Mutual influence between perception and language in artificial agents

Xenia Ohmer, Michael Marino, Michael Franke*, and Peter König*
*Authors contributed equally.

Institute of Cognitive Science, Osnabrück University, Osnabrück, Germany
{xenohmer,mimarino,michael.franke,pkoenig}@uni-osnabrueck.de

Keywords: AI; Language Emergence; Categorization; Similarity

1 Motivation

Artificial agents usually have separate modules for processing visual and linguistic information. While an abundance of work focuses on improving or analyzing the agents' behavior, the role of interactions between the two modalities in joint architectures is often overlooked. In this work, we focus on deep multi-agent reinforcement learning models, used in modern language emergence research, to study such interactions. These models are based on the idea that language derives its meaning from its use, and accordingly many of its aspects cannot be captured by supervised learning. Instead, agents are trained to solve a cooperative task, which requires the creation of a communication protocol. We use such a setup to explore the role of perception-language interactions for artificial agents whose communication is grounded in the visual world.

2 Setup

Fig. 1 visualizes the general setup. We use a subset of the *3Dshapes* data set [1]. The objects in our world can have four different shapes, sizes, and colors, respectively, resulting in 64 different objects (=classes). There are different images of the same object with varying viewpoint and background appearance. The model consists of a sender and a receiver agent playing a reference game. In each round of the game, the sender sees a random target object and produces a discrete message. Based on that message, the receiver tries to identify the target among several distractors. If the receiver is successful, both agents receive a positive reward, $R = 1$, else zero reward, $R = 0$. Weights are updated using *reinforce*, a policy gradient algorithm. We set message length $L = 3$ and vocabulary size $|V| = 4$, which means the agents can use 64 distinct messages.

Each agent has a vision module, implemented as a Convolutional Neural Network (CNN), and a language module, implemented as a Gated Recurrent Unit (GRU). The agents use the vision module to extract object representations, and the language module to generate (sender) or interpret (receiver) messages. The CNN is pretrained on a supervised object classification task, and the agents use the output of the penultimate layer as object representation.

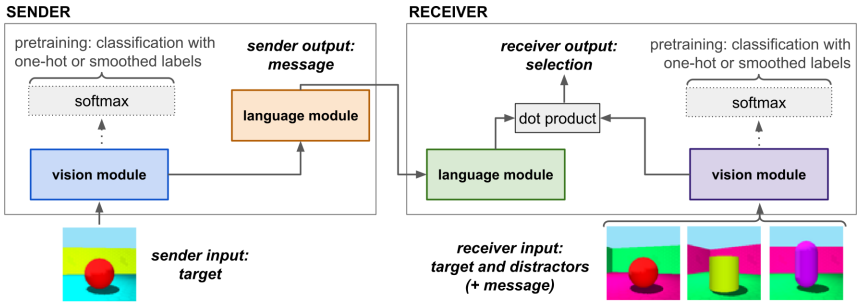

Fig. 1. Visualization of agent architecture and communication game.

3 Introducing visual biases

We introduce systematic visual biases by applying *Relational Label Smoothing* [3] to the CNN pretraining. Next to the unmanipulated *default*, we create four conditions, for details see [2]. In the *color*, *size*, and *shape* conditions, the agents are particularly good at perceiving similarity relationships between objects with respect to the corresponding feature, $f \in \{color, size, shape\}$. In an additional *all* condition, we enforce object similarities with respect to all three features simultaneously. Given a pretrained CNN, we quantify the perceptual bias for each feature, using *Representational Similarity Analysis* (RSA) [4]: We calculate the pairwise similarities between objects with respect to f, as well as the pairwise similarities between the corresponding visual representations, and correlate the two similarity matrices. Differences in these feature-wise RSA scores indicate that the vision module represents variations in a specific feature better than variations in other features, i.e. the system is biased. The analysis shows that relational label smoothing induces the intended biases in the four manipulated conditions, see Fig. 2.A. Interestingly, the *default* network exhibits an inherent color bias, probably due to easy color access via the three image channels.

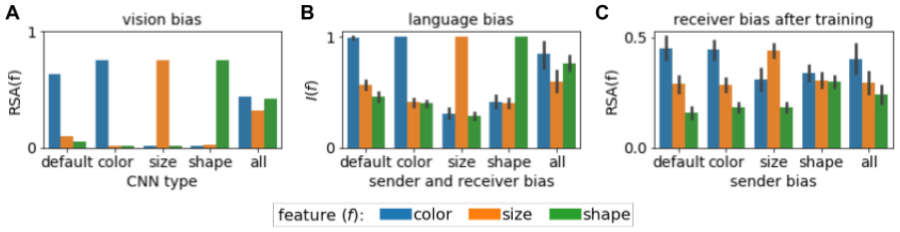

Fig. 2. A Visual bias for each CNN pretraining condition. **B** Linguistic bias for sender-receiver pairs with the same visual bias. **C** Visual bias of a default receiver after learning the languages developed by the agents in B. For B and C we report means and standard deviations across ten runs.

4 Mutual influence between perception and language

First, we study the emergent language when visually biased agents play the communication game and only the language modules are trained. For each feature f, we quantify how much information about the objects' values for f, F, is contained in the messages, M, using a normalized conditional entropy score, $I(f) = 1 - {}^{H(F|M)}/_{H(F)}$. If sender and receiver have a *color*, *size*, or *shape* bias, the messages contain mostly information about the respective feature, see Fig. 2.B. *All* agents not only cover all relevant features in their messages but also achieve higher rewards than agents in the other conditions, including *default*.

Having shown that perceptual biases lead to linguistic biases, we investigate the reverse direction. We take the sender from the trained sender-receiver pair of each condition described above, and freeze all of its weights, such that the language is fixed. Then we combine each of these senders with a default receiver and train the receiver on the communication game. In order to allow for changes in the receiver's perception, now also the vision module weights are trainable. The vision module is trained on a joint objective from the communication game and the classification task, to ensure that the receiver's perception does not deteriorate to processing only aspects relevant to the communication game. The default receiver starts out with an inherent perceptual color bias. Fig. 2.C shows how, after training, its perceptual bias has shifted towards the linguistic bias it encountered in the game.

5 Conclusion

Our work shows that there is a mutual influence between perception and language in artificial agents. For one thing, we show that perceptual biases of image processing networks—the color bias being one of them—are reflected in visually grounded language, but also that increasing perceptual sensitivity to relevant features can improve communication. For another thing, we show that an agent's perceptual bias adapts to that of its communication partner, if information from a downstream communication task is backpropagated through the vision module.

References

1. Burgess, C., Hyunjik, K.: 3D Shapes Dataset. https://github.com/deepmind/3d-shapes (2018)
2. Ohmer, X., Marino, M., Franke, M., König, P.: Why and how to study the impact of perception on language emergence in artificial agents. In Proceedings of the 43rd annual meeting of the Cognitive Science Society (CogSci) 2021
3. Marino, M., Nieters, P., Heidemann, G., Hertzberg, J.: Manipulating class relationships via relational label smoothing. Manuscript in preparation (2021)
4. Kriegeskorte, N., Mur, M., Bandettini, P.: Representational similarity analysis - connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience **2**(4) (2008)