Reinforcement Learning
PPO GRPO DAPO

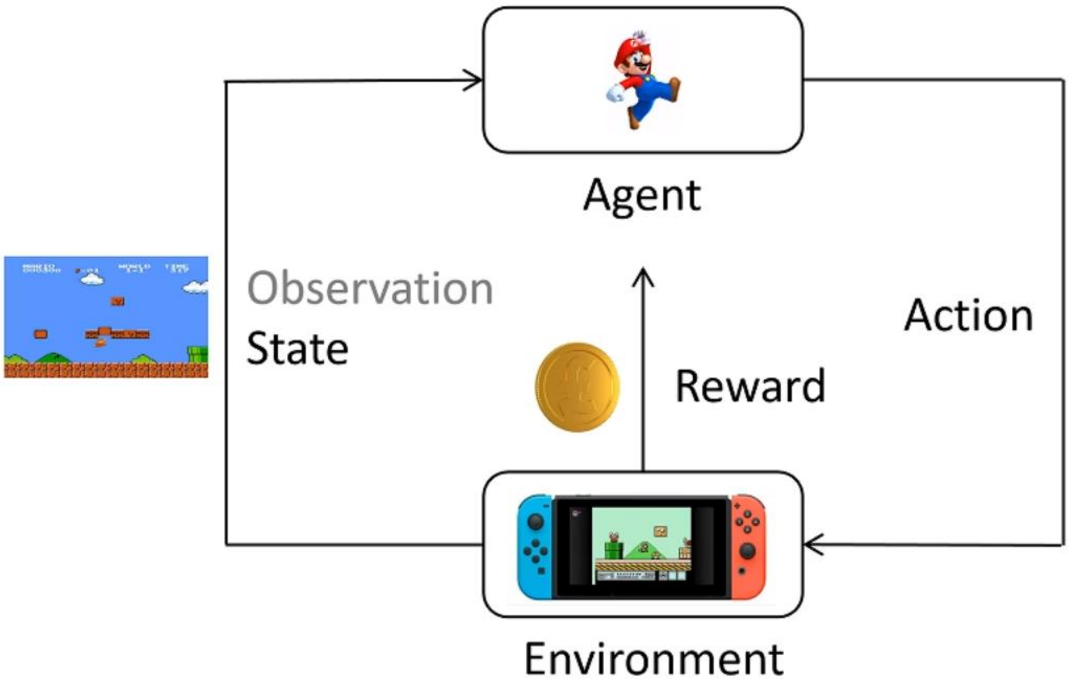
Name： Hongbo Wang

25-04-14

Contents

- Basic Concepts
- PPO
- GRPO
- DAPO

Basic Concepts



Agent: Agent, adopts a policy

State: Current state

Action: Action

Policy: Policy Π, the probability of taking each action

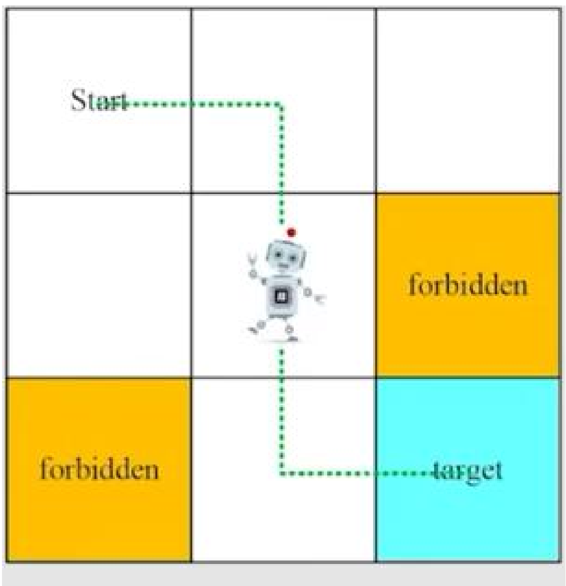
Reward: Reward

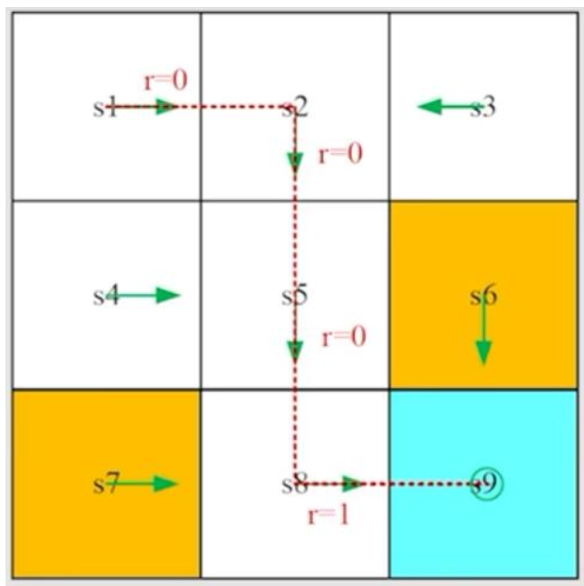
Trajectory: Multi-step trajectory

Return: Cumulative reward of the trajectory

The goal of reinforcement learning is to train a policy that **maximizes the expected return** over all possible trajectories, i.e., to maximize the average return across the entire path.

Basic Concepts





Agent: LLM

State: Current output token state

Action: Output token

Policy: Policy Π, how to select the output token

Reward: reward model
Trajectory: Complete output sentence

The goal of reinforcement learning is to train a policy that **maximizes the expected return** over all possible trajectories, i.e., to maximize the average return across the entire path.

Expected Value Calculation Formula

$$E(x)_{x\sim p(x)} = \sum_x x * p(x) \approx \frac{1}{n}\sum_{i=1}^{n} x \qquad x\sim p(x)$$

Expectation Maximization Derivation

$$E(R(\tau))_{\tau\sim P_\theta(\tau)} = \sum_\tau R(\tau)P_\theta(\tau)$$

$$\begin{aligned}
\nabla E(R(\tau))_{\tau\sim P_\theta(\tau)} &= \nabla \sum_\tau R(\tau)P_\theta(\tau) && \mathbf{1} \\
