中国科学: 生命科学 2024 年 第 54 卷 第 3 期: 482 ~ 506
SCIENTIA SINICA Vitae lifecn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS

评 述

人工智能大模型医学应用研究

郭华源[1,2], 刘盼[2], 卢若谷[2], 杨菲菲[3], 徐洪丽[2], 庄严[2], 黄高[4], 宋士吉[4], 何昆仑[2*]

1. 中国人民解放军总医院医学创新研究部医学人工智能研究中心, 北京 100853;
2. 中国人民解放军总医院医学创新研究部医学大数据研究中心, 北京 100853;
3. 中国人民解放军总医院第四医学中心心血管内科, 北京 100048;
4. 清华大学自动化系, 北京 100084
* 联系人, E-mail: kunlunhe@plagh.org

收稿日期: 2023-02-05; 接受日期: 2023-06-16; 网络版发表日期: 2024-01-22
工业和信息化部科技司2020年产业技术基础公共服务平台项目(批准号: 2020-0103-3-1, 2020-0103-3-1-8)资助

摘要 近年来, 以自然语言处理和视频图像分析为主的人工智能大模型技术得到快速发展, 其基本特征是聚焦相关应用领域的共性需求, 通过大数据、强算力和复杂算法的高效协同与深度融合, 构建通用预训练模型, 广泛适配下游任务, 有力提高模型的处理性能与研发效率. 因此, 大模型技术为医学人工智能高质量发展提供了难得契机. 本文通过全面梳理国内外大模型的研究进展、关键技术与核心算法, 分析总结生物医学领域一系列标准数据集和预训练模型的发展特点, 结合医学人工智能的研发实践, 深入剖析医学领域大模型构建的应用需求、解决思路与研发经验, 助力推动医学大模型创新发展.

关键词 医学, 人工智能, 大模型, 自然语言处理, 医学图像分析

近年来, 随着人工智能技术的不断进步, 语言表征预训练模型(bidirectional encoder representations from transformers, BERT)[1], 生成式预训练(generative pre-training, GPT)[2], 对比语言-图像预训练(contrastive language-image pre-training, CLIP)[3], DALL-E[4]等模型获得飞速发展, 并在自然语言处理(natural language processing, NLP)、机器翻译、文本图像分类与生成、视频图像理解、短视频推荐等领域得到推广应用. Google, 微软, 阿里, 华为等公司, 以及北京大学、清华大学、北京智源人工智能研究院等院校纷纷联合攻关, 集中力量研制参数量更大、通用性更广、准确性更高的超级大模型. 于是, M6[5], PanGu[6], 女娲[7], BioBART[8]等模型相继问世(图1), 引发业界对大模型技术的广泛关注和深入探索.

目前, 大模型既是人工智能技术发展历程上的一次重要突破, 又是加速人工智能赋能产业的一个全新引擎, 还是“AI+”创新模式全面落地的一项基础支撑, 已成为国内外各大新ICT企业纷纷抢占行业发展先机的一块主要阵地.

2017年以来国务院先后发布《新一代人工智能发展规划》[9]和《国家新一代人工智能标准体系建设指南》[10], 都将医疗作为其中一个重要的应用领域. 医疗卫生事业关系国计民生与经济发展, 关乎到人民生命健康和社会安全稳定, 是国家高质量发展的重要体

引用格式: 郭华源, 刘盼, 卢若谷, 等. 人工智能大模型医学应用研究. 中国科学: 生命科学, 2024, 54: 482–506
Guo H Y, Liu P, Lu R G, et al. Research on a massively large artificial intelligence model and its application in medicine (in Chinese). Sci Sin Vitae, 2024, 54: 482–506, doi: 10.1360/SSV-2022-0298

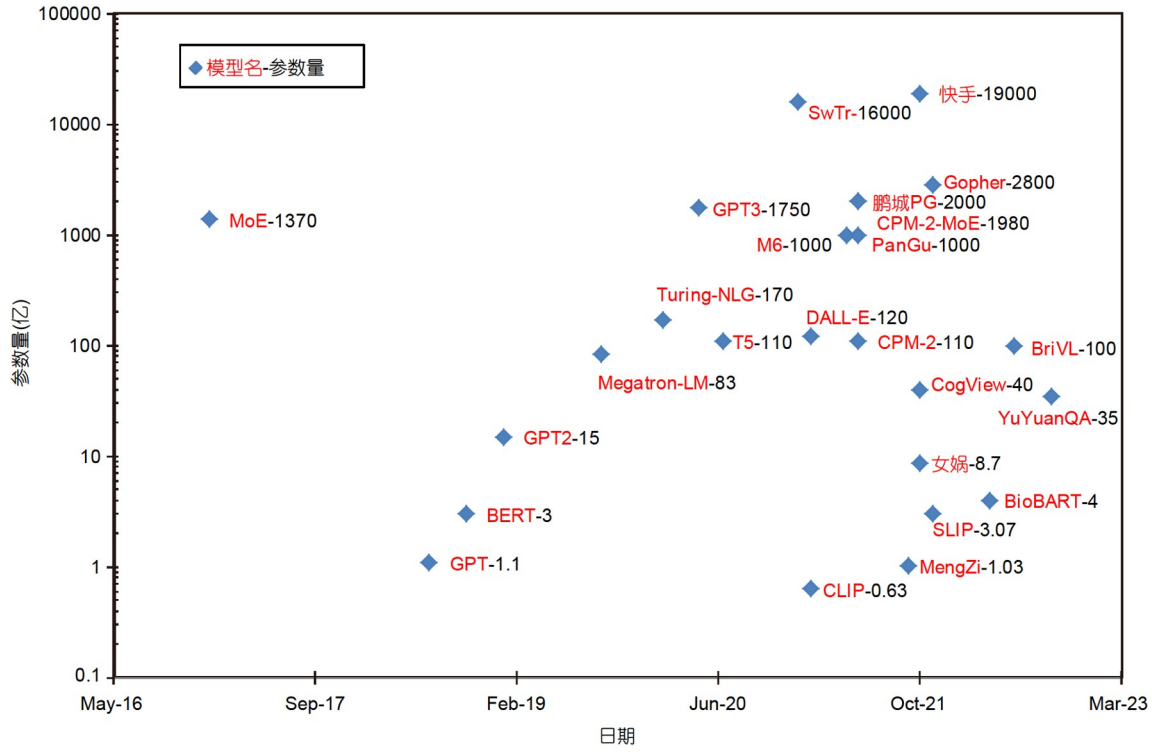

图 1　著名大模型概况

Figure 1　Overview of popular massively large AI models

现. 当前, 国内医学人工智能得到快速发展与应用推广, 但从整体上看技术研发仍处在跟跑阶段[11]. 在此情况下, 我国医学领域人工智能大模型如何规划布局, 怎样扬帆起航, 快速构建高质量创新发展的新格局, 已经成为当前一项紧迫重要的任务[12].

1　基本概念

近几年, 以自然语言处理为代表的大模型技术发展十分迅速, 模型参数量以每年1~2个数量级的速度在不断增长. 通过综合采用数据并行、模型并行以及混合并行等多种策略, 充分发挥GPU, TPU, 昇腾加速卡等不同硬件的加速性能, 模型训练规模和计算量持续递增, 模型体量与主要性能快速攀升, 使得大模型具备更强大的语义理解和更精准的计算推理等能力.

鉴于当前大模型还没有一个明确概念, 根据对现有大模型相关文献的调研与梳理, 我们认为大模型(massively large model)是一个历史的、动态的、发展的事物. 其中, BERT和GPT等早期大模型立足于自然语言处理领域, 具有数亿级参数, 具备先进性、基础性与通用性等特点, 为大模型的发展奠定基础、积累经验. 当前大模型通常是指具有百亿级以上参数规模, 致力于人工智能典型应用的巨型模型, 包括M6, GPT-3, PanGu等. 这些模型在大规模数据集上经过充分训练, 可广泛适配下游任务, 又被称为基础模型(foundation model)[13].

同时, 这类模型大都需要经过两阶段训练, 即首先在预训练阶段对大样本、广谱性数据进行预先训练(pre-training), 为后续下游任务训练出一个底层、共性的通用模型. 然后, 在第二阶段训练中加入下游任务各自领域内的小样本数据, 经进一步精调训练(fine-tuning)后获得面向各种具体场景的专用模型. 这种两阶段训练模式不仅可以提高资源使用效率和模型精准度, 而且有利于在专业领域和技术实力各不相同的多个单位之间协同推进大模型的研发与应用. 因此, 这类模型也称为预训练模型(pre-training model).

近期, 出现一类基于Zero-shot learning的通用生成式大模型[14]. 这些预训练模型采用相似度计算、目标

迁移等技术，不需要fine-tuning或者使用fine-tuning但不用下游任务领域的数据进行训练，却又能够识别仅在测试中出现的数据类别. 类似方法还包括one-shot learning, few-shot learning等[15]，这些方法使得通用大模型具有识别新事物的能力，从而进一步提高模型的适用度和可靠性.

2 主要需求

2012年以来, 随着图像识别AlexNet模型、深层卷积神经网络和高性能GPU计算技术的兴起，以深度学习为代表的人工智能技术得到飞速发展，并在计算机视觉、机器翻译、语音识别、图像处理等领域获得广泛应用, 从而引发了新一轮人工智能发展大潮. 在此过程中, 深度学习、对抗学习、半监督/弱监督学习等人工智能技术不断演化发展, 模型性能持续提高, 应用领域逐渐扩大, 人工智能进入一个新的繁荣时期, 但同时也遇到一些严峻挑战.

在理论层面，人工智能的基础理论、建模机理和系统框架亟待完善与创新[11]. “电脑-人脑”之间的性能鸿沟、功耗鸿沟和应用鸿沟愈发突显[16]，数据驱动、监督反馈、参数迭代等现有智能计算模式的性价比低、能耗过高，数据密集型、计算密集型、IO密集型处理需求失衡, 仍未有效摆脱粗放扩张的发展格局.

在技术层面，基于数据驱动的人工智能技术存在诸多问题. 例如, 深度学习的可解释性不强, 模型性能与样本数据的质量分布紧密相关，多模态数据标注、模型调参和算法优化的智能化程度较低[17]，仍需投入大量的人力与时间. 小样本学习、弱监督训练等建模技术的可靠性和泛化性亟待增强，多目标多任务模型结构复杂、训练耗时，针对跨平台、瘦客户机和移动端等复杂应用场景[18]，缺乏精简统一、协同高效的模型构建与训练技术.

在应用层面，不仅基础数据、领域知识和算子算力等模型构建的必备资源大都高度集中在行业巨头手中，而且模型训练所需的标注数据与核心算法仍未充分公开共享，极易形成新的技术壁垒，有悖于公平竞争和均衡发展.

在医学应用领域，人工智能也面临医疗数据使用难、数据标准化不高、疑难杂症处理困难、智能化应用不足等突出问题.

第一, 经过多年信息化建设, 许多医院已经积累了丰富的临床医疗数据[19]，但由于患者隐私保护及信息安全等原因，这些宝贵的数据资源难以有效共享. 而大量医疗AI公司由于缺少优质充足的临床数据，难以有效推进医疗AI模型算法的研发创新和推广应用. 第二，多源异构、多种模态的临床医疗数据在编码、格式、类型、内容上都存在较大差异，面向医疗健康AI创新应用的临床医疗数据使用亟需统一标准，这在一定程度上也桎梏了医疗AI行业的融合发展. 第三，针对“同病异影”“同影异病”以及“多病共存”等疑难复杂问题，缺乏高质量的样本数据、高性能的处理框架和高效率的模型算法. 第四，面向不同的病种、疾病、部位以及器官、组织等，已陆续研发出许多人工智能模型算法. 但如何将这些功能各异、种类繁多的医疗AI模型与临床工作紧密结合，与现有流程充分适配，与HIS(hospital information system)系统无缝衔接，实现操作便捷、交互顺畅、管理高效等目标，仍需各方持续努力. 此外，医学人工智能还面临医疗服务安全、个人隐私保护、伦理法律约束等特殊要求，也亟需妥善解决.

在此背景下，近年来一种构建大规模、高性能人工智能系统的新模式悄然出现并日渐成熟，即对大量数据进行训练产生一个基础通用模型，该模型能够广泛适配下游任务. 这种新型模型不仅深刻改变人工智能系统的构建形式、研发模式和推广方式，还将产生重大、深远、广泛的社会影响.

3 国外进展

国际上, 美国、欧洲等在大模型研究上起步早, 投入大, 成果多. 尤其是美国, 在大模型的理论研究和实践探索等方面一直走在世界前列，具有雄厚的研究基础和突出的先发优势.

第一，美国的一些大学、研究机构和科技公司率先研制了BERT和GPT等早期大模型, 抢先掌握早期大模型构建的核心原理与关键技术，同时不断投入持续推进，相继研发出Switch Transformer[20]，ChatGPT (https://openai.com/blog/chatgpt)，StyleGAN-T[21]等一批架构新颖、功能强大的新模型，展现出强大的理论创新与研发实力.

第二，Google，NVIDIA和微软等著名企业不仅聚

焦模型架构、算法结构等软件创新，而且深入研究加速器件、训练算力等硬件结构. 譬如, 基于CUDA架构的NVIDIA Tesla图形处理器GPU，基于脉动矩阵、统一缓冲与混合精度的张量处理器TPU[22]，基于尖峰频率适应可提高能效4~16倍的英特尔神经形态芯片Loihi5[23]. 此外，还积极探索基于软硬结合的建模新技术, 例如, 基于并行重叠原理的多节点参数分片(parameter sharding)[4]，以及面向并行处理的自注意力机制(self-attention)[24]等.

