北大王立威：理论视角看大模型，为什么 AI 既聪明又愚蠢

发布时间：2024-09-13

以英伟达为代表，近期美股科技巨头市值蒸发超过万亿，引发了市场对 AI 泡沫破裂的担忧，特别引发焦虑的是大模型领域，甚至有人将其与互联网泡沫相提并论。

我们惊叹于当前 AI 的成果，但若深究其过程则往往感到失落。在生成式 AI 盛行的当下，这种矛盾心理尤为突出。

大语言模型（LLM）的通用能力是一个意外的收获，为了改进机器翻译序列处理而提出的 Transformer，性能是如此强大，已经成为语音、文本、图像领域事实上的基础架构，并且展现出一统模态的巨大潜力。从 GPT-3 到 GPT-3.5（即 ChatGPT），模型能力似乎有了质的飞跃，但二者在训练方式上并没有本质区别，这是否意味着更多的数据、更大的模型是通往智能的正确路径，更好地「预测下一个词」最终能让我们创造出会思考的机器？

今天，大模型已经开始走向产品阶段，人工智能正渗透到千行百业，我们在享受智能化所带来的便利的同时，也面临一系列现实问题。现有的理论还难以解释深度学习的许多重要问题，导致实践无法系统且高效的进行。大模型的出现，给机器学习理论提出了全新的课题。在技术创新飞速发展、知识创造相对滞后的当下，理论研究该如何应对挑战、抓住机遇？

本期机器之心《智者访谈》邀请到北京大学智能学院教授王立威，从机器学习理论视角看大模型的能力边界，探讨理论对 AI 未来发展的影响。

王立威教授指出，很多人都将今天的人工智能与工业革命相类比，但我们是否想过，蒸汽机虽是传世的发明，却鲜有与其设计相关的理论流传下来。如果仅仅只停留在解释具体现象的层面，如今的机器学习理论研究是否也会面临同样的命运？

当 AI 技术实践不断突破而理论认知未能同步提升时，技术创新的风险也将被放大，甚至阻碍其真正价值的实现。

王立威教授鼓励青年学者挑战现有框架，探索未知领域，大模型不是人工智能的全部，机器学习也不止一条路径，只有看得更深、更加本质，才能发现足以传世的「AI 领域的能量守恒定律」，进而指导未来的研究和实践。

他说，探索需要勇气，承担一定风险，很多事情都无法预测，但这也正是探索的乐趣。

访谈文字整理

机器之心：王立威教授好，非常高兴您做客机器之心的《智者访谈》。我们知道您深耕机器学习领域多年，尤其关注基础理论方面的研究。在如今这个技术飞速发展、应用日新月异的时代，对基础理论的洞察尤其重要，我们希望今天能与您探讨机器学习理论相关的内容，以及理论对于未来 AI 领域发展的影响。

王立威：很高兴参加机器之心的活动，分享一些我个人的看法。

为什么如今的 AI 既聪明又愚蠢？

机器之心：都说现在的 AI 聪明得惊人又蠢得出奇。我们见到了有 Google DeepMind 研发的 AlphaGeometry 系统，能够解奥赛级别的几何证明题，并且获得了 IMO 银牌。同时前段时间热议的，很多大模型连 9.11 和 9.9 在数值上的大小都分不清，为什么会出现这种情况，您能从原理上给我们解释一下吗？

王立威：首先我想跟大家澄清一点，今天的 AI 系统，我们应该具体地去看，而不是笼统地去看。比如你刚才举的两个例子，一个是 Google DeepMind 研发的以 AlphaGeometry 以及后来的 AlphaProof 为代表的，这是一类系统，还有另一类是以 OpenAI 的 ChatGPT 这种语言大模型为代表的系统。

这两类系统虽然都是 AI 系统，但它们是截然不同的，无论从自身的结构、原理到进行机器学习的方式，再到处理的问题，各自的擅长与弱点，都非常不一样。大家可能用过 OpenAI 的 ChatGPT 或者其他的一些语言大模型，这类 AI 系统主要处理的是语言对话，而且是非常宽泛、普适的场景，其优点是可以处理大量的问题，但缺点和不足是逻辑性稍有欠缺，对于需要严密逻辑推理的问题，比如说数学或一些科学问题，这不是它的所长，也跟这类系统的原理密切相关。

刚才讲的另一类系统，比如说 AlphaGeometry，以及后来的 AlphaProof，用的是深度强化学习这种方法，而深度强化学习不是今天语言大模型的主要技术方案。它们还有一大特点，也是区别于语言大模型的，是专注于解决特定类型的封闭世界问题（closed-world problem）。选择封闭世界问题，并利用深度强化学习去解决，这套思路与 DeepMind 在几年前用 AlphaGo 下围棋的方法一脉相承。今天我们已经开发出很多的这种解决特定问题的 AI 系统，它们各自拥有独特的优势和技术路线，在功能和应用上也有所区别。

机器之心：后来 DeepMind 又推出了升级版的 AlphaGeometry 2，说是基于 Google 的 Gemini 大模型进行了训练，并且性能得到了提升。在您看来，这个具体提升在哪里呢？

