Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

News & Highlights

数学推理挑战人工智能

Sean O'Neill

Senior Technology Writer

深度神经网络形式的人工智能（AI）在从图像识别到游戏，从自然语言翻译到语音合成的各种高质量使用情景中均有着出色的表现[1–4]。但是，至少在当前，它在数学推理中的表现显得不那么尽如人意，而数学推理被认为是人类智能的一项核心能力。

2019年4月，谷歌母公司Alphabet旗下聚焦AI的企业DeepMind Technologies在伦敦的团队发表了研究结果，探讨最先进的通用神经网络执行数学运算的能力[5]。为了提供一个易于理解的评估，DeepMind的最佳表现模型在“考试”中测试了40道题，这些题是从英国16岁学生的公开数学考试中抽取的。在40分总分中，DeepMind得到了14分，并没有达到及格的分数，导致媒体发布了如“DeepMind AI在高中数学考试中失败”的头条新闻[6]。

“这种评价可能有点不公平，不过结果并不令人惊讶”，总部位于旧金山，专注于AI的公司OpenAI的首席科学家Ilya Sutskever说[7]。微软已同意向OpenAI投资10亿美元，“主要目的是研究当前常用的神经网络的功能”，Sutskever说，他参与了DeepMind团队之前的一些工作[8,9]。“结果表明，在数学推理方面，这些神经网络模型遭遇了困难。”

这种推理对人工系统具有挑战性，因为它不仅仅涉及处理数字，还需要一套认知能力，包括学习基本公理以及以正确的顺序进行推理、计划和做事的能力，当然，AI首先需要读懂问题。“任何非常有用的AI系统都需要能够处理数学、推理和计算，并在现实世界中灵活地运用这些技能，”Sutskever说，“因此，明智的做法是尝试教AI数学。”

DeepMind数据集基于英国国家学校数学课程，并包含诸如代数、算术、微积分、比较、多项式和概率之类的模块。对于每个模块，该团队生成了200万个问题（输入）和答案（输出），对它们进行了神经网络模型的训练，并生成了10万个问题，随后用这些问题对它们进行了测试。

虽然最近的一些研究已经探索了AI解决代数语言问题的能力，如西雅图的华盛顿大学艾伦人工智能研究所的欧几里德项目[10]，但DeepMind数据集更侧重于数学推理，而不是对问题的语言理解。为此，它涵盖了更多的数学领域，但在如何提出问题的角度上变化较小。DeepMind论文的第一作者、研究工程师David Saxton说：“如果我们能开发出更先进、能够很好地解决此数据集中的问题的模型，这些模型很可能使用一些通用技巧，那么这些技巧也能够很好地解决AI中的其他难题。”

经过测试，性能最佳的模型是Transformer，该模型由机器学习小组GoogleBrain的Ashish Vaswani及其同事和美国加利福尼亚州山景城的Google Research在2017年推出[11]。它是长短期记忆（LSTM）模型的变异模型，其性能与其余模型相比相同甚至更好。Saxton说：“最令人惊讶的是，开箱即用的语言模型，如Transformer，在许多类型的数学问题上表现很好。”例如，在涉及舍入和比较数量级的问题上，它获得了近乎完美的分数。

对于AI来说，最难回答的问题是那些需要更多理论和程序知识的问题，如对人类来说也很难的因子分解。这个判断是有道理的：“似乎仅从输入/输出示例本身无

英文原文: Engineering 2019, 5(5): 817–818
引用本文: Sean O'Neill. Mathematical Reasoning Challenges Artificial Intelligence. *Engineering*, https://doi.org/10.1016/j.eng.2019.08.009

法推断出合成规则”，爱丁堡大学贝叶斯中心校长、伦敦艾伦·图灵研究所的常任理事Michael Rovatsos说。而爱丁堡大学贝叶斯中心和伦敦艾伦·图灵研究所是人工智能和数据科学领域的顶尖研究院所。

Transformer模型正确回答了“加或减数”和“乘或除数”模块上90%或更多的问题。但是，在涉及使用括号将四则运算混合在一起的问题上，其正确率降低到了50%。在他们的论文中，作者推测产生较差结果的原因是，尽管基本运算可以以相对线性、直接的方式执行，但“没有捷径可以用来评估含有括号的算术表达式，因为这需要计算中间值”。而一些具备数学基础知识的人会知道该怎么做。研究人员以此为依据，证明这些模型没有学到任何代数或算法对值的操作，而是“学习相对较浅的技巧”以获得答案。

