Embodied Affordance Grounding using Semantic Simulations and Neural-Symbolic Reasoning: An Overview of the PlayGround Project

Andreas Persson[1,†], Amy Loutfi[1]

[1]*Center for Applied Autonomous Sensor Systems (AASS), Department of Science and Technology, Örebro University, SE-701 82 Örebro, Sweden*

Abstract

In this paper, we present a synopsis of the PlayGround project. Through neural-symbolic learning and reasoning, the PlayGround project assumes that high-level concepts and reasoning processes can be used to advance both symbol grounding and object affordance inference. However, a prerequisite for reasoning about objects and their affordances is integrated object representations that concurrently maintain symbolic values (e.g., high-level concepts), and sub-symbolic features (e.g., spatial aspects of objects). Integrated representations that, preferably, should be based upon neural-symbolic computation such that neural-symbolic models can, subsequently, be used for high-level reasoning processes. Nevertheless, reasoning processes for symbol grounding and affordance inference often require multiple inference steps. Taking inspiration from the cognitive prospects in simulation semantics, the PlayGround project further presumes that these reasoning processes can be simulated by neural rendering complementary to high-level reasoning processes.

Keywords

Symbol Grounding, Semantic World Modeling, Affordance Inference, Semantic Simulation, Neural-Symbolic Reasoning

1. Introduction

The meaning of an object goes beyond the symbol used to refer to it. One of the tenets of intelligence is the ability to solve the symbol grounding problem, also known as the representation grounding problem. In *symbol grounding* [1, 2], it is argued that purely computational symbol manipulation cannot gain true meaning without reference to an agent's embodied interaction with the world. Said differently, the symbol grounding problem describes the ability to map words in the language to aspects of the external world. State-of-the-art research can produce computational models of language and vision that enable us, to some extent, to caption images, answer questions about images, and generate an image from a natural language description.

[†] Corresponding author.

andreas.persson@oru.se (A. Persson); amy.loutfi@oru.se (A. Loutfi)

https://www.oru.se/english/employee/andreas_persson (A. Persson); https://www.oru.se/english/employee/amy_loutfi (A. Loutfi)

CEUR Workshop Proceedings (CEUR-WS.org)

Significant progress has been made in each of these tasks. However, one essential task for robotic systems – and the "holy grail" of language grounding – is to be able to discern impossible goals from possible ones. For instance, a statement like *"the coffee inside the mug"* is clearly discernable from *"the mug inside the coffee"*. This is what we call *affordability inference* as what is discernable is determined by affordances.

Figure 1: Typical examples of affordability inference, given various perplexity: **1)** an *"apple"* that affords being cut into smaller sized wedges, **2)** a *"mug"* that affords the containment of smoking hot beverages, and **3)** an *"apple-mug"*, which by reasoning based on the previous two examples, would afford both the containment of smoking hot beverages, as well as being cut into wedges.

Affordability inference is a task that comes relatively easily to us humans. Our ability to perform affordability inferences stems from the fact that we are able to index words and phrases to objects in the world to prototypical symbols of those objects. Once we derive affordances from those objects, such affordances constrain the way objects can be coherently combined – they determine what is possible and what is not. Affordability inference is, however, a task that, in many ways, is challenging, saying the least, for robotic systems, as exemplified in Figure 1. Cognitive scientists have long advocated that the mechanisms by which affordability inference is possible are through a process called *simulation semantics* – the process by which we understand and *reason* about utterances by *simulating* their content, using similar constructs to both perception and control. Reasoning about affordances can, arguably, proceed at a high level using symbolic representations. However, sub-symbolic representations are needed to capture the 3D spatial aspect of objects, as 2D representations are inherently limited in both capturing adequate dimensional features and features that are invariant to camera motions [3].

An underlying assumption of the PlayGround project is that high-level concepts and reasoning processes can be leveraged to improve symbol grounding and affordance inference. This assumption implies integrated representations, administrated by a symbolic – sub-symbolic framework, that can cope with both the high-level concepts (symbolic), as well as 3D spatial aspects of objects (sub-symbolic). Furthermore, this symbolic – sub-symbolic framework should, intuitively, be based upon neural-symbolic computation such that neural-symbolic models, subsequently, can be used for high-level reasoning processes. Based on the observation that reasoning processes for symbol grounding and affordance inference often require multiple inference steps, a further assumption of the PlayGround project is that these processes can be simulated (in alignment with the cognitive perspective on simulation semantics). It is, therefore, natural to develop a simulation framework for reasoning about symbol grounding and affordances. The simulation framework will use neural rendering to generate possible 3D scenes and sequences of such scenes, given language statements and instructions. This semantic simulation

renders low-level 3D scenes and is, hence, complementary to the high-level reasoning processes, supported by the neural-symbolic approach.

