文章类型：论文|刊号（ISSN）：2972-4007(P) / 2972-4015(O)

重新思考大语言模型中的调优与对齐

文木源
GPT DESK PTE LTD
DOI:10.12238/deitar.v1i2.6554

[摘　要] 调优大语言模型面临两大挑战：数据质量和遗忘问题。首先,模型的性能极大地受到训练数据质量的影响。低质量数据导致的问题可以通过提高对齐数据的质量来缓解,例如通过更好的数据清洗、采购高质量数据集或使用迭代优化的半监督学习技术。其次,大语言模型面临的遗忘问题,即在针对新任务进行微调时,模型可能会忘记之前学到的信息。虽然提出了如弹性权重合并(EWC)和渐进神经网络等技术,但这些解决方案并不完美。为了应对这些挑战,提出了迭代改进模型和超越两阶段训练的概念。迭代模型训练方法涉及数据集和模型之间的循环相互改进,这可以使模型和数据集随时间不断优化。而超越两阶段训练,则建议采用持续完善数据和模型参数的迭代方法,其中模型的输出用于为下一阶段的数据管理提供信息。OpenChat框架,配备了条件化强化学习微调(C-RLFT),是微调语言模型的一个例子。这个框架利用混合质量数据进行微调,结合最大似然估计(MLE)和强化学习(RL),以提高模型的预测能力和自适应能力。OpenChat旨在通过类别条件和粗粒度奖励来完善模型的预测能力,同时生成更符合人类偏好的“人类的语言”。总体来看,OpenChat框架代表了向更复杂、更细致和更有效的语言模型训练迈出的关键一步,旨在处理各种数据质量并生成更可靠和通用的语言模型。
[关键词] 数据质量；遗忘问题；迭代改进；双阶段训练；条件化强化学习微调；OpenChat框架
中图分类号：TV149.2　**文献标识码**：A

Rethinking of Tuning and Alignment in Large Language Models

Muyuan Wen
GPT DESK PTE LTD

[Abstract] Tuning large language models faces two major challenges: data quality and forgetting problems.First, the performance of a model is greatly affected by the quality of training data. Problems caused by low- quality data can be mitigated by improving the quality of aligned data, for example through better data cleaning, sourcing high-quality datasets, or using iteratively optimized semi-supervised learning techniques. Secondly, large language models face the forgetting problem, that is, when fine-tuning for new tasks, the model may forget previously learned information. Although techniques such as Elastic Weight Combination (EWC) and Progressive Neural Networks have been proposed, these solutions are not perfect. To address these challenges, the concepts of iterative model improvement and beyond two-stage training are proposed. Iterative model training methods involve cyclical mutual improvements between the data set and the model, which allows the model and data set to be continuously optimized over time. Beyond two-stage training, it is recommended to adopt an iterative approach that continuously improves data and model parameters, where the output of the model is used to inform the next stage of data management. The OpenChat framework, equipped with Conditioned-RLFT, is an example of fine-tuning language models. This framework leverages mixed-quality data for fine-tuning, combining maximum likelihood estimation (MLE) and reinforcement learning (RL) to improve the model's predictive and adaptive capabilities. OpenChat aims to improve the model's predictive capabilities through category conditions and coarse-grained rewards, while generating "human language" that is more inline with human preferences. Overall, the OpenChat framework represents a key step toward more complex, detailed, and efficient language model training, designed to handle a variety of data qualities and

Copyright © This work is licensed under a Commons Attibution-Non Commercial 4.0 International License.

文章类型：论文|刊号（ISSN）：2972-4007(P) / 2972-4015(O)

produce more reliable and versatile language models.
[Key words] Data Quality; Forgetting Problem; Iterative Improvement; Two-Stage Training; Conditioned Reinforcement Learning Fine-Tuning; OpenChat Framework

一般来说,我们将大语言模型的训练分为两个阶段:一个是训练基础模型,另一个是调优。目前这两个阶段非常有用,而且非常有效。

市面上已经有一些很好的基础模型,其中相当一部分是开放源代码的。目前大家可以得到的一些通用的经验包括:

