

AI能否突破大语言模型和图像识别的天花板 智能进化,从空间开始

面对席卷全球的AI浪潮,人工智能探索的先锋人物李飞飞再次将目光投向更远的前沿——她坚信,未来十年AI的真正突破将来自于“世界模型”与“空间智能”的构建。

■ 潮声 | 执笔 严粒粒

当人工智能在文生文、文生图、文生视频领域大展拳脚时,科学家们却开始焦虑起它的未来。

李飞飞,美国国家工程院、美国国家医学院、美国艺术与科学院三院院士,深耕人工智能领域前沿研究20多年,是现代人工智能的关键催化剂——大数据库ImageNet的创建者。近日,她在社交媒体公开发布的万字长文《从语言到世界:空间智能是AI的下一个前沿》,在学术界、产业界、投资界引起广泛讨论。

文中,她尖锐指出:当前以大语言模型和图像分类器为主导的人工智能时代已经达到了极限,而具备空间智能的AI将突破这一瓶颈。

超越“黑暗中的文字匠人”

万字长文中,李飞飞将人工智能划分为两种:说话的智能和做事的智能。

而今,各种大语言模型输出可靠文本的能力已经非常强大,但也许你已经注意到一个问题——AI有时比人还“蠢”。

在社交平台上,许多人将人工智能调侃为“人工智障”。

一位网友“喂”给豆包一张3个小朋友在沙漠的合照,要求在沙漠的远处添加两只清晰写实的骆驼。很快,一张令人啼笑皆非的效果图出来了:两只骆驼站在孩子身后,大得像猛犸象。豆包还顺便“奉送”了一只脖子和长颈鹿一样长的苍狐,以及一只兔子。

甚至连最新的多模态大模型,在这一方面的表现也不尽如人意。

Sora是OpenAI旗下的文生视频大模型。这个让许多动画、影视行业从业者大呼“抢饭碗”的人工智能,也闹出过大笑话。2024年春节期间,Sora首发。项目研发团队成员阿迪蒂亚·拉梅什(Aditya Ramesh)发布了一段蚂蚁在蚁巢内部移动的视频。粗看效果惊艳,细看啼笑皆非——蚂蚁居然只有四条腿。

这回,杨立昆(Yann LeCun)和加里·马库斯(Gary Marcus)这对在人工智能领域观点对立的学者也统一战线,分别在社交媒体发表评价。

他们一个揶揄:“嗨,Aditya,蚂蚁有6条腿,不是吗?”一个深深担忧:“今天它弄错的可能是一只蚂蚁,明天就是月球的轨道。”

前不久,升级后的Sora2面世。画面流畅度再一次提高,可各种细节衔接却还是生硬。官方放出的最新演示视频中,男子挥舞的棍子上的流苏,一会儿有一会儿没,不符合常识。

是哪里出了问题?

在早前的一次播客访谈中,李飞飞这样解释:语言从根本上说,是人工信号,输入输出过程难免有信息损耗;但是真实世界遵循物理规律,自成体系,所见即所知。

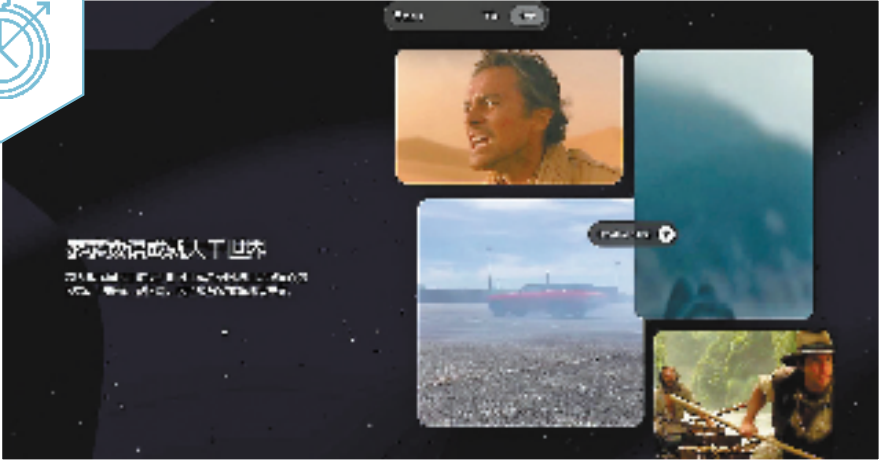
正如她在万字长文中的开宗明义:如今,诸如大语言模型等领先的人工智能技术已经开始改变我们获取和处理抽象知识的方式。然而,它们仍然是身处黑暗中的文字匠人;它们能言善辩却缺乏经验,知识渊博却缺乏根基。