王立威：我个人认为大模型在里边应该没有起到什么太重要或者本质的帮助，可能更多是吸引大家关注，因为毕竟大模型现在是一个热点。

AlphaGeometry 其实是基于我们中国已故的著名数学家吴文俊先生所做的「数学机械化」方法，去做平面几何的定理自动证明。有很多研究者都在从事这方面的工作，比如中国科学院数学研究所的高小山老师等等，他们已经深耕了很多年。

AlphaGeometry 是建立在这样一系列工作的基础上，又做了一定的改进，你可以把这些改进概括为「神经符号系统」这样的名词，但其本质还是使用 DeepMind 所擅长的那套较为标准和成熟的深度强化学习方法。AlphaGeometry 的论文已经正式发表，它相较于吴方法已经做到一个什么水平，例如在 f 值、m 值之上加了几个新的成分，每一个成分加进去可以提升多少，都有非常清楚的描述。

所以我觉得 AlphaGeometry 好，首先在于选择了平面几何这个很对的研究问题。但是，平

面几何早在吴文俊先生那个时代我们就已经知道，这条路是可以走，并且可以走得很好的，今天 AlphaGeometry 只是把它做到更好，接近完美的一个水平。

使用机器学习解决数学和科学问题的潜力

机器之心：您近年来也关注使用机器学习方法解决数学和科学问题，显然看中了其潜力，您能再展开谈一谈吗？

王立威：用机器学习、人工智能的方法解决数学或者科学问题，在我看来确实非常具有潜力。当然我们也要辩证地看这个问题，不是说有了机器学习和人工智能就能包打天下，就能替代我们的科学家、数学家去解决科自然科学、数学领域的问题。

应该说今天的机器学习、人工智能在这方面是一个有力的工具，但在可预见的未来还无法替代人类。我个人认为未来发展路径可能是：人类科学家仍然要做顶层设计，但是其中的某些环节或部分可以用机器学习和人工智能方法更高效地完成，因为很多时候需要处理大量的数据，尤其是一些不是很规律的、很复杂的表示。

我经常和我组里面的学生讲，我用一个词叫 regular，就是有规律性，人类比较擅长发现或处理一些特别 regular 的对象。今天的机器学习可能在处理一些没那么 regular 的对象，甚至发现一些没那么 regular 的规律方面比人更强一点，但如果是非常伟大的发现，我觉得可能单纯靠今天的机器学习困难是很大的，人和机器学习必须要更有机地结合起来才行。

机器之心：说到用机器学习解决数学问题，我们很容易想到陶哲轩教授，他认为 AI 一定能为我们带来巨大的突破。对此您是怎么看的呢？

王立威：今天用机器学习和人工智能去解决数学问题，实际上有几个不一样的技术路线，应该说是非常不一样的技术路线，一种就是刚才我们谈到的 Google DeepMind，他们用以深度强化学习为代表的方法去解决一些非常特定的领域里面的问题。

以陶哲轩为代表，还有很多数学家，包括另一位著名的菲尔兹奖得主舒尔茨，他们其实在做一个叫「数学形式化」的工作，形式化本身并不涉及 AI，没有 machine learning，它其实是想把今天人类在写数学论文时所用的数学语言，翻译成一种非常标准的，每一步都按部就班的，甚至类似于代码的这样一种语言，其好处是由于人在写数学证明的时候其实是容易犯错的，甚至中间有一些 gap 数学家自己都意识不到。但是如果翻译成形式化的语言，每一步可以自动地由计算机去验证，这样就能保证数学证明里不会存在潜在的漏洞。

在这个过程当中，既然计算机可以直接去读，直接去验证，甚至直接去进行一些逻辑上的推演，那么这个时候 AI、machine learning 就有可能进入进来。实际上早在几十年前就有一个领域叫做定理的自动证明，目标就是希望用计算机来自动完成定理的证明。

今天由于有了机器学习和人工智能，所以大家希望从这条路去做一些事情，在形式化后，是不是有可能通过机器学习的方式，对于一个想要证明的定理，自动地去发现它的证明过程，更准确地说，是在证明的过程中，每一次我走到一步，下一步应该去做什么、去证明什么，这样一步一步从命题到最终结论，全部自动完成。

这是陶哲轩等人在探索的技术路径。就我个人而言，我倾向于认为形式化加定理自动证明这条路，需要很长的时间去走，而且有很大的难度，不仅是技术上的难度，还有很多条件上的难度，比如说数据等问题。

今天的语言大模型，无论是 ChatGPT 还是其他模型，实际上已经把互联网上几乎所有的数据全部用到了。然而，在数学或者一些自然科学领域，我们需要的并非海量的简单文本，而是高质量的专业数据，比如人类数学家撰写的数学论文和与之对应的形式化语言表达这样的配对。就好比机器翻译，今天大模型在自然语言翻译上取得显著成果，其根源在于大量的双语语料库，比如中文和英文的配对。

