Toward a Goal-Directed Construction of State Spaces

Sohrab Saeb and Cornelius Weber
Frankfurt Institute for Advanced Studies, Goethe Universität, Frankfurt/Main, Germany

FIAS Frankfurt Institute for Advanced Studies

Introduction

Reinforcement learning of complex tasks presents at least two major problems. The first problem is caused by the presence of sensory data that are irrelevant to the task. It will be a waste of computational resources if an intelligent system represents information that are irrelevant, since in such a case state spaces will be of high dimensionality and learning will become slow. Therefore, it is important to represent only the relevant data. Unsupervised learning methods such as independent component analysis can be used to encode the state space [1]. These methods are able to separate sources of relevant and irrelevant information in certain conditions, however, all data are represented.

The second problem arises when information about the environment is incomplete as in so-called partially observable Markov decision processes. This leads to the perceptual aliasing problem, where different world states appear the same to the agent even though different decisions have to be made in each of them. To overcome this problem, one should constantly estimate the current state based also on previous information. This estimation process is traditionally performed using Bayesian estimation approaches such as Kalman filters and hidden Markov models [2].

The above-mentioned methods for solving these two problems are merely based on the statistics of sensory data without considering any goal-directed behaviour. Recent findings from biology suggest an influence of the dopaminergic system on even early sensory representations, which indicates a strong task influence [3,4]. Our goal is to model such effects in a reinforcement learning approach.

Model Architecture

- An input sensory later, a hidden state layer and an output action layer.
- Soft-max activation functions for each neuron, with the inverse temperature parameter β that controls the amount of exploration vs. exploitation.
- Stochastic action selection based on output layer activation.

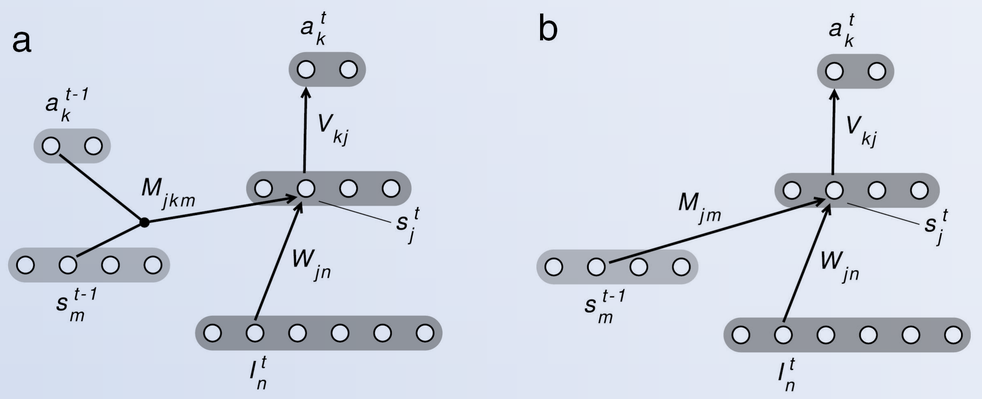

Figure 1. Model architecture. (a) Extended recurrent architecture includes two memory layers which hold the state and the action vectors of the previous time step. (b) Simple recurrent architecture includes only one memory layer for the previous state vector.

Temporal Difference Learning

- Learning is performed by gradient descent on the following energy function:

$$E^t = \frac{1}{2} \sum_{\mathbf{s}^t, \mathbf{a}^t} \mathrm{P}^{\Pi}(\mathbf{s}^t, \mathbf{a}^t)(Q^{\Pi}(\mathbf{s}^t, \mathbf{a}^t) - Q(\mathbf{s}^t, \mathbf{a}^t))^2$$

where:

- $\mathrm{P}^{\Pi}(\mathbf{s}^t, \mathbf{a}^t)$ is the policy(Π)-dependent distribution of state action pairs;
- $Q(\mathbf{s}^t, \mathbf{a}^t)$ is the current estimate of the value function that is estimated by the network:

$$Q(\mathbf{s}^t, \mathbf{a}^t) = \sum_{j,k} V_{kj}\, a_k^t\, s_j^t$$

- $Q^{\Pi}(\mathbf{s}^t, \mathbf{a}^t)$ is the true policy-dependent value function that is estimated as:

$$Q^{\Pi}(\mathbf{s}^t, \mathbf{a}^t) \simeq r^t + \gamma Q(\mathbf{s}^{t+1}, \mathbf{a}^{t+1})$$

where r^t is the reward signal and γ is the future value discount factor.

