M2 internship
Multimodal representation learning

Context

There is “a fundamental misalignment between human and typical AI representations: while the former are grounded in rich sensorimotor experience, the latter are typically passive and limited to a few modalities such as vision and text” [3]. We propose in the MeSMRise (Multimodal deep SensoriMotor Representation learning) project to take inspiration from the way human babies learn to explore their environment through actions that shape their multimodal experience [7]. Especially, the sensorimotor contingencies (SMC) theory [5] combines coherent pieces of evidence from neuroscience, psychology, etc. of human perception and learning in a unified framework. The key claims are the learning of SMCs defined as “the structure of the rules governing the sensory changes produced by various motor actions” [5] and active perception as the “organism’s exploration of the environment that is mediated by knowledge of SMCs” [4]. Some models implementing this theory are able to learn complex concepts such as containment [3] for instance.

Inspired by the SMC theory, the main objective of the project is to study how action can structure the multimodal representations, learned with self-supervised learning (SSL) methods. This will be applied to 3D objects, perceived by vision and point cloud, and manipulated in virtual environments. Specifically, we target the following properties:

- generalization to unknown environments and contexts
- robustness, e.g. to the orientation, background, shape ... of the object
- adaptability via the capacity of the model to autonomously find relevant information
- generality by using similar architectures and principles for all research questions

Subject

This intership takes place in the first work package of the project (that will also includes a PhD position opening on September 2024). The aim of this work package is to extend contrastive architectures by introducing action in the learned representations:

1. In [1] we introduced the parameters of the transformations in the representation, via an equivariance module, to improve visual classification of SoTA architectures. We will go further by introducing the consequence of action in the representation to learn a simple predictive model of the world. Moreover, we want to study more precisely the impact of using manipulation actions with these predictive representations, as we already showed that it improves performance when learning invariance to transformation [6].
2. We will study how the action, that can have various consequences depending on the modality, impacts contrastive multimodal learning. This raises the question of aligning the various modal manifolds that can be partially related. Moreover, in [2] we proposed a fusion method that considers the relevance of each modality depending on the precision of its sensor, based on a simple topological learning of the input density. We will adapt this method to the manifold learned by contrastive methods and extend it to consider richer relevance evaluation, such as a contextual one, to improve object recognition by focusing on the right modality at the right time (related to other work packages of the project).

Evaluation will include learning of predictive multimodal (vision, point cloud) representation of objects, especially their transfer to unknown environments and classification with little labeled data. Preliminary results could also be evaluated on more classical image datasets.

Profile

The following skills are mandatory:

- master's degree in artificial intelligence / machine learning or equivalent
- good programming skills (Python, Pytorch/Tensorflow)
- autonomy
- scientific curiosity

Duration

The internship will start in February-March and last 5-6 months.

Gratification

4.05€/h, 35h/week (i.e. around 580€/month)

Localisation

LIRIS laboratory, Lyon, France. Some travel to Clermont-Ferrand may also be organised.

Advisors

- Mathieu Lefort: associate professor at LIRIS, Lyon (`http://perso.liris.cnrs.fr/mathieu.lefort/`)
- Stefan Duffner: associate professor at LIRIS, Lyon (`http://duffner-net.de`)
- Jochen Triesch: visiting professor at Institut Pascal, Clermont-Ferrand - professor at FIAS, Frankfurt (`http://www.fias.science/en/life-and-neurosciences/research-groups/jochen-triesch/`)

Application

Please send a CV, cover letter and transcripts of your current and previous years' results to Mathieu Lefort (mathieu.lefort@liris.cnrs.fr).

References

[1] Alexandre Devillers and Mathieu Lefort. Equimod: An equivariance module to improve visual instance discrimination. In *The Eleventh International Conference on Learning Representations*, 2022.

[2] Simon Forest, Jean-Charles Quinton, and Mathieu Lefort. Combining manifold learning and neural field dynamics for multimodal fusion. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.

[3] Nicholas Hay, Michael Stark, Alexander Schlegel, Carter Wendelken, Dennis Park, Eric Purdy, Tom Silver, D Scott Phoenix, and Dileep George. Behavior is everything: Towards representing concepts with sensorimotor contingencies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[4] Erik Myin and J Kevin O'Regan. Perceptual consciousness, access to modality and skill theories. a way to naturalize phenomenology? *Journal of consciousness studies*, 9(1):27–46, 2002.

[5] J Kevin O'regan and Alva Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, 24(5):939–973, 2001.

[6] Felix Schneider, Xia Xu, Markus R Ernst, Zhengyang Yu, and Jochen Triesch. Contrastive learning through time. In *SVRHM 2021 Workshop@ NeurIPS*, 2021.

[7] Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2):13–29, 2005.