然而，数学是一个高度专门化且深奥的领域，尽管我们有大量的数学论文，但与之对应的形式化语言表达却非常匮乏，因为将自然语言的数学论文转换为形式化语言，需要耗费大量的人工成本，并且必须由数学领域的专家来完成。我知道有很多学者正致力于这方面的研究，他们尝试通过人工、半自动或自动化的方法，将人类的数学语言转换为形式化的数学语言，但这需要一个长期的积累过程。

机器之心：报道称 AlphaGeometry 使用合成数据，从头开始训练，您如何看待合成数据的前景？

王立威：我自己也曾尝试利用合成数据来提升标准自然数据的表现。然而，这一方法的关键在于，即使生成了新的数据，仍然需要人工介入，运用专业知识进行校对和纠正，这样才能真正输入新的信息。熟悉信息论的听众应该了解，单纯的合成数据并不能提供任何新的信息量，除非有新的 input，那么这种新的 input 是什么呢？就可能是专家对合成数据进行的检验和校正。因此，我认为利用合成数据是一个可行的方向，但单纯依靠合成数据是难以取得突破的。

机器之心：这跟 AlphaGo 自我对弈并从中学习的区别是什么呢？

王立威：AlphaGo 解决的是围棋问题，自我对弈之所以能够带来新的信息，是因为每一盘棋结束后，胜负结果都可以根据规则明确判断，而每一次的胜负结果都提供了新的信息。因此，如果我们所研究的问题也能够产生类似的反馈机制，那么利用合成数据并结合这种反馈，就有可能取得成功。

理论视角看思维链：Transformer 是一种电路

机器之心：回到大语言模型，您团队在 NeurIPS 2023 上面有一篇 oral 论文，首次从理论视角研究了思维链（Chain of Thought，CoT）提示的作用。您能谈一谈这篇论文的结论和启示吗？

王立威：好的，我从几个方面来谈。首先，无论是解决数学任务还是进行逻辑推理，大语言模型最终都需要完成特定任务。我们可以从几个层面来理解模型是如何完成任务的。

第一个层面是模型本身的结构，比如我们刚才谈到的 Transformer。除了结构之外，如何使用结构也很重要，思维链本质上就是一种使用 Transformer 这种结构的方式。无论是模型结

构本身，还是使用结构的方式，都与模型的表达能力密切相关。大家可以想象，如果 Transformer 或者说大模型的神经网络结构过于简单，那它的表达能力必然很弱，很多复杂逻辑或运算就无法表达。所以，我们这篇论文就是从表达能力的角度出发，研究思维链与 Transformer 结合后的效果。

我们的主要结论是，如果只用 Transformer 而不使用思维链，那么 Transformer 这种结构的表达能力实际与电路非常接近。电路大家都很熟悉，比如逻辑电路、数字电路，它们由一些逻辑门组成，例如与门、或门、非门等，逻辑门之间通过线路连接。

我们可以将神经网络与逻辑电路进行类比：神经网络中的神经元对应电路中的逻辑门，神经元之间的连接对应电路中的连线。两者唯一的区别在于，逻辑门的计算操作和神经元的计算操作有所不同。但我们的研究发现，这种区别并不本质，它们之间可以相互转化，因此可以近似地认为两者是差不多的。

所以，如果只是一个单纯的深度神经网络，比如 Transformer，我们就可以将其视为一种电路，并从电路的角度来分析它的计算能力，也就是它处理数学问题的能力。早在上世纪 70 年代，人们就已经对各种电路的计算能力进行了深入的研究。因此，我们可以很清楚地说，如果仅仅使用 Transformer 神经网络来处理数学问题，至少从表达能力的角度来看，它的能力是有限的。

但是，我们的论文进一步分析了，如果引入思维链，情况就会发生变化。思维链相当于让神经网络进行一步一步的推演，每一步的输出都会作为下一步的输入，形成一种循环迭代。这种循环迭代相当于反复利用了神经网络，在某种意义上可以认为是扩大了神经网络的规模，从而提升了它的表达能力和计算能力。因此，使用了思维链的 Transformer 神经网络在处理数学问题时，就能够解决更复杂、更困难的问题。

但是，要构建一个真正能够解决很多数学问题的大模型，仅仅依靠表达能力是不够的，还需要考虑模型的学习能力，包括如何从数据中学习，还有泛化能力，也即模型能否能够把从已有数据中学习到的知识应用到新的、没有见过的数据上。我们这篇论文还没有涉及这些方面的内容，但这对于大模型能否成功解决数学或逻辑推理问题至关重要，也是未来研究的重要方向。

机器之心：Transformer 本身表达能力有限，但堆叠到万亿乃至十万亿、百万亿等更大的规模后，模型的表达能力是否足以解决数学或者科学问题呢？

王立威：虽然现在的大模型已经达到万亿参数级别，但很多人认为，与人脑相比仍然相差甚远，可能还有几个数量级的差距。不过，如果从理论角度来分析，我们不能简单地用参数量来衡量模型的能力，还有一个重要的指标是模型的增长速度，看模型的复杂度是呈多项式级别增长，还是指数级别增长。

