论文阅读

11.02 A Survey of Resource-efficient LLM and Multimodal Foundation Models

Abstract

Large foundation models, including large language models (LLMs), vision transformers (ViTs), diffusion, and LLM-based multimodal models, are revolutionizing the entire machine learning lifecycle, from training to deployment. However, the substantial advancements in versatility and performance these models offer come at a significant cost in terms of hardware resources. To support the growth of these large models in a scalable and environmentally sustainable way, there has been a considerable focus on developing resource-efficient strategies. This survey delves into the critical importance of such research, examining both algorithmic and systemic aspects. It offers a comprehensive analysis and valuable insights gleaned from existing literature, encompassing a broad array of topics from cutting-edge model architectures and training/serving algorithms to practical system designs and implementations.
The goal of this survey is to provide an overarching understanding of how current approaches are tackling the resource challenges posed by large foundation models and to potentially inspire future breakthroughs in this field.

大型基础模型，包括大型语言模型（LLM）、视觉转换器（ViT）、扩散和基于 LLM 的多模态模型，正在彻底改变从训练到部署的整个机器学习生命周期。 然而，这些模型在通用性和性能方面的巨大进步需要付出巨大的硬件资源代价。 为了以可扩展和环境可持续的方式支持这些大型模型的发展，人们一直非常关注开发资源高效型战略。 本调查报告深入探讨了此类研究的关键重要性，从算法和系统两方面进行了研究。 它提供了从现有文献中收集到的全面分析和宝贵见解，涵盖了从前沿模型架构和训练/服务算法到实用系统设计和实现的广泛主题。 本调查报告的目的是让人们全面了解当前的方法是如何应对大型基础模型带来的资源挑战的，并为该领域未来的突破提供潜在的灵感。

单词：

1. substantial 重大的，坚实的
2. deployment 部署
3. versatility 多功能性，通用性
4. scalable 可扩展的
5. delves into 深入研究
6. valuable insights 宝贵见解
7. encompass 拥有，包含
8. a broad array of 各种各样的
9. cutting-edge model architectures 前沿的模型架构
10. overarching 总体的
11. tackling 解决，应对

_Keywords Foundation Model(基础模型) · Large Language Model(大语言模型) · Vision Transformer(视觉 transformer) · Diffusion Model(扩散模型) · Multimodal LLM(多模态LLM) · Model Compression(模型压缩) · Machine Learning System(机器学习系统) · Serving System(服务系统) · Pre-training(预训练) · Fine-tuning(微调) · Edge Intelligence(边缘智能)

1 INTRODUCTION

In the rapidly evolving field of artificial intelligence (AI), a paradigm shift is underway. We are witnessing the transition from specialized, fragmented deep learning models to versatile, one-size-fits-all foundation models. These advanced AI systems are capable of operating in an open-world context, interacting with open vocabularies and image pixels for unseen AI tasks, i.e., zero-shot abilities. They are exemplified by (1) Large Language Models (LLMs) such as GPTs 41 that can ingest almost every NLP task in the form as a prompt; (2) Vision Transformers Models (ViTs) such as Masked Autoencoder 141 that can handle various downstream vision tasks; (3) Latent Diffusion Models (LDMs) such as Stable Diffusion 336 that generate high-quality images with arbitrary text-based prompts; (4) Multimodal models such as CLIP 321 and ImageBind 123 that map different modal data into the same latent space and are widely used as backbone for cross-modality tasks like image retrieval/search and visual-question answering. Such flexibility and generality marks a significant departure from the earlier era of AI, setting a new standard for how AI interfaces with the world.

在快速发展的人工智能（AI）领域，模式正在发生转变。 我们正在见证从专业化、碎片化的深度学习模型向多功能、一刀切的基础模型过渡。 这些先进的人工智能系统能够在开放世界的环境中运行，与开放的词汇表和图像像素进行交互，以完成未曾见过的人工智能任务，即零拍摄能力。 它们的例子有：(1) 大型语言模型（LLMs），如 GPTs [41]，能以提示形式摄取几乎所有 NLP 任务；(2) 视觉转换器模型（ViTs），如 Masked Autoencoder [141]，能处理各种下游视觉任务；（3）潜在扩散模型（LDM），如Stable Diffusion336，可生成高质量图像，并带有任意文本提示；(4) 多模态模型，如 CLIP [321] 和 ImageBind [123]，将不同模态数据映射到同一潜在空间，被广泛用作图像检索/搜索和视觉问题解答等跨模态任务的支柱。 这种灵活性和通用性标志着与早期人工智能时代的重大不同，为人工智能与世界的交互方式设定了新标准。

