Referring as a collaborative process: learning to ground language through language games

Dominik Künkele[1] and **Simon Dobnik**[1,2]
Department of Philosophy, Linguistics and Theory of Science[1]
Centre for Linguistic Theory and Studies in Probability (CLASP)[2]
University of Gothenburg, Sweden
dominik.kuenkele@outlook.com and simon.dobnik@gu.se

Abstract

How do artificial agents based on neural networks coordinate on a new language through referential games over 3-d scenes? We extended a popular CLEVR dataset to control for different combinations of features of target and distractor objects and examine the success of referential grounding learned by the agents.

1 Introduction

Agents interact with the physical world through their actions and perception, and with other agents through language. Their sensors and actuators 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, the language that is used to communicate with other agents is based on representations that are composed of a limited set of discrete and arbitrarily chosen symbols. How can both domains and representations arising from these interactions be combined? How are the ranges of measurements expressed in a continuous domain mapped to discrete linguistic labels? How is ambiguity and underspecification resolved? How can agents achieve it through interactive grounding (Regier, 1996; Roy, 2005; Cooper, 2023)?

In this paper we explore how agents based on artificial neural networks learn referential grounding of entities in images of 3-dimensional scenes through language games (Clark, 1996; Bartlett and Kazakov, 2005; Kirby et al., 2008; Steels and Loetzsch, 2009; Zaslavsky et al., 2018). One agent is describing the entities represented as features within bounding boxes of objects, inventing new vocabulary as necessary. The other agent learns to interpret the reference of symbols by identifying one of the bounding boxes based on object attributes such as shape, colour and size. Both agents learn through the success of interaction. The novelty of our work, compared with the previous work with this setup (Kharitonov et al., 2019; Lazaridou et al., 2017), consists the extension of the popular CLEVR dataset (Johnson et al., 2016) with new artificially generated 3-d scenes of objects. These can be referred to based on attributes such as *shape*, *colour* and *size* and discriminated based on different overlaps of these attributes between the target and the distractor objects.

2 CLEVR-Dale-2 and Dale-5

We extend the CLEVR dataset (Johnson et al., 2016) by dividing the objects into one *target object* and *distractors* and by controlling for the number of shared attributes between these groups as in the GRE algorithm in (Dale and Reiter, 1995). The target object is always unique, because at least one attribute is different from the distractors. Each distractor can share a maximum of two attributes with the target object. There is no restriction on the relation between distractors, hence it is possible to have multiple identical distractors in one image. Given the ranking of features in the original GRE algorithm, the target object is therefore identifiable either by the *shape* (1), the *shape* and *colour* (2) or the *shape*, *colour* and *size* (3). For each image, fixed-size bounding boxes are extracted around the centre-point of each object. 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. Examples are shown in Appendix A.

3 Language games

The language games were developed and run in the EGG framework (Kharitonov et al., 2019).[1] Both our sender and receiver have a similar architecture to the *agnostic sender* and *receiver* of (Lazaridou et al., 2017), as shown in Appendix B. One central

[1] https://github.com/DominikKuenkele/MLT_Master-Thesis

difference is the production of the message. As we focus on sequences of referring expressions, made-up of different attributes, our models produce sequences of symbols for the message instead of a single symbol to refer to an image. This is done by using an encoder LSTM (sender) and a decoder LSTM (receiver) to encode language descriptions. Another difference is that both sender and receiver receive visual input as segmented objects rather than as two images. The order of the objects is random, except that the first object for the sender is always the target object to be referred to. For the sender, the images are passed through *ResNet101* (He et al., 2016) and a following linear layer that reduces the dimensions to an embedding size e_s. All embedded images are concatenated and passed through another linear layer to reduce the dimensions to the hidden size h_s. This is then used as the initial state of the encoder LSTM. After, the sequence is created through Gumbel-Softmax relaxation (Jang et al., 2017). The receiver also encodes all images using *ResNet101* with a following linear layer, reducing it to e_r. The sequence, received by the sender is the input for its decoder LSTM, where the hidden state with a dimension of h_r is randomly initialized. After each step of the LSTM, the receiver calculates the dot product between the hidden state and all of its image encodings. The receiver then 'points' to one of the images by applying the softmax function over the results of the dot products. The loss is calculated using the NLL-loss. Following, the losses for all steps are summed up, and all weights of the receiver as well as the sender are updated based on this summed loss.

4 Experiments and results

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 $|V|$. Table 1 shows the accuracy of the models calculated on the success of communication if the receiver can identify the target object. A random guess corresponds to 50% in the *Dale-2* dataset and 20% in the *Dale-5* dataset.

