Avoiding Side Effects in Complex Navigation Environments

Aseem Saxena (saxenaa@oregonstate.edu)
Devin Crowley (crowleyd@oregonstate.edu)

Oregon State University

Background

What's the problem?

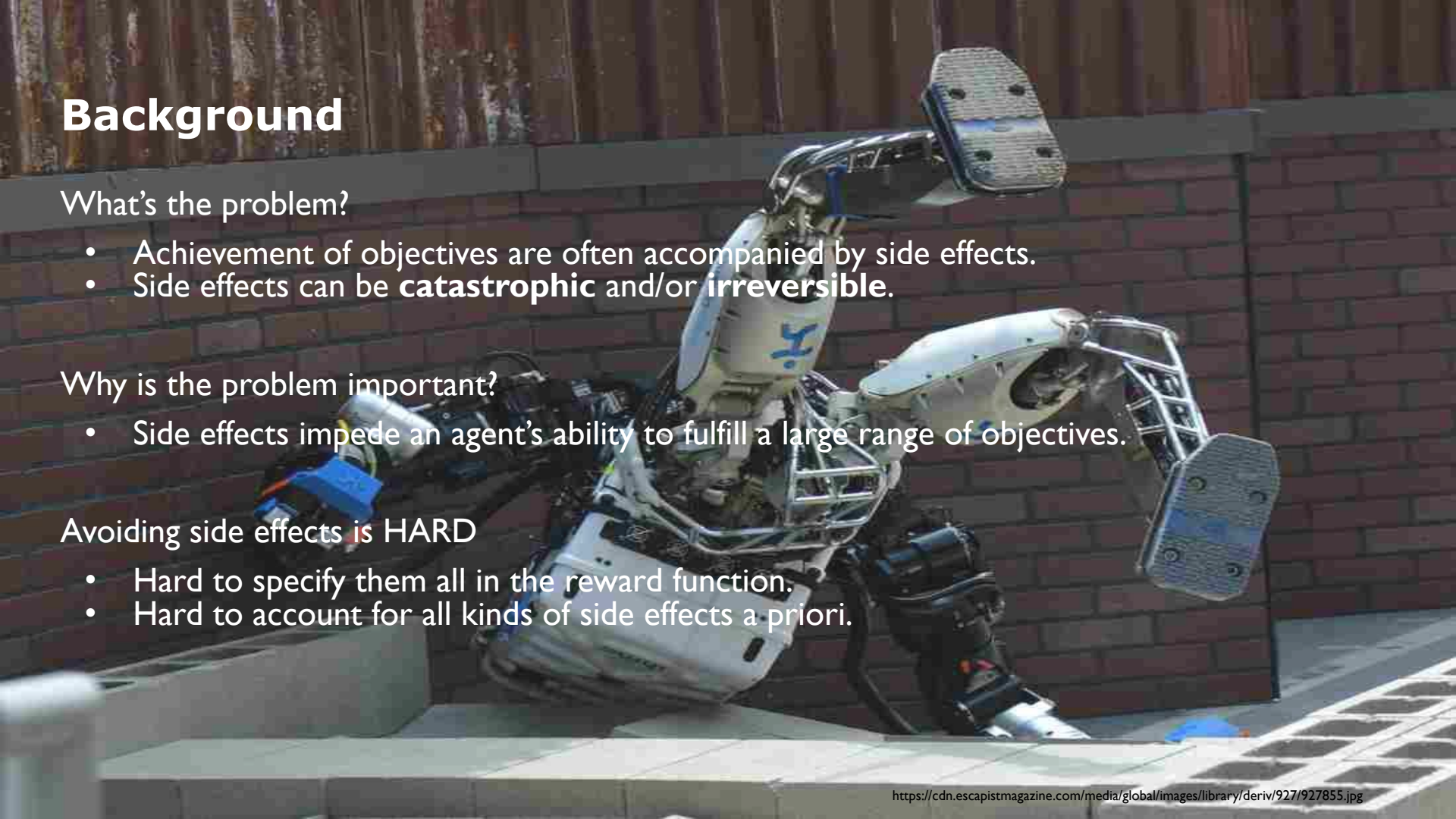
- Achievement of objectives are often accompanied by side effects.
- Side effects can be **catastrophic** and/or **irreversible**.

Why is the problem important?

- Side effects impede an agent's ability to fulfill a large range of objectives.

Avoiding side effects is HARD

- Hard to specify them all in the reward function.
- Hard to account for all kinds of side effects a priori.



https://cdn.escapistmagazine.com/media/global/images/library/deriv/927/927855.jpg

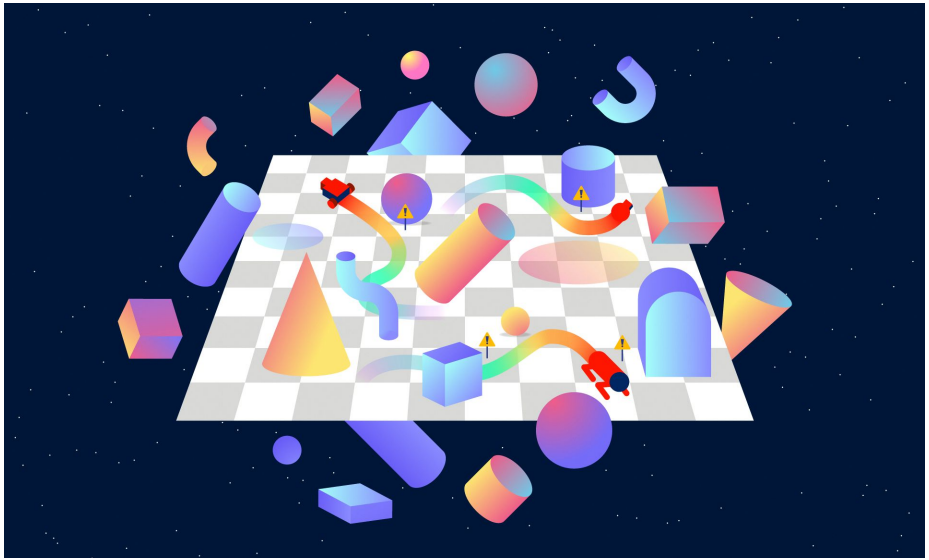
Background

The "True Objective"

- The behavior humans would like our agents to embody.
- Fully specifying this reward function to our agents, very hard.
  - Even then it may be difficult to optimize over.

