大模型（LLMs）训练集面

来自： AiGC面试宝典

宁静致远

2023年12月24日 00:33

1. SFT（有监督微调）的数据集格式?

一问一答

2. RM（奖励模型）的数据格式?

一个问题 + 一条好回答样例 + 一条差回答样例

3. PPO（强化学习）的数据格式?

理论上来说，不需要新增数据。需要提供一些prompt，可以直接用sft阶段的问。另外，需要限制模型不要偏离原模型太远（ptx loss），也可以直接用sft的数据。

4. 找数据集哪里找?

推荐Alpaca-COT，数据集整理的非常全，眼花缭乱。

5. 微调需要多少条数据?

取决于预训练数据和微调任务的数据分布是否一致，分布一致，100条就够，分布差异大就需要多些数据，千条或者万条以上为佳。
自己的任务复杂或者下游任务行业比较冷门，如药品名称识别任务，则需要较多监督数据。还有微调大模型时，一遍是记不住的。100条的微调数据，epochs=20才能稳定拟合任务要求。

6. 有哪些大模型的训练集?

预训练数据集togethercomputer/RedPajama-Data-1T「红睡衣」开源计划总共包括三部分：

- 高质量、大规模、高覆盖度的预训练数据集；
- 在预训练数据集上训练出的基础模型；
- 指令调优数据集和模型，比基本模型更安全、可靠。

预训练数据集RedPajama-Data-1T已开源，包括七个子集，经过预处理后得到的token数量大致可以匹配Meta在原始LLaMA论文中报告的数量，并且数据预处理相关脚本也已开源。
完整的RedPajama-Data-1T数据集需要的存储容量为压缩后3TB，解压后5TB。
CoT微调数据集：Alpaca-CoT 里面包括常用的alpaca，CoT等数据集，有中文的。

7. 进行领域大模型预训练应用哪些数据集比较好?

通过分析发现现有的开源大模型进行预训练的过程中会加入数据、论文等数据。主要是因为这些数据的数据质量较高，领域相关性比较强，知识覆盖率（密度）较大，可以让模型更适应考试。给我

们自己进行大模型预训练的时候提供了一个参考。同时领域相关的网站内容、新闻内容也是比较重要的数据。

8. 如何选取和构建大模型微调数据？

- 动机：在 微调大模型时，首先需要解决的问题是“选取和构建大模型微调数据”，那如何选择呢？
- 问题一：什么样的 数据 才是 最优的 大模型微调数据？

1. 数据的多样性：

一般情况下我们数据的分布都是符合一个长尾分布的。主要的几个类别数据占据了90%的数据量，剩下的90%的类别只有10%的数据量。

举个栗子：小红书上，query的意图识别里，美食，穿搭，旅游攻略类非常多，但是还有一些同学去搜大模型微调的数据技巧。

如果说我们直接采样一批线上的图文文本，直接送给标注的话，会存在一个严重的问题：他们标注的数据大部分都是攻略类，技术类比较少，标了3个月才攒了几千条大模型技术文本，但是攻略类已经成几万了。

这样搞肯定是不行的，人力成本方面的消耗是在是太大了，并且模型因为数据平衡的问题也没有特别好

1. 数据的标注质量；
2. 数据的不确定性；

- 问题二：如何构建 大模型微调数据？

3. 方法一：“self-instruct”的框架，通过自我生成来提升指令跟随能力。文章的流程是从语言模型中生成指令、输入和输出样本，然后在使用这些数据微调原始模型之前进行清洗。
4. 方法二：“主动学习”

主动学习有两个基本原则，在监督训练的时候，注意主动发现数据的两个方面，一个是数据多样性，另外一个是数据的不确定性。这样讲是比较抽象的概念，那我们在大模型实践中如何体现呢？

第一，数据的多样性。

多样性即为数据的去重，去重这件事的核心是相似度度量，现在的相似度度量方法大家用的比较多的是基于对比学习构造的语义向量这套思路，当然简单的基于词袋或者tfidf的方案也是可以的。有了核心的相似度度量方法后，我们可以使用简单的onepass聚类方法进行过滤，考虑复杂一点的话，我们可以使用带优化目标的聚类：比如K-Center-Greedy算法，其约束条件是在最大化多样性的情况下，使指令数据集最小。

另外，如果我们已经有了一批已经去重的人工处理过的高质量数据，那么我们如何寻找与这批数据不一样的数据呢？

这里有一个非常简单实用的方案，并且这个方案可以用在很多其他的地方。

我们简单地把已有的数据全部当成正样本打上1，然后待筛选的数据全部当成负样本打上0，我们使用deberta等构建二分类模型，并进行K-fold的交叉验证，在交叉验证过程中，选出每一个fold过程中的测试集合里概率接近于0的样本。

通过这样的操作，就能把长得与已有数据不一样的数据给选出来了，并且这个过程是半监督的。

这套方案也可以用在很多其他地方，比如数据质量选择，只要我们有一批已经确定标签/结果/标注的种子数据，就能通过这样的方法选出与种子数据长得比较像的，长得不像的。

第二，数据的不确定性。

数据的不确定性主要体现数据的质量筛选上，选取模型学的不那好的数据，模型没有把握的数据。

最简单的，我们可以选出模型对应PPL值比较差的那批数据。如果是指令数据的话，比如大模型做题和对应的答案。我们可以把所有选项对应的概率之和计算出来，然后过滤出概率和比较低的那一批数据，这批数据就是模型“不太肯定”的样本，我们需要加强针对性的训练。

当然这样可能有一个副作用，就是这批数据是质量比较差而不是模型学的不太好的。

为此，我们还要借助reward model，这个reward model是广义的，他是一个质量的二分类模型。可以祭出我们的deberta，继续用标注数据进行做二分类，进行数据质量的判断。

有了质量打分模型后，我们就可以判断一些指令数据的质量高低，并且据此选出模型真正不确定的数据。

这个过程类似于手动的拒绝采样，核心是选择“模型不确定”+“数据质量达标”的那部分数据。

- 总结一下:监督学习中主动学习的两个基本原则是寻找多样性的数据，模型不确定性的数据，在寻找的过程中，我们使用了一些小技巧，比如聚类去重，对抗半监督过滤，自建reward二分类等方法。这几个小技巧，学术上没有什么高深莫测的东西，都是实践中总结出来的好用的方法。