研究人员利用人工智能的深度学习方法，发现了38个新的强引力透镜候选体，为研究天体物理学问题提供了新的“宇宙探针”候选体。


视觉中国供图

帮天文学家“大海捞针” 人工智能有了新办法

本报记者 **赵汉斌** 通讯员 **陈 艳**

近年来，随着技术日益进步，天文学研究中产生了海量数据。天文学家要想从郭守敬望远镜、“中国天眼”FAST、LSST大型综合巡天望远镜等遍布世界的大型望远镜捕获的海量数据中找出有价值的信息以资研究，无异于大海捞针。

如何高效地处理这些数据，已成为现代天文学面临的一项重要挑战。由于人工智能在海量数据分析和处理方面所具有的突出优势，它也很自然地走入了天文学家的视野。

日前，中国科学院云南天文台丽江天文观测站龙潜研究员与云南大学中国西南天文研究所宇宙学研究组尔欣中教授团队合作，利用人工智能深度学习的方法，发现了38个新的强引力透镜候选体，为研究天体物理学问题提供了新的可靠的“宇宙探针”候选体。英国《皇家天文学会月刊》发表了这项研究成果。

天文观测产生海量数据 用机器学习给天体分类已十分普遍

随着下一代大规模测光巡天项目的开展，人们期待发现数以万计的强引力透镜系统。但如何在海量天体图像中快速地找到强引力透镜候选体？近年来，人工智能的快速发展，给人类提供了一种新的可能。

以2009年发射升空的世界首个用于探测太阳系外类地行星的飞行器开普勒太空望远镜为例，仅在起初3年半的任务期内，就监控了超过15万个恒星系统，同时也产生了海量数据。这些数据通常要经由计算机处理，但当计算机识别出一定的信号时，又必须依靠人类分析，判断其是否是行星轨道所产生的，这项巨大的筛查工作单靠美国国家航空航天局（NASA）的科学家或科学小组，是无法有效完成的。

“如此大的数据量，人工分析在很多时候已经达不到所需要的速度。借助人工智能的优势，我们可以极大地提升对数据的分析速率。”龙潜向科技日报记者介绍，人工智能展现出来的效率和准确性远高于传统方法。

龙潜研究员长期从事人工智能深度学习方面的研究。近期，他与尔欣中教授团队合作，构建并训练了一个卷积神经网络，用来寻找强引力透镜系统。他们把这个网络应用到欧洲南方天文台2.6米巡天望远镜（VST）千平方度巡天数据，并找到了38个新的强引力透镜候选体。此次构建的神经网络，也可应用于其他大型望远镜的巡天数据。

“在这项工作中，我们用计算机分别模拟了强引力透镜图像和非强引力透镜图像，从而来训练计算机。我们发现，在准备训练计算机的图像时，非强引力透镜图像比强引力透镜更加重要。”尔欣中说，开始的分析中，他们使用简单的规则星系图像作为非强引力透镜训练样本，发现结果正确率非常低。只有把各种可能的非引力透镜图像都考虑进来之后，才能得到比较好的结果。

“这就像在教电脑认识什么是狗的时候，还要告诉它猫、羊、牛等都不是狗。而如果你只告诉它猫不是狗，电脑有非常大的概率把羊、牛认成狗。”龙潜说，目前利用机器学习来对天文学中各种天体分类已经非常普遍，最简单的是把恒星和星系分开，或者把不同行态的星系进行分类，以及利用星系的多重颜色来估计星系的距离等。

每秒可识别上万张照片 新型神经网络便于实时修改、训练和测试

人眼看强引力透镜系统的图像，最快就是每秒钟看一张图。而计算机每秒钟可以识别成千上万张图片。

龙潜研究员和尔欣中教授团队此番训练的这个卷积神经网络，可以充分利用GPU进行并行加速，通过装备更多或更强的GPU，系统可以根据实际需要极大提升搜索速度和效率。

“这个神经网络的训练，主要使用模拟数据，只使用了很少的人工标注数据，由于模拟数据可以任意生成，因此多样性远大于人工标注数据，进一步根据数据的特点调节训练参数和训练算法，使神经网络的泛化能力得到了极大的提高。”龙潜说，此外，研究人员使用新型科学计算语言Julia完全自定义网络结构，由于Julia语言兼具速度和灵活性，使得神经网络在CPU和GPU上都有良好的性能，并且可以任意切换，因此非常有利于研究人员实时修改、训练和测试。

