Energy and Policy Considerations for Modern Deep Learning Research

Emma Strubell
Facebook AI Research
strubell@fb.com

Ananya Ganesh
University of Massachusetts Amherst
aganesh@cs.umass.edu

Andrew McCallum
University of Massachusetts Amherst
mccallum@cs.umass.edu

Abstract

The field of artificial intelligence has experienced a dramatic methodological shift towards large neural networks trained on plentiful data. This shift has been fueled by recent advances in hardware and techniques enabling remarkable levels of computation, resulting in impressive advances in AI across many applications. However, the massive computation required to obtain these exciting results is costly both financially, due to the price of specialized hardware and electricity or cloud compute time, and to the environment, as a result of non-renewable energy used to fuel modern tensor processing hardware. In a paper published this year at ACL, we brought this issue to the attention of NLP researchers by quantifying the approximate financial and environmental costs of training and tuning neural network models for NLP (Strubell, Ganesh, and McCallum 2019). In this extended abstract, we briefly summarize our findings in NLP, incorporating updated estimates and broader information from recent related publications, and provide actionable recommendations to reduce costs and improve equity in the machine learning and artificial intelligence community.

Introduction

Recent advances in methodology and computational hardware have enabled exciting advances across many application areas of artificial intelligence, such as game playing (Silver et al. 2017; OpenAI 2018), natural language processing (Devlin et al. 2019; Aharoni, Johnson, and Firat 2019), computer vision (Chollet 2017; Brock, Donahue, and Simonyan 2019) and robotics (Agostinelli et al. 2019). Many of these impressive results depend on training large models on considerable quantities of data, incurring substantial financial and environmental costs due to the energy required to perform this computation. Whereas a decade ago most AI research could be performed on a commodity desktop computer, modern deep learning research increasingly requires access to a cluster containing specialized tensor processing hardware such as GPUs and TPUs, and obtaining state-of-the-art performance on common benchmarks requires days or weeks of training on tens or hundreds of these nodes.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

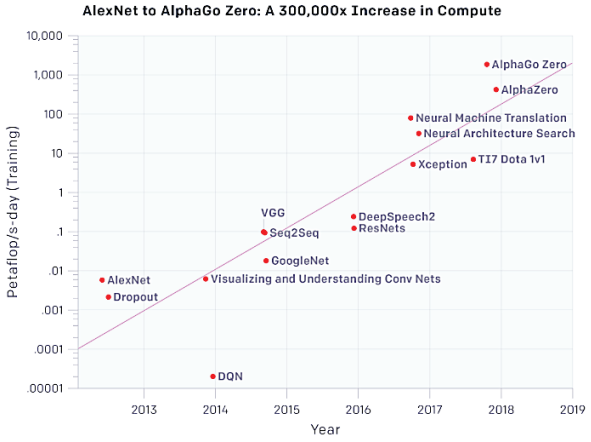

Figure 1: Training time (petaflop/s-day; log scale) of notable AI models from 2012–2018. Compute used to train the largest models continues to grow exponentially, exceeding the rate of Moore's Law by a wide margin. Figure and analysis from Amodei and Hernandez (2018).

In this article, we summarize previous work characterizing the energy required to train and develop recent deep learning models for NLP, and share conclusions and recommendations inspired by those results that apply broadly to artificial intelligence researchers and practitioners.

Case study 1: Training

To quantify the computational and environmental cost of training deep neural network models for NLP, we first perform an analysis of the energy required to train four popular off-the-shelf NLP models. We do this by training the models described below using the default settings provided, and sample power consumption during training using readily available command-line tools. We then estimate the time to train to convergence using wall-clock training times and hardware reported in the original papers, and combine training time with power draw as described under Methods below to estimate total energy consumption and corresponding carbon footprint during training.

Consumer	Renewable energy consumption
China	22%
Germany	40%
United States	17%
Amazon AWS	50%
Google†	100%
Microsoft	50%

Table 1: Percent renewable energy (e.g. hydroelectric, solar, wind) for the top 3 cloud compute providers compared to the United States, China, and Germany. Country percentages taken from Strubell, Ganesh, and McCallum (2019), while corporate numbers have been updated to the latest available information. † indicates that this includes purchases of renewable used to offset non-renewable energy used at locations and times where renewable energy is unavailable.

