Referential Communication in Heterogeneous Communities of Pre-trained Visual Deep Networks

Extended Abstract

Matéo Mahaut
Universitat Pompeu Fabra
Barcelona, Spain
mateo.mahaut@upf.edu

Francesca Franzon
Universitat Pompeu Fabra
Barcelona, Spain
francesca.franzon@upf.edu

Roberto Dessì
Universitat Pompeu Fabra and Meta AI
Barcelona, Spain
rdessi@meta.com

Marco Baroni
Universitat Pompeu Fabra and ICREA
Barcelona, Spain
marco.baroni@upf.edu

ABSTRACT

As large pre-trained image-processing neural networks are being embedded in autonomous agents such as self-driving cars or robots, the question arises of how such systems can communicate with each other about the surrounding world, despite their different architectures and training regimes. As a first step in this direction, we explore the task of *referential communication* in a community of state-of-the-art pre-trained visual networks, showing that they can develop a shared protocol to refer to a target image among a set of candidates. Such shared protocol, induced in a self-supervised way, can to some extent be used to communicate about previously unseen object categories. Finally, we show that a new neural network can learn the shared protocol developed in a community with remarkable ease, and the process of integrating a new agent into a community more stably succeeds when the original community includes a larger set of heterogeneous networks.

KEYWORDS

Emergent deep net communication; deep visual nets; multi-agent communication

ACM Reference Format:
Matéo Mahaut, Francesca Franzon, Roberto Dessì, and Marco Baroni. 2023. Referential Communication in Heterogeneous Communities of Pre-trained Visual Deep Networks: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

With deep neural networks being deployed industrially in a range of disciplines, many concurrent architectures will soon be functioning in the same spaces. We come to wonder how they could communicate about the surrounding visual world, as seen through the lenses of their respective core visual components. A new line of research has recently emerged, focusing on methods to let deep networks develop a shared communication protocol [7]. Specifically, the area of *deep net emergent communication* [3, 4] has scaled this line of research to larger networks and datasets.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

Table 1: Pre-trained visual architectures employed

Architecture	Type	Training	Parameters
ResNet152 [5]	CNN	Supervised	60.2M
Inception [14]	CNN	Supervised	27.2M
VGG 11 [12]	CNN	Supervised	132.9M
ViT-B/16 [6]	Attention	Supervised	86.6M
ViT-S/16 [2]	Attention	Self-supervised	21M
Swin [8]	Attention	Supervised	87.7M

2 SETUP

Inspired by the core communicative task of *reference* (e.g. [13]), agents play the referential communication game. A *sender* is given a target input (e.g., a picture) and issues a message (e.g., a small vector). A *receiver* is exposed to a set of inputs, including the target, and must correctly point to the latter based on the message it receives from the sender. In our setup both sender and receiver are neural networks as described in Fig. 1. We use as vision modules widely used state of the art vision models (Table 1). Note that no parameters are shared between sender and receiver, except those of the frozen visual modules in the case in which the two agents are using homogeneous visual architectures. Only the communication and mapper modules are trained. All networks were pre-trained on ILSVRC2012 ImageNet data [11]. We explore communication across architectures, evaluating referential ability, generalization capacity, and the learnability of communication strategies.

Training: We train the communication and mapper modules using the Imagenet 1k validation set, which has not been seen by vision modules during pre-training. We reserve 10% of the validation set for testing. Agents are also tested on an out-of-domain (*OOD*) dataset containing classes from the larger ImageNet-21k repository, as pre-processed by [9]. We selected 52 new classes among those that are neither hypernyms nor hyponyms of imagenet1k classes, and we made sure that no OOD class had a WordNet path similarity score [1] above 0.125 with any imagenet1k class.[1]

[1]We provide scripts to reproduce our imagenet1k and OOD datasets at https://github.com/mahautm/emecom_pop_data

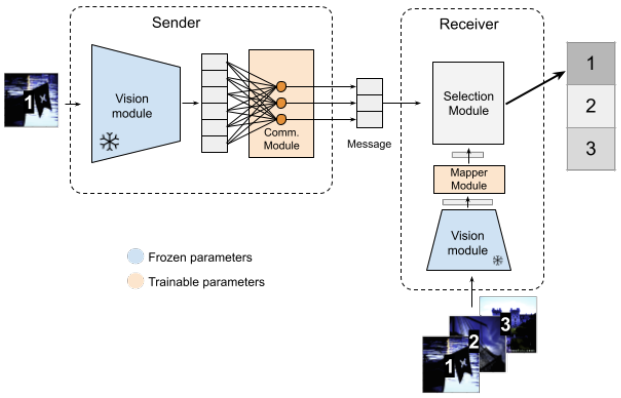

Figure 1: A target image is input to the sender (left), that extracts a vector representation of it by passing it through a pre-trained frozen visual network This vector representation is fed to the single layer feed-forward *Communication Module* that generates a message (a 16 dimension continuous vector). This message vector is input to the Receiver (right). The receiver processes each of 64 candidate images in turn by passing it through a pre-trained frozen visual network (which is either the same (homogeneous) or another (heterogeneous) architecture), obtaining a set of vector representations. These are fed to a *Mapper Module*, another two layered fully connected feed-forward component that maps them to vectors in the same space as the sender message embedding. The *Selection Module* of the receiver simply consists in a parameter-free cosine similarity computation between the message and each image representation, followed by Softmax normalization. The receiver is said to have correctly identified the target if the largest value in the resulting probability distribution corresponds to the index of the target in the candidate array.

