Journal of Machine Learning Research 24 (2023) 1-23 Submitted 4/21; Revised 3/23; Published 5/23

A First Look into the Carbon Footprint of Federated Learning

Xinchi Qiu[⋆] xq227@cam.ac.uk
Titouan Parcollet[†,⋆] titouan.parcollet@univ-avignon.fr
Javier Fernandez-Marques[‡] javier.fernandezmarques@cs.ox.ac.uk
Pedro P. B. Gusmao[⋆] pp524@cam.ac.uk
Yan Gao[⋆] yg381@cam.ac.uk
Daniel J. Beutel[§,⋆] db849@cam.ac.uk
Taner Topal[§,⋆] taner@flower.dev
Akhil Mathur[▽] akhil.mathur@nokia-bell-labs.com
Nicholas D. Lane[⋆,§] ndl32@cam.ac.uk

[⋆]*Department of Computer Science and Technology, University of Cambridge*
15 JJ Thomson Ave, Cambridge CB3 0FD, United Kingdom

[†]*Laboratoire Informatique d'Avignon, Avignon Université*
339 Chemin des Meinajaries, 84000 Avignon, France

[‡]*Department of Computer Science, University of Oxford*
15 Parks Rd, Oxford OX1 3QD, United Kingdom

[§]*Flower Labs GmbH*
Winterhuder Weg 29, 7. Stock, 22085 Hamburg, Germany

[▽]*Nokia Bell Labs*
21 JJ Thomson Avenue, Cambridge CB3 0FA, United Kingdom

Editor: Qiaozhu Mei

Abstract

Despite impressive results, deep learning-based technologies also raise severe privacy and environmental concerns induced by the training procedure often conducted in data centers. In response, alternatives to centralized training such as Federated Learning (FL) have emerged. FL is now starting to be deployed at a global scale by companies that must adhere to new legal demands and policies originating from governments and social groups advocating for privacy protection. *However, the potential environmental impact related to FL remains unclear and unexplored. This article offers the first-ever systematic study of the carbon footprint of FL.* We propose a rigorous model to quantify the carbon footprint, hence facilitating the investigation of the relationship between FL design and carbon emissions. We also compare the carbon footprint of FL to traditional centralized learning. Our findings show that, depending on the configuration, FL can emit up to two orders of magnitude more carbon than centralized training. However, in certain settings, it can be comparable to centralized learning due to the reduced energy consumption of embedded devices. Finally, we highlight and connect the results to the future challenges and trends in FL to reduce its environmental impact, including algorithms efficiency, hardware capabilities, and stronger industry transparency.

Keywords: federated learning, carbon footprint, energy analysis, green AI, on-device AI

©2023 Xinchi Qiu, Titouan Parcollet, Javier Fernandez-Marques, Pedro P. B. Gusmao, Yan Gao, Daniel J. Beutel, Taner Topal, Akhil Mathur, Nicholas D. Lane.

License: CC-BY 4.0, see `https://creativecommons.org/licenses/by/4.0/`. Attribution requirements are provided at `http://jmlr.org/papers/v24/21-0445.html`.

1. Introduction

Atmospheric concentrations of carbon dioxide, methane, and nitrous oxide are at levels not seen in the last $800,000$ years (IPCC, 2014). Together with other anthropogenic drivers, their effects have been detected throughout a network of distributed systems and are extremely likely to have been the dominant cause of the observed global warming since the mid-20^{th} century (Pachauri et al., 2014; Crowley, 2000). Unfortunately, deep learning (DL) algorithms keep growing in complexity, and numerous "state-of-the-art" models continue to emerge, each requiring a substantial amount of computational resources and energy, resulting in clear environmental costs (Strubell et al., 2019). Indeed, these models are routinely trained for thousands of hours on specialized hardware accelerators in data centers that are extremely energy-consuming (Berriel et al., 2017). As Amodei and Hernandez (2018) showed, the amount of computing used by the largest machine learning (ML) training has been exponentially increasing and grown by more than $300,000\times$ from 2012 to 2018, which is equivalent to a 3.4-months doubling period – a rate that dwarfs the well-known Moore's 2-year doubling period. Even though the amount of energy per FLOPS has been exponentially decreasing over time, making the deep learning model more and more computationally efficient, the carbon footprint of ML models is still one of the big concerns in society.

The data centers that enable DL research and commercial operations are not often accompanied by visual signs of pollution. In a few isolated cases, they are even powered by environmentally friendly energy sources (Google, 2020b; AWS, 2020). Still, they are responsible for an increasingly significant carbon footprint. Each year data centers use 200 terawatt-hours (TWh), which is more than the national electricity consumption of some countries, representing 0.3% of global carbon emissions (Nature, 2018; Andrae and Edler, 2015). In comparison, the entire information and communications technology ecosystem accounts for 2%. To put this issue in a more human perspective, each person on average on the planet is responsible for 5 tonnes of emitted CO_2-equivalents (CO_2e) per year (Strubell et al., 2019), while training a large Natural Language Processing (NLP) transformer model with neural architecture search may produce 284 tonnes of CO_2e (Strubell et al., 2019). Even for smaller deep neural networks and routine research experiments, Parcollet and Ravanelli (2021) demonstrated that the training process necessary to create a state-of-the-art speech recognizer could produce more than 0.1 tonnes of CO_2e with consumer-grade hardware. Even though the number refers to one of the largest ML models, given the increasing interest in Large Language Models (LLM), it is likely that this trend will continue and possibly expand to tasks besides NLP. Understanding the carbon footprint of ML training will play a paramount role in allowing people to develop more carbon-efficient models and hardware, making the emission more transparent, and choosing renewable energy where possible.

Decentralized alternatives to a data center-based DL and other forms of machine learning are emerging. Among these, the most prominent to date is *Federated Learning (FL)*, first formalized by McMahan et al. (2017). Under FL, training of models primarily occurs in a distributed scenario, either across a large number of personal devices (*cross-device*), such as smartphones; or across a small number of institutions that cannot share data among themselves (*cross-silo*), such as private hospitals. Devices collaboratively learn a global model but do so without uploading to a data center any of the locally stored sensitive data. Then they send the locally trained models to a central server, where models get aggregated following a strategy such as FedAVG (McMahan and Ramage, 2017; Kairouz et al., 2019; Konečnỳ et al., 2015). While FL is still a maturing technology, it is already being used by millions of users on a daily basis; for example, Google uses FL to train models for: predictive keyboard, device setting recommendation, and hot keyword personalization on phones (McMahan and Ramage, 2017).

At present, data owners are holding more and more sensitive information, such as individual activity data, life-logging videos, email conversations, and others (Nishio and Yonetani, 2019), so keeping personal medical and healthcare data private recently became one of the major ethical concerns (Kish and Topol, 2015). To this extent, and in response to an increasing number of such privacy issues, policy-makers have responded with the implementation of data privacy legislation such as the European General Data Protection Regulation (GDPR) (Lim et al., 2020). Due to these regulations, moving data across national borders becomes subject to data sovereignty law, making centralized training infeasible in some scenarios (Hsieh et al., 2020).

Furthermore, there are nearly seven billion connected Internet of Things (IoT) devices (Lim et al., 2020) and three billion smartphones around the world, potentially giving access to an astonishing amount of training data and decentralized computing power for meaningful research and applications. sing mobile sensing and smartphones to boost large-scale health studies, such as in Pryss et al. (2015), Barrett et al. (2020) and Shen (2015), has caused increased interest in the healthcare research field, and privacy-friendly framework including FL are potential solutions to answer this demand.

Despite FL privacy being under great scrutiny from the scientific community, we currently have little to no understanding of its impact on carbon emissions. This is a worrying situation, given the increasing interest in this technology. Therefore, the carbon footprint of FL needs to be assessed before vast systems are further deployed.

