南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

高效大模型策略

魏明强、宫丽娜
计算机科学与技术学院

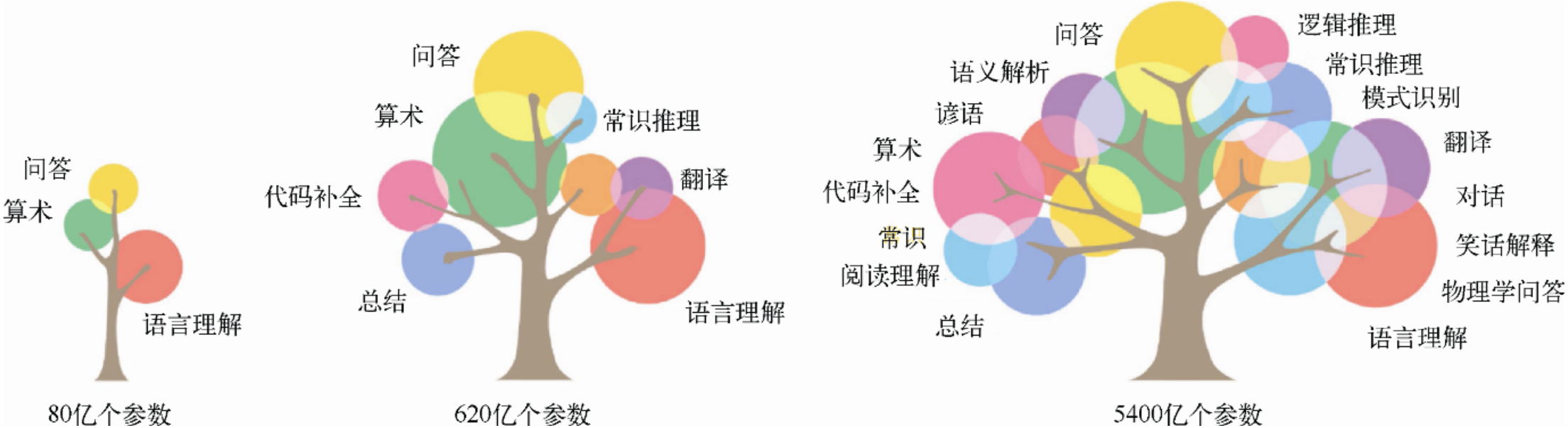
目录

目录

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

大模型效率面临的问题

- 由于更大的参数规模需要更高的计算成本和内存需求，大模型的训练和微调会受到严重限制
- 训练这些模型需要大量的数据和资源，给数据获取、资源分配和模型设计带来挑战，探索不同架构或策略的成本变得过高
- 大规模参数使大模型不适合部署在资源受限的环境中，如边缘设备



随着模型参数规模的增大，大模型不仅提高了现有任务的性能，而且还出现了很多新功能

目录

大模型效率及其评估指标

- 本章将“大模型效率”定义为大模型产生特定性能时所需的资源，与性能成正相关，与资源成负相关

- 高效大模型策略旨在不影响模型性能的情况下优化计算和内存资源，这些评估指标将是高效大模型策略的重要依据和体现

- **评估大模型效率的关键指标**
  - 参数数量
  - 模型大小
  - 浮点运算次数
  - 推理时间/token生成速度
  - 内存占用
  - 碳排放

