CARVING NATURE AT ITS JOINTS AND CARVING JOINTS INTO NATURE: HOW LABELS AUGMENT CATEGORY REPRESENTATIONS

GARY LUPYAN

Carnegie Mellon University Department of Psychology and
Center for the Neural Basis of Cognition
Pittsburgh, PA 15213 USA
glupyan@cnbc.cmu.edu

Using words to label categories is a true human universal. In addition to their public function in communication, labels may also serve private functions in shaping how concepts are represented. The present work explored the effects of assigning category labels on perceptual representations. A connectionist simulation is presented that examines the effects of labels on learning different types of categories. It is found that labels can augment perceptual information, and play an especially important role in shaping representations of entities whose perceptual features alone are insufficient for reliable classification.

1. Introduction

The phrase carving nature at its joints goes back to the ancient Greeks, who noticed that properties in the world are not equally distributed, but rather come in clusters. For instance, things that have beaks also tend to fly, so the property [has beak] is more likely to occur with [flies] than not. In forming a concept [bird], we pay heed to this correlation. But concept formation is not just a passive process of observing correlations. As the only species to evolve language, the human process of concept formation is bi-directional—not only do we learn to represent the natural joints of nature, we continuously carve our own joints using words. The present work explores through a connectionist simulation what happens when, in learning to form a concept, we are presented with not just the correlated perceptual properties of a concept like [bird], but also with the corresponding category label “bird.”

The most obvious function of words is to communicate information between individuals. One provocative consequence is that once the meaning of words is shared in a community, their use allows for the learning of novel combinations of concepts through manipulation of labels rather than trial-and-error learning (Cangelosi, Greco, & Harnad, 2000). But in addition to this public function of labels, they may also have a private function. Words may help to cohere concepts, reinforcing the correlated perceptual features that comprise them, and aid in contrasting similar, but functionally different concepts. The goal of the first simulation was to serve as a broad testing platform for the argument that labels have an effect on representation of categories and that the effect depends on how well-formed the categories are. Specifically, the presence of labels may help in categorizing exemplars in high-variability categories, but not in low-variability ones. Consider categorizing apples versus strawberries on the one hand, and tables versus chairs on the other. Apples and strawberries are members of what can justifiably be called *low-variability* categories. These categories have high internal coherence (Gentner, 1981), having both a high within-category similarity and a high between-category contrast. On the other hand, categorizing objects like chairs and tables is made more difficult by both their high perceptual within-category variability (tables and chairs can take a wide variety of shapes) and the fuzzy boundary between the categories (some tables are quite chair-like, and vice-versa). Including labels in one's experience of such *high-variability* categories may facilitate accurate identification of exemplars. On the other hand, including labels as part of training on low-variability categories is not predicted to make a difference due to the perceptual well-formedness of such categories. If the effect of labels is to increase the internal coherence of a category (Kotovsky & Gentner, 1996; Kersten & Smith, 2002), then in doing so, representations of exemplars from one category will be made more similar, while representations of items from separate categories will be made more distinct.

A secondary aim of this simulation was to further explore the consequences of introducing labels: if labels improve categorization accuracy, does this improvement come at a cost? A basic tenet of categorization is that placing an item into a category involves highlighting some properties while abstracting over others (e.g., Pothos, 2004). For instance, color is probably unimportant to the category *chair*, and so should be suppressed in the representation of a particular chair. However, color is important for the category *banana*, and so should be highlighted in its representation. If resources are abundant and each item can be memorized, abstraction and highlighting of particular dimensions may speed responses, but is otherwise not crucial. If one can afford to encode accurately every observable property of an object, categorization is not necessarily compromised by representing all the dimensions. However, when resources are limited, preventing memorization of individual exemplars (as was the case in the current simulation), the properties that are important to a category will be represented at the expense of properties irrelevant to the category. Making a representation more categorical (i.e., moving it closer to the prototype) will therefore lead to better categorization (smaller categorization errors), but at the same time, greater difficulty with reconstructing the exact item with all its original features (i.e., larger reconstruction error). Furthermore, this effect should be larger for high-variability categories since it is for these categories that labels are predicted to have the greatest effects.

2. Method

2.1. *Network Architecture*

The simulations used a standard feedforward network configured as an auto-associator. This particular framework is used in the present model for several reasons. First, architectures of this type are quite familiar, commonly used to model category learning (e.g., Plunkett, Sinha, Moller, & Strandsby, 1992; Mareschal, French, & Quinn, 2000). Auto-associators perform a type of principal component analysis, first picking up on the broad distinctions between exemplars, and then on the finer differences. With training, representations in the hidden layer come to reflect the category structure present in the training set. Because these networks are self-supervised, no external teaching signal is necessary—the structure of the training set itself guides learning.

