Deep Recurrent Q-Learning for Partially Observable MDPs

Matthew Hausknecht and Peter Stone

AAAI Fall Symposium on Sequential Decision Making for Intelligent Agents (AAAI-SDMIA15)

Why Do We Need Recurrency?

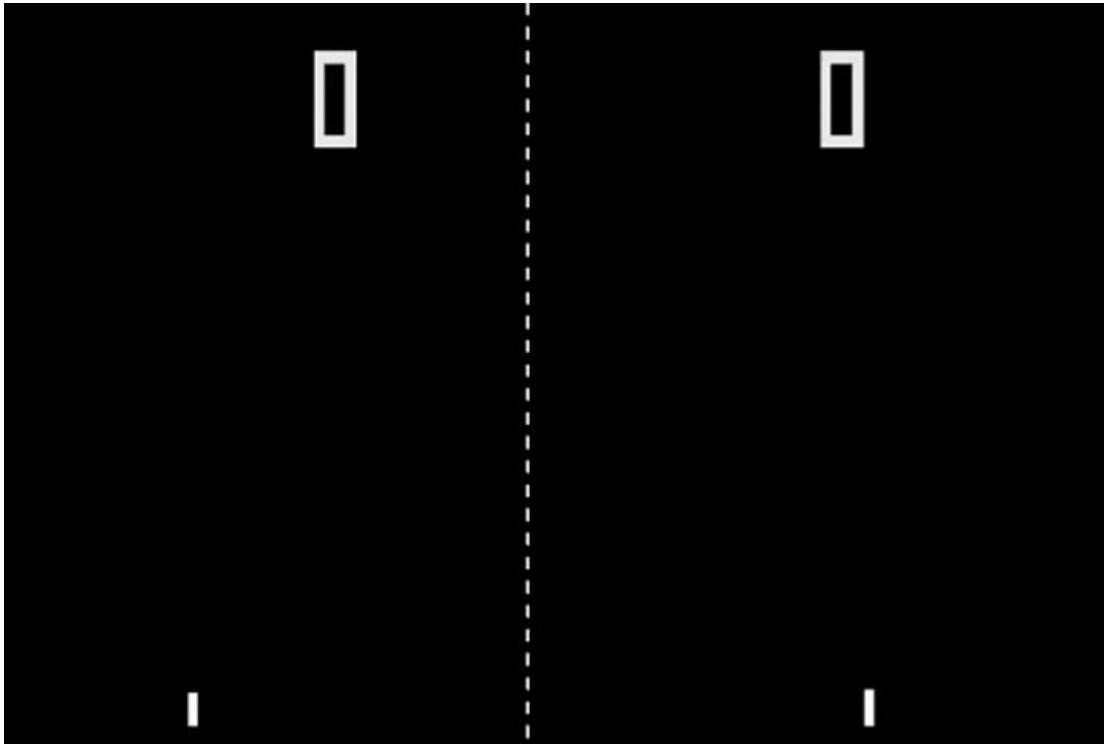


Image retrieved from: link

Why Do We Need Recurrency?

