An Actor Critic with An Internal Model

Farzaneh S. Fard (fard@cs.dal.ca)
Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada

Abraham Nunes (nunes@dal.ca)
Department of Psychiatry, Dalhousie University, Halifax, Nova Scotia, Canada
Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada

Thomas Trappenberg (tt@cs.dal.ca)
Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada

Abstract

Current evidence suggests that the brain uses multiple systems for instrumental control; these systems are known as *model-based* and *model-free*. The former predicts action-outcomes using an internal model of the agent's environment, while the latter learns to repeat previously rewarded actions. This paper proposes a neural architecture comprised of both model-free and model-based reinforcement learning systems, and tests this model's ability to perform target-reaching with a simulated biarticulate robotic arm. Target-reaching conditions included (A) both static and dynamic target properties, (B) slowly changing robotic arm kinematics, and (C) absence of visual inputs. The proposed model rapidly learns an internal model of environmental dynamics, shows target-reaching performance superior to an existing state of the art model, and successfully performs target-reaching without visual input.

Keywords: Machine Learning; Deep Reinforcement Learning; Model-free; Model-based; Predictive Learning;

Introduction

Current evidence posits the existence of multiple behavioural control systems in the brain (Daw & O'Doherty, 2013). One of these, the model-based system, employs an internal representation of the world and its body (Kawato et al., 2003), which is advantageous for multiple reasons. These include (A) endowment of the agent with information for decision-making when sensory inputs are lacking, and (B) predicting the future states of the world given the agent's actions (Kawato et al., 2003; Wolpert, Miall, & Kawato, 1998). These benefits are exemplified during arm-reaching in the absence of visual input.

We have previously shown (using dynamic neural fields) that adaptive internal representations of environmental dynamics could provide reliable information for target-reaching during occluded vision (Fard, Hollensen, Heinke, & Trappenberg, 2015). As such, in the present study we expand the deep deterministic policy gradient (DDPG) method (Lillicrap et al., 2015) by combining it with a capacity to learn internal models of environmental dynamics during control of a simulated robotic-arm. The result is an architecture that can learn an internal model to predict and plan movements. We compared our model's ability to perform target-reaching (using a simulated two-joint arm) with that of DDPG.

Proposed model

Figure 1 illustrates our proposed model. Input to the actor includes (1) the current location of the robot arm joints, and (2) the current target location. Actor output is a (continuous) change in shoulder (α) and elbow (β) joint angle. Given these actions and the present state, the forward model predicts the next location of the arm. The forward model is trained using the Euclidean distance between actual and predicted arm positions at the next time step. The reward function defined as the negative Euclidean distance between the hand and target. The integrator component integrates predicted future location with the real future location (visual information). The output of the integrator component updates the critic through TD-error component. To train the actor, the ∇ component provides a gradient derived from either the forward model or the critic; here, precedence is given to the forward model error signal, which if unavailable is substituted by the critic-derived error signal.

Experiments and Results

We tested our model on a target reaching task. We examined our model compared to the DDPG under 4 different circumstances (static/changing kinematics and static/changing target locations). *Changing kinematics* are modeled by increasing the length of arm segments by 0.001 cm in every step after episode 100. The *changing target* is modeled by (randomly) relocating the target in every episode. The initial arm lengths were set to 8 cm and 5 cm for distal and proximal components, respectively. To facilitate statistical comparisons, we ran 20 experimental comparisons of our model with DDPG: each experiment runs for 1000 episodes, with a maximum of 30 time-steps per episode. Episodes were terminated upon target reaching—defined as a distance of less than 0.5 cm between the robot ”hand” and the target—or lapse of 30 time steps (whichever occurred first). Performance was quantified as the proportion of episodes in which the target was successfully reached, and this was compared between models using a two-sample z-test for proportions. Table 1 shows these results under 4 different conditions, averaged over 20 different runs. Figure 2 shows performance during reaching of a changing target with changing arm kinematics (averaged over 20 runs). These data show statistically significant performance improvement with the proposed model.

An important contribution of the actor-critic with internal models (our proposed architecture) is the ability to reach the target despite occluded vision. Conversely, DDPG and other model-free solutions are dependent on visual input. After 900 episodes of training, we tested the ability of our model to reach 20 sequentially presented targets. After presenting the target location at the initial time-step, no further visual feedback was available. The first outcome measure was the hand-to-target distance calculated once the blinded arm stopped at the location it predicted to be sufficiently close to the target (see Table 2). Second, successful target reaching was defined as a distance less than 5mm between the robot hand and target.

