ESTIMATING THE SUSTAINABILITY OF AI MODELS BASED ON THEORETICAL MODELS AND EXPERIMENTAL DATA

ESTIMATING THE SUSTAINABILITY OF AI MODELS BASED ON THEORETICAL MODELS AND EXPERIMENTAL DATA

Ralf Gitzel[1]*, Marie Platenius-Mohr[1] and Andreas Burger[1]

[1] ABB Corporate Research
* Correspondence: ralf.gitzel@de.abb.com

Abstract: As deep learning AI becomes more and more common in business and even in our daily lives, it is important to understand what the carbon impact of this type of software is. Recent papers have shown that it can be quite great, i.e., the training of a single high-end model can result in emissions of more than 500t of CO_2eq. In this article we describe a life-cycle-focused framework to estimate the carbon drivers of a new deep learning model. We experimentally verify some claims in the literature and provide suggestions on how to reduce the carbon footprint of a deep learning-based offering. The article should enable developers and managers to make informed and meaningful decisions to minimize their ML projects' sustainability impact.

Keywords: AI, Sustainability, Energy efficiency, Deep learning, Neural networks

1. Introduction

Artificial intelligence (AI), more specifically machine learning (ML), is increasingly becoming part of our lives. Many-layered neural networks (deep learning models) have brought us technological wonders. Facial recognition allows us to conveniently protect our smart phones. Natural language processing models understand human speech and turn it into commands for smart home applications. Companies use AI extensively in industrial applications ranging from the interpretation of infrared images of machinery to the analysis of production-related data. There is a strong competition to improve performance which leads to larger models that are trained longer. This in turn implies a greater energy consumption and thus more CO_2 emissions (cf. [1], [2] and [3]).

Yet, as concern over the climate crisis increases, more thought is given to the carbon footprint of AI models. Each model generation seems to grow in size and models start to consume energy on a massive scale for training alone. Critics of this trend cite examples such as GPT-3 [4], a deep NLP model with 175B parameters that writes human-like texts and needed 1'287 MWh for training. This corresponds to 552 t of CO_2, which is equal to the annual emission of 276 average-style cars [2]. On the other hand, many AI models used today are much smaller. In fact, Patterson et al feel that some studies exaggerate the scope of the problem (cf. [5]). In the end, many AI providers are not sure what their models' carbon footprint is and how to reduce it.

The purpose of this article is to give managers and data scientists some guidance to understand exactly what impact their individual models have on the environment. For this purpose, a simple but comprehensive framework is presented that explains the key carbon drivers. We will give some advice as to how to reduce those drivers. Finally, some experimental results test the statements in the literature and challenge the current recommendations related to transfer learning.

2. The Carbon Footprint of AI – State of the Art

There is already a body of related work that has analyzed potential drivers for the carbon footprint of deep learning AI models. For example, there is a series of theoretical models that can be used *ex ante* to estimate the carbon impact of a new AI model based on its architecture (mainly layer types and size), training approach, and use for inference (see [2], [3], [6], and [7]). Between these approaches, there is some discussion about which metrics drive the carbon footprint and which are potentially deceiving ([3], [8], [9], [10]). Since most of these models put focus on a particular step in the ML lifecycle and they use different input metrics, a consolidated view is hard to achieve.

On the other hand, the carbon footprint of ML models can be measured *ex post* with software tools to document the impact of development or use (*carbon accounting*). Some tools are web-based and use key metrics such as training time, energy mix, and hardware information to estimate the carbon footprint of a model ([6], [15]). Other tools such as *energyusage* or *codecarbon* integrate directly with the ML code ([6], [12], [16], [17]). Often, CPU power usage is computed using the *RAPL (Running Average Power Limit)* interfaces found on Intel processors. Tools like *nvidia-smi* can be used to make an estimate for computations run on the GPU. Comparison of different types of hardware or types of models in benchmark experiments is an important aspect of research ([11], [12], [11], [13], [14]).

