Image classification pressures in language emergence

Barbara Brocades Zaalberg **Błażej Dolicki**
Msc Artificial Intelligence
University of Amsterdam
{barbara.brocadeszaalberg, blazej.dolicki}@student.uva.nl

Abstract

Studying the emergence of language can help to better understand human language and improve natural language understanding in systems such as chatbots or virtual assistants. We use the Lewis' signaling game (Lewis, 1969) as a simple environment where two agents communicate about images to achieve a common goal. We propose two pressures that leverage image classification to incentivize communication about high-level features rather than pixel values which should allow better generalization: MultiLabel Binary image classification (MLB) and MultiClass Target image classification (MCT). We show that while MLB performs similarly or worse than the baseline, MCT obtains substantially better results and thus we conclude that the latter successfully pressures the model to include high-level features in the messages. Our code is publicly available [1]

1 Introduction

The ability to understand and communicate through natural language is widely considered as a requirement for achieving human-like intelligence (Mikolov et al., 2016; Turing, 1950). However language knowledge in humans is often abstract and unconscious (Ellis, 2005), they are able to use and produce natural language but are unable to explain how the system works. Studying the emergence of language can give insight into how this complex system arises which can simultaneously help to improve natural language understanding in machines, for example in applications like chatbots and virtual assistants.

The field of language emergence uses simple environments to examine the language that materializes during communication of multiple agents that cooperate to achieve a common goal. An example of such an environment is the Lewis's signaling game (Lewis, 1969) where one agent, the Sender, observes an image and sends a message to the second agent, called the Receiver, who describes the image. Afterwards, the Receiver is presented with multiple images and should select the image send by the Sender, if it selects the right image the communication was successful. Ideally the message conveyed by the Sender describes high-level concepts of an image rather than pixel-level differences as this will lead to better generalization and robustness - small perturbations in input images should also lead to small differences in the Receiver's outputs. Unfortunately, Bouchacourt and Baroni (2018) showed that currently the former is the case. To enforce the agents to communicate about high-level features, we propose to extend the signaling task with an image classification task. In image classification a model has to learn high-level features of images since instances of the same class can vastly differ on pixel-level (e.g. images with a label "dog" can contain dogs of different breeds, in different backgrounds etc.) however they share the same high-level features (characteristic dog nose, four legs, eyes). Our main research goal in this work is to incorporate image classification as an additional task to pressure the model to include high-level features in its communication protocol and simultaneously retain high communication success.

In Section 5.1 we show that both pressures retain high communication success while optimizing for the additional image classification task. To evaluate whether our extensions indeed have the intended effect we additionally compute their accuracy on noise images and topographic similarity between messages and input image representations (Section 5.2 and Section 5.3). While MLB performs similarly or worse than the baseline, MCT obtains

[1] https://github.com/blazejdolicki/nlp2-language-emergence

substantially better results and thus we conclude that the latter successfully pressures the model to include high-level features in the messages.

2 Related Work

Lazaridou et al. (2017) proposed a variation of Lewis' signaling game (Lewis, 1969) with real images using single-symbol messages. Havrylov and Titov (2017) extended this task with variable-length messages which are more similar to human language. Additionally, they only supply the sender agent with the target image without showing other images seen by the receiver (*distractors*). Both publications attempted to apply pressures (i.e. add additional constrains to the models) to obtain a message protocol similar to natural language. Lazaridou et al. (2017) leverages the image classification to force the sender to use object labels as messages while simultaneously optimizing for the initial objective. We do not attempt to map the image labels one-to-one to the message but rather indirectly affect the message protocol to make it more reliant on high-level concepts rather than low-level features. From an architectural perspective, in our work the receiver predicts image labels based on the hidden state of the receiver (the message) as opposed to the sender layers. This also allows us to use an arbitrary vocabulary size instead of having a vocabulary size that is tied to the number of classes in the image classification dataset. Havrylov and Titov (2017) tried minimizing the Kullback-Leibler (KL) divergence between the distribution of the message protocol and the natural language based on image descriptions. After Bouchacourt and Baroni (2018) showed that the agents communicate based on pixel values rather than concept-level features, Rodríguez Luna et al. (2020) attempted to mitigate this issue by perturbing the location and the position of the objects in images from a synthetic dataset. Our goal is similar to Rodríguez Luna et al. (2020), but our different approach allows us to address the problem using real images instead of synthetic dataset and it doesn't require additional image preprocessing.