Models

We analyze four representative models in this case study, which are described in more detail in (Strubell, Ganesh, and McCallum 2019), as well as in the original papers.

Tensor2Tensor (**T2T**) introduced multi-head self-attention for machine translation (Vaswani et al. 2017). The **T2T**$_{base}$ model comprises 65M parameters and the **T2T**$_{big}$ model contains 213M parameters. We also estimate the cost of training T2T using neural architecture search (**NAS**; So, Liang, and Le 2019), a scaled up tuning procedure that consists of training many model architecture variants to find one that performs best on held-out data.

ELMo is a large language model based on stacked bidirectional LSTMs (Peters et al. 2018). Replacing context-independent pre-trained word embeddings with ELMo contextualized word representations has been shown to increase performance on downstream tasks such as named entity recognition, semantic role labeling, and coreference.

BERT is another large language model, based on multi-head self-attention and trained with a different objective (Devlin et al. 2019). BERT contextualized word representations substantially improve accuracy on tasks requiring sentence-level representations such as question answering and natural language inference. **BERT**$_{base}$ has 110M parameters and **BERT**$_{large}$ has 340M parameters. We focus analysis here on BERT$_{base}$ as we were encountered memory limitations with our hardware when trying to train BERT$_{large}$ with the same settings as reported.

GPT-2 is also a large language model using multi-head self-attention, consisting of more parameters and trained for longer on more data than ELMo or BERT (Radford et al. 2019). The large GPT-2 model has 1542M parameters.

Method

We calculate the power consumption in kilowatt-hours (kWh) with the following methodology. Let p_c be the average power draw (watts) from all CPUs during training, let p_r be the average power draw from all DRAM (main memory), let p_g be the average power draw of a GPU during training, and let g be the number of GPUs used to train. We estimate total power consumption as the sum of GPU, CPU and DRAM draw, then multiply this by Power Usage Effectiveness (PUE), which accounts for the additional energy required to support the compute infrastructure (mainly cooling). We use a PUE coefficient of 1.58, the 2018 global average for data centers[1] (Ascierto 2018). It follows that the total power p_t draw at a given instance during training is given by:

$$p_t = \frac{1.58t(p_c + p_r + gp_g)}{1000} \quad (1)$$

The U.S. Environmental Protection Agency (EPA) reports the average CO_2 produced (in pounds per kilowatt-hour) for power consumed in the U.S. (EPA 2018) as: 0.954. Strubell, Ganesh, and McCallum (2019) use this conversion without modification to convert kilowatt-hours to carbon footprint, since that article was based on a 2016 source that cited the renewable energy use by the largest cloud services provider, Amazon Web Services (AWS), as comparable to that of the United States overall. To account for updated reports that AWS sources 50% renewable energy, we cut this number in half. Thus we convert power to estimated CO_2 emissions as follows:

$$\mathrm{CO_2e} = 0.477p_t \quad (2)$$

This conversion makes the assumption that the 50% non-renewable energy used by cloud providers comes from the same relative proportions of different energy sources (natural gas, coal, nuclear) as consumed to produce energy in the United States. As far as we are aware, none of the cloud providers considered in this paper report a detailed breakdown of the sources of energy powering their compute, so we believe this is a reasonable assumption for U.S. workloads. Table 1 lists the relative energy sources for China, Germany and the United States compared to the top three cloud service providers. Note that although Google purchases enough renewable energy to equal its non-renewable use, resulting in effectively 100% renewable energy use in its datacenters, due to technological and geographic limitations, Google still relies on some amount of non-renewable energy to fuel computation, and thus does leave a tangible but not publicly available carbon footprint.

Results

Table 2 lists the estimated cost of training NLP models in terms of kilowatt-hours, carbon emissions, and cloud compute cost. TPUs are more efficient than GPUs for models that are designed for that hardware (e.g. BERT), resulting in lower costs. This finding supports the development of specialized hardware for AI models as one avenue to reduce consumption. We also see that models emit non-trivial carbon emissions. So, Liang, and Le (2019) report that NAS achieves a new state-of-the-art BLEU score of 29.7 for English to German machine translation, an increase of just 0.1 BLEU at the cost of at least $150k in on-demand compute time and potentially substantial carbon emissions.

