Multi-Modal Alignment for Robotic Episodic Memory using Auto-Encoders

Timo Birr, Joana Plewnia, and Tamim Asfour

High Performance Humanoid Technologies Lab (H^2T), Karlsruhe Institute of Technology (KIT)

Background. In our robot software framework [1], we develop a cognitive architecture centered around a biologically inspired memory system [2]. This architecture distinguishes between working memory for short-term processing and episodic long-term memory for persistent storage of experiences.
Our long-term memory combines two complementary approaches: A file-system-based storage for raw data and a learned Deep Episodic Auto-Encoder (DEAE) for compact, structured representations of experience [3] [4]. The DEAE allows us to compress so-called memory snapshots — representations of the world at a specific point in time — into a latent space that supports efficient storage and predictive modeling. Each snapshot captures the state of a particular entity in the world (e.g., an object, a robot, a human, an image). Entities are dynamic and can evolve over time, and their changing state is captured by sequences of snapshots.

Figure 1: generated by ChatGPT

Idea. The current design of our memory system focuses on encoding snapshots of individual entities over time using separate auto-encoders. While this approach is effective for capturing the evolution of isolated objects or agents, it does not exploit the rich relationships and dependencies that exist between different entities within a scene. In particular, it lacks the ability to create joint representations of complete memory snapshots that reflect the state of the entire environment at a given moment.
To overcome this limitation, the goal of this project is to investigate how multiple modalities — such as images, text descriptions, trajectories, joint angles, and semantic scene graphs — can be aligned and integrated into a shared latent space within a single multi-modal auto-encoder architecture. This would enable us to encode complete memory snapshots of the full scene, rather than treating each entity in isolation. Such a representation would not only improve the compression efficiency of our long-term memory, but also facilitate cross-modal prediction, relational reasoning, and a more holistic understanding of complex robotic environments.

Requirements. The project is suitable for students with a strong interest in machine learning and robotics, particularly in the areas of representation learning and neural networks. Experience with Python and machine learning frameworks such as PyTorch is highly recommended, as the project will involve the design and implementation of deep learning models. Furthermore, the project requires the ability to work independently on a research-oriented task, including reading and analyzing scientific literature. The student should also be motivated to write a scientific publication

Contact: timo.birr@kit.edu, joana.plewnia@kit.edu www.humanoids.kit.edu

based on their results and be comfortable with structured scientific writing, as the end-goal will be a publication in a renowned conference like ICRA, Humanoids or IROS.

[1] Vahrenkamp, N., Wächter, M., Kröhnert, M., Welke, K. and Asfour, T., The Robot Software Framework ArmarX, Information Technology, vol. 57, no. 2, pp. 99-111, 2015

[2] Peller-Konrad, F., Kartmann, R., Dreher, C. R. G., Meixner, A., Reister, F., Grotz, M. and Asfour, T., A memory system of a robot cognitive architecture and its implementation in ArmarX, Robotics and Autonomous Systems, vol. 164 , pp. 1-20, 2023

[3] Rothfuss, J., Ferreira, F., Aksoy, E. E., Zhou, Y. and Asfour, T., Deep Episodic Memory: Encoding, Recalling, and Predicting Episodic Experiences for Robot Action Execution, IEEE Robotics and Automation Letters (RA-L), vol. 3, no. 4, pp. 4007-4014, October, 2018

[4] Bärmann, L., Peller-Konrad, F., Constantin, S., Asfour, T. and Waibel, A., Deep Episodic Memory for Verbalization of Robot Experience, IEEE Robotics and Automation Letters (RA-L), vol. 6, no. 3, pp. 5808-5815, 2021