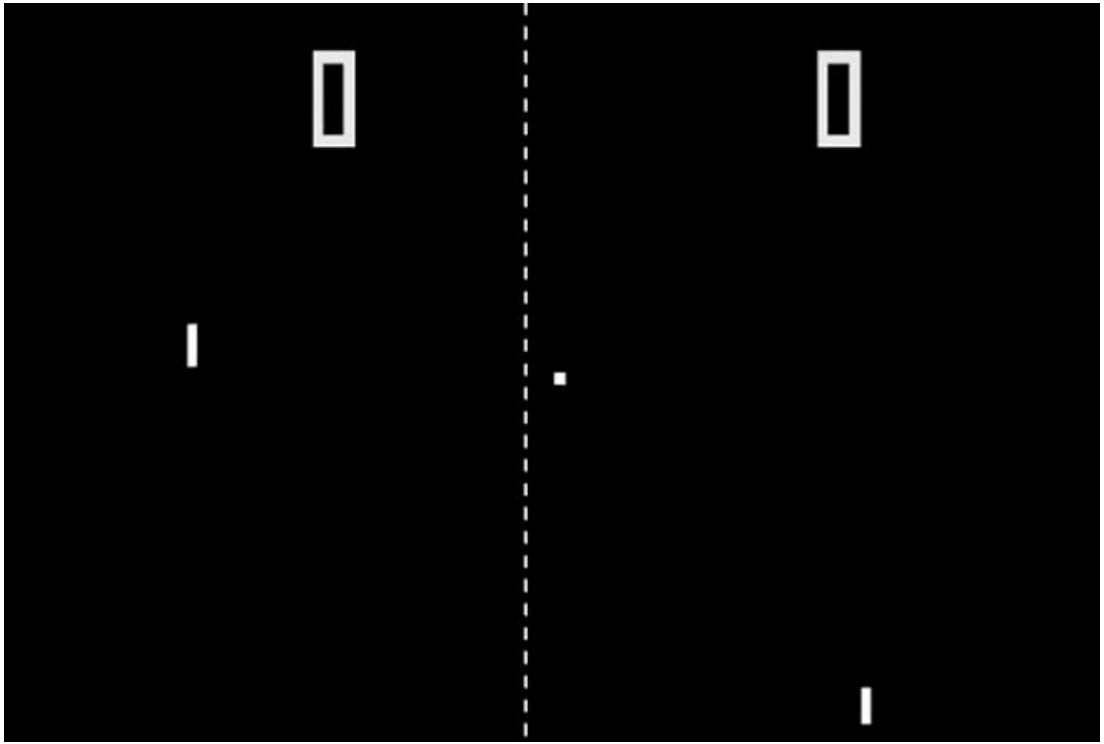


Image retrieved from: link

Why Do We Need Recurrency?

Image retrieved from: link

Outline

1. ~~Why use Recurrency?~~
2. MDPs and POMDPs
3. Deep Q-Learning
4. Putting the R in DRQN
5. Atari Experiments
6. Experiment Review
7. Conclusion

MDPs and POMDPs

Markov Decision Process

- Agent receives state **s**
- **s** is the true system state

Partially Observable Markov Decision Process

- Agent receives observation **o**
- **o** only partial description of system state

Recurrency helps to “narrow the gap” between **Q(s,a)** and **Q(o,a)**

Q-Learning

Learn

$$Q(s, a)$$

Using

$$Q(s,a) := Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s', a') - Q(s,a)\right)$$