3 Approach

This section will discuss the approach taken to pressure the agents to communicate about high level concepts. As our baseline we play a variable-length signaling game similar to Havrylov and Titov (2017). The sender observes only the target image while the receiver is exposed to the target image and the distractors. We use a multiclass cross-entropy loss L_{sg}.

3.1 Pressures

We apply two visual pressures [2]: MultiLabel Binary image classification (MLB) and MultiClass Target image classification (MCT).

3.1.1 MultiLabel Binary image classification

In MultiLabel Binary image classification (MLB), the model additionally predicts for each image (the target image and distractors) whether it is of the same class as the target image using multilabel binary cross-entropy loss L_{MLB}. The total loss is:

$$L = L_{sg} + \alpha \cdot L_{MLB}$$

where α is a weight determining the extent to which the additional loss impacts the total loss.

3.1.2 MultiClass Target image classification

In MultiClass Target image classification (MCT) the model additionally predicts the class of the target image using a cross-entropy loss. The total loss is:

$$L = L_{sg} + \alpha \cdot L_{MCT}$$

where α is a weight determining the extent to which the additional loss impacts the total loss.

3.2 Evaluation

We evaluate our method with the following experiments:

3.2.1 Signaling game accuracy

The standard signaling game accuracy shows how many times the receiver correctly selected the target image. When this metric reaches proximity of 100% we can conclude that the agents communicate successfully. It is required that applying pressures doesn't jeopardize the original task performance. Therefore, this accuracy should be comparable between the baseline and the pressures.

3.2.2 Image classification accuracy

The aim of adding the additional training task is to impact the models and the message protocol in

[2] By "pressure" we mean additional task/loss for which the model will optimize and which incentivizes the model to exhibit desired behaviour. In this work, it encourages the model to communicate about high-level features.

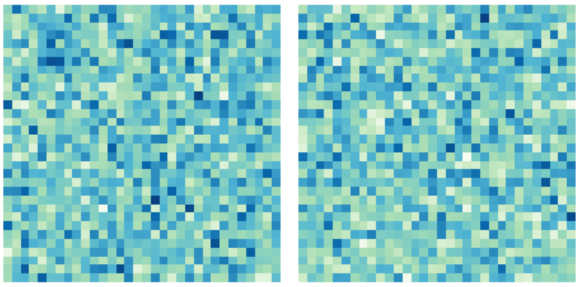

Figure 1: Noise images presented to the model at validation time

the aforementioned way. Image classification accuracy shows the extent to which the signaling game is affected by the additional pressure, therefore it should be as high as possible.

3.2.3 Signaling game accuracy on Gaussian noise images

Bouchacourt and Baroni (2018) showed that a trained signaling game model is able to distinguish between synthetic images containing random noise which clearly shows that the model does not communicate about high-level concepts but rather particular pixel values. Ideally our pressures decrease the accuracy on noise images to a random guess. See Figure 1 for an example of such noise images.

3.2.4 Topographic Similarity (TS) between messages and input image representations

Topographic Similarity is a measure based on the intuition that differences between representations of objects in the input image representation space should be similar to the differences between representations of objects in the message representation space (Lazaridou et al., 2018). High topographic similarity gives strong evidence that the additional tasks achieve the goal of making the communication more focused on high-level concepts in the images.