Whilst the carbon footprint for centralized learning has been studied in many previous works (Anthony et al., 2020; Lacoste et al., 2019; Henderson et al., 2020; Uchechukwu et al., 2014), the energy consumption and carbon footprint related to FL remains virtually unexplored. This article provides the key step in attempting to fill this void by giving a first look into the carbon analysis of FL. It expands upon our initial treatment of the area (Qiu et al., 2021) with a more comprehensive study; our original paper, and this article, have also prompted significant subsequent investigations within the community (Kim and Wu, 2021; Savazzi et al., 2023; Pilla, 2023). Studies of this kind are essential because state-of-the-art results in deep learning are usually determined by metrics such as the accuracy of a given model or model size, while energy efficiency is often overlooked. Whilst accuracy remains crucial, we hope to encourage researchers to also focus on other metrics that are in line with the increasing societal global warming awareness. Recent research (Patterson et al., 2022) indicates the approaches to reduce energy and carbon emissions in centralized training in data centers. By quantifying carbon emissions for FL and demonstrating that very specific FL setups may lead to a decrease of these emissions, we encourage the integration of the released CO_2e as a crucial metric to the FL deployment. The scientific contributions of this work are as follows:

- **Analytical Carbon Footprint Model for FL.** We provide the first quantitative CO_2e emissions estimation method for FL (Section 3), including emissions resulting from both hardware training and communication between server and clients.

- **Extensive Experiments.** Carbon sensitivity analysis is conducted with this method on real FL hardware under different settings, strategies, and tasks (Section 4). We demonstrate that CO_2e emissions depend on a wide range of hyper-parameters and that emissions derived from communication between clients and server can represent from 0.7% up to more than 96% of total emission. When compared to centralized training, we show that for different tasks and settings, FL can emit from 72% to hundreds of times more carbon than its centralized version.

- **Analysis and Roadmap towards Carbon-friendly FL.** We provide a comprehensive analysis and discussion of the results to highlight the challenges and future research directions in developing carbon-friendly federated learning (Section 5).

2. Federated Learning Background

Traditional machine learning involves using a central server that hosts the machine learning models and all the data in one place. In contrast, in FL frameworks client devices collaboratively learn a shared global model using their own local data. FL has distinct privacy advantages over centralized training as the data are not transferred to the central server for training. In fact, the only information transferred from clients to the server is their respective updated model parameters obtained after each local training. To further limit the leakage of client's information in the model update, several mechanisms have been proposed over the years including Secure Aggregation (Bonawitz et al., 2016) and Differential Privacy (McMahan et al., 2018).

FL training occurs over multiple communication rounds. During each round, a fraction of the clients are selected and receive the global model from the server. Those selected clients then perform local training with their local data before sending the updated models back to the central server. Finally, the central server aggregates these updated models, resulting in a new global model. Then, this three-stage process is repeated for a fixed number of rounds.

There exists several aggregation strategies targeting to solve different FL problems. The most widely adopted one is FedAvg (McMahan et al., 2017), in which the central server aggregates the models by performing a weighted sum of the received parameters based on the number of samples in each local dataset. More advanced strategies inspired by adaptive momentum-based gradient descent optimizers have also been proposed e.g., FedADAM (Reddi et al., 2021).

In addition, FL settings can be classified as either *cross-silo* or *cross-device.* In a *cross-silo* scenario, clients are generally few, with high availability during all rounds, and are likely to have similar data distribution for training, e.g. consortium of hospitals. This scenario serves as motivation to consider Independent and Identically Distributed (IID) distributions. On the other hand, a *cross-device* system will likely encompass thousands of clients having very different data distributions (non-IID) participating in just a few rounds, e.g. training of next-word prediction models on mobile devices. In practice, non-IID datasets not only means class imbalance, but also feature imbalanced among clients. Indeed, many latent factors can change such as the voice timbre in speech recognition (Gao et al., 2022).

3. Quantifying CO_2e emissions

Two major steps can be followed to quantify the carbon footprint of training deep learning models either in data centers or on the edge. First, we perform an analysis of the energy required by the method (Section 3.1), mostly accounting for the total amount of energy consumed by the hardware. It includes training energy for centralized learning and training and communication energy for FL (Section 3.2). Then, the latter amount is converted to CO_2e emissions (Section 3.3) based on geographical locations which, as it will be presented, vary significantly depending on the sources of energy. This study does not include emissions related to hardware manufacturing as such information is still largely unavailable.

3.1 Training Energy Consumption

First, we consider the energy consumption coming from GPU and CPU, which can be measured by sampling GPU and CPU power consumption at training time (Strubell et al., 2019). For NVIDIA-based hardware, we can repeatedly query the NVIDIA System Management Interface (NVIDIA-smi) to sample the GPU power consumption and report the average over all processed samples while training. In the context of FL, not all clients are equipped with a GPU, and this part can thus be removed from the equation if necessary. To this extent, we propose to consider e_{clt} as the power of a single client combining both GPU and CPU measurements. Then, we can connect these measurements to the total training time of the model. We define $T_{FL}(e, N, R)$ to be the total training energy consumption consisting of a total of N clients in the pool with hardware power e for a total of R rounds in FL setup:

$$\mathrm{T}_{FL}(e, N, R) = \sum_{j=1}^{R} \sum_{i=1}^{N} \mathbb{1}_{\{Clt_{i,j}\}} \cdot t_i \cdot e_{client,i}, \tag{1}$$

where $\mathbb{1}_{\{Clt_{i,j}\}}$ is the indicator function indicating if client i is chosen for training at round j, t_i the wall clock time per round and $e_{clt,i}$ the power of client i.

Hardware components, such as system memory and storage, are also responsible for energy consumption. According to Hodak et al. (2019), one may expect a variation of around 10% while considering these parameters. However, they are also highly dependent on the infrastructure considered and the device distribution that is unfortunately unavailable. We exclude the energy costs of powering such components since they account for a small portion of the total energy consumption during training.

The particular case of cooling in centralized training. Cooling in data centers accounts for up to 40% of the total energy consumed (Capozzoli and Primiceri, 2015). While this parameter does not exist for FL, it is crucial to consider it when estimating the cost of centralized training. Such estimation is particularly challenging as it depends on the data center efficiency. To this extent, we consider the use of Power Usage Effectiveness (PUE) ratio. As reported in the *2019 Data Center Industry Survey Results* (UptimeInstitue, 2019), the world average PUE for the year 2019 is 1.67. As expected, observed PUE strongly varies depending on the considered company. For instance, *Google* declares a comprehensive trailing twelve-month PUE ratio of 1.11 (Google, 2020a) compared to 1.2 and 1.125 for *Amazon* (AWS, 2020) and *Microsoft* (Microsoft, 2015) respectively. We also report a PUE ratio of a University-scale cluster (Avignon University, France) as an example. The PUE ratio is reported to be 1.55 for a cluster containing 17 computing nodes with 4 to 8 GPUs each. Therefore, Eq. (1) is adapted to centralized training setting as:

$$\mathrm{T}_{center} = \mathrm{PUE} \cdot (t \cdot e_{center}), \tag{2}$$

with e_{center} representing the power combining both GPUs and CPUs in a centralized training setup, and t stands for the total training time.

3.2 Wide-area-networking (WAN) Emission

As clients continue to perform individual training on local datasets, their models begin to diverge. To mitigate this effect, model aggregation must be performed by the server in a process that requires frequent exchange of models between clients and the server.

According to Malmodin and Lundén (2018), the embodied carbon footprint for Information and Communication Technology (ICT) network operators is mainly related to the construction and deployment of the network infrastructure including digging down cable ducts and raising antenna towers.