Deep Q-Learning

Learn

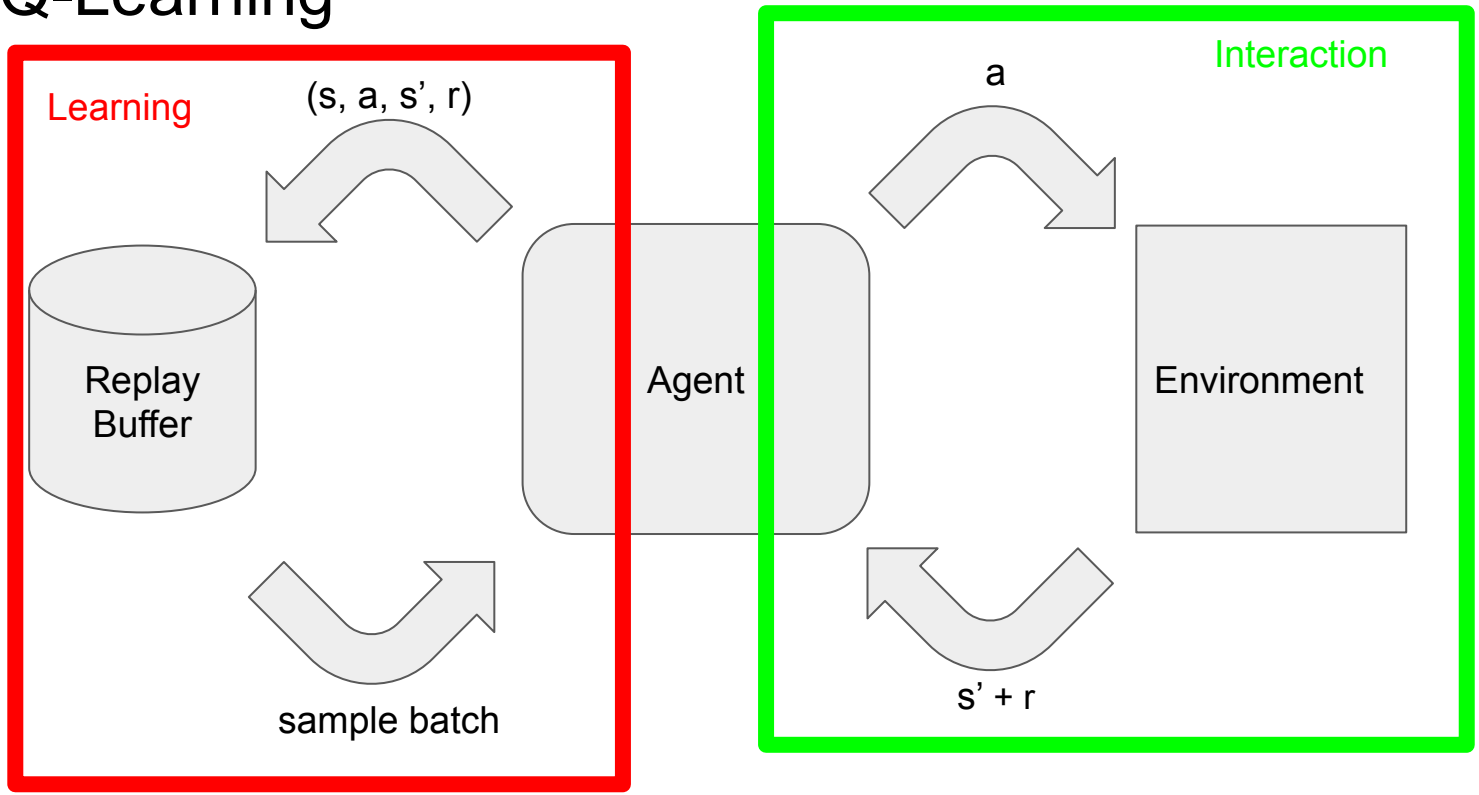
$$Q(s, a|\theta_i)$$

Using

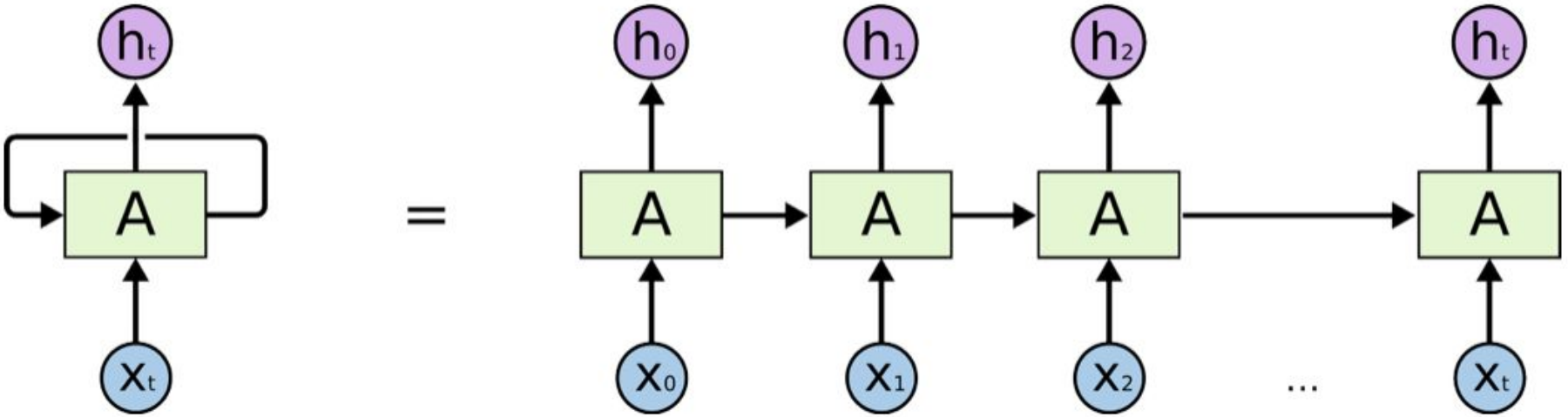
$$L(s, a|\theta_i) = \left(r + \gamma \max_{a'} Q(s', a'|\theta_i) - Q(s, a|\theta_i)\right)^2$$

