THE COMPUTATIONAL LIMITS OF DEEP LEARNING

By Neil C. Thompson, Kristjan Greenewald, Keeheon Lee, Gabriel F. Manso

IN THIS BRIEF

- The computational demands of deep learning applications in areas such as image classification, object detection, question answering, and machine translation are strongly reliant on increases in computing power–an increasingly unsustainable model.
- Deep learning is intrinsically more dependent on computing power than other techniques because these models have more parameters, and require more data to train.
- The strong reliance on computing power for deep learning reveals that progress is rapidly becoming economically, technically, and environmentally unsustainable as limits are stretched.
- Continued progress will require dramaticlly more computationally efficient methods either from changes to deep learning itself, or by moving to other machine learning methods.

Deep learning's recent history has been one of achievement: from triumphing over humans in the game of Go to world-leading performance in image and voice recognition, translation, and other tasks. But this progress has come at a steep cost: a voracious appetite for computing power.

This article reports on the computational demands of deep learning applications in five prominent application areas: Image classification; object detection; question answering; named entity recognition, and machine translation. It shows that progress in all five is strongly reliant on increases in computing power. Extrapolating forward, this reliance reveals that progress along current lines is rapidly becoming economically, technically, and environmentally unsustainable. Continued progress will require dramatically more computationally efficient methods, which either will have to come from changes to deep learning itself, or from moving to other machine learning methods.

Even when the first neural networks were created, performance was limited by available computation. In the past decade, these constraints have relaxed along with specialized hardware (e.g. GPUs) and a willingness to spend more on processors. However, because the computational needs of deep learning scale so rapidly, they are quickly becoming burdensome again.

DEEP LEARNING REQUIREMENTS IN THEORY

The relationship between performance, model complexity, and computational requirements in deep learning is still not well understood theoretically. Nevertheless, there are important reasons to believe that deep learning is intrinsically more dependent on computing power than other techniques, particularly because of the role of "over-parameterization" – when a model has more parameters than data points – and how this scales. An example of large-scale over-parameterization is the current state-of-the-art image recognition system, NoisyStudent, which has 480M parameters for Imagenet's 1.2M data points.

The challenge with deep learning is that both the size of the network and the number of data points must grow rapidly to improve performance.

Since the cost of training a deep learning model scales with the product of the number of parameters and the number of data points, computational requirements apparently grows as the square of the number of data points in the over-parameterized setting. This quadratic scaling, however, is an underestimate of how fast deep learning networks must grow to improve performance, since the amount of training data must scale much faster than linearly in order to get linear performance improvements.

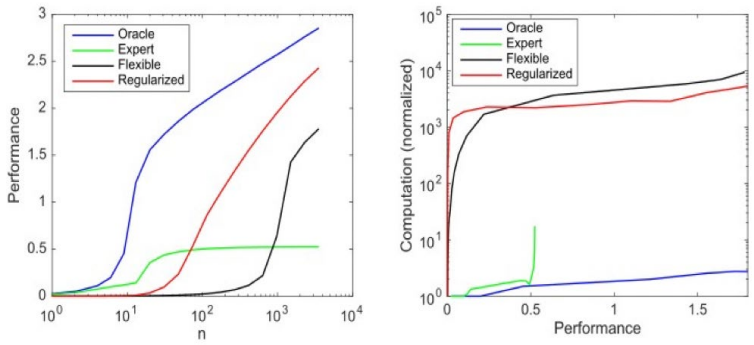

Figure 1: The effects of model complexity and regularization on model performance (measured as the negative log10 of normalized mean squared error of the prediction compared to the optimal predictor) and on computational requirements, averaged over 1000 simulations per case. (a) Average performance as sample sizes increase. (b) Average computation required to improve performance.

Figure 1 generalizes an insight attributed to Andrew Ng: That traditional machine learning techniques do better when the amount of data is small, but that flexible deep learning models do better with more data. We argue that this is a more general phenomenon of flexible models having greater potential, but also having vastly greater data and computational needs. In our example, 1,500 observations are needed for the "flexible" model to reach the same performance as the "oracle" with 15. Regularization helps with this, dropping the data need to 175, but it is much less helpful with computational costs, as Figure 1(b) shows.

In sum, deep learning performs well because it uses over-parameterization to create a highly flexible model and uses (implicit) regularization to make the complexity tractable. At the same time, however, deep learning requires vastly more computation than more efficient models. Paradoxically, the great flexibility of deep learning inherently implies a dependence on large amounts of data and computation.

