检查更新

大规模人工智能语言系统显示一个涌现的有能力理由类比

类比推理是人类智能的标志，因为它使我们能够灵活地解决新问题，而无需大量练习。通过广泛的测试，我们证明了大规模人工智能语言模型 GPT-3 能够以与人类水平相当的水平解决困难的类比问题。

这是以下内容的摘要：
韦伯，T.等人。大型语言模型中的紧急类比推理。纳特。哼。行为。 https://doi.org/10.1038/s41562-02301659-w（2023）。

在线发布：2023 年 8 月 4 日

问题

人类推理具有无需大量实践或训练即可灵活解决新问题的独特能力。这种能力在很大程度上取决于类比推理——通过将不熟悉的问题与熟悉的问题进行比较来理解它的能力。认知科学的一个主要问题是人脑如何执行这一过程。这个问题与人工智能（AI）系统是否以及如何获得类似的类比推理能力的问题密切相关。特别是，对于深度学习系统（从经验中逐渐学习的神经元类单元的多层网络）如果接触到足够的训练数据，是否最终会发展出这种灵活推理的能力，存在着大量的争论。

发现

我们测试了 Generative Pre-trained Transformer 3 (GPT-3) 的一个版本，这是一个大规模深度学习系统，经过训练可以在一系列广泛的类比问题上生成类似人类的文本。最值得注意的是，该集合包括 Raven 渐进矩阵的新颖的基于文本的版本（图 1a、b），这是一种视觉类比问题集，通常被视为衡量人类解决问题能力的最佳指标之一。在原始（视觉）问题中，呈现了包含空白条目的几何图形数组（图 1a）。任务是使用此模式从一组潜在答案中选择正确的缺失图块来“填补空白”。我们通过将几何图形转换为数字来创建这些问题的基于文本的版本（图 1b）。我们还在其他类比问题上测试了 GPT-3，其中一些涉及现实世界的概念（例如，“爱：恨：：丰富：？”）或整个故事（其中的任务是确定两个目标故事中的哪一个）与源故事最相似）。重要的是，我们在没有对这些问题进行任何直接训练的情况下测试了 GPT-3，以反映人类在没有大量实践的情况下解决类比问题的能力。

我们发现，GPT-3 在大多数任务设置中都匹配或超过了大学生的表现（图 1c）。此外，GPT-3 表现出与人类参与者非常相似的错误率模式，因为它在解决人类往往认为困难的问题时表现出更大的难度（例如，受多个规则控制或需要更大程度的问题）抽象）。这些结果表明，作为学习生成类人的结果文本中，GPT-3 已经获得了解决类比问题的高度通用能力。

影响

许多认知科学家和人工智能研究人员的一个共同观点是，深度学习系统需要大量的特定任务训练，并且在训练过程中经历的条件之外的泛化能力有限。这些结果挑战了这种广泛持有的观点，并表明，如果对通用任务（即学习生成类似人类的文本）进行足够广泛的训练，深度学习系统就有可能获得推理新问题的能力。

重要的是要注意这些结果的一些局限性。 GPT-3是纯粹基于文本的，因此不能直接从视觉输入解决类比问题（如图1a所示的问题）。 GPT-3 还缺乏长期记忆能力，并且表现出较差的推理解决物理问题的能力（其无法解决涉及工具使用的简单构造问题就证明了这一点）。我们还发现，GPT-3 很难（相对于人类参与者）识别故事之间更抽象的类比，尽管 GPT-4 的表现比 GPT-3 更好。

这项研究提出的一个主要问题是 GPT-3 是否使用与人脑相似的机制来解决类比问题。尽管 GPT-3 等深度学习系统松散地受到大脑的启发（因为它们由类似神经元的处理单元组成，以多层的层次结构排列），但目前尚不清楚这些系统如何执行计算操作被认为是人类类比推理的基础。然而，与人脑不同的是，GPT-3 等系统的内部机制至少在原则上是可以直接探测的。这一特征表明，此类系统可以作为一种“模型有机体”来理解高阶认知过程的神经基础。为了实现这一目标，开发可供认知科学家研究的这些系统的开源版本以及评估它们所需的资源非常重要。这一进展还将使科学家能够提高我们对此类系统的优点和局限性的理解，并确保它们安全地部署到社会中。

