Language Emergence in Referential Games: Encoding Relative Dimensions

Yke Rusticus
University of Amsterdam
yke.rusticus@student.uva.nl

Maarten Peters
University of Amsterdam
maarten.peters@gmail.com

Abstract

The study of language emergence in referential games captures an essential aspect of natural language: communication. In such games, two agents communicate through messages in a visual environment, and tend to come up with a language that captures visual elements of the images presented. Work in this field, however, often focuses on the descriptions of single images. In this work we took into account a "reference" image, as to investigate whether the agents could come up with relative terms such as "larger" or "smaller". We provide both qualitative and quantitative analyses of the language and show that this task is harder for the agents than expected.

1 Introduction

The study of artificial intelligence and machine learning has a long history and often relies on large datasets and supervised learning to *teach* machines a large collection of tasks. For linguistics, this results in a statistical distribution of language usage, but fails to encode the functional aspects of a language. Consequently the meaning of language becomes misrepresented by failing to capture common knowledge.

Also one might argue that in order to *understand* a language, one most be able to use it for communication, requiring some form of interaction. This is the premise of the research-field of language emergence in multi-agent systems (Lazaridou et al., 2016; Mikolov et al., 2016; Crawford and Sobel, 1982). These networks represent a simplified environment where agents communicate using their own developed languages, with the goal to develop a common descriptive language of the task at hand. A classic example is the Lewis signalling game (Lewis, 2008), where two agents exchange signals in order to accomplish a common task without sharing their state. In visually grounded emergent linguistics this is generally represented by picking a target image based on the senders description. Because the emergence of language is a core component for the study of artificial intelligence (Kirby, 2002), this task is well suited to study language, the influence of grammar and syntax, and even the translation between existing languages (Lee et al., 2017).

For this study, we consider whether agents in a game of this setup can also capture relative information, i.e. spatial transformations, based on visual input. Next sections discuss related work, the methods used, and the experiments, in which we provide both qualitative and quantitative analysis of the emerged language. The given task showed to be difficult to solve for the agents, the reason for which is discussed further in the final sections.

2 Related Work

The study of emergent languages are commonly aimed at representing natural language properties, such as the use of discrete symbols (Havrylov and Titov, 2017) or descriptions of concepts (Lazaridou et al., 2016). Whether the emergent languages contain characteristics like compositionality and (human-)interpretability is still an area of debate (Kottur et al., 2017), there is no question that they can perform the Lewis signalling game accurately (Lewis, 2008).

But most of these studies focus on distinct concepts, like hand-written digits (LeCun, 1998) and (non-)living objects (McRae et al., 2005). This paper poses the question how they perform on recognizing relative changes in the form of spatial transformations. Recent related work has focused on the language evolution of trajectory encoding (Chaabouni et al., 2019), but without a visually grounded task and using stricter rules for communication. Although this study is based on a

multi-agent environment, its mechanics rely on agent evolution through parent-child communication (Sutskever et al., 2014), rather than using continuous dialogue between two static agents. This study teaches the agents a trajectory along a grid, which shows similarities to our study, encoding relative spatial information, but with a purely visual grounding, as illustrated in figure 1.

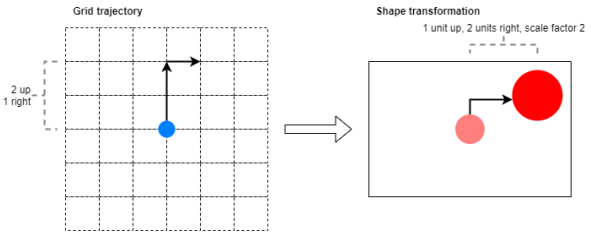

Figure 1: Information comparison to Chaabouni (2019).

A more critical work by Kottur et al. (2017) used a similar signalling game set up, but researched the emergent language for different image attributes, like *shape*, *color* and *style*. This study will strike a balance between both papers (Kottur et al., 2017; Chaabouni et al., 2019) and establish whether emergent language can encode relative change between two images.

