
AI “模型崩溃” 风险需警惕

作者：writer 来源：科学网

本文原地址：<https://www.iikx.com/news/progress/29632.html>

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

AI “模型崩溃” 风险需警惕。



图片来源：美国《The Week》

?

从客户服务到内容创作，人工智能（AI）影响了众多领域的进展。但是，一个日益严重的被称为“模型崩溃”的问题，可能会使AI的所有成就功亏一篑。

“模型崩溃”是今年7月发表在英国《自然》杂志上的一篇研究论文指出的问题。它是指用AI生成的数据集训练未来几代机器学习模型，可能会严重“污染”它们的输出。

多家外媒报道称，这不仅是数据科学家需要担心的技术问题，如果不加控制，“模型崩溃”可能会对企业、技术和整个数字生态系统产生深远影响。天津大学自然语言处理实验室负责人熊德意教授在接受科技日报记者采访时，从专业角度对“模型崩溃”进行了解读。

“模型崩溃”是怎么回事

大多数AI模型，比如GPT-4，都是通过大量数据进行训练的，其中大部分数据来自互联网。最初，这些数据是由人类生成的，反映了人类语言、行为和文化的多样性和复杂性。AI从这些数据中学习，并用它来生成新内容。

然而，当AI在网络上搜索新数据来训练下一代模型时，AI很可能会吸收一些自己生成的内容，从而形成反馈循环，其中一个AI的输出成为另一个AI的输入。当生成式AI用自己的内容进行训练时，其输出也会偏离现实。这就像多次复制一份文件，每个版本都会丢失一些原始细节，最终得到的是一个模糊的、不那么准确的结果。

美国《纽约时报》报道称，当AI脱离人类输入内容时，其输出的质量和多样性会下降。

熊德意解读称：“真实的人类语言数据，其分布通常符合齐普夫定律，即词频与词的排序成反比关系。齐普夫定律揭示了人类语言数据存在长尾现象，即存在大量的低频且多样化的内容。”

熊德意进一步解释道，由于存在近似采样等错误，在模型生成的数据中，真实分布的长尾现象逐渐消失，模型生成数据的分布逐渐收敛至与真实分布不一致的分布，多样性降低，导致“模型崩溃”。

AI自我“蚕食”是坏事吗

对于“模型崩溃”，美国《The Week》杂志近日刊文评论称，这意味着AI正在自我“蚕食”。

熊德意认为，伴随着这一现象的出现，模型生成数据在后续模型迭代训练中占比越高，后续模型丢失真实数据的信息就会越多，模型训练就更加困难。

乍一看，“模型崩溃”在当前似乎还是一个仅需要AI研究人员在实验室中担心的小众问题，但其影响将是深远而长久的。

美国《大西洋月刊》刊文指出，为了开发更先进的AI产品，科技巨头可能不得不向程序提供合成数据，即AI系统生成的模拟数据。然而，由于一些生成式AI的输出充斥着偏见、虚假信息和荒谬内容，这些会传递到AI模型的下一版本中。

美国《福布斯》杂志报道称，“模型崩溃”还可能会加剧AI中的偏见和不平等问题。

不过，这并不意味着所有合成数据都是不好的。《纽约时报》表示，在某些情况下，合成数据可以帮助AI学习。例如，当使用大型AI模型的输出训练较小的模型时，或者当可以验证正确答案时，比如数学问题的解决方案或国际象棋、围棋等游戏的最佳策略。

AI正在占领互联网吗

训练新AI模型的问题可能凸显出一个更大的挑战。《科学美国人》杂志表示，AI内容正在占领互联网，大型语言模型生成的文本正充斥着数百个网站。与人工创作的内容相比，AI内容的创作速度更快，数量也更大。

OpenAI首席执行官萨姆·奥特曼今年2月曾表示，该公司每天生成约1000亿个单词，相当于100万本小说的文本，其中有一大部分会流入互联网。

互联网上大量的AI内容，包括机器人发布的推文、荒谬的图片和虚假评论，引发了一种更为消极的观念。《福布斯》杂志称，“死亡互联网理论”认为，互联网上的大部分流量、帖子和用户都已被机器人和AI生成的内容所取代，人类不再能决定互联网的方向。这一观念最初只在网络论坛

上流传，但最近却获得了更多关注。

幸运的是，专家们表示，“死亡互联网理论”尚未成为现实。《福布斯》杂志指出，绝大多数广为流传的帖子，包括一些深刻的观点、犀利的语言、敏锐的观察，以及在新背景下对新生事物的定义等内容，都不是AI生成的。

不过，熊德意仍强调：“随着大模型的广泛应用，AI合成数据在互联网数据中的占比可能会越来越高，大量低质量的AI合成数据，不仅会使后续采用互联网数据训练的模型出现一定程度的‘模型崩溃’，而且也会对社会形成负面影响，比如生成的错误信息对部分人群形成误导等。因此，AI生成内容不仅是一个技术问题，同时也是社会问题，需要从安全治理与AI技术双重角度进行有效应对。”

作者：张佳欣 来源：科技日报

更多 科学进展 请访问 <https://www.iikx.com/news/progress/>

本文版权归原作者所有，请勿用于商业用途，[爱科学iikx.com](https://www.iikx.com)转发