【知识福利社】你的私人书院，微信：312005650 全网热门课程众筹共享！

02 开悟和涌现：AI能力的三个境界

万维钢·AI前沿
今天

进入课程 >

时长：12分56秒 大小：8.87MB

00:00 / 12:56

转述：怀沙

你好，这里是《万维钢·AI前沿》课。这一讲，我们来说说大型语言模型为什么这么厉害。

现在最流行的几个生成性AI，包括ChatGPT和画画的AI，背后都是「大型语言模型（Large Language Models，简称LLM）」。这大约也是通往AGI的技术路线。

我先举个例子。我问ChatGPT：棒球棒能被藏进人的耳朵里吗？它说不能，因为人的耳朵是非常小的，棒球棒的大小和形状超出了耳朵所能容纳的范围……很有条理。

我又问它：为什么金箍棒能被藏进孙悟空的耳朵里？它回答说因为那是虚构的故事，金箍棒的形状和大小可以随意改变……

棒球棒能被藏进人的耳朵里吗？

不，棒球棒无法被藏进人的耳朵里。人的耳朵内部是非常小且复杂的结构，只能容纳适合大小的物品，例如耳机或耳塞。即使是较小的物品，如小型器械或药物，也需要通过特殊的工具才能安全地放入耳朵内部。棒球棒的大小和形状显然超出了耳朵所能容纳的范围，此外，把物品塞进耳朵可能会导致耳朵内部组织受到损伤，引起疼痛、感染或其他健康问题。因此绝对不应该尝试这种行为。

为什么金箍棒能被藏进孙悟空的耳朵里？

金箍棒是《西游记》中孙悟空所使用的武器，它的形状和大小可以随意改变，甚至能够缩小到可以被藏在孙悟空的耳朵里。

然而，这只是一个虚构的故事情节，不具备现实可行性。在现实中，任何物体都有一定的大小和形状，无法无限制地缩小或扩大，而人的耳朵内部也有固定的大小和结构，不能容纳大的物体。

因此，孙悟空将金箍棒藏进耳朵里这一情节是一个神话和传说中的故事，不应该被看作是现实世界中的真实情况。

你仔细想想的话，这两个回答非常了不起。很多人说语言模型都是基于经验的，只能根据词汇之间的相关性输出答案，根本没有思考能力……但是从这两个问答来看，ChatGPT是有思考能力的。

谁会写一篇文章讨论棒球棒能否被藏进人的耳朵里呢？ChatGPT之所以能给出答案，肯定不是因为它之前听过这样的说法，而是因为它能进行一定的推理。它考虑到并且知道棒球棒和耳朵的相对大小，它还知道金箍棒和孙悟空是虚构的。

它这些思维是怎么来的呢？

你可能没想到，这些能力，并不是研发人员设计的。

研发人员并没有要求语言模型去了解每种物体的大小，也没有设定让它们知道哪些内容是虚构的。像这样的规则是列举不完的，那是一条死胡同。

ChatGPT背后的语言模型，GPT-3.5，是完全通过自学，摸到了这些思考能力。以及别的能力——你列举都列举不出来的能力。连开发者都说不清楚它到底会多少种思考能力。

语言模型之所以有这样的神奇能力，主要是因为它们足够大。

✳

GPT-3有1750亿个参数。Meta刚刚发布了一个新语言模型叫LLaMA，有650亿个参数。Google在2022年4月推出一个语言模型叫PaLM，有5400亿个参数；之前Google还出过有1.6万亿个参数的语言模型。据OpenAI的CEO山姆·阿尔特曼（Sam Altman）说，GPT-4的参数并不会比GPT-3多很多；但大家猜测，GPT-5的参数将会是GPT-3的100倍。

这是只有在今天才能做到的事情。以前不用说算力，光是存储训练模型的语料的花费都是天文数字。1981年，1GB的存储成本是10万美元，1990年下降到9000美元，而现在也就几分钱。你要说今天的AI科学跟过去相比有什么进步，计算机硬件条件是最大的进步。

今天我们做的是「大」模型。

大就是不一样 。当然语言模型有很多高妙的设计，特别是我们一再提到的transformer就是一个最关键的架构技术，但主要区别还是在于大。当你的模型足够大，用于训练的语料足够多，训练的时间足够长，就会发生一些神奇的现象。