In summary, the overall goal of the PlayGround project is to *"[...] contribute novel techniques for affordance inference and for symbol grounding that are based on **1)** an integrated symbolic – sub-symbolic framework, and **2)** a semantic simulation framework."*

2. Fundamentals and Related Work

Symbol grounding for physically embedded systems has followed several different tracks. One track learns the meaning of words in the sensorimotor space of the robot using neural networks. Typically, features are extracted from the perceptual data, and the output of the network is tightly coupled to the exact motor configuration of the robot [4]. Another approach is to manually create a symbol system and structures for maintaining the percept-symbol correspondence [5]. *Semantic perception* is a further topic aiming to augment sensor data into semantic representations. Today, semantic perception is dominated by fundamental topics such as 2D/3D semantic object recognition and semantic mapping. For example, the task-planning ability of a robot ultimately presupposes that the symbols that a symbolic planner uses are *anchored* in the physical world. Another practical approach with the aim to model semantically meaningful object representations is *semantic world modeling*. Initially presented in association with probabilistic multiple hypothesis anchoring [6], semantic world modeling promotes the modeling of object structures that captures object properties beyond only numeric properties. Further explored in [7], which argued that – unlike multiple hypothesis target tracking – semantic world modeling should also incorporate specific domain characteristics, e.g., objects can have features besides location, which makes them distinguishable from each other in general, and most object states do not change over short periods of time. Symbol grounding and, to some extent, semantic world modeling are essential for PlayGround. However, PlayGround differs from this body of literature in robotics by focusing on affordances, neural-symbolic learning and reasoning, and semantic simulation to determine object feasibility.

Learning *object affordances* has been reported in correlation to both semantic object recognition [8], as well as computer vision [9]. In [8], object properties, learned from RGB-D sensory data, were utilized to identify objects based on natural language queries that contained appearance and name properties. The work presented in [9] promoted, instead, a probabilistic approach to track the relations between objects and human hand actions to learn the function of objects. However, in the context of PlayGround, we need to approach the problem differently as we approach the problem from the relational setting between all objects (and not just hand activities) to extract relational affordances. There is also a growing interest in learning visual concepts from descriptive language in the machine learning community. For example, network architectures for neural attentions [10], or visually grounded question-answer pairs [11]. Another interesting architecture is the one for learning disentangled representations where visual and language features are broken down and learned as separate dimensions [3]. In PlayGround, we will examine how perceptual systems can learn disentangled representations between naming objects, observing the actions performed on objects, and generating the effects of those actions through simulation.

3. Preliminary and Previous Results

This section presents a selection of previous results. Results that have paved the way for the PlayGround project, which we therefore also count as preliminary results of PlayGround.

3.1. Symbolic – Sub-symbolic Framework

In previous work [12], we have presented ProbAnch – a modular data-driven probabilistic anchoring framework. The novelty of this framework is the integration of data-driven *bottom-up anchoring* [13], together with *probabilistic reasoning* based on dynamic distributional clauses (DDC)[14]. This integration allows the ProbAnch framework not only to create and maintain representations of objects (i.e., *anchors*) based on perceptual observations (derived from sub-symbolic sensory data), but also to reason about objects in the absence of perceptual inputs (e.g., in the case of object occlusions), using a combination of logical, probabilistic, and neural-symbolic methods. In other words, ProbAnch is a framework for handling *semantic world modeling* with an extension for *semantic relational object tracking* [15, 16], as seen in Figure 2.