For the *Dale-2* dataset it can be clearly seen that an embedding size and hidden size that are as high as the vocabulary size are beneficial for identifying the correct object. The receiver identifies 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 vocabulary size of 10 is enough to communicate a meaningful message for the *Dale-2* dataset. Using *Dale-5* with four distractors and with low hidden, embedding and vocabulary sizes, the agents barely pass the random baseline with 23%. Only increasing the vocabulary size to 100 raises the accuracy by almost 20% to 43% which is still considerably lower than the 95% of the *Dale-2* dataset.

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: Results: h are different hidden sizes, e embedding sizes and $|V|$ vocabulary sizes.

5 Discussion and future work

Unsurprisingly, the agents have a much higher difficulty to discriminate a target object from four instead of one distractor. Since we discriminate objects based on properties that are also distinguished in human cognition (colour, size, shape), we expect that the vocabulary onto which the agents converge reflects these categories and is therefore close to human vocabulary. There are 48 possible combinations of attributes. Still, for Dale-2, 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 discriminative attributes between the objects. The need for a more detailed description of discriminative attributes is higher when more distractors are involved. Therefore, the models need to learn more combinations of symbols in order to attest to this higher level of detail and especially how to relate them to features in the images.

In our ongoing work we are investigating deeper the emerged language and the new vocabulary, in particular whether it uses similar categories as human language and how its words are combined to form complete messages. In future work we will also extend the learning to the relations between entities and the features required to capture them.

References

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

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. *Cognitive science*, 19(2):233–263.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. *arXiv*, arXiv:1611.01144 [stat.ML]:1–13.

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.

Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. 2019. EGG: a toolkit for research on emergence of language in games. *arXiv*, arXiv:1907.00852 [cs.CL]:1–6.

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. 2017. Multi-agent cooperation and the emergence of (natural) language. *arXiv*, arXiv:1612.07182v2 [cs.CL]:1–11.

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 Extended CLEVR datasets

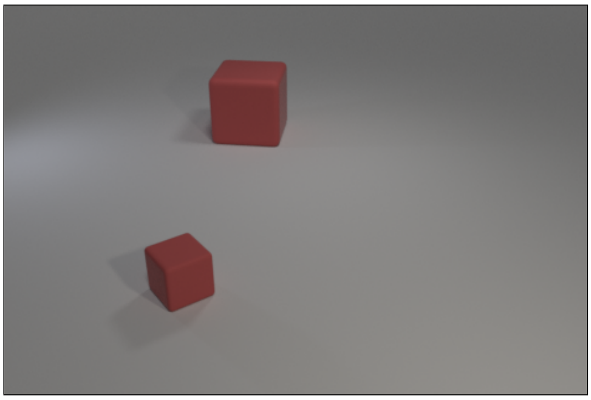


Figure 1: An example from the Dale-2 dataset

In Figure 1, the small red cube is the target object. Since all attributes except for the size are shared with the distractor, all three attributes are necessary, to identify it following Dale and Reiter (1995)'s rules, namely the *small red cube*.

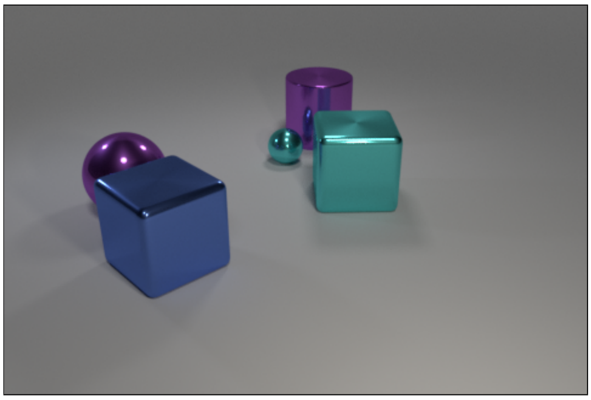
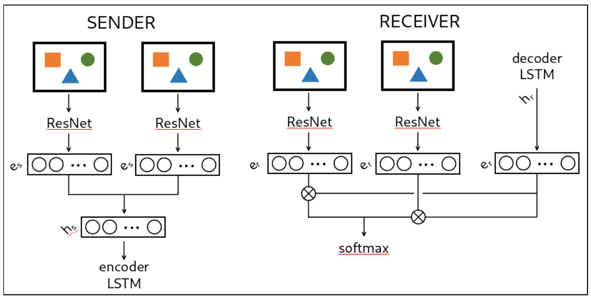


Figure 2: An example from the Dale-5 dataset

The target object in Figure 2 is the purple cylinder. It shares the same colour and size with the purple sphere, the same size with the two cubes and no attribute with the turquoise sphere. It can be uniquely identified as the *cylinder*.

B Setup of the language game


SENDER
RECEIVER
decoder LSTM
h_r
ResNet
ResNet
ResNet
ResNet
e_s
e_s
e_r
e_r
e_r
h_s
encoder LSTM
softmax