The solution

- Don't try to encode the True Objective in a reward function.
- Learn to avoid unspecified side effects



https://github.com/openai/safety-gym/blob/master/safety_gym.png

Prior work

- **State reversibility**
  - "Irreversible actions are detrimental to other tasks and are characteristic of negative side effects."[a]
- **Implicit preference in initial state**
  - "When a robot is deployed in an environment that humans act in, the state of the environment is already optimized for what humans want."[b]
- **Conservative Agency**
  - "An approach that balances optimization of the primary reward function with preservation of the ability to optimize (uninformative) auxiliary reward functions."[c]

AUP

AUP[c,d] is an approach that penalizes a reward function based on the ability to optimize uninformative reward functions:

$$R_{AUP}(s,a) := R(s,a) - \lambda * \frac{PENALTY(s,a)}{SCALE(s)}$$

By optimizing over both the agent's objective and one or several unrelated (and uninformative) objectives, the agent learns to prefer actions that lead to minimal side effects without having those side effects specified.

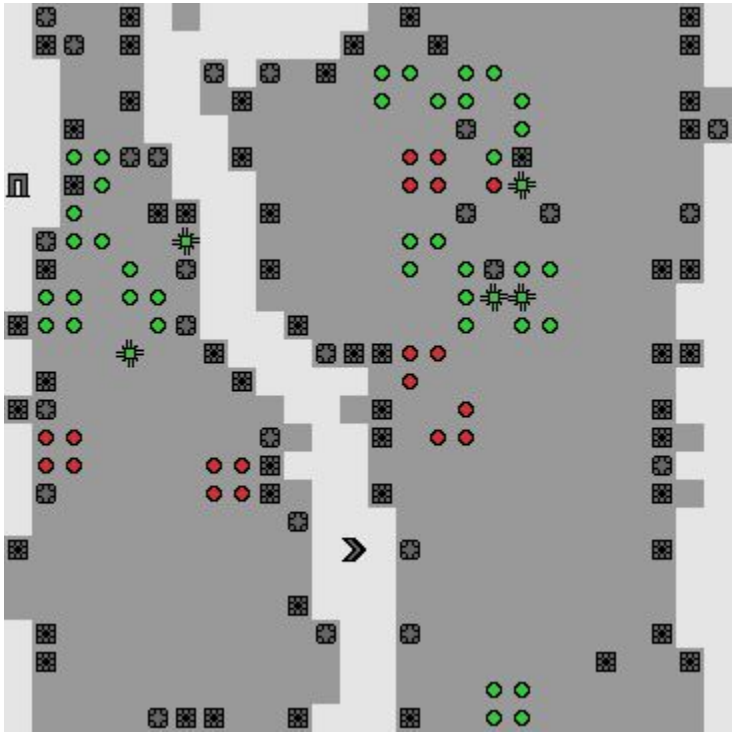
This maximizes the ability to achieve arbitrary objectives.

SafeLife Environment

- Publicly available reinforcement learning environment focusing on safety of RL agents.
- Complex Dynamics following Conway's Game of Life
- Procedurally generated and fixed navigation levels.

Link : https://arxiv.org/pdf/1912.01217.pdf

Code : https://github.com/PartnershipOnAI/safelife



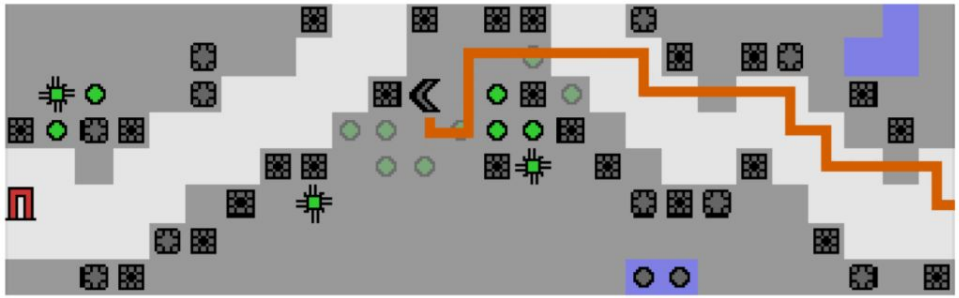
Pruning task: eliminate red, don't disturb green

SafeLife Environment Rules

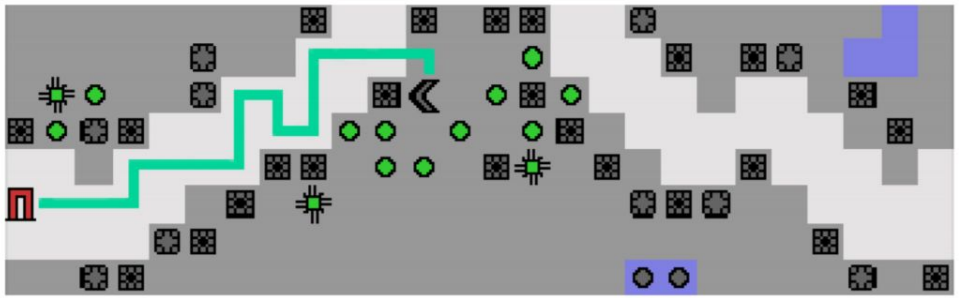
- Rules of Conway's Game of Life.
- 5 actions → go straight, turn clockwise, turn anticlockwise, destroy a green cell, do nothing.
- Reaching the goal state yields a sparse reward of 1.
- Side effect score = Wasserstein metric between the initial and final green cells.