2021年，OpenAI的几个研究者在训练神经网络过程中有一个意外发现 。

我给你打个比方，比如说你在教一个学生即兴演讲。他什么都不会，所以你找了很多现成的素材让他模仿。在训练初期，他连模仿这些素材都模仿不好，磕磕巴巴说不成句子。随着训练加深，他可以很好地模仿现有的演讲了，很少犯错误。可是如果你给他出个没练过的题目，他还是说不好。于是你就让他继续练。

继续训练好像没什么意义，因为现在只要是模仿他就都能说得很好，只要是真的即兴发挥他就不会。但你不为所动，还是让他练。

就这样练啊练，突然有一天，你惊奇地发现，他会即兴演讲了！给他一个什么题目，他都能现编现讲，发挥得很好！

这个过程就是下面这张图——

红色曲线代表训练，绿色曲线代表生成性发挥。训练到一千步乃至一万步，模型对训练题的表现已经非常好了，但是对生成性题目几乎没有能力。练到10万步，模型做训练题的成绩已经很完美，对生成题也开始有表现了。练到100万步，模型对生成性题目居然达到了接近100%的精确度。

这就是最新研究的成果。研究者称为格（Grokking）。

内容禁止商业，仅供社群内部免费交流使用
【知识福利社】微信:312005650

✳

开悟，到底是发生了什么呢？

先别急，我再举个例子。ChatGPT有个很关键的能力叫做「少样本学习（Few-Shot Learning）」，就是你给它一两个例子，它就能学会你的意思并且提供相似的输出。

比如我让ChatGPT模仿我给出的例题，再出几道小学数学题。我的例题是"小明有3个苹果，妈妈又给了他2个苹果，现在他有几个苹果？"ChatGPT马上就出了五道题，全是这个风格——比如"小李有5支笔，他送出了3支笔，还剩下几支笔？"

给一个小学生出几道数学题，模仿下面这个例子：小明有3个苹果，妈妈又给他2个苹果，现在他有几个苹果？

好的，以下是几道数学题，适合小学生做练习：

1. 小燕有4个橙子，她吃掉了2个橙子，还剩几个橙子？
2. 小李有5支笔，他送出了3支笔，还剩下几支笔？
3. 小华有6个糖果，他把其中3个糖果分给了小明，还剩几个糖果？
4. 小红有7元钱，她去买了一本书花了3元钱，还剩下几元钱？
5. 小张有8个球，他把其中4个球送给了小李，还剩几个球？

希望这些题目能够帮助小学生练习基本的数学计算和问题解决能力。

简直就跟对对联一样。少样本学习是个关键能力，你可以利用这个能力让ChatGPT帮你做很多事情。那这个能力是怎么出来的呢？

来自更多的参数和训练。看下面这张图——

图中说的是GPT-2和GPT-3模型的演化。参数越多，少样本学习的能力就越强。

而少样本学习只是其中一项能力。还有很多别的能力也是如此：大了，它们就出来了。

✳

这个现象，其实就是科学家之前一直说的「涌现（Emergence）」。涌现的意思是当一个复杂系统复杂到一定的程度，就会发生超越系统元素简单叠加的、自组织的现象。比如单个蚂蚁很笨，可是蚁群非常聪明；每个消费者都是自由的，可是整个市场好像是有序的；每个神经元都是简单的，可是大脑产生了意识……

万幸的是，大型语言模型，也会涌现出各种意想不到的能力。

2022年8月，谷歌大脑研究者发布一篇论文，专门讲了大型语言模型的一些涌现能力，包括少样本学习、突然学会做加减法、突然之间能做大规模、多任务的语言理解、学会分类等等……而这些能力只有当模型参数超过1000亿才会出现——

我再强调一遍：研究者并没有刻意给模型植入这些能力，这些能力是模型自己摸索出来的。

就如同孩子长大往往会出乎家长的预料。

✳

当然你也得先把模型设计好才行。Transformer架构非常关键，它允许模型发现词与词之间的关系——不管是什么关系，而且不怕距离远。但是当初发明transformer的研究者，可没想到它能带来这么多新能力。

