大模型与智能运维

中国通信企业协会高级专家
中国移动设计院原副院长
刘　涛

大模型概述

- **LLM (Large Language Model，大型语言模型)** 是指一种拥有数十亿参数的模型，通过对互联网上的大规模数据（10~100TB）进行训练而得到，可用于实现自然语言处理的多种应用。研究发现，当模型参数和用于训练的数据量达到一定规模时，LLM将出现**智能涌现**，具备一定的生成、理解和创作能力
- 大模型事实上就是算法、数据、算力上的有效结合。
- 大模型代表了新一代AI技术，带来了里程碑式的变化，解决了三个核心问题。一是，一个通用的算法就可以应对过去数十种场景，落地空间变得更大。二是，数据无需人工标注，大量数据可以得到自动训练，而且模型结构是统一的。三是，语音、图像、自然语言可以实现多模态融合，三者互相关联
- 模型能力不仅与模型大小有关，还与数据大小和总计算量有关。同时，预训练数据的质量对取得良好的性能起着关键作用， 因此在扩展预训练语料库时，数据收集和清洗策略是非常重要的考虑

大模型概述

- 数据规模大，通过大量的数据提高模型的泛化能力和性能
- 大规模并行计算能力，随着计算硬件的不断进步，如GPU和TPU的普及，大规模并行计算能力的提升使得训练和推理大模型成为可能
- 更“大”模型复杂性：大模型具备更深层次、更复杂的网络结构，可以捕捉更丰富的特征和关系，提高了模型的表达能力
- 训练大模型，不仅需要充足的计算资源和数据，还需要深厚经验和技能，还需要一定的耐心和定力，就像“炼丹”一样
- OpenAI在开发ChatGPT的过程中，它还请了很多高水平的专家，写了数万个各领域的问答对子，这些也是重要的数据来源

大模型概述

- 大模型两条路径发展：第一条是通用类大模型持续拓展应用领域，打造跨行业通用化的人工智能能力平台，其他行业应用借助于此平台开发；第二条有针对地开发垂直领域专业类大模型
- 多模态具身智能，是探索AGI的重要内容。
- 大模型适用于需要一定容错率的行业，大模型应用在要求100%准确的行业目前难度比较大，适用于一定的容错率而非确定性的决策工作

大模型概述

- 算力是通过对信息数据进行处理，实现目标结果输出的计算能力
- 算力主要包括三种类型，分别是基础算力、智能算力和超算算力。基础算力主要用于传统的计算应用，如计算机科学、数值计算和物理模拟等领域，智能算力是指基于GPU、FPGA、ASIC等可以加速AI计算的服务器平台提供的算法，主要用于人工智能应用，如图像识别、语音识别和自然语言处理等领域
- 根据微软官方的数据，ChatGPT从无到有消耗的算力是3640PF-days，就是说即便每秒进行千万亿次计算，也要花3640天。训练ChatGPT的庞大算力，由微软Azure的超算提供，配置至少有，285000 个 CPU，10000 个 GPU，每秒 400G 的网络。
- 其中这台超级计算机拥有28.5万个CPU核心，超过1万颗GPU（英伟达 V100 GPU）；按此规格，如果自建IDC，以英伟达A100 GPU芯片替代V100 GPU芯片，依照性能换算，大约需要3000颗A100 GPU芯片

大模型概述

- 大模型是深度学习，模型中的知识和能力是人通过精选的数据和巧妙的训练方法授予的。三条路径的融合使大模型成为当前实现人工智能的最强大手段
- 大模型本身具备类推推理（analogical reasoning）的能力，但不具备逻辑推理（logical reasoning）的能力（逻辑推理是指基于三段论的推理）
- 如何优化模型：也就是降低训练和使用成本，同时扩大可处理问题的规模。
- 如何保证模型生成内容的真实性：也就是避免幻觉
- 如何构建可信赖大模型：也就是保证模型生成结果的有用性，安全性等

大模型概述

- OpenAI公司已经吸引全球大约200万开发人员围绕GPT进行应用开发，超过92%的世界500强公司成为旗下产品用户，而旗下对话机器人ChatGPT的周活跃用户已经达到1亿人
- 北京时间11月7日凌晨2点，OpenAI在美国旧金山举行首场开发者大会。在会上，该公司将GPT大模型更新到GPT-4 Turbo版本，并进一步完善大模型开发的业务架构，包括推出吸引软件开发者的“商店”产品、推出版权盾、提供API开发助手等
- OpenAI此次推出的GPT-4 Turbo，重点突出“加量、提速、降价”三大特点。2020年6月，发布GPT-3模型； 2022年11月发布GPT-3.5 ，2023年3月，发布GPT-4

