Internship project, master level, 2022

Title: Self-organization of shared graphical languages in groups of agents using multimodal contrastive deep learning mechanisms

Supervision: Pierre-Yves Oudeyer (Inria and Microsoft Research Montreal), Romain Laroche (Microsoft Research Montreal) and Clément Moulin-Frier (Inria)

Host team: Flowers team, Inria Bordeaux, https://flowers.inria.fr/ (and involves a collaboration with MSR Montreal (possibilities to visit Montreal can be discussed)

Duration: 6 months, around march - august 2022

How to apply: contact t-poudeyer@microsoft.com , romain.laroche@microsoft.com and clement.moulin-frier@inria.fr (send one email to all) with a CV and letter of motivation

Keywords: emergence of language; multi-agent multimodal learning; deep learning; Transformers; contrastive learning

Context: Computational models of the formation of communication systems in groups of agents have enabled to improve our understanding of the evolution of human language (Cangelosi et al., 2012; Kirby et al., 2014). For example, the naming game model (Steels and Loetzsch, 2012) showed how a population of agents can self-organized a culturally shared lexicon without centralized coordination (see the Talking Heads experiments, Steels, 2015 https://www.youtube.com/watch?v=n6876onk7sI, or this more recent work from OpenAI: https://openai.com/blog/learning-to-communicate/). They have also provided new perspectives in AI for building machines capable of fluid and adaptive language communication with humans (Lazaridou et al., 2020) and among each other, e.g. using mechanisms for joint attention (Kaplan and Hafner, 2006), grounding or perspective taking (Cangelosi et al., 2010). Recently, a new wave of models has been leveraging deep learning methods, and in particular multi-agent deep reinforcement learning (Foerster et al., 2016; Sukhbaatar et al., 2016; Lazaridou et al., 2018; Mordatch et al., 2017; see Lazaridou and Baroni, 2020 for a review, and Moulin-Frier and Oudeyer, 2020 for open questions). This has enabled to scale up language game models to environments where linguistic conventions are jointly learned with visual representations of raw image perception, as well as to environments where emergent communication is used as a tool to achieve joint cooperative tasks.

However, there are still several important open questions raised by these deep learning-based models. First, while such models make few assumptions on conceptual representations (which are jointly learned from raw perception), they are still far from enabling convergence to efficient shared communication systems in ecological settings, i.e. through decentralized learning. Second, such models have considered so far only idealized symbolic communication channels: yet, it remains an open question to understand how communication could emerge in agents producing signs with a constrained sensorimotor system. This could be a vocal production system or a gestural/writing system enabling to produce continuous graphical signs (e.g. related to the sign language self-organized by deaf Nicaraguan children). In this context, there is one very important open question: under what conditions such systems could self-organize a discrete combinatorial system of language signs?

Finally, from a technical perspective, such deep learning-based models have not so far leveraged recent advances in approaches to vision and language learning based on Transformers (Vaswani et al., 2017), especially multimodal and contrastive approaches such as the CLIP architecture (Radford et al., 2021). It would be highly interesting to evaluate the use of such techniques in the context of multi-agent self-organization of language systems.

Project:

This internship project aims at addressing one or several of these limits, based on a first stage of literature review.

A first objective could be to take inspiration from contrastive learning dynamics of early simple naming games models, that enabled robust convergence to efficient shared communication systems, the project will study the design of a new model leveraging contrastive deep learning techniques. Several possibilities will be considered, ranging from energy based models to multimodal transformers using a contrastive loss function, such as the CLIP model. A first round of experimentation will study how such models could enable more robust communication to shared communication systems, while letting the system learn its own representations of visual concepts (from images).

A second possible perspective for this project (possibly using the contrastive models mentioned above) will be to consider agents that use a constrained sensorimotor system to produce signs. In particular, we will consider the use of a drawing sensorimotor ability, enabling agents to draw shapes and use them as signs in language games (such a modality could facilitate analysis and interpretation). One possibility would be for example to consider the production of movements of a simulated hand using dynamic motion primitives (adapted from the robotics litterature Schaal, 2006). We may consider either the possibility for other agents to perceive directly the motor parameters used to produce the drawing, or to perceive only the final image of the drawing. The same analysis could be made on the production perspective: it would also be interesting to see whether it is more efficient (from the language emergence/learning perspective) for them to produce through drawing or to produce through image generation. Evaluation will potentially consider two dimensions. One will be the study of the convergence dynamics of such groups of agents with a sensorimotor system. Another will be the study of the structure of emergent graphical sign systems, and in particular a study of when and how they could become discrete and combinatorial (similarly to human speech or writing systems).

