南京航空航天大學

Nanjing University of Aeronautics and Astronautics

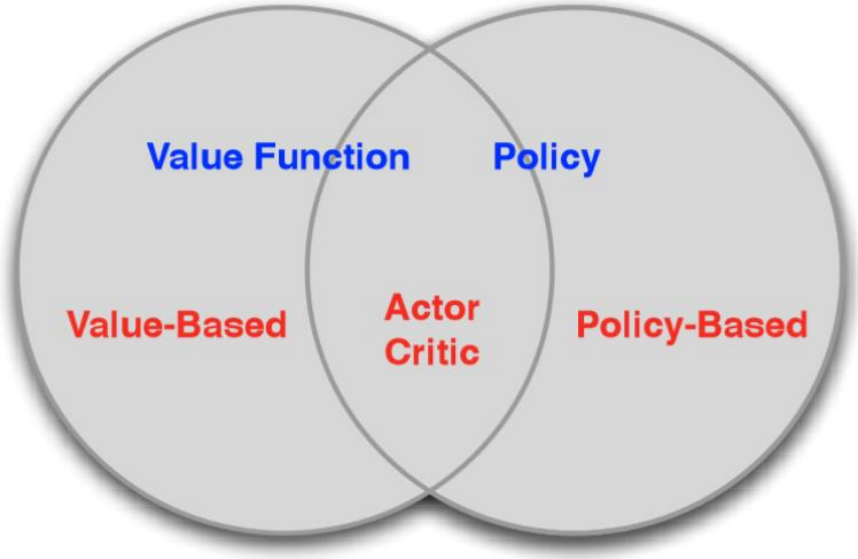
Mastering Complex Control in MOBA Games with Deep Reinforcement Learning

Deheng Ye, Zhao Liu, Mingfei Sun etc.

AAAI 2019

background - model free RL

- Value Based
  - Learnt Value Function
  - Implicit policy (e.g. ϵ-greedy)
- Policy Based
  - No Value Function
  - Learnt Policy
- Actor-Critic
  - Learnt Value Function
  - Learnt Policy



background - Policy gradient

$$\pi^* = \arg\max_{\pi} E_{\tau\sim\pi}[R(\tau)] \qquad \tau = s_1, a_1, r_1, s_1, a_2 \ldots \qquad R(\tau) = \sum_i r_i$$

$$J(\theta) = \sum_{\tau} \pi_\theta(\tau) R(\tau)$$

$$\nabla_\theta J(\theta) = \sum_{\tau} \nabla_\theta \pi_\theta(\tau) R(\tau)$$

$$= \sum_{\tau} \nabla_\theta \pi_\theta(\tau) \frac{\pi_\theta(\tau)}{\pi_\theta(\tau)} R(\tau)$$

$$= \sum_{\tau} \pi_\theta(\tau) \nabla_\theta \log \pi_\theta(\tau) R(\tau)$$

$$= E[\nabla_\theta \log \pi_\theta(\tau) R(\tau)] \qquad \pi_\theta(\tau) = \pi(s_1)\pi(a_1|s_1,\theta)\pi(r_1, s_2|s_1, a_1)\pi(a_2|s_2,\theta)\ldots$$

$$= E_t[\nabla_\theta \log \pi_\theta(a_t|s_t) R(\tau)]$$

$$J(\theta) = E_t[\log \pi_\theta(a_t|s_t)(R(\tau) - b)]$$

background - Actor-Critic

Advantage function: $A(s,a) = Q(s,a) - V(s)$

Advantage actor-critic: $J(\theta) = E_t[\log \pi_\theta(a_t|s_t) A_t]$ on-policy

Importance Sampling : $E_{x\sim p}[f(x)] = \int f(x)p(x)dx = \int f(x)\frac{p(x)}{q(x)}q(x)dx = E_{x\sim q}[f(x)\frac{p(x)}{q(x)}]$

$$J(\theta) = E_t[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}A_t]$$ off-policy

TRPO : $\underset{\theta}{\text{maximize}}\, \hat{\mathbb{E}}_t\left[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}\hat{A}_t - \beta\,\text{KL}[\pi_{\theta_{\text{old}}}(\cdot \mid s_t), \pi_\theta(\cdot \mid s_t)]\right]$

PPO : $L^{CLIP}(\theta) = \hat{\mathbb{E}}_t\left[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)\right]$

introduction

Two-agent, one vs. another
Many game units : turrets, creeps, heros, etc.

