FINALE : RL

REINFORCEMENT LEARNING

martin @ reddragon.ai
sam @ reddragon.ai

27 November 2017

`WiFi : SG-Guest`

Problems with Installation? **ASK!**

RED DRAGON AI

PLAN OF ACTION

TODAY

- Reinforcement Learning
- ~1 minute summaries
- Group picture!
- Finalize Projects

PLAN OF ACTION

30-NOV

- Project Deadline
- Feedback Forms!

RED DRAGON AI

REINFORCEMENT LEARNING

- Learning to choose actions ...
- ... which cause environment to change

REINFORCEMENT LEARNING

- Techniques that focus on decision-making processes ...
  - ... where each decision/action affects the future options available
- Standard setting :
  - Playing Checkers & Backgammon
  - ~~Playing Chess~~
  - Playing Atari 2600
  - Playing Go
  - Playing Poker + Dota2 + Starcraft

RED DRAGON AI

GOOGLE DEEPMIND'S ALPHAGO

First 99 moves

Moves 100–199 (118 at 107, 161 at ▢)

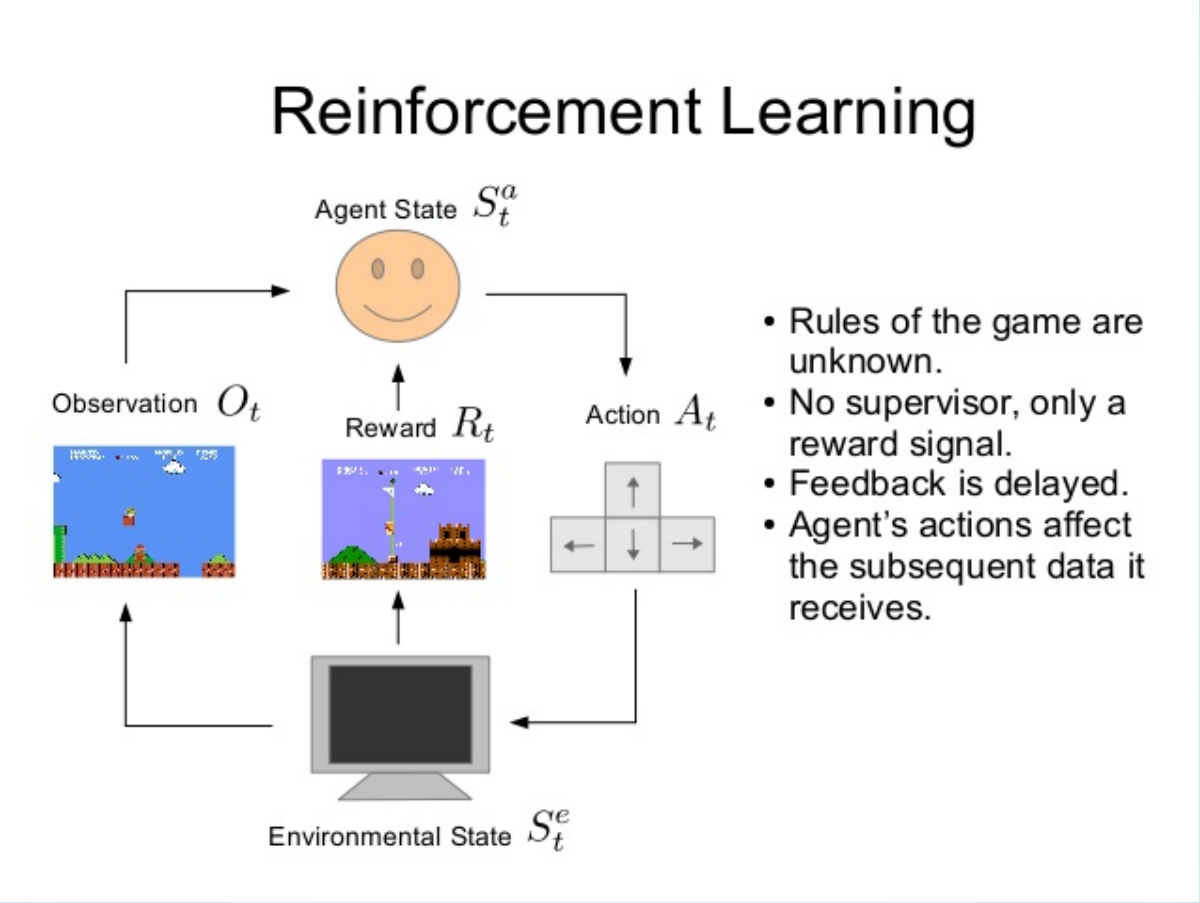
Moves 200–280 (240 at 200, 271 at ▢, 275 at ▢, 276 at △)

