The Effect of Labels on Category Learning

Mira Partha
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA, USA
mirap@mit.edu

***Abstract*—It has been shown that perception is affected by grouping perceptual stimuli into categories. Adding labels to the categories appears to facilitate human category learning, as shown in "Language Is Not Just for Talking: Redundant Labels Facilitate Learning of Novel Categories" (Lupyan, Rakison, and McClelland, 2007). In this work, a computational model was developed which attempted to replicate the previously demonstrated effects of labels on category learning. The model learned two categories from stimuli with two relevant dimensions by looking at one example at a time, choosing a category, and then receiving feedback. To understand how labels can facilitate category learning, the model was then adapted to incorporate label-based feedback. Both unlabeled and labeled conditions were tested to see whether they demonstrated the same learning differences (that is, better classification accuracy for label-based learning) observed by Lupyan et al.**

Keywords—learning, verbal labels, perception, categorization, hierarchical models, Bayesian inference

I. Introduction

An interesting phenomenon is that people are faster at distinguishing colors across categories than colors within categories. Why is this the case? Research has shown that grouping perceptual stimuli into categories affects their perception. This effect appears whether or not categories are labeled; however, further research has demonstrated that labels facilitate category learning. Labels act as indicators of shared features, thus encouraging identically labeled objects to be grouped into shared categories. Thus, words can transform category learning into a supervised task. The goal of this project was to develop a computational model to represent the effects of labels on category learning, by using a Bayesian framework to quantify how labels constrain inference and change how prior category distributions are learned.

II. Background

Lupyan et al. (2007) conducted experiments to assess whether the presence of labels facilitated category learning. The first of these experiments observed performance of subjects learning to categorize aliens into two groups, those to be approached and those to be avoided. The stimuli set was comprised of sixteen images of aliens from the YUFO stimulus set, divided into two categories of eight stimuli each. The two categories differed in two relevant dimensions, roundness versus flatness of base, and smoothness versus ridgedness of head.

The training phase was conducted in nine blocks of sixteen trials each. In each trial, one alien from the stimulus set was presented, and subjects chose whether to approach or avoid the alien. After a short delay, auditory feedback on the correctness of the response was provided. For the label condition, an additional printed label, denoting the aliens as either "leebish" or "gracious," was provided for additional feedback before the conclusion of each trial. The results of the experiment are shown below.

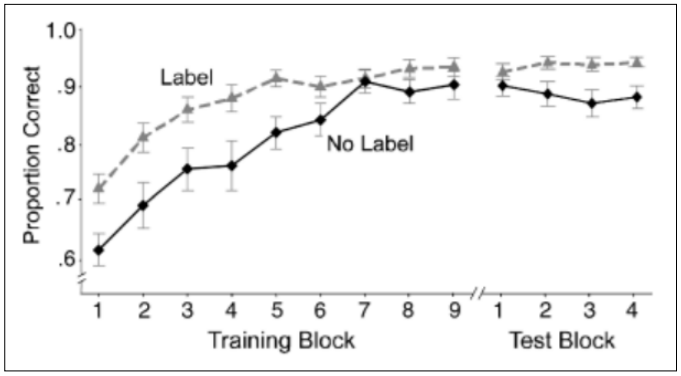

Fig. 1. Mean classification accuracy in the training and test phases of Experiment 1 from Lupyan et al. (2007).

In summary, both conditions saw improvements in performance over time. Significantly, the label group was demonstrably more accurate, and also learned the categories much more quickly.

In this work, a computational model is presented, that attempts to replicate the learning differences found above.

III. Model Implementation

To model the category learning, a hierarchical generative model was used. The model uses observed examples to infer category parameters, that is, what a prototypical example for each category would look like. Using this inferred distribution, the model could predict a category given a feature vector. The model would then perform a Bayesian update by conditioning on the received feedback. This procedure was conducted iteratively to create the training phase results.

