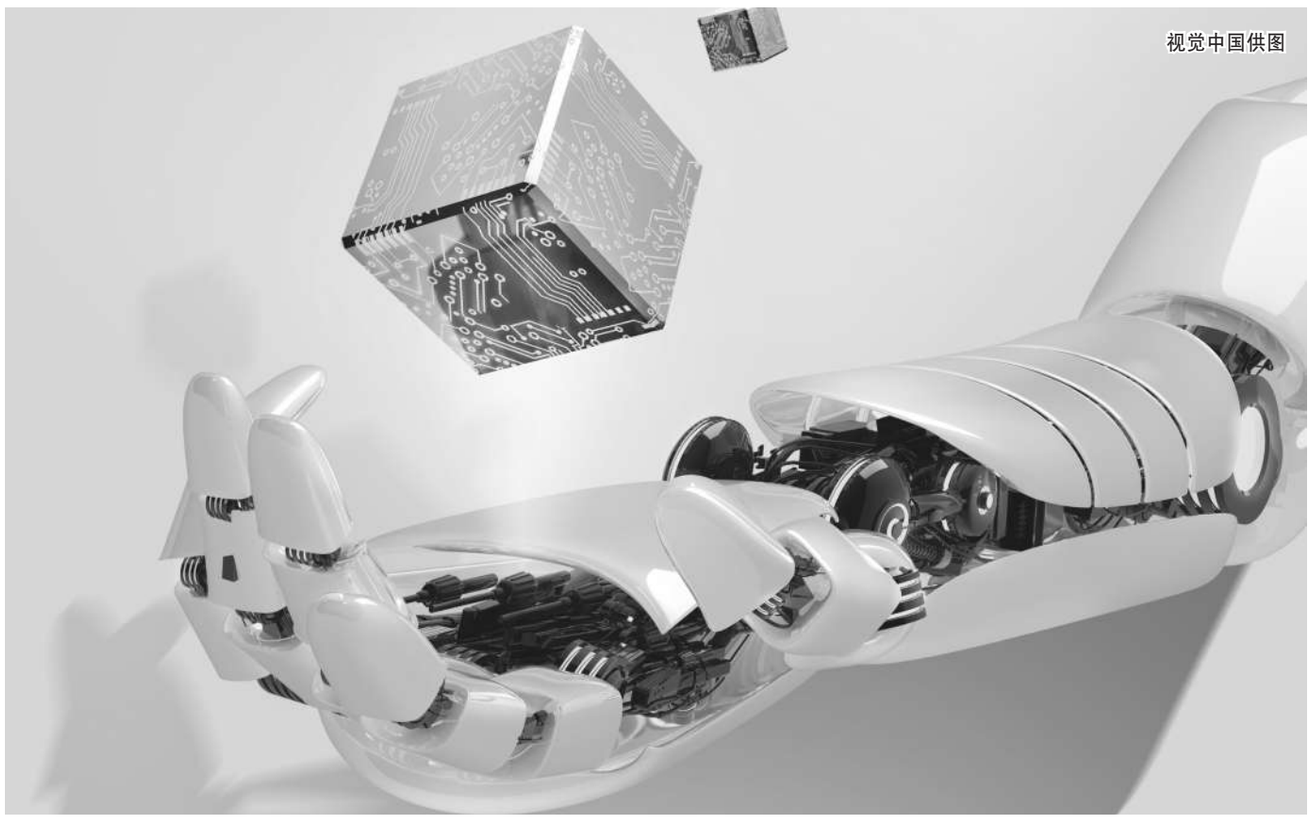
视觉中国供图

> 产业的需求决定了要完成的人工智能任务越来越复杂，轻量化人工智能必须通过加速运算效率、提高计算密度才能实现极致的效率。
>
> **冷聪**
> 中国科学院自动化研究所副研究员

AI进入下半场 目标瞄准“举重若轻”“大材小用”

