Interactive Learning through Dialogue for Multimodal Language Grounding

Yanchao Yu
Interaction Lab
Heriot-Watt University
Edinburgh,UK
yy147@hw.ac.uk

Oliver Lemon
Interaction Lab
Heriot-Watt University
Edinburgh,UK
o.lemon@hw.ac.uk

Arash Eshghi
Interaction Lab
Heriot-Watt University
Edinburgh,UK
arash.eshghi@hw.ac.uk

Abstract

We present initial work addressing the problem of interactively learning perceptually grounded word meanings in a multimodal dialogue system. We design an incremental dialogue system using Type Theory with Records (TTR) semantic representations for learning about visual attributes of objects through natural language interaction. This paper explores the use of multi-label visual attribute classification models (TRAM and MLKNN) for such a system. However, these models are found not to perform adequately for this task, so we suggest future directions.

1 Introduction

Learning to identify and talk about objects/events in the surrounding environment is a key capability for intelligent, goal-driven systems what interact with other agents and external world, e.g. smart phones and robots. There has recently been a surge of works and significant progress made on generating image descriptions, identifying images/objects using text descriptions, as well as classifying/describing novel objects using low-level concepts (e.g. colour and shape) (Farhadi et al., 2009). However, most systems rely on pre-trained data of high quality and high quantity without possibility of online error correction. Furthermore, they are unsuitable for robots and multi-modal systems that continuously, and incrementally learn from the environment, and may encounter objects they haven't seen in training data. These limitations may be alleviated if systems can learn concepts from situated dialogue with humans. NL interaction enables systems to take initiative and seek the particular information they need or lack by e.g. asking questions with the highest information gain (see e.g. (Skocaj et al., 2011), and Fig. 1).

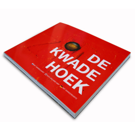
Dialogue	Image	Final semantics
S: Is this a green mug? T: No it's red S: Thanks.		$\begin{bmatrix} x_{=o1} & : & e \\ p2 & : & red(x) \\ p3 & : & mug(x) \end{bmatrix}$
T: What can you see? S: something red. What is it? T: A book. S: Thanks.		$\begin{bmatrix} x1_{=o2} & : & e \\ p & : & book(x1) \\ p1 & : & red(x1) \\ p2 & : & see(sys, x1) \end{bmatrix}$

Figure 1: Example dialogues & resulting semantic representations

We present the first step in a larger programme of research with aim of developing dialogue system what learns (visual) concepts – word meaning – through situated dialogues with humans. We integrate a basic dialogue system using DS-TTR (Eshghi et al., 2012), with two multi-label classification models (MLkNN and TRAM) to simulate the interactive learning process. In effect, the dialogue with a tutor continuously provides semantic information about objects in the scene which is then fed to an online classifier in the form of training instances. Conversely, the system can utilise the grammar and existing knowledge base to make references and formulate questions related to different objects' attributes identified in the scene. For evaluating the performance of situated dialogue on attribute-based recognition, we compare the performance of two learning models as more training instances are presented to them.

2 System Architecture

The architecture of the system (see Fig. 2) contains two main modules: a vision module for visual feature extraction and classification; and a dialogue system module using DS-TTR. We assume access to logical semantic representations by DS-

Proceedings of the 19th Workshop on the Semantics and Pragmatics of Dialogue, September 24-26, 2015, Gothenburg, Sweden.

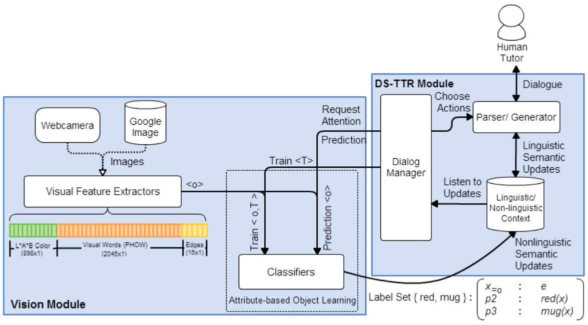


Figure 2: Architecture of the teachable system

TTR parser/generator as a result of processing dialogues with a human tutor.