[1]Many cloud providers report an average PUE below 1.2, but specialized hardware such as GPUs generate up to 66% more heat than standard CPU-based data centers, so we split the difference and use 1.58.

Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e	Cloud compute cost
$T2T_{base}$	P100x8	1415.78	12	27	13	$41–$140
$T2T_{big}$	P100x8	1515.43	84	201	96	$289–$981
ELMo	P100x3	517.66	336	275	131	$433–$1472
$BERT_{base}$	V100x64	12,041.51	79	1507	719	$3751–$12,571
$BERT_{base}$	TPUv2x16	—	96	—	—	$2074–$6912
NAS	P100x8	1515.43	274,120	656,347	313,078	$942,973–$3,201,722
NAS	TPUv2x1	—	32,623	—	—	$44,055–$146,848
GPT-2	TPUv3x32	—	168	—	—	$12,902–$43,008

Table 2: Estimated cost of training a model in terms of CO_2 emissions (lbs) and cloud compute cost (USD) (Strubell, Ganesh, and McCallum 2019). Power and carbon footprint are omitted for TPUs due to lack of public information on power draw.

Case study 2: Hyperparameter tuning

A substantial but often under-reported aspect of the computation required for training is due to hyperparameter tuning. To quantify the computational requirements of developing a new model, in this case study we analyze the logs of all training required to develop Linguistically-Informed Self-Attention (Strubell et al. 2018), a multi-task model that performs four related natural language tasks. This model makes for an interesting case study as the four tasks represent a typical NLP pipeline, and the paper was awarded Best Long Paper at EMNLP 2018.

Results

The project required a total of 9998 days (27 years) of GPU time, or about 60 GPUs running throughout the duration of the 6 month project. Table 3 lists upper and lower bounds of the estimated cost in terms of Google Cloud compute and raw electricity required to develop and deploy this model.[2] Though training a single model is relatively inexpensive, the cost of tuning a model for a new dataset, which we conservatively estimate here as 24 jobs, or performing the full research and development cycle to develop this model, quickly becomes prohibitively expensive.

		Estimated cost (USD)	
Models	Hours	Cloud	Electric
1	120	$52–$175	$5
24	2880	$1238–$4205	$118
4789	239,942	$103k–$350k	$9870

Table 3: Estimated cost of training: (1) a single model (2) a single tune and (3) all models trained during R&D (Strubell, Ganesh, and McCallum 2019).

Conclusions

We conclude by providing actionable recommendations to the community based on our analysis. See Strubell, Ganesh, and McCallum (2019) for a more detailed discussion of the first three conclusions summarized below.

[2]Based on average U.S cost of electricity of $0.12/kWh.

Authors should report training time and sensitivity to hyperparameters.

Our experiments suggest that it would be beneficial to directly compare models not just in terms of accuracy on benchmark data, but also in terms of efficiency using a standard metric. See Schwartz et al. (2019) for more discussion of standard metrics for reporting efficiency, and Dodge et al. (2019) for further analysis and concrete methods for reporting tuning and hyperparameter sensitivity.

Academic researchers need equitable access to computation resources.

Recent advances in available compute come at a high price not attainable to all who desire access. Limiting this style of research to the wealthiest labs hurts the AI research community by stifling creativity and prohibiting certain types of research on the basis of access to financial resources. The prohibitive start-up cost of building in-house infrastructure forces resource-poor groups to rely on cloud compute services, though in-house compute is less expensive in the long term. All of the above serves to further entrench the already problematic "rich get richer" cycle of research funding.

Researchers should prioritize computationally efficient hardware and algorithms.

We recommend a concerted effort by industry and academia to promote research and development of more computationally efficient algorithms, as well as hardware that requires less energy. Making efficient algorithms readily available in popular software should also be a priority. Figure 2 depicts the number of papers focusing on accuracy or efficiency at four top AI conferences, labeled from a random sample of 20 papers from each conference. There is a clear bias towards research focused on obtaining higher accuracy. AAAI also follows this trend: a quick search for "efficien" in AAAI 2019 accepted technical track paper titles yields 40 out of 1149 total papers, or about 3.5%. AAAI 2019 also held a computational sustainability track last year, comprising 0.4% of technical track papers.