Related topics are the use of deep learning to solve sustainability problems (e.g., [18]) and ML applications not related to deep learning. However, these are beyond the scope of this paper.

3. Drivers of Deep Learning AI Models' Carbon Footprints

The carbon footprint (CO_2eq) of deep ML is driven by the energy use of the models (in kWh) and the carbon intensity of the energy source (lbs/kWh). Various decisions at the different life cycle stages influence a model's energy use and location and sourcing influence the carbon intensity. Among the most expensive examples given in the literature are GTP-3 (1'214'400 lbs CO_2eq) [2] or NAS (626'155 lbs CO_2eq) [7]. However, the carbon footprint of other high-performing models is a lot lower, e.g., $BERT_{base}$ has a footprint of 1'438 lbs CO_2eq [7]. For comparison, an average car emits 11'000 lbs per year.

There are three important life cycle phases to consider when estimating the impact of a deep learning model (cf. [2], [3]): First a model is designed (Model Architecture Search). Next, it is trained with data (Training) and finally, it is run by its users (Inference).

3.1. Inference

Inference is the "usage stage" of a model and thus the last life cycle stage in ML While a single inference is quite cheap, inference is executed many times in the field and is estimated to cause 80-90% of a model's total energy use (see [2], [19], or [20] and section 4.3). Also, the other life cycle phases execute inferences multiple times to optimize various parameters. Thus, the energy use of inference needs to be explained first as it influences Training and Model Architecture Search.

In essence, inference is the application of a complex mathematical formula using learned parameters to transform an input vector into an output vector. The output could be an image, a time series, a predicted value, or interpreted as a classification of the input. For example, a model could take a matrix of float values (representing a grey-scale image) as input and return a number that is close to 1.0 if there is a car shown in the image and near 0.0 if not.

In the simplest case shown in **Figure 1** (i.e., a model consisting only of dense layers, the most basic of all layer types), the mathematical operations for a layer consist of a matrix multiplication as well as the application of a simple activation function such as $f(x) = \max(0, x)$, called ReLU, to the result. The output of one layer acts as the input for the next layer leading to a series of matrix multiplications. These mathematical operations need a certain amount of energy while being executed.

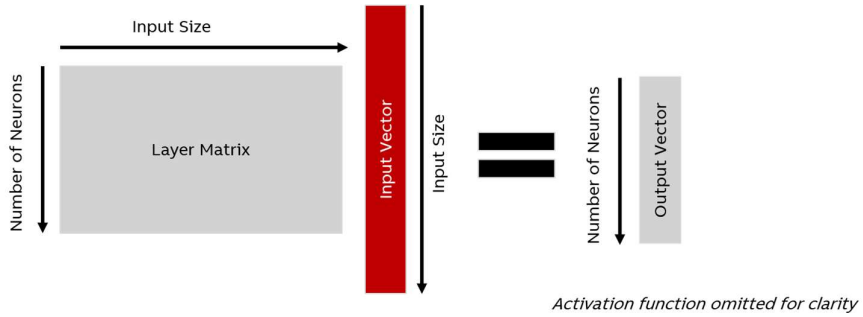

Figure 1. *Inference on a single layer expressed as a matrix multiplication*

Inference energy use depends on the model architecture (M), i.e., layer types, their order, and their size, as well as the type and quantity of processing units (*PT*), the main types being CPUs, GPUs, and TPUs. The energy use is further influenced by the power usage effectiveness (*PUE*) of the data center or similar infrastructure [21]. Thus, the energy cost *I* of an inference can be described as:

$$I = f(M, PT) \cdot PUE$$

It is not easy to determine a simple approximation for *f* because the exact way the hardware performs computations and uses memory can differ greatly. Also, more specialized layers used in modern models differ quite a bit from the basic principle described above, introducing more complexity. Due to this heterogeneity, attempts to replace M with substitutes such as the number of trainable parameters [23] is problematic [8]. It might seem disheartening that even the basic building block of a model's carbon footprint is not easy to calculate. However, using the software tools mentioned above, it is easy and inexpensive to measure *I* and use it in calculations to determine the total life cycle carbon footprint of the model.

