

## 世界首款真开源类ChatGPT大模型Dolly 2.0，可随意修改商用

机器之心 2023-04-13 13:07 发表于北京

机器之心报道

编辑：泽南、蛋酱

我们鼓励员工手搓了一个数据集，训练 LLM 还把它开源。

众所周知，在 ChatGPT 的问题上 OpenAI 并不 Open，从 Meta 那里开源的羊驼系列模型也因为数据集等问题「仅限于学术研究类应用」，在人们还在因为寻找绕过限制方法的时候，主打 100% 开源的大模型来了。

4 月 12 日，Databricks 发布了 Dolly 2.0，这是两周前发布的类 ChatGPT 人类交互性（指令遵循）大语言模型（LLM）的又一个新版本。

Databricks 表示，Dolly 2.0 是业内第一个开源、遵循指令的 LLM，它在透明且免费提供的数据集上进行了微调，该数据集也是开源的，可用于商业目的。这意味着 Dolly 2.0 可用于构建商业应用程序，无需支付 API 访问费用或与第三方共享数据。



- 项目链接：<https://huggingface.co/databricks/dolly-v2-12b>
- 数据集：<https://github.com/databrickslabs/dolly/tree/master/data>

根据 Databricks 首席执行官 Ali Ghodsi 的说法，虽然已有其他大模型可以用于商业目的，但「它们不会像 Dolly 2.0 那样与你交谈。」而且基于 Dolly 2.0 模型，用户可以修改和改进训练数据，因为它是在开源许可下免费提供的。所以你可以制作你自己的 Dolly 版本。

Databricks 还发布了 Dolly 2.0 在其上进行微调的数据集，称为 databricks-dolly-15k。这是由数千名 Databricks 员工生成的超过 1.5 万条记录的语料库，Databricks 称这是「第一个开源的、人工生成的指令语料库，专门设计用于让大型语言能够展示出 ChatGPT 的神奇交互性。」

### Dolly 2.0 是怎么诞生的

在过去的两个月里, 业界、学界纷纷追赶 OpenAI 提出了一波遵循指令的类 ChatGPT 大模型, 这些版本被许多定义视为开源 (或提供某种程度的开放性或有限访问)。其中 Meta 的 LLaMA 最受人关注, 它引发了大量进一步改进的模型, 如 Alpaca、Koala、Vicuna 以及 Databricks 的 Dolly 1.0。

但另一方面, 许多这些「开放」模型都处于「工业限制」之下, 因为它们接受了旨在限制商业用途的条款的数据集的训练 —— 例如来自 StanfordAlpaca 项目的 5.2 万个问答数据集, 是根据 OpenAI 的 ChatGPT 的输出进行训练的。而 OpenAI 的使用条款包括一条规则, 即你不能使用 OpenAI 的服务反过来与其竞争。

Databricks 思考了解决这个问题的方法: 新提出的 Dolly 2.0 是一个 120 亿参数的语言模型, 它基于开源 EleutherAI pythia 模型系列, 专门针对小型开源指令记录语料库进行了微调 (databricks-dolly-15k), 该数据集由 Databricks 员工生成, 许可条款允许出于任何目的使用、修改和扩展, 包括学术或商业应用。

到目前为止, 在 ChatGPT 的输出上训练的模型一直处于合法的灰色地带。「整个社区一直在小心翼翼地解决这个问题, 每个人都在发布这些模型, 但没有一个可以用于商业用途,」 Ghodsi 表示。「这就是我们非常兴奋的原因。」

「其他人都想做得更大, 但我们实际上对更小的东西感兴趣,」 Ghodsi 在谈到 Dolly 的微缩规模时说。「其次, 我们翻阅了所有的答案, 它是高质量的。」

Ghodsi 表示, 他相信 Dolly 2.0 将启动「雪球」效应, 让人工智能领域的其他人加入并提出其他替代方案。他解释说, 对商业用途的限制是一个需要克服的大障碍:「我们现在很兴奋, 因为我们终于找到了一个绕过它的方法。我保证你会看到人们将这 15000 个问题应用于现有的每一个模型, 他们会看到这些模型中有多少突然变得有点神奇, 你可以与它们互动。」

