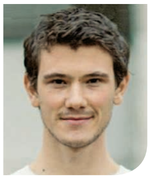
Alexandra Luccioni

Alexandre Lacoste

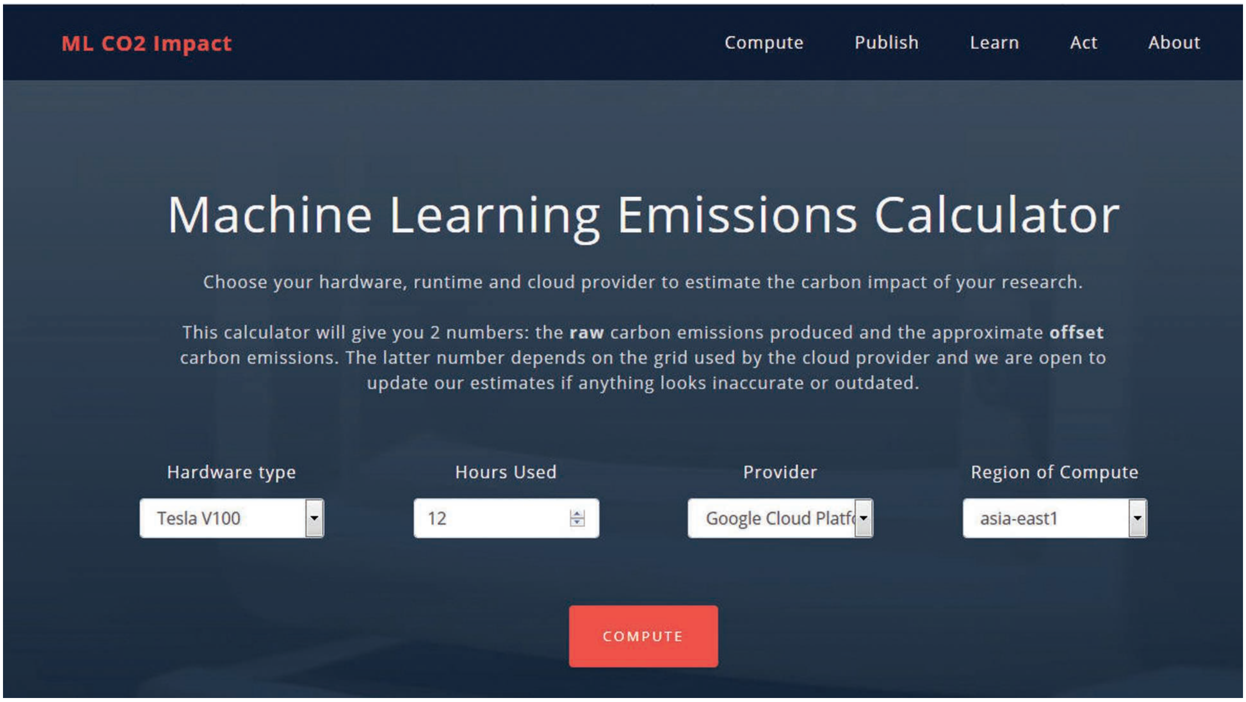


Victor Schmidt

Estimating Carbon Emissions of Artificial Intelligence

Advances in Artificial Intelligence — and Machine Learning (ML) in particular — have resulted in amazing progress in the last years and decades, democratizing technology and making many aspects of our lives simpler, faster, and less complicated. While many of us hear about the latest and greatest breakthrough in AI technology, what we hear less about is its environmental impact. In fact, much of AI's recent progress has required ever-increasing amounts of data and computing power. And this all comes at a cost — while currently cloud computing represents roughly 0.5% of the world's energy

Digital Object Identifier 10.1109/MTS.2020.2991496
Date of current version: 18 June 2020



Machine Learning Emissions Calculator

Choose your hardware, runtime and cloud provider to estimate the carbon impact of your research.

This calculator will give you 2 numbers: the **raw** carbon emissions produced and the approximate **offset** carbon emissions. The latter number depends on the grid used by the cloud provider and we are open to update our estimates if anything looks inaccurate or outdated.

Hardware type	Hours Used	Provider	Region of Compute
Tesla V100	12	Google Cloud Platf	asia-east1

LUCCIONI

consumption, that percentage is projected to grow beyond 2% in the coming years [1].

We believe that tracking and communicating the environmental impact of ML should be a key part of the research and development process, and have developed a tool for estimating the carbon impact of this process, the Machine Learning Emissions Calculator (see Figure 1). We present the tool and its importance in the present article, and explore related issues and challenges.

