MapFormers: Transformers with Cognitively-Plausible Memory Systems

Main advisors: Victor Rambaud, Yair Lakretz

Lab: Laboratoire de Sciences Cognitives et Psycholinguistiques (LSCP), ENS-Paris

Duration: 6 months - full time

Transformers [1] have been the most widely used architecture in AI since their introduction in 2017, allowing the training of increasingly large Large Language Models, but despite their success, Transformer-based models learn incoherent world models [2] and are unable to generalize out-of-distribution (OOD) to non-regular formal tasks [3]. One explanation of this is the incapacity of the Transformer architecture to learn *cognitive maps* [4], abstract relational models which factorize content and structure, disentangling them, giving humans and animals the flexibility to adapt to new situations.

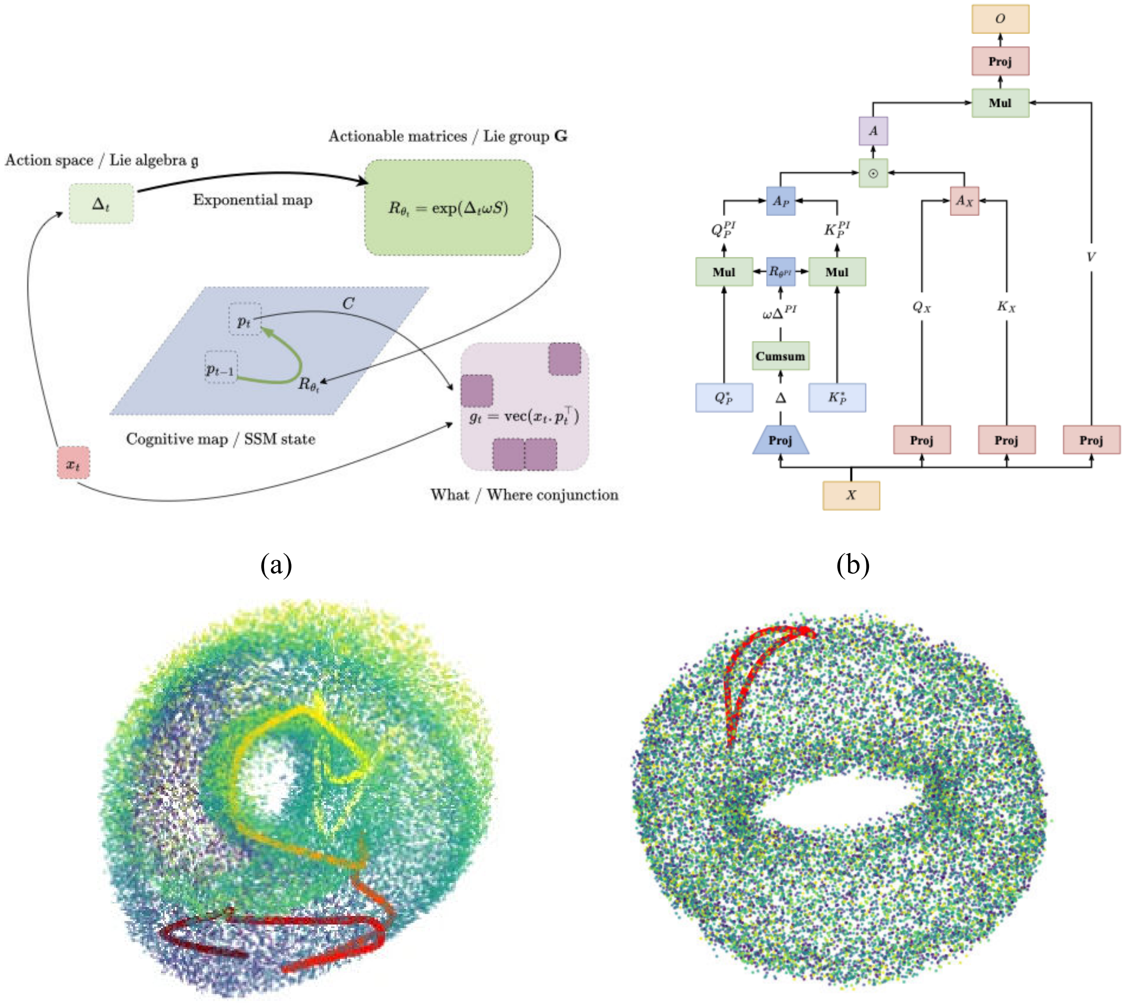
Recently, we have introduced *MapFormers* [5], a Transformer-based architecture with *input-dependent* matrices in order to update absolute or relative positional encoding, respectively as models of Episodic Memory (EM) and Working Memory (WM) [6]. *MapFormers* learn cognitive maps with no supervision, which provides these models with strong (OOD) generalization capacities on structure-dependent tasks such as 2D navigation.

