GPT-3诞生，Finetune也不再必要了！NLP领域又一核弹！

原创 rumor酱 夕小瑶的卖萌屋 5月30日

来自专辑

卖萌屋@自然语言处理

一只小狐狸带你解锁**炼丹术&NLP秘籍**

2018年10月推出的BERT一直有着划NLP时代的意义，然而还有一个让人不能忽略的全程陪跑模型——OpenAI GPT（Generative Pre-Training）在以它的方式坚持着，向更通用的终极目标进发。

最初的GPT只是一个12层单向的Transformer，通过预训练+精调的方式进行训练，BERT一出来就被比下去了。之后2019年初的GPT-2提出了meta-learning，把所有NLP任务的输入输出进行了整合，全部用文字来表示，比如对于翻译任务的输入是“英翻法：This is life”，输出是“C'est la vie”。直接把任务要做什么以自然语言的形式放到了输入中。通过这种方式进行了大规模的训练，并用了15亿参数的大模型，一举成为当时最强的生成模型。

遗憾的是，GPT-2在NLU领域仍并不如BERT，且随着19年其他大模型的推出占据了下风，年初微软推出的Turing-NLG已经到达了170亿参数，而GPT-2只有15亿。这些模型的尺寸已经远远超出了大部分公司的预算和调参侠们的想象。。。已经到极限了吗？

不，“极限挑战”才刚刚开始，OpenAI在十几个小时前悄然放出了GPT第三季——《**Language Models are Few-Shot Learners**》。

Language Models are Few-Shot Learners

Tom B. Brown* **Benjamin Mann*** **Nick Ryder*** **Melanie Subbiah***

Jared Kaplan† **Prafulla Dhariwal** **Arvind Neelakantan** **Pranav Shyam** **Girish Sastry**

Amanda Askell **Sandhini Agarwal** **Ariel Herbert-Voss** **Gretchen Krueger** **Tom Henighan**

Rewon Child **Aditya Ramesh** **Daniel M. Ziegler** **Jeffrey Wu** **Clemens Winter**

Christopher Hesse **Mark Chen** **Eric Sigler** **Mateusz Litwin** **Scott Gray**

