Keynote 2

Learning and Evolution for Real Robots

by external keynote speaker Dr. Eiji Uchibe, Okinawa Institute of Science and Technology

Abstract: Our daily behaviours are guided by rewards in multiple ways, such as appetitive, aversive, sexual, and social rewards. What is the origin of such multiple reward systems? The goal of the Cyber Rodent project (http://www.nc.irp.oist.jp/crp/) is to explore the design principles of the reward systems for artificial agents to realize self-preservation and self-reproduction, and thereby try to better understand the origins of reward systems of biological agents. Reinforcement learning (RL) is an attractive learning framework with a wide range of possible application areas. The RL framework has been conceived as a model of animal and robot behavioural learning. However, critical unsolved problems in the real-world applications of RL are the choices of state representations, learning algorithms, reward functions, and meta-parameters. To deal with these problems, we introduce two learning frameworks called CLIS and CPGRL.

CLIS: Our brain can be seen as a heterogeneous mixture of multiple agents: simple, hard-wired controllers in the spinal cords and the brainstem to highly adaptive functions the cerebrum and the cerebellum. A recent brain imaging experiment suggests that there are parallel reinforcement learning pathways in the human brain, each specialized for reward prediction at different time scales. CLIS, derived as a practical means for accelerating learning by maximally utilizing limited number of experiences, might give a clue for understanding the parallel learning mechanism of the brain. The robot possesses multiple modules with different state representations, learning algorithms, and meta-parameters and improves their multiple policies simultaneously to accomplish a particular task. Some of the modules can have fixed hand-coded policies. The CLIS framework can select an appropriate module for action and accurately improve the policies of all learning modules, including those that are not selected, based on the method of importance sampling. It is possible to obtain purposive behaviours efficiently and rapidly because the modules that are not selected can learn from the experience derived by the actions of another module.

CPGRL and Embodied Evolution: Reward functions can usually be classified into two types: those directly representing the successful achievement of the task and those aimed for facilitating efficient and robust learning. We assume that the former, “extrinsic rewards”, are fixed for a given task and consider how the latter, “intrinsic rewards”, can be optimized by the robots during their lifetime or within their colony by evolution. Typical examples of intrinsic rewards are the curiosity for novelty that promote exploration and innate preference for certain sensory features that promote goal-directed search. CPGRL maximizes the average of intrinsic rewards within the bounds specified by the extrinsic rewards. This enables optimization of intrinsic rewards without compromising the main task goals specified by the extrinsic rewards. For optimization of intrinsic rewards, we take the embodied evolution approach. Each agent in a colony has a genetic code for its own intrinsic rewards and exchanges the codes with fellow agents as they pass by. Additional information about the fitness, which is closely related to the task achievement represented by extrinsic rewards, specifies the intrinsic rewards in the new generation.