RED DRAGON AI

REINFORCEMENT LEARNING

- Other application examples :
  - Deciding which advertisements to show
  - Dynamic pricing policies
  - Control of unknown `plant' (e.g. air conditioning)
  - Robots "learning-by-example"

AGENT LEARNING SET-UP



RED DRAGON AI

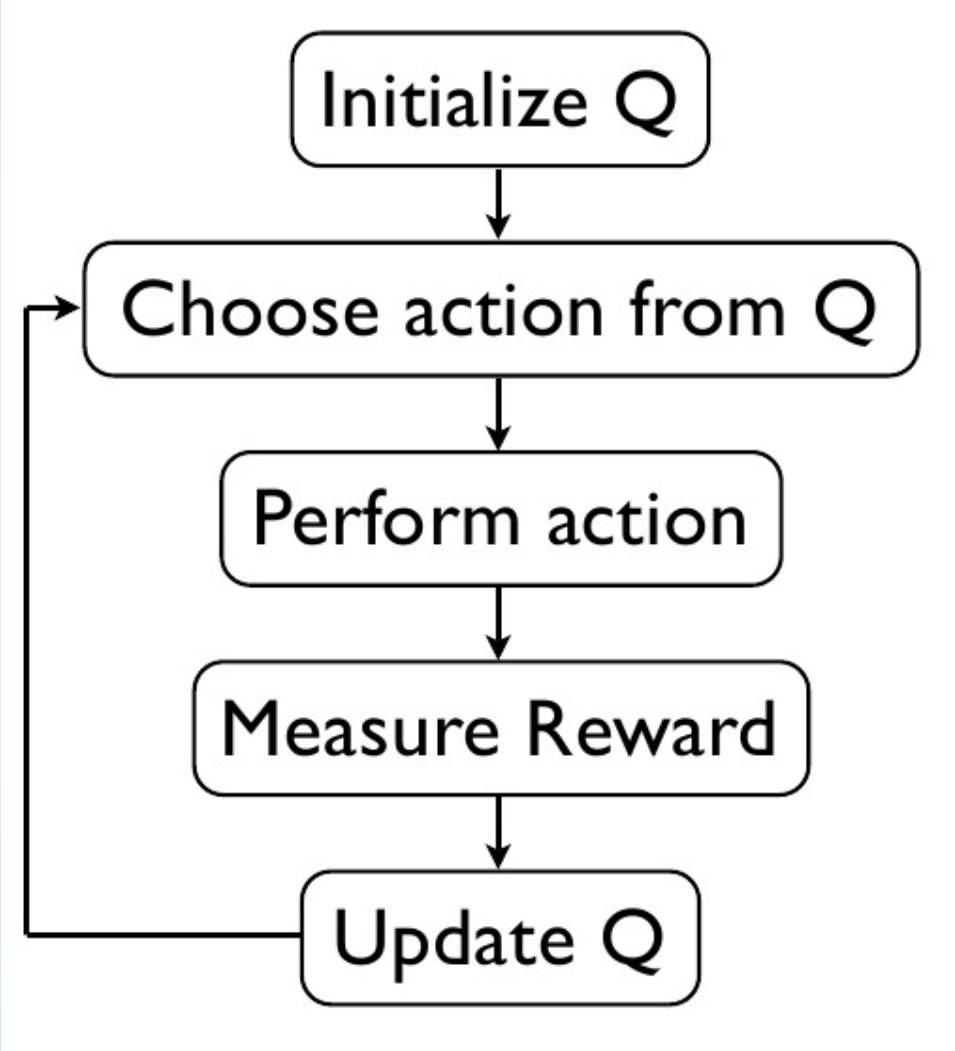
Q-LEARNING 1

- Estimate value of entire future from current state
- ... to estimate value of next state, for all possible actions
- Determine the 'best action' from estimates

Q-LEARNING 2

- ... do the best action
- Observe rewards, and new state
- * Update Q(now) to be closer to R+Q(next) *