所谓多项式复杂度，指的是随着输入规模的增大，模型规模的增长速度可以用一个关于输入规模的多项式来描述。比如，如果输入规模为 x，那么模型规模的增长速度可能是 x^2 或 x

^{3} 等等。而指数复杂度指的是模型规模的增长速度随着输入规模的增大呈指数级增长，比如 2 的 x 次方。

一般从理论角度认为，如果模型复杂度是多项式级别的增长，那么模型的规模是可以控制的，因为多项式级别的增长速度远低于指数增长。但是，如果模型复杂度是指数级别的增长，那么模型的规模将会非常庞大，实现起来非常困难。因此，在讨论模型规模的时候，通常会限定在多项式复杂度增长的范围内。

我们之前的分析表明，如果仅仅使用一个规模按照多项式级别增长的 Transformer 模型，那么很多数学问题是无法解决的。但是，如果允许模型规模以指数级别增长，理论上模型可以处理任何问题。但你可以想象一下，在指数级别增长的情况下，如果模型的输入是一本数学教材，那么模型的规模就不是万亿参数级别，可能要在后面加上很多个零。

我们这篇关于思维链的论文想要说明的是，即使模型规模的增长速度是多项式级别的，也就是实际中大家认为可接受、可实现的，用上思维链以后，模型也可以表达和处理那些复杂的数学问题。

机器之心：这对于我们有什么启示？

王立威：我认为主要的启示是，我们需要不断探索更高效、更有效的模型结构和方法。思维链是一种方式，但未必是最优的一种方式，甚至现有的 Transformer 架构加上思维链也未必是最佳方案。

Transformer 只是众多优秀模型结构中的一种，它不是唯一的，可能还有大量的其他结构，跟 Transformer 一样好，甚至更好也是有可能的。只不过今天大家都在进行超大规模的实验，大模型、大数据，训练一次模型的代价太大了，我们没有能力做大量的实验，但是我相信存在很多不同的模型结构都有很好的性能。

目前的大模型普遍采用 token 进行表示。但如果要处理逻辑性强、严谨性高的问题，例如数学问题，仅仅依靠现有的表示方式是否足够？这一点尚不明确。我不确定是否有学者对此进行过深入研究。毕竟日常对话中的逻辑关系和复杂度相对有限，而在学术领域，尤其是数学领域，一个概念可能是基于其他非常多概念的基础之上，一个概念跟其他概念之间有着非常复杂而深刻的联系，如何有效地表示这些概念以及它们之间的关系，是值得深入探讨的。用今天的这种狭义的神经网络结构能否很好地表示这些复杂的概念和关系，我自己是觉得不能完全确定。

从 2012 年 Hinton 等人的突破性工作算起，深度学习的发展仅仅经历了 12 年的时间。人工智能领域在这 12 年间的飞速发展，在其他领域中实属罕见。但毕竟时间尚短，大家探索的内容仍然有限。

我们今天已经有大模型，能处理非常多的日常问题，这确实很了不起了。可是我们必须承认，科学问题，包括一些数学问题，在难度和深度上肯定远超日常的问题，复杂程度也要高很多。我们不能想当然地以为，现有的模型结构和模式在自然语言对话上取得了成功，沿着这条路

线走下去就能解决科学问题，这也不一定。

机器之心：总体看来您很谨慎，但感觉挺悲观的。

王立威：哈哈，不是悲观，我只是觉得需要时间。其实我觉得这就是探索，探索的乐趣就在于你事先没有办法确定，很多是偶然的，这也是做研究的乐趣。

大模型并不存在所谓的「涌现」

机器之心：ChatGPT 之所以惊艳世人，就在于什么呢？当时有一个词叫「涌现」，是说当模型大到一定规模之后，就能够完成很多种不同的任务，这是在小模型上不曾观察到的。您是怎么看待大模型的这种涌现能力或现象的？

王立威：首先，目前这些千亿甚至万亿参数级别的大模型，与一亿参数以下的小模型相比，在能力上确实存在着本质区别。但谈到「涌现」，我们需要明确其定义。物理学中的涌现与相变的概念相关，相变通常指存在一个阈值，低于阈值时现象完全不存在，超过阈值后现象就会突然出现，例如物理学中的超导现象。

我倾向于认为，在今天我们讨论的大模型中，并不存在相变意义上的涌现，也就是说，不存在一个明确的阈值，例如 500 亿参数的模型没有某种能力，而 501 亿参数的模型就突然具备了这种能力。现在看模型能力的提升应该是一个循序渐进的过程。只不过，现在的模型规模比过去大了几个数量级，所以与之前的小模型相比，差异才显得如此巨大。

机器之心：我一直很好奇，现在可以先训练一个大模型，然后通过一些方法将其压缩成小模型，这个小模型能够实现与之前大模型相似的效果。那么，这个压缩后的小模型与一开始的小模型之间有什么区别呢？因为压缩后的小模型显然具备了之前小模型不具备的能力，您能解释一下其中的原理吗？