3 Approach

First of all, a set of artificial images was created. Each image in this dataset consists of a square (ones) on an empty background (zeros), which could vary in three attributes: x-/y-coordinates and size.

The general game setup (see Fig. 2) consists of two agents: the Sender sees a target and a reference image, which differs from the target in only one attribute, and can send a message to the Receiver; the Receiver sees the the target image and several distractors, and needs to pick the target image based on the message and the reference image. Specifically, the Receiver maps the reference image in combination with the message to a representation space that is shared with the encoded target and distractor images. The task is successful if the encoded message + reference image has a smaller MSE with the encoded target image than with the distractors. In mathematical terms, we optimize for the loss

$$\mathcal{L} = -\text{MSE}(x, d_0) + \log \sum_j \exp(-\text{MSE}(x, d_j)), \tag{1}$$

where x represents the message + reference representation, d_0 the target image and $d_{>0}$ the distractor images. Eq. 1 is a cross entropy loss with the negative MSE between the representations as logits.

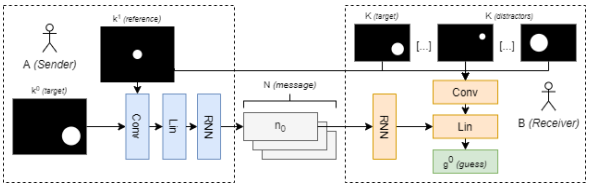

Figure 2: Game setup.

In each case we used six distractors, five of which varied in one attribute from the reference image in each direction (larger or smaller), except for the direction and attribute that was already taken by the target image. The sixth distractor was the reference itself, which was to make sure that the Receiver would not just take the reference itself as basis, while ignoring the message. The choice of distractors was supposed to encourage the agents to find the distinction between the images, which were their attributes and their values: larger or smaller.

The model architectures of the Sender and Receiver are similar. The Sender consists of a CNN vision module, a linear layer, and an RNN with gated recurrent units (GRU, (Cho et al., 2014)) to produce the messages. The Receiver consists of a vision module and a linear layer to combine the reference image and the message into a single representation. The Receiver uses the same (however different from the Sender) vision module to process the reference, target and distractor images.

We used the EGG toolkit to implement the game (Kharitonov et al., 2019), with as starting point the pre-implemented tutorial for a referential game[1]. Within this framework, we trained the models using Gumbel-Softmax. Moreover, we used an effective vocabulary size of 3 and a message length of 2, such that each attribute and direction could be encoded by the model.

We used 10K distinct training samples and 2K test/validation samples. Variations in attribute val-

[1] https://github.com/facebookresearch/EGG/blob/master/tutorials/EGG%20walkthrough%20with%20a%20MNIST%20autoencoder.ipynb

Game	Accuracy	Probe: attribute	Probe: attribute + direction	Best captured	Worst captured
Without reference	0.41 ± 0.03	-	-	-	-
With reference	0.55 ± 0.01	0.5	0.66 ± 0.01	”more to the left”	”scale down”
With reference + pretrained vision	0.54 ± 0.02	0.57 ± 0.03	0.70 ± 0.06	”scale up”	”scale down”

Table 1: Performance and analysis for each of the tested game setups. Accuracies were averaged over three runs. The probe accuracies were averaged over the classification accuracy of each of the separate cases within the category. ”Best captured” and ”worst captured” show the classes which had the highest and lowest classification accuracies respectively. Values show one standard deviation error, except if the error was smaller than 10^{-3}.

ues were randomly sampled. Still, given the number of samples we expected there to be overlap between training and test images, but we did not attempt to change that as we were mostly interested in the language, not the performance. We used a batch size of 64 for training.

3.1 Language Evaluation

In order to qualitatively measure consistency across situations and resulting messages, we created a simple test environment. In this environment we showed the Sender samples of targets and references that all differed in the same attribute in equal amount. For each sample we counted which message was produced by the Sender. If for each sample the Sender uses the same message, it is said to be consistent. We did this for each attribute change in each direction, larger and smaller (see figure 3).

