Iterative Model-based Transfer in Deep Reinforcement Learning

Jelmer L. A. Neeven
Thesis supervisor: Dr. Kurt Driessens

Maastricht University
Department of Data Science and Knowledge Engineering
Maastricht, the Netherlands

Keywords: Transfer Learning · Deep Reinforcement Learning · Model-based Reinforcement Learning · Continual Learning · Life-long Learning

In recent years, advances in the field of Deep Reinforcement Learning (DRL) have enabled artificial agents to obtain superhuman performance on various tasks, given enough interactions with the environment [7,4,12]. While the state of the art in DRL keeps improving rapidly, most algorithms result in agents that generalize badly, performing well only on the single task they were trained on [9]. Simultaneously, while most existing DRL transfer learning literature considers *model-free* RL algorithms [15,10,11,9], its counterpart *model-based* RL, in which agents explicitly model their environment rather than directly predicting state-action values or policies, has shown great successes, especially in recent years [3,2,8,1,5]. Despite these successes, however, the feasibility of transfer learning with these approaches remains relatively under-explored. As it follows from intuition that representing multiple environments in a single model may help express their similarities and differences and may therefore benefit transfer learning, this thesis explicitly investigates the feasibility of model-based DRL as a basis for (life-long) transfer learning.

As a starting point for this investigation, a state-of-the-art model-based approach [2] is extended to an iterative version that continually alternates between a model training and data collection phase, dubbed *Iterative World Models* (IWM). It first collects a very limited set of observations using a random policy, and then supervisedly trains a VAE (to compress the observations) and LSTM (to predict the next observation given the current compressed observation and chosen action). The predictions of these *World Models* are then used to train a deep Q-Network [7] to choose the appropriate action. The updated models are used to collect more on-policy observations from the environment, after which the cycle is repeated.

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-2491/abstract6.pdf

This IWM agent is trained sequentially on different variations of the *Catcher* [13] environment, each varying the angular direction of the falling fruit, each time re-training the weights obtained from the previous environment.
Following [14], five different metrics are reported, most importantly the *maximum obtained reward*, the *cumulative reward* over each experiment and the *time to threshold*, measuring the average number of steps required for the agent to reach respectively 75% and 95% of the maximum possible reward. As a baseline for comparison, a separate IWM agent is trained from scratch on each environment individually.

As expected, re-training the existing agent models on a new environment facilitates transfer, since the environment dynamics largely remain similar. In particular, an average increase in cumulative reward of 136% is observed relative to training an agent from scratch on each environment, and an average reduction of 72% is observed for the *time to threshold* (i.e. a 72% increase in sample efficiency). No transfer is observed in terms of maximum obtained reward, as both the agent trained from scratch and the sequentially trained agent eventually obtain the maximal reward.
However, if the agent is then again re-trained on a different variation of the environment with both visual differences and an inverted reward structure (where the player has to dodge the fruit rather than catching it), no significant transfer is observed at all. Investigations indicate that the necessary re-training of the VAE effectively destroys all previous knowledge. To overcome this, an extension to the algorithm is introduced, *Iterative World Models with Persistent Memory* (IWM-PM), which trains the agent not only on the current environment, but also on its "memories" of all previous environments combined. Experiments show this extension indeed has the desired effect, substantially decreasing the amount of change in VAE encodings after training on a new environment. Subsequently, significant transfer is then observed for this new environment, albeit negative with a 36% decrease in cumulative reward and 50% increase in *time to threshold*, which is not surprising given the inverted reward structure of this new environment relative to all previous environments.
Additionally, on the environments with different angles, this extension further increases cumulative reward by 158% and reduces *time to threshold* by 88% on average compared to an agent trained from scratch.

In conclusion, because the proposed IWM-PM algorithm can be considered a combination of several state-of-the-art model-based algorithms [2,3,6,5], the conducted experiments show that model-based DRL indeed has strong potential for transfer learning. While the experiments conducted for this thesis were limited and additional research is required before any strong conclusions can be drawn, recent work on model-based DRL gives no reason to believe that these results cannot be extended to more complex environments. Additionally, model-based transfer learning approaches may have several preferable properties to prominent model-free transfer algorithms such as Actor-Mimic and Progressive Neural Networks [15,10,9,11].

References

1. Gu, S., Lillicrap, T., Sutskever, I., Levine, S.: Continuous deep Q-learning with model-based acceleration. In: International Conference on Machine Learning. pp. 2829–2838 (2016)
2. Ha, D., Schmidhuber, J.: Recurrent world models facilitate policy evolution. In: Advances in Neural Information Processing Systems 31, pp. 2450–2462 (2018)
3. Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., Davidson, J.: Learning latent dynamics for planning from pixels (2019), arXiv preprint, http://arxiv.org/abs/1811.04551
4. Hessel, M., Modayil, J., Van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., Horgan, D., Piot, B., Azar, M., Silver, D.: Rainbow: Combining improvements in deep reinforcement learning. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
5. Kaiser, L., Babaeizadeh, M., Milos, P., Osinski, B., Campbell, R.H., Czechowski, K., Erhan, D., Finn, C., Kozakowski, P., Levine, S., et al.: Model-based reinforcement learning for atari (2019), arXiv preprint, http://arxiv.org/abs/1903.00374
6. Ketz, N., Kolouri, S., Pilly, P.: Continual learning using world models for pseudo-rehearsal (2019), arXiv preprint, http://arxiv.org/abs/1903.02647
7. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529 (2015)
8. Nagabandi, A., Kahn, G., Fearing, R.S., Levine, S.: Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 7559–7566 (2018)
9. Parisotto, E., Ba, J.L., Salakhutdinov, R.: Actor-mimic: Deep multitask and transfer reinforcement learning (2015), arXiv preprint, http://arxiv.org/abs/1511.06342
10. Rusu, A.A., Colmenarejo, S.G., Gulcehre, C., Desjardins, G., Kirkpatrick, J., Pascanu, R., Mnih, V., Kavukcuoglu, K., Hadsell, R.: Policy distillation (2015), arXiv preprint, http://arxiv.org/abs/1511.06295
11. Rusu, A.A., Rabinowitz, N.C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., Hadsell, R.: Progressive neural networks (2016), arXiv preprint, http://arxiv.org/abs/1606.04671
12. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms (2017), arXiv preprint, http://arxiv.org/abs/1707.06347
13. Tasfi, N.: Pygame learning environment. https://github.com/ntasfi/PyGame-Learning-Environment (2016)
14. Taylor, M.E., Stone, P.: Transfer learning for reinforcement learning domains: A survey. Journal of Machine Learning Research **10**, 1633–1685 (2009)
15. Teh, Y., Bapst, V., Czarnecki, W.M., Quan, J., Kirkpatrick, J., Hadsell, R., Heess, N., Pascanu, R.: Distral: Robust multitask reinforcement learning. In: Advances in Neural Information Processing Systems 30, pp. 4496–4506 (2017)