◎本报记者 **陆成宽**

人工智能算法的复杂度急剧攀升，神经网络计算的能耗代价越来越高，数据洪水式地涌积堰塞……这些年，人工智能的发展遇到了越来越多的瓶颈。如何将人工智能模型及其计算载体前端化、轻量化成为亟待解决的问题。最新兴起的轻量化人工智能被寄予厚望，以“轻量化”为赛点的人工智能竞赛下半场已经来临。为此，4月27日，科技日报记者采访了中国科学院自动化研究所的相关专家。

轻量化成为人工智能下半场赛点

人工智能技术在行业应用中，大多依赖海量的训练数据和大规模服务器的算力支持。

然而，近年来，随着信息技术领域的摩尔定律逐步放缓，硬件的发展越来越难以满足当前人工智能模型动辄万亿级规模的存储和算力需求，数据堰塞、存储暴涨、隐私泄露、能耗高企等问题随之而来。

“当前，对于人工智能设备和应用的快速响应、隐私保护以及节能减排的需求越发凸显，轻量化人工智能应运而生，并被寄予厚望。2020年，《麻省理工科技评论》将轻量化人工智能列为‘全球十大突破性技术’。”中国科学院自动化研究所研究员程健说。

所谓轻量化人工智能，是指以一系列轻量化技术为驱动提高芯片、平台和算法的效率，在更紧密的物理空间上实现低功耗的人工智能训练和应用部署，不需要依赖与云端的交互就能实现智能化操作的人工智能。

轻量化人工智能被评入“全球十大突破性技术”，《麻省理工科技评论》给出的评选理由是，轻量化智能使现有的服务，比如语音助手、手机拍照等，变得更好更快，不必每次都需要连接云端才能运行深度学习模型；此外，轻量化人工智能也将使新的应用成为可能，比如基于移动端的医学检测分析、对反应时间要求更快的自动驾驶汽车；此外，本地化的人工智能更利于隐私保护，用户的数据不再需要离开设备就能实现服务功能的进化。

“更重要的是，轻量化人工智能将人工智能推向更主流，它大大降低了人工智能系统的部署难度和成本，把人工智能从一场高门槛的科技巨头竞赛变成更容易普惠民生的智能生态。”程健说，在人工智能领域的角逐中，以轻量化为赛点的下半场已经来临。

极致效率、极低能耗是最终追求

在表现上，轻量化人工智能是在做减法，降低能耗、降低对硬件平台性能指标的要求、降低与云端的通讯需求等。

然而，“实质上，轻量化的内核却是在做加法。”中国科学院自动化研究所副研究员冷聪说，产业的需求决定了要完成的人工智能任务越来越复杂，轻量化人工智能必须通过提高运算速度、计算密度才能实现极致的效率。

在程健看来，在精度接近无损的前提下，将人工智能模型及其计算载体轻量化，是一个极具挑战性的任务。

解决这一问题，需要对神经网络进行轻量化设计、计算加速以及设计新的计算架构以实现模型的硬件化，这需要从软件和硬件两方面来着手。

在软件上，进行模型和算法创新，通过轻量化模型设计、矩阵分解、稀疏表示、量化计算来实现模型的微型化和计算加速；而在硬件上，则要通过流水线设计、存储模式设计等手段进行硬件架构的创新，通过软硬协同设计和优化实现人工智能的轻量化。

“虽然执行神经网络计算的是硬件，但神经网络结构和人工智能平台决定了计算量的大小和运算方式。”冷聪坦言，所以极致的轻量化必须是软件和硬件的协同轻量化——基于复杂的人工智能应用场景，将芯片、平台和算法充分结合以联合加速。

作为人工智能的硬件载体，人工智能芯片必须达到更高的性能、更高的效率、更低的功耗和更小的体积。这样才能有足够平价高效的计算平台满足产业需求，承载复杂的人工智能任务，并且使推理和运算从云端迁移到终端。

同时，轻量化的人工智能平台要以更低的功耗来训练和运行人工智能算法，最大化的发掘硬件的能力。更重要的是，应用轻量化技术的神经网络模型要小规模、少运算量并保持良好的精度。