第三，美国还率先启动对大模型的理论、工程、应用、伦理等全方位研究布局，试图牢牢占据大模型研发、应用与推广的主阵地(https://crfm.stanford.edu/workshop.html)(表1), 具体包括五大举措.

(1) 成立专门研究机构. 2021年8月18日, 美国斯坦福大学成立基础模型研究中心(Center for Research on Foundation Models, CRFM), 汇集大学10多个系170余名研究人员, 专门研究基础模型的技术原理, 致力为未来人工智能系统的发展寻找新动力，构建新框架. CRFM尝试建立一种新的跨学科协作攻关模式，即以斯坦福大学“以人为本”人工智能研究所(Human-Centered Artificial Intelligence, HAI)为基础, 在技术研发、成果推广以及生态培育等层面持续发力，争取在基础模型的研究、开发和部署等方面取得重大进展.

(2) 全面布局研究工作. 为更好地适应人工智能研究范式的转变, CRFM研究人员将集智攻关, 重点研究未来人工智能系统的以下三方面. (ⅰ) 底层技术, 包括模型架构、训练算法、数据系统以及评估理论等; (ⅱ) 应用潜力，包括面向医疗保健、生物医学、法律和教育等领域的发展潜能; (ⅲ) 社会影响, 包括经济与环境影响, 法律与道德考虑, 以及隐私安全风险等.

另外，CRFM还将针对基础模型的训练和评估问题，研制开放共享、易于使用的开发工具，建立跨学科、跨领域的遴选规则，以便吸纳更多来自不同领域的研究者，推广应用基础模型，提出颇有价值的改进意见.

(3) 组织撰写技术论著. CRFM从多学科交叉角度, 广泛召集计算机科学、经济学、社会科学、法律、哲学等领域的专家学者, 一起研究和构建基础模型, 并以“基础模型的机遇和风险”为题撰写技术报告[13]，全面分析基础模型的核心功能、主要应用、关键技术及其深远影响.

(4) 召集举办学术研讨. CRFM在2021年8月23~24日举办“基础模型研讨会”，来自学术界和工业界具有不同学科背景的专业人员分别从不同层面与角度分析基础模型的机遇、挑战等问题, 提出各自见解.

(5) 预测未来发展趋势. 此次研讨会还大胆预测未来的人工智能系统将会愈发倚重基础模型，积极倡导制定更为严格的规则指南，加强规范基础模型的开发、部署等工作，同时归纳总结出基础模型的十大问题.

第四, 在生物医学领域, 美国IBM公司和德克萨斯大学在2011年采用机器学习、自然语言处理等技术联合研制出智能问答系统Watson[39]，努力使机器能够理解人类的自然语言，实现既能答题又能看病等功能, 从而引起全世界极大关注. 近年美国又联合英国、韩国等国将ELMo[40]和BERT等上下文词嵌入模型应用到医疗临床语料处理，不断提高临床自然语言处理性能[41]，并发布一系列面向生物医学与临床文本处理的基准数据集[42](表2)和预训练模型(表3).

其中, 用于文本处理的ClinicalBERT, 用于文本挖掘的BioBERT，用于知识蒸馏的BioNLP，用于医学问答、实体识别与关系抽取的BioMegatron以及用于医学理解、医学知识检索与推理应用的PaLM等模型, 多为BERT或GPT两类经典模型的变种. 它们在继承经典模型优点的同时，还采用生物医学语料库和基准数据集进行特定领域的预训练，有力提升模型的端到端关系抽取、基于上下文的医学问答、基于文本信息的文档分类以及基于Prompt的优化精调等性能.

第五, 在医学大模型的构建中, 医学文献、电子病历、知识图谱和专家经验等都是重要的数据源，尤其电子病历(electronic health record, EHR)具有内含丰富、数量庞大、私密度高等特点，是不可或缺的一类数据资源. 为提高对EHR的分析处理，先后出现将单词、字符、上下文等不同语言单元转换为对应向量进行统一计算的处理方法. 其中，图嵌入(graph embedding)技术颇具代表性，通过将高维稠密的图结构数据映射为低维稠密的向量表示，有效表征图的拓扑结构、顶点关系等相关信息，从而较好解决图数据难以高效输入机器学习算法的问题. 图嵌入可进一步细分为无监督、有监督两类方法[55].

其中，无监督方法既可基于随机遍历生成的路径来预测图中邻居(DeepWalk[56], node2vec[57]), 又可通过

表 1 著名大模型一览表

Table 1 Comparison list of popular massively large artificial intelligence models

名称	发表时间	参数量(亿)	训练数据量	语种	研制单位	主要架构	训练策略	支撑硬件	应用领域
MoE[25]	2017.1	1370	新闻语料库1000亿词	英语	Google	基于稀疏门控的专家层混合体, 由多达数千个前馈子网络组成	引入稀疏门控专家混合层, 每一个样本都有一个可训练门控网络计算稀疏组合	16~32 Tesla K40 GPUs	机器翻译 多语言翻译
GPT[2]	2018.7	1.1	语料库7000本书	英语	OpenAI	基于Transformer的特征抽取器, 保留有标注的多头注意力	两阶段训练, 分别采用无监督预训练语言模型和有监督的下游任务fine-tuning	GPU	文本生成
BERT[1]	2018.10	1~3	开源语料+维基百科33亿词	英语	Google	多层双向转换编码器+双向自注意力机制	使用特征提取器Transformer和语言模型实现预训练的深度双向表示	TPU	语言表征 文本分类
GPT-2[26]	2019.1	15	Web Text 40 GB	英语 法语	OpenAI	由多层单向Transformer的解码器构成一种自回归模型	以对数线性方式进行多任务小样本学习	GPU	机器问答 阅读理解 文本生成
Megatron-LM[27]	2019.9	83	聚集语料174 GB	英语	NVIDIA	PyTorch Transformer+模型并行+数据并行	采用简单高效的层内模型并行方法, 并在原生PyTorch中插入通信指令	GPU V100 (512块)	自然语言处理
Turing-NLG[28]	2020.2	170	文本	英语	微软	基于Transformer的生成语言模型, 通过生产单词完成开放式文本任务	基于NVIDIA DGX-2和Infini-Band实现快速通信, 降低模型并行复杂度	GPU V100 (256块)	编写故事 生成答案 文本摘要
BioBERT[29]	2020.4	–	生物医学180亿词	英语	高丽大学	与BERT结构基本相同, 多层双向转换编码器+双向自注意力机制	基于NSML架构, 进行“BERT通用领域语料+BioBERT生物医学语料”双层预训练	GPU V100 (8块)	实体识别 关系抽取 知识问答
GPT-3[30]	2020.5	1750	文本570 GB	英语	OpenAI	采用可替代密集和局部带状稀疏注意力机制, 构建自回归语言模型	基于小样本学习策略, 在每个矩阵相乘的内部以及不同网络层之间实现混合并行	GPU集群V100	文本生成 机器问答
Oscar[31]	2020.7	–	图像文本对650万条	英语	微软 华盛顿大学	基于图像文本表示的多层Transformer建立视觉语言预处理模型	提取图像静态对象的标记信息, 进行图像-文本对的表征学习	GPU Cluster	视觉语言理解 与生成等
T5[32]	2020.7	110	文本750 GB	英语	Google	为NLP预训练提供一个文本到文本的通用框架	基于模型并行和数据并行, 构建无监督/有监督的文本生成预训练任务	Cloud TPU (1024块)	语言理解任务
Switch Transformer[20]	2021.1	16000	爬虫语料库	英语	Google	基于稀疏路由结构, 每轮迭代触发部分Expert计算	采用稀疏激活模型有利于TPU计算, 简化MoE路由算法	TPU	自然语言处理 机器翻译
DALL-E[4]	2021.2	120	文本图像对330万	英语	OpenAI	采用dVAE编解码+16位混合精度架构	采用两阶段训练, 离散变分自编码+自回归转换器	GPU V100 (1024块)	文本图像生成
CLIP[3]	2021.2	0.63	文本图像对4亿	英语	OpenAI	通过预训练图像编码和文本编码, 建立零样本分类器	利用对比式训练、自监督学习等技术减少人工标注, 将图像分类转为图文匹配	GPU V100 (592块)	文本图像分类
M6[5]	2021.5	1000	1.9 TB图像+292 GB文本	中文	阿里 清华大学	基于多GPU架构, 在每个MoE层嵌入多个Expert以充分使用内存	基于自关注+大规模预训练, 建立多模态、多任务、百万级Transformer模型	GPU A100 (128块)	图像文本转换 多模态文本转换

(表1续1)

名称	发表时间	参数量(亿)	训练数据量	语种	研制单位	主要架构	训练策略	支撑硬件	应用领域
PanGu-α[6]	2021.6	1000	文本语料库100 GB	中文	华为	采用基于Transformer的自回归语言模型	支持数据并行、操作级模型并行、流水线模型并行、优化器模型并行和重栅格化	Ascend 910 (1024块)	文本摘要 自动问答 对话生成等
鹏城PanGu-α	2021.6	2000	文本语料库1.1 TB	中文	华为	采用基于Transformer的自回归语言模型	支持数据并行、操作级模型并行、流水线模型并行、优化器模型并行和重栅格化	Ascend 910 (2048块)	知识问答 知识检索 知识推理
CPM-2[33]	2021.6	110	悟道语料2.3 TB	中文	清华大学	引入知识继承, 利用现有预训练语言模型加速训练	将自注意力和前馈层分离, 使用prompt tuning减存特定的任务参数	GPU	语言理解 文本生成 机器翻译
CPM-2-MoE[33]	2021.6	1980	悟道语料2.3 TB 300 GB	中文 英文	清华大学	引入中英文知识继承, 利用预训练语言模型加速训练	基于模型并行性分离自注意力和前馈层, 使用prompt tuning减存任务参数	GPU	语言理解 文本生成 机器翻译
MengZi[34]	2021.10	1.03	维基百科、新闻和爬网300 GB	中文	上海交通大学等	基于RoBERTa的轻量级预训练模型	通过大模型蒸馏和小模型初始化, 突出多模态、辨别力、生成性与领域特定性	Tesla 3090 Ti (32块)	语言理解 文本生成 视觉应用
CogView[35]	2021.11	40	文本图像对3千万	中文	清华大学	基于自回归模型, VQ-VAE和Transformers相结合的文本图像转换框架	采用PB-Relaxation和Sandw- ich-LN机制	GPU	文本图像生成
快手推荐模型	2021.11	19000	WebText+视频	中英文	快手	基于数据、特征、网络结构以及单列用户交互等特点, 设计多任务学习新框架	基于MMoE的多目标预估优化算法, 提取视频类、上下文及用户历史行为等特征	CPU+GPU	短视频推荐
女娲[7]	2021.11	8.7	语音+图像+视频	英语	微软 北京大学	基于邻近稀疏注意机制, 综合利用空间-时间轴的邻近特性	基于3D Transformer的通用编解码, 进行自回归视觉合成预训练	GPU A100 (64块)	多模态图像合成 视觉合成
SLIP[36]	2021.12	3.07	YFCC15M +CC3M +CC12M	英语	Facebook	采用无分类标注的语言监督和图像自监督, 联合学习视觉特征	将图像自监督学习与CLIP多任务学习相结合	GPU V100 (64块)	视觉识别任务
Gopher[37]	2021.12	2800	多源文本库10.5 TB	英语	Google	采用自回归Transformer 架构, RMSNorm和相对位置编码机制	使用JAX的pmap转换, 高效表示数据和模型并行性	TPU v3	阅读理解 事实核查 语言识别
BioBART[8]	2022.4	4	生物医学预料41 GB	英语	清华大学 IDEA	面向生物医学领域, 采用双向自回归Transformer构建生成式语言模型	基于DeepSpeed架构, 采用领域自适应策略, 训练生成式语言预训练模型	GPU A100 (16块)	医学语言生成 对话、总结 命名实体识别
悟道文澜BriVL[38]	2022.6	百亿级	图像+文本6.5亿	中英文	中国人民大学	针对弱语义相关数据特点, 采用自监督“视觉-语言”桥接技术构建多模态基础模型	采用双塔结构对图像和文本输入进行独立编码, 并基于DeepSpeed架构进行训练	GPU A100 (112块)	文本图像生成 遥感场景分类 中文新闻分类

表 2 著名生物医学基准数据集

Table 2 Popular biomedical benchmark datasets

名称	中文全称	主要内容
MedQA[43]	医学问题自由选择数据集	从专业医学委员会考试中收集医学问题, 涵盖三种语言: 英语、简体中文和繁体中文
MedMCQA[44]	大规模多选择问答数据集	收集美国医学入学考试问题, 超过19.4万个高质量AIIMS和NEET PG入学考试MCQ, 涵盖2400个医疗主题和21个医学主题
PubMedQA[45]	生物医学问答数据集	是第一个从PubMed摘要中收集的QA数据集, 需要对生物医学研究文本进行推理, 以及根据其定量内容回答问题
MedicationQA[46]	药物问题解答标准语料库	回答消费者关于药物的健康问题, 由674个问答对组成, 并标注问题焦点、类型, 以及答案来源等
MMLU[47]	多任务文本模型准确性测试	涵盖57项任务, 包括基础数学、美国历史、计算机科学、法律等, 要求模型必须具有广泛的世界知识和解决问题的能力
MultiMedQA[48]	开放式医学问答数据集	结合六个现有的开放式医学问答数据集, 涵盖专业医学考试、科学研究和消费者查询等内容
HealthSearchQA[48]	在线医疗问题搜索与响应数据集	构建一个基于真实性、准确性、可能的危害与偏见等多维模型答案的人类评估框架
MIMIC-III[49]	重症监护医学信息数据库	包括人口统计学信息、实验室检验信息、患者用药信息、护理相关信息、患者检查成像报告、患者每次出入院信息等

