VOILA: An Optimised Dialogue System for Interactively Learning Visually-Grounded Word Meanings (Demonstration System)

Yanchao Yu
Interaction Lab
Heriot-Watt University
y.yu@hw.ac.uk

Arash Eshghi
Interaction Lab
Heriot-Watt University
a.eshghi@hw.ac.uk

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

Abstract

We present VOILA: an optimised, multi-modal dialogue agent for interactive learning of visually grounded word meanings from a human user. VOILA is: (1) able to learn new visual categories interactively from users from scratch; (2) trained on real human-human dialogues in the same domain, and so is able to conduct natural spontaneous dialogue; (3) optimised to find the most effective trade-off between the accuracy of the visual categories it learns and the cost it incurs to users. VOILA is deployed on Furhat[1], a human-like, multi-modal robot head with back-projection of the face, and a graphical virtual character.

1 Introduction

As intelligent systems/robots are brought out of the laboratory and into the physical world, they must become capable of natural everyday conversation with their human users about their physical surroundings. Among other competencies, this involves the ability to learn and adapt mappings between words, phrases, and sentences in Natural Language (NL) and perceptual aspects of the external environment – this is widely known as *the grounding problem*. Our work is similar in spirit to e.g. (Roy, 2002; Skocaj et al., 2011) but advances it in several aspects (Yu et al., 2016).

In this demo paper, we present a dialogue agent that learns visually grounded word meanings interactively from a human tutor, which we call: VOILA (Visually Optimised Interactive Learning Agent). Our goal is to enable this agent to learn to identify and describe objects/attributes (colour and shape in this case) in its immediate visual environment through interaction with human users, incrementally, over time. Unlike a lot of past work (Silberer and Lapata, 2014; Thomason et al., 2016; Matuszek et al., 2014), here we assume that the agent is in the position of a child, who does not have any prior knowledge of perceptual categories. Hence, the agent must learn from scratch: (1) the perceptual/visual categories themselves; and (2) how NL expressions map to these; and in addition, (3) as a standard conversational agent, the agent much also learn to conduct natural, spontaneous conversations with real humans.

In this demonstration, *VOILA* plays the role of an interactive, concept learning agent that takes initiative in the dialogues and actively learns novel visual knowledge from the feedback from the human tutor. What sets VOILA apart from other work in this area is:

- VOILA's dialogue strategy is *optimised* via Reinforcement Learning to achieve an optimal trade-off between the accuracy of the concepts it learns/has learnt from users, and the effort that the dialogues incur on the users: this is a form of active learning where the agent only asks about something if it doesn't already know the answer with some appropriate confidence (see (Yu et al., 2016) for more detail).

- VOILA is trained on a corpus of real Human-Human conversations (Yu et al., 2017), and is thus able to process *natural* human dialogue, which contains phenomena such as *self- corrections, repetitions and restarts, pauses, fillers, and continuations*

VOILA is deployed onto Furhat, a human-like robot head with a custom back-projected face, built-in stereo microphones, and a Microsoft

[1] http://www.furhatrobotics.com/

Proceedings of the SIGDIAL 2017 Conference, pages 197–200,
Saarbrücken, Germany, 15-17 August 2017.

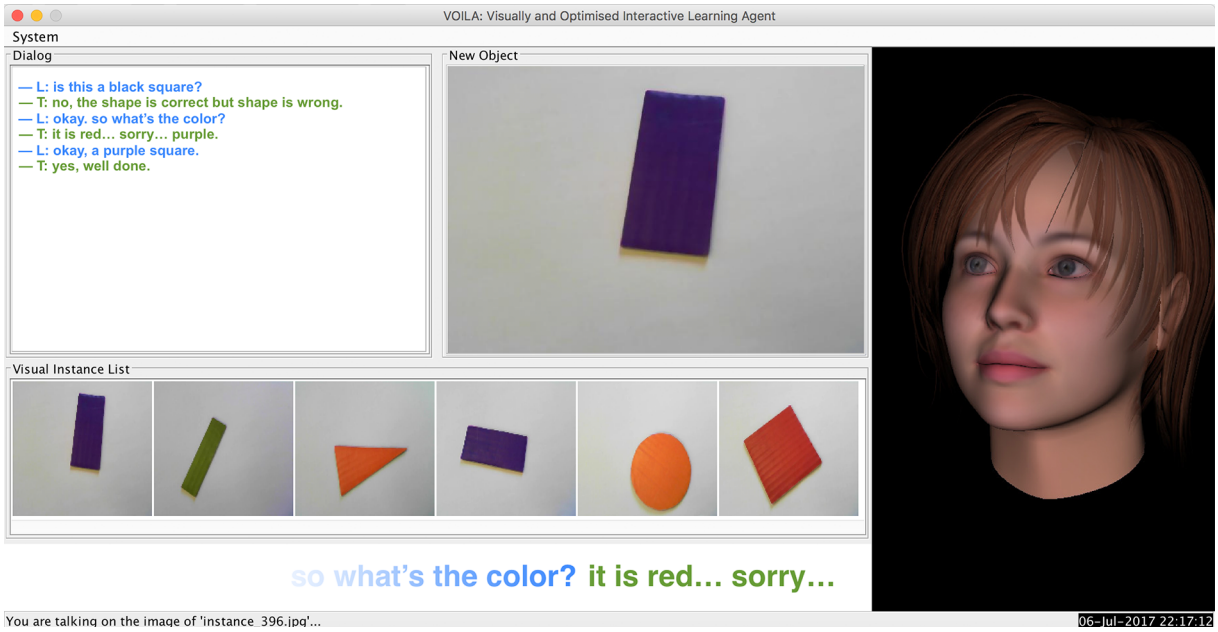

Figure 1: Interactive Visual Concept Learning in the VOILA Agent (Screenshot)