“我们还通过对引力透镜数据的研究，定制了有针对性的小型网络，有效地抑制了过拟合现象，同时实验证明该网络具有与大型网络相似的准确率。相比大型网络，小型网络在普通计算机终端就可以训练和测试，不需要依赖大型GPU集群，这为天文工作者使用和改进网络提供了便利。”龙潜说。

目前，随着技术与装备水平快速发展，人工智能在天文学上的应用还会越来越多。“我们计划对一些变源的多波段光变曲线来进行机器的快速分类，这样在实施大样本巡天的时候，电脑可以自动对所发现的变源进行筛选，并对我们感兴趣的天体做出提示，以便进一步开展后续研究工作。”尔欣中说，正因为人工智能的帮助，天文研究者得以从耗时单调的数据筛查分析中解脱出来，当人力“大海捞针”难以招架之日，正是人工智能大显身手之时。

延伸阅读

AI从旧数据中识别出50颗新行星

近日，由英国华威大学的大卫·阿姆斯特朗（David Armstrong）领导的研究团队开发了一项新的机器学习算法，可以从NASA的旧数据中识别出系外行星——即太阳系外的行星。该团队已通过这一工具对一批潜在行星进行了识别，并从这些天体中确认出了50个新的行星。该研究的论文发表在《皇家天文学会月刊》上。

天文学家有2种方法可以用来探测系外行星。一种是径向速度方法，它用来监测恒星是否有行星引力引起的小反运动。第二种是更敏感的技术，也是凌日系外行星巡天卫星和开普勒采用的技术，它主要依靠宿主星的亮度变化。如果一个恒星系的平面对准正确，从我们的角度看，它的行星就会在恒星前面过境。通过监测这些亮度的变化，我们可以很有把握地推断出系外行星的存在。问题是，第二种方法产生了大量恒星的亮度数据，其中许多恒星不会有可见的系外行星。这就需要计算机分析和人工相结合，才能确定候选星并确认它们的存在。

论文作者在摘要中写道：“我们的模型只要短短几秒就能对数千个肉眼看不见的候选行星进行识别，确认其是否真的是行星。”考虑到许多天文学数据库的规模都大得惊人，该方法有望大大提高人们探索世界的效率。

这一算法的原理是将真假行星区分开来。阿姆斯特朗在一份声明中说：“我们现在不仅能说哪些候选行星‘更可能’是行星，而是可以用确切的数据说明这种可能性有多大：如果候选天体是‘假行星’的可能性小于1%，就可以被确认为是真正的行星。”

研究人员并不是随便打开一个开关，就能让人工智能通过数据筛选来发现行星。他们必须用已确认的系外行星和假阳性的数据来训练神经网络，这样它才能在新的数据中识别出那些明显的迹象。华威大学确认的50颗系外行星中，从海王星大小的气体巨行星到比地球还小的岩石世界，无所不包。而使用传统方法确认较小的行星存在一定困难，这也说明了人工智能在确认较小行星方面的潜力。

根据新的研究，在所有确认的系外行星中，大约有三分之一是用单一的分析方法确认的，这并不理想。科学家们说，即使现有的技术能够发现所有可观测到的系外行星，我们也应该有更多的选择。他们希望新的机器学习系统在检测更多行星的过程中不断发展，成为系外行星探索过程中的重要组成部分。

一天在屏幕上标200万个点

数据标注师：我们就像AI的“幼儿教师”

新华社记者 **马晓媛 梁晓飞**

“都说数据是人工智能（AI）时代的石油，我们的工作就是把原油炼成汽油。”

“我们就像一个‘幼教’，教AI更好地认识数据。”

……

聊起数据标注师这份职业，“90后”李宇龙显得格外兴奋。虽然从业仅4年，但他已经是一名资深的数据标注师。

数据标注师是随着人工智能的发展而出现的新职业。人工智能练习认知需要大量经过标注的数据，数据标注工作最早由AI工程师完成，随着人工智能所需数据量的不断增加，数据标注逐渐独立成为新的工种。

“数据标注有时候就像玩游戏。”李宇龙最近正做一个自动驾驶的数据标注项目，工作内容是对照一张2D街景照片，在相应的3D点云图上框选打点。

“你看，把汽车框起来，都打成白色的点，就代表这是一个障碍物。”随着鼠标快速滑动，屏幕上的点云图不断翻转，一个个针尖大的数据点被标注在图中不同物体上——蓝色是路面、绿色是绿植、红色是路沿、白色是障碍物。

