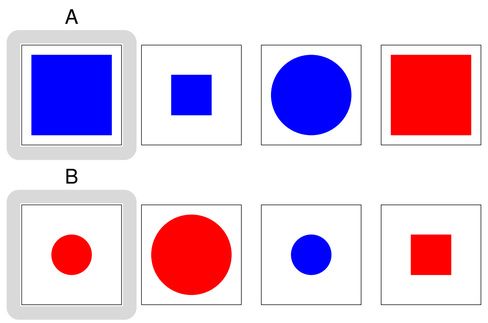
Levels of representation in a deep learning model of categorization
https://dx.doi.org/10.1101/626374

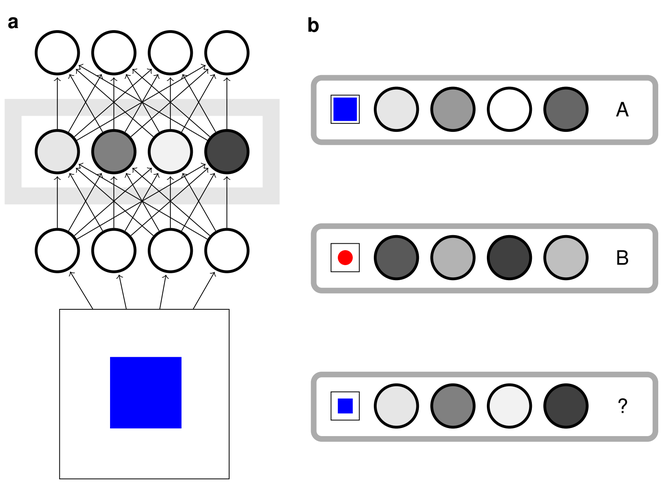
Olivia Guest[1,2] and Bradley C. Love[2,3]

[1]RISE, Cyprus [2]Experimental Psychology, UCL, UK [3]The Alan Turing Institute, UK

Deep neural network plus exemplar model



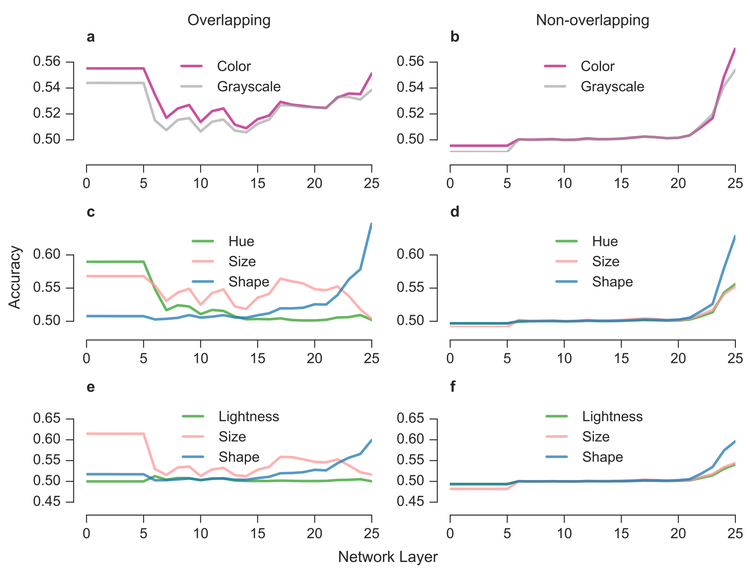
The 3-dimensional binary-valued stimulus set. Stimuli are members of the category for which they match the prototype on 2 or more of the three features (size, shape, and color). The prototype for each category is shown with a gray outline in the first column.



A schematic of the DCNN-plus-exemplar model. On the left in **a**), a stimulus is presented to the pre-trained DCNN; and on the right in **b**), the exemplar model has two previously stored exemplars, one each from category "A" and "B".

