Abstractions with MDP homomorphisms enable **transfer** of skills.

Online Abstraction with MDP Homomorphisms for Deep Learning

Ondrej Biza and Robert Platt

Objective

We aim to learn abstractions for simple robotic manipulation tasks, such as the puck stacking task depicted in Figure 1, bottom. Abstractions can be create through *partitioning*. For instance, the puck stacking abstraction (Figure 1, top) partitions the continuous state space (Figure 1, bottom) into three blocks: "no stack, hand empty", "no stack, puck in hand" and "stack of 2, hand empty". However, state partition alone does not help as much because we still have thousands of actions for each state.

Hence, our abstractions should partition both the state and action spaces; they should also be learned *online* (i.e. from a stream of experience).

Methods

Our algorithm is based on the theory of Markov Decision Process (MDP) homomorphisms [1]. The main idea is to iteratively split the *state-action space* of the task until we arrive at a partition that captures all the important dynamics of the original problem. The splitting is accomplished through a fully convolutional network (Figure 1, left) that predicts the *outcome* of each action. The network is iteratively re-trained each time we split new state-action blocks.

The resulting state-action partition can be converted into a new, abstract, Markov Decision Process (e.g. Figure 1, top). We can then plan the optimal policy in the abstract MDP and map it to the original MDP.

Experiments

We test the learned abstractions in a pucks world domain (Figure 2). The abstractions are create for an initial task and then transferred to a new task. We measure the speed-up in learning given the abstractions (Table 1). The **Options** agent has an option (a temporally extended action, [2]) for reaching each *abstract state* from the original task (e.g. the three abstract states in Figure 1, top). It can execute these options at the start of the episode to reach a desirable state in the environment; the values of options are learned. **Baseline** is a deep Q-network and **baseline, weight sharing** is a deep Q-network initialized with the weights learned in the initial task.

References

[1] Balaraman Ravindran. 2004. An Algebraic Approach to Abstraction in Reinforcement Learning. Ph.D. Dissertation.
[2] Richard S. Sutton, Doina Precup, and Satinder Singh. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning.

