Conjunctive Coding of Color and Shape in Convolutional Neural Networks

JohnMark Taylor (johnmarktaylor@g.harvard.edu)
Psychology Department, Harvard University
33 Kirkland St., Cambridge, MA 02139 USA

Yaoda Xu (yaoda.xu@yale.edu)
Psychology Department, Yale University
1 Prospect St., New Haven, CT 06520 USA

Abstract

Understanding how the visual system conjunctively codes color and shape has long fascinated cognitive psychologists, cognitive neuroscientists and neurophysiologists. Recent developments in convolutional neural networks (CNNs) provide us with an excellent opportunity to examine how color and shape conjunctions may be coded in an artificial, feedforward system only trained to perform object recognition. To determine whether CNNs encode color and shape independently or in an interactive manner, we used representational similarity analysis to characterize the responses of Alexnet to a collection of 540 different objects, each presented in 36 different colors. We found that whereas lower layers of Alexnet encode colors in a similar manner across different objects, in higher layers the color spaces associated with different objects are more distinct. Interestingly, the similarity between the color spaces of different objects was only weakly (though significantly) associated with the objects' shape similarity. These results demonstrate that rather than being encoded in an orthogonal manner, color and shape processing becomes increasingly interactive in higher layers of a CNN, suggesting that feedforward networks optimized for object recognition will naturally develop conjunctive coding of color and shape.

Keywords: color; shape; convolutional neural nets; conjunctive coding; binding

Introduction

The human visual system must successfully process different features, like color and shape, that each impose their own unique processing demands, such as inferring color constancy or computing 3D shape. At the same time, it must successfully bind together different features belonging to the same object, raising the question of what sort of overall processing architecture might be able to perform these complementary demands of both segregating and integrating the processing of different features.

Influential accounts of visual feature binding in the human brain have argued that whereas single features are encoded in a fast, bottom-up, parallel manner, successfully binding different features into objects requires focused spatial attention (Treisman & Gelade, 1980). That said, several lines of behavioral evidence suggest that in certain cases, feature conjunctions can be encoded in a fast, bottom-up manner, such as for familiar color/shape conjunctions (Rappaport, Humphreys, & Riddoch, 2013; Reavis, Frank, Greenlee, & Tse, 2016). It remains unknown what neural mechanisms could undergird these patterns of behavioral data. Some theories of color/shape binding emphasize the role of feedback or recurrent connections in feature binding, whereas others focus on the role of bottom-up computations (Riesenhuber & Poggio, 1999; Singer, 1999; Zhang et al., 2014).

In order to better understand how color and shape processing might interact in a purely feed-forward architecture, excluding any recurrent or feedback processing, we examined the responses of Alexnet, a convolutional neural network (CNN) solely optimized for object recognition, to a collection of object stimuli that were each presented in several different colors (Krizhevsky, Sutskever, & Hinton, 2012). While the input to Alexnet's first layer consists of an image with separate RGB channels, the model has no built-in constraints on how color is subsequently processed by the network, allowing us to examine what color representations naturally emerge as a byproduct of training the network to perform object recognition. In particular, we used representational similarity analysis (RSA) to characterize how color is encoded for different shapes across different layers of Alexnet (Kriegeskorte, Mur, & Bandettini, 2008). To the extent that color and shape are encoded orthogonally in CNNs, then the pattern of dissimilarities between different colors should be invariant across different objects; conversely, if color encoding is contingent on shape coding, then the pattern of color dissimilarities should vary among different objects. We additionally examined whether differences in the color spaces for different objects are predicted by differences in the shape representations of these objects in different layers.

Stimuli

As stimuli, we used a stimulus set consisting of 540 different objects from Brady et al. (2013). Following the manipulation employed in a recent neurophysiology study (Chang, Bao, & Tsao, 2017), to vary the color in these stimuli, we converted the RGB images to the LUV color space, which is designed such that equally similar stimuli in this color space are also equally perceptually discriminable. We then adjusted the stimuli such that they all had equal mean luminance and saturation, and had 36 evenly spaced hues comprising a circle in color space. We constructed the stimuli using two different methods. In the first method, we retained the original textures of the images; in the other, we filled each image with a uniform color, forming a “silhouette” for each object. We performed this manipulation to examine whether our results depend on the internal texture of the objects, or only their overall shape envelope. Figure 1 depicts examples of these stimuli.

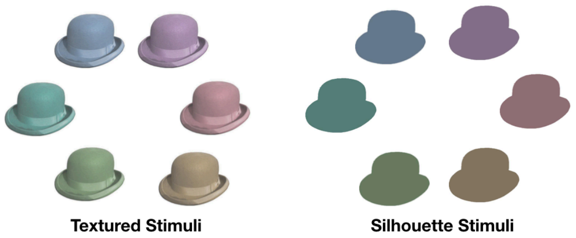

Figure 1. Six (out of 36) example colors of one example stimulus, for the two stimulus types.