A. Stimuli Set

Analogous to the subset of the YUFO stimuli set utilized by Lupyan et al., a set of sixteen bi-dimensional feature vectors was generated. The two dimensions of the feature vectors were intended to correspond to the roundness versus flatness of base, and smoothness versus ridgedness of head, of the YUFO stimuli. For each feature, values were randomly sampled from Gaussian distributions with the same standard deviation but a different mean for each category. Each feature vector was associated with the category from whose distributions it was drawn, either "approach" or "avoid", eight to each category. In the label condition, each feature vector was additionally associated with one of two nonsensical labels, "leebish" or "grecious," corresponding to the category it belonged to.

B. No Label Model

Data was modelled using a mixture model, learning the parameters for each category while receiving feedback on the assignment of stimuli to categories in each trial. To start, the model was constructed assuming that each stimulus was drawn from one of $K = 2$ categories, with uniform prior category probabilities. z_i indicates which category object i belongs to. Categories and labels were drawn from Categorical distributions, with concentration parameter αs set to 0.1 (near-zero, since tautologically objects of the same category should be categorized together; and the labels, as category names, functioned identically).

$$\vec{\pi}_{category} \sim \text{Dirichlet}(\vec{\alpha} = [1,1])$$

$$\alpha_{cat} = 0.1$$

$$\vec{\pi}_{cat} \sim \text{Dirichlet}(\vec{\alpha} = [\alpha_{cat}, \alpha_{cat}])$$

$$z_i \sim \text{Categorical}(\vec{p} = \vec{\pi}_{category}, \vec{v} = [0,1,2])$$

$$\mu_{feature\ 1} \sim \text{Normal}(\mu = 0, \sigma = 2)$$

$$\mu_{feature\ 2} \sim \text{Normal}(\mu = 0, \sigma = 2)$$

$$\text{feature 1} \sim \text{Normal}(\mu = \mu_{feature\ 1}, \sigma = 0.2)$$

$$\text{feature 2} \sim \text{Normal}(\mu = \mu_{feature\ 2}, \sigma = 0.2)$$

$$\text{category} \sim \text{Categorical}(\vec{p} = \vec{\pi}_{cat}, \vec{v} = [\text{"approach"},\text{"avoid"}])$$

C. Incorporating Label-Based Feedback

To accommodate the label assignments and label feedback, the model for the label condition additionally included the following:

$$\alpha_{label} = 0.1$$

$$\vec{\pi}_{label} \sim \text{Dirichlet}(\vec{\alpha} = [\alpha_{label}, \alpha_{label}])$$

$$\text{label} \sim \text{Categorical}(\vec{p} = \vec{\pi}_{label}, \vec{v} = [\text{"leebish"},\text{"grecious"}])$$

Given the hierarchical set of priors and the likelihood, the model infers which aliens are in which categories, the parameters of the distributions that define each category, and the distributions over the category parameters.