李宇龙说，像这样一张普通的点云图，大约要标注18万个点，一个熟练的数据标注师只用半个多小时就能完成，“这样算下来，一天标200万个点不成问题”。

李宇龙原本在一家印制电路板的外资企业工作，偶然机会下接触了数据标注行业，便投身其中。他说，与传统产业相比，这份职业有种“科幻感”：传统行业的原料、产品都看得见、摸得着，而数据标注师只需要一台电脑、一根网线，原料是数据，产品也是数据。

然而，这份“科幻”的职业却实实在在地改变着现实生活。自动驾驶、人脸支付、智慧医疗、智能家居……人工智能正在给生活带来越来越多的便利，这背后都有着数据标注师的功劳。

“虽然我们从事的是人工智能领域最基础的工作，却经常能体会到价值感。”李宇龙说，新冠肺炎疫情期间他和同事做了一个医疗项目，是在肺部CT片上标注病灶数据，以提高人工智能对病毒的识别能力。“平常医生看一张CT片需要几分钟，如果用改进后的人工智能算法作为辅助，几秒钟就能初步判断一张CT上是否存在疑似病毒。”

从事数据标注需要每天对着电脑，不免让人觉得枯燥。但李宇龙却说，数据标注为他打开了更大的世界，因为经常接触不同的项目，每个项目涉及的领域也不同，会经常带来新鲜感。

“更重要的是，这会是一个持续发展的行业。”李宇龙说，随着人工智能进入越来越多的行业领域，对数据标注的需求会更多、要求也会更高，数据标注行业的前景无限。

如今，仅李宇龙所在的百度（山西）人工智能基础数据产业基地，就有35家数据标注企业、2300多名数据标注师。百度智能云数据众包则拥有超过5万名线下采集员，超过2000万名众包互联网用户。

“每当有人问起我的职业，我回答数据标注师的时候，对方的脸上总是写满了问号。”李宇龙说，期待有越来越多人了解这个行业，“希望有一天，大家提起数据标注师，就像提起教师、医生一样。”

相关链接

山村里的他们，成了“机器人饲养员”

新华社记者 **梁爱平**

27岁的白莹莹是初中学历，她从来没有想过，工作会和“高大上”的人工智能沾上边儿。在陕西省榆林市清涧县城一栋办公楼的大平面，白莹莹坐在电脑前，鼠标“咔哒”几下，一个选项标注完成。曾有人问她做什么工作，她回答“我是机器人的老师”。

其实，白莹莹是“机器人饲养员”，又名人工智能训练师，这是一个“国家认定”的新职业。

白莹莹的家在清涧县双庙河乡安家畔村，作为建档立卡贫困户，她和丈夫曾经去煤矿打过工，卖过菜……她也很想通过自己的努力改变生活。2019年12月，清涧县从阿里巴巴和蚂蚁集团协调引进“AI豆计划”人工智能产业扶贫孵化项目，成立了县政府直属国有企业——清涧县爱豆科技有限公司，并开始招兵买马。

白莹莹一直认为“高大上”的人工智能“无所不知”。上班后才知道，人工智能之所以“聪明”是因为背后有许多像她这样的“老师”在手把手地训练它。比如，人工智能进行垃圾分类，用手机扫一扫某种垃圾，软件就能自动识别并帮助人们进行分类。其实，这背后是“机器人饲养员”们对玻璃、塑料、纸袋、调料罐等各类垃圾的图片进行分类标注后，人工智能垃圾分类模型才能不断提升识别的准确率，做好垃圾分类。

工作8个多月，白莹莹平均每月收入2000多元，最高一个月还拿到了3700元。其实，在清涧县爱豆科技有限公司的113名员工中，像白莹莹一样的困难群众有65人，他们因从事人工智能数据标注工作而改变了自己的生活。同时，这一新兴产业还吸引了不少外出务工人员返乡工作。

清涧县爱豆科技有限公司总经理鱼涛说，人工智能数据标注工作做得好、做得多就赚得多。目前公司员工的平均月薪超过了3000元。“他们可能并不懂人工智能，但他们知道，人工智能可以帮他们实现梦想。”

对话

时越：棋手与AI正在“教学相长”