Regarding FL, we estimate the energy required to transferring model parameters between the server and the clients following two parts. The first part is the energy consumed by routers throughout the FL communication process, while the second part is the energy consumed by the hardware when downloading and uploading the model parameters. We propose to use country-specific download and upload speed as reported on *Speedtest* (Speedtest) and router power reported on *The Power Consumption Database* (Database). Due to the rapid development of ICTs, we propose to use the median power obtained from all data submitted during 2021 to the database. We also take idle power consumption of hardware into consideration while they are communicating model parameters. Let us define D and U the download and upload speeds expressed in Mbps respectively. The communication energy per round is defined as:

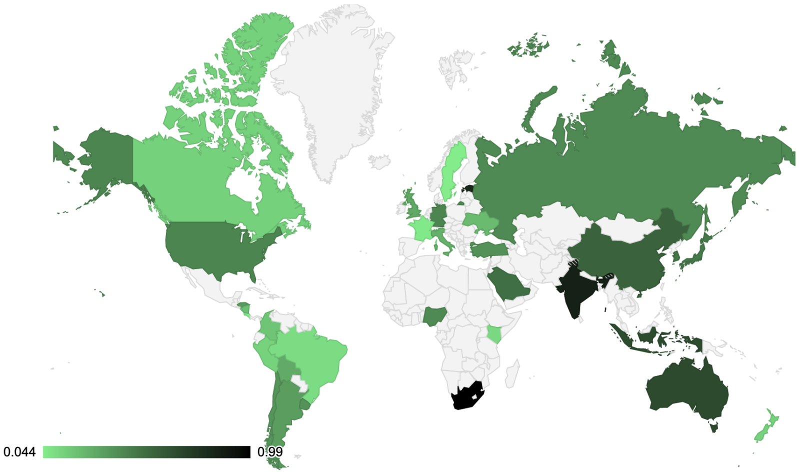
$$\mathrm{C}(e, N, R) = \sum_{j=1}^{R} \sum_{i=1}^{N} \mathbb{1}_{\{Clt_{i,j}\}} \cdot S \cdot \left(\frac{1}{D} + \frac{1}{U}\right) \cdot (e_r + e_{idle,i}), \tag{3}$$

with S the size of the model in Mb, e_r the power of the router, and e_{idle} the power of the hardware of the idle clients.

3.3 Converting to CO$_2$e emissions

Realistically, it is challenging to compute the exact amount of CO$_2$e emitted in a given location since the information regarding the energy grid, *i.e.*, the conversion rate from energy to CO$_2$e, is rarely publicly available (Lacoste et al., 2019; Hodak et al., 2019). Therefore, we assume that all data centers and the edge devices are connected to their local grid directly linked to their physical location. Electricity-specific CO$_2$e emission factors are obtained from official governmental websites and reports. Out of all these conversion factors expressed in $kg\ CO_2e/kWh$, we picked three of the most representative ones averages over a one year-period: Australia (0.656)[1], the United Kingdom (0.281) [2] and France (0.054)[3]. The estimation methodology provided takes into accounts both transmission and distribution emission factors (*i.e.* energy lost when transmitting and distributing electricity) and the efficiency of power plants. As expected, countries relying on carbon-efficient productions are able to lower their corresponding emission factor (*e.g.* France, Canada). A heatmap demonstrating different levels of conversion rates in various countries can be found in Fig. 1.

Figure 1: Global heat map of electricity to CO_2e conversion rate (in kg/kWh). The conversion rate are obtained from governmental sources or on the website Climate Transparency [4]



Therefore, the total amount of CO_2e emitted in kilograms for FL (E_{FL}) and centralized training (E_{center}) obtained from Eq. 1, 2 and 3 are:

$$\mathrm{E}_{FL} = \mathrm{c}_{rate} \cdot [\mathrm{T}(e, N, R) + \mathrm{C}(e, N, R)], \tag{4}$$

$$\mathrm{E}_{center} = \mathrm{c}_{rate} \cdot \mathrm{T}_{center}, \tag{5}$$

where c_{rate} is the conversion rate factor. It is worth noticing that when dealing with non-IID partitions, the total training energy consumption ($\mathrm{T}(e, N, R)$) and energy for communication cost ($\mathrm{C}(e, N, R)$) will often be larger than IID partitions, as it usually requires larger number of communication rounds to reach certain model performance Karimireddy et al. (2020); Zhao et al. (2018), and it is also shown in our experiments later in Section 4. In general, c_{rate} will depend on the physical location of the hardware where the training takes place, and it is possible that c_{rate} is not unique across the FL settings as clients can be scattered around the globe. We will need to adjust the c_{rate} for each client based on their physical locations. In our experiments, we assume that all FL clients are located at the same physical locations for ease of comparison.

Carbon emissions may be compensated by carbon offsetting or with the purchases of Renewable Energy Credits (RECs, in the US) or Tradable Green Certificates (TGCs, in

[1] source:https://www.climate-transparency.org/countries/asia/australia
[2] source:https://www.climate-transparency.org/countries/europe/the-united-kingdom
[3] source:https://www.climate-transparency.org/countries/europe/france

the EU). Carbon offsetting allows polluting actions to be mitigated directly via different investments in *carbon-friendly* projects, such as renewable energies or massive tree planting (Anderson, 2012). RECs and TGCs (Bertoldi and Huld, 2006), on the other hand, guarantee that specifics volumes of electricity are generated from renewable energy sources. However, in our analysis, carbon rates are obtained at country level and do not integrate industry level carbon offsetting schemes or RECs.

4. Experiments

This article provides extensive estimates across different types of tasks and datasets, including image classification with CIFAR10 (Krizhevsky et al., 2009), FEMNIST (LeCun, 1998; Cohen et al., 2017), and ImageNet (Russakovsky et al., 2015), speech processing with keyword spotting on Speech Commands (Warden, 2018) and speech recognition with CommonVoice (Ardila et al., 2020). First, we provide an estimate of the carbon footprint following different realistic FL setups. Then, we conduct an in-depth analysis of these results to highlight the differences observed.

4.1 Experimental Protocol

Experiments are built on top of PyTorch (Paszke et al., 2019) and SpeechBrain (Ravanelli et al., 2021). We make use of the Flower framework (Beutel et al., 2020) to implement and parameterized different FL training pipelines. In addition to the carbon model (Section 3), results are influenced by configurations of the hardware and systems of datacenters and FL respectively.

Centralized training hardware. We run our experiments on a server equipped with two Xeon 6152 22-core processors and NVIDIA Tesla V100 32GB GPUs. The CPU and GPU have TDP of 240W and 250W, respectively. We use a single GPU per experiment and measure the power drawn by both CPU and GPU through *nvidia-smi* monitoring and the cross-platform *psutil* tools.

Federated learning hardware. We consider the use of NVIDIA Tegra X2 (Smith, 2017) and Jetson Xavier NX (Smith, 2019) devices as our FL clients. These devices can be viewed as a realistic pool of FL clients since they can be found embedded in various IoT devices including cars, smartphones, and video game consoles. NVIDIA Tegra X2 offers two power modes with theoretical power limits of $7.5W$ and $15W$ and Xavier NX offers $10W$ and $15W$. Across our different runs, we use the lower power mode for each device, and we employ the built-in utility *tegrastats* to report the overall power consumption. For both power consumption and training time, we report averaged values across several FL rounds for each experiment. We also measure the idle power consumption for both devices, which was recorded as $1.35W$ and $2.25W$ for TX2 and NX respectively.