The **Vision Module** is implemented with two multi-label classification algorithms (MLkNN and TRAM) for learning/classifying low-level object-based attributes: a) MLkNN (Zhang and Zhou, 2007), as a supervised learning model, predicts potential label sets of unseen objects using k-nearest neighbour algorithm; b) TRAM (Kong et al., 2013) proposes a semi-supervised model that predicts the binary label set of a novel instance based on utilized information from both seen and unseen objects. For learning new multi-label classifiers, we build a pair of inputs – a 2960-dimensional visual feature vector from each object using features from (Farhadi et al., 2009) and an i-dimensional binary label vector for each instance (where the $i - th$ attribute takes the value of 1 if it belongs to the instance and -1 otherwise).

The **Dialogue System Module** implements DS-TTR, which is a word-by-word incremental semantic parser/generator for dialogue, based around the Dynamic Syntax (DS) grammar framework (Cann et al., 2005), in which interlocutors interactively construct contextual and semantic representations (Purver et al., 2011). The contextual representations afforded by DS-TTR are of the fine-grained semantic content that is jointly negotiated/agreed upon by the interlocutors, as a result of processing questions and answers, clarification requests, corrections, acceptances, etc (see (Eshghi et al., 2015) and the first row of Fig. 1).

3 Results & Future work

We evaluated the performance of two multi-label classification models (MLkNN and TRAM) for attribute classification of object images. TRAM outperforms MLkNN and both models improve on classifying attributes for which they receive more training instances. However, the results show that both models are not ideal approaches to our problem, since for good performance they require many more training examples than can be provided in an interactive teaching session with a human. What we need are learning methods which can operate effectively on small numbers of samples, and which can improve performance robustly while continuously learning new examples. These properties are know as "zero-shot" and "incremental" learning respectively. We will explore these two approaches in future work.

Acknowledgments

This research is partially supported by the EPSRC, under grant number EP/M01553X/1 (BABBLE project[1]).

References

Ronnie Cann, Ruth Kempson, and Lutz Marten. 2005. *The Dynamics of Language*. Elsevier, Oxford.

Arash Eshghi, Julian Hough, Matthew Purver, Ruth Kempson, and Eleni Gregoromichelaki. 2012. Conversational interactions: Capturing dialogue dynamics. In S. Larsson and L. Borin, editors, *From Quantification to Conversation: Festschrift for Robin Cooper on the occasion of his 65th birthday*, volume 19 of *Tributes*, pages 325–349. College Publications, London.

A. Eshghi, C. Howes, E. Gregoromichelaki, J. Hough, and M. Purver. 2015. Feedback in conversation as incremental semantic update. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS 2015)*, London, UK. Association for Computational Linguisitics.

Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. 2009. Describing objects by their attributes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR*.

Xiangnan Kong, Michael K. Ng, and Zhi-Hua Zhou. 2013. Transductive multilabel learning via label set propagation. *IEEE Trans. Knowl. Data Eng.*, 25(3):704–719.

Matthew Purver, Arash Eshghi, and Julian Hough. 2011. Incremental semantic construction in a dialogue system. In J. Bos and S. Pulman, editors, *Proceedings of the 9th International Conference on Computational Semantics*, pages 365–369, Oxford, UK, January.

Danijel Skocaj, Matej Kristan, Alen Vrecko, Marko Mahnic, Miroslav Janícek, Geert-Jan M. Kruijff, Marc Hanheide, Nick Hawes, Thomas Keller, Michael Zillich, and Kai Zhou. 2011. A system for interactive learning in dialogue with a tutor. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, pages 3387–3394.

Min-Ling Zhang and Zhi-Hua Zhou. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048.

[1] `https://sites.google.com/site/hwinteractionlab/babble`