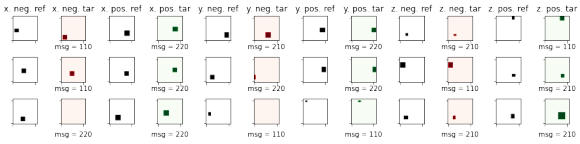

Figure 3: Examples of the data used for evaluation.

We also implemented a simple linear classifier, which had the goal to predict which attribute change was encoded by the Sender. The better this classifier is able to succeed at this task, the more we can say that the model captures the correct attribute changes, and thus captures relative terms such as ”larger”, ”smaller” or ”more to the left”.

4 Experiments and Results

We compared three game setups in performance. As a baseline we implemented the game described in Section 3, but without reference images. In the second setup, reference images where present and each parameter in the model was trained simultaneously. Lastly, we used a fixed vision module for both the Sender as Receiver that was pre-trained to recognize specific attribute changes between two images. The accuracy over training is showed in figure 4. We can see that the baseline model has large fluctuations in performance during training, indicating that without reference the agents are difficult to train. The model using an untrained vision module is much more stable, even with different message lengths, but reaches a plateau of performance. The model with a pre-trained vision module reaches similar accuracy, but seems to show promise by having slightly smaller performance fluctuations and does not flatten out at 40 epochs of training.

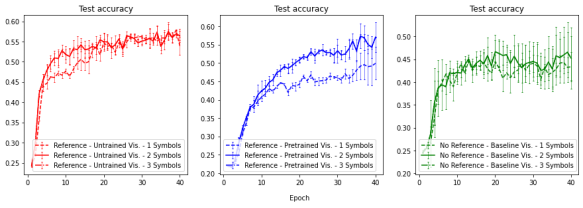

Figure 4: Model accuracy was evaluated during training at each epoch and averaged over three runs for each of the message lengths.

Final test results are summarized in Table 1. The following subsections describe the evaluation of the emerged language.

4.1 Qualitative analysis

From figure 5 we can clearly see there is no single message that represents a positive (green) or negative (red) direction change for a specific attribute. It is clear that the different vision modules result in different message distributions, but they do not have a clearer information encoding related to the original attributes. Given these distributions we can see both models have two sets of similar distributions: a flat distribution and a slanting distribution. Each model actually contains approximately four unique distributions, indicating that from the original attributes point of view it could positively identify these transformations, but not the others. This could partly explain the model could not im-

prove over > 0.5 accuracy, failing to identify at least 33% of the attributes uniquely.

More striking is the fact that the third symbol was not used. Also the "12" message was used, but the "21" was not. This indicates that there might have been implementation errors or the model simply disregarded these as viable options for the information encoding. Within this study we could not find any errors in the implementation, but a more detailed investigation might be necessary.

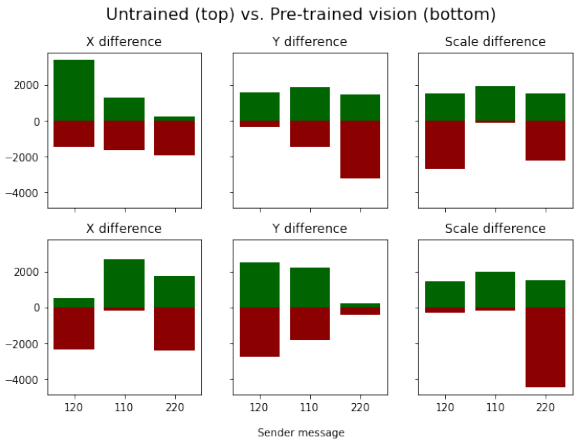

Figure 5: Qualitative analysis results over 5K samples for each direction, performed for the untrained and pre-trained vision modules.