AUP (example)

Figure: https://arxiv.org/pdf/2006.06547.pdf



(a) Baseline trajectory

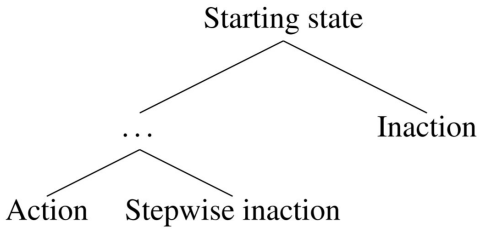


(b) AUP trajectory

(a) The baseline trajectory is trained using only the SafeLife reward.
(b) The AUP trajectory is trained using an additional single random auxiliary reward function.

This is the precise AUP reward function. Note that the severity of the penalty is taken with respect to choosing inaction from the same state.

$$R_{\text{AUP}}(s,a) := R(s,a) - \frac{\lambda}{|\mathcal{R}|} \sum_{R_i \in \mathcal{R}} \left| Q_i^*(s,a) - Q_i^*(s,\varnothing) \right|$$



Our Project

Problem → Navigate to a goal position with the smallest impact on the green dots, without specifying the impact on the green dots in the reward function.

Approaches:

- Baseline DQN
- DQN w/ AUP
- DQN-MT (multi-task)
- DQN-PE (penalty)
- DRQN w/ AUP (memory) [e]

DRQN w/ AUP

DRQN[e] network structure:

- Convolutional layers with ReLU activations
- LSTM layer
- Fully-connected layer → Q-values

We hypothesize that it should confer a greater ability to associate actions to more indirect, far-reaching side effects.

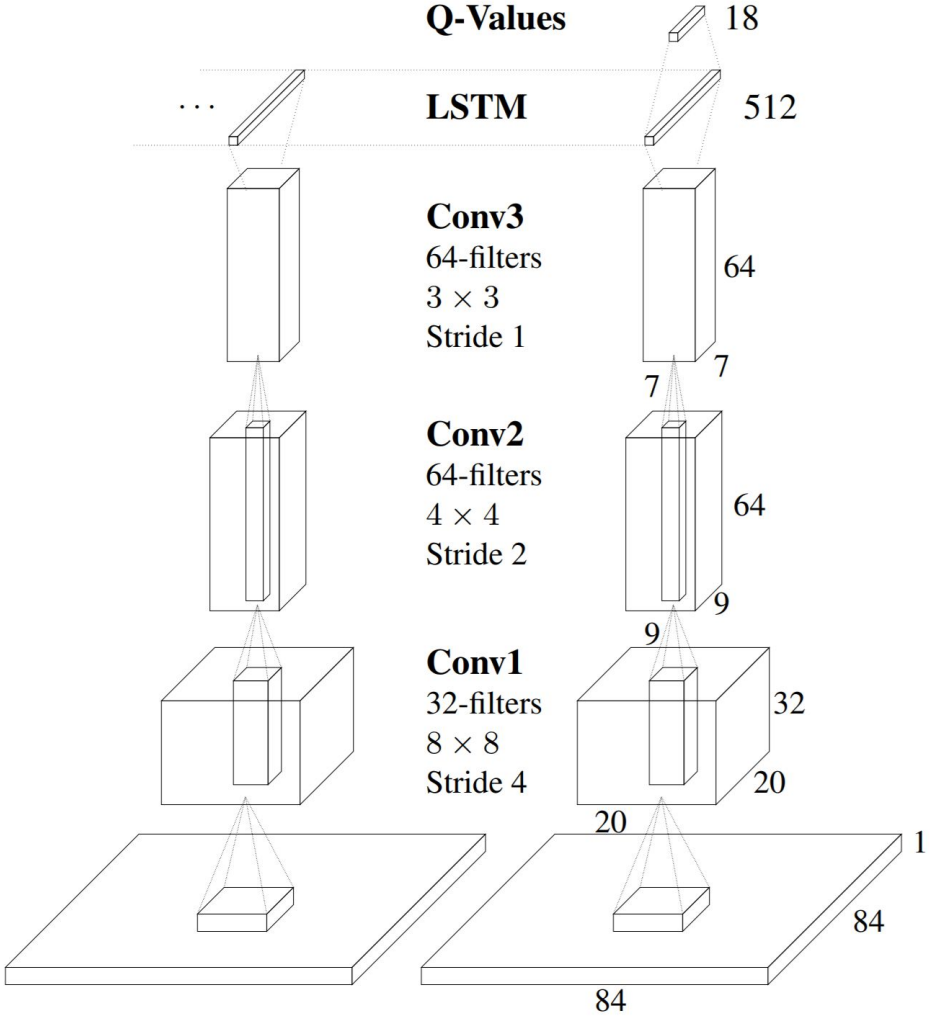


Figure: https://arxiv.org/pdf/1507.06527.pdf

MT-DQN

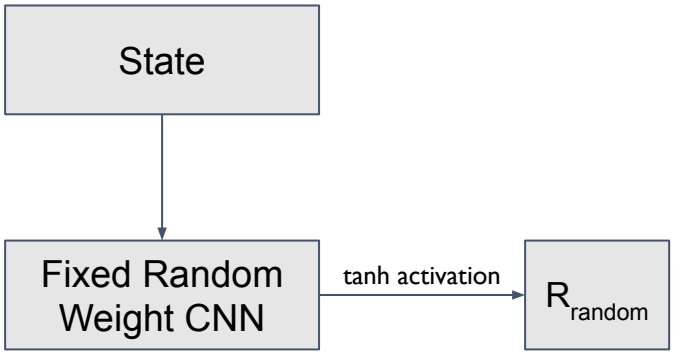
Multi-Task variant of DQN.

