SFT Memorizes, RL Generalizes A Comparative Study of Foundation Model Post-training

Yang

2025 年 6 月 30 日

1 Introduction

当下模型在进行 post-training 的时候会使用 SFT 与 RL，但是它们在对模型泛化提升上各自的分工尚未明确，难点在于将 memorization 和 transferable principles 分离，前者倾向于 SFT，后者倾向于 RL 用于泛化

测试项目：基于文本的和基于视觉的，General Points(类似于 24 点) 和 V-IRL(类似于现实中导航问题)

结论：SFT 倾向于记忆，而 RL 倾向于泛化，且基于结果的 RL 奖励学习可以提高模型的视觉识别能力，SFT 有助于规范输出格式，是 RL 前的必要步骤，可以使 RL 获得更好的性能提升。增加最大步骤数来扩展推理时间的计算可以有更好地泛化能力

SFT：显著提高 zero-shot 模型处理未见过的任务得能力，且可以提高 format

RL：使得模型与人类偏好一致且训练基础模型解决特定任务

Memorization：表现在模型记住训练数据的能力

Generalization：表现在模型输出分布与与训练数据之间的 divergence(散度)

LLM 在知识密集简单的任务上表现出过拟合，而在复杂推理任务上表现出泛化

distribution：指模型在给定输入的情况下输出的概率分布，比如给定输入一个词，那么它输出的词可能有什么存在概率分布

divergence：衡量两个概率分布之间的不同，常见为 KL 散度 $D(P||Q) = \sum Plog\frac{P}{Q}$

2 Preliminaries

2.1 RL

令 S 表示 state space，A 表示 action space，T 表示每个 episode(从任务开始到结束比如执行任务开始到执行任务结束) 用到的最大步数。定义 reward function $r : S * A \to R$，

最后目的是为了选择一个 policy $\pi : S \to A$，其中 $\pi(a|s)$ 表示在 s 的状态下选择 a 的概率，最大化 $max_\pi E_\pi \sum_{t=0}^{T} r(s_t, a_t)$

理解：π 是一个 policy，决定在给定状态下下一步的 action，存在概率分布，对于一个完整的 episode 会执行许多操作，每一个操作都会进入一个不同的状态，那么对于一个确定的流程的分支，其奖励总和在以概率作为权重之后可以表示为 $\sum_{t=0}^{T} \pi(a_t|s_t) r(s_t, a_t)$，然后把所有可能的分支加起来得到的就是在 policy 为 π 的情况下 $E \sum_{t=0}^{T} r(s_t, a_t)$，然后以 policy 作为参数进行得到的 reward 最大化

2.2 将 RL 运用到 LLM 与 VLM

令 V 表示离散的 token 空间，设 m 和 n 分别代表输入和输出的最大 token 数，则输入文本空间可以定义为 V^m，输出为 V^n，令 O 代表所有的 RGB 图像，对于 S，LLM 和 VLM 分别是 V^m 和 $V^m * O$，$A = V^n$，则定义 verifier：$V^n \to R * V^k$，通过评估输出给出一个 reward function r 和文本信息。实例：$VER(v_t^{out}) = (r_t, v_t^{ver})$，同样定义一个 policy $\pi_\theta : S \to V^n$

verifier 的作用是对于这一步的输入进行打分和提示，帮助模型在中间步骤作出更好地修改，相当于中间过程的评价老师。回到上面的 RL 基础知识，上面需要最大化分数期望，而此处 verifier 的评分就是需要最大化的指标

在 $t = 0$ 的时候，v^{in} 是 system prompt，在之后 v^{in} 是之前输入和输出影响下的 input prompt，这个过程称为 sequential revision

3 Expriment

3.1 General Point

每个状态是四张卡片，要求使用者四张卡片中的全部仅一次通过某种数学运算凑出给定的和，对于 VLM 有额外的要求是记住这四张卡片并能辨别，每次尝试可能可以成功，也可能会失败，因此会使用 verifier 做出判断并影响下一次的输入

