LLM Alignment Week 7 Progression

Yuanqi Zhao

University of Oxford

25.Sep.2024

本周进度总结

- 已获得服务器权限（VPN+SSH），目前正在配置环境。
- 相关文献阅读，找寻所提供的 code。
- 将 LLM distillation 与 Alignment 相关知识点，与看过论文中的 idea 做成思维导图（正在制作），理顺相关知识点与该领域目前的发展趋势，能更快验证我的想法是否有效或者可做。

相关文献阅读

- 本周论文阅读总结：

  Distillation 领域存在的主要问题——“分布不匹配”：指在训练学生模型时使用的固定数据集（通常是教师模型生成的完美序列或真实标注数据）与学生模型在推理过程中自回归生成的序列之间存在差异。强化学习（LLM Alignment）可以缓解此现象。

  Distillation 中，forward KL-divergence 倾向于使得学习的分布的均值趋向于原分布，reverse KL-divergence 则倾向于使得学习的分布的形状趋向于原分布。在 distillation 任务中使用后者可以使得学生模型更好的学习到教师模型概率输出的多样性（也就是教师模型概率高与低的地方保持一致）。

  2024 ICLR 第一次提出 LLM Alignment + Distillation 的方法，来以此缓解分布不匹配问题。

本周所思

- 是否存在其他的 distribution similarity metric 更适合 Model distillation 的应用场景？
- 在联合训练过程中，根据学生模型的表现动态调整蒸馏和强化学习的权重是否会有更好的结果？
- 是否可以用 GAN 框架来加强模型蒸馏效果？用类似 VITS 的模型架构（GAN+VAE+FLOW）（还在构思是否可行）

下周所做

- 验证一些想法是否可行，以及配置服务器环境，开始相关实验。