Teaching Robots through Situated Interactive Dialogue and Visual Demonstrations*

Jose L. Part and **Oliver Lemon**
HRI Lab, Heriot-Watt University
Edinburgh Centre for Robotics
Edinburgh, Scotland, UK
jose.part@ed.ac.uk, o.lemon@hw.ac.uk

Abstract

The ability to quickly adapt to new environments and incorporate new knowledge is of great importance for robots operating in unstructured environments and interacting with non-expert users. This paper reports on our current progress in tackling this problem. We propose the development of a framework for teaching robots to perform tasks using natural language instructions, visual demonstrations and interactive dialogue. Moreover, we present a module for learning objects incrementally and on-the-fly that would enable robots to ground referents in the natural language instructions and reason about the state of the world.

1 Introduction

In recent years, robots have started to become more common in a wide variety of environments. Despite coming in many shapes and covering a wide range of applications, such as virtual assistants, vacuum cleaners, smart toys and flying drones, they are still quite limited in what they can do. This is especially true for robots that are expected to operate in unstructured and unconstrained environments like a home, an office or a hospital. Hence, it is necessary to equip robots with mechanisms to incorporate new knowledge and adapt to the users' preferences.

In this project we address the problem of teaching robots about objects in an incremental online manner and about tasks through situated interactive dialogue and visual demonstrations. We are interested in the particular case of the home scenario since it is one of the most challenging, *e.g.* it is highly unstructured, there is a lot of variability from one home to another, and the number of users is quite limited which poses a hard constraint in the amount of available feedback.

The contributions of this project are:

- The development of an incremental online object learning module [Part and Lemon, 2016].

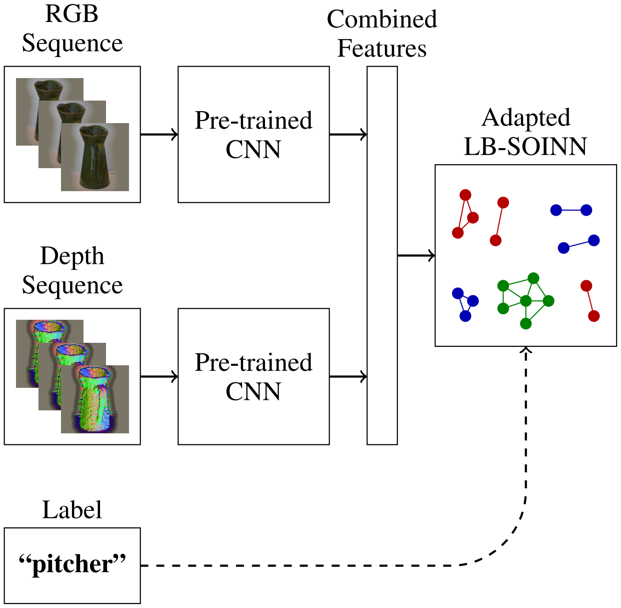

Figure 1: Incremental online object learning module. Features are extracted from preprocessed RGB images and colourized surface normals of the target objects, and are combined and used to train an adapted version of a Load-Balancing Self-Organizing Incremental Neural Network (LB-SOINN) classifier. During training, each sample is accompanied by its label which can be obtained from interacting with a human tutor.

- The development of algorithms for learning hierarchical task structures from natural language instructions and visual demonstrations.
- The integration of the aforementioned modules and their implementation and evaluation on a real robotic platform.

2 Incremental Online Object Learning

One of the goals of this project is to allow a robot to incorporate new knowledge that may be required for performing a task, *i.e.* for an assembly task the robot may need to learn the tools that are involved in the task execution quickly and efficiently. Hence, the aim of our object learning module is to be able to learn new objects on-the-fly from very few examples

[*] J. L. Part is supported by a James-Watt scholarship from the School of Mathematical and Computer Sciences at Heriot-Watt University. This work is partially supported by the ROBOTARIUM Grant (EPSRC Grant No. EP/J015040/1).

without requiring any prior knowledge. Moreover, it should be able to detect when a new object is encountered and trigger a self-driven learning mechanism.

Figure 1 illustrates our current object learning module, which is an improved version of the one described by [Part and Lemon, 2016]. A sequence of RGB images and the corresponding colourized surface normals are fed into a pre-trained Convolutional Neural Network (CNN). This produces two feature vectors that are combined and used to train a Load-Balancing Self-Organizing Incremental Neural Network (LB-SOINN). The surface normals are obtained from the depth maps. We preprocessed the input images by squaring and resizing them to the format expected by the CNN and we reduced the influence of the background by fading it into the mean image of the dataset originally used for training the CNN. By providing the label through dialogue, the system can determine whether it has a symbolic representation for the object and update its knowledge base accordingly.

