Mastering Tetris with Reinforcement-Learning

B.Sc. Thesis Presentation

Shengdi Chen shenchen@ethz.ch

2022-08-09

Mastering Tetris
with

Reinforcement-Learning

Shengdi Chen shenchen@ethz.ch

Synopsis

Quick-Start

Setup

Three Agen

.....

.

- 1. Quick-Start
- 2. From Zero to gym
 - 2.1 MDP
 - 2.2 gym-adaptation
- 3. Setup
 - 3.1 Training
 - 3.2 Testing

- 4. Result
 - 4.1 SB3
 - 4.2 DQN
 - 4.3 Reference
- 5. Summary

Synopsis

Synopsis

Quick-Start

Status-Quo

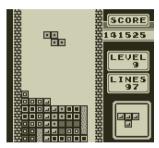
Overviev

etup

Three Agent

Close

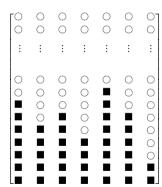






The Field

Just some 0s and 1s



Mastering Tetris with

Reinforcement-Learning

The Field

Just some 0s and 1s

The Field

$$\iff$$

$$\iff \qquad \left(\{\bigcirc, \blacksquare\} \equiv \mathsf{bool}\right)^{\mathsf{#rows} \times \mathsf{#columns}}$$

Indexing by numpy

Representatio

The Field

iece Config: loordinates

ordinates ce

P

.... Ato

loser

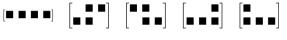
 $\begin{cases} & \text{#rows} := 20 \\ & \text{#columns} := 10 \end{cases}$

Piece Confias















pid = 6

Representation

Piece Configs

Piece Configs
Coordinates

Piece

MDP gym-adaptatio

Three Agents

a. de

losor



$$pid \in \{0, 1, \dots, 6\}$$

 $\triangleq range(7)$



etup

. Poprorontatio

ne Field

Piece Confias

oordinates

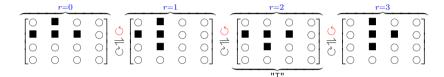
rce

DP

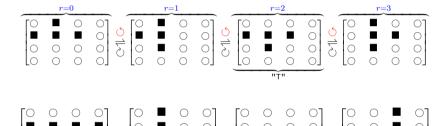
gym-adaptatio

nree Agents

esult



Piece Confias



"Ĭ"

etun

Representatio

e Field

Piece Configs Coordinates

ce

P . adaptation

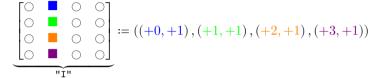
Three Agente

...--..

esult

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

Coordinates



epresentatio

riece Configs

Coordinates Piece

MDP

Three Agent

esult

Closer

$$\begin{bmatrix}
\bigcirc & \bullet & \circ & \circ \\
\circ & \bullet & \circ & \circ \\
\circ & \bullet & \circ & \circ \\
\circ & \bullet & \circ & \circ
\end{bmatrix} := ((+0, +1), (+1, +1), (+2, +1), (+3, +1))$$

$$\underbrace{(+0, +1), (+1, +1), (+2, +1), (+3, +1)}_{\text{TI}"}$$

.....

Setup

presentatio no Field

ece Configs

Coordinates Piece

)P

ym-adaptatior

ree Agents

esult

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

etup Representat

he Field

Coordinates Piece

MDP

gymadaptation

ree Agents

. . .

esult

Closer

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

.....

$$\left[A-COORD \right] := pos(\star) \widetilde{+} \left[R-COORD \right]$$

Synopsis

Quick-Start

Representation

The Field

iece Confias

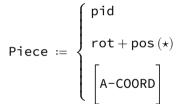
oordinates

Piece

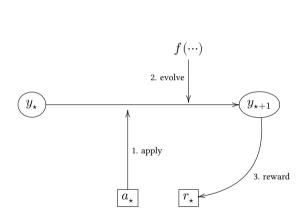
ИDP

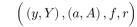
Tl.... A......

Result



Overview





Synopsis

Quick-Sta

Setup

Representa

/IDP

Overview

State

ction

volution

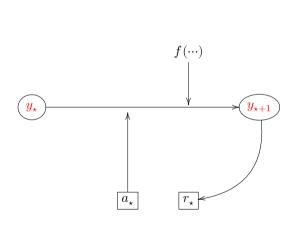
ym-adaptation

hree Agents

mee Agem

.....

loser



 $\left(\left(\boldsymbol{y},\boldsymbol{Y}\right),\left(a,A\right),f,r\right)$

Mastering Tetris with

Reinforcement-Learning

shenchen@ethz.ch

Synopsis

Quick-Star

Setup

Representation

MDP

Overview

State

ction

volution

Reward

gym-adaptatio

ree Agents

nree Agents

esult

Close

Synopsi

Quick-Star

Setup

Representatio

ИDP

State

Action

olution

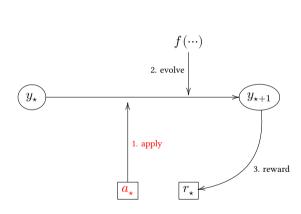
eward

m-adaptation

hree Agents

logult.