Kinect for skeletal tracking and processing non-verbal signals. A graphical version of the character can also bee used (see 1).

2 Interactive Multimodal Framework

We developed a multimodal framework in support of building an interactive learning system, which loosely follows that of Yu et al. (2016). The framework consists of two core modules:

Vision Module The vision module produces visual attribute predictions, using two base feature categories: the HSV colour space for colour attributes, and a 'bag of visual words' (i.e. PHOW descriptors) for the object shapes/class. It consists of a set of binary classifiers - Logistic Regression SVM classifiers with Stochastic Gradient Descent (SGD) (Zhang, 2004) – to incrementally learn attribute predictions. The visual classifiers ground visual attribute words such as 'red', 'circle' etc. that appear as parameters of the Dialogue Acts used in the system.

Dialogue Module This module relies on a classical architecture for dialogue systems, composed of Dialogue Management (DM) and Natural Language Understanding (NLU), as well as Generation (NLG) components. These components interact via Dialogue Act representations (Stolcke et al., 2000), e.g. *inform(color=red)*, *ask(shape)*. The Natural Language Understanding component processes user utterances by extracting a sequence of key patterns, slots and values, and then transforming them into dialogue-act representations, following a list of hand-crafted rules. The NLG component makes use of a template-based approach that chooses a suitable learner utterance for a specific dialogue act, according to the statistical distribution of utterance templates from dialogue examples. Finally, the DM component is implemented with an optimised learning policy using Reinforcement Learning (see Section 3). This optimised policy is trained to: (1) conduct interaction with human partners, and (2) achieve an optimum balance between classification performance and the cost of the dialogue to the tutor in the interactive learning process.

3 Learning How to Learn

In this section, we briefly describe our method for optimising the dialogue agent with Reinforcement Learning and in interaction with a simulated tutor, itself built from the BURCHAK human-human dialogue corpus[2] (Yu et al., 2017) within a simulated learning environment (see Fig. 2).

Given the visual attribute learning task, a smart agent must learn novel visual objects/attributes as accurately as possible through natural interactions with real humans, but meanwhile it should attempt

[2]BURCHAK is freely available at `https://sites.google.com/site/hwinteractionlab/babble`

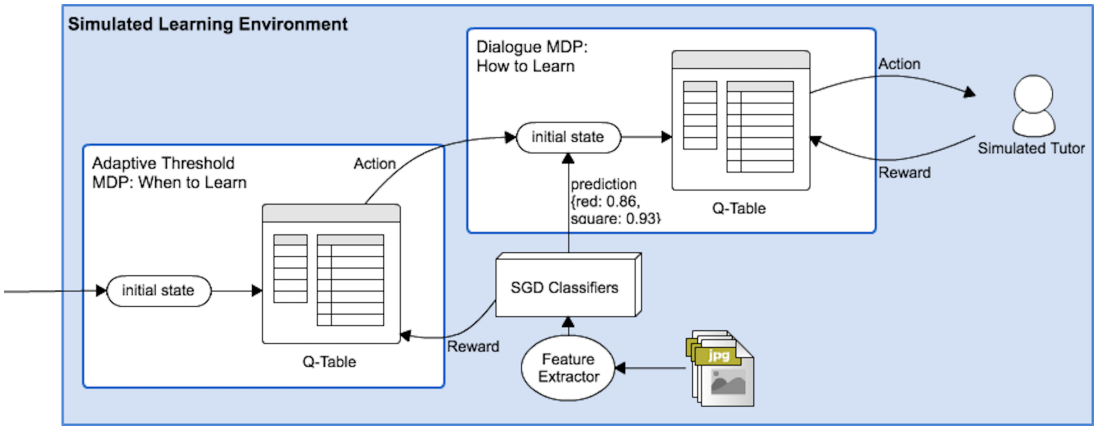

Figure 2: Architecture of Optimised Learning Policy with a Hierarchical MDP

to minimise the human involvement as much as possible in this life-long learning process. Here, we divide this interactive learning task into two sub-tasks, modeled as a hierarchical Markov Decision Process, consisting of two interdependent MDPs in charge of decisions about: *"when to learn"* and *"how to learn"*.