$$\theta_{i+1} = \theta_i + \alpha \nabla_\theta L(\theta_i)$$

Deep Q-Learning



Putting the R in DRQN

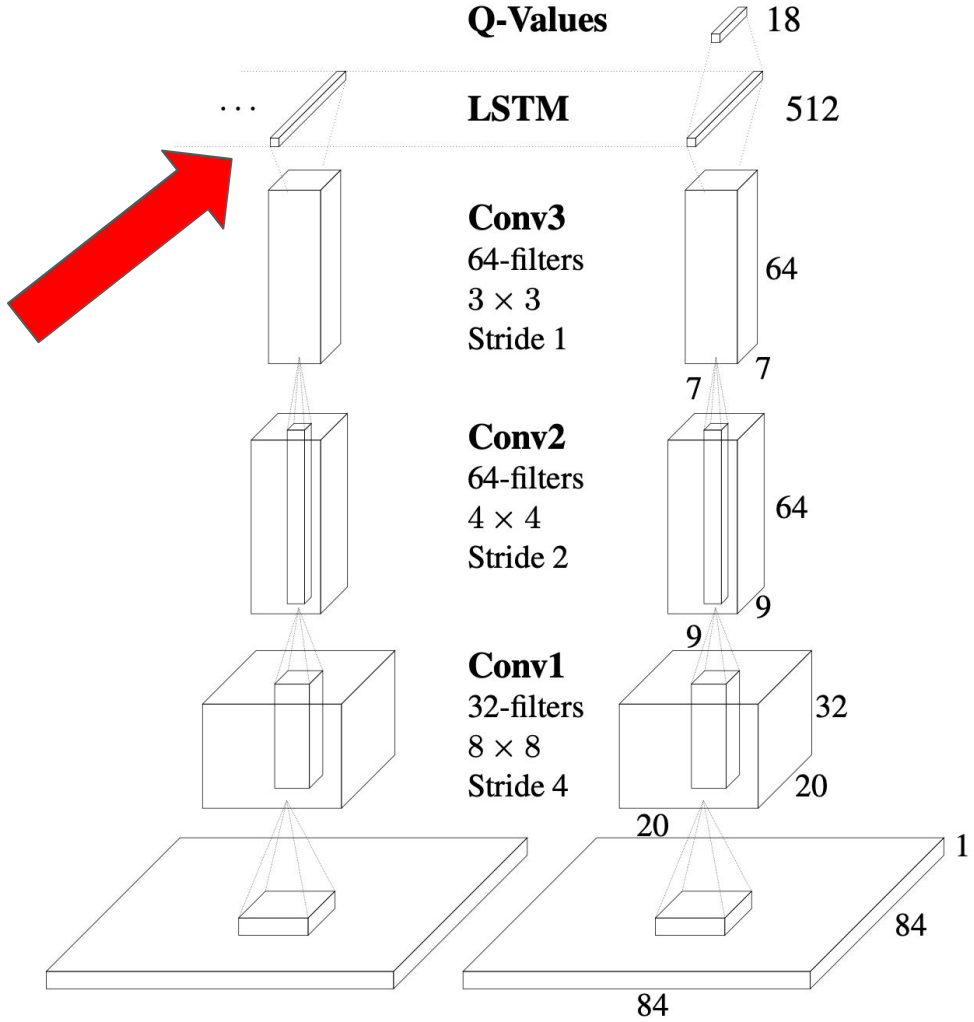


An unrolled recurrent neural network.

Image retrieved from: [link]

