大模型面经——超细节大模型训练与微调实操经验总结（上）

原创 喜欢卷卷的瓦力 瓦力算法学研所 2024年05月11日 09:14 广东

◈ 技术总结专栏 ◈

作者：vivida

瓦力算法学研所
我们是一个致力于分享人工智能、机器学习和数据科学方面理论与应用知识的公众号。我...
117篇原创内容

公众号

本篇主要从训练数据预处理、模型结构、训练参数设置与错误处理四大角度比较细节地分享大模型微调经验。

大模型的训练和微调过程相对于以前NLP中fine-tuning模式存在一些新的坑，并且做一些简单的消融实验相对于以前的模式试错成本也更高；此外目前很多算法工程师更多精力都放在了处理数据上， 工作之余很难有精力去做探索实验。

所以小伙伴们在实践前可以多看看一些通用的实践经验，带着一些先验知识去探索，尽量规避自己陷入一些无意义的坑中。

本篇将开启一个新系列，尽量细节的讲讲大模型中训练和微调的经验。

本篇主要从**训练数据预处理、模型结构、训练参数设置与错误处理**四大角度来谈经验，下面是一个问题的快捷目录。

1. 拿到业务产生的一批新的对话数据，需要进行SFT，怎样对这批数据进行优化?
2. 模型训练时，历史对话长度是不是设置得越长越好，一般设置多少?
3. 模型训练样本量规模增大，导致训练任务直接报OOM了，该怎么办?
4. 微调大模型的时候在模型结构方面有哪些经验?
5. 微调大模型的时候训练配置一般是怎样的?
6. 微调大模型时出现错误崩溃该怎么办?

拿到业务产生的一批对话数据，需要进行SFT，怎样对这批数据进行优化?

1. 上下文内容处理

考虑具体模型历史对话长度，输入历史对话数据进行左截断， 保留最新的对话记录。

2. 语句顺滑处理

把一些口语化的语气词、语法错误等进行顺滑，如嗯嗯、呃、啊啊之类的口语词。

3. 去掉一些敏感或不合适的内容

这里可以从整句和词的角度来考虑。

- **整句**

可以基于如fasttext等模型训练一个简单的文本分类模型，把价值观不正确的或不合适的样本数据筛出来；

还可以训练一个奖励模型，奖励模型怎么训练可以参考大模型强化学习实操（一）——如何训练一个自己偏好的大模型（附代码）这篇内容

- **词**

这里比较直接，可以设置一个敏感词列表。

4. 扩充用户特征标签

基于年龄、性别、地域、人群等，针对对话的用户做一个特征标签，可以便于后期分析，做其他实验等。

模型训练时，历史对话长度是不是设置得越长越好，一般设置多少？

这个消融实验是这么设计的，选同一个模型，分别用两种方案训练，变量是max_source_length和max_target_length，对训练好之后的模型从Loss、Bleu指标、离线人工评估等角度进行对比分析。

下面直接附上结论：

基于现有显存条件，从人工评估少量样本以及loss下降来看，历史对话长度设置得越长越好。历史对话长度**1024比512长度好**，后续如果训练可能上线模型，**可以扩大到1024长度。**

模型训练样本量规模增大，导致训练任务直接报OOM了，该怎么办？

1. 方案

对数据并行处理，核心思想是**使数据向量化耗时随处理进程的增加线性下降**，训练时数据的内存占用只和数据分段大小有关，可以根据数据特点，灵活配置化。

2. 具体操作

- 均分完整数据集到所有进程（总的GPU卡数）；
- 每个epoch训练时整体数据分片shuffle一次，在每个进程同一时间只加载单个分段大小数据集；
- 重新训练时可以直接加载向量化后的数据。

微调大模型的时候在模型结构有哪些经验?

- 模型结构：目前都用Causal Decoder + LM。有很好的zero-shot和few-shot能力，涌现效应
- Layer normalization：使用Pre RMS Norm
- 激活函数: 使用GeGLU或SwiGLU
- Embedding层后不添加layer normalization，否则会影响LLM的性能
- 位置编码: 使用ROPE或ALiBi。ROPE应用更广
- 去除偏置项: 去除dense层和layer norm的偏置项，有助于提升稳定性

微调大模型的时候在训练配置方面有哪些经验?

- Batch size: 大模型在硬件显存满足的情况下，一般batch size越大越好， 建议选用很大的batch size; 后期动态地增加batch size的策略，GPT3逐渐从32K增加到3.2M tokens。
- 学习率设置: 先warmup再衰减。学习率先线性增长，再余弦衰减到最大值的10%。最大值一般在5e-5到1e-4之间。
- 梯度裁剪: 通常将梯度裁剪为1.0。
- 权重衰减: 采用AdamW优化器，权重衰减系数设置为0.1Adamw相当于Adam加了一个L2正则项。
- 混合精度训练: 采用bfloat16，而不是foat16来训练。

微调大模型时出现错误崩溃该怎么办?

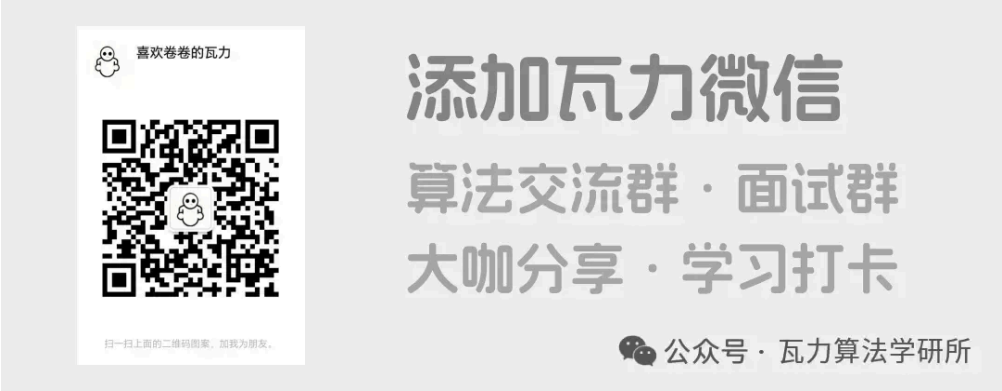
前面都好好的，过某个shard的时候突然崩溃了大概率是数据问题。

选择一个好的断点，跳过训练崩溃的数据段，进行断点重训。

选择一个好的断点的两点标准:

- 损失标度 lossscale>0;
- 梯度的L2范数<一定值 且 波动小。

本系列将会持续更新，想要获取面经资料的同学欢迎关注公众号，进群一起交流~



修改于2024年05月11日