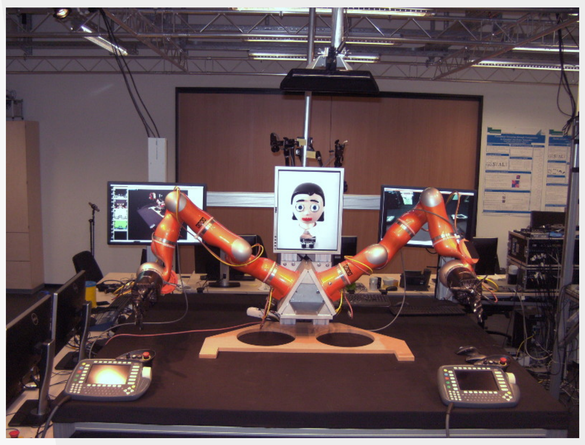
Universität Bielefeld
Maximilian Panzner

A Deep Reinforcement Learning Based Model Supporting Object Familiarization

Motivation



FAMULA

Deep Familiarization and Learning Anthropomorphic Robotic Platform

An important ability of cognitive systems is the ability to familiarize themselves with the properties of objects and their environment as well as to develop an understanding of the consequences of their own actions on physical objects.

Challenges

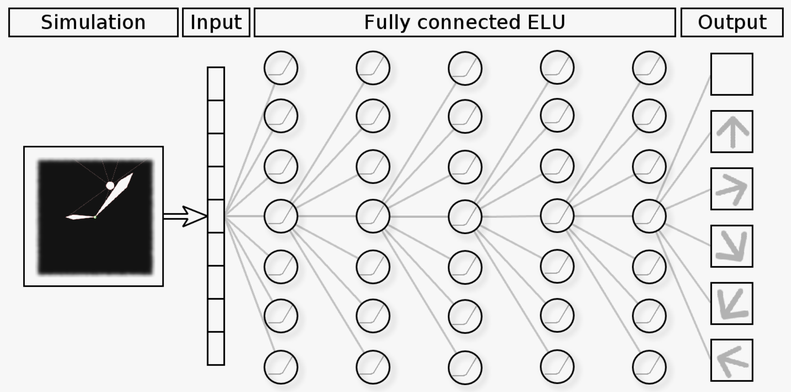
- The system has to learn meaningful ways to manipulate objects without prior knowledge
- Explore the state-action space and find good tradeoff between exploration and exploitation
- Structure the state-action space by exploiting reproducible action-reaction patterns
- Find good utility rewards to drive the learning

Environment

Simulation environment for robotic table-top object manipulation tasks.

- Simulated toy clock for the agent to familiarize with
- Actuator modelled after the fingertip of the robotic hand
- Partial observation of the environment
- Faster than realtime simulation on multiple servers for faster experimentation compared to a physical robotics platform
- Accurate physics

Network Layout



- Deep learning and reinforcement based model that allows a system to familiarise itself with consequences of actions performed on an object
- Exploration based emergence of state-action space and policy based on repeatable action-reaction patterns

Reward Functions

Reward functions support the system in distinguishing helpful from unhelpful actions.

- Rewards are given only at certain points in time and it is the task of the learner to identify the key decisions made in the past that led to achieving the reward.
- Evaluation of three different reward combinations of: goal derived, tutoring and penalty rewards

Exploration Strategies

The system learns by structuring the state-action space through exploration. An exploration strategy mediates between exploration and exploitation of already learned policy

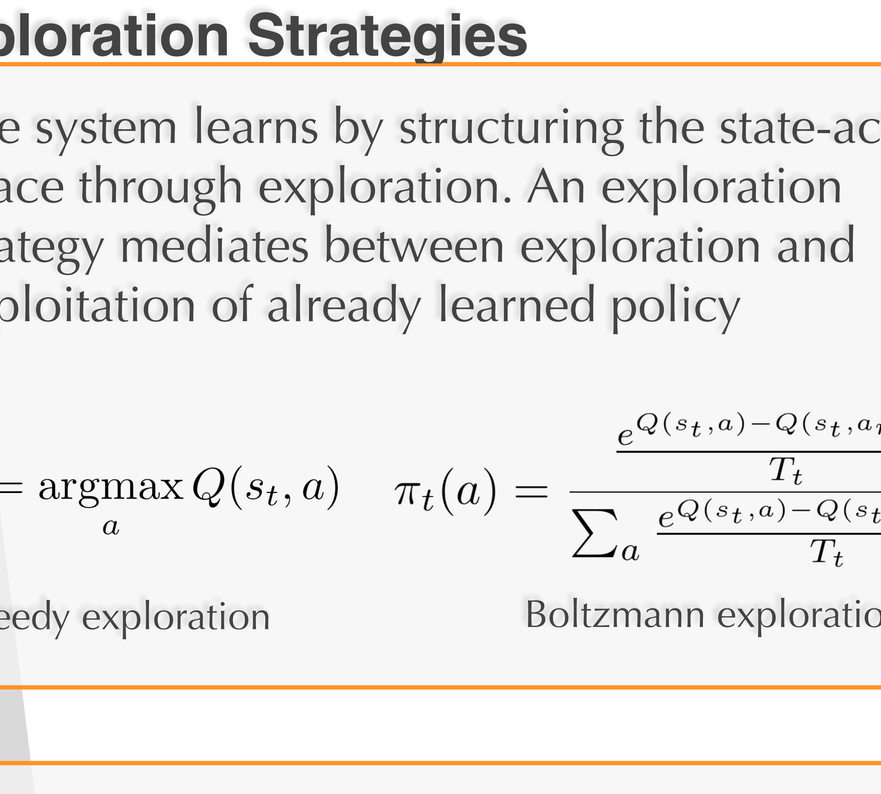
$$\pi_t = \underset{a}{\operatorname{argmax}}\, Q(s_t, a) \qquad \pi_t(a) = \frac{\frac{e^{Q(s_t,a)-Q(s_t,a_{max})}}{T_t}}{\sum_a \frac{e^{Q(s_t,a)-Q(s_t,a_{max})}}{T_t}}$$

ϵ-greedy exploration — Boltzmann exploration

Results

Method	$score_{max}$	$score_{avg}$	σ_{avg}
Goal derived reward	47	38	4.0
Goal derived + tutoring	69	53	6.8
Goal derived + tutoring + penalty	65	59	3.16
ϵ-greedy exploration	46	39	3.5
Boltzmann exploration	65	59	3.16
No regularization	60	53	7.4
L_1 Regularization	65	59	3.16
Dropout (ρ=.5)	10	5	1.8

Evaluation of method combinations with their score (percentage of time-steps the object was moving) and standard deviations across evaluation runs. Bold results represent the same configuration



Evaluation of the systems score after every training episode with all rewards in combination with Boltzmann exploration and L_1 regularization. Shaded area corresponds to standard deviation