Putting the R in DRQN

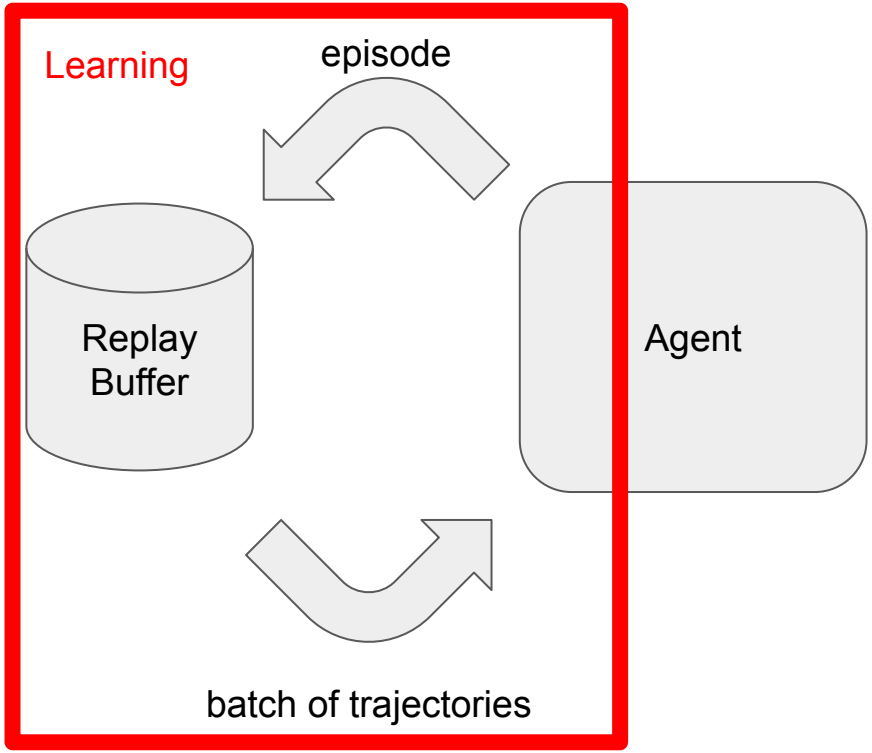
first fully connected layer is replaced with LSTM



Putting the R in DRQN

Bootstrapped Random Updates

- Sample sequences of length k
- Hidden state zeroed at start of update

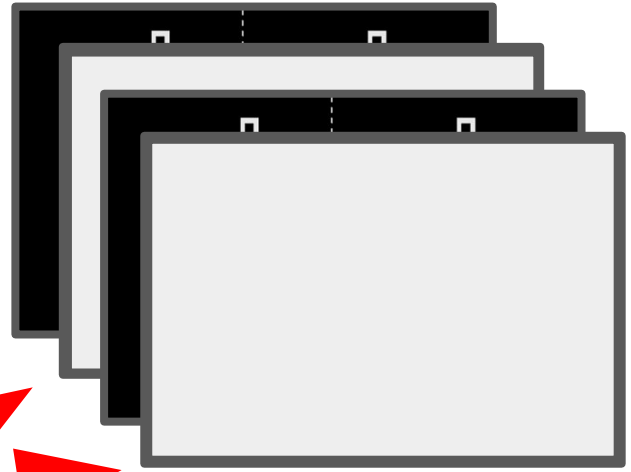


Outline

1. ~~Why use RNNs?~~
2. ~~MDPs and POMDPs~~
3. ~~Deep Q-Learning~~
4. ~~Putting the R in DRQN~~
5. Atari Experiments
6. Experiment Review
7. Conclusion

Questions?