新华社记者 **王浩宇 王镜宇**

围棋进入人工智能时代前，世界冠军时越九段每盘棋之后，最困扰的是如何找出输赢的关键。

“常常一盘棋不知道输在哪，或者是在哪个地方走错了，我如何能够去找出自己的问题所在，这个是AI出来之前我一直很有困扰的地方，有时候赢棋也不知道怎么赢的。”时越说。

自2006年在全国围棋甲级联赛首秀起，到如今正在进行的2020新赛季，时越在围甲留下了293战182胜的战绩。“我肯定希望能够下得更长久一点，”时越说，“希望能够去探知这个棋盘里面更多的东西。”

人工智能的出现让时越的探知过程中，在棋盘上发现了更多的可能。

“很多招法以前大家认为是很必然的一种想法，现在思路打开了，很多招法不像以前的固有定式，不是说这个局部下完一定要把它定型。现在AI就是很多招法是按这个全局（的考虑）来的，在一个局部下没有一个固定的招。这个是一个认识上的提升。”

如果以目数来算，时越认为棋手们比人工智能时代之前，实力普遍“能长个三四目棋”。“比如说现在的我和以前的我下，现在可能布局就会便宜不少。以前的招法以现在的认识来看，就有局限性了。现在大家肯定都涨棋了，自己的棋理能够涨，就是对棋的理解能够提升一个台阶。”

日本围棋大师藤泽秀行曾言：“棋道一百 我只知七”。在人工智能时代到来之后，大家意识到这并不完全是谦辞，时越甚至认为人对围棋的认知还到不了百分之七。

当如今的AI能让先甚至让两子击败人类顶尖高手，每一步棋都能通过AI的精确计算反应在胜率的变化上，是否意味着人工智能已能找到对弈中的最优解？

时越认为还远远达不到，“围棋的变化太多了。电脑其实只是比我们强，但是他远远到不了说把围棋里面的所有变化都给解析出来的。AI的训练和算法越多，可能水平越高，但距离围棋上帝、绝对真理还是很遥远的。”

“比如比赛中盘，不可能用AI把所有变化都去算清，布局可能大家都差不多，主要是拼中盘，中盘的战斗力是实打实的个人水平，比的就是大家各自对棋的不同理解。”

这位曾经的中国围棋第一人已年近30，他对围棋的理解，也已经从比赛胜负延伸至生活之中。

“万物都是有联系的，我是觉得围棋能够指导我的生活。比如，可能见到某一个棋形或者某一个局面下的一手棋，结合到生活中会有感触，互相印证。围棋没有绝对的正解，你会觉得这个时候，这步棋当前是合适的。在生活中遇到一些事，能不能找到一个合适的处理方式？围棋中很多种局面下的不同的选择，就能映射出这个人在生活中的样子或者他的一种风格。”

“一些围棋爱好者他不希望出现有一个标准答案式的东西，我认为这步就是最好的，你不要跟我讲AI怎么下，我第一步喜欢下天元，虽然你告诉我胜率跌了，但无所谓，我也不在乎。其实这个状态也挺好。”

情报所

广西首个无人驾驶轨道交通车辆基地建成

新华社讯（记者**齐中熙**）记者从中国铁建股份有限公司获悉，近日，由中铁十八局集团等单位参建的广西南宁轨道交通5号线全自动地铁车辆基地建成并投入使用。这是广西首个无人驾驶轨道交通车辆基地。

据中铁建北部湾公司南宁地铁5号线指挥长李小平介绍，该基地总建筑面积15.03万平方米，共铺设轨道51条。基地主要运用智能化、信息化、生态化等新技术，实现列车无人驾驶和车站机电设备的节能运行，大幅降低系统整体能耗和人力成本，提升运转效率。

李小平说，作为轨道交通的“神经中枢”，全自动驾驶作业综合管理系统全面提升了该基地的人工智能化水平，对推动广西轨道交通向智能化、生态化发展具有重要的示范带动作用。

图说智能

数字·健康小镇开园

近日，浙江省杭州市余杭区未来科技城举办2020生命健康未来峰会暨中国（杭州）数字·健康小镇开园仪式。

小镇重点布局基于人工智能、5G、区块链等技术赋能的生命健康产业，规划面积3.2平方公里，此次开园的小镇启动区块共10栋建筑，包含小镇客厅、企业研发总部、成果转化区、商业配套等。

图为在中国（杭州）数字·健康小镇的成果转化区，嘉宾在了解一套染色体人工智能诊断系统。 新华社记者 **黄宗治**摄