西湖大学博士生导师、空间智能和机器人实验室负责人刘沛东认为,“当前AI系统的核心瓶颈之一,正是对物理世界缺乏深度的、常识性的理解。这一根本缺陷直接体现在模型的输出上:无论底层是纯文本还是多模态模型,其生成内容都可能因为缺乏物理逻辑这个‘内在锚点’而变得不合常理。”

“人工智能发展的终极愿景之一,是能够响应‘去买杯咖啡’这样的高层指令。这意味着机器需要具备自主拆解任务、规划步骤并应对环境变化的能力。”刘沛东说。

那么,人工智能版图缺的这块拼图是什么?

李飞飞认为,是空间智能。



△ 去年文生视频大模型Sora一上线,便引来全球从事文字、音视频以及科技的工作人士的热议。
Sora网站截屏

◁ 空间智能一旦实现突破,被视为将在自动驾驶等领域得到广泛应用。图为在深圳市坪山区道路上行驶的无人驾驶车辆。
新华社发

让AI读懂物理世界

空间智能的基本概念,最初来源于教育心理学。

上世纪八十年代,霍华德·加德纳(Howard Gardner)在多元智能理论中指出,空间智能指对空间信息进行感知、理解与操作的综合智能,涉及三维思维、图形敏感性及空间想象等维度。这是一种人类从婴儿阶段就开始发展的本能。古希腊的埃及托色尼将影子转化为几何测量,通过亚历山大与赛伊尼两地的日影夹角计算出地球周长;哈格里夫斯发明的“珍妮纺纱机”通过将纺锤并列放置的空间构想,使单人工效提升八倍,革新纺织业;沃森与克里克通过操纵三维分子模型,让碱基对的空间排列豁然开朗,进而揭开了DNA结构之谜……

在人类文明进步的许多关键时刻,总有空间智能的身影。它让人类不仅“看见”,还能“看懂”,进而驱动想象、推理并完成创造。

后来,人们将空间智能概念迁移至科技界,希望机器也能拥有这项能力。

“其实从历史发展维度看,空间智能一直以来都是AI研究的重要方向,科学家们的探索早就开始了。”群核科技首席科学家周子寒举例,“早在20世纪70年代,计算神经学的创始人大卫·马尔(David Marr)的遗作《视觉》就总结了其研究成果,并首次系统提出了3D世界表征的框架,定义了从图像恢复三维世界的一系列核心任务。”2022年,在空间智能概念逐渐升温之时,这本书的中文版面世了。

目前,这一技术早已在实验室酝酿,并走向产业。

手机内置的“AR测量”工具、沉浸式交互工具VR眼镜、扫地机器人,甚至连可以和人握手、打招呼的机器人和机器狗都具备了某种程度的空间智能。不过,它们或需要人为操控,或对空间适应能力不强——例如,环境杂乱、物体识别不清等因素,都会直接造成扫地机器人传感器误判,导致猫屎糊一地、猛撞落地窗等尴尬问题。

