预测蛋白质结构的难度跟人工智能下棋绝不是一个数量级的,深度学习并不是所有难题的答案。有时候,方法思路比深度学习更重要。

生物界"AlphaGo"来袭?

AlphaFold"大胜"人类的"戏码"不太足

■本报记者 赵广立

继围棋、国际象棋等竞技项目之后,谷歌旗下专注于人工智能(AI)前沿技术的DeepMind团队展示了一项新成就,将其AI技术应用的边际拓宽至基础科学研究领域——蛋白质结构预测。AlphaGo家族也再次"扩军",迎来新成员"AlphaFold"。

当地时间12月2日,两年一度的国际蛋白质结构预测挑战竞赛(CASP)在墨西哥岛城坎昆举办。正是在这次大赛上,AlphaFold一鸣惊人。在与来自世界各地数百支参赛队伍的"较量"中,DeepMind团队以AlphaFold领衔的"A7D"参赛队在43个参赛蛋白中拿到25个单项最佳模型,并获得总分第一名,将第二名(该团队拿到3个单项最佳)远远抛诸身后。

你可能以为,这是"AlphaGo大胜人类围棋冠军李世石"在"蛋白质结构预测"领域的一场重演。但《中国科学报》记者经过寻访了解到,将神经网络模型、深度强化学习等AI技术用于未知蛋白质结构解析,DeepMind团队既非首创、亦非个例,甚至在此次大赛中排名前五的参赛团队中,都不同程度地使用了深度学习算法。

那么,AlphaFold缘何脱颖而出?此外,有专业人士指出,AlphaFold的此番"大胜"并不完美,那么,AI用于蛋白质结构预测,还有哪些不尽如人意之处?

AlphaFold虽然胜出,但它距离解决折叠问题还有差距。 图片来源:SELF格致论道官网

并不是一次完美的"大胜"

在结果揭晓的当天,谷歌同时发布了一篇供媒体参考的新闻稿件,标题醒目:《AlphaFold:用AI进行科学探索》(英文题目:*AlphaFold: Using AI for scientific discovery*)。

"DeepMind团队(的新闻稿)并没有披露,此次挑战赛的前五名都使用了深度学习技术,并且其他加入了深度学习的解构预测模型也很好。"巴黎笛卡尔大学前沿生物学博士郭昊天对《中国科学报》举例说,今年参赛的队伍中,很多都使用了CNN等深度学习方法,如拿到了第二名的密歇根大学的Yang Zhang(音:张扬)团队,就在其开发的I-TASSER结构预测模型(近10年来最流行的结构计算模型之一)的基础上,将卷积神经网络(CNN)与之整合、优化,提高了预测准确率。

"该方法与AlphaFold相比,准确率的差别其实相当微弱——前者预测结构和真实结构相符的概率为85.1%,只比AlphaFold的87.9%相差不到3%。考虑到DeepMind的计算力,这个边际效应非常小。"郭昊天说。

曾从事蛋白结构信息学及基于蛋白组学的系统生物学研究的国家纳米科学中心研究员方巧君也告诉《中国科学报》,从与待测蛋白质真实结构的相符程度来看,前几名的差别并没有那么大。

也就是说,AlphaFold之于其他团队的真实优势,并非如"25个单项最佳"与"3个单项最佳"这种数字上的反差那么强烈。

郭昊天告诉记者,早在AlphaFold面世之前,就有一些学者尝试使用神经网络和强化学习来完成模型预测中的"模拟退火"过程。

那么,是什么让此次AlphaFold能够在本次比赛中脱颖而出呢?

"谷歌有钱有TPU!"郭昊天略带调侃地说,DeepMind团队的优势在于"硬件的胜利",本次蛋白质结构预测挑战赛,确切地说应该是AlphaFold"大胜没钱的研究机构的其他深度学习算法"。

郭昊天解释说,DeepMind可以动用几千片TPU(张量处理单元,谷歌专为机器学习而定制的芯片,笔者注),这是一般科研团队所难以比拟的。"如果使用DeepMind的资源,重新训练模型,或许Zhang团队得到的结果比AlphaFold更好也未可知。"

距离"成功预测蛋白质结构"还差得远

同样参加本次挑战赛的英国科学家Liam McGuffin,观察到许多工作组使用各种机器学习方法试图预测蛋白结构,表达了他对AI给这个领域带来的推动的乐观情绪:"这几年来AI给这个领域带来了惊人的推动,也许在2020年左右,我们就可以基本上解决蛋白结构预测的问题。"

基于此,有评论称:"结构生物学的春天来了。"

方巧君有着不同的看法。"AlphaFold虽然胜出,但是我们也看到它距离解决折叠问题、距离实际运用还有差距。"她告诉记者,实际中待解的蛋白分子都比较大,而比赛中看到的蛋白质只有100个左右氨基酸,"说到2020年就可以基本解决问题有点太乐观了"。