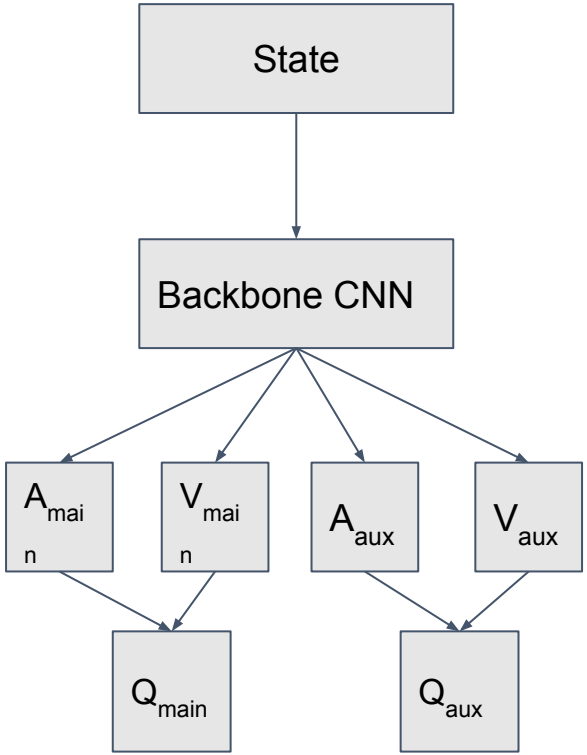
Q = A + V - mean(A) (A⇔ Advantage, V⇔ Value)

Target for main task ⟶ R_{sim} + gamma*max$Q_{main}(s_{next})$

Target for auxiliary tasks ⟶ R_{random} + gamma*max$Q_{aux}(s_{next})$

The random instantaneous come from a randomly initialized network with fixed weights.





MT-DQN Network Architecture

PE-DQN

Builds over MT-DQN, Outputs Q_{main} and AUP penalty term. Learn AUP penalty term directly instead of auxiliary Q functions.

Q = A + V - mean(A) (A⇔ Advantage, V⇔ Value)

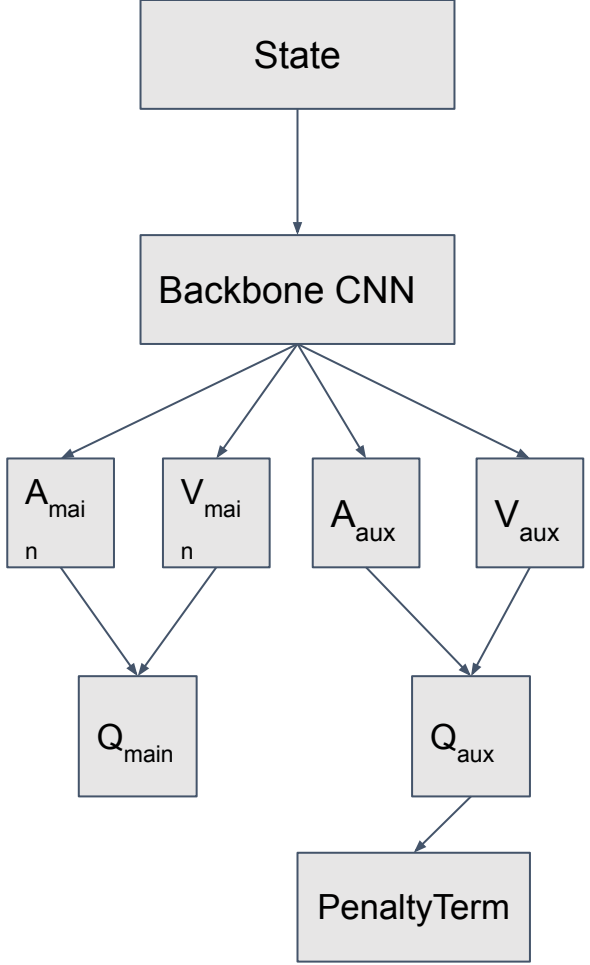
Target for main task → R_{sim} + gamma*max$Q_{main}(s_{next})$

'a' is the epsilon-greedy action wrt Q,main

Penalty term is given by

$$\frac{\sum |Qaux(a) - Qaux(noop)|}{|R|}$$

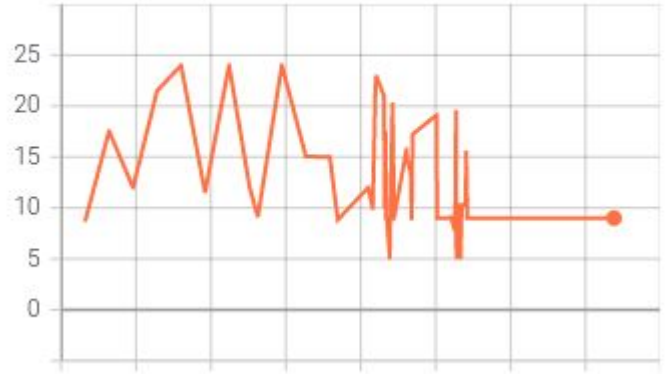
Prediction → Q_{main} - lambda*penaltyTerm