大模型效率及其评估指标

评估大模型效率的关键指标

- 参数数量

参数数量是直接影响模型学习能力和复杂性的关键因素。

这些参数包括权重和偏差等参数，在训练或微调阶段是可以学习的。

更大的参数数量通常使模型能够学习到更复杂的数据模式和新功能，但会影响训练和推理计算的时间。

- 模型大小

模型大小定义为存储整个模型所需的磁盘空间，通常以千兆字节（GB）或兆字节（MB）等为单位。

模型大小会受到多个因素的影响，其中最主要的因素是参数数量，其他因素有参数数据类型和特定的体系结构。

模型大小会直接影响存储需求，提前考虑模型大小对在存储受限环境下的部署尤其重要。

- 浮点运算次数

浮点运算次数是指单次前向传播过程中浮点运算（如加减乘除法）的次数（计算量），用于估算大模型的计算复杂度。

较高的浮点运算次数通常意味着模型有着更高的计算要求，在资源有限的环境中部署这种模型将是一个挑战。

系统的并行优化以及不同的架构也都会影响最终的整体计算效率。

大模型效率及其评估指标

评估大模型效率的关键指标

- 推理时间/token生成速度

推理时间也称为延迟，是大模型在推理阶段从输入到生成响应所需的时间，单位通常为毫秒或秒。

推理时间是在实际部署的设备上进行评估的，考虑了特定的硬件和优化条件，提供了现实世界性能的实用衡量标准。

token生成速度是指模型在每秒内可以处理的token数，它能够用来规范推理时间，是反映模型速度和效率的关键性能指标。

- 内存占用

内存占用是指在推理或训练期间加载和运行模型所需的随机存取存储器的内存大小，通常以MB或GB为单位。

内存占用的内容不仅包括模型参数，还包括其他运行时必需数据，如中间变量和数据结构。

较大的内存占用会限制模型的可部署性，尤其是在资源受限的环境中，需要优化技术来降低占用，如模型剪枝或量化。

- 碳排放

碳排放通常以模型从训练到推理的过程中排放的二氧化碳量来衡量，反映了训练和运行该模型对环境的影响。

碳排放受到各种因素的影响，包括所用硬件的能源效率、电力来源，以及模型训练和运行的持续时间。

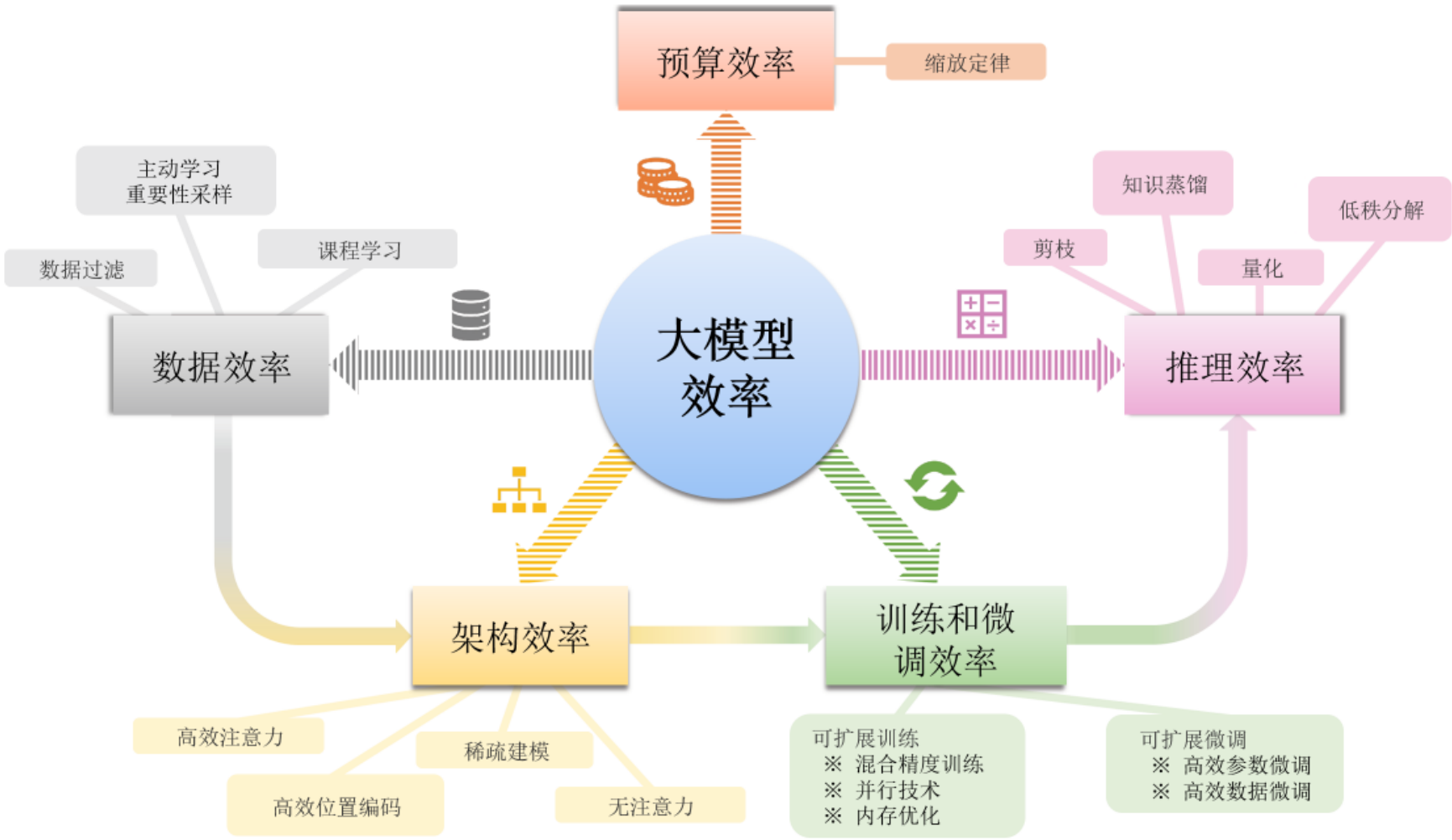
可以通过模型优化、硬件加速和算法改进等方式提高能效，还可以为数据中心（如苹果公司的云上贵州数据中心、腾讯的七星洞数据中心）选择更环保的能源，从而减少碳排放。

目录

高效大模型策略

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

- **提高大模型效率的关键方向**
  - 预算效率
  - 数据效率
  - 架构效率
  - 训练效率
  - 推理效率
  - 微调效率



目录