算法概览图符号表

符号	英文含义	中文含义及其说明
t	timestep in real environment	在真实环境中的时间步
$\gamma \in [0,1]$	discount factor	折扣因子
k	hypothetical step	“假设步骤”是指为了考虑某种情况下可能的结果在模型中采取的假设性动作。
ot	observation at timestep t	在时间步 t 的观测
u_t	reward at timestep t	在时间步 t 的奖励
a_t	action at timestep t	在时间步 t 的动作
s_t	hidden state at timestep t	在时间步 t 的 hidden state
r_t	predicted reward at timestep t	在时间步 t 的预测奖励
p	predicted policy logits	预测的策略logits
v	predicted value	预测的值
h	Representation Network	表征网络。$s^0 = h(o^0)$ 是 the initial hidden state，
g	Dynamics Network	动力学网络。$r^k, s^k = g(s^{k-1}, a^k)$
f	Prediction Network	预测网络。$p^k, v^k = f(s^k)$
vp	predicted value prefix	预测的值前缀。仅在EfficientZero中使用。
r_h^k	Reward hidden state at hypothetical step k	在假设步 k 的reward hidden state。仅在EfficientZero中使用。
g^{rnn}	(recurrent) Dynamics Network	（循环）动力学网络。$vp^k, s^k, r_h^k = g(s^{k-1}, a^k, r_h^{k-1})$。仅在EfficientZero中使用。
β	Proposal policy distribution	提议策略分布。仅在Sampled MuZero中使用。

$\hat{\beta}$	empirical policy distribution	经验策略分布。仅在Sampled MuZero中使用
$\tau = \{s_0, a_0, s_1, a_1, ...\}$	trajectory	轨迹
$R(\tau) = \sum_{t=1}^{T} r_t$	return	轨迹的奖励
$G_t = \sum_{k=0}^{T-t-1} \gamma^k r_{t+k}$	Discounted return	回报（累计折扣奖励）
$\pi_\theta(a_t \vert s_t)$	stochastic policy	随机性策略（其中 θ 是网络参数）
$V_\pi(s) = \mathbb{E}_\pi[G_t \vert s_t = s]$	state value fuction	状态价值函数
$Q_\pi(s, a) = \mathbb{E}_\pi[G_t \vert s_t = s, a_t = a]$	state-action value fuction	动作价值函数

(表1：算法概览图的符号说明。)