

大语言模型技术研究与应用探讨

张文强

武汉学院, 湖北 武汉 430000

DOI:10.61369/MEC.2025010014

摘 要 : 大语言模型作为当下人工智能领域一项关键的技术成果, 有规模庞大、理解能力较强、生成能力颇为突出等特性, 在自然语言处理的各个子领域当中呈现出了广泛的应用潜力。本文对大语言模型的发展现状以及所面临的主要挑战做了简要探讨, 还分析了其核心技术, 覆盖模型训练以及预处理、结构优化、生成策略和实际应用等方面。期望能给语言模型技术的研究与实践带来一定参考。

关 键 词 : 大语言模型; 自然语言处理; 深度学习; 语言生成; 模型训练

Research and Application of Large Language Model Technology

Zhang Wenqiang

Wuhan College, Wuhan, Hubei 430000

Abstract : As an important technical advancement in current artificial intelligence technology, the large language model possesses large scale, high comprehension ability, and first-rate generation capability with vast application potential in other natural language process subfields. The paper briefly elaborates on the development process and important concerns of the large language model and analyzes its main technologies in terms of model training and preprocessing, structural optimization, generation strategy, and practical application in order to provide a reference for language model technology research and application practice.

Keywords : large language model; natural language processing; deep learning; language generation; model training

引言

随着人工智能技术持续发展, 自然语言处理作为关键分支受到科学界普遍关注, 大语言模型通过深度学习以及海量资料的训练能力, 在语言理解和生成方面呈现出强大性能。近些年来, 从 BERT、GPT 到 ChatGPT 等有代表性的模型陆续推出, 促使了人机交互、内容创作、智能搜索等多个领域发生革新。相较于传统模型, 大语言模型有更强的上下文理解能力以及语义生成能力, 已然成为人工智能研究和应用的核心方向之一, 但是模型规模不断扩大也引发了训练成本高、推理效率低以及生成可信度不稳定等一系列挑战。系统探讨大语言模型的技术路径、应用场景与存在问题, 对推动其发展有着意义。本文将围绕其关键技术展开剖析, 为相关研究与实践提供理论支撑。

一、大语言模型技术的现状与挑战

(一) 技术现状

如今大语言模型技术获得了一些发展成果, 已被广泛应用于机器翻译、智能问答、文本生成、代码辅助等众多领域。有代表性的模型如 OpenAI 的 GPT 系列、Google 的 PaLM、Meta 的 LLaMA 以及国内的 ChatGLM、文心一言等, 均基于 Transformer 架构, 有多轮对话、上下文理解以及高质量生成等

能力。在训练层面, 主流大模型运用大规模语料库以及高性能计算资源, 结合无监督学习与指令微调等策略, 实现从泛化到定制的跨任务迁移能力, 开源社区的发展促使模型结构优化以及工具链完善, 比如 Hugging Face 平台提供了丰富的预训练模型和推理框架, 便于研究人员与开发者开展二次开发和应用部署^[1]。

(二) 面临挑战

尽管大语言模型发展速度较快, 然而其依然面临着众多挑战。一是模型规模不断扩大, 这给予了极高的训练成本以及部署

课题信息: 依托大学生科研创新团队: 批准号: XST202404 智空安防团队, XST202406 三维机器视觉与逆向工程检测创新团队, XST202310 人工智能研究团队, 202413634001 智影空探——轻量级共轴无人机, 202413634008S 无人机起飞空域管理系统——自动生成姿态通行令牌, 202413634009 基于 5G+北斗的智能温湿度监测装置, XCDC202404 蓝航守望者——北斗集成水质监测技术系统, XCDC202436 基于北斗卫星的近海水质智能监测养殖系统

作者简介: 张文强, 本科, 研究方向: 软件工程。

成本，对计算资源以及能源消耗提出了非常严苛的要求；二是语言模型存在“幻觉”现象，也就是说生成的内容有可能出现事实方面的错误或者逻辑不一致的情况，这会影响其在高可靠性领域的应用。另外模型输出可能包含偏见、歧视或者敏感信息，带来伦理与监管方面的风险，在数据层面，预训练语料的质量以及多样性对模型性能有着关键影响，但是当前训练数据的来源往往缺乏透明度，这影响了可控性与解释性。在推理效率方面，大模型的响应速度比较慢，很难契合边缘部署或者实时交互的应用需求。怎样在提升能力的同时降低成本、提高可信度和可控性，成为了当前研究的重点方向^[2]。

二、大语言模型的关键技术

（一）模型训练与预处理技术

大语言模型构建依靠海量语料与高效训练流程，训练阶段和预处理阶段对模型语言理解深度、生成质量及泛化能力有决定性影响，整个流程主要有数据预处理、预训练、微调 and 评估四个关键步骤，其中有技术性和策略性方面的细节优化工作。