Closer



 $\Big(\left(y,Y\right),\left({\color{red}a,A}\right),f,r\Big)$

```
Synopsi
```

Setup

Representation

Overview

State

Action

volution

Reward

gym-adaptatio

hree Agents

..........

Rest

Close

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

.

Representatio

MDP

Overview State

ction

Evolution

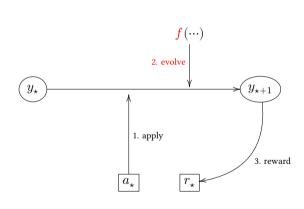
reward

ym-adaptation

Three Agents

. .

locar



 $\Big(\left(y,Y\right),\left(a,A\right),{\color{red}f},r\Big)$

Setup

Representation

Overview

State Action

Action Evolution

Reward

ym-adaptation

hree Agents

Pocult

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

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

.

Representatio

IDP

State

ction

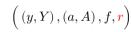
Reward

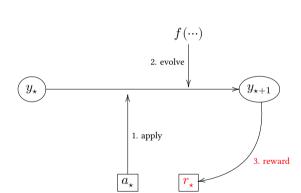
/m-adaptation

hree Agents

...--..

.





Reward

$$r_\star \coloneqq r_\star^{\rm GO} + r_\star^{\rm LC}$$

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

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

Mastering Tetris with Reinforcement-Learning

Shenadi Chen

Synopsis

Quick-Star

Setup

Representation

tepresentation

Overview

Action

ction

Reward

vm-adaptati

hree Agents

.

iesuit

Reward

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

$$r^{\text{LC}}\left(n\right) \propto k\left(n\right)$$

$$:= 10 \cdot k\left(n\right)$$

OpenAl, 2016

Quick-Star

Setup

Representat

DP

n-adaptation

Introduction Field-based

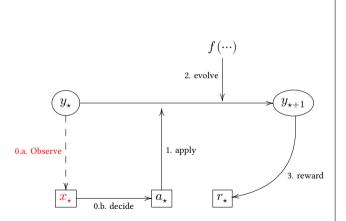
thers

our Modes

ree Agents

seult.

locar



[System Internals]

[Agent Perception]

Shengdi Chen shenchen@ethz.ch

ynopsis

Quick-Start

Setup

Representation

DP

ym-adaptation

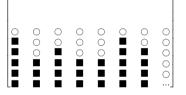
Field-based

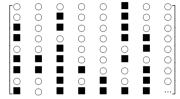
Four Modes

Three Agents

Result

oser





Synopsis

Quick-Start

Setup

Representation

IDP

Introduction

Field-based

Four Modes

hree Agents

esult

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

Column-wise Observation

Field of 20×10

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

gym.MultiDiscrete
$$\left(\left[\underbrace{21...21 | 21...21 | 20...20}_{\text{len} = 29} \right] \right)$$

Mastering Tetris with Reinforcement-Learning

Shengdi Chen

ynopsis

Quick-Start

etup

epresentation

adaptation

Introduction Field-based

Others

Four Modes

ree Agen

esult

3 values (used to be 29)

IDP

m-adaptation troduction

Field-based Others

Four Modes

ree Agents

. .

suit

Observations	gym-translation			
Height $x^{H, \mathbb{C}}$	range(201)			
Elevation $x^{E, \mathbb{C}}$	range(181)			
Holes $x^{N, \mathbb{C}}$	range(191)			
MultiDiscrete([201, 181, 191])				

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

etup

epresentation

IDP

Introduction

Field-based

Others

... A....ta

ree Agent

esult

oser

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

Synopsis

Quick-Start

etup

Representation

1DP

-adaptation

ntroduction

Others

Four Modes

ree Agents

.

sult

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

etup

Representation

ym-adaptati

ntroduction

Field-based

Four Modes

ree Agents

ilee Ageilts

sult

1----

Variation-CodeField $(x^{\{H|E|N\}})$ pid (x^{pid}) DefaultCompactIncludedVariation-1 (Min.)CompactOmittedVariation-2 (Max.)FullIncludedVariation-3FullOmitted

Catura

Three Ag

SB3

Overview

Specials

dapted-DQ

.