表 3 重要生物医学预训练模型

Table 3 List of important biomedical pretraining models

名称	发表时间	主要功能
ClinicalBERT[41]	2019.7	研制面向临床的BERT模型, 用于一般临床文本和出院总结的自然语言文本分析, 该特定领域模型可提高NLP处理性能
BioBERT[29]	2020.4	基于大规模生物医学语料库, 进行预先训练的特定领域语言表示模型, 使用NLP技术从生物医学文献中提取有价值信息
BioNLP[50]	2020.11	针对18个生物医学和临床NLP任务, 从开源生物医学和临床NLP模型中, 挑选适合不同环境的模型, 并分析各种预训练设计对模型下游性能的影响
BioMegatron[51]	2020.11	实证评估影响领域语言应用性能的因素, 包括子词词汇集、模型大小、预训练语料库和领域迁移等, 提高在生物医学领域进行预先训练的语言模型性能
ScholarBERT[52]	2022.5	将14个基于Transformer的模型应用于11项科学任务, 评估训练数据、模型大小、预训练时间、微调长度等因素如何影响下游性能, 是迄今规模最大、种类最多的科学语言模型
BioGPT[53]	2022.10	由于BERT和GPT模型(含变种)在下游生物医学任务中缺乏生成能力, 现对大规模生物医学文献进行预先训练, 研制特定领域生成式Transformer语言模型
PaLM[48]	2022.12	通过调整模型量表和指令提示, 提高理解、知识回忆和医学推理等能力, 突显评估框架和开发方法对大规模语言模型(PaLM及变体Flan-PaLM, Med-PaLM)的重要支撑
ChatDoctor[54]	2023.4	通过使用在线医疗咨询网站10万余次真实世界的医患对话数据, 对大语言模型进行微调, 显著提升模型在患者需求理解、知情建议提供以及用户响应准确性等方面性能

保持图中节点的一阶、二阶(甚至更高阶)邻域拓扑结构, 最大程度保留网络结构信息(LINE[58]). 有监督图嵌入算法则是通过利用图神经网络, 不仅可保留图结构信息, 而且能利用额外的边信息保持节点/边的属性, 包括图神经网络GraphSAGE[59]以及图注意力网络GAT[60]等.

但是研究表明, 上述方法应用到EHR数据时仍有不足, 其原因一是这些方法原为通用的图应用而开发, 并不能很好适配医学图应用的特征, 二是所有医疗服务都有时间戳属性, 而这些方法却缺乏处理时间序列的有效机制. 因此, Choi等人提出Med2Vec[61]和graph convolutional transformer(GCT)[62]等方法, 可在每次处理中通过诊断代码与治疗代码的显式或隐式分类, 来学习和构建医疗预测任务中的多级医疗嵌入特征, 提

高处理性能. 只是这类方法没有充分考虑个人医疗服务所特有的时域特征, 无法解决患者旅程中遇到的非规则时间间隔问题. 于是, Wu等人[55]提出一个基于图的分层医疗实体嵌入框架ME2Vec, 即先采用分层结构依次嵌入医疗服务、医生、患者三类实体信息, 然后利用EHR数据特点, 构造最适合每种特定实体的嵌入机制.

在此基础上, 为了在数据有限的情况下也能使用深度学习进行EHR数据预测建模, Rasmy等人[63]基于上下文嵌入技术和结构化数据, 构建一个大型预训练语言模型Med-BERT, 使用小型本地EHR训练数据集进行疾病预测. 同时, Hong等人[64]基于稀疏嵌入回归(knowledge extraction via sparse embedding regression, KESER)算法, 开发知识提取系统支持特征选择和综合网络分析, 通过多中心大规模的代码嵌入, 识别感兴趣疾病的相关特征, 从而有效表达多种临床知识, 推动基于多中心的EHR应用研究进程.

4　国内现状

近几年, 国内相关研究团队在大模型的样本数据处理、模型算法设计和算力资源构建等关键领域加大投入, 奋起直追. 并且, 通过紧扣重点行业的应用需求, 不断加强基础创新, 持续推进核心技术的原创性、通用性与本土化, 产出一批具有重要影响的大模型成果. M6, PanGu和女娲等模型相继被研制出来(表1), 在信息检索、商品推介、智能问答、药物研发等行业应用中显露锋芒.

4.1　大模型构建

针对大模型的研发, 国内整体呈现出良好的发展态势, 共同推动数据预处理、大模型构建、算法优化与算力集成等关键技术取得全栈式发展、协同化进步. 主要表现如下.

第一, 训练样本多源异构、巨量分散. 譬如M6[25], PanGu[6], 悟道文澜BriVL[38], 封神榜[65]等大模型分别收集整理数百GB乃至数个TB以上的图像、文本、网页等信息, 用于模型训练、验证与测试.

当前, 人工智能模型逐渐从以数据驱动、深度学习、监督机制等为主要特征的阶段演进到以“数据和知识”双轮驱动、生成式、弱监督为标志的新阶段, 模型优异的可靠性、精准的识别率和良好的泛化性都需要足量、优质的样本数据为基本前提. 因此, 这些多模态、跨尺度、宽领域的样本数据为模型性能的大幅跃升奠定了坚实的数据基础.

第二, 训练算力异构虚拟、分布并行. 目前, 大模型研究主要针对自然语言理解、图像融合分析、文本图像生成等问题求解, 数据量大, 复杂度高. 同时, 大模型训练需要进行高维空间搜索、大规模矩阵运算和深层次迭代优化, 对算力资源需求很大.

因此, 大模型训练普遍基于分布式、并行化的计算架构, 结合采用模型并行、数据并行等策略构建大算力集群系统, 不断增强大模型的训练支撑实力. 同时, 深入挖掘模型架构与算力优化之间的巨大潜力, 努力为模型训推提供高性能、可扩展的算力支撑[66].

第三, 算法结构高效精巧、融合创新. 大模型研发起源于常规模型, 聚焦于基础性、普适性与协同化的高精度数值模型的构建与优化. 一方面通过大批量增设独立参数, 缩短迭代步长, 加大隐含层深, 扩大搜索空间, 提高并行占比, 以及优化模型结构等措施, 大幅提高模型参数规模和应用性能[67].

另一方面, 基于系统工程视角研制新的加速器件、算法框架和模型结构, 譬如基于达芬奇架构(Da-Vinci)的AI芯片, 基于MoE的高性能推理框架IN-FMoE[33], 以及基于Transformer的文本分析和语义处理模型等, 努力从软硬结合、存算一体[17]、云网融合等角度提升模型精度和算法性能.

第四, 基础研究方兴未艾、成绩显著. 国内一些研究团队对脑科学[68]、认知计算[69]等相关基础理论[70]与前沿技术[34]展开深入研究, 取得重要进展. 例如, 通过构建一个两阶段学习模式, 探索医疗知识的准确表示, 然后基于先易后难、先局部后全局的决策模式, 实现面向诊断的复杂推理, 使得“智医助理”机器人能以高分通过国家医师资格考试[71]. 此外, 基于联邦学习和隐私保护等技术, 在去中心化的用户数据上联合训练图神经网络, 实现模型准确性、隐私保护与通信成本之间的平衡, 并应用于智慧医疗等场景[72].

第五, 语言模型不断涌现, 应用广泛. 2021年11月, 粤港澳大湾区数字经济研究院(International Digital Economy Academy, IDEA)陆续推出“封神榜”系列开源大模型, 包括用于自然语言理解的“二郎神”, 解决通用任务的“燃灯”以及面向医疗领域的“余元”等模

型[65]. 其中, 具有13亿参数的“二郎神-1.3B”大模型在中文语言理解权威评测基准FewCLUE上登顶, 在CHID(成语填空)、TNEWS(新闻分类)两项上表现超过人类; 拥有35亿参数的“余元-3.5B”大模型在医疗事实判断上的准确率接近90%; 而具有7.7亿参数的“燃灯-770M”大模型能够很好地完成自然语言生成和理解任务.

2022年11月以来, 随着ChatGPT模型的出现, 国内一些研究机构和公司陆续推出多个大语言模型(表4), 涵盖语言、文学、金融、法律、医疗等多个应用领域.

这类语言模型普遍使用人类输入提示(prompt)、人类反馈强化学习(rein-forcement learning from human feedback, RLHF)以及基于思维链(chain of thought, CoT)的应用推理等技术, 不断提升模型学习、理解与推理等能力[48]. 其中, 提示工程允许用户通过输入一组提示用语, 有效引导和控制语言模型的输出, 使其生成符合特定需求的文本结果. 而人类反馈强化学习技术, 则是基于人类反馈机制进行学习, 使系统能更好地遵循指令和提供帮助. 思维链技术通过模仿人类推理过程, 先将问题分解为一系列推理步骤, 然后对多个链条进行采样, 并通过投票机制得到最终答案, 从而有效提高大语言模型在复杂情况中的推理性能.

此外, 针对不同大模型在多学科中文知识理解、推理评测等方面需求, Huang等人[73]首次提出一个大模型高级知识和推理能力评估套件C-EVAL, 共包括52个学科的中国文化背景, 并对GPT-4, ChatGPT, Claude, LLaMA, Moss等9个国内外大模型在中文学科问题上的性能进行评估. 结果显示, 在所有参评大模型中, 只有GPT-4的平均准确率超过60%(为68.7%), 而国产模型中表现最好的MiniMax模型准确率只有49%, 存在较大提升空间.

同时, 为便于对17个聊天模型进行提问检测, 我国研究人员研制一款名为“齐叨”(ChatALL)的应用软件(https://github.com/sunner/ChatALL), 支持中英德三种语言, 主要功能包括快速提问、对话信息留存, 自动保持连线, 以及多列视图便捷切换等, 并提供网页访问、API访问等可选模式, 有力推动语言模型的融合发展.

4.2 医学信息处理

近年来, 为提高对中文医学信息的识别、理解与应用等能力, 国内围绕医学实体标注、中文医学知识图谱构建、中文语言理解评价基准及评测数据集的研制, 以及面向生物医学的大语言模型研究等方面开展技术攻关.

第一, 医学实体标注规范与知识图谱构建. 目前, 医学共享语料库仍很稀缺, 面向通用场景的大规模医

表 4 近期国内发布的重要语言模型一览表

Table 4 Summary of important language models recently released in China

主研单位	模型名称	发布时间	访问链接
元语智能	Chatyuan	2023-02-07	https://www.clueai.cn/
复旦大学	Moss	2023-02-21	https://moss.fastnlp.top/
澜舟科技	孟子	2023-03-14	https://www.langboat.com/potal/Mengzi-model
百度	文心一言	2023-03-16	https://yiyan.baidu.com/
清华大学	ChatGLB-6B	2023-03-28	https://github.com/THUDM/chatGLM-6B
阿里巴巴	通义千问	2023-04-07	https://tongyi.aliyun.com/
360	360智脑	2023-04-10	https://www.360dmodel.com/
科大讯飞	讯飞星火认知	2023-05-06	https://xinghuo.xfyun.cn/
–	中文法律知识	2023-05-13	https://github.com/pengxiao-song/LaWGPT/
度小满	轩辕金融大模型	2023-05-19	https://github.com/Duxiaoman-DI/XuanYuan
天津超算中心	天河天元	2023-05-20	https://www.nscc-tj.cn/index/
面壁智能等	中文语言模型	2023-05-23	https://github.com/thunlp/WebCPM
云知声	山海大模型	2023-05-24	https://shanhai.unisound.com

学文本标注数据也非常缺乏，但利用自然语言处理技术进行医学信息处理的需求却日益增长. 因此，张欢等人[74]参考UML定义的语义类型, 提出面向医学文本信息处理的医学实体标注规范，涵盖疾病、临床表现、医疗病程等11种医学实体，并基于规范构建了医学实体标注语料库. 随后，中文医学知识图谱CMeKG[75]，CPubMed-KG，OMAHA领域知识库等系统陆续出现，有力提升了中文医学信息处理的标准化和智能化水准.

第二，中文语言理解评估基准及评测数据集研制. GLUE和SuperGLUE等自然语言理解基准系统，为语言模型的性能评估提供了统一框架，极大促进了英文自然语言处理的深入研究与广泛应用. 受此启发，Xu等人[76]提出一个汉语理解评估基准CLUE. 这是一个面向中文文本信息处理的开放式、统一驱动的评价系统, 共汇集了单句分类、句对分类以及机器阅读等9项中文语言理解任务. 随后，出现FewCLUE[77]评测基准以及CBLUE[78], PMC-Patients[79]等数据集.

其中，FewCLUE是CLUE推出的一项中文小样本学习评测基准，用于评估模型是否能通过极少样本的学习来掌握特定的自然语言处理技能. CBLUE是开源的中文医疗信息处理评测基准数据集，由医学文本信息抽取、医学术语标准化、医学文本分类、医学句子语义关系判定、医学对话理解与生成等5类14个子任务组成. PMC-Patients则是一个大规模英文病历摘要及关系标注数据集，通过从PubMed Central中直接提取病历摘要，并以引入文章的metadata和文章间引用关系作为数据标注, 共包括16.7万余例病历摘要.