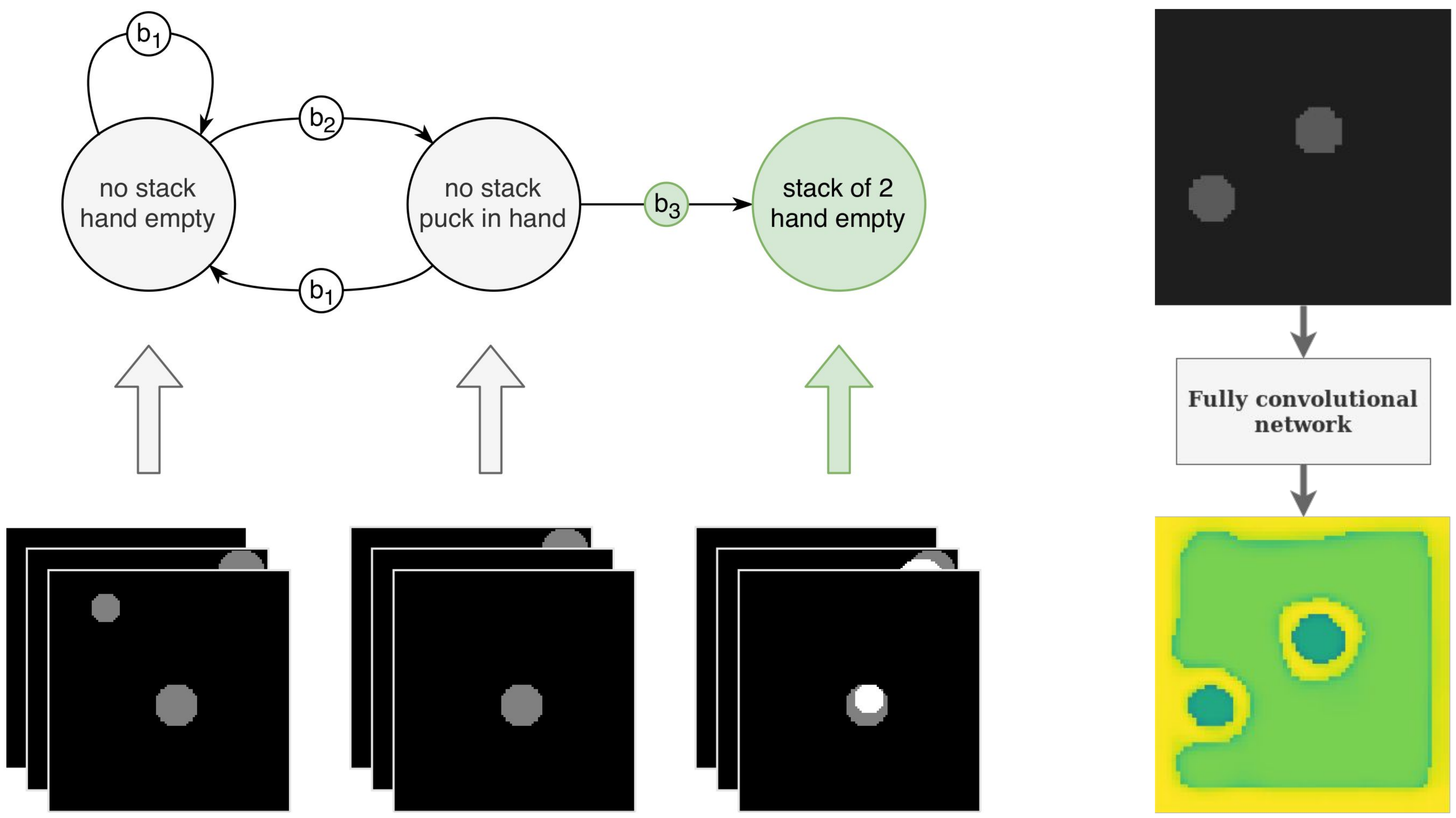

Figure 1: Abstraction of state and action space (left), fully convolutional network for state-block prediction (right).

Task	Options	Baseline	Baseline, share weights
2 puck stack to 3 puck stack	$\mathbf{2558 \pm 910}$	5335 ± 1540	10174 ± 5855
3 puck stack to 2 and 2 puck stack	$\mathbf{2382 \pm 432}$	-	3512 ± 518
2 puck stack to stairs from 3 pucks	$\mathbf{2444 \pm 487}$	4061 ± 1382	4958 ± 3514
3 puck stack to stairs from 3 pucks	$\mathbf{1952 \pm 606}$	4061 ± 1382	5303 ± 3609
2 puck stack to 3 puck component	2781 ± 605	3394 ± 999	6641 ± 5582
stairs from 3 pucks to 3 puck stack	3947 ± 873	5335 ± 1540	6563 ± 4299
stairs from 3 pucks to 2 and 2 puck stacks	5552 ± 3778	-	5008 ± 1998
stairs from 3 pucks to 3 puck component	3996 ± 2693	3394 ± 999	4856 ± 3600
3 puck component to 3 puck stack	$\mathbf{3729 \pm 742}$	5335 ± 1540	8540 ± 4908
3 puck component to stairs from 3 pucks	3310 ± 627	4061 ± 1382	2918 ± 328

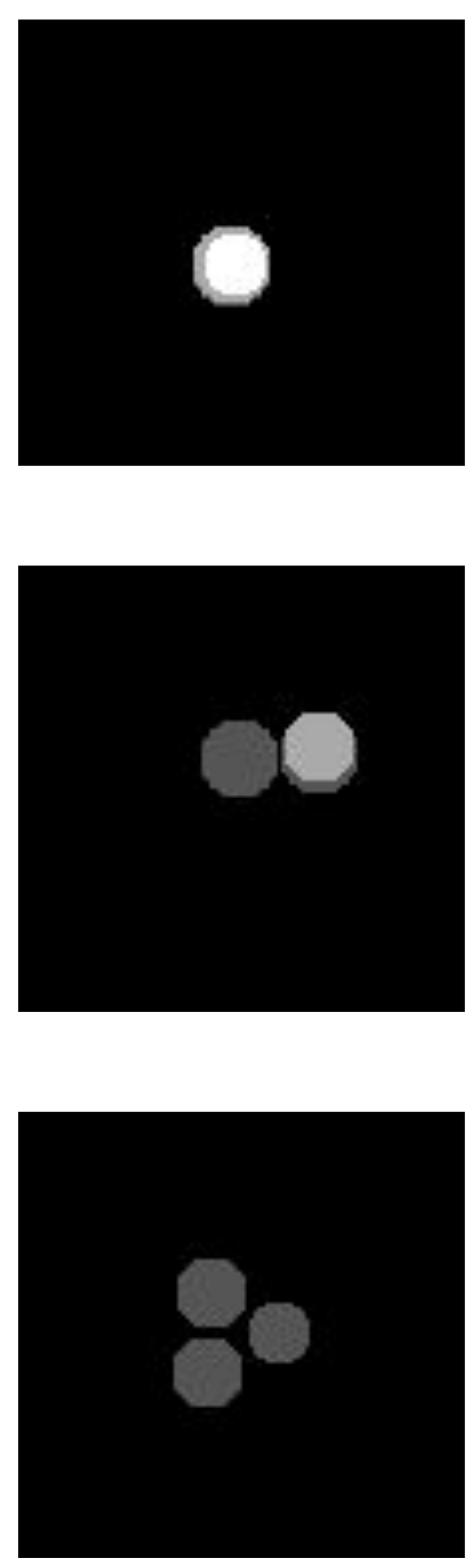

Table 1: Abstraction transfer experiments in the pucks world domain; we measure the number of episodes required to learn the new task. **Figure 2**: Goal states of stacking pucks, building stairs and arranging a connected component.

CTU CZECH TECHNICAL UNIVERSITY IN PRAGUE

Khoury College

This research was partially supported by grant no. GA18-18080S of the Grant Agency of the Czech Republic, grant no. EF15_003/0000421 of the Ministry of Education, Youth and Sports of the Czech Republic and by the National Science Foundation through IIS-1427081, IIS-1724191, and IIS-1724257, NASA through NNX16AC48A and NNX13AQ85G, ONR through N000141410047, Amazon through an ARA to Platt, and Google through a FRA to Platt.

Take a picture to **download** the **full paper**