Result

Closer

 $\begin{array}{c} \mathsf{Action} \\ \mathsf{Obs} \end{array} \} \in \mathsf{MultiDiscrete} \qquad \Longrightarrow \qquad \mathsf{sb3.} \begin{cases} \mathsf{PPO} \\ \mathsf{A2C} \\ \mathsf{DQN} \end{cases}$

sb3 := stable_baselines3

src/rl/shetris/env/shenv.py

Specials

```
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))
```

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

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

shenchen@ethz.ch

Synopsis

Quick-Start

etup

Three Age

nee Agent

dapted-DQN

Theory

Theory

Adaptations

eference

esult

oser

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

$$\pi^* \coloneqq \argmax_{\pi} \left\{ Q_{k:T} \right\}$$

Reinforcement-Learning

Theory

Theory

```
Q^* \stackrel{	ext{Q-Learning}}{:=}
                              \max_{\pi} \left\{ Q_{k:T} \right\}
```

1000

Theory

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

Obs-Variation	First-Layer
Default	5
Var-1 (Min.) Var-2 (Max.)	$\frac{4}{31}$
Var-3	30

Synopsis

Quick-Star

Caren Sta

Three Aa

Three Age

SB3

dapted-DQN --

Tuning

Adaptations

acult

oser

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

$$A^{\mathrm{raw}} \coloneqq egin{bmatrix} A^{\mathrm{rot}} &\coloneqq \mathtt{range}(4) \ A^{\mathtt{posl}} &\coloneqq \mathtt{range}(10) \end{bmatrix}$$

Piece-specific Tailoring!















Mastering Tetris with

Reinforcement-Learning

Illustration	pid	$A^{ m raw}$	
	0	4	
$[\blacksquare \blacksquare \blacksquare \blacksquare]$	[1;3]	4	
	[4;6]	4	

Illustration		pid	A^{raw}	A^{tuned}
		0	4	1
$[\blacksquare \blacksquare \blacksquare \blacksquare]$		[1; 3]	4	2
	$\left[\begin{smallmatrix}\bullet&\bullet\\\bullet&\bullet&\bullet\end{smallmatrix}\right]$	[4; 6]	4	4

 $A^{
m raw} \equiv \left[egin{array}{c} {
m range}(4) \ {
m range}(10) \end{array}
ight]$

Synopsis

Quick-Start

Catum

Three Age

nree Agei

13

apted-DQN

Tuning

daptations

Result

oser

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

$$A^{ ext{tuned}} \coloneqq egin{bmatrix} A^{ ext{rot}} \left(ext{pid}
ight) \ A^{ ext{pos1}} \left(ext{pid}, ext{rot}
ight) \end{bmatrix}$$



Quick-Star

etun

hree Ager

B3

eference

lesult!

oser

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

.....

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

Python >=3.10

```
Synopsis
```

Quick-Star

Setup

Three Agents

Result Setup

Training

Testin SB3

DQN SB3

Closer

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

\$ pipenv shell

```
[stable-baselines 3]
gym
(numpy)
```

Quick-Sta

Three Agents

Setup

Training

SB3

DQN

Close

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	
[20; 50]	[]	Insufficient
[50; 100]	Positive	
> 100		Successful

Mastering Tetris with <u>Reinf</u>orcement-Learning

Shengdi Chen

Synopsis

Quick-Start

Setup

Three Agents

esult

aining

Testing SB3

DQN B3

loser

sb3: 3 Algorithms, 4 modes

 $n_P \coloneqq \mathsf{Number} \; \mathsf{of} \; \mathsf{Pieces}$

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

Mastering Tetris with

Reinforcement-Learning

Synopsis

Quick-Sta

C - - - - -

Three Agents

etup

etup

SB3 Details

rtails inclusion

onclusion

locar

Closer

Quick-Start

Setup

Inree Agents

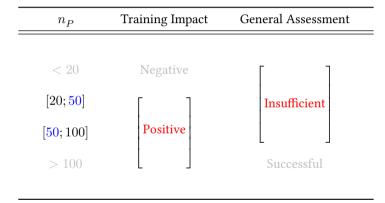
Setup

Details

Conclusion

.

Closer



sb3: positive training, but insufficient perf.

PPO + Var-2 (max. observation)

Reinforcement-Learning

Shengdi Chen
shenchen@ethz.ch

Mastering Tetris

with

.

Three Agents

esult etup

B3

Details Conclusion

Conclusion ON

Closer

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	

 $n_L \coloneqq \mathsf{num}.$ lines cleared

etup B3

OQN

Training

Generator:

loser

 \hookrightarrow

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

Successful!

