重磅！Meta发布LLaMA2，最高700亿参数~完全免费可商用！

数据学习者 2023-07-19 18:35

本文来自DataLearner官方博客：重磅！Meta发布LLaMA2，最高700亿参数，在2万亿tokens上训练，各项得分远超第一代LLaMA~完全免费可商用！ | 数据学习者官方网站(Datalearner)

LLaMA是由Meta开源的一个大语言模型，是最近几个月一系列开源模型的基础模型。包括著名的vicuna系列、LongChat系列等都是基于该模型微调得到。可以说，LLaMA的开源促进了大模型在开源界繁荣发展。而刚刚，微软官方宣布Azure上架LLaMA2模型！这意味着LLaMA2正式发布！

LLaMA2比LLaMA1多40%的训练数据，性能更加强大，但是依然完全免费可商用！

LLaMA2简介和参数

根据官方的介绍，Meta和Microsoft准备将LLaMA2引入到Azure公有云以及Windows本地上。目前已经在AzureAI上开放了LLaMA2系列的6个模型供大家使用。

LLaMA2模型的**参数范围从70亿到700亿不等**，在**超过2万亿tokens数据集**上训练。官方对齐微调的结果称为LLaMA2-Chat系列，专门针对场景优化。

LLaMA2-Chat模型在微软测试的大多数基准测试中胜过开源聊天模型，并在人工评估中在实用性和安全性方面与一些流行的闭源模型如ChatGPT和PaLM相当。

LLaMA2具体的模型信息如下：

	参数	上下文长度	GQA	训练tokens	学习率
LLaMA2-7B	7B	4k	✗	2万亿	3.0 x 10-4
LLaMA2-13B	13B	4k	✗	2万亿	3.0 x 10-4
LLaMA2-700B	70B	4k	✔	2万亿	1.5 x 10-4

LLaMA2的训练信息

所有模型都使用全局批量大小为4M tokens进行训练。更大的700亿参数模型使用Grouped-Query Attention（GQA）来提高推理可扩展性。

LLaMA2的**训练时间为2023年1月至2023年7月**。且是一个**纯文本模型**，仅接受文本输入和文本的输出。

预训练过程中，Meta估计使用了总计**33万GPU小时**的计算，硬件类型为A100-80GB（功耗为350-400W）。

LLaMA2模型架构

LLaMA2是一种优化的自回归语言变换器。微调版本使用监督微调（SFT）和人工反馈强化学习（RLHF）来对齐人类对实用性和安全性的偏好。

LLaMA2训练数据

LLaMA2是在来自公开可用来源的2万亿tokens数据上进行的预训练。微调数据包括公开可用的指令数据集，以及**超过100万个新的人工注释示例**。预训练和微调数据集均不包含Meta用户数据。