王立威：这是一个很好的问题。我先问你一个问题，你会骑自行车吗？

你有没有意识到，当一个人刚开始学自行车的时候，骑不太好的时候，你感觉全身都投入到骑自行车这件事情上，您感觉你的大脑已经完全被骑自行车这件事给占据了。但是，当你学会骑自行车之后，你发现你的大脑可能只需要分出很小一部分用来骑自行车就行了，你可以一边骑车一边和别人聊天，还可以看风景。

实际上，机器学习在你刚才提到的这个问题上与人类学习非常相似。从学术角度来讲，当我们需要从零开始学习时，可能需要一个大模型，但是当我们学会了之后，就可以把大模型蒸馏成一个小模型。

更具体地说，为什么在学的时候必须用大模型呢？

在理论上已经有人证明，如果想从零开始学习，使用小模型很可能找不到正确的路径，而使用大模型则更容易找到从初始状态到目标状态的正确路径。找到正确路径后，我们会发现其实并不需要这么大的模型，再把真正有用的部分抽取出来即可。但是，如果直接使用小模型，就很难找到那条正确的路径，学习难度会大大增加。

关于幻觉：如今的大模型基于统计而非逻辑，永远无法保证 100% 正确

机器之心：我们的终极目标是希望 AI 能够独立完成数学证明，并且像伟大的科学家比如爱因斯坦那样，发现新的科学理论。为了实现这个目标，还需要克服哪些主要困难？

王立威：这个问题非常困难。首先，我们来看目前取得了相当成功的语言大模型。它们成功的关键在于，在训练过程中接触了海量的问题和解决方案。以 GPT 为代表的这类方法，将许多不同类型的任务都转化为自然语言的形式进行描述和学习。

如果我们希望大模型在数学或自然科学领域也能自主解决新的问题，那么它首先必须要见过数学和自然科学领域里面大量不同的问题，以及解决这些问题的方式方法。然而，目前这方面的数据非常匮乏。现有的数据大多是一些习题级别的内容，例如中小学习题、大学本科习题，甚至奥赛习题，但科研层面的数据还非常之少，而且科研层面的数据往往是不完整的。我举个例子，科学家在发表论文时，通常只会呈现最终的发现和结论，而不会详细描述整个思维过程。

越是那些最高水平的科研成果，越是精炼，越没有去写研究人员的思维过程。阿贝尔曾说，高斯就像一只狡猾的狐狸，把自己走过的脚印都抹掉了。实际上，很多科学家都会做类似的事情。他们在研究过程中使用的草稿纸是以千记的，但最终发表的论文可能只有几十页。除非你能把那些草稿纸全部找到，当成训练数据。

机器之心：不过现在科研已经基本数字化了，接下来还有没有这种可能呢？

王立威：我们刚才一直讨论的是从数据中学习，但这只是机器学习和人工智能解决科学问题方法中的一部分。我个倾向于认为，只通过从数据中学习是不能完全解决用机器学习和 AI 处理数学和自然科学问题的。为什么呢？

因为真正的科学研究不仅仅是从数据中学习，更重要的是创造和验证。科学家在进行研究时，会产生许多想法和假设。这些想法的产生过程与现在大模型的 next-token prediction 模式类似，都是基于过去的经验和观察去生成新的内容。

但是仅有这种生成是不够的，即使是最伟大的科学家，产生的 100 个 idea 中，可能有 98 个都是错误的，必须要进一步严格地去验证，发现错误之后，还要想办法如何去修正和改进，这才是科学研究的关键。

我觉得今天的大模型产生幻觉，跟人类产生想法的机制非常类似，只不过今天的大模型产生了想法，next-token prediction 之后就直接输出了，就把 next-token prediction 的结果作为答案交给人了。如果未来大模型能在验证、判断和纠错方面做得更好，相信效果会比现在更好。

机器之心：所以说幻觉在您看来是大模型的一种固有特性？

王立威：对，我认为幻觉是大模型一种内在的、应该存在的一种性质。

今天的大模型都是采用从数据中学习的方式，本质上是一种基于统计的方法。既然是基于统计而不是基于逻辑，就永远无法保证 100% 正确。当然我前面说过，幻觉的存在是有其意义的，而且我认为不应该把它完全抹杀。我们应该允许模型生成一些并不一定 100% 正确的内

容，然后人类再从中进行筛选。

只要大模型仍然采用从数据中学习、去做 next-token prediction，如果只做到这一步，那幻觉就是无法消除的。如果想要消除幻觉，就必须在后面增加检验、纠错等机制。

机器之心：那么现有的机器学习方法，或者说更广泛的人工智能方法，能够进行这样的验证或纠错吗？

王立威：这就回到了我们刚开始讨论的内容，现在的机器学习不止一条路径。比如我们前面谈到的 AlphaGeometry，它和语言大模型走的就是完全不同的路线。AlphaGeometry 在每个环节都需要进行验证，确保自身的正确性，但它在内容生成方面的能力可能不如语言大模型。

