Using features from deep neural networks to model human categorization of natural images

Ruairidh Battleday (battleday@berkeley.edu)
Helen Wills Neuroscience Institute
University of California, Berkeley
Berkeley, CA 94720, U.S.A.

Joshua Peterson (jpeterson@berkeley.edu)
Department of Psychology Room 3210, Tolman Hall
University of California, Berkeley
Berkeley, CA 94720-1650, U.S.A.

Thomas Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology Room 3210, Tolman Hall
University of California, Berkeley
Berkeley, CA 94720-1650, U.S.A.

Abstract

An open question is whether recent advances in machine learning can be used to study human intelligence. Recent work from neuroscience has shown that stimulus representations from convolutional neural networks (CNNs) are our best currently available predictors of several visual areas in the brain, and behavioral studies have found they best explain human similarity judgments of natural object images. A prime candidate to continue this avenue of investigation is categorization, a fundamental cognitive function. Although a range of high-precision formal models of categorization exist, they are limited to simple, artificial stimuli because representing more realistic stimuli is difficult. We show that representations derived from CNNs can be used to extend these models to natural images, and that a group of cognitive models based on these representations capture human behavior over a novel crowdsourced database of $>$500,000 human classification decisions. Successful models include both exemplar and prototype models, contrasting with the dominance of exemplar models in previous work. We also find that performance is improved by using representations from networks that score more highly at ground-truth prediction.

Keywords: Categorization, Deep Neural Networks, Cognition.

Introduction

In this paper, we investigate whether stimulus representations from convolutional neural networks (CNNs) (LeCun, Bengio, & Hinton, 2015) can be used in the study of categorization, inspired by recent findings that they are our best predictors of visual cortex voxel activity (Agrawal, Stansbury, Malik, & Gallant, 2014) and human similarity judgments between natural images (Peterson, Abbott, & Griffiths, 2016). Categorization has been well-studied in cognitive science, yielding a range of high-precision formal models of behavior (Griffiths, Canini, Sanborn, & Navarro, 2007). However, experiments have mostly been limited to simple, artificial domains because of issues of how to represent more complex stimuli (McKinley & Nosofsky, 1995). We use CNN representations to extend these models to natural images, enabling human categorization to be studied over the complex visual domain in which it evolved and developed.

Methods

We collected a novel dataset of $>$500,000 human classifications of thousands of natural images using Amazon Mechanical Turk. We then extract stimulus representations for the same images from a range of discriminative and generative CNNs previously trained to classify ground-truth labels. Finally, we fit a range of prototype and exemplar models from psychology to our behavioral data using the CNN stimulus representations as approximations of inaccessible human mental ones.

Results

We find that a range of models capture human behavior well, near the reliability of human judgments, and better than the classifications by the original deep CNNs. Interestingly, this group contains both exemplar and prototype models, at odds with what might be expected if extrapolating from studies involving simple artificial stimuli. We also find that representations from better-performing networks improve the performance of all categorization models.

References

Agrawal, P., Stansbury, D., Malik, J., & Gallant, J. L. (2014). Pixels to voxels: modeling visual representation in the human brain. *arXiv preprint arXiv:1407.5104*.

Griffiths, T. L., Canini, K., Sanborn, A., & Navarro, D. (2007). Unifying rational models of categorization via the hierarchical dirichlet process. In *Proceedings of the 29th annual conference of the cognitive science society.*

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(1), 128.

Peterson, J., Abbott, J., & Griffiths, T. (2016). Adapting deep network features to capture psychological representations. In *Proceedings of the 38th annual conference of the cognitive science society.* Austin, TX.