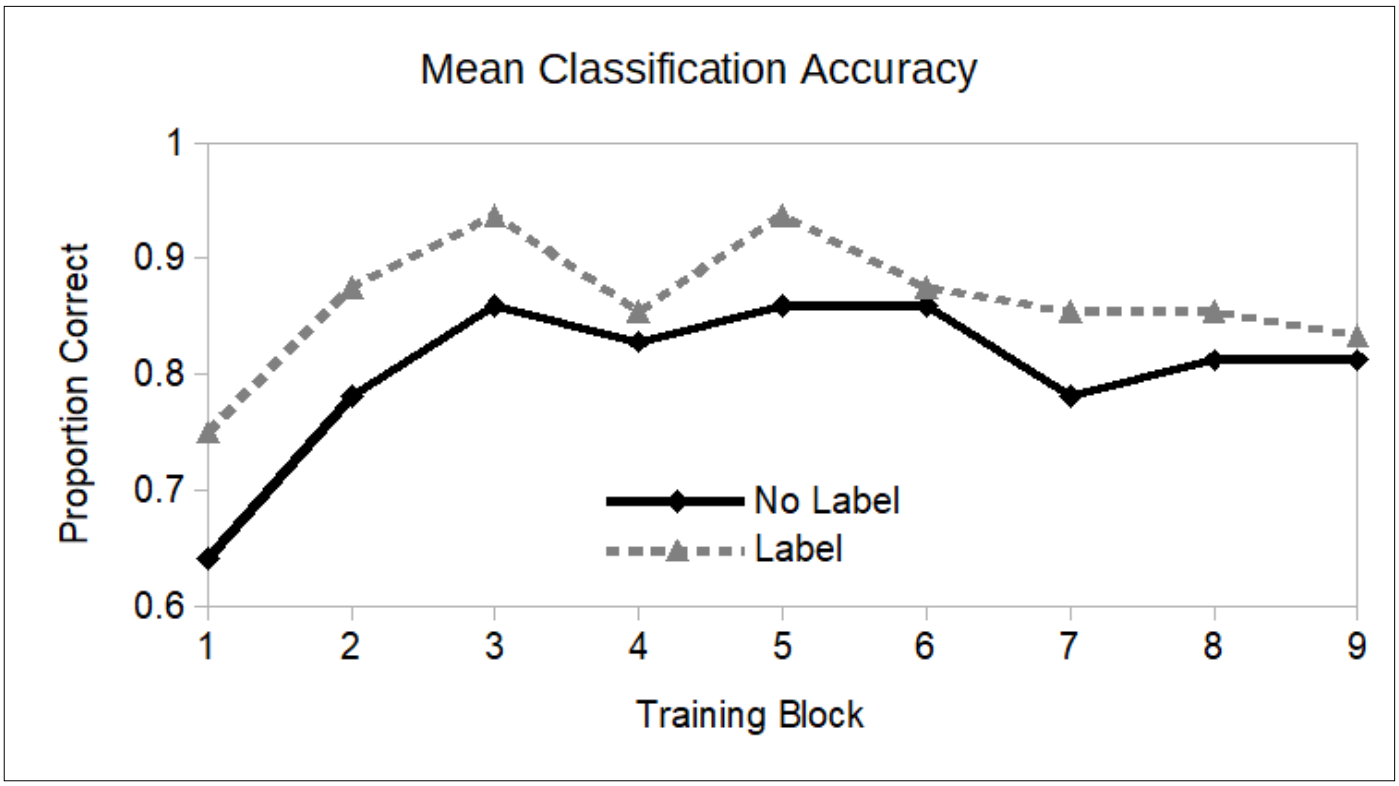

Fig. 2. Mean classification accuracy of the implemented computational model, showing learning differences between the label and no label conditions.

IV. Results

Similar to the results found by Lupyan et al., the label condition was overall more accurate than the no-label condition, though the difference was less significant. The label condition also learned categories faster; for instance, the label condition reached the 86%-correct performance threshold in approximately sixteen trials fewer than the no-label condition.

Peculiarly, however, classification accuracy did not continue to rise with the latter half of the training blocks, even descending slightly before plateauing in both conditions.

Groups of correlated associations reinforce each other. In this model, labels were associated with stimuli, but also completely correlated with the categorization ('approach' or 'avoid'). Thus, the labels would have reinforced the association between the stimuli and the categorizations. This impact of labels on category learning manifested in the results produced by the computational model, though not as starkly as in the human experiments.

V. General Discussion

The results demonstrated by the implemented computational model are limited in scope. For one, generalization to previously unseen stimuli remained untested (no testing phase). However, it is expected that the label condition would continue to perform better than the no label condition, even without receiving additional feedback.

There are a number of aspects of the implementation of this computational model, that render it representationally different from the human experiments conducted by Lupyan et al. (2007). One, the stimuli set was designed with much less nuance. The YUFO stimuli used in the human experiments were complex images, and human subjects would begin by considering an immense number of features before discerning which features were most relevant to the categorization problem at hand. For the computational model, the stimuli set used was represented as simple bi-dimensional feature vectors, with the only two relevant features ever

being used. However, this probably would not impact the results, since irrelevant features would remain uniform through the Bayesian update process; only computation time would be impacted.