未来轻量化人工智能将赋能万物

程健介绍，中国科学院自动化研究所是轻量化人工智能的先行者，很早就开始了软硬协同轻量化的技术研究，并走在国际前列。

早在2016年，卷积神经网络大规模迈向应用之初，中国科学院自动化研究所就在国际人工智能顶级期刊发表了多篇神经网络模型轻量化领域的重要论文，成为国际上最早开始人工智能轻量化研究的机构之一，相关成果引起了国内外诸多专家的广泛关注。

“我们设计开发的轻量化人工智能平台QEngine及轻量化算法已经在数十万终端上部署。2019年，在国际神经信息处理系统大会的微型网络挑战竞赛中，我们与ARM、IBM、高通、Xilinx等国际一流芯片公司同场竞技，获得了轻量化神经网络架构图像类的双冠军。”程健表示。

2020年，中国科学院自动化研究所自主研发的全球首款极低比特量化神经处理芯片（QNPU）成功流片，绕开了芯片计算领域备受关注的“内存墙”难题，在芯片成本、功耗、计算结构、边缘计算等方面实现革命性的变革。

“该芯片的面世，也标志着自动化研究所成为了全球为数不多的拥有‘人工智能芯片—平台—算法’全栈轻量化人工智能技术的机构之一。”冷聪说。

未来，以人工智能驱动的小型化设备会越来越多出现在我们身边。由人工智能芯片、平台和算法组成的轻量化人工智能终端将在越来越多的场景中应用。

“比如，在电力行业，我国的输电线路覆盖广，野外自然环境复杂，检修维护作业危险系数高、难度大，我们设计的自主巡检无人机、缺陷识别分析便携终端、通道可视化智能感知摄像头具备多种智能识别、检测和分析功能，能够保障输配电线路的安全和电力系统稳定。”程健举例说。

同时，在消费电子行业，暗光增强、超分辨率等自动化所设计的轻量化算法及轻量化神经网络计算架构，也为手机终端、安防终端提供了影像增强效果。

程健表示，轻量化人工智能未来将赋能万物，让每个设备都具有环境感知、人机交互、决策控制的能力。

情报所

AI识别新冠肺炎重症患者预判病程精确到“天”

◎本报记者 **张 晔**

4月25日，《自然》子刊《NPJ数字医学》（NPJ Digital Medicine）发表了我国学者的一项研究成果——基于人工智能图像分析技术的新冠肺炎快速风险分层系统。这项研究能在救治新冠肺炎患者方面实现优化医疗资源调度、及时介入治疗，减少不良结局发生率，从而挽救更多生命。

这项将人工智能技术运用于新冠肺炎诊断的研究成果由解放军东部战区总医院放射诊断科卢光明、张龙江教授，联合斯坦福生物医学信息研究中心、深睿医疗等机构联合完成。“精准预测新冠肺炎患者病情进展并实现快速分诊，是此项研究成果的重要价值。”卢光明表示，他们从医学影像、临床资料、实验室检查等诸多维度入手，通过人工智能机器学习模型筛选出新冠肺炎不良进展的危险因素，精准识别28天内可能需要重症监护、机械通气，甚至发生死亡的危险患者，同时预测这些高危患者从入院至出现各种不良结局的具体天数。

世界卫生组织数据显示，截至2021年4月25日，全球累计新冠肺炎感染患者达到1.46亿人，其中超过300万死亡病例。新冠病毒持续暴发吞没全球医疗资源，尤其造成重症监护病房（ICU）超负荷和机械通气（呼吸机）短缺，导致医疗挤兑。因此，准确识别高风险患者并预测不良结局的发生时间，对优化医疗资源分配和临床分层诊疗具有重要意义。

据卢光明介绍，目前，重症监护病房和机械通气等支持性治疗资源的短缺是造成新冠肺炎病人死亡的主要因素。而当前临床常用急性生理和慢性健康评估Ⅱ评分、中性粒细胞/淋巴细胞比值（NLR）等实验室指标，以及以计算机断层扫描（CT）为主的影像学手段，来评估感染性肺炎患者的病情严重程度。“但这些评估手段主观性强、耗时长、不够全面，且无法对病情进展情况进行预测。”