Flickering Atari



P(obscured) = 0.5

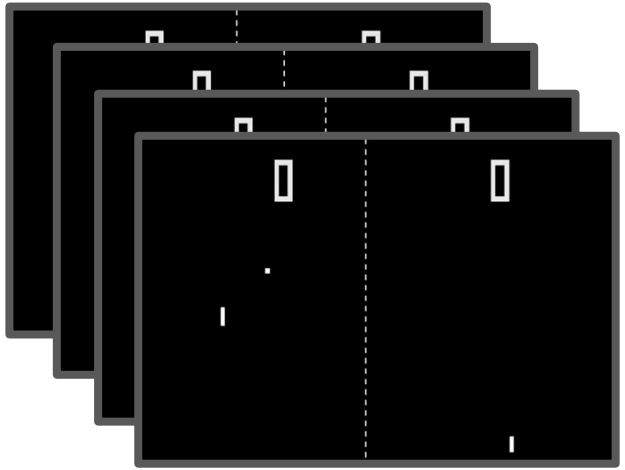
DQN Agent

DRQN Agent

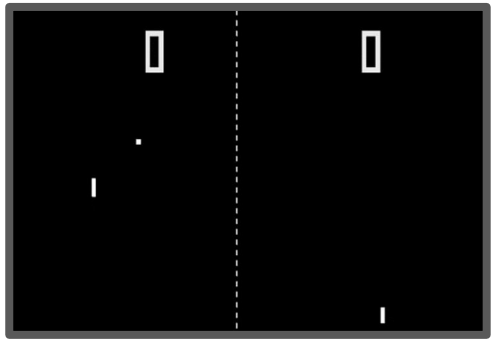
Results

Flickering	DRQN $\pm std$	DQN $\pm std$
Asteroids	1032 ($\pm$410)	1010 ($\pm$535)
Beam Rider	618 ($\pm$115)	**1685.6** ($\pm$875)
Bowling	65.5 ($\pm$13)	57.3 ($\pm$8)
Centipede	4319.2 ($\pm$4378)	5268.1 ($\pm$2052)
Chopper Cmd	1330 ($\pm$294)	1450 ($\pm$787.8)
Double Dunk	-14 ($\pm$2.5)	-16.2 ($\pm$2.6)
Frostbite	414 ($\pm$494)	436 ($\pm$462.5)
Ice Hockey	-5.4 ($\pm$2.7)	-4.2 ($\pm$1.5)
Ms. Pacman	1739 ($\pm$942)	1824 ($\pm$490)
Pong	**12.1** ($\pm$2.2)	-9.9 ($\pm$3.3)

