腾讯“绝悟”论文披露技术细节。

小小何先生
东北大学 信息科学与工程学院硕士在读

【论文阅读】Mastering Complex Control in MOBA Games with Deep Reinforcement Learning

Mastering Complex Control in MOBA Games with Deep Reinforcement Learning

Deheng Ye [1], Zhao Liu [1], Mingfei Sun [1 *], Bei Shi [1], Peilin Zhao [1], Hao Wu [1 *], Hongsheng Yu [1], Shaojie Yang [1], Xipeng Wu [1], Qingwei Guo [1], Qiaobo Chen [1], Yinyuting Yin [1], Hao Zhang [1], Tengfei Shi [1], Liang Wang [1], Qiang Fu [1], Wei Yang [1], Lanxiao Huang [2]

[1] Tencent AI Lab, Shenzhen, China
[2] Tencent Timi Studio, Chengdu, China
{dericye,ricardoliu,mingfeisun,beishi,masonzhao,alberthwu,yannickyu,shaojieyang,[illegible],
ciaochen,mailyyin,howezhang,francisshi,enginewang,leonfu,willyang,jackiehuang}@tencent.com

这个算法运用强化学习框架，在多人在线战术竞技游戏（MOBA）中1v1击败职业选手。

绝悟难在哪里？

谷歌DeepMind早在2015年用深度Q网络就攻破了Ataria游戏，在2016年更是基于监督学习和强化学习自我博弈训练AlphaGo攻破了人类最后一道防线围棋。而这次腾讯AILab提出来的算法在1v1的MOBA中战胜人类顶级职业选手，也是在像即时策略游戏这种高度复杂的控制游戏中的一个突破吧。

为了对比围棋和MOBA 1v1有啥不同，其论文在**动作空间**、**状态空间**、以及**收集的人类数据**和**游戏本身特性**方面做了对比，如表1所示：

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

在MOBA中智能体还得学会**规划**、**运营**、**攻击**、**防御**、**连招**等等。这一系列的问题需要智能体在长期的序贯决策过程中必须学会精确的动作控制响应。

并且在王者荣耀中你控制的**英雄会有不同的技能属性，不同的攻击属性，不同的控制对象智能体要有不同的玩法**，这就需要你的算法具备充分的鲁棒性。

最后说一下，这里腾讯做的是1v1，这将比5v5更难获得监督的数据资源，因为大家玩王者荣耀好像都是玩的5v5的吧。

绝悟怎么做到的呢？

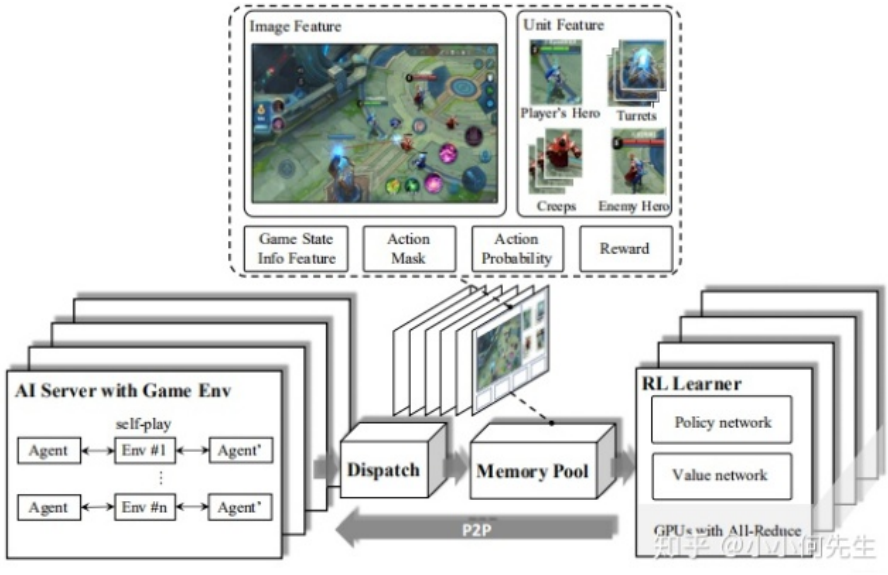
设计了多模态的编码输入、关联动作的解耦、探索剪枝机制和攻击注意机制。

所作出的贡献如下：

- 设计了一个大规模、Off-Policy训练方法。
- 设计了一个神经网络用于MOBA动作控制，也就是设计了一个属于MOBA控制的神经网络。
- 神经网络的优化目标是一个多标签的proximal policy algorithm（PPO）目标。输入给神经网络的特征必须具备支持动作解耦的特性，注意力机制用于目标选择，LSTM网络学连招，并且把PPO改进了，改成了dual-clip PPO来确保其收敛。
- 最终的结果就是用不同类型的英雄可以在王者荣耀中击败职业选手。