Datasets. We conduct our estimations on three image classification tasks of different complexity both in terms of the number of samples and number of classes: CIFAR10, FEMNIST, and ImageNet. FEMNIST, federated extended MNIST, is built by partitioning the data in Extended MNIST (EMNIST) according to writer ids. It contains 671K 28×28 images of digits and letters. In addition, we also perform analysis on speech processing

[4] Climate Transparency: https://www.climate-transparency.org/

following the same complexity concern with the Speech Commands dataset for keyword spotting and Common Voice for automatic speech recognition (ASR). Speech Commands contains $65K$ 1-second long audio clips of 30 keywords, with each clip consisting of only one keyword. Following the setup described in Zhang et al. (2017), we train the model to classify the audio clips into one of the 10 keywords - "Yes", "No", "Up","Down", "Left", "Right", "On", "Off", "Stop", "Go", along with "silence" (*i.e.* no word spoken) and "unknown" word, representing the remaining 20 keywords from the dataset. The training set contains a total of $56,196$ clips with $32,550$ (57%) samples from the "unknown" class and around 1800 samples (3.3%) from each of the remaining classes, hence the dataset is naturally unbalanced. Also, we used the Common Voice Italian (CV Italian) dataset (version 6.1) containing a total of 84K utterances (132 hours) which were recorded by more than 10K Italian-speaking participants. The train set consists of 748 speakers (89 hours of speech), while both valid and test sets contain around 22 hours of speech from 1219 and 3404 speakers respectively.

Model Architectures. For CIFAR10 and ImageNet we make use of ResNet-18 (He et al., 2016). For FEMNIST, we choose a much shallower CNN as proposed by (Caldas et al., 2018). These architectures are kept the same for both centralized and FL experiments. These models are trained with SGD but only the centralized setting makes use of momentum. For the sake of completeness, we choose to use different deep learning model for the Speech Commands dataset. We employ 4 layers of LSTM each with 256 nodes. The models are trained using *Adam* optimization. Also, the hyper-parameters, such as learning rates, are set to be the same as centralized learning without further tuning. For ASR task on CV Italian dataset, the experiments are based on a encoder-decoder model trained with the joint connectionist temporal classification (CTC)-attention objective (Kim et al., 2017). A typical ASR model includes three modules: the encoder, the decoder and the attention mechanism. The encoder has the following architecture: CNN — LSTM — DNN, and the decoder is a single hidden layer GRU. Models are jointly trained with CTC and cross entropy (CE) loss. Note that the federated training for ASR task starts from a pre-trained initialized model since all the existing FL aggregation methods fail to converge without pre-training (Gao et al., 2022; Dimitriadis et al., 2020).

Data partition methodology. As mentioned in Section 2, FL settings can usually be classified as *cross-silo* or *cross-devices*. In *cross-silo* settings, data distribution in each client will be the same as the global data distribution, hence training energy should be very close to centralized training with additional communication cost. In this work, we focus the experiments on *cross-device* settings, and the IID partition provides the best-case scenarios and the baselines for comparison between centralized and FL settings.

We simulate different level of non-IID data distribution following the latent Dirichlet allocation (LDA) partition method (Reddi et al., 2021; Yurochkin et al., 2019; Hsu et al., 2019) ensuring that each client gets allocated the same number of training samples. Each sample is drawn independently with class labels following a categorical distribution over N classes parameterized with a vector $\mathbf{q}$ ($q_i \geq 0$, $i \in [1, m]$ and $\sum q_i = 1$ for a total of m classes from the dataset). Thus, to simulate the partition, we draw $\mathbf{q} \sim Dir(\alpha \mathbf{p})$ from a Dirichlet distribution, where $\mathbf{p}$ stands for the prior distribution of the dataset, and α stands for the concentration which controls the level of heterogeneity of the partition. As $\alpha \rightarrow \infty$, the partition becomes more uniform (IID), and as $\alpha \rightarrow 0$, the partition becomes more heterogeneous. As the dataset is balanced across classes for both CIFAR10 and ImageNet, the prior distribution $\mathbf{p}$ is uniform. For ImageNet, we chose $\alpha = 1000$ for the IID dataset partition and $\alpha = 0.5$ for non-IID following Yurochkin et al. (2019) and Hsu et al. (2019). For CIFAR10, we choose $\alpha = 0.1$ following the same protocol as Reddi et al. (2021). As for Speech Commands, in light of the unbalanced nature of the dataset, we propose to change the prior of LDA from uniform distribution to multinomial distribution. Hence the LDA can be summarized as:

$$\mathbf{p} = \left(\frac{N_1}{N}, \frac{N_2}{N}, ..., \frac{N_m}{N}\right) \tag{6}$$

$$\mathbf{q} \sim Dir(\alpha\mathbf{p}), \tag{7}$$

where N_i stands for the number of data from class i, N stands for total number of data in the dataset. According to Yurochkin et al. (2019); Hsu et al. (2019), α is commonly set as 0.5 for a non-IID partition of balanced dataset. Given the aforementioned unbalanced nature of the dataset, we propose to match the variance of 10 keywords classes with multinomial prior to the variance of 10 keywords classes with a uniform prior by changing α to 1.0.

In practice, a non-IID dataset can mean both class-imbalance and feature-imbalance among clients. Other latent factors can change such as the user accent or voice timbre in speech recognition or different calligraphy styles in hand-written text. Therefore, we also include two naturally partitioned datasets FEMNIST and CV Italian to capture the feature imbalanced datasets.

For CV Italian, we first pre-train the model on half of the data samples in a centralized fashion. We do this by partitioning the original dataset into a small subset of speakers (99) for centralized training and a larger subset of speakers (649) for the FL experiment. Then, we simulate a scenario of single speaker using their individual devices by naturally dividing the training sets based on users ID into 649 partitions. We followed the paritioning methodology in Caldas et al. (2018) to extract the FEMNIST dataset from EMNIST following a natural partitioning by writer id.

Client pool. Following Reddi et al. (2021), we consider a pool of 500 client for CIFAR10 with 10 active clients training concurrently per round.We split ImageNet and SpeechCommands into 100 clients and randomly select 10 clients per round. As for FEMNIST and CV Italian, there are 3597 and 649 natural clients respectively, and we select 35 and 10 clients in each communication round.

FL strategy. To better reflect realistic FL scenarios, we propose to investigate the energy consumption with the common FedAVG strategy (McMahan et al., 2017), and the more complex FedADAM strategy (Reddi et al., 2021). For CIFAR10, we follow the experimental protocol proposed in Reddi et al. (2021) considering the suggested best values for η, η_l, and τ in almost every experiment except for FedAVG, where we had to lower the value of η_l to $10^{-3/2}$ to allow training. All other experiments used a server learning rate $\eta = 0.1$ and $\tau = 0.001$.

Local epoch (LE). We also propose to vary the number of local epochs done on each client to better highlight the contribution of the local computations to the total emissions. To be consistent, we choose to do 1 and 5 local epochs across all tasks except ASR task (insisting with 5 local epochs to obtain acceptable performance).

Target accuracies. To make fair comparisons between different setups, we set the target accuracies for each tasks and report the respective carbon emission. This is a common procedure when evaluating FL workloads. We set the target accuracies for CIFAR10, FEMNIST and ImageNet to be 70%, 80% and 50% top-1 accuracy respectively. For Speech Commands, the threshold is set to 70%, and for CV Italian, the target is set to be 25% of Word Error Rate (WER).

4.2 Experimental Results

This section presents the experimental results. Power consumption and training times obtained for all FL and centralized setups are reported in Table 1. Table 1 also shows the power measurement and energy consumption for each setups. Both power usage and training time per epoch reflect the mean value for each training tasks. The total energy is calculated as the energy per device multiplied by the number of selected clients per communication round for FL. In the centralized scenario it is equal to the energy per device. The numbers of communication rounds required by each setup to reach their target accuracies are summarized in Table 2. Table 3 shows the carbon emission for each training task in every experimental setups, calculated by adding the energy consumption for communication and convert the energy consumption to carbon emission by multiplying the country-specific conversion factor as explained in Eq (4) and Eq (5).

As shown in Table 1, it is worth noting that centralized training (V100) took solely 2 epochs to achieve the target accuracy for CIFAR10, and 8 epochs for ImageNet, 1 for FEMNIST and 10 for CV Italian. This translates to 48 seconds for CIFAR10, 3.2 hours for ImageNet, 19 seconds for FEMNIST and 1.4 hours for CV Italian.

