Supplemental Material: Learning Inductive Biases with Simple Neural Networks

Reuben Feinman (reuben.feinman@nyu.edu)
Center for Neural Science
New York University

Brenden M. Lake (brenden@nyu.edu)
Department of Psychology and Center for Data Science
New York University

Brenden: To make things clearer, can you divide into section numbers (like SM1, SM2, SM3, etc. that we can reference in main text?

Experiments 1 & 2

Selection of architectures and training parameters

Reuben: TODO - explain how we came to the architectures that we did for both the MLP and the CNN, explain that L2 regularization was important, explain selection of training parameters.

Experiment 2: perceptual sensitivity tests

As in Experiment 1, for Experiment 2 we parametrically manipulate the stimuli to analyze the network's sensitivity to changes along different stimulus dimensions, using a CNN trained with N=30 & K=10. Distance in shape space is quantified as the Modified Hausdorff Distance (Dubuisson & Jain, 1994) between the shape pair. In color space, physical distance is quantified using the cosine similarity of the RGB vector pair. Beginning with an exemplar object stimuli, we sample 50 secondary shapes and order them by their distance from the exemplar. We then modify the shape of the exemplar parametrically by stepping along this list, recording network similarities between the original and modified versions in each case. A mirroring experiment is then performed with color; in each case, only 1 attribute is altered at a time. Results are shown in Fig. 4. As with the MLP, our CNN's selection preferences show a clear parametric dependency on shape, and a much weaker dependency on color

Experiment 2: color bias training

For the sake of comparison, in Experiment 2 we also trained our CNN to label objects with names organized by color. Our goal was to compare the required sample complexity for color bias training with that of shape bias training, and to evaluate whether color bias development follows a similar 2-step process. All dataset parameters mirrored those of shape training, except that the object labels were aligned with the color attribute of each training image. Performance on the generalization tests was measured as the fraction of trials for which the network selects the color match. Results for CNN color bias training are shown in Supp. Fig. 1. Notably, the color-trained CNN requires a smaller sample complexity to achieve 0.7 accuracy on the 2$^{\text{nd}}$-order test, reaching a score of 0.73 with N=2 & K=3. Furthermore, this network does not appear to follow the 2-step process of bias development; results for 1$^{\text{st}}$- and 2$^{\text{nd}}$-order generalizations look near-identical to one another. In order to identify a stimulus as a member of a particular color category, the network needs only to find a single pixel of that color, a task that is much simpler than representing and identifying shape. Representing color requires a simple 3D space. By learning to isolate and preserve this space in the hidden layers, the network can easily generalize to novel colors, hence the early 2$^{\text{nd}}$-order results. We inspected the learned representations of both a shape-trained and a color-trained CNN, trained with N=50 & K=18, by visualizing the first-layer convolution filters of each network (Supp. Fig. 2). As we would expect, filters of the shape-trained CNN look identical across R, G and B channels, as this network needs no sensitivity to color. In contrast, filters of the color-trained CNN vary across channels, indicating that the network has learned a selectivity for color.

Experiment 3

Network architecture

In Experiment 3, we use a slightly modified version of the CNN from Experiment 2 (see Supp. Fig. 3). We train our CNN to simultaneously label the object's name, which correlates with shape, as well as its color and texture names. The CNN thus has 3 softmax layers, each of which extends from the same fully-connected layer. The training loss is computed as a weighted average of the 3 softmax losses Brenden: negative loglikelihood loss, I think is the right term here?, with weights of 0.6, 0.2 and 0.2 assigned to shape, color and texture, respectively:

$$Loss = 0.6 * shape\_loss + 0.2 * color\_loss + 0.2 * texture\_loss$$

Correlation 3

Brenden: My preference would be to cut correlation 3 entirely. it's harder to compare than the others As a 3$^{\text{rd}}$ metric of the dependency between shape bias acquisition and vocabulary acceleration in toddlers, Gershkoff-Stowe & Smith (2004) computed a correlation of 0.85 between the first session in which a child shows a "systematic" shape bias and the first session in which she shows a "substantive" increase in vocabulary size, computed across participants. They define the former as the first session in which a child exhibits a performance on the shape bias test that, if the child were selecting matches in this test at random, would only occur with probability 0.1. In our framework, we have 1500 test trials during each session. Using a binomial test, we can reject the null hypothesis that the network is responding randomly with $p < 0.1$ given 523 or more shape choices out of the 1500 trials. The authors define a "substantive" increase in vocabulary size as the first session that vocabulary size increases by 10 from the previous session. Since they choose threshold 10 for a maximum vocabulary of 100 words, we use threshold 4 (i.e.