第三，中文预训练语言理解模型研究. 截至目前，已出现PCL-MedBERT，MC-BERT，eHealth[80]，HuaTuo[81]等多个中文生物医学预训练语言模型. 其中，PCL-MedBERT和MC-BERT是中文医疗领域首批模型, 在医学和通用领域使用的效果不是很明显. 而作为一个基于多层次文本辨析构建的中文生物医学语言模型——eHealth模型在CBLUE的11项医学NLP任务中，包括CMeEE(中文医学命名实体识别)、CMeIE(中文医学文本实体关系抽取)、CDN(临床术语标准化任务)、STS(疾病问答迁移学习)、QIC(医疗搜索检索词-意图分类)、QTR(医疗搜索查询词-页面标题相关性)、QQR(医疗搜索查询词-查询词相关性)等处理效果优于PCL-MedBERT和MC-BERT模型. eHealth模型采用大量生物医疗数据预训练，由生成和判别两部分结构组成，并在ELECTRA基础上把判别模型细分为token和sequence两个层面，不依赖外部资源，便于模型精调.

HuaTuo模型——是采用开源LLaMA[82]基础模型, 通过整合中文医学知识图谱CMeKG的结构化和非结构化医学知识, 并利用基于知识的指令数据且以QA问答形式进行精调，使模型具有较丰富的医学领域专业知识, 从而提高智能诊断的专业化与准确性.

此外, 智能预咨询(intelligent pre-consultation, IPC)是一种部署在移动终端上，用于提前收集咨询患者信息的新型应用. 由于大多数患者缺乏医学背景，往往倾向于使用口语来描述他们的症状. 因此，Zhang等人[83]将症状检测表述为一个检索问题，提出一种双向硬负强制噪声对比估计方法(bi-hardNCE)，与常用检索模型相比, 该方法显著提高了症状检测性能.

上述技术交织互补, 协同创新, 共同推动国内大模型算法有效支撑TB级以上样本数据的训练测试，百亿级以上模型参数的迭代优化以及多样化、宽领域、全场景的应用适配与融合集成，全力释放人工智能对科学进步、经济转型和社会发展的巨大潜力.

5　模型构建

大模型构建涉及多模态数据处理、大规模分布并行训练、模型参数调优、跨平台模型推理应用以及伦理法律规范和个人隐私保护等多个方面，研究人员提出一系列以图计算为主要特点的新方法，包括分布式图数据库[84]、流式图计算[85]以及可解释图网络[86]等技术. 通过把数据抽象成图, 方便复杂关系的关联、还原和可视化处理，使得图计算技术在对多源、多维、异构数据进行深度下钻、关联、归因等分析时具有比传统关系型数据库更佳的性能[87].

同时，由表1分析可知，以自然语言处理、文本图像生成等应用为代表的大模型已经从学习任务特定的表示和设计任务特定的体系结构，逐渐转变为使用任务无关的预训练和任务无关的体系结构，用于适配大规模、全场景、宽领域的应用需求. 并且，这种变革伴随着Transformer, MoE(mixture of experts), MMoE(multi-gate mixture of experts), 以及并行分布式训练、模型压缩加速等建模优化技术的快速发展.

5.1 转换器模型

转换器(Transformer)是Google在2017年提出用于自然语言处理的一种经典神经网络结构[24]. 与传统卷积神经网络(convolutional neural network, CNN)和循环神经网络(recurrent neural network, RNN)网络不同, Transformer网络结构主要由Encoder, Decoder以及全连接部件Attention三部分组成, 通过采用自注意力机制并行训练网络, 不仅便于模型学习全局信息, 而且利于提高训练速度. 同时, Transformer使用多头(头部个数常为8)注意力机制, 每头参数均独立不共享. 这种多头注意力机制善于感知不同关注点处的区域信息, 可以明显增强模型对不同位置的信息获取和特征识别能力.

如图2所示, 不管是双向式的BERT模型, 生成式的GPT模型, 还是自回归的DALL-E模型等, 都是基于Transformer块进行功能结构扩展. 在这些模型中, Transformer块的迭代计算通常都可转化成大规模矩阵的并行处理, 因而可使用GPU等器件进行硬件加速, 如表5所示.

目前, Transformer技术还在快速发展演化, 有两点动向特别值得关注.

第一, 通过交叉融合, 扩展应用范围. 由于注意力机制能够有效捕获文本数据的全局信息, 可以弥补CNN因感受野小而未能有效提取图像全局特征的缺陷, 因而出现一些基于CNN与Transformer相结合的视觉Transformer新模型ViT[88], VAN[89]等, 用于计算机视觉和图像处理领域. 同样, 针对Transformer的内存瓶颈和二次方扩展等不足, 出现一种基于RNN和Transformer的新架构RWKV[90], 能够弥补神经网络架构在计算效率和表达能力之间的差距, 从而为构建下一代更可持续、计算效率更高的序列处理模型铺平道路. 此外, 将BERT模型的双向编码器与GPT模型自左至右单向解码器相结合, 构建一种基于Transformer结构、序列到序列的去噪自编码预训练BART模型[91], 经模型微调可有效提高文本生成与理解能力, 并随之出现适用生物医学领域的生成式语言模型BioBART[8]等.

第二, 改进模型结构, 提升处理性能. 在经典Transformer模型中, 基于自注意力的token mixer对模型贡献突出. 但是, PoolFormer, MetaFormer等新模型通过使用简单池化操作替代attention模块实现最基本的token mixing功能, 可以在多个视觉处理任务中表现更为优异[92]. 同样, 视觉友好的转换器Visformer在ImageNet分类精度方面, 优于基于Transformer和基于卷积的模型, 尤其当模型复杂度较低或训练集较小时, 优势变得更加显著[93]. 此外, 针对视频和文本信息的高效处理, VideoFormer, TextFormer和BridgeFormer等模型采

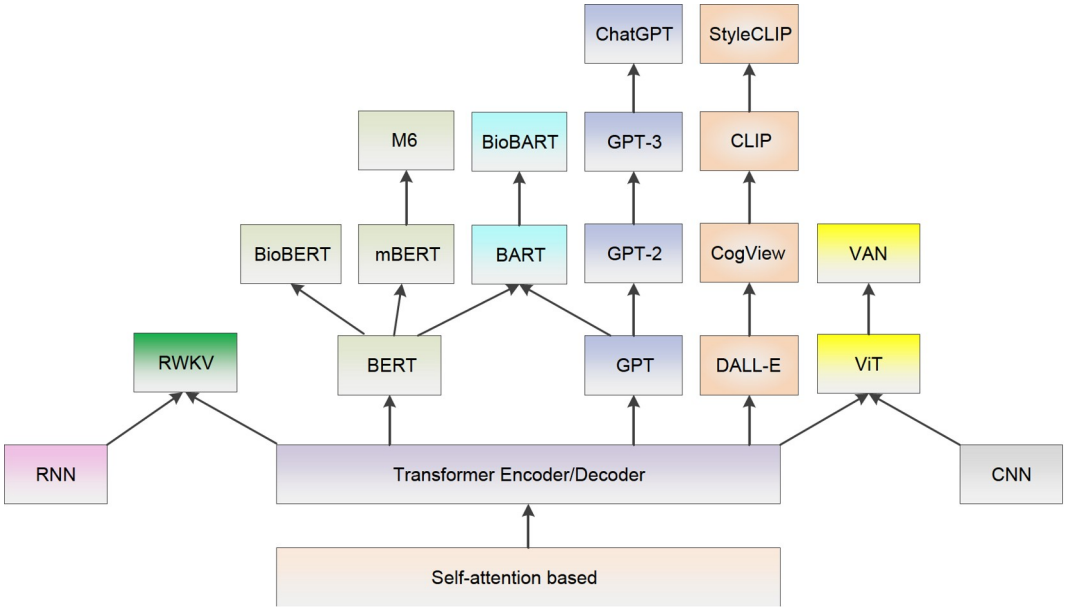

图 2 基于Transformer结构的常见模型分类

Figure 2 Model classification diagram based on a Transformer structure

表 5　部分大模型的结构参数

Table 5　Structural parameters of some massively large artificial intelligence models

大模型名称	模型层数	隐含状态数	多头注意力	参数总量(亿)	主要特点
CLIP	12	512	8	0.63	Transformer采用小写的字节对编码(BPE)表示
MengZi	12	768	12	1.03	采用轻量化的Transformer
GPT	12	768	12	1.1	采用有标注自注意力机制的解码Transformer
BERT-BASE	12	768	12	1.1	采用双向Transformer
BERT-LARGE	24	1024	16	3.4	采用双向Transformer
GPT-2	48	1600	32	15	采用多层单向Transformer(解码端), 多任务预训练
CogView	48	2560	40	40	采用单向Transformer
Megatron-LM	72	3072	32	83	基于PyTorch Transformer的神经语言模型
DALL-E	64	3968	62	120	采用自回归Transformer
T5	48	1024	128	110	使用标准Encoder-Decoder Transformer模型
CPM-2	24	10240	64	110	基于双向编码器、单向解码器的Transformer架构
M6	24	1024	16	1000	由自关注Transformer和逐点式前馈网络(FFN)组成
CPM-2-MoE	24	10240	64	1980	基于双向编码器、单向解码器的Transformer架构

用多层双向Transformer编码器原理以及多头注意力机制, 能够更有效地利用语言模态, 从更少的视频信息中获得更强的视频表示[94].

5.2　专家混合范式

专家混合范式(mixture of experts, MoE)是一种将多个模型整合到一个单独任务中的神经网络, 由多个专家(expert)模块组成, 每个专家都由一个简单的前馈神经网络和一个可训练的选通网络构成, 并由一个总的门控模块用于选择专家组合来处理输入数据, 详见文献[25].

首先, 与CNN和RNN等常规网络不同, MoE能够使用数据分离策略训练出多个模型, 并采用各种线性或非线性函数, 将多个模型整合到一个单独任务中, 实现多任务、多目标学习, 被称为“One Model To Learn Them All”.

其次, 通过简化MoE模型的路由机制, 基于Transformer MoE网络结构的Switch Transformers模型可在降低计算量的同时大幅提升模型性能[20]. 同时, 以MoE模型为基础衍生出一些扩展模型, 可进一步提高多任务学习和推理效果. 例如, 通过综合考虑数据、特征、embedding、网络结构以及单列用户交互等多项特点, 快手针对个性化推荐应用提出一个用于多目标预估优化的MMoE多任务学习框架[95], 致力于提高多任务学习的性能.

此外, 针对MoE模型推理应用, 通过设计新颖的MoE架构和模型压缩算法, 微软提出一个端到端的MoE训练与推理系统DeepSpeed MoE[96], 其中Pyramid-MoE通过在浅层设置少量experts, 而在深层使用更多experts, 形成Pyramid形状来减少参数量. 其次, Residual-MoE则利用其已填experts对首个expert有“纠错”能力的特点, 采用直连和选跳策略构建一种类似ResNet的新MoE结构, 提高CPU内存与GPU内存的协同工作效率, 大幅降低内存占用、处理延迟和整体成本.

5.3　专用训练框架

2014年以来, Tensorflow, PyTorch, Caffe, CNTK等一批深度学习框架陆续开源, 极大推动了人工智能技术的应用与发展. 但是, 该领域快速增长的模型复杂度、参数海量化和任务多样性等趋势, 给算法开发带来一些新挑战, 使得这些通用型学习框架难以满足新需求. 譬如, 更灵活的学习网络设计, 更精准的模型参数设置, 分布式环境下的快速计算, 以及不同硬件、不同平台之间算法框架的兼容性和移植性等.

因此, 一些研究团队采用“并行加速+结构优化”等策略, 为大模型训练研制专用的算法框架, 进一步提升软硬件资源的兼容性与高效性, 详见表6.

表 6 大模型训练的主要框架

Table 6 Key frameworks for massively large model training

框架名称	发表时间	主研单位	功能特征	应用案例
Mesh-Tensorflow[97]	2018.11	Google	基于数据并行策略, 建立一种分布式张量计算的通用框架, 将数据并行视为沿batch维度进行拆分的张量操作	Switch-Transformer[20] T5[32]
Megatron-LM[27]	2019.9	NVIDIA	基于模型并行策略, 建立一种PyTorch框架, 用于训练巨型语言Transformer模型	GPT-2[26] BERT[1]
OneFlow[67]	2020.7	一流科技	基于SBP抽象和actor模型的一种新型异构分布式训练框架, 降低数据并行、模型并行、多机多卡等编程难度, 实现高效管理	OneBrain[67] OneAgent[67]
DeepSpeed[98]	2020.8	Microsoft	基于零优化内核、分布式培训和混合精度等技术, 小幅更改PyTorch模型就可实现模型训练的并行加速	Turing-NLG[28], MengZi[34] BioBART[8]
MindSpore	2020.9	华为	五大演进: 从深度学习框架到张量可微计算, 从手工并行到自动并行, 从图算分离到联合优化, 从端云分离到端云统一, 从消费级AI到企业级AI	PanGu[6] 鹏城PanGu
NNL[99]	2021.2	SONY	一种分布式开源神经网络库, 包含深度学习 Python API 与嵌入式C++ API, 支持静态和动态两种计算图模式, 兼容台式电脑、HPC集群, 嵌入式等应用	NNABLA[99]
Whale[100]	2021.8	阿里	基于Tensorflow的统一分布式训练框架, 支持多种并行无缝衔接, 在计算效率、通信效率、显存消耗等方面深度优化, 使框架高效易用	M6[5]
VISSL	2021.10	Facebook	基于PyTorch的自监督学习库, 具有可扩展和模块化等特点, 加快自监督学习	SLIP[36]
Persia[95]	2021.11	快手	通过优化算法和分布式系统架构的同步设计, 提出一种混合训练新算法, 其嵌入层和密集神经网络由各自同步机制处理	快手推荐

5.4 网络模型优化

为了提高大规模神经网络模型的学习效率，减少训练耗时，优化算法结构，通常可从框架、数据和结构等不同层面推进模型算法的优化工作.

