The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink

David Patterson, Google and University of California, Berkeley

Joseph Gonzalez, University of California, Berkeley

Urs Hölzle, Quoc Le, Chen Liang, and Lluis-Miquel Munguia, Google

Daniel Rothchild, University of California, Berkeley

David R. So, Maud Texier, and Jeff Dean, Google

Machine learning (ML) workloads have rapidly grown, raising concerns about their carbon footprint. We show four best practices to reduce ML training energy and carbon dioxide emissions. If the whole ML field adopts best practices, we predict that by 2030, total carbon emissions from training will decline.

Over the past few years, a growing number of papers have highlighted the carbon emissions of machine learning (ML) workloads. While this work has been instrumental in rightfully elevating the discussion around carbon emissions in ML, some studies overestimated actual emissions. For example, Thompson et al.[1] extrapolated future ML energy use from what turned out to be faulty estimates in another paper[2] and concluded:

> "The answers are grim: Training such a model would cost US$100 billion and would produce as

Digital Object Identifier 10.1109/MC.2022.3148714
Date of current version: 4 July 2022

much carbon emissions as New York City does in a month."

Recent work highlights the complexities and nuances associated with carbon accounting for ML and, more broadly, computing workloads.[3–5] In this article, we contribute the following:

- We describe four practices that reduce the energy use and carbon emissions of ML workloads by orders of magnitude relative to traditional choices.
- We show that these practices have helped to keep ML under 15% of Google's total energy use for the past three years.
- We explain why published estimates were 100–100,000× higher than real carbon footprints.

Responsible artificial intelligence (AI) is a broad topic; we focus on a single issue that has received much attention from the ML community and the public: carbon emissions from ML training. Emissions can be classified as follows:

- *Operational*: the energy cost of operating ML hardware, including data center overheads
- *Lifecycle*: the embedded carbon emitted during the manufacturing of all components, from chips to data center buildings.

Like prior work, we focus on operational emissions; estimating lifecycle emissions is a larger, future study.

We identified best practices that can reduce energy use by up to 100× and carbon emissions by up to 1,000× when compared to four orthodox choices: model, machine, mechanization, and map (4Ms), as follows:

1. *Model:* Selecting efficient ML model architectures while advancing ML quality, such as sparse models versus dense modes, can reduce computation by factors of ~5–10.
2. *Machine*: Using processors optimized for ML training, such as tensor processing units (TPUs) and recent GPUs (for example, the V100 and A100), versus general-purpose processors can improve performance/watt by factors of 2–5.
3. *Mechanization:* Computing in the cloud rather than on premise improves data center energy efficiency, reducing energy costs by a factor of 1.4–2. (The cloud uses custom warehouses designed for energy efficiency, whereas on-premise data centers are inefficiently located in smaller, older spaces intended for other purposes.)
4. *Map*: Moreover, cloud computing enables ML practitioners to pick the location with the cleanest energy, further reducing the gross carbon footprint by factors of 5–10. (Most data transmission power is for the network equipment of the Internet even when idle[3]; in comparison, shipping photons over fiber optics is relatively trivial. Using carbon-neutral clouds, such as Facebook and Google, further reduces the footprint to zero because the services match 100% of their operational energy use with renewable energy; we exclude those offsets. High-performance computing data centers are efficient but cannot enable shifting to green locations.)

SELECTING EFFICIENT ML MODEL ARCHITECTURES WHILE ADVANCING ML QUALITY, SUCH AS SPARSE MODELS VERSUS DENSE MODES, CAN REDUCE COMPUTATION BY FACTORS OF ~5–10.

Figure 1 illustrates how four good choices together reduce energy consumption by 83× and carbon dioxide (CO_2) emissions by 747× over four years while maintaining the same quality. The original modeled estimate represents training the Transformer model in 2017 on an ML-oblivious GPU (the 2016 NVIDIA P100 was optimized for graphics, not ML) in a typical data center using an average energy mix (such as in Strubell et al.[2]). The yellow line shows optimizations possible in 2019; the green-line optimizations are possible today. In both cases, optimized ML hardware reduces energy consumption significantly, with newest-generation hardware (TPU v4) providing an additional 2.4× over the 2019 hardware (TPU v2).

Using efficient cloud data centers and a low-carbon data center region per Google's 24/7 carbon-free energy (CFE) methodology further reduces the carbon footprint by another order of magnitude (note the log scale *y*-axis), resulting in a 747-fold reduction in the carbon footprint compared to the original estimate. In this article, gross CO_2 emissions are the carbon emissions resulting from a workload in a particular location before any compensating actions.