There are several ways to optimize the carbon footprint of inference. PT and PUE can be optimized by choosing highly efficient data centers and/or hardware. Hardware optimized for matrix operations can reduce energy use. For deep learning applications, a GPU is 10 times more efficient than a CPU. A TPU is 4 to 8 times more efficient than a GPU [6]. The energy cost of memory (DRAM) access and storage is non-negligible but difficult to model ([16], [7], [22], [9]), so it is not easy to say what exact impact it has and how to minimize it.

Selecting a good M is a bit more complex. A good choice of M can reduce energy use without significant sacrifices to performance (cf. [23], [12], [17]), lowering computation effort by factors 5-10 ([2] , [5]) or even 40 for CNNs [23]. One rule of thumb is to reduce model size. Recommended techniques for size reduction are pruning, adding sparsity, quantization, or knowledge distillation ([2], [25]). The latter trains a smaller model with random data classified by a larger model trained for the actual problem (cf. [26]).

3.2. Training

Energy use for training depends on the time and number of processors [2]. It is driven by three factors: The energy cost of a single inference (I), the size of the training data set (D) and the number of epochs (E) used to optimize the model weights. There is also significant overhead for the loss function and backpropagation step that is expressed as a constant θ (see section 4.3 for a possible estimate).

$$T \propto E \cdot D \cdot I \cdot \theta$$

The formula implicitly considers the PUE and type of processor via *I* but ignores static power consumption. Theoreticaly, training energy can be reduced by transfer learning (cf. [2], [3], [6], or [27]). However, our empirical analysis adds some caveats (section 4.1).

3.3. Model Architecture Search

Different model architectures can be used for the same task with different accuracies. At the Model Architecture Search (MAS) [2] stage, many different architectures are trained, and the best solution is selected for final training. Today, the optimization criterion is primarily performance but there is no reason why energy consumption cannot be included in the search.

The cost at this stage (CT) is proportional to two factors (cf. [3]): The cost of training *T* and the number of times the hyperparameters are tuned (H). Some of *T*'s components, i.e., *I*, *E*, and *D* (see above), might vary for each tuning step resulting in different values of *T* for each step in the tuning.

$$CT \propto \sum_{h=1}^{H} T_h \cdot$$

Trying many different variants (*H*) multiplies the energy use, so good search strategies are important. The worst approach is to use grid search which systematically compares many very similar architectures with little improvement making even random search preferable [6]. Also, starting with a good architecture (e.g., when applying transfer learning) can cut down or even eliminate MAS [2].

3.4. Life cycle Energy Use and Carbon Footprint

The total life cycle energy use depends on the energy cost of all three life cycle phases (*CT*, *T*, *I* and the expected number of inference calls *e*):

$$E_{life} = CT + T + I \cdot e$$

.

The conversion of energy use into Co_2eq is done by multiplying the energy cost with the carbon emission factor (*EF*):

$$CO_2eq = E_{life} \cdot EF$$

EF is a critical factor for the carbon footprint that can vary greatly depending on the source of the energy used. Even in North America, *EF* ranged from 20g CO_2eq/kWh (Quebec) to 736.6g CO_2eq/kWh (Iowa) in 2019 [6].

Optimizing location and scheduled execution time can reduce energy use by up to 80% according to Xu [28]. Therefore, choosing the right location "*is likely the easiest path for ML practitioners to reduce CO_2eq*" [2].
Using the inputs from the prior steps, the framework can provide a reasonable estimate of a deep learning model's carbon footprint. It ignores some aspects like static energy consumption and excludes the impact of the original hardware production from its scope (cf. [16]). Also, compared to some other existing models, it sacrifices accuracy for the sake of ease of use.