第一，框架层面是指采用分布式、并行化等技术加速模型训练. 并行化处理是大规模机器学习的一项关键技术，具体包括将数据分块计算的数据并行，将任务分解处理的任务并行，将模型拆分训练的模型并行，以及基于多种并行方式交叉组合的混合并行等技术[5]. 并行分布式计算框架的多样性给大模型训练带来很多机遇和挑战，需综合考虑数据结构、网络拓扑[101]、内存优化[102]以及设备分布[103]等因素，对现有算法进行并行化改造[104].

第二，数据层面是在采集更多样本数据用于模型训练的同时，更需提高训练样本质量，通过增强数据的代表性与多样化提升模型的精准度、可靠性和泛化性，避免出现数据依赖、分布同质化和信息茧房等问题[105]. 同时，要妥善处理可能因数据规模增大导致的维度灾难、梯度消失、梯度爆炸、模式崩塌以及模式丢弃等异常情况. 此外, 还积极研究混合精度、参数量化等技术[106]，有实验发现使用Int8量化推理会大幅减少Transformer模型的内存占用，却不会降低模型的预测性能[107], 从而有力推动大模型在轻量级硬件平台上的部署使用.

第三，结构层面就是一方面通过模型压缩、网络剪枝、知识蒸馏等技术[108]，不断精简模型结构，减少不必要的计算，及时剪除参与度低、影响性小的分支结构[109]；另一方面引入动态网络[110]、可变形注意力[111]、视觉Transformer[112]以及视觉注意力[89]等机制, 增强模型对多场景、多目标和多任务的适用性, 努力降低计算压力、存储空间和传输负载，提高训练质量、模型性能和整体效率, 详见表7.

5.5 工作流程制定

医学大模型涉及大数据预处理、大算力建设与复杂算法构建等技术，需要医疗、信息、大数据、人工智能等相关专业团队通力协作与集智攻关，加快构建关键技术体系，统筹制定研发工作流程(图3)，进一步

表 7　模型优化的主要技术

Table 7　Key technologies for model optimization

分类	名称	特征
结构精简	模型压缩[113,114]	通过修剪模型分支结构, 降低模型复杂度, 提高模型准确率, 减少模型计算、存储与传输压力, 加速模型训练与推理, 实现模型的轻量、精简与高效, 包括前端、后端等多种压缩方法
	网络剪枝[115]	通过剔除权重矩阵相对“不重要”的权值, 根据尺度因子修剪特征图通道, 以及剪除幅值过小的梯度等, 重新微调网络结构, 减少模型参数, 降低网络复杂度, 增加泛化性, 加速模型部署
	知识蒸馏[116]	是一种前端压缩方法, 基于教师网络(复杂、预测精度高)的相关指标构造总体损失函数, 诱导学生网络(精简、适合推理)通过迭代训练实现知识迁移、联合训练与模型精简
	参数量化[117]	将神经网络的浮点计算转换为定点, 支持手机、平板等资源受限设备的实时应用, 主要策略包括低比特量化、总体训练加速量化和分布式训练梯度量化等, 最大限度减小模型对计算空间和时间的消耗
	低秩近似[118]	秩就是矩阵行或者列向量极大无关组中所含向量的个数, 通过低秩分解、SVD奇异值分解等方法将权重矩阵稀疏化, 以此减小模型计算和存储的开销
功能扩展	动态网络[110]	与传统静态网络相比较, 动态网络可根据不同样本调节神经网络结构与参数, 包括动态深度、动态宽度和超网络动态路由等, 进一步提高模型运算效率、特征表达和自适应推理能力
	可变形注意力[111]	通过数据相关方式选择自注意力中键值对的位置, 使自注意力模块专注于相关区域, 捕获更多信息特征, 这种具有可变形注意力的通用主干网络模型适用于图像分类和密集预测等任务
	视觉Transformer[112]	利用屏蔽语言建模能够将文本分词为多个有意义语义片段的特性, 建立一个基于在线分词器执行屏蔽预测的自监督训练框架, 实现在线分词器的图像BERT预训练, 提高检测精度和鲁棒性
	视觉注意力[89]	针对视觉任务需求, 结合卷积和自注意力的优点, 包括局部结构信息、长期依赖性和适应性等, 构建一种自适应和长程相关的大核注意力网络, 在图像分类、目标检测、语义分割等应用中表现优异

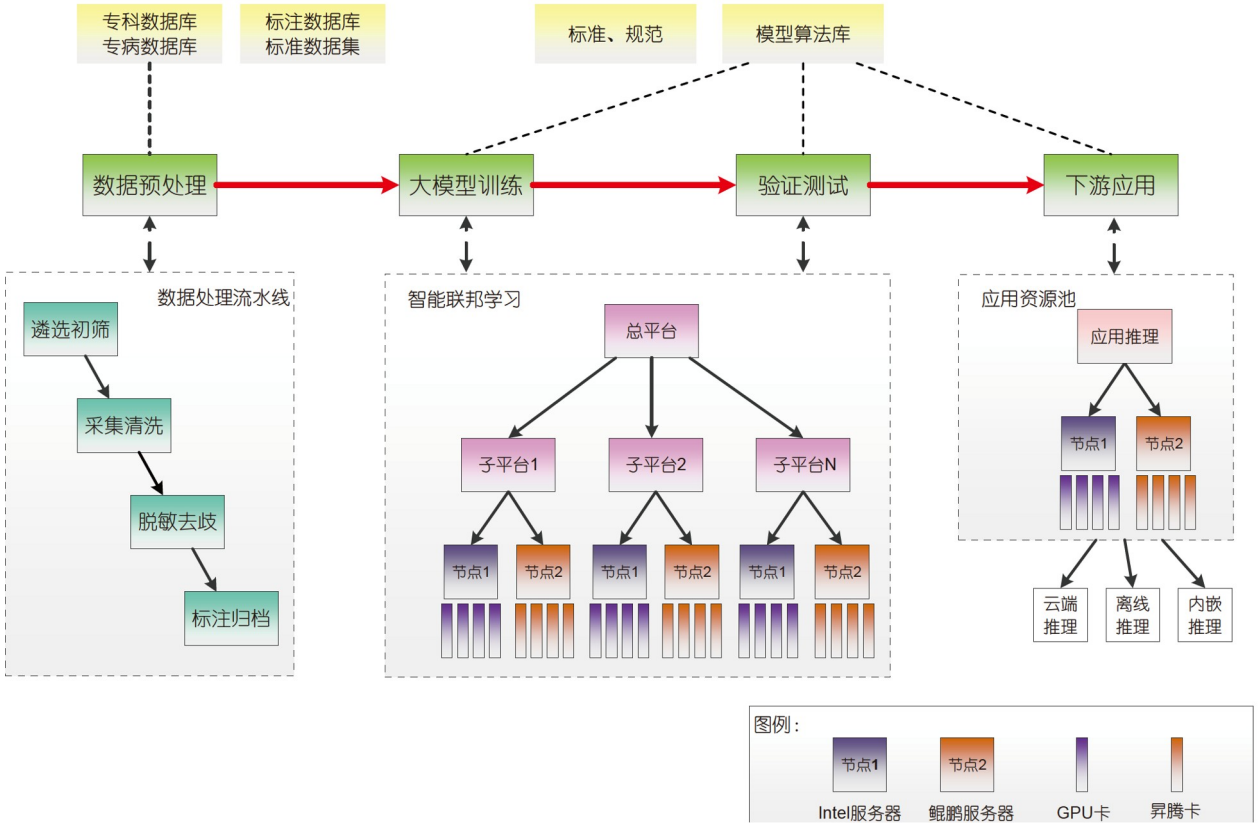

图 3　医学大模型研发工作流程

Figure 3　Massively large artificial intelligence medical model research and development workflow

提高团队编组、协同攻关的质量与效率.

其中, 大数据预处理需要对多源异构分散的医疗数据统一进行遴选初筛、采集清洗、脱敏匿名、标注归档与编码压缩等治理, 确保用于医学大模型训练、测试与验证的数据安全高效、合规可信、可管可控. 参考国际通行的预训练基准数据集(表2), 结合使用图数据库[119]、联邦学习[120]、自然语言处理等技术, 构建面向医学大模型研制的电子病历数据库、专科专病数据库、标注数据集等[121].

大算力建设是指坚持自主可控、异构兼容等原则, 通过搭建高性能、并行化、分布式的大规模集群系统或者是大容量、虚拟化、边云协同的云平台, 高效集成图形处理器GPU、张量处理器TPU以及昇腾处理器等异构硬件加速卡, 为医学大模型研究提供集中统一的大算力资源视图和虚拟透明的交互式应用平台[66].

复杂算法构建就是瞄准医学大模型的应用场景、功能特点和关键指标等需求, 一方面抓紧研究联邦学习架构、预训练机制[122]等, 针对小样本、自监督、多目标、多任务等学习算法特性, 不断研制结构精巧、性能优良、稳定高效的创新算法[123]. 另一方面在模型结构、计算平台等要素基本保持不变, 通过精准识别和有效消除通信、内存与计算等方面的性能瓶颈[124], 高效使用CPU, GPU或TPU等不同计算内核, 降低资源消耗, 减少运算耗时[125], 扩大适用范围[34], 提高医学大模型的训练质量、运行效率和生态效益.

5.6 伦理法律规范

目前, 以深度学习、神经网络为代表的人工智能技术存在模型可解释性不强, 优化调控效率不高, 人工智能生成内容(AI generated content, AIGC)不托底, 生成式模型结果不可控等问题, 甚至还有可能触犯伦理底线和法律红线.

因此, 在医学领域开展相关科学研究时, 通常需要先进行项目伦理审查工作(表8), 以确保项目研究遵守伦理规范, 恪守法律准则.

2023年4月11日, 国家互联网信息办公室起草的《生成式人工智能服务管理办法(征求意见稿)》向社会公开征求意见[126]. 征求意见稿对生成式人工智能服务提出了明确要求, 包括“遵守法律法规的要求, 尊重社会公德、公序良俗”“承担该产品生成内容生产者的责任”“承担个人信息处理者的法定责任, 履行个人信息保护义务”“对生成式人工智能产品的预训练数据、优化训练数据来源的合法性负责”“利用生成式人工智能产品向公众提供服务前, 应当按照《具有舆论属性或社会动员能力的互联网信息服务安全评估规定》向国家网信部门申报安全评估”等. 根据征求意见稿, 违反规定的服务提供者, 如果构成犯罪, 可依法追究刑事责任.

2023年5月16日, 世界卫生组织公开呼吁谨慎使用人工智能生成的大型语言模型工具. 世界卫生组织强调, 虽然借助ChatGPT, BARD这类工具能方便地获取健康信息, 甚至可在资源不足的地区提高诊断能力, 但使用时必须认真分析其风险(包括输入信息不实、模型响应错误、数据未经授权、敏感数据保护、语言模型滥用、生成虚假信息等), 否则可能导致医护人员出错, 给患者带来伤害, 降低人们对人工智能的期盼, 从而影响人工智能未来的整体发展.

5.7 个人隐私保护

在医学大模型的训练、验证与测试中, 需要使用大量医疗数据, 不可避免地涉及患者隐私和敏感信息. 根据《个人信息保护法》《电子病历基本规范》等法规要求, 需要在数据预处理阶段对所有样本数据中涉及的敏感和隐私信息进行脱敏、匿名化等处理, 确保患者合法权益. 因此, 可将个人隐私保护范畴划分为两类: 直接隐私保护和间接隐私保护.

表 8 伦理审查相关依据

Table 8 Laws and regulations related to ethical reviews

序号	部门/组织	文件名称	备注
1	国家卫生健康委、教育部、科技部、中医药局	《涉及人的生命科学和医学研究伦理审查办法》	2023年2月实施
2	世界医学协会	《赫尔辛基宣言》	1964年6月通过
3	世界卫生组织	《生物医学研究审查伦理委员会操作指南》	2000年发布
4	国际医学科学组织委员会	《涉及人的健康相关研究国际伦理准则》	2016年发布

第一，直接隐私保护是指数据中直接包含有个人隐私的相关信息，需要对这些显性隐私信息进行去标识和匿名化等处理. 我国《个人信息保护法》明确规定: (ⅰ) 立法目的是为了保护个人信息权益，规范个人信息处理活动, 促进个人信息合理利用; (ⅱ) 个人信息是以电子或者其他方式记录的与已识别或者可识别的自然人有关的各种信息，不包括匿名化处理后的信息; (ⅲ) 敏感个人信息是一旦泄露或者非法使用，容易导致自然人的人格尊严受到侵害或者人身、财产安全受到危害的个人信息，包括生物识别、宗教信仰、特定身份、医疗健康、金融账户、行踪轨迹等信息, 以及不满十四周岁未成年人的个人信息; 等等.

并且, 国家标准《信息安全技术 个人信息安全规范》(GB/T 35273—2020)进一步明确: (ⅰ) “个人信息”具体包括姓名、出生日期、身份证件号码、个人生物识别信息、住址、通信通讯联系方式、通信记录和内容、账号密码、财产信息、征信信息、行踪轨迹、住宿信息、健康生理信息、交易信息等; (ⅱ) 匿名化: 通过对个人信息的技术处理，使得个人信息主体无法被识别或者关联，且处理后的信息不能被复原的过程. (ⅲ) 去标识化: 通过对个人信息的技术处理，使其在不借助额外信息的情况下，无法识别或者关联个人信息主体的过程.

第二，间接隐私保护是指数据不直接含有显性的个人隐私信息，但是通过对多条相关数据进行关联分析与融合计算, 可从中获得个人隐私信息. 譬如, 近年来发展较快的数据画像技术，就存在较大的间接隐私保护风险.