单词

1. a paradigm shift is underway 模式转变正在进行中
2. specialized, fragmented deep learning models to versatile, one-size-fits-all 专业化、碎片化的深度学习模型向多功能、一刀切
3. image pixels 图片像素
4. They are exemplified by (1) Large Language Models (LLMs) 具体例子有 (1) 大语言模型
5. GPTs that can ingest almost every NLP task in the form as a prompt; GPT 可以将表单中的几乎所有 NLP 任务进行提取作为提示；
6. latent space 潜空间
7. backbone 主干
8. cross-modality tasks 跨模态任务
9. image retrieval/search and visual-question answering. 图像检索/搜索和视觉问题解答
10. marks a significant departure(离开，出发，离职) 标志着重大转变

The success of these foundation models is deeply rooted in their scalability: unlike their predecessors, these models' accuracy and generalization ability can continuously expand with more data or parameters, without altering the underlying simple algorithms and architectures. An impressive evidence is the scaling law [177]: it describes how the performance of transformer-based models can predictably improve with more model size and data volume; until today, the scaling law stands still. This scalability is not just a matter of model size; it extends to their ability to tackle increasingly complex tasks, making them a cornerstone in the journey towards artificial general intelligence (AGI).

这些基础模型的成功深深植根于它们的可扩展性：与它们的前辈不同，这些模型的准确性和概括能力可以随着数据或参数的增加而不断扩大，而无需改变底层的简单算法和架构。 一个令人印象深刻的证据是 缩放定律[177]：它描述了基于transformer的模型的性能如何随着模型规模和数据量的增加而得到可预测的提高；直到今天，缩放定律仍未改变。 这种可扩展性不仅仅是模型大小的问题，它还扩展到了处理日益复杂任务的能力，使其成为人工通用智能（AGI）的基石。

单词

1. predecessors 前任，前辈
2. volume 卷、音量、量
3. cornerstone 基石
4. AGI artificial general intelligence 人工通用智能

However, the scalability comes at a cost of huge resource demand. Foundation models, by their very nature, are resource-hungry for training and deployment. These resources encompass not only the computing processors like GPUs and TPUs, but also the memory, energy, and network bandwidth. For example, the pre-training of LLaMa-2-70B takes 1.7× millions of GPU hours and consumes 2.5×10^12 Joules of energy. The estimated total emissions were 291 tons of CO2 equivalent. Beyond training, the data processing, experimentation, and inference stages consume comparable or even more electricity according to Meta AI [416]. A recent analysis [81] reveals that, to satisfy the continuation of the current trends in AI capacity and adoption, NVIDIA needs to ship 1.5 million AI server units per year by 2027. These servers, running at full capacity, would consume at least 85.4 terawatt-hours of electricity annuall – more than what many countries like New Zealand and Austria use in a whole year, as illustrated in Figure 1. Since foundation models proceed growth in size and complexity, their resource requirements escalate, often exponentially, posing a significant challenge in their development and deployment.

然而，可扩展性的代价是巨大的资源需求。 基础模型就其本质而言，在训练和部署时需要大量资源。 这些资源不仅包括 GPU 和 TPU 等计算处理器，还包括内存、能源和网络带宽。 例如，LLaMa-2-70B 的预训练需要 1.7× 百万 GPU 小时，消耗 2.5×10^12 焦耳能量。 估计总排放量为 291 吨二氧化碳当量。 据 Meta AI称，除训练外，数据处理、实验和推理阶段消耗的电力相当甚至更多。 最近的一项分析显示，为了满足当前人工智能容量和应用趋势的持续发展，英伟达公司需要在2027年前每年出货150万台人工智能服务器。 如图1所示，这些满负荷运行的服务器每年将消耗至少 85.4 太瓦时的电力，超过新西兰和奥地利等许多国家全年的用电量。 由于基础模型的规模和复杂性不断增长，其资源需求也随之增加，通常呈指数级增长，这给基础模型的开发和部署带来了巨大挑战。