### 手搓数据集

要下载 Dolly 2.0 模型的权重, 只需访问 Databricks Hugging Face 页面, 并访问 databricks-labs 的 Dolly repo, 下载 databricks-dolly-15k 数据集。

「databricks-dolly-15k」数据集包含 15000 个高质量的人类生成的 prompt / 回复对, 由 5000 多名 Databricks 员工在 2023 年 3 月和 4 月期间撰写, 专门设计用于指令调优大型语言模型。这些训练记录自然、富有表现力, 旨在代表广泛的行为, 从头脑风暴、内容生成到信息提取和总结。

根据该数据集的许可条款 (Creative Commons Attribution-ShareAlike 3.0 Unported License), 任何人都可因任何目的使用、修改或扩展这个数据集, 包括商业应用。

目前, 这一数据集是首个开源的、由人类生成的指令数据集。

为什么要创建这样一个数据集? 团队也在博客中解释了原因。

创建 Dolly 1.0 或任何遵循 LLM 的指令的一个关键步骤是, 在指令和回复对的数据集上训练模型。Dolly 1.0 的训练费用为 30 美元, 使用的是斯坦福大学 Alpaca 团队用 OpenAI API 创建的数据集。

在 Dolly 1.0 发布之后, 就有很多人要求试用, 此外还有一部分用户希望在商业上使用这个模型。

但是训练数据集包含 ChatGPT 的输出，正如斯坦福大学团队所指出的，服务条款试图阻止任何人创建一个与 OpenAI 竞争模型。

此前，所有的知名指令遵循模型（Alpaca、Koala、GPT4All、Vicuna）都受到这种限制：禁止商业使用。为了解决这个难题，Dolly 团队开始寻找方法来创建一个没有商业用途限制的新数据集。

具体而言，团队从 OpenAI 公布的研究论文中得知，最初的 InstructGPT 模型是在一个由 13000 个指令遵循行为演示组成的数据集上训练出来的。受此启发，他们开始研究是否可以在 Databricks 员工的带领下取得类似的结果。

结果发现，生成 13000 个问题和答案比想象中更难。因为每个答案都必须是原创的，不能从 ChatGPT 或网络上的任何地方复制，否则会「污染」数据集。但 Databricks 有超过 5000 名员工，他们对 LLM 非常感兴趣。因此，团队进行了一次众包实验，创造出了比 40 位标注者为 OpenAI 创造的更高质量的数据集。

当然，这项工作耗时耗力，为了激励大家，团队设立了一个竞赛，前 20 名的标注者将获得惊喜大奖。同时，他们也列出了 7 项非常具体的任务：

- 公开问答：例如「为什么人们喜欢喜剧电影？」或「法国的首都是什么？」在某些情况下，没有一个正确的答案，而在其他情况下，需要借助于整个世界的知识；
- 封闭式问答：这些问题只用一段参考文献中的信息就可以回答。例如，给定维基百科中关于原子的一段，人们可能会问：「原子核中质子和中子的比例是多少？」；
- 从维基百科中提取信息：在这里，标注者会从维基百科上复制一个段落，并从该段落中提取实体或其他事实信息，如重量或测量；
- 总结维基百科上的信息：对于这一点，注释者从维基百科上提供了一段话，并被要求将其提炼为一个简短的摘要；
- 集思广益：这项任务要求进行开放式的构思，并列出相关的可能选项。例如「这个周末我可以和我的朋友做哪些有趣的活动？」；
- 分类：在这项任务中，标注者被要求对类别成员进行判断（例如，列表中的项目是动物、矿物还是蔬菜），或者判断一段短文的属性，例如电影评论的情绪；
- 创意写作：这项任务将包括写一首诗或一封情书等内容。

以下是一些示例：

**Example of Open QA in databricks-dolly-15k**

**How do I build a campfire?**

Safety should always come first when starting a campfire. Ensure you have water or another way to extinguish your fire readily available. Use a campfire ring, pit, or clear a large area around your campfire location. Also, make sure campfires are permitted in your area.

With safety out of the way, all a fire needs are heat, fuel, and oxygen. Any combustible material will start a fire, but good campfire fuel sources include tinder, kindling, and dry logs.