在科学家们看来,自动驾驶是空间智能目前规模最大、最成熟的应用之一。可即便最新的自动驾驶系统,也极易在交错复杂的城市街道上造成安全隐患。

“理想的空间智能,构建于‘感知—推理—行动’的完整行为链之上。我认为,当前技术在感知(识别颜色、形状等物理属性)和执行(具身机器人的基础动作)层面已相对成熟,真正的挑战在于中间的‘推理’环节——即缺乏一个能够理解物理规则、进行因果预测并做出合理规划的内部模型。这正是当前研究与产业攻关的焦点。”刘沛东说。

如果用更形象的例子解释,那就是



李飞飞在2017年中国计算机大会上做特邀报告。

新华社发



大卫·马尔(David Marr)的《视觉》中文版

自动驾驶的汽车可以识别常规路障、红绿灯、移动的行人和车,却难以在摄像头和雷达“看到”路上突然出现的一个小球时,“推测”出下一秒的潜在危险——球绝不会凭空出现,大概率是人踢出来的。人类驾驶员此时便会预测有小朋友马上跑上马路,进而紧急刹车。

上一轮AI公司做的空间智能,更多的是基于图像或视频理解生成的智能。周子寒认为,“李飞飞说的‘空间智能’,补上了当前以‘语言+图像’为主的大模型在‘物理世界理解与创造’方面的短板。如果真的能实现突破,将在自动驾驶、具身智能、创意设计、虚拟现实、科学模拟等领域有广泛的应用空间。”

“世界模型”成必争之地

学界普遍认为,推动人工智能发展的“三驾马车”是数据、模型、算力。

如何让人工智能在空间智能上得到

突破?

是数据吗?

专家们早有共识:能训练出更优性能的高质量语言数据可能在2026年左右被耗尽。

是算力吗?

今天,训练一个GPT-4(OpenAI在2023年为ChatGPT发布的语言模型)需要成千上万个高端GPU运行数月,成本可能超过1亿美元。可人工智能发展史上,“开启深度学习时代新纪元”的Alexnet模型,不过是用2块英伟达的消费级GPU、训练了三五天的结果。

在算力和数据有限的前提下,模型成为关键。

在人工智能领域,许多重要的模型都是模仿人类的某种能力或大脑的某种机能建立的。例如,卷积神经网络的灵感来自生物视觉皮层感受域的工作方式,Transformer模型则借鉴了人类注意力的聚焦机制。

同样,2018年才提出的“世界模型”,灵感也源自人脑一种重要机制——在大脑中构建并运用“内部世界”的能力。通俗地讲,这是一种让人工智能像人类一样能提取、整合感官获取的抽象信息,进而对物理世界进行推理和预测的工具。换言之,理想状态下,拥有了这套工具,人工智能完全可以靠直觉预判咖啡杯滑落轨迹,并敏锐地伸手凌空截住,而不一定要“过”一道复杂的信息提取、概率计算和统计。

世界模型,进而成为近年来科学与产业界投身的全新“必争之地”——

1月,芯片巨头英伟达CEO黄仁勋身着新皮衣亮相2025年美国拉斯维加斯消费电子展(CES),除了推出最新的GPU,还宣布推出可预测和生成“物理感知”视频的Cosmos世界模型。

3月,拥有全球最大室内场景空间数据集的空间智能企业群核科技,开源了空间语言模型SpatialLM,旨在帮助机器具备空间认知与推理能力。SpatialLM开源后便登上全球领先的AI开

源社区平台HuggingFace模型趋势榜第二位。

8月,谷歌DeepMind发布了支持“实时交互”的世界模型Genie-3,为智能体训练提供了更广阔的模拟空间,还为游戏开发、教育和创意设计等领域带来了新的可能性。

就在前不久,李飞飞参与创建的公司World Lab发布了首款多模态世界模型产品Marble,支持用户通过输入文本、照片、视频、3D布局图或全景图,生成可编辑、可下载的3D环境。