4. Empirical Evaluation and Carbon Footprints of Different Models

The framework shown in the previous section is based on statements taken from the literature. We have conducted a series of experiments to test the underlying assumptions. Clearly, these experiments cannot serve as definite proof, but they add some further arguments for or against some aspects of statements in the literature. Especially the findings regarding pretrained models suggest a reexamination of that topic. The code (Keras/Python) was tested on a PC with a *GeForce RTX 2080 Ti* GPU and 32 GB RAM. For emission calculations, the energy mix of Germany was assumed.

4.1. Training Set Size, Epochs, and the Use of Pretrained Models

While training set size and the number of epochs are related to model accuracy, they also drive carbon footprint. In an experiment, the number of training samples linearly increased the carbon footprint of the model used. The same happend when the number of epochs was increased for two models of different sizes (for which we tested 100 and 50 layers of size 25).
A solution to reduce the number of epochs as well as training set size is to use pre-trained models (cf. [2], [3], [6], [27]). In the example experiment conducted by Walsh et al, using a pretrained model was almost 15 times as energy efficient as training a model with the same architecture from scratch. The example used was the Xception model repurposed for classification on the "cats vs dogs" data set [27]. While we could confirm the lowered energy use (the 0.337 kWh in our case is in the same region as the 0.32 kWh in the paper), we would like to point out a critical problem with this example. Since Xception is already capable of distinguishing dogs and cats (even different breeds)[1], there is little point in doing a retraining. Instead, we chose another domain (MNIST) and used Xception as a pretrained model[2]. The new model (6 epochs) requires only 0.451 kWh, a lot less than full Xception. However, a smaller, dedicated MNIST model with even better accuracy requires even less energy (0.005 kWh). Using transfer learning for problems of lesser complexity can actually *increase* carbon footprint, not even accounting for the inference phase.

4.2. Impact of Model Size

While the connection between model size and energy use is not straightforward, larger models *generally* need more energy than smaller models, especially if the model properties are mostly the same otherwise. Our experiments are in line with this statement (**Figure 2**). In the experiments, layer size varied between 25 and 100 nodes and the number of layers ranged from 10 to 170. The models were trained for 10 epochs with 100'000 samples. Except for an odd dent at the end, the growth seems almost linear.

[1] https://gist.github.com/yrevar/942d3a0ac09ec9e5eb3a

[2] Slightly adapting the cats_v_dogs code used for the TLR experiment

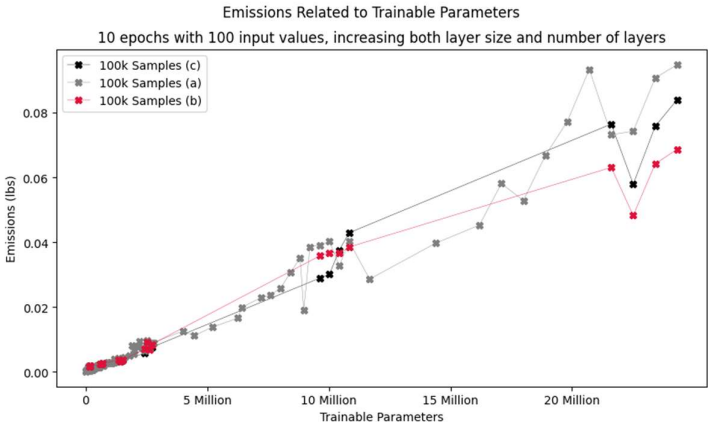

Figure 2. *Influence of model size on energy consumption*

As stated before, however, the literature advises against using the number of trainable parameters in the model as a carbon driver. In fact, when we compare wide and narrow models with the same number of trainable parameters, there is a great divergence in energy use due to the way training works at the lowest levels (**Figure 3**). The energy consumption of deep nets is a lot higher than that of wide nets of the same size. However, one should not draw the conclusion that wide networks are preferable to deep ones. In fact, Zhou et al have shown that depth is far better at increasing expressive power of a neural network than width [29].