The network was configured with the following set of connections: A 30-unit input layer connected to a 6-unit hidden layer that connected to a second 6-unit hidden layer, which in turn connected to a 30-unit output layer.[i] In the labeling conditions, the second hidden layer also connected to a 4-unit labeling output layer. Because labels were only presented as outputs, they could have an effect only during training. Labels were not included as inputs because it was discovered that doing so did not change the results qualitatively, while greatly complicating the training procedure. If labels were to have an effect on representations, it was having to represent the labels as outputs that would make the difference.

Due to the difference between the size of the two output layers (30 versus 4 units), the error derivatives of the labeling output layer were multiplied by a constant factor to 10. This approximately equated the contributions of the two layers.

2.2. *Stimuli*

The training corpora consisted of 30-unit binary patterns. First, four category prototypes were generated with each one overlapping on 24 out of 30 bits with the other prototypes. Training items were generated by randomly flipping a set number of bits: 5 out of 30 for the *low-variability* training set and 10 out of 30 for the *high-variability* set. Each category consisted of 8 exemplars for a total of 32 (4*8) exemplars per training set. The category labels were modeled by 4-bit orthogonal vectors (e.g., 1000 for all examples from the first category, 0100 for the second category, etc.).

2.3.*Training procedure*

The networks were trained in Lens v.2.4 with cross entropy error using “Doug’s Momentum,” a modified version of steepest gradient descent (Rohde, 1999). The learning rate was set to 0.1, and momentum to 0.9. Training was done in batch mode—weights were updated only after presenting all the examples.

On each training trial, the outputs of the input layer were clamped to a randomly selected example. Since the network’s task was simply to auto-associate the inputs, the targets of the main output layer were identical to the network inputs. In the labeling conditions, the second output layer provided the additional targets corresponding to the category labels. Twenty randomly-initialized networks were run in each of the four condition (*labeled low-variability*, *unlabeled low-variability*, *labeled high-variability*, and *unlabeled high-variability*).

2.4. *Testing procedure*

Testing was conducted by running a forward pass through the network for each item after training. Results are reported for 200 epochs of training. Three measures were computed to gauge the networks’ performance. First, an accuracy measure was calculated to answer the question “is the network’s response closest to the correct category?” This was done by computing the Euclidean distance between the produced output and the 4 category prototypes. The network’s response was marked correct if the minimum distance corresponded to the correct category. The responses were translated to a percentage correct measure over the 32 testing items (8 for each category). The second measure was the cross-entropy error of the training items, and answers the broad question “how good is the network at representing the trained patterns?” Since the networks’ hidden layers were kept deliberately small to prevent the network from memorizing the exact patterns, this error is not expected to be close to 0. The third measure corresponded to the error of the mapping from the trained or novel items to the category prototype from which the item was derived. This answers the question "how categorical is the network's representation?" In other words, how close to the correct category is the network's response?

3. Results

The simulation results confirmed both predictions. Labels did not affect categorization performance for the low-variability corpus, *t*(38) = 1.16, ns, with performance very close to 100% regardless of labels. Labels did affect categorization for the fuzzier category structure of the high-variability corpus *t*(38)=22.57, *p*<.0001, creating a highly significant interaction between labeling condition and category structure as a predictor of categorization accuracy, *F*(3,76)=487.18, *p*<.0001. Performance was very close to 100% for the high-variability corpus when trained with labels, but dropped to an average of 55.6% when trained without labels (chance=25%; see Figure 1).

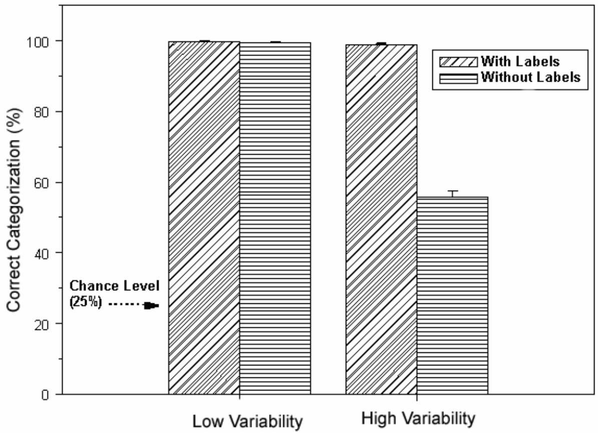

Figure 1: Labels improve categorization accuracy for high-variability categories, but not low-variability ones.

