Mastering Tetris with Reinforcement-Learning

B.Sc. Thesis Presentation

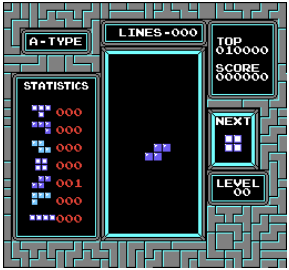
Shengdi Chen
shenchen@ethz.ch

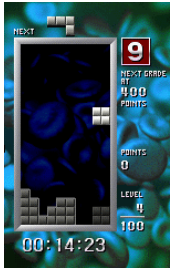
2022-08-09

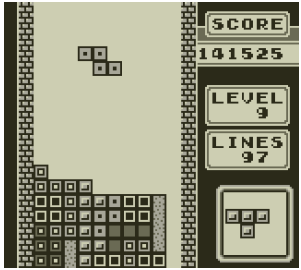
Synopsis

Throw-backs

Shengdi Chen
shenchen@ethz.ch

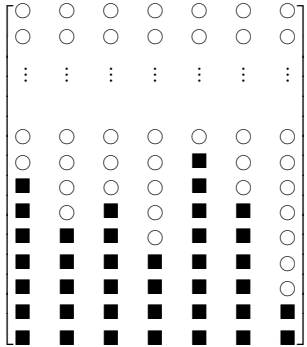






The Field

Just some 0s and 1s

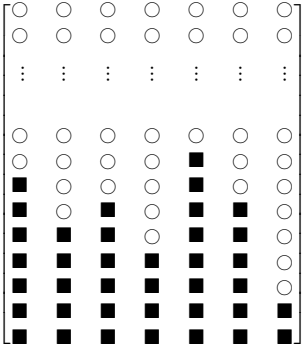


Shengdi Chen
shenchen@ethz.ch

The Field

Just some 0s and 1s

Shengdi Chen
shenchen@ethz.ch



$$\Longleftrightarrow \quad \left(\{\bigcirc, \blacksquare\} \equiv \texttt{bool}\right)^{\#\text{rows} \times \#\text{columns}}$$

The Coordinates

Indexing by numpy

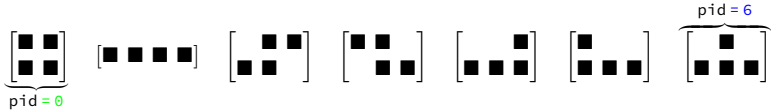
$$\begin{bmatrix} (0,0) & (0,1) & \cdots & (0,9) \\ (1,0) & (1,1) & & \\ \vdots & & & \\ \vdots & & & \\ \vdots & & \ddots & \\ (19,0) & & & (19,9) \end{bmatrix}$$

$$\begin{cases} \#\text{rows} & := & 20 \\ \#\text{columns} & := & 10 \end{cases}$$