3.1 When to Learn: Adaptive Confidence Threshold

The first MDP performs a kind of *active learning*: the learner/agent only acquires the feedback from humans about a visual attribute if it is not confident enough already about its own predictions. Following previous work (Yu et al., 2016), here we use a positive confidence threshold, which determines when the agent believes its own predictions. For instance, the learner can ask either polar or WH-questions about an attribute if its confidence score is higher than a certain threshold; otherwise, there should be no interaction about that attribute. But as Yu et al. (2016) point out the confidence score from a classifier is not reliable enough at the early stages of learning, so in order to find an optimum dialogue policy, a threshold should be able to dynamically adjust according to the previous learning performance of the agent. We therefore assign a separate but dependent component MDP for adjusting the threshold dynamically in order to optimise the trade-off between accuracy and cost. Note now that the adjusted confidence threshold will affect the agent's dialogue behaviour, modeled in the other MDP presented in the next section (natural interaction with humans).

3.2 How to Learn: Natural Interaction with Humans

The second MDP, as a purely conversational agent, aims at managing natural, spontaneous conversation with human partners or other agents to achieve the final goal, i.e. gain useful information about visual attributes. The initial state in this MDP is determined by a combination of the adjusted threshold from the former MDP and the visual predictions from the color and shape classifiers that ground NL attributes terms ('red', 'square', etc): either the color or shape status can be assigned to: 0, if the learner has a low confidence on its predictions (i.e. the confidence score is lower than 0.5); or, 1 if the confidence score is higher than 0.5, but lower than the positive threshold; or else, 2. This together with the previous dialogue act constitutes the state space of this MDP. The agent is then trained to choose the correct dialogue action to achieve a state in which both shape and color of the current object are known with certainty (with status = 2), either through feedback from the user, or through the agent's own existing visual knowledge. Of course, the agent must also learn to produce coherent dialogues by responding to questions at the right time, giving feedback at the right time, asking for feedback at the right time, etc.

4 Demonstration

As noted, VOILA has been deployed onto Furhat: a human-like Robot Head (Moubayed et al., 2011), which provides an interaction framework for the management of multi-party, multi-modal interactions, and which employs a Microsoft Kinect for skeletal tracking. In the demonstration,

the VOILA agent will randomly choose 20 visual objects, and then learn to describe them using their low-level visual attributes (e.g. color and shape) image-by-image through interaction with users. As mentioned above, we assume that VOILA is in the position of a child learning from scratch, but instead of complex real objects with noisy backgrounds, we use a set of simple toy objects (see dialogue example in Fig. 1), but without annotations or labels. It is essential to highlight that the VOILA agent would only start a conversation with a human partner when it isn't confident about its own attribute predictions.

5 Conclusion

We have presented a multi-modal learning agent – VOILA – that can learn grounded visual-concept meanings through interaction with human tutors incrementally, over time. The agent is deployed with an *adaptive* dialogue policy (optimised using Reinforcement Learning), which has learned to (1) process natural, coherent conversations with humans and (2) achieve comparable learning performance to a hand-crafted system, but with less tutoring effort needed from humans. Recently, we also extended the VOILA agent to learn real visual object classes instead of toy objects by integrating with a Load-Balancing Self-Organizing Incremental Neural Network (LB-SOINN) (Zhang et al., 2014) for object classification.

Acknowledgements

This research is supported by the EPSRC, under grant number EP/M01553X/1 (BABBLE project[3]).

References

Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. 2014. Learning from unscripted deictic gesture and language for human-robot interactions. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*. pages 2556–2563.

Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. 2011. Furhat: A back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive Behavioural Systems - COST 2102 International Training School, Dresden, Germany, February 21-26, 2011, Revised Selected Papers*. pages 114–130.

Deb Roy. 2002. A trainable visually-grounded spoken language generation system. In *Proceedings of the International Conference of Spoken Language Processing*.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, volume 1, pages 721–732.

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 *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2011, San Francisco, CA, USA, September 25-30, 2011*. pages 3387–3394. https://doi.org/10.1109/IROS.2011.6094926.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca A. Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *CoRR* cs.CL/0006023.

Jesse Thomason, Jivko Sinapov, Maxwell Sevtlik, Peter Stone, and Raymond J. Mooney. 2016. Learning multi-modal grounded linguistic semantics by playing "i spy". In *To Appear: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-16, New York City, USA, July 9-15, 2016*.

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2016. Training an adaptive dialogue policy for interactive learning of visually grounded word meanings. In *Proceedings of SIGDIAL 2016, 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Los Angeles, pages 339–349.

Yanchao Yu, Arash Eshghi, Gregory Mills, and Oliver Lemon. 2017. *Proceedings of the Sixth Workshop on Vision and Language*, Association for Computational Linguistics, chapter The BURCHAK corpus: a Challenge Data Set for Interactive Learning of Visually Grounded Word Meanings, pages 1–10. http://aclweb.org/anthology/W17-2001.

Hongwei Zhang, Xiong Xiao, and Osamu Hasegawa. 2014. A Load-Balancing Self-Organizing Incremental Neural Network. *IEEE Transactions on Neural Networks and Learning Systems* 25(6):1096–1105.

Tong Zhang. 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*. ACM, page 116.

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