

UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Can Abstract Categories Be Represented by Shared Features in Concept Bottleneck Models?

Permalink

<https://escholarship.org/uc/item/8vw6p3n7>

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors

Xie, Haodong

Wang, Xuena

Maharjan, Rahul Singh

et al.

Publication Date

2025

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Can Abstract Categories Be Represented by Shared Features in Concept Bottleneck Models?

Haodong Xie

University of Manchester, Manchester, United Kingdom

Xuena Wang

Tsinghua University, Beijing, China

Rahul Singh Maharjan

University of Manchester, Manchester, United Kingdom

Federico Tavella

University of Manchester, Manchester, United Kingdom

Angelo Cangelosi

University of Manchester, Manchester, United Kingdom

Abstract

Learning abstract concepts is a core component of human cognition, yet remains challenging for artificial intelligent. We present a computational model that investigates whether abstract categories can be acquired through shared perceptual features, using a Label-Free Concept Bottleneck Model (CBM) trained to induce basic-level concepts using shared features. Concepts are represented through intermediate concepts layer, enabling the model to form grounded representations of basic-level categories. To evaluate conceptual robustness beyond surface-level accuracy, we conduct a series of generalization and ablation experiments. These assess whether the model forms robust conceptual representations rather than merely mapping inputs to labels. Our results show that the CBM achieves high accuracy on a dataset comprising four basic-level classes and twelve subordinate Image-Net subclasses, while also yielding interpretable intermediate representations. This framework demonstrates that abstract categorization can emerge through feature based induction, and suggests a pathway for cognitive models of concept learning.