Shengdi Chen
shenchen@ethz.ch

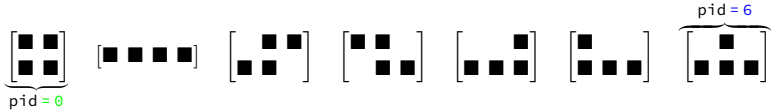
The shapes

range() as in python



The shapes

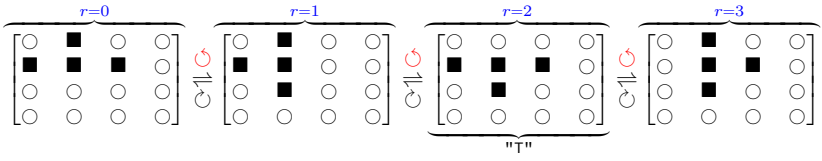
range() as in python



$$\texttt{pid} \in \{0, 1, \cdots, 6\} \triangleq \texttt{range(7)}$$

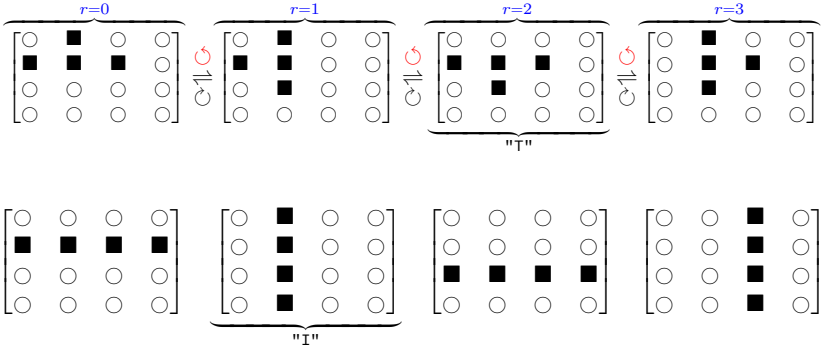
The Rotations

aka, the »Super-Rotation-System«



The Rotations

aka, the »Super-Rotation-System«

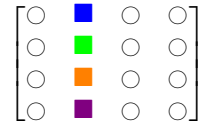


Relative Coordinates

$$\left[\texttt{R-COORD}\right] := \begin{bmatrix} \star & (+0,+1) & (+0,+2) & (+0,+3) \\ (+1,+0) & (+1,+1) & \cdots & \cdots \\ (+2,+0) & \vdots & \ddots & \\ (+3,+0) & \vdots & & (+3,+3) \end{bmatrix}$$

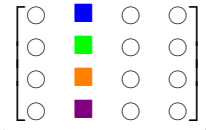
Some Examples

Shengdi Chen
shenchen@ethz.ch

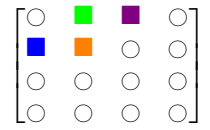


$$:= ((+0,+1),(+1,+1),(+2,+1),(+3,+1))$$

Some Examples



"I" $:= ((+0,+1),(+1,+1),(+2,+1),(+3,+1))$



"S" $:= ((+1,+0),(+0,+1),(+1,+1),(+0,+2))$

Absolute Coordinates

$$\begin{bmatrix} (\star_0, \star_1) & (\star_0 + 0, \star_1 + 1) & (\star_0 + 0, \star_1 + 2) & (\star_0 + 0, \star_1 + 3) \\ (\star_0 + 1, \star_1 + 0) & & \cdots & \cdots \\ (\star_0 + 2, \star_1 + 0) & \vdots & \ddots & \\ (\star_0 + 3, \star_1 + 0) & \vdots & & (\star_0 + 3, \star_1 + 3) \end{bmatrix}$$

Shengdi Chen
shenchen@ethz.ch

Absolute Coordinates

$$\begin{bmatrix} (\star_0, \star_1) & (\star_0 + 0, \star_1 + 1) & (\star_0 + 0, \star_1 + 2) & (\star_0 + 0, \star_1 + 3) \\ (\star_0 + 1, \star_1 + 0) & & \cdots & \cdots \\ (\star_0 + 2, \star_1 + 0) & \vdots & \ddots & \\ (\star_0 + 3, \star_1 + 0) & \vdots & & (\star_0 + 3, \star_1 + 3) \end{bmatrix}$$

$$\begin{bmatrix} \texttt{A-COORD} \end{bmatrix} := \texttt{pos}\,(\star) \mathbin{\widetilde{+}} \begin{bmatrix} \texttt{R-COORD} \end{bmatrix}$$

Full Piece Specification

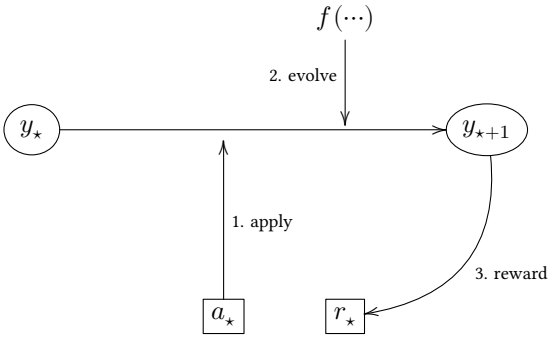
Shengdi Chen
shenchen@ethz.ch

$$\texttt{Piece} := \begin{cases} \texttt{pid} \\ \texttt{rot} + \texttt{pos}\,(\star) \\ \begin{bmatrix} \texttt{A-COORD} \\ \end{bmatrix} \end{cases}$$