Q-LEARNING DIAGRAM



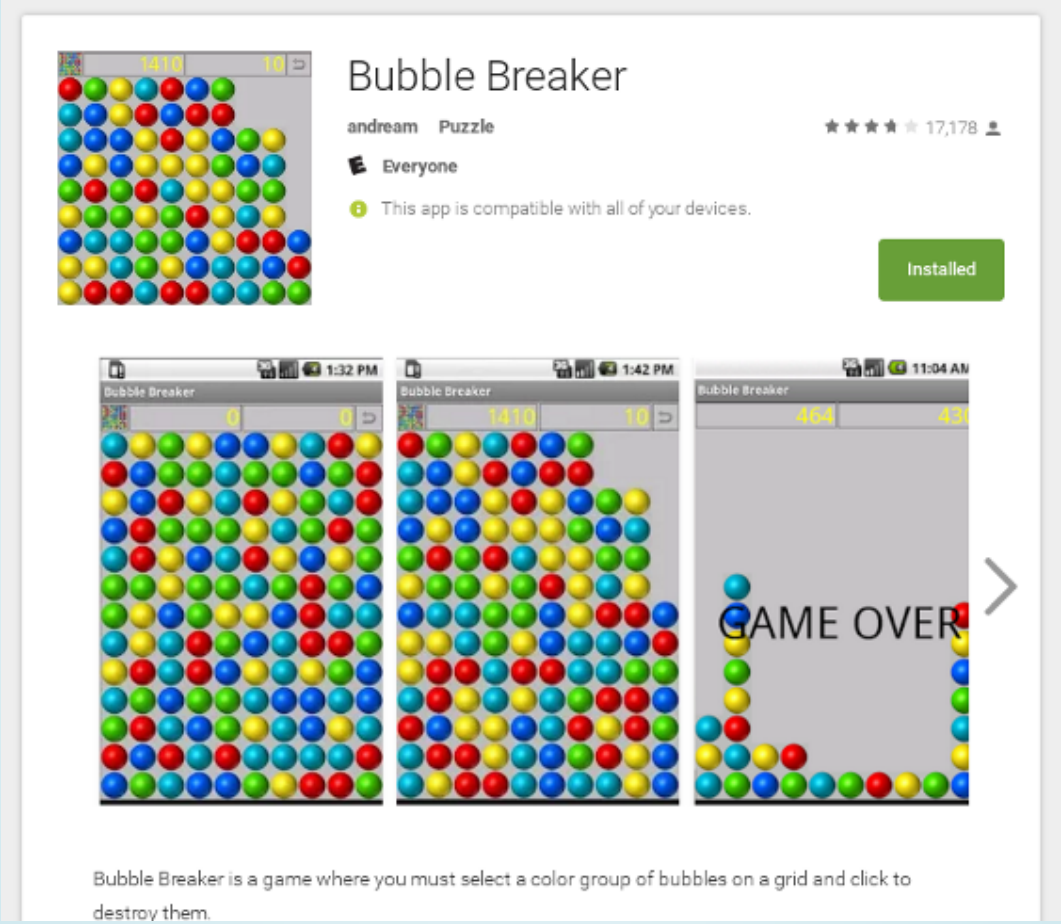
Q is a measure of what we think about the future

RED DRAGON AI

DEEP Q-LEARNING

- Set Q() to be the output of a deep neural network
- ... where the input is the state
- Train network input/output pairs from observed steps
- ... over *many* games

