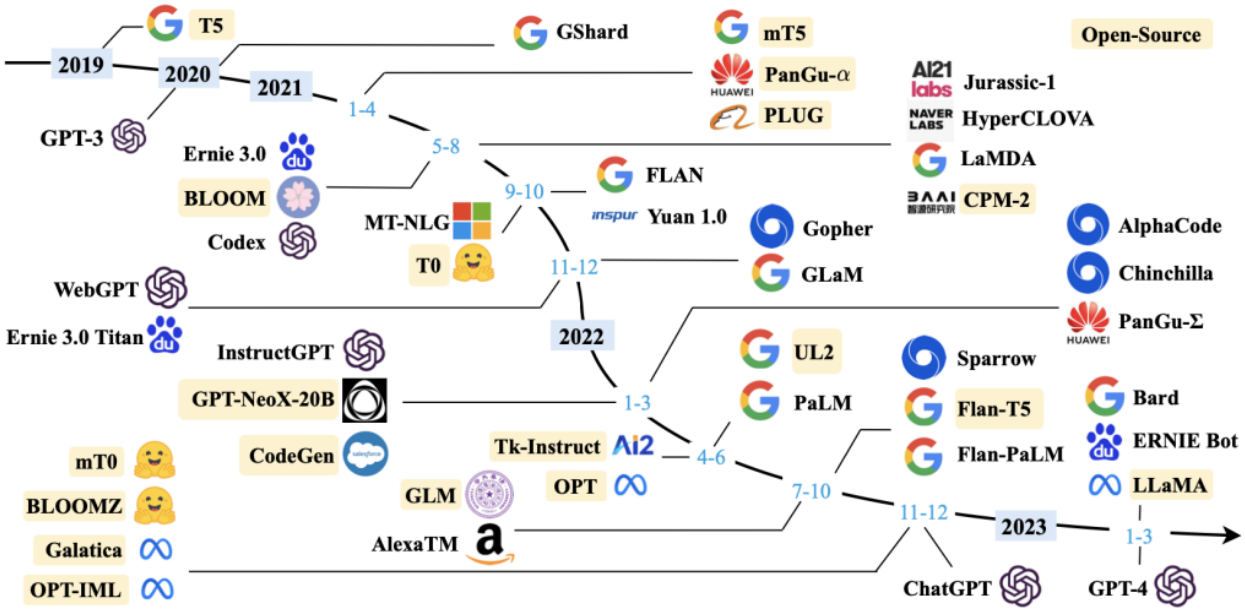
训练ChatGPT的必备资源：语料、模型和代码库完全指南

原创 python 夕小瑶科技说 2023-04-10 12:06 发表于四川



文 | python

前言

近期，ChatGPT成为了全网热议的话题。ChatGPT是一种基于大规模语言模型技术（LLM，large language model）实现的人机对话工具。但是，如果我们想要训练自己的大规模语言模型，有哪些公开的资源可以提供帮助呢？在这个github项目中，人民大学的老师同学们从模型参数（Checkpoints）、语料和代码库三个方面，为大家整理并介绍这些资源。接下来，让我们一起来看看吧。

资源链接：

https://github.com/RUCAIBox/LLMSurvey

论文地址：

https://arxiv.org/pdf/2303.18223.pdf

模型参数

从已经训练好的模型参数做精调、继续训练，无疑可以极大地降低计算成本。那目前有哪些开源的大模型参数，可以供我们选择呢？

第一类是100~1000亿参数的模型。这类模型除了LLaMA（650亿）之外，参数范围都集中在100~200亿之间。具体而言，包括：LLaMA[1], mT5[2], T0[3], GPT-NeoX-20B[4], CodeGen[5], UL2[6], Flan-T5[7], mT0[8], PanGu-α[9]。

