Constitutional AI

沙之洲 2020012408

Constitutional AI 这篇工作提出了一种让 LLM follow 给定的 constitutions 的一种方法。具体来说，首先让 LLM 生成对于一些可能的 harmful 问题的回答；接着，从若干人为给定的 constitutions 里边sample一些作为 prior prompt，让模型去 revise 回答；通过这种方式构造出 harmless 的回答，将 harmless 回答作为 ground truth 来进一步 finetune LLM，最后得到遵循 constitutions 的 AI，也即更 harmless 的 AI

这篇工作的优点在于，进一步推动了 RLAIF 的实际应用。RLAIF，人类标注员不需要再去花费大量精力去标注海量样本，而只需要着重关注哪些 constitutions 能够带来人们想要的结果，比如说，一些 constitutions 可能能够让模型更 helpful，而另一些 constitutions 能够让模型更加 harmless，对于模型使用者不同的需求，本文提出的方法可以以更小的成本让模型达到预期的效果。

这篇工作的不足之处在于，我认为是没有对 helpful 进行一个比较合理的评价。在文中实验结果的曲线来看，helpful 和 harmless 呈现负相关的趋势。而实际上，对于 harmful 问题的 helpful 的回答对于现实世界是没有正向贡献的，所以 helpful 的指标应该将那些 harmful 问题的回答排除在外，甚至应该对 harmful 问题的 helpful 回答采取一定的惩罚。通过这种方法，或许能够进一步缓解本文中提到的 helpful 和 harmless 之间的张力。

对于这篇工作，有两个课上有趣的观点值得记录。首先，在 Constitutional AI 训练的过程中，RLAIF 过程上会用了 PPO 算法，而其中的 KL 散度作用和一般意义上 PPO 中 KL 散度的作用略有不同。在 RLAIF 过程中，PPO 的 KL 散度保证了 finetune 模型的参数不能偏移原始模型参数太多，这主要是考量到在 finetune 的过程中，不能因为为了追求更好的 harmlessness 而损失原本模型的语言能力。但是这种方法未免过于保守，很大程度上限制了模型 harmless 的能力。现在对于模型能力的保证是基于在参数上没有太大的调整，这种方法未免过于直接粗暴，是否可以考虑以另外更合理的度量方法保证模型的语言能力不衰减，比如说定期在 corpus 上对模型的能力进行测试，如果测试不通过（模型语言能力下降）则将模型参数回滚。

其次，再次回顾整个 Constitutional AI 的 pipeline，可以认为这个 pipeline 提供了一种将自然语言规则嵌入到模型参数中的一种方法。也即，首先通过原始模型 sample 出 answers，再通过加入 constitutions 让模型 revise，最后用 revised answers 去 finetune model。如此一来，关于 constitutions 的知识被加入到了模型的参数当中。这给未来的工作指出了一个方向，比如说，我们在模型的回答中发现模型缺少某一方面的知识，可以通过这个工作的 pipeline 将这一方面的知识嵌入到模型当中，也即，提供了一种显式地向模型输入人类知识的方法。