

[618提前购]云产品限时特惠

立即抢购 >

备案

控制台

学习

实践

活动

专区

工具

更多

文章/答案/技术大牛

写文章

提问

登录/注册

社区首页 > 专栏 > 小七的各种胡思乱想 > 正文

解密Prompt系列6. lora指令微调扣细节-请冷静,1个小时真不够~

原创

发布于 2023-04-29 14:37:55

1.3K

0

举报

上一章介绍了如何基于APE+SELF自动化构建指令微调样本。这一章咱就把微调跑起来，主要介绍以Lora为首的低参数微调原理，环境配置，微调代码，以及大模型训练中显存和耗时优化的相关技术细节

标题这样写是因为上周突然收到了一周内上线一版chatbo的命令，原因无它领导们都刷到了《一个小时你也可以拥有ChatGPT》，《100美金训练ChatGPT》，《仅训练3小时超越ChatGPT》，《人人都可以拥有ChatGPT》。。。领导说人人都有了为啥我没有呀？！！真诚呼吁标题党们手下留情，留人一命！于是这里我换个标题来Debuff！Debuff！

看到这里本文最重要的部分已经说完了，累了的小伙伴可以撤退了，五一快乐~



低参数微调原理

LORA: LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS

原理: INTRINSIC DIMENSIONALITY EXPLAINS THE EFFECTIVENESS

OF LANGUAGE MODEL FINE-TUNING

前人的肩膀: Adapter: Parameter-Efficient Transfer Learning for NLP

我们之前在解密Prompt系列3. 冻结LM微调Prompt介绍过一些soft-prompt，包括P-Tuning和Prompt-Tuning也属于低参数微调。这些方案是通过参数拼接的方案引入额外参数。这里介绍另一类方案，同样是冻结LLM的参数，通过参数相加的方案引入额外参数。相较soft-prompt最明显的优势，就是不会占用输入token的长度。

LoRA的原理比较简单，原始全量微调其实就是在原始模型参数上通过微调加入增量 $W = W_0 + \Delta W$ ，那我们可以通过冻结原始参数 W_0 ，并且把增量部分通过低秩分解方式进一步降低参数量级 $\Delta W = A * B^T$ ，原始参数的维度是 $d * d$ ，则低秩分解后的参数量级是 $2 * r * d$ ，因为这里的 $r \ll d$ ，因此可以起到大幅降低微调参数量级的效果，如下图

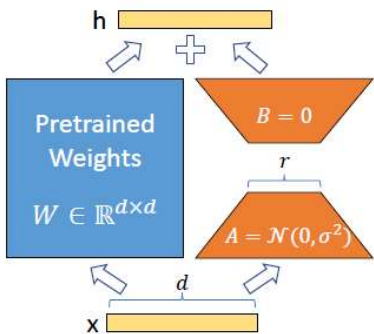


Figure 1: Our reparametrization. We only train A and B.

核心代码如下

```
1  ## 初始化低秩矩阵A和B
2  self.lora_A.update(nn.ModuleDict({adapter_name: nn.Linear(self.in_features, r, bias=False)}))
3  self.lora_B.update(nn.ModuleDict({adapter_name: nn.Linear(r, self.out_features, bias=False)}))
4  self.scaling[adapter_name] = lora_alpha / r
5
6  ## 向前计算
7  result = F.linear(x, transpose(self.weight, self.fan_in_fan_out), bias=self.bias)
8  result += (
9      self.lora_B[self.active_adapter](
```

关于作者

风雨中的小七

LV2

71

64.9K

203

55

文章

累计阅读量

获赞

作者排名

关注

前往专栏

社区活动

有奖征文 | 玩转Cloud Studio

参与活动，赢取限量周边礼品

立即参加

邀请好友加入自媒体分享计划

邀请好友，同享奖励 30 / 100 / 180 元云服务器代金券

立即邀请

精选专题

腾讯云原生专题

云原生技术干货，业务实践落地。

广告

腾讯云

GPU云服务器限时秒杀

4元/天 起

立即抢购

GPU

关注 - 腾讯云 开发者 公众号

将获得

10元无门槛代金券

洞察腾讯核心技术

剖析业界实践案例



切换旧版

领券

```
10         self.lora_A[self.active_adapter](self.lora_dropout[self.active_adapter](x))
11     )
12     * self.scaling[self.active_adapter]
13 )
```