DEEP LEARNING'S REQUIREMENTS IN PRACTICE

Early on it was clear that computational requirements limited what neural networks could achieve. In 1960, when Frank Rosenblatt wrote about a three-layer neural network, there were hopes that it had "gone a long way toward demonstrating the feasibility of a perceptron as a pattern-recognizing device." But as Rosenblatt recognized, "as the number of

connections in the network increases, the burden on a conventional digital computer soon becomes excessive."

In1969, Minsky and Papert noted a potential solution: Introducing longer chains of intermediate units (that is, by building deeper neural networks). Despite this potential workaround, much of the academic work in this area was abandoned because there simply wasn't enough computing power available.

In the decades that followed, improvements in computer hardware provided, by one measure, a ≈ 50,000× improvement in performance and neural networks grew their computational requirements proportionally. Since the growth in computing power per dollar closely mimicked the growth in computing power per chip, this meant that the economic cost of running such models was largely stable over time. Despite this large increase, deep learning models in 2009 remained "too slow for large-scale applications, forcing researchers to focus on smaller-scale models, or to use fewer training examples."

The turning point seems to have been when deep learning was ported to GPUs, initially yielding a 5 – 15× speed-up. By 2012 the increase grew to more than 35×, which led to the important victory of AlexNet at the 2012 Imagenet competition. But image recognition was just the first of these benchmarks to fall. Soon, deep learning systems also won at object detection, named-entity recognition, machine translation, question answering, and speech recognition.

The introduction of GPU-based (and later ASIC-based) deep learning led to widespread adoption of these systems. But the amount of computing power used in cutting-edge systems grew even faster--at approximately 10× per year from 2012 to 2019. This rate far exceeded the ≈ 35× total improvement gained from moving to GPUs-- meager improvements from the last vestiges of Moore's Law – or the improvements in neural network training efficiency. Instead, much of the increase came from a less economically attractive source: Running models for more time on more machines. It turns out that scaling deep learning computation by increasing hardware hours or number of chips is problematic because it implies that costs scale at roughly the same rate as increases in computing power, which will quickly make it unsustainable.

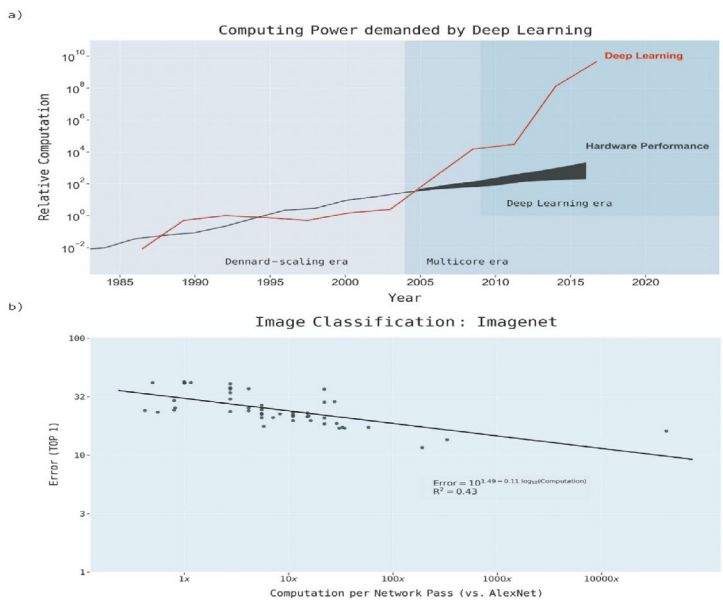

Figure 2: Computing power used in: (a) deep learning models of all types (as compared with the growth in hardware performance from improving processors, as analyzed by and), (b) image classification models tested on the ImageNet benchmark (normalized to the 2012 AlexNet model).

METHODOLOGY

For our research, we perform two separate analyses of computational requirements reflecting the two types of information available: (1) Computation per network pass (the number of floating point operations required for a single pass in the network, also measurable using multiply-adds, or the number of parameters in the model), and (2) Hardware burden (the computational capability of the hardware used to train the model, calculated as #processors × ComputationRate × time), which is shown in the full paper.

First, we demonstrate our analysis in the area with the most data and longest history: image classification. Here, the relevant performance metric is classification error rate. We find that the computation requirements scale as O(Performance9) for image recognition, and with exponents between 7.7 and 50 for other areas. Collectively, our results make it clear that progress in training models has depended on large increases in the amount of computing power used—a dependence not unique to deep learning, but seen in other areas such as weather prediction and oil exploration.