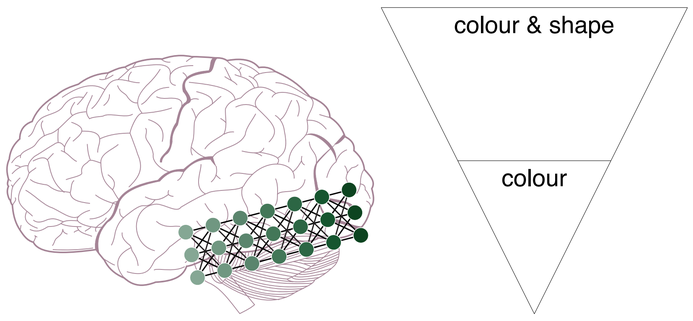
$$\mathrm{sim}(S_i, S_j) \equiv \rho(S_i, S_j) + 1$$

$$p(R_j|S_i) = \frac{\left(\sum_{j \in C_J} \mathrm{sim}(S_i, S_j)\right)^{\gamma}}{\sum_{K=1}^{m} \left(\sum_{k \in C_K} \mathrm{sim}(S_i, S_k)\right)^{\gamma}}$$



Accuracy (with the γ decision parameter set to 1) for the exemplar model using various network layers to provide exemplar representations.

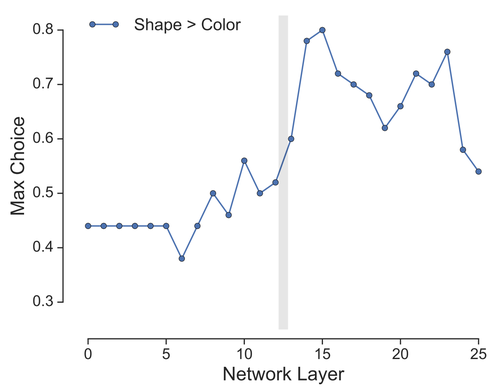
Hierarchy or inverse pyramid?



Shape bias

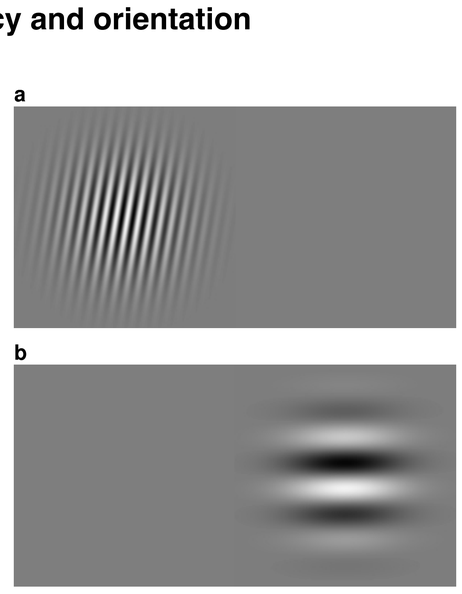
$$\mathrm{sim}(x, y) \equiv \rho(x, y) + 1$$

Stimuli from Linda Smith's lab:
http://www.indiana.edu/~cogdev/SB_testsets.html

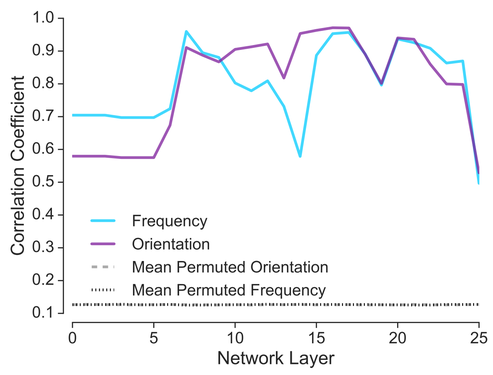


Preference for shape over color match in the triplet task for each layer of the network. $\chi^2(1,\ N = 1300) = 63.34,\ p < 0.0001$

Frequency and orientation



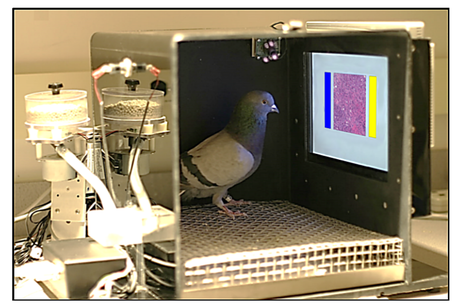
Examples of the Gabor patch stimuli that are non-overlapping. The Gabor patch on top in **a**) has a higher spatial frequency and more vertical orientation than the Gabor patch in **b**) shown below.



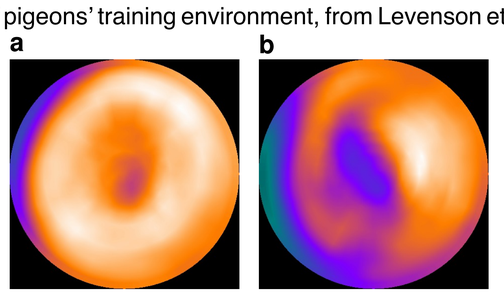
Network activity resulting from viewing Gabor patches varying in orientation and frequency was subjected to PCA.