The MDP

and the 4-tuple

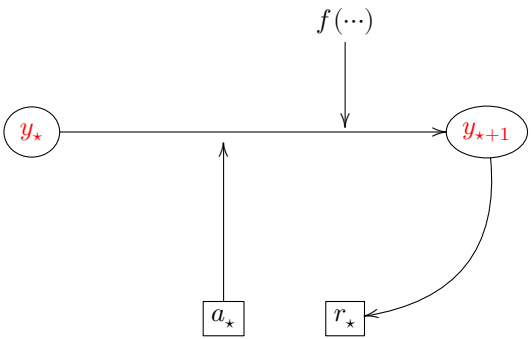
Shengdi Chen
shenchen@ethz.ch



$$\Big((y, Y), (a, A), f, r \Big)$$

State y

Shengdi Chen
shenchen@ethz.ch



$$\Big((y, Y), (a, A), f, r \Big)$$

State y

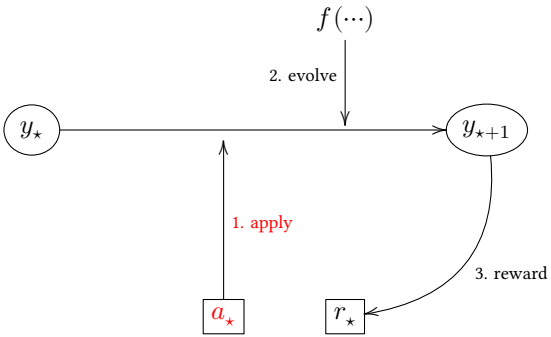
$t_\star \triangleq \star \in \mathbb{N}$

$$y_\star := \left[\begin{array}{ll} y^F & := \texttt{Field} \\ y^P & := \texttt{Piece} \end{array}\right]\Bigg|_{t=t_\star}$$

Shengdi Chen
shenchen@ethz.ch

Action a

Shengdi Chen
shenchen@ethz.ch



$$\Big((y, Y), (a, A), f, r \Big)$$

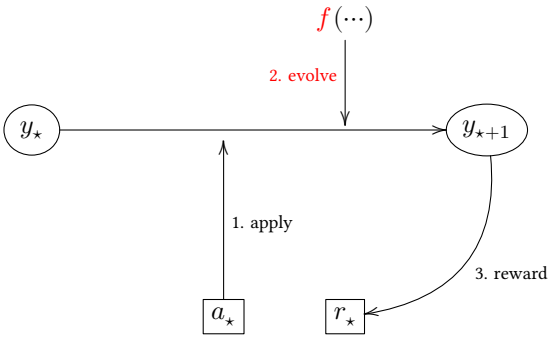
Action a

Time-invariant!

$$\underbrace{\begin{bmatrix} a^{\texttt{rot}} \\ a^{\texttt{pos0}} \\ a^{\texttt{pos1}} \end{bmatrix}}_{a} \in \underbrace{\begin{bmatrix} A^{\texttt{rot}} & := \texttt{range(4)} \\ A^{\texttt{pos0}} & := \emptyset \\ A^{\texttt{pos1}} & := \texttt{range(10)} \end{bmatrix}}_{A}$$