大模型概述

- GPT-4 Turbo文本解读长度从32k升级到了128k，这相当于10万个英文词汇或300页英文文本内容，每分钟的处理速率也增加了一倍。此外，新模型的输入单价将从3美分降至1美分，输出单价从6美分降至3美分
- 外部文档和数据库的截止日期更新到了 2023 年 4 月。与之相比，GPT-4 的知识库截止日期为 2021 年 9 月
- OpenAI也建立起更完善的业务架构，新推出的“GPT商店” 将提供API助手，模式和苹果公司十分相似，或也将吸引更多开发者围绕GPT模型开发应用
- GPTs，让用户们无需代码，结合自己的指令、外部知识和能力创建自定义版本的 ChatGPT
- OpenAI将旗下多款明星产品在GPT-4 Turbo中进行结合，包括文生图工具DALL·E和语音识别工具Whisper等。这意味着这个大模型不仅能“说”，还能“看”、能“听”

大模型概述

- 已经学完世界5000年以来的所有知识，一直到2023年，已经没有他不知道的任何知识
- 已经为任何人及任何公司建立垂直小模型(中国几万家垂直细分的AI公司将100%全部倒闭)
- 多模态:文字图片语音视频都完备了
- 赚钱商业模式已经完成
- 任何人都可以用gpt写代码与程序

大模型概述

- 11月13日，英伟达在国际超算大会SC23上推出新一代AI计算平台NVIDIA HGX H200
- 相比H100，H200的推理速度提升60%-90%，带宽提升1.4倍，显存容量提升1.8倍
- 对于用户来说非常重要的推理能耗，H200相比H100直接腰斩
- H200与H100完全兼容，意味着将H200添加到已有系统中不需要做任何调整
- H200将于2024年第二季度向全球系统制造商和云服务提供商供货
- 英伟达将于2024年发布新一代旗舰AI芯片B100，持续突破性能与效率的极限，性能已经望不到头了

国内大模型发展现状

- 国内正式通过《生成式人工智能服务管理暂行办法》备案，向全社会开放的大模型
- 第一批：华为、腾讯、科大讯飞星火大模型、百度文心一言、抖音云雀、百川智能百川大模型、清华智谱华章智谱清言、中科院紫东太初、商汤“商量SenseChat”、MiniMaxABAB大模型、上海人工智能实验室的书通用大模型，共11个
- 第二批：网易有道“子曰”大模型、蚂蚁集团百灵大模型、面壁智能“面壁露卡Luca”、出门问问“序列猴子”、昆仑万维“天工”大模型、美团模型、知乎“知海图AI”模型、月之暗面moonshot、金山办公WPSAI、好未来MathGPT大模型、360“奇元大模型”，共11个
- 第一批主要为通用大模型，第二批多为垂直领域的头部企业

国内大模型发展现状

- 2023年中，三大电信运营商相继发布了大模型产品
- 中国移动建立近1500人核心技术团队自主研发的“九天”1+N大模型，已发布九天·众擎基座大模型+N个行业大模型，已正式发布的行业大模型包括：九天·海算政务大模型、九天·客服大模型、九天·网络大模型、九天·企业通话大模型以及九天·川流出行大模型。医疗大模型、司法大模型、能源大模型、运输大模型、航空大模型等正在合作共建中
- 中国电信于2022年成立中国电信数字智能科技分公司聚焦大数据和AI核心能力建设。目前已发布一款大语言模型-TeleChat大模型以及一款视觉大模型-“星河”通用视觉大模型2.0。此外，其语音大模型和多模态大模型也预计在2023年底商用
- 中国联通借助外部力量快速布局多模态大模型，从公众线的增值业务场景出发，打造面向C端市场的多模态大模型-鸿湖图文大模型，鸿湖图文大模型是首个面向运营商增值业务的大模型

国内大模型发展现状

- 三大运营商发布的大模型具有相似之处，它们都采用了基座大模型与行业大模型相结合的技术路线。在内部应用方面，它们主要集中在电信客服、云网运营和安全管理领域
- 中国移动和中国电信主要以政企市场作为切入点，中国联通的大模型则以公众线增值服务为切入点
- 中国电信同时布局了大语言模型、视觉大模型、语音大模型和多模态大模型，而中国移动则主要使用大语言模型，中国联通则采用多模态大模型作为基座大模型
- 中国电信整合并成立针对AI能力建设的分公司，中国移动主要依靠内部专业公司和研究院的研发力量进行AI能力建设，而中国联通更多地借助外部力量来加快研发进程