MOBA 1v1 games is more appropriate to study the problem of complex control than 5v5 games.

introduction

- **Complexity of the game**
  - Enormous action space
  - Enormous state space
  - Real time
  - Playing method
  - Target selection
  - Little high-quality human data

Table 1: Comparing Go and MOBA 1v1

Game	**Go 1v1**	**MOBA 1v1**
Action space	$250^{150} \approx 10^{360}$ (250 pos available, 150 decisions per game on average)	10^{18000} (100+ discretized actions, 9,000 frames per game)
State space	$3^{361} \approx 10^{170}$ (361 pos, 3 states each)	$2^{2000} \approx 10^{600}$ (2 heroes, (1000+ pos)*(2+ states))
Human player data	rich, high-quality	little
Peculiarity	long-term tactics	real-time, complex control

Large-scale system for exploration 、 Unified modeling、 Self-play

System Design

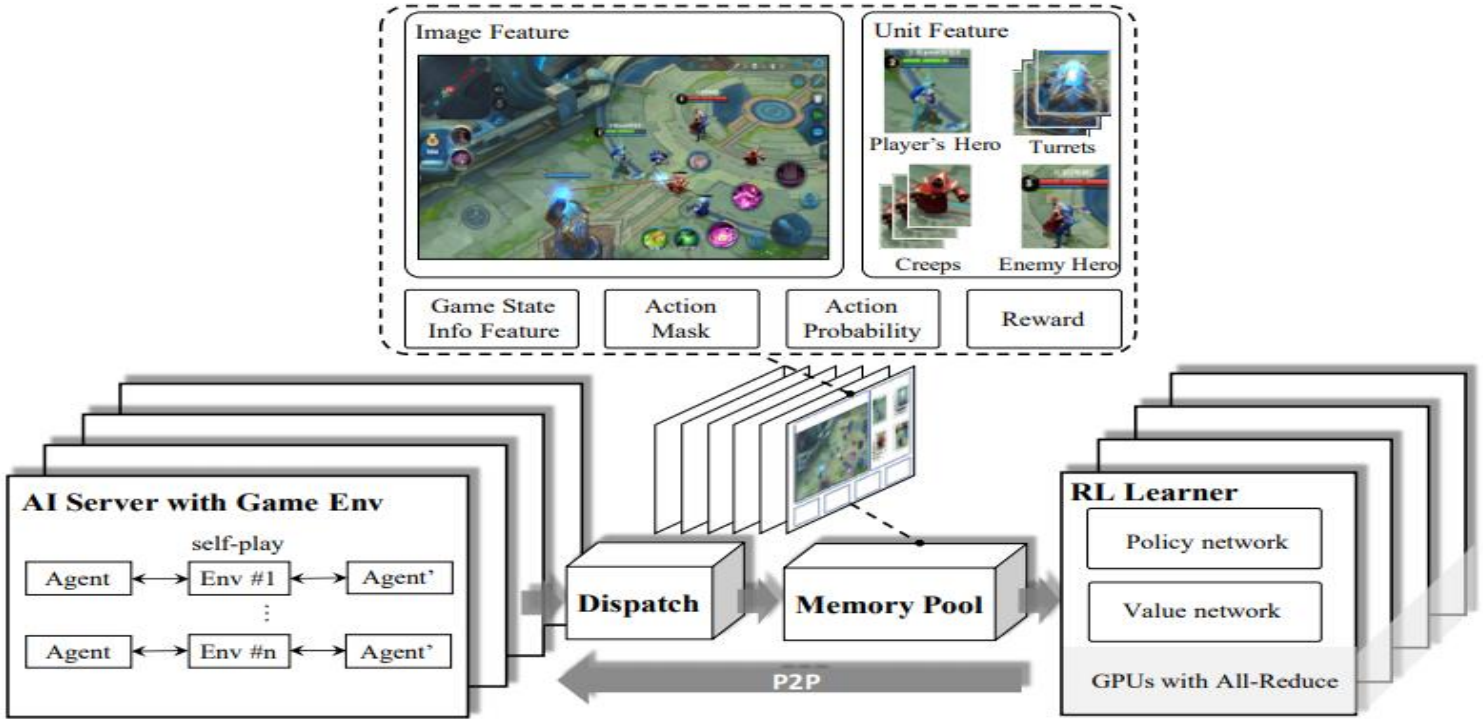


Figure 1: Overview of our System Design

- **Large-scale**

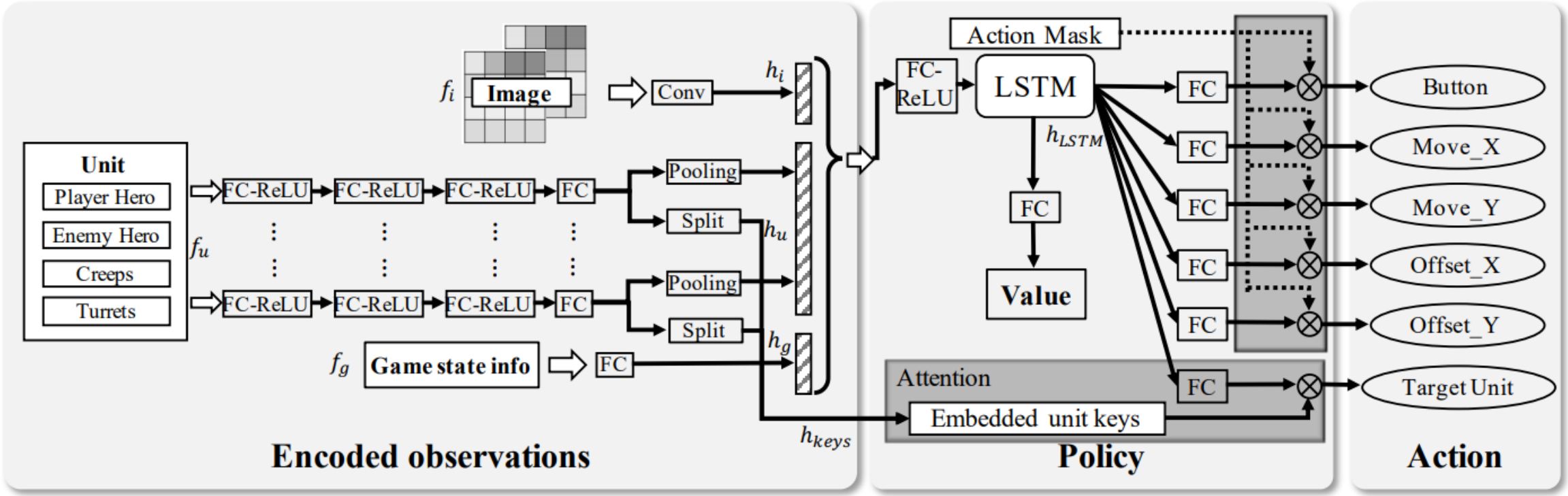