Analyses and Results

The stimuli were run through Alexnet, using the Pytorch implementation (Paszke et al., 2017). We then extracted the unit activations from each layer of Alexnet. We retained the spatial dimensions of the various layers instead of averaging across space within each kernel in order to better approximate the methods used in neural data analysis, which have no clear analogue of performing kernel-wise averaging. After flattening these activations into 1D vectors, we computed a representational dissimilarity matrix (RDM) for the 36 different colors of each of the 540 objects, for each layer of the network. We then computed a second-order RDM (540*540) for each layer that measures the correlations in the color spaces across the different objects (Figure 2), and examined how the distribution of correlation values in this matrix evolve across the layers of the network (Figure 3). The correlations in the color spaces among different objects decline over the course of the network, suggesting that color and shape are processed in an increasingly interactive manner in Alexnet. These results held for both the textured stimuli and the silhouette stimuli.

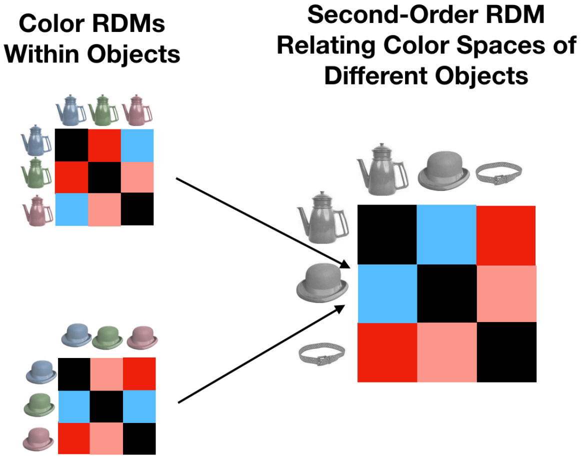

Figure 2. For each layer, color RDMs were computed for each object (only 3 of 36 colors shown) to measure its color similarity space. These RDMs were then correlated across objects to measure the similarity of the color representations for different objects (only 3 of 540 objects shown).

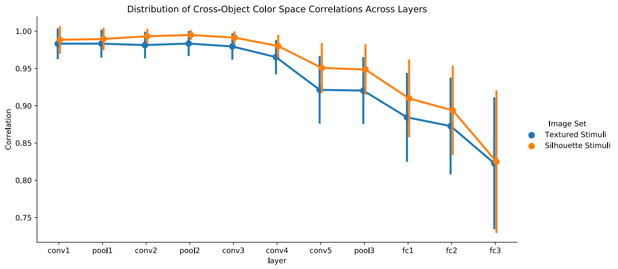

Figure 3. Distribution of cross-object color space similarities across layers. Error bars show SD.

We then examined whether objects with a similar color space at the beginning of the network also had a similar color space at the end of the network. To do this, we correlated the second-order RDMs from the previous analysis between layers conv1 and fc2, finding only a modest, though significant, correlation (r = .17). This suggests that the differences between objects with respect to their color spaces evolve over the course of the network, with the color spaces of some object pairs converging, and others diverging in higher layers.

To what extent do these differences in color spaces between objects reflect differences in the shape representations for these objects? In other words, do objects with similar shapes tend to have similar color representations? To answer this, for each layer we extracted the activations to grey-scale versions of all the object stimuli, and constructed an RDM capturing the differences in non-chromatic shape between objects. For each layer, we then correlated this RDM with the second-order RDM containing the differences in color-space between different objects. This allows us to examine whether objects with more different shape representations also encoded color in a more different manner. Across the layers, we found modest, though significant, correlations ranging from .1 to .3, suggesting that differences in shape representation between different objects only partially track differences in their color space representation.

Discussion

Collectively, our results suggest that rather than encoding color and shape in an orthogonal manner, CNNs optimized for object recognition naturally represent these features in an interactive manner, such that color coding varies across different shapes and varies across the different CNN layers. These results therefore represent an existence proof that conjunctive coding of color and shape can naturally arise in the absence of feedback or recurrent mechanisms in a CNN trained only for object recognition.

Acknowledgments

This work was funded by a National Science Foundation Graduate Research Fellowship to J.T.

References

Brady, T. F., Konkle, T.F., Gill, J., Oliva, A. and Alvarez, G.A. (2013). Visual long-term memory has the same limit on fidelity as visual working memory. *Psychological Science*, 24(6), 981-990.

Chang, L., Bao, P., & Tsao, D. Y. (2017). The representation of colored objects in macaque color patches. *Nature communications*, *8*(1), 2064.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 4.

Krizhevsky, A., Sutskever, I., & Hinton, G.E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM, 60*, 84-90.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. & Lerer, A. (2017). Automatic differentiation in pytorch.

Rappaport, S. J., Humphreys, G. W., & Riddoch, M. J. (2013). The attraction of yellow corn: Reduced attentional constraints on coding learned conjunctive relations. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(4), 1016.

Reavis, E. A., Frank, S. M., Greenlee, M. W., & Tse, P. U. (2016). Neural correlates of context-dependent feature conjunction learning in visual search tasks. *Human brain mapping*, *37*(6), 2319-2330.

Riesenhuber, M., & Poggio, T. (1999). Are cortical models really bound by the “binding problem”?. *Neuron*, *24*(1), 87-93.

Singer, W. (1999). Neuronal synchrony: a versatile code for the definition of relations?. *Neuron*, *24*(1), 49-65.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*(1), 97-136.

Zhang, X., Qiu, J., Zhang, Y., Han, S., & Fang, F. (2014). Misbinding of color and motion in human visual cortex. *Current Biology*, *24*(12), 1354-1360.