我想借此机会澄清一点，在自然科学或数学研究领域，存在着各种各样的问题，它们的类别也是不一样的，不同类别的问题由于自身的特殊性，需要机器学习如何参与，或者说需要机器学习参与进来用什么样的技术路线可能是千差万别。对于那些拥有海量数据的自然科学问题，例如在化学和一些生物学领域，已经积累了极多的观察数据，这时我们就可以把数据交给模型去学习，例如之前的 AlphaFold。但在某些领域，人类经过几百年的科学研究，已经发现了一些重要的规律，这时我们就不能完全放弃这些规律，而应该将知识与数据结合起来。所以，我想并不存在一种包打天下的办法，机器学习也是如此。我们需要根据具体的问题和条件，设计相应的解决方案。

机器之心：假如我是一名自然科学领域的研究人员，比如物理或化学，但我对人工智能方法了解不多，我该如何选择适合我的方法呢？

王立威：我的建议是要么从头开始学习，要么找一位机器学习专家进行合作。在我的研究小组里，有一些本科学习自然科学的博士生，他们在加入我的团队后，继续学习了人工智能相关的知识。同时，我的组里面也有一些机器学习和 AI 背景的同学，他们在做 AI for Science 研究时，也必须学习相关的自然科学知识。如果只是把机器学习当作一个封装好的现成工具去使用，我认为很难在 AI for Science 领域里做出比较重要的贡献。

机器之心：所以说，一方面要对人工智能和机器学习方法有深刻的理解，另一方面也要对自己要解决的问题本身以及需要什么样的方法有深刻的理解。

王立威：是的，我甚至认为，未来我们应该注重培养同时具备这两种能力的青年人才，这是 AI for Science 未来发展的重要方向。

The Bitter Lesson & Scaling Law

机器之心：Richard Sutton 教授在 2017 年发表了《The Bitter Lesson》，文中讨论了计算能力和数据的重要性，结合到现在以 OpenAI 为代表，他们推崇依靠数据和扩大规模带来性能的提升。您怎么理解 Sutton 教授的 bitter lesson？您又怎么看 Scaling Law 和算法创新之间的关系？

王立威：我之前看过 Sutton 写的《The Bitter Lesson》，我是感同身受，因为我做机器学习

也有 20 多年的时间了，在 2010 年之前，也就是深度学习和 ImageNet 崛起之前，当时的机器学习研究主要在一个叫做 UCI Repository 的数据集上进行，UCI Repository 包含几百个数据集，但大部分数据集都只有几百个数据，以现在的眼光来看，这是难以想象的小数据。

当时大家提出一个新算法后，通常会在这些只有几百个数据的小数据集上进行验证。从今天的角度看，这种验证得到的结论是完全靠不住的。所以，无论是 Rich Sutton 的这篇文章，还是现在大家谈论的 Scaling Law，都在告诉我们——数据的规模和数据的多样性至关重要。2010 年之前，有成千上万篇论文都陷入了这种小数据验证的陷阱。我们应该从中吸取教训，认识到使用大规模的数据进行学习和验证的重要性。这是过去十几年一个重大的认识上的收获。这一点我完全同意。

但这并不意味着我们只需要追求数据、算力和模型规模就够了。**Scaling Law 更准确的含义是，能否通过设计模型和算法，在大规模的时候取得好的效果，**而不是说只是无脑地去把规模增大，因为当数据、算力或模型规模达到一定程度后，不同的模型和方法之间在性能上仍然可能存在本质上的差距，我们仍然需要去做非常多的设计。

大家可能知道，神经网络，不是深度神经网络，其实早在上世纪就已经展开研究了，甚至在上世纪八九十年代的时候，还是一个对神经网络研究的高潮，只不过当时研究的主要是浅层神经网络，因为一些算法、算力和数据方面的限制，没有能够做到深层的神经网络。

到了 2010 年以后，随着技术的发展，大家逐渐去把网络做深了，一个自然而然的问题就是：深层网络和浅层网络相比，究竟哪个更好？今天大家可能觉得答案显而易见，肯定是深的网络更好。但这种说法并不严谨，更严谨的问法应该是：如果两个网络的神经元数量相同，也就是说网络规模相同，但网络结构不同，例如一个是浅而宽的网络，另一个是窄而深的网络，那么哪个网络的表达能力更强？

我们组大概从 2017 年提出这个问题并进行研究，一直到去年，一组以色列的机器学习理论研究者终于回答了这个问题，他们从理论上、在数学上严格证明了：宽度合理、深度也合理的网络表达能力是最强的，明显强于浅而宽的网络。所以，**即使你把网络规模增加到很大，也需要合理的结构才能发挥最佳性能。**

increase the number of parameters over the target network.