Evolution f

Shengdi Chen
shenchen@ethz.ch



$$\Big((y, Y), (a, A), f, r \Big)$$

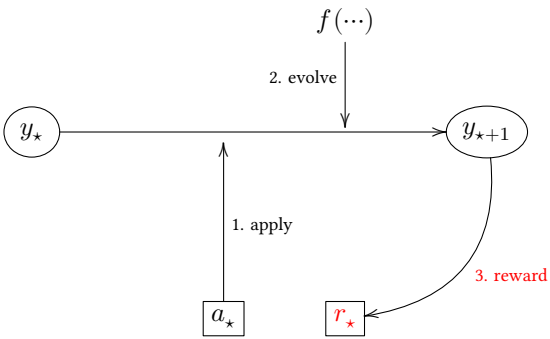
Evolution f

$$\begin{bmatrix} y^P_{\star+1} \\ \tilde{y}^P \\ y^F_{\star+1} \end{bmatrix} := \begin{bmatrix} f^{\text{gen}} \to y^P_\star \\ f^{\text{fall}} \to (a_\star \to y^P_\star) \\ f^{\text{line-clear}} \to \tilde{y}^P \end{bmatrix}$$

$$y := \begin{bmatrix} y^P \\ y^F \end{bmatrix}$$

Reward r

Shengdi Chen
shenchen@ethz.ch



$$\Big((y, Y), (a, A), f, r \Big)$$

Two-Part Reward

$$r_\star := r_\star^{\mathrm{GO}} + r_\star^{\mathrm{LC}}$$

$r^{\mathrm{GO}} :=$ Game-Over

$r^{\mathrm{LC}} :=$ Line-Clears

Shengdi Chen
shenchen@ethz.ch

Game over?

aka, r^{GO}

Shengdi Chen
shenchen@ethz.ch

$$r^{\text{GO}} := \begin{cases} +1 & \text{if episode not terminated} \\ -2 & \text{otherwise} \end{cases}$$

How many full lines n?

aka, r^{LC}

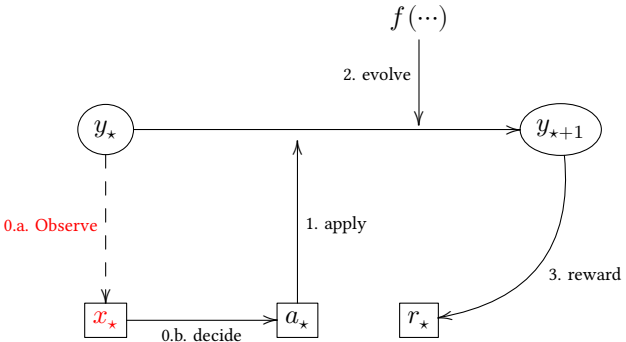
$$k\Big(n \in \{0,1,2,3,4\}\Big) := \begin{cases} 0 & n = 0 \\ 1 & n = 1 \\ 3 & n = 2 \\ 5 & n = 3 \\ 8 & n = 4 \end{cases}$$

$$\begin{aligned} r^{\text{LC}}(n) &\propto k(n) \\ &:= 10 \cdot k(n) \end{aligned}$$

Interfacing with gym

OpenAI, 2016

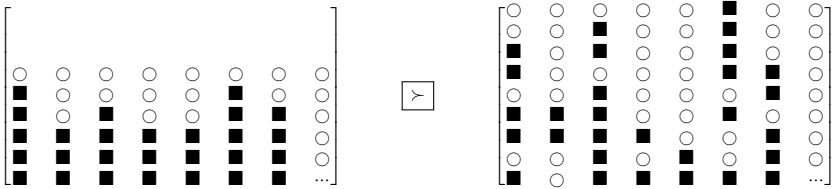
Shengdi Chen
shenchen@ethz.ch



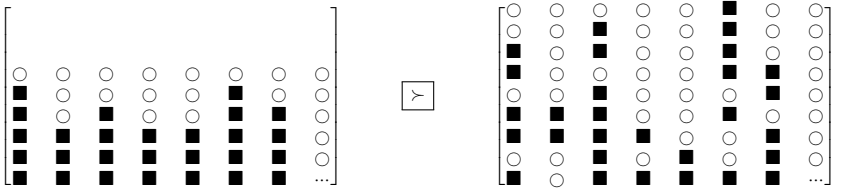
[System Internals]

[Agent Perception]

Good or Bad?