国外运营商大模型发展现状

- 美国运营商在AI大模型领域主要扮演使用者的角色，利用OpenAI、谷歌、Meta等科技巨头的大模型优化自身业务并提高内部运营效率
- 大部分欧洲运营商虽然正在采用大模型技术，但并未推出自己的大模型产品。他们更倾向于利用现有的技术和合作伙伴关系来提供AI增值服务，有一些欧洲运营商通过合作的形式来打造行业级别的大模型
- SKT自2022年起在AI领域进行了一系列的投资并购以快速构建其AI能力

国内大模型发展现状

- 中国企业对外发布的大模型普遍是单模态的，且普遍面临GPU算力不足的挑战
- 中美在算力方面的差距，由于GPU芯片等问题，在一定程度上国内算力已被卡脖子了。即使国内头部公司，算力上的差距也是比较明显的
- 大模型训练需要处理高颗粒度的信息，对云端训练芯片的芯片处理信息的精细度和算力速度要求更高，现阶段国产GPU大多还不具备支撑大模型训练所需的能力
- 不同于多媒体和图形处理的单精度浮点计算（FP32）计算需求，在超算领域，双精度浮点计算能力FP64是进行高算力计算的硬性指标。英伟达的A100同时具备上述两类能力，而国内GPU芯片的云端训练公司，大多只能处理单精度浮点计算

国内大模型发展现状

- 云厂商购买A800的到手价通常是十万元一枚，如果以一万枚A800，搭配两万枚AMD霄龙7642构建智能计算集群，即使不计算其他硬件的成本，硬件采购的成本就高达12亿元，服务器的采购成本通常是数据中心建设成本的三分之一，数据中心的建设成本就超过了三十亿。
- 按照一枚GPU搭配两枚CPU来算，根据产品信息上的功率是700瓦，就算是没有任何能量浪费，一万枚GPU和两万枚CPU每天消耗的电量也是十六万八千千瓦时，根据中国城乡居民的平均用电量，每人每月用电26度，一户三口之家用电取80度。这个数据中心运行一天，CPU和GPU消耗的电就可以供一个三口之家用2100个月，约175年。
- 这还只是用电量。还要考虑访问量带来的网络、服务器负担。这些软性成本难以估计

网络（系统）维护的主要内容

- 网络（系统）运维包括对网络整体表现、产品运营表现、业务使用体验、对网络资源健康度进行管理、监控、分析，通过被动的监控和处理
- 或者通过对故障报警和性能劣化的主动感知分析以及自动化的资源调整实现网络、业务的恢复
- 通过售前、售中、售后的端到端支撑能力，提供贯穿于运维各项生产环节的自动化运维感知和决策信息的流转能力
- ITU-T定义了电信管理网络的五大基本域，故障（Fault）管理、配置（Configuration）管理、计费（Accounting）管理、性能（Performance）管理和安全（Security）管理简称FCAPS
- 随着SDN和NFC技术的出现，传统面向FC APS五元组管理被控制（Control），编排（Orchestral）管理、策略（Policy）分析（Analysis）。新五元组所替代简称COMPA

网络运维基本概念

- 网络拓扑管理
- 网络资源管理
- 业务配置管理
- 业务自动割接
- 业务生存保障
- 网络告警管理
- 网络性能管理
- 网络流量管理
- 网络自动巡检

网络运维面临的挑战

- 业务难感知
- 故障难定位
- 故障恢复慢
- 安全运行
- 人力紧缺

网络运维发展的阶段

- 从传统的人工维护，发展到智能维护，以后发展到智慧维护
- 智能维护开始应用人工智能技术
- 智慧维护广泛应用大模型技术

智能运维的概念和现状

- 智能运维（AIOps）将人工智能应用于运维领域，基于已有的运维数据（日志、监控信息、应用信息等），通过人工智能的方式来进一步解决自动化运维无法解决的问题。
- 不依赖于人为指定规则，由人工智能算法自动地从海量运维数据中不断地学习，不断地提炼并总结规则。
- 智能运维以系统所采集的运维大数据为基础，利用人工智能算法对运维数据进行深入分析