4.2 Quantitative analysis

We implemented a linear classifier (similar to (Hupkes et al., 2018)) that we trained to classify which attribute and which direction (larger or smaller) was referred to in messages sent by the Sender. As ground truth we took the actual attribute changes between the target and reference images from the test set. We trained for each specific class (attribute only or attribute + direction) a separate classifier, such that the classification task was binary. Either a message encodes a specific change (class 1), or it does not (class 0). Note that the majority of samples lies in class 0. We found that the classifier could not distinguish these classes based on all data, and would classify each case as class 0. So, in order to create a fair class balance, we took all cases of class 1, and randomly sampled an equal number of cases of class 0. The random baseline for each case thus had an classification accuracy of 0.5. Table 1 shows that only for the standard game with reference, the classifier was not able to distinguish which attribute was referred to in the messages. Surprisingly though, for the same game it was able to predict better than random when a specific attribute change with direction was referred to in the messages. This suggests that an abstract language was used that was not consistent over attribute changes independent from the direction of change. Rather, the language was dependent on only the combination of attribute and direction of change. Carefully, we could therefore exclude that compositionality was present in this language. Furthermore, we found that the game played with a pre-trained vision module showed slightly more distinguishable messages. However, this was not a significant difference in the case for attribute change and direction combined.

5 Discussion

In conclusion, the agents were not able to perform well on the given task. Although they achieved quite high accuracies (> 0.5) for a referential game with 6 distractors, they only used a small part of the given vocabulary. The reason for this remains unclear. Possibly part of the problem was the way the data was generated. For example, the size of an object was slightly dependent on the xy-coordinates, as the objects were not allowed to fall off an image. This could have resulted in smaller objects around the edges, with additional effect on what kind of distractors would be generated. The model might have exploited these biases in the data, such that it prioritized learning these over learning relevant messages. Chaabouni et al. (2019) showed successes with a game that also implemented spatial transformations, however in the form of non-visual inputs. We found that in a visual setting, our agents struggle to capture the relative spatial transformations. We investigated whether agents would be able to encode differences between two images in messages that were similar in function to, for example, what we know as "more to the left", "further up" or "further down". Probe models showed that they did capture these terms to some extent, however not impressively. To relate these results to Kottur et al. (2017), we can again confirm that natural language, this time in relational terms, did not come naturally. Nevertheless, it would be interesting to further improve this experiment, as referential frames come as natural to humans. First steps of future work would therefore be to make the data less biased, and to dig deeper into the way we humans interpret relative differences.

References

Rahma Chaabouni, Eugene Kharitonov, Alessandro Lazaric, Emmanuel Dupoux, and Marco Baroni. 2019. Word-order biases in deep-agent emergent communication. *CoRR*, abs/1905.12330.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Vincent P Crawford and Joel Sobel. 1982. Strategic information transmission. *Econometrica: Journal of the Econometric Society*, pages 1431–1451.

Serhii Havrylov and Ivan Titov. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in neural information processing systems*, pages 2149–2159.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. 2019. Egg: a toolkit for research on emergence of language in games. *arXiv preprint arXiv:1907.00852*.

Simon Kirby. 2002. Natural language from artificial life. *Artificial life*, 8(2):185–215.

Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. 2017. Natural language does not emerge ńaturallyín multi-agent dialog. *arXiv preprint arXiv:1706.08502*.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2016. Multi-agent cooperation and the emergence of (natural) language. *arXiv preprint arXiv:1612.07182*.

Yann LeCun. 1998. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

Jason Lee, Kyunghyun Cho, Jason Weston, and Douwe Kiela. 2017. Emergent translation in multi-agent communication. *arXiv preprint arXiv:1710.06922*.

David Lewis. 2008. *Convention: A philosophical study*. John Wiley & Sons.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559.

Tomas Mikolov, Armand Joulin, and Marco Baroni. 2016. A roadmap towards machine intelligence. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 29–61. Springer.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.