&= \sum_\tau R(\tau)\nabla P_\theta(\tau) \\
&= \sum_\tau R(\tau)\nabla P_\theta(\tau)\frac{P_\theta(\tau)}{P_\theta(\tau)} && \mathbf{2} \\
&= \sum_\tau P_\theta(\tau)R(\tau)\frac{\nabla P_\theta(\tau)}{P_\theta(\tau)} \\
&= \sum_\tau P_\theta(\tau)R(\tau)\frac{\nabla P_\theta(\tau)}{P_\theta(\tau)} \\
&\approx \frac{1}{N}\sum_{n=1}^{N} R(\tau^n)\frac{\nabla P_\theta(\tau^n)}{P_\theta(\tau^n)} && \mathbf{3} \\
&= \frac{1}{N}\sum_{n=1}^{N} R(\tau^n)\nabla \log P_\theta(\tau^n) && \mathbf{5}
\end{aligned}$$

$$\nabla \log f(x) = \frac{\nabla f(x)}{f(x)} \qquad \mathbf{4}$$

$$\tau\sim P_\theta(\tau)$$

Expectation Maximization Derivation

$$= \frac{1}{N}\sum_{n=1}^{N} R(\tau^n)\nabla\log P_\theta(\tau^n) \quad 5$$

$$= \frac{1}{N}\sum_{n=1}^{N} R(\tau^n)\nabla\log\prod_{t=1}^{T_n} P_\theta(a_n^t|s_n^t) \quad 6$$

$$= \frac{1}{N}\sum_{n=1}^{N} R(\tau^n)\sum_{t=1}^{T_n} \nabla\log P_\theta(a_n^t|s_n^t)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} R(\tau^n)\nabla\log P_\theta(a_n^t|s_n^t)$$

Common Forms of Loss Functions

$$\text{Loss} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} R(\tau^n)\log P_\theta(a_n^t|s_n^t)$$

Replace with Advantage Function

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n}\boxed{R(\tau^n)}\nabla\log P_\theta(a_n^t|s_n^t)$$

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n}\boxed{A_\theta^{GAE}(s_n^t,a_n^t)}\nabla\log P_\theta(a_n^t|s_n^t)$$