Standard Atari



DQN Agent

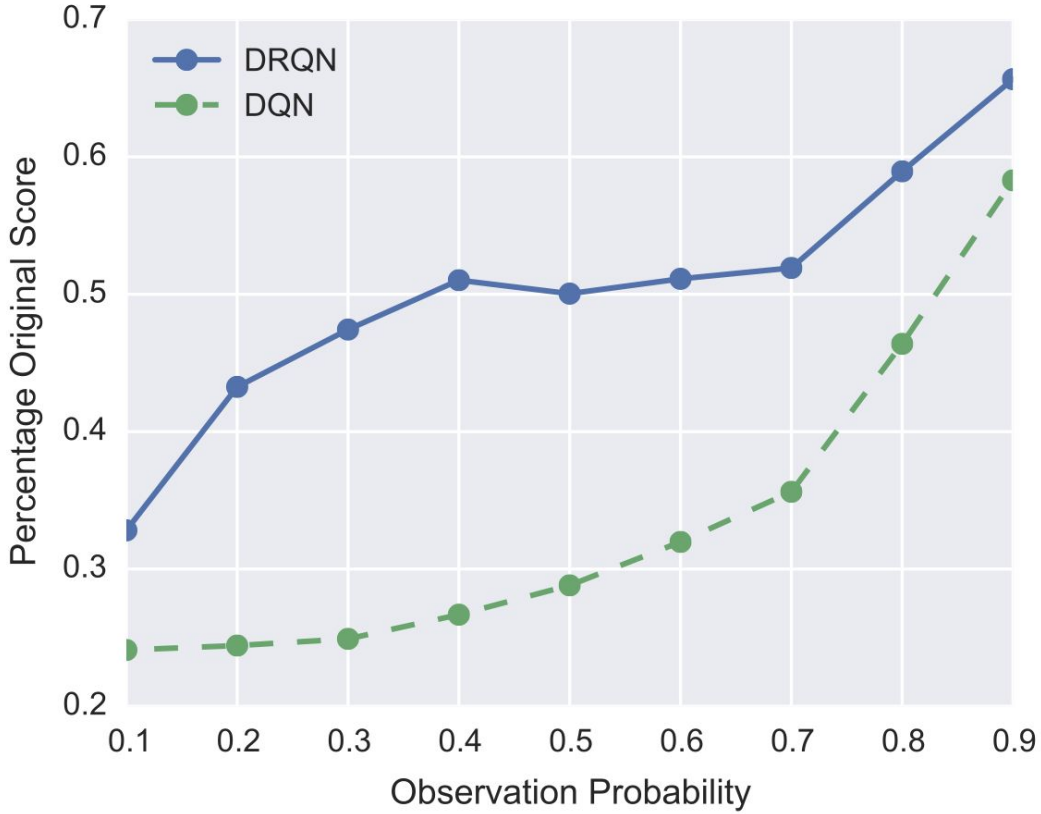


DRQN Agent

Standard Atari Results

	DRQN $\pm std$	DQN $\pm std$	
Game		Ours	Mnih et al.
Asteroids	1020 ($\pm$312)	1070 ($\pm$345)	1629 ($\pm$542)
Beam Rider	3269 ($\pm$1167)	**6923** ($\pm$1027)	6846 ($\pm$1619)
Bowling	62 ($\pm$5.9)	72 ($\pm$11)	42 ($\pm$88)
Centipede	3534 ($\pm$1601)	3653 ($\pm$1903)	8309 ($\pm$5237)
Chopper Cmd	2070 ($\pm$875)	1460 ($\pm$976)	6687 ($\pm$2916)
Double Dunk	**-2** ($\pm$7.8)	-10 ($\pm$3.5)	-18.1 ($\pm$2.6)
Frostbite	**2875** ($\pm$535)	519 ($\pm$363)	328.3 ($\pm$250.5)
Ice Hockey	-4.4 ($\pm$1.6)	-3.5 ($\pm$3.5)	-1.6 ($\pm$2.5)
Ms. Pacman	2048 ($\pm$653)	2363 ($\pm$735)	2311 ($\pm$525)

MDP to POMDP Generalization



Benjamini-Hochberg Procedure

Problem: Multiple hypothesis testing can result in false positives

Solution: Adjust p-values so that the % of false positives is controlled

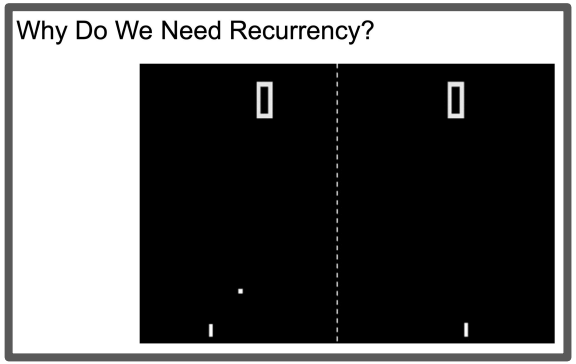
B-H with threshold of 0.05 means that 5% of **significant** results are false positives

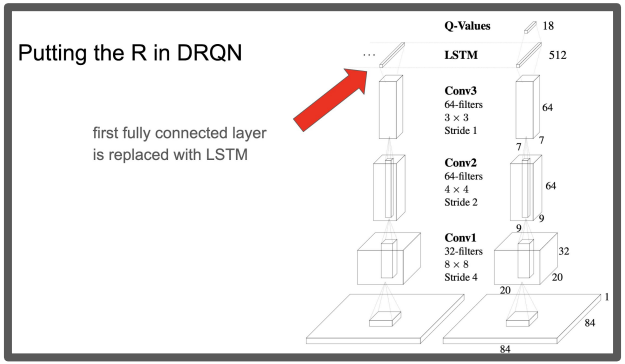
Experiment Review

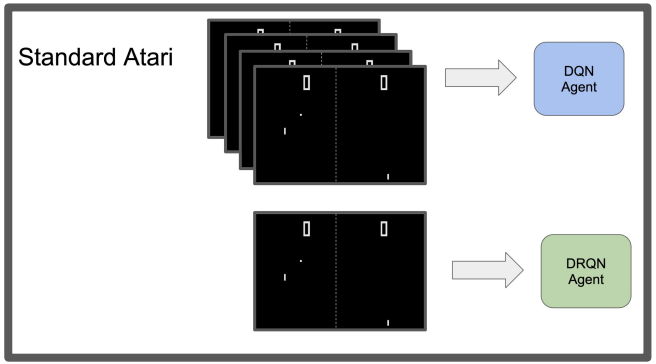
Good	Needs Improvement
● Agents evaluated offline at intervals as recommended in [1] ● Mostly small adjustments from DQN paper ● BH procedure to control false positive rate	● Individual t-tests assume normality ● Compare against original DQN results but make modifications ● Hyperparameter selection not explained ● Ambiguous experiment details ● Missing error bars ● Some conclusions seem unjustified

[1]:https://arxiv.org/abs/1904.06979

Conclusion


Why Do We Need Recurrency?


Putting the R in DRQN
first fully connected layer
is replaced with LSTM
Q-Values
18
LSTM
512
Conv3
64-filters
3 × 3
Stride 1
64
7
7
Conv2
64-filters
4 × 4
Stride 2
64
9
9
Conv1
32-filters
8 × 8
Stride 4
32
20
20
1
84
84


Standard Atari
DQN Agent
DRQN Agent