测试还产生了一些意想不到的结果（图1）。在一个问题上，训练有素的Transformer模型正确地回答了“Caculate 17×4.”，其结果为68。同一问题但没有句点，得出的答案为69。其他测试问题为1+1+…+1，其中n表示1出现的次数。对于$n \leqslant 6$，LSTM和Transformer模型均正确回答；对于$n=7$，模型回答6；在$n>7$的情况中，它们以其他不正确的值响应。

该研究的重要贡献之一是模块化的数据集，因此易于扩展。作者写道：“我们希望该数据集将成为开发具有更多功能的模型的可靠的可分析基准。”他们指出，未来将扩展数据集以包括更大的语言复杂性和视觉问题，如几何形状。对于神经网络本身，Saxton说，DeepMind团队下一步将开发可以通过学习在代数/符号推理任务中表现出色的模型。

但是，可能最重要的是确定模型得出错误答案的原因（图2）。Sutskever说：“在有可靠的工具告诉我们为什么神经网络会产生它的答案之前，还有很长的路要走。”

图1. 在DeepMind的数据集上进行数学推理训练和测试的神经网络有时会以意外和令人惊讶的方式失败。模型确实解决了这个普遍的问题（1 +1 = 2），以及1 +1+…+1的相关问题，其中1出现n次，直到n =6。但是，对于n = 7，模型以6作答；对于n> 7，它们以其他不正确的值响应。图片来自：Pexels（公共领域）。

对于某些研究人员而言，这个神秘的、神经网络的“黑箱”元素——无法理解它们是如何做出决定的——代表了通用人工智能（AGI）技术发展过程中的一个关键问题。“令人担心的是，我们专注于量化的绩效而不是可理解性，”Rovatsos说，“如果AI真的开始发展人类的智能并被广泛应用于日常使用中，我们就要仔细检查并更正这些系统，以确保其行为符合我们的社会规范和道德价值观。在我看来，我们正在制造‘赛车’，而不是能够安全带我们到达目的地的车辆。”

图2. 在解决涉及AI的问题时，深层神经网络建立的联系通常是“黑箱”之谜，神经网络不容易理解，难以确定它们出了什么问题以及如何纠正故障。图片经许可来自：DeepMind。

References

[1] Lee TB. How computers got shockingly good at recognizing images [Internet]. Ars Technica; 2018 Dec 18 [cited 2019 Jul 24]. Available from: https://arstechnica.com/science/2018/12/how-computers-got-shockingly-good-atrecognizing- images/.

[2] Stokel-Walker C. DeepMind AI thrashes human professionals at video game StarCraft II [Internet]. London: New Scientist; 2019 Jan 24 [cited 2019 Jul 24]. Available from: https://www.newscientist.com/article/2191910-deepmind-aithrashes- human-professionals-at-video-game-starcraft-ii/.

[3] Joshi P. A must-read NLP tutorial on neural machine translation—the technique powering Google Translate [Internet]. Medium; 2019 Jan 31 [cited 2019 Jul 24]. Available from: https://medium.com/analytics-vidhya/a-must-readnlp- tutorial-on-neural-machine-translation-the-technique-powering-googletranslate- c5c8d97d7587.

[4] Wang X, Takaki S, Yamagishi J. Neural source-filter-based waveform model for statistical parametric speech synthesis. In: Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing; 2019 May 12–17; Brighton, UK. Piscataway: IEEE; 2019. p. 5916–20.

[5] Saxton D, Grefenstette E, Hill F, Kohli P. Analysing mathematical reasoning abilities of neural models. 2019. arXiv:1904.01557.

[6] Tian R. DeepMind AI flunks high school math test [Internet]. Medium; 2019 Apr 5 [cited 2019 Jul 24]. Available from: https://medium.com/syncedreview/deepmind-ai-flunks-high-school-math-test-2e32635c0e2d.

[7] Nellis S. Microsoft to invest $1 billion in OpenAI [Internet]. London: Reuters; 2019 Jul 22 [cited 2019 Jul 24]. Available from: https://www.reuters.com/article/ us-microsoft-openai/microsoft-to-invest-1-billion-in-openai-idUSKCN1UH1H9.

[8] Kaiser L, Sutskever I. Neural GPUs learn algorithms. 2015. arXiv:1511.08228.

[9] Zaremba W, Sutskever I. Learning to execute. 2014. arXiv:1410.4615.

[10] Euclid [Internet]. Seattle: Allen Institute for Artificial Intelligence; [cited 2019 Jul 24]. Available from: http://allenai.org/euclid/.

[11] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA. p. 5998–6008.