TODAY'S STRATEGY GAME



Classic game : No superfluous features

RED DRAGON AI

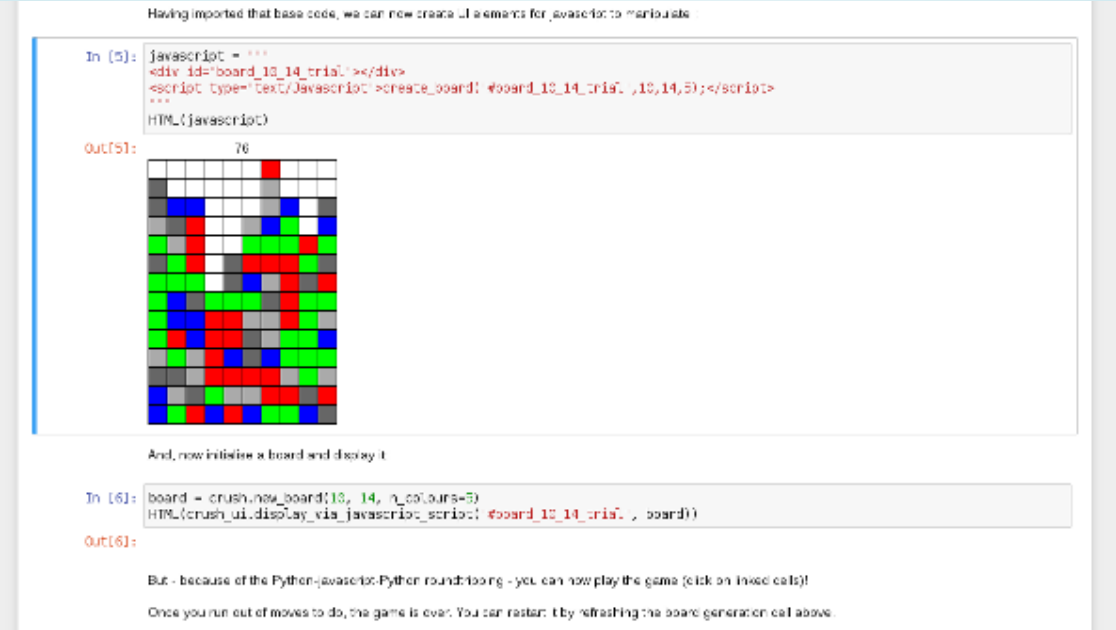
BUBBLE BREAKER : YOU

- How-to-play
- 5 mins test...

BUBBLE BREAKER : YOU

- Clicking on 'joined' bubbles kills group
- Bubbles fall down from the top to fill space
- Empty columns filled by shifting columns over from left
- No special bubbles : 5 colours only
- Game ends when there are no moves left

REINFORCEMENT LEARNING NOTEBOOK



Having imported that base code, we can now create all elements for javascript to manipulate

In [5]:
```
javascript = '''
<div id='board_13_14_trial'></div>
<script type='text/Javascript'>create_board('board_13_14_trial',13,14,5);</script>
'''
HTML(javascript)
```

Out[5]:

76

And, now initialise a board and display it

In [6]:
```
board = crush.new_board(13, 14, n_colours=5)
HTML(crush_ui.display_via_javascript_script('board_13_14_trial', board))
```

Out[6]:

But - because of the Python-javascript-Python roundtripping - you can now play the game (click on linked cells)!

Once you run out of moves to do, the game is over. You can restart it by refreshing the board generation cell above.

Deep Reinforcement Learning for Bubble Breaker

RED DRAGON AI

BUBBLE BREAKER LESSONS

- Planning
- Strategies
- Failure modes

BUBBLE BREAKER (RL)

- Turning the Board into Features
- Choosing which move to make
- Choosing a reward function
- Batch Learning

BOARD → FEATURES

- Using colours of blobs as features is possible
- ... but wasteful, due to symmetry
- Encode position as several feature layers:
  - Board silhouette
  - colour[i, j] == colour[i+a, j]
  - colour[i, j] == colour[i, j+b]
- Symmetry speedup : 120x (=5!)

CHOICE OF MOVE

- Game code can 'run' an action against the board
- Evaluate each separate resulting board
- Choose from ranked list :
  - Exploit : Choose best move
  - Explore : Choose random move (10%)

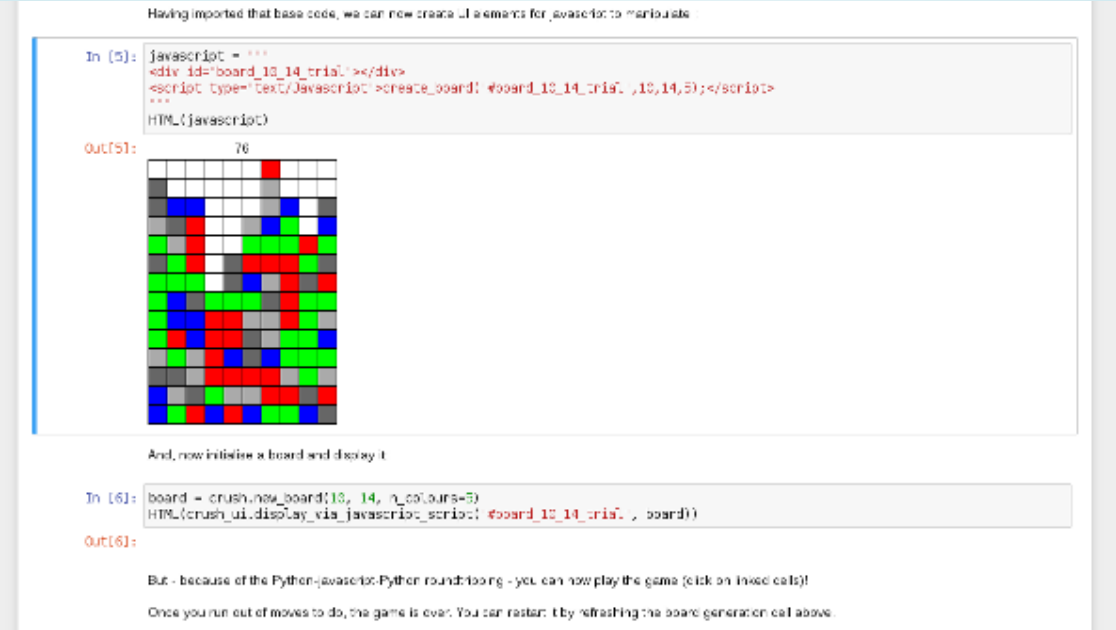
REWARD FUNCTION

- Pros/cons of using 'change in score' :
  - Using the ‘score’ promotes short-term gains
  - Using new-columns-added leads to ‘better’ play