$A_\theta(s,a) = Q_\theta(s,a) - V_\theta(s)$ 在state s下，做出Action a，比其他动作能带来多少优势。

$$A_\theta(s_t,a) = r_t + \gamma * V_\theta(s_{t+1}) - V_\theta(s_t)$$

$$A_\theta^{GAE}(s_t,a) = (1-\lambda)(A_\theta^1 + \lambda * A_\theta^2 + \lambda^2 A_\theta^3 + \cdots)$$

Add Reference Policy
Off policy

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_{\theta}^{GAE}(s_n^t, a_n^t)\boxed{\nabla \log P_{\theta}(a_n^t|s_n^t)}$$

$$=\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_{\theta'}^{GAE}(s_n^t, a_n^t)\boxed{\frac{P_{\theta}(a_n^t|s_n^t)}{P_{\theta'}(a_n^t|s_n^t)}\nabla \log P_{\theta}(a_n^t|s_n^t)}$$

$$=\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_{\theta'}^{GAE}(s_n^t, a_n^t)\frac{\nabla P_{\theta}(a_n^t|s_n^t)}{P_{\theta'}(a_n^t|s_n^t)}$$

$$Loss_{ppo}=-\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_{\theta'}^{GAE}(s_n^t, a_n^t)\frac{P_{\theta}(a_n^t|s_n^t)}{P_{\theta'}(a_n^t|s_n^t)}+\beta KL(P_{\theta}, P_{\theta'})$$

$$\boxed{E(R(\tau))_{\tau\sim P_{\theta}(\tau)}=\sum_{\tau} R(\tau)P_{\theta}(\tau)}$$

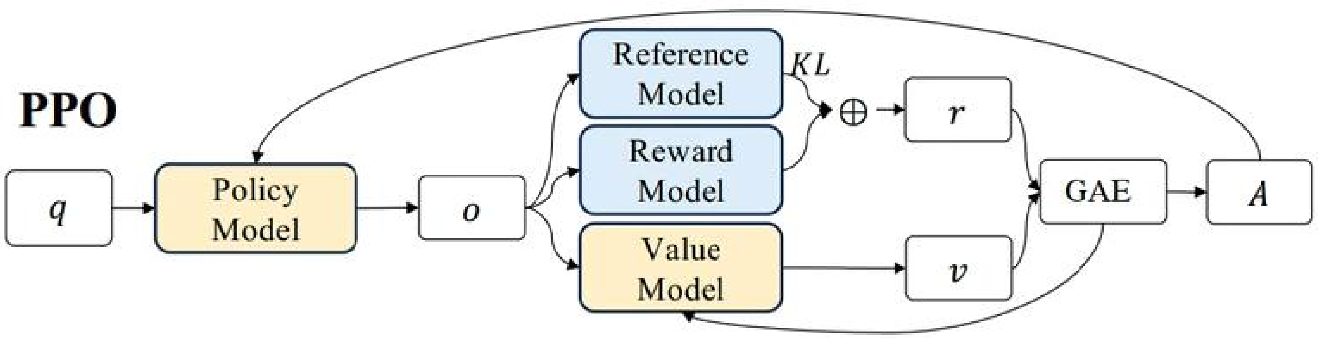
PPO

$$Loss_{ppo} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_{\theta'}^{GAE}(s_n^t, a_n^t)\frac{P_\theta(a_n^t|s_n^t)}{P_{\theta'}(a_n^t|s_n^t)} + \beta KL(P_\theta, P_{\theta'})$$

$A_\theta(s,a) = Q_\theta(s,a) - V_\theta(s)$ 在state s下，做出Action a，比其他动作能带来多少优势。

$$A_\theta(s_t, a) = r_t + \gamma * V_\theta(s_{t+1}) - V_\theta(s_t)$$

$$A_\theta^{GAE}(s_t, a) = (1-\lambda)(A_\theta^1 + \lambda * A_\theta^2 + \lambda^2 A_\theta^3 + \cdots)$$



Limitation of PPO

Q **用户问题:** 什么是数据库?

