ERCIM "ALAIN BENSOUSSAN" FELLOWSHIP PROGRAMME

ERCIM
European Research Consortium for Informatics and Mathematics

Scientific Report

First name / Family name	Davide Zambrano
Nationality	Italian
Name of the *Host Organisation*	Centrum Wiskunde & Informatica (CWI)
First Name / family name of the *Scientific Coordinator*	Gunnar Kalu
Period of the fellowship	1/5/2014 to 30/4/2015

I – SCIENTIFIC ACTIVITY DURING YOUR FELLOWSHIP

How to decide what to do next? A monkey exposed to visual stimuli on a screen decides which actions give the reward: juice. The monkey does not know the rules of the task, nor when the task starts or ends or when new task rules apply. Yet, it attends to the task while its brain processes thousands of asynchronous events and asynchronously coordinates many muscles. This is an example of the complexity of the environment that we live in, where many unknown and unexpected events have to be recognized and processed rapidly and continuously.

Despite advances in machine learning, current robotic systems are not able to rapidly and efficiently respond in the real world: the challenge is to learn to recognize both *what* is important, and also *when* to act. Reinforcement Learning (RL) algorithms are commonly used as a learning paradigm to learn *what* to respond to in complex environments. In RL, the agent changes its behaviour according to experience collected during the exploration of the world. In typical RL tasks however, ad hoc abstractions are used: the actual relevant events are provided in compact representations. Much attention has been given recently to learning compact state representations from high-dimensional observations, where deep learning is the most well-known approach for this, including approaches like Long Short-Term Memory (LSTM) that capture state from both present and past observations in memory structures. Memory allows such networks to solve the

what problem as posed by working memory tasks, by transforming certain classes of partially observable Markov decision problems (POMDPs) into Markov Decision Problems (MDPs). Effectively, working memory learns to extract *what*.

In standard RL not only the representation is abstracted, but also the *timing* of events that is sampled in the ordered presentation. State-transitions in RL are defined in discrete steps and the agent state is updated every step. Effectively, the agent is given the information on *when* a decision has to be taken. To respond quickly however, the environment has to be sampled often. A programmer has to decide on the step-size as a time-representation, choosing between a fine-grained representation of time or to a coarse temporal resolution. The former corresponds to many state-action transitions that are difficult to learn, while in the latter correct action sequences are easier to learn but lack precise timing. For a learning self-driving car, this means it will either be very difficult to learn to avoid unexpected obstacles, or it will respond too late and hit the man.

AuGMEnT (Attention-Gated MEmory Tagging), developed by Rombouts (CWI), Dr. Bohte (CWI) and Prof. Roelfsema (NIN), is a recent biologically plausible neural network framework that is trained with on-policy SARSA. AuGMEnT includes working memory and shares a number of features with Long Short-Term Memory (LSTM). During my period at CWI, I derived a continuous-time version of AuGMEnT. This solution is based on the idea that in biological brains, instantaneous actions of infinitesimal duration are actually impossible. We introduced an action selection system that controls the action execution, by keeping active the selected action for the needed time. In biological brains, an action selection network can be mapped onto specific neural substrates, including the basal ganglia. Current actions can be interrupted if another is more important or urgent.

Moreover the exploratory system has been re-defined allowing enough time for execution and thus for spatial and temporal credit assignment. The continuous-time framework changes the way standard RL problems are presented: it defines *time* as an intrinsic property of the task and it considers unavoidable delays in action selection and execution. Several case studies, including the monkey's task, have been tested, showing the ability of the network to reach convergence within a remarkable short number of epochs.

II – PUBLICATION(S) DURING YOUR FELLOWSHIP

Davide Zambrano, Pieter R Roelfsema, and Sander M Bohte, "**Continuous-Time on-Policy Neural Reinforcement Learning of Working Memory Tasks**," International Joint Conference on Neural Networks, 2015.