Table 2 reports the numbers of communication rounds required by each setup to reach their target accuracies. We can see that standard FedAVG failed to converge within the allotted 2000 rounds in the non-IID setting when using only 1 local epoch for CIFAR10, while the more sophisticated FedADAM strategy was able to reach the target. For Speech Commands experiments, it is interesting that FedAVG needs even more rounds for IID than non-IID if we only do one local epoch, which might be due to the dataset being naturally unbalanced. Similar as FEMNIST, CV Italian is a naturally partitioned dataset, so there only exist non-IID results. As settings with only 1 local epoch does not converge, we only show settings with 5 local epochs in the tables.

From Table 3 we can see that for image classification task (CIFAR10, ImageNet and FEMNIST) we observe the centralized settings generally consume less energy compared to their FL counterparts. The difference is the biggest when we compare CIFAR10 non-IID with 1 local epoch settings with centralized training. In this comparison, FL emits more than 10 times more carbon than centralized training. The difference is smaller when we perform 5 local epochs in FL. However, for ImageNet, the outcome is the other way around. FL with 5 local epochs emits more carbon compared to 1 local epoch settings. The difference between FL and centralized training for ImageNet is smaller than CIFAR10. For FEMNIST, as the dataset is naturally partitioned, there are only non-IID results. Similar as CIFAR10, 1 local epoch settings emits more carbon compared with 5 local epochs settings, and they both more emits higher carbon compared to centralized training. More surprisingly is the slower convergence rate of FedADAM for CIFAR10 and ImageNet to reach the specified target accuracy. However, FedADAM often performed better in the longer term resulting in higher final accuracies. For Speech Commands experiments, 3 also highlights the setups

						Costs to Reach Threshold Accuracy			
Dataset	**Training Strategy**	**HW**	**Power Usage (W)**	**Local Epochs**	**Time per Epoch(s)**	**Num. Rounds**	**Time(s)**	**Energy per device (Wh)**	**Total Energy (Wh)**
CIFAR10	Centralized	V100	160+42	1	24	2	48	2.7	**2.7**
	FedAVG	TX2	4.7	5	0.8	580	2320	3.03	30.3
		NX	6.3		0.6		1740	3.05	30.5
	FedAdam	TX2	4.7	1	0.8	1800	1440	1.88	18.8
		NX	6.3		0.6		1080	1.89	18.9
ImageNet	Centralized	V100	220+84	1	1,440	8	11,520	973	**971**
	FedAVG	TX2	6.5	1	474	339	160,686	290	2,901
		NX	9.7		273		92,547	249	2,494
	FedAdam	TX2	6.5	1	474	590	279,660	504	5,049
		NX	9.7		273		161,070	434	4,340
FEMNIST	Centralized	V100	96+20	1	19	1	19	0.6	**0.6**
	FedAVG	TX2	2.4	1	0.24	205	29	0.03	1.1
		NX	2.7		0.15		18	0.02	0.8
	FedADAM	TX2	2.4	1	0.24	60	14	0.01	0.3
		NX	2.7		0.15		9	0.007	0.2
Speech Commands	Centralized	V100	68+56	1	52	6	312	10.7	10.7
	FedAVG	TX2	5.7	5	1.6	140	1,120	1.8	17.7
		NX	7.9		0.9		630	1.4	13.8
	FedAdam	TX2	5.7	1	1.6	193	309	0.5	4.9
		NX	7.9		0.9		174	0.4	**3.8**
CV Italian	Centralized	V100	170 + 48	1	509	10	5090	308.2	**308**
	FedAVG	TX2	6.7	5	76	50	19,000	35.4	354
		NX	9.8		48		12,000	32.7	327

Table 1: Energy consumption of centralized training using GPUs against FL settings where each client trains on a small dataset partition using low-power GPU-enabled edge devices. For FL rows, each strategy reports the lowest total energy among 1 and 5 local epochs for non-IID partitions. For centralized setting, one "Local Epoch" is one standard epoch using the entire dataset and, "Power Usage" is reported as GPU+CPU. The "Time" column reports the total training time required, which is calculated by multiplying the "Time per Epoch" and the "Number of Rounds". For FL rows, the "Total Energy" is obtained by multiplying the "Energy per Device" by the number of clients participating in each round. Despite edge devices consuming an order of magnitude less power, the total energy required for FL is often greater (but of the same order of magnitude) than centralized training. For Speech Commands, a very lightweight workload, FL can reach the target accuracy while requiring little energy. For datasets with variable amount of data per client (e.g. FEMNIST, CV Italian), we report the time taken to train a client that contains the average data samples observed in the whole dataset.

when FL emits less carbon compared to centralized training, which happens in France when FL performs 5 local epochs. For the CV Italian experiments, it is worth noticing that all FL settings emits less carbon when compared with centralized training in the data centers with

Dataset	Training Strategy	Local Epochs	Partition	
			IID	non-IID
CIFAR10	FedAVG	1	480	>2000
		5	180	580
	FedAdam	1	580	1800
		5	250	800
ImageNet	FedAVG	1	232	339
		5	95	114
	FedAdam	1	550	590
		5	180	200
FEMNIST	FedAVG	1	-	205
		5	-	120
	FedAdam	1	-	60
		5	-	40
Speech Commands	FedAVG	1	>1000	770
		5	119	140
	FedAdam	1	140	193
		5	53	66
CV Italian	FedAVG	5	-	50

Table 2: Number of FL rounds needed for each dataset-strategy pair to reach the target accuracy when data is partitioned in IID and non-IID fashion. Note that there is only non-IID partition for FEMNIST and CV Italian, as both datasets are naturally partitioned. We observe that increasing the number of local epochs always results in fewer FL rounds to reach convergence. However, this does not guarantee a smaller overall energy consumption.

the averaged PUE ratio of 1.67. It is even less than centralized training in the data centers with PUE ratio of 1.55 in France.

5. Carbon Footprint of Federated Learning

5.1 CO$_2$e Analysis

So far we have considered the energy required to achieve a given accuracy on different tasks for various sets of hyper-parameters and optimizers. We now turn our attention to how this translates into carbon emissions.

The first thing to notice is that there are some settings with Speech Commands and CV Italian where FL emits slightly less carbon compared with centralized training. For Speech Commands, the model architecture is light-weighted, hence both training using TX2 and NX consumed much less energy, as shown in Table 1. Since the model only has 5.3 million parameters, communication did not consume much energy either. As for CV Italian, training energy for FL and centralized is about the same as shown in Table 1. Since the process only requires 50 communication rounds to reach our target accuracy and because we need to take into account the PUE ratio for data centers, the overall carbon emission for FL, in this specific scenario, can be lower than centralized training. Therefore, emissions from centralized and federated learning can be more comparable when using lightweight models, typically of cross-device setups.

