



解决AI“高分低能” 需要升级的是考核机制

◎本报记者 华凌

近日,有媒体报道,目前部分人工智能沉迷刷榜,在基准测试时高分通过,表现优异,但实际应用中却还会犯一些非常基础的错误。

AI模型哪家好,基准测试来说话

AI模型应该如何衡量其性能?
“目前AI模型能力的高低取决于数据,因为AI的本质是学习数据,输出算法模型。为了公平衡量AI能力,很多机构、企业甚至科学家会收集、设计不同的数据集,其中一部分喂给AI训练,得到AI模型,另外一部分数据用于考核AI模型的能力,这就是基准测试。”近日,西安电子科技大学电子工程学院教授吴家骥接受科技日报记者采访时表示。
吴家骥介绍说,机器学习越来越多地用于各种实际应用,例如图像和语音识别、自动驾驶汽车、医学诊断等。因此,了解其在实践中的行为和性能变得非常重要。其鲁棒性和不确定性的高质量估计对于许多功能至关重要,尤其是在深度学习领域。为掌握模型的行为,研究人员要根据目标任务的基础来衡量其性能。
2010年,基于ImageNet数据集的计算机视觉竞赛的推出,激发了深度学习领域一场算法与数据的革命。从此,基准测试成为衡量AI模型性能的一个重要手段。微软公司的计算机科学家马塞

洛·里贝罗表示,基准测试应该是从业者工具箱中的一个工具,人们用基准来代替对于模型的理解,通过基准数据集来测试“模型的行为”。
例如,在自然语言处理领域,GLUE科研人员让AI模型在包含上千个句子的数据集上训练,并在9个任务上进行测试,来判断一个句子是否符合语法,并分析情感,或者判断两个句子之间是否是逻辑蕴含等,一度难倒了AI模型。随后,科研人员提高了基准测试难度,一些任务要求AI模型不仅能够处理句子,还要处理来自维基百科或新闻网站的段落后回答阅读理解问题。仅经过1年的发展,AI模型的性能从不70分轻松达到90分,超越了人类。
吴家骥表示:“科学研究要有科学问题、方法、计算、试验对比等要素。因此在进行科学研究,包括人工智能的科研中,也必须有计算与试验对比,也就是说AI算法的能力应该是可测量的,目的是验证研究方法的可行性、有效性。因此,基准测试很有必要,这样才能公平验证AI算法能力的高低好坏,避免各说各话,‘王婆卖瓜自卖自夸’。”

应用广泛,需尽快建立国家标准

美国麻省理工学院 Cleanlab 实验室的研究指出,常用的10个作为基准的数据集中,有超过3%的标注是错误的,基于这些基准跑分的结果则无参考意义。
“如果说,基准测试堪称人工智能领域的‘科举制’,那么,‘唯分数论’输赢,是不可能训练出真正的好模型。要打破此种现象,一方面需要采用更全面的评估方法,另一方面可以考虑把问题分而治之,比如用多个AI模型解决复杂问题,把复杂问题转化为简单确定的问题。简单且经过优化的基线模型往往优于更复杂的方法。谷歌的研究人员为常见的AI任务引入了不确定性基线库,来更好评估AI应用的稳健性和处理复杂不确定性的能力。”远望智库人工智能事业部部长、图灵机器人首席战略官谭茗洲指出。
虽然行业正在改变对于基准的态度,但目前基准测试研究仍然是一个小众研究。谷歌在一份研究中采访了工业界和学术界的53位AI从业者,其中许多人指出,改进数据集不如设计模型

特殊的软硬件设置对模型进行调整和处理,让AI在测试中表现出色,但这些性能在现实世界中却无法施展。”西安电子科技大学研究员尚坤指出。
在智能手机领域,我们谈及手机的使用体验时一般都不会涉及手机的性能表现,这些性能通常会用跑分成绩来表现。然而,我们常常会遇到一款手机的跑分成绩处于排行榜领先水平,但是在实际使用过程中却出现动画卡顿、页面滑动卡顿、应用假死等的现象。全球顶级评测网站 AnandTech 的一篇报道曾对这种现象提出质疑,指出某品牌手机跑分时启动了“性能模式”,而在平时的使用中“性能模式”很少被调用开启。这种处理方式虽然能够获得高分,但是不能模拟用户真实的使用情景,这让基准测试不具有参考意义。
尚坤认为,针对上述问题,改进基准的方法主