因此, 为做好间接隐私保护工作, 也需从多个层面采取防范措施: (ⅰ) 根据《个人信息保护法》在收集个人信息时要遵循“最小范围”原则，在处理个人信息时应遵循“最小影响”原则，不能随意扩大数据收集和处理范围; (ⅱ) 在进行数据画像处理之前，要全面分析研判画像可能产生的隐私保护风险，做到未雨绸缪, 防患未然; (ⅲ) 研究数据画像调控技术，通过噪声增强、随机采样等策略准确控制和有效降低画像结果中敏感信息的表示精度, 从而能够更好地保护个人隐私.

6　研发实践

解放军总医院作为全国数字化建设示范医院之一, 经过数十年持续的信息化建设, 已基本建立功能完整、数据共享与业务协同的医疗信息系统，积累了种类繁多、质量优异、数量庞大的临床数据资源，为医学人工智能研发奠定了坚实基础，也提供了广泛的“AI+”应用场景和优越的适配环境[19].

为此，项目组围绕医学影像AI筛查、辅助诊断与临床决策支持等应用开展了深入研究，先后申请授权20余项发明专利，编写4部国家/行业标准[127~130]，取得一批较高水平研究成果. 例如，基于3D肺CT图像开发一种DenseNet-264模型COVIDNet, 实现对新冠肺炎病灶的快速识别, 准确率为94.3%[131]; 利用深度学习提取心脏超声彩色多普勒信息, 自动识别心脏瓣膜病, 测量关键参数进行病程分级[132]; 开发一种基于深度学习的自动分类管理框架，能够准确模拟患者的疾病进展风险和健康经济评估状况，通过揭示疾病进展与医疗花费之间的关系，帮助患者在经济条件许可范围内进行更为有效的治疗, 获得更好疗效[133].

同时，解放军总医院牵头承担以工业和信息化部科技司2020产业技术基础公共服务平台“面向医疗健康行业的人工智能筛查和辅助诊断公共服务平台”为代表的一批重大项目，针对心肺等脏器、脑、眼、血管等疾病，建立新一代人工智能筛查和辅助诊断关键技术体系，搭建面向医疗、体检、养老等基层机构的人工智能筛查和辅助诊断公共服务平台，探索发展人工智能产业新模式.

通过上述研发实践，我们深刻认识到在医学人工智能领域，针对复杂场景的医学AI建模训练与推理应用在准确性、可靠性、泛化性等方面存在许多困难和挑战, 亟待理论创新与技术进步[134]. 因此, 在医疗健康行业面向人工智能筛查、辅助诊断和临床决策支持等重要需求, 研制医学大模型将具有突出的应用价值, 重要的科学意义和深远的社会影响. 同时, 需要在政策支撑、协同创新、技术攻关、组织实施等关键要素上进一步做好铺垫衔接, 持续使能赋能, 推动医学大模型的研发应用不断实现新跨越, 取得新突破.

6.1　加强政策支撑

医疗健康大数据是国家重要的基础性战略资源. 研制医学大模型，不仅有力推动医疗健康大数据的创新发展，推进医疗健康服务模式的深刻变革，激发深化医药卫生体制改革的动力和活力，提升医疗健康服

务效率和质量, 而且有利于培育医疗健康人工智能的新生态、新业态和经济增长点.

近年来, 针对健康管理、医疗服务、日常保健、疫情防控等重大需求, 国家已密集出台多项政策(表9), 大力倡导、积极推进对医疗健康大数据进行安全有序、高质高效的开发应用.

并且, 为有效保护患者隐私, 保证医疗安全, 政府发布一系列法律法规和标准规定(分别如表10和11所示), 明确大数据使用的红线和底线, 同时也为医学大模型的研究与应用保驾护航.

此外, 针对跨网跨域、安全防护、伦理审查和转化推广等突出问题, 还需进一步加强政策扶持与技术指导, 为医学大模型的研究探索和先行先试提供不可或缺的制度保证, 保障研究工作顺利开展.

6.2 推进协同创新

当前, 医疗AI标准数据集共享协议、网信安全核心技术、相关行业监管规范、产业交叉融合机制等相关工作仍亟待加强. 在医疗机构内部、各医疗机构之间以及医疗机构与AI企业之间等仍然难以有效充分共享AI数据、算力、算法、模型等关键性资源. 同时, 在医疗AI领域中, 不仅“云大物移”新ICT技术尚未得到充分应用与示范, 而且针对基础芯片、先进模型、算法算力以及云边端架构的原创性探索与系统化研究等工作也需抓紧推进.

因此, 在研发医学大模型时, 需要大力构建“产学研用”于一体, “理工医信”相结合的协同创新机制, 尽快消除“信息孤岛”“服务壁垒”等顽疾, 医疗机构、高等院校、AI企业和主管部门等需各司其职, 紧密协作, 共同发力, 突破行业壁垒, 消除服务盲区, 战胜困难挑战, 推动领域创新.

当前, 推进医学大模型的研发主要有两大策略, 如表12所示. 其中, 迁移训练和原生训练各具特色, 互为补充, 可分别适用于不同应用场景.

(1) 迁移训练. 以自然语言处理、视频图像分析等领域现有通用大模型为基础, 瞄准应用场景的具体需求, 采集整理医学相关领域的多模态数据, 包括电子病历、医学影像、临床检验等信息, 对现有大模型进行再训练, 生成针对特定医学领域新的大模型, 并面向医疗、保健、科研等领域进行持续赋能.

(2) 原生训练. 指充分借鉴和吸收现有大模型研发精髓, 对语义理解、文本生成、图文互转、图像分类、机器翻译等技术进行不断创新, 大量采集整理医学及其相关领域的文本、图像、波形、视频等多模态数据, 持续推进高性能算子、大规模算力以及多任务算法等要素的深度融合, 统一构建医学领域基础大模型.

6.3 聚力技术攻关

2022年11月以来, 国内外相关机构已经陆续发布多个重要大语言模型, 为医学大模型的研究奠定了坚实基础和有益参考, 分别如表3和4所示. 因此, 采用迁移训练方法, 在功能先进、开源共享的大语言模型(包括LLaMA[82], CPM[33], ChatGLM-6B[135], VisualGLM-6B[135]等)基础上, 加入精心挑选的医疗领域数据进行模型预训练, 从而生成医学大模型底座系统, 正成为

表 9 健康大数据相关政策

Table 9 Policies related to big data for health

序号	主管部门	文件名称	备注
1	国务院	《关于印发促进大数据发展行动纲要的通知》	国发〔2015〕50号
2	国务院	《关于促进和规范健康医疗大数据应用发展的指导意见》	国办发〔2016〕47号
3	国家卫生健康委员会	《国家健康医疗大数据标准、安全和服务管理办法(试行)》	2018-07-12施行
4	国家卫生健康委员会	《关于落实卫生健康行业网络信息与数据安全责任的通知》	国卫办规划函〔2019〕8号
5	中央网络安全和信息化委员会	《关于做好个人信息保护利用大数据支撑联防联控工作的通知》	2020-02-04下发
6	国务院	《关于推动公立医院高质量发展的意见》	国办发〔2021〕18号
7	中共中央 国务院	《关于构建数据基础制度更好发挥数据要素作用的意见》	2022-12-02下发
8	中共中央 国务院	《数字中国建设整体布局规划》	2023-02-27发布

表 10　法律法规

Table 10　Relevant laws to security and privacy

法律名称	施行日期
《中华人民共和国国家安全法》	2015-07-01
《中华人民共和国网络安全法》	2017-06-01
《中华人民共和国数据安全法》	2021-09-01
《中华人民共和国个人信息保护法》	2021-11-01

一条切实可行的研发路径. 例如, 在医疗人工智能应用系统研发中, 临床电子病历是一类不可或缺的基础数据, 有着广泛的应用场景. 为了促进电子病历数据的融合共享与赋能反哺, 可基于迁移训练模式构建一个电子病历大模型, 构建流程如图4所示.

首先, 通过全面梳理电子病历大模型的应用需求和关键指标, 分析出训练样本的覆盖范围与数据规模, 初步设计大模型训练算法的核心框架和主要功能.

然后, 根据上述要求采集清洗多源、异构、分散的电子病历数据(包括结构化的检验数据、门诊处方, 半结构化的检查数据、麻醉记录, 非结构化的文本数据、波形影像等), 以及相关医学文献等资料, 分门别类地统一各种数据的语义、格式与编码, 按照相关要求进行数据标注.

同时, 针对不同结构化程度的电子病历数据, 分别采用擦除、替换、过滤等技术进行脱敏、匿名化处理, 并加强相关人员的法规学习和督导落实, 建立个人隐私、信息安全和医学伦理的保护机制.

最后, 将全部数据输入到基础大模型, 依托大算力资源进行大规模预训练. 最后输出大模型训练结果, 并结合下游多种应用场景分别进行AI赋能. 同时, 还可有针对性地收集整理下游应用的各种反馈信息, 形成反馈强化学习闭环, 促进大模型持续健康发展.

在研制医学大模型的技术攻关中, 不仅要突破一批平台底座、联网协议、参数调优、软硬协同和安全防护的核心技术[136], 研究一系列有关医学大模型构建、训练、验证、测试与应用评价的新理论、新框架和新算法[137]; 而且需要制定医学人工智能相关的国家标准、行业规范和应用指南, 组织撰写医学大模型研发、应用与推广的理论专著和技术手册.

表 11　标准规定

Table 11　Relevant standards and regulations

编码	名称	实施日期
GB/T 22239—2019	《信息安全技术 网络安全等级保护基本要求》	2019-12-01
GB/T 25058—2019	《信息安全技术 网络安全等级保护实施指南》	2020-03-01
GB/T 22240—2020	《信息安全技术 网络安全等级保护定级指南》	2020-11-01
GB/T 35273—2020	《信息安全技术 个人信息安全规范》	2020-10-01
GB/T 39725—2020	《信息安全技术 健康医疗数据安全指南》	2021-07-01
GJB 7561—2012	《军队办公网络及设备安全防护通用要求》	2012-09-01
–	《中国人民解放军计算机网络安全保密规定》	2012-12-01
–	《军队互联网媒体管理规定》	2018-02-01

表 12　两种医学大模型研发策略的对比

Table 12　Comparison of two development strategies for massively large medical artificial intelligence models

对比内容	迁移训练	原生训练
数据需求	较大	很大
算力资源	较多	很多
模型算法	在原有基础上升级、改造、扩展	构建完整、创新的模型算法
技术要求	较高	很高
研发投入	较大	很大
团队协作	需求很大	需求很大
应用结合	较紧密	较宽广

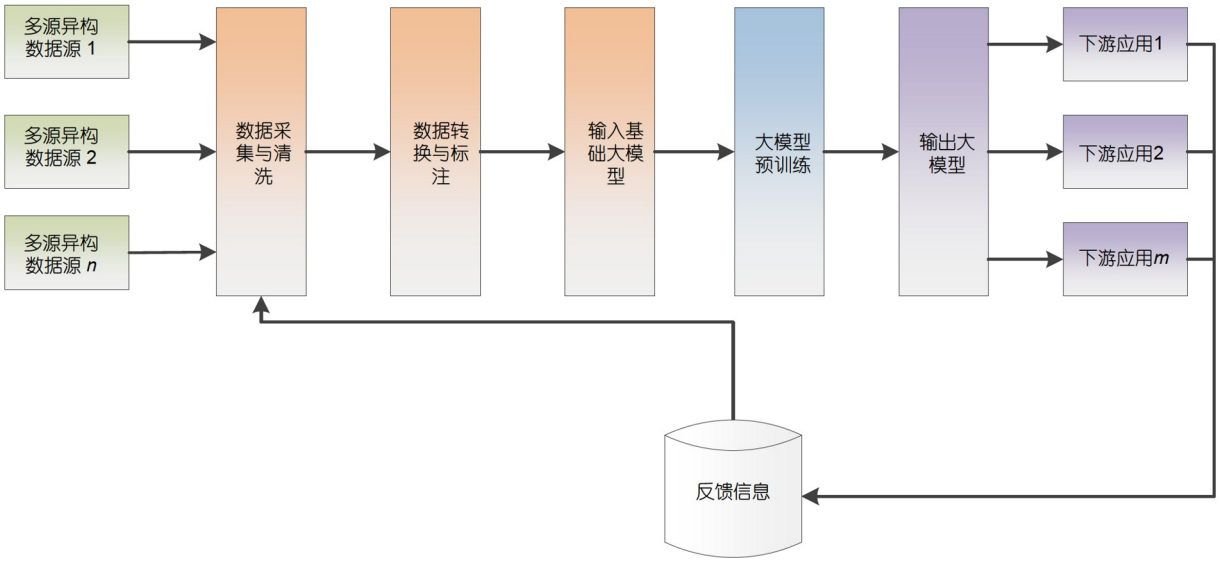

图 4 电子病历大模型构建流程示意图
Figure 4 Flowchart for building massively large EHR-based artificial intelligence models

6.4 强化组织实施

医学大模型研制既是一场总体战，也是一场攻坚战，需要不同领域、不同单位、不同专业的人员主动参与, 高效协同.

首先，参研单位要强强联合，既需有国内知名医院，其综合实力、专科特色、技术优势、声誉口碑居业界一流；又要有医学人工智能、云计算、网络通信、芯片研发等领域的标杆企业；还要有人工智能、信息科学等领域的著名高校和科研院所以及不断推进医学大模型标准化进程的综合性国家级检验机构等单位.

其次，需要具备业内领先的技术研发实力、产品创新活力和产业服务能力，全力构建涵盖硬件、软件、算法、模型、数据、芯片等各层面完全自主可控的医学大模型产业服务平台，努力在模式创新、产业升级、平台共享、应用转化等方面逐步形成军民融合、深度发展的崭新格局，助力构建医学大模型研究、应用、服务与保障一体化的国家战略体系.