Supported by the results in Figure 1 and in the "Case Study 1: Transformer Versus Evolved Transformer Versus Primer" section, we predict that if ML communities embrace these 4M best practices, the carbon footprint of ML training will shrink during this decade, as summarized in the following:

- Two studies show the impact of best practices: a 750× emissions reduction without a loss of accuracy from the Transformer (Figure 1) and a 14× emissions reduction from Generative Pretrained Transformer 3 (GPT-3) by the larger Generalist Language Model (GLaM), which improves accuracy.
- Location choices, even within one country, can significantly impact the carbon footprint.
- We provide the first report by a hyperscaler company of the percentage of its overall energy use devoted to ML training and inference.
- We show that the carbon footprint of searching for better ML models can reduce the impact of downstream ML tasks by much more than the cost of the search.
- We describe how following the 4M best practices reduced the energy consumption and carbon footprint of training significantly compared to the faulty estimates commonly cited.[2,6,7]

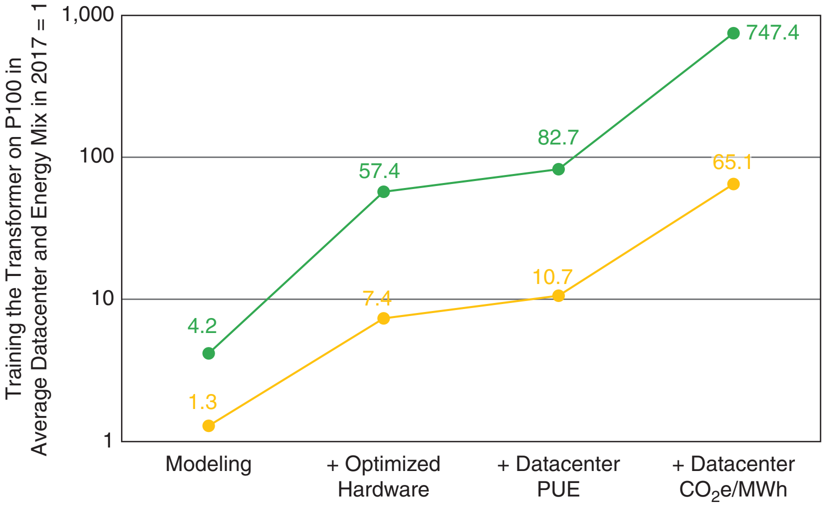

FIGURE 1. The reduction in gross carbon dioxide (CO_2) emissions since 2017 by applying the 4M best practices (see the "Case Study 1: Transformer Versus Evolved Transformer Versus Primer" section). It shows large end-to-end improvements, broken down into four factors. The gross CO_2 emissions here exclude Google's carbon-neutral and 100% renewable energy credits and reflect its 24/7 carbon-free energy methodology.[5] The yellow line is for the Evolved Transformer[7] on TPU v2s in 2019, and the green line is for the Primer[8] on TPU v4s in 2021; both types run in Google data centers. PUE: power usage effectiveness; CO_2e: CO_2-equivalent emissions.

OVERVIEW OF ENERGY AND CO_2-EQUIVALENT EMISSIONS FOR ML TRAINING

We estimate energy and carbon footprints using the following terms:

- CO_2-equivalent emissions (CO_2e) account for CO_2 and all the other greenhouse gasses: methane, nitrous oxide, and so on.
- Metric tons are the common CO_2e unit of measure, abbreviated tCO_2e, representing 1,000 kg (2,205 lb).
- Megawatt hours measure energy; 1 MWh equals 1 million W of electricity used continuously for 1 h. One terawatt hour equals 1 million MWh.
- Power usage effectiveness (PUE) is the industry standard metric of data center efficiency, defined as the ratio between total energy use (including all overheads, such as cooling) divided by the energy directly consumed by a data center's computing equipment. The average industry data center PUE in 2020 was 1.58 (58% overhead), while cloud providers had PUEs of ~1.1.[5]
- Carbon intensity (metric tons per megawatt hour) is a measure of the cleanliness of a data center's energy. The average data center carbon emissions in 2020 was

0.429 tCO_2e/MWh, but the gross CO_2e per megawatt hour can be 5× lower in some Google data centers.

The energy consumption of the servers performing a training task is proportional to the number of processors used and the duration of the training run:

$$\text{MWh} = \text{hours to train} \times \text{number of processors} \times \text{average power per processor}.$$

We include all server components in "processors" (including local memory, network links, and so on). Additionally, a data center consumes energy to power and cool hardware (for example, voltage transformation losses and cooling equipment), which is captured by the PUE. Thus, the final formula for energy consumption:

$$\text{MWh} = (\text{hours to train} \times \text{number of processors} \times \text{average power per processor}) \times \text{PUE}.$$

We can then turn energy into carbon by multiplying it with the carbon intensity of the energy supply:

$$tCO_2e = \text{MWh} \times tCO_2e \text{ per MWh}.$$

The real-world values for many factors are readily available. ML practitioners usually publish the number and type of processors and hours to train, and the power consumption of most hardware components is well known and can be accurately measured. Many cloud companies publish the PUE of their data centers. In comparison, carbon intensity is harder to obtain. For this article, we use the carbon intensity of Google data centers, derived from Figure 2. We hope other providers will publish metrics so that carbon intensity can be compared across data centers.