  1000+GPU cards

  500,000+CPUs

- **off policy**

  Actor highly decoupled from Learner

network



Input: observations/features

Internal: neural network model

Output: hero actions

Output:

- Hierarchical, multi-label
  - First, predict **which action** to take, i.e., Button
    - E.g., move
  - Second, predict **how to execute** that action
    - E.g., the direction to move

algorithm

Multi-label PPO (proximal policy optimation)

$$Maximize_{\theta} \sum_{label\_i} E_{s_t, a_t \sim \pi_{\theta_{old}}} [\frac{\pi_{\theta}(a_t^{label\_i} | s_t)}{\pi_{\theta_{old}}(a_t^{label\_i} | s_t)} (R - V_{\theta_{old}}(s_t))] - \frac{1}{2} E_{s_t \sim \pi_{\theta_{old}}} [(R - V_{\theta}(s_t)]^2$$

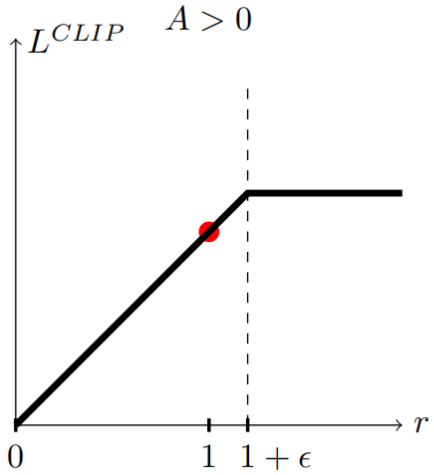
Label treated independently

ppo

value loss

disadvantage of standard PPO

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t\left[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)\right]$$



