Instructions for *ACL Proceedings

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

TODO:

- how can models reidentify attributes (and objects) from learned encodings in an artificial language **(done)**

Language games with a sender and receiver offer an opportunity on to look at how neural models condense seen information into a distinct representation of symbols. The receiver needs to interpret the meaning of these symbols and combine it with other information. This paper studies how the agents can encode real-looking objects into distinct symbols and decode this message to reidentify this object. Specifically, the study explores discrimination games using an extended version of the CLEVR dataset, focusing on the identification and differentiation of objects based on their attributes such as shape, color, and size.

2 Background and previous work

3 Materials and methods

TODO:

- creation of dataset (CLEVR) **(done)**
  - multiple 'real' objects in scene **(done)**
  - 3 attributes (color, size, shape) differentiate objects **(done)**
  - using 'dale' setup to uniquely identify target object **(done)**
  - extracting bounding boxes **(done)**
- building a language game using EGG **(done)**
- based on feature extractors ResNet **(done)**
- setup of discriminating game of objects in image **(done)**
- message encoder/decoder is auto-encoder

3.1 Dataset

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

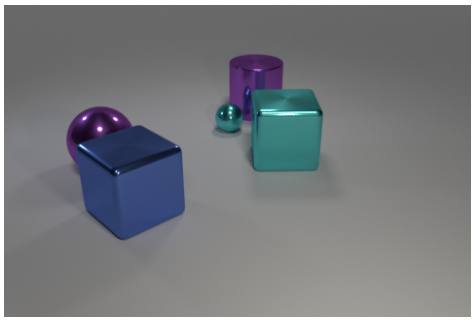

Figure 1: Example of a generated image with the extended code

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 *DaleTwo* dataset contains one target object and one

[1]The extended source code can be found on GitHub: https://github.com/DominikKuenkele/MLT_Master-Thesis_clevr-dataset-gen

distractor, while the *DaleFive* 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.

3.2 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 $|V|$.

Dataset	h_s	e_s	h_r	e_r	$\|V\|$	**Acc.**
DaleTwo	10	10	10	10	10	95%
DaleTwo	50	50	128	128	10	50%
DaleFive	10	10	10	10	10	23%
DaleFive	10	10	10	10	20	23%
DaleFive	10	10	10	10	100	41%

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

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

4 Results

TODO:

- DaleTwo: **(done)**
  - small hidden/embedding dims, small vocab -> high accuracy
  - high hidden/embedding dims, small vocab -> low accuracy
- DaleFive: **(done)**
  - small hidden/embedding dims, small vocab -> low accuracy
  - small hidden/embedding dims, bigger vocab -> higher accuracy
- ... test different hidden/embedding dims
- ... test 3/4 objects

The results of the experiments are summarized in Table 1. For the *DaleTwo* 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 *DaleTwo* dataset.

The result change, when using the *DateFive* 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 *DaleTwo* dataset.

5 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

6 Conclusions and further work

Acknowledgements

References

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.

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

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

A Example Appendix