We extrapolate the estimates from each domain to understand the projected computational power needed to hit various benchmarks. To make these targets tangible, we present them not only in terms of the computational power required, but also in terms of the economic and environmental cost of training such models on current hardware (using the conversions from). Because the polynomial and exponential functional forms have roughly equivalent statistical fits – but quite different extrapolations – we report both in Figure 3.

Benchmark	Error rate	Polynomial			Exponential		
		Computation Required (Gflops)	Environmental Cost (CO_2)	Economic Cost ($)	Computation Required (Gflops)	Environmental Cost (CO_2)	Economic Cost ($)
ImageNet	Today: 11.5%	10^{14}	10^{6}	10^{6}	10^{14}	10^{6}	10^{6}
	Target 1: 5%	10^{19}	10^{10}	10^{11}	10^{27}	10^{19}	10^{19}
	Target 2: 1%	10^{28}	10^{20}	10^{20}	10^{120}	10^{112}	10^{112}
MS COCO	Today: 46.7%	10^{14}	10^{6}	10^{6}	10^{15}	10^{7}	10^{7}
	Target 1: 30%	10^{23}	10^{14}	10^{15}	10^{29}	10^{21}	10^{21}
	Target 2: 10%	10^{44}	10^{36}	10^{36}	10^{107}	10^{99}	10^{99}
SQuAD 1.1	Today: 4.621%	10^{13}	10^{4}	10^{5}	10^{13}	10^{5}	10^{5}
	Target 1: 2%	10^{15}	10^{7}	10^{7}	10^{23}	10^{15}	10^{15}
	Target 2: 1%	10^{18}	10^{10}	10^{10}	10^{40}	10^{32}	10^{32}
CoLLN 2003	Today: 6.5%	10^{13}	10^{5}	10^{5}	10^{13}	10^{5}	10^{5}
	Target 1: 2%	10^{43}	10^{35}	10^{35}	10^{82}	10^{73}	10^{74}
	Target 2: 1%	10^{61}	10^{53}	10^{53}	10^{181}	10^{173}	10^{173}
WMT 2014 (EN-FR)	Today: 54.4%	10^{12}	10^{4}	10^{4}	10^{12}	10^{4}	10^{4}
	Target 1: 30%	10^{23}	10^{15}	10^{15}	10^{30}	10^{22}	10^{22}
	Target 2: 10%	10^{43}	10^{35}	10^{35}	10^{107}	10^{99}	10^{100}

Figure 3: Implications of achieving performance benchmarks on the computation (in Gigaflops), carbon emissions (lbs.), and economic costs ($USD) from deep learning based on projections from polynomial and exponential models.

We do not anticipate that the computational requirements implied by the targets in Figure 3 will be hit. The hardware, environmental, and monetary costs would be prohibitive. Moreover, as we note, enormous effort is going into improving scaling performance. Nonetheless, these projections provide a scale for the efficiency improvements that would be needed to hit these performance targets. For example, even in the more-optimistic model, it is estimated to take an additional 105× more computing to get to an error rate of 5% for ImageNet. Hitting this in an economical way will require more efficient hardware, more efficient algorithms, or other improvements such that the net impact is this large a gain.

The rapid escalation in computing needs in Figure 3 also makes a stronger statement: Based on current trends it will be impossible for deep learning to hit these benchmarks. Instead, fundamental re-architecting is needed to lower the computational intensity so that the scaling of these problems becomes less onerous. And there is promise that this could be achieved. Theory tells us that the lower bound for the computational intensity of

regularized flexible models might be as low as O(Performance4), which is much better than current deep learning scaling. Encouragingly, there is historical precedent for algorithms improving rapidly.

The economic and environmental burden of hitting the performance benchmarks noted previously suggest that deep learning is facing an important challenge: Either find a way to increase performance without increasing computing power, or have performance stagnate as computational requirements become a constraint. Approaches to address this challenge include increasing computing power through hardware accelerators, reducing computational complexity through network compression and acceleration, and finding high-performing small deep learning architectures.

CONCLUSION

The explosion in computing power used for deep learning models has set new benchmarks for computer performance on a wide range of tasks. However, deep learning's prodigious appetite for computing power imposes a limit on how far it can improve performance in its current form, particularly in an era when improvements in hardware performance are slowing.