要有成就感

要有:一种是增加更多的数据集,让基准变得困难。用没有见过的数据测试,这样才能判断AI模型是否能够避免过拟合。研究人员可创建一个动态数据集收集和基准测试平台,针对每个任务,通过众包的方式,提交他们认为人工智能模型会错误分类的数据,成功欺骗到模型的样例被加入基准测试中。如果动态地收集数据增加标注,同时迭代式的训练模型,而不是使用传统的静态方式,AI模型应该可以实现更实质的进化。
尚坤说,另一种是缩小实验室数据和现实场景之间的差距。基线测试无论分数多高,还是要用实际场景下的数据来检验,所以通过对数据集进行更贴近真实场景的增强和扩容使得基准测试更加接近真实场景。如ImageNet-C数据集,可根据16种不同的实际破坏程度对原有的数据集进行扩充,可以更好模拟实际数据处理场景。
谭茗洲表示,AI应用基准研究是构建国内统一大市场的内在需要,当前AI已经在国计民生的各类领域中得到广泛应用,更需要设立标准对AI模型进行全面有效的评估,片面追求和采用高分AI模型,可能会让模型在复杂极端场景下出现“智障”行为,并且可能由于训练和推理性能的低效,造成不良社会影响、经济损失和环境破坏。
谭茗洲强调,AI应用基准研究关乎国家战略。针对重要领域,建立我国自己的AI基准测试标准,AI数据集,AI模型评估标准等迫在眉睫。
据了解,西安电子科技大学的DvcLab也在AI基准测试这个领域进行了前瞻性研究,特别是针对AI应用基准测试中数据集的整体质量与动态扩展两个关键问题,正在开发可在线协作的数据标注与AI模型研发托管项目,并计划今年陆续开源,正在为构建国家AI基准评估标准体系进行积极探索。

算法最终服务实践,而非刷榜

有人说,高分是AI模型的兴奋剂。于是,有的人工智能为了取得好成绩而频频刷榜。
微软公司2020年发布报告指出,包括微软、谷歌和亚马逊在内的各种sota模型包含诸多隐含错误,比如把句子里的“what's”改成“what is”,模型的输出结果就会截然不同,而在此前,从没

有人意识到这些曾被评价还不错的商业模型竟会在应用中如此糟糕。显然,这样训练出的AI模型就像一个只会考试、成绩优异的学生,可以成功通过科学家设置的各种基准测试,却不懂为什么。
“为了获得好成绩,研究人员可能会使用

实时生成海量数字盲文,AI帮视障人士无障碍阅读

◎洪恒飞 周立超 本报记者 江耘

“西湖十景位于浙江省杭州市西湖区境内……”近日,在浙江杭州举行的之江实验室专场开放活动上,视障志愿者陆林松通过现场展示的智能盲文阅读器,准确读出了自己触摸到的盲文内容。
他所体验的盲文阅读器,是由之江实验室科研团队联合多家机构共同研发的集视、触、听等多类型感知为一体的盲文数字化智能设备,与之一并展出的还有盲文教学机及英语机,涵盖盲人学习、阅读、公共服务等三大核心应用场景。
“这三款盲文机搭载了团队全球首创的密集点阵电磁驱动毫米级精准触觉再现技术,以及国内领先的盲文数字资源AI实时生成技术、视听触同步感知技术。”之江实验室类人感知研究中心研究专家杨文珍介绍道,团队盲文数字化科研工作始于2011年,经过4轮技术攻关和更新迭代,产品性能日益完善,已经在全国21个省份地区试点推广。
一套盲文读物相比同样内容的普通读物,价格可贵数倍乃至更多。记者现场看到,一本盲文版的小学语文教材,不但纸张宽大,厚度也堪比一本字典。
“盲文书制作工序复杂,出版周期很长,信息承载量少,且易磨损不耐用。”之江实验室类人

感知研究中心工程专员陶文韬说,纸质盲文书依然是国内盲人学习知识和获取信息的主要工具,盲文数字化设备普及率非常低。
相关统计显示,我国视障人士人数达到1700多万,其中800多万为全盲,占世界失明人口的20%,每年会出现新盲人约45万。由于多方面原因,我国视障群体不会盲文的现象较为普遍,盲文师资同样短缺。
2017年,财政部、中国残联等多部门联合启动“盲人数字阅读推广工程”,促成1000台盲文电脑和盲文电子显示器被配置到全国100所盲人教育机构,但仍难以满足我国广大视障群体的巨大需求。
“在知识爆炸且快速迭代的信息时代,盲文数字资源严重缺乏,视障群体的数字鸿沟问题越显突出。”杨文珍表示,目前将人工智能技术用于盲文信息无障碍产品的还很少。结合我国视障群体的国情,团队应用自研的柔性界面触觉计量测试技术、视触融合感知技术以及盲文AI实时翻译算法等,专门研制了三款功能各有侧重的产品。
杨文珍介绍,团队提出了一种视觉、听觉和触觉同步刺激的数字化盲文学习方法,能够提高视障人士的盲文学习效率,设计了一种多感知信息匹配AI算法,能够输出文字、声音和盲文点位相同内容的信息,为视障人士无障碍学习盲文提供条件。
智能系统提高翻译准确率