单词

1. by their very nature 就他们本质而言
2. estimated 估计
3. comparable 相当
4. adoption 采用，应用
5. ship 船运
6. terawatt-hours 兆兆瓦时
7. as illustrated in Figure 就像图所示
8. escalate 升级
9. exponentially 指数地

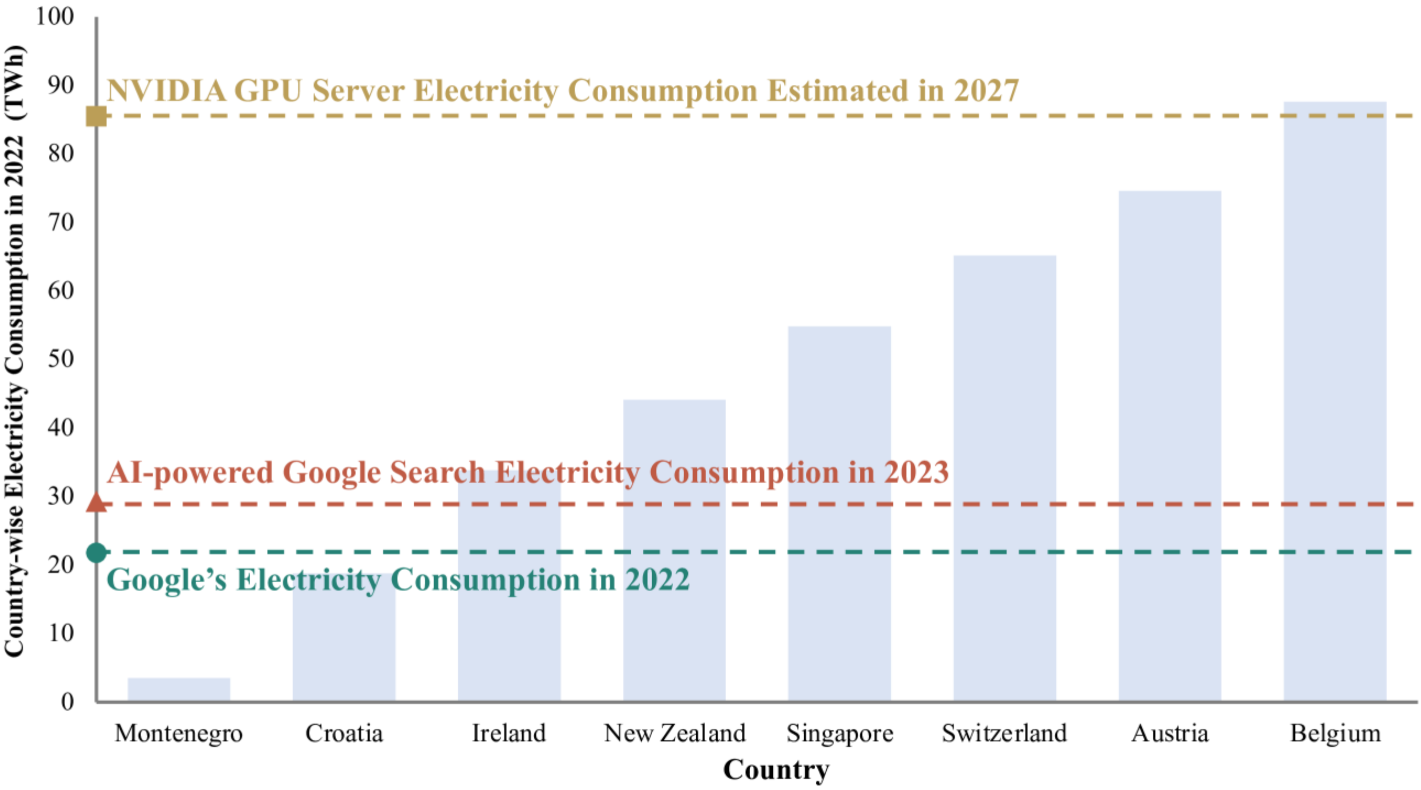

Figure 1: The electricity consumption comparison between countries and AI. Data source: [81].

