Energy Costs trends and Optimization in AI

Over the past decade, the use and development of machine learning has grown exponentially. And in the past few years the extreme increase in energy consumption in the field has raised the question of whether or not the practices as of now are sustainable. Because of the world's current electricity infrastructure, a large majority of energy sources leave a carbon footprint, and as a result energy intensive domains such as machine learning are leaving a large carbon footprint. A study conducted by MIT in 2020 did an exhaustive examination of thousands of papers in machine learning research to attempt to answer the question of why machine learning has become so data hungry. The researchers claim that much of the advancements made over the past decade have been a result of harnessing more computational power and energy [1]. They take a theoretical and statistical perspective to analyze why deep learning is so costly, but also note how energy costs are scaling faster than theoretical lower bounds would indicate. It is noted that the existence of this difference implies that there are lots of potential optimizations to be made, and establishing it in a quantifiable way makes room for improvements.

The first practice which is noted is overparameterization of models as it is theoretically sound and practical to do so for deep learning [1]. To overparemetize means to have more parameters than data points in the model, and this also entails that as you increase the number of data points you have to add more parameters to the model. A simple mathematical estimate based on this fact would suggest that the cost of training a deep learning model is quadratic, as it is proportional to the input size times the number of parameters [1]. These quadratic scalings can be somewhat acceptable in some applications, but are not great when it comes to handling big data [1]. It is important to note that the estimates made are conservative ones, because although the polynomial modeling of the data was the best fit for the machine learning models

which were examined, it is entirely possible that in reality the relationship is exponential. And perhaps this is the case since in reality the scaling seen is much faster [1].

The use of large parameter models was also observed by another paper published in 2020 by Schwartz et al, mostly coming from models developed by large companies and organizations. Some notable examples at the time were Google's BERT-large8 and T5-11B36, which contained roughly 350 million parameters and 175 billion parameters respectively; OpenAI's openGPT2-XL model35 contained 1.5 billion parameters; AI2's Grover contained 1.5 billion parameters; NVIDIA's Megatron-LM contained over 8 billion parameters; openAI's openGPT-3 contained 175 billion parameters [2]. Most of these were natural language models, as language is usually considered both abstract and harder for a machine to extract features from. In the past, it was actually quite common for natural language processing models to be trained on a commodity laptop or remote server [4]. However since then the paradigm has shifted completely. The heightened popularization of large transformer neural networks in the domain of natural language processing has produced results, but at the cost of a large number of parameters. Even when performing "small" tests for optimization, the Google Brain team tests on models with parameter sizes varying from 500 million to 1.9 billion [3].

The study conducted by MIT also devoted a lot of attention to its manual examination of machine learning papers in various research subdomains. These included image classification, object detection, question answering, named-entity recognition, machine translation, speech recognition, face detection, image generation, and pose estimation. The reason for choosing these subfields was that they all have well established and well quantified benchmarks for accuracy, whereas other domains such as language or audio generation do not have this benefit. The first set of results the MIT study presents is the data in the domain of image classification, as it has an extremely long history in the field of AI and deep learning and likewise

has lots of papers to examine. To measure performance, they used image classification error rates and found that on average 5,000 times the computing power is needed to half the error rate [1]. With a 95% confidence interval, the lower estimate is 1,700 and the upper estimate is 17500 as much computation needed [1]. The researchers also mention that their analysis can measure algorithmic improvements and optimizations, and found that it accounted for 10 times the reduction of computational cost over the course of 3 year time periods. This suggests that although improvements are being made in terms of cost, it is not solving nor providing an effective remedy for the fundamental problem. The results were also examined when looking at the top 10% models in terms of efficiency, and the results were that even among those models a 3,800 times increase in cost was necessary to cut the error in half [1]. Similar results were found for the other areas, although it is emphasized that their stability is not as great since only the area of image classification had enough papers which reported the number of floating point operations and thus cost had to be measured in terms of hardware burden (which can fluctuate from machine to machine) [1].