与方巧君持同样观点的还有哈佛大学医学科学博士袁博以及在美国布鲁克海文国家实验室"用机器学习做生物信息"的在读博士Z。他们认为,AlphaFold距离"成功预测蛋白质结构"还差得远。

"在结构生物学领域,这毫无疑问是一项巨大的突破,但也掀起了很多质疑和担忧的声音。事实上,AlphaFold的模型还没有达到极高的准确率,在一些传统模型可以解决的案例中,反而达不到预期的效果。"袁博对《中国科学报》说,AlphaFold对某些蛋白的预测甚至没有达到平均水平。他认为该模型对于"什么样的蛋白分子更有效?为什么更有效""这样的模型可以被用来实际应用帮助药物开发吗"这些问题都还未详细研究,还存在不少问号。

"深度学习虽然是个'神器',但跟任何机器学习一样,深度学习必须依赖足够数据。目前来看,AlphaFold样本数量少得可怜。"Z表示,仅仅大致勾勒出蛋白质结构是远远不够的,人们需要依靠一种可靠性高的蛋白质结构预测手段,而所谓可靠性高,"必须精准预测才行"。因此他认为,模型的预测分辨率必须非常高才有较大实际作用。

方法思路有时比深度学习更重要

就AlphaFold的表现而论,郭昊天谈道:"把深度学习引入蛋白质结构预测是大势所趋,没有道理不用,也没有道理不好用。"他认为,以一个特定的算力,一定存在一个很好的处于"平衡点"的算法——混合了深度学习和基于人类知识的传统方法。对于一般团队的算力而言,DeepMind开发的AlphaFold肯定不是那个平衡点;甚至他们所采用的方法也未必在那个平衡点上。

郭昊天言外之意,在DeepMind可调动的资源范围内,AlphaFold的表现仍有提升空间。

不过,尽管AI在蛋白质结构预测乃至生物信息学领域的潜力仍待进一步挖掘,但依靠越来越智能的计算来解决生物学问题正变得越来越重要,已是大势所趋。甚至在北京大学生物化学与分子生物学教授昌增益看来,蛋白质预测本质上"一直就是一种人工智能的应用,好像不能算是一件新鲜事"。

"我的认识不一定准确,但几年前有学者帮我们预测过蛋白质结构,他们给我的印象就是如此。"昌增益说。

Z也表示,深度学习在生物信息领域里"绝对不是什么新鲜事,现在很多paper都用上了深度学习"。

不过,生物信息学领域的特点,也让AI技术难以尽情施展。"生信领域复杂度太高、可训练的样本太小,这特别不利于设计模型结构和调参。"郭昊天认为,国际蛋白质结构库(PDB)所有物种的蛋白加在一起(含大量衍生同种型蛋白质)只有不到15万个可搜索的结构,这种训练样本显然不合AI的胃口。

"按现在的路子,恐怕很难提高准确率。"郭昊天说,可能过不了多久,就会有一个新的模型超越AlphaFold。

"预测蛋白质结构的难度跟人工智能下棋绝不是一个数量级的,深度学习并不是所有难题的答案。有时候,方法思路比深度学习更重要。"Z对记者如此说道。

延伸阅读

蛋白质结构解析:生物学的『圣杯』

■赵广立

美国密歇根大学计算医学与生物信息学中心教授安布里什·罗伊(Ambrish Roy)曾于2010年在*Nature Protocol*发文,称"我有一个要研究的蛋白,但我不知道它的结构和功能"是几乎所有分子和细胞生物学家每天面临的最大难题之一。无怪乎罗伊发此喟叹,当年的统计数字显示,只有0.6%的已知蛋白序列被解析出了相应的结构。

不过,自从美国科学家克里斯蒂安·安芬森(Christian B. Anfinsen)提出"蛋白质的高级空间结构由蛋白质的氨基酸序列决定"后(他也因此获得1972年诺贝尔化学奖),人们开始寻找一种能够预测蛋白质结构的算法,可以精确地从蛋白质的氨基酸序列,利用计算机预测出其复杂的空间结构,甚至其由结构决定的功能。

值得一提的是,尽管随着氨基酸测序技术的发展,越来越多的蛋白质序列得以被高通量的读取,但是从解析一维序列到能够解析实际三维结构,仍然还有很大的距离。

"这不但是生物信息学,也是整个生物学中的一个重要的'圣杯'。"巴黎笛卡尔大学前沿生物学博士郭昊天如此说道。毕竟,要研究蛋白质的功能或是设计靶向药物,蛋白质结构都是非常重要的一环。

