Are deep communicating agents efficient encoders?

Rahma Chaabouni
Facebook AI / LSCP
rchaabouni@fb.com

Eugene Kharitonov
Facebook AI
kharitonov@fb.com

Emmanuel Dupoux
Facebook AI / LSCP
dpx@fb.com

Marco Baroni
Facebook AI / ICREA
mbaroni@fb.com

Studying whether controlled models can develop human-like language in interactive setups enables us to examine the prerequisites for natural language development. Moreover, learning to communicate through interaction, rather than relying on explicit supervision, is often considered a prerequisite for developing AI. For these considerations, the study of emergent communication has recently seen revived interest in the machine learning community. In particular, modern neural network models have been used in different settings to analyze the emergence of efficient communication systems (e.g, Lazaridou et al. [2016]). However, this field lack from theoretical framing leading to a poor understanding of communicating agents' behavior (Bouchacourt and Baroni [2018]).

To contribute to a better general understanding of the emerging properties of language among interacting deep agents, we ask the basic question of whether such language exhibits the *optimal coding* properties that natural language approximates (Zipf [1949]). In particular, we study an environment in which a first agent, the *speaker*, is provided with a one-hot vector input and allowed to send a sequence of discrete symbols to a second agent, called *listener*. The *listener* must then rely on this sequence of symbols to reconstruct the *speaker*'s input. The described communication task can be interpreted as a discrete auto-encoder (AE) where the encoder represents the *speaker* sending a discrete message to the decoder representing the *listener*. Specifically, we use modern recurrent networks to implement the agents.

In this basic communication task, given a distribution of inputs, we know from coding theory what is the optimal encoding for a certain length and vocabulary size. This allows us to analyze to what extent the communicating agents come up with a code that is approximating optimality for different inputs' distribution. Interestingly, we observe that, even if the AE solves the task perfectly by reconstructing the inputs, the inputs' distribution does not affect the encoding characteristics. In other words, the *speaker* would not describe frequent inputs with shorter messages, but would have a random mapping from the meaning space to the message space. This suggests that, without further constraints, such agents have no "innate" priors towards optimal coding. As a way to address this lack of bias, we introduce a regularizer on the encoding's length to create a penalty on the length of sent messages. We investigate to which extend this additional loss can enforce an optimal coding. We finally study the impact of a noisy channel looking at whether and how agents invent a noise-robust encoding.

References

Diane Bouchacourt and Marco Baroni. How agents see things: On visual representations in an emergent language game. *CoRR*, 2018.

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

George Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Boston, MA, 1949.