The huge resource footprint of large foundation model also hinders its democratization. Till the end of 2023, there are only a few major players capable of training and deploying the state-of-the-art foundation models, who thereby have powerful control over the public and can potentially manipulate them in a way they prefer. The models are served on clouds instead of devices as many lightweight DNNs do [434, 476]; it makes data privacy preservation almost impossible. Though recently, smartphone vendors have been boasting about running large foundation models locally and some pioneering engines are developed for on-device LLMs [121, 11, 10], the models demonstrated are limited to relatively small scale (e.g., <10B) [266] and have not yet seen real-world deployment.

大型基金会模式的巨大资源占用也阻碍了其民主化。 到 2023 年底，只有少数几个主要公司有能力训练和部署最先进的基础模型，从而对公众拥有强大的控制权，并有可能按照自己喜欢的方式操纵这些模型。 正如许多轻量级 DNN 所做的那样 [434, 476]，这些模型在云端而非设备上提供服务；这使得数据隐私保护几乎成为不可能。 尽管最近智能手机供应商一直在吹嘘可以在本地运行大型基础模型，而且一些开创性的引擎也是为设备上的 LLMs 开发的[121, 11, 10]，但所展示的模型仅限于相对较小的规模（例如，小于 10B）[266]，而且尚未在现实世界中部署。

单词

1. smartphone vendors 智能手机供应商

Thereby, a significant amount of research has been dedicated to enhance the efficiency of these foundation models. These efforts span a wide range of approaches, from optimizing algorithms to system-level innovations, focusing on reducing the resource footprint of these models without compromising their performance. This survey aims to delve into these research efforts, exploring the diverse strategies employed to make foundation models more resource-efficient. We will examine advancements in algorithmic efficiency, system optimizations, data management techniques, and the development of novel architectures that are less resource-intensive. The survey also spans from clouds to edge and devices, where the large foundation models gain dramatic attentions as well. Through this exploration, we aim to provide a comprehensive understanding of the current state and future directions of resource-efficient algorithms and systems in the realm of foundation models.

因此，大量研究致力于提高这些基础模型的效率。 这些努力涵盖了从优化算法到系统级创新等多种方法，重点是在不影响性能的前提下减少这些模型的资源占用。 本调查旨在深入探讨这些研究工作，探索为提高基础模型的资源效率而采用的各种策略。 我们将考察在算法效率、系统优化、数据管理技术以及开发资源密集度更低的新型架构方面取得的进展。 调查范围还包括云、边缘和设备，其中大型基础模型也受到了极大关注。 通过这些探索，我们旨在全面了解基础模型领域资源高效算法和系统的现状和未来发展方向。

单词

1. be dedicated to 致力于
2. reducing the resource footprint of these models 减少这些模型的资源占用
3. compromise 妥协，折中
4. resource-intensive 资源密集型
5. span 跨越，范围，包括。涵盖
6. comprehensive 综合的；所有的；综合性的(接收各种资质的学生)；全部的；
7. in the realm(领域) of foundation models. 在基础模型领域

Scope and rationales. The scope of this survey is mainly defined by following aspects. (i) We survey only algorithm and system innovations; we exclude a huge body of work at hardware design, which is equally important but has been already wrapped up well [192, 185]. (ii) The definition of resource in this survey is limited to mainly physical ones, including computing, memory, storage, bandwidth, etc; we exclude training data (labels) and privacy that can also be regarded as resources; (iii) We mainly survey papers published on top-tier CS conferences, i.e., those included in CSRankings.