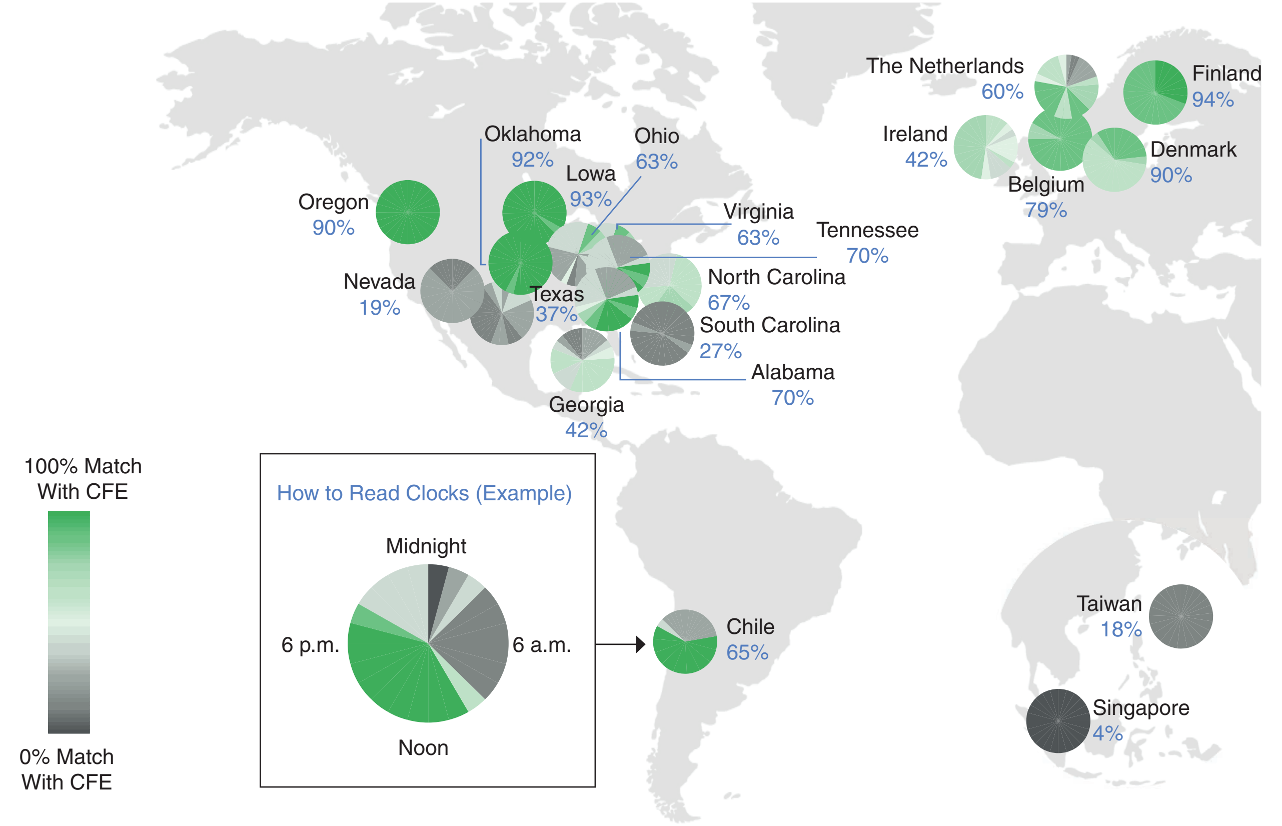

FIGURE 2. The percentage of CFE by Google Cloud location in 2020. The map shows the percentage and how it changes by time of day. Chile has a high CFE percentage from 6 a.m. to 8 p.m. but not at night. The U.S. examples range from 19% CFE in Nevada to 93% in Iowa, which has strong prevailing winds during the night and day. (Source: https://sustainability.google/progress/energy/.)

CASE STUDY 1: TRANSFORMER VERSUS EVOLVED TRANSFORMER VERSUS PRIMER

Many of the headline-grabbing advances in AI stem from deep neural networks (DNNs); indeed, three DNN leaders shared the 2018 Association for Computing Machinery A.M. Turing Award. DNN computations have two phases: training, which constructs accurate models through an intensive computational process involving the iterative updating of parameters, and inference, which uses trained models to generate outputs from model inputs. ML practitioners use different models for various tasks: object recognition, language translation, and so on. Training "learns" parameters that raise the likelihood of correctly mapping from input to result. Unlike in traditional computing, the actual DNN code is relatively small. The "smarts" come from training DNNs from millions of labeled examples versus writing millions of lines of code.

> CLOUD COMPUTING ENABLES ML PRACTITIONERS TO PICK THE LOCATION WITH THE CLEANEST ENERGY, FURTHER REDUCING THE GROSS CARBON FOOTPRINT BY FACTORS OF 5–10.