4 Experiments

The precise game setup and hyperparameters are as follows. To create image embeddings we use a vision module that is a pretrained ResNet56 model (He et al., 2015) that maps an image to a 64-dimensional vector. The sender and receiver are implemented as GRU cells (Cho et al., 2014) with cross-entropy loss. To optimize the loss we use the Gumbel-Softmax estimator (Havrylov and Titov, 2017). To train and validate the model we use the Cifar-100 dataset, which contains images categorised in 100 classes where each class contains 600 images (Krizhevsky et al., 2009). The training set size is 50000, the validation set size is 10000. We create the noise validation set, also containing 10000 images, from random Gaussian noise with mean equal to 0 and standard deviation equal to 1. The number of images presented to the receiver is two, one target image and one distractor image. The maximum message length that the sender sends to the receiver is 10 tokens and the vocabulary size is 100. We experiment with three probabilities that a distractor image will be of the same class as the target image, $p_{class} = \{0.0, 0.5, 1.0\}$. This is necessary since the used dataset contains 100 different classes. Always drawing the distractor image from a random class would make the classification task too easy for the model, since predicting that the distractor image is of a different class than the target image would be correct most of the time. Training was done for a maximum of 10 epochs, with an early stopping validation accuracy of 0.97. For all sender/receiver model hyperparameters see Table 3 in appendix A.

pressure	**$\mathbf{p}_{\mathbf{class}}$**	**acc**	**img acc**
baseline	0.0	0.976	-
	0.5	0.955	-
	1.0	0.945	-
MLB	0.0	0.973	0.918
	0.5	0.923	0.856
	1.0	0.935	1.000
MCT	0.0	0.972	0.379
	0.5	0.918	0.455
	1.0	0.905	0.414

Table 1: Validation signaling game accuracy and image classification accuracy for each pressure, averaged over 3 seeds.

5 Results

In this section we present the results of our evaluation methods.

5.1 Signaling game and classification accuracies

Table 1 shows the signaling game accuracy results and the image classification accuracy results on the validation set for the different tasks with different class probabilities. We see that for $p_{class} = 0$, the signaling game accuracy is comparable across all

pressure	acc	noise acc	TS
baseline	0.955	0.809	0.119
MLB	0.923	0.746	0.049
MCT	0.918	0.594	0.176

Table 2: Comparison of noise validation accuracies between the baseline and the pressures as well as the topographic similarities (TS) between messages and input image representations for the different tasks

pressures. For $p_{class} = 0.5$ we see a small performance drop in the signaling game accuracy of 0.032 for MLB and 0.037 for MCT when compared to the baseline. For $p_{class} = 1.0$ we see a performance drop of 0.01 for MLB and 0.041 for MCT. Furthermore we see that MLB achieves high image classification accuracy across all three class probabilities. The image classification performance of MCT is not high, however it is far above random (random would be somewhere around 0.01 since there are 100 classes). The image classification accuracy for MLB is noticeably higher than for MCT. This is probably because MLB is a relatively simple binary classification task whereas MCT is a multi-label classification problem.

5.2 Noise accuracy

Table 2 compares the noise validation accuracies between the baseline and the visual pressures. All results reported in Table 2 are with $p_{class} = 0.5$. For every pressure, we see a performance drop when evaluated on the noise validation set. Both pressures obtain a lower noise accuracy than the baseline - for MLB it only differs by 0.063, while MCT obtains almost random choice score of 0.594.

5.3 Topographic Similarity

Table 2 also shows the topographic similarity between messages and input image representations for the different tasks. We see that for MLB the topographic similarity score is 0.049, which is much lower than that of the baseline. The topographic similarity score of MCT is 0.176, 0.057 higher than that of the baseline.