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

Setup

Three Agents

Result Setup

SB3

ΩN

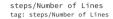
Training

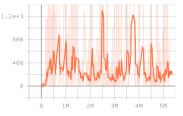
Generator

loser

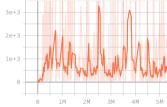
Best Perf'er	Tuned-DQN
Obs-Variant	Var-1 (Min.)
Obs-Size	4
n_P	669.04

sb3
Var-2 (Max.)
$\frac{31}{55}$





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



Mastering Tetris with

Reinforcement-Learning

Shengdi Chen shenchen@ethz.ch

Synopsis

Quick-Star

Satur

Three Agents

Result

SB3

DQN Training

Training Generator:

oser

steps/Number of Lines

Synopsis

Quick-Star

C - 4----

Three Agents

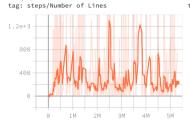
Setup SB3

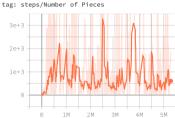
DQN

Training Generator

Generator: Testing

ser





steps/Number of Pieces

.....

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

The bags of the pid-Generator

 $perm() \triangleq source of randomness$

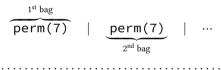
```
perm(7)
```

Mastering Tetris with

Reinforcement-Learning

Generators

 $perm() \triangleq source of randomness$



Says the Tetris-Guideline...

Mastering Tetris with

Reinforcement-Learning

shenchen@ethz.ch

Synopsis

Quick-Sta

Setup

Three Agents

Popult

Setup

DON SB3

Training

Generators

Testing

loser

```
ynopsis
```

Quick-Star

.

Three Agents

```
Result
```

SB3

NDC

Training

Generators Testing

oser

 $\mathsf{pid}\text{-Bag} \in \left\{ \begin{array}{l} \{ \texttt{"0", "I"} \} \\ \\ \{ \texttt{"0", "I", "T"} \} \\ \\ \text{7-bag} \ \triangleq \ \mathsf{range}(7) \end{array} \right.$

```
Synopsis
```

Quick-Star

etup

Three Agent

Result Setup

SB3

QN Training

Generators

esting

ioser

```
Bag (random) \in \begin{cases} \{"0", "I"\} \\ \{"0", "I", "T"\} \end{cases}
7-bag
```

.....

```
Bag (non-random) \in \begin{cases} "0" \\ "I" \\ \{"0", "I"\} \\ \{"0", "I", "T"\} \end{cases}
```

pid-Bag	Tuned DQN (ours)
"0"	17
"I"	> 500
{"0", "I"}	> 500
{"0", "I", "T"}	307
Full-7	97

Reference (theirs)
> 500
> 500
> 500
416
> 500

Mastering Tetris with

Reinforcement-Learning

Synoneie

Quick-Sta

etup

Three Agents

sult

rtup 33

žΝ

aining

Generator Testing

oser

The Benchmark: n_P

Quick-Sta

.

Three Agent

esult

etup B3

an QN

aining

Generator: Testing

sting

loser

Random Bag	Avg.	Max.	Min.
{"0", "I"}	> 500	> 500	> 500
{"O", "I", "T"}	> 500	> 500	143
Full-7	215.3	> 500	41

I	Ref. (theirs)
	> 500
	> 500
:	min = 311

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

shenchen@ethz.ch

Synopsis

Quick-Start

Setup

hree Ager

Result

esult

Closer

Look-Back Look-Ahead

ımmary

4 D > 4 P > 4 E > 4 E > E 990

ree Age

Three Agen

Result

Look-Back Look-Ahead

```
12 weeks \triangleq \begin{cases} \sim 14 \text{ GB} & \text{saveload data} \\ 300 & \text{node-hours} \end{cases}
```

.....

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

Quick-Sta

Setup

Three Agen

Closer

Look-Back

Look-Ahead

General (Raw)

Maximal

55

Insufficient

Quick-Star

hree Aae

IIIIco Agoi

Closer

Look-Back Look-Ahead

ummary

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

What remains to be explored...

Shengdi Chen	

Quick-Star

Thurs An

Three Age

esult

Closer

Look-Ahead

ummary

pid-previewing agent-on-agent	- 41 1 9- 1:1 CDI
pid-swapping Versus-Mode	other algs & libs of RL (Semi-) Superv'd, etc.

.....

Semester-Thesis?

;)

Quick-Sta

etup

Three Agen

an all

Closer

Look-Back Look-Ahead Summary

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

Quick-Star

Cotup

Three Agen

. .

Closer

Look-Back Look-Aheac

Look-Ahead Summary

B.Sc. Thesis @ ISE-Group	Prof. Buhmann; Ivan, Ami, Eugene
Mastering Tetris	Modeling and MPD
with Reinforcement-Learning	Benchmarking Framework
Raw sb3	5 insufficient performance
(PPO + A2C + DQN)	4 Invariant Action-Sp.

setup

Three Agen

Closer

Look-Back

Look-Ahead

Prof. Buhmann; Ivan, Ami, Eugene
Modeling and MPD
Benchmarking Framework
4 insufficient performance
5 Invariant Action-Sp.
✓ successful results
√ Suggestion for future work

Closer

Summary