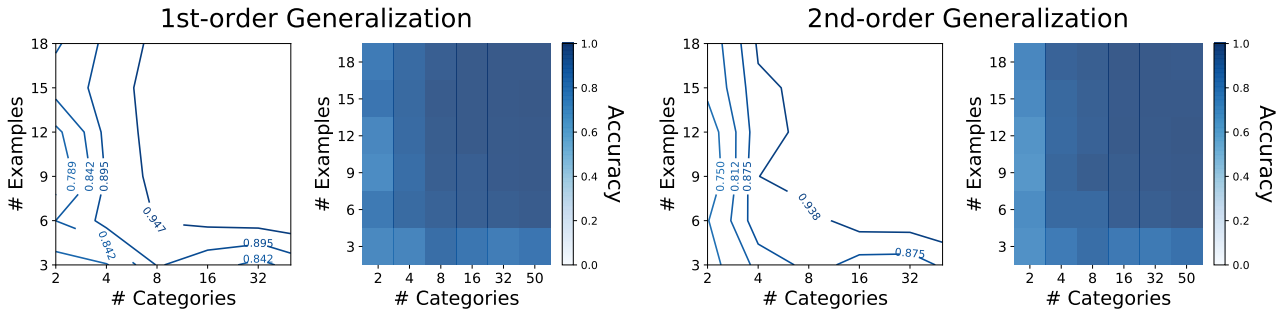
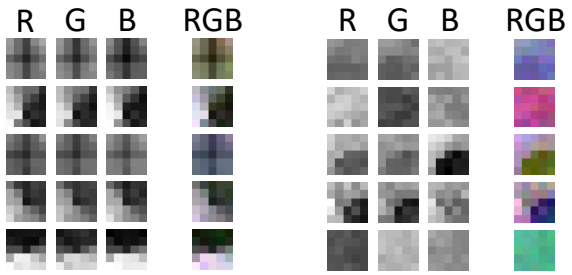


Figure 1: CNN generalization results for color bias training. The network is trained to label objects with category names based on color. In this case, the generalization tests evaluate the fraction of times that the color match is selected. Results in each grid show the average of 10 training runs.

3.6) for our 36-word vocabulary. With these thresholds, we find a correlation of 0.52 ($p < 0.015$).



(a) Shape (b) Color

Figure 2: Visualizing RGB channels of learned first-layer convolution filters. (a) shows the filters of our CNN trained with explicit shape bias training (N=50 & K=18). Each row corresponds to 1 of the 5 filters. The first 3 channels are shown in the 'R', 'G' and 'B' columns, respectively. These 3 channels are shown together in a 4th column, labeled 'RGB'. (b) shows mirroring filters for our CNN trained to label objects with category names based on color. In both (a) and (b), only channels 1-3 of the 4 are shown.

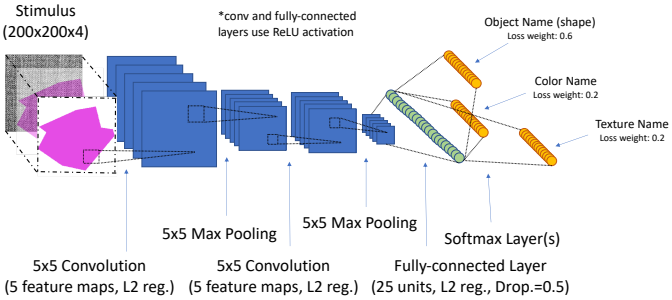


Figure 3: Reuben: caption TODO

References

Dubuisson, M., & Jain, A. K. (1994). A modified hausdorff distance for object matching. In *Proceedings of the international conference on pattern recognition* (pp. 566–568).

Gershkoff-Stowe, L., & Smith, L. B. (2004). Shape and the first hundred nouns. *Child Development*, *75*(4), 1098–1114.