Expert-augmented actor-critic
for Vizdoom and Montezuma's Revenge

Michał Garmulewicz[1,2], Henryk Michalewski[1,3], Piotr Miłoś[1,3,4]

University of Warsaw[1], Brain Corp.[2], deepsense.io[3], IMPAN[4]

mgarmulewicz@mimuw.edu.pl
https://github.com/ghostFaceKillah/expert

TL;DR;

Problem Montezuma's Revenge and Vizdoom navigation are too difficult for vanilla reinforcement learning without curiosity or expert data. Behavioral cloning (surprisingly?) also does not work.
Solution We augment natural gradient actor-critic with expert trajectories to get good performance.

State-of-the-art - Montezuma's Revenge

Approach	Mean score	Max score	Trans. $\times 10^6$	Methods used
ExpAugAC (Garmulewicz et al. [2018])	27,052	804,900	200	Expert loss based on expert data, approx. natural policy gradient
DQfD (Hester et al. [2017])	4,740	-	200	750k batches of expert pretraining, 3 additional loss terms, prioritizing expert data replay
Behavioral cloning (from Hester et al. [2017])	575	-	24	–
Ape-X DQfD (Pohlen et al. [2018])	29,384	-	2,500	Methods from DQfD, temporal consistency loss, transformed Bellman operator
Playing hard YT (Aytar et al. [2018])	41,098	-	1,000	Auxiliary reward encouraging imitation of videos of expert gameplays
Learning MR from a Single Demo (Salimans et al. [2018])	-	74,500	50,000	Decomposing task into a curriculum of shorter subtasks, assumes ability to set env. state
Unifying Count-Based Exploration (Bellemare et al. [2016])	3,439	6,600	100	Exploration-based auxiliary reward based on pseudo-count derived from density model to measure uncertainty, mixing Double Q-learning target with MC return
Count-based exploration (Ostrovski et al. [2017])	3705	3705	150	Advanced neural density model for images used to generate extrinsic reward based on pseudo-count.
Self-imitation learning (Oh et al. [2018])	1,100	2,400	50	Past good experience imitation loss terms based on transitions sampled from a replay buffer.
Exploration by Random Network Distillation (Burda et al. [2018])	11,347	17,500	2,000	Exploration bonus equal to the error of a NN predicting features of the observations given by a fixed random neural network.
Learning to Control Visual Abstractions (Ionescu et al. [2018])	2350	2450	250	Discrete pixel grouping model used to derive geometric intrinsic reward and learn policies to control them.

Table 1: The state of the art for Montezuma's Revenge.

Relation to Self-Imitation Learning

Actor-critic methods have been combined in [1] with prioritized replay of past good trajectories. Authors of this work theoretically justify that policy gradient estimator, in conjunction with the value function estimator, jointly expressed by loss

$$\mathcal{L}^{\text{sil}} = \mathbb{E}_{s,a,R\in\mathcal{D}}\left[-\log\pi_\theta(a|s)(R-V_\theta(s))_+ + \beta^{\text{sil}}||(R-V_\theta(s))_+||^2\right]$$

are related to a lower bound of the optimal Q-function under the entropy-regularized RL formalism. Although in the presented work we do not use this value function estimator, a similar off-policy value function estimator has proved to be beneficial in our new experiments, focused on Pitfall!. This theoretical justification partially explains why we ignore that expert samples are off-policy.

Acknowledgments

We would like to thank the PL-Grid Infrastructure for enabling extensive use of the Prometheus supercomputer, located in the AGH University of Science and Technology in Kraków, Poland.