智能运维的概念

- 智能运维的核心价值就在于由人工智能取代人力决策，快速给出故障处理建议，或者提前规避故障
- 智能运维是以大数据平台和算法平台为核心。
- 智能运维需要从各个监控系统中抽取数据、面向用户提供服务、并有执行智能运维产生决策模型的自动化系统
- 智能运维和自动化运维有天差地别
- 智能运维作为一个新型的技术领域，目前正在快速发展，越来越多的研究者将机器学习、自然语言处理、数据挖掘等技术应用于智能运维领域

智慧运维的功能

- 网络异常智能感知
- 网络智能规划
- 故障智能溯源
- 健康分析
- 网络流量预测

智慧运维应用大模型技术优势

- 借助于大模型具有的更渊博知识和更强大的解决问题的能力，智慧运维可以以更高的精度分析解决复杂的问题
- 预测维护：使用大模型技术分析历史数据和实时监测数据，可以预测设备或系统可能出现的故障或异常，并提前采取维护措施
- 故障诊断：大模型技术可以分析设备传感器数据、日志数据等信息，识别设备故障的根本原因
- 资源优化：大模型技术可以对设备的运行状态进行实时监测和分析，以优化设备的使用和配置。以达到节能减排和资源的合理调度和利用的目的
- 风险预警：大模型技术可以对设备或系统可能出现的风险进行预警和识别

智慧运维应用大模型技术优势

- 决策支持：利用大模型技术对运维数据进行分析和挖掘，可以提供有价值的运维决策支持
- 借助大模型，可以使集成和部署效率提升 10 倍以上
- 大模型可以帮助自动生成测试用例
- 大模型可以通过学习领域知识和历史缺陷数据，辅助测试人员制定更加精准的测试计划，识别并智能检测出潜在的缺陷，减少漏测和误测的情况
- 大模型还可以根据缺陷类型和严重程度进行排序，提高缺陷修复的优先级和效率

智慧运维应用大模型技术优势

- 通过大模型智能引擎可以通过问答方式或数据交互方式快速得到网络节能优化策略以及网络能耗vs网络性能vs用户感知的分布地图
- 大模型可以较为完美的代替目前的知识百科、专家经验库、甚至知识图谱等系统及技术，用户可以自由就关心的问题提问，并得到较为准确的回答，同时可以较为高效的成为专业技能培训系统，短时间快速提升员工技能
- 大模型可以有效管理各类用户需求，并及时导入专网运营平台
- 基于大模型形成面向投诉的智能问答处理系统，识别用户投诉的语义、情感，并形成处理策略

智慧运维应用大模型技术优势

- 行业数据的优势
- 行业算力的优势
- 行业算法模型的优势
- 行业专家的优势

智慧运维与智能运维

- 智慧运维注重对设备的实时监控和数据分析，智能运维更强调对运维过程的智能化和自动化
- 两者的最大区别就是智慧运维比智能运维更加智能，更节省人力
- 智慧运维除了能够比智能运维更精准地进行分析和预测外，还可以准确地生成各种场景
- 大模型的广泛应用，是实现智慧运维的重要基础和前提条件
- 智慧运维既可以基于智能运维，也可以直接实现

智慧运维中应用大模型途径

- 使用大模型技术建立智能运维平台，采集有关海量数据，采用大模型技术，提供智慧运维服务
- 使用大模型建立智能运维工具，专注提供模块化智慧运维服务

大模型在智慧运维中应用

- 高质量数据是网络的本源，网络在运行过程中会产生大规模、复杂类型的数据，除了这些数据外，还包括通信行业数据，这些数据来自于内部收集、公共数据集和第三方数据提供商
- 资源配置数据
- 客户业务数据网络，承载的客户业务信息
- 运行状态数据
- 运维操作数据
- 来自于第三方系统，如光缆哑资源信息故障、工单信息、机房环境温度信息等
- 通信行业属性数据
- 客户服务对话数据。
- 行业标准和法规文件
- 通信技术数据

大模型在智慧运维中应用

- 对这些数据进行清洗，去除重复、冗于、不一致的数据。提取高质量的核心网络数据集
- 算法模型是挖掘网络知识，实现网络智慧的关键
- 由于从零开始训练一个大模型需要大量的数据资源和算力资源，行业大模型通常基于已有的通用大模型继续训练得到。可基于开源开放的通用大模型进行训练
- 强大算力是智慧网络的基础，由需要处理的数据量和算法决定

大模型在智慧运维中应用亟待解决的问题

- 海量数据的清洗和标注，特别是数据标准化
- 急需专门人才，以解决模型的结构设计，以及怎么训练、推理
- GPU智算能力不足
- 如何处理好与数据安全的关系

结束语

- 大模型技术将极大地促进智慧运维的发展
- 任重道远