Factors Involved in Estimating ML Carbon Emissions

Neural networks are essentially complicated computer architectures with thousands — sometimes millions — of connections and weights, and millions of parallel calculations that have to be carried out both during the training of the network (when it learns a task, for instance classifying images into classes) and during inference time (when the result of the training is applied on a new sample, i.e., an image that was not seen during training). In fact, training neural networks is a complicated balance of parameter tuning, optimization, and often much trial and error. For each successful training of a network, which ends up having the network successfully do the task that it was meant to do, there are dozens and even hundreds of failed experiments. This means that while, in itself, a single training procedure of a given neural network is not necessarily very energy-consuming (and carbon-emitting), if all of the experiments are taken into account, this can quickly add up to a significant amount of emissions. There are a few factors that have the biggest impact on these emissions, however, and we will discuss these below.

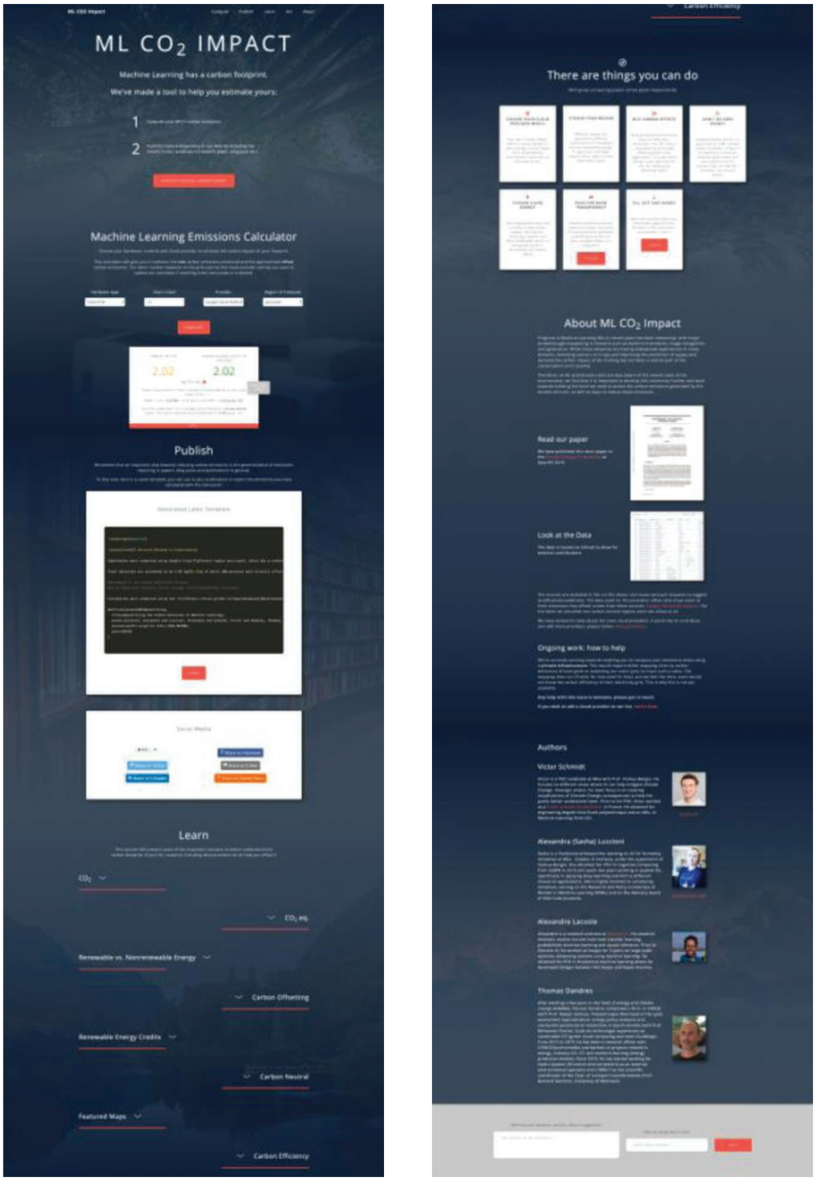

ML CO₂ IMPACT
Machine Learning has a carbon footprint.
We've made a tool to help you estimate yours:
1
2
Machine Learning Emissions Calculator
2.02
2.02
Publish
Learn
CO₂
CO₂ eq
Renewable vs. Nonrenewable Energy
Carbon Offsetting
Renewable Energy Credits
Carbon Neutral
Featured Maps
Carbon Efficiency
There are things you can do
About ML CO₂ Impact
Read our paper
Look at the Data
Ongoing work: how to help
Authors
Victor Schmidt
Alexandra (Sasha) Luccioni
Alexandre Lacoste
Thomas Dandres

FIGURE 1. Machine Learning Emissions Calculator: Web Interface.

