Kickstarting Deep Reinforcement Learning*

Simon Schmitt, Jonathan J. Judson, Augustin Zidek et al.

2018

1 What

Method that uses previously learned agent as a teacher, leveraging policy distillation [Parisotto et al., 2015, Rusu et al., 2015] and population-based training [Jaderberg et al., 2017] ideas.

There are no architectural constrains on students or teachers. There is an automatic teacher influence decreasing as training proceeds via loss weight annealing (look at the reward more than at what your teacher tells you). The authors use an IMPALA [Espeholt et al., 2018] agent for all the experiments, so, you won't be able to repeat this at home x_X.

Using DeepMind lab as an environment, the authors show that:

- kickstarting with a single teacher leads to 1.5x speedup over an agent trained from scratch
- student can outperform its teacher
- kickstarting with multiple task-specific teachers leads to 9.5x speedup and surpassing the performance by 42.2%.

2 Why

RL is super data inefficient. Hence, it would be great if we can leverage other agent experience to make learning faster in terms of environment interaction. The paper focuses on a setting when some pre-trained agents are readily available.

3 How

Usual policy distillation objective:

$$l_{\text{distill}}(\omega, x, t) = \mathrm{H}(\pi_T(a|x_t) \| \pi_S(a|x_t, \omega)), \tag{1}$$

*Notes by Vitaly Kurin `https://yobibyte.github.io/`

where $\mathrm{H}(\cdot\|\cdot)$ is the cross-entropy.

The authors don't want a learner blindly follow the expert, they want it to use it merely as a proxy to maximise the reward:

$$l^k_{\text{kick}}(\omega, x, t) = \ell_{\text{RL}}(\omega, x, t) + \lambda_k \mathrm{H}(\pi_T(a|x_t)\|\pi_S(a|x_t, \omega)), \tag{2}$$

The authors look at the auxiliary loss from the perspective of entropy regularisation in A3C [Mnih et al., 2016]: minimisation of negative entropy is equivalent to minimising $\mathrm{D}_{\text{KL}}(\pi_S(a|x_t, \omega)\|U)$, where U is a uniform distribution over actions. Same for $l^k_{\text{kick}}(\omega, x, t)$, it is equivalent to the $KL(teacher||student)$. (*Don't fully get it. Isn't entropy regularisation for encouraging exploration when the latter is just to be close to an expert?*)

Using the same notation, let's rewrite A3C loss and add cross entropy to it:

$$\begin{aligned}\ell_{\text{A3C}}(\omega, x, t) = &\log \pi_S(a_t|x_t, \omega)(r_t + \gamma v_{t+1} - V(x_t|\theta)) \\ &- \beta \mathrm{H}(\pi_S(a|x_t, \omega) + \lambda_k \mathrm{H}(\pi_T(a|x_t)\|\pi_S(a|x_t, \omega)).)\end{aligned}$$

IMPALA's V-trace needs off-policyness correction:

$$\rho_t \nabla_\omega \log \pi_S(a_t|x_t, \omega)\big(r_t + \gamma v_{t+1} - V(x_t|\theta)\big). \tag{3}$$

PBD is used in a way so that each agent gets other agent's params and checks whether they are significantly (*what's the metric?*) better than their own. If this is true, it adopts the weights. Moreover, the params are slightly altered for exploration.

4 Evaluation

Evaluation is done on DMLab-30 (30 stands for 30 tasks).

Still don't get why people use mean score instead of median score to mitigate the following problem: if you are much better at one task, but get no better at the others, usual mean will show that you're super star, but this should not be the case (in Atari I've also seen using median as a metric [1]).

I really like λ weight analysis which provides insights on how the method works (I understand that it should always be taken with a grain of salt, but anyway). And this is just beautiful:

... We find this a wonderful parallel with how the best human educators teach: not telling the student what to think, but simply putting the student in a fruitful position to learn for themselves.

I wish, the next thing would be more clear (may be I'm just missing it): what is the amount of frames the teacher was trained on before they started to use it for kickstarting?

[1] https://yobibyte.github.io/atari-eval.html

5 Comments

- It would be interesting to compare this method to what Reptile [Nichol and Schulman, 2018] calls 'pre-initialisation' of the network

- the authors separate their approach from imitation learning, saying that they do not use pre-recorder data to train, but rather ask 'what would my teacher do if it were in my shoes?'. How does this compare with Dagger or DART?

- It would be interesting to see how learner depends on quality of the teacher, what if teacher is really bad[2]

- Can we learn the same task from two teachers?

- An approach looks similar to DQfD [Hester et al., 2018] to me (merging Q-learning a Behaviour Cloning via using expert policy as a proxy for Q), however the authors do not mention it at all.

- Regarding the previous point, in [Hester et al., 2018] the authors explain why cross-entropy is bad in learning from demonstration: *This difference is likely because the cross-entropy loss is less compatible with the Q-learning loss as it pushes the action values as far apart as possible.* (for some reason, this paragraph exists only in v1 version of Arxiv paper). How does kickstarting deals with it? The only thing if found is: *This ensures a dense learning signal, and does not have to be fully Kickstarting Deep Reinforcement Learning aligned with the RL objective. Then through adaptation of k during the course of learning, the agent is able to shift its optimization focus on the (potentially sparse) reward sig- nal rt, similar to how continuation methods relax optimisa- tion problems to make finding the solution easier.*

References

[Espeholt et al., 2018] Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., et al. (2018). Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. *arXiv preprint arXiv:1802.01561.*

[Hester et al., 2018] Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., Horgan, D., Quan, J., Sendonaris, A., Dulac-Arnold, G., et al. (2018). Deep q-learning from demonstrations. In *Proceedings of the Conference on Artificial Intelligence (AAAI).*

[2] Found in the paper: *In fact, we find the student 'ignores' the (incompentent) 1-bot expert...*

[Jaderberg et al., 2017] Jaderberg, M., Dalibard, V., Osindero, S., Czarnecki, W. M., Donahue, J., Razavi, A., Vinyals, O., Green, T., Dunning, I., Simonyan, K., et al. (2017). Population based training of neural networks. *arXiv preprint arXiv:1711.09846*.

[Mnih et al., 2016] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pages 1928–1937.

[Nichol and Schulman, 2018] Nichol, A. and Schulman, J. (2018). Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*.

[Parisotto et al., 2015] Parisotto, E., Ba, J. L., and Salakhutdinov, R. (2015). Actor-mimic: Deep multitask and transfer reinforcement learning. *arXiv preprint arXiv:1511.06342*.

[Rusu et al., 2015] Rusu, A. A., Colmenarejo, S. G., Gulcehre, C., Desjardins, G., Kirkpatrick, J., Pascanu, R., Mnih, V., Kavukcuoglu, K., and Hadsell, R. (2015). Policy distillation. *arXiv preprint arXiv:1511.06295*.