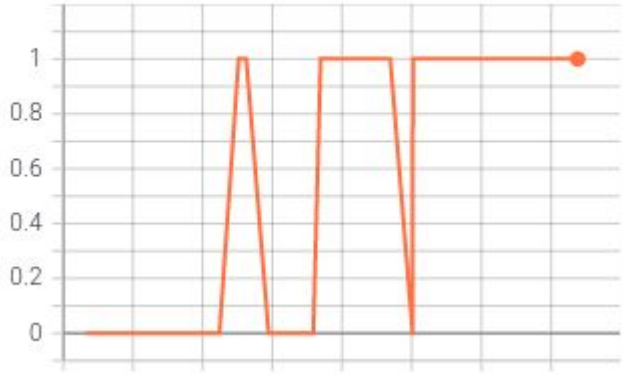
Results - Baseline DQN

Minimum valid Side Effect Score Achieved = 5.165

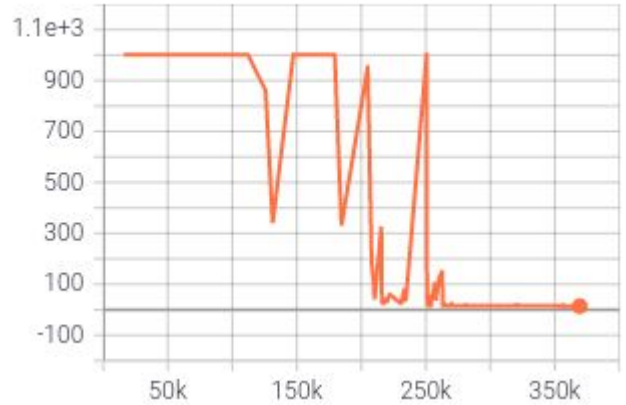
training/side_effect
tag: episodes/training/side_effect



training/reward
tag: episodes/training/reward



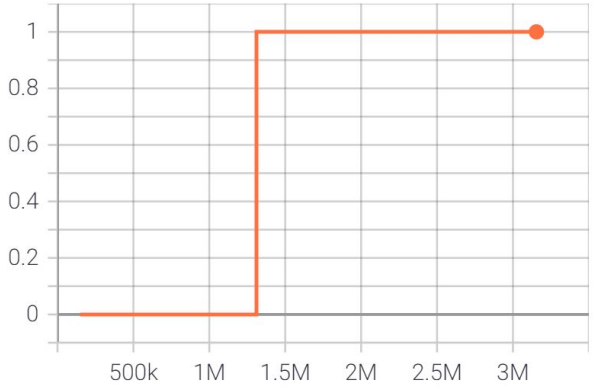
training/length
tag: episodes/training/length