The Transformer model debuted in 2017 and is used primarily for natural language processing (NLP). Its distinguishing feature is focusing attention on portions of its input. Two years later, So et al. used neural architecture search (NAS) to discover the Evolved Transformer model, which matched the Transformer's quality scores but was ~1.3× faster.[8] In 2021, a different NAS found the Primer model, which again matched the quality scores but was 4.2× faster than the original Transformer.[9]

Figure 1 plots the end-to-end reduction in CO_2e by applying the best practices from the beginning of the article. The reference point is the Transformer model trained on a P100 GPU in an average on-premise data center with the average PUE of 1.6 in 2017 and using the average 0.488 tCO_2e/MWh. The practices (the 4Ms) are given in the following:

1. *Model*: In 2019, the best model was the Evolved Transformer; in 2021, it was the Primer.
2. *Machine*: Compared to the unoptimized P100s from 2017, the ML-optimized TPU v2 in 2019 and TPU v4 in 2021 reduced energy consumption by 5.7 and 13.7×, respectively. This reduction was a function of improved logic (more specialized hardware), newer chip fabrication technology, and more efficient mapping of the training task to hardware (better utilization of the functional units).[10]
3. *Mechanization*: The third point shows a reduction of 1.4× from the better PUE of Google's cloud data center versus the average data center.
4. *Map*: A big surprise was how much the location of a data center affected carbon intensity (Figure 2). In 2019, the data center in the U.S. region with the highest CFE score was in Oklahoma, with a score of 96%, and in 2020, it was in Iowa, at 93%.

To summarize, following the 4M best practices yielded a 65× reduction in CO_2e two years after the Transformer was introduced. Two years after that—with ML model, hardware, and energy mix improvements—another 11× was possible, for an overall reduction of 747×. These drastic improvements, as well as their trajectory through time, suggest that extrapolating current parameters to predict future CO_2e is fraught with peril.

CASE STUDY 2: GPT-3 VERSUS GLAM

Next is a large NLP model that received considerable attention in the ML community and the press in 2020: GPT-3 is an autoregressive language model with 175 billion parameters, 10× more than any nonsparse language model at the time, and 100–1,000× more than most other ML models.[11] To put GPT-3 into perspective, its predecessor, GPT-2, had 1.5 billion parameters, and the Transformer models used ≤0.2 billion. Developed by OpenAI, GPT-3 was trained on 10,000 V100 GPUs in a Microsoft cloud data center (the 2017 NVIDIA V100 is optimized for ML). A winner of the best paper award at the Conference and Workshop on Neural Information Processing Systems (NeurIPS), a recent GPT-3 paper already has >3,500 citations and made mainstream media headlines. One benefit of large models such as GPT-3 is that they do not need to be retrained for every new task—called *few-shot generalization*—unlike smaller models.

GLaM is a new language model using 7× more parameters than GPT-3. It is a mixture-of-experts model that selectively activates experts based on the input so that no more than 95 billion parameters (8%) are active per input token. The dense GPT-3 activates all 175 billion parameters on every token. More parameters and sparsity enable GLaM to exceed GPT-3 on quality and efficiency.[12] Figure 3 compares them. GPT-3 took 405 V100 years to train in 2020. OpenAI trained in the Microsoft cloud to leverage a low PUE but with an energy mix that matched the U.S. data center average.[5] In comparison, GLaM trained on TPU v4s in 2.8× fewer accelerator years, using 2.8× less energy than GPT-3. Additionally, GLaM ran in the Oklahoma data center, where the CO_2e per megawatt hour were ~5× lower (0.088 versus 0.429). The Evolved Transformer and Primer improve energy use and CO_2e while maintaining quality scores, but GLaM betters all three metrics.

ML researchers are continuously improving the efficiency of large language models through innovations in algorithms and model architectures. Only 18 months after GPT-3, GLaM can reduce the gross carbon footprint by ~14× despite raising accuracy. These drastic improvements again show that extrapolating current ML trends to predict future ML energy use and CO_2e can greatly overestimate consumption, as there are continuous, significant improvements in algorithms and hardware.

OVERALL ML ENERGY CONSUMPTION

The preceding sections investigated the energy consumption of a single training task. Here, we discuss the overall footprint of all ML workloads at a major user, Google. Many hyperscalers regularly publish their energy consumption metrics. According to their sustainability reports, the annual energy consumption in 2020 was 15.4 TWh for Google and 10.8 TWh for Microsoft. These reports put the training energy of large models into perspective. Training GPT-3 was ~0.012% of Microsoft's energy consumption in 2020, and GlaM was ~0.004% of Google's. For further comparison, the portion of the 22,000 people from 68 countries who in 2019 flew to attend the two major ML conferences (NeurIPS and the Conference on Computer Vision and Pattern Recognition) collectively had a CO_2e impact that was likely ~10–100× higher than that of training all the ML models in this article.[5]

While Google's overall energy consumption increases as usage rises, our data show that despite the growth of ML applications, the ML portion of the company's overall energy consumption is not expanding. To estimate that fraction, we measured the energy consumption (including data center overheads) of the following components:

- *All TPUs and GPUs in Google data centers, including associated dedicated servers and networking equipment*: Virtually all ML training executes on TPUs and GPUs, and most inference, as well. We can differentiate training versus inference runs on TPUs and GPUs.
- *Any CPU consumption attributable to ML inference*: No significant training was done solely on CPUs.