A parallel pattern of performance was observed when the network was tested on novel items. When tested on previously unseen items, performance was predictably lower overall, but labels still did not affect categorization accuracy for the low-variability corpus *t*(38)=0.6, while significantly aiding performance in the high-variability corpus *t*(38)=7.06, *p*<.0001, with correct categorization averages of 54.75% and 37.20% with and without labels, respectively. The interaction remained highly significant: *F*(3,76)=368.27, *p*<.0001. As a way to ensure that the null effect of labels on the low-variability corpus was not due to a ceiling effect, an additional analysis examined performance over time. Figure 2 shows that for the low-variability condition, labels did not affect performance at any point during training *t*(24)=.65. For the high-variability corpus, labels improved categorization accuracy throughout training both for the trained exemplars *t*(24)=7.30, *p*<.0001) and for the novel items *t*(24)=8.03, *p*<.0001.

The second measure looked at reconstruction error—how efficient was the network at auto-associating the input patterns. Figure 3 shows that exemplars from the high-variability categories were more difficult to reconstruct accurately *F*(1,78)=553.24, *p*<.0001. More importantly, there was a differential effect of labels on the two corpora *F*(3,76)=550, *p*<.0001. The presence of labels did not affect reconstruction accuracy for the low-variability set, *t*(38)=-.15, but led to a significant increase in the error for the high-variability set *t*(38)=14, *p*<.0001.

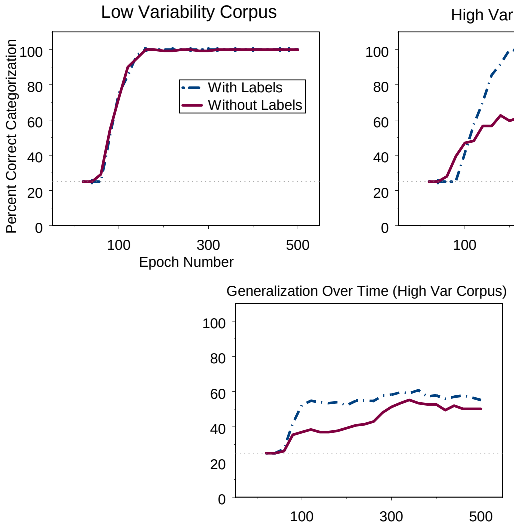

Figure 2: Averaged training profiles for the high and low variability corpora. At no point in training, do labels improve categorization performance for the low-variability categories.

To examine what caused this increase in reconstruction error, the final measure looked at the degree to which the network's representations were clustered around the prototype. This measure was the error between the network's output and the category prototype. Here, labels led to a smaller error (more categorical responses) for both training corpora *F*(1,78)=224.38, *p* < .0001; however, there was also a significant interaction that indicated that labels had more effect for the high-variability corpus than the low-variability one *F*(3,76)=2087, *p*<.0001. The effect of labels on the degree of clustering can be observed more directly by looking at principal component analysis (PCA) plots of the hidden unit activations of the trained network. Figure 4 shows the effect of labels on forming more categorical representations for the high- and low-variability sets. The degree to which exemplar representations were categorical was enhanced for both corpora, but since low-variability unlabeled representations were already quite categorical, the inclusion of labels did not alter the response accuracy.

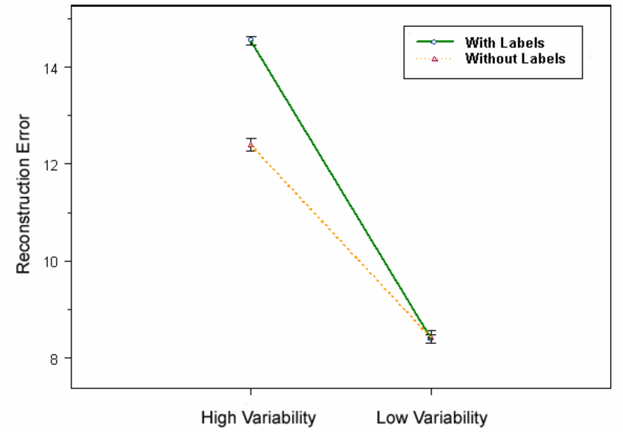

Figure 3: The improved categorization accuracy comes at a cost to accurate representation.

