训练 GPT-2

卢艳峰

August 18, 2023

前言

- 移交之前工作的代码，向师弟师妹介绍项目结构和代码接口，帮助搭建开发环境。
- 训练 **GPT-2** 模型。
- 阅读 **InstructGPT**（**ChatGPT** 的姊妹模型）原论文。

GPT-2 训练数据

- 维基百科中文数据集包含 **1,043,224** 个词条。
- 纯文本大小为 **1.2G**。

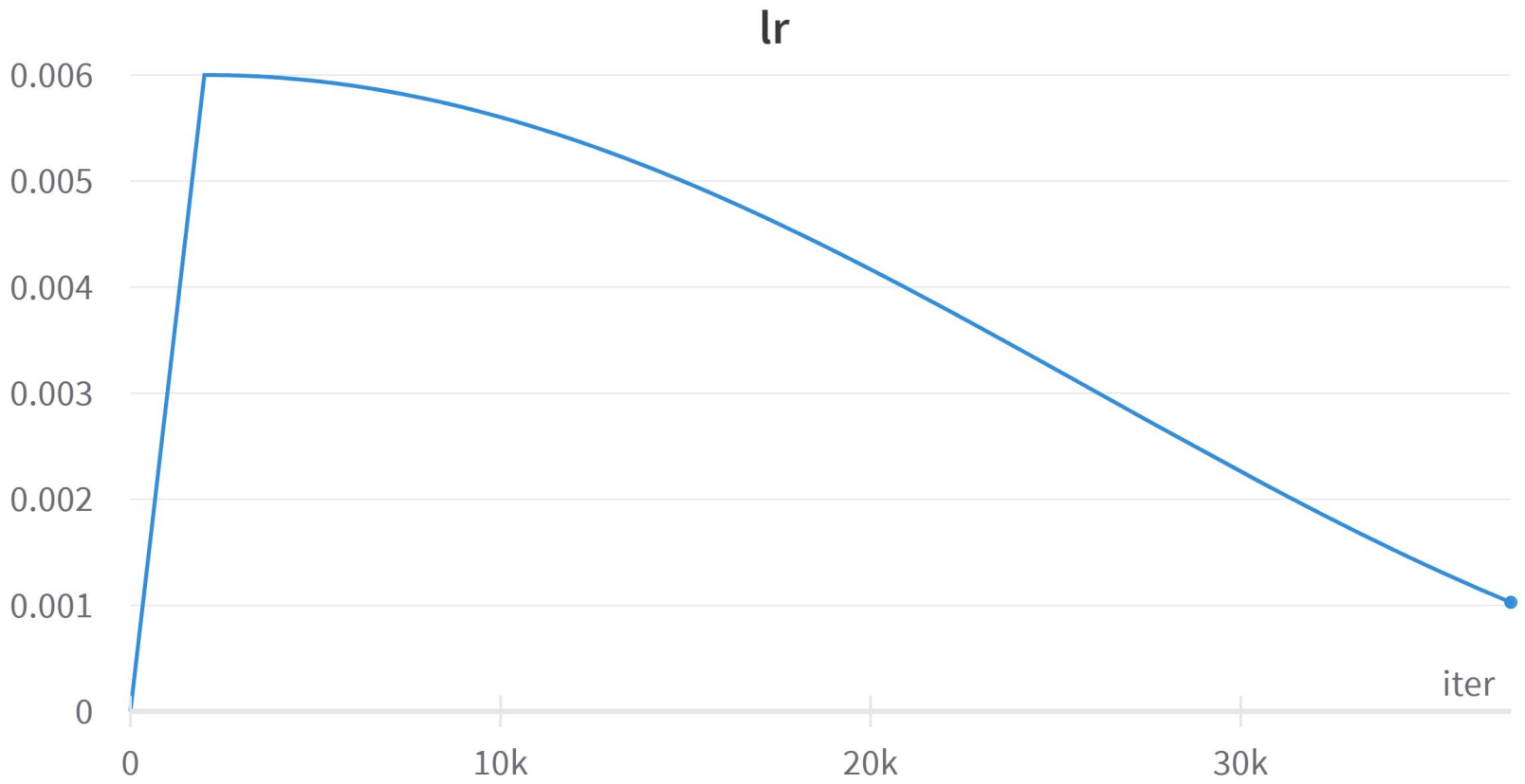
训练 Tokenizer

- 选择 **Byte-Pair Encoding (BPE)** 算法作为分词方法。
- **min_frequency** 设置为 **10**。
- 需要 **60G** 内存。