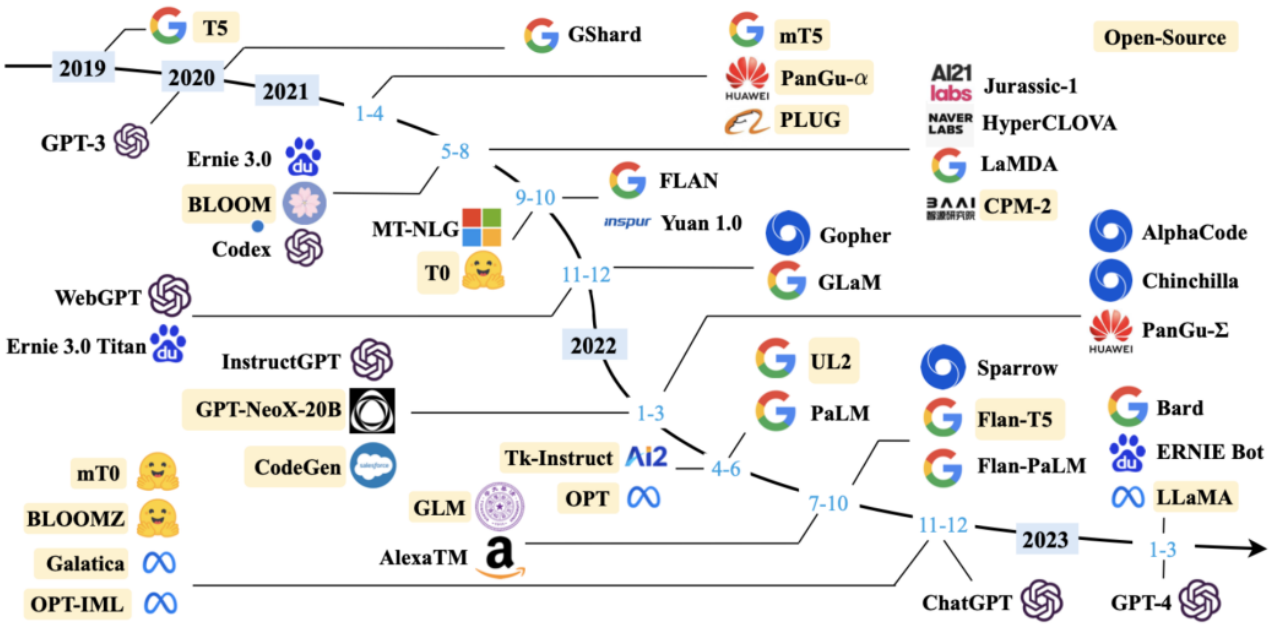
其中，Flan-T5经过instruction tuning的训练；CodeGen专注于代码生成；mT0是个跨语言模型；PanGu-α有大模型版本，并且在中文下游任务上表现较好。

第二类是超过1000亿参数规模的模型。这类模型开源的较少，包括：OPT[10], OPT-IML[11], BLOOM[12], BLOOMZ[13], GLM[14], Galactica[15]。参数规模都在1000亿~2000亿之间。

其中，OPT是专为开源和大模型复现提出的；BLOOM 和 BLOOMZ具有跨语言能力；Galactica, GLM, 和 OPT-IML都是经过instruction tuning的。

这些模型参数大多使用几百到上千块显卡训练得到。比如GPT-NeoX-20B（200亿参数）使用了96个A100-SXM4-40GB GPU，LLaMA（650亿参数）使用了2048块A100-80G GPU学习了21天，OPT（1750亿参数）使用了992 A100-80GB GPU，GLM（1300亿参数）使用了768块DGX-A100-40G GPU训练了60天。

除了这些可供公开下载参数的模型之外，OpenAI还提供在他们的服务器上精调GPT-3模型的服务，可以选择的初始模型参数包括babbage（GPT-3 1B），curie（GPT-3 6.7B）和 davinci（GPT-3 175B）。



上图中，标黄的模型均为开源模型。

语料

训练大规模语言模型，训练语料不可或缺。主要的开源语料可以分成5类：书籍、网页爬取、社交媒体平台、百科、代码。

书籍语料包括：BookCorpus[16] 和 Project Gutenberg[17]，分别包含1.1万和7万本书籍。前者在GPT-2等小模型中使用较多，而MT-NLG 和 LLaMA等大模型均使用了后者作为训练语料。