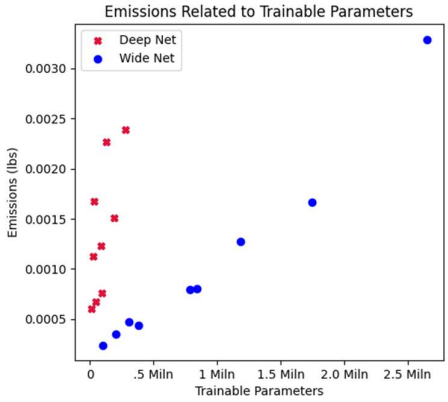

Figure 3. *Energy Footprint of varying shapes (10 epochs, 100k samples with 100 values)*

Certain papers (e.g. [9] or [23]) stress the different behavior of specific layer types such as convolutional layers. To get a rough impression of the impact of layer type, two groups of models were compared. One is a series of wide models with dense layers, the other is a series of similarly-shaped convolutional layers (where the "width" is represented by the number of filters). **Figure 4** compares the models by trainable parameters. As can be seen, purely convolutional models with the same number of trainable parameters consume a lot more energy.

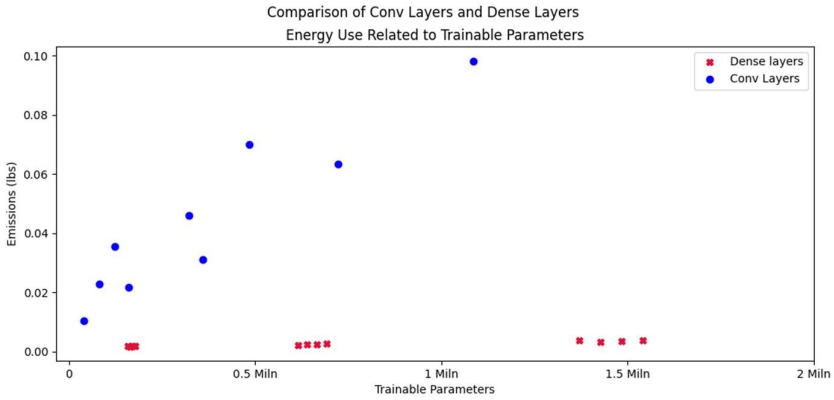

Figure 4. *Energy consumption of different layer types*

4.3. Training vs Inference

One final experiment we conducted was to find out what the overhead of training vs basic inference is and whether it is a constant value that can be used in the formula in section 3.2. In the experiment we compared pure inference on the training set to actual training on the same set for models of various layer sizes and depths. When expressed as a percentage of the total training emissions, the overhead appears to asymptotically approach a value of about 80%. Figure 5 shows this overhead for the experiments from above as well as the same models trained with 10000 samples.

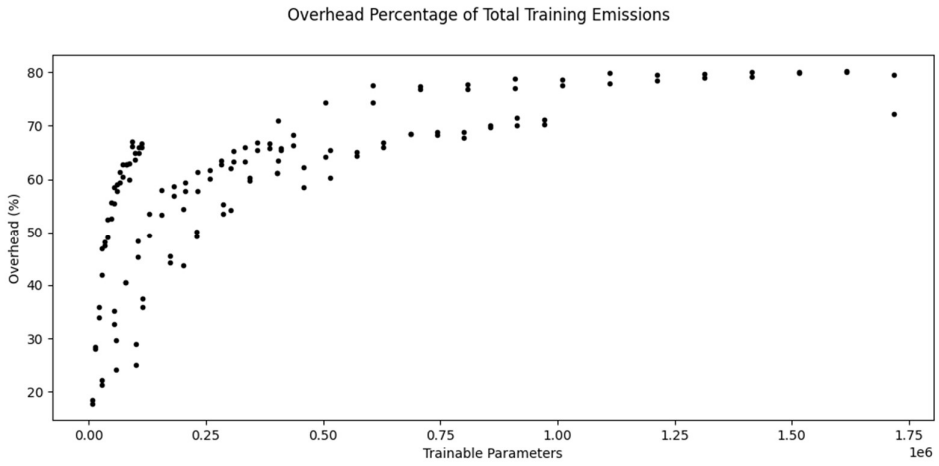

Figure 5. Overhead percentage

5. Conclusions

In this paper, we have examined theoretical frameworks and empirical studies to predict the carbon footprint of machine learning models. We ran our own experiments to test some of the statements in the literature.

