Interactive visual grounding with neural networks

José Miguel Cano Santín[1] **Simon Dobnik**[1,2] **Mehdi Ghanimifard**[1,2]
[1]Department of Philosophy, Linguistics and Theory of Science (FLoV)
[2]Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg, Sweden
[1]jmcs990@gmail.com [2]{simon.dobnik,mehdi.ghanimifard}@gu.se

Abstract

Training strategies for neural networks are not suitable for real time human-robot interaction. Few-shot learning approaches have been developed for low resource scenarios but without the usual teacher/learner supervision. In this work we present a combination of both: a situated dialogue system to teach object names to a robot from its camera images using Matching Networks (Vinyals et al., 2016). We compare the performance of the system with transferred learning from pre-trained models and different conversational strategies with a human tutor.

1 Introduction

Robotic systems need to acquire constantly new knowledge about their environment and the objects present in it as well as knowledge that they receive by interacting with their conversational partners. In grounding what they see in language they can benefit a lot by taking into account *how* such information is presented to them by the context of their perceptual environment and a human tutor (Skočaj et al., 2010). To this end, our situated dialogue system implements (i) different *dialogue interaction strategies* that exploit linguistic and interactive information from the dialogue and (ii) different *perceptual classification strategies* that exploit the interaction of the agent with its environment, namely the context of the previous objects and exposure to background knowledge. While our system includes both (i) and (ii) the focus of this paper is (ii). In particular, we examine how a deep neural network model can benefit from knowledge pre-trained on large data offline to learn new objects with little online data but with contextually provided background categories.

2 Method

Our situated dialogue system is based on the KILLE setup designed by Dobnik and de Graaf (2017). It consists of a stationary Kinect v1 sensor in front of which a small individual object can be presented. The camera is connected to a Robot Operating System (ROS) framework (Quigley et al., 2009) running on Ubuntu 16.04 system using the *Freenect* driver. Our Python script which is implemented as a node within the ROS community takes care of both the object recognition and the dialogue management.

Dialogue Interaction Strategies Our situated dialogue system uses two different strategies to learn objects. The human tutor can present the object (e.g. *This is an apple*), in which case it will save the object in the dataset and, if necessary, it will retrain the model to learn a new label or ask for more about the object if it does not have enough images to learn. The robot can also be queried (e.g. *What is this?*). In this case, the robot will attempt to recognise the object presented and answer with the highest scoring label and the certainty of the guess depending on the score of that label (e.g. *This is an apple*; *I think this is an apple*; *I don't know what is this. Please, tell me.*). Then, the user can give feedback and confirm if the guess is true or tell the system the true label.

Interactive Visual Classification Our deep neural model consists of two independent modules. Firstly, there is a stack of the VGG16 CNN layers (Simonyan and Zisserman, 2014) pre-trained on ImageNet (Russakovsky et al., 2015). Here we test the advantages and disadvantages of the learning transferred from a large dataset to encode the images that the robot will perceive with its camera. Secondly, we present our implementation of Matching Networks (Vinyals et al., 2016) in a robot scenario with interactive grounding, which is the main contribution of this work. The idea of the matching network algorithm is to rapidly train a neural network from few observations. Each training instance consists of few (k)