A1 **大模型回答1:** 数据库是一个有组织的数据集合，允许高效的数据存储、检索和管理。

A2 **大模型回答2:** 数据库用于存储数据。

Reward Score1 9.7

Reward Score2 3.8

Reward 模型

Prompt+Response	什么	是	数据库	?	数据库	用于	存储	数据	。

Policy model	训练模型	[0.11,0.02,0.02,0.32,0.04...]	[0.11,0.02,0.02,0.32,0.04...]	[0.11,0.02,0.02,0.32,0.04...]	[0.11,0.02,0.02,0.32,0.04...]	[0.11,0.02,0.02,0.32,0.04...]
Reference model	参考模型	[0.11,0.02,0.02,0.32,0.04...]	[0.11,0.02,0.02,0.32,0.04...]	[0.11,0.02,0.02,0.32,0.04...]	[0.11,0.02,0.02,0.32,0.04...]	[0.11,0.02,0.02,0.32,0.04...]
Score		0	0	0	0	3.8

In the reward model, a sentence outputs a reward value, and the previous values are **all 0**. Using 0 to calculate the advantage function does not make much sense.

GRPO
Group-relative Policy Optimization

PPO

Prompt+ Response	什么	是	数据库	？	数据库	用于	存储	数据	。
Score					0	0	0	0	3.8

GRPO

什么	是	数据库	？	数据库	用于	存储	数据	。			3.8
什么	是	数据库	？	数据库	是	一个	有	组织	的	数据	集合	5.2
什么	是	数据库	？	数据库	是	用来	高效	存取	数据	的	软件	6.1

Score

数据库	用于	存储	数据	。			
-1.06	-1.06	-1.06	-1.06	-1.06			
数据库	是	一个	有	组织	的	数据	集合
0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14
数据库	是	一个	有	组织	的	数据	集合
0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92

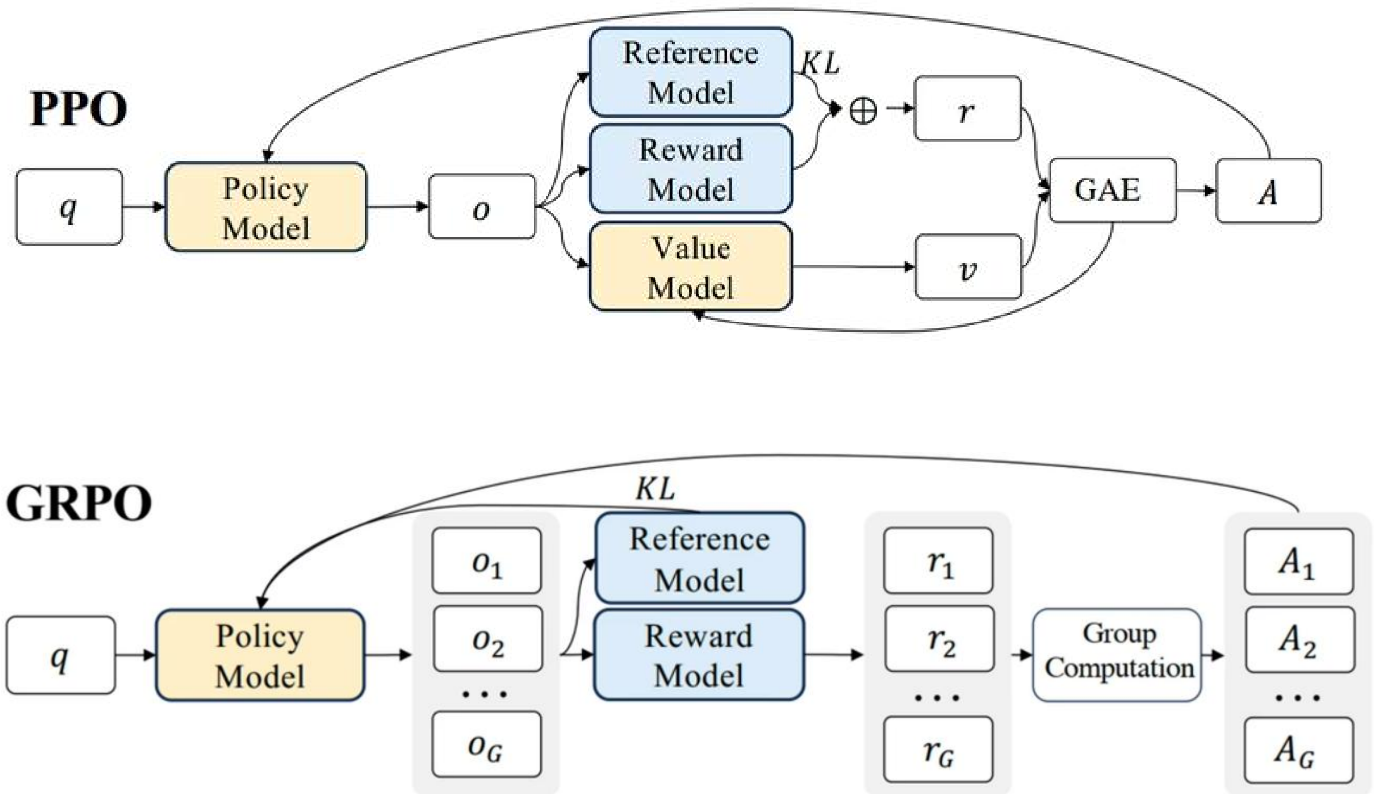
$r_1 = 3.8 \quad r_2 = 5.2 \quad r_1 = 6.1$

