Supplementary material:
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

1. Game setup

This work explores the effect of population heterogeneity on the emergence of a visual communication protocol. The agents are trained on a referential communication game in which a target image has to be referenced and differentiated among a set of possible candidates (Havrylov & Titov, 2017; Mihai & Hare, 2021). For this study, a set of 50 candidates, 1 target and 49 distractors, is used. The drawing game setup was introduced in Mihai and Hare (2021). The communication is done through a black-and-white, 20-line sketch. The sender sees an image and needs to communicate through drawing such that the receiver correctly identifies the target from the set of candidates. The sender does not know what other images the receiver can choose from. Therefore, the sender needs to learn to extract and communicate meaningful information. For this study, we are interested in whether the information expressed through the visual means of drawing is more of an icon than a symbol when the communication is shaped by a population instead of a (homogeneous) pair.

2. Model architecture

The populations studied in this work are made out of 4, i.e. 2 senders and 2 receivers, or 6 agents, i.e. 3 senders and 3 receivers. At each time step of training, a pair is sampled from the population, one sender and one receiver to play the game.

The model architecture of a sender-receiver pair for agents that communicate through drawing is defined in Mihai and Hare (2021). As shown in Fig. 1, in the heterogeneous setup we consider, each agent, sender and receiver, has its separate visual encoder, pretrained and frozen, but they may be different. This is unlike Mihai and Hare (2021), where the visual system was shared between the two agents.

Each agent has some specialised components based on the activity it performs. The sender agent, on top of a visual encoder, has a drawing module to create the sketch. This drawing module maps the feature vectors to line coordinate encodings which are then rasterised into the actual sketch. The Rasteriser in Fig. 1 does not have learnable parameters but allows backpropagation of the error signal through to the learnable components. The Drawing module is unique for each sender in the population, it is not shared.

The receiver agent needs to process the sketch and the possible target images using its visual system (also pre-trained and frozen) and compute scores that in combination with a multi-class hinge loss would give an error which can be back-propagated through the whole system (trainable parts) to update the agents' learnable modules.

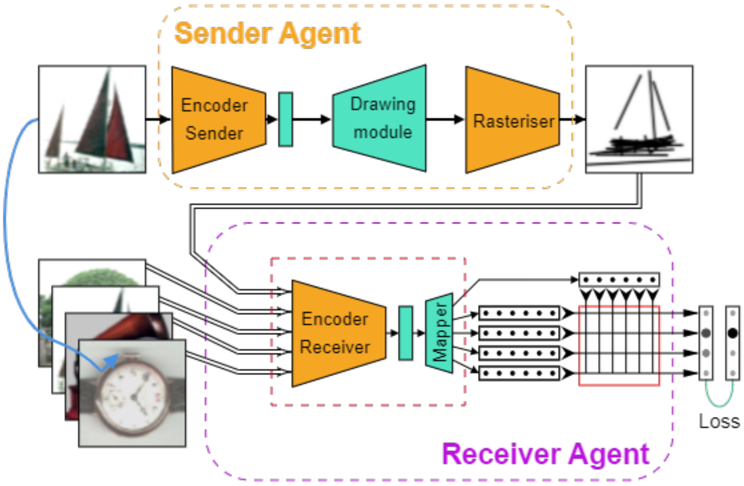

Figure 1. A sender-receiver pair as defined in Mihai and Hare (2021). Orange components are frozen, while green ones are trained.

3. Population heterogeneity

This paper addresses the topic of population heterogeneity with deep-network-modelled agents. The heterogeneity is introduced by defining agents that have different visual systems (pretrained and fixed visual encoders). The choice to model heterogeneity by defining agents with distinct architectures of pretrained and fixed image encoding networks, distinguished also by pretraining data and objectives, was motivated by the belief that this would, to some extent, approximate the different visual experiences humans have in the world. As is the case in a population of human agents, each has a unique cognitive architecture and different experiences during one's lifetime. There are no two humans who have seen and visually experienced the same things.

This work models populations of up to 6 agents, which would be up to 3 senders and 3 receivers. These three sender agents and three receiver agents are differentiated by three distinct visual systems: VGG16 (Simonyan & Zisserman, 2015) trained on Stylized-ImageNet (Geirhos et al., 2019) that introduces a shape bias, ResNet18 (He et al., 2016) trained on ImageNet and Vision Transformer (ViT) (Radford et al., 2021) trained on a set of (image, text) pairs. Each sender and each receiver has on top of the visual system a task-specific module. It is worth noting that in this work, an agent is defined such that it can only either communicate or comprehend (i.e. interpret the communication), but not both. A sender agent can *only communicate* in the form of sketches, and a receiver agent can *only interpret* sketches and target candidates and decide which are the correct targets. In traditional agent-based modelling, however, agents may both speak and comprehend.