Benjamin Chess **Jack Clark** **Christopher Berner**

Sam McCandlish **Alec Radford** **Ilya Sutskever** **Dario Amodei**

OpenAI

paper链接：https://arxiv.org/abs/2005.14165

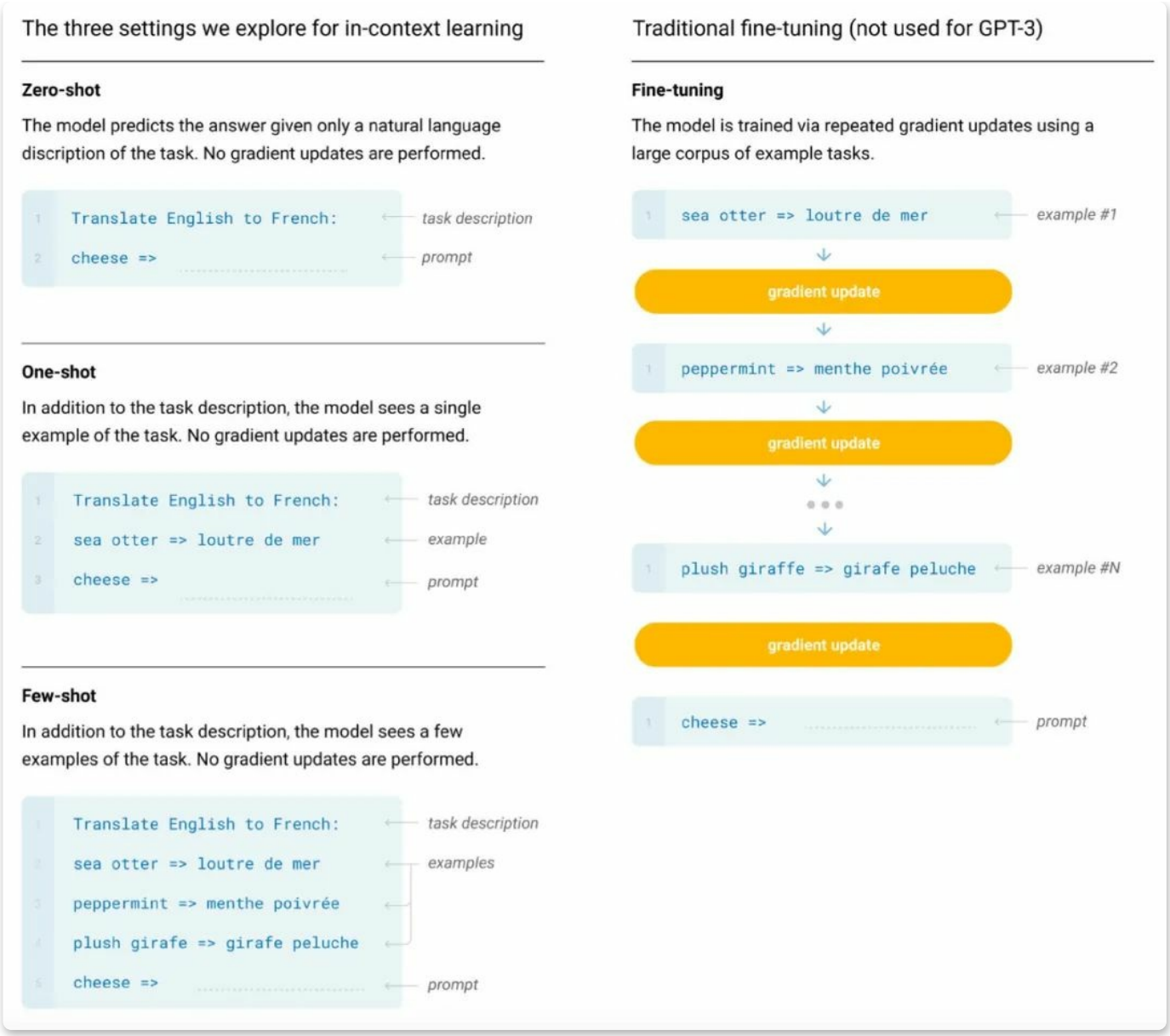
github链接：https://github.com/openai/gpt-3

GPT-3依旧延续自己的单向语言模型训练方式，只不过这次把模型尺寸增大到了**1750亿，**并且使用*45TB*数据进行训练。同时，GPT-3主要聚焦于更通用的NLP模型，解决当前BERT类模型的两个缺点：

1. **对领域内有标签数据的过分依赖：**虽然有了预训练+精调的两段式框架，但还是少不了一定量的领域标注数据，否则很难取得不错的效果，而标注数据的成本又是很高的。
2. **对于领域数据分布的过拟合：**在精调阶段，因为领域数据有限，模型只能拟合训练数据分布，如果数据较少的话就可能造成过拟合，致使模型的泛化能力下降，更加无法应用到其他领域。