Neural evidence of spatial cognitive maps have been identified in rodents in the hippocampal region, with *place cells* firing every time the animal came back to a specific location [7], while *grid cells* fire periodically as an animal is navigating a room [8]. In the Prefrontal Cortex (PFC), intracranial recordings have shown that the neural code of a macaque memorizing and repeating a sequence was sequentially shifted into orthogonal neural subspaces [9].

This phenomena can be mathematically explained by the formalism of *MapFormers* and their *input-dependent* matrices, however, one key difference between AI models and biological systems is that biological neurons cannot fire negatively. This constraint, coupled with bounded synaptic weights and neural activity can result in disentangled neural representations [10] and in the case of 2D navigation, explain geometric phenomena observed in *grid cells* [11].

Main Goals of the Intership:

- **Develop and extend *MapFormers* to learn complex relational structure**s: Extending our previous work – this involves investigating novel architectures or modifications to enhance the capacity of MapFormers to learn more complex cognitive mapping.
- **Conduct a neural analysis of learned representations in *MapFormers* and compare them with human and animal data**. This involves:
  - Implementing and evaluating the ability of *MapFormers* to learn complex recursive structures.
  - Studying the internal dynamics and comparing the learned neural representations with established biological data (e.g., place/grid cells, and sequential coding in the PFC).
  - Implementing and quantifying the impact of biologically-inspired constraints (such as non-negative neural firing, following).



(a) (b)

(c) (d)

(a) Schematic view of cognitive maps in Episodic Memory (EM) models. (b) Implementation of EM in transformers: *MapFormer*-EM, with a dedicated pool of neurons for position, acting as *grid cells* [8]. (c-d) Example of neural analysis comparison between animals and *MapFormers*. (c) Recorded neural activity of *grid cells* in a rodent's Entorhinal Cortex (EC) lies on a torus [12]. (right) PCA of samples trajectories of *MapFormer*-EM model, where the red curve corresponds to the model performing a square trajectory.

Prerequisites: Strong background and experience with PyTorch and Transformers, strong mathematical background. Finally, mastering Python and visualization tools is mandatory. Familiarity and interest with cognitive sciences is also desired.

Salary: based on the ENS grid for internships (~650 Euros per month).

Please send your CV, grade lists and relevant projects to:

victor.rambaud@gmail.com and yair.lakretz@gmail.com

References:

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

[2] Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model, 2024.

[3] Grégoire Deletang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, and Pedro A Ortega. Neural networks and the chomsky hierarchy. In The Eleventh International Conference on Learning Representations, 2023.

[4] Timothy E.J. Behrens, Timothy H. Muller, James C.R. Whittington, Shirley Mark, Alon B. Baram, Kimberly L. Stachenfeld, and Zeb Kurth-Nelson. What is a cognitive map? organizing knowledge for flexible behavior. Neuron, 100(2):490–509, 2018.

[5] Victor Rambaud, Salvador Mascarenhas, Yair Lakretz. MapFormer: Self-Supervised Learning of Cognitive Maps with input-dependent positional embeddings, 2025.

[6] James C.R. Whittington, William Dorrell, Timothy E.J. Behrens, Surya Ganguli, and Mohamady El-Gaby. A tale of two algorithms: Structured slots explain prefrontal sequence memory and are unified with hippocampal cognitive maps. Neuron, 113(2):321–333.e6, Jan 2025.

[7] J. O'Keefe and L. Nadel. The hippocampus as a cognitive map. Clarendon Press, Oxford, United Kingdom, 1978.

[8] Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I. Moser. Microstructure of a spatial map in the entorhinal cortex. Nature, 436(7052):801–806, Aug 2005.

[9] Yang Xie, Peiyao Hu, Junru Li, Jingwen Chen, Weibin Song, Xiao-Jing Wang, Tianming Yang, Stanislas Dehaene, Shiming Tang, Bin Min, and Liping Wang. Geometry of sequence working memory in macaque prefrontal cortex. Science **375**, 632-639, 2022.

[10] James C. R. Whittington, Will Dorrell, Surya Ganguli, and Timothy E. J. Behrens. Disentanglement with biological constraints: A theory of functional cell types, 2023.

[11] William Dorrell, Peter E. Latham, Timothy E. J. Behrens, and James C. R. Whittington. Actionable neural representations: Grid cells from minimal constraints, 2023.

[12] Richard J. Gardner, Erik Hermansen, Marius Pachitariu, Yoram Burak, Nils A. Baas, Benjamin A. Dunn, May-Britt Moser, and Edvard I. Moser. Toroidal topology of population activity in grid cells. Nature, 602(7895):123–128, Feb 2022.