BATCH LEARNING

- Normally, networks train on same data repeatedly
- But past actions may become irrelevant to training
- Retain some memory of previous actions
- But 'roll forward' with newer examples continuously

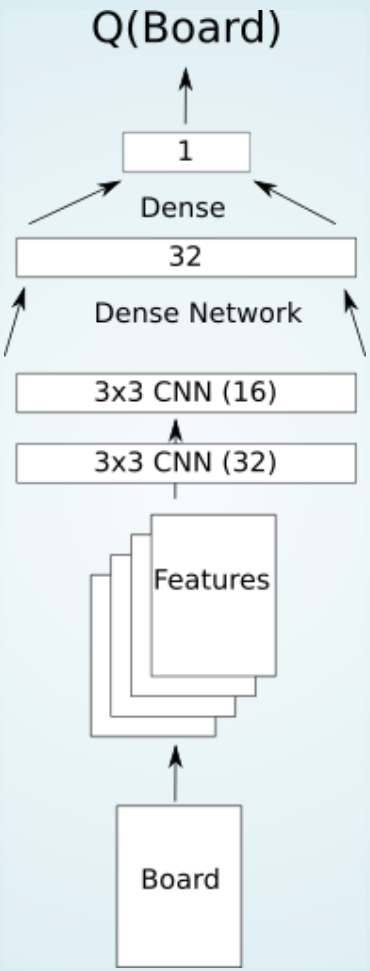
REINFORCEMENT LEARNING DEMO



Deep Reinforcement Learning for Bubble Breaker

RED DRAGON AI

NETWORK PICTURE



RED DRAGON AI

ALPHAGO RECORD

- May 2016 : Defeat of Lee Sedol
- Jan 2017 : 'Master' played 60 games online
- May 2017 : Defeat of Ke Jie
- Aug 2017 : AlphaGo Zero is better
- ... retired from match-play

ALPHAGO EXTRAS

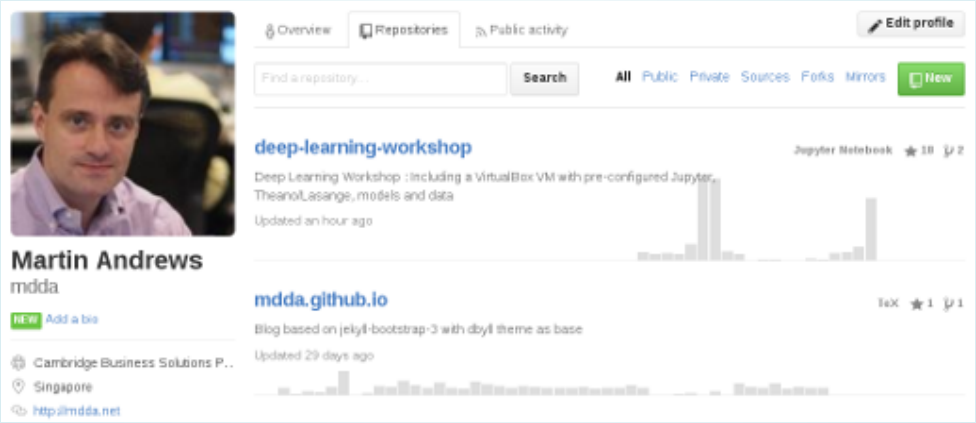
- Monte-Carlo Tree Search
- Policy Network to hone search space
- Self-play
- ... and running on 1202 CPUs and 176 GPUs

ALPHAGO ZERO

- Only self-play
- Policy network and value network share weights
- Stability problems didn't affect learning
- ... and running 4 TPUs

WRAP-UP

- Explore structure vs accuracy tradeoffs
- Even tiny models work 'well enough'
- Lots more behind all this



* Please add a star... *

RED DRAGON AI

- QUESTIONS -

MARTIN.ANDREWS @
REDDRAGON.AI

My blog : http://mdda.net/

GitHub : mdda

RED DRAGON AI