The Dark Side of Machine Learning: What to Do with the Growing Concerns Over Energy Consumption

QiLin Xue (*Word Count: 1994*)

(Dated: November 17, 2021)

In this paper, we outline key environmental issues involved in machine learning. The energy usage during training and inference of many models are growing exponentially, for diminishing increases in performance. Many groups recommend that researchers should be more transparent about the energy costs of their model. However, we find that this may be unrealistic, and we focus our attention on the data centers which provide the hardware for training and inference. If we encourage data centers to be more energy efficient and more transparent, it will allow for a push by the machine learning community to be more environmentally conscious and friendly.

I. INTRODUCTION

Machine learning is transforming the way we live. It has found its way into dozens of areas, from marketing to medicine. Consider, for example, the rise of image recognition in the past decade. It powers applications like Google Photos, Facebook, Snapchat, and Pinterest, among many others. It's also behind voice recognition, search engines, and even the object recognition on your phone.

More impressively, this entire introduction (except this sentence) was written by a model developed by OpenAI named GPT-3. GPT-3 is a deep learning system that can write simple sentences. It might not have captured the imagination of the general public yet, but it could in the near future.

It's easy to see how AI could be increasingly beneficial in the future. It's also easy to see how it could be abused. That's what makes AI so interesting. It's one of the most important technologies being developed today, and it represents the cutting edge of technology innovation.

A. Scope

Ethics in machine learning and artificial intelligence is a hot topic, with several conferences dedicated to discussing how this technology should be used and regulated. Instead, this paper will focus on an aspect of machine learning that is often neglected: the environmental impacts of running machine learning models.

B. Background

Machine learning models consist of several interconnected parameters, known as neurons, that perform a simple mathematical operation. When put together, these neurons are able to simulate extremely complex tasks. Training such a model consists of systematically adjusting what mathematical operation each neuron performs, and if there are a lot of neurons, this can be computationally expensive.

For example, GPT-3 consists of 175 billion parameters. Training just one model has the same carbon footprint as the lifetime carbon footprint of five cars, including production[1][2]. A model may be trained many times during development, and while the company has not openly stated

how many times the model was trained, other researchers estimate it to be in the thousands[2].

II. ISSUES WITH ENERGY CONSUMPTION

A. Exponential Growth in Size

Machine learning models have seen an exponential growth in size. The size of a typical deep learning model has grown from tens of thousands to hundreds of billions of parameters in just a matter of years. This is due to the fact that more and more data is being collected, which in turn leads to better models. However, this also means that training these models becomes increasingly difficult as they grow larger. In order to train these models, it is necessary to have large amounts of computing power. The bigger the model grows, the more computing resources are required to train it[3].

The trend first began with AlexNet, which created a revolution by winning the 2012 ImageNet Large Scale Visual Recognition Challenge by a huge margin through the use of deep learning. Since then, models have grown in size and complexity, and we're now at the stage where state-of-the-art models such as ResNet use millions of parameters and require huge amounts of computational power[4].

B. Diminishing Returns

One may argue that the energy costs may be a necessary sacrifice in order to obtain models that have great socioeconomic benefits for humanity. However, increasing performance by increasing the number of parameters is foolish.

This is because the performance of a model does not scale linearly with the amount of computations that the model performs. Theory suggests that to double the performance, one needs to increase the computations by at least 16 times. However in practice, this may be much higher. In fact, using recent models as a guideline, researchers estimate that if we continue current trends, the number of computations that deep neural networks make needs to increase by 128 to 16384 times in order to just double the performance[5].

As an example, in the image classification challenge ImageNet, the models that have the most success are the ones that are bigger. In order to half the error, the number of computations need to increase by nearly 1000 times[5].

The increased number of computations leads to higher energy usage in not only the training phase, where a model may be trained thousands of times, but also after the model has been deployed.

C. Energy Costs of Inference

When machine learning models are used by the client, known as inference, i.e. during a Google Search, the computations don't occur on one's personal device. Instead, they run on a data center run by GPUs. There are two reasons for this:

First, many modern models are proprietary software, so corporations would want to control when it is called. Second, as seen in the GPT-3 example, models can have hundreds of billions of parameters. Most personal use devices simply do not have the hardware requirements to carry out these computations in a relevant amount of time.

This is a little-known but surprising fact. Not only does the usage of phones and computers contribute to e-waste, but it also contributes to the production of carbon dioxide in data centers across the world. In fact, the computer systems company Nvidia which designs

GPUs estimates that 80-90% of the cost in neural networks reside in inference instead of training[6].

D. Energy Costs of Data Centers

Data centers is a broad term to encompass a building that houses a collection of hardware designed to perform a certain task. In 2020, data centers accounted for 1% of the world's electricity consumption and 0.3% of the carbon emissions[7]. They use a lot of energy because not only do they need to store data and carry out the necessary computations, cooling mechanisms need to be in place to ensure computers do not overheat.

The abundance of data centers means that there are a variety of different data centers with different energy performances.

III. RECOMMENDATION

This paper will first present two suggestions, which, although they have good intentions, this paper does not believe they will play a significant part in making machine learning more environmentally friendly.

A. What not to do: Push for Researchers to be Transparent

First, many people suggest that researchers should be transparent with energy costs in their papers. One group of researchers suggest that papers should report the training time and the computational resources required in hopes that this transparency will raise awareness for the environmental impacts of training deep learning models as well as encouraging models that use less resources[8].