Use stuff like dry pine needles, dry leaves, grasses, and wood shavings for tinder. You could also use newspaper (or any paper), dryer lint, or cardboard. For kindling, use smaller pieces of wood, such as twigs and sticks no thicker than 1 inch in diameter (about the thickness of your thumb).

Dry or well-seasoned logs will ensure a good burn. Stack the wood in a way that guarantees oxygen can easily reach your flame. Many prefer a teepee or log cabin stacking design. Now, use a heat source to create a flame or spark and apply this to your tinder. Once the tinder combusts, it should light the kindling, and the kindling should eventually ignite your logs. If all goes well, you should now be enjoying your campfire.

Example of Brainstorming in databricks-dolly-15k

Instead of making a peanut butter and jelly sandwich, what else could I combine peanut butter with in a sandwich? Give five ideas.

Instead of jelly, try one of the following with peanut butter in a sandwich:

1. Honey
2. Raisins
3. Fruit preserves
4. Bananas
5. Sliced apples

最开始，团队对于是否能达到 10000 个结果持怀疑态度。但通过每晚的排行榜游戏，一周内就成功地突破了 15000 个结果。

随后，出于对「占用员工生产力」的担心，团队关闭了比赛（这很合理）。

商业化的可行性

在数据集火速创建完成之后，团队开始考虑商业应用的问题了。

他们想制作一个可在商业上使用的开源模型。尽管 databricks-dolly-15k 比 Alpaca（训练 Dolly 1.0 的数据集）小得多，但基于 EleutherAI pythia-12b 的 Dolly 2.0 模型却表现出高质量的指令遵循行为。

事后看来，这并不令人惊讶。毕竟最近几个月发布的许多指令调优数据集包含合成数据，这些数据往往包含幻觉和事实错误。

另一方面，databricks-dolly-15k 是由专业人士生成的，质量很高，而且包含大多数任务的长篇答案。

以下是 Dolly 2.0 用于总结和内容生成的一些例子：

Dolly 2.0 summarizes Databricks documentation

Instruction	Response
<p>What is the default configuration for new DBSQL warehouses?</p> <p>Input:</p> <p>Databricks SQL Serverless supports serverless compute. Admins can create serverless SQL warehouses (formerly SQL endpoints) that enable instant compute and are managed by Databricks. Serverless SQL warehouses use compute clusters in your Databricks account. Use them with Databricks SQL queries just like you normally would with the original customer-hosted SQL warehouses, which are now called classic SQL warehouses. Databricks changed the name from SQL endpoint to SQL warehouse because, in the industry, endpoint refers to either a remote computing device that communicates with a network that it's connected to, or an entry point to a cloud service. A data warehouse is a data management system that stores current and historical data from multiple sources in a business friendly manner for easier insights and reporting. SQL warehouse accurately describes the full capabilities of this compute resource. If serverless SQL warehouses are enabled for your account, note the following: New SQL warehouses are serverless by default when you create them from the UI. New SQL warehouses are not serverless by default when you create them using the API, which requires that you explicitly specify serverless. You can also create new pro or classic SQL warehouses using either method. You can upgrade a pro or classic SQL warehouse to a serverless SQL warehouse or a classic SQL warehouse to a pro SQL warehouse. You can also downgrade from serverless to pro or classic. This feature only affects Databricks SQL. It does not affect how Databricks Runtime clusters work with notebooks and jobs in the Data Science &amp; Engineering or Databricks Machine Learning workspace environments. Databricks Runtime clusters always run in the classic data plane in your AWS account. See Serverless quotas. If your account needs updated terms of use, workspace admins are prompted in the Databricks SQL UI. If your workspace has an AWS instance profile, you might need to update the trust relationship to support serverless compute, depending on how and when it was created.</p>	<p>By default, when you create a new DBSQL warehouse using the UI, it will be a serverless SQL warehouse. However, when creating the same using the API, you will have to explicitly specify that it is a pro SQL warehouse.</p>