泛化能力一方面体现在将"J" 等扑克牌识别为"11"，也可以通过改变卡片的颜色等方式，检测 arithmetic reasoning ability

3.2 V-IRL

同样也是分为纯文本和文本与视觉输入相结合，主要任务是根据给出的信息导航路线，视觉输入的难点在于识别 landmark。鉴别是单纯记忆还是拥有泛化能力的指标是第一个变量统一为绝对方向 (东南西北)，而第二个变量统一为相对方向 (左右)，视觉泛化体现在对于不同位置地标的识别能力

3.3 Result

指标分成两个部分：ID(in distribution) 即见过相同分布的训练数据和 OOD(out of distribution) 即未见过相同分布的数据，属于不同领域

对于 General Point，ID 处理"J", "Q", "K" 为 10，而 OOD 处理为 11，12，13

对于 V-IRL，ID 处理为绝对方向，而 OOD 处理为相对方向

变量因素，模型训练的时候要么只 SFT，要么在 SFT 基础上进行 RL，模型的类型有 LLM 和 VLM

结果：RL 显著提高了各种模型的 OOD 能力，则说明提高了泛化能力，对于额外的视觉识别功能 RL 也有一定的提升。而在纯文本的 ID 中，SFT 的表现优于 RL，体现 SFT 集中于 memorization

3.4 OOD 中的视觉泛化

实验变量：对于 General Point，训练时使用黑桃梅花，测试时使用红桃方块。对于 V-IRL，训练时使用 New York，测试时使用别的环境，规则保持不变，不像上一个部分进行了更改，作控制变量

结论：RL 提升了视觉部分的泛化能力

3.5 为什么 RL 能够提高 visual generalization

scale up RL 指将原来应用于小规模数据场景的 RL 推广到大数据规模量级

基于 General Point 和 V-IRL 进行试验，结果发现 Scaling RL 可以提高视觉识别的准确性，而 Scaling SFT 会降低识别准确性，进一步降低整体性能

3.6 是否还需要 SFT

进行试验只使用 RL，不进行 SFT 的预处理会出现 poor instruction following 的问题若主干模型 (backbone model 不能够遵循指令)，也就是出现给定一个指令，但是结果却胡乱回答的情况。原因在于模型不具备天然根据要求回答的能力，SFT 很好的引导模型面对一类问题该进行什么样的回答，RL 只是强化了这方面的能力并可以做到能力泛化应用到更多场景

3.7 中间提高使用 verifier 的次数可以有效提高泛化能力

3.8 Summary

本篇论文主要研究了 SFT 和 RL 在模型性能提升的两个方面：Memorization 和 Generalization 上的贡献，最后得出的结论是 SFT 负责前者，RL 负责后者，且 RL 训练也需要 SFT 作预处理。实验有两个，分别是 General Point 和 V-IRL，前者是扑克牌点数

运算得到指定的和，后者为导航识别。实验对象为 LLM 和 VLM，对于 LLM，一切信息都以文本呈现，而对于 VLM，信息除了文本还有图像输入，因此还可以考察 SFT 和 RL 对于视觉识别能力的影响。进行的实验有两类变量：规则和图像。对于规则，GP 是对于"J"，"Q"，"K" 扑克牌点数意义的概念，是都认为是 10 还是分别是 10、11、12，V-IRL 是绝对方向和相对方向；对于图像，GP 是训练用黑色，测试用红色花色，而 V-IRL 是训练用纽约街景，测试用别的。显然图像变量主要是针对在 VLM 观察 RL 的影响，影响结果分为两类 ID(侧重于 Memorization)，OOD(侧重于 Generalization)，因此共 8 中实验结果对比图。其中着重对 RL 的奖励策略进行分析，在专业术语中会要求最大化策略带来的步骤奖励和的期望，即 $max_\pi E_\pi \sum_{t=0}^{T} r(s_t, a_t)$，而实际应用到 LLM 和 VLM 的时候会考虑引入 verifier 作为奖励评定的模型，verifier 会影响下一步的决定，同时会给当前决定评分并给出建议。在实验中也把 verifier 的使用次数作为变量观察 verifier 对训练结果的影响，更多地使用 verifier 可以帮助模型在中间步骤进行更快的调整，快速走上正轨。没有 verifier 指导，下一步的输入只会受到历史输入和历史输出的影响，而如果有 verifier 进行评定，则还会加上 verifier 的文本输入