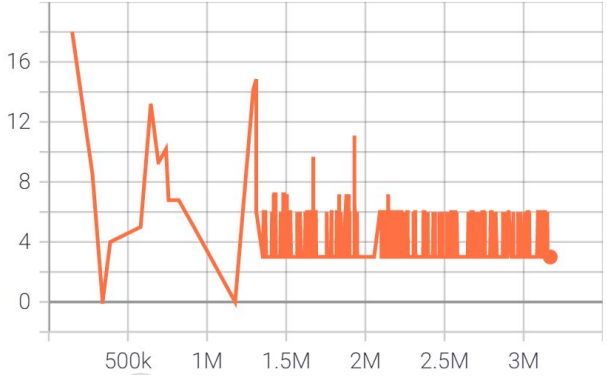
Results - DQN w/ AUP

Minimum valid Side Effect Score Achieved = 3

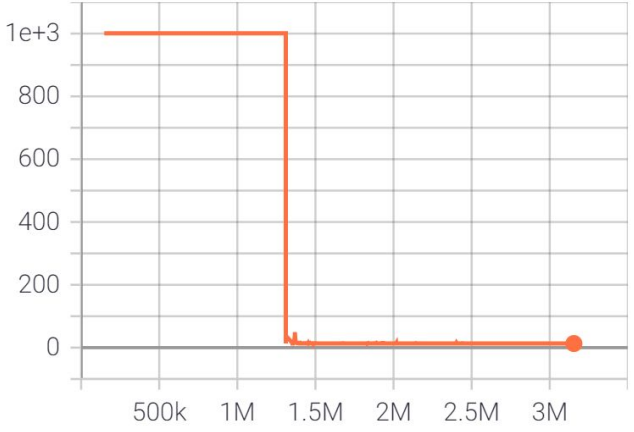
Reward



Side Effects

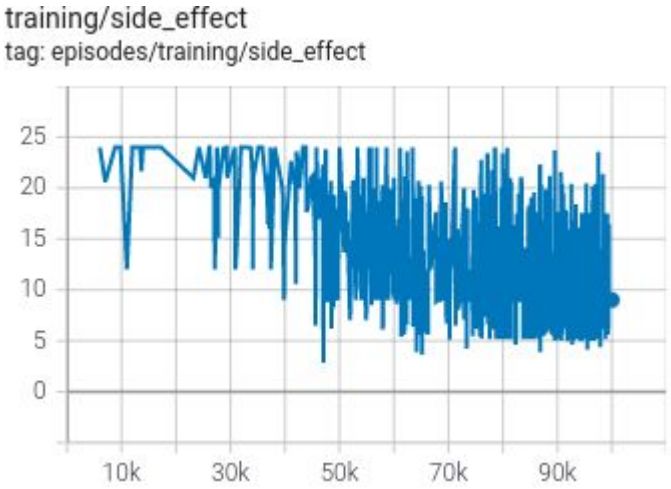


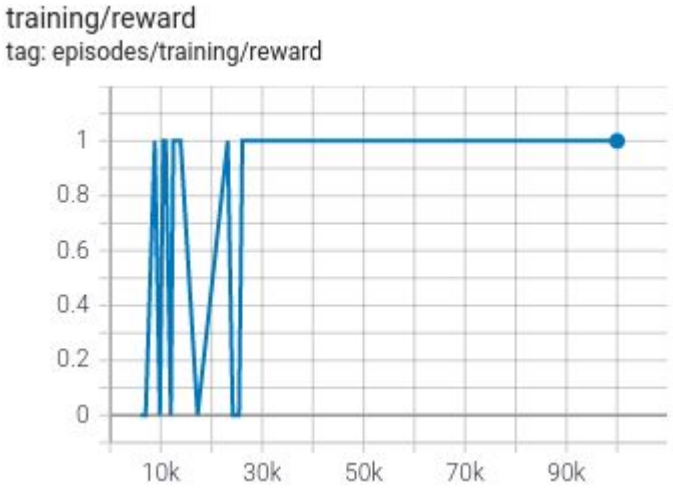
Episode Length

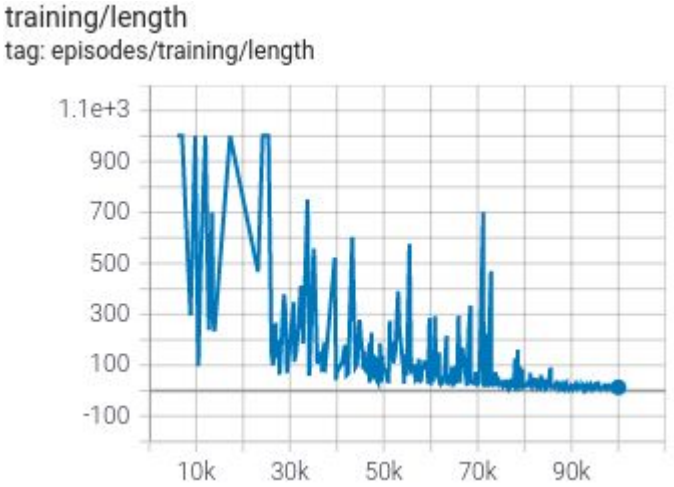


Results - MT-DQN (|R|=1)

Minimum valid Side Effect Score Achieved = 2.861



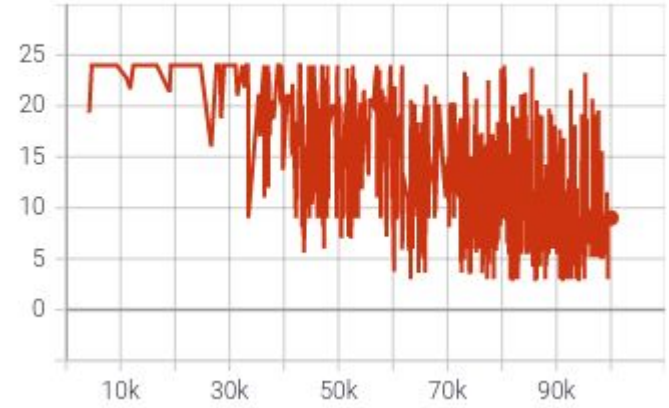




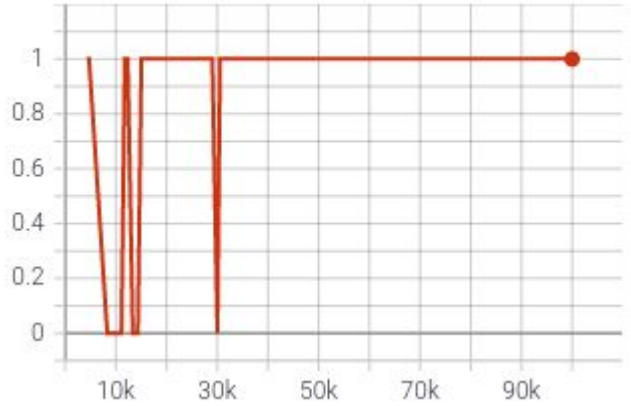
Results - MT-DQN (|R|=2)

Minimum valid Side Effect Score Achieved = 2.806

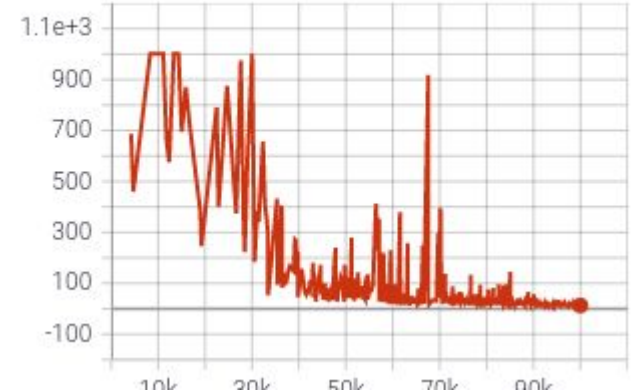
training/side_effect
tag: episodes/training/side_effect



training/reward
tag: episodes/training/reward



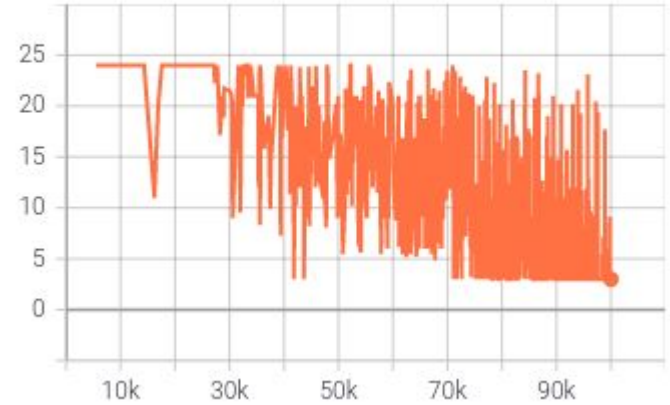
training/length
tag: episodes/training/length



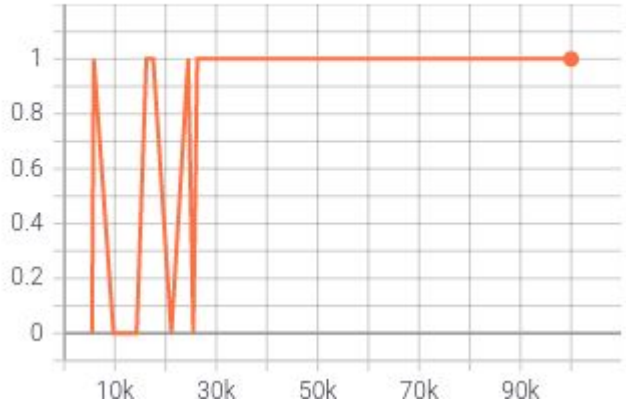
Results - MT-DQN (|R|=4)

