Discriminating 3D-objects using

Anonymous ACL submission

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)?

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. One agent is describing the entities, more specifically bounding boxes of objects and the other agent learns to interpret the reference of symbols by identifying and referring to one of the bounding boxes 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; Lazaridou et al., 2016) 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 as one of the potential bounding boxes.

2 Our datasets: CLEVR-Dale-2 and Dale-5

For these experiments, we extend the CLEVR dataset, by dividing the objects into a *target object* and *distractors*.[1] 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 between these groups is defined by rules researched in Dale and Reiter (1995). The target object is always unique and at least one attribute is different from the distractors. First, a random target object is created. In the next step, the distractors are created at a random location in the scene. Each of the distractors can share a maximum of two attributes with the target object. There

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

is no restriction on the relation between distractors, namely it is possible to have multiple identical distractors in one image. 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. The way, how the distractors are created is the same. 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.

Figure 1 shows an example of a generated image of the Dale-5 dataset. Both datasets also contain ground truth information about each image, including attributes and locations for all objects.

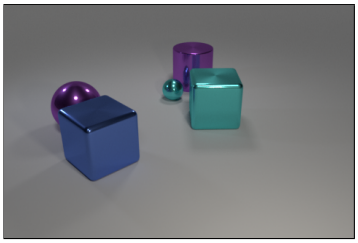

Figure 1: Example of a generated image in the Dale-5 dataset with the extended code

3 Language game

The language game was developed and run in the EGG framework (Kharitonov et al., 2019). [2] 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. 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) instead of a single symbol (which corresponded to different concepts of the images). This is done using an encoder LSTM (sender) and a decoder LSTM (receiver).

The target object is always the first image that is passed to the sender. The order of the objects for the receiver is random. Figure 2 shows the architectures of the sender and the receiver. 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. The sequence is then created through Gumbel-Softmax relaxation (Jang et al., 2016).

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.

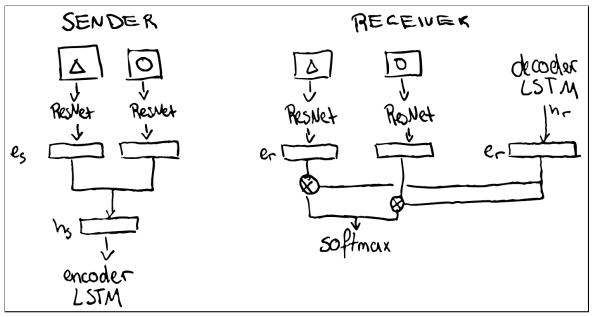

Figure 2: Architectures of the sender and receiver

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.

4 Results

The results of the experiments are summarized in Table 1. For the *Dale-2* dataset it can be clearly seen that an embedding and hidden size that are as

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

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

high as the vocabulary size 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 vocabulary size of 10 is enough to communicate a meaningful message for the *Dale-2* dataset.

The results 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.

5 Discussion

Two interesting conclusions can be drawn. First, the hidden as well as the embedding sizes need to be close to the vocabulary size. This even applies for very low vocabulary sizes, which means that the image encodings need to be compressed to the same low dimensions. 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) as opposed to learn how to extract the important information from a vector.

The second conclusion that can be drawn looks at the differences between the two datasets. Unsurprisingly, the agents have a much higher difficulty to discriminate a target object from four instead of one object. Since we discriminate objects based on properties that are also distinguished in human cognition (color, size, shape) we expect that the vocabulary onto which the agents will converge will reflect these categories and therefore be 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 to attest to this higher level of detail and especially how to relate them to features in the images.

6 Conclusions and further work

In this research, we introduced an extended version of the CLEVR dataset that offers a high control over the relations between generated objects. This should help to study, how models can learn these relations and refer to them as well as the objects. In our research we used this dataset to study, how agents can learn to communicate and refer to these objects in a language game setup. We found that a bigger number of distractors requires a higher vocabulary size. This hints to the fact that the agents learned to communicate discriminative attributes between the images.

In future research, we want to investigate deeper the emerged language and the new vocabulary. Especially, we want to test if our hypothesis that this emerged language uses similar categories as human language holds true and how these words are combined to form complete messages.

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.

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. 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