•更好的基础模型训练数据可以产生更好的对齐结果

•数据集质量对基础模型影响很大

•更多高质量的对齐数据产生更好的结果

全世界成千上万的技术团队正在添加越来越多的高质量数据来进行对齐。现在一次对齐实验需要数千个GPU小时,这已经达到了训练一个小型基础模型的水平。

我们无法区分很多改进是由“更好的对齐”引起的,或者我们只是简单地将模型拟合到更好的数据集上,从而改进了“基础模型”。

但对齐会遇到遗忘问题,并且对齐任务往往并不完美:

•当基础模型与特定数据集进行微调(对齐)时,它往往会“忘记”未包含在对齐数据集中的信息。这是神经网络训练中一个众所周知的问题(称为灾难性遗忘)的表现。

•在模型训练中,保留一般能力和优化特定子任务的性能之间存在“固有的权衡”。如果对齐过于狭隘地关注子任务,模型可能会失去其泛化能力,而恰恰是这一能力对于在更广泛的未来任务中表现良好至关重要。

•思想链(CoT)推理、记忆和总结等能力对于大语言模型的实用性尤其重要,尤其是当模型遇到对齐数据中未涵盖的任务时。

那么,如何将大型语言模型与特定任务结合起来而不影响其一般能力呢？常见的解决方案有:

•PPO-ptx[1]：解决遗忘问题的一种方法是将预训练数据纳入强化学习(RL)过程。这种方法使用近端策略优化(PPO)来微调模型,同时尝试保留预训练数据中的知识。然而,这种方法复杂、耗时且占用资源。

•在对齐数据中包含相关任务：更直接的方法是将相关任务添加到对齐数据集中。这种方法类似于基础模型的初始训练,有助于模型在各种任务上保持其性能。本质上,它涉及不断扩展对齐数据集以涵盖更多任务,这可能成为一个广泛且持续的过程。

当训练损失接近零时,模型发生的变化较少,遗忘问题变得不那么明显。在这种状态下,模型的参数相对稳定,表明它已经将任务学习到了令人满意的水平。然而,在不损失泛化能力的情况下达到如此低损失的状态是具有挑战性的。

提高大语言模型在特定任务上的表现同时保持其广泛能力有相当的难度。我们需要在对齐过程中仔细平衡,并考虑各种策略以防止遗忘,同时又不牺牲模型的灵活性和适应性。

1 调优大语言模型的关键挑战

调优大语言模型的挑战,我认为可以总结成两大关键问题:

1.1数据质量问题:

基础模型的性能很大程度上受到其训练数据质量的影响。机器学习领域有一句老话,“垃圾进,垃圾出”,形象生动地反映了这样一个现实：使用低质量数据训练的模型几乎必然产生糟糕的结果。

提高对齐数据的质量可以有效缓解此问题。已经有许多方法可以有效提高输入数据的质量,例如更好的数据清洗、采购高质量的数据集,甚至可以使用迭代优化数据的半监督学习技术。

1.2遗忘问题:

大语言模型中的熵,意味着模型行为的随机性或不可预测性。当模型针对新任务进行微调时,它往往会忘记以前学到的信息。

这是一个复杂的问题,没有简单的解决方案。人们已经提出了诸如弹性权重合并(EWC)或渐进神经网络之类的技术来对抗遗忘,但它们还远远称不上完美。对这些问题的反思导致我们考虑采用更具迭代性的模型训练方法:

(1)迭代改进:在数据集和模型之间迭代不失为一种动态训练的好方法,其中的模型和数据集在一个循环中相互改进。经过初步训练后,该模型可用于评估甚至增强数据集的质量,然后用于进一步的训练。从理论上讲,这个过程可以使模型和数据集随着时间的推移不断改进。

(2)超越两阶段训练:当前的标准方法通常涉及两阶段策略：预训练,然后进行微调。为什么不采用不断完善数据和模型参数的迭代模型。这可能涉及训练阶段和数据细化阶段,其中模型的输出用于为下一阶段的数据管理提供信息,反之亦然。

