专访唐杰|我国首个超大智能模型悟道发布，迎接基于模型 AI 云时代

https://baijiahao.baidu.com/s?id=1695640933687104913&wfr=spider&for=pc
2021/03/30

唐杰认为，超大规模预训练模型的出现，很可能改变信息产业格局，继基于数据的互联网时代、基于算力的云计算时代之后，接下来可能将进入基于模型的 AI 时代。智源研究院致力于成为这样一个时代的引领者，集聚各方资源力量，构建一个超大规模智能模型技术生态和开放平台，供北京乃至全国的研究人员、开发者和企业使用。

自 2018 年谷歌发布 BERT 以来，预训练模型（Pre-trained Models，PTMs）逐渐成为自然语言处理（NLP）领域的主流。

2020 年 5 月，OpenAI 发布了拥有 1750 亿参数量的预训练模型 GPT-3。作为一个语言生成模型，GPT-3 不仅能够生成流畅自然的文本，还能完成问答、翻译、创作小说等一系列 NLP 任务，甚至进行简单的算术运算，并且其性能在很多任务上都超越相关领域的专有模型，达到 SOTA 水平。

很快，OpenAI 便开始了 GPT-3 的商业化探索，并催生了一系列落地应用，微软的巨额投资也立马跟进。同样看中 PTM 潜力的谷歌，在 2021 年初推出超级语言模型 Switch Transformer，将参数量提升至万亿级别。

以 GPT-3 为代表的超大规模预训练模型，不仅以绝对的数据和算力优势彻底取代了一些小的算法和模型工程，更重要的是，它展示了一条探索通用人工智能极富潜力的路径。然而，作为全球使用人数第一的语言，中文 PTM 寥寥可数。在这样的发展态势下，构建以中文为核心的超大规模预训练模型及生态势在必行。

2021 年 3 月 20 日，北京智源人工智能研究院（下称「智源研究院」）发布了我国首个超大规模智能模型系统「悟道」的第一阶段成果。「悟道」由智源研究院牵头，汇聚清华、北大、人大、中科院等高校院所，以及诸多企业的 100 余位 AI 领域专家共同研发，从基础性能、有效使用到预训练模型扩展，提出一系列创新解决方法，取得多项国际领先的 AI 技术突破和多个世界第一。

机器之心专访了智源研究院学术副院长、清华大学教授唐杰。作为悟道项目负责人，唐杰分享了团队关于超大规模预训练模型的技术思考和战略布局，以及智源研究院作为新一代 AI 研究机构的优势。

智源研究院学术副院长、清华大学教授唐杰

唐杰认为，超大规模预训练模型的出现改变了 AI 产业格局，继基于数据的互联网时代、基于算力的云计算时代之后，接下来可能将进入基于模型的 AI 时代。而智源研究院要做的，则是致力于成为这样一个时代的引领者，集聚各方资源力量，构建一个超大规模智能模型技术生态和开放平台，供北京乃至全国的研究人员、开发者和企业使用。

今后越来越多的人会使用云上的超大规模预训练模型作为其 AI 研究和应用的基础。超大规模预训练模型系统将成为一种 AI 基础设施，推动理论研究和技术应用更上一层。

超大模型势在必行，迎接基于模型的 AI 时代

AI 模型越做越大这件事不是最近才发生的。早在 3 年前便有人统计指出，计算机视觉领域的 SOTA 模型体积越来越大 [1]。

NLP 领域亦然，从最早的 ELMo（5 亿参数）到后来的 Turing NLG（170 亿参数），GPT-3 更是将模型的体积和复杂度拔升至一个全新的境界。美国大规模在线预测征求和汇总引擎 Metaculus 曾做过一项调研，参加者预计 GPT-4 参数量的中位数大约在 2.5 万亿 [2]。

唐杰表示，大模型可以包含更多数据，表示更多信息，模型往超大规模发展是一个必然的趋势。目前有很多团队都在做万亿级模型，国外有 DeepMind、谷歌 Brain，国内有华为、快手等，研究成果各有千秋。

「谷歌在今年 1 月就已经推出了万亿参数模型，但精度上并没有提升很多。」因此，他推测 GPT-4 的参数规模很有可能上万亿，不仅如此，OpenAI 还会强调模型在众多任务上精度的提高。

智源也在布局万亿级模型，包括配套的高性能算力平台。不过，唐杰表示，由于万亿级模型参数量过于庞大，模型设计非常复杂，训练耗时长，直接使用还存在一定困难，很多时候反而不如百亿级的模型。在现阶段的实际应用中，充分利用数据，参数规模更小的模型常常能实现更好的性能。

目前，悟道团队一方面扩大模型的规模，让模型的表示能力更强，一方面针对实际应用，提高精度。此外，还在模型微调算法上进行创新，希望早日打通百亿级模型和万亿级模型的桥梁。