Abstract. As living organisms, one of our primary characteristics is the ability to rapidly process and react to unknown and unexpected events. To this end, we are able to recognize an event or a sequence of events and learn to respond properly. Despite advances in machine learning, current cognitive robotic systems are not able to rapidly and efficiently respond in the real world: the challenge is to learn to recognize both *what* is important, and also when to act. Reinforcement Learning (RL) is typically used to solve complex tasks: to learn the *how*. To respond quickly - to learn *when* - the environment has to be sampled often enough. For "enough", a programmer has to decide on the step-size as a time-representation, choosing between a fine-grained representation of time (many state-transitions; difficult to learn with RL) or to a coarse temporal resolution (easier to learn with RL but lacking precise timing). Here, we derive a

continuous-time version of on-policy SARSA-learning in a working-memory neural network model, AuGMEnT. Using a neural working memory network resolves the *what* problem, our *when* solution is built on the notion that in the real world, instantaneous actions of duration *dt* are actually impossible. We demonstrate how we can decouple action duration from the internal time-steps in the neural RL model using an action selection system. The resultant CT-AuGMEnT successfully learns to react to the events of a continuous-time task, without any pre-imposed specifications about the duration of the events or the delays between them.

Davide Zambrano, Pieter Roelfsema and Sander Bohte **Learning continuous time representations of tasks** (in preparation)

III – ATTENDED SEMINARS, WORKHOPS, CONFERENCES

CAMP 2014 Computational Approaches to Memory and Plasticity, National Centre for Biological Sciences, Bangalore, India

Dutch-Belgian Reinforcement Learning Workshop 2014, November 28, Brussels, Belgium. http://ai.vub.ac.be/node/1308
Talk. Davide Zambrano: Toward a Continuous Time AuGMEnT.

Symposium on Intelligent Machines http://www.snn.ru.nl/v2/ml2015.php. 17 March 2015 in Nijmegen, Netherlands
Poster presented. D. Zambrano, P. Roelfsema and S. Bohte. Continuous-time neural reinforcement learning of working memory tasks.

The Integrated Systems Neuroscience Workshop 2015
http://www.isn2015.ls.manchester.ac.uk. 23-24 March 2015 in Manchester, United Kingdom
Poster presented. D. Zambrano, P. Roelfsema and S. Bohte. Learning continuous time representations of tasks.

IV – RESEARCH EXCHANGE PROGRAMME (REP)

Visiting Prof. Frederic Alexandre, Director of Research, INRIA-Bordeaux, Bordeaux, France.
Activity description.
How do we learn when to explore? Most reinforcement learning (RL) algorithms, including AuGMEnT and more importantly in CT-AuGMEnT, use a constant exploration. This means that, during the trial, in every moment there is the same probability to explore other non-greedy actions. This is exacerbated in continuous-time system where every agent update has the same exploration probability. However, we are likely to explore when we are unconfident about inputs. Several studies have shown that the activity of the amygdala is correlated with the confidence with the environment. Bakker in 2002 proposed a similar approach by training a separate network on the prediction of the time different error. However, this system was not defined in

continuous-time and it is not explicitly linked with the amygdala activity.
In our work, the amygdala will learn to associate a conditioned stimuli (CS) with an unconditioned stimuli (US). A model of amygdala that solves distal reward problems in continuous-time has not been developed yet. A possible approach could be the work proposed by Ludvig in 2008. The INRIA's group will work on that by providing a simplified amygdala model based on their recent studies, able to inform about the confidence with the input. To be consistent with their current model, the confidence with the input can only be computed as the difference between the amygdala activity in response to CS and the experienced reward (US). Thus, only at the end of the trial. That error will be used to evaluate the uncertainty about the amygdala prediction in the next trial, during the next presentation of the CS.
My contribution was to demonstrate that the decrease of the exploration probability in specific moments of the trial improves convergence rate. The exploration rate decreases in case of low uncertainty (e.g. if a reward has been given in a previous trial). As expected, the results demonstrate an improvement in the convergence rate, however further studies are needed to merge the activity from the amygdala into the AuGMEnT framework.