预训练数据的截止日期为2022年9月，但某些微调数据更近，最新的可达到2023年7月。

LLaMA2的评估结果

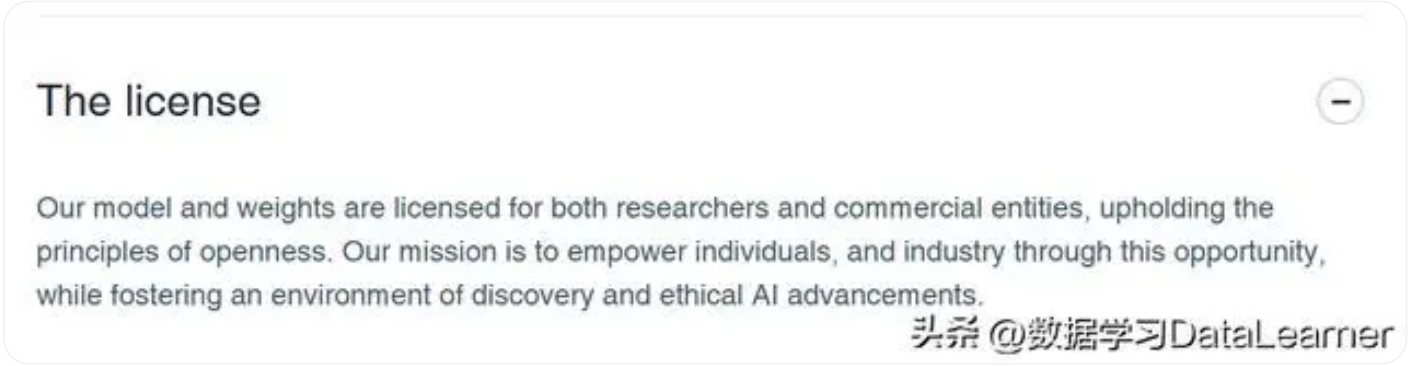
官方给出了详细信息：

Model	Size	Code	Commonsense Reasoning	World Knowledge	Reading Comprehension	Math	MMLU	BBH	AGI Eval
LLaMA 1	7B	14.1	60.8	46.2	58.5	6.95	35.1	30.3	23.9
LLaMA 1	13B	18.9	66.1	52.6	62.3	10.9	46.9	37.0	33.9
LLaMA 1	33B	26.0	70.0	58.4	67.6	21.4	57.8	39.8	41.7
LLaMA 1	65B	30.7	70.7	60.5	68.6	30.8	63.4	43.5	47.6
LLaMA 2	7B	16.8	63.9	48.9	61.3	14.6	45.3	32.6	29.3
LLaMA 2	13B	24.5	66.9	55.4	65.8	28.7	54.8	39.4	39.1
LLaMA 2	70B	**37.5**	**71.9**	63.6	**69.4**	**35.2**	**68.9**	**51.2**	**54.2**

可以看到，LLaMA2在各方面都超过第一代很多。尤其是数学、文本理解等方面。代码方面的得分也不错。

LLaMA2的开源协议

官方表示LLaMA2完全开源可商用，都没写啥具体协议：



不过，使用依然需要填写表单申请，需要审核通过后才可以下载。

LLaMA2的实测样例

官方给出了实际测试结果：
输入：

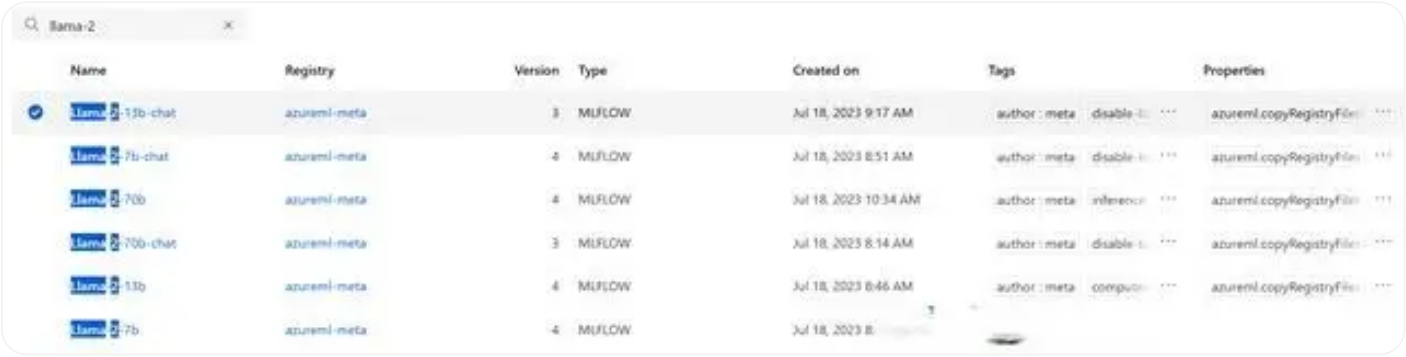
```
{  "input_data": {    "input_string": [      {        "role": "us
```

输出结果：

```
{  "output": "There are many reasons why the Eiffel Tower is one
```

LLaMA2的总结和使用

目前LLaMA2模型首先上架了Azure的模型服务。包括如下几个：



关键微软发布的信息，LLaMA2支持聊天应用也支持微调部署~未来也会在Windows本地引入该模型。只是，微软与Meta走得这么近，OpenAI咋办呢~~

LLaMA2的开源地址：
https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md
LLaMA2的下载地址：https://ai.meta.com/resources/models-and-libraries/llama-downloads/
LLaMA2的官方博客地址：https://ai.meta.com/resources/models-and-libraries/llama/