AI researchers and practitioners should be mindful of energy sources powering their compute.

Do you know whether your flops are fuelled by coal or hydroelectric power? As we see in Table 1, not all cloud ser-

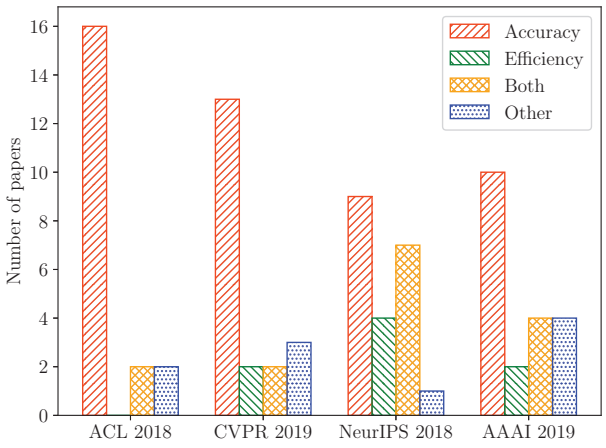

Figure 2: Distribution of papers targeting accuracy, efficiency, both, or neither labeled from a sample of 80 papers at four top AI conferences. Current trends focus on accuracy over efficiency. Figure based on (Schwartz et al. 2019).

vices provide equally sustainable compute, and the exact breakdown of energy source and thus carbon footprint varies widely based on geographic location. The same is true of in-house resources; with today's renewable resources and grid technology, it is simply not possible for all regions to source renewable energy all of the time. See Google's recent whitepaper (Google 2018) for a deeper discussion of some of the challenges in attaining 100% renewable energy in datacenters across the globe, and (Kim and Pierce 2018) for further reading on the nuances of purchasing carbon offsets. Lacoste et al. (2019) recently published an online calculator[3] that provides geographically-aware estimates of effective carbon emissions for users of Google, Amazon and Microsoft cloud resources. We strongly encourage ML researchers to analyze, audit and report the carbon footprint of their research using this valuable tool.

Acknowledgments

We are grateful to Roy Schwartz and Jesse Dodge for sharing the raw data for their figures, and to John Platt for helpful feedback on the original paper.

References

Agostinelli, F.; McAleer, S.; Shmakov, A.; and Baldi, P. 2019. Solving the rubik's cube with deep reinforcement learning and search. *Nature Machine Intelligence* 1(8):356–363.

Aharoni, R.; Johnson, M.; and Firat, O. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3874–3884. Minneapolis, Minnesota: Association for Computational Linguistics.

Amodei, D., and Hernandez, D. 2018. AI and Compute.

Ascierto, R. 2018. Uptime Institute Global Data Center Survey. Technical report, Uptime Institute.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large scale GAN training for high fidelity natural image synthesis.

Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 1251–1258.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

Dodge, J.; Gururangan, S.; Card, D.; Schwartz, R.; and Smith, N. A. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of EMNLP*.

EPA. 2018. Emissions & Generation Resource Integrated Database (eGRID). Technical report, U.S. Environmental Protection Agency.

Google. 2018. Achieving our 100% renewable energy purchasing goal and going beyond.

Kim, R., and Pierce, B. C. 2018. Carbon Offsets: An Overview for Scientific Societies.

Lacoste, A.; Luccioni, A.; Schmidt, V.; and Dandres, T. 2019. Quantifying the carbon emissions of machine learning.

OpenAI. 2018. OpenAI Five.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *NAACL*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners.

Schwartz, R.; Dodge, J.; Smith, N. A.; and Etzioni, O. 2019. Green AI. *arXiv preprint arXiv:1907.10597*.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. 2017. Mastering the game of go without human knowledge. *Nature* 550(7676):354.

So, D. R.; Liang, C.; and Le, Q. V. 2019. The evolved transformer. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Strubell, E.; Verga, P.; Andor, D.; Weiss, D.; and McCallum, A. 2018. Linguistically-Informed Self-Attention for Semantic Role Labeling. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Strubell, E.; Ganesh, A.; and McCallum, A. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650. Florence, Italy: Association for Computational Linguistics.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *31st Conference on Neural Information Processing Systems (NIPS)*.

[3] https://mlco2.github.io/impact/