绝悟系统设计

动作空间太大方差会比较大，腾讯设计了一个scalable and loosely-coupled(高可扩展低耦合)结构，主要由四块组成：**Reinforcement Learning (RL) Learner** (强化学习智能体), **Artificial Intelligence Server**(人工智能服务器), **Dispatch Module**(调度模块) and **Memory Pool**(记忆库)。



在AI服务器里面智能体与环境进行交互，分发模块简单地收集、压缩、传输数据。

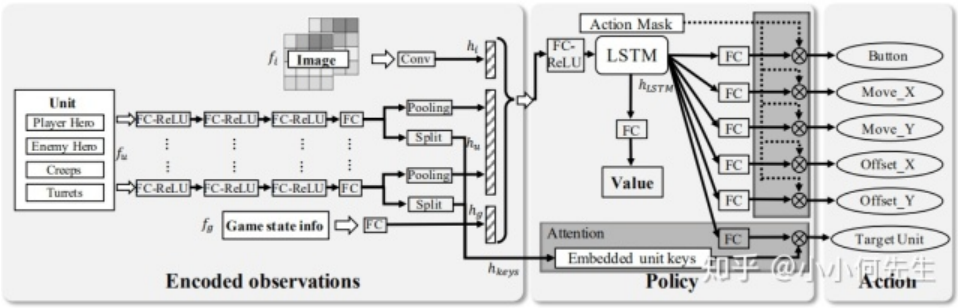
AI server generates episodes via **self-play** with **mirrored policies** (Silver et al. 2017). The opponent policy sampling is similar to (Bansal et al. 2017). Based on the features extracted from game state, hero action is predicted using **Boltzman exploration** (Cesa-Bianchiet al. 2017), i.e., sampling based on **softmax distribution**。

为了加快每回合的推理速度，它们采用了FeatherCNN[3]。官方介绍如下：

FeatherCNN is a state-of-the-art inference engine for mobile devices: github.com/Tencent/Feat...。

分发模块就从AI服务器里面拿数据，组成奖励、特征、和动作概率送给记忆库。之后用于训练。The gradients in the RL learners are averaged through **the ring allreduce algorithm** (Sergeev and Balso 2018)。智能体使用分享内存而不用Socket与记忆库通信，减少IO开销，提速。

绝悟算法设计



- `target attention mechanism` 机制用于帮助神经网络选择目标。
- `LSTM` 用于帮助AI学习**连招**造成有效的高伤害。
- 控制端输出解耦，形成了一个**多标签PPO优化目标**。
- 一种 `game-knowledge-based` 的剪枝算法 `action mask` 被设计出来用于交互过程中更好地探索。
- dual-clipped PPO算法用于保证收敛。

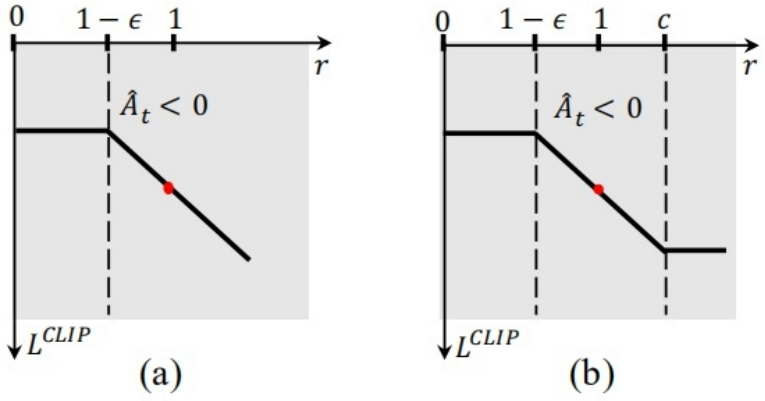
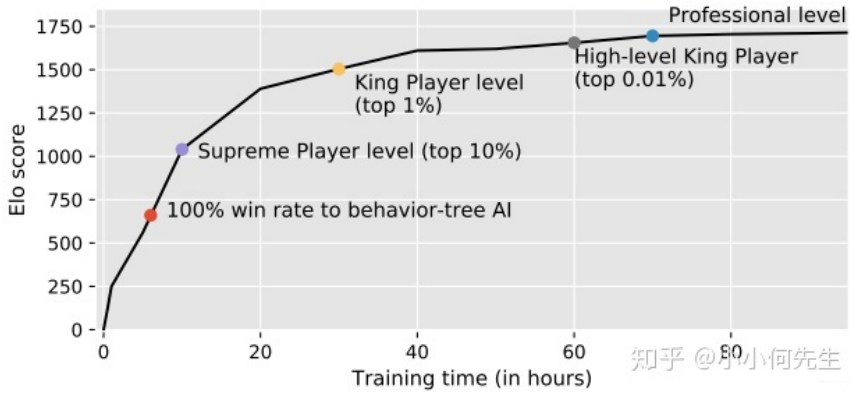