Minimum valid Side Effect Score Achieved = 2

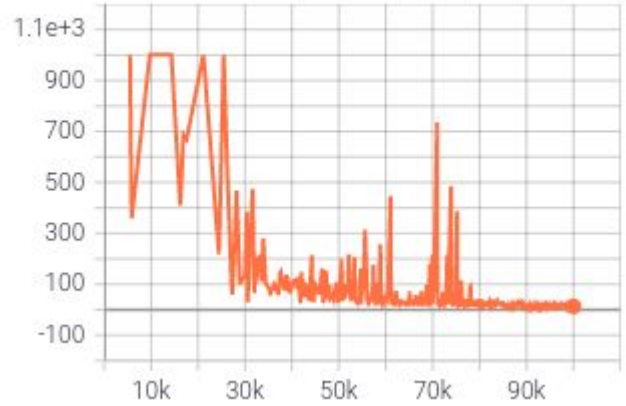
training/side_effect
tag: episodes/training/side_effect



training/reward
tag: episodes/training/reward



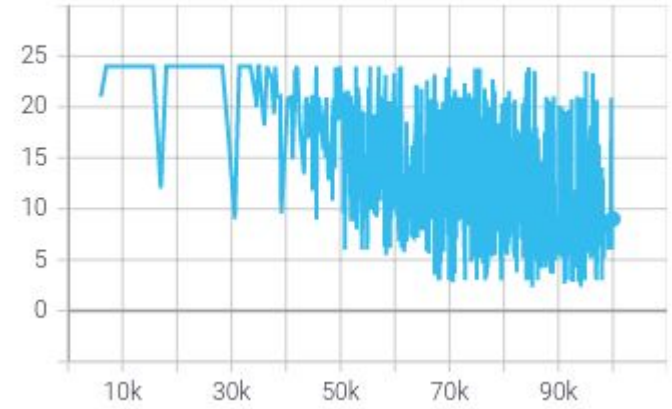
training/length
tag: episodes/training/length



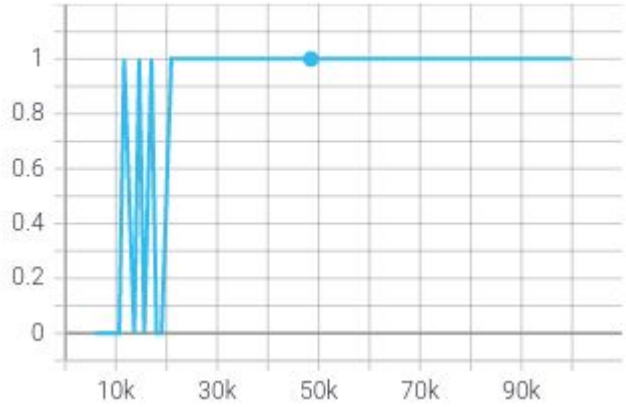
Results - MT-DQN (|R|=8)

Minimum valid Side Effect Score Achieved = 2

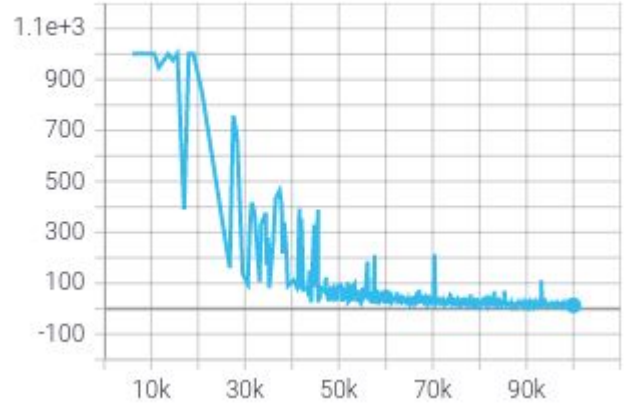
training/side_effect
tag: episodes/training/side_effect



training/reward
tag: episodes/training/reward



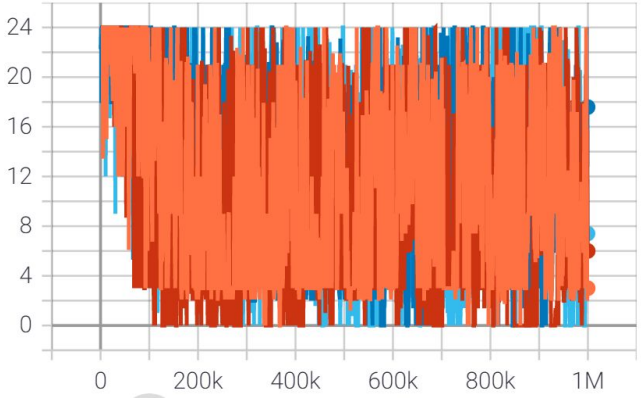
training/length
tag: episodes/training/length



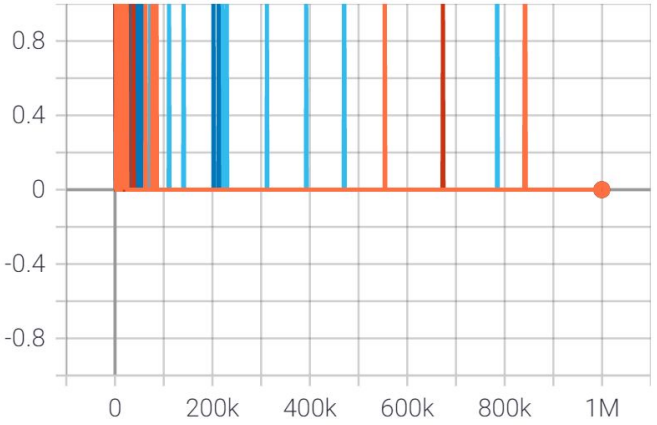
Results - PE-DQN

Most episodes time out before reaching the goal.
Number of aux rewards: 1, 2, 4, & 8