4. Discussion

As predicted, labels improved categorization accuracy of items organized in high-variability, but not low-variability categories. Although training with labels resulted in more categorical representations regardless of the training set, the effect was much greater for the fuzzier categories, suggesting that it is these categories that benefit most from having some of their features abstracted. By associating a single label with numerous exemplars, the resulting category representations of initially ill-formed categories became less idiosyncratic, resulting in improved categorization, and better generalization to unseen exemplars.[ii] In accord with the present findings, Tijsseling and Harnad (1997) found that category labels had no effect on representations when categories were initially very discriminable (e.g., apples and cars). This does not mean, however, that such labels play no role in shaping these category representation. While perceptual properties are quite sufficient to discriminate apples and cars, calling different-looking cars by the same name may help cohere the category, helping individuals to find the features common to its members. The present simulation provides evidence that when presented with a category containing highly-variable exemplars, labels help to highlight common features at the cost of abstracting irrelevant ones.

An intuitive interpretation of these results is that labels made the representations more categorical because training with labels meant that networks were performing a categorization task in addition to the auto-encoding one. One may therefore object to my referring to the category output nodes in the simulation as labels. Although the function of nodes in connectionist simulations is largely a matter of interpretation, the act of assigning a discrete code to a range of exemplars as done here and in other similar situations (Harnad, Hanson, & Lubin, 1991; Tijsseling & Harnad, 1997) may be particularly characteristic of linguistic labels. While categories can be formed without labels, as evidenced by a common response of an organism to a range of stimuli (e.g., Cangelosi, Greco, & Harnad, 2002), such common behavior does not imply a shared representation among the category members. For instance, a well-behaved cat may be said to have formed a category [surfaces in the house that may be scratched]. However, one cannot infer from the cat’s behavior that it is not treating each instance of the category on an individual basis. The learning of a unique category label encourages the formation of a shared (rather than idiosyncratic) representation of the category. Words also allow the combining of concepts. Cangelosi et al. (2000) provide a compelling illustration of the advantage of combining existing grounded symbols into propositions through “symbolic theft” over learning the perceptual properties through “sensorimotor toil.” For instance, knowing [horse] and [stripes] permits a working definition of [zebra] (horse+stripes) without

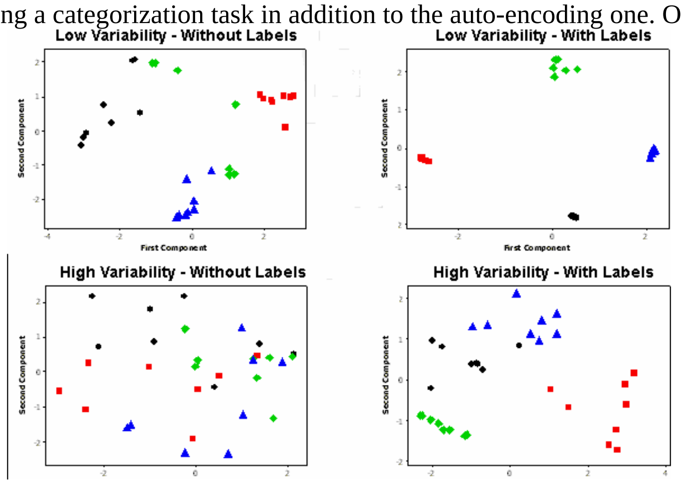

Figure 4: PCA plots of the hidden unit activations for the two traininng corpora and labeling conditions. The different symbols represent the four categories used in training.

having to experience one. The present work suggests that even in the absence of the combinatorial powers of words, providing discrete category labels facilitates the highlighting of relevant features and abstracting of irrelevant ones.

The reason experience with labels led to more categorical representations in the model is simply that each act of labeling was also an instance of categorization. Merely perceiving an object does not require categorizing it. In contrast, naming an object (whether to communicate to another individual or for your own benefit) does require placing it into a category. The repeated experience with categorizing exemplars through labeling has the effect of gradually altering the category representations to be more in line with the named categories. This does not mean that words and language are required to entertain certain thoughts or to hold particular representations (this is an open question, e.g., Hermer-Vazquez, Spelke, & Katsnelson, 1999). Indeed, not making the labels available to the models at test was done expressly to observe the process by which labels may affect representations of exemplars as distinct from using them during recognition.