6 Discussion

In this work we proposed an extension of the standard signaling game setup used to study language emergence. We saw that applying visual pressures leads to a comparable signalling game accuracy when compared to the baseline accuracy, which means that the agents still achieve high communication success. The image classification accuracies are high for MLB and substantially higher than random for MCT, which means the additional tasks are learned and impact the model behaviour. Part of our research was aimed at improving (i.e. decreasing) the noise accuracy results by Bouchacourt and Baroni (2018). We were not able to reproduce the baseline results reported in their paper so we can't make a meaningful comparison. Our noise validation accuracy is 0.809, the noise validation accuracy by Bouchacourt and Baroni (2018) was 0.950. When evaluating our own noise experiment results in combination with the topographic similarity scores, we can conclude two things. Firstly, we see that the noise accuracy for MLB is lower than that of the baseline. Moreover, since the topographic similarity score for this pressure is much lower than that of the baseline and the other visual pressure, we conclude that MLB is not a effective pressure to apply for our purposes. Secondly, we see that the MCT pressure obtains a noise accuracy that is close to random. MCT also has a substantially higher topographic similarity score than the baseline and MLB. This leads us to conclude that MCT successfully pressures the model to include high-level features in the messages.

In future work it could be interesting to look at other evaluation metrics to assess the effectiveness of the MCT pressure. The fact that MCT achieves near random validation accuracy on noise images and that it has the highest topographic similarity score gives a strong indication that the pressure is effective, however these measures do not give conclusive evidence that the agents are now communicating about conceptual features of the images. To achieve the same baseline results as Bouchacourt and Baroni (2018) it would be a good idea to train the model on a different dataset, e.g. ImageNet (Deng et al., 2009), since their model was trained and evaluated using this dataset. Lastly, to investigate the potential of MCT further, it could be interesting to modify the classification task a bit. In addition to the 100 normal classes the Cifar-100 dataset contains, there are 20 superclasses. An additional classification task could be to predict the superclasses of the target and distractor images, or to predict if the distractor image is of the same superclass as the target image.

References

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

Diane Bouchacourt and Marco Baroni. 2018. How agents see things: On visual representations in an emergent language game. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 981–985, Brussels, Belgium. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Nick C Ellis. 2005. At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in second language acquisition*, pages 305–352.

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*, volume 30. Curran Associates, Inc.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images.

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. 2018. Emergence of linguistic communication from referential games with symbolic and pixel input. *CoRR*, abs/1804.03984.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language.

David Lewis. 1969. *Convention*. Harvard University Press, Cambridge, MA.

Tomas Mikolov, Armand Joulin, and Marco Baroni. 2016. A roadmap towards machine intelligence.

Diana Rodríguez Luna, Edoardo Maria Ponti, Dieuwke Hupkes, and Elia Bruni. 2020. Internal and external pressures on language emergence: least effort, object constancy and frequency. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4428–4437, Online. Association for Computational Linguistics.

Alan Turing. 1950. Computing machinery and intelligence. *Mind*, 59(236):433.

Jerrold H Zar. 1972. Significance testing of the spearman rank correlation coefficient. *Journal of the American Statistical Association*, 67(339):578–580.

A Hyperparameters

See Table 3 for all hyperparameters used in this research.

Hyperparameter	Value
Embedding size	64
Hidden size sender	200
Hidden size receiver	200
Temperature for GS	1.0
Learning rate	1e-3
Learning decay factor	0.90
Batch size	64
Early stopping accuracy	0.97
Loss weight	1.0
Vocabulary size	100
Message length	10
Number of distractors	1
Max epochs	10

Table 3: Sender/Receiver model hyperparameters

B Correlation between validation and noise accuracy

Looking at the results in 5.2, one could argue that lower noise accuracy is due to lower validation accuracy. To verify that we calculated the Pearson correlation coefficient (Benesty et al., 2009) and the Spearman's rank correlation coefficient (Zar, 1972) between them for 27 runs summarized in Table 1. The results of this are 0.0830 and 0.0128 respectively (with p-values equal to 0.68 and 0.95). The fact that both metrics are close to zero shows that there is no significant correlation between the validation and noise accuracies.