A Social Neuro-AI approach to Multi-Agent Reinforcement Learning

Zahra Sheikhbahaee
CHUSJ Research Center
Mila - Quebec AI Institute,
University of Montreal
zahra.sheikh@ppsp.team

Juan-David Vargas
CHUSJ Research Center
Mila - Quebec AI Institute,
University of Montreal
juan.vargas@ppsp.team

Samuele Bolotta
CHUSJ Research Center
University of Montreal
samuele.bolotta@ppsp.team

Adam Safron
Center for Psychedelic
& Consciousness Research
John Hopkins University
asafron1@jhmi.edu

Leonardo Christov-Moore & Nicco Reggente
Institute for Advanced Consciousness Studies
leo@advancedconsciousness.org
nicco@advancedconsciousness.org

Dianbo Liu
Mila - Quebec AI Institute,
University of Montreal
National University of Singapore
dianbo.liu@alumni.harvard.edu

Guillaume Dumas
CHUSJ Research Center
Mila - Quebec AI Institute,
University of Montreal
guillaume.dumas@ppsp.team

1 Introduction

Inspired by cognitive science, we integrate a causal modular architecture in a multi-agent reinforcement learning (MARL) featuring specific constraints (*i.e.* recurrent neural networks), representing knowledge systems that can learn environment dynamics effectively [1].

These modules, interacting sparsely through attention bottlenecks, allow for specialization and enable agents to dynamically adjust attention and resources [2]. We additionally represent the state and evolution of individual objects, based on a factorization of declarative (object properties) and procedural (object behaviour) knowledge, enhancing the system's ability to track and interact with multiple entities and systematically model complex environments [3]. Aligning with the view that intelligence measures an agent's ability to achieve goals in diverse contexts, our approach emphasizes generalization and adaptability in a wide range of social settings [4].

Attention schema theory proposes that internal control of attention states is necessary to predict others' behaviors [5, 6], facilitating social cognition and emergent consciousness [7, 8]. Applying this attention-focused approach may amplify the efficiency of the training process [9]. Inductive biases, manifesting similarly to attention [10, 11] may act as concealed forms of training data, compensating for the absence of robust initial assumptions, and are, in some cases, equivalent to having more training data [12].

Complex cognitive functions develop through social interaction. Neurodynamical systems in MARL could emulate sociocognitive learning processes within interaction, enhancing the adaptability and effectiveness of artificial agents in complex social environments [13, 14].

We attempt to address three questions: (1) How do attention mechanisms influence inter-agent communication and enhance cooperative strategies? (2) Can modularity within an agent's architecture improve adaptability and generalization in rapidly changing scenarios? (3) What implications do these architectural choices have for the development of prosocial behaviours in complex multi-agent settings?

Workshop on Melting Pot Competition at NeurIPS 2023.

We evaluate our proposed architecture in varied, unpredictable social settings within Melting Pot 2.0 [15], which provides a comprehensive evaluation protocol measuring agents' ability to generalize to novel social partners in diverse test scenarios with various social interdependencies and mixed incentives [16]. This approach aligns with our objective to investigate multi-agent dynamics in complex, socially-rich environments, emphasizing the importance of adaptability and generalization in the presence of unseen policies.

2 Methods

Our on-policy multi-agent algorithm draws inspiration from the pioneering work of Yu et al. 2022 who specifically tailored Proximal Policy Optimization (PPO) for multi-agent settings [17]. We leverage a modified version of shared parameter PPO combined with a centralized value function in a multi-agent context, termed MAPPO. Additionally, we employ decentralized, independent PPO (IPPO) to ensure our agents remain both collaborative and autonomous to the degrees required for skillful performance in various situations. In our benchmark environments, using MAPPO, our agents share a common global observation for the value network, and local information is fed to each agent's policy network. Previous studies have observed that using centralized value functions— due to the inclusion of the full global state —can facilitate value learning (e.g. by making credit assignment problems more tractable). Additionally, we incorporate Generalized Advantage Estimation (GAE) [18], advantage normalization, value-clipping, and value normalization, to enhance the stability of value learning.