刘沛东推断:“大语言模型在抽象推理上的成功,验证了‘基础模型’技术路线的巨大潜力。这促使业界开始瞄准下一个前沿——赋予AI对物理世界的理解与交互能力,即空间智能。而‘世界模型’被视为实现这一目标的根本路径,因此吸引了产业界的扎堆投入。”

从图灵在1950年提出“如果计算机能让与之进行打字对话的人相信它是人类,这可能就是智能的标志”算起,到ChatGPT等各种大语言模型大爆发,时间跨度70多年。

其间,因为人们对人工智能发展速度持有不切实际的幻想,导致人工智能两度陷入低迷,直到近10年才成为一种全球现象,以及技术、经济乃至地缘政治的转折点。

世界模型的搭建,目前仍有许多亟待突破的难点。

科学家们认为,真实世界具有极大的复杂性与不确定性,许多现象是由众多因素相互作用的结果,目前的数学模型很难全面地处理这种复杂的多因素系统;

现有数据具有局限性,如对地球内部的构造和运动等难以取得的数据,会限制模型的准确性和完整性;

物理学本身也具有局限性,暗物质和暗能量等许多物理现象难以直接观测和测量,会使得构建真实世界模型时缺乏足够的实验数据支持;

……

科技的进步,总是要历经坎坷的。

打造一个像人类一样去干活的人工智能,愿景美好,但挑战才刚开始。

链接

AI的下一颗北极星

“空间智能”并非李飞飞一夜之间想出的概念。在她去年出版的自传《我看见的世界》中,就已对当时如火如荼的大语言模型发展提出了深层思考。我们节选了该书《下一颗北极星》章节中的观点,有助于大家理解她提出“空间智能”的思想脉络:

每一个故事都代表着一个变化——我们对数据的看法发生了变化,对数据的期望也发生了变化。我们曾经试图给算法类似百科全书式的意识,希望算法可以识别所有的类别和事物,而现在,我们的目标更为广泛。我们想对万事万物所蕴含的空间、时间甚至意义有更深入的了解。我们的目标不仅仅是数量上的增加,还有细节和细微差别的扩展。新的数据处理方法不仅仅是简单的整理和编目,而是要模拟整个环境,模拟在环境中展开的行动。这就是为什么随着技术复杂性出现爆炸式增长,我们研究背后的人本主义也在不断发展。要形成对现实生活的整体观,为了创造比以往任何时候都更加真实的世界表征,我们需要深度和保真度,而在我看来,即使目前最先进的技术也达不到这种需求。因此,我们再次热血沸腾,迎接挑战。我们再次需要进化。

……

大型语言模型,即使是多模态的大型语言模型,可能也并不具备真正意义上的“思考”能力。看看就知道了:大型语言模型很容易出现荒谬的概念性失误,也乐于编造听起来合理但实际上毫无意义的胡言乱语。了解这些事实有助于我们避免过分迷恋模型的能力。然而,随着大型语言模型生成的文本、图像、语音和视频越来越复杂,真与假之间的界限愈加模糊。越来越多的评论家开始质疑,为我们敲响警钟:作为个人、机构,甚至社会,我们究竟有没有能力区分真实和虚构?当人们意识到这一切还只是1.0版本时,这种发问尤其令人警醒。

就这样,科技不断发展。算法语言表达的高级程度已逼近人类水平。机器人正在逐渐学会应对真实的环境。视觉模型不仅可以通过照片进行训练,还可以在全三维世界中进行沉浸式实时训练。人工智能能够识别内容一样流畅地生成内容。与此同时,伦理问题在我们周围不断涌现,与人类经济社会发展的关联也日益紧密。但这就是科学一直以来的样子。随着旅程的展开,前路只会变得更漫长、更复杂。无穷无尽的分叉、不断扩大的视野、新的发现、新的危机、新的争论,故事永远处于第一幕。

(整理自《我看见的世界:李飞飞自传》)