Candidates will have the possibility to propose their own directions and ideas of approaches.

Requirements: We are looking for motivated MSc students (Master II) with solid expertise in machine learning, especially deep learning algorithms and associated software tools (pyTorch, tensorFlow, etc). Solid skills in mathematics or physics, especially for studying the dynamics of complex dynamical systems, would be particularly welcome.

References

Cangelosi, A., & Parisi, D. (Eds.). (2012). *Simulating the evolution of language*. Springer Science & Business Media.

Cangelosi, A., Metta, G., Sagerer, G., Nolfi, S., Nehaniv, C., Fischer, K., ... & Zeschel, A. (2010). Integration of action and language knowledge: A roadmap for developmental robotics. IEEE Transactions on Autonomous Mental Development, 2(3), 167-195.

Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. 2021. Communicating artificial neural networks develop efficient color-naming systems. Proceedings of the National Academy of Sciences, 118.

de Boer, B., and Zuidema, W. 2010. Multi-Agent Simulations of the Evolution of Combinatorial Phonology. Adaptive Behavior 18(2):141–154.

Foerster, J.; Assael, Y. M.; de Freitas, N.; and Whiteson, S. 2016. Learning to communicate with deep multiagent reinforcement learning. In Advances in Neural Information Processing Systems, 2137–2145.

Kaplan, F., & Hafner, V. V. (2006). The challenges of joint attention. *Interaction Studies*, *7*(2), 135-169.

Kaplan, F. (2001). *La naissance d'une langue chez les robots*. Hermès Science Publications.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current opinion in neurobiology*, *28*, 108-114.

Lazaridou, A., Potapenko, A., & Tieleman, O. (2020). Multi-agent communication meets natural language: Synergies between functional and structural language learning. ACL 2020

Lazaridou, A., & Baroni, M. (2020). Emergent multi-agent communication in the deep learning era. arXiv preprint arXiv:2006.02419.

Lazaridou, A.; Hermann, K. M.; Tuyls, K.; and Clark, S. 2018. Emergence of Linguistic Communication from Referential Games with Symbolic and Pixel Input. In Sixth International Conference on Learning Representations (ICLR 2018).

Mordatch, I., and Abbeel, P. 2017. Emergence of Grounded Compositional Language in Multi-Agent Populations. In Thirty-Second AAAI Conference on Artificial Intelligence.

Moulin-Frier, C.; Diard, J.; Schwartz, J.- L. J.-L.; and Bessiere, P. 2015. COSMO ('Communicating about ` Objects using Sensory-Motor Operations'): a Bayesian modeling framework for studying speech communication and the emergence of phonological systems. Journal of Phonetics 53:5–41.

Moulin-Frier, C., & Oudeyer, P. Y. (2020). Multi-Agent Reinforcement Learning as a Computational Tool for Language Evolution Research: Historical Context and Future Challenges. *arXiv preprint arXiv:2002.08878*.

Oudeyer, P.-Y. 2006. Self-Organization in the Evolution of Speech, volume 6 of Studies in the Evolution of Language. Oxford University Press.

Portelance, E., Frank, M. C., Jurafsky, D., Sordoni, A., & Laroche, R. (2021). The Emergence of the Shape Bias Results from Communicative Efficiency. 5th Conference on Computational Natural Language Learning (CoNLL)

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020.

Schaal, S. (2006). Dynamic movement primitives-a framework for motor control in humans and humanoid robotics. In *Adaptive motion of animals and machines* (pp. 261-280). Springer, Tokyo.

Steels, L., & Loetzsch, M. (2012). The grounded naming game. *Experiments in cultural language evolution*, *3*, 41-59.

Steels, L., & Kaplan, F. (2000). Aibo’s first words: The social learning of language and meaning. Evolution of communication, 4(1), 3-32.

Steels, L. (2015). The Talking Heads experiment: Origins of words and meanings (Vol. 1). Language Science Press.

Sukhbaatar, S.; Szlam, A.; and Fergus, R. 2016. Learning Multiagent Communication with Backpropagation. In Proceedings of the 30th International Conference on Neural Information Processing Systems.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).