Engineering 5 (2019) 817–818

Contents lists available at ScienceDirect

Engineering

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

News & Highlights

Mathematical Reasoning Challenges Artificial Intelligence

Sean O'Neill

Senior Technology Writer

Artificial intelligence (AI) in the form of deep neural networks excels in a variety of high-profile settings, from image recognition and gaming to natural language translation and voice synthesis [1–4]. However, its performance in mathematical reasoning, considered a core ability of human intelligence, appears somewhat less impressive—at least at the present.

In April 2019, a London-based team at DeepMind Technologies, the AI-focused enterprise owned by Alphabet, Inc., the parent company of Google, published research exploring the ability of state-of-the-art, general purpose neural networks to perform mathematical reasoning [5]. To provide an easily understood assessment, DeepMind's best-performing model was tested on an "exam" of 40 questions drawn from publicly available mathematics exams for British 16-year-olds. It scored 14 out of 40, equivalent to a failing grade, leading to media headlines such as "DeepMind AI flunks high school math test" [6].

That assessment may be a little unfair, as the outcome is not that surprising, said Ilya Sutskever, Chief Scientist at OpenAI, a San Francisco-based, AI-focused company in which Microsoft has agreed to invest $1 billion USD [7]. "The goal was mainly to investigate the capabilities of current, commonly used neural networks," said Sutskever, who co-authored previous work that was cited by the DeepMind team [8,9]. "The results demonstrated that when it comes to mathematical reasoning, these neural network models struggle."

Such reasoning is challenging to artificial systems because it involves more than merely crunching numbers: It requires a suite of cognitive abilities, including learning the underlying axioms and the ability to infer, plan, and do things in the right order, and, of course, understand the question in the first place. "Any very useful AI system will need to be able to deal with math, reasoning, and calculation, and flexibly apply these kinds of skills in real-world settings," said Sutskever. "So, it is sensible to see if AI can be taught mathematics."

The DeepMind data set was based on the UK national school mathematics curriculum and included modules such as algebra, arithmetic, calculus, comparisons, polynomials, and probability. For each module, the team generated 2 million questions (inputs) and answers (outputs) on which the neural network models were trained, and 100 000 questions on which the trained models were subsequently tested.

While some recent research has explored AI's ability to solve algebraic word problems, such as work done by the Euclid project at the Allen Institute for Artificial Intelligence at the University of Washington in Seattle [10], the DeepMind data set focused more on mathematical reasoning than on the linguistic comprehension of the questions. To this end, it covered more areas of mathematics, with less variation in how the questions themselves were posed. "If we can develop more sophisticated models that are good at solving the problems in this data set, then these models would likely be using general skills that would be good at solving other hard problems in AI as well," said the first author of the DeepMind paper [5], Research Engineer David Saxton.

Upon testing, the best-performing model was the Transformer, introduced in 2017 by Ashish Vaswani and colleagues at Google Brain, a machine-learning group, and Google Research in Mountain View, California, USA [11]. The model did as well or better than the rest, which were variations on the long short-term memory (LSTM) model. "Most surprising was how well an out-of-the-box language model—the Transformer—could do across many types of mathematics questions," said Saxton. It achieved nearly perfect scores, for example, on questions involving rounding and comparing magnitudes.

The hardest questions for the AI to answer included those that required more theoretical and procedural knowledge, such as factorization, which are hard for humans too. That makes sense: "It seems extraordinarily hard to infer the compositional rules themselves just from input/output examples," said Michael Rovatsos, who directs the University of Edinburgh's Bayes Centre and is also affiliated with the Alan Turing Institute in London, both leading institutes in artificial intelligence and data science.

The Transformer model correctly answered 90% or more of the problems on the modules "add or subtract several numbers" and "multiply or divide several numbers." But on problems that involved the mixing of all four operations together using parentheses, performance decreased to 50% correct. In their paper, the authors speculated that the poorer outcome occurs because while basic operations can be performed in a relatively linear, straightforward manner, there are "no shortcuts to evaluating arithmetic expressions with parentheses, where intermediate values need to be calculated," something someone with basic knowledge of mathematics would know how to do. The researchers took this as evidence that the models had not learned any algebraic/algorithmic manipulation of values but were instead "learning relatively shallow tricks" to obtain answers.

The testing also produced some unexpected results (Fig. 1). On one question a trained Transformer model correctly answered

https://doi.org/10.1016/j.eng.2019.08.009

Fig. 1. The neural networks trained and tested on DeepMind's data set for mathematical reasoning sometimes failed in unexpected and surprising ways. The models did correctly solve this proverbial problem (1 + 1 = 2), as well the related problems of 1 + 1 + ··· + 1, where 1 occurs n times, up to n = 6. However, for n = 7, the models answered 6, and for n > 7, they responded with other incorrect values. Credit: Pexels (public domain).

"Calculate 17 × 4." as 68. The same question, but without the period, resulted in the answer 69. Other test questions posed 1 + 1 + ··· + 1, where 1 occurs n times. For $n \leq 6$, both the LSTM and Transformer models answered correctly. For n = 7, the models answered 6. For n > 7, they responded with other incorrect values.

One important contribution of the research is the modular, and therefore easily extendable, data set. "We hope this data set will become a robust analysable benchmark for developing models with more abilities," the authors wrote. Useful future work, they note, would be to extend the data set to include greater linguistic complexity, and visual problems, such as geometry. In terms of the neural networks themselves, Saxton said the next step for the DeepMind team is to develop models that can learn to do well across algebraic/symbolic reasoning tasks.

But what may be perhaps most important to determine is how the models arrived at their wrong answers (Fig. 2). "We've got a long way to go before we have reliable tools that can tell us why a neural network produced the answer it did," said Sutskever.

For some researchers, this mysterious, "black box" element of neural networks—not being able to understand how they arrive at decisions, represents a key issue as the technology moves toward artificial general intelligence (AGI). "It worries me that we are focusing on quantitative performance rather than intelligibility," said Rovatsos. "Should AI really start to develop human-level intelligence and be widely adopted for everyday use, we would want to scrutinise and correct these systems to ensure their behavior complies with our social norms and moral values. It seems to me we're building 'racing cars,' rather than vehicles that can take us safely to our destination."

Fig. 2. The connections deep neural networks make when tackling a problem involving artificial intelligence are often a "black-box" mystery—not easily understood, making it hard to determine where they go wrong and how to correct for the failure. Credit: courtesy of DeepMind, with permission.

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-at-recognizing-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-ai-thrashes-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-read-nlp-tutorial-on-neural-machine-translation-the-technique-powering-google-translate-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.