论文测试了在多数场景下适当的LORA微调和全量微调的效果不相上下。一个可能原因是INTRINSIC DIMENSIONALITY论文中提出，虽然语言模型整体参数空间很大，但具体到每个任务其实有各自的隐表征空间(intrinsic dimension)，这个隐表征空间的维度并不高,因此在微调过程中加入低秩分解并不一定会影响微调效果。使用LORA微调有以下几个细节

1. 对哪些参数进行微调：基于Transformer结构，LORA只对每层的Self-Attention的部分进行微调，有 W_q, W_k, W_v, W_o 四个映射层参数可以进行微调。消融实验显示只微调 W_q 效果略差，微调 W_q, W_v 的效果和微调 W_q, W_k, W_v, W_o 的效果相似。需要注意不同模型参数名称不同，像chatglm对应的参数名称就是query_key_value
2. Rank的选取：Rank的取值作者对比了1-64，效果上Rank在4-8之间最好，再高并没有效果提升。不过论文的实验是面向下游单一监督任务的，因此在指令微调上根据指令分布的广度，Rank选择还是需要8以上的取值进行测试。
3. alpha参数：alpha其实是个缩放参数，本质和learning rate相同，所以为了简化我默认让alpha=rank，只调整lr，这样可以简化超参
4. 初始化：A和Linear层的权重相同Uniform初始化，B是zero初始化，这样最初的Lora权重为0。所以Lora参数是从头学起，并没有那么容易收敛。

Lora的优点很明显，低参数，适合小样本场景；可以拔插式的使用，快速针对不同下游任务训练不同的lora权重；完全没有推理延时，这个在后面代码中会提到推理时，可以预先先把lora权重merge到原始权重上。

但Lora微调虽好，个人在尝试中感受到的局限性就是adapter类的微调方案可能更适合下游单一任务类型/生成风格。至于是否适合作为通用指令微调的解决方案，有个问题我也没有搞懂，就是通用的指令样本是否真的有统一的低秩空间表征？这个表征又是什么含义？因为指令微调阶段的样本其实是混合的多任务指令样本，这种情况下lora是否合适，感觉需要更全面的评估（当前出来的众多LLama们都缺少合理统一全面可比的Evaluation），当前就我们的尝试情况lora的效果并不及预期。

环境配置

GPU 云服务厂商对比

我用了featurize和揽睿星海。云服务厂商的选择主要看是否有jupyter，存储够大，下载快，能连git，有高配torch环境。这两家在众多小厂里脱颖而出，4090的卡一个小时也就3块钱，来来来盆友辛苦把推广费结一下~

强调下环境配置，想跑通微调，搞定环境你就成功了80%！运气好1分钟，运气差1天都在原地打转

1. 实例环境：TRX4090 + py38 + torch2.0 + CUDA12
2. python环境：主要坑在transformers和peft，几个相关issue包括：[llama tokenizer special token有问题](#)，[peft adapter.bin微调不更新](#)，[Bug with fan_in_fan_out](#)。我一个不差都踩中了。。。

```
1  # 以下配置可能会随时间变化，出了问题就去issue里面刨吧
2  # 要相信你不是唯一一个大冤种！
3  accelerate
4  appdirs
5  loralib
6  bitsandbytes
7  black
8  black[jupyter]
9  datasets
10 fire
11 transformers>=4.28.0
12 git+https://github.com/huggingface/peft.git
13 sentencepiece
14 gradio
15 wandb
16 cpm-kernel
```

模型初始化

以下代码主要整合自alpaca-lora和chatglm-finetune。其实lora微调的代码本身并不复杂，相反是如何加速大模型训练，降低显存占用的一些技巧大家可能不太熟悉。模型初始化代码如下，get_peft_model会初始化PefitModel把原模型作为base模型，并在各个self-attention层加入lora层，同时改写模型forward的计算方式。

主要说下load_in_8bit和prepare_model_for_int8_training，这里涉及到2个时间换空间的大模型显存压缩技巧。

```
1  from peft import get_peft_model, LoraConfig, prepare_model_for_int8_training, set_peft_model_state_dict
2  from transformers import AutoTokenizer, AutoModel
3
4  model = AutoModel.from_pretrained("THUDM/chatglm-6b", load_in_8bit=True, torch_dtype=torch.float16)
5  tokenizer = AutoTokenizer.from_pretrained("THUDM/chatglm-6b", trust_remote_code=True)
6  model = prepare_model_for_int8_training(model)
7
8  lora_config = LoraConfig(
9      task_type=TaskType.CAUSAL_LM,
10     inference_mode=False,
11     r=8,
12     lora_alpha=8,
13     lora_dropout=0.05,
14 )
15 model = get_peft_model(model, lora_config)
16 model.config.use_cache = False
```