(3)迭代优势:迭代模型可以使大语言模型的知识更加最新和全面,通过不断重新访问各种数据点来减少遗忘的影响,并逐步完善模型跨任务泛化的能力。

从本质上讲,我们需要一种更细致、更连续的方法来培训大语言模型,这种方法将数据质量和模型培训视为一个持续过程中相互关联的部分,而不是离散的步骤,从而可以带来更强大、适应性更强和知识更丰富的模型。

2 重新思考大语言模型的调优

最新的发现是,大语言模型的“条件”对调优有很大帮助。

•开放源代码的大语言模型OpenChat提出了一种名为“条件强化学习微调”(C-RLFT)的方法[2],该方法在微调数据集中添加条件并帮助网络学习价值函数,从而在具有相同参数大小的所有模型中获得了最好的结果。

Copyright © This work is licensed under a Commons Attibution-Non Commercial 4.0 International License.

文章类型：论文|刊号（ISSN）：2972-4007(P) / 2972-4015(O)

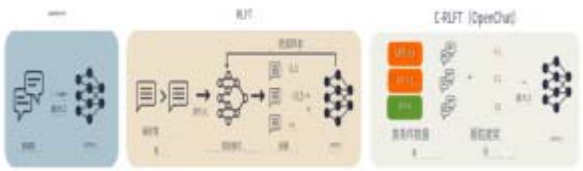

图1：我们提出的带有条件 RLFT 的 OpenChat 框架，与之前的监督微调 (SFT) 方法和强化学习微调 (RLFT) 方法相比，可利用混合质量数据推进开源语言模型微调。MLE 和 RL 分别表示最大似然估计和强化学习。

上图中展示了OpenChat with Conditioned- RLFT(强化学习微调)的框架,旨在使用混合质量数据改进开源语言模型的微调。我们不妨将该框架与之前的监督微调(SFT)和强化学习微调(RLFT)方法进行比较：

2.1 SFT(监督微调)：

它从一个数据集开始,通过最大似然估计(MLE)方法输入大语言模型(LLM)。这种传统方法使用数据集来训练模型,以根据输入预测下一个单词或序列。

2.2 RLFT(强化学习微调)：

使用偏好数据(代表人类偏好或对模型输出的判断)来训练奖励模型,奖励模型根据完成情况(模型输出)与偏好数据的匹配程度来输出奖励。最大似然估计和强化学习都有助于训练LLM,并通过强化学习方面的奖励来提供信息。

2.3 C-RLFT(条件强化学习微调)：

OpenChat使用两个版本的生成式预训练Transformer模型：GPT-3.5和GPT-4并以它们的输出结果作为基础。然后,以某些类或类别为条件的数据集与GPT-3.5结合使用以生成样本补全。OpenChat使用了粗粒度奖励,用于通过MLE方法微调GPT-4,这表明奖励并不像RLFT方法中那样精细区分。

带有Conditioned-RLFT的OpenChat框架旨在利用混合质量数据进行微调。该框架的定位是通过将条件数据和强化学习纳入训练过程来改进SFT和RLFT。

OpenChat框架旨在利用MLE的预测能力和RL的自适应能力,通过使用质量不同的数据更有效地微调语言模型。这种方法使得模型通过从更细致和更现实的数据集中学习来更好地理解和生成类似人类的文本。事实上,我们观察到我们的模型通过这些粗粒度的奖励来学习价值函数,并且响应的质量可以通过类别条件进行调整。

带有Conditioned-RLFT的OpenChat框架代表了一种微调语言模型的创新方法。该框架旨在通过更有效地利用混合质量数据的方式结合最大似然估计(MLE)和强化学习(RL)来改进传统的监督微调方法。通过使用类条件数据集和粗粒度奖励,OpenChat与C-RLFT旨在完善模型的预测能力,同时仍然符合人类偏好,从而更好地调整语言模型以生成“人类的语言”。

总体而言,OpenChat框架是朝着更复杂、更细致、更有效的语言模型训练迈出的关键一步,可以处理各种数据质量并生成更可靠和通用的语言模型。

[参考文献]

[1]Zheng,Rui,etal.”Secrets of RLHF in large language models part I:PPO.”arXiv preprint arXiv:2307.04964(2023).

[2]https://arxiv.org/abs/2309.11235.

Copyright © This work is licensed under a Commons Attibution-Non Commercial 4.0 International License.