InstructGPT summary

沙之洲 2020012408

InstructGPT 这篇文章提出了一种将语言模型和人类意图对齐的方法。具体来说，分为了三个步骤。首先，labeler 写一些对于特定 prompt 的人工回答，并用这些人工回答对模型进行 fine tune，也即 SFT；接着，labeler 对模型的输出进行排序，训练一个 reward model，之后用这个 reward model 对模型的输出进行排序；最后利用 PPO 算法，将 GPT 的输出用 reward model 进行排序之后，作为奖励反馈给模型。

InstructGPT 这篇文章的优点在于，首次提出了将 LLM 和人类意图进行对齐的想法，同时实现的结果也非常优秀。也正是因为这种对齐，使得 NLP 领域的模型能够大规模落地，为后续的工作拓宽了实际应用场景。这篇工作把强化学习引入到 LLM 的训练当中，开创了在 NLP 领域中使用控制论的先河。

对于这篇文章的缺点，首先是从这篇文章开始，openai 的工作开始不开源了，这对于其他 NLP 研究者是非常大的打击。其次，缺少对于 InstructGPT 训练的三个步骤的重要性评估，无法知道其中具体哪一个步骤在最终的模型性能提升上起到了关键作用。除此之外，这篇文章中对于训练细节的涉及非常少，在 finetune LLM 的时候，一些关键训练细节的确实会导致最终模型的性能大幅度下降，这也对后续想要复现或者 follow 这篇工作的研究者造成了非常大的困扰。

一些对于课上讨论的总结和对这篇工作未来的展望。课上的讨论中提到，PPO 的步骤可以理解为“帮模型学会什么东西是不好的”，因为 SFT finetune 只能让模型知道什么样的输出是好的，但是这个过程中没有负样本，所以模型会缺乏鉴别好坏的能力。而 reward model 通过学习对输出进行排序，然后模型通过 PPO 间接地学习到了 reward model 鉴别好坏的能力。除此之外，我认为 reward model 还实现了对于人类意图放大的作用，由于 labeler 标注的数据是有限的，可能没法完整地反应他们的意图，而通过 reward model 对于人类的行为进行拟合，可以通过有限的人工标注数据挖掘到一些深层次的意图信息。最后再将 LLM 和这些意图信息对齐，使得模型的输出更能符合人类的偏好。

最后是对于 InstructGPT 后续的一些思考。实际上 InstructGPT 是将 LLM 和 general 的人类意图进行对齐，这个对齐过程花费了很多 labeler 标注的数据以及 finetune 的时间。那么如果后续将 模型 从 general 的人类意图对齐到每一个用户具体的意图，所需要的时间和数据量是否可以更少？这样每个用户用到的模型都是一个很符合他们自己“口味” 的模型，这会大大提升用户的满意程度。同时，InstructGPT 中的 reward model 需要很多人类标注的排序数据才能逐渐模仿人类的偏好，是否可以通过 pre-train + finetune 的方式，用更少的数据模拟用户的偏好？最后，考虑到给每一个用户单独运行一个模型显然是不划算的，是否可以通过为每一个用户设计特殊 prompt 的方式，让模型针对于每一个用户，给出符合用户偏好的反馈？