The architectures of both actors and critics are specially designed for processing input images. This processing begins with convolutional layers, essential for extracting spatial feature hierarchies from the images. Following this, the extracted features undergo positional encoding using high-fidelity Fourier features. This step is instrumental in embedding a detailed representation of the position in the input units. This method ensures that each unit's activity is meaningfully correlated with a particular semantic or spatial context. The image encoders in both architectures embed the pixels of local (O_t^l) and global (O_t^g) observations into low dimensional features. Subsequently, the architecture channels output from this process into a Recurrent Independent Mechanism (RIM) [19] or SChema/Object-File Factorization (SCOFF) [20], both of which are cognitively inspired methods that disentangle different objects and operations to improve the models' learning efficiency and generalization (see details in Algorithm 1). We do not apply masking to these parts of our architecture.

3 Results

Our preliminary results are on two environments of Melting Pot 2.0: Harvest and Territory Rooms, associated respectively with cooperative and competitive social behaviors. The Figure 1 shows the evolution of reward for three models of increasing complexity, from the traditional baseline to architecture using attention. According to our hypothesis, the latter performs the best. Interestingly, the model with attention seems to perform even better in the cooperative environment (Harvest).

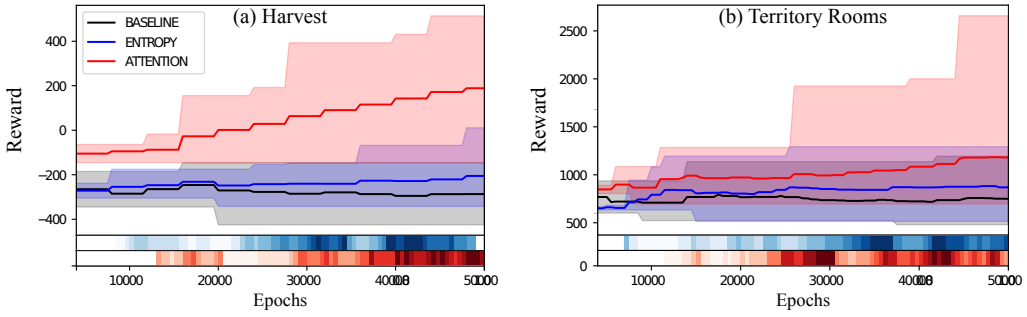

Figure 1: Evolution of reward during learning for three illustrative models and two substrates of Melting Pot 2.0. BASELINE (black) is the classic architecture. ENTROPY (blue) is like the BASELINE model but with a term in the loss to nudge exploration in agents. Finally, ATTENTION (red) is the model with attention mechanisms. Thick lines indicate the average rewards across agents. Shaded areas indicate the minimum and maximum rewards. Bottom blue and red stripes indicate the statistical differences with BASELINE reward for respectively ENTROPY and ATTENTION models.

4 Conclusion

In summarizing our work, the advanced architecture of our MARL model, incorporating attention mechanisms and recurrent neural networks, demonstrates a promising approach to designing social artificial agents. Due to these complexities and our present resource constraints, we were unable to exhaustively explore the full potential of our model within the competition deadline timeframe. However, moving forward, we intend to conduct a rigorous and comprehensive investigation of its capabilities, aiming to uncover and leverage its full spectrum of functionalities. We hope that this research yields deeper insights and enhanced performance, aligning with our commitment to trying to contribute to the field of multi-agent reinforcement learning and social Neuro-AI. We eagerly anticipate engaging with the NeurIPS and broader machine learning community to share our progress, foster collaborations, and drive forward the frontiers of cooperative AI.

