Human-Interpretable Grounded Language Processing

Author: Liesbet De Vos[1]
Supervisors: Katrien Beuls[2] and Paul Van Eecke[3]

[1] Catholic University of Leuven
liesbet.devos@hotmail.com

[2] University of Namur
katrien.beuls@unamur.be

[3] Artificial Intelligence Lab Brussels
paul.ai.vub.ac.be

Grounded language processing is a central component in many artificial intelligence systems as it allows agents to communicate about their physical surroundings. Given its importance, tasks involving this issue are researched extensively, typically using deep learning techniques that perform end-to-end mappings between natural language expressions and representations grounded in the environment. A popular task for measuring grounded language processing is visual question answering, as it requires grounding of natural language questions into their accompanying image. While neural architectures achieve high state-of-the-art results on VQA benchmarks, they are often criticized for their reliance on large amounts of training data, their closed nature, and their lack of interpretability (Agrawal et al., 2016). As an alternative, this thesis proposes a fully explainable, data-efficient architecture for open-ended grounded language processing that can be applied to visual question answering.

The architecture is based on two main components. The first component comprises an inventory of human-interpretable concepts learned through task-based communicative interactions, as proposed by Nevens et al. (2020). In each interaction, a tutor and learner agent work together to complete a common task and reach communicative success. To do so, the learner must identify the topic-object in the scene, referred to by the tutor using a visual property that maximally discriminates it from the other objects in the scene. Through these interactions, the learner learns which feature streams are important for each concept, and what their prototypical values are. The resulting concepts therefore connect the continuous sensorimotor experiences of an agent to meaningful symbols that can be used for reasoning operations. The second component concerns a computational construction grammar that maps between natural language expressions and procedural semantic representations. This grammar was implemented by Nevens et al. (2019) using the framework of fluid construction grammar (FCG), and contains 170 constructions that were tailored specifically for the CLEVR visual question answering benchmark (Johnson et al., 2017). The outputted representations are grounded in the environment through their integration with the learned inventory of human-interpretable concepts.

The CLEVR-dataset (Johnson et al., 2017) provides a benchmark for visual question answering that limits statistical bias and measures true visual reasoning. It does so by rendering synthetic images of geometric scenes, where co-occurences of different shapes, colors or materials are evenly balanced (Johnson et al., 2017). To evaluate the architecture proposed in this thesis, a variation on this CLEVR benchmark was used where each visual scene is represented as a combination of continuous features. These features were either extracted using a Mask-R-CNN or simulated using the ground truth annotations provided by CLEVR. Since these simulated features are not based on the images themselves, the obtained results in this condition cannot be compared directly to others on the CLEVR benchmark. Instead, they serve as an indicator for the model's performance in an ideal and noise-free environment.

Results of the evaluation show high results on the simulated features (96% accuracy). In contrast to most neural architectures, our system can abstain itself from answering questions it does not know the answer to, which increases its reliability. This is shown by the fact that in 84% of all incorrect cases, the system did not provide an answer. The high results on the simulated features demonstrate that the integration of a computational construction grammar with an inventory of explainable grounded concepts can effectively achieve fully human-interpretable grounded language processing.

Bibliography

Agrawal, A., Batra, D., and Parikh, D. (2016). Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. (2017). CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2910.

Nevens, J., Van Eecke, P., and Beuls, K. (2019). Computational construction grammar for visual question answering. *Linguistics Vanguard*, 5(1):20180070.

Nevens, J., Van Eecke, P., and Beuls, K. (2020). From continuous observations to symbolic concepts: A discrimination-based strategy for grounded concept learning. *Frontiers in Robotics and AI*, 7:84.