3 Situated Interactive Task Learning

Over the last few years, there has been some work on mapping instructions into commands [Tellex *et al.*, 2011; Antunes *et al.*, 2016], task learning from videos [Yang *et al.*, 2015] and jointly learning from instructions and demonstrations [Liu and Yang, 2016]. The main drawback of these approaches is that they either require significant knowledge to be available in advance or rely on large sets of pre-acquired labelled data and offline model learning.

[Forbes *et al.*, 2015] proposed a programming by demonstration paradigm where the robot infers object manipulation actions based on demonstrations given as natural language commands and the state of the world. However, they do not provide visual demonstrations to ground the commands.

[Mohseni-Kabir *et al.*, 2015] learn Hierarchical Task Networks from a single demonstration through interaction. However, their focus is on evaluating the grouping of primitive actions into abstract tasks through the incorporation of suggestions made by the system to the user. They do not perform visual grounding and hence, their system does not require to engage in dialogue other than for making suggestions.

As opposed to the previous approaches, we aim at learning hierarchical task representations that can be guided by natural language instructions and grounded on visual demonstrations. Our goal is to learn from a single user interaction rather than generalizing over multiple examples, as [Liu and Yang, 2016] proposed. In this scenario, the robot is able to engage in clarifying dialogue with the user to resolve references, correct mistakes and guide the inference and grounding processes.

4 Discussion and Future Work

Currently, our object learning module relies on a single SOINN network and treats each class as a single entity in the feature space. This poses several issues due to how real-world data is distributed. For example, in the context of category recognition we found that many clusters exist for any one class and they can be sparsely distributed across the feature space. Moreover, when treating the classes in this manner, some of them may present overlaps or interfere with each other during the learning process, which affects the recognition performance of our method and the potential identification of unknown objects.

In order to address these issues, we are looking into defining independent networks for each class and treating the different clusters present in these networks as individual instances. In this manner, it will also be possible to define a heuristic based on similarity with respect to the individual instances to identify unknown objects. We believe this is very important in the context of robot learning since such feature would enable the detection of knowledge gaps and elicit a self-driven mechanism for addressing them.

A major contribution of this project is the learning and grounding of tasks through situated interactive dialogue and visual demonstrations. In order to address this problem, we are looking into building on works like the ones described by [Mohseni-Kabir *et al.*, 2015] and [Liu and Yang, 2016]. In addition, we will integrate the resulting module with our incremental online object learning module to allow for grounding and inference over the objects involved in the task that is being learnt.

Finally, we aim to evaluate our full integrated system on a real task on a PR2 robot.

References

[Antunes *et al.*, 2016] Alexandre Antunes, Lorenzo Jamone, Giovanni Saponaro, Alexandre Bernardino, and Rodrigo Ventura. From Human Instructions to Robot Actions: Formulation of Goals, Affordances and Probabilistic Planning. In *ICRA*, pages 5449–5454, 2016.

[Forbes *et al.*, 2015] Maxwell Forbes, Rajesh P. N. Rao, Luke Zettlemoyer, and Maya Cakmak. Robot Programming by Demonstration with Situated Spatial Language Understanding. In *ICRA*, pages 2014–2020, 2015.

[Liu and Yang, 2016] Changsong Liu and Shaohua Yang. Joint Learning Task Structures from Language Instruction and Visual Demonstration. In *EMNLP*, pages 1–10, 2016.

[Mohseni-Kabir *et al.*, 2015] Anahita Mohseni-Kabir, Charles Rich, Sonia Chernova, Candace L. Sidner, and Daniel Miller. Interactive Hierarchical Task Learning from a Single Demonstration. In *HRI*, pages 205–212, 2015.

[Part and Lemon, 2016] Jose L. Part and Oliver Lemon. Incremental On-Line Learning of Object Classes using a Combination of Self-Organizing Incremental Neural Networks and Deep Convolutional Neural Networks. In *Workshop on Bio-inspired Social Robot Learning in Home Scenarios (IROS)*, Daejeon, Korea, October 2016.

[Tellex *et al.*, 2011] S. Tellex, T. Kollar, S. Dickerson, M.R. Walter, A.G. Banerjee, S. Teller, and N. Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI*, 2011.

[Yang *et al.*, 2015] Yezhou Yang, Yi Li, Cornelia Fermuller, and Yiannis Aloimonos. Robot Learning Manipulation Action Plans by ”Watching” Unconstrained Videos from the World Wide Web. In *AAAI*, pages 3686–3692, 2015.