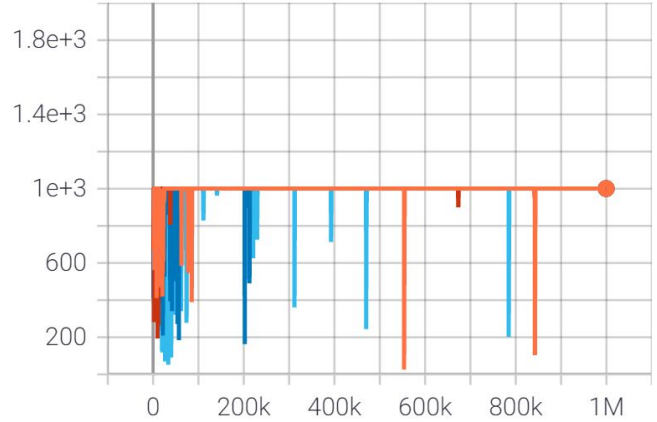
Side Effects



Reward



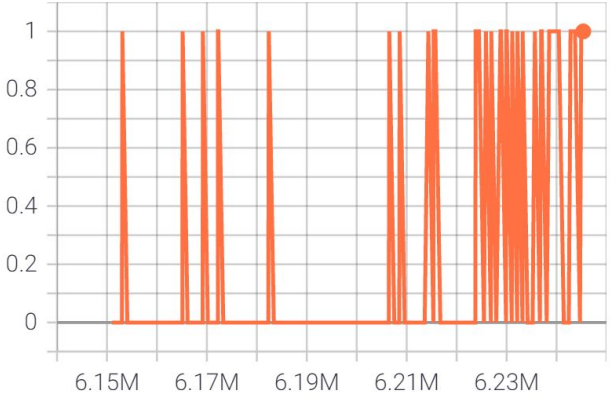
Episode Length



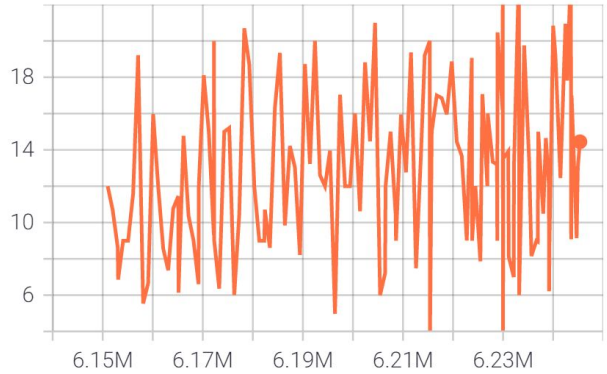
Results - DRQN w/ AUP (|R|=1)

Note: x-axis labels artificially start at 6.15M, this should be 0.
Represented are just shy of 100,000 steps

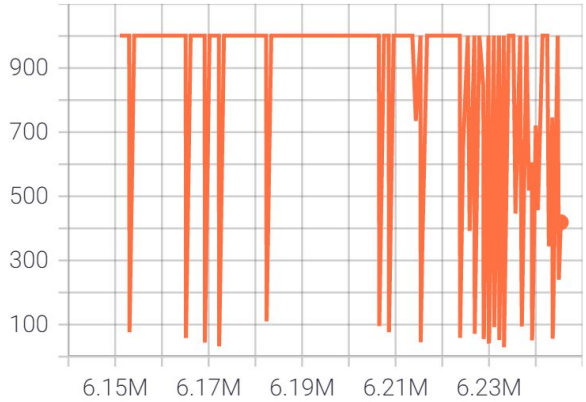
Reward



Side Effects



Episode Length



Our Contributions

- MT-DQN as a single stage alternative to multi stage AUP.
  - **Improve Side effect score** over DQN-AUP (2<3)
  - **Converges in100k steps compared to 2M** for DQN-AUP (1M for aux and 1M for aup)
  - Avoids the requirement of a ‘no-op’ action.
  - Can be scaled to sufficiently large number of auxiliary rewards without significantly increasing training time.

- DRQN: Incorporating LSTM into DQN
  - Learns on trajectories rather than individual transitions
  - Maintains DQN’s decorrelated training data by sampling random subsequences from replay buffer
  - Better able to capture far-reaching (non-immediate) side effects

Discussion and Future Work

MT-DQN

- Do a more rigorous evaluation and comparison on more difficult navigation tasks.
- Explore how MT-DQN performs on append and prune SafeLife tasks.

DRQN w/ AUP

- DRQN for side effects seems compelling
- Unable to produce representative results w/ limited computational resources
- 2 implementations
  - First: improved side effect score but failed to learn navigation task
  - Second: learned navigation task, insufficient data to determine impact on side effects
- Next step: reduce memory consumption, train 20x longer

References

a Victoria Krakovna, Laurent Orseau, Miljan Martic, and Shane Legg. Measuring and avoiding side effects using relative reachability. CoRR, abs/1806.01186, 2018.

b Rohin Shah, Dmitrii Krasheninnikov, Jordan Alexander, Pieter Abbeel, and Anca Dragan. The implicit preference information in an initial state. In International Conference on Learning Representations, 2019

c Alexander Matt Turner, Dylan Hadfield-Menell, and Prasad Tadepalli. Conservative agency via attainable utility preservation. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pages 385–391, 2020.

d Alexander Matt Turner, Neale Ratzlaff, Prasad Tadepalli. Avoiding Side Effects in Complex Environments In Advances in Neural Information Processing Systems 33 (NeurIPS 2020)

e Matthew Hausknecht, Peter Stone. Deep Recurrent Q-Learning for Partially Observable MDPs. In AAI Fall Symposium on Sequential Decision Making for Intelligent Agents (AAAI-SDMIA15)

Questions?

Email us!
(saxenaa, crowleyd)