Automatically Mapped Transfer Between Reinforcement Learning Tasks via Three-Way Restricted Boltzmann Machines[1]

Haitham Bou Ammar[a] Decebal Constantin Mocanu[a] Matthew E. Taylor[b]
Kurt Driessens[a] Gerhard Weiss[a] Karl Tuyls[a]

[a] *Maastricht University, Department of Knowledge Engineering, Netherlands*
[b] *School of Electrical Engineering and Computer Science, Washington State University, USA*

Reinforcement learning (RL) has become a popular framework for autonomous behaviour generation from limited feedback [2, 3], but RL methods typically learn *tabula rasa*. Transfer learning (TL) aims to improve learning by providing informative knowledge from a previous (source) task or tasks to a learning agent in a novel (target) task. If the agent is to be fully autonomous, it must: (1) automatically select a source task, (2) learn how the source task and target tasks are related, and (3) effectively use transferred knowledge when in the target task. While fully autonomous transfer is not yet possible, this paper advances the state of the art by focusing on part (2). In particular, this work proposes methods to automatically learn the relationships between pairs of tasks and then use this learned relationship to transfer effective knowledge.

In TL for RL, the source task and target task may differ in their formulations. In particular, when the source task and target task have different state and/or action spaces, an *inter-task mapping* [4] that describes the relationship between the two tasks is needed. While there have been attempts to discover this mapping automatically, finding an optimal way to construct this mapping is still an open question. Existing techniques either rely on restrictive assumptions made about the relationship between the source and target tasks or adopt heuristics that work only in specific cases.

This paper introduces an autonomous framework for learning inter-task mappings based on *restricted Boltzmann machines* (RBMs) [1]. RBMs provide a powerful but general framework that can be used to describe an abstract common space for different tasks. This common space is then used in turn to represent the inter-task mapping between the two tasks and to transfer knowledge about transition dynamics between the two tasks.

The contributions of this paper are summarised as follows. First, a novel RBM is proposed that uses a three-way weight tensor (i.e., TrRBM). Since this machine has a computational complexity of $\mathcal{O}(N^3)$, a factored version (i.e., FTrRBM) is then derived that reduces the complexity to $\mathcal{O}(N)$. Learning in this factored version can not be done with vanilla Contrastive Divergence (CD). The main reason is that if CD divergence was used as is, FTrRBM will learn to correlate random samples from the source task to random samples in the target. To tackle this problem, as well as ensure computational efficiency, a modified version of CD is proposed. In Parallel Contrastive Divergence (PCD), the data sets are first split into batches of samples. Parallel Markov chains run to a certain number of steps on each batch. At each step of the chain, the values of the derivatives are calculated and averaged to perform a learning step. This runs for a certain number of epochs. At the second iteration the same procedure is followed but with randomised samples in each of the batches.

[1]This paper was published in full in the Proceedings of the European Conference on Machine Learning 2013, pp 449–464.

After learning, the FTrRBM encodes an inter-task mapping from the source to the target task. This encoding is then used to transfer (near-)optimal sample transitions from the source task, forming sample transitions in the target task. Given a (near-)optimal source task policy, $\pi_S^\star$, the source task is sampled greedily according to $\pi_S^\star$ to acquire optimal state transitions. The triplets are passed through the visible source layer of FTrRBM and are used to reconstruct initial target task samples at the visible target layer, effectively transferring samples from one task to another. If the source and target task are close enough then the transferred transitions are expected to aid the target agent in learning an (near-)optimal behaviour. They are then used in a sample based RL algorithm, such as LSPI to learn an optimal behaviour in the target task (i.e., $\pi_T^\star$).

Experiments showing that the proposed method is capable of successfully learning a useful inter-task mapping were conducted using the standard RL tasks of inverted pendulum (shallow transfer) and mountain car (deep transfer) as source tasks and cart-pole as the target task. Specifically, the results demonstrate that FTrRBM is capable of:

1. Automatically learning an inter-task mapping between different MDPs.
2. Transferring informative samples that reduce the computational complexity of a sample-based RL algorithm.
3. Transferring informative instances which reduce the time needed for a sample-based RL algorithm to converge to a near-optimal behaviour.

Although successful, the approach is not guaranteed to provide useful transfer. To clarify, the reward was not included in the definition of the inter-task mapping, but when transferring near-optimal behaviours sampled according to near-optimal policies such rewards are implicitly taken into account and thus, attaining successful transfer results. However this means that negative transfer may occur if the rewards of the source task and target tasks were highly dissimilar. A solution to this potential problem is left for future work, but will likely require incorporating the sampled reward into the current approach. A second potential problem may occur during the learning phase of FTrRBM, which could be traced back to quality of the random samples. If the number of provided samples is low and are very sparse — areas of the state space are not sufficiently visited — the learned mapping may be uninformative. The solution of this problem is also left for future work, but could possibly be tackled by using a deep belief network to increase the level of abstraction.

References

[1] H. Ackley, E. Hinton, and J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, pages 147–169, 1985.

[2] L. Buşoniu, R. Babuška, B. De Schutter, and D. Ernst. *Reinforcement Learning and Dynamic Programming Using Function Approximators*. CRC Press, Boca Raton, Florida, 2010.

[3] Richard S. Sutton and Andrew G. Barto. Reinforcement learning: An introduction, 1998.

[4] Matthew E. Taylor, Peter Stone, and Yaxin Liu. Transfer learning via inter-task mappings for temporal difference learning. *Journal of Machine Learning Research*, 8(1):2125–2167, 2007.