Comprehensive Assessment Methods are Key to Progress in Deep Learning

M. W. Spratling (michael.spratling@kcl.ac.uk)
King's College London, Department of Informatics, Bush House, 30 Aldwych, London, UK.

Abstract
Bowers etal. eloquently describe issues with current deep neural network (DNN) models of vision, claiming that there are deficits both with the methods of assessment, and with the models themselves. I am in agreement with both these claims, but propose a different recipe to the one outlined in the target article for overcoming these issues.

The target article proposes that DNNs be assessed using controlled experiments that evaluate changes in model behaviour as all but one variable is kept constant. Such experiments might provide information about the similarities and differences between brains and DNNs, and hence, spur development of DNNs better able to model the biological visual system. However, in reality work in deep learning is concerned with developing methods that work, irrespective of the biological-plausibility of those methods: deep learning is an engineering endeavour driven by the desire to produce DNNs that perform the "best". Even in the sub-domain where brain-like behaviour is a consideration (Schrimpf etal., 2020) the desire is to produce DNNs that produce the best performance. Hence, if controlled experiments were introduced, the results would almost certainly be summarised by a single value so that the performance of competing models could be ranked, and as a consequence there would be little to distinguish these new experimental methods from current ones.

What is meant by "best" performance, and how is it assessed, is the key issue. While training samples and supervision play a role in deep learning analogous to nurture during brain development, assessment plays a role analogous to that of evolution: determining which DNNs are seen as successful, and hence, which will become the basis for future research efforts. The evaluation methods accepted as standard by a research community thus have a huge influence on progress in that field. Different evaluation methods might be adopted by different fields, for example classification accuracy on unseen test data might be accepted in computer vision, while Brain-Score or the sort of controlled experiments advocated by the target article might be used to evaluate models of biological vision. However, as is comprehensively catalogued in the target article, current DNNs suffer from such a range of severe defects that they are clearly inadequate either as models of vision or as reliable methods for computer vision. Both research agendas would, therefore, benefit from more rigorous and comprehensive evaluation methods that can adequately gauge progress.

Given the gross deficits of current DNNs, it seems premature to assess them in fine-detail against psychological and neurobiological data. Rather, their performance should be evaluated by testing the ability to generalise to changes in viewing conditions (Michaelis etal., 2019; Mu and Gilmer, 2019; Hendrycks and Dietterich, 2019; Shen etal., 2021), the ability to reject samples from categories that were not seen during training (Hendrycks and Gimpel, 2017; Vaze etal., 2022), the ability to reject exemplars that are unlike images of any object (Nguyen etal., 2015; Kumano etal., 2022), and robustness to adversarial attacks (Szegedy etal., 2014; Biggio and Roli, 2018; Croce and Hein, 2020).

Methods already exist for testing generalisation and robustness of this type, the problem is that they are not routinely used, or that models are assessed using one benchmark but not others. The latter is particularly problematic, as there are likely to be trade-offs between performance on different tasks. The trade-off between adversarial robustness and clean accuracy is well known (Tsipras etal, 2019), but others are also likely to exist. For example, improving the ability to reject unknown classes is likely to reduce performance on classifying novel samples from known classes, as such exemplars are more likely to be incorrectly seen as unknown. Hence, efforts to develop a model that is less deficient in one respect, may be entirely wasted as the resulting model may be more deficient in another respect. Only when the community routinely requires comprehensive evaluation of models for generalisation and robustness will progress be made in reducing the range of deficits exhibited by models. Once such progress has been made it will be necessary to expand the range of assessments performed in order to effectively distinguish the performance of competing models and to spur further progress to address other deficiencies. The range of assessments might eventually be expanded to include neurophysiological and psychophysical tests.

The assessment regime advocated here can only be applied to models that are capable of processing images, and hence, would not be applicable to many models proposed in the psychology and neuroscience literatures. The target article advocates expanding assessment methods to allow such models to be evaluated and compared to DNNs. However, the ability to process images would seem to me to be a minimum requirement for a model of vision, and models that can not be scaled to deal with images are not worth evaluating.

