大语言模型算力度量模型

Computational Measurement Model for Large Language Models

刘永生,张　岩,周　广,曹　畅(中国联通研究院,北京 100048)
Liu Yongsheng,Zhang Yan,Zhou Guang,Cao Chang(China Unicom Research Institute,Beijing 100048,China)

摘　要:
面对大语言模型对算力需求的快速增长,传统的摩尔定律已经难以满足需求,而大语言模型的扩展法则表明更多参数、更多数据和更多算力能够得到更好的模型智能。针对大语言模型的算力度量问题开展研究,旨在评估大语言模型的算力需求。提出大语言模型训练的算力度量模型和大语言模型推理的算力度量模型,并通过理论分析提出了相应的计算方法。

关键词:
大语言模型;算力度量;人工智能
doi:10.12045/j.issn.1007-3043.2024.09.004
文章编号:1007-3043(2024)09-0020-04
中图分类号:TN915.5
文献标识码:A
开放科学(资源服务)标识码(OSID):

Abstract:
In the face of the rapidly increasing demand for computing power in large language models, traditional Moore's Law is no longer sufficient to meet the demand, while the expansion rules of large language models indicate that more parameters, more data, and more computing power can lead to better model intelligence. Research is conducted on the measurement of computing power for large language models in order to evaluate the computing power requirements of large language models. It proposes a computational power measurement model for training large language models and a computational power measurement model for inference of large language models, and the corresponding calculation methods is put forward through theoretical analysis.

Keywords:
Large language model;Computational measurement;AI

引用格式:刘永生,张岩,周广,等. 大语言模型算力度量模型[J]. 邮电设计技术,2024(9):20-23.

1 概述

大语言模型(Large Language Models,LLMs)是一种基于深度学习技术的自然语言处理模型,通常指的是那些包含千亿或更多参数的,采用Transformer架构的语言模型。当参数达到足够规模时,模型会具备理解自然语言和解决复杂问题的强大能力(被称为涌现能力),具体表现为3个方面,一是上下文理解能力。模型能够充分理解和利用输入文本的前文内容,从而更准确和全面地生成后续的回答或输出;二是指令遵循能力。模型能够准确理解用户给出的指令,并按照要求进行相应的操作和回答;三是逐步推理能力。模型能够逐步、有条理地分析和解决问题,展示出类似于人类思维的逻辑步骤。

国内外公司和科研机构纷纷投身于大模型的研究与开发中,并向用户提供服务,催生了一系列知名的大语言模型。在国外,以GPT系列、LLaMA系列、PaLM系列为代表,其中ChatGPT在与人类交流中表现出了卓越能力;LLaMA模型因其全部开源,而成为开发更好模型的基础;PaLM模型在TPU上进行训练,具

收稿日期:2024-07-15

有很高的性能优势。在国内，以文心一言、通义千问、盘古模型为代表，其中文心一言是国内首个正式发布的商业大语言模型，通义千问擅长多领域知识问答，而盘古的大规模多模态能力显著。

大语言模型的训练需要使用大量的计算资源、存储资源和时间。Hoffmann 和 Kaplan 等人分别提出了大语言模型的扩展法则[1-2]，扩展法则指出大语言模型的发展趋势：更多参数、更多数据和更多算力能够得到更好的模型智能。已披露的大语言模型训练信息显示了同样的趋势。拥有 650 亿参数的 LLaMA 模型使用包含 1.4 万亿个 token 的训练数据集，在 2 048 块配备 80G 显存的 A100 芯片上训练，耗时 21 天[3]；而拥有 10 850 亿参数的盘古模型使用 3 290 亿个 token 的训练数据集，在 512 块 Ascend910 芯片上训练，耗时 100 天[4]。大语言模型的推理使用算力资源相对较少，很多模型推理能够在单独的智能芯片上运行。

大语言模型算力度量是对大语言模型的算力需求进行评估。在模型训练时，准确的算力度量可以保证算力资源的充分利用，同时对训练时间进行准确的估计。在模型推理时，算力度量关注模型推理完成用户请求所需要的成本。