Algorithm 1 RIM (SCOFF)-MAPPO

Initialize θ, the parameters for policy π and ϕ, the parameters for critic V, using Orthogonal initialization
Set learning rate α
while $\text{step} \leq \text{step}_{\max}$ **do**
  set data buffer $D = \{\}$
  for actor $i = 1, \ldots,$ batch_size **do**
    $\tau = []$ empty list
    Encode observation $z_t^l = \psi_\pi(O_t^l)$, positional encoding and initialize $h_{0,\pi}^{(1)} \cdots h_{0,\pi}^{(n)}$ actor RIM (SCOFF) hidden states
    Encode observation $z_t^g = \psi_V(O_t^g)$, positional encoding and initialize $h_{0,V}^{(1)} \cdots h_{0,V}^{(n)}$ critic RIM (SCOFF) hidden states
    for $t = 1 \cdots T$ **do**
      for all agent a **do**
        $p_t^{(a)}, h_{t,\pi}^{(a)} = \pi\big(z_t^l(a), h_{t-1,\pi}^{(a)}; \theta\big)$
        $u_t^{(a)} \sim p_t^{(a)}$
        $v_t^{(a)}, h_{t,V}^{(a)} = V\big(z_t^g(a), h_{t-1,V}^{(a)}; \phi\big)$
      end for
      Execute action u_t, observe r_t, $\boldsymbol{z}_{t+1}$, $\boldsymbol{O}_{t+1}$
      $\tau += [z_t^g, z_t^l, \boldsymbol{O}_t, h_{t,V}, h_{t,\pi}, u_t, r_t, z_{t+1}^g, z_{t+1}^l, \boldsymbol{O}_{t+1}]$
    end for
    Compute advantage estimate $\hat{A}$ via GAE on τ, using PopArt
    Compute reward-to-go $\hat{R}$ on τ and normalize with PopArt
    Split trajectory τ into chunks of length L
    for $l = 0, \cdots, T//L$ **do**
      $D = D \bigcup \big(\tau\big[l : l + T, \hat{A}[l : l + L], \hat{R}[l : l + L]\big]\big)$
    end for
  end for
  for mini-batch $k = 1, \cdots, K$ **do**
    $b \leftarrow$ random mini-batch from D with all agent data
    for each data chunk c in the mini-batch b **do**
      update RIM (SCOFF) hidden states for π and V from first hidden state in data chunk
    end for
  end for
  Adam update θ on $L(\theta)$ with data b
  Adam update ϕ on $L(\phi)$ with data b
end while

Acknowledgement

Zahra Sheikhbahaee and Adam Safron were supported by the Survival and Flourishing Fund (SFF-2023-H1) through the Institute for Advances Consciousness Studies (IACS). Guillaume Dumas was supported by the Institute for Data Valorization, Montreal (IVADO; CF00137433), the Fonds de recherche du Québec (FRQ; 285289), the Natural Sciences and Engineering Research Council of Canada (NSERC; DGECR-2023-00089), and the Azrieli Global Scholars Fellowship from the Canadian Institute for Advanced Research (CIFAR) in the Brain, Mind, & Consciousness program. Compute was enabled in part by support provided by Calcul Québec and Digital Research Alliance of Canada. In addition, the authors gratefully acknowledge the helpful feedback provided by colleagues, particularly Eugene Vinitsky and Anirudh Goyal. The author has no conflicts of interest to disclose.

References

[1] Spelke, E.S. and Kinzler, K. *Core knowledge*. Developmental Science 10:1, pp 89–96, 2007.

[2] Goyal, A. and Lamb, A. and Hoffmann, J. and Sodhani, S. and Levine S. and Bengio, Y. and Schölkopf, B., *Recurrent independent mechanisms*. In Proceedings of the International Conference on Learning Representations* (ICLR), 2021.

[3] Goyal, A., Lamb, A., Gampa, P., Beaudoin, P., Blundell, C., Levine, S., Bengio, Y. and Mozer, M. *Factorizing Declarative and Procedural Knowledge in Structured, Dynamical Environments*, In Proceedings of the International Conference on Learning Representations* (ICLR), 2021.