最常用的网页爬取语料是CommonCrawl[18]。不过该语料虽然很大，但质量较差。大模型大多采用从其中筛选得到的子集用于训练。常用的4个子集包括：C4[19], CC-Stories, CC-News[20], 和 RealNews[21]。CC-Stories的原版现在已不提供下载，一个替代选项是CC-Stories-R[22]。

社交媒体平台语料主要获取自Reddit平台。WebText包含了Reddit平台上的高赞内容，然而现在已经不提供下载，现在可以用OpenWebText[23]替代。此外，PushShift.io[24]提供了一个实时更新的Reddit的全部内容。

百科语料就是维基百科（Wikipedia[25]）的下载数据。该语料被广泛地用于多种大语言模型（GPT-3, LaMDA, LLaMA 等），且提供多种语言版本，可用于支持跨语言模型训练。

代码语料主要来自于GitHub中的项目，或代码问答社区。开源的代码语料有谷歌的BigQuery[26]。大语言模型CodeGen在训练时就使用了BigQuery的一个子集。

除了这些单一内容来源的语料，还有一些语料集。比如 the Pile[27]合并了22个子集，构建了800GB规模的混合语料。而 ROOTS[28]整合了59种语言的语料，包含1.61TB的文本内容。

Corpora	Size	Source	Latest Update Time
BookCorpus [100]	5GB	Books	Dec-2015
Gutenberg [101]	-	Books	Dec-2021
C4 [71]	800GB	CommonCrawl	Apr-2019
CC-stories-R [102]	31GB	CommonCrawl	Sep-2019
CC-NEWS [27]	78GB	CommonCrawl	Feb-2019
REALNEWs [103]	120GB	CommonCrawl	Apr-2019
OpenWebText [104]	38GB	Reddit links	Mar-2023
Pushift.io [105]	-	Reddit links	Mar-2023
Wikipedia [106]	-	Wikipedia	Mar-2023
BigQuery [107]	-	Codes	Mar-2023
the Pile [108]	800GB	Other	Dec-2020
ROOTS [109]	1.6TB	Other	Jun-2022

上图统计了这些常用的开源语料。目前的预训练模型大多采用多个语料资源合并作为训练数据。比如GPT-3使用了5个来源3000亿token（word piece）,包含开源语料CommonCrawl, Wikipedia 和非开源语料（WebText2，Books1, Books2）。

代码库

使用代码库，可以帮助你快速搭建模型结构，而不用一个个矩阵乘法地搭建transformers结构。具体而言，包括以下7个：

1. Transformers[29]是Hugging Face构建的用来快速实现transformers结构的库。同时也提供数据集处理与评价等相关功能。应用广泛，社区活跃。

2. DeepSpeed[30]是一个微软构建的基于PyTorch的库。GPT-Neo，BLOOM等模型均是基于该库开发。DeepSpeed提供了多种分布式优化工具，如ZeRO，gradient checkpointing等。

3. Megatron-LM[31]是NVIDIA构建的一个基于PyTorch的大模型训练工具，并提供一些用于分布式计算的工具如模型与数据并行、混合精度训练，FlashAttention与gradient checkpointing等。

4. JAX[32]是Google Brain构建的一个工具，支持GPU与TPU，并且提供了即时编译加速与自动batching等功能。

5. Colossal-AI[33]是HPC-AITech开发的一个大模型训练工具，支持并行化训练。最近有一个基于LLaMA训练的对话应用ColossalChat就是基于该工具构建的。

6. BMTrain[34] 是 OpenBMB开发的一个大模型训练工具，强调代码简化，低资源与高可用性。在其ModelCenter中，已经构建好如Flan-T5 与 GLM等模型结构可供直接使用。

7. FastMoE[35] 是一个基于pytorch的用于搭建混合专家模型的工具，并支持训练时数据与模型并行。

结束语

通过使用以上提到的模型参数、语料与代码，我们可以极大地方便自己实现大规模语言模型，并搭建出自己的对话工具。但是，尽管数据资源相对容易获取，计算资源却十分稀缺。想要获得足够的显卡资源以训练/调整大规模模型，仍然是一件非常困难的事情。因此，私有化

