Learning Abstract Categories through Linguistic Feature Explanations

Haodong Xie[1], **Xuena Wang**[2], **Rahul Singh Maharjan**[1], **Federico Tavella**[1] **& Angelo Cangelosi**[1]
[1]Centre for Robotics and AI, The University of Manchester
[2]The School of Psychology and Cognitive Science, East China Normal University

Abstract

Abstract categories often lack a single, bounded perceptual referent, making them harder to learn than highly concrete categories. Explanations that identify properties shared across category members may support category learning by highlighting features that generalise beyond individual examples. This study investigates the role of explanatory concepts in category learning using a label-free Concept Bottleneck Model framework. We generate human-interpretable feature concepts for both basic-level and super-ordinate categories, then learn a projection from visual representations into this concept space. A classifier then predicts category labels from these concept activation to evaluate whether explanatory features support category learning. We evaluate the approach on CIFAR-100, CUB-200, and ImageNet, organising labels into basic-level and super-ordinate categories. Across datasets, concept-based models achieve accuracy comparable to standard classifiers while providing interpretable concepts. Visualisations reveal distinct category clusters in concept space, and evaluations on held-out subordinate classes suggest that explanatory concepts transfer effectively to previously unseen subclasses.

Keywords: Concept Learning; Abstract Concept; Concept Bottleneck Model

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*