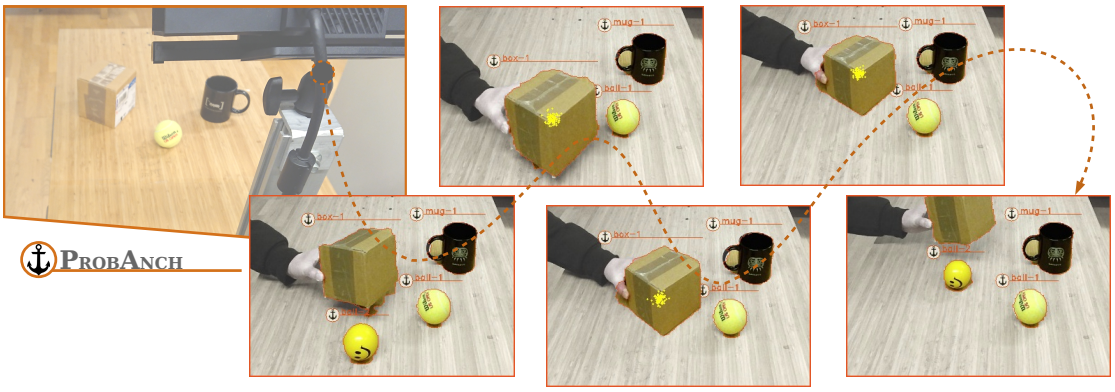

Figure 2: Example of how the ProbAnch framework handles relational object tracking. Occluded objects are tracked via their relationship with observed objects using logical rules, i.e., the position of an occluded object is logically inferred through the position of the occluding object.

3.2. Semantic Simulation Framework

In parallel with the work that resulted in ProbAnch, we have additionally presented initial work on learning generative image manipulations from language instructions using a semantic simulation framework [17]. This work has explored whether a perceptual visual system can simulate human-like cognitive capabilities by training a computational model to predict the output of actions expressed through language instructions. Using a combination of language instructions and images pairs of objects before and after state, as the effects of manipulation actions, the computational model was trained in the settings of a *generative adversarial network* (GAN)[18] in order to generate simulated images that visualize the *effect* of an *action* on a given *object*, i.e., a synthetic generated image that demonstrates the effect of a certain basic manipulation action (e.g., *move*, *remove*, *add*, and *replace*). Aiming to bridge the gap between

simulation and the real world, the trained computational model was subsequently tested in real-world scenarios, as illustrated in Figure 3.

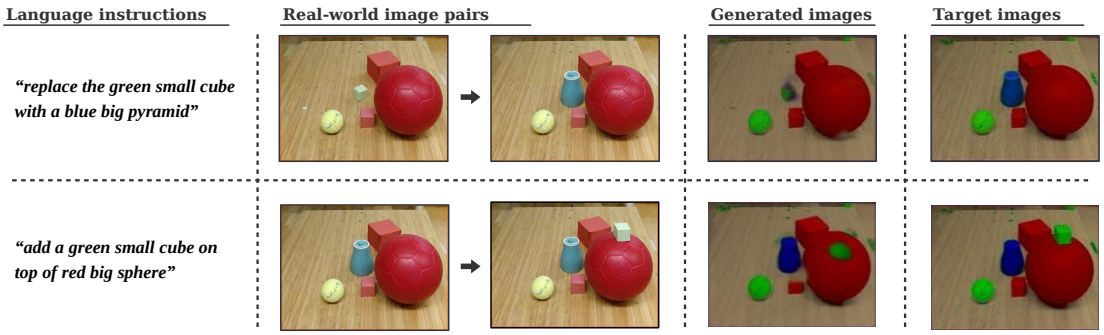

Figure 3: Examples of synthetically generated images as predictions based on given language instructions and real-world images as input. Target images are included for reference.

4. Future Work and Objectives

A natural direction of future work within the PlayGround project would be to integrate both frameworks, as outlined in the previous section. Such integration would allow the symbolic – sub-symbolic framework to utilize the semantic simulator, in combination with language instructions, to predict the future whereabouts of objects and thereby inject positional probability distributions to support the subsequent *anchoring* of the objects (once the objects are perceived through sensory observations). However, the data used to train the generative model of the semantic simulator was neither realistic (image-wise), nor expressive (language-wise). As a result, the model failed to generate representative images while tested in real-world settings (as seen by the generated images in Figure 3). Taking inspiration from the CLEVR [19] and CLEVRER [20] approaches to rendering visual representations together with language questions, an initial objective of PlayGround is to develop a *synthetic generator* for generating realistic synthetic scenarios. This generator should be qualified to generate realistic (and expressive) scenarios so that the scenarios can be transferred to real-world settings. Essential for the PlayGround project is that the generator is also incorporating the notion of affordances, both in terms of affordances given visual representations, as well as affordances in language instructions. Generated synthetic scenarios can, thereby, be used to advance the development of novel techniques for both affordance inference and symbol grounding. Given a qualified synthetic generator, the long-term objectives of PlayGround are, subsequently, to develop integrated symbolic – sub-symbolic representations for supporting high-level reasoning processes, as well as a semantic simulator utilizing neural rendering techniques.