Dolly 2.0 summarizes a customer support ticket

Instruction	Response
<p>Give me a list of the main complaints in this customer support ticket. Do not write a reply.</p> <p>Input:</p> <p>I am writing to express my deep disappointment and frustration with the iPhone 14 Pro Max that I recently purchased. As a long-time Apple user and loyal customer, I was excited to upgrade to the latest and greatest iPhone model, but unfortunately, my experience with this device has been nothing short of a nightmare.</p> <p>Firstly, I would like to address the issue of battery life on this device. I was under the impression that Apple had made significant improvements to their battery technology, but unfortunately, this has not been my experience. Despite using the phone conservatively, I find that I have to charge it at least twice a day just to ensure it doesn't die on me when I need it the most. This is extremely inconvenient and frustrating, especially when I have to carry around a bulky power bank or constantly hunt for charging outlets.</p> <p>Furthermore, I am extremely disappointed with the camera quality on this device. Despite Apple's claims of improved camera technology, I have found that the photos I take on this phone are often blurry or grainy, and the colors are not as vibrant as I would like. This is especially disappointing considering the high price point of the iPhone 14 Pro Max, which is marketed as a premium smartphone with a top-of-the-line camera.</p> <p>Overall, I feel as though I have been let down by Apple and their latest iPhone offering. As a loyal customer who has invested a significant amount of money into their products over the years, I expect better from a company that prides itself on innovation and customer satisfaction. I urge Apple to take these concerns seriously and make necessary improvements to the iPhone 14 Pro Max and future models.</p> <p>Thank you for your attention to this matter.</p>	<p>The customer is unhappy about the battery life and the camera quality of the iPhone 14 Pro Max.</p>

Dolly 2.0 generates content for a tweet

Instruction	Response
<p>Write me a tweet about the launch of Dolly 2.0, our new LLM.</p>	<p>Dolly 2.0 is here 🥳</p> <p>We've upgraded our LLM, making it more efficient, more powerful, and more accessible to a broader audience.</p> <p>Learn more 📄</p> <p><a href="https://t.co/KxzT3HXCPb">https://t.co/KxzT3HXCPb</a></p>

Dolly 团队表示，根据最初的客户反馈，像这样的能力可在整个企业中进行广泛的应用。因为很多企业希望拥有自己的模型，以此为自己的特定领域应用创建更高质量的模型，而不是将自己的敏感数据交给第三方。

Dolly 2 的开源为构建更好的大模型生态开了一个好头。开放源代码的数据集和模型鼓励评论、研究和创新，有助于确保每个人都从人工智能技术的进步中受益。Dolly 团队期望新模型和开源数据集将作为众多后续工作的种子，帮助引导出更强大的语言模型。

参考内容：

<https://venturebeat.com/ai/databricks-releases-dolly-2-0-the-first-open-instruction-following-llm-for-commercial-use/>

<https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm>



机器之心

Talent AI

专注于人工智能领域人才服务

**TalentAI**是机器之心AI人才招聘服务，推荐人工智能领域技术类岗位，如果你正在找工作或看机会，欢迎关注。

### 近期在招企业

**BAAI**  
智源研究院  
上海人工智能实验室  
Shanghai Artificial Intelligence Laboratory**達摩院**  
ALIBABA DAMO ACADEMY**JDT**  
京东科技**小红书**  
同花顺  
科大讯飞  
iFLYTEK  
蚂蚁集团  
ANT GROUP  
miHoYo  
TECH STATUS SAVE THE WORLD**MEGVII** 旷视

### 岗位方向包括

大模型 | AIGC | 多模态 | 数据科学  
强化学习 | 计算机视觉 | NLP | 运筹优化  
智能驾驶 | 高精定位 | 多传感器融合  
药物研发 | 蛋白质设计 | 量化研究.....

校招全职、社招全职与实习生  
关注TalentAI公众号  
查看岗位详情



转载请联系本公众号获得授权

投稿或寻求报道：content@jiqizhixin.com

喜欢此内容的人还喜欢

开发者笑疯了！ LLaMa惊天泄露引爆ChatGPT平替狂潮，开源LLM领域变天  
新智元



对标GPT-4！非盈利机构LAION重磅开源OpenFlamingo：基于LLaMA的多模态大模型  
机器学习算法工程师



Llama-X开源！呼吁每一位NLPer参与推动LLaMA成为最先进的LLM  
PaperWeekly

