, we present another visual task that our deep convolutional neural network is able to successfully learn. However, we illustrate the pitfalls arising from interpreting the decision strategies of a machine from a human-centred perspective and ways to overcome this bias.

Closed Contour Detection — Models Do Not Necessarily Learn Human-like Concepts

Closed contours are thought to be prioritised by the human perceptual system and to be important in perceptual organisation (Elder & Zucker, 1993; Koffka, 2013; Kovacs & Julesz, 1993; Tversky, Geisler, & Perry, 2004). Here, we tested how well humans and a neural network could separate closed from open contours. To this end, we developed our own data set of images consisting of a main open or closed contour and additional task-irrelevant flankers.

We found that both humans and our ResNet50-based model could reliably tell apart images containing a closed contour from images containing an open contour (Fig. 3A, first column). In addition, our model also performed well on many variations of the contour task without further fine-tuning, including on dashed or asymmetric flankers or curvy instead of straight lines (Fig. 3A). These results suggest that our model did, in fact, learn the concept of open and closed contours and that it performs a similar contour integration-like process as humans.

However, our model did not learn this concept. For one, our model did not generalise to other variations such as different line colours or thicknesses (Fig. 3B). Second, very small and almost imperceivable changes in the brightness values of the image changed the decision of the model (Fig. 3C).

Instead of contour integration, our model might rely on

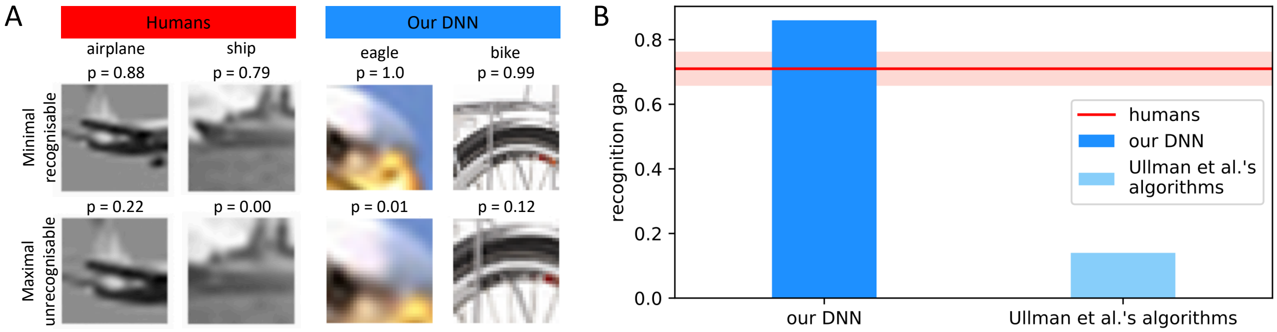

Figure 2: A: Minimal recognisable (upper row) and maximal unrecognisable (lower row) images to humans (tested by and reproduced from Ullman et al. (2016), left column) as well as to our DNN (right column). Titles of patches indicate recognition probabilities p. B: Recognition gaps for machine algorithms (blue) and humans (red).

much simpler features that are easily overlooked by humans. To analyse this, we made use of a recently introduced interpretable model, BagNet-33 (Brendel & Bethge, 2019), that sums up evidence from local patches of size 33 x 33 pixels to reach an image-level decision and is unable to perform any kind of non-linear integration on larger length scales (such as those needed for contour integration). This architecture let us trace exactly if the task could be solved with local features and which patches were most informative. Indeed, BagNet-33 reached close to 90% classification accuracy and used distinct local statistics, in particular multiple edges close to the endpoints of an open contour, to solve the task without any kind of contour integration as believed to be adopted by the human visual system (Fig. 3D).

The three techniques employed here (testing generalisation, adversarial perturbations and BagNets) provide complementary ways to investigate the strategies learned by the machine learning model in order to better understand differences in inferential processes compared to humans. To avoid premature conclusions about what models did and did not learn, we advocate for the routine use of various analytic techniques.

Conclusions

In this paper, we described notorious difficulties that arise when comparing humans and machines. First, negative results are not sufficient to conclude that particular network architectures cannot perform well at a task in principle. Second, when comparing humans and machines, equivalent experimental settings are crucial in order to make claims regarding the existence of phenomena in either system. Finally, even if a model does achieve high performance on a task, this does not mean its decision-making process is human-like. Different analysis tools such as generalisation tests, adversarial examples and tests with constrained networks can reveal insights into the models' inner workings.

Acknowledgments

We thank Alexander S. Ecker for helpful discussions.

The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting JB and CMF. We acknowledge support from the German Federal Ministry of Education and Research (BMBF) through the competence center for machine learning (FKZ 01IS18039A) and the Bernstein Computational Neuroscience Program Tübingen (FKZ: 01GQ1002), the German Excellence Initiative through the Centre for Integrative Neuroscience Tübingen (EXC307), and the Deutsche Forschungsgemeinschaft (DFG; Projektnummer 276693517 SFB 1233).

References

Alhazen, I. (1083). Book of optics. *The Optics of Ibn al-Haytham two volumes; translation by A I Sabra (1989; London: Warburg Institute)*.

Brendel, W., & Bethge, M. (2019). Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, *40*(4), 834–848.

Eigen, D., & Fergus, R. (2015). Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the ieee international conference on computer vision* (pp. 2650–2658).

Elder, J., & Zucker, S. (1993). The effect of contour closure on the rapid discrimination of two-dimensional shapes. *Vision Research*, *33*(7), 981–991.

