LEARNING GENERAL, ADAPTIVE AND HUMAN-INTERPRETABLE CONCEPTS THROUGH SITUATED INTERACTIONS

JENS NEVENS*[1], PAUL VAN EECKE[1], and KATRIEN BEULS[1]

*Corresponding Author: jens@ai.vub.ac.be
[1]Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Brussels, Belgium

For communicating and reasoning about the world in which we live, we rely on a repertoire of concepts that form a symbolic abstraction layer over our continuous sensori-motor experiences. For example, the cones in our eyes can convert the whole visible colour spectrum into nerve impulses, but we communicate and reason about abstractions over this spectrum, using concepts like RED, YELLOW and GREENISH BLUE. There is overwhelming evidence that these concepts are not universal or innate (Levinson, 2000), but that they can emerge and evolve through experiences and interactions in the world (Steels & Belpaeme, 2005; Bleys, 2015). The research on which we report here contributes to this view, by presenting computational simulations of how meaningful concepts can be distilled from streams of sensori-motor data through a series of situated communicative interactions. Crucially, the concepts that are learned are interpretable, adaptive to changes in the world, and general enough to be applicable to previously unseen objects.

Our approach builds further on earlier work within the language game paradigm (Steels, 2001). In this work, the concepts that were learned were either limited to continuous data on a single feature channel, such as colour (Bleys, 2015) or spatial position (Spranger, 2012), or to non-continuous data on multiple feature channels (Wellens, 2008). Here, we lift both limitations at the same time and investigate how concepts can be distilled from a larger number of continuous-valued feature channels. Our approach radically differs from other recent work, which applies deep learning techniques to concept learning, e.g. (Dolgikh, 2018; Shi, Xu, Yao, & Xu, 2019). The models resulting from these techniques are often high-performing, but require huge amounts of training data, yield concepts that are not human-interpretable, and require a partial or complete re-training of the neural network in order to adapt to changes in the world.

For our purposes, we set up a series of tutor-learner experiments, which each explore a different concept learning strategy. The experiments consist of a large number of tutor-learner interactions, which are set in a world based on the CLEVR dataset (Johnson et al., 2017). This world consists of scenes containing geometrical objects, which differ in horizontal and vertical position, colour, material, shape and size. During each interaction, the tutor uses a single word to refer to an object in the scene. The task of the learner is to point to the object to which the tutor referred. At the end of the interaction, the learner receives feedback on the outcome of the task, and the tutor points to the correct object if the learner was wrong.

The learner observes the scenes through continuous-valued human-interpretable feature channels, such as ‘area’, ‘width-height-ratio’, or ‘position-on-x-axis’. Depending on the specific experiment, the feature values are obtained through simulation or via a neural network model for object detection and segmentation (He, Gkioxari, Dollár, & Girshick, 2017; Yi et al., 2018). For each concept, the learner must simultaneously learn which feature channels are important to which extent, and what the prototypical value for each channel is. Figure 1 (a) and (b) present concepts that were learned using simulated and extracted data respectively, showing the weight and prototypical value of each relevant feature channel. Figure 1 (c) shows how the communicative success increases with the number of interactions that take place, using simulated (green line) and extracted (yellow line) features. After around 1000 interactions, a stable conceptual system is in place, achieving communicative success in 100% of the interactions using the simulated data and in nearly 85% of the interactions using the extracted data.

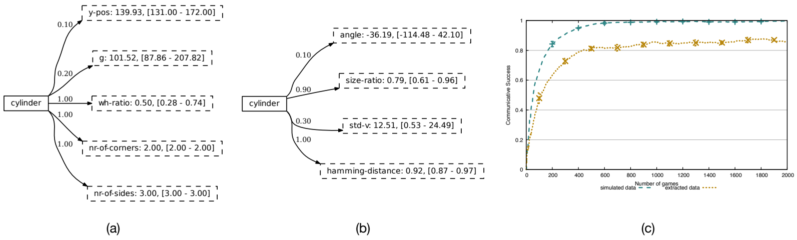

Figure 1. Learned concepts for “cylinder” using simulated (a) and extracted (b) data; communicative success over time, using simulated (green) and extracted (yellow) features (c).

In order to investigate the generality and adaptivity of the concepts that are learned, we have carried out an additional experiment. In the first phase of the experiment, the tutor and learner were placed in an environment that exhibited certain biases (condition A - e.g. cubes are always red or brown). After the conceptual system of the learner had stabilised, we then changed the biases in the world (condition B - e.g. cylinders are always blue or green). The results show no drop in communicative success when transitioning from A to B, indicating that the learned concepts are not affected by these environmental co-occurrences.

In sum, we have used computational simulations to show how human-interpretable concepts can be distilled from parallel streams of continuous sensori-motor data through repeated communicative interactions, and have demonstrated that these concepts can adapt to changing environmental conditions.

References

Bleys, J. (2015). *Language strategies for the domain of colour.* Berlin: Language Science Press.

Dolgikh, S. (2018). Spontaneous concept learning with deep autoencoder. *International Journal of Computational Intelligence Systems*, *12*(1), 1–12.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *Proceedings of the ieee international conference on computer vision* (pp. 2961–2969). Honolulu, Hawaii.

Johnson, J., Hariharan, B., Maaten, L. van der, Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2901–2910). Honolulu, Hawaii.

Levinson, S. C. (2000). Yélî dnye and the theory of basic color terms. *Journal of Linguistic Anthropology*, *10*(1), 3–55.

Shi, J., Xu, J., Yao, Y., & Xu, B. (2019). Concept learning through deep reinforcement learning with memory-augmented neural networks. *Neural Networks*, *110*, 47–54.

Spranger, M. (2012). Potential stages in the cultural evolution of spatial language. In *The evolution of language* (pp. 328–335). World Scientific.

Steels, L. (2001). Language games for autonomous robots. *IEEE Intelligent systems*, *16*(5), 16–22.

Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and brain sciences*, *28*(4), 469–488.

Wellens, P. (2008). Coping with combinatorial uncertainty in word learning: A flexible usage-based model. In *The evolution of language* (pp. 370–377). World Scientific.

Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., & Tenenbaum, J. (2018). Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *Advances in neural information processing systems* (pp. 1031–1042). Montreal, Canada.