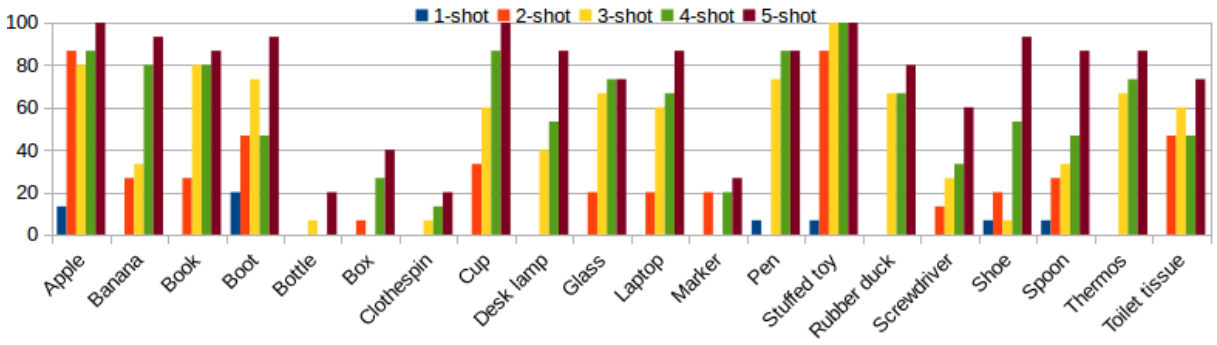

Figure 1: Results on learning new labels. The k-shot learned label is specified under the x axis and each of the bars represent the accuracy of the classification from 1-shot (left) to 5-shot (right).

images of each labelled class (n) that make up the support set S and a target image t belonging to one of these classes. The objective of training is to predict the class of the target image and therefore learn how to discriminate images of different classes. The contribution and the novelty of our work is the application of this model in an interactive robot training scenario where the classes of objects in the support set can be thought of as contextually present objects related to the current task in which the robot is being trained.

3 Experiments

Baseline In order to validate our implementation of the matching networks we use 20 categories from the test split of miniImageNet, which was created from ImageNet by Vinyals et al. (2016). Figure 2 shows that the accuracy of our system increases considerably when adding more images per label, as well as it becomes more difficult to classify correctly with more labels. An important aspect that needs to be considered in an interactive scenario is how long training and application of the model takes. Encode time does not seem to increase much with more images, while train time of the matching network is more clearly affected. However, being able to train a model for 20 labels in about 15 seconds and achieving an accuracy of 74.2% seems very reasonable.

Learning a New Class of Objects The objective of this experiment is to test how many images the system needs to see to learn new class labels. We collected a new small dataset of 20 categories and 20 images per category taken with the robot's camera as the support set and target images. We simulate the learning process by training matching networks on 19 labels with five images each, which represent the categories that the robot already knows, and then adding the remaining label to the support set for training each model which is learned with 1 to 5 images in each case. Then, we evaluate the recognition accuracy of the new label on the remaining 15 images of the same label for each of the models. Figure 1 shows that four to five images are necessary for most of the labels to have a reliable object recognition. Also, some labels are clearly easier to learn than others, e.g. bottle, box, clothespin and marker *vs.* apple, book and stuffed toy, which did not need more than three images to get to about 80% accuracy.

5 labels	**1-shot**	**5-shot**	**10-shot**
Accuracy	75.8%	89.8%	98.8%
Encode time	1.12s	1.63s	2.15s
Train time	1.43s	3.57s	7.27s
20 labels	**1-shot**	**5-shot**	**10-shot**
Accuracy	52.5%	74.2%	82.6%
Encode time	1.41s	1.93s	2.39s
Train time	3.26	12.15s	25.99s

Figure 2: Baseline results on miniImageNet. Encode time is the number of seconds to encode the support set (S) images with VGG16. Train time is the number of seconds to train the matching networks.

4 Conclusion and Future Work

The observed results are promising and the system could be extended in multiple ways, which is the focus of our ongoing work. For instance, by using offline pre-trained knowledge also for the matching networks, new interactive strategies with the robot, attention over the visual regions objects to avoid the influence of the background objects and trying different techniques for selecting the images of the support set.

Acknowledgments

The research reported in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

Simon Dobnik and Erik de Graaf. 2017. Kille: a framework for situated agents for learning language through interaction. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 162–171, Gothenburg, Sweden. Association for Computational Linguistics.

Morgan Quigley, Ken Conley, Brian P. Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y. Ng. 2009. Ros: an open-source robot operating system. In *ICRA Workshop on Open Source Software*.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv e-prints*, page arXiv:1409.1556.

Danijel Skočaj, Miroslav Janiček, Matej Kristan, Geert-Jan M. Kruijff, Aleš Leonardis, Pierre Lison, Alen Vrečko, and Michael Zillich. 2010. A basic cognitive system for interactive continuous learning of visual concepts. In *ICRA 2010 workshop ICAIR - Interactive Communication for Autonomous Intelligent Robots*, pages 30–36, Anchorage, AK, USA.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching Networks for One Shot Learning. *arXiv e-prints*, page arXiv:1606.04080.