Model Extrapolation Expedites Alignment 方法的复现与扩展技术报告

臧一赫

2025年10月

目录

1 背景与目标

CoAI 团队在论文《Model Extrapolation Expedites Alignment》中提出了**Model Extrapolation (ExPO)**，本实验复现原论文的方法，在 DPO 模型为 5% 训练量时验证效果（原文使用 0.1, 0.2, 0.4, 1.0 训练量）；提出新的 **Layerwise Extrapolation** 扩展方法；尝试进行多模态扩展，应用在 RLHF-V 数据集上以降低模型幻觉。代码与结果见 `github.com/assassinlike/ExPO-reproduction-and-extension`。

2 复现设计与结果

2.1 模型与数据

- **SFT 模型：** 与论文相同的监督微调模型 `zephyr-7b-sft-full`；
- **Aligned 模型 (DPO)：** 训练了 5% 数据的模型 `chujiezheng/zephyr_0.05`；
- **评测集：** AlpacaEval；
- **硬件环境：** 阿里云 GPU 实例（NVIDIA A10）。

2.2 参数与方法

项目	描述
α 选择	论文中：10%→8.5，20%→2.0；本文使用 $\alpha = 12$ 对应 5% 训练
内存控制	使用分块加载防止 OOM

2.3 复现结果

在 AlpacaEval 数据集上，ExPO 模型表现出如下特征：

- 模型倾向于在输出中复述指令后再回答问题；
- 复述指令后的回答质量正常；
- 在部分样本中仅出现复述（未作答）；

3 扩展方法：Layerwise Extrapolation

为改善直接外推的不稳定性，本文提出 **Layerwise Extrapolation**：

3.1 方法描述

将 Transformer 拆分为三个大层组，对每个层组使用不同外推系数 α：

层级	范围	α 值
Early Layers	前 1/3 层	5
Middle Layers	中间 1/3 层	10
Late Layers	后 1/3 层	15

3.2 实验结果

- **稳定性提升**：未再出现“仅复述不回答”的样本；
- **回答质量提高**：内容更自然、逻辑性更强；
- 仍存在复述指令的现象。

4 多模态扩展

RLHF-V 是一种用于降低多模态模型幻觉的对齐方法，本节将 ExPO 应用到这种方法，训练只使用 20% 数据的模型作为 aligned model，应用 ExPO。

4.1 实验设置

选取 `clip-vit-base-patch16` 作为视觉编码器，`open_llama_3b_v2` 作为大语言模型。使用线性层将 CLIP 的输出向量映射到隐藏层维度，连同 prompt 的特征向量一起作为 LLM 的输入。在进行训练之前，已使用 `init_infer.py` 进行测试，确保模型具有 MLLM 的推断能力，能输出一定质量的回答。参数选择见表1。

4.2 训练逻辑

`train_ddpo.py` 实现了完整的端到端训练逻辑：

1. 加载 HuggingFace 格式的数据集，选取 20% 数据；
2. 通过 CLIP 提取图像 embedding，并使用线性层映射到 LLM 隐藏维度；
3. 拼接图像 prefix embedding 与文本 embedding；
4. 分别计算 chosen/rejected 样本的 log 概率；

参数	值	说明
学习率	2×10^{-6}	
epoch	3	
global_batch_size	8	有效批大小
micro_batch	2	每步显存可承载的 batch
gradient_accumulation_steps	4	梯度累计次数
weight_decay	0.01	
warmup_ratio	0.03	学习率热身比例
β	0.1	DPO 损失温度系数
γ	2.0	corrected segment 权重
max_length_prompt	256	prompt 最大长度
max_length_response	256	response 最大长度

表 1: 多模态训练主要参数

5. 使用带 γ 加权的 segment 平均 $\log \pi$；
6. 基于 DPO 损失更新 LoRA 参数；
7. 使用 AMP + 梯度累积与 OOM 恢复机制，确保显存稳定；
8. 每个 epoch 保存 LoRA adapter 权重与 tokenizer。

4.3 问题与现状

由于多处代码细节处理不当，训练未能正常进行，出现了具体如混合精度缩放失败、张量尺寸不匹配等问题。不过多模态输入逻辑（CLIP $\rightarrow$ prefix embedding）已在推理阶段 `init_infer.py` 成功验证，实现了“图像 + prompt”双模态条件输入。

5 评估与限制

- **评估限制：** 由于支付账户注册限制，未能使用 GPT-4-turbo 进行自动评估，也即缺失win rate；
- **后续工作：**
  - 使用开源评测模型进行客观对齐评估；
  - 在 Layerwise Extrapolation 中进一步细化层级划分；
  - 尝试通过正则化或提示修正减少复述问题；

在多模态扩展方面：

- 修改多模态扩展的代码问题，完成 20% 数据量的 DDPO 模型训练；
- 得到“部分对齐”的模型后，应用 ExPO 方法。

6 结论

1. 成功复现了 CoAI 团队的 ExPO 方法；
2. 验证了在仅 5% 训练步数下仍可通过大 α 获得合理的对齐效果；
3. 提出了 **Layerwise Extrapolation** 扩展，提升生成稳定性；
4. 实现了多模态 DDPO 训练框架，确认 LLM 实际接收了图像与文本双输入；