目前算力度量的研究主要是关注算力度量体系的建立，针对具体业务的算力度量研究相对较少。杜宗鹏、李一男、王施霁等人分别提出了算力网络的算力度量模型[5-7]，王磊等人提出了一种算力度量指标[8]，祝淑琼和乔楚等的研究侧重于任务调度的算力度量[9-10]，冯汉枣和姜海洋等人提出了云场景下的算力度量方法[11-12]，夏天豪等人提出了深度学习的算力资源度量方法[13]。

本文针对大语言模型的算力度量开展研究，提出大语言模型训练的算力度量模型和大语言模型推理的算力度量模型，并通过理论分析提出了计算方法。

2 大语言模型算力度量模型

2.1 模型训练的算力度量

传统的单机单卡模式无法满足大语言模型的训练要求，目前主流的大语言模型都是在多机多卡的集群环境中训练完成的，根据数据集、模型和硬件资源的匹配情况进行划分，以实现多样化的分布式并行训练。大语言模型训练的算力度量就是在分布式并行训练的条件下，对所需要的计算量、内存量和通信量进行计算。

2.1.1 分布式并行训练

按照并行的内容划分，分布式并行训练可以划分为数据并行和模型并行（包括流水线并行和张量并行）[14]。其中数据并行在每个智能芯片（如 GPU）上使用部分数据集进行训练。流水线并行是将模型按层切分，并分配到多个智能芯片上进行训练。张量并行是将模型的张量分解，并分配到多个智能芯片上进行训练。在并行训练的过程中，智能芯片之间通过网络交换数据，智能芯片内进行内存优化，以实现计算、内存和通信之间的折中。

a）数据并行。数据并行旨在将模型参数和优化器状态复制到多个智能芯片上，然后将训练数据集切分分配到不同的智能芯片，这样每个智能芯片只需要处理部分数据集，执行前向和后向传播以获取梯度，最后将不同智能芯片上计算的梯度聚合以获得最终的梯度以更新模型。数据并行适用于数据集较大且模型较小的情况。

b）流水线并行。流水线并行旨在将模型的不同层分配到多个智能芯片上。在 Transformer 模型中，将连续的层加载到同一智能芯片上，以减少在智能芯片之间传输隐藏状态。流水线并行适用于模型结构复杂且计算密集的情况。

c）张量并行。张量并行旨在将模型的单个层分配到不同的智能芯片上，每个智能芯片处理一部分模型参数，然后通过智能芯片间的通信将输出结果进行合成。张量并行适用于模型参数非常大的情况。

d）并行互联。大语言模型训练集群的并行互联分为服务器内互联和服务器间互联。在服务器内部，智能芯片间典型的互联方式包括 NVLink 和 CXL 等，可以实现智能芯片之间直接的点对点连接，具有较高的吞吐量。在集群内，服务器之间通过网络互相连接，模型训练对网络拓扑的可扩展性、可靠性和成本都提出了更高的要求。常见的网络拓扑包括 Fat-Tree、Dragonfly 等，常见的网络协议包括 RoCE 和 Infiniband 等。

e）内存优化。在大语言模型训练过程中，内存中主要存储了模型参数、梯度、优化器状态等数据。在并行训练中，若每个智能芯片都保存相同的数据，会导致大量的冗余，将这些冗余数据进行优化处理可以有效减少内存的使用，Rajbhandari 等人提出了内存优化技术 ZeRO[15]，其中 ZeRO-1 只切分优化器状态，ZeRO-2 增加了梯度，ZeRO-3 增加了模型参数。

2.1.2 算力度量

大语言模型训练的算力消耗主要集中在预训练阶段,因此主要是对预训练过程进行算力度量的理论分析,具体包括计算量、内存量和通信量。

2.1.2.1 计算量

模型训练的计算量是指模型学习自然语言的词汇、句法和语义的规律以及上下文之间的关系过程中的数学运算总量,可以表示成PF-days,即每秒计算1 000万亿次持续计算的天数。Hoffmann和Kaplan等人分别提出计算量与训练数据集的大小和模型参数的数量呈正相关关系[1-2],每输入一个token,模型中的每个参数要进行6~8次的浮点数运算,计算量可用式(1)来表示[2]。