- The following weight update rules emerge as a result:

$$\Delta V_{kj} \propto \delta^t\, a_k^t\, s_j^t$$

$$\Delta W_{jn} \propto \delta^t\, \beta\, s_j^t I_n^t \left(V_{ij} - \sum_p V_{ip} s_p^t \right)$$

$$\Delta M_{jkm} \propto \delta^t\, \beta\, s_j^t\, a_k^{t-1} s_m^{t-1} \left(V_{ij} - \sum_p V_{ip} s_p^t \right)$$

Reinforcement Learning Scenario

- Four visual features consisting of four bars
- Four actions to move the visual field, with periodic boundary condition
- 10% probability for each action to fail
- 50% probabily for each bar to be invisible

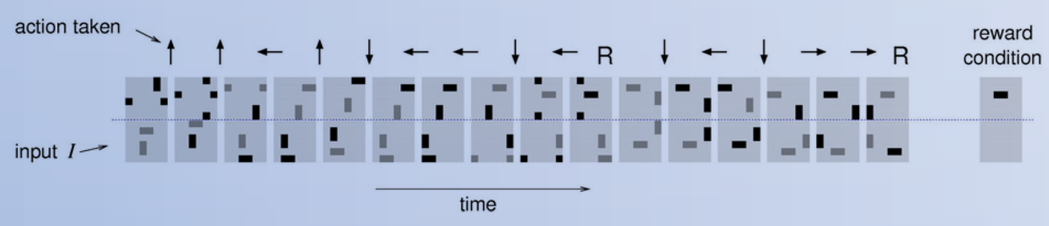

Figure 2. Reinforcement learning scenario. Visual features are upper horizontal (UH), upper vertical (UV), lower horizontal (LH), and lower vertical (LV) bars. Black color indicates visible and gray color invisible bars, which appear as the background to the network. An action moves the features into one of four directions from one time step to the next, as indicated by the arrows. A reward R is given when the ``rewarded feature" (UH) appears at a specific location (shown as ``reward condition").

Experiments and Results

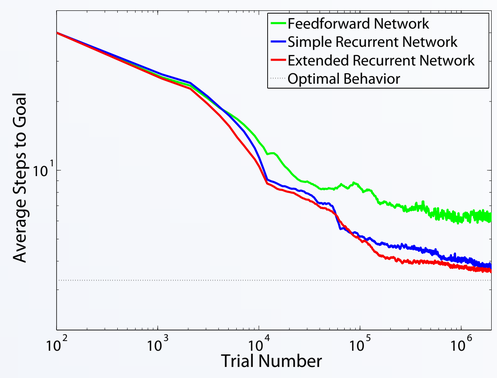

Figure 3. (Above) Action performance for simple and extended recurrent architectures, as well as the memory-less feedforward network. The performance is defined as the average number of steps taken by the agent to reach the goal. The optimal performance is 3.3 (dotted horizontal line).

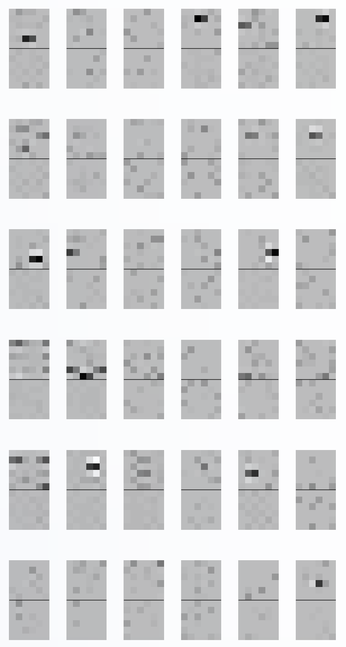

Figure 4. (Right) Receptive fields emerging from first layer connections (right). Obviously the weights are coding the relevant feature (UH).