Type of Energy Used

Few energy consumers have the ability to choose how much carbon they will produce when plugging in a device to a power outlet in their home or office; in most cases, this is defined by the energy source that the outlet is connected to. In the case of a physical device, this means that the energy is taken from the local energy grid of the socket's location, which can be generated from coal, hydroelectricity, solar, nuclear, or any combination of these (and others). These sources can be broadly divided into *renewable energy,* collected from sources that are naturally replenished relatively quickly (e.g., wind, solar, hydro, tidal, etc.), and *nonrenewable energy sources,* which can take millions of years to be formed again (this is the case for coal).

In the case of Machine Learning models trained locally on a server connected to the power socket in the nearest wall, not much can be done to select an energy source. However, since increasing quantities of models are trained on the cloud, it is entirely possible and even very easy to choose the location of a server used for training and, therefore, the energy source it is connected to. Therefore, while practically speaking it is hard to calculate precisely how much carbon is emitted by a model during its training on the cloud, this amount can be estimated using information about the local energy grid if we assume that all servers are connected to local grids at their physical location and no alternative source of energy exists.[1]

Cloud Providers

When running on the cloud, it is necessary to choose a cloud provider, whose servers are scattered throughout the word and therefore connected to grids with different emissions factors. In order to create our ML emissions calculator, we gathered as much publicly-available data as we could regarding the carbon emissions of different energy grids, at varying levels of granularity, from regional to national. We then cross-referenced these with cloud server locations from the three major cloud providers: Google Cloud Platform, Microsoft Azure, and Amazon Web Services. We found that emissions can vary by up to a factor of forty depending on the energy mix of the grid. For example, in a region like Québec, Canada, which relies mostly on hydroelectricity, this can be roughly 20 grams of carbon per kWh, whereas in a place like Queensland, Australia, this can go up to 800 grams of carbon, since Queensland's power grid relies almost solely on fossil fuels. The carbon emissions can really add up for a big neural network trained on a few graphical processing units (GPUs) for several weeks, resulting in almost a ton of carbon, or the equivalent of a transatlantic flight [2].

Something to keep in mind is that some major cloud providers are already *carbon neutral*, meaning that they have a net zero carbon footprint. This is done via carbon offsetting and renewable energy credits (RECs), which involve purchasing an equivalent amount of renewable energy for the amount of non-renewable energy used. This does not mean, however, that the associated cloud computing does not emit carbon — it just means that there is a significant effort to balance the carbon emitted by incentivizing and democratizing the use of renewable energy. There is also a general tendency towards more transparency and accountability, which can help when choosing a given cloud provider when several operate in the same region. For instance, Microsoft recently announced that they will be "carbon negative" by 2030, meaning that they will remove all the carbon that they have emitted either directly or by electrical consumption since their founding in 1975.[2]

Hardware and Training Time

While computing hardware has been getting more powerful, capable of carrying out more calculations in less time, efficiency is being reduced because computing is being used for longer and longer to learn more and more complex tasks. For instance, MegaFace, a popular facial recognition dataset, has almost 5 million images [5], and training a state-of-the-art facial recognition model with millions of parameters can take months on even the most powerful hardware. While this case should be considered common practice, there is a general trend that can be observed towards more data, more powerful computing hardware, and longer training time. Recent data published by OpenAI shows that the computing power required for key ML tasks has doubled every 3 months or so, increasing 300 000 times between 2012 and 2018. At this rate, not only are emissions from AI going to rise exponentially, but so will the barriers to entry in the domain, since few individuals and smaller companies and research labs will be able to afford the sheer computational power required to innovate and to beat existing leaderboards.

As we mentioned above, while the initial network training procedure definitely accounts for a large portion of its carbon emissions, there are also other factors to consider: for instance, whether it is necessary to fine-tune the network and to what extent. In fact, in a recent article comparing the carbon emissions produced by a neural network to the average lifetime emissions of five cars, only a small fraction of those emissions were produced by a single model training, whereas the majority was a result of the architecture search and hyperparameter tuning [6]. This is an important factor to keep in mind, since many recent powerful models in areas like machine translation and image recognition have been shared online with the broader research community, meaning that it is no longer necessary to train models from scratch, and it is possible to take a pre-trained model and use it as is, or train it for less time on a more specific task. This means that, for example, if a neuro-linguistic programing (NLP) model was initially trained to translate English to French texts on a huge corpus with millions of documents and shared publicly, it can then be customized on a smaller set of documents to carry out

[1] This is not always the case, since sometimes data centers rely in part or in full on local energy sources and are not connected to the energy grid, e.g., Google's St. Ghislain data center, which has solar panels on its roof (https://blog.google/around-the-globe/google-europe/time-shine-new-solar-facility-and-additional-data-center-belgium/).