Due to the large difference between electricity-specific CO$_2$e emission factors among countries, the carbon footprint of both centralized training and FL can be highly dependent on the geolocation of hardware. Training in France always has the lowest CO$_2$e emissions given their use of nuclear energy with the lowest energy to CO$_2$e conversion rate. Geolocation

CIFAR10	**Centr.**			**IID 5LE**				**non-IID 1LE**				**non-IID 5LE**			
Country/	**PUE**			**FedAVG**		**FedADAM**		**FedAVG**		**FedADAM**		**FedAVG**		**FedADAM**	
CO$_2$e(g)	1.67	1.55	1.11	TX2	NX	TX2	NX	TX2	NX	TX2	NX	TX2	NX	TX2	NX
Australia	3.0	2.7	**2.0**	70.6	78.1	98.1	108.5	>730	>813	656.7	731.6	227.5	251.7	313.8	347.2
UK	1.3	1.2	**0.8**	29.4	32.5	40.8	45	>303	>337.7	272.8	303.9	94.7	104.8	130.6	144.5
France	0.2	0.2	**0.2**	2.1	2.3	3.0	3.2	>19	>21	17.4	19.3	6.9	7.5	9.5	10.4

ImageNet	**Centr.**			**IID 5LE**				**non-IID 1LE**				**non-IID 5LE**			
Country/	**PUE**			**FedAVG**		**FedADAM**		**FedAVG**		**FedADAM**		**FedAVG**		**FedADAM**	
CO$_2$e(g)	1.67	1.55	1.11	TX2	NX	TX2	NX	TX2	NX	TX2	NX	TX2	NX	TX2	NX
Australia	1066	989	**708**	2701	2330	5117	4415	2025	1771	3524	3083	3241	2796	5686	4905
UK	457	424	**303**	1156	998	2191	1890	866	757	1507	1317	1388	1197	2435	2100
France	88	81	**59**	220	190	418	359	160	138	278	240	265	228	464	399

FEMNIST	**Centr.**			**non-IID 1LE**				**non-IID 5LE**			
Country/	**PUE**			**FedAVG**		**FedADAM**		**FedAVG**		**FedADAM**	
CO$_2$e(g)	1.67	1.55	1.11	TX2	NX	TX2	NX	TX2	NX	TX2	NX
Australia	0.7	0.6	**0.4**	140.9	156.9	41.2	45.9	84.2	93.1	28.1	30.1
UK	0.3	0.3	**0.2**	58.5	65.1	17.1	19.1	35.0	38.7	11.7	12.9
France	0.1	0.1	**0.03**	3.6	4.0	1.1	1.2	2.3	2.5	0.8	0.8

SpeechCmd	**Centr.**			**IID 5LE**				**non-IID 1LE**				**non-IID 5LE**			
Country/	**PUE**			**FedAVG**		**FedADAM**		**FedAVG**		**FedADAM**		**FedAVG**		**FedADAM**	
CO$_2$e(g)	1.67	1.55	1.11	TX2	NX	TX2	NX	TX2	NX	TX2	NX	TX2	NX	TX2	NX
Australia	11.8	10.9	**7.8**	30.5	30.8	13.6	13.7	146.4	159.1	36.7	39.9	35.9	36.2	16.9	17.1
UK	5.0	4.7	**3.4**	12.8	12.9	5.7	5.7	60.9	66.2	15.3	16.6	15.1	15.1	7.1	7.1
France	1.0	0.9	0.6	1.3	1.2	0.6	**0.5**	4.4	4.6	1.1	1.2	1.6	1.4	0.7	0.7

CV Italian	**Centr.**			**non-IID 5LE**	
Country/	**PUE**			**FedAVG**	
CO$_2$e(g)	1.67	1.55	1.11	TX2	NX
Australia	337.7	313.4	224.4	330.3	**324.0**
UK	144.6	134.2	96.1	140.2	**137.3**
France	27.8	25.8	18.5	21.6	**20.4**

Table 3: CO$_2$e emissions (expressed in grams, i.e **lower is better**) for both centralized learning and FL when they reach the target accuracies, with different tasks and setups. The tables report results of both FedAvg and FedADAM in both IID and non-IID partitions. As non-IID is more realistic, we report both 1 and 5 local epochs experiment results for this setup only. Results in bold indicate lower carbon emissions overall.

also impacts the carbon footprint of training in FL via communication speed. If the physical location has a slower Internet connection, the total time for communicating model parameters back and forth from the clients to the server will be longer, hence more energy is consumed.

Hardware efficiency is also a critical factor when estimating the total carbon footprint. As new AI applications for consumers are created every day, it is realistic to assume that novel versions of chips like Tegra TX2 will soon be embedded in numerous devices, including smartphones, tablets, and others. However, such specialized hardware is certainly not an exact estimate of what is currently being used for FL. Therefore, to facilitate carbon impact estimations of large-scale FL deployment, the industry must increase its transparency with respect to its devices' distribution over the market. As we can see from the results, even though NX requires less training time compared to TX2, it also consumes more power both during training and in an idle state. This leads to a trade-off between high-power hardware and actually training consumption. For example, training FEMNIST with 1 local epoch with FedAVG in TX2 emits more carbon compared to NX, but it emits less carbon compared to NX when we switch to FedADAM.

As explained in our estimation methodology, FL will always have an advantage in the respect that FL does not require cooling as opposed to centralized learning in the data centers. In fact, even though GPUs or even TPUs are getting more efficient in terms of computational power delivered by the amount of energy consumed, the need for strong and energy-consuming cooling remains; thus, the FL can always benefit more from the hardware advancement. On the other hand, FL always has a drawback of communication when the model parameters are communicated between clients and the central server.

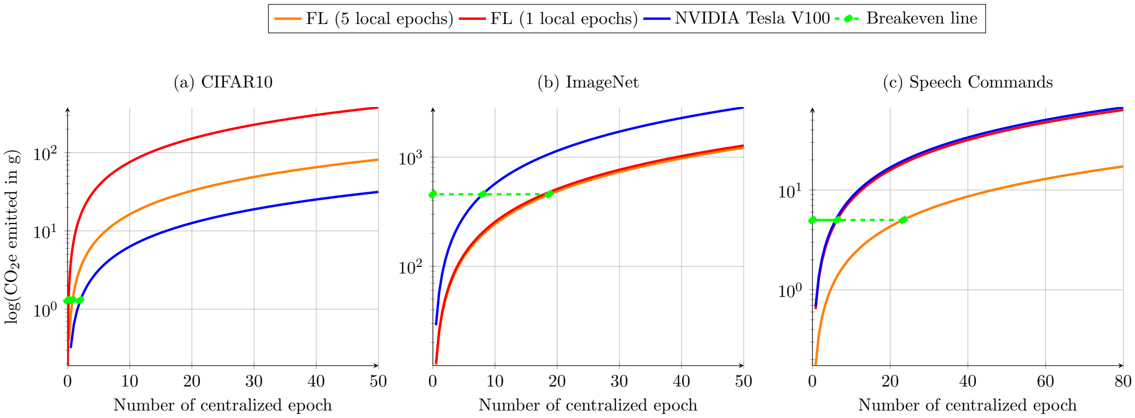


Figure 2: Growth of CO$_2$e emissions (in log scale) with a $PUE = 1.67$ for centralized learning and TX2 devices for FL in the UK (expressed in grams, i.e **lower is better**). Communication rounds in FL are converted to centralized epochs for a fair comparison, and CO$_2$e emissions are linearly dependent on the number of centralized epochs. The break-even line is chosen at the level when the centralized training reaches target accuracy. (a) For CIFAR10, 1 centralized epoch is equivalent to 50 communication rounds for 1 Local Epoch (LE) and 10 for 5 LE, as there are a total of 500 clients. The green line shows that with an equal amount of emissions between FL and V100, FL would train for 7.5 rounds with 5LE and 8 rounds with 1LE. (b) For ImageNet, due to the smaller size of the total client pool (100), 1 centralized epoch is equivalent to 2 communication rounds with 5LE and 10 rounds with 1LE. The green line shows that with an equal amount of emissions, FL can only train for 178 rounds with 1LE and 38 rounds with 5LE. (c) For Speech Commands, the total number of clients is also 100. The green line shows that with an equal amount of emissions, FL can only train for 47 rounds with 1LE and 64 rounds with 5LE.

Furthermore, CO$_2$ emissions depend on the distribution of clients' datasets. Our results show that realistic training conditions for FL (*i.e.* non-IID data) are largely responsible for longer training times, which in turn translates to a high level of CO$_2$e emissions. While it is well known that the simpler aggregation form of FL (*e.g.* FedAVG) performs reasonably well on IID data, it definitely struggles with non-IID partitioned data in terms of accuracy (Li et al., 2020; Qian et al., 2020). Interestingly, more complex strategies such as FedADAM can enable a decrease of up to 75% and 70% of the emitted CO$_2$e on Speech Commands and FEMNIST, respectively, compared to FedAvg. It is worth pointing out that for CIFAR10, FEMNIST, and Speech Commands non-IID partitions running 1LE produces more carbon than 5LE regardless of the aggregation strategy or devices. This is because communication consumption plays a big role in the total energy consumption, and 5LE communication costs less than using 1LE as fewer communication rounds are required.