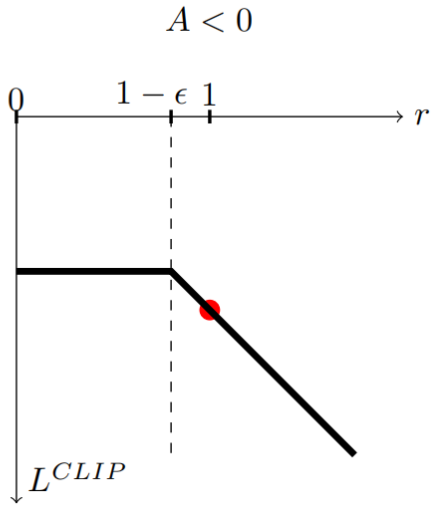


Table 1: Comparing Go and MOBA 1v1

Game	**Go 1v1**	**MOBA 1v1**
Action space	$250^{150} \approx 10^{360}$ (250 pos available, 150 decisions per game on average)	10^{18000} (100+ discretized actions, 9,000 frames per game)
State space	$3^{361} \approx 10^{170}$ (361 pos, 3 states each)	$2^{2000} \approx 10^{600}$ (2 heroes, (1000+ pos)*(2+ states))
Human player data	rich, high-quality	little
Peculiarity	long-term tactics	real-time, complex control

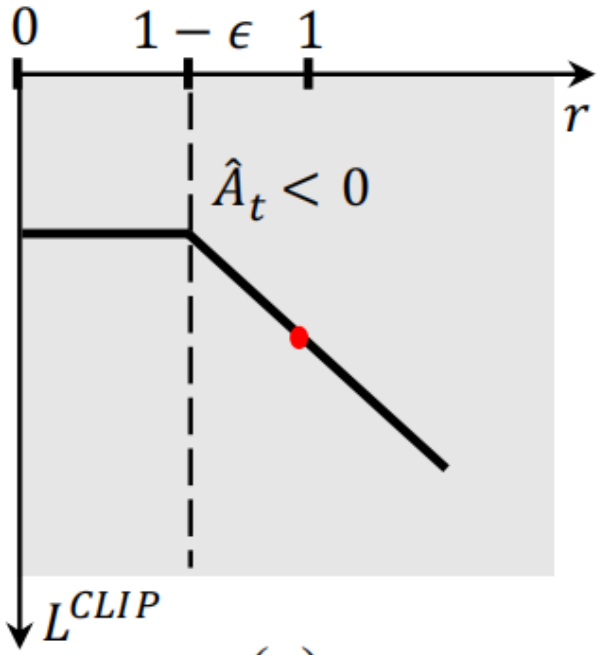
The problem:

- **large-scale & off-policy** setting → policy deviations

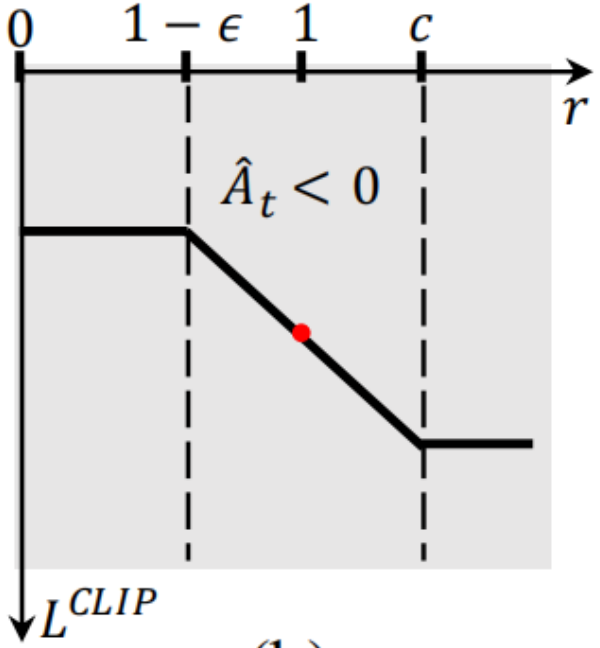
when $\pi_\theta(a_t^{(i)}|s_t) \gg \pi_{\theta_{\text{old}}}(a_t^{(i)}|s_t)$ and $\hat{A}_t < 0$ $\Longrightarrow$ $\dfrac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}\hat{A}_t \ll 0$