To perform well in terms of generalisation and robustness it seems likely that DNNs will require new mechanisms. As Bowers etal say, it is unclear if suitable mechanisms can be learnt purely from the data. Indeed, even a model trained on 400 million images fails to generalise well (Radford etal., 2021). The target article also points out that biological visual systems do not need to learn many abilities (such as adversarial robustness, tolerance to viewpoint, etc.), and instead these abilities seem to be "built-in". Brains contain many inductive biases: the nature side of the nature-nurture cooperation that underlies brain development. These biases underlie innate abilities and behaviours (Zador, 2019, Malhotra etal., 2022) and constrain and guide learning (Zaadnoordijk etal., 2022, Johnson, 1999). Hence, as advocated in the target article, and elsewhere (Hassabis etal. 2017; Zador, 2019, Malhotra etal., 2020), biological insights can potentially inspire new mechanisms that will improve deep learning. However, work in deep learning does not need to be restricted to only considering inductive biases that are biologically-inspired, especially as there are currently no suggestions as to how to implement many potentially useful mechanisms which humans appear to use. Indeed, if better models of biological vision are to be developed it is essential that work in neuroscience and psychology contribute useful insights. Unfortunately, the vast majority of such work so far has concentrated on cataloguing "where" and "when" events happen (where an event might be a physical action, neural spiking, fMRI activity, etc.). Such information is of no use to modellers who need information about "how" and "why".

References

Biggio, B. and Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84:317–31. doi:10.1016/j.patcog.2018.07.023.

Bowers, J. S., Malhotra , G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton R. F., Evans, B. D., Mitchell, J. and Blything, R. (2023). Deep problems with neural network models of human vision. Behavioral and Brain Sciences (in press).

Croce, F. and Hein, M. (2020). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In Proceedings of the International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 2206–16. arXiv:2003.01690.

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. Neuron, 95:245–58. doi:10.1016/j.neuron.2017.06.011.

Hendrycks, D. and Dietterich, T. G. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In Proceedings of the International Conference on Learning Representations. arXiv:1903.12261.

Hendrycks, D. and Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In Proceedings of the International Conference on Learning Representations. arXiv:1610.02136.

Johnson, M. H. (1999). Ontogenetic constraints on neural and behavioral plasticity: evidence from imprinting and face recognition. Canadian Journal of Experimental Psychology, 53:77–90.

Kumano, S., Kera, H., and Yamasaki, T. (2022). Are DNNs fooled by extremely unrecognizable images? arXiv:2012.03843.

Malhotra, G., Evans, B. D., and 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. doi:10.1016/j.visres.2020.04.013.

Malhotra, G., Dujmović, M., and Bowers, J. S. (2022). Feature blindness: A challenge for understanding and modelling visual object recognition. PLoS Computational Biology, 18(5):e1009572. doi:10.1371/journal.pcbi.1009572.

Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A. S., Bethge, M., and Brendel, W. (2019). Benchmarking robustness in object detection: Autonomous driving when winter is coming. arXiv:1907.07484.

Mu, N. and Gilmer, J. (2019). MNIST-C: A robustness benchmark for computer vision. arXiv:1906.02337.

Nguyen, A., Yosinski, J., and Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. arXiv:1412.1897.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. arXiv:2103.00020.

Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R. and DiCarlo, J. J. (2020) Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence, Neuron, 108(3):413-423 https://www.cell.com/neuron/fulltext/S0896-6273(20)30605-X

Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., and Cui, P. (2021). Towards out-of-distribution generalization: A survey. arXiv:2108.13624

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. (2014). Intriguing properties of neural networks. In Proceedings of the International Conference on Learning Representations. arXiv:1312.6199.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. In Proceedings of the International Conference on Learning Representations, 2019. arXiv:1805.12152.

Vaze, S., Han, K., Vedaldi, A., and Zisserman, A. (2022). Open-set recognition: a good closed-set classifier is all you need? In Proceedings of the International Conference on Learning Representations. arXiv:2110.06207.

Zaadnoordijk, L., Besold, T. R., and Cusack, R. (2022). Lessons from infant learning for unsupervised machine learning. Nature Machine Intelligence, 4:510–20. doi:10.1038/s42256-022-00488-2.

Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. Nature Communications, 10(3770). doi:10.1038/s41467-019-11786-6.