To estimate the CPU portion of inference, we inspected Google-Wide Profiling results to measure the CPU

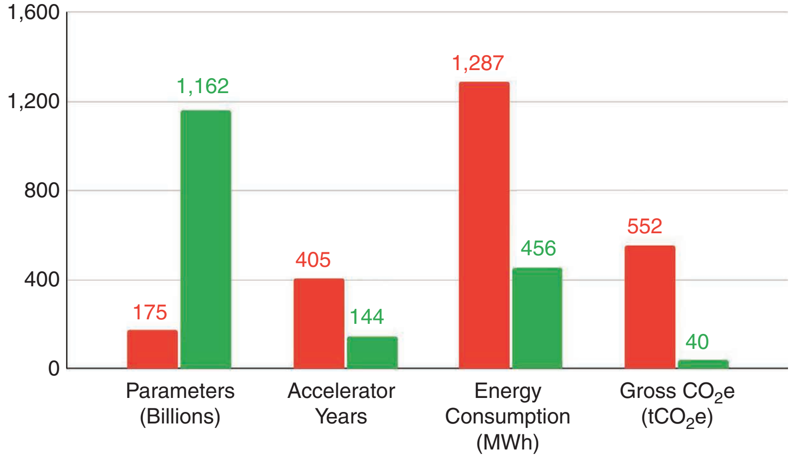

FIGURE 3. The parameters, accelerator years of computation, energy consumption, and gross CO_2e for GPT-3 (V100 in 2020, in red) and GLaM (TPU v4 in 2021, in green). If instead of outperforming GPT-3 on quality scores, GLaM were trained only to match, it would halve the time, energy, and CO_2e. Google's renewable energy purchases further reduce the impact to zero.

consumption of the libraries used for ML inference. We then converted CPU utilization into energy consumption by using sensors that measure server power. Our numbers likely overestimate because some libraries are used in non-ML cases, as well. Also, we may double count some host CPUs that are already accounted for in the TPU/GPU measurements, and some GPU use is for graphics.

We retroactively performed these calculations based on data for one week of April in 2019, 2020, and 2021. Each time, the ML portion was 10–15% of Google's total energy consumption, despite ML representing 70–80% of the FLOPS at Google. While ML use certainly increased during those years, algorithmic and hardware improvements kept that expansion to a rate comparable with overall energy growth at the company. Across all three years, about three-fifths of the ML energy use was for inference, and two-fifths were for training. These measurements include all ML energy consumption: research, development, testing, and production. Consequently, we take the stable fraction for ML as a strong indication that despite ML's increasing popularity, when following the 4M best practices, its energy consumption is not skyrocketing, contrary to commonly expressed fears. This stability may reflect economic factors in addition to technical ones: after all, everything has a budget, and budget limits encourage the efficient use of ML resources.

GLOBAL DATA CENTER ENERGY CONSUMPTION INCREASED BY ONLY 6% FROM 2010 TO 2018 DESPITE DATA CENTER COMPUTING CAPACITY GROWING 550% THROUGH THE SAME PERIOD.

Worldwide data center energy consumption is not growing quickly either. Masanet et al.[4] observe that global data center energy consumption increased by only 6% from 2010 to 2018 despite data center computing capacity growing 550% through the same period and contrary to 2010 predictions of a 70% increase by 2018. One key factor was the shift from conventional data centers to cloud data centers. Not only are cloud data centers often more efficient, cloud servers typically have significantly higher utilization than on-premise ones. That enables the same workloads to be served with less hardware and thus less energy, just as books purchased for libraries are more frequently read than those bought for home use. As of 2021, only 15–20% of all workloads have moved to the cloud,[13] so there is plenty of headroom for cloud growth to replace inefficient on-premise data centers.

Finally, most cloud companies compensate at least partially for their carbon emissions. In particular, Google (since 2017) and Facebook (since 2020) annually purchase enough renewable energy to match 100% of their use, so each megawatt hour of new consumption is offset by one 1 MWh of new renewable energy, albeit not necessarily in the same location. Microsoft's similar goal is for 2025. Thus, the net carbon impact of ML computations for some companies could be considered zero. Such multibillion-dollar direct energy purchases by hyperscalers have substantially spurred the growth of renewable energy: in some countries, the companies are more significant investors in renewable energy than government subsidies.[14]

ADDITIONAL FACTORS

For completeness, we will briefly address two other concerns about ML energy use: the impact of NAS, which may execute thousands of training runs as part of a single search—potentially exploding overall energy consumption—and ML's impact on client-side energy use. A commonly expressed concern is that automated methods might increase training energy consumption. As the name implies, NAS employs computers to find models, with higher quality and efficiency than human experts can achieve. NAS is generally not performed once per model training but once per problem domain + architectural search space combination. The Evolved Transformer and Primer are examples of the benefits of NAS.[8,9] It has also been applied to find models that have better quality and run faster by adapting them to a given processor.[15]

The NAS producing the Evolved Transformer used 7.5 MWh. The use of the Evolved Transformer while training the large Meena model saved 15× the energy cost of the NAS.[5] Finding the even faster Primer used only 6.2 MWh. Overall, NAS is a net environmental gain if a discovered model is trained more than a few times. Often, the more efficient models found by NAS are open sourced and reused hundreds and even thousands of times.[5] Consequently, as a whole, it is likely that NAS reduces total ML energy consumption by producing more efficient models whose downstream use more than compensates for the initial search effort.

