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

1 Aims

- Visual **grounding** of objects descriptions.
- Learning to recognise objects in **interaction** (Skočaj et al., 2010):
  - human tutor;
  - situated dialogue system.
- **Few-shot learning** with neural networks: object categories from few samples.
- **Transfer learning**: pre-trained knowledge on large offline dataset.

2 Robot setup

- Based on the Kille framework (Dobnik and de Graaf, 2017).
- Microsoft Kinect v1 RGB-D sensor using *Freenect* drivers.
- Robot Operating System (ROS) framework (Quigley et al., 2009).
- Python scripts implemented as nodes within the ROS community take care of
  - object recognition;
  - dialogue management.

3 Visual classification

- The goal is to train neural network models for image classification which are suitable for online interaction with a robot.
- Therefore, we need:
  - very fast training;
  - learning from few observations.
- A neural network which consists of two modules.
- Image encoder with VGG16 CNN layers (Simonyan and Zisserman, 2014).
  - Pre-trained on ImageNet (Russakovsky et al., 2015).
  - Test transfer learning from a large dataset.
- Matching Networks (Vinyals et al., 2016) in a robot scenario.
  - Neural network algorithm designed for one-shot learning.
  - Fast learning from few examples.
- Each training instance consists of:
  - Few (k) images of each labelled class (n) that make up the support set S.
  - A target image t belonging to one of these classes.
- The objective is to predict the class of t and therefore learn how to discriminate images of different classes.

4 Experiments

Baseline

- How well does the system manage to recognise object categories of the miniImageNet corpus (Vinyals et al., 2016)?
  - 5 or 20 labels are presented in each round, each with 1, 5 or 10 images.
  - Evaluate the accuracy of the object recognition on the rest of the images of each label.
  - The time that the system takes to encode the images and to train the matching network.

5 labels	1-shot	5-shot	10-shot
Accuracy	75.8%	89.8%	98.8%
Encode time	1.12s	1.63s	2.15s
Training 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
Training time	3.26	12.15s	25.99s

Figure 1: *Baseline results on miniImageNet. Encode time is the number of seconds to encode the support set (S) images with VGG16. Training time is the number of seconds to train the matching network.*

- The system achieves very good results for 5-shots or more and it is fast to encode images and train a new model.
- However, training time increases significantly when adding more categories and images per category.

Learning a new class of objects

- How many examples are required to learn a new label?
  - Images collected from our robot domain: 20 categories with 20 images per category.
  - Each round has 19 categories already learned (5-shot) and one new label with 1 to 5 images in each round.
  - Evaluate the performance on the new label: the rest of the images of this label.

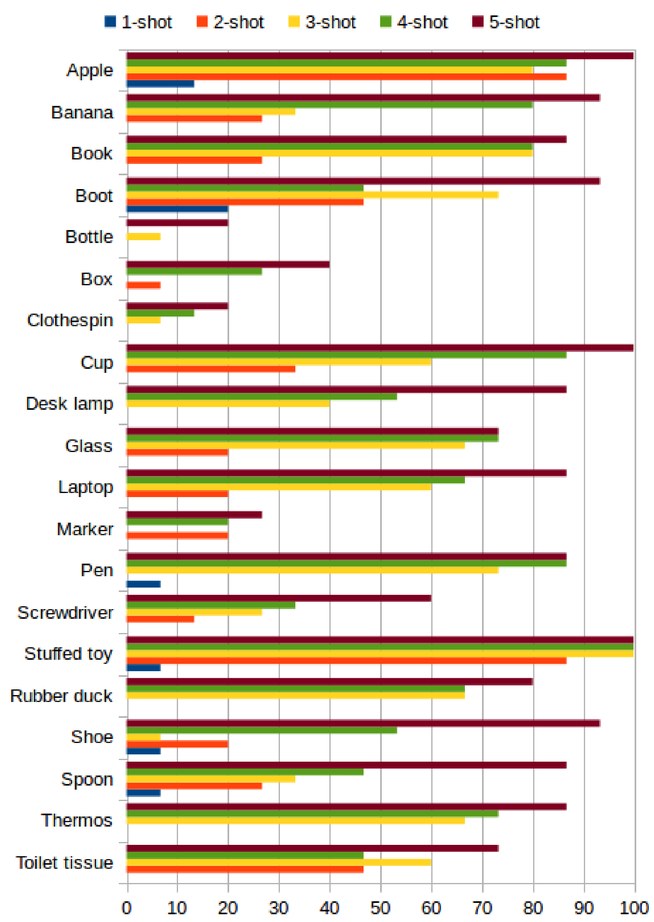

Figure 2: *Results on learning new labels. The k-shot learned label is specified on the y axis and each of the bars represent the accuracy of the classification from 1-shot to 5-shot.*

- 4-5 images are necessary for most of the labels to recognised reliably.
- Some labels are clearly easier to learn than others.

5 Dialogue interaction strategies

- **Present an object and teach it**

  User: *This is an apple*
  System: [Observes the object and checks how many images of "apple" has in its database.]
  – More than 5 images.
  S: *I am updating my systems on apple.*
  – Exactly 5 images.
  S: *I am learning apple.*
  – Less than 5 images.
  S: *Please, show me more examples of apple.*

- **Confident object recognition**

  U: *What is this?*
  S: [Observes the object and recognises it with high confidence.]
  S: *This is a rubber duck. Am I right?*
  U: *Yes.*
  S: *Great!*

- **Wrong recognition**

  U: *What is this?*
  S: [Observes the object and recognises it, but it is not sure.]
  S: *I think this is a boot. Am I right?*
  U: *No.*
  S: *What is this, then? Write only the name of the object.*
  U: *Shoe.*
  S: [Checks how many images of "shoe" has in its database.]

- **Very low confidence recognition**

  U: *What is this?*
  S: [Observes the object and fails to assign it a category: the confidence score is very low.]
  S: *I don't know what this is. Please, tell me.*

6 Conclusions

- The observed results are promising.
- The system could be extended in several ways.
  - Using offline pre-trained knowledge also for the matching networks.
  - New interactive strategies with the robot.
  - Attention over the visual regions of the objects to avoid the influence of the background.
  - Trying different techniques for selecting the images in the support set.

CLASP centre for linguistic theory and studies in probability

LondonLogue – Semdial 2019: The 23rd Workshop on the Semantics and Pragmatics of Dialogue, Queen Mary University of London, 4–6 September 2019