$$\tilde{r_i} = \frac{r_i - mean(\boldsymbol{r})}{std(\boldsymbol{r})}$$

$\tilde{r_1} =- 1.06 \quad \tilde{r_2} = 0.14 \quad \tilde{r_3} = 0.92$

Compared to PPO, GRPO no longer requires using the entire action; it only uses the reward to calculate the advantage function, and no longer needs a value model.

GRPO
Group-relative Policy Optimization



GRPO can calculate the advantage function without the need for a value model.

$$Loss_{ppo2} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} min(\boxed{A_{\theta'}^{GAE}}(s_n^t, a_n^t)\frac{P_\theta(a_n^t|s_n^t)}{P_{\theta'}(a_n^t|s_n^t)}, clip(\frac{P_\theta(a_n^t|s_n^t)}{P_{\theta'}(a_n^t|s_n^t)}, 1-\varepsilon, 1+\varepsilon)A_{\theta'}^{GAE}(s_n^t, a_n^t))$$

$$J_{GRPO} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} min(\boxed{A_{\theta'}^{GRPO}}(s_n^t, a_n^t)\frac{P_\theta(a_n^t|s_n^t)}{P_{\theta'}(a_n^t|s_n^t)}, clip(\frac{P_\theta(a_n^t|s_n^t)}{P_{\theta'}(a_n^t|s_n^t)}, 1-\varepsilon, 1+\varepsilon)A_{\theta'}^{GRPO}(s_n^t, a_n^t)) - \beta KL(P_\theta, P_{\theta'})$$

Standard Formula in Papers

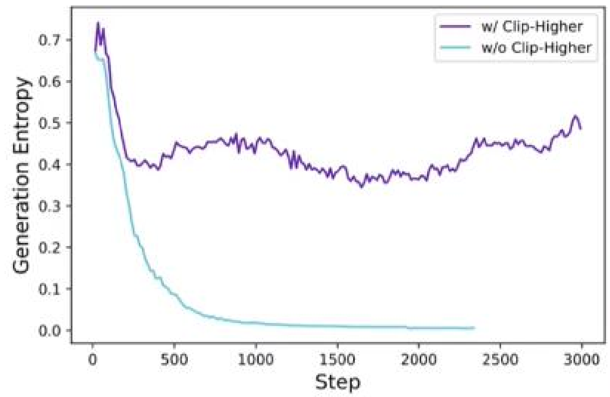
$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D}, \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left(\min\left(r_{i,t}(\theta)\hat{A}_{i,t},\ \mathrm{clip}\left(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\right) - \beta D_{\mathrm{KL}}(\pi_\theta||\pi_{\mathrm{ref}})\right)\right],$$

where

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\mathrm{old}}}(o_{i,t} \mid q, o_{i,<t})}.$$