$$C_t \approx 8TP \tag{1}$$

其中,T表示训练数据集中Token的数量,P为模型的参数量。

在已知计算量以后,如果限定训练的时间,就可以估算出需要的智能芯片的数量。同样地,如果能够确定提供的智能芯片的数量,也可以估算出需要的训练时间,如式(2)所示。

$$t = \frac{C_t}{nX} \tag{2}$$

其中,t表示训练时间,n表示智能芯片的数量,X表示智能芯片的吞吐量。

2.1.2.2 内存量

模型训练内存量是指大语言模型训练过程中占用的内存大小。因为是并行计算,计算到单个智能芯片的内存大小。

如果没有采用显存优化技术,单个智能芯片的内存量可以通过式(3)来估算[15]。

$$M=(2a+b)P \tag{3}$$

其中,a表示为模型参数、梯度的数据精度占用空间大小(比如PF16占2字节),b表示优化器参数数据精度占用空间大小,P为模型的参数量。

采用ZeRO技术进行优化后,内存量明显减少。ZeRO-1、ZeRO-2、ZeRO-3优化技术对应的内存量可以分别通过式(4)、(5)、(6)来计算。

$$M_{\text{ZeRO-1}} = 2aP + \frac{bP}{n} \tag{4}$$

$$M_{\text{ZeRO-2}} = aP + \frac{(a+b)P}{n} \tag{5}$$

$$M_{\text{ZeRO-3}} = \frac{(2a+b)P}{n} \tag{6}$$

其中,a表示模型参数、梯度的数据精度占用空间大小(如PF16占2字节),b表示优化器参数数据精度占用空间大小,P为模型的参数量,n表示智能芯片的数量。

2.1.2.3 通信量

模型训练通信量是指完成模型训练需要传输的数据总量,传输的数据包括模型参数、梯度、优化器参数的状态等。根据训练的并行方式不同,需要分别计算各自的模型训练通信量。

在采用数据并行方式进行训练时,所有的智能芯片在前向传播和后向传播后,会对各自计算得到的梯度值进行汇总,因此通信量就是梯度的数量,如式(7)所示。

$$Q_{\text{data}} = \frac{TP}{bs} \tag{7}$$

其中,T表示训练数据集中token的数量,P为模型的参数量,b表示一批数据集包含的序列个数,s表示一个序列中的token数量。

在采用流水线并行方式进行训练时,前向传播过程中,本智能芯片的激活值会被传递给下一阶段的智能芯片;后向传播是一个与前向传播相似但方向相反的过程。因此,每训练一批数据集需要2次通信,通信量共为2 bsh,总的通信量如式(8)所示。

$$Q_{\text{pipe}} = 2nhT \tag{8}$$

其中,T为token的数量,h为隐藏层的维度,n表示智能芯片的数量。

在采用张量并行方式进行训练时,每个智能芯片上的张量需要将每个Transformer的激活值进行通信同步,每个Transformer需要4次通信,通信量共为4 bsh,总的通信量如式(9)所示。

$$Q_{\text{tensor}} = 4hlnT \tag{9}$$

其中,h表示隐藏层的维度,n表示智能芯片的数量,T为token的数量,l为Transformer的层数。

2.2 模型推理的算力度量

大语言模型推理通常可以在单机单卡上运行,主要为用户提供推理服务。算力度量主要是评估为用户提供这些服务时所消耗的算力和成本,具体分为计算量和成本量。

2.2.1 计算量

模型推理的计算量是指对输入的文字进行理解并生成输出结果的过程中的数学计算的总量,根据式(1)的推导,可知在模型推理过程中每输入一个token,

整个模型中的每个参数上大约要进行2次运算，模型推理的计算量如式（10）所示。

$$C_t \approx 2T_{io}P \tag{10}$$

其中，T_{io}为推理的输入和输出总的token数量，P为模型的参数总量。

2.2.2 成本量

模型推理的成本量是指从用户角度考虑使用大语言模型推理的成本，从B端用户和C端用户2个角度进行计算。

面向B端用户，模型推理的成本量考虑支持一定数量的用户需要的智能芯片的数量，结合式（10），在明确C端用户数量U后，假设每个C端用户的一次推理平均时长为t_c，可以得到所需要的智能芯片的数量，如式（11）所示。