Figure 2 shows the growth of carbon emission when the number of centralized epochs increases. We first see that for CIFAR10, FL with 1LE has the highest slope, while for the other two datasets, centralized learning has the highest slope. Normally centralized learning should exhibit stronger slopes as TDPs of centralized learning hardware are much higher than for FL. However, for CIFAR10, and due to the large model size, the communication consumption is much higher than the actual training consumption, resulting in a very steep slope suggesting that employing a complex model does not benefit FL.

In Figure 3, we compare equivalent carbon budgets on CIFAR for FL and V100s. The former would only be able to train for 7.5 and 8 rounds with 5 and 1 local epochs, respectively, resulting in degraded performances. The same goes for ImageNet. Indeed, FL would only train for 178 rounds with 1 LE and 38 rounds with 5 LE. On the other hand, in Speech Commands, FL did not outperform in the UK. Hence would only train for 64 rounds with 1 LE and 47 with 5 LE. However, we can see that the difference between the break-even rounds and actual rounds required, as shown in Table 2, is much smaller than other tasks. It is also interesting to note that, for ImageNet, the communication cost is negligible, hence both FL curves look very similar.

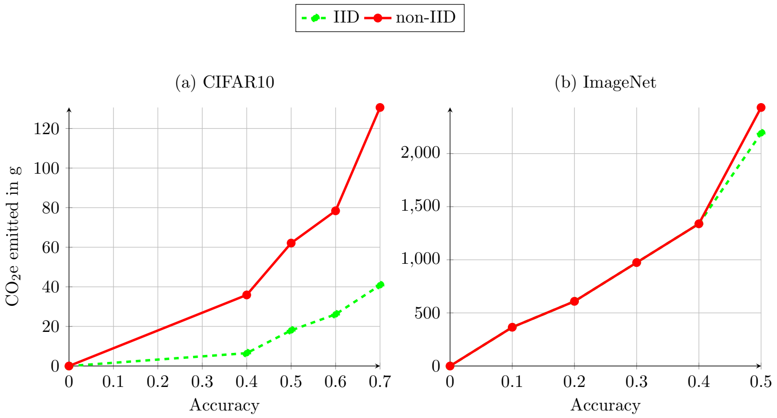


Figure 3: Growth of CO$_2$e emissions in the UK using TX2 for CIFAR10 and ImageNet respects to accuracies. The reported CO$_2$e emissions is for FedADAM with 5 local epochs.

Furthermore, Fig. 3 demonstrates the growth of carbon emission with respect to accuracies. Non-IID partitioning generally emits more carbon as it requires larger numbers of communication rounds. Fig. 3 also shows that the marginal carbon emission for additional accuracy gains is increasing exponentially. However, it is interesting to notice that carbon emissions of IID and non-IID at the beginning of ImageNet training overlap, as they require a similar number of rounds to reach certain accuracies.

Finally, it is also worth noticing that the percentage of CO$_2$e emission resulting from WAN changes across the dataset and FL setups. It highly depends on the size of the model, the size of the dataset in each client, and the energy consumed by clients during training. More precisely, communications accounted for up to 0.7% (ImageNet with 5 local epochs) and 96% (CIFAR10 with 1 local epoch) of the total emissions. With CIFAR10 tasks, communication actually emits much more CO_2e than training, while on the other hand, WAN plays a very small role for ImageNet.

5.2 Road-map for FL

FL is still a maturing framework with a lot to improve in a different aspect. We would like to highlight a few challenges and future research directions based on our analysis.

First, as carbon footprint largely depends on the physical location of hardware, either in terms of training or communication, carbon emission can be immensely reduced by selecting clients from greener locations or with faster internet connections. Obviously, there will be practical concerns in choosing clients in certain locations more often. For example, clients from greener locations might not have enough data samples for training or might represent a skewed data distribution. This, however, could lead to a demographic bias and needs further investigation.

Also, industrial statistics on the available fleet of devices are crucial to optimize the carbon emissions of FL. Indeed, in the real world, hardware efficiency can vary vastly from client to client. Similarly to the physical location, we would also like to choose clients with more efficient hardware and comparable computing capability and such a selection also induces potential biases.

As is the case in centralized training, hyper-parameter tuning is of great importance in reducing training times. In our experiments, we decided only to modify optimizer-related parameters (e.g. learning rate, momentum) to ensure a fair comparison and a sufficient level of performance. Further tuning can be done to facilitate the training convergence of FL. Nevertheless, with FL, hyper-parameter tuning becomes a more arduous task as it potentially involves hundreds of different models (i.e. local clients models), each making use of a small dataset that is likely to follow a very skewed distribution. In addition to client-side tuning, the aggregation strategy (e.g. FedADAM) might also offer further parameterization, therefore increasing the complexity of the tuning process. Therefore, novel hyper-parameter tuning algorithms should carefully be designed to minimize carbon emission by jointly maximizing the accuracy and minimizing the released CO_2e.

The number of local epochs is also an important hyper-parameter that can surely impact the overall carbon emission. As seen in Table 3, 5 local epochs settings often emit less carbon than 1 local epoch settings for non-IID, apart from the ImageNet task. This is easily explained by the hidden communication cost. Indeed, a single local epoch implies more communication rounds and, therefore, energy to converge compared to five local epochs. Furthermore, the number of communication rounds required for five local epochs usually is less than five times the number of communication rounds required for one local epoch. In the context of ImageNet, things are completely different as the local training becomes much more energy-demanding. Therefore, simply finding the right number of local epochs also clearly appears as a critical point in reducing FL carbon emissions.

Finally, carbon emission also depends on aggregation strategies. With more advanced aggregation strategies, the number of communication rounds can be reduced, hence reducing the overall carbon emissions.

In summary, we quantify the carbon footprint based on the training energy consumption and communication energy consumption, which depend on the physical locations of the hardware, hardware efficiency, training hyper-parameters, and FL strategies. We found that the carbon footprint of FL is hard to assess compared to centralized training without context, due to the inherent complications in how FL is currently performed. The complications might include data heterogeneity, client geographic distribution, and system heterogeneity. We provide a comprehensive analysis in this section and highlight the challenges and future research directions toward a more carbon-friendly FL.

6. Conclusion

Federated learning is an upcoming paradigm in the ML world that is often proposed as an alternative to an already carbon-emitting centralized training. A number of recent studies have begun to detail the environmental costs of their novel deep learning methods, sometimes even integrating CO_2e emissions as an objective to be minimized. Following this important trend, this article takes a first look into the carbon footprint of an increasingly deployed training strategy known as federated learning. In particular, this work introduces a generalized methodology to systematically compute the carbon footprint of many FL setups and conducts extensive experiments on real FL hardware under different settings, models, strategies, and tasks. We highlight the carbon footprint of FL from different perspectives and demonstrate that each element of FL can have an impact on the total CO_2e emission, including physical location, deep learning tasks, model architecture, FL aggregation strategies, and hardware efficiency. Finally, we hope to emphasize the importance of taking the carbon footprint into consideration for future research, and innovative research for both deep learning and FL can integrate the carbon footprint as a novel metric.

Acknowledgments

This work was supported by the UK's Engineering and Physical Sciences Research Council (EPSRC) with grants EP/M50659X/1 and EP/S001530/1 and the European Research Council via the REDIAL project.

References

Dario Amodei and Danny Hernandez. Ai and compute. *Heruntergeladen von https://openai.com/blog/ai-and-compute/*, 2018.