Limitation of GRPO
Introdution to DAPO

1.The entropy decreases too quickly, leading to rapid convergence and entropy collapse, which causes many parameters to remain suboptimal — set a higher Clip ε_{high}



2.Dynamic Sampling -- In the group response for reward sampling, all correct = all incorrect

$$\text{s.t.} \quad 0 < \left|\{o_i \mid \texttt{is\_equivalent}(a, o_i)\}\right| < G. \qquad R(\hat{y}, y) = \begin{cases} 1, & \texttt{is\_equivalent}(\hat{y}, y) \\ -1, & \text{otherwise} \end{cases}$$

什么	是	数据库	?	数据库	用于	存储	数据	。				1
什么	是	数据库	?	数据库	是	一个	有	组织	的	数据	集合	1
什么	是	数据库	?	数据库	是	用来	高效	存取	数据	的	软件	1

Delete Sample

3.Imbalanced reward for long and short sentences -- Token-level optimization

DAPO (Decoupled Clip and Dynamic Sampling Policy Optimization)

- **1.The entropy decreases too quickly, leading to rapid convergence and entropy collapse, which causes many parameters to remain suboptimal — set a higher Clip ε_{high}**
- **2.Dynamic Sampling -- In the group response for reward sampling, all correct = all incorrect**
- **3.Imbalanced reward for long and short sentences -- Token-level optimization**

$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left(\min\left(r_{i,t}(\theta)\hat{A}_{i,t},\ \mathrm{clip}\left(r_{i,t}(\theta),1-\varepsilon,1+\varepsilon\right)\hat{A}_{i,t}\right)-\beta D_{\mathrm{KL}}(\pi_\theta||\pi_{\mathrm{ref}})\right)\right],$$

$$\mathcal{J}_{\mathrm{DAPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|q)} \left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\left(r_{i,t}(\theta)\hat{A}_{i,t},\ \mathrm{clip}\left(r_{i,t}(\theta),1-\varepsilon_{\mathrm{low}},1+\varepsilon_{\mathrm{high}}\right)\hat{A}_{i,t}\right)\right]$$

➢ 3 ➢ 1

$$\text{s.t.}\quad 0<\left|\{o_i \mid \texttt{is\_equivalent}(a,o_i)\}\right|<G.$$

➢ 2

$$R(\hat{y},y)=\begin{cases}1, & \texttt{is\_equivalent}(\hat{y},y)\\ -1, & \text{otherwise}\end{cases}$$

DAPO (Decoupled Clip and Dynamic Sampling Policy Optimization)

$$\min\left(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\Big(r_{i,t}(\theta), 1-\varepsilon_{\text{low}}, 1+\varepsilon_{\text{high}}\Big)\hat{A}_{i,t}\right)$$