The topic of population heterogeneity needs a more granular definition in the field of multi-agent learning. While the architecture of the visual encoding system (Mahaut et al., 2023) and the learning rate (Rita et al., 2022) can be used as proxies for speaker heterogeneity, so does the data used for pertaining. Visual features can be experienced differently in distinct parts of the world; for example, having a wider variety of words for a specific colour due to its more frequent presence in the environment. Likewise, the amount of pretraining can be a factor of differentiation between agents in a population. This could resemble the scenario of adults having a longer and more complex visual experience of the world than children. Other specific cognitive conditions like colour blindness, which reflects a differently limited experience of the visual world, can be considered too when modelling agent heterogeneity. Future work should consider a systematic and extensive analysis of the ways heterogeneity can be defined.

4. Results - comparison between heterogeneous and homogeneous agents

While the main paper presents the most important part of the results, this section includes a wider range of agent populations/pairs and reports the test task success, along with a comparison of the sketches based on the iconicity measure approximated through the perceptual similarity metric of Zhang et al. (2018). It is worth noting that the reported results are from playing the game on a separate evaluation set, consisting of images unseen during training.

4.1. *Heterogeneous populations*

For populations, test communication success averaged across all possible agent pairings is reported. For example, the VGG-Res-ViT Population in Table 1 means there are 6 agents: 3 senders and 3 receivers, one with each of the 3 previously mentioned pretrained visual systems. This population has a total of 9 possible pairs to be sampled from at each training step.

Table 1. Average population communication success. The communication is averaged across all possible agent pairs in a population, the value in brackets represents the standard deviation from the mean across pairs in the population evaluated on the same test set.

	VGG-Res-ViT Pop	VGG-ViT Pop	VGG-Res Pop	Res-ViT Pop
250 epoch training	$0.26(\pm 0.43)$	$\mathbf{0.4}(\pm\mathbf{0.49})$	$0.35(\pm 0.47)$	$0.13(\pm 0.33)$
1000 epoch training	$0.28(\pm 0.45)$	$\mathbf{0.43}(\pm\mathbf{0.49})$	$0.42(\pm 0.49)$	$0.16(\pm 0.37)$

As seen in Table 1, overall population success can improve with more training steps. However, when analysing the communication success of individual pairs in the population it is clear some are more successful than others. For example in the case of the 6-agent population, the average communication rate is lowered due to the performance of Res-ViT and ViT-Res pairs which achieve less about 10% success. This is also illustrated in the Res-ViT population setup which has the lowest score across all tested populations.

Lastly, it is worth noticing that while in the case of discrete or continuous token-based communication (Chaabouni et al., 2022; Rita et al., 2022; Mahaut et al., 2023), population training might not negatively impact task accuracy, this is not the case for a visual channel (i.e. sketch) in which population communication success significantly drops compared to homogeneous pairs (i.e. agents having the same visual system). This could be due to the images being encoded to representations that vary considerably across model architectures, pretraining regimes, and pretraining data. Consequently, image encodings used by the sender to produce sketches might not be compatible with the receiver's visual system. For example, sketches produced by a VGG sender agent might not contain relevant information for a ResNet receiver.

4.2. *Heterogeneous pairs*

The following results illustrate the scenario of heterogeneous pairs, which means that the sender and receiver have distinct visual systems and the pair is fixed throughout training. Each pair is trained separately for 250 epochs, unlike the population setup in which, at each time step, a pair is sampled to play the game and hence will result in a lower number of training interactions per pair.

These results, once more, confirm that certain agent pairings work better than others. For example, pairing agents with VGG and ResNet visual encoders work better than other combinations (see VGG-Res Pair and Res-VGG Pair in Table 2, compared to ViT-Res Pair for example).

Table 2. Test communication success of heterogeneous pairs; A-B Pair notation represents a pair in which the sender has visual encoder A and receiver B. Each pair was trained separately for 250 epochs. In brackets, the standard deviation across games is reported.