This reasoning relates directly to work on Categorical Perception (CP) (Harnad, 1987), a phenomenon in which the repeated placement of exemplars into categories exaggerates the perceptual difference between categories (acquired distinctiveness) and collapses perceptual differences within a category (acquired similarity). Categorical perception, unlike language, is not unique to humans (e.g., categorical perception of phonemes in chinchillas, Kuhl & Miller, 1978), and so would appear to not require language. Indeed, the literature on CP rarely makes any mention of the role of words or language. While numerous experiments by Goldstone and colleagues (e.g., Goldstone, 1998 for review) provide evidence that categorization in the laboratory can alter perceptual representations of both familiar and novel dimensions, the role of linguistic labels in this process is not fully specified. However, recent studies have shown that under some circumstances verbal interference can extinguish CP of colors and facial expressions (Roberson & Davidoff, 2000), implicating verbal coding of some kind in the functioning of CP in human adults. Özgen and Davies (2002) provided evidence of how experience with categorization can create new boundaries in the domain of color perception, arguing that lifetime experience with color words heavily influences the resulting color categories. While CP can occur in the absence of language, our ubiquitous experience with named categories suggests that further study of the role of language in category formation will be a useful and productive line of research.

Acknowledgments

Sincere gratitude to Jay McClelland and David Rakison for useful discussion. This material is based upon work supported by a National Science Foundation Graduate Research Fellowship.

References

Cangelosi, A., Greco, A., & Harnad, S. (2000). From robotic toil to symbolic theft: grounding transfer from entry-level to higher-level categories. *Connection Science, 12,* 143-162.

Cangelosi, A., Greco, A., & Harnad, S. (2002). Symbol Grounding and the Symbolic Theft Hypothesis. In A.Cangelosi & D. Parisi (Eds.), *Simulating the Evolution of Language* (London: Springer.

Gentner, D. (1981). Some Interesting Difference between Nouns and Verbs. *Cognition and Brain Theory, 4,* 161-177.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology, 49,* 585-612.

Harnad, S. (1987). *Categorical Perception: The Groundwork of Cognition*. New York: Cambridge University Press.

Harnad, S., Hanson, S. J., & Lubin, J. (1991). Categorical Perception and the Evolution of Supervised Learning in Neural Nets. In D.W.Powers & L. Reeker (Eds.), *Working Papers of the AAI Spring Symposium on Machine Learning of Natural Language and Ontology* (pp. 65-74). Stanford University.

Hermer-Vazquez, L., Spelke, E. S., & Katsnelson, A. S. (1999). Sources of flexibility in human cognition: Dual-task studies of space and language. *Cognitive Psychology, 39,* 3-36.

Homa, D. & Cultice, J. (1984). Role of Feedback, Category Size, and Stimulus Distortion on the Acquisition and Utilization of Ill-Defined Categories. *Journal of Experimental Psychology-Learning Memory and Cognition, 10,* 83-94.

Kersten, A. W. & Smith, L. B. (2002). Attention to novel objects during verb learning. *Child Development, 73,* 93-109.

Kotovsky, L. & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development, 67,* 2797-2822.

Kuhl, P. K. & Miller, J. D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stim. *The Journal of the Acoustical Society of America, 63,* 905-917.

Mareschal, D., French, R. M., & Quinn, P. C. (2000). A connectionist account of asymmetric category learning in early infancy. *Developmental Psychology, 36,* 635-645.

Ozgen, E. & Davies, I. R. L. (2002). Acquisition of categorical color perception: A perceptual learning approach to the linguistic relativity hypothesis. *Journal of Experimental Psychology-General, 131,* 477-493.

Plunkett, K., Sinha, C., Moller, M. F., & Strandsby, O. (1992). Symbol Grounding or the Emergence of Symbols? *Connection Science, 4,* 293-312.

Pothos, E. M. The Rules versus Similarity Distinction. *Behavioral and Brain Sciences,* (in press).

Roberson, D. & Davidoff, J. (2000). The categorical perception of colors and facial expressions: The effect of verbal interference. *Memory & Cognition, 28,* 977-986.

Rohde, D. L. T. (1999). *Lens: The light, efficient network simulator* (Rep. No. CMU-CS-99-164). School of Computer Science, Carnegie Mellon University.

Tijsseling, A. & Harnad, S. (1997). Warping Similarity Space in Category Learning by BackProp Nets. In M.Ramscar, U. Hahn, E. Cambouropolos, & H. Pain (Eds.), *Proceedings of SimCat 1997: Interdisciplinary Workshop on Similarity and Categorization* (pp. 263-269). Edinburgh University.

[i] The reported pattern of results does not change if only one hidden layer is used. The architecture made use of two hidden layers in order to investigate the effects of selective lesioning of one layer versus the other. These results are not discussed here.

[ii] Interestingly, in a study the author was not aware of at the time, Homa and Cultice (1984) reported a strikingly similar pattern of results in a behavioral experiment using adult participants. In investigating the role of feedback on learning categories of varying levels of well-formedness, Homa and Cultice (1984) found that while feedback provided little benefit for well-formed categories, learning ill-defined categories was only possible with category feedback.