Presenting the results of this study on their own do beg the question of how efficiency and cost was measured, which although was elaborated before the results in the study, was left out because there are lots of points for discussion on cost quantification. To associate the accuracy (performance) of the models with the cost, the researchers measured error rate against both hardware computational capacity and floating point operations [1]. Using the latter as a way to measure cost was also argued for in the paper by Schwartz et al. While there may be other metrics which may seem more obviously correlated to economic and environmental costs, a lot of them have issues of portability. For instance, it may seem most intuitive to directly measure carbon emissions that result from training and testing a machine learning model. However this metric is largely specific to the power infrastructure of the location in which the model was trained [2]. Similarly, measuring electricity needed by a machine learning model is

hardware dependent. Even measuring based on model parameters of the machine learning application fails to take into account the complexity of the operations performed on the data [2]. Meanwhile, floating point operations is the driving force behind all the aforementioned metrics: intuitively more operations means more processing time and hardware burden and thus energy cost. Another benefit of this metric is that it is comparable across hardware a model is trained on, the location in which it is done and so on [2]. Thus the researchers argued this metric would also allow for experimental claims about efficiency to be replicated and measured comparably.

That being said, using floating point operations to measure efficiency is not without its flaws either. For instance, Henderson et Al note that floating point operations do not correlate to speed nor energy. On the contrary, they note instances where convolution calculations amount to approximately 3% of the floating point operations, yet utilizing over 80% of training time [5]. There have also been networks designed in the literature which get increased efficiency in terms of energy consumption by increasing floating point operations [5]. The reason for this discrepancy is that the floating point operations still have differentials in their complexity. On traditional scales of computing these differences are negligible, but when it comes to the computation heavy and large scaling of machine learning, the differences between floating point operations becomes more pronounced [9]. Thus Henderson et Al suggest taking into account a large number of factors to get a holistic view of efficiency and cost [5]. These included metrics ranging from disk write speeds, to total memory utilization, to basic information about the hardware. However not all this information is readily available to all researchers, and this has to do with how fast the field is being pushed. Software and hardware tools need to be designed in order to measure the aforementioned factors scientifically, however lots of new processors are coming out due to the demand in machine learning and other fields such as cryptocurrency, making it hard to keep up [6].

There have also been more complex, yet theoretical frameworks lately which have been focused on weighing energy efficiency metrics and combining them. Li et al are one of the few researchers who acknowledge the importance of the data movement cost of machine learning workloads. Simply preprocessing and moving large amounts of data to and from the processor takes a lot of energy, and is perhaps often eluded because it is inherent to the process [9]. Their main takeaways were that there is always a disconnect between associating the model's responsibility for the energy consumption and the reality. Processors and their arithmetic units can differ slightly, which can be exaggerated at large computation scales, and so experimental measurements become hard to tie to the theoretical predictions [9]. Their framework for analysis takes into account both of these ends through simulation, and even if it is not adopted it is indicative of the shift in practice still needed in the research community [9].

In terms of actuable results, when measuring energy consumption and emissions directly, there are numerous examples of optimizations which have been developed after this study that achieve more promising results. Techniques such as reducing floating numbers from 32 to 16 to even 8 bit provided the same performance with 2.4 times less power consumption, while algorithmic optimizations have resulted in factors 12 times less power consumption [8].

Schwartz and al also point out the increasing energy costs machine learning has tolled over the past decade, although with a bit less granularity. For example, similar to the notion mentioned in MIT's study, the paper outlines that the cost of training a neural network is proportional to the number of hyperparameter experiments, training data set size, and the time it takes to process one data point [2]. The important addition here is that this model takes into account the number of experiments it takes to reach a particular model, whereas often in many papers only the cost of the final model is reported, if at all. Instead of doing an exhaustive analysis of research papers, the researchers looked at 60 papers randomly sampled from top AI

conferences and found that only 10-20% of them claimed better efficiency as one of their results [2]. Moreover many AI research endeavors which are known for pushing scientific boundaries in reality take a great deal of computational power: Deepmind's AlphaGo ran on 1920 CPUs and 280 GPUs to play a game of Go, and it has been estimated that the cost to reproduce their results experimentally would take $35,000,000 [2]. The paper refers to these sorts of projects as Red AI, as they are economically and environmentally costly to produce and in a way gate the research community by "buying results" [2]. Despite the negative connotation attached to the label, they also do acknowledge the importance of Red AI as it empirically validates the use of AI at the large scale and points out targets for improvement and optimization [2]. That being said, the researchers conclude that Red AI is disproportionately dominant in the research community as of now, given the literature's direction towards more performance and high computing costs, as opposed to light weight, efficient AI [2]. They note that one of the benefits of correcting this imbalance is that it allows for the research community to be more inclusive. Lower training costs and a focus on efficiency allows smaller researcher groups and motivated individuals to contribute to the body of research. Researchers Xu et al also build on the motivation established by Scharwtz et al expand on the areas for optimization. The main points of interest are training on smaller datasets, by implementing some forms of transfer learning (reusing already trained networks essentially); and being able to prune large networks such that they retain same accuracy, but have less parameters, thus greatly reducing the costs they incur while in production settings [11].