To estimate ML energy use on client devices, Patterson et al.[16] studied mobile phones. Most modern phones have ML accelerators; for example, the Google Pixel 6 has an edge TPU, which runs most of the ML workload. Their upper bound for ML energy use on today's mobile phones is 1.5%.

The estimated global energy use of the 6.6 billion mobile phones in 2022 was 40 to 58 TWh, assuming nightly charging and accounting for charger inefficiency.[16] The upper bound for ML on mobile phones is, then, 0.6 to 0.9 TWh. Google's ML server energy use in 2020 was ~3 to 4× higher than this conservative estimate of ML on all mobile phones. This calculation does not include the energy consumption of ML at other cloud companies, so server-side ML energy use clearly dominates client-side use.

RELATED WORK

Henderson et al.[17] conducted a similar study that provides a framework to understand the potential climate impacts of ML research. They also offered a leaderboard to foster competitions on reducing the CO_2e of ML and a tool to collect energy use and CO_2e from the preliminary training runs. Patterson et al,[5] the authors of this article, produced a 22-page technical report that goes into greater detail on many of the issues discussed here.

Schwartz et al.[18] warn of the danger of "Red AI," which focuses on model quality gains regardless of the training cost and CO_2e. They encourage embracing "Green AI," where the emphasis is on computing efficiency as well as model quality. Arguing that it can be difficult to measure energy and CO_2e, they recommend minimizing the number of floating-point operations (FLOPs) to train a model. Alas, FLOPs are not a good metric, for time and energy can be uncorrelated with them. For example, automated ML found faster models that used 2.4× as many FLOPs.[15] An underlying reason is that main memory accesses are much slower and consume significantly more energy than FLOPs today. A dynamic random-access memory access is ~6,000× the energy of 16-bit FLOPS (1,300 versus 0.21 pJ).[10] Another reason is that scaling up the FLOPS per second is much easier for ML accelerators than scaling up memory bandwidth. To improve efficiency further, ML practitioners should focus more on reducing memory accesses than FLOPs. More successful attempts to simplify the calculation of energy are online calculators, such as the ML Emissions Calculator.[5,19]

The opening quote in this article is based on a 2019 project from the University of Massachusetts Amherst that estimated the environmental impact of training.[2] More than 1,250 papers cite Strubell et al.[2] as the source for the impact ML models have on carbon emissions, including Bender et al.,[20] Freitag et al.,[7] Schwartz et al.[18] Thompson et al,[1] and Thompson et al.[6] The study calculated the energy consumed and the carbon footprint of the NAS by So et al.[8] that led to the Evolved Transformer. Their estimate (they did not run the NAS code) was 284 tCO_2e for NAS; the actual number was only 3.2 tCO_2e, a factor of 88 smaller. The reasons for the overshoot include the following:

GOOGLE (SINCE 2017) AND FACEBOOK (SINCE 2020) ANNUALLY PURCHASE ENOUGH RENEWABLE ENERGY TO MATCH 100% OF THEIR USE.

1. Since the authors of the original NAS paper didn't include energy and emissions for Google systems, their estimate was based on older GPUs not optimized for ML instead of TPU v2 and on the average data center PUE and U.S. average carbon intensity instead of the real numbers for a Google data center (they used the P100; the most recent GPU available was the V100, which was much faster, in part because it was optimized for ML, unlike the P100). This difference explains 5×.
2. There was also confusion about the computational cost of NAS. Described subtly in So et al.,[8] the Evolved Transformer NAS used a small proxy task to search for the best models to save time,

money, and energy and then scaled up the found models to full size. However, Strubell et al.[2] assumed the search was done with full-size tasks. The resulting NAS computation estimate was another 18.7× too high.

The actual overshoot was 18.7× for computation and 5× for Google versus the average data center, so the real emissions for the one-time search were 88× less (3.2 versus 284 tCO2e). The faulty estimates are understandable given the lack of access to internal information. It is likewise understandable that those estimates were propagated in other papers. Unfortunately, many papers confuse the one-time cost of the NAS of So et al.[16] with the relatively tiny "every time" cost that is incurred from training (the NASs for the Evolved Transformer and Primer produce 1,347 and 1,618× more CO_2e, respectively, than their training).

This confusion led them to believe Evolved Transformer used more than 2 million GPU hours to train, cost millions of dollars, and its emissions were five times the lifetime of a car (284,019 kg).[1,6]

In reality, the training cost of the medium Evolved Transformer, which achieves the same accuracy level as the Transformer-big model,

- 120 TPU v2 hours, not 2 million GPU hours, which is >15,000× less
- US$40 to train on Google Cloud (four TPU v2s cost US$1.35/h), not millions of dollars, which is >50,000× less
- Fewer than 2.4 kg of CO_2e, or 0.00004 car lifetime emissions, not 284,019 kg and five car lifetimes, translating to 120,000× less.

The gap is nearly as large as confusing the CO_2e from manufacturing a car with the CO_2e from driving a car and then overestimating the production cost by ~100×. The gap between these quotes and actual measurements illustrates the importance of authors calculating and publishing energy consumption and carbon footprints, as accuracy is difficult if estimated retrospectively.