Kevin Anderson. The inconvenient truth of carbon offsets. *Nature*, 484(7392):7–7, 2012.

Anders SG Andrae and Tomas Edler. On global electricity usage of communication technology: trends to 2030. *Challenges*, 6(1):117–157, 2015.

Lasse F Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. *arXiv preprint arXiv:2007.03051*, 2020.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222, 2020.

AWS. Aws and sustainability. https://aws.amazon.com/about-aws/sustainability/, 2020.

E. K. Barrett, C. M. Fard, H. N. Katinas, C. V. Moens, L. E. Perry, B. E. Ruddy, S. D. Shah, I. S. Tucker, T. J. Wilson, M. Rucker, L. Cai, L. E. Barnes, and M. Boukhechba. Mobile sensing: Leveraging machine learning for efficient human behavior modeling. In *2020 Systems and Information Engineering Design Symposium (SIEDS)*, pages 1–7, 2020. doi: 10.1109/SIEDS49339.2020.9106648.

R. F. Berriel, A. T. Lopes, A. Rodrigues, F. M. Varejão, and T. Oliveira-Santos. Monthly energy consumption forecast: A deep learning approach. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 4283–4290, 2017.

Paolo Bertoldi and Thomas Huld. Tradable certificates for renewable electricity and energy savings. *Energy policy*, 34(2):212–222, 2006.

Daniel J Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Hei Li Kwing, Titouan Parcollet, Pedro PB de Gusmão, and Nicholas D Lane. Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*, 2020.

K. A. Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for federated learning on user-held data. In *NIPS Workshop on Private Multi-Party Machine Learning*, 2016. URL https://arxiv.org/abs/1611.04482.

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings, 2018.

Alfonso Capozzoli and Giulio Primiceri. Cooling systems in data centers: state of art and emerging technologies. *Energy Procedia*, 83:484–493, 2015.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pages 2921–2926. IEEE, 2017.

Thomas J Crowley. Causes of climate change over the past 1000 years. *Science*, 289(5477): 270–277, 2000.

The Power Consumption Database. http://www.tpcdb.com/list.php?page=1&type=11.

Dimitrios Dimitriadis, Ken'ichi Kumatani, Robert Gmyr, Yashesh Gaur, and Sefik Emre Eskimez. A federated approach in training acoustic models. In *Interspeech*, pages 981–985, 2020.

Yan Gao, Titouan Parcollet, Salah Zaiem, Javier Fernandez-Marques, Pedro PB de Gusmao, Daniel J Beutel, and Nicholas D Lane. End-to-end speech recognition from federated acoustic models. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7227–7231. IEEE, 2022.

Google. Efficiency-data centres. `https://www.google.co.uk/about/datacenters/efficiency/`, 2020a.

Google. 24/7 carbon-free energy by 2030. `https://www.google.com/about/datacenters/cleanenergy/`, 2020b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning. *The Journal of Machine Learning Research*, 21(1):10039–10081, 2020.

Miro Hodak, Masha Gorkovenko, and Ajay Dholakia. Towards power efficiency in deep learning on data center hardware. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1814–1820. IEEE, 2019.

Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, pages 4387–4398. PMLR, 2020.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

IPCC. Climate change 2014 synthesis report. 2014.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4835–4839. IEEE, 2017.

Young Geun Kim and Carole-Jean Wu. Autofl: Enabling heterogeneity-aware energy efficient federated learning. In *MICRO-54: 54th Annual IEEE/ACM International Symposium on*

Microarchitecture, MICRO '21, page 183–198, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385572. doi: 10.1145/3466752.3480129. URL `https://doi.org/10.1145/3466752.3480129`.

Leonard J Kish and Eric J Topol. Unpatients—why patients should own their medical data. *Nature biotechnology*, 33(9):921–924, 2015.

Jakub Konečný, Brendan McMahan, and Daniel Ramage. Federated optimization: Distributed optimization beyond the datacenter. *arXiv preprint arXiv:1511.03575*, 2015.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 2020.

Jens Malmodin and Dag Lundén. The energy and carbon footprint of the global ict and e&m sectors 2010–2015. *Sustainability*, 10(9):3027, 2018.

Brendan McMahan and Daniel Ramage. Federated learning: Collaborative machine learning without centralized training data. *Google Research Blog*, 3, 2017.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR. URL `http://proceedings.mlr.press/v54/mcmahan17a.html`.

H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=BJ0hF1Z0b`.

Microsoft. Datacenter fact sheet - microsoft download center. `http://download.microsoft.com/download/8/2/9/8297f7c7-ae81-4e99-b1db-d65a01f7a8ef/microsoft_cloud_infrastructure_datacenter_and_network_fact_sheet.pdf`, 2015.

Nature. How to stop data centres from gobbling up the world's electricity. `https://www.nature.com/articles/d41586-018-06610-y`, Sep 2018.

T. Nishio and R. Yonetani. Client selection for federated learning with heterogeneous resources in mobile edge. In *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, pages 1–7, 2019.

Rajendra K Pachauri, L Gomez-Echeverri, and K Riahi. Synthesis report: summary for policy makers. 2014.

Titouan Parcollet and Mirco Ravanelli. The Energy and Carbon Footprint of Training End-to-End Speech Recognizers. In *Proc. Interspeech 2021*, pages 4583–4587, 2021. doi: 10.21437/Interspeech.2021-456.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Hung Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeffrey Dean. The carbon footprint of machine learning training will plateau, then shrink. 2022.

Laércio Lima Pilla. Scheduling algorithms for federated learning with minimal energy consumption. *IEEE Transactions on Parallel and Distributed Systems*, 34(4):1215–1226, 2023.

R. Pryss, M. Reichert, J. Herrmann, B. Langguth, and W. Schlee. Mobile crowd sensing in clinical and psychological trials – a case study. In *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*, pages 23–24, 2015.

Jia Qian, Xenofon Fafoutis, and Lars Kai Hansen. Towards federated learning: Robustness analytics to data heterogeneity. *arXiv preprint arXiv:2002.05038*, 2020.

Xinchi Qiu, Titouan Parcollet, Daniel J. Beutel, Taner Topal, Akhil Mathur, and Nicholas D. Lane. Can federated learning save the planet?, 2021.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*, 2021.

Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *International Conference on Learning Representations (ICLR)*, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Stefano Savazzi, Vittorio Rampa, Sanaz Kianoush, and Mehdi Bennis. An energy and carbon footprint analysis of distributed and federated learning. *IEEE Transactions on Green Communications and Networking*, 7(1):248–264, 2023.

Helen Shen. Smartphones set to boost large-scale health studies. *Nature News*, 2015.

Ryan Smith. Nvidia announces jetson tx2: Parket comes to nvidia's embedded system kit. `https://www.anandtech.com/show/11185/ia-announces-jetson-tx2-parker`, mar 2017.

Ryan Smith. Nvidia gives jetson agx xavier a trim, announces nano-sized jetson xavier nx. `https://www.anandtech.com/show/15070/nvidia-gives-jetson-xavier-a-trim-announces-nanosized-jetson-xavier-nx`, nov 2019.

Speedtest. `https://www.speedtest.net/global-index`.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019.

Awada Uchechukwu, Keqiu Li, Yanming Shen, et al. Energy consumption in cloud computing data centers. *International Journal of Cloud Computing and Services Science (IJ-CLOSER)*, 3(3):145–162, 2014.

UptimeInstitue. 2019 data center industry survey results. `https://uptimeinstitute.com/2019-data-center-industry-survey-results`, 2019.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning*, pages 7252–7261. PMLR, 2019.

Yundong Zhang, Naveen Suda, Liangzhen Lai, and Vikas Chandra. Hello edge: Keyword spotting on microcontrollers. *arXiv preprint arXiv:1711.07128*, 2017.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.