以色列魏茨曼科学研究所的研究团队发现，对于 ReLU 神经网络的表达能力而言，深度比宽度更重要。地址：https://proceedings.mlr.press/v178/vardi22a/vardi22a.pdf

关于可解释性

机器之心：随着大模型越来越广泛的应用，如何解释模型的行为也得到了越来越多的重视，包括您所从事的医疗相关的研究，为此我们需要在理论方面取得哪些突破？

王立威：我来分享一下我对可解释性的一些看法。我觉得今天的模型实际上要从不同的层次来看，或者说模型和数据要放在一起，从不同的层次来分析。

这里面有一些非常底层的信号，比如说人看到一只猫，能够识别出它是一只猫，这就是一些比较底层的信号，一些很底层的视觉信号。当人去研究一些逻辑性问题的时候，思维方式又会是另外一个层次，和刚才的视觉识别是不一样的。**实际上，在不同的层次上，对于可解释性的要求，甚至模型是否可解释，都是不一样的。**在一些更偏底层的问题上，也许没有办法去解释，因为它们就是很复杂。但是对于一些更高层次的任务，有一部分是可解释的，是可以把逻辑写出来的。所以我觉得要分层次去看待可解释性这个问题。

另一方面，我觉得可解释性也许不完全是一个客观的问题，它可能跟人的心理因素也有关系。例如下围棋，自从 AlphaGo 出现之后，用机器、用机器学习系统去下围棋，已经远远超过了今天人类顶尖棋手的水平。我自己也是个围棋爱好者，虽然自从 AlphaGo 出现之后，我就不再下围棋了。

其实，对于 AlphaGo 以及其他一些现在最具代表性的机器学习围棋系统，人类的看法也是经历了一个过程，这里面也体现了可解释性的问题。在 AlphaGo 出现的初期，人类顶尖棋手一直想理解机器为什么这样下棋。机器走的一步棋，人类棋手之前可能根本就不会想到，他们非常想理解为什么机器要这么下，需要开发团队告诉他们，这个东西怎么解释，这一步棋怎么解释。开发团队后来想了一些办法，比如告诉你，这步棋下在每一个不同位置，最终估计的这盘棋的胜率是多少，那这是不是一种解释？

机器之心：不是我们想要的那种解释。

王立威：那还有没有别的解释？最后发现人类没办法从机器那里得到想要的解释。对于机器来讲，它就是经过了大量的训练之后，对棋局有了自己的理解和判断。在当前的局面下，它认为应该下在哪里，并通过大量的计算，最终得到了一个结果，人类是没办法理解的。我相信现在绝大部分的职业棋手，都不会再去问这个系统，为什么要下这步棋，你给我解释一下这步棋要下在哪里。

我还可以举一个更极端的例子，在上世纪 90 年代到大概 2010 年，围棋界排名第一的选手是一位韩国棋手，他的外号叫「石佛」李昌镐。他曾经就对机器下围棋，也就是现在以 AlphaGo 为代表的这种机器下围棋的一步棋，发表过评论。当时机器走了一步棋，叫做「点三三」，这是一个围棋术语。之前的人类职业棋手都认为这是一步很差的棋，谁下出来肯定被老师骂的。所以李昌镐说，在他理解机器为什么下点三三这步棋之前，他是不会下这步棋的。所以现在的情况就是，他不下，但是其他所有职业棋手都下。因此，现在李昌镐下不过其他人了。

所以我想总结一下，刚才是讲了一些趣事趣闻，就是可解释性有人类的心理因素在里面。今天机器下围棋已经远远超过了人类顶尖棋手，他们可能再也不问可解释性的问题了。在其他的一些领域，像刚才你提到的医疗，现在医疗 AI 的水平可能跟顶尖的医生相比还没有达到，或者说没有超过人类顶尖医生的水平，所以自然而然地，我们人类在心理上，就会想要问机器，为什么要做出这样的判断。但是，如果未来每一次机器做出的判断都比人事后验证更准

确的时候，也许人就不再问了。

机器之心：您能够预见这样子的未来吗？

王立威：这取决于具体是什么问题。因为刚才说的是下围棋，最终有胜和负，这是一个新的信息，我们也认为它是一种金标准，最终就是谁赢谁输了。在这样的一些问题上，机器确实能够超越人类。但也不是所有的问题都有这样的金标准，有的时候机器仍然是从人类标注的数据中去学习，那么这个时候它可能最好也只能学到人类的顶尖水平。

重新定义泛化

机器之心：泛化能力是衡量模型性能的一个重要指标。过往我们研究泛化，主要是去考量是什么因素控制了泛化能力。大模型时代，我们是否需要重新考虑对泛化能力的定义？

王立威：对，这是一个非常好的问题。我觉得在过去讨论泛化和今天大模型时代讨论泛化，可能具体的定义不太一样。我先澄清一点，过去我们讨论泛化，是在一个比较狭义的意义上去讨论，比如说我固定了一个任务，就是去做一个分类问题，那么对于这个分类问题，我有一些训练数据，可以用这些数据训练模型，并得到一个训练的准确率。但还有一些在训练的时候没有见过的新数据，这些数据可能是在未来实际应用或者测试的时候才会遇到。那么模型在这些新数据上的性能，我们就称之为泛化性能。但此时讨论的都是一个非常确定的任务，就是去分类、去识别。在这样一个很狭义的意义上，过去机器学习理论做了很多工作，也建立了一套理论的体系。