Good or Bad?



$$\begin{bmatrix} x^H \\ x^E \\ x^N \end{bmatrix} := \begin{bmatrix} \text{Height} \\ \text{Elevation} \\ \text{Num-Holes} \end{bmatrix}$$

Column-wise Observation

Field of 20×10

Shengdi Chen
shenchen@ethz.ch

Observations	numbers of	discrete-range	gym-translation
Height x^H	10	$[0; 20]$	`range(21)`^{10}
Elevation x^E	9	$[0; 20]$	`range(21)`^{9}
Holes x^N	10	$[0; 19]$	`range(20)`^{10}

$$\texttt{gym.MultiDiscrete}\Big(\texttt{[}\underbrace{\texttt{21...21|21...21|20...20}}_{\texttt{len}=29}\texttt{]}\Big)$$

Compact Observation

3 values (used to be 29)

Observations	gym-translation
Height $x^{H,C}$	`range(201)`
Elevation $x^{E,C}$	`range(181)`
Holes $x^{N,C}$	`range(191)`

`MultiDiscrete([201, 181, 191])`

$$x^{*,C} := \sum_{\text{columns}} x^{\{H \mid E \mid N\}}$$

Two more things…

Observations	numbers of	discrete-range	gym-translation
Lines-cleared x^{LC}	1	$[0; 4]$	`range(5)`
`pid` $x^{\mathtt{pid}}$	1	$[0; 6]$	`range(7)`

4 Variations of Observation

Shengdi Chen
shenchen@ethz.ch

Variation-Code	Field ($x^{\{H\|E\|N\}}$)	`pid` ($x^{\texttt{pid}}$)
Default	Compact	Included
Variation-1	Compact	Omitted
Variation-2	Full	Included
Variation-3	Full	Omitted

Observing more or less?

Variation-Code	Field ($x^{\{H \mid E \mid N\}}$)	`pid` ($x^{\texttt{pid}}$)
Default	Compact	Included
Variation-1 (Min.)	Compact	Omitted
Variation-2 (Max.)	Full	Included
Variation-3	Full	Omitted

Algorithms

Hill, A. *et al.* (2018)

$$\left.\begin{matrix}\texttt{Action}\\ \texttt{Obs}\end{matrix}\right\} \in \texttt{MultiDiscrete} \quad \Longrightarrow \quad \texttt{sb3.}\begin{cases}\texttt{PPO}\\ \texttt{A2C}\\ \texttt{DQN}\end{cases}$$

$$\texttt{sb3} := \texttt{stable\_baselines3}$$

Adaptations: DQN by sb3

src/rl/shetris/env/shenv.py

Shengdi Chen
shenchen@ethz.ch


```python
class ShetrisEnv:
    ...

    def _action_int_to_np(self, action: int) -> np.ndarray:
        width = self.engine.field.size[1]
        return np.array((action // width, action % width))
```

$$\underbrace{\tilde{a}_{\texttt{sb3-DQN}}}_{\mathbb{N}^1} \longmapsto \begin{bmatrix} a^{\texttt{rot}} \\ a^{\texttt{pos1}} \end{bmatrix} \in \mathbb{N}^2$$

Q-Learning

Watkins, C. *et al.* (1992)

$$\begin{cases} R_{k:T} & := \sum_{t \in [t_k; t_T]} \gamma^{t-k} r_t \\ Q_{k:T} & := \mathbb{E}\left[R_{k:T} \mid x_k, a_k, \pi\right] \end{cases}$$