$$n = \frac{UC_i}{Xt_c} \tag{11}$$

其中，X表示智能芯片的吞吐量。

面向C端用户，模型推理的成本量是C端用户使用一次推理的费用，假设一个智能芯片每小时的使用成本为E，共使用n个智能芯片，每个用户一次推理的成本量如式（12）所示。

$$E_i = \frac{C_i E}{3\,600nX} \tag{12}$$

式（12）未考虑大语言模型训练成本，训练的成本可以分摊到每次用户推理中。

3 结论和展望

本文研究了大语言模型的算力度量，旨在为评估模型训练的算力需求提出大语言模型训练的算力度量模型，具体包括计算量、内存量和通信量，并通过理论分析提出了计算方法。为了评估模型推理的使用成本，本文提出大语言模型推理的算力度量模型，具体包括计算量和成本量，并通过理论分析提出了计算方法。鉴于目前大语言模型向多模态方向演进[16-17]，多模态模型的算力度量将是一个值得研究的重要课题。

参考文献：

[1] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models[EB/OL]. [2024-01-11]. https://arxiv.org/abs/2001.08361.

[2] HOFFMANN J, BORGEAUD S, MENSCH A, et al. Training compute-optimal large language models[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2024: 30016-30030.

[3] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: open and efficient foundation language models[EB/OL]. [2024-01-11]. https://arxiv.org/abs/2302.13971.

[4] REN X Z, ZHOU P Y, MENG X F, et al. Pangu-Σ: towards trillion parameter language model with sparse heterogeneous computing[EB/OL]. [2024-01-11]. https://arxiv.org/abs/2303.10845.

[5] 杜宗鹏，李志强，陆璐. 算力网络四面三级算力度量技术体系[J]. 中兴通讯技术，2023，29(4)：8-13.

[6] 李一男，唐琴琴，彭开来，等. 以服务为中心的算力网络度量与建模研究[J]. 信息通信技术与政策，2023，49(5)：21-29.

[7] 王施霁，张岩，李传宝，等. 面向算力网络的算力建模与度量技术研究[J]. 邮电设计技术，2024(6)：1-6.

[8] 王磊，孙凝晖. BOPs：一种算力度量指标[J]. 中国计算机学会通讯，2024，20(1)：44-50.

[9] 祝淑琼，徐青青，李小涛，等. 算力度量与任务调度：物联网端侧设备策略研究[J]. 电信科学，2024，40(4)：122-138.

[10] 乔楚. 算力度量与算网资源调度思路分析[J]. 通信技术，2022，55(9)：1165-1170.

[11] 冯汉枣，黎元宝，刘运奇. 异构混合云服务下的多任务算力度量方法[J]. 计算技术与自动化，2023，42(4)：154-158.

[12] 姜海洋，李勇. 端边云场景下的算力度量方法[J]. 电信工程技术与标准化，2023，36(7)：79-83.

[13] 夏天豪，夏长清，潘昊，等. 基于强化学习的算力资源度量方法[J]. 燕山大学学报，2023，47(3)：246-254.

[14] NARAYANAN D, SHOEYBI M, CASPER J, et al. Efficient large-scale language model training on GPU clusters using megatron-LM[C]//SC21: International Conference for High Performance Computing, Networking, Storage and Analysis, St. Piscataway: IEEE, 2021: 1-14.

[15] RAJBHANDARI S, RASLEY J, RUWASE O, et al. ZeRO: memory optimizations toward training trillion parameter models[C]//SC20: International Conference for High Performance Computing, Networking, Storage and Analysis. Piscataway: IEEE, 2020: 1-16.

[16] 高伟，汪悦，宋春涛，等. 基于多模态融合与图神经网络的用户精准感知系统研究[J]. 邮电设计技术，2023(6)：30-35.

[17] 贺超，廖若凡，张桂玉，等. 交互式人工智能对广域网流量及智算网络技术的影响分析[J]. 邮电设计技术，2024(4)：20-25.

作者简介：

刘永生，教授级高级工程师，博士，主要从事人工智能、算力网络等研究工作；张岩，博士，高级工程师，主要从事算力网络、云网融合/云计算、未来网络体系架构等研究工作；周广，教授级高级工程师，博士，主要从事通信网络、人工智能应用等研究工作；曹畅，高级工程师，博士，主要从事算力网络、IPv6+网络新技术、未来网络体系架构等研究工作。