指令微调

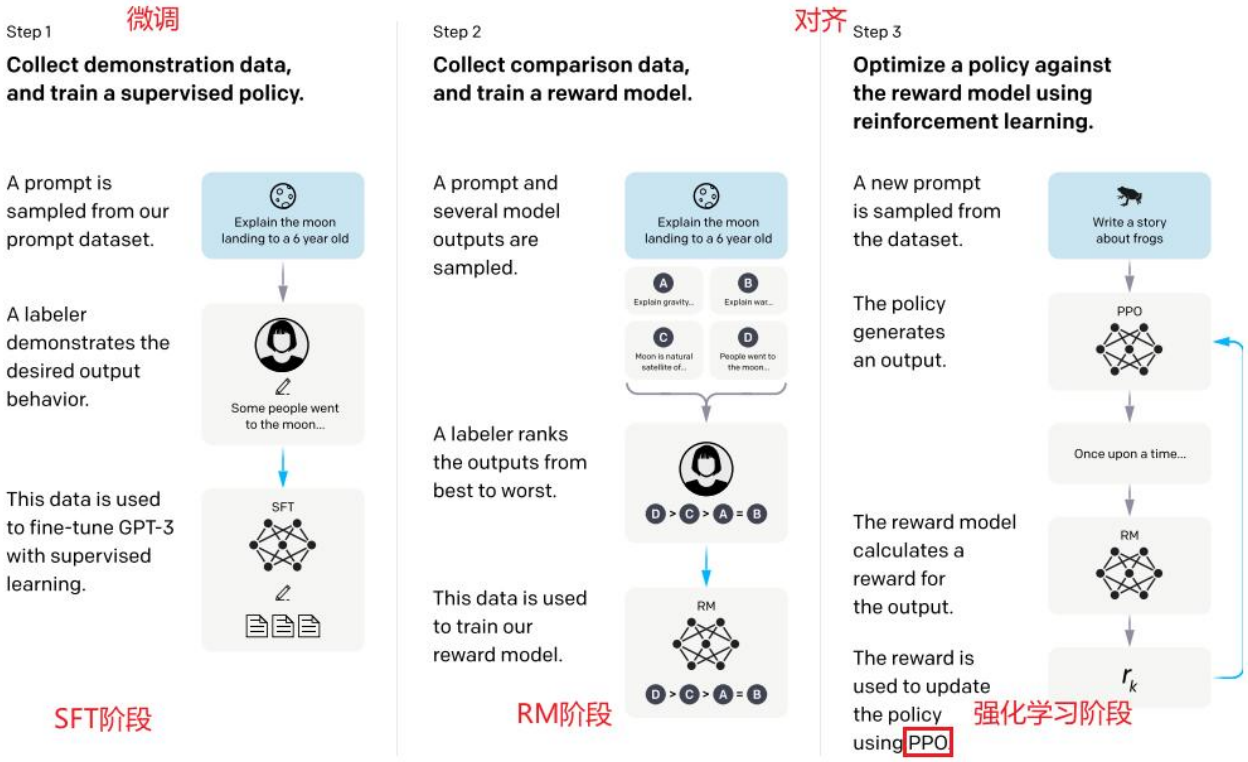
这样做的原因：预训练过程中的任务是预测下一个token，而不是依据指令生成回答。

一个典型的指令跟随数据集如下：

```
[
   {"role": "system", "content": "You are a helpful assistent"},
   {"role": "user", "content": "Hello, how are you?"},
   {"role": "assistant", "content": "I'm doing great. How can I help you today?"},
   {"role": "user", "content": "I'd like to show off how chat templating works!"},
]
```

InstructGPT

来自b站李沐——论文精读：InstructGPT



三阶段：

1、得到SFT模型

2、得到reward model

3、继续微调第一阶段得到的SFT，使得生成的答案经过reward model的打分能够尽量得到一个比较高的分数，这里用到了强化学习RLHF

数据集怎么准备?

1、从用户log里面收集（常用），对每个用户id最多采样200个问题，划分训练集验证集测试集时是依据用户id来划分的，这样做是为了不污染数据，因为一个用户可能会问一些类似的问题，如果出现了人名之类的用户信息要过滤掉。

2、标注人员写了很多问题

Plain：让标注人员想尽可能多样化的任务

Few-shot：针对一个Instruction，让标注人员想多个问答对作为few shot

User-Based：从用户log里面收集问题，标注人员去写对应的答案

这样就得到了三个数据集：

1、SFT数据集，来自人工标注和API，13k

2、RM数据集，来自人工标注和API，33k

3、PPO数据集，只来自API，31k

model

1）SFT模型：微调GPT3得到的，训练了16个epochs，其实1个epoch就过拟合了，但是也没有问题，因为这个模型也不是直接拿来用，而是用来初始化后面的模型

2）RM模型：将SFT模型最后一层替换为投影层得到的，输入是问答对，输出是一个分数。这里用的是6B的小模型，没有选择用175B的，因为大的不稳定。

损失函数使用的是排序任务中常用的Pairwise Ranking Loss：

这里取的k=9（对于每个问题生成9个答案），因此就有了$C(9,2)=36$对用于计算loss的问答对。

$$\text{loss}(\theta) = -\frac{1}{\binom{K}{2}} E_{(x,y_w,y_l)\sim D}\left[\log\left(\sigma\left(r_\theta\left(x,y_w\right) - r_\theta\left(x,y_l\right)\right)\right)\right]$$

x表示问题，如果一个答案的排序（y_w)比另一个答案的排序(y_l)高的话，那么就尽量使得他们之间的奖励分数差得比较远。

假定现在有一个排好的序列：A > B > C >D。

我们需要训练一个打分模型，模型给四句话打出来的分要满足 r(A) > r(B) > r(C) > r(D)。

loss应该为：

```
loss = r(A) - r(B) + r(A) - r(C) + r(A) - r(D) + r(B) - r(C) + ... + r(C) - r(D)
loss = -loss
```

为了更好的归一化差值，我们对每两项差值都过一个 sigmoid 函数将值拉到 0 ~ 1 之间。

可以看到，loss 的值等于排序列表中所有「排在前面项的reward」减去「排在后面项的reward」的和。

而我们希望模型能够「最大化」这个「好句子得分」和「坏句子得分」差值，而梯度下降是做的「最小化」操作。

因此，我们需要对 loss 取负数，就能实现「最大化差值」的效果了。

3）RL模型：用到了强化学习里的优化算法PPO

损失函数有两个：

- 在每个token上都计算一个和SFT模型之间的KL-Divergence，目的是希望在强化学习的过程中不要太过于偏离最开始的生成模型
- PPO-ptx，在训练的同时加入一些通用预训练任务，以维持在通用NLP任务上的性能

把上面两个函数相加：

$$\begin{aligned}\text{objective}\,(\phi) = & E_{(x,y)\sim D_{\pi_\phi^{\mathrm{RL}}}}\left[r_\theta(x,y) - \beta\log\left(\pi_\phi^{\mathrm{RL}}(y\mid x)/\pi^{\mathrm{SFT}}(y\mid x)\right)\right] + \\ & \gamma E_{x\sim D_{\text{pretrain}}}\left[\log(\pi_\phi^{\mathrm{RL}}(x))\right]\end{aligned}$$