Language emergence and caption generation using an adversarial data set

August, 2019

1 Emergent language games

Emergent language games are experimental protocols designed to model how communication may arise among a group of agents. They typically involve agents that have to send each other messages in order to achieve some goals (common or not). The syntax and semantics of these messages are unspecified at the beginning of the experiment, and have to be determined by the agents. One of the goals of such experiments is to study under which conditions the multiple agents manage to develop an efficient protocol of communication, and when so, the properties of this protocol.

Much recent work on the topic [HT17; BB18; Laz+18] involve two agents: a *sender* and a *receiver*. The sender is exposed to some data (e.g., an image or an array of images), then produces a message that is transmitted to the receiver.[1] The receiver has then to answer a question related to the data that the sender was exposed to. Both agents share the common goal of the sender answering correctly to the question.

In this proposal, we focus on the work of [Laz+18]. In their setup, the sender is exposed to a picture drawn from a data set and generates a sequence of atomic symbols. After reading this sequence, the receiver sees different images from the data set, among which is the original one (the others are called "distractors"), and produces a distribution of probability from which one of the images is selected. Both agents are trained to maximise the probability of the correct image (the one shown to the sender) to be selected.[2] To solve this game, the sender must learn to produce a message describing the image it has been exposed to, distinguishing it from all other examples of the data set, and the receiver must learn to correctly interpret this message.

[1] This message might be atomic, composed of a sequence of symbol, or involve continuous types of information.

[2] The learning process is based on the REINFORCE algorithm [Wil92]. Other kinds of methods, however, are available to handle the kind of non-differentiablity involved by the transmission of symbolic messages, such as the continuous relaxation used by [HT17].

2 Difficulties of current methods

While experiments of this kind do succeed in making efficient communication emerge, it has been observed (in particular by [BB18]) that the messages generated tend to describe low-level (pixel) features of the images. In contrast, humans would describe the same scenes by referring to the involved objects, their shapes, colours and relative positions. For example, it is quite straightforward for a neural network to compute the average intensity of an image and such information (even discretised, if using small enough bins) is sufficient to identify the image with very high accuracy. As a consequence, these emergent languages have a very simple structure and have little in common with natural languages.

For a complex language to emerge, the communication between agents must arguably encode complex, structured information.[3] Our intuition is that the problem with most current experimental protocols is that the discrimination task assigned to the receiver is too easy to encourage such communication. A common way to increase the difficulty of the discrimination task is to increase the number of distractors. Doing so has experimentally proven unsatisfying. As natural images form only a tiny subset of all possible images, low-level pixel features will indeed intuitively suffice to discriminate between any two natural images, even when they appear indistinguishable to the human eye. It is likely that such basic methods for increasing the difficulty of the discrimination task merely shape the sender as a more and more complex hash function, with low collision probability but encoding arbitrary information.

3 Adversarial data sets

To generate the pressure to transmit structured high-level information, we propose instead to supplement the natural data set with images specifically designed to make low-level features useless for the discrimination task. A first approach is to corrupt the images before showing them to the sender and/or the receiver, e.g., by adding random noise to each pixel — as a picture of an apple on a table with random noise is generally still interpreted as a picture of an apple on a table, even if all pixels differ slightly. However, we hypothesise that such ad hoc data corruption processes will not be in general effective, as there is a practically infinite number of simple combinations of low-level features that could be left invariant (and thus still efficient for the discrimination task).

A second approach is to develop an *adversarial data set*. The idea of an adversarial data set is to adapt automatically to the agents in order to counteract any trivial strategy from their part. We propose to implement this adversarial data set using a third agent, the *adversary*. The adversary, after reading the description produced by the reader, generates an image that is shown to the

[3] While a language is also shaped by constraints on the length of messages, the cognitive power they require to be produced and interpreted, as well as they resistance to noise, we believe that the primary issue for emergent language games is semantics.

receiver. The adversary is trained to maximise the probability that, based on the description, the receiver selects the newly generated image rather than the ground truth one. The setup is represented in Figure 1. We expect that most low-level features are easy for the adversary to simulate, leading to a pressure on the sender to encode more and more complex information in its messages.[4] Finally, in order to both ease the training process and increase the interpretability of the emergent languages, we consider additionally (pre-)training the sender to produce and/or the receiver to understand natural descriptions using caption data sets [Vin+14].

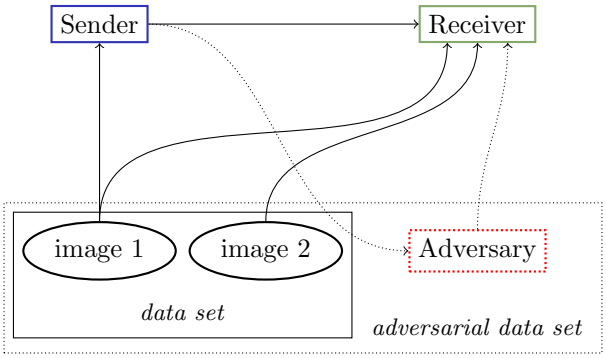

Figure 1: The *basic setup* is composed of a sender, which produces a description for any given image, and a receiver, which predicts whether any given image corresponds to any given description. The common goal of both agents is that the receiver be able to identify any image in the data set based on its description by the sender. The *proposed setup* extends the data set in an adversarial fashion. Another agent, the adversary, is introduced. Based on a description produced by the sender, the goal of the adversary is to fool the receiver, that is to say, to produce an image that the receiver also judges compatible with the description.

References

Bouchacourt, Diane and Marco Baroni. "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*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 981–985. DOI: 10.18653/v1/D18-1119. URL: https://www.aclweb.org/anthology/D18-1119.

[4] As some relatively low-level features, such as the sharpness of the pictures, are potentially difficult to simulate, we also include natural distractors.

Havrylov, Serhii and Ivan Titov. "Emergence of Language with Multi-Agent Games: Learning to Communicate with Sequence of Symbols". English. In: *5th International Conference on Learning Representations (ICLR 17, workshop track)* (Mar. 2017). URL: https://openreview.net/forum?id=SkaxnKEYg (visited on 04/25/2018).

Lazaridou, Angeliki et al. "Emergence of Linguistic Communication from Referential Games with Symbolic and Pixel Input". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. 2018. URL: https://openreview.net/forum?id=HJGv1Z-AW.

Vinyals, Oriol et al. "Show and Tell: A Neural Image Caption Generator". In: *CoRR* abs/1411.4555 (2014). arXiv: 1411.4555. URL: http://arxiv.org/abs/1411.4555.

Williams, Ronald J. "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning". In: *Mach. Learn.* 8.3-4 (May 1992), pp. 229–256. ISSN: 0885-6125. DOI: 10.1007/BF00992696. URL: https://doi.org/10.1007/BF00992696.