ML workloads have rapidly grown in importance, raising legitimate concerns about their energy use. Fortunately, the real-world energy use trend of ML is fairly boring. While overall energy use at Google grows annually with greater consumption, the percentage for ML held steady for the past three years, representing <15% of total energy use. Inference represents about three-fifths of the total ML energy use at Google, owing to the many-billion-user services that incorporate ML. GLaM, the largest natural language model trained in 2021, improved model quality yet produced 14× less CO_2e than training the previous state-of-the art model from 2020 (GPT-3) and accounted for only 0.004% of Google's annual energy.

Furthermore, we illustrated that in large-scale production ML deployments, minimizing emissions from training is not the ultimate goal. Instead, the combined emissions of training and serving need to be minimized. Approaches such as NAS increase emissions but lead to more efficient serving and a strong overall reduction of the ML carbon footprint. Another perspective is that some consider the carbon footprint to be erased entirely if a cloud provider matches 100% of its energy consumption with renewable energy, as Google and Facebook have done and as Microsoft will soon do.

While ML workloads exploded over the past decade, and while the number of computations per training run has similarly increased by orders of magnitude, our data show that technology improvements have largely compensated for this greater load. We believe that this consistent overall low percentage is testimony to the benefits of the following the 4M best practices:

- Data center providers should publish the PUE, CFE percentage, and CO_2e per megawatt hour per location so that customers who care can understand and reduce their energy consumption and carbon footprint.
- ML practitioners should train using the most effective processors in the greenest data centers they have access to, which today is often in the cloud.
- ML researchers should continue to develop more efficient ML models,[8,9] such as by leveraging sparsity[12] and integrating retrieval into smaller models. They should also publish their energy consumption and carbon footprint to foster competition on more than just model quality and ensure accurate accounting of their work, which is difficult to do with precision post hoc.

These numbers may vary across companies, but the 4M practices we have identified are applicable to virtually every ML training workload and open to all to use. As a result, we predict that if all ML communities embrace these 4M best practices, we can create a virtuous circle that will bend the curve so that in this decade we will see the total carbon footprint of ML training at first plateau and then shrink. Finally, we showed that

ABOUT THE AUTHORS

DAVID PATTERSON is a distinguished engineer in the Google Brain project, Mountain View, California, 94043, USA; the vice-chair of the RISC-V Foundation's board of directors; the director of the RISC-V International Open Source Laboratory; and a professor emeritus at the University of California, Berkeley. His research interests include domain-specific computer architectures and open instruction set architectures. Patterson received a Ph.D. in computer science from the University of California, Los Angeles. He is a Life Fellow of IEEE. Contact him at pattrsn@berkeley.edu.

JOSEPH GONZALEZ is a professor of computer science at the University of California, Berkeley, Berkeley, California, 94720, USA. His research interests include the design of systems for machine learning as well as efficient neural architectures. Gonzalez received a Ph.D. in machine learning from Carnegie Mellon University. Contact him at jegonzal@berkeley.edu.

URS HÖLZLE is the senior vice president of operations at, and a fellow of, Google, Mountain View, California, 94043, USA. His research interests include large-scale clusters, cluster networking, Internet performance, and data center design. Hölzle received a Ph.D. in computer science from Stanford University. He is a Member of IEEE. Contact him at urs@google.com.

QUOC LE is a principal scientist in the Google Brain project, Mountain View, California, 94043, USA. His research interests include artificial intelligence, automated machine learning, natural language understanding, and computer vision. Le received a Ph.D. in computer science from Stanford University. Contact him at qvl@google.com.

CHEN LIANG is a researcher in the Google Brain project, Mountain View, California, 94043, USA. His research interests include automated machine learning, neural symbolic methods, natural language understanding, and program synthesis. Liang received a Ph.D. in computer science from Northwestern University. Contact him at crazydonkey200@gmail.com.

LLUIS-MIQUEL MUNGUIA is a senior software engineer at Google, Mountain View, California, 94043, USA, where he works on codesign for deep learning accelerators. His research interests include the performance analysis of special-purpose computer architectures, power efficiency, and high-performance computing. Munguia received a Ph.D. in computational science and engineering from Georgia Institute of Technology. Contact him at llmunguia@google.com.

DANIEL ROTHCHILD is a Ph.D. student at the University of California, Berkeley, Berkeley, California, 94720, USA, advised by Joseph Gonzalez. His research interests include distributed and federated learning and machine learning for drug discovery and materials design. Rothchild received an MPhil in astronomy from the University of Cambridge. Contact him at drothchild@berkeley.edu.

DAVID R. SO is a staff research engineer in the Google Brain project, Mountain View, California, 94043, USA. His research interests include language modeling, automated machine learning, and improving deep learning efficiency. So received a B.S. in computer science from Columbia University. Contact him at davidso@google.com.

MAUD TEXIER is the head of energy development at Google, Mountain View, California, 94043, USA. Her research interests include carbon abatement technologies, carbon-free energy technologies, and grid system modernization and decarbonization. Texier received an M.S. in engineering in energy and power systems from École Centrale Paris. Contact her at maudt@google.com.