在数据预处理阶段，语料质量把控很关键。一般大语言模型要用 PB 级甚至 EB 级文本数据，数据来源有维基百科、新闻语料、社交媒体内容、书籍、论坛及程序代码等。为保证数据多样性和代表性，研究者要对数据做清洗操作，比如剔除 HTML 标签、去除广告、过滤重复文本、纠正乱码以及屏蔽低质量段落等，清洗工作还包括识别和剔除敏感信息、非法内容及过度偏倚语句，以保障训练过程伦理性及模型生成内容社会接受度^[3]。

文本编码属于预处理的关键环节，由于自然语言有着语法规则、表达多样的特性，采用固定词表会引发词汇表爆炸以及 OOV 问题，主流模型采用子词级编码方式，例如 BPE、Word-Piece 或者 SentencePiece。这些方法可把词分解成更稳定的子单元，有效压缩输入长度、减少词表规模，并且提高处理多语言文本的通用性，例如 GPT 系列采用 BPE 编码，而 T5 和 mT5 则采用 SentencePiece，这些编码策略在模型训练效率和语义表达能力之间达成了平衡。

在训练机制方面，大语言模型主要采用两种训练范式：自回归语言模型和掩码语言模型。前者像 GPT 系列，借助预测下一个词来进行序列生成学习，适用于生成任务，后者如 BERT，依靠掩盖部分词汇并预测其内容来进行学习，擅长理解与判别类任务，当下许多模型采用统一架构来进行多任务学习，让模型有更广泛的适应性。为提升模型行为的可控性以及实际交互质量，近年来人类反馈强化学习得到广泛应用。该技术依靠引入人类打分系统或奖励模型，引导模型生成更符合人类偏好的内容，是 ChatGPT 等对话模型背后的关键机制，指令微调在模型迁移中也扮演着关键角色，能使模型理解自然语言中的“任务指令”结构^[4]。

训练进程里会涉及优化器的选择（如 AdamW）、学习率调度（如 warm-up 策略）、梯度裁剪、正则化等机制，来提高模型的收敛速度以及泛化能力。在大模型训练期间，混合精度训练、数

据并行与模型并行等工程方面的优化措施也被广泛应用，降低了硬件负载以及训练时间。

（二）模型结构与优化技术

大语言模型的核心结构大多是基于 Transformer 架构构建的，自 2017 年该架构被提出以后，因其并行计算能力较强、建模长距离依赖能力较为突出等优势，渐渐取代了传统的循环神经网络也就是 RNN 以及卷积神经网络也就是 CNN，成为构建大规模自然语言模型的主要框架。Transformer 的基本单元覆盖了自注意力机制也就是 Self-Attention、多头注意力也就是 Multi-Head Attention、前馈神经网络也就是 Feedforward Neural Network 以及残差连接和层归一化机制，这些组成部分共同发挥作用，让模型可有效地提取序列中的上下文信息，并且在深层网络中维持梯度稳定^[5]。

在结构演化这个层面上，为了可适配规模更大的数据以及参数，大语言模型正逐步朝着“深层 + 宽层”的方向不断发展，即通过增加 Transformer 的层数、每层的隐藏维度以及注意力头数来提高模型容量，以 GPT-3 为例，它拥有 1750 亿个参数、96 个 Transformer 层以及 12288 维隐藏单元。而像 GPT-4 或者 PaLM 2 这类规模更大的模型，突破了千亿参数规模，呈现出了更强的语义推理以及跨领域泛化能力。

随着模型规模呈指数形式增长，结构优化成为控制资源消耗以及提升推理效率的关键方向，其中稀疏化结构是一种典型的优化路径，凭借对注意力矩阵进行稀疏计算，像 Longformer、BigBird 等模型可有效降低序列处理复杂度，从 $O(n^2)$ 降低至 $O(n \log n)$ 甚至更低，节省了内存和计算资源，分层编码结构则用于在长文本处理中提升建模效率，适用于文档级语言建模^[6]。

为了达成高效训练与部署，参数高效的结构改进策略也得到了广泛研究，轻量化技术比如 LoRA、Adapter Tuning 和 Prompt Tuning 可以在冻结主模型参数的情况下，仅训练少量附加模块，实现快速微调与任务迁移，这些方法降低了训练成本，还提高了模型的可重用性和部署灵活性，在垂直行业应用中，轻量微调技术极大地减少了对高性能硬件的依赖。

模型压缩与蒸馏同样是优化大模型结构的关键方式，知识蒸馏借助训练一个小模型去模仿大模型的输出，使小模型在保留大部分性能的同时降低了计算开销^[7]。

（三）推理与生成技术

推理以及生成技术是大语言模型应用可落地的关键所在，直接对模型的输出质量以及响应速度产生影响，生成策略属于其中的关键部分，最基础的贪婪搜索尽管速度快，不过容易致使内容出现重复或者缺乏多样性，研究者提出了多种改进方式，例如 Beam Search、Top-k 采样以及 Top-p 采样，这些策略在准确性与多样性之间达成了更好的平衡，是当前主流大模型生成的核心技术。

