This page left blank on purpose
Learning to Reinforcement
Learn
Learn
Paper by
Wang, Jane X; Kurth-Nelson, Zeb; Tirumala, Dhruva; Soyer, Hubert; Leibo, Joel Z; Munos, Remi; Blundell, Charles; Kumaran, Dharshan; Botvinick, Matt

Presented by: Aqeel Labash
Background

“Study the past if you would define the future.”

Confucius
Background

Artificial Neural Network

A) Neuron diagram showing dendrites, cell body, and terminal axon.

B) Schematic of a function $f(x)$ with inputs $x_1, x_2, \ldots, x_n$ and an output $y$.

C) Additional neuron diagram highlighting synapses.

D) Diagram of a neural network with input and output layers connected by synapses.
Background

Deep learning

Reinforcement learning
Background

Recurrent Neural Network

General

RNN

LSTM
Advantage Actor-critic

Use critic (value function) to update actor (policy)
Asynchronous advantage actor-critic
The Problem

“Every problem is a gift - without problems we would not grow.”

Anthony Robbins
Problems

Amount Of data and Generalization
Methods

“Now that we are all so smart, we don’t easily find resolutions.”

Dejan Stojanovic, The Sun Watches the Sun
Methods

Network Architecture

(a) LSTM A2C
(b) LSTM A3C
(c) Stacked-LSTM A3C
The Experiments

“The best way to show that a stick is crooked is not to argue about it or to spend time denouncing it, but to lay a straight stick alongside it”

D.L. Moody
Experiments

Exp 1. Bandit with two independent arms $D_i$ (Bernoulli dis)
20K train, 300 test, 100 Trial each. Probability change each Episode

$P_1, P_2 \in [0, 1]$
Regret: Expected reward from optimal arm - current reward.

\[ \mu^*(b) = \max_a \mu_a(b) = \mu_{a^*}(b)(b) \]

\[ R_T(b) = \sum_{t=1}^{T} \mu^*(b) - \mu_{a_t}(b) \]
Exp 1. Bandit with two independent arms $D_i$ (Bernoulli dis)
20K train, 300 test, 100 Trial each. Probability change each Episode
Exp 2. Bandit with two dependent arms (Bernoulli dis)
20K train, 300 test, 100 Trial each. Probability change each Episode

\[ P_1 \in [0, 1] : P_2 = (1 - P_1) \]

\[ D_u : P_1 \sim U([0, 1]) \]

\[ D_e : P_1 \sim U(\{0.1, 0.9\}) \]

\[ D_m : P_1 \sim U(\{0.25, 0.75\}) \]

\[ D_h : P_1 \sim U(\{0.4, 0.6\}) \]
Exp 2. Bandit with two dependent arms (Bernoulli dis)
Probability change each Episode

(c) Testing: Dependent Uniform
(d) Testing: Easy

<table>
<thead>
<tr>
<th>LSTM A2C “Dependent Uniform”</th>
<th>Gittins</th>
<th>Thompson</th>
<th>UCB</th>
</tr>
</thead>
</table>

| LSTM A2C “Medium” | Gittins | Thompson | UCB |

Cumulative Regret vs Trial #
Exp 2. Bandit with two dependent arms (Bernoulli dis)
Probability change each Episode

![Graph showing cumulative regret over trials and cumulative regret table across different training and testing conditions.](image-url)
Experiments

We might have the right answer but the wrong way.
Exp 3. Bandit with 11 arms (informative, target, less reward)
Target arm change per episode while informative arm is the same
Exp 3. Bandit with 11 arms (informative, target, less reward)
Target arm change per episode while informative arm is the same

Experiments

![Graph showing cumulative regret over trials]

- LSTM A2C
- Optimal
- Thompson
- UCB
Exp 4. Bandit with 2 arms
Probabilities change during the episode.

\[ P_1 \in [0, 1] : P_2 = (1 - P_1) \]
Exp 5. Two-step Task

MDP (Markov Decision Process)
Exp 5. Two-step Task

MDP (Markov Decision Process)
Exp 6. Select the right image.
Select Image, back to the center

(a) Fixation  (b) Image display  (c) Right saccade and selection
Exp 6. Select the right image.
Select Image, back to the center

(d) Training performance
(e) Robustness over random seeds
(f) One-shot learning
Exp 7. I-maze
Find Food in 3D world after Exploration.

(a) Labryinth I-maze  
(b) Illustrative Episode
Experiments

Exp 7. Select the right image.
Select Image, back to the center

(c) Performance
(d) Value Function
Conclusion
Conclusion

Ingredients
Conclusion

New son is born
Questions

“He who has a why to live for can bear almost any how.”

Friedrich Nietzsche
Credits

Presentation template:
- Presentation template: SlidesCarnival
- Photographs: Unsplash
- Ornaments: Briar Press
- Methodology: Michal Alley

Content:
- Arthur Juliani

Critic:
- Daniel.