Discriminating 3D-objects using

Dominik Künkele
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
dominik.kuenkele@outlook.com

Simon Dobnik
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
simon.dobnik@gu.se

Abstract

1 Introduction

Agents interact with the physical world through their action and perception and with other agents through language. Their sensors and actuators (if they are artificial agents, but the same also holds for natural agents) allow them to sample the world and their own state using measures that are continuous in nature such as intensity of light, distance, angles, velocity and others which can be measured with a high degree of accuracy. On the other hand language that is used to communicate with other agents is based on representations that are composed of a limited set of discrete symbols. How can both domains and representations arising from these interactions be combined? How are ranges of measurements expressed in a continuous domain mapped to discrete linguistic labels? How is ambiguity and underspecification resolved? How can agents achieve this through interactive grounding (Regier, 1996; Roy, 2005; Cooper, 2023)?

2 Background and previous work

Language games with a sender and a receiver (Clark, 1996; Bartlett and Kazakov, 2005; Kirby et al., 2008; Steels and Loetzsch, 2009; Zaslavsky et al., 2018) offer an opportunity to examine how neural models condense seen information into a distinct representation of symbols through linguistic interaction. A sender describes a perceptual situation that can also be seen by a receiver starting with a random label while the receiver attempts to interpret the meaning of these symbols and combine it with other information. It then sends feedback to the sender about its interpretation. Interaction is optimized based on the communicative success of the sender and the receiver. Gradually, both agents converge on a set of symbols that they can use to refer to situations in jointly attended perceptual scenes (Chai et al., 2016; Kelleher and Dobnik, 2020). Communication is successful and grounding converges because learning is constrained by joint attention and the agents are playing following the rule of the communicative games.

In this paper we explore how agents based on artificial neural networks learn referential grounding of entities in images of 3-dimensional scenes where one agent is describing the entities and the other agent learns to interpret the reference of symbols in the scene either by identifying the object in the scene as a bounding box or location of the object based on object attributes such as shape, color and size. The novelty of our work compared with the previous work with this setup (Kharitonov et al., 2019) is that (i) we extend the popular CLEVR dataset (Johnson et al., 2016) with new artificially generated 3-d scenes of objects that can be referred to and discriminated based on the attributes such as shape, color and size whereby the discrimination is based on different overlaps of these attributes between the target and the distractor objects; and (ii) we implement a focused referring to objects either as one of the potential bounding boxes or location of the target.

3 Our dataset: CLEVR-Dale-2 and Dale-5

The dataset used for these experiments is an extended version of the CLEVR dataset (Johnson et al., 2016).[1] This dataset includes 3D-generated images depicting scenes with different kinds of objects. Each of these objects has different combinations of attributes, such as *shape*, *color* and *size*. The objects are randomly placed into the scene and assigned with random attributes. Figure 1 shows an example of a generated image of this CLEVR dataset. The dataset also contains ground truth in-

[1] https://github.com/DominikKuenkele/MLT_Master-Thesis_clevr-dataset-gen

formation about each image, including attributes and locations for all objects.

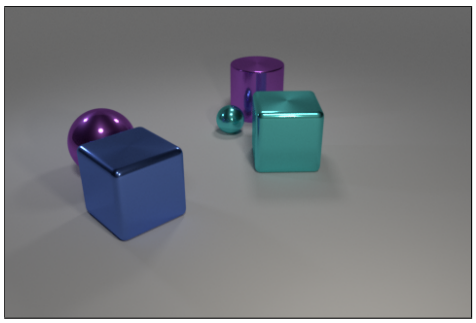

Figure 1: Example of a generated image in the Dale-5 dataset with the extended code

The extended version of the dataset divides the objects into *target object* and *distractors*. The target object is the main object in the scene that should be identified and communicated by the agents. The distractors contain similar objects to the target object. The number of shared attributes is defined by rules researched in Dale and Reiter (1995). The target object is therefore always identifiable using either only the *shape* (1), the *shape* and *color* (2) or the *shape*, *color* and *size* (3).

For this research, two datasets are created. The *Dale-2* dataset contains one target object and one distractor, while the *Dale-5* dataset contains one target object and four distractors. Both datasets contain 10.000 images. For each image fixed size bounding boxes around each object are extracted. These bounding boxes are the input for the sender and receiver in the language game.

4 Language game

The language game was developed and run in the EGG framework (Kharitonov et al., 2019). The setup of the game is based on the game described in the paper by Lazaridou et al. (2016). The same multiple images are passed to a sender and a receiver. The sender must communicate the target image to the receiver with a message, who needs to identify this target image. Instead of distinguishing two images from different concepts, in this experiment the agents need to distinguish two objects with shared attributes. Both sender and receiver have a similar architecture to the *agnostic sender* and *receiver* in Lazaridou et al. (2016). One central difference is the production of the message. Their paper focuses on the classification of a concept for the input image and therefore produces only one-symbol messages that should correspond to these concepts. This research focuses on the identification of attributes for the objects and their combination, which is why our models produce a sequence of symbols for the message (which may correspond to the attributes). This is done using an encoder LSTM (sender) and a decoder LSTM (receiver) (see Figure **??**).

The initial state of the encoder LSTM is the image, passed through ResNet101 and a following linear layer that reduces the dimensions. The sequence is then created through Gumbel-Softmax relaxation (Jang et al., 2016).

