Categories with Mismatching Feature Saliency and Diagnosticity: How a Label Helps Learning

Arthur Capelier-Mourguy*[†], Katherine E. Twomey[†], Vanja Kovic[‡] and Gert Westermann[†]
[†] Department of Psychology, Lancaster University
Emails: a.capelier-mourguy@lancaster.ac.uk
k.twomey@lancaster.ac.uk g.westermann@lancaster.ac.uk
[‡] Department of Psychology, University of Belgrade
Email: vanja.kovic@f.bg.ac.rs

From a very young age infants form categories to simplify the world they encounter. They do so by relying both on the physical features of objects and the labels with which they are paired. A key question in understanding the formation of labeled categories both in children and adults is how perceptual features and labels interact. Here we explore the mechanisms of these interactions depending on the saliency of diagnostic features. We used a simple auto encoder neurocomputational model to capture a recent empirical study [1] of adult categorization. In this task adults were given a 5-4 categorization task [2] with animal-like drawings with either salient or non-salient diagnostic features. Labels supported categorization in both conditions, while in the non-labeled condition learning was significantly more difficult when the diagnostic features were less salient. In other words, adding a label improved learning only when the diagnostic features were of low saliency. Recent theoretical accounts of the status of labels in categorization can explain these results. On the *Labels-as-Features* (LaF) [3] account, labels are part of object representations in the same way as color or shape. LaF therefore predicts that labels will interact with other features depending on their diagnosticity and saliency, boosting categorization when diagnosticity is low. In contrast, the *Label-as-Referent* (LaR) [4] account assumes that labels shape perceptual representations in a top-down manner. Specifically, LaR argues that a label will drive attention towards diagnostic features, highlighting these features as an "invitation to form categories". In the context of the current task this process will be redundant when features are already highly salient but will lead to significant improvements in categorization when they are of lower saliency.

To tease apart these competing explanations we implemented both views in an auto-encoder model of category learning. As is typical in such models, we recorded the Sum Squared Error (SSE) of the model to determine how well the category was learnt. We used the number of units encoding a feature as a proxy for its saliency: the more units code for a feature in the model, the more impact this feature will have, thus being more "salient ".

For the LaF version of the model we set the label as an input and an output feature, with the exact same status as other features, differing only in the number of units use to encode it. Specifically, the label was as salient as the most salient visual feature, in line with the current research suggesting that labels — and possibly all auditory features — are particularly salient [5], [6]. For the LaR model, we provided the network with a "goal" label unit on its output layer only. Initially, this allowed the model to learn to associate a set of features with the given label, the label acting as a unifying goal for different exemplars of the same category. At test, the label is activated by an exemplar of the corresponding category. In line with LaR, in this implementation of the label remains qualitatively different from the other features.

Only the LaF model could replicate the pattern of results in [1], with no difference in looking time to a category prototype whether the diagnostic features were salient or not. Overall, the current study suggests that adults, at least in the experimental conditions in [1], exhibit categorization capacities suggesting that they represent labels as features. In contrast to studies which suggest that adults consider labels category markers [7], this result is consistent with a recent study showing that some adults might as well consider labels as features [8].

References

[1] V. Kovic, K. Plunkett, and G. Westermann, "Heads or Labels? Labels and visual features act together in category formation," *Cognitive Science*, under review.
[2] D. L. Medin and M. M. Schaffer, "Context theory of classification learning." *Psychological review*, vol. 85, no. 3, 1978.
[3] V. M. Sloutsky and A. V. Fisher, "Induction and Categorization in Young Children: A Similarity-Based Model." *Journal of Experimental Psychology: General*, vol. 133, no. 2, 2004.
[4] S. Waxman, "Words as Invitations to Form Categories: Evidence from 12- to 13-Month-Old Infants," *Cognitive Psychology*, vol. 29, no. 3, Dec. 1995.
[5] V. M. Sloutsky and A. C. Napolitano, "Is a Picture Worth a Thousand Words? Preference for Auditory Modality in Young Children," *Child Development*, vol. 74, no. 3, May 2003.
[6] V. M. Sloutsky and C. Robinson, "The Role of Words and Sounds in Infants' Visual Processing: From Overshadowing to Attentional Tuning," *Cognitive Science: A Multidisciplinary Journal*, vol. 32, no. 2, Mar. 2008.
[7] T. Yamauchi, N. Kohn, and N.-Y. Yu, "Tracking mouse movement in feature inference: Category labels are different from feature labels," *Memory & Cognition*, vol. 35, no. 5, Jul. 2007.
[8] W. Deng and V. M. Sloutsky, "Carrot Eaters or Moving Heads: Inductive Inference Is Better Supported by Salient Features Than by Category Labels," *Psychological Science*, vol. 23, no. 2, Feb. 2012.