范围和理由 本次调查的范围主要由以下几个方面确定。
(i) 我们只调查算法和系统创新；我们排除了硬件设计方面的大量工作，这些工作同样重要，但已被很好地总结[192, 185]。
(ii) 本调查中的资源定义主要限于物理资源，包括计算、内存、存储、带宽等；我们不包括也可视为资源的训练数据（标签）和隐私；
(iii) 我们主要调查在顶级 CS 会议上发表的论文，即 CSRankings 中收录的论文。

单词

1. Scope and rationales. 范围和理由。

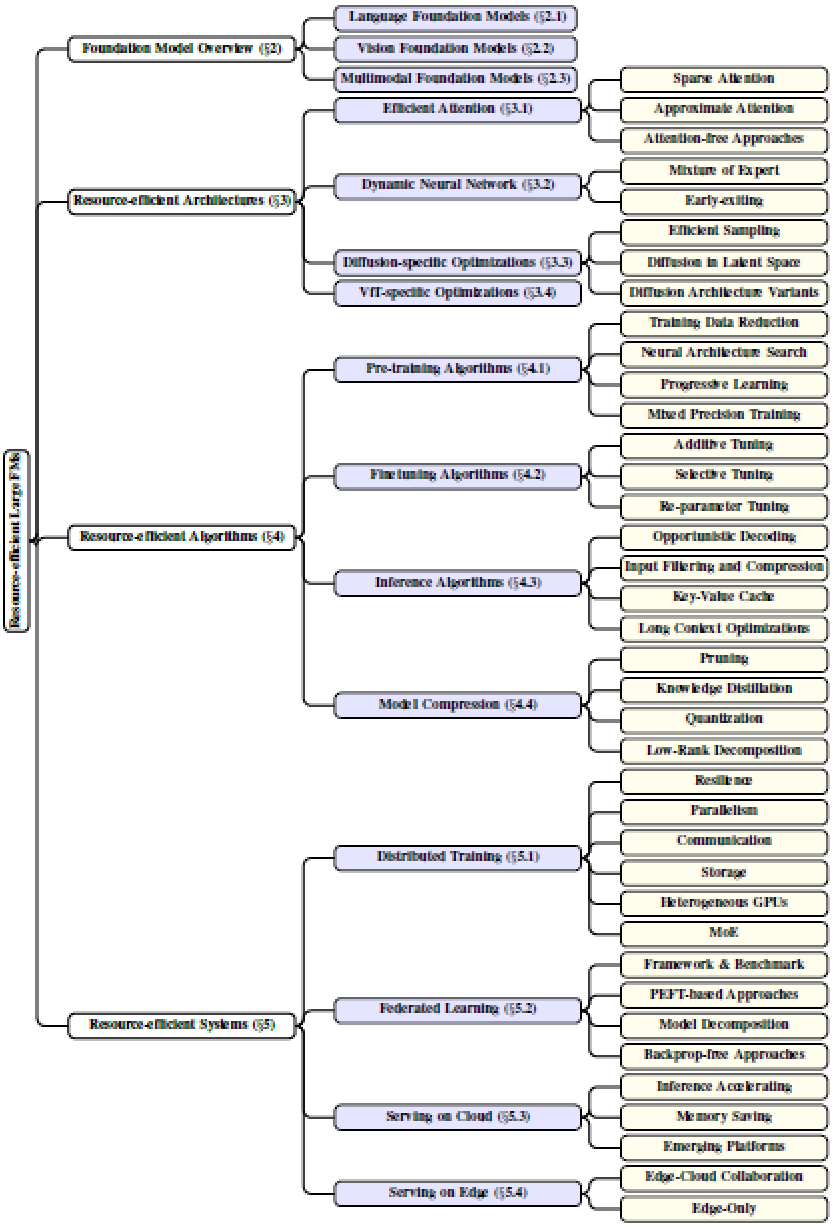
Resource-efficient Large FMs
Foundation Model Overview (§2)
Language Foundation Models (§2.1)
Vision Foundation Models (§2.2)
Multimodal Foundation Models (§2.3)
Resource-efficient Architectures (§3)
Efficient Attention (§3.1)
Sparse Attention
Approximate Attention
Attention-free Approaches
Dynamic Neural Network (§3.2)
Mixture of Expert
Early-exiting
Diffusion-specific Optimizations (§3.3)
Efficient Sampling
Diffusion in Latent Space
Diffusion Architecture Variants
ViT-specific Optimizations (§3.4)
Resource-efficient Algorithms (§4)
Pre-training Algorithms (§4.1)
Training Data Reduction
Neural Architecture Search
Progressive Learning
Mixed Precision Training
Finetuning Algorithms (§4.2)
Additive Tuning
Selective Tuning
Re-parameter Tuning
Inference Algorithms (§4.3)
Opportunistic Decoding
Input Filtering and Compression
Key-Value Cache
Long Context Optimizations
Model Compression (§4.4)
Pruning
Knowledge Distillation
Quantization
Low-Rank Decomposition
Resource-efficient Systems (§5)
Distributed Training (§5.1)
Resilience
Parallelism
Communication
Storage
Heterogeneous GPUs
MoE
Federated Learning (§5.2)
Framework & Benchmark
PEFT-based Approaches
Model Decomposition
Backprop-free Approaches
Serving on Cloud (§5.3)
Inference Accelerating
Memory Saving
Emerging Platforms
Serving on Edge (§5.4)
Edge-Cloud Collaboration
Edge-Only

