Visual communication in heterogeneous populations of artificial communicating agents

Daniela Mihai*[1]

*Corresponding Author: adm1c22@soton.ac.uk
[1]Vision, Learning and Control Group, University of Southampton, Southampton. UK

Innovations in artificial neural networks, deep and reinforcement learning techniques have led to research on communication protocols emergent from multi-agent interactions in the form of gameplay (Chaabouni et al., 2020; Guo, 2019; Lazaridou et al., 2018; Havrylov & Titov, 2017; Nölle et al., 2018; De Boer & Zuidema, 2010). Referential communication games (Lewis, 1969) have often been used as a simple, yet effective task that allows agents to develop their language through cooperation. These communication protocols evolved by artificial agents can be of various types, discrete or continuous, symbolic as in token-based (Havrylov & Titov, 2017) or iconic (Mihai & Hare, 2021). Recent studies looked into the effect of population size (Chaabouni et al., 2022) and heterogeneity (Rita et al., 2022; Mahaut et al., 2023) on token-based communication protocols.

In this work, we explore how population heterogeneity impacts a visual communication protocol of agents interacting through drawing. Using the drawing game environment of Mihai and Hare (2021), we study populations of agents with different pretrained visual encoding networks. Different model architectures are used as a proxy for agent heterogeneity to account for different visual experiences during the lifetime (see the supplementary for further discussion). The communication is through sketches made of 20 black lines. We explore how the graphical protocol changes under social pressures as agents are part of a community and need to adapt to more communicating partners. Previous works showed that pairs of agents playing a referential game tend to develop symbolic representations of the world specific to their interaction partner (Hawkins et al., 2023). Fay et al. (2014) also suggest that iconic representations become symbolic through repetitive interaction. Conversely, if the graphical communication evolves as the population stochastically participates in the games, then all participants will shape the final protocol. We hypothesise that the drawings of agents from population-based training will be more generalised across the population, and hence more iconic and less abstract than those emerging from homogeneous pair-wise interactions.

The setup involves a sender communicating through drawing about an image from STL-10 dataset (Coates et al., 2011). The game's goal is, based on the sketch

that the sender produces, for the receiver to correctly distinguish the target image from 50 candidates. The populations we test are of 4 and 6 agents (i.e. 2, or 3 respectively, senders and receivers) to sample from at each time step. It is worth noting that what we refer to as a sender agent can only produce drawings, and a receiver only interprets. Depending on the population, an agent can be instantiated with one of the 3 visual feature extraction modules: VGG16 (Simonyan & Zisserman, 2015), ResNet18 (He et al., 2016) and Vision Transformer (ViT) (Radford et al., 2021). The features are extracted after the last convolution layer and are passed through an additional batch normalisation layer before being fed into the sketch drawing module. The population model architecture, detailing each agent's learnable and pretrained modules can be found in the supplementary.

Preliminary experiments show that some populations are more successful than others. For example, the population of agents with VGG and ViT visual systems overall better solve the task. Due to the stochastic nature of the training, the communication rate can vary considerably throughout training as different agents interact and establish the convention at each step. In Fig. 1 Right we report average population accuracy, i.e. the task success averaged across all possible agent pairs in the population tested on the same evaluation set. In the supplementary material, we compare the communication success of heterogeneous populations with that of homogeneous pairs and observe the former are more difficult to train. Although additional training steps can sometimes improve average task success over the population, the iconicity of sketches does not significantly change.

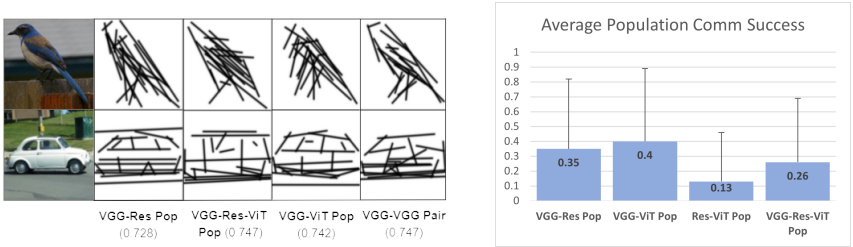

Figure 1. Left: Sketches produced by a VGG-sender agent trained in different population/pair configurations with *perceptual similarity* score as a measure of iconicity - lower means more iconic. Right: Average communication success of different populations on the test set after 250 epoch-training. The line represents the standard deviation from the mean accuracy across pairs in the population.

Fig. 1 Left compares sketches and reports perceptual similarity, as defined by Zhang et al. (2018), between the target image and sketches produced by VGG-senders trained in different configurations. This measure computed across deep features of a pretrained network is used as a proxy for iconicity - the more similar the representation (sketch) is to the real object (image), the more iconic (Peirce, 1867). These preliminary results point towards the hypothesis that sketches produced by population agents are more *iconic* than those of homogeneous agents.

References

Chaabouni, R., Kharitonov, E., Bouchacourt, D., Dupoux, E., & Baroni, M. (2020). Compositionality and generalization in emergent languages. *arXiv preprint arXiv:2004.09124*.

Chaabouni, R., Strub, F., Altché, F., Tarassov, E., Tallec, C., Davoodi, E., Mathewson, K. W., Tieleman, O., Lazaridou, A., & Piot, B. (2022). Emergent communication at scale. In *International conference on learning representations.*

Coates, A., Ng, A., & Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (pp. 215–223).

De Boer, B., & Zuidema, W. (2010). Multi-agent simulations of the evolution of combinatorial phonology. *Adaptive Behavior*, *18*(2), 141–154.

Fay, N., Ellison, M., & Garrod, S. (2014). Iconicity: From sign to system in human communication and language. *Pragmatics & Cognition*, *22*(2), 244–263.

Guo, S. (2019). Emergence of numeric concepts in multi-agent autonomous communication. *arXiv preprint arXiv:1911.01098*.

Havrylov, S., & Titov, I. (2017). Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 2149–2159). Curran Associates, Inc.

Hawkins, R. D., Sano, M., Goodman, N. D., & Fan, J. E. (2023). Visual resemblance and interaction history jointly constrain pictorial meaning. *Nature Communications*, *14*(1), 1–13.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Lazaridou, A., Hermann, K. M., Tuyls, K., & Clark, S. (2018). Emergence of linguistic communication from referential games with symbolic and pixel input. In *International conference on learning representations.*

Lewis, D. K. (1969). *Convention: A philosophical study.* Wiley-Blackwell.

Mahaut, M., Franzon, F., Dessì, R., & Baroni, M. (2023). Referential communication in heterogeneous communities of pre-trained visual deep networks. *arXiv preprint arXiv:2302.08913*.

Mihai, D., & Hare, J. (2021). Learning to draw: Emergent communication through sketching. In *Advances in neural information processing systems.*

Nölle, J., Staib, M., Fusaroli, R., & Tylén, K. (2018). Environmental and social factors motivate the emergence of systematic categories and signs. In C. Cuskley, M. Flaherty, H. Little, L. McCrohon, A. Ravignani, & T. Ver-

hoef (Eds.), *The evolution of language: Proceedings of the 12th international conference (evolangxii)*. NCU Press.

Peirce, C. S. (1867). Five hundred and eighty-second meeting. may 14, 1867. monthly meeting; on a new list of categories. *Proceedings of the American Academy of Arts and Sciences*, *7*, 287–298.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Rita, M., Strub, F., Grill, J.-B., Pietquin, O., & Dupoux, E. (2022). On the role of population heterogeneity in emergent communication. In *The international conference on learning representations (iclr) 2022*.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International conference on learning representations*.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *CoRR*, *abs/1801.03924*.