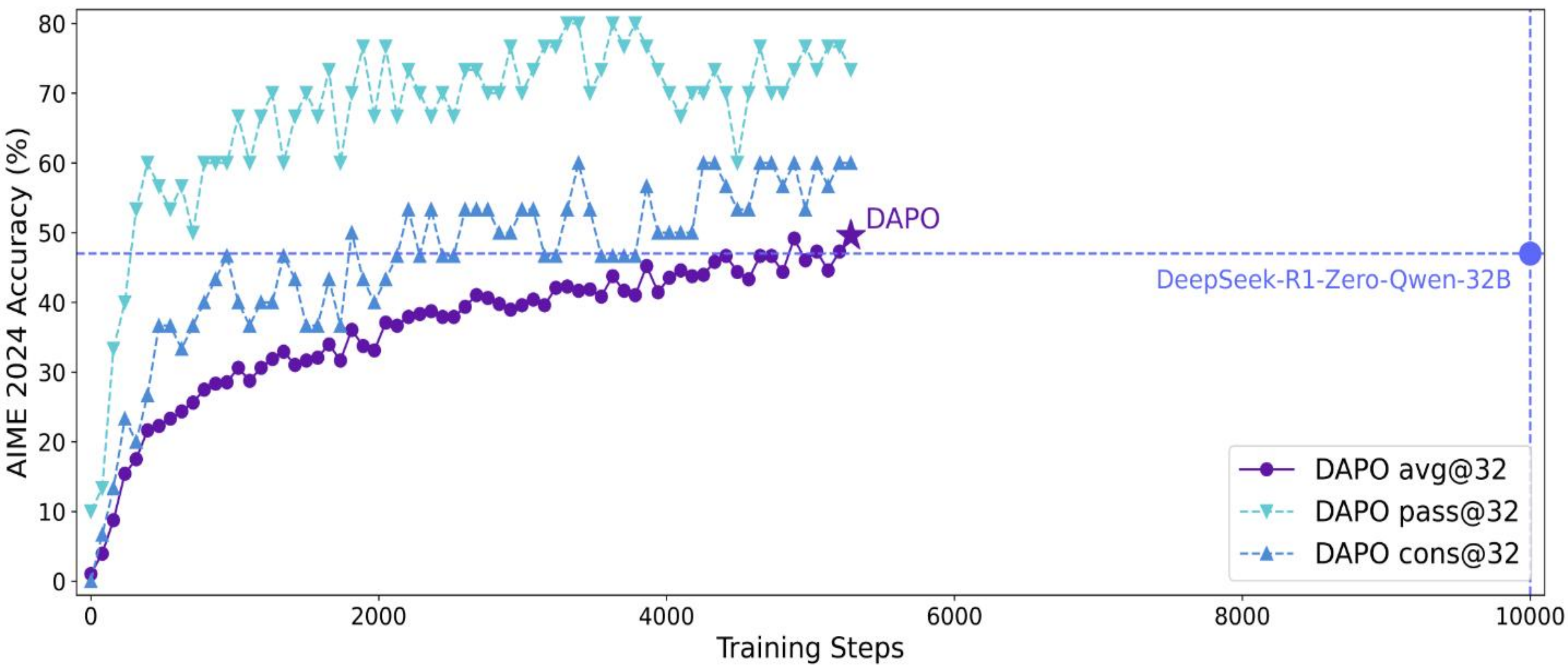
```
actor_rollout_ref:
  actor:
    clip_ratio_low: 0.2
    clip_ratio_high: 0.28
```

```
pg_losses1 = -advantages * ratio
pg_losses2 = -advantages * torch.clamp(ratio, 1 - cliprange_low, 1 + cliprange_high)
pg_losses = torch.maximum(pg_losses1, pg_losses2)
```

$$\text{s.t.}\quad 0 < \Big|\{o_i \mid \texttt{is\_equivalent}(a, o_i)\}\Big| < G.$$

```
data:
  gen_batch_size: 1536
  train_batch_size: 512
algorithm:
  filter_groups:
    enable: True
    metric: acc # score / seq_reward / seq_final_reward / ...
    max_num_gen_batches: 10 # Non-positive values mean no upper limit
```

DAPO (Decoupled Clip and Dynamic Sampling Policy Optimization)



In terms of inference performance, DAPO outperforms the GRPO framework's Deepseek-R1-Qwen-32B, with a **50% reduction in resources.**

Reference

零基础学习强化学习算法：ppo
https://www.bilibili.com/video/BV1iz421h7gb/?spm_id_from=333.1391.0.0&vd_source=6374cb49b8a20f29e86ab9de12471afa

代码实现大模型强化学习(PPO)，看这个视频就够了
https://www.bilibili.com/video/BV1rixye7ET6?spm_id_from=333.788.videopod.sections&vd_source=6374cb49b8a20f29e86ab9de12471afa

DeepSeek-GRPO
hhttps://www.bilibili.com/video/BV1enQLYKEA5/?spm_id_from=333.1391.0.0&vd_source=6374cb49b8a20f29e86ab9de12471afa

[文章解读]字节跳动与清华发布强化学习新算法DAPO，性能超越DeepSeek-R1(GRPO)
https://www.bilibili.com/video/BV1gmXFYxEKT/?spm_id_from=333.1391.0.0&vd_source=6374cb49b8a20f29e86ab9de12471afa

Thanks!

大连理工大学

信息检索研究室

搜人搜物搜信息 重情重義重認知

http://ir.dlut.edu.cn