但是在今天，由于大模型的出现，我们讨论的任务和之前不一样了。今天我们的大模型能够处理的任务是非常之多的，不再是一个固定的，像图像分类这样的单个任务。所以我们在讨论泛化的时候，已经不是过去那种狭义的泛化了，甚至我们今天讨论的泛化，是指给大模型一个全新的任务，看它能不能够把这个任务也解决好。所以从这个层面上说，过去的理论就显得比较局限了。那么有没有更新的理论，能够在刚才说的任务这个层面上去分析泛化，现在这方面的工作还比较少，也是未来可以去研究的一个关注点。

机器之心：关于如何评估大模型的性能，也是一个热点问题。现在的很多 benchmark 都已经被刷烂了，或者说不具备跟以往相比那么强的指示性。在这种情况下，如何去评估一个模型的性能，您是怎么看的呢？

王立威：今天的大模型，已经有相当一部分走到了产品这个层次。那么今天对大模型的评估，就应该用一种评估产品的方式。对产品最好的评估方式就是交给用户去使用，让用户用他们的体验，最后用脚来投票。所有在 benchmark 上的测试，都只是一种内部的测试，只是一些中间结果。

因为大模型最终面对的是用户，是人，那么它好不好是由人的体验说了算。当然，如果你的机器学习模型所处理的任务，确实存在着一个客观的评判标准，其中没有人主观因素的干扰，那么完全可以通过 benchmark 来评判。

这也是一个我觉得思维模式需要转变的地方。因为过去几十年机器学习的研究，还基本上停

留在学术的范畴，所以有 benchmark 这样的指标是有助于学术研究的。但是，真的到了产品阶段，没有任何一个产品是用 benchmark 作为最终衡量标准的。

大模型时代的理论研究

机器之心：您作为理论研究者，如何看待大模型时代机器学习理论的价值和前景？

王立威：我经常听到有人把今天的人工智能和工业革命做类比。我们可以一起来设想，第一次工业革命的代表就是发明和改良蒸汽机。如果我们回过头来看，有没有什么理论工作是关于蒸汽机的设计的？

蒸汽机无疑是传世的工作，也许当时的确有一些关于蒸汽机理论工作，但并没有流传下来。我们再来看一看今天的机器学习和深度学习，其实也有很多的理论工作，有一些对实际的模型和算法设计也起到了帮助作用。但是这些工作能不能传世呢？我必须要打一个问号，虽然我自己也是做机器学习理论的。

让我们再回到蒸汽机的例子，其实是有相关的传世理论的，比如能量守恒定律。这是一个伟大的理论发现，当人们知道了能量守恒之后，就再也不用白费力气设计永动机了。其实在一两百年前，有无数的人去设计永动机。所以，能量守恒就是一个典型的传世理论。

我认为在过去大概十来年的时间里，大家做了很多关于深度学习、强化学习的机器学习理论研究，有一些工作非常出色，但可能还没有达到能够传世的水平。如果我们想做出传世的机器学习理论，可能需要看得更深入，需要去问一些更新的问题，而不是仅仅关注今天大家研究的这些问题。

机器之心：比如说哪些问题呢？

王立威：哈哈，如果我要能回答这是什么问题，可能就已经解决一半了。我只能说一说我自己一些非常模糊的想法。其实在过去这几年，大家对现在的机器学习理论，也有一些意见，有一些不同的看法。今天的机器学习理论可能太过于追求去解释机器学习里面的一些实验现象。也许我们应该走得更深，去看一些更本质的问题，这些问题不一定要和我们现在实验中的现象完全对应起来。就像我刚才举的例子，能量守恒和如何设计蒸汽机可能并不直接相关，但它更本质。所以我建议，特别是我们国内的这些年轻学者，可以尝试从不同的角度，更深入地去思考这些问题。

可以更多一些探索，少一些束缚，不用太被今天大家对热点问题的关注所束缚住。

因为很多时候研究是没有办法预测的，深度神经网络和大模型完全有可能只是一个局部的极值，真正的全局最优可能还需要我们退回去，再走另一条路才能找到。所以应该有更多的学者，特别是青年学者，去做一些探索。毕竟理论研究也不需要那么多的资源，它可能需要的资源相对比较少。所以，多做一些自己感兴趣的事情，希望大家有这个勇气，这确实也需要一定的勇气，承担一定的风险。

嘉宾简介

王立威，北京大学智能学院教授，研究兴趣为机器学习。长期从事机器学习基础理论研究，

为设计更有效的新算法提供理论指导，并开发基于机器学习的医疗影像诊断算法与系统。近来致力于通过机器学习方法解决科学与数学领域重大基础问题。

王立威教授已在 NeurIPS、ICML、TPAMI 等国际顶级期刊和会议上发表论文 150 余篇，其中关于图神经网络表示理论的两篇工作分获 ICLR 杰出论文奖与提名奖。担任 TPAMI 编委，并长期担任 NeurIPS、ICML、ICLR 等机器学习顶会的领域主席 / 高级领域主席。此外，入选 AI's 10 to Watch，是首位获此殊荣的亚洲学者。

来源：机器之心