泰勒·韦伯
加州大学心理学系，美国加利福尼亚州洛杉矶

专家意见

“类比是人类的核心认知能力，被认为需要特定的表征形式或定制的计算架构。这些结果挑战了这一立场，因为所考虑的计算模型是为序列学习和语言设计的通用神经网络

因此，它应该对认知科学和人工智能社区产生影响。实验严谨且有原则，这项工作将成为机器学习领域良好科学实践的灯塔

研究。” Felix Hill，谷歌 DeepMind，英国伦敦。

数字

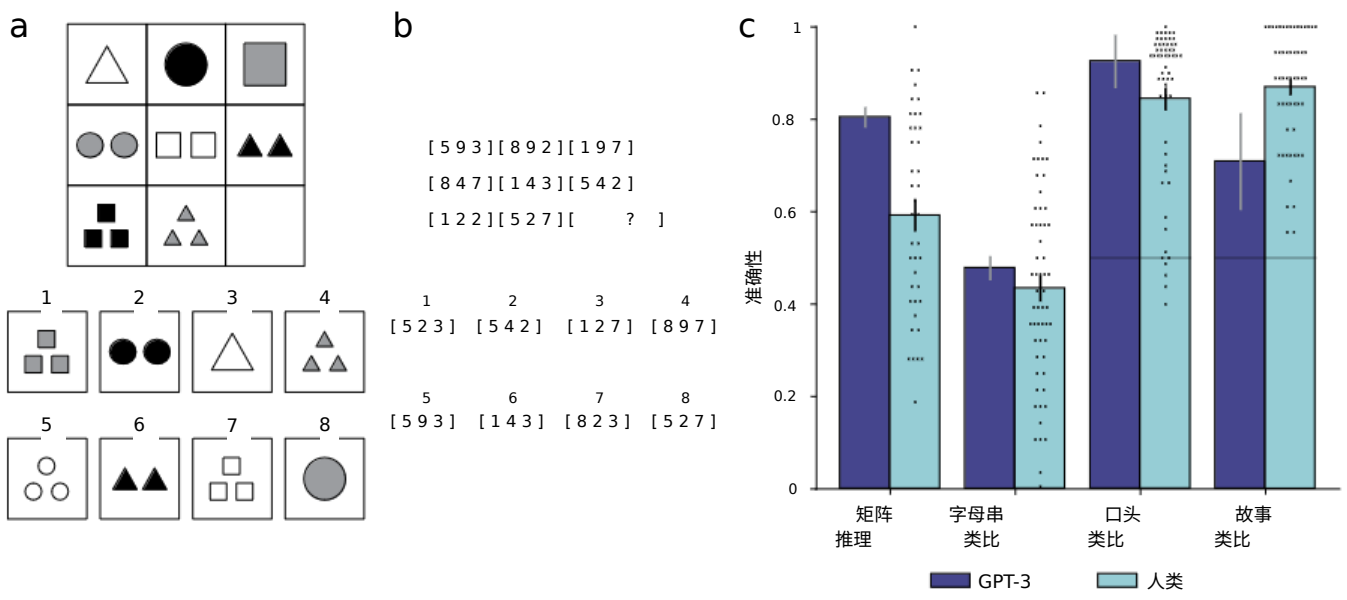

图1| GPT-3 和人类参与者的类比测试结果。 a,b，描述问题的示例

Raven 渐进矩阵的结构 (a) 和用于评估 GPT-3 的基于文本的版本 (b)。正确答案是5(a)和7(b)。 c，GPT-3在基于文本的矩阵推理问题上超越了人类参与者（大学生）的平均表现，并且在其他类比问题上也表现出了很强的表现。点代表个体参与者；条形表示平均值±s.e.m。 © 2023，Webb, T. 等人。

论文的背后

我和我的合著者对支持抽象推理的计算和神经机制有着长期的兴趣，特别是对神经网络模型（例如 GPT-3）的优点和局限性感兴趣。在之前的工作中，我们强调需要将标准神经网络方法与更结构化的推理操作相结合，以匹配人类推理的灵活性。因此，我们对 GPT-3 解决类比问题的能力感到非常惊讶，

这似乎表明抽象推理能力的出现。为了更好地理解这一观察结果，我们开发了新颖的测试材料（重要的是这些材料是新颖的，以确保 GPT-3 没有接受过针对它们的训练），并与人类行为进行比较来系统地评估 GPT-3。结果证实了这一初步结论，GPT-3 和 GPT-4 的后续变体表现出了更强的性能。 T.W.

来自编辑

“韦伯等人。表明新的人工智能语言模型，例如 GPT-3，能够以类似人类的性能水平解决类比推理问题。这一结果值得注意，因为这些模型从未经过明确的训练来执行此类任务，并且类比推理被广泛认为是人类的核心部分

智力。” Jamie Horder，《自然人类行为》高级编辑。

参考

1. 霍利奥克，K.J.牛津思维与推理手册（Holyoak，K. J. 编辑）
   ＆莫里森，R.G.）234–259（牛津大学）
   出版社，2012）。
   书籍章节总结了认知科学中类比推理的工作。

2. 布朗，T.等人。语言模型是小样本学习者。
   在高级中。神经信息处理系统 33
   （Larochelle, H. 等编辑）1877–1901
   （Curran Associates, 2020）。

   本文描述了 GPT-3，即当前工作中评估的人工智能系统。

3. Raven，J.C. 渐进矩阵：智力知觉测试，个体形式（Lewis Raven，1938）。

   视觉类比问题集，通常用作测试解决问题的能力。

4. Lake，B.M.等人。建造像人一样学习和思考的机器。行为。脑科学。 40，E253（2017）。

   描述深度学习系统的一些局限性的回顾和观点。

5. 米切尔，M。人工智能中的抽象和类比。

   安.纽约学院。科学。 1505, 79–101 (2021)。
   总结人工智能类比推理工作的评论。

6. Lu, H., Ichien, N. 和 Holyoak, K. J. 语义关系网络的概率类比映射。心理。牧师。

   129, 1078 (2022).
   将深度学习与结构化推理操作相结合的工作示例。