Big research groups, usually affiliated with large corporations, invest lots of experiments into tuning hyperparameters and training lots of neural networks to determine which parameters work best for the job. As alluded to previously, many examples of this can be found in the area of transformer neural networks and natural language processing. For instance, a paper published in September 2021 by Google Brain presented a new architecture named 'Primer' for

transformer neural networks, claiming it to be more compute and power efficient than previous designs. Their comparisons were done taking into account different datasets, different size models, different hardware platforms and different points at compute scaling (time). And most notably, the gains as computer cost increases do not diminish and increase as costs increase [3].

The results may sound overwhelmingly positive, the experiments are still costly, and the paper is idiomatic of the Red AI research which Schwartz and al outline. Although the final model which is settled on is demonstrated to be efficient, it took lots of experimentation to get to there. For starters, the researchers used a genetic algorithm to construct the model [3]. This means that they started with a set of transformer architectures and then mixed and matched their features: testing the efficiency with millions of data points on each iteration. The researchers do not allude to how expensive the experiments were or how many iterations they used. Additionally, this is not just an expense with the specific methodology employed by Google. An internal study done by Facebook describes the process as one where "...researchers design, implement and evaluate the quality of proposed algorithms, model architectures, modeling techniques, and/or training methods for determining model parameters" and how "...This model exploration process is computationally-intensive" [8]. Moreover the study also notes how despite the optimizations mostly being with respect to training time and cost, a lot of the costs incurred and associated with machine learning happen during inference, or in other words when a model is in deployment [8]. And to this end the literature is generally unclear on even the estimations of these costs, although as mentioned previously there is some work to reduce them heuristically.

Lots of the literature more recently does point towards hopeful expectations when it comes to leaving less of a carbon footprint, although to what extent is not exactly clear. For

example a paper jointly published by researchers at Google and University of California Berkeley identified 4 key points in which reducing the carbon footprint of machine learning workloads is possible: model choice, machine choice, mechanization and mapping [7]. The first point directly relates to the research around Primer, as it demonstrates how model choice can affect the energy efficiency of machine learning models. And although the authors do acknowledge the heavy cost associated with actually discovering these more efficient methods through experimentation, it is claimed that this cost is greatly outweighed by the savings incurred by finding a more efficient model. The reasoning for this is that machine learning models are often open sourced, trained and retrained dozens, if not hundreds of times [7]. The second observation the authors make is that hardware choice makes a large difference in energy consumption of neural network training. It is observed that processing on TPUs or GPUs tailored towards machine learning workloads can achieve the same performance using 2-5 times less watts [7]. It has also been pointed out that deploying across multiple GPUs can achieve the same performance with about 10 times less power as well [8]. The third and fourth factors identified can be paired together as they both point towards the optimizations that come with computing on cloud datacenters. For starters, neural network training is not usually latency bound: companies can run training in cloud regions geographically far away since training models usually do not require round trip communication requirements [5].

This means that several benefits of cloud computing can be extracted into the domain of machine learning: namely inherit infrastructure efficiency, renewable energy matching and locational carbon footprint minimization [7]. Cloud service users are essentially able to choose where their computing and data processing takes place, meaning that areas with efficient energy infrastructure can be chosen. The researchers show that this, combined with the intricate structure and efficiency of cloud data centers, can achieve the same performance with 7 to 20 times less watts [7]. This largely seems to contradict the common perception that these

large corporations and their machine learning models are leaving a large carbon footprint. While it is true that the machine learning workloads of companies like Google have increased, it only accounted for 10-15% of their energy consumption [7]. Moreover data centers worldwide have only seen a 6% power usage increase over the past decade, despite data center usage increasing by 500% from 2010-2018 [7]. These numbers may differ a little after the pandemic but the main point still holds. The large companies which back these data centers also match a lot of their energy consumption with renewable energy purchases and it is noted that in some countries they even invest more in renewable energy than the government subsidies provided. Of course there is a bit of conflict of interest here, it is mentioned in the paper that Google has had this policy for years and that Microsoft is also looking to achieve similar goals by 2025 [7]. Even companies like Facebook have been employing this renewable energy matching policy in order to reach net zero carbon emissions [8]. The authors also address client side machine learning and model usage. It is pointed out that the use of machine learning in mobile phones accounts for only 1.5% of the power usage in many contexts. Thus the conclusion is drawn that machine learning during the training process and server side exacts a much heavier toll on energy [7].