Q-Learning

Watkins, C. *et al.* (1992)

$$\begin{cases} R_{k:T} & := \sum_{t \in [t_k; t_T]} \gamma^{t-k} r_t \\ Q_{k:T} & := \mathbb{E}\left[R_{k:T} \mid x_k, a_k, \pi \right] \end{cases}$$

$$\pi^* := \arg\max_{\pi} \left\{ Q_{k:T} \right\}$$

Q-Learning

Watkins, C. *et al.* (1992)

$$Q^* \overset{\text{Q-Learning}}{:=} \max_{\pi} \{Q_{k:T}\}$$

Deep Q-Learning

Mnih, V. *et al.* (2013)

$$Q^* \overset{1992}{:=} \max_{\pi} \{Q_{k:T}\}$$
$$\overset{2013}{:\approx} \max \left\{ \hat{Q}\left(x, a; \theta\right) \right\}$$

Obs-Variation	First-Layer
Default	5
Var-1 (Min.)	4
Var-2 (Max.)	31
Var-3	30

Action-Space Tuning

Raw $\Rightarrow$ Time-invariant

Shengdi Chen
shenchen@ethz.ch

$$A^{\text{raw}} := \begin{bmatrix} A^{\texttt{rot}} & := \texttt{range(4)} \\ A^{\texttt{posi}} & := \texttt{range(10)} \end{bmatrix}$$

Action-Space Tuning

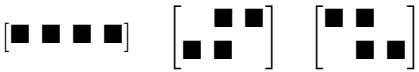
Raw ⇒ Time-invariant

Shengdi Chen
shenchen@ethz.ch

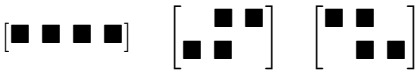
$$A^{\text{raw}} := \begin{bmatrix} A^{\texttt{rot}} & := \texttt{range(4)} \\ A^{\texttt{pos1}} & := \texttt{range(10)} \end{bmatrix}$$

Piece-specific Tailoring!

Tuning Example: $A^{\texttt{rot}}$

Illustration	pid	A^{raw}
	0	4
	[1; 3]	4
	[4; 6]	4

Tuning Example: $A^{\texttt{rot}}$

Illustration	pid	A^{raw}	A^{tuned}
	0	4	1
	$[1;3]$	4	2
	$[4;6]$	4	4

Action-Space Tuning, cont'd

Shengdi Chen
shenchen@ethz.ch

$$A^{\text{raw}} \equiv \begin{bmatrix} \texttt{range(4)} \\ \texttt{range(10)} \end{bmatrix}$$

Action-Space Tuning, cont'd

$$A^{\text{raw}} \equiv \begin{bmatrix} \texttt{range(4)} \\ \texttt{range(10)} \end{bmatrix}$$

$$A^{\text{tuned}} := \begin{bmatrix} A^{\texttt{rot}}\left(\texttt{pid}\right) \\ A^{\texttt{pos1}}\left(\texttt{pid, rot}\right) \end{bmatrix}$$

$$\subsetneq \quad A^{\text{raw}}$$

Reference Solution

Lee, 2018

$$\begin{bmatrix} x^{H,C} \\ x^{E,C} \\ x^{N,C} \\ x^{\mathrm{LC}} \end{bmatrix} =: x_\star \xrightarrow{\pi} a_\star := \arg\max\left(r_\star\right)$$

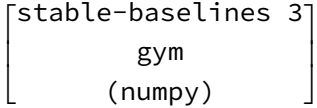
$$r := \begin{bmatrix} -0.51 & -0.18 & -0.35 & +0.76 \end{bmatrix}^\top \circ x$$

Quick Setup

Python >=3.10


```
# install packages
$ pipenv run; pipenv install

$ pipenv shell
```



Training Constraints

$$\text{Euler:} \begin{cases} 40 \text{ hours} \\ 10 \text{ mil. steps} \end{cases}$$

Benchmark: How many pieces n_P

Field of 20×10

n_P	Training Impact	General Assessment
< 20	Negative	Insufficient
$[20; 50]$	Positive	
$[50; 100]$		
> 100		Successful

sb3: 3 Algorithms, 4 modes

$n_P :=$ Number of Pieces

Variation	PPO	A2C	DQN
Default	42	37	27
Variation-1	27	27	22
Variation-2	55	46	45
Variation-3	31	34	43

Shengdi Chen
shenchen@ethz.ch

Recall: benchmark n_P

n_P	Training Impact	General Assessment
< 20	Negative	Insufficient
$[20; 50]$	Positive	
$[50; 100]$		
> 100		Successful

Shengdi Chen
shenchen@ethz.ch

sb3: positive training, but insufficient perf.