Expert-augmented ACKTR

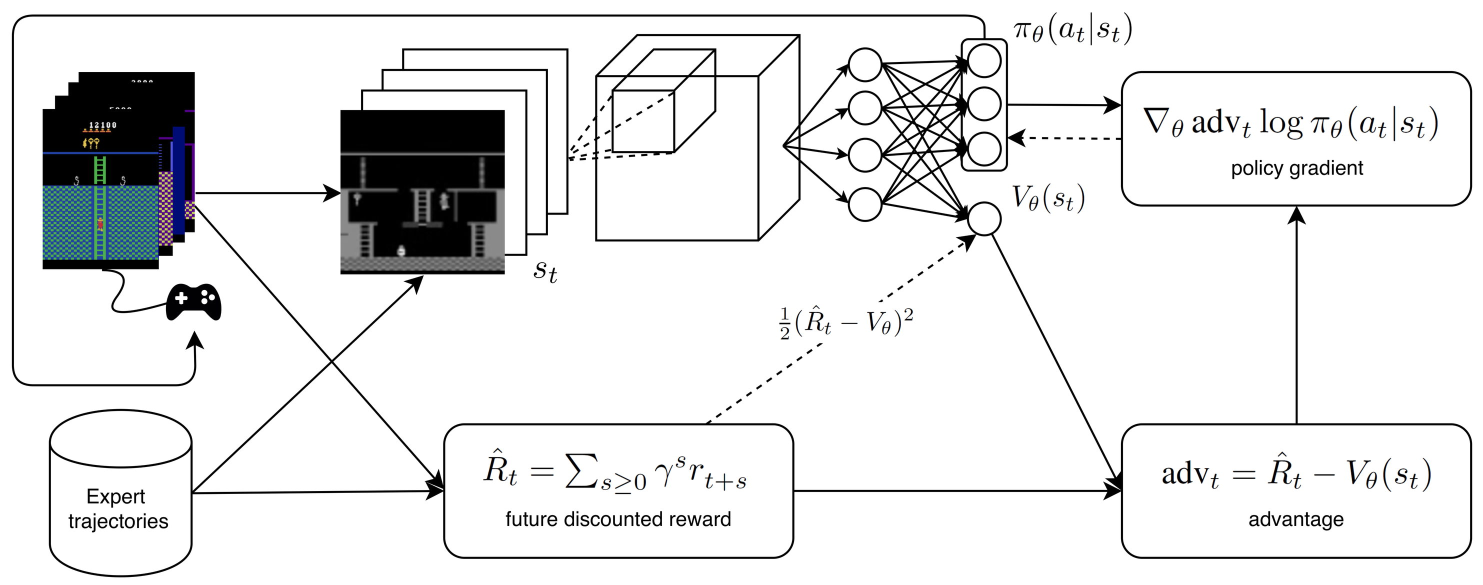

Figure 1: Visual representation of the algorithm.

We introduce a new term $L_t^{\text{expert}}(\theta)$ to the loss function of ACKTR:

$$L_t(\theta) = \underbrace{\mathbb{E}_{\pi_\theta}\left[-\text{adv}_t\log\pi_\theta(a_t|s_t) + \frac{1}{2}(R_t - V_\theta(s_t))^2\right]}_{L_t^{\text{A2C}}(\theta)} \underbrace{- \frac{\lambda}{k}\sum_{i=1}^{k}\text{adv}_i^{\text{expert}}\log\pi_\theta(a_i^{\text{expert}}|s_i^{\text{expert}})}_{L^{\text{exp}}(\theta)}.$$

We consider three variants of the expert advantage:

$$\textit{reward:}\ \text{adv}_t^{\text{expert}} = \sum_{s\geq 0}\gamma^s r_{t+s}^{\text{expert}}\quad \textit{critic:}\ \text{adv}_t^{\text{expert}} = \left[\sum_{s\geq 0}\gamma^s r_{t+s}^{\text{expert}} - V_\theta(s_t)\right]_+\quad \textit{simple:}\ \text{adv}_t^{\text{expert}} = 1.$$

where $[x]_+ = \max(x, 0)$.

Data: Parameter vector θ;
Dataset of expert transitions $(s_t^{\text{expert}}, a_t^{\text{expert}}, s_{t+1}^{\text{expert}}, r_t^{\text{expert}})$
for $iteration \leftarrow 1$ **to** *max steps* **do**
  for $t \leftarrow 1$ **to** T **do**
    Perform action a_t according to $\pi_\theta(a|s_t)$
    Receive reward r_t and new state s_{t+1}
  end
  for $t \leftarrow 1$ **to** T **do**
    Compute discounted future reward: $\hat{R}_t = r_t + \gamma r_{t+1} + \ldots + \gamma^{T-t+1}r_{T-1} + \gamma^{T-t}V_\theta(s_t)$
    Compute advantage: $\text{adv}_t = \hat{R}_t - V_\theta(s_t)$
  end
  Compute A2C loss gradient $g_{A2C} = \nabla_\theta\sum_{t=1}^{T}\text{adv}_t\log\pi_\theta(a_t|s_t) + \frac{1}{2}(\hat{R}_t - V_\theta)^2$
  Sample mini batch of k expert state-action pairs
  Compute expert advantage estimate $\text{adv}_t^{\text{expert}}$ for each state-action pair.
  Compute expert loss gradient $g_{\text{expert}} = \nabla_\theta\frac{1}{k}\sum_{i=1}^{k}\text{adv}_i^{\text{expert}}\log\pi_\theta(a_i^{\text{expert}}|s_i^{expert})$
  Update ACKTR inverse Fisher estimate.
  Plug in gradient $g = g_{A2C} + \lambda_{\text{expert}}g_{\text{expert}}$ into ACKTR Kronecker optimizer.
end