Consolidated model: We created a consolidated model that describes the factors that influence the carbon footprint at each life cycle stage of a machine learning model. This model helps understand the benefit of various rule of thumbs to reduce carbon footprint (e.g., transfer learning or model distillation). However, it is not suitable for an exact prediction. The experiments have confirmed that there are no simple metrics and formulas that work correctly.

Experimental confirmation of literature: Many of the statements found in the literature could be confirmed by our experiments. Epochs and training set size are of vital importance. Trainable parameters are a basic indicator but only if comparing models that share many properties such as general shape and type of layers. However, we find that the arguments for transfer learning can be challenged and need further investigation.
Carbon footprint is a problem but often not a huge problem: While there are high-end models that use a lot of energy during architecture search and training, the typical use case will not have a significant impact in that regard. If the model is successful and sees a lot of use, the inference phase is more important than the initial phases. Thus, it would seem to be a good practice to optimize the inference stage once a good model is achieved. Since a single inference is not very expensive, tests can be run to understand the cost and to test reduction methods. Nevertheless, even though most models are not a problem at the moment, any concept that relies on an increasing carbon footprint for improvement ("*red AI*" [3]) needs close monitoring and corrective steps.
We hope that this paper enables ML practitioners as well as managers to choose the right decisions during the design and deployment of their models. Further, it should encourage more research into prediction models and reduction techniques for AI's carbon footprint.

6. References

[1] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis and K. Taha, "Efficient Machine Learning for Big Data: A Review.," *Big Data Research 2 (3),* p. 87–93, 2015.

[2] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia and D. e. a. Rothchild, "Carbon emissions and large neural network training.," in *In arXiv preprint arXiv:2104.10350.*, 2021.

[3] R. Schwartz, J. Dodge, N. A. Smith and O. Etzioni, " Green AI," *Commun. ACM 63 (12) ,* p. 54–63, 2020.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell and S. Agarwal, "Language models are few-shot learners," *Advances in neural information processing systems,* vol. 33, pp. 1877-1901, 2020.

[5] D. Patterson, J. Gonzalez, U. Hölzle, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. R. So, M. Texier and J. Dean, "The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink," *IEEE Computer,* vol. 55, no. 7, pp. 18-28, 2022.

[6] A. Lacoste, A. Luccioni, V. Schmidt and T. Dandres, "Quantifying the Carbon Emissions of Machine Learning: arXiv. Available online at doi:10.48550/ARXIV.1910.09700.," 2019.

[7] E. Strubell, A. Ganesh and A. McCallum, "Energy and Policy Considerations for Deep Learning," in *NLP: arXiv. Available online at doi:10.48550/ARXIV.1906.02243*, 2019.

[8] L. Lai, N. Suda and V. Chandra, "Not All Ops Are Created Equal!," in *SysML Conference*, Stanford, CA, USA, 2018.

[9] T.-J. Yang, Y.-H. Chen, J. Emer and V. Sze, "A method to estimate the energy consumption of deep neural networks," in *20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)*, 2020.

[10] A. Gupta, "Beyond Single-Dimensional Metrics for Digital Sustainability," [Online]. Available: https://greensoftware.foundation/articles/beyond-single-dimensional-metrics-for-digital-sustainability. [Accessed 10 07 2022].

[11] W. Yuxin, W. Qiang, S. Shaohuai, H. Xin, T. Zhenheng, Z. Kaiyong and C. Xiaowen, "Benchmarking the performance and energy efficiency of AI accelerators for AI training," in *20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID). IEEE, 2020.*, 2020 .