The receiver on the other side takes the sequence as input for its decoder LSTM. The hidden state is randomly initialized. After each step of the LSTM, the receiver calculates the dot product between the hidden state and its own image encoding (calculated as the sender's image encoding). The receiver then 'points' to one of the images, while applying the softmax function over the results. The loss is calculated using the NLL-loss. Following, the losses for each token in the message sequence are summed up, and all weights of the receiver as well as the sender are updated, based on this summed loss.

There are five variables in the experiments that are adjusted: (1) the image embedding size for the sender e_s, (2) the LSTM hidden size for the sender h_s, (3) the image/message embedding size for the receiver e_r, (4) the LSTM hidden size for the receiver h_r and (5) the size of the vocabulary $|V|$.

The results will be evaluated using the accuracy if the receiver could identify the target object. A random guess corresponds to 50% in the *Dale-2* dataset and 20% in the *Dale-5* dataset.

5 Results

TODO:

- Dale-2: **(done)**
  - small hidden/embedding dims, small vocab -> high accuracy
  - high hidden/embedding dims, small vocab -> low accuracy
- Dale-5: **(done)**
  - small hidden/embedding dims, small vocab -> low accuracy
  - small hidden/embedding dims, bigger vocab -> higher accuracy

Dataset	h_s	e_s	h_r	e_r	$\|V\|$	Acc.
Dale-2	10	10	10	10	10	95%
Dale-2	50	50	128	128	10	50%
Dale-5	10	10	10	10	10	23%
Dale-5	10	10	10	10	20	23%
Dale-5	10	10	10	10	100	41%

Table 1: Result of the experiments with different hidden sizes, embedding sizes and vocabulary sizes

- ... test different hidden/embedding dims
- ... test 3/4 objects

The results of the experiments are summarized in Table 1. For the *Dale-2* dataset it can be clearly seen that very small embedding and hidden sizes are beneficial for identifying the correct object. The receiver identifies indeed almost every sample correctly with all sizes of 10. When the hidden and embedding sizes are increased, the guesses by the receiver are random. Interestingly, a vocab size of 10 is enough to communicate a meaningful message for the *Dale-2* dataset.

The result change, when using the *Date-5* dataset with four distractors. With low hidden, embedding and vocab size, the agents barely pass the random baseline with 23%. Only increasing the vocabulary size to 100 raises the accuracy by almost 20% to 43%. This accuracy is still far lower than the 95% with the *Dale-2* dataset.

6 Discussion

TODO:

- maybe calculation of loss (multiplicating instead of summing loss per token), unlikely, since sequence length short -> shouldn't result in big differences
- reducing dims of image better the increasing dims of message, increasing dims is not learnable for models
- Vocab:
  - vocabulary could describe attributes of target image (non-discriminative) or describe only differences (discriminative)
  - in second case, two images is a far easier task than five images. Hence, much lower accuracy

Two interesting conclusions can be drawn. First, the hidden as well as the embedding sizes need not be bigger than the vocabulary size. Even though that means that the image encodings of 4096 or even higher dimensions need to be compressed to fit the vocabulary size, this will still help to produce and interpret more meaningful messages. The reason for this is very likely that neural models have difficulties to upscale from lower dimensions (e.g. from low h_r to high e_r).

The second conclusion that can be drawn looks at the difference between the two datasets. Unsurprisingly, the agents have a much higher difficulty to discriminate a target object from four instead of one object. Still the difference in the necessary vocabulary size is striking. There are 48 possible objects. Still a vocabulary size of only 10 is enough for an almost perfect accuracy with two objects. This hints to the fact, that the agents don't describe the complete target object, but only rely on discriminating attributes between the objects. This gets more complex when the agents need to discriminate between five objects. A bigger vocabulary offers a bigger space

7 Conclusions and further work

Acknowledgements

References

Mark Bartlett and Dimitar Kazakov. 2005. The origins of syntax: from navigation to language. *Connection Science*, 17(3-4):271–288.

Joyce Y Chai, Rui Fang, Changsong Liu, and Lanbo She. 2016. Collaborative language grounding toward situated human-robot dialogue. *AI Magazine*, 37(4):32–45.

Herbert H. Clark. 1996. *Using language*. Cambridge University Press, Cambridge.

Robin Cooper. 2023. *From Perception to Communication: A Theory of Types for Action and Meaning*, volume 16 of *Oxford Studies in Semantics and Pragmatics*. Oxford University Press Press.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2016. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *CoRR*, abs/1612.06890.

John D. Kelleher and Simon Dobnik. 2020. Referring to the recently seen: reference and perceptual memory in situated dialogue. pages 41–50.

Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. 2019. Egg: a toolkit for research on emergence of language in games.

Simon Kirby, Hannah Cornish, and Kenny Smith. 2008. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31):10681–10686.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2016. Multi-agent cooperation and the emergence of (natural) language.

Terry Regier. 1996. *The human semantic potential: spatial language and constrained connectionism*. MIT Press, Cambridge, Massachusetts, London, England.

Deb Roy. 2005. Semiotic schemas: a framework for grounding language in action and perception. *Artificial Intelligence*, 167(1-2):170–205.

Luc Steels and Martin Loetzsch. 2009. Perspective alignment in spatial language. In Kenny R. Coventry, Thora Tenbrink, and John. A. Bateman, editors, *Spatial Language and Dialogue*. Oxford University Press.

Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. 2018. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942.

A Example Appendix