Figure 2: The organization of this survey.

We also manually pick related and potentially high-impact papers from arXiv.
(iv) We mainly survey papers published after the year of 2020, since the innovation of AI is going fast with old knowledge and methods being overturned frequently.

Organization. Figure 2 illustrates the organization of this survey.

Full open-source. All materials of this survey are freely available at:

https:github.com/UbiquitousLearning/Efficient_Foundation_Model_Survey

我们还人工从 arXiv 中挑选相关的、潜在的高影响力论文。 (iv) 我们主要调查 2020 年以后发表的论文，因为人工智能的创新速度很快，旧的知识和方法经常被推翻。

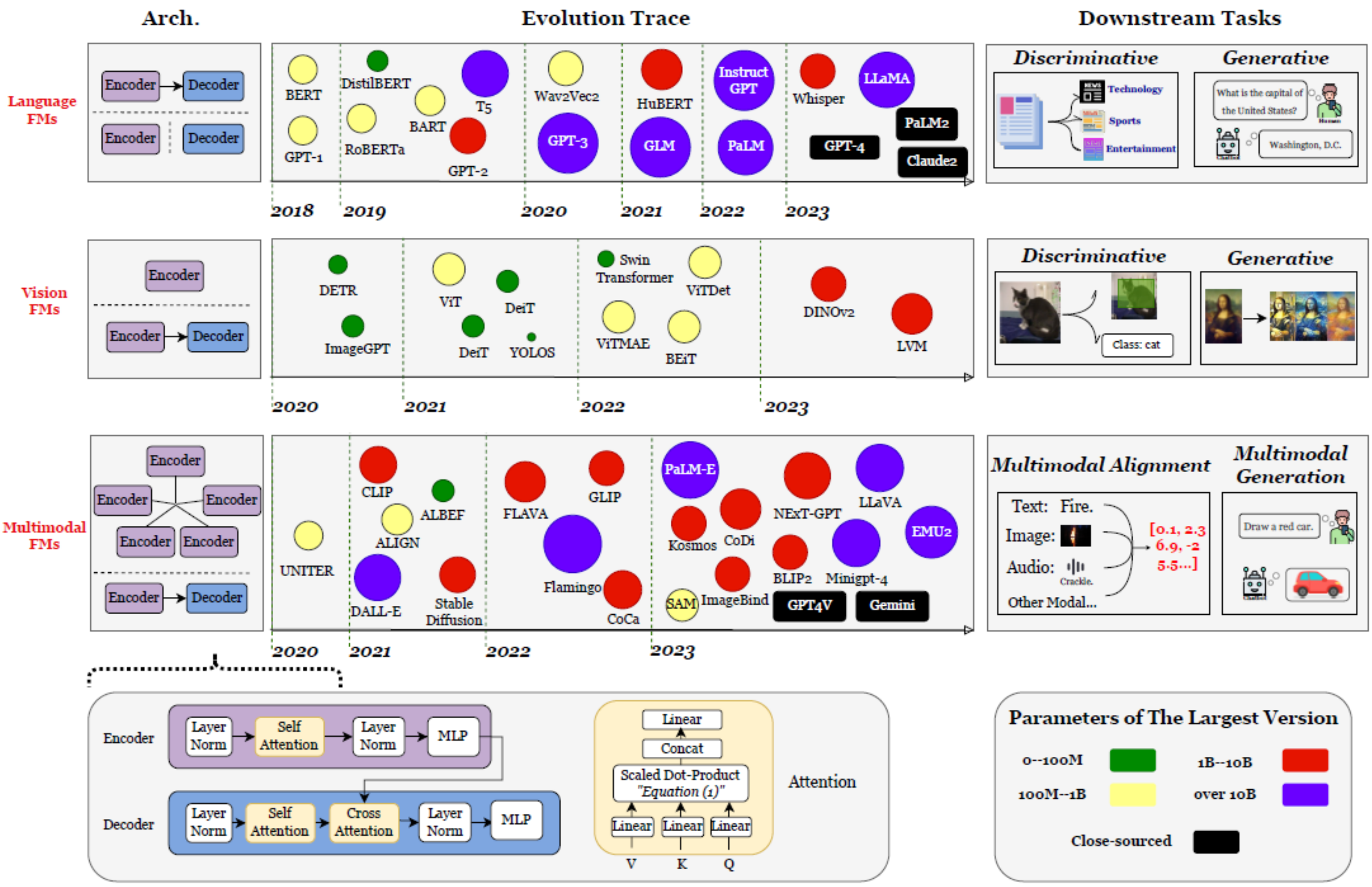

Figure 3: The evolutionary trace of foundation models.

2 FOUNDATION MODEL OVERVIEW

2.1 Language Foundation Models 语言基础模型

This section includes a discussion of both text-based and speech-based language models, highlighting their key architecture and milestone models.

本节将讨论基于文本和基于语音的语言模型，重点介绍其关键架构和里程碑模型。

2.1.1 Model Architectures

Transformer pipeline. Vaswani et al. 388 introduced the attention-based Transformer architecture, a foundational element in the development of most Large FMs. As depicted in Figure 3, the process initiates by converting input words into high-dimensional vectors through an embedding layer. During processing, attention mechanisms assign varying weights to different segments of these input vectors. Following attention, layer normalization is applied to the output, ensuring stabilization and standardization of the activations. Subsequently, each position-wise vector undergoes transformation through a feedforward network, introducing non-linearity and enabling the model to capture complex data patterns. Through