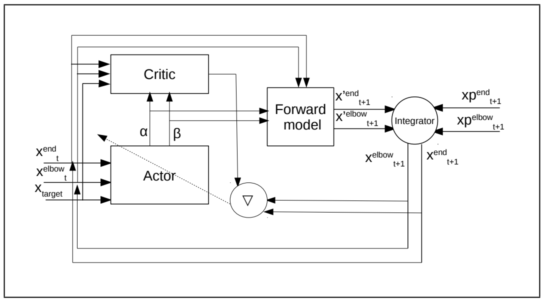

Figure 1: The proposed model architecture. The locations of arm (x^{end}, x^{elbow}) and the target (x_{target}) define the current state. Actions α and β represent joint angle changes. The *Integrator* integrates the predicted (x'^{end}, x'^{elbow}) and proprioceptive (xp^{end}, xp^{elbow}) signals, and ∇ uses the error signal to train the actor.

Table 1: Comparison of average performance between our proposed model and DDPG during 1000 episodes over 20 runs.

Target/Kinematics	DDPG (SD)	Ours (SD)	P-value
Static/Static	94.5 (2.5)	95.7 (7.6)	0.6
Static/Changing	87.5 (12.0)	96.4 (4.1)	0.02
Changing/Static	50.7 (15.9)	83.0 (2.4)	< 0.001
Changing/Changing	46.4 (10.7)	83.3 (2.6)	< 0.001

Discussion

Learning an internal representation of the environment is important for flexible decision making in complex environments with noisy (or absent) sensory inputs. Our data show that combination of such a model with actor-critic (model-free) control enables learning of flexible decision-making policies. In the face of changing reward contingencies (e.g. variable target positions), this is akin to expression of goal-directed control. Moreover, the proposed architecture demonstrates that behavioral control guided by an internal model of the world can account for adaptive behaviour in the absence of continual sensory feedback. Further work must better elucidate the mechanisms of arbitration between model-based and model-free control systems, and incorporate Pavlovian-instrumental interactions.

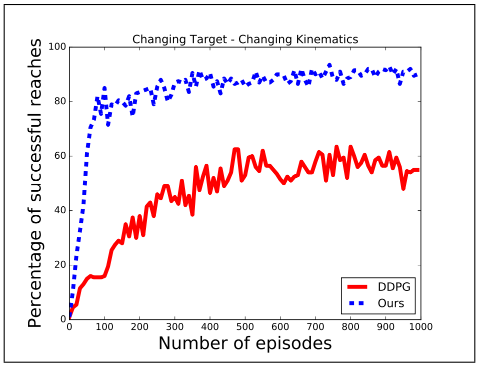

Figure 2: Percentage of successful reach over episodes for DDPG and the proposed model.

Table 2: Performance of the proposed model with occluded vision during 10 sequential target-reaching episodes after 900 episodes of training. *Reached* denotes number of successful reaches. *Steps* denotes the average number of time steps until the model terminated reaching movement. This test was thrice repeated and and p-values computed by binomial test.

Target/Kinematics	Reached	Steps (SD)	P-value
Static/Static	20	1.01 (0.12)	< 0.001
Static/Changing	20	1.65 (0.5)	< 0.001
Changing/Static	19	2.06 (2.4)	< 0.001
Changing/Changing	16	2.73 (4.06)	< 0.001

References

Daw, N. D., & O'Doherty, J. P. (2013). Multiple Systems for Value Learning. In P. W. Glimcher & E. Fehr (Eds.), *Neuroeconomics: Decision making and the brain: Second edition* (pp. 393–410). Elsevier Inc.

Fard, F. S., Hollensen, P., Heinke, D., & Trappenberg, T. P. (2015). Modeling human target reaching with an adaptive observer implemented with dynamic neural fields. *Neural Networks*, *72*, 13–30.

Kawato, M., Kuroda, T., Imamizu, H., Nakano, E., Miyauchi, S., & Yoshioka, T. (2003). Internal forward models in the cerebellum: fmri study on grip force and load force coupling. *Progress in brain research*, *142*, 171–188.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., ... Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Wolpert, D. M., Miall, R. C., & Kawato, M. (1998). Internal models in the cerebellum. *Trends in cognitive sciences*, *2*(9), 338–347.