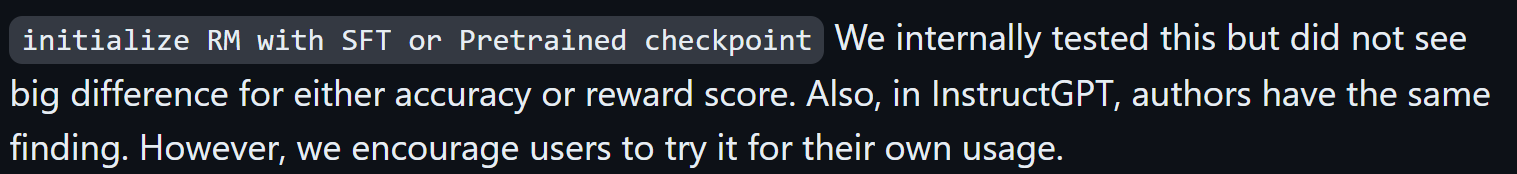
텍스트, 스크린샷, 폰트, 정보이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명



**RLHF training 은 쉽게 발산 가능. 아래와 같이 설정해야함.**

* **Weight decay:** 꺼야함
* **Dropout:** 꺼야함
* **Training epochs:** reward score가 쉽게 platou가 되기 때문에 1 epoch도 충분하다
* **Ema checkpoint:** actor 모델의 큰 움직임을 막기 위해 새로운 모델 사용
* **Mix unsupervised training:** PPO와 unsupervised training을 섞는 방법. 하지만 사용하면 모델 학습이 수렴하지 않아서 사용하면 안됨.
* --per\_device\_train\_batch\_size와 -–per\_device\_mini\_batch\_size는 같은 크기여야함.
* --ppo\_epochs, --generation\_batch\_numbers는 1을 넘으면 안됨.

**Deep speed chat이 GPU 메모리를 줄이는 방법:**

* ZeRO optimization을 통해서 model parameter와 optimizer를 여러 GPU에 골고루 나눠 써서 병목 현상을 없애 준다.
* Actor모델은 GPU에 Reference 모델은 CPU에 올리는 offload 기술을 제공한다.
* LoRA 방식을 사용한다.
* **DeepSpeed Hybrid Engine**
  + 하나의 node에서 Training과 inference를 번갈아 가면서 할 때는 원래 resource가 낭비가 되지만 빠르게 해준다.

**A2C**

**텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명**

* DeepMind used a similar reward setup for Gopher but used [synchronous advantage actor-critic](http://proceedings.mlr.press/v48/mniha16.html?ref=https://githubhelp.com) (A2C) to optimize the gradients, which is notably different but has not been reproduced externally.

텍스트, 폰트, 스크린샷, 영수증이(가) 표시된 사진

자동 생성된 설명