ChatGPT的道路任重而道远。在计算资源相对匮乏的情况下，我们更是要利用好手头的模型参数、语料与代码等资源，以有限的计算量取得最好的表现。

卖萌屋作者：**python**

北大毕业的NLP博士。日常写点论文，码点知乎，刷点leetcode。主要关注问答、对话、信息抽取、预训练、智能法律等方向。力扣国服第一python选手（经常掉下来）。知乎 ID 是 Erutan Lai， leetcode/力扣 ID 是 pku_erutan，欢迎没事常来逛逛。

作品推荐

1. 恕我直言，你的实验结论可能严重依赖随机数种子！
2. AllenAI 发布万能问答系统 MACAW！各类题型样样精通，性能大幅超越 GPT-3！
3. 吐血整理：论文写作中注意这些细节，能显著提升成稿质量
4. 恕我直言，你的模型可能并没看懂 prompt 在说啥

后台回复关键词【入群】

加入卖萌屋**NLP**、**CV**、搜推广与求职讨论群

夕小瑶科技说

更快的AI前沿，更深的行业洞见。一线作者均来自清北、国外顶级AI实验室和互联网大...

527篇原创内容

公众号

参考文献

[1]https://github.com/facebookresearch/llama

[2]https://huggingface.co/google/mt5-xxl/tree/main

[3]https://huggingface.co/bigscience/T0

[4]https://huggingface.co/EleutherAI/gpt-neox-20b/tree/main

[5]https://huggingface.co/Salesforce/codegen-16B-nl

[6]https://github.com/google-research/google-research/tree/master/ul2

[7]https://github.com/google-research/t5x/blob/main/docs/models.md#flan-t5-checkpoints

[8]https://github.com/bigscience-workshop/xmtf

[9]https://openi.pcl.ac.cn/PCL-Platform.Intelligence/PanGu-Alpha

[10]https://github.com/facebookresearch/metaseq/tree/main/projects/OPT

[11]https://huggingface.co/facebook/opt-iml-30b

[12]https://huggingface.co/bigscience/bloom

[13]https://github.com/bigscience-workshop/xmtf

[14]https://github.com/THUDM/GLM-130B

[15]https://huggingface.co/facebook/galactica-120b

[16]https://huggingface.co/datasets/bookcorpus

[17]https://www.gutenberg.org/

[18]https://commoncrawl.org/

[19]https://www.tensorflow.org/datasets/catalog/c4

[20]https://huggingface.co/datasets/cc_news

[21]https://github.com/rowanz/grover/tree/master/realnews

[22]https://huggingface.co/datasets/spacemanidol/cc-stories

[23]https://skylion007.github.io/OpenWebTextCorpus/

[24]https://files.pushshift.io/reddit/

[25]https://dumps.wikimedia.org/

[26]https://cloud.google.com/bigquery/public-data?hl=zh-cn

[27]https://pile.eleuther.ai/

[28]https://arxiv.org/abs/2303.03915

[29]https://huggingface.co/

[30]https://github.com/microsoft/DeepSpeed

[31]https://github.com/NVIDIA/Megatron-LM

[32]https://github.com/google/jax

[33]https://github.com/hpcaitech/ColossalAI

[34]https://github.com/OpenBMB/BMTrain

[35]https://github.com/laekov/fastmoe

文章已于2023-04-14修改

喜欢此内容的人还喜欢

陈丹琦团队新作：单卡A100可训300亿参数模型啦！

夕小瑶科技说

GPT-4知道它是不是“胡说八道”吗？一篇关于大模型“自知之明”的研究

夕小瑶科技说

GPT-4使用效果不好？美国奥本大学提出Prompt分类法，另辟蹊径构建Prompt设计指南

夕小瑶科技说