最后，研制开源共享的医学大模型将有助于破解行业痛点，有力推动我国面向医疗健康行业的人工智能产业新模式和新生态的茁壮成长，努力探索一条具有中国特色、优质高效、自主可控、绿色低碳的医学大模型高质量发展之路.

7 总结

近年来，国内外连续推出了多个功能强大、特色鲜明的大模型，模型体量和参数规模获得快速跃升，模型功能与技术指标得到不断突破，引起社会广泛关注. 近日，北京市发布《促进通用人工智能创新发展的若干措施》[12]，强调要抢抓大模型发展机遇，重视通用人工智能发展，充分发挥政府引导作用和创新平台催化作用，整合创新自由，加强要素配置，推动通用人工智能技术在医疗、政务、金融等多个领域的示范应用和创新发展. 同时, 天津、上海等地推出加快算力基础设施建设、增强人工智能公共算力资源、支撑千亿级参数量的大语言模型、多模态大模型、大规模精细神经网络模拟仿真模型研发等政策，努力改善大算力紧缺、大模型稀缺的局面.

因此, 人工智能大模型浪潮正奔涌而来, 相关行业的转型重塑也在加速推进. 国内大模型研发和应用已进入发展的快车道，这也给医学人工智能大模型的发展带来新的契机. 本文通过全面梳理和对比分析，清晰勾勒出大模型的技术脉络、发展路径和演进趋势; 并且结合医学人工智能领域的实际需求，着重阐述了医学大模型的构建思路与研发经验，希望以此推动国内医学大模型的繁荣发展.

参考文献

1 Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv, 2019, 1810.04805

2 Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. Available from URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

3 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. arXiv, 2021, 2103.00020

4 Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation. arXiv, 2021, 2102.12092

5 Lin J Y, Men R, Yang A, et al. M6: A Chinese multimodal pretrainer. arXiv, 2021, 2103.00823

6 Zeng W, Ren X Z, Su T, et al. PANGU-α: large-scale autoregres-sive pretrained Chinese language models with auto-parallel computation. arXiv, 2021, 2104.12369

7 Wu C F, Liang J, Ji L, et al. NÜWA: visual synthesis pre-training for neural visUal world creAtion. arXiv, 2021, 2111.12417

8 Yuan H Y, Yuan Z, Gan R Y, et al. BioBART: pretraining and evaluation of a biomedical generative language model. arXiv, 2022, 2204.03905

9 The State Council. Development Plan for The New Generation of Artificial Intelligence (in Chinese). Available from URL: http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm [国务院. 新一代人工智能发展规划. http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm]

10 Standardization Administration, Office of the Central Cyberspace Affairs Commission, National Development and Reform Commission, Ministry of Science and Technology of the People's Republic of China, Ministry of Industry and Information Technology of the People's Republic of China. Guidelines for the Construction of the National New Generation Artificial Intelligence Standard System (in Chinese). Available from URL: https://www.ncsti.gov.cn/kjdt/ztbd/gjjcyfw/rengongzhineng/rengongzhinengzczc/202109/P020210927407819249513.pdf [国家标准化管理委员会, 中央网信办, 国家发展改革委, 科技部, 工业和信息化部. 国家新一代人工智能标准体系建设指南. https://www.ncsti.gov.cn/kjdt/ztbd/gjjcyfw/rengongzhineng/rengongzhinengzczc/202109/P020210927407819249513.pdf]

11 Strategic Consulting Center of Chinese Academy of Engineering Global Engineering Frontier. Global Engineering Frontier (2019) (in Chinese). Beijing: Higher Education Press, 2019. 33–34+173–174+176–178+182–184+192–194+217–218 [全球工程前沿(2019版). 中国工程院战略咨询中心. 北京: 高等教育出版社, 2019. 33–34+173–174+176–178+182–184+192–194+217–218]

12 Beijing Municipal Commission of Science and Technology, Zhongguancun Science and Technology Park Management Committee. Several Measures for Promoting the Innovative Development of General Artificial Intelligence in Beijing (2023-2025) (Exposure Draft) (in Chinese). Available from URL: http://kw.beijing.gov.cn/art/2023/5/12/art_2418_4626.html [北京市科学技术委员会, 中关村科技园区管理委员会. 关于对《北京市促进通用人工智能创新发展的若干措施(2023-2025年)(征求意见稿)》公开征集意见的公告. http://kw.beijing.gov.cn/art/2023/5/12/art_2418_4626.html]

13 Bommasani R, Hudson D A., Adeli E, et al. On the opportunities and risks of foundation models. arXiv, 2021, 2108.07258

14 Sanh V, Webson A, Raffel C, et al. Multitask prompted training enables zero-shot task generalization. arXiv, 2021, 2110.08207

15 Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. arXiv, 2020, 2005.14165

16 Sharir O, Peleg B, Shoham Y. The cost of training nlp models: a concise overview. arXiv, 2020, 2004.08900

17 Strategic Consulting Center of Chinese Academy of Engineering. Global Engineering Frontier (2021) (in Chinese). Beijing: Higher Education Press, 2021. 35–37+48–49 [中国工程院战略咨询中心. 全球工程前沿(2021版). 北京: 高等教育出版社. 35–37+48–49]

18 Tulloch A, Jia Y Q. High performance ultra-low-precision convolutions on mobile devices. arXiv, 2017, 1712.02427

19 Ren L Z, Chen Y J, Guo X, et al. HIS Kernel Design Guide-Systematic Thinking about Planning and Design of Hospital Information System (in Chinese). Beijing: China Industrial and Information Technology Publishing Group, 2021 [任连仲, 陈一君, 郭旭, 等. HIS内核设计之道——医院信息系统规划设计系统思维. 北京: 中国工信出版集团, 2021]

20 Fedus W, Zoph B, Shazeer N, et al. Switch Transformers: scaling to trillion parameter models with simple and efficient sparsity. arXiv, 2021, 2101.03961

21 Sauer A, Karras T, Laine S, et al. StyleGAN-T: unlocking the power of gans for fast large-scale text-to-image synthesis. arXiv, 2023, 2301.09515

22 Ouyang W, Wang G Y. TPU: analysis of Google artificial intelligence chip structure (in Chinese). Dev Appl High Perform Comput, 2018, 62: 27–32 [欧阳伟, 王广益. TPU: Google人工智能芯片结构浅析. 高性能计算发展与应用, 2018, 62: 27–32]

23 Rao A, Plank P, Wild A, et al. A long short-term memory for AI applications in spike-based neuromorphic hardware. Nat Mach Intell, 2022, 4:

467–479
24 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv, 2017, 1706.03762
25 Shazeer N, Mirhoseini A, Maziarz K, et al. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. arXiv, 2017, 1701.06538
26 Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. 2019. Available from URL: https://gwern.net/doc/ai/nn/transformer/gpt/2019-radford.pdf
27 Shoeybi M, Patwary M, Puri R, et al. Megatron-LM: training multi-billion parameter language models using model parallelism. arXiv, 2019, 1909.08053
28 Rosset C. Turing-NLG: a 17-billion-parameter language model by Microsoft. 2020. Available from URL: https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/
29 Lee J, Yoon W, Kim S D, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. arXiv, 2019, 1901.08746
30 Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners. arXiv, 2020, 2005.14165
31 Li X J, Yin X, Li C Y, et al. Oscar: object-semantics aligned pre-training for vision-language tasks. arXiv, 2020, 2004.06165
32 Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv, 2019, 1910.10683
33 Zhang Z Y, Gu_Y X, Han X, et al. CPM-2: large-scale cost-effective pre-trained language models. arXiv, 2021, 2106.10715
34 Zhang Z S, Zhang H Q, Chen K M, et al. Mengzi: towards lightweight yet ingenious pre-trained models for Chinese. arXiv, 2021, 2110.06696
35 Ding M, Yang Z Y, Hong W Y, et al. CogView: mastering text-to-image generation via transformers. arXiv, 2021, 2105.13290
36 Mu N, Kirillov A, Wagner D, et al. SLIP: self-supervision meets language-image pre-training. arXiv, 2021, 2112.12750
37 Rae J W, Borgeaud S, Cai T, et al. Scaling language models: methods, analysis & insights from training gopher. arXiv, 2021, 2112.11446
38 Fei N, Lu Z, Gao Y, et al. Towards artificial general intelligence via a multimodal foundation model. Nat Commun, 2022, 13: 3094
39 Tesauro G, Gondek D C, Lenchner J, et al. Simulation, learning, and optimization techniques in Watson's game strategies. IBM J Res Dev, 2012, 56: 16:1–16:11
40 Peters M E, Neumann M, lyyer, M et al. Deep contextualized word representations. arXiv, 2018, 1802.05365
41 Alsentzer E, Murphy J R, Boag W, et al. Publicly available clinical BERT embeddings. arXiv, 2019, 1904.03323
42 Gu Y, Robert T, Cheng H, et al. Domain-Specific language model pretraining for biomedical natural language processing. arXiv, 2020, 2007.15779
43 Jin D, Pan E, Oufattole N, et al. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci, 2021, 11: 6421
44 Pal A, Umapathi L K, Sankarasubbu M. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering. arXiv, 2022, 2203.14371
45 Jin Q, Dhingra B, Liu Z, et al. PubMedQA: a dataset for biomedical research question answering. arXiv, 2019, 1909.06146
46 Abacha A B, Mrabet Y, Sharp M, et al. Bridging the gap between consumers' medication questions and trusted answers. Stud Health Technol Inform, 2019, 264: 25–29
47 Hendrycks D, Burns C, Basart S, et al. Measuring massive multitask language understanding. arXiv, 2021, 2009.03300
48 Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. arXiv, 2022, 2212.13138
49 Johnson A E, Pollard T J, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data, 2016, 3: 160035
50 Lewis P, Ott M, Du J F, et al. Pretrained language models for biomedical and clinical tasks: understanding and extending the state-of-the-art. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop. Online. Association for Computational Linguistics, 2020. 146–157
51 Shin H C, Zhang Y, Bakhturina E, et al. BioMegatron: larger biomedical domain language model. arXiv, 2020, 2010.06060
52 Hong Z, Ajith A, Pauloski G, et al. ScholarBERT: bigger is not always better. arXiv, 2022, 2205.11342
53 Luo R Q, Sun L A, Xia Y C, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. arXiv, 2022, 2210.10341
54 Li Y X, Li Z H, Zhang K, et al. ChatDoctor: a medical chat model fine-tuned on llama model using medical domain knowledge. arXiv, 2023, 2303.14070
55 Wu T, Wang Y L, Wang Y, et al. Leveraging graph-based hierarchical medical entity embedding for healthcare applications. Sci Rep, 2021, 11:

5858
56 Perozzi B, Al-Rfou R, Skiena S. Deepwalk: online learning of social representations. arXiv, 2014, 1403.6652
57 Grover A, Leskovec J. node2vec: scalable feature learning for networks. arXiv, 2016, 1607.00653
58 Jian T, Meng Q, Wang M, et al. LINE: large-scale information network embedding. arXiv, 2015, 1503.03578
59 Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. arXiv, 2017, 1706.02216
60 Velikovi P, Cucurull G, Casanova A, et al. Graph attention networks. arXiv, 2017, 1710.10903
61 Choi E, Xiao C, Stewart W F, et al. MiME: multilevel medical embedding of electronic health records for predictive healthcare. arXiv, 2018, 1810.09593
62 Choi E, Xu Z, Li Y, et al. Graph convolutional transformer: learning the graphical structure of electronic health records. arXiv, 2019, 1906.04716
63 Rasmy L, Xiang Y, Xie Z Q, et al. Med-BERT: pretrained contextualized embeddings on large scale structed electronic health records for disease prediction. arXiv, 2020, 2005.12833
64 Hong C, Rush E, Liu M, et al. Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data. NPJ Digit Med, 2021, 4: 151
65 Wang J J, Zhang Y X, Zhang L, et al. Fengshenbang 1.0: being the foundation of Chinese cognitive intelligence. arXiv, 2022, 2209.02970
66 Zhu S Q, Yu T, Xu T, et al. Intelligent computing: the latest advances, challenges and future. arXiv, 2022, 2211.11281
67 Yuan J H, Li X Q, Cheng C, et al. OneFlow: redesign the distributed deep learning framework from scratch. arXiv, 2021, 2110.15032
68 Fan J T, Suo J L, Wu J M, et al. Video-rate imaging of biological dynamics at centimetre scale and micrometre resolution. Nat Photon, 2019, 13: 809–816
69 Wu J M, Lu Z, Jiang D, et al. Iterative tomography with digital adaptive optics permits hour-long intravital observation of 3D subcellular dynamics at millisecond scale. Cell, 2021, 184: 3318–3332
70 Zeng Y, Zhao D C, Zhao F F, et al. BrainCog: a spiking neural network based brain-inspired cognitive intelligence engine for brain-inspired ai and brain simulation. arXiv, 2022, 2207.08533
71 Wu J, Liu X, Zhang X, et al. Master clinical medical knowledge at certificated-doctor-level with deep learning model. Nat Commun, 2018, 9: 4352
72 Wu C, Wu F, Lyu L, et al. A federated graph neural network framework for privacy-preserving personalization. Nat Commun, 2022, 13: 3091
73 Huang Y Z, Bai Y Z, Zhu Z H, et al. C-Eval: a multi-level multi- discipline chinese evaluation suite for foundation models. arXiv, 2023, 2305.08322
74 Zhang H, Zong Y, Chang B B, et al. Medical entity annotation standard for medical text processing (in Chinese). In: Proceedings of the 19th Chinese National Conference on Computational Linguistics, Chinese Information Processing Society of China. Haikou. 2020. 561–571 [张欢, 宗源, 常宝宝, 等. 面向医学文本处理的医学实体标注规范. 见: 第十九届中国计算语言学大会论文集. 海口. 2020, 561–571]
75 Institute of Computational Linguistics, Peking University, Natural Language Processing Lab, Zhengzhou University, PengCheng Laboratory. Release of China Medical Knowledge Atlas CMeKG2.0 (in Chinese). Available from URL: http://www5.zzu.edu.cn/nlp/info/1018/1785.htm [北京大学计算语言学研究所, 郑州大学自然语言处理实验室, 鹏城实验室. 中文医学知识图谱CMeKG2.0版. http://www5.zzu.edu.cn/nlp/info/1018/1785.htm]
76 Xu L, Hu H, Zhang X W, et al. CLUE: a Chinese language understanding evaluation benchmark. arXiv, 2020, 2004.05986
77 Xu L, Lu X J, Yuan C Y, et al. FewCLUE: a Chinese few-shot learning evaluation benchmark. arXiv, 2021, 2107.07498
78 Zhang N Y, Chen M S, Bi Z, et al. CBLUE: a Chinese biomedical language understanding evaluation benchmark. arXiv, 2021, 2106.08087
79 Zhao Z Y, Jin Q, Chen F Y, et al. PMC-Patients: a large-scale dataset of patient summaries and relations for benchmarking retrieval-based clinical decision support Systems. arXiv, 2022, 2202.13876
80 Wang Q, Dai S T, Xu B F, et al. Building Chinese biomedical language models via multi-level text discrimination. arXiv, 2021, 2110.07244
81 Wang H C, Liu C, Xi N W, et al. HuaTuo: tuning LLaMA model with Chinese medical knowledge. arXiv, 2023, 2304.06975
82 Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. arXiv, 2023, 2302.13971
83 Zhang S W, Sun J C, Huang Y, et al. Medical symptom detection in intelligent pre-consultation using bi-directional hard-negative noise contrastive estimation. In: KDD '22: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery, 2022. 4551–4559