国际蛋白质结构预测挑战竞赛(CASP)应运而生。自首届CASP于1994年在美国加州举办以来,20多年间科学家们开发出许多用于蛋白质结构预测的计算模型,这些计算模型主要分成三大"流派":演化流、比对流和从零开始的ab initio流——ab initio就是拉丁语里"从最初开始"的意思。

演化流的核心概念是寻找演化历史上同源或者近似同源的氨基酸序列,从它们的结构出发预测新的目标蛋白;比对流则不一定要求演化上同源,直接将目标序列中的片段和曾解析出来的三维结构进行匹配和比对,由此来预测新的蛋白;而最难也最关键的ab initio流,则是完全从零开始预测那些完全找不到相似性的蛋白序列。

随着CASP挑战的持续进行,这些流派之间的界限逐渐变得模糊,越来越多的科研团队开始把这三种流派整合到一个模型之中,融合成一个更加准确的预测模型。而在对模型的优化过程之中,CNN、RNN(循环神经网络)、DNN(深度神经网络)、强化学习等技术也在不断地被调用于各个计算环节。

一个有趣的工作是,华盛顿大学David Baker团队于1999年开发了一款基于ab initio流派的Rossetta模型,利用此模型该团队先后成功预测了长度100个氨基酸左右的若干蛋白和一段长度93个氨基酸的人工合成序列。2005年,Baker团队突发奇想,开发出屏保程序Rosetta@home,使用PC端在闲置时帮助Rossetta服务器进行结构解析的模拟运算。借用这种分布式计算的形式,Rossetta模型调用众多闲置个人计算资源,取得了极好的效果。

按图索"技"

①"猪脸识别"融合图像识别、算法分析等处理技术,为养猪企业提供智能化养猪解决方案。

②"微笑支付"系统通过人脸识别技术完成支付结算。

东西方科技对话:人工智能深入传统行业

2017年,中国农村互联网用户达2.09亿,农村互联网普及率达35%。科技巨头和创业公司抓住机遇,推动发展中国偏远地区的网络教育,在全国范围尤其是偏远地区保障高效医疗卫生服务,在弥补教育短板方面发挥了重要作用。

近日在广州举行的"东西方科技对话"上,商界领袖、投资者和专家热议的话题是中国如何能有效地在全国实现服务的"均等化",高效平衡地普及医疗、教育以及科技服务的问题。

风险投资公司500Startups今年7月发布的报告称,随着技术的发展,网络在线学习平台为农村学生提供了更多受教育的机会,弥补了偏远和贫困地区师资的不足。

报告称,自中国教育部要求各级政府将至少8%的教育预算用于数字化教育后,科技产业开始大举进军教育领域。报告还指出,目前中国农村地区的5500万名学生可以通过网络直播课程学习。

如今在任何科技活动中,人工智能(AI)都是一个热门词汇。医疗健康公司星创视界CEO王智民认为,未来数十年,人工智能将普及高水平的医疗服务。

他表示,人工智能可能会缩小中国的城乡差距。"在过去的30至40年里,由于医疗资源不足(分配不均)出现了一些误诊案例。医疗人工智能将缓解这一形势。例如,我们可以培训人工智能来辅助农村医生。"王智民表示,在中国,家庭医生和转诊系统还不够完善。每个人都会去大医院,导致大医院非常拥挤。他希望人工智能可以降低人们去城市大医院的需求。

中国对人工智能的应用并不局限于医疗行业。风险投资公司500Startups的大中华区负责人、《中国互联网报告》的作者杨珮珊谈到人工智能如何帮助农民时说:"腾讯和阿里巴巴都开始与农民合作,人工智能养猪和养鸡,使用物联网设备记录猪的生命体征,确保食品安全。"

Modex公司负责区块链和人工智能研发的首席技术官阿林·伊夫泰米称,阿里巴巴的应用是个有趣的例子。"阿里巴巴专注于动物保护,因为算法可以预测动物的疾病。而人工智能可以在更多方面大幅降低中国贫困农村的农业成本。"(贡晓丽)

岩土工程,将技术与艺术相结合

■本报记者 贡晓丽

岩土工程是地下空间开发利用的基石,是保障21世纪我国资源、能源、生态安全可持续发展的重要基础领域之一。

"岩土工程目前来看,最大的问题在于地下水的控制和对周围环境的影响。"在近日于京举行的荣创岩土院士专家工作站授牌仪式之后,中国工程院院士龚晓南接受《中国科学报》采访时表示。

"在认知岩土体继承性和岩土工程复杂多变性的基础上,新时期岩土工程师应创新理论体系、技术装备和工作方法,发展智能、生态、可持续岩土工程,服务国家战略和地区发展。"

以上是近日于深圳举行的2018全国岩土工程师论坛取得的共识之一,也是龚晓南一直以来思考的重点。在对岩土工程的问题思考中,他认为,"每一个有关土力学的实际问题至少有某些特点是没有先例的,创新空间很大。"

减小分析误差

岩土体是自然、历史的产物,有着复杂的特性。著名土力学家太沙基曾表示,"岩土工程是门应用科学,更是一门艺术"。这是否可理解为对岩土工程学科特点的阐述?