[4] Chollet, F., *On the Measure of Intelligence*, arXiv preprint arXiv:1911.01547, 2019.

[5] Graziano, M. S. A., and Kastner, S. *Human consciousness and its relationship to social neuroscience: A novel hypothesis*. Cognitive Neuroscience, 2(2): 98–113, 2011.

[6] Liu, D. and Bollotta, S. and Zhu, H. and Bengio, Y. and Dumas, G., *Attention Schema in Neural Agents*, arXiv preprint arXiv:2305.17375, 2023.

[7] Carlson, R.A., Dulany, D.E. *Conscious attention and abstraction in concept learning*. J. Exp. Psychol.: Learn. Mem. Cogn., vol. 11, pp. 45, 1985.

[8] Newman, J., Baars, B.J., Cho, S.B. *A neural global workspace model for conscious attention*. Neural Networks, vol. 10, pp. 1195–1206, 1997. doi: 10.1016/S0893-6080(97)00060-9

[9] Goyal, A. and Bengio, Y. *Inductive biases for deep learning of higher-level cognition*. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 478, no. 2266, pp. 20210068, 2022.

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I. *Attention is all you need*. In Advances in Neural Information Processing Systems, pages 5998–6008, 2017.

[11] Bahdanau, D., Cho, K., Bengio, Y. *Neural machine translation by jointly learning to align and translate*. arXiv preprint arXiv:1409.0473, 2014. [Online]. Available: `http://arxiv.org/abs/1409.0473`

[12] Welling, M. *Do we still need models or just more data and compute?* University of Amsterdam, April 20, 2019.

[13] Bolotta, S. and Dumas, G. *Social Neuro AI: Social Interaction as the "Dark Matter" of AI*. Frontiers in Computer Science, Volume 4, 2022.

[14] Christov-Moore, L., Reggente, N., Vaccaro, A., Schoeller, F., Pluimer, B., Douglas, P. K., Iacoboni, M., Man, K., Damasio, A., & Kaplan, J. T. *Preventing Antisocial Robots: A Pathway to Artificial Empathy*. Science Robotics 8:80, eabq3658, 2023.

[15] Agapiou, J. P. and Vezhnevets, A. S. and Duéñez-Guzmàn, E. A. and Matyas, J. and Mao, Y. and Sunehag, P. and Köster, R. and Madhushani, U. and Kopparapu, K. and Comanescu, R. and Strouse, D. J. and Johanson, M. B. and Singh, S. and Haas, J. and Mordatch, I. and Mobbs, D. and Leibo, J. Z., *Melting Pot 2.0*, arXiv preprint arXiv:2211.13746, 2023.

[16] Leibo, Joel Z. and Dueñez-Guzman, Edgar A. and Vezhnevets, Alexander and Agapiou, John P. and Sunehag, Peter and Koster, Raphael and Matyas, Jayd and Beattie, Charlie and Mordatch, Igor and Graepel, Thore. *Scalable evaluation of multi-agent reinforcement learning with melting pot*. In International Conference on Machine Learning, pp. 6187–6199. PMLR, 2021.

[17] Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., and Wu, Y. *The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games*, Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022), 2022.

[18] Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. *High-dimensional continuous control using generalized advantage estimation*. International Conference on Learning Representations, 2016.

[19] Goyal, A., Lamb, A., Hoffmann, J., Sodhani, S., Levine, S., Bengio, Y., and Schölkopf, B. (2019). *Recurrent independent mechanisms*. arXiv preprint arXiv:1909.10893.

[20] Goyal, A., Lamb, A., Gampa, P., Beaudoin, P., Levine, S., Blundell, C., Bengio, Y. and Mozer, M., 2020. *Object files and schemata: Factorizing declarative and procedural knowledge in dynamical systems*. arXiv preprint arXiv:2006.16225.