84 Zhu X W, Chen W G, Zheng W M, et al. Gemini: a computation-centric distributed graph processing system. In: OSDI′16: Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation. Berkeley: USENIX Association, 2016. 301–316

85 Davoudian A, Chen L, Tu H, et al. A workload-adaptive streaming partitioner for distributed graph stores. Data Sci Eng, 2021, 6: 163–179

86 Battaglia P W, Hamrick J B, Bapst V, et al. Relational inductive biases, deep learning, and graph networks. arXiv, 2018, 1806.01261

87 Stanton I, Kliot G. Streaming graph partitioning for large distributed graphs. In: KDD ′12: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: Association for Computing Machinery, 2012. 1222–1230

88 Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv, 2020, 2010.11929

89 Guo M H, Lu C Z, Liu Z N, et al.Visual attention network. arXiv, 2022, 2202.09741

90 Peng B, Alcaide E, Anthony Q, et al. RWKV: reinventing RNNs for the transformer era. arXiv, 2023, 2305.13048

91 Lewis M, Liu Y H, Goyal N, et al. BART: denoising sequences-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv, 2019, 1910.13461

92 Yu W H, Luo M, Zhou P, et al. MetaFormer is actually what you need for vision. arXiv, 2022, 2111.11418

93 Chen Z S, Xie L X, Niu J W, et al. Visformer: the vision-friendly transformer. arXiv, 2021, 2104.12533

94 Ge Y Y, Ge Y X, Liu X H, et al. BridgeFormer: bridging video-text retrieval with multiple choice questions. arXiv, 2022, 2201.04850

95 Lian X R, Yuan B H, Zhu X F, et al. Persia: an open, hybrid system scaling deep learning-based recommenders up to 100 trillion parameters. arXiv, 2021, 2111.05897

96 Rajbhandari S, Li C L, Yao Z W, et al. DeepSpeed-MoE: advancing mixture-of-experts inference and training to power next-generation AI scale. arXiv, 2022, 2201.05596

97 Shazeer N, Cheng Y L, Parmar N, et al. Mesh-TensorFlow: deep learning for super-computers. arXiv, 2018, 1811.02084

98 Rasley J, Rajbhandari S, Ruwase O, et al, DeepSpeed: system optimizations enable training deep learning models with over 100 billion parameters. In: KDD ′20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: Association for Computing Machinery, 2020. 3505–3506

99 Narihira T, Alonsogarcia J, Cardinaux F, et al. Neural network libraries: a deep learning framework designed from engineers' perspectives. arXiv, 2021, 2102.06725

100 Jia X Y, Jiang L, Wang A, et al. Whale: a unified distributed training framework. arXiv, 2020, 2011.09208

101 Zhao W J, Jiao X W, Hu M Q, et al. Communication-efficient terabyte-scale model training framework for online advertising. arXiv, 2022, 2201.05500

102 Rajbhandari S, Ruwase O, Rasley J, et al. ZeRO-Infinity: breaking the GPU memorywall for extreme scale deep learning. arXiv, 2021, 2104.07857

103 Yi X D, Luo Z Y, Meng C, et al. Fast training of deep learning models over multiple GPUs. In: Middleware ′20: Proceedings of the 21st International Middleware Conference. New York: Association for Computing Machinery, 2020. 105–118

104 Bekkerman R, Bilenko M, Langford J. Scaling Up Machine Learning: Parallel and Distributed Approaches (in Chinese). In: Liu Z, Wang Y G, Zhang J T, et al., transl. Beijing: National Defense Industry, 2021 [罗恩•贝克曼, 米哈伊尔•比伦科, 约翰•兰福特.大规模机器学习: 并行和分布式技术. 见: 柳征, 王莹桂, 张建廷, 等, 译. 北京: 国防工业出版社, 2021]

105 Zhang X G, Qin H T, Ding Y F, et al. Diversifying sample generation for accurate data-free quantization. arXiv, 2021, 2103.01049

106 Dettmers T, Lewis M, Shleifer S, et al. 8-bit optimizers via block-wise quantization. arXiv, 2021, 2110.02861

107 Dettmers T, Lewis M, Belkada Y, et al. LLM.int8(): 8-bit matrix multiplication for transformers at scale. arXiv, 2022, 2208.07339

108 Denil M, Shakibi B, Dinh L, et al. Predicting parameters in deep learning. arXiv, 2013, 1306.0543

109 Misha Denil, Babak Shakibi, Laurent Dinh, et al. Predicting parameters in deep learning. arXiv, 2013, 1306.0543

110 Han Y Z, Huang G, Song S J, et al. Dynamic neural networks: a survey. arXiv, 2021, 2102.04906

111 Xia Z F, Pan X R, Song S J, et al. Vision transformer with deformable attention. arXiv, 2022, 2201.00520

112 Zhou J H, Wei C, Wang H Y, et al. iBOT: image BERT pre-training with online tokenizer. arXiv, 2021, 2111.07832

113 Cheng J, Wang P S, Li G, et al. Recent advances in efficient computation of deep convolutional neural networks. arXiv, 2018, 1802.00939

114 Cheng Y, Wang D, Zhou P, et al. A survey of model compression and acceleration for deep neural networks. arXiv, 2017, 1710.09282

115 Li H, Kadav A, Durdanovic I, et al. Pruning filters for efficient convNets. arXiv, 2016, 1608.08710

116 Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention. arXiv, 2020, 2012.12877

117 Micikevicius P, Narang S, Alben J, et al. Mixed precision training. arXiv, 2017, 1710.03740

118 Denton E, Zaremba W, Bruna J, et al. Exploiting linear structure within convolutional networks for efficient evaluation. arXiv, 2014, 1404.0736

119 Deghmani F, AmineAmarouche I. Graph databases and big data technologies in healthcare: a gap analysis. In: Proceedings of the Advances of Decisional Systems. Marrakech. 2018

120 Tong Y, Pan X, Zeng Y, et al. Hu-Fu: efficient and secure spatial queries over data federation. Proc VLDB Endow, 2022, 15: 1159–1172

121 Yang J F, Guan Y, He B, et al. Corpus construction for named entities and entity relations on Chinese electronic medical records (in Chinese). J Software, 2016, 27: 2725–2746 [杨锦锋, 关毅, 何彬, 等. 中文电子病历命名实体和实体关系语料库构建. 软件学报, 2016, 27: 2725–2746]

122 Wang S G, Jiang S S. Optimal hyperparameters and structure setting of multi-objective robust CNN systems via generalized taguchi method and objective vector norm. arXiv, 2022, 2202.04567

123 Paren A, Berrada L, Poudel R, et al. A stochastic bundle method for inter-polating networks. arXiv, 2022, 2201.12678

124 Gusak J, Cherniuk D, Shilova A, et al. Survey on large scale neural network training. arXiv, 2022, 2202.10435v1

125 Pinto D, Arnau J M, González A. Mixture-of-Rookies: saving DNN computations by predicting ReLU outputs. arXiv, 2022, 2202.04990

126 Cyberspace Administration of China. Management Measures for Generative Artificial Intelligence Services (Exposure Draft) (in Chinese). Available from URL: http://www.cac.gov.cn/2023-04/11/c_1682854275475410.htm [国家互联网信息办公室. 关于《生成式人工智能服务管理办法(征求意见稿)》公开征求意见的通知. http://www.cac.gov.cn/2023-04/11/c_1682854275475410.htm]

127 National Medical Products Administration. Quality requirements and evaluation of artificial intelligence medical devices - Part 1: Terminology (in Chinese), YY/T 1833.1-2022, Released on July 1, 2022 [国家药品监督管理局. 人工智能医疗器械 质量要求和评价 第1部分: 术语. YY/T 1833.1-2022, 2022-07-01发布]

128 National Medical Products Administration. Quality requirements and evaluation of artificial intelligence medical devices - Part 2: General requirements for datasets (in Chinese), YY/T 1833.2-2022, Released on July 1, 2022 [国家药品监督管理局. 人工智能医疗器械 质量要求和评价 第2部分: 数据集通用要求. YY/T 1833.2-2022, 2022-07-01发布]

129 National Medical Products Administration. Quality requirements and evaluation of artificial intelligence medical devices - Part 3: General requirements for data annotation (in Chinese), YY/T 1833.3-2022, Released on Aug 17, 2022 [国家药品监督管理局. 人工智能医疗器械 质量要求和评价 第3部分: 数据标注通用要求. YY/T 1833.3-2022, 2022-08-17发布]

130 National Medical Products Administration. Performance testing method for algorithms in artificial intelligence medical device lung imaging assisted analysis software (in Chinese), YY/T 1858-2022, Released on Aug 17, 2022 [国家药品监督管理局. 人工智能医疗器械 肺部影像辅助分析软件 算法性能测试方法. YY/T 1858-2022, 2022-08-17发布]

131 Liu B, Liu P, Dai L, et al. Assisting scalable diagnosis automatically via CT images in the combat against COVID-19. Sci Rep, 2021, 11: 4145

132 Yang F, Chen X, Lin X, et al. Automated analysis of doppler echocardiographic videos as a screening tool for valvular heart diseases. JACC Cardiovasc Imag, 2022, 15: 551–563

133 Zhong Q, Li Z, Wang W, et al. Integrated medical resource consumption stratification in hospitalized patients: an Auto Triage Management model based on accurate risk, cost and length of stay prediction. Sci China Life Sci, 2022, 65: 988–999

134 Lipkova J, Chen R J, Chen B, et al. Artificial intelligence for multimodal data integration in oncology. Cancer Cell, 2022, 40: 1095–1110

135 Zeng A H, Liu X, Du Z X, et al. GLM-130B: an open bilingual pre-trained model. arXiv, 2022, 2210.02414

136 Aghajanyan A, Gupta S, Zettlemoyer L. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. arXiv, 2020, 2012.13255

137 Yang J F, Jin H Y, Tang R X, et al. Harnessing the power of LLMs in practice: a survey on ChatGPT and Beyond. arXiv, 2023, 2304.13712

Research on a massively large artificial intelligence model and its application in medicine

GUO HuaYuan[1,2], LIU Pan[2], LU RuoGu[2], YANG FeiFei[3], XU HongLi[2], ZHUANG Yan[2], HUANG Gao[4], SONG ShiJi[4] & HE KunLun[2]

1 Medical Artificial Intelligence Research Center, Department of Medical Innovation Research of PLA General Hospital, Beijing 100853, China;
2 Medical Big Data Research Center, Department of Medical Innovation Research of PLA General Hospital, Beijing 100853, China;
3 Department of Cardiology, the Fourth Medical Center of Chinese PLA General Hospital, Beijing 100048, China;
4 Department of Automation, Tsinghua University, Beijing 100084, China

Recent years have witnessed rapid advancements in massively large artificial intelligence (AI) models based on natural language processing and video image analysis. In order to meet the requirements of relevant application fields, the universal pretraining model is developed by the efficient collaboration and deep integration of big data, large-scale computing power, and complex algorithms. The model demonstrates adaptability to a wide range of downstream tasks. In addition, the massively large model presents considerable opportunities for advancing the quality of medical AI development. Therefore, this paper comprehensively analyzes the progress of massively large models within domestic and international contexts in recent years, with an emphasis on their key technologies and algorithmic framework. Meanwhile, the developmental characteristics of a series of standard datasets and pretraining models in the biomedical field have been presented in detail. Incorporating our team's practical experience in medical AI research and development, we undertake a comprehensive analysis of the application requirements, our solutions and experiences related to the construction of massively large models in the medical field and persistently promote innovation and development within the realm of large-scale medical models.

medicine, artificial intelligence, massively large model, natural language processing, medical image analysis

doi: 10.1360/SSV-2022-0298