PPO + Var-2 (max. observation)

Shengdi Chen
shenchen@ethz.ch

Variation	PPO	A2C	DQN	Best Alg.
Default	42	37	27	PPO
Variation-1	27	27	22	PPO
Variation-2	55	46	45	PPO
Variation-3	31	34	43	DQN
Best Var.	Var-2	Var-2	Var-2	

Self-Tuned DQN

$n_L :=$ num. lines cleared

Variation	n_P	n_L
Default	367.94	134.78
Variation-1	669.04	255.43
Variation-2	154.05	48.03
Variation-3	180.14	58.60

Successful!

$\hookrightarrow \quad n_P^{\mathsf{sb3}} \leq 55$

Which Observation?

Best Perf'er	Tuned-DQN	`sb3`
Obs-Variant	Var-1 (Min.)	Var-2 (Max.)
Obs-Size	4	31
n_P	669.04	55

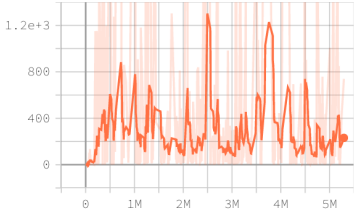
Shengdi Chen
shenchen@ethz.ch

Graph during training

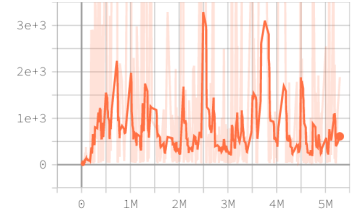
Variation-1 (Min. Observation)

Shengdi Chen
shenchen@ethz.ch

steps/Number of Lines
tag: steps/Number of Lines



steps/Number of Pieces
tag: steps/Number of Pieces

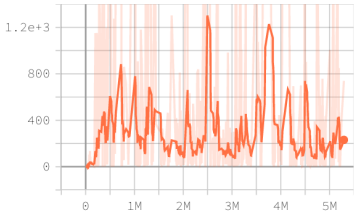


Graph during training

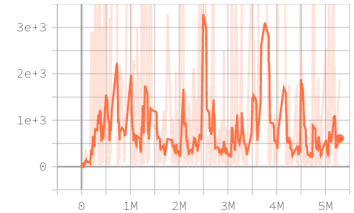
Variation-1 (Min. Observation)

Shengdi Chen
shenchen@ethz.ch

steps/Number of Lines
tag: steps/Number of Lines



steps/Number of Pieces
tag: steps/Number of Pieces



$$\text{Fluctuations} \triangleq \begin{cases} \text{High Peaks} \\ + \\ \text{Low Trenches} \end{cases}$$

The bags of the `pid`-Generator

`perm()` $\triangleq$ source of randomness

$$\overbrace{\texttt{perm(7)}}^{1^{\text{st}}\text{ bag}} \;\Big|\; \underbrace{\texttt{perm(7)}}_{2^{\text{nd}}\text{ bag}} \;\Big|\; \cdots$$

The bags of the `pid`-Generator

`perm()` $\triangleq$ source of randomness

$$\overbrace{\texttt{perm(7)}}^{1^{\text{st}}\text{ bag}} \;\Big|\; \underbrace{\texttt{perm(7)}}_{2^{\text{nd}}\text{ bag}} \;\Big|\; \cdots$$

Says the Tetris-Guideline...