The values for $\alpha_{category}$ and α_{label} were required to be slightly above zero. This is contrary to human cognition, wherein subjects would know with complete certainty that categories are comprised of all aliens to approach or all aliens to avoid, and similarly with the labels of "leebish" or "grecious." However, the design of this particular computational model involved a non-human feature that may have offset this difference slightly. For the update process, rather than updating the category prototypes and distribution parameters by conditioning with just the current observed stimulus, the model instead re-generated these based on all past observed examples, during each trial. This would be equivalent to a human subject with infinite memory (and impressive computational capability – the closest human analogue is, perhaps, card-counting.) The utilized inference methods also had to be MCMC, rather than enumeration, due to the infinite support of Gaussian distributions. This probably contributed to more non-human results.

The computational model was also limited in that it could not distinguish between verbal versus non-verbal labels. It therefore cannot be leveraged to support the linguistic relativity hypothesis. However, it did allow for isolation of effects specific to labels rather than learning any additional association because of the near-zero α_{label} value, which constrained categories to consist of only one label value.

The evidence of learning differences between label and no label conditions produced by the implemented computational model lend credence to particular theories over others. The correlation of a simplified perceptual distinction with a complex multi-dimensional feature space is suitable for human experiments, but less relevant to a computational model. This implementation of the computational model, with labels that were purely redundant to the category feedback, also provided no evidence for the law of dissociation by varying concomitants. The application of Miller and Dollard's hypothesis (1941) to this model becomes somewhat more interesting, though. Experiments to test their hypothesis failed to attribute increased discriminability definitively to one of either learned associations between labels and stimuli, or increased experience with the stimuli. The curious performance decreases that appeared in the computational model's results might suggest that (at least for this particular case) the learned associations between labels and stimuli somehow counteract the additional experience with the stimuli.

Overall, however, from the results of the computational model, it appears conclusive that labels contribute to more robust category attractors, because they become associated with the most relevant features as they are paired with individual examples. Even when the labels provide no additional information, the labeled categories appear easier to learn than unlabeled categories.

Acknowledgments

Many thanks for the guidance of Anna Ivanova, graduate student of the MIT Department of Brain and Cognitive Sciences. Also, many thanks to the course staff of 9.660: Professor Tenenbaum, Luke Hewitt, Matthias Hofer, Jon Gauthier, and Max Nye.

References

[1] Cibelli E, Xu Y, Austerweil JL, Griffiths TL, Regier T (2016) The Sapir-Whorf Hypothesis and Probabilistic Inference: Evidence from the Domain of Color. PLoS ONE 11(7): e0158725. https://doi.org/10.1371/journal.pone.0158725

[2] Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. Psychological Review, 116(4), 752-782. http://dx.doi.org/10.1037/a0017196

[3] L Gilbert, Aubrey & Regier, Terry & Kay, Paul & Ivry, Richard. (2006). Whorf hypothesis is supported in the right visual field but not the left. Proceedings of the National Academy of Sciences of the United States of America. 103. 489-94. 10.1073/pnas.0509868103.

[4] Holmes, K. J., & Wolff, P. (2012). Does categorical perception in the left hemisphere depend on language? Journal of Experimental

Psychology: General, 141(3), 439-443. http://dx.doi.org/10.1037/a0027289

[5] Kemp, Charles & Perfors, Amy & B Tenenbaum, Joshua. (2007). Learning Overhypotheses with Hierarchical Bayesian Models. Developmental science. 10. 307-21. 10.1111/j.1467-7687.2007.00585.x.

[6] Lupyan, Gary & H Rakison, David & Mcclelland, James. (2008). Language is not Just for Talking Redundant Labels Facilitate Learning of Novel Categories. Psychological science. 18. 1077-83. 10.1111/j.1467-9280.2007.02028.x.

[7] N. D. Goodman and A. Stuhlmüller (electronic). The Design and Implementation of Probabilistic Programming Languages. Retrieved from http://dippl.org.