From Continuous Observations to Symbolic Concepts: A Discrimination-Based Strategy for Grounded Concept Learning

Jens Nevens, Paul Van Eecke, and Katrien Beuls

Artificial Intelligence Lab, Vrije Universiteit Brussel
Pleinlaan 2, B-1050 Brussels, Belgium
{jens|paul|katrien}@ai.vub.ac.be

In order to reason and communicate about their environment, autonomous agents need to be able to distill meaningful concepts from the observed streams of continuous sensori-motor data. In this paper, we report on computational simulations of how such concepts are distilled through a series of situated communicative interactions. Our approach builds further on earlier work within the language game paradigm [5], where concepts were either limited to continuous data on a single feature channel (e.g. [1]) or to non-continuous data on multiple feature channels (e.g. [6]). We lift both restrictions at the same time. Through a tutor-learner scenario, our novel method allows an agent to construct meaningful concepts which are formed by discriminative combinations of prototypical values on human-interpretable feature channels. Most current approaches that bridge between the continuous and symbolic domain make use of deep learning techniques (e.g. [2]). These approaches often achieve high levels of accuracy but they rely on large amounts of training data, the resulting models lack transparency and they require partial or complete re-training to accommodate changes in the environment.

The experiments are set in an environment based on the CLEVR dataset [4]. This environment consists of scenes with geometrical objects of different colours, shapes, sizes and materials. In each interaction, the tutor uses a single word to refer to one of the objects, e.g. *"sphere"*. The learner observes the scene through continuous-valued and human-interpretable feature channels, such as 'area', 'number-of-corners' or 'width-height-ratio'. These features are obtained through simulation (simulated world setting) or object detection, segmentation and feature extraction techniques [3] (noisy world setting). The task of the learner is to point out the object meant by the tutor. It does so by computing the similarity between each object and the current representation of the concept that is associated with the word form uttered by the tutor. At the end of the interaction, the learner receives feedback on whether or not it was correct and the tutor points out the correct object. Using this information, the learner can update the concept it used. More specifically, the learner rewards the most discriminating subset of feature channels and punishes the others. Additionally, all prototypical values are shifted slightly towards the object. For each concept, the learner must simultaneously learn which feature channels are important and what their prototypical values should be. Figure 1 shows the communicative success of the agents and an example of a learned concept.

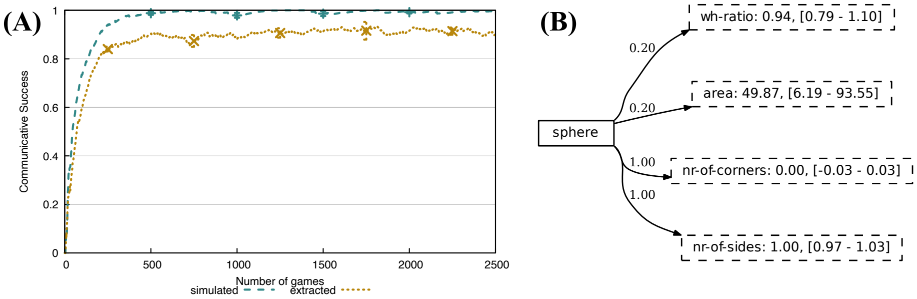

Fig. 1. **(A)** The agent achieves 100% communicative success in the simulated world and 91% in the noisy world **(B)** Concepts are represented through a weighted set of attributes. The weight corresponds to the certainty of an attribute belonging to the concept. Each attribute is modelled as a normal distribution that keeps track of its prototypical value (the mean) and the standard deviation. The weighted sets capture discriminative combinations of attributes. The concept SPHERE focusses on attributes related to shape.

Through a range of experiments, we showcase several desirable properties of our approach. The first experiment shows that the agent rapidly adapts to changes in the environment and the approach allows for incremental learning. In the second experiment, we demonstrate that the concepts generalise well to unseen settings. Finally, we show that the concepts can be learned even when combined compositionally. These properties, combined with fast and data-efficient learning and human-interpretable representations, make our approach well-suited to be used in robotic agents for mapping continuous sensory input to grounded, symbolic concepts. These can in turn be used for higher-level reasoning tasks, such as navigation, (visual) question answering and action planning.

References

1. Bleys, J.: Language strategies for the domain of colour. Language Science Press, Berlin (2015)
2. Dolgikh, S.: Spontaneous concept learning with deep autoencoder. International Journal of Computational Intelligence Systems **12**(1), 1–12 (2018)
3. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969. Honolulu, Hawaii (2017)
4. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: 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 (2017)
5. Steels, L.: Language games for autonomous robots. IEEE Intelligent systems **16**(5), 16–22 (2001)
6. Wellens, P.: Coping with combinatorial uncertainty in word learning: A flexible usage-based model. In: The Evolution Of Language, pp. 370–377. World Scientific (2008)