Fleuret, F., Li, T., Dubout, C., Wampler, E. K., Yantis, S., & Geman, D. (2011). Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences*, *108*(43), 17621–17625.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245–258.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee*

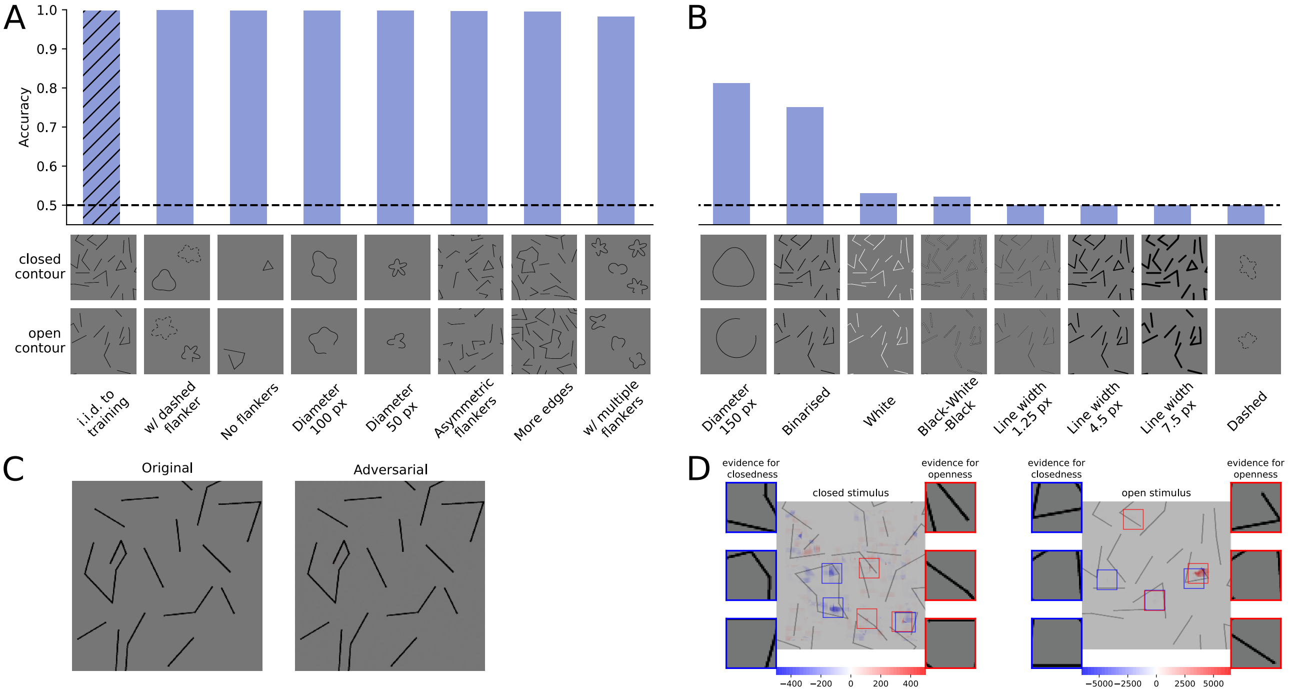

Figure 3: A: Our ResNet50-based model generalised well to many data sets, suggesting it would be able to distinguish closed and open contours. B: However, the poor performance on many other data sets shows that our model did *not* learn the concept of closedness. C: We generated adversarial examples for images of the closed contour data set. If the network used similar features as humans to discriminate closed from open contours, then adversarial images should swap the class label for humans. However, they appear identical to the original images. D: The heatmaps of our BagNet-based model show which parts of the image provide evidence for openness or closedness.

conference on computer vision and pattern recognition (pp. 770–778).

Helmholtz, H. v. (1925). Treatise on physiological optics. translated from the 3rd german edition (1910), southall, j. p. c (ed.). *Optical Society of America*.

Howard, I. P. (1996). Alhazen's neglected discoveries of visual phenomena. *Perception*, *25*(10), 1203-1217.

Jozwik, K., Kriegeskorte, N., Cichy, R. M., & Mur, M. (2019). Deep convolutional neural networks, features, and categories perform similarly at explaining primate high-level visual representations.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral streams execution of core object recognition behavior. *Nature neuroscience*, *22*(6), 974.

Kim, J., Ricci, M., & Serre, T. (2018). Not-so-clevr: learning same–different relations strains feedforward neural networks. *Interface focus*, *8*(4), 20180011.

Koffka, K. (2013). *Principles of gestalt psychology*. Routledge.

Kovacs, I., & Julesz, B. (1993). A closed curve is much more than an incomplete one: Effect of closure in figure-ground segmentation. *Proceedings of the National Academy of Sciences*, *90*(16), 7495–7497.

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature neuroscience*, 1.

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

Kümmerer, M., Theis, L., & Bethge, M. (2014). Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Stabinger, S., Rodríguez-Sánchez, A., & Piater, J. (2016). 25 years of cnns: Can we compare to human abstraction capabilities? In *International conference on artificial neural networks* (pp. 380–387).

Tversky, T., Geisler, W. S., & Perry, J. S. (2004). Contour grouping: Closure effects are explained by good continuation and proximity. *Vision Research*, *44*(24), 2769–2777.

Ullman, S., Assif, L., Fetaya, E., & Harari, D. (2016). Atoms of recognition in human and computer vision. *Proceedings of* *the National Academy of Sciences*, *113*(10), 2744–2749.