multiple layers that incorporate these components, the Transformer learns hierarchical representations of the input data. In the final stage, the output from the last Transformer layer is directed into a linear layer, culminating in the final prediction. We briefly outlines the key components of Large FMs as follows:

transformer管道。 Vaswani 等人388介绍了基于注意力的 Transformer 架构，这是开发大多数大型基础模型的基础元素。 如图3所示，处理过程的第一步是通过嵌入层将输入词转换为高维向量。 在处理过程中，注意力机制会为这些输入向量的不同片段分配不同的权重。 注意之后，对输出进行层归一化处理，确保激活的稳定和标准化。 随后，每个位置向量通过前馈网络进行转换，引入非线性，使模型能够捕捉复杂的数据模式。 通过包含这些组件的多个层，transformer可以学习输入数据的分层表示。 在最后阶段，最后一个transformer层的输出被导入线性层，最终得出预测结果。 我们简要概述大型基础模型的关键组成部分如下：

单词

1. "FMs"指"Foundation Models"，即基础模型。
2. mechanisms 机制
3. Subsequently 随后
4. undergo 经历，经受
5. a feedforward network 一个前馈神经网络
6. Through multiple layers that incorporate these components 通过包含这些组件的多个层 incorporate 使并入；包含；合并；注册成立；吸收；
7. the Transformer learns hierarchical representations of the input data transformer可以学习输入数据的分层表示 hierarchical 等级制的；等级制度的；按等级划分的
8. culminate 达到顶点；(以某种结果)告终；

Embedding. Initially, the input word is transformed into a sequence of tokens by a tokenizer. Commonly used tokenizers, such as wordpiece and byte-pair encoding, are frequently employed in this process [380]. Following tokenization, a learned embedding layer converts these tokens into a sequence of vectors. In such sequences, the order of words is essential for meaning. To address this, position encoding is incorporated into the embeddings, infusing them with positional information. This addition is critical for capturing the sequential nature of the input, ensuring that the model accurately interprets word order and context.