Dual-clip PPO

$$\hat{\mathbb{E}}_t\left[\max\left(\min\left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right), c\hat{A}_t\right)\right]$$



(a)



(b)

Experiments

Table 3: Match Statistics of our AI vs. Professional Players on Different Types of Heroes

Hero	DiaoChan	DiRenjie	LuNa	HanXin	HuaMulan
Hero Type	Mage	Marksman	Warrior+Mage	Assassin	Warrior
Score (BO5)	3:0 (AI:Professional)	3:0 (AI:Professional)	3:0 (AI:Professional)	3:1 (AI:Professional)	3:0 (AI:Professional)
Kill	5.0:1.3	2.3:0.7	2.7:1.0	2.5:1.5	4.0:1.3
Game Length	6'56"	6'23"	7'53"	6'41"	6'48"
Gold/min	852.7:430.6	869.3:606.6	969.7:724.0	954.1:754.2	945.2:654.2
Exp/min	900.0:573.0	895.3:661.7	979.0:817.2	965.4:802.5	921.4:723.1

Table 4: Results of AI vs. Various Top Human Players

Hero Name	Hero Type	#Matches	#Win	Rate
DiaoChan	Mage	445	445	100%
DiRenJie	Marksman	264	264	100%
HuaMuLan	Warrior	256	256	100%
HanXin	Assassin	221	220	99.55%
LuNa	Warrior+Mage	260	260	100%
HouYi	Marksman	79	78	98.70%
LuBan	Marksman	354	354	100%
SunWukong	Assassin	221	219	99.09%
		2100	2096	99.81%

Table 5: Results of Ablation Experiments

Item	Win rate vs Base	Time to converge
Base	-	80 h
Base + AM	50.5%	**65 h**
Base + TA	**75%**	90 h
Base + LSTM	73%	100 h
Full version	90%	80 h

Experiments

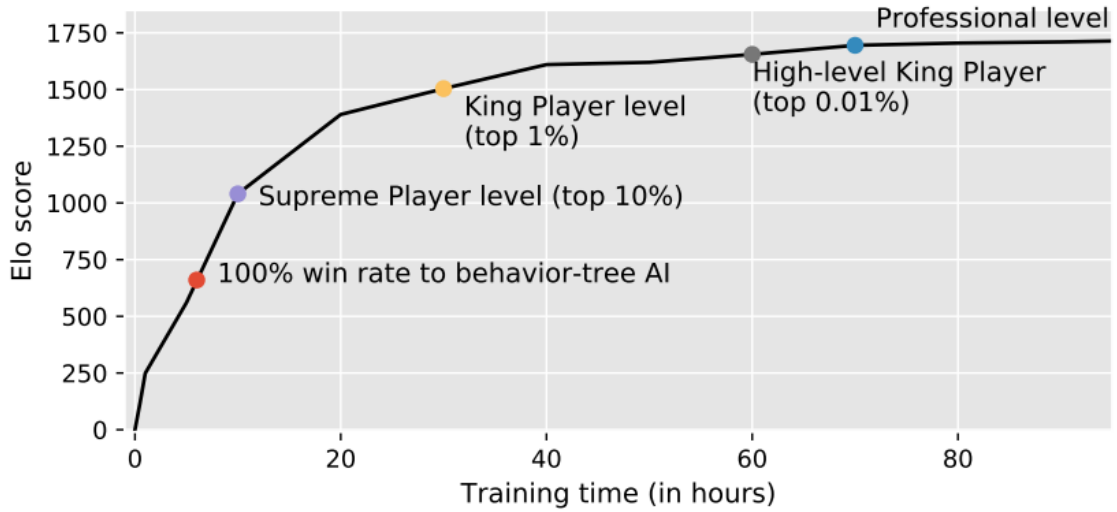


Figure 5: Elo Change during Training

Comparison with Baselines(MCTS and its variants):