Figure 3: (a) Standard PPO (clip with ϵ); (b) Our Proposed Dual-clip PPO (clip with ϵ and c when $\hat{A}_t < 0$)

这里动作解耦这一块感觉还是可以，感兴趣的可以阅读原文，仔细揣摩，看个热闹的到这就可以了，基本思想以及算法大概框架也差不多了。

绝悟训练细节概要

- a total number of **600,000 CPU** cores encapsulated in Dockers。
- **1,064 Nvidia GPUs** (a mixture of Tesla P40 and V100)。
- **1600 vector features** containing observable unit attributions and game information, and 2 channels of image features read from gamecore (the obstacle channel and the hero position channel).
- we have experiences collected **per day per hero is about 500 years human data** in the 1v1 mode of Honor of Kings.
- We use generalized **advantage estimation (GAE) (Schulman et al. 2015) for reward calculation**。



文章中还有更多细节，这里我也不一一抠了，以后要是这篇文章代码开源了(官方说要开源)，有空了再说吧，我复习考试去了，哭了。

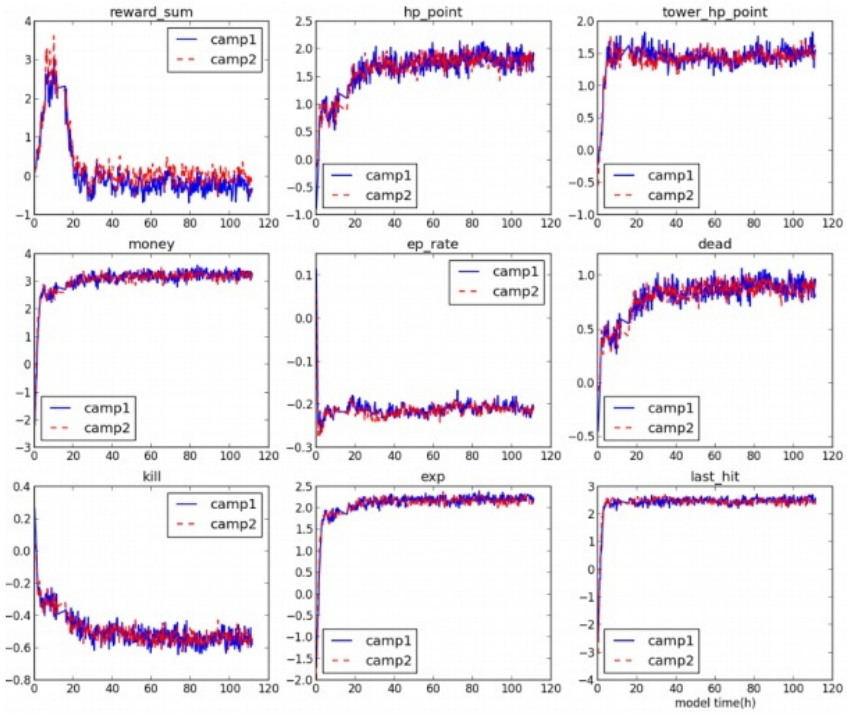


Figure 7: A case of reward change during training; the x-axis is the training time in hours, the y-axis is reward; camp1 and camp2 refer to the two camps (teams) in MOBA games.

看完的感觉就是，没有预期想象感觉中的那样精妙绝伦，那种感觉还是保留在AlphaZero那。但是这个多目标PPO优化感觉还可以。具体的训练细节没扣，准备考试了去了，哭了。

原论文链接：

arxiv.org/abs/1912.0972...

发布于 2019-12-24

深度学习（Deep Learning） 强化学习 (Reinforcement Learning) 即时战略游戏（RTS）

文章被以下专栏收录

强化学习分享交流
希望可以交流一些强化学习和深度学习的知识，共同进步

进入专栏

强化学习前沿
读呀读paper

进入专栏