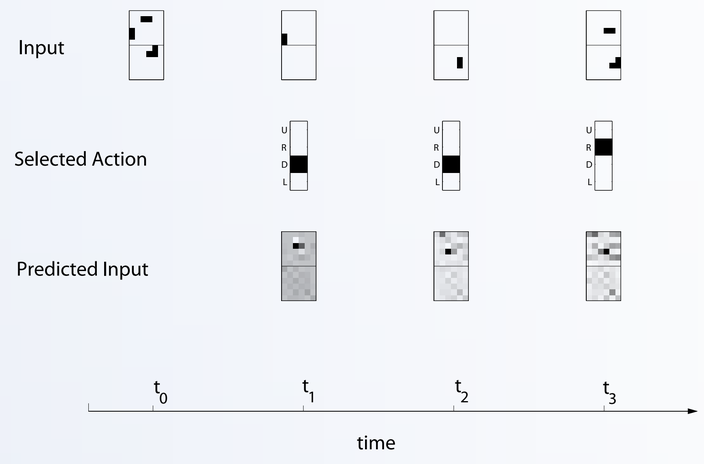

Figure 5. Predicted network input compared to real input for three successive time steps that lead to the goal (the final position of the upper horizontal bar is the goal position). Note that features are randomly switched off, and the network tries to predict the current position of the relevant feature regardless of other features' positions. The action space consists of going up (U), down (D), right (R) and left (L).

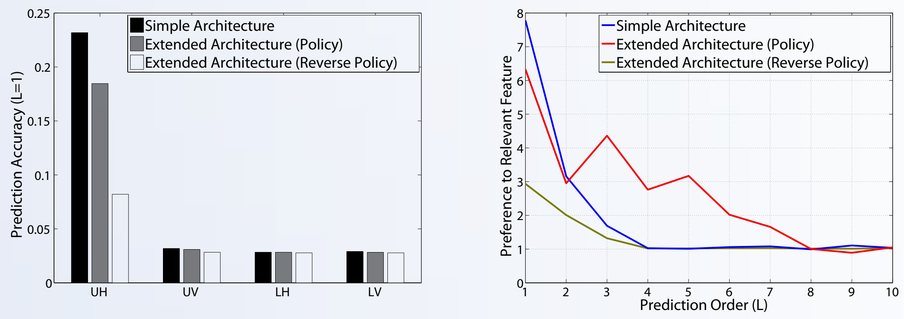

Figure 6. (Left) Prediction accuracies for UH, UV, LH and LV features, calculated for the first-order prediction task. (Right) Preference to the relevant feature (UH) versus prediction order. For all cases the preference approaches 1 for overly long-lasting trials, meaning that for high prediction orders the networks do not show preference to predict the relevant feature.

Discussion

Our proposed model learns both, to ignore task-irrelevant and to predict missing relevant features of sensory information. It learns a bars task which is challenging even for humans: when a reward is given whenever one of several features appears at a specific position, it is hard to track which feature consistently appears at the same position over several trials. This scenario becomes even more complicated when the bars randomly disappear from the visual field.

The experimental setup used in this study is one step toward getting reinforcement learning to deal with realistic image data with pixel information. Many reinforcement learning approaches still rely on hand-crafted, abstract state spaces to model learning mechanisms of the brain, whereas in our model a goal-directed state space is autonomously shaped.

Acknowledgement

This work was supported by EU projects "PLICON" and "IM-CLeVeR", and by the Hertie Foundation.

References

[1] Independent component analysis: a new concept? P. Comon. Signal Processing, 36(3):287-314 (1994).
[2] Planning and acting in partially observable stochastic domains. L. P. Kaelbling, M. L. Littman and A. R. Cassandra. Artificial Intelligence, 101:99-134 (1995).
[3] Practising orientation identification improves orientation coding in V1 neurons. A. Schoups, R. Vogels, N. Qian and G. Orban. (2001). Nature, 412: 549-53 (2001).
[4] Reward-dependent modulation of working memory in lateral prefrontal cortex. S. W. Kennerley, and J. D. Wallis. J. Neurosci, 29(10): 3259-70 (2009).