REINFORCEMENT LEARNING WITH UNSUPERVISED AUXILIARY TASKS

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki
Tom Schaul, Joel Z Leibo, David Silver, Koray Kavukcuoglu

DeepMind

25th April 2017, ICLR
**Motivation**

Deep reinforcement learning is very data hungry.
- Many steps to find sparse rewards.
- Scalar supervision (returns).
- Neural networks with many parameters.
- Slow stochastic gradient descent.

This work — augment an RL agent with auxiliary prediction and control tasks.

This provides extra supervision to train feature extractors resulting in
- **10x improvement in data efficiency** over A3C on 3D DeepMind Lab.
- **60% improvement in final scores** over A3C.
UNREAL augments an LSTM A3C agent with 3 auxiliary tasks.

Can be used on top of DQN, DDPG, TRPO, or other agents.
BASE POLICY
Base policy is an LSTM agent trained with A3C [Mnih 2016].

Advantage actor-critic algorithm with multiple asynchronous workers.

$$\nabla_\theta \log \pi(a_t|s_t; \theta)(R_t - V(s_t; \theta_v))$$

$$\sum_{i=0}^{k-1} \gamma^i r_{t+i} + \gamma^k V(s_{t+k}; \theta_v)$$
Base policy is an LSTM agent trained with A3C [Mnih 2016]. Advantage actor-critic algorithm with multiple asynchronous workers.

UNREAL always acts in the environment with the base A3C policy.
UNREAL always acts in the environment with the base A3C policy.

Base policy is an LSTM agent trained with A3C [Mnih 2016].

Advantage actor-critic algorithm with multiple asynchronous workers.

UNREAL always acts in the environment with the base A3C policy.

Stores sequences of transitions in a replay buffer.
Sampling sequences of transitions from replay allows us to perform auxiliary tasks off-policy for feature learning — no need to for off-policy correction.

Networks used for auxiliary tasks are weight shared to agent's network.
Augment A3C with many auxiliary control tasks.

Learning to control many aspects of the environment.

**Pixel Control** — learn to maximally change parts of the visual input.

**Feature Control** — learn to control the internal representations.
**Pixel Control** — learn to maximally change parts of the visual input.

- Divide observation into cells.
- Learn a per-cell policy — dueling deconv net [Wang 2016].
- Reward is absolute change in average pixel intensity in cell between time steps.
- Optimise with n-step Q-learning.

Augment A3C with many auxiliary control tasks.

Learning to control many aspects of the environment.
**Reward Prediction** — learn to classify the next step reward.

- Sample mini-sequences — skewed sampling so 50% sequences end in non-zero reward.
- Encode observations with agent’s CNN.
- Concatenated encodings are used to classify +ve, zero, or -ve reward in subsequent frame.

Focus agent’s features on rewarding events.

Shape agent’s CNN through prediction task.
Value Function Replay

Reuse experience to faster train value function.

Better features to predict long term reward.

Value Function Replay — further training of value function.

- Sample sequences and perform value regression.
- Bootstraps from values which are not seen on-policy.
UNREAL optimises all 4 objectives simultaneously.

Minimal tuning on task balancing.

\[
\mathcal{L}_{UNREAL}(\theta) = \mathcal{L}_{A3C} + \lambda_{VR} \mathcal{L}_{VR} + \lambda_{PC} \sum_c \mathcal{L}_Q^{(c)} + \lambda_{RP} \mathcal{L}_{RP}
\]
EXPERIMENTS

Unsupervised Learning for RL

UNREAL optimises all 4 objectives simultaneously.

Minimal tuning on task balancing.
DeepMind Lab EXPERIMENTS

Tested on a suite of 13 levels. Lasertag, procedural mazes, apple foraging. Compared to human performance.

Performance
Avg. TOP 3 agents

10x improvement in data efficiency.
60% improvement in final performance.

Robustness

Significantly more robust to hyper parameters.
Control vs prediction vs reconstruction

Input **reconstruction** can help initially but often harms policy.
Input **change prediction** helps much more.
Input **control** works best.
Feature control can work as well as pixel control. Promising future direction.
Unsupervised Learning for RL

DeepMind Lab EXPERIMENTS

Live Play

Reward

Auxiliary Tasks

Pixel Control

Q value for action

Reward Prediction

Value Function Replay

Actions Value Function

Reward
No Reward
ATARI EXPERIMENTS

State of the art results on Atari:
1453% mean, 331% median of human performance across suite of 57 games.
In 24 hours, UNREAL is close to solving nav_maze_static_01 while A3C is just starting to learn.
SCALING RL

Overnight, in 8 hours, UNREAL starts learning. Add more workers?
4x workers means that overnight UNREAL is solving the level.
With 512 workers, can reach human performance in 2 hours.
SCALING RL  
Atari 1k workers

Alien

Frostbite

Pong

Space Invaders

Evolution Strategies (ES) [Salimans 2017]
CONCLUSION

Unsupervised auxiliary tasks drastically improve RL training
- Pixel Control + Reward Prediction + Value Function Replay.
- Improve speed of convergence, robustness to hyperparameters, and results in better performance.
- 10x faster learning and 60% higher performance on DM Lab.
- State of the art results on Atari.

RL scales with CPUs
- Adding more asynchronous workers to A3C-style algorithms greatly speeds up learning.
- Super-human performance on complex domains in a matter of hours.