VGG-Res Pair	VGG-ViT Pair	ViT-VGG Pair	ViT-Res Pair	Res-VGG Pair	Res-ViT Pair
$0.43(\pm 0.49)$	$0.26(\pm 0.44)$	$0.35(\pm 0.47)$	$0.13(\pm 0.33)$	$0.4(\pm 0.49)$	$0.1(\pm 0.3)$

Table 3. Test communication success of homogeneous pairs, i.e. sender and receiver that have the same pretrained and fixed visual system. In brackets, the standard deviation across games is reported.

VGG-VGG Pair	ViT-ViT Pair	Res-Res Pair
$0.96(\pm 0.19)$	$0.75(\pm 0.43)$	$0.1(\pm 0.3)$

4.3. *Homogeneous pairs*

As shown in Table 3, homogeneous pairs, in which both sender and receiver have the same visual system, reach higher task accuracy compared to the same sender-receiver pairs trained in a population setting (see Table 1) or heterogeneous pairs (see Table 2). The only exception is the Res-Res pair, which represents the pair using the same pretrained and frozen ResNet18 visual encoder for both sender and receiver. This pair performs worse than some heterogeneous pairs in which one of the agents has ResNet as encoder, for example, Res-VGG Pair in Table 2. Further work would be needed to determine the cause of this behaviour.

4.4. *The iconicity of graphical protocols*

Lastly, we provide an iconicity measure of the visual communication protocol and sample sketches produced by sender agents trained in several settings: homogeneous pairs and different heterogeneous populations. In this study, the iconicity of the graphical protocol is measured based on the sketch's likeness to the target image. The perceptual similarity is computed using the method of Zhang et al. (2018), based on deep neural network features (from pretrained AlexNet), it has been shown to approximate human perceptual similarity judgements. The value under each sample sketch represents the mean perceptual similarity averaged across sketches produced from *test* images by that particular sender. A lower score means the sketch produced by the specific sender trained in that configuration is on average more similar to the target, and hence more iconic.

In the case of a VGG-based sender, the most iconic sketches (lowest perceptual similarity score) are produced when the agent is trained in a population of agents with VGG16 and ResNet18 visual systems (see Fig. 2). So is the case for a ResNet-based sender (see Fig. 3). For a sender with a Visual Transformer encoder, however, the most iconic sketches emerge in the 6-agent population setting (see Fig. 4).

It is worth mentioning that the method of measuring iconicity employed in this study is only one of the many, and that other methods such as human evaluation would provide a more reliable comparison of what humans perceive as iconic, but it would also be significantly more time-consuming.

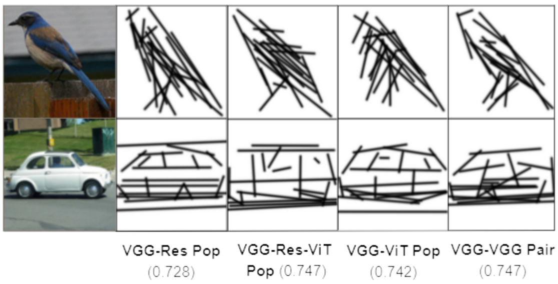

Figure 2. Iconicity, measured through perceptual similarity between test images and sketches produced by a VGG-sender trained in different population/pair configurations.

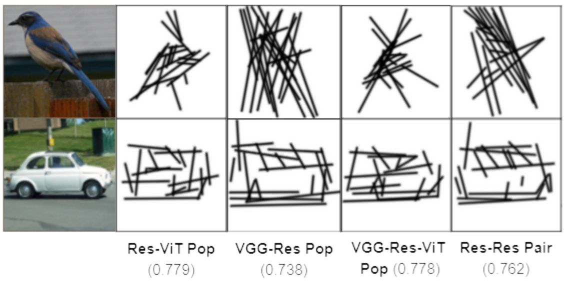

Figure 3. Iconicity, measured through perceptual similarity between test images and sketches produced by a ResNet-sender trained in different population/pair configurations.

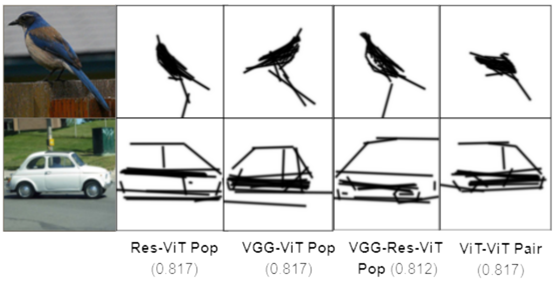

Figure 4. Iconicity, measured through perceptual similarity between test images and sketches produced by a ViT-sender trained in different population/pair configurations.

References

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.*

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations.*

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.

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).

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.*

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.*