JEFF DEAN is a senior fellow of, and the senior vice president of research at Google, Mountain View, California, 94043, USA, where he cofounded the Google Brain project. His research interests include large-scale distributed systems, machine learning, applications of machine learning, information retrieval, microprocessor architecture, and compiler optimizations. Dean received a Ph.D. in computer science from the University of Washington. Contact him at jeff@google.com.

published studies overestimated the cost and carbon footprint of ML training by 100–100,000× because they didn't have access to the right information or because they extrapolated point-in-time data without accounting for algorithmic and hardware improvements.

Climate change is important, so we must get the numbers right to ensure that we work on the biggest challenges. Many efforts are underway to reduce the operational energy and CO_2e of ML training, as illustrated by the 4Ms. Thus, within information technology, we believe the biggest climate change challenge is not the operational cost of ML but more likely the lifecycle cost of manufacturing computing equipment of all types and sizes: IT manufacturing for 2021 included 1.54 billion smartphones, 0.34 billion PCs, and 0.01 billion data center servers.

ACKNOWLEDGMENTS

We had a great deal of help from others along the way for an earlier study [16] that eventually led to this version of the article. Emma Strubell made several suggestions for the prior article, including the recommendation to examine giant NLP models. Christopher Berner, Ilya Sutskever, OpenAI, and Microsoft shared information about GPT-3. Dmitry Lepikhin and Zongwei Zhou did a great deal of work to measure the performance and power of GPUs and TPUs in Google data centers. Hallie Cramer, Anna Escuer, Elke Michlmayr, Kelli Wright, and Nick Zakrasek helped with the data and policies for energy and CO_2e at Google. Talia Ringer provided helpful suggestions on how to better present related work.

REFERENCES

1. N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, "Deep learning's diminishing returns: The cost of improvement is becoming unsustainable," *IEEE Spectr.*, vol. 58, no. 10, pp. 50–55, 2021, doi: 10.1109/MSPEC.2021.9563954.
2. E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 13,693–13,696.
3. J. Koomey and E. Masanet, "Does not compute: Avoiding pitfalls assessing the Internet's energy and carbon impacts," *Joule*, vol. 5, no. 7, pp. 1625–1628, 2021, doi: 10.1016/j.joule.2021.05.007.
4. E. Masanet, A. Shehabinuoa, L. Smith, and J. Koomey, "Recalibrating global datacenter energy-use estimates," *Science*, vol. 367, no. 6481, 2020, doi: 10.1126/science.aba3758.
5. D. Patterson *et al.*, "Carbon emissions and large neural network training," 2021, *arxiv:2104.10350*.
6. N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, "The computational limits of deep learning," 2020, *arxiv:2007.05558*.
7. C. Freitag, M. Berners-Lee, K. Widdicks, B. Knowles, G. S. Blair, and A. Friday, "The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations," *Patterns*, vol. 2, no. 9, p. 100,340, 2021, doi: 10.1016/j.patter.2021. 100340.
8. D. R. So, C. Liang, and Q. V. Le, "The evolved transformer," in *Proc. Int. Conf. Mach. Learn.*, pp. 6010–6022.
9. D. R. So, W. Mańke, H. Liu, Z. Dai, N. Shazeer, and Q. V. Le, "Primer: Searching for efficient transformers for language modeling," in *Proc. Conf. Neural Inf. Process. Syst.*, 2021.
10. N. Jouppi *et al.*, "Ten lessons from three generations shaped Google's TPUv4i," in *Proc. Int. Symp. Comput. Arch.*, 2021, pp. 1–14.
11. T. B. Brown *et al.*, "Language models are few-shot learners," in *Proc. Conf. Neural Inf. Process. Syst.*, pp. 1877–1901, 2020.
12. N. Du *et al.*, "GLaM: Efficient scaling of language models with mixture-of-experts," 2021, *arxiv:2112.06905*.
13. B. Evans, "Amazon shocker: CEO Jassy says cloud less than 5% of all IT spending," Cloudwars, 2021. https://cloudwars.co/amazon/amazon-shocker-ceo-jassy-cloud-less-than-5-percent-it-spending/
14. S. Schechner, "Amazon and other tech giants race to buy up renewable energy," *The Wall Street Journal*, Jun. 23, 2021. [Online]. Available: https://www.wsj.com/articles/amazon-and-other-tech-giants-race-to-buy-up-renewable-energy-11624438894
15. S. Li *et al.*, "Searching for fast model families on datacenter accelerators," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2021.
16. D. Patterson *et al.*, "Energy analysis of smartphones and machine learning's role," submitted for publication.
17. P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, and J. Pineau, "Towards the systematic reporting of the energy and carbon footprints of machine learning," *J. Mach. Learn. Res.*, pp. 8085–8095, 2020.
18. R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *Commun. ACM*, vol. 63, no. 12, pp. 54–63, 2020, doi: 10.1145/3381831.
19. A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, "Quantifying the carbon emissions of machine learning," 2019, *arxiv:1910.09700*.
20. E. M. Bender, T. Gebru, A. Mcmillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proc. ACM Conf. Fairness, Accountability, Transparency*, 2021, pp. 610–623, doi: 10.1145/3442188.3445922.