How similar identical, but non-spatially overlapping, Gabor patches are to one another relative to other stimuli. The 50*th* percentile indicates chance performance in which the matching Gabor patches are as similar to each other as would be expected for randomly selected stimuli.The 100*th* percentile indicates perfect discrimination.

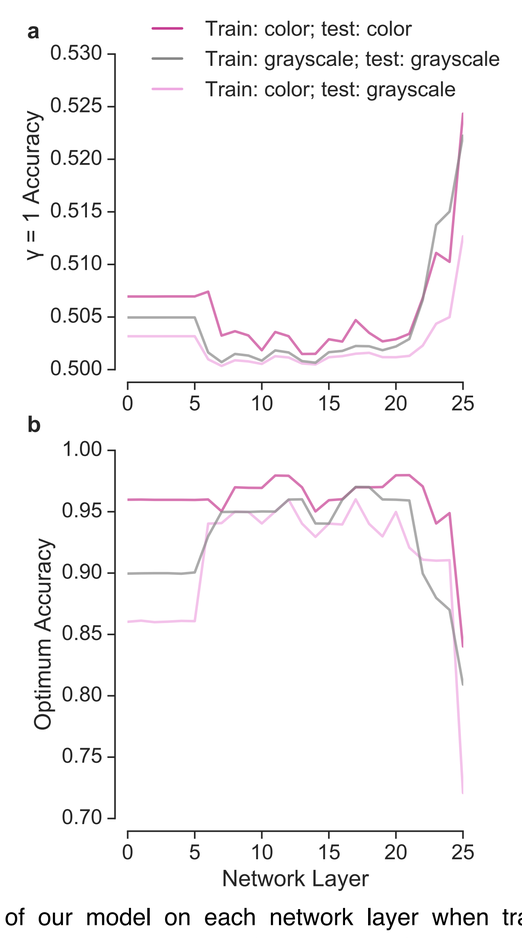
Pigeon categorization



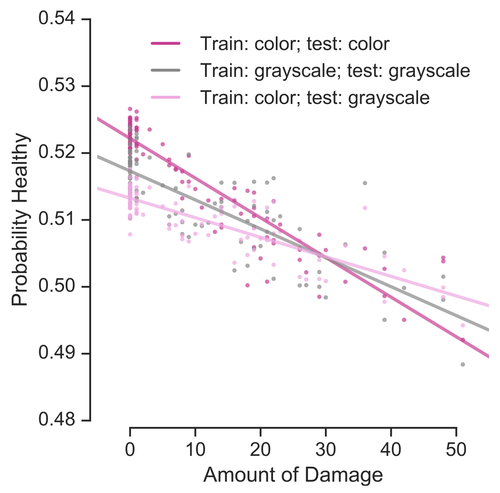
Example of pigeons' training environment, from Levenson et al. (2015).



Two examples cardiogram stimuli: **a**) a normal cardiogram without any perfusion damage; and **b**) an abnormal cardiogram with total perfusion damage 20 (of a maximum of 51).



Accuracy of our model on each network layer when trained and tested in a manner analogous to the pigeons. The qualitative pattern of performance observed at the lowest network layers mirrors the performance of the pigeons. In panel **a**), γ is set to 1, whereas in panel **b**) an optimal boundary (determined from the training set) is used.



When relying on very low-level DCNN representations, the model, like the pigeons, shows a gradient of responding depending on the level of damage. As shown by the regression lines, the model displays the same difficulty ordering when generalizing to novel test stimuli.

References

Guest, O. and Love, B. C. (2017). What the success of brain imaging implies about the neural code. *Elife*, 6:e21397.

Landau, B., Smith, L. B., and Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive development*, 3(3):299–321.

Levenson, R. M., Krupinski, E. A., Navarro, V. M., and Wasserman, E. A. (2015). Pigeons (columba livia) as trainable observers of pathology and radiology breast cancer images. *PLoS One*, 10(11):e0141357.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, 115(1):39.

Ritter, S., Barrett, D. G., Santoro, A., and Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2940–2949. JMLR. org.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Yamins, D. L. K., and DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*.

Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the national academy of sciences*, 104(15):6424–6429.