模型显存占用分成两个部分，一部分是静态显存基本由模型参数量级决定，另一部分是动态显存在向前传播的过程中每个样本的每个神经元都会计算激活值并存储，用于向后传播时的梯度计算，这部分和batchsize以及参数量级相关。以下8bit量化优化的是静态

切换旧版

领券

显存，而梯度检查优化的是动态显存。

1. 8bit Quantization

<https://huggingface.co/blog/hf-bitsandbytes-integration>

[点击展开阅读全文](#)

原创声明：本文系作者授权腾讯云开发者社区发表，未经许可，不得转载。
如有侵权，请联系 cloudcommunity@tencent.com 删除。

promptnlpchatgpt

[登录](#) 后参与评论

相关文章

计算机专用英语词汇1695个词汇表

特别感谢： 不愿意透露姓名的小虾同学提供的音标部分 1.单词说明：command n. 命令，指令 [kə'mɑ:nd 单词拼写 名词 单词含义 音标 （发...

2017-12-272.9K

JavaScript学习总结(一)——ECMAScript、BOM、DOM（核...

2018-01-042.2K

Faster RCNN：RPN，anchor，sliding windows

2018-01-091.7K

Java面试笔试题大汇总(最全+详细答案)

其他

声明：有人说,有些面试题很变态，个人认为其实是因为我们基础不扎实或者没有深入。本篇文章来自一位很资深的前辈对于最近java面试题目所做的总结归纳，有170道题目，知识面很广，而且这位前...

汤高2018-01-1126.7K0

linux运维中的命令梳理（四）

其他

-----管理命令----- ps命令：查看进程 要对系统中进程进行监测控制，查看状态，内存，CPU的使用情况，使用命令：/bin/ps （1） ps：是显示瞬间进程的状态，并不动态连续； （2） top：如果想对进程运行时间监控，应该用 top 命令； ...

洗尽了浮华2018-01-235.6K0

C++、腾讯、小米背后的科技原力

人工智能

原力有多少人是通过《原力觉醒》这部电影听到这个词的？这部电影来自“星球大战”系列，在许多人心 中埋下了深厚的情结，更让大家知道了原力。原力没有固定的定义，我们可以把它当做万物生长和所...

腾讯技术工程官方号2018-01-294970

BAT机器学习面试1000题系列（第76~149题）

机器学习

76、看你是搞视觉的，熟悉哪些CV框架，顺带聊聊CV最近五年的发展史如何？深度学习 DL应用 难 原 英文： adeshpande3.github.io 作者： Adit Deshpande，UCLA CS研究生 译者： 新智元闻菲、胡祥杰...

用户13324282018-03-091.2K0

Java面试复习大纲2.0（持续更新）

java

想要成为合格的Java程序员或工程师到底需要具备哪些专业技能，面试者在面试之前到底需要准备哪些东西呢？本文陈列的这些内容既可以作为个人简历中的内容，也可以作为面试的时候跟面试官聊的东...

切换旧版

领券

Java帮帮	2018-03-15	1.4K	0
--------	------------	------	---

经典Java面试题收集

java			
1、面向对象的特征有哪些方面？ 答：面向对象的特征主要有以下几个方面： 抽象：抽象是将一类对象的共同特征总结出来构造类的过程，包括数据抽象和行为抽象两方面。抽象只关注对象有哪些属性和...			
nnngu	2018-03-15	1.1K	0

Java面试复习大纲更新1.0（持续更新）

java			
1、背熟你的简历 原因：面试的第一个问题，一般都是让你简单介绍下你自己，或者介绍一下你最近的项目，而一个面试官，如果连自己的简历都无法熟知，对里面提到的项目、技术都无法描述清楚的话...			
Java帮帮	2018-03-15	923	0

R语言函数的含义与用法，实现过程解读

r语言	数据处理	大数据	
R的源起 R是S语言的一种实现。S语言是由 AT&T贝尔实验室开发的一种用来进行数据探索、统计分析、作图的解释型语言。最初S语言的实现版本主要是S-PLUS。S-PLUS是一个商业 软件， 它基于S语言， 并由MathSoft公司的统计科学部进一步完善。后来...			
学到老	2018-03-19	1.5K	0