盲文AI实时翻译算法是本软件的核心技术之一,可以实时生成海量盲文数字资源。其运行逻辑是首先将中文、英文等文字编码格式统一转换为Unicode格式,然后选择盲文类型,再调用盲文AI实时翻译算法,就可以输出国家通用盲文、现行盲文、全拼盲文、双拼盲文。
杨文珍
之江实验室类人感知研究中心研究专家

之一,可以实时生成海量盲文数字资源。”杨文珍介绍说,其运行逻辑是首先将中文、英文等文字编码格式统一转换为Unicode格式,然后选择盲文类型,再调用盲文AI实时翻译算法,就可以输出国家通用盲文、现行盲文、全拼盲文、双拼盲文。
汉盲翻译是把汉字源文本自动转化为对应的盲文文本,目前存在多音字混淆、未登录词不能增加、不符合盲文分词连写规则等问题。记者

情报所

高精度智能线虫诞生 可蠕动前行

◎本报记者 付丽丽

近日,在2022智源大会开幕式上,北京智源人工智能研究院(以下简称智源研究院)发布了天演团队最新科研成果高精度智能线虫天宝1.0。画面中,一条“秀丽隐杆线虫”(以下简称秀丽线虫)正在蠕动前行,就像真的一样。这一幕,让屏幕前的很多观众瞪大了眼睛。

智源研究院生命模拟研究中心负责人马雷介绍,通过计算秀丽线虫神经电动力学信号,将其传到肌肉模型,通过肌肉模型不断地传导到前面与环境互动的部分,然后通过线虫的肌肉和水流环境的互动,就可以实现蠕动前行。

“现在,它只是初步表现出类似生物线虫的趋利(食物)避害(毒物)能力,下一阶段将逐步实现拐弯、避障、觅食等复杂智能行为。”马雷说。

据了解,作为一种无毒害、可独立生存的最小模式动物,秀丽线虫结构简单却功能完备,仅约1毫米体长、302个神经元,却足以完成感知、逃逸、觅食、交配等一系列通用智能行为。

这只透明小生物,是科学界研究的“常客”,近20年来3次诺贝尔奖都与它有关。它是“最简单的生命智能体”,也是通过生物神经机理模拟实现通用人工智能的最小载体。

2020年10月,《自然》子刊发表了一篇关于神经元的论文,媒体在报道时使用的标题是“19个神经元控制了一辆自动驾驶汽车”。

“这虽然有些夸张,但也反映了生物神经元计算的巨大潜力。”马雷说,之所以要做高精度的秀丽线虫神经元模型,是因为生物神经元表征生物智能性,精细化程度越高,意味着智能水平也越高,通过实现更高精度的秀丽线虫神经系统,可帮助深挖生物智能更多潜力。

“从事智能生命的模拟,可以说是一个长远但是又充满希望的工程。”智能研究院院长黄铁军说。

目前,高精度秀丽线虫模型已突破了神经系统与肌肉动力学结合的关键技术,有望推动生物神经元精细模拟进入新领域。“我们的目的是通过构建生命智能模型挖掘生物智能机理,启发和探索新一代人工智能。”黄铁军强调。

这是目前生物精度最高的仿真秀丽线虫。马雷介绍,他和团队已完成了秀丽线虫全部302个神经元及其连接关系的精细建模,对其中106个感知和运动神经元进行了高精度建模,整体精细程度达到国际最高水平;神经元结构模拟采用多胞室模型,单神经元胞室最多2313个,神经元生理模拟支持14种离子通道,实现了亚细胞级别的精细突触位点连接。

“环境是智能的摇篮。”马雷说,极致的环境模拟对于数字智能生命研究至关重要。环境变了,对应的生命体的结构、行为、智慧,复杂程度都不一样。
为此,团队还实现了仿真秀丽线虫与仿真环境的全闭环模拟,训练出由高精度神经系统控制的、与环境实时交互的仿真秀丽线虫,使其能够像真实线虫一样嗅探并控制身体蠕动到感兴趣的目标。

“高精度智能线虫的诞生,也许只是向生命智能迈出的第一步,却代表了一种计算智能生命的诞生。”黄铁军说。

黄铁军表示,未来,天演团队还将致力实现多个神经系统规模的生命数字孪生,从果蝇、斑马鱼、小鼠、直至最高智慧的人类大脑,沿着模式动物的神经系统规模进化路线,把生命模拟的规模做得越来越大。

“当然,规模大仅仅是一方面,更重要的是精细,要贴近生物,逼近生物,才能重新认识生物智能机理,打造从生命智能到通用人工智能的可行性路径,从而对新一代人工智能真正起到支撑作用。”黄铁军说。

图说智能

“镜像”人机



受疫情影响中断两年线下活动后,今年汉诺威工博会恢复线下举行。氢能飞机、外骨骼技术、智能机械臂等科技前沿产品格外引人注目。图为近日在德国汉诺威工业博览会上,一款智能机器人与观众互动。该智能机器人可感知一定空间范围内人的手势,并作出相同的手势。
新华社记者 任鹏飞摄