There are opposing sides, to the aforementioned claims, throughout the literature however. A point is made by the Facebook study mentioned previously that infrastructure for carbon limited energy sources are limited, and that encouraging the heavy use of them in the field of machine learning may overload them. Moreover the speed of machine learning research and its development in industry will likely outpace the time and cost to create carbon free infrastructure [8]. On top of this, a large portion of the public is against complete data centralization for the reason of privacy, and for good reason. Thus machine learning on personal devices is an equally growing paradigm as computing on the cloud [8]. And although some preliminary results from the previous paper showed that on device costs were negligible, there

are cases where the complete opposite is true. The researchers demonstrated that non-trivial, small machine learning tasks can have the same energy usage as a model which is “orders of magnitude” larger but is centralized [8].

Because of these considerations, a distributed methodology for machine learning at scale dubbed federated learning has also emerged. Specifically in the medical and healthcare fields, centralization of data processing, even if more efficient, would be legally dubious [10]. A paper published by Qiu et al looks to characterize and evaluate the carbon efficiency of this new distributed machine learning methodology within actual applications. Federated learning essentially takes place over a network of personal devices in which data cannot be shared between them, except for the network parameters which are being adjusted after processing a set of data [10]. The researchers make it clear that the bandwidth and networking costs associated with distributing the process were expected to be as large as the costs associated with doing the computing across multiple devices. The reasoning for this being that the emissions associated with networking and bandwidth are structural [10]. However their results showed the contrary and that on certain workloads which required lots of communication, it could account for up to 96% of the emissions. The researchers used several common and large datasets to measure the energy costs, and then converted it to carbon emissions under the assumption that each distribution used its local power grid [10]. Thus the results also demonstrated that, like in cloud computing, geographic location is vital to reducing carbon emissions in workloads where the cost of data distribution is not overwhelmingly dominant [10]. It is important to note that this study, nor the one done by University of California Berkeley and Google, take into account structural emissions, or the emissions which are associated with creating hardware, data center infrastructure, etc. The paper highlights the extensive measures needed in order to properly quantify and characterize the energy cost and carbon footprint of machine learning workloads.

References

[1] Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2020). The Computational Limits of Deep Learning. arXiv. https://doi.org/10.48550/ARXIV.2007.05558

[2] Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2019). Green AI. arXiv. https://doi.org/10.48550/ARXIV.1907.10597

[3] So, D. R., Mańke, W., Liu, H., Dai, Z., Shazeer, N., & Le, Q. V. (2021). Primer: Searching for Efficient Transformers for Language Modeling. arXiv. https://doi.org/10.48550/ARXIV.2109.08668

[4] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. arXiv. DOI:https://doi.org/10.48550/ARXIV.1906.02243

[5] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. arXiv. DOI:https://doi.org/10.48550/ARXIV.2002.05651

[6] Eva García-Martín, Crefeda Faviola Rodrigues, Graham Riley, and Håkan Grahn. 2019. Estimation of energy consumption in machine learning. Journal of Parallel and Distributed Computing 134, (2019), 75–88. DOI:https://doi.org/https://doi.org/10.1016/j.jpdc.2019.07.007

[7] D. Patterson et al., "The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink," in Computer, vol. 55, no. 7, pp. 18-28, July 2022, doi: 10.1109/MC.2022.3148714.

[8] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga Behram, James Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin S. Lee, Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, and Kim Hazelwood. 2021. Sustainable AI: Environmental Implications, Challenges and Opportunities. arXiv. DOI:https://doi.org/10.48550/ARXIV.2111.00364

[9] Chen Li, Antonios Tsourdos, and Weisi Guo. 2022. A Transistor Operations Model for Deep Learning Energy Consumption Scaling. ArXiv abs/2205.15062, (2022)

[10] Xinchi Qiu, Titouan Parcollet, Javier Fernández-Marqués, Pedro Porto Buarque de Gusmão, Daniel J. Beutel, Taner Topal, Akhil Mathur, and Nicholas D. Lane. 2021. A first look into the carbon footprint of federated learning. CoRR abs/2102.07627, (2021). Retrieved from https://arxiv.org/abs/2102.07627.

[11] Jingjing Xu, Wangchunshu Zhou, Zhiyi Fu, Hao Zhou, and Lei Li. 2021. A Survey on Green Deep Learning. CoRR abs/2111.05193, (2021). Retrieved from https://arxiv.org/abs/2111.05193