37.Django1.11.6文档

django	数据库	python	
第一步 入门 检查版本 python -m django --version 创建第一个项目 django-admin startproject mysite 运行 python manage.py runserver 更改端口 python manage.py runserver 8080 更改IP python...			
zhang_derek	2018-04-11	8.4K	0

iOS面试中被面试官问到的问题答案(一)

ios	单元测试	mvc	
2016-03-1016:30:14 发表评论 1,091°C热度 1.请你谈谈static和宏定义的区别。什么时候用static什么时候用宏定义。 让你声明的常量只在你声明的文件里有作用要编译器会保存 2.你是怎么看代理和通知的 他们有什么区别？ 3.说说你对内存管理的理解。 4....			
timhbw	2018-05-03	1.4K	0

iOS开发学习路线

ios	开源		
这里筑梦师,是一名正在努力学习的iOS开发工程师,目前致力于全栈方向的学习,希望可以和大家一起交流技术,共同进步,用简书记录下自己的学习历程.			
筑梦师winston	2018-05-14	1.3K	0

「我是可微分编程的粉丝」， Gary Marcus再回应深度学习批判言论

其他			
选自Medium 作者： Gary Marcus 机器之心编译 近日， Gary Marcus 针对各研究者与开发者的评论作出了回应，他从什么是通用人工智能开始回应了常见的 14 个质疑或问题，其中就包括 LeCun 所说的「mostly wrong」。此外， Marcus 还重申了他对深度学...			
机器之心	2018-05-11	512	0

[Java面试三]JavaWeb基础知识总结.

java			
1.web服务器与HTTP协议 Web服务器 I WEB，在英语中web即表示网页的意思， 它用于表示Internet主机上供外界访问的资源。 I Internet上供外界访问的Web资源分为： * 静态web资源（如html 页面）： 指web页面中供人们浏览的数据始终是不变。 * 动态w...			
一枝花不算浪漫	2018-05-18	1.5K	0

入侵总统DNA，未来犯罪新手段

大数据	网络安全		
作者 Andrew Hessel,Marc Goodman,Steven Kotler 来源： 译言 网站： http://select.yeeyan.org 据传， 美国政府一方面正在秘密收集各国领导人的DNA， 另一方面也在不遗余力保护巴拉克·奥巴马的DNA。一旦这些基因密码被破解， 就可能泄露一些对当事人...			
大数据文摘	2018-05-04		

切换旧版


领券

大数据文摘	2018-05-21	748	0
手把手教你从零搭建深度学习项目（可下载PDF版）			
其他			
第一部分：启动一个深度学习项目 1. 应该选择什么样的项目？ 很多人工智能项目其实并没有那么严肃，做起来还很有趣。2017 年初，我着手启动了一个为日本漫画上色的项目，并作为我对生成对抗网络（GAN）研究的一部分。这个问题很难解决，但却很吸...			
昱良	2018-06-25	740	0

手把手教你从零到一搭建深度学习项目			
深度学习	nat		
在学习了有关深度学习的理论之后，很多人都会有兴趣尝试构建一个属于自己的项目。本文将会从第一步开始，告诉你如何解决项目开发中会遇到的各类问题。			
数据派THU	2018-07-30	524	0

手把手教你从零搭建深度学习项目（附链接）			
深度学习	nat		
本文共1万+字，建议阅读10+分钟。 本文将会从第一步开始，教你解决项目开发中会遇到的各类问题。			
数据派THU	2018-07-30	551	0

点击加载更多

社区		活动		资源		关于		腾讯云开发者	
专栏文章		自媒体分享计划		技术周刊		社区规范		 扫码关注腾讯云开发者	
阅读清单		邀请作者入驻		社区标签		免责声明			
互动问答		自荐上首页		开发者手册		联系我们			
技术沙龙		技术竞赛		开发者实验室		友情链接			
技术视频									
团队主页								扫码关注腾讯云开发者 领取腾讯云代金券	
腾讯云TI平台									
热门产品		域名注册 云存储	云服务器 视频直播	区块链服务	消息队列	网络加速	云数据库	域名解析	
热门推荐		人脸识别 SSL 证书	腾讯会议 语音识别	企业云	CDN加速	视频通话	图像分析	MySQL 数据库	
更多推荐		数据安全 网站监控	负载均衡 数据迁移	短信	文字识别	云点播	商标注册	小程序开发	

切换旧版

领券