嵌入。 首先，输入的单词会被标记化器转换成一串token。 在这一过程中，常用的标记化器，如词片和字节对编码，经常被使用[380]。 标记化之后，学习嵌入层将这些标记转换成向量序列。 在这些序列中，词的顺序对意义至关重要。 为了解决这个问题，嵌入层中加入了位置编码，为它们注入了位置信息。 这种添加对于捕捉输入的顺序性至关重要，可确保模型准确解释词序和上下文。

单词

1. infuse 注入；输注（药物等）；使具有(某特性)；
2. sequential nature 顺序性

Attention. Attention mechanisms play a crucial role in capturing the relationships between words in a sequence. The calculation of attention can be represented as:

$$A(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

注意机制在捕捉序列中单词之间的关系方面发挥着至关重要的作用.

where Q, K, and V represent the query, key, and value, respectively; each derived by multiplying the input vector with a distinct weight matrix, and d_k denotes the dimension of these vectors. Self-attention, a specific form of attention where queries, keys, and values all originate from the same input sequence, enables the model to focus on different segments of the input for each position. In contrast, multi-head attention, a variation of self-attention, permits simultaneous attention to information from diverse representation subspaces at different positions. Other variants, such as sparse attention [36] and multi-query attention [346], are tailored for efficiency or various downstream tasks. These variants are further detailed in §3.1, §4.3.3 and §4.3.4.

其中，Q、K 和 V 分别代表查询、键和值；每个向量都是通过将输入向量与一个不同的权重矩阵相乘而得到的，d_k 表示这些向量的维度。 自注意力是一种特定形式的注意力，其中查询、键和值都源自相同的输入序列，它能使模型关注每个位置的不同输入片段。 相比之下， 多头注意力是自我注意力的一种变体，它允许在不同位置同时关注来自不同表征子空间的信息。 其他变体，如稀疏注意力和多查询注意力，则是为 提高效率或完成各种下游任务而量身定制的 。这些变体在§[3.1](https://arxiv.org/html/2401.08092v2#S3.SS1 “3.1 高效注意力‣ 3 资源高效架构‣ 资源高效 LLM 和多模态基础模型概览”)、§[4.3.3](https://arxiv.org/html/2401.08092v2#S4.SS3.SSS3 "4.3.3 键值缓存‣ 4. 3 推论算法 ‣ 4 资源高效算法 ‣ 资源高效 LLM 和多模态基础模型概览"）和§[4.3.4](https://arxiv.org/html/2401. 08092v2#S4.SS3.SSS4 “4.3.4 长语境 ‣ 4.3 推论算法 ‣ 4 资源节约型算法 ‣ 资源节约型 LLM 和多模态基础模型概览”）。

单词

1. respectively 分别，各自
2. each derived by multiplying the input vector with a distinct weight matrix 每个都是通过将输入向量与一个不同的权重矩阵相乘而得出的
3. derived 获得;取得;得到
4. d_k denotes the dimension of these vectors. d_k表示这些向量的维数。denotes 表示，标志，象征
5. a variation of self-attention 自注意力机制的 变体
6. simultaneous 同时

Encoder-decoder architecture. The standard Transformer architecture consists of two main components: an encoder and a decoder. Encoder processes the input sequence through self-attention mechanisms, allowing the model to assign varying weights to different segments of the input sequence based on their relative importance. This feature is crucial for discerning complex patterns and dependencies within the input data. In contrast, the decoder is responsible for generating the output sequence. Decoder utilizes self-attention mechanisms to understand the relationships within the generated output so far. Additionally, the decoder incorporates cross-attention mechanisms, focusing on the input sequence to extract relevant information for each token in the output sequence. This part of the architecture is autoregressive, generating tokens sequentially. The production of each token depends on the tokens generated previously, unlike the parallel processing approach of the encoder.

编码器-解码器架构 标准转换器架构由两个主要部分组成：编码器和解码器。编码器通过自我关注机制处理输入序列，允许模型根据输入序列不同片段的相对重要性为其分配不同的权重。这一功能对于识别输入数据中的复杂模式和依赖关系至关重要。相反，解码器负责生成输出序列。解码器利用自我注意机制来理解迄今为止生成的输出中的关系。此外，解码器还采用交叉注意机制，重点关注输入序列，以提取输出序列中每个标记的相关信息。架构的这一部分是自回归的，按顺序生成标记。每个标记的生成都取决于之前生成的标记，这与编码器的并行处理方法不同。