5. Conclusions

In this paper, we have presented a summary of the PlayGround project. In PlayGround, we emphasize symbol grounding and affordance inference using neural-symbolic reasoning and

semantic simulation. Inspired by CLEVR [19] and CLEVRER [20], we promote the use of a synthetic generator for generating scenarios representative of symbol grounding and affordance inference problems. Based on generated scenarios, the grander ambition of PlayGround is thereafter to develop both a symbolic – sub-symbolic learning and reasoning framework (i.e., a neural-symbolic framework), and a semantic simulator framework. Furthermore, as both frameworks are tightly connected and likewise intended for symbol grounding and affordance inference, we expect to additionally be able to exploit synergies between neural-symbolic reasoning and semantic simulation.

Acknowledgments

First of all, we would like to acknowledge the larger consortium of this project that, besides the authors of this short paper, consists of *Prof. Luc De Raedt* and two senior researchers, *Marjan Alirezaie* and *Martin Längkvist*. Thank you for all the work on the proposal that realized the PlayGround project. In addition, this project is funded and supported by the Swedish Research Council (sv. Vetenskapsrådet), grant number: 2021-05229.

References

[1] J. R. Searle, Minds, brains, and programs, Behavioral and brain sciences 3 (1980) 417–424.
[2] S. Harnad, The symbol grounding problem, Physica D: Nonlinear Phenomena 42 (1990) 335–346.
[3] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, J. Wu, The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision, arXiv preprint arXiv:1904.12584 (2019).
[4] D. Marocco, A. Cangelosi, K. Fischer, T. Belpaeme, Grounding action words in the sensorimotor interaction with the world: experiments with a simulated icub humanoid robot, Frontiers in neurorobotics 4 (2010) 7.
[5] S. Coradeschi, A. Saffiotti, Perceptual anchoring of symbols for action, IJCAI International Joint Conference on Artificial Intelligence (2001).
[6] J. Elfring, S. van den Dries, M. van de Molengraft, M. Steinbuch, Semantic world modeling using probabilistic multiple hypothesis anchoring, Robotics and Autonomous Systems 61 (2013) 95–105.
[7] L. L. Wong, L. P. Kaelbling, T. Lozano-Pérez, Data association for semantic world modeling from partial views, The International Journal of Robotics Research 34 (2015) 1064–1082.
[8] Y. Sun, L. Bo, D. Fox, Attribute based object identification, in: Robotics and Automation (ICRA), 2013 IEEE International Conference on, IEEE, 2013, pp. 2096–2103.
[9] H. Kjellström, J. Romero, D. Kragić, Visual object-action recognition: Inferring object affordances from human demonstration, Computer Vision and Image Understanding 115 (2011) 81–90. doi:http://dx.doi.org/10.1016/j.cviu.2010.08.002.
[10] J. Yang, S. E. Reed, M.-H. Yang, H. Lee, Weakly-supervised disentangling with recurrent transformations for 3d view synthesis, Advances in neural information processing systems 28 (2015).

[11] I. Vendrov, R. Kiros, S. Fidler, R. Urtasun, Order-embeddings of images and language, arXiv preprint arXiv:1511.06361 (2015).
[12] A. Persson, P. M. Zuidberg Dos Martires, L. De Raedt, A. Loutfi, Probanch: a modular probabilistic anchoring framework, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence„ International Joint Conferences on Artificial Intelligence, 2020, pp. 5285–5287.
[13] A. Loutfi, S. Coradeschi, A. Saffiotti, Maintaining coherent perceptual information using anchoring, in: Proc. of the 19th IJCAI Conf., Edinburgh, UK, 2005, pp. 1477–1482.
[14] D. Nitti, T. De Laet, L. De Raedt, Probabilistic logic programming for hybrid relational domains, Machine Learning 103 (2016) 407–449.
[15] A. Persson, P. Z. Dos Martires, L. De Raedt, A. Loutfi, Semantic relational object tracking, IEEE Transactions on Cognitive and Developmental Systems 12 (2020) 84–97.
[16] P. Zuidberg Dos Martires, N. Kumar, A. Persson, A. Loutfi, L. De Raedt, Symbolic learning and reasoning with noisy data for probabilistic anchoring, Frontiers in Robotics and AI 7 (2020) 100.
[17] M. Längkvist, A. Persson, A. Loutfi, Learning generative image manipulations from language instructions, in: Concepts in Action: Representation, Learning, and Application (CARLA 2020), Virtual workshop, September 22-23, 2020, 2020.
[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 27, Curran Associates, Inc., 2014, pp. 2672–2680. URL: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.
[19] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, R. Girshick, Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2901–2910.
[20] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, J. B. Tenenbaum, Clevrer: Collision events for video representation and reasoning, in: International Conference on Learning Representations, 2019.