Algorithm 1: Expert-augmented ACTKR

Results - Montezuma's Revenge

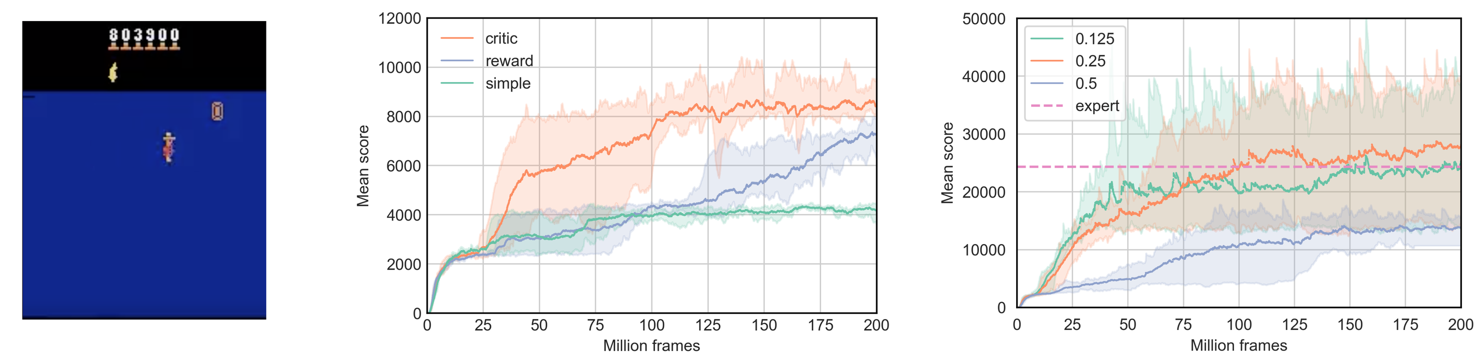

Figure 2: Left: interestingly, our algorithm discovered a bug, which manifests through scores exceeding $800,000$ in some evaluation rollouts (these are excluded when we calculate the mean evaluation score). Center: a performance comparison between various advantage estimators. Right: scores for optimized hyper-parameters, that is we apply $\gamma = 0.995$, the critic advantage estimator and selected values of λ_{expert}.

Results - ViZDoom

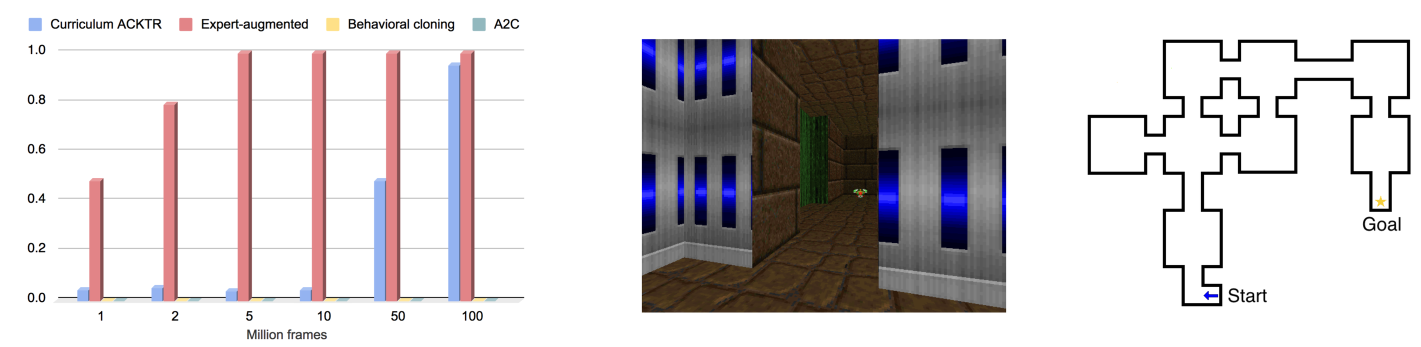

Figure 3: Left: In ViZDoom behavior is similar for all expert advantage estimators. Right: performance of a curriculum learning in MyWayHome compared to our expert-augmented algorithm. The curriculum consists in re-spawning the agent in random locations. We experimentally verified that the ACKTR algorithm without the curriculum was unable to solve the MyWayHome task. This echoes observations made for another actor-critic model-free algorithm in [2]. Our behavioral cloning experiments also failed to solve this task. The curiosity-based method described in [2] achieves an average score 0.7 after 10M of frames.

Conclusions

Based on the experimental results we claim that the algorithm presented in this work is a practical method of getting good performance in cases when multiple interactions with the environment are possible and good quality expert data is available. It could be particularly useful in settings such as Montezuma's Revenge, where neither supervised learning from expert data nor random exploration yield good results.

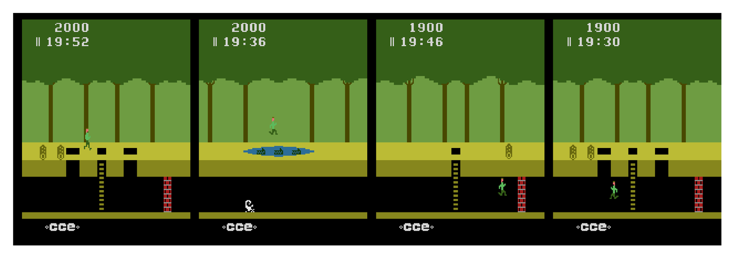

Figure 4: Left: our agent in Pitfall! trained on an expert trajectory which achieves the perfect score. The expert trajectory and our agent run only above the ground. Right: the agent falls into the lower level and is unable to return to the expert trajectory.

References

[1] Junhyuk Oh, Yijie Guo, Satinder Singh, and Honglak Lee. Self-imitation learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 3875–3884, 2018.

[2] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 2778–2787, 2017.