「如果能用万亿级模型在一些任务上取得性能的显著提升，这将是一个里程碑式的进步。」唐杰说。

随着算力的不断提升，我们现在可以训练越来越大的模型。或许有一天，真能出现与人脑突触量级相当的 100 万亿参数模型。即便这样的模型真能做出来，训练也势必花费巨资，动辄数十亿美元。

超大规模预训练模型只能是有钱人的游戏吗？小团队如何创新？

对此，唐杰的看法是，人工智能发展可以分为这样几个阶段：继基于数据的互联网时代、基于算力的云计算时代之后，接下来可能将进入基于模型的 AI 时代，相当于把数据提升为超大规模预训练模型。未来，研究人员可以直接在云模型上进行微调，很多公司甚至不用维护

自己的算法研发团队，只需要简单的应用工程师就行。

超大规模预训练模型系统的开放，小团队可以说是最大的受益者，大家不必从零开始，预训练基线智能水平大幅提升，平台多样化、规模化，大家在云上可以找到自己所需的模型，剩下的就是对行业、对场景的理解。这将给 AI 应用创新带来一个全新的局面。

至于基础研究，唐杰说：「理论上可以研究得更深、更系统了，以前研究这个模型使用这种数学方法好，现在可以摆到台面上、扩大到更广的范围来。」

「数据规模化的使用，将促使业界和有关机构更深入地讨论哪些内容可以学、哪些内容不能学，更加注重 AI 伦理、数据隐私、保密和安全等问题。」

智源悟道 1.0 阶段性成果发布，取得多项世界第一

智源研究院自 2020 年 10 月正式启动超大规模智能模型「悟道」项目，悟道 1.0 已启动了 4 个大模型的开发，取得多项国际领先 AI 技术突破，持续填补我国研究领域空白：

悟道 · 文汇——首个面向认知的超大规模新型预训练模型

该模型在多项任务中表现已接近突破图灵测试，通过简单微调即可实现 AI 作诗、AI 作图、AI 制作视频、图文生成、图文检索和一定程度的复杂推理。尤其是 AI 作诗方面，已接近诗人水平，并能首次实现根据现代概念生成古体诗。文汇的最终目标是研发出更通用且性能超越国际水平的预训练模型，搭建预训练模型体系，形成认知智能的生态。

悟道 · 文澜——首个超大规模多模态预训练模型

该模型基于从公开来源收集并脱敏的 5000 万个图文对上进行训练，性能已达国际领先水平，在中文公开多模态测试集 AIC-ICC 的图像生成描述任务中，得分比冠军队高出 5%；采用双塔模型，在图文互检任务中，得分比目前最流行的 UNITER 模型高出 20%。最终目标是生成产业级中文图文预训练模型和应用。目前，文澜模型已对外开放 API。

悟道 · 文源——首个以中文为核心的超大规模预训练模型

该模型目前参数量 26 亿，预训练数据规模 100 GB，具备识记、理解、检索、多语言等多种能力，并覆盖开放域回答、语法改错、情感分析等 20 种主流中文自然语言处理任务，技术能力已与 GPT-3 实现齐平。最终目标是构建完成全球规模最大的、以中文为核心的预训练语言模型，探索具有通用能力的自然语言理解技术，进行脑启发的语言模型研究。

悟道 · 文溯——超大规模蛋白质序列预测预训练模型

该模型已在蛋白质方面完成基于 100GB UniParc 数据库训练的 BERT 模型，在基因方面完成基于 5-10 万规模的人外周血免疫细胞（细胞类型 25-30 种）和 1 万耐药菌的数据训练，同时搭建训练软件框架并验证其可扩展性。最终目标是以基因领域认知图谱为指导，研发出可以处理超长蛋白质序列的超大规模预训练模型，在基本性能、可解释性和鲁棒性等多

个方面达到世界领先水平。

同时，悟道数据团队还构建并开放了全球最大中文语料数据库 WuDaoCorpora，数据规模达 2TB，超出之前全球最大中文语料库 CLUECorpus2020 十倍以上。该数据库不仅为悟道项目提供了数据支撑，由于来源广泛及多样性，可广泛用于中文 NLP 领域中多种任务的模型训练，并使模型具有更好的泛化性。数据经过了专门的清洗，确保隐私和安全及保密问题。

为进一步实现模型规模和性能的扩增中面临的挑战，悟道系统团队还开源了 FastMoE，作为首个支持 PyTorch 框架的高性能 MoE 系统，打破了行业研究受制于谷歌的局限，支持多种硬件，只需一行代码即可完成 MoE 化改造，相比 PyTorch 朴素实现速度提升 47 倍。