Tweaking the generator

with the bag

Shengdi Chen
shenchen@ethz.ch

$$\texttt{pid-Bag} \in \begin{cases} \texttt{\{"O", "I"\}} \\ \texttt{\{"O", "I", "T"\}} \\ \text{7-bag} \triangleq \texttt{range(7)} \end{cases}$$

Tweaking the generator

with the bag & stochasticity

$$\text{Bag}\,(\text{random}) \in \begin{cases} \texttt{\{"O", "I"\}} \\ \texttt{\{"O", "I", "T"\}} \\ \text{7-bag} \end{cases}$$

$$\text{Bag}\,(\text{non-random}) \in \begin{cases} \texttt{"O"} \\ \texttt{"I"} \\ \texttt{\{"O", "I"\}} \\ \texttt{\{"O", "I", "T"\}} \\ \text{7-bag} \end{cases}$$

Non-Random Generator

The Benchmark: n_P

`pid`-Bag	Tuned DQN (ours)	Reference (theirs)
`"O"`	17	> 500
`"I"`	> 500	> 500
`{"O", "I"}`	> 500	> 500
`{"O", "I", "T"}`	307	416
Full-7	97	> 500

Shengdi Chen
shenchen@ethz.ch

(Generally) Successful!

The Benchmark: n_P

Random Bag	Avg.	Max.	Min.	Ref. (theirs)
`{"O", "I"}`	> 500	> 500	> 500	> 500
`{"O", "I", "T"}`	> 500	> 500	143	> 500
Full-7	215.3	> 500	41	$\min = 311$

Shengdi Chen
shenchen@ethz.ch

What was done…

$$12\ \text{weeks} \triangleq \begin{cases} \texttt{~14 GB} & \text{saveload data} \\ 300 & \text{node-hours} \end{cases}$$

What was done…

$$12 \text{ weeks} \triangleq \begin{cases} \texttt{~14 GB} & \text{saveload data} \\ 300 & \text{node-hours} \end{cases}$$

$$4 \text{ obs-modes } @ \begin{cases} \texttt{sb3} := (\texttt{PPO+A2C+DQN}) \\ + \\ \text{Tuned DQN} \end{cases}$$

What is now known…

	Tuned-DQN	`sb3`
Action-Space	TETRIS-Tailored	General (Raw)
Obs-Variation	Minimal	Maximal
Training n_P	669.04	55
Verdict	Successful	Insufficient

What remains to be explored…

TETRIS-Internals	Testing	Algorithms
`pid`-previewing	agent-on-agent	other algs & libs of RL
`pid`-swapping	Versus-Mode	(Semi-) Superv'd, etc.

What remains to be explored…

Tetris-Internals	Testing	Algorithms
`pid`-previewing	agent-on-agent	other algs & libs of RL
`pid`-swapping	Versus-Mode	(Semi-) Superv'd, etc.

Semester-Thesis?

;)

The Gist

B.Sc. Thesis @ ISE-Group	Prof. Buhmann; Ivan, Ami, Eugene
Mastering Tetris with Reinforcement-Learning	Modeling and MPD Benchmarking Framework

The Gist

B.Sc. Thesis @ ISE-Group	Prof. Buhmann; Ivan, Ami, Eugene
Mastering Tetris with Reinforcement-Learning	Modeling and MPD Benchmarking Framework
Raw `sb3` (`PPO + A2C + DQN`)	↯ insufficient performance ↯ Invariant Action-Sp.

The Gist

B.Sc. Thesis @ ISE-Group	Prof. Buhmann; Ivan, Ami, Eugene
Mastering Tetris with Reinforcement-Learning	Modeling and MPD Benchmarking Framework
Raw `sb3` (`PPO + A2C + DQN`)	↯ insufficient performance ↯ Invariant Action-Sp.
Self-implt'd DQN (Tuned Action-Sp. + Min. Obs-Sp.)	✓ successful results ✓ Suggestion for future work

Open (Randomized) Discussion

Shengdi Chen
shenchen@ethz.ch