Many journals already require researchers to be transparent about certain information, such as their model accuracy and architecture in order for others to reproduce the result. However, we find requiring researchers to also report the environmental cost of their models to be unrealistic and problematic for two reasons.

First, a group of researchers from the ByteDance AI Lab rightly describe the issues with such an approach. If everyone runs the code under the same hardware and software, then the above suggestion would be a great idea. However, this is hardly the case. For example, carbon emissions are dependent on the local infrastructure and some hardware may be more efficient than others. There are too many variables that may influence the environmental impact of training a model[9].

Second, even if there was a method to take into account all the various factors, the environmental impact during training is only part of the story. As mentioned in a previous section, a big portion of the energy usage comes from inference. Some studies have used the number of floating-point operations (FLOPs) to measure how computationally intensive a model is. This metric is related to both inference and training with the added benefit of being almost independent of hardware and software. However, this is only a theoretical metric and experiments have shown it is not always a good representation[9].

B. Greener Models come Naturally

In fact, we suggest that unlike heavy polluting industries such as the oil and gas sector, there is an inherent incentive for corporations to develop greener models without the intervention of the government or other third parties. This is because while the cheaper option in many industries involves the use of fossil fuels, the cheaper option in machine learning is the greener one.

Corporations have an internal incentive to

develop more efficient algorithms to shorten both the training time and the energy costs. We are already seeing a trend. Recently, a team of researchers has developed a technique to reduce the energy cost of large models. Their model M6 has 10 trillion parameters, which is 80x the size of GPT-3, yet the energy cost is only 1% of GPT-3[10].

The currently being developed successor to GPT-3, called GPT-4, was initially planned to have 100 trillion parameters[11]. However in a September Q&A with Sam Altman, the CEO of OpenAI, Altman claimed that GPT-4 will be much smaller, as they have found a way of reducing the size while still improving the quality[12].

There seems to already be a shift in the development of large models even without external incentives, whether it is in the form of more energy efficient algorithms or finding ways to improve quality with fewer parameters.

C. Push for Greener Data Centers Instead

The efficiency of data centers vary a lot. Many cloud data centers can be up to 2 times as efficient as traditional data centers. This increase in efficiency is due to the fact that cloud data centers can be located in areas of the world where renewable energy is readily available and in colder climates to reduce the energy required to cool down systems[13].

For example, Google and Microsoft have been carbon neutral since 2012, meaning that the emissions they create are equal to the emissions they reduce somewhere else, such as by investing in renewable energy[14]. In the last month, Google Cloud has also initiated a change in which it notifies users of their energy consumption and carbon footprint[15].

With the rise of machine learning and dominance of inference, companies like Alibaba, Amazon, Google, and NVIDIA have all created data centers specifically tailored towards inference, which can be 2x to 5x times more energy efficient than traditional data centers[13].

Recommendation: We recommend that independent researchers as well as corporations choose cloud data centers for training and specially designed machine learning accelerators for inference. For accountability, we encourage data centers to also be transparent with customers about the carbon footprint of their usage.

One possible drawback is that many traditional data center companies may lose customers, thus losing money. However, traditional data centers have other customers as well and the shift to greener data processing imposed by the machine learning community will encourage companies to either use more renewable energy, invest in cloud data centers, and/or redesign the infrastructure to be tailored towards training or inference.

IV. CONCLUSION

Machine learning is at the forefront of the next big revolution we are currently going into. While work is being done to discuss the equity concerns this technology may bring, it is also important to acknowledge the environmental impact, often done in data centers hidden from the public view.

We recommend that data center companies push towards using cleaner energy and be transparent with their energy usage. We ask that machine learning researchers push to use greener data centers for training and inference, in order to put pressure on data centers to be more transparent.

[1] T. Brown and e. a. Mann, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds.), vol. 33, pp. 1877–1901, Curran Associates, Inc., 2020.

[2] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in nlp," 2019.

[3] "Machine learning is getting big (part i)," Jan 2020.

[4] D. Gershgorn, "The data that transformed ai research-and possibly the world," Jul 2017.

[5] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, "The computational limits of deep learning," 2020.

[6] M. I. a. Strategy, "Google cloud doubles down on nvidia gpus for inference," May 2019.

[7] N. Jones, "How to stop data centres from gobbling up the world's electricity," Sep 2018.

[8] L. F. W. Anthony, B. Kanding, and R. Selvan, "Carbontracker: Tracking and predicting the carbon footprint of training deep learning models," 2020.

[9] J. Xu, W. Zhou, Z. Fu, H. Zhou, and L. Li, "A survey on green deep learning," 2021.

[10] J. Lin, A. Yang, J. Bai, C. Zhou, L. Jiang, X. Jia, A. Wang, J. Zhang, Y. Li, W. Lin, J. Zhou, and H. Yang, "M6-10t: A sharing-delinking paradigm for efficient multi-trillion parameter pretraining," 2021.

[11] A. Romero, "Gpt-4 will have 100 trillion parameters - 500x the size of gpt-3," Sep 2021.

[12] "Sam altman qa recap: Gpt and agi," Sep 2021.

[13] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, "Carbon emissions and large neural network training," 2021.

[14] D. Oberhaus, "Amazon, google, microsoft: Here's who has the greenest cloud," Dec 2019.

[15] B. N. . O. 12, "Google will show its cloud customers their carbon footprint," Oct 2021.