所有的 NLP 任务都是生成任务

唐杰认为，超大规模预训练模型有三个关键：首先，模型本身，这也是团队智慧的体现；其次，大算力；第三，高质量的数据。

目前，悟道团队在模型设计上：第一，针对复杂任务设计模型，通过记忆机理或者类似于推理的机理，把一些更远的上下文信息加入到预训练中；第二，在把模型做大的过程中，要能加速模型收敛性；第三，在后端的微调算法上探索，提高模型的可用性，把下游任务的精度大大提高。

在此次发布的多项突破中，由唐杰率领的悟道文汇团队提出全新的预训练范式 GLM，以生成为核心，打破 BERT 和 GPT 瓶颈，同时在语言理解、生成和 Seq2Seq 任务上取得最佳性能。

文汇团队还提出了基于连续向量的微调算法 P-Tuning，首次实现自回归模型在理解任务上超越自编码模型，并在知识抽取 (LAMA)、少样本学习 (Superglue Fewshot) 等 10 多个任务上取得世界第一，性能提升超 20%。

GLM：基于生成的通用预训练框架

谈到 GLM 的技术实现思路，唐杰表示，基于双向模型 BERT 和 GPT 各自在理解和生成上的优势，团队便思考如何将这两个模型的优点融合在一起。随着研究的进行，他们修改了优化结合的方式，在优化目标函数上做了尝试。再后来发现，auto-encoder、seq-seq 以及填空任务等都可以整合到生成模型中，所有的 NLP 任务都可以被视为生成任务，统一在一个通用框架下。

唐杰表示，机器学习的传统上可以分为判别模型和生成模型，这两大派系也在不断融合。当数据量少的情况下，判别模型的效果会更好；而生成模型则比较复杂，需要在「理解」的基础上进行判别，而大数据、大模型、大算力的到来，为生成模型提供了基础，计算机可以实现基于大参数的「理解」，这也是如今生成式方法成为机器学习大态势的原因。

至于是否可以将生成看作是「理解」，「其实这是一个哲学问题」，唐杰说。

计算机到底需不需要「理解」，人类「理解」的本质又是什么？对此，悟道团队做了很多的思考。

最简化地讲，人类的理解分三个层次：第一种可以叫做人脑知识 query，把已经记住的知识

查取出来；第二种叫 case based，基于以前的认知和经验来完成新的任务；第三种叫随机推理，也叫试错性推理。

人类的这三种推理方式，其实计算机都可以实现。唐杰认为，当有一天计算机在众多任务上通过了图灵测试，就可以把计算机「理解」问题的引号去掉了。

数据和知识双轮驱动的通用 AI 之路

假设有一个囊括全世界所有数据的模型，我们想要完成什么任务，给它输入，模型返回多个候选结果，人类在此基础上进行调整完善，再将结果反馈给模型，让其优化。与此同时，模型自身也能不断地从网络上抓取数据进行自我学习…… 长此以往，最终获得的模型，是否就是通用 AI 呢？

唐杰说，「这其实也涉及到一个哲学问题」。关于计算机能否像人一样思考，甚至超越人类智慧，「很多人包括我自己在内，都是不相信，或者说不敢这样认为的。但是，现在我的想法转变了，我认为计算机实现乃至超越人类智能是可以实现的。」

悟道大规模预训练模型系统的目标，便是从更本质角度进一步探索通用人工智能，让机器像人一样思考，让模型具有认知能力。对于神经科学和人脑的思维方式，唐杰表示自己的发言权十分有限，但总的来讲，如果可以用计算机模型实现人类认知的 9 个准则，那么他认为计算机就可以被称为具有认知能力。

认知 AI 需要具有的 9 大能力

但他也补充说，如果那一天实现了，也不代表计算机就把人脑颠覆了，也许到那一天，我们人脑也会进步。「人的思维，包括我们的学习能力和进化能力，尤其是当人类处于压力情况

下，我们会往前大大进化一步。而且，人的思维方式和思维的本质目前也没有真正得到一个结论。」

像刚才说的那样，让模型包含尽可能多的数据，并从数据中提出内容，一般被称为人工智能研究的「纯学习派」。同时，还有另一个派系，也就是传统「符号 AI」，认为只需要把知识表示出来，计算机做搜索、匹配就可以了。

悟道团队走的是将知识与数据相结合的路线，这也是张钹院士在几年前提出的看法。「悟道在用两条腿走路」，唐杰说：「一条腿是数据模型，另一条腿是知识图谱。」一方面把知识图谱做得非常大，另一方面，把知识图谱放到预训练模型中，抽取知识图谱反哺模型，进行双轮驱动，「我认为这是当前实现通用人工智能最有前景的方法」。