在维持上下文连贯性这方面，模型一般运用固定窗口机制来处理历史输入，因为上下文长度受到限制，长文本生成的时候可能会出现信息遗漏或者重复的情况。近些年来引入了位置编码优化、KV 缓存技术以及更高效的注意力机制，以此提升长文本推理

性能。在多轮对话以及复杂问答任务当中，通过引入提示工程和对话状态建模技术，模型可理解任务意图以及语境变化，提高生成的一致性与逻辑性。

另外，推理效率是衡量大模型可用性的核心指标，为了加快部署，常用方法包含混合精度计算、图优化、模型量化以及编译加速工具比如 ONNX 和 TensorRT，这些技术可降低内存占用以及响应时延，适用于边缘设备和移动端部署。

生成内容的安全性也引起了关注，为了防止输出虚假信息或者不当言论，需要结合关键词屏蔽、人类反馈校验以及响应过滤机制，对生成结果进行实时监控与控制。推理与生成技术持续地演进，正在推动大语言模型朝着更高效、更智能、更安全的方向发展^[8]。

（四）可视化与交互技术

在大语言模型投入应用的进程里，生成内容的可视化以及交互设计，对提升用户体验以及系统可控性颇为关键，以往的语言模型大多是以纯文本形式输出，然而现代系统变得日益重视借助图形化界面、交互式操作以及可视化反馈，来强化用户理解和操作灵活性。

在可视化方面，模型生成的结果可经由富文本、颜色高亮、层级结构展示等办法，提高语义的可读性。例如，在文本摘要、情感分析、代码生成等任务当中，采用高亮关键词、边界框、语法结构图等形式，可协助用户迅速辨别模型输出的重点信息。对

于开发者和研究人员而言，还可引入注意力权重可视化工具。例如 BertViz、ExBERT，以此观察模型内部的注意机制，方便调试以及解释模型行为。

在交互方面，当下主流的大语言模型大多集成于 Web 端应用，比如 ChatGPT、文心一言等，支持多轮对话、语境保持以及指令式交互，通过把提示构造与参数可调节选项相结合，例如温度、Top-p 值，用户可灵活把控生成内容的风格、长度以及语气，另外运用 Gradio、Streamlit 等开源框架，开发者可快速搭建交互式演示界面，达成语料输入、输出预览以及响应调优的一体化^[9]。

三、结束语

大语言模型在自然语言理解、生成以及人机交互等领域呈现出优势，已成为人工智能应用里的关键支撑技术。随着模型规模持续扩大以及优化算法不断演进，其在文本生成、对话系统、辅助编程等任务中发挥着变得日益关键的作用。但是模型训练成本高、推理效率低、生成内容可信度有限这类问题仍旧有待解决。未来发展趋势会聚焦于结构精简、推理加速、可控性提高以及部署多样化等方向，达成更高效、更安全且更具普适性的语言智能系统^[10]。

参考文献

- [1] 俞扬, 陆怡昕, 杨强. 大语言模型发展现状与未来展望. 中国科学: 信息科学, 2023 53(1), 1-15.
- [2] 薛瑞尼, 张永红. 从 GPT 到 ChatGPT: 大语言模型的发展脉络与应用趋势. 计算机学报, 2023 46(5), 943-956.
- [3] 刘聪, 宋强, 刘知远. 中国大语言模型的演进与挑战: ChatGLM 的研究进展. 中文信息学报, 2023 37(3), 1-11.
- [4] 马少平, 李佳芮. 面向类人对话的预训练语言模型技术综述. 软件学报, 2022 33(8), 2802-2821.
- [5] 陈启东, 张敏, 张超. 国产大语言模型现状、关键技术与发展建议. 电子科技导报, 2023 21(6), 59-67.
- [6] 罗焕坤; 葛一峰; 刘帅. 大语言模型在数学推理中的研究进展 [J]. 计算机工程, 2024(09): 1-17.
- [7] 卡祖铭; 赵鹏; 张波; 傅晓宁. 面向大语言模型的推荐系统综述. 计算机科学, 2024(S2): 11-21.
- [8] 舒文韬; 李睿潇; 孙天祥; 黄萱菁; 邱锡鹏. 大型语言模型: 原理、实现与发展 [J]. 计算机研究与发展, 2024(02): 351-361.
- [9] 王耀祖; 李擎; 戴张杰; 徐越. 大语言模型研究现状与趋势 [J]. 工程科学学报, 2024(08): 1411-1425.
- [10] 李松柏; 李岳洋; 李佳航. 大语言模型发展现状与应用趋势. 辽宁经济, 2024(11): 92-96.