龚晓南表示赞同。他认为,这里的艺术(art)不同于一般绘画、书法等艺术。"岩土工程分析在很大程度上取决于工程师的判断,具有很高的艺术性。岩土工程分析应将艺术和技术美妙地结合起来。"

关于研究方法,龚晓南表示,需要定性分析和定量分析相结合,进行综合判断来减小分析误差。"岩土工程分析中误差大是普遍存在的问题,误差主要来自哪个环节?最有潜力减少误差的是哪个环节?如何减少分析误差?应该采取的对策是什么?这些都是值得思考的,需要从工程问题变成物理模型、数学问题再求解。"

据介绍,岩土工程里面的基本问题包括稳定问题、变形问题、渗流问题等,不同类别的工程,遇到的岩土工程问题也不一样,比如建筑地基工程、基坑工程、道路工程等。

"岩土工程分析对自然条件的依赖性和条件的不确知性、计算条件的模糊性和信息的不完全性、测试方法的多样性、结果参数的不确定性,这些都导致单纯的力学计算不能解决实际问题。"龚晓南表示,需要定性分析和定量分析相结合,进行综合工程判断。"不求计算精确,只求判断正确。"

他表示,工程师的判断在岩土工程设计里占重要地位,要具体问题具体分析,因地制宜,抓主要矛盾,宜粗不宜细,宜简不宜繁。

在岩土工程稳定分析中应遵循"四匹配原则",即稳定的分析方法、计算参数、测定方法、安全系数取值,四者应相互匹配。

地下水控制是工程关键

关于进驻荣创岩土院士专家工作站,龚晓南表示,希望今后与荣创岩土在潜孔冲击高压旋喷桩技术方面有合作机会。荣创岩土以岩土工程施工为主业,自主研发的潜孔冲击高压旋喷桩技术,攻克了卵砾石、抛石填海等复杂地层条件下施工复合桩和止水帷幕的难题,"有一定的创新性"。

据了解,高压旋喷技术自20世纪70年代在中国开始应用和研究,被广泛用于建筑地基的加固处理、基坑和水利工程的防渗截水,经过多年的实践应用和不断改进,该技术已有多种工法问世并应用于不同的工程,具有工艺简单、快速、经济等方面的优势。

进入21世纪后,我国建筑、地铁及水利行业的发展均进入快车道,涉及的地基处理项目和基坑工程大幅增长。"由于我国地域辽阔、地质条件复杂多样,导致地基处理难度加大;基坑工程也不断加深,地下水控制问题已成其质量与安全的决定性因素。"龚晓南表示。

大量的工程实践证明截水帷幕是一种能够实现既能保证基坑和地下工程的顺利施工,又不浪费和污染地下水资源的优选方案。常见截水帷幕可分为墙式和桩式,分别以地下连续墙和高压旋喷桩为典型代表。前者整体效果好,但施工难度大、造价高;后者多采用高压旋喷桩技术,与钻孔灌注桩咬合搭接形成的"止水+支护"联合体,相对经济和高效。

新工艺解决旧问题

荣创岩土董事长张亮于2017年发表的文章《潜孔冲击高压旋喷桩工法原理及特性研究》提到,传统的高压旋喷桩工法优点是设备和工艺都比较简单、造价较低,因此使用范围广,但在工程实践应用中存在以下不足之处:在复杂地层中成孔难;容易产生注浆盲区;钻杆垂直度易产生较大偏差;施工效率低;材料利用率低、环境污染严重。

潜孔冲击高压旋喷桩是从设备、工艺和组织管理等多个角度进行综合研究,通过大量的理论研究和工程实践,形成的一套适用性广、能力强、效率高的新型高压旋喷桩施工技术,可以解决传统工法存在的诸多问题。

龚晓南介绍,潜孔冲击高压旋喷桩工法特点包括:施工效率高;成桩质量高、止水效果好;节能、环保效果明显;对周边环境影响小。

上述研究文章也指出,潜孔冲击高压旋喷桩施工技术创造性地运用"潜孔锤高频振动冲击+高压水、气切割土体"成孔和"气爆"技术,破解了传统高压旋喷桩无法应用于复杂地层和场地条件的难题,拓展了止水帷幕桩的应用范围;同时,由于对原状土进行充分搅拌并混合水泥浆,与传统工法的水泥浆置换原状土机理不同,产生弃土和废浆非常少,可以显著节约水泥、减少废浆排放和提高施工效率,优化现场作业环境,实现文明施工。