Table 2: Percentage accuracy of agents playing the referential communication game on both datasets

	Imagenet-1k	OOD
Homogeneous	100 ± 0	92 ± 5
Heterogeneous	97 ± 2	61 ± 16
Population	98 ± 1	66 ± 15

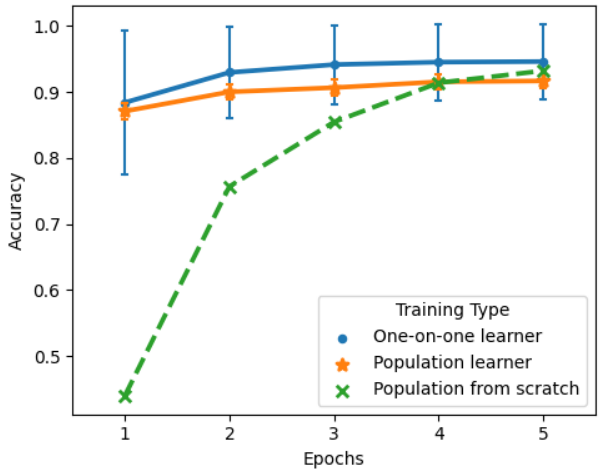

Figure 2: Test accuracy and learning speed of learner agents. Blue line: learning curve on test data for learner agent added to a communicating pair, averaged across all possible heterogeneous triples. Orange line: learning curve for learner agent added to an existing community, averaged across all possible leave-one-out cases. Vertical bars indicate standard deviation across cases. As a baseline for learning speed, the dashed green line shows the learning curve when training the whole 6x6 populations at once from scratch.

3 RESULTS

If different networks were pre-trained independently can they still agree on reference? With a very small communication channel (16 dimensions), models rapidly converge to near perfect accuracy on the referential task (first column of Table 2). This remains true for all 6 tested models, in both homogeneous (both sender use the same vision module, e.g. a ResNet) and heterogeneous (sender and receiver have different vision modules, e.g. a ResNet sender and VGG receiver) cases, across architecture types, and training types. OOD performance, while not being as good as in-domain, confirms the notable generalization capabilities of the emerged language. It should be noted that all results using continuous communication widely outperform those obtained with discrete communication (not reported here for space reasons), in training speed (x9), accuracy (+30%), and generalisation capabilities (+30%).

Population: Like [3, 10] we study the impact of training agents in a population setup. At every trial, two agents are randomly sampled to play the referential game together, so that throughout training any of the 6 senders will be paired with each of the 6 receivers. Table 2 shows that performance is very close to that obtained when training agent in fixed pairs (one-on-one). It is therefore possible to make agents use messages that are understood by all architectures, with little to no loss in performance. We leave extensive message analysis to future works.

Learning a communication strategy: We finally imagine a scenario where a group of agents has already been deployed, and there is a need to add a new agent to the group: can it learn to communicate with its peers ? Once a group of agents has converged upon a communication strategy that reaches high accuracy on the referential task, we investigate how easily a new untrained agent can learn to use it. New agents perform very well very fast (Fig. 2), with accuracy reaching more than 85% by the first epoch. We show that communication induced once by training, can be learnt, to a lesser cost than if it had to be re-developed from scratch.

ACKNOWLEDGMENTS

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101019291). This

paper reflects the authors' view only, and the funding agencies are not responsible for any use that may be made of the information it contains.

REFERENCES

[1] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of ICCV*. Online, 9650–9660.

[3] Rahma Chaabouni, Florian Strub, Florent Altché, Eugene Tarassov, Corentin Tallec, Elnaz Davoodi, Kory Wallace Mathewson, Olivier Tieleman, Angeliki Lazaridou, and Bilal Piot. 2022. Emergent communication at scale. In *Proceedings of ICLR*. Online. Published online: https://openreview.net/group?id=ICLR.cc/2022/Conference.

[4] Roberto Dessì, Eleonora Gualdoni, Francesca Franzon, Gemma Boleda, and Marco Baroni. 2022. Communication breakdown: On the low mutual intelligibility between human and neural captioning. In *Proceedings of EMNLP*. Abu Dhabi, United Arab Emirates. In press.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of CVPR*. Las Vegas, NV, 770–778.

[6] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of ICLR*.

[7] Angeliki Lazaridou and Marco Baroni. 2020. Emergent multi-agent communication in the deep learning era. https://arxiv.org/abs/2006.02419.

[8] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.

[9] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik. 2021. ImageNet-21K pretraining for the masses. In *Proceedings of NeurIPS Datasets and Benchmarks Track*. Online. Published online: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021.

[10] Mathieu Rita, Florian Strub, Jean-Bastien Grill, Olivier Pietquin, and Emmanuel Dupoux. 2022. On the role of population heterogeneity in emergent communication. In *Proceedings of ICLR*. Online. Published online: https://openreview.net/group?id=ICLR.cc/2022/Conference.

[11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

[12] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of ICLR Conference Track*. San Diego, CA. Published online: http://www.iclr.cc/doku.php?id=iclr2015:main.

[13] Brian Skyrms. 2010. *Signals: Evolution, learning, and information.* Oxford University Press, Oxford, UK.

[14] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2014. Going deeper with convolutions. http://arxiv.org/abs/1409.4842.