事后分析，涌现新能力的关键机制，叫做「思维链（Chain-of-Thought）」。

简单说，思维链就是当模型听到一个东西之后，它会嘟嘟囔囔自说自话地，把它知道的有关这个东西的各种事情一个个说出来。

比如你让模型描写一下"夏天"，它会说："夏天是个阳光明媚的季节，人们可以去海滩游泳，可以在户外野餐……"等等。

思维链是如何让语言模型有了思考能力的呢？也许是这样的。比如我们前面说的那个棒球棒问题。模型一听说棒球棒，它就自己跟自己叙述了棒球棒的各个方面，其中就包括大小；那既然你的问题中包括"放进耳朵"，大小就是一个值得标记出来的性质；然后对耳朵也是如此……它把两者大小的性质拿出来对比，发现是相反的，于是判断放不进去。

只要思考过程可以用语言描写，语言模型就有这个思考能力。

再看下面这个实验——

给模型看一张图片——皮克斯电影《机器人总动员》的一张剧照——问它是哪个制片厂创造了图中的角色。如果没有思维链，模型会给出错误的回答。

怎么用思维链呢？可以先要求模型自己把图片详细描述一番，它说"图中有个机器人手里拿了一个魔方，这张照片是从《机器人总动员》里面来的，那个电影是皮克斯制作的……"。这时候你简单重复它刚说的内容，再问它那个角色是哪个制片厂创造的，它就答对了。

既然如此，只要我们设置好让模型每次都先思考一番再回答问题，它就能自动使用思维链，它就有了思考能力。

有人分析 ，思维链很有可能是对模型进行编程训练的一个副产品。我们知道现在GPT-3是可以帮程序员编程的。在还没有接受过编程训练的时候，它的原理大概是先把一切媒体都转化成语言，再用语言模型处理。多模态模型可以做像下面这样的"看图片找规律"的智商测验题——

前面那个《机器人总动员》剧照的例子就来自这篇论文，演示了看图说话的思维链。论文里有这样一个例子，在我看来相当惊人——

给模型看一张图既像鸭子又像兔子的图，问它这是什么。它回答说这是个鸭子。你说这不是鸭子，再猜是什么？它说这像个兔子。你问它为什么，它会告诉你，因为图案中有兔子耳朵。

这个思维过程岂不是跟人一模一样吗？

✳

我看《荀子·劝学篇》中的一段话，正好可以用来描写AI能力的三个境界——

第一境界是「积土成山，风雨兴焉」。参数足够多，训练达到一定的积累，你就可以做一些事情。比如AlphaGo下围棋。

第二境界是「积水成渊，蛟龙生焉」。模型再大到一定程度，就会涌现出一些让人意想不到的神奇功能。比如AlphaZero不按人类套路下围棋、大型语言模型的思维链。

第三境界是「积善成德，而神明自得，圣心备焉」。这就是AGI了，它产生了自我意识，甚至有了道德感……

古往今来那么多人读《劝学》，也不知有几人真按照荀子的要求去学了……但是我们现在知道，AI肯定听进去了。你给它学习材料，它是真学。

✳

好，我们来总结一下这一讲。因为开悟和涌现，AI现在已经获得了包括推理、类比、少样本学习等等思考能力。

我们不得不重新思考以前对AI做出的各种假设——什么AI做事全靠经验、AI不会真的思考、AI没有创造力，包括"AI会的都是用语言可以表达的东西"这一条，现在我也不敢肯定了。

如果AI通过思维链能达到这样的思考水平，那人又是怎么思考的？我们的大脑是不是也有意无意也在使用了思维链呢？如果是这样，人脑跟AI到底有什么本质区别？

这些问题都在呼唤全新的答案。

以上就是这一讲的全部内容。下一讲，咱们聊聊AI能力的局限。

划重点　添加到笔记

1.AI能力的三个境界——
第一境界是「积土成山，风雨兴焉」。参数足够多，训练达到一定的积累，你就可以做一些事情。
第二境界是「积水成渊，蛟龙生焉」。模型再大到一定程度，就会涌现出一些让人意想不到的神奇功能。
第三境界是「积善成德，而神明自得，圣心备焉」。它产生了自我意识，甚至有了道德感。
2.因为开悟和涌现，AI现在已经获得了包括推理、类比、少样本学习等等思考能力。

首次发布: 今天

上一篇　01 大变局：一个新智基形态的产生
下一篇　03 底牌和命门：AI能力的局限

可免费赠送给20位朋友
微信　朋友圈

用户留言

最新　最热　只看作者回复

提交留言可与专栏作者互动

【知识福利社】你的私人书院，微信：312005650 全网热门课程众筹共享！