“临床急需能预测患者病情进展情况的快速分诊方法，以判断需要ICU病房、呼吸机等医疗资源的病人数量及需要的具体时间，这也是优化并保障医疗资源供应、平衡疫区ICU负荷、实现及时救治的关键。”卢光明团队自2020年3月起，致力于创建全新的结合多种数据类型的综合评估手段，以预测患者病情进展情况及进展时间。

他们集中内地39家医院3522例新冠肺炎患者的数据建立人工智能模型，用人工智能软件处理并提取患者CT图像特征，结合血液检测实验室指标和临床资料，对患者在未来28天内是否会发展为重症并需要送入ICU进行预测，然后对可能出现重症的患者进行分层，预测他们是否会发生呼吸衰竭等需要机械通气的情况，以及是否会发生死亡。此外，模型还会预测高危病人发生每项关键事件的时间点，如：入院后多少天需要送入ICU或机械通气、入院后多少天会出现死亡。

“我们的系统预判准确率达到97.9%。”卢光明说，当患者呼吸系统情况恶化时，通常只有极为有限的时间来挽救生命，我们的研究通过全新的多类型数据融合实现了对患者病情发展的精准预判。

卢光明表示，可以利用预测结果来优化不同的医疗中心的资源分配，如提前调配医生、ICU病床或呼吸机去疫情严重的医院，或转运患者以平衡ICU负载。另外，在患者入院时即预测其对呼吸机的需求，可实现对高危患者更密切的监护和病情评估。

此外，新成果将对医疗资源需求与死亡的预测结果相结合，可以在资源极为短缺时，为最有可能受益的患者分配资源，“这有助于新冠肺炎流行期间的优先配给策略制定。”

深圳首个智能网联汽车应用示范许可发出

科技日报讯（记者**刘传书**）4月23日，L4级自动驾驶解决方案提供商深圳元戎启行科技有限公司（以下简称元戎启行）获得深圳市智能网联汽车道路测试联席工作小组发出的《智能网联汽车应用示范通知书》。元戎启行成为第一家能够在深圳开展自动驾驶载人应用示范的企业。

根据规划，元戎启行将从小区域定向邀请开始，逐步扩大载人应用示范规模，在今年年中于深圳市中心向公众开放自动驾驶出行服务。

根据《深圳市关于推进智能网联汽车应用示范的指导意见》，开展载人应用示范的自动驾驶车辆，应在申请应用示范所在区域道路测试累计不低于每车1000公里，且无交通违法或有责任交通事故。

近日，深圳市人大公开征求意见的《深圳经济特区智能网联汽车管理条例（征求意见稿）》（以下简称《征求意见稿》）明确了自动驾驶道路测试和示范应用、准入和登记、道路运输、交通事故及违章处理的相关条例。

《征求意见稿》进一步放宽了自动驾驶道路测试和示范应用相关条件，深圳特区的高速公路和城市快速路将允许开展道路测试和示范应用，车路协同设施较完善的行政区域可展开全域开放的自动驾驶道路测试、示范应用和商业化试点。针对出行服务，《征求意见稿》鼓励智能网联汽车提供定制出行、社区出行、夜间出行、应急保障等多样化服务，并允许合法收取服务费用。

此外，《征求意见稿》还对全无人自动驾驶的测试和应用进行了规定，满足相应条件的智能网联汽车，在进行道路测试和示范应用时，可不配备驾驶人。

根据天眼查数据显示，深圳共有约800家自动驾驶相关企业，占全国20%，拥有较完整的自动驾驶相关产业链。随着法规和政策的不断细化和完善，自动驾驶有望在深圳实现最全面的商业化试点。

