人工智能的极限：智能拼卡时代

作者：张军平 来源：科学网博客

本文原地址：https://www.iikx.com/news/topnews/26900.html

近两年人工智能的发展不可谓不快，在聊天式人工智能、文生图、文生视频、文生音乐方面，一个接一个新的成果被快速推出，ChatGPT, Sora, Suno。2024年1月份学术圈还在讨论4秒的文生视频如何拓展到15秒，3月份OpenAI公司的Sora就直接生成了1分钟的视频。而最近Anthropic公司的Claude3以及Meta的LIama3也在陆续刷新大语言模型的性能。比如Claude3甚至有“ 意识 ”发现自己处在模拟环境中，正在接受某种测试。

伴随而来的，大家也发现LLM对能源的需求越来越大。据说，训练GPT6需要的电能由于过大，以至于不能将硬件全部放在美国的单个城市里，只能分散到不同地方，因为可能会导致当地居民无法正常用电。也有人预测，按这个趋势，到2035年，人工智能在美国的耗电量将占到全部总量的20-30%。

造成这样的局面，从大的框架来看，主要是因为对大数据、大模型以及硬件环境的严重依赖。

大数据的标注能帮助人
工智能形成好的监督或有教师指导的学习。所
以，
大量使用人力进行数
据的标注，便成为人工智能研究者们第一
反应要做的事
。不过，这事也在变化。因为过大的模型需要的数据，光靠人标还不够，用机器巨量生成的效率会更高。结果，有些数据标注已经发展成为机器生成、机器标注、机器自动评判标注的有效性。而生成的数据规模，也远超人类文明历史所有，比如图像，人工智能生成的数据可能已经超过200亿张。如此大规模的数据，要学习起来，已经非人力可为，只有高效的机器能处理，毕竟它们不太需要休息。

模型的参数已经奔着8000亿去了。但如果按以往人工智能研究者对数据和模型参数的关系来看，数据的规模要随模型参数呈指数级增长，才能保证模型得到充分训练。它意味着两种可能，要么是数据还不够多，不足以喂饱大模型;要么是数据够了，但要寻优，弄个过完备的大模型更稳妥些。前者是有一定道理的，因为我们仍然能看到数据标注方面的大幅投入，不管是人力的，还是虚拟的。后者也有可能，因为过完备空间是常见操作，早在小波流行的时代，大家就知道扩充维度，能帮助更好地找最优值。把寻找最优值的空间变得足够大，虽然大部分地方是空的，但只要有耐心，总能找到最优解或近似解。当然这个耐心是需要用显卡和能源消耗来代偿的，毕竟天下没有免费的午餐，时间换空间、空间换时间，你想得到某样东西，就一定会失去某些东西。结果，我们看到了显卡公司股份的一路飚升，电量消耗让众多大模型公司都有些吃不消。

从工业落地的角度来看，无可厚非。因为公司级的AI产品最好是独孤求败型，只要能训练出具有优异性能的唯一模型即可，哪怕其中存在大量不方便“ 复现 ”的工程技巧。只考虑可复现性、反而会缩短与竞争公司的差距，导致前期投入打水漂，除非是遥遥领先，或者自我感觉良好。

但从学术界的角度来看，跟着这股
风去做科研，其实不太明智
。一是本来就没这么强的算力。以国内高校的显卡数量来看，能有近千块A100显卡的屈指可数。而公司级别玩大模型的可能都在十倍甚至百倍的规模。二是没这么多的数据。比如互联网大数据，这是天然就不在学校手里。而现在相当多的应用都依赖于源自企业的大数据。三是设计大模型的人力成本也没有。公司层面吸引AI人才的力度，也不可能是学校发助学费或劳务费就能超越的。即使有，那也只可能是凤毛麟角。所以，如果沿着大模型方向走，要么只能玩点玩具级或被缩水过(如蒸馏学习)的大模型，要么就融入到大公司或有充足资金硬件支持的大实验室，草船借箭式发展。

实际上，
现在对显卡的依赖也或多或少影响
对人工智能创新方法的评价
。有些人可能以为，只有模型大才需要大量显卡。而实际上，只要有(深度)模型，只要无闭式解，就需要调参，需要一组参数一组参数的调整。在这种情况下，卡对科研是永远不够的。试想，有两个学生同时找到一个简单且直接的创新点，需要对其进行60组参数的调参，才能确定最优性能的参数。A学生实验室有60张卡可供其使用，B学生实验室只有1张卡可用。跑一组参数需要1天时间。那么，结果是显而易见的。A学生用一天时间就找到最优参数，并立马可以开始撰写论文补充各种实验。而B学生花两个月跑完实验，正准备写论文时，在arXiv上却发现A学生已经把完整的论文挂在线了。

不仅如此，很多大模型一旦跑起来，由于参数量过于庞大，即使中间出错也不太愿意停下重来，因为重新训练的成本太高，只能将错就错，期望后续的训练过程能纠正一些。

所以，
这些问题导致大
家在AI科研上有两个主要的使力
模式，所谓“ 穷调参，富买卡 ”
。由于卡数量上的差异，也导致不少富组跑出来的结果，穷组在时间和技术层面都无法复现。

除了卡的问题，大模型的发展方向依然有不少问题难以解决。比如幻觉问题、一本正经的胡说八道。就我个人的理解，它本质上还是70年代莱特希尔报告里指出的人工智能问题的再现，即组合爆炸问题。一是规则以外总有例外，二是大模型对问题的理解和回答是基于概率而非真正人的理解模式，这就决定了它始终会出现幻觉和胡说八道的笑话。它也意味着随着人类参与相关模式对话的深度和广度加深后，这类现象会层出不穷。

另外，
目前的大模型策略有明
显同质化的倾向，几乎已被生成式的策略一统
天下
。自2017年transformer提出以来，我们能看到的大多是它的变体。即便最新推出的文生视频Sora，从公开的报道来看，也是将扩散模型里的U-Net网络改成了Transformer。甚至人类的活动也开

始被人工智能同质化。比如近年来围棋选手在比赛时，已经出现差不太多的开局模式。因为如果不按阿尔法狗给的开局模式来落子，输棋可能是大概率的。

尽管近两年人工智能的确取得
了令人瞩目的成绩，需要提醒的是，
大模型和人类在处理问题的方式上是显然不同。一是能量消耗上，人类没有这么大的耗能。二是知识储备上，人类更是没有
。但，也许人类的这种智能才是自然的最优选择，才是更为绿色低碳。

而从科研的角度来看，
当科研创新变成一种套路，甚至转为依赖工程技巧时，当钱烧到看不到希望的时候，它的转向就只是迟早问题了
。只不过，希望转向别再次伤害那么本来就在扎实做探索的人工智能科研工作者。

张军平

2024年4月22日