[2] Source: https://blogs.microsoft.com/blog/2020/01/16/microsoft-will-be-carbon-negative-by-2030/.

translations in the legal domain, even if there were no legal documents in the initial training texts [3].

Action Items

We do not pretend to offer specific guidelines that anyone can follow to reduce the carbon emissions of their models; however, given the factors described above, there are some concrete actions that can be taken to reduce the carbon footprint of ML:

1) **Choose Your Cloud Providers Wisely**: all of the major cloud providers have information regarding their sustainability efforts on their websites and there are third-party resources that endeavor to compare their environmental footprints.
2) **Select Data Center Location**: When requesting a cloud GPU or CPU, it is very easy to request a server in a specific location (e.g., U.S.-East, Europe-West, etc.) in order to choose the least carbon-intensive location. There is freely available emissions factor documentation that provides the carbon emitted by energy grids worldwide.
3) **Reduce Wasted Resources**: Carrying out a literature review before starting experimentation, using pretrained models when possible, and using random search instead of grid search can reduce the quantity of failed experiments needed to obtain the best results, and therefore the footprint of the model as a whole.
4) **Choose More Efficient Hardware**: Recent generations of computing hardware such as GPUs and tensor processing units (TPUs) have been specifically designed for the parallel computations involved in training neural networks. Using this hardware instead of traditional chips can improve the efficiency of training ML models, as well as reducing training time and therefore, energy usage.
5) **Use our ML Emissions Calculator**: By inputting details regarding the training of an ML model, such as the region of the server, the type of GPU, and the training time, our tool gives as output the approximate amount of CO_2eq produced. Using our tool can give a good estimate of the order of magnitude of emissions produced by a given ML experiment or set of experiments.[3]
6) **Disclose the emissions associated with published ML results**: Nowadays, few papers disclose the specifics of their training approach, i.e., what infrastructure they used for training and how long it took to obtain their results. Publishing both these details and the overall emissions generated by ML experiments is important for raising awareness around the environmental footprint of ML research.

Insight into Environmental Impact of ML

We have discussed some major factors and considerations in the current article, enabling Machine Learning practitioners to have some insight regarding the environmental impact of training of their models. We realize that it is not always possible to take all of these factors into consideration during ML practice, which brings with it constraints such as privacy and data accessibility, but these considerations are useful to keep in mind as a guiding thread towards more sustainable ML research and practice.

Our hope is that estimating and disclosing the quantity of carbon emissions produced by ML models will increasingly become a more mainstream phenomenon and part of the ML research process, similar to the way in which sharing code and data has increasingly become the norm in recent years. In order to facilitate this even further, we are currently working to create an easy-to-install python package that will allow seamless tracking during experimentation time. We believe that our work, along with that of others, will open the door to measuring the environmental impact of our field, and for making positive changes in order to reduce those impacts.

Author Information

Alexandra Luccioni is with Mila, and with the Université de Montréal, Montreal, Quebec, Canada.

Alexandre Lacoste is with Element AI, Montreal, Canada.

Victor Schmidt is with MILA, Montreal, Canada.

References

[1] D. Guyon, A.C. Orgerie, C. Morin, and D. Agarwal, "How much energy can green HPC cloud users save?," in *Proc. 2017 25th Euromicro Int. Conf. Parallel, Distributed, and Network-Based Processing (PDP)*. IEEE, Mar. 2017, pp. 416-420.

[2] N. Hill, C. Dun, R. Watson, and K. James, "2015 Government GHG conversion factors for company reporting: Methodology paper for emission factors," final rep., *Department of Energy and Climate Change (DECC)*, London, U.K., 2015.

[3] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *arXiv preprint arXiv:1801.06146*, 2018.

[4] J. Koomey, "Growth in data center electricity use 2005 to 2010," *Analytical Press*, report completed at the request of *The New York Times*, vol. 9, p. 161, 2011.

[5] I. Kemelmacher-Shlizerman, S.M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 4873-4882.

[6] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," *arXiv preprint arXiv:1906.02243*, 2019.

[3] While we endeavor to estimate the quantity of CO_2 produced by the energy usage involved in training an ML model, this is a simplified estimation of the total CO_2 produced by a given ML model, which would also need to include an extensive Life Cycle Assessment of the hardware used during the process, as well as the emissions produced during inference time.