人工智能：应用门槛降低，技术红利变现

◎本报记者 **王祝华**

新冠肺炎疫情期间，人工智能应用突飞猛进，社会各界对人工智能技术也寄予厚望，并期待人工智能可以成为疫情之后数字经济的核心驱动力之一。在3月发布的“十四五”规划和2035年远景目标纲要中，我国将“新一代人工智能”列为科技前沿攻关的七大领域之一，人工智能已上升为国家战略。

4月19日，博鳌亚洲论坛2021年年会特别关注“后疫情时代的人工智能”，多位专家受邀，线上线下共同探讨人工智能的产业应用、参与社会治理面临的机遇与挑战。

人工智能技术亟待突破

2015年图灵奖获得者、美国著名密码学与安全技术专家惠特菲尔德·迪菲在互联网安全领域颇有建树，被誉为“公钥加密技术之父”，曾是斯坦福大学人工智能实验室一员的他表示，人工智能必须要有更多的计算力才能达成我们想看到的人工智能发展远景。目前，人类对人工智能计算力的预测还过于保守。

上海智臻智能网络科技股份有限公司创始人、董事长袁辉表示：“从技术层面讲，人工智能面临‘卡脖子’问题，这不是卡哪个国家的脖子，而是卡人工智能体系、人工智能技术发展核心的‘脖子’。”他认为，目前人工智能技术体系和技术框架并没有摆脱或者突破过去60多年的技术积累。“人工智能技术在未来应该实现更大的突破，这是全世界科学家所面临的共同挑战。”

开放开源让应用效率提升

百度集团首席技术官王海峰表示，未来人工智能行业会出现越来越多的融合创新。

他说，人工智能技术发展到今天，从科学研究的角度讲，面临的问题越来越复杂，但是应用门槛会越来越降低，更多人可以不用关心核心算法，通过一个好的平台或者模型库就能解决应用问题。

王海峰在介绍百度自主研发的我国首个开源开放、功能完备的产业级深度学习平台飞桨时提到，飞桨目前已经有260多万开发者，这些开发者不需要每个人都从第一行人工智能的算法代码开始写起，而是直接调用框架。在类似平台上，人工智能应用门槛大幅降低，与此同时，这也推进了人工智能更快地实现应用推广，更快地推进产业智能化。

近年来，深度学习开源平台正在成为中国各行各业迅速布局人工智能的重要选择。最新数据显示，在中国深度学习平台市场综合份额中，谷歌、百度、脸书稳居前三，占据70%以上市场份额。其中，百度占比提升3.38%增速第一，综合市场份额位列第二，与位列第一的谷歌几乎持平。

> 未来的5至10年是人工智能非常大的红利变现期。这一阶段，各行业将充分应用过去65年的技术红利，争相进行人工智能的落地，从而推进各个产业的发展。
>
> **袁辉**
> 上海智臻智能网络科技股份有限公司创始人、董事长

推广落地促进各产业发展

在行业应用方面，袁辉表示：“未来的5至10年是人工智能非常大的红利变现期。”这一阶段，各行业将充分应用过去65年的技术红利，争相进行人工智能的落地，从而推进各个产业的发展，包括医疗、教育等。

“在新冠肺炎疫情期间，当人与人无法面对面交流时，人工智能应用解决了人们生产生活中的刚需问题。”科大讯飞高级副总裁杜兰介绍，作为人工智能龙头厂商，在疫情期间收获人工智能红利的科大讯飞备受关注，在国内疫情期间，第二波输入性病例增加的时候，常常能看见机场、海关工作人员和社区工作人员拿着科大讯飞翻译器和国外来宾进行沟通。科大讯飞电话机器人还参与了武汉900万人的排查，并还用于韩国的疫情防控。

“人工智能的技术与应用很大程度上提升了数据的价值。”杜兰特别强调了人工智能的数据安全问题。她表示，目前个人信息保护法和数据安全法从讨论的阶段，已经慢慢过渡到正在制定和出台的阶段，这将带来非常好的产业发展前景。

“目前人工智能接触的主要是生物特征相关数据，这些数据应该提前做好分层管理，并且由具有涉密资质的机构规范操作，经过脱敏化处理后再进行分享。”杜兰说。