因此GPT-3的主要目标是**用更少的领域数据、且不经过精调步骤去解决问题**。

为了达到上述目的，作者们用预训练好的GPT-3探索了不同输入形式下的推理效果：

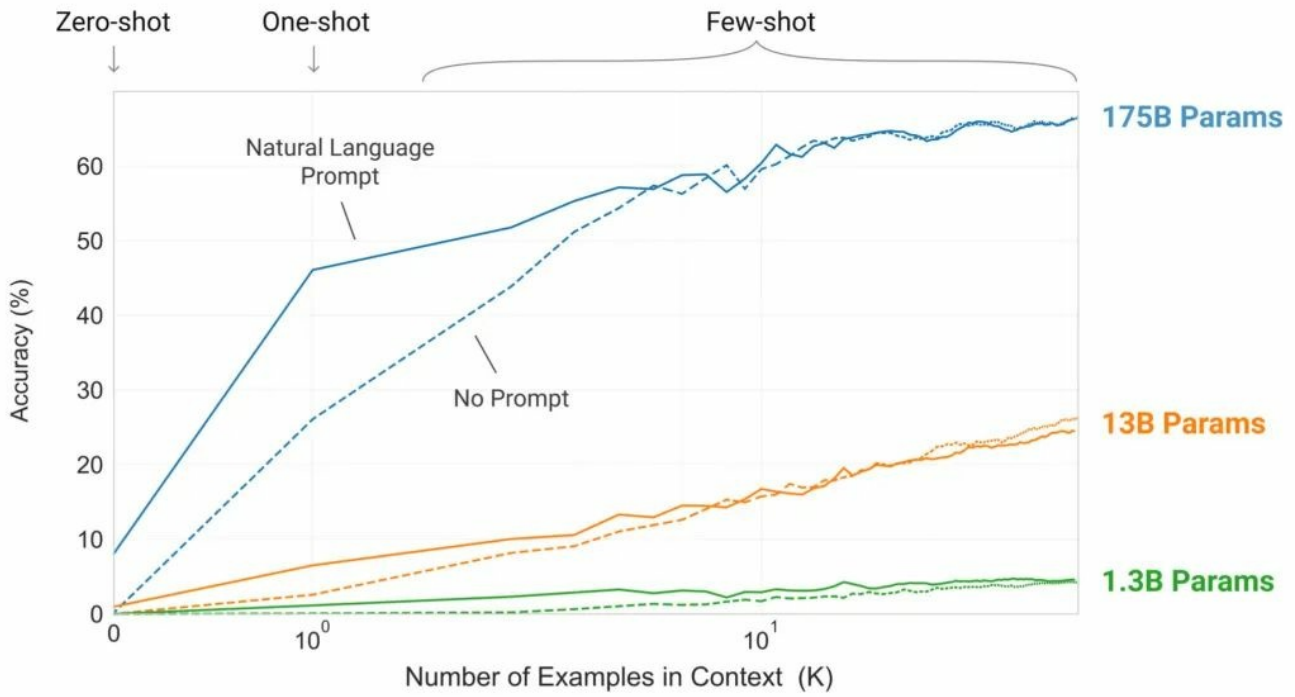


这里的Zero-shot、One-shot、Few-shot都是完全不需要精调的，因为GPT-3是单向transformer，在预测新的token时会对之前的examples进行编码。

作者们训练了以下几种尺寸的模型进行对比：

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or “GPT-3”	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

实验证明Few-shot下GPT-3有很好的表现：



最重要的是，**GPT-3在Few-shot设定下，在部分NLU任务上超越了当前Fine-tuning的SOTA**。该论文长达72页（Google T5是53页），第10页之后都是长长的实验结果与分析。需要的同学们可以在公众号后台回复➡0529⬅获取下载链接!

显然，GPT-3的模型参数、训练数据和工作量都是惊人的，论文署名多达31个作者，所有实验做下来肯定也耗费了不少时间。虽然一直都存在对于大模型的质疑声音，但我们确确实实从T5、GPT-3这样的模型上看到了NLP领域的进步，众多业务也开始受益于离线或者线上的BERT。**事物的发展都是由量变到质变的过程，感谢科研工作者们的不懈努力和大厂们的巨额投入，奥利给。**

本文收录于原创专辑：《卖萌屋@自然语言处理》

重磅惊喜：卖萌屋小可爱们苦心经营的 自然语言处理讨论群 成立三群啦！扫描下方二维码，后台回复「 入群」即可加入。众多顶会审稿人、大厂研究员、知乎大V以及美丽小姐姐（划掉 ♀）等你来撩噢~（手慢无

夕小瑶的卖萌屋

关注&星标小夕，带你解锁AI秘籍

订阅号主页下方「撩一下」有惊喜

可能喜欢

- ACL2020 | 线上搜索结果大幅提升！亚马逊提出对抗式query-doc相关性模型
- 别再蒸馏3层BERT了！变矮又能变瘦的DynaBERT了解一下
- All in Linux：一个算法工程师的IDE断奶之路
- 卖萌屋算法岗面试手册上线！通往面试自由之路
- 巨省显存的重计算技巧在TF、Keras中的正确打开方式
- 硬核推导Google AdaFactor：一个省显存的宝藏优化器

文章已于修改