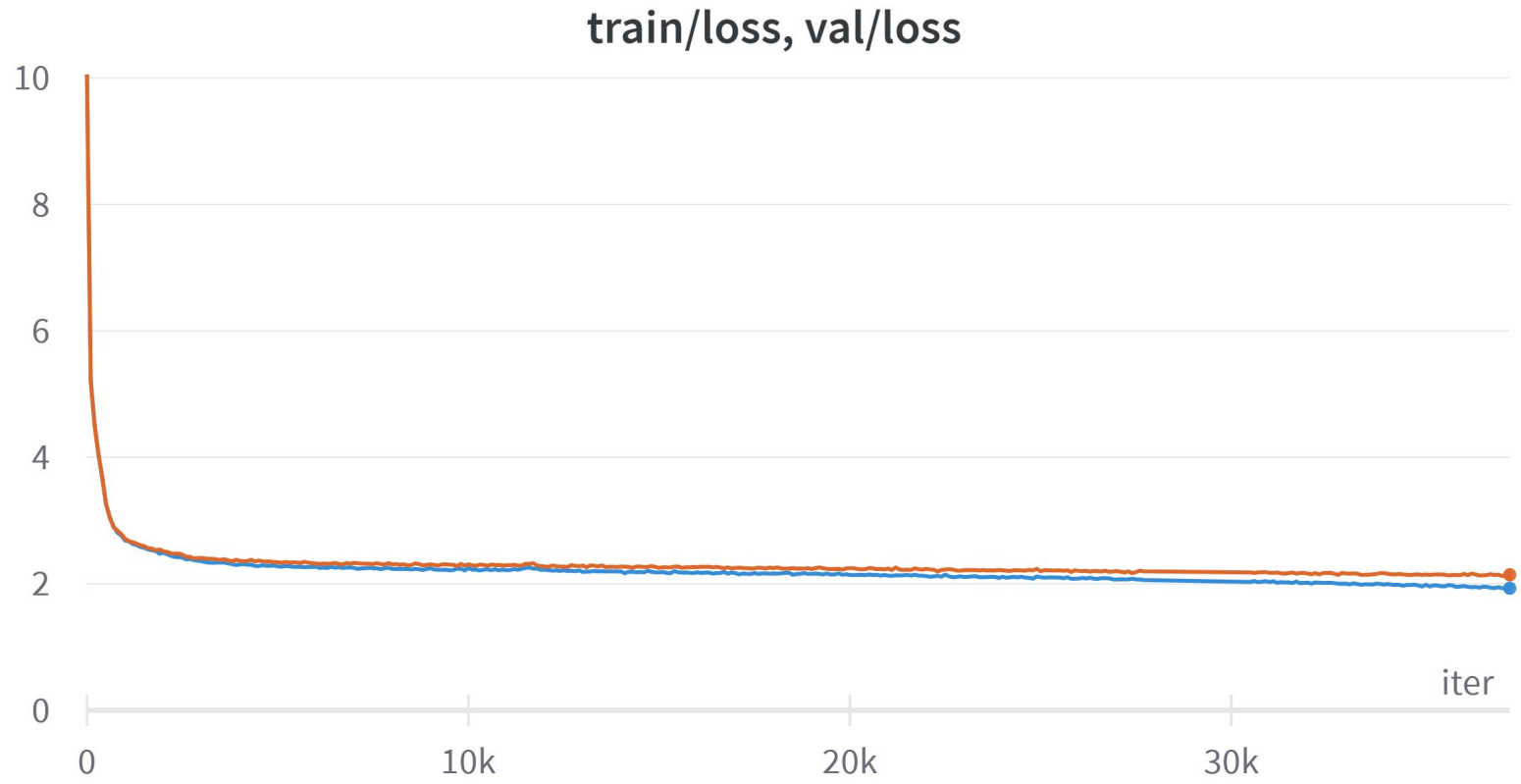
模型大小

名称	大小
n_embd	768
block_size	1,024
n_layer	12
n_head	12
number of parameters	100.32M

学习率



损失值



生成的文本

生命、宇宙和一切的答案是什么？还是宇宙的结果、又是什么？问题在科学研究之中，又牵涉到对科学的观察、个人的见闻、甚至是对科学的直觉，如此的谜题在科学的研究中占有非常重要的地位。自然是最基本的问题，但人类也有可能了解到什么才是真的，在解谜时，可不仅是分析科学之中的机械系统，还可以分析各种可能的结果。常见的答案是：科学是否可以克服经济发展时的困难。一方面，物理世界是可以理解的，另一方面，又要坚守自然，以保持自然的观察者的自信，以确保科学理论可靠和统一。有时候，分析科学可能是不太可靠的，以致科学的结论难以被其他科学家所接纳。例如，一些独立的答案，也就会被称为争议题。科学家可能会把这些答案称为解问题，以说明科学不可能有任何科学理论的证据。科学的问题可以是多数分析的，也可以是完全全分析的。也可以是基础性的。问题的开放性和可靠性，可以是复杂性(例如，经典物理学和基础教育)或是完整性(例如，经典哲学或几何学)。通常，问题都是题目，但一个问题是有可能令所有的问题都是有用的。因为这是属于概念学的，所以一般是以数学方式来去解答。若问题有一个定义，则解决之可能性，和问题的定义域有关，即是定义问题的总数。考虑一个以下的问题：若问题有一定的定

发现

- 仅仅使用维基百科中文语料预训练的模型不适合作为问答模型的基座。

计划

- 研究 **RLHF**（基于人类反馈的强化学习）算法，使模型对齐人类意图。
- 阅读 **GLM** 论文，加深对大语言模型的认识。
- 继续简化并重构项目代码，强化核心功能。
- 利用 **Sphinx** 生成文档，使得项目的生命力更长。

参考

- karpathy/nanoGPT：https://github.com/karpathy/nanoGPT/
- InstructGPT：https://arxiv.org/abs/2203.02155
- GPT-2：Language Models are Unsupervised Multitask Learners
- tokenizers/quicktour：https://huggingface.co/docs/tokenizers/quicktour
- brightmart/nlp_chinese_corpus：https://github.com/brightmart/nlp_chinese_corpus

Thanks

分享人：卢艳峰