[12] K. Lottick, S. Susai, S. A. Friedler and J. P. Wilson, "Energy Usage Reports: Environmental awareness as part of algorithmic accountability," *arXiv,* 2019.

[13] R. Selvan, N. Bhagwat, L. F. W. Anthony, B. Kanding and E. B. Dam, "Carbon Footprint of Selecting and Training Deep Learning Models for Medical Image Analysis," *arXiv preprint arXiv:2203.02202,* 2022.

[14] L. Heim, A. Biri, Z. Qu and L. Thiele, " Measuring what Really Matters: Optimizing Neural Networks for TinyML," *arXiv,* 2021.

[15] L. Lannelongue, J. Grealey and M. Inouye, "Green Algorithms: Quantifying the carbon footprint of computation," *arXiv,* 2020.

[16] P. Henderson, J. Hu, J. Romof, E. Brunskill, D. Jurafsky and J. Pineau, "Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning.," *Journal of Machine Learning Research,* p. 1–43, 2020.

[17] X. Z. L. L. Y. W. a. W. S. Mohit Kumar, "Energy-Efficient Machine Learning on the Edges," in *IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, 2020.

[18] F. D. H., "Recent Advances in AI for Computational Sustainability," *IEEE Intelligent Systems,* vol. 31, no. 4, p. 56–61, 2016.

[19] T. Trader, "AWS to Offer Nvidia's T4 GPUs for AI Inferencing," 2019. [Online]. Available: https://www.hpcwire.com/2019/03/19/aws-upgrades-its-gpu-backed-ai-inference-platform/. [Accessed 17 08 2022].

[20] J. Barr, "Amazon EC2 Update – Inf1 Instances with AWS Inferentia Chips for High Performance Cost-Effective Inferencing," 2019. [Online]. Available: https://aws.amazon.com/blogs/aws/amazon-ec2-update-inf1-instances-with-aws-inferentia-chips-for-high-performance-cost-effective-inferencing/. [Accessed 17 08 2022].

[21] E. Jaureguialzo, "PUE: The Green Grid metric for evaluating the energy efficiency in DC (Data Center). Measurement method using the power demand," in *IEEE 33rd International Telecommunications Energy Conference (INTELEC)*, 2011.

[22] M. Horowitz, "1.1 Computing's energy problem (and what we can do about it)," *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC),* pp. 10-14, 2014.

[23] E. Cai, D.-C. Juan, D. Stamoulis and D. Marculescu, "NeuralPower: Predict and Deploy Energy-Efficient Convolutional Neural Networks," in *PMLR (Proceedings of Machine Learning Research, 77)*, Seoul, Korea, 2017.

[24] A. Gupta, "Why should sustainability be a first-class consideration for AI systems?," 2021. [Online]. Available: https://greensoftware.foundation/articles/why-should-sustainability-be-a-first-class-consideration-for-ai-systems. [Accessed 17 08 2022].

[25] D. J. J. G. O. J. F. a. J. G. Blalock, " What is the state of neural network pruning?," in *Proceedings of machine learning and systems 2* , 2020.

[26] G. Hinton, O. Vinyals and J. Dean, "Distilling the Knowledge in a Neural Network," *arXiv,* 2015.

[27] W. Paul, B. Jhilam, S. V. Saujanya, K. Vikrant, R. R. M and R. O. (Eds.), Sustainable AI in the Cloud: Exploring Machine Learning Energy Use in the Cloud, doi:10.1109/ASEW52652.2021.00058, 2021.

[28] T. Xu, "These simple changes can make AI research much more energy efficient (MIT Technology Review)," 2022. [Online]. Available: https://www.technologyreview.com/2022/07/06/1055458/ai-research-emissions-energy-efficient/?truid=&utm_source=the_download&utm_medium=e. [Accessed 17 08 2022].

[29] Z. Lu, H. Pu, F. Wang, Z. Hu and L. Wang, "The expressive power of neural networks: A view from the width," in *NIPS*, 2017.