How Labels Augment Category Representations: A connectionist model

Gary Lupyan
Carnegie Mellon University and the Center for the Neural Basis of Cognition
glupyan@cnbc.cmu.edu

Using words to label categories is a true human universal. The present work reviews converging evidence that, beyond communication, verbal labels play important roles in concept formation. I present a connectionist simulation exploring the effect of labels on learning different types of categories. I argue that labels both supplement and augment perceptual information, and play an especially important role for entities whose perceptual features alone are insufficient for reliable classification.

In order to explore the idea that labels have reorganizing effects on category representations, a simple auto-encoder model was implemented which contrasted auto-encoding items with and without labels. The goal of this simulation was to serve as a broad testing platform for the argument that labels have an effect on representation of items generated from prototypes and that labels have a differential effect depending on how well-formed the resulting category structure is. Specifically, the presence of labels might help in classifying exemplars that come categories with fuzzy boundaries, but not ones that are in already well-defined categories.

The presence of labels had no effect on classification accuracy for categories with already well-defined boundaries (the *low-variability* training set), but greatly increased accuracy for categories with fuzzy boundaries (the *high variability* set) (see Figure 1). Importantly, these benefits were seen when labels were present only during training, but never at test, further arguing that labels may play an important role in shaping representations directly rather than resulting only in strategic biases (e.g., Goldstone, Lippa, & Shiffrin, 2001 for a similar proposal).

Representing an object as part of a category typically involves highlighting some features while abstracting others (e.g., Pothos, 2004). In accordance with this basic principle of categorization, the more clustered (i.e., well-separated) a category structure became, the more feature abstraction took place. In short, unless allowed to simply memorize all the items, labels improve categorization accuracy at the cost of reconstruction accuracy (see Figure 2). Especially for the high-variability training sets, representations of labeled items drifted towards category prototypes compared to when labels were absent. This finding makes the provocative prediction that in the context of category-learning, labeling exemplars will lead to greater abstraction (i.e., reduced memory for exact attributes) in favor of a more categorical representation of those items. We are currently getting ready to test this prediction in human participants.

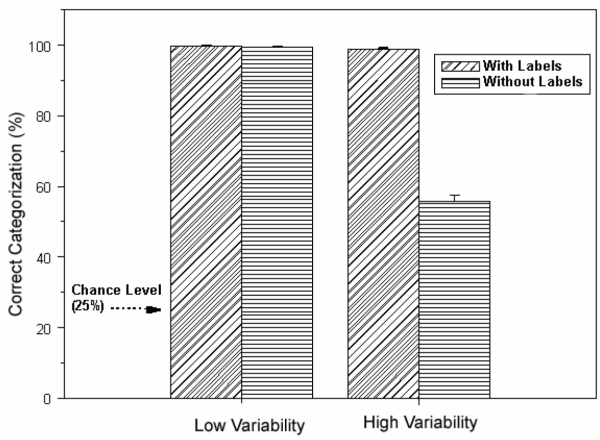

Figure 1: Labels improve categorization accuracy for high-variability categories, but not low-var ones.

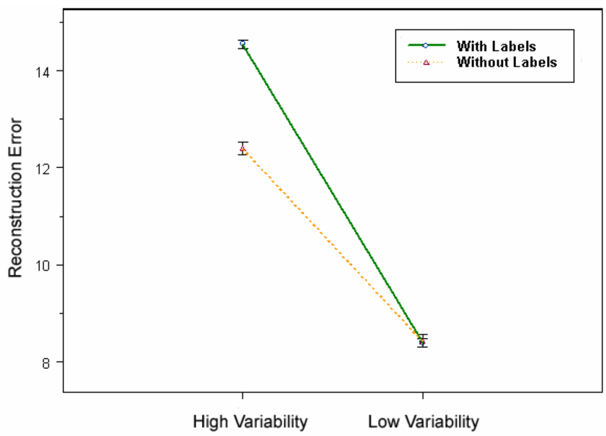

Figure 2: The improved categorization accuracy comes at a cost to accurate representation.