This paper shows that the computational limits of deep learning will soon be constraining for a range of applications, making the achievement of important benchmark milestones impossible if current trajectories hold. Finally, we have discussed the likely impact of these computational limits: Forcing deep learning toward less computationally intensive methods of improvement, and pushing machine learning toward techniques that are more computationally efficient than deep learning.

REPORT

The full research paper can be found here.

ACKNOWLEDGEMENT

The authors would like to acknowledge funding from the MIT Initiative on the Digital Economy and the Korean Government's National Research Foundation.

ABOUT THE AUTHORS

Neil C. Thompson is a Research Scientist at MIT's Computer Science and Artificial Intelligence Lab and a Visiting Professor at the Lab for Innovation Science at Harvard.

Kristjan Greenewald is an artificial intelligence research scientist at the MIT-IBM Watson AI Lab in Cambridge Massachusetts, a joint lab between IBM Research AI and MIT.

Keeheon Lee is an Assistant Professor of Creative Technology Management Professor Lee majored in four different areas that dealt with strategic management, scientometrics, data science, and business intelligence.

Gabriel F. Manso Gabriel F. Manso is a Research Assistant at MIT's Computer Science and Artificial Intelligence Lab and at the Initiative on the Digital Economy. He is also a senior student of Software Engineering at the University of Brasília (UnB).

REFERENCES

Dario Amodei and Danny Hernandez. Ai and compute. 2018.

NVIDIA Corporation. Tesla P100 Performance Guide - HPC and Deep Learning Applications. NVIDIA Corporation, 2017.

Andrew Danowitz, Kyle Kelley, James Mao, John P. Stevenson, and Mark Horowitz. CPU DB: Recording microprocessor history. Queue, 10(4):10:10–10:27, 2012.

J Gambetta and S Sheldon. Cramming more power into a quantum device. IBM Research Blog, 2019.

John L. Hennessy and David A. Patterson. Computer Architecture: A Quantitative Approach. Morgan Kaufmann, San Francisco, CA, sixth edition, 2019.

Danny Hernandez and Tom Brown. Ai and efficiency. 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems, pages 1097–1105, 2012.

Karin Kruup. Clearing the buzzwords in machine learning, May 2018.

Charles E. Leiserson, Neil C. Thompson, Joel Emer, Bradley C. Kuszmaul, Butler W. Lampson, Daniel Sanchez, and Tao B. Schardl. Theres plenty of room at the top: What will drive growth in computer performance after Moore's law ends? Science, 2020.

Marvin Minsky and Seymour Papert. Perceptrons: An Introduction to Computational Geometry. MIT Press, Cambridge, MA, USA, 1969.

Rajat Raina, Anand Madhavan, and Andrew Ng. Large-scale deep unsupervised learning using graphics processors. Proceedings of the 26th International Conference on Machine Learning, 2009.

Frank Rosenblatt. Perceptron simulation experiments. Proceedings of the IRE, 48(3):301– 309, March 1960.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy,

Aditya Khosla, Michael Bernstein, Alexander Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3):211–252, 2015.

Yash Sherry and Neil Thompson. How fast do algorithms improve? Mimeo, 2020.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. arXiv preprint arXiv:1906.02243, 2019.

Neil Thompson, Shuning Ge, and Gabriel Filipe. The importance of (exponentially more) computing power. Mimeo, 2020.

Neil Thompson and Svenja Spanuth. The decline of computers as a general purpose technology: Why deep learning and the end of Moores law are fragmenting computing. Available at SSRN 3287769, 2018.

Qizhe Xie, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Self-training with noisy student improves Imagenet classification. arXiv preprint arXiv:1911.04252, 2019.

MIT INITIATIVE ON THE DIGITAL ECONOMY

The MIT IDE is solely focused on the digital economy. We conduct groundbreaking research, convene the brightest minds, promote dialogue, expand knowledge and awareness, and implement solutions that provide critical, actionable insight for people, businesses, and government. We are solving the most pressing issues of the second machine age, such as defining the future of work in this time of unprecendented disruptive digital transformation.

SUPPORT THE MIT IDE

The generous support of individuals, foundations, and corporations are critical to the success of the IDE. Their contributions fuel cutting-edge research by MIT faculty and graduate students, and enables new faculty hiring, curriculum development, events, and fellowships. Contact Devin Wardell Cook (devinc@mit.edu) to learn how you or your organization can support the IDE.

TO LEARN MORE ABOUT THE IDE, INCLUDING UPCOMING EVENTS, VISIT IDE.MIT.EDU