唐杰表示，我们应该允许机器犯错，犯错不可怕，最关键是要知道错误的原因。人的认知中有一个试错过程，意识到错误会反馈修改。「什么叫做『创新』？人通过试错，如果试对了，就是一种『创新』。」

尽管在受限领域，计算机已经可以自我纠错，比如 AlphaZero，在下棋过程中会感知自己走错了，然后进行反馈，自我进化。但在通用领域，计算机是没有这个反馈的，它错了以后没法修正，甚至不知道自己错了。

那把受限领域都集中到一起，是否就能让机器在通用领域自我纠错了呢？唐杰指出，这是数据和知识的一个悖论，人总觉得自己的知识是无限扩张的，人每天都可以创造新的知识，无法把所有知识都装在机器里。

而机器生成的内容，很多人不认为是知识或者「创新」，而只是一种组合。「如果有一天机器发现的东西获得了诺贝尔奖，那我认为就可以视机器能够『创新』。」

科学没有高下之分，只看能在多大程度上解决 Why 与 How

「哲学」这个词在采访中多次出现；超大规模预训练模型的出现，让唐杰从不相信、不敢认为，到相信机器的智能可能超越人类。

但是，也有观点认为大规模预训练模型是大数据、大算力之下的暴力美学，缺乏对世界本源的理解。唐杰认为，这个世界上科学就两种，一种是回答 Why，一种是 How。而回答 Why 有两个范畴，一个叫做基础理论科学，另一个叫做工程科学，两者没有高下之分。

至于 How，则是看研究成果应用范围有多广，以及真正能推动哪些产业进步。具体到超大规模预训练模型，唐杰认为模型上云是一个大的方向，将来谁可以成为模型上云引领者，推动整个产业的发展，谁就是最终的成就者，「这就是所说的 how 以及谁能做这个事」。

而探究人脑思维则是在回答 Why。「科学的本质是什么？为什么人脑的思维就一定要强过计算机？对此我们可以大胆质疑，小心求证，大家说人类智能比机器好，我们可以反过来问，为什么机器的智能不能比人好？这是回答 Why 的过程。」

唐杰表示，科研成果的评价指标需要根据不同的行业、不同的场景来判别，归根结底是看能在多大程度上解决了 Why 与 How，是否真正推进了社会的进步。就像万亿级参数模型，可能这个世界上 99% 的公司都用不上，但是作为科研探索很重要。

要做就做最难的、对标最好的

智源悟道 1.0 的发布，标志着「智源模式」取得阶段性实质进展。

作为新型的 AI 研究机构，智源研究院聚焦原始创新与核心技术，致力于建立自由探索与目标导向相结合的科研体制。作为北京市 AI 战略科技平台，智源从创立以来，在科研机制上进行了多种尝试，比如「智源学者计划」，支持科学家勇闯无人区，「就是想做什么就做什么，」唐杰说：「只要够牛，要么回答了 how，要么回答了 why，而且是别人做不到的。」

同时，智源研究院也会围绕目标明确、有战略意义的大项目，灵活机动地组织跨学科、跨机构的专业研究和工程人员，组成紧密协作的大规模团队，共同攻关，比如这次的超大规模智能模型系统项目。

「GPT-3 出来以后，我们看到市场未来产业化的发展，从数据云到计算云到模型云，这是一个大的趋势，智源研究院有义务、也有能力来引领，因此迅速确定目标，组织团队。」唐杰说：「每个参与方，包括高校、企业和研究院所，都是带有目标、带有资源、带有情怀的，因此能够通力协作。」

唐杰介绍说，悟道 1.0 只是一个阶段性的成果，今年 6 月将会有一个更大、更高的智慧模型发布。第一，模型规模会有实质性的进展；第二，模型会在更多任务上突破图灵测试；第三，把应用平台做得更加夯实。后续悟道模型将以开放 API 的形式对外提供服务，用户通过申请并经授权后，可以基于模型 API 开发各类智能化应用。另外，也会开源模型的社区版本，服务我国 AI 科研发展。

「我们希望每一个我们做的东西一定是世界上最好的，如果不能做到最好，那就不做了。或者，如果很多人都能做得比较好，我们也不做，我们就要做最难的，对标最好的，包括我自己的定位。」

「此外，光盯着现在的事情我们也不做，我们要瞄向下一步，十年以后、二十年以后人工智能是什么样子，我们觉得能做就会去做。认知 AI 是我特别看好的，预训练模型和知识数据双轮驱动，是实现通用 AI 的其中一个办法。我非常坚信，十年、二十年以后，计算机在很多任务上就能突破图灵测试。」

参考链接：

[1] https://heartbeat.fritz.ai/deep-learning-has-a-size-problem-ea601304cd8
[2] https://www.metaculus.com/questions/4852/how-many-parameters-will-gpt-4-have-if-it-is-released-in-billions-of-parameters/