A Carbon Tracking Model for Federated Learning: Impact of Quantization and Sparsification

Luca Barbieri[1,2], Stefano Savazzi[2], Sanaz Kianoush[2], Monica Nicoli[1], Luigi Serio[3]
[1] *Politecnico di Milano, Milan, Italy*, [2] *Consiglio Nazionale delle Ricerche, Milan, Italy*,
[3] *Technology Department, CERN - 1211 Geneva 23 - Switzerland*

***Abstract*—Federated Learning (FL) methods adopt efficient communication technologies to distribute machine learning tasks across edge devices, reducing the overhead in terms of data storage and computational complexity compared to centralized solutions. Rather than moving large data volumes from producers (sensors, machines) to energy-hungry data centers, raising environmental concerns due to resource demands, FL provides an alternative solution to mitigate the energy demands of several learning tasks while enabling new Artificial Intelligence of Things (AIoT) applications. This paper proposes a framework for real-time monitoring of the energy and carbon footprint impacts of FL systems. The carbon tracking tool is evaluated for consensus (fully decentralized) and classical FL policies. For the first time, we present a quantitative evaluation of different computationally and communication efficient FL methods from the perspectives of energy consumption and carbon equivalent emissions, suggesting also general guidelines for energy-efficient design. Results indicate that consensus-driven FL implementations should be preferred for limiting carbon emissions when the energy efficiency of the communication is low (i.e., < 25 Kbit/Joule). Besides, quantization and sparsification operations are shown to strike a balance between learning performances and energy consumption, leading to sustainable FL designs.**

***Index Terms*—Federated Learning, Consensus, Energy Consumption, Green Machine Learning, Internet of Things.**

I. Introduction

Data centers are today a key component of many Artificial Intelligence of Things (AIoT) services, which rely on the network for data sharing and Artificial Intelligence (AI) for analytics. They contribute 0.3 % of the global equivalent Green House Gas (GHG) emissions (and about 15% of the emissions of the entire Information and Communication Technology (ICT) ecosystem), which will further increase in the years to come [1], [14]. Federated Learning (FL) [2] is emerging as a promising alternative to centralized AIoT, especially for training tasks where the data privacy needs to be protected (i.e., medical data): it distributes the computing tasks across many edge devices possibly characterized by a more efficient use of the energy compared with data centers [4], [14]. Combined with a judicious design of networking and training stages, FL is expected to bring significant benefits in terms of environmental impact, obviating in many cases the need for a large centralized infrastructure for cooling or power delivery,

This paper is funded by the EU in the call HORIZON-HLTH-2022-STAYHLTH-01-two-stage under grant agreement No 101080564. The work is also supported by the Italian National Recovery and Resilience Plan (NRRP) of NextGeneration EU, partnership on "Telecommunications of the Future" (PE0000001 program "RESTART").

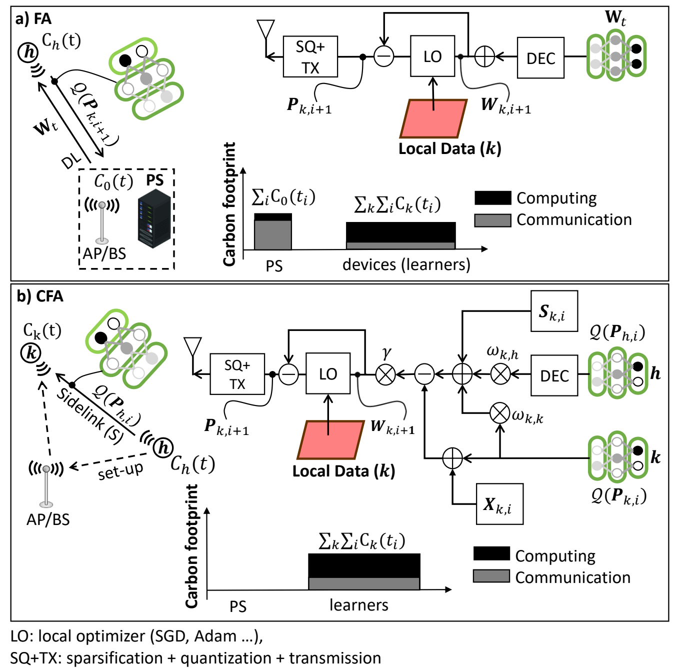

Fig. 1. Federated Averaging (FA) relying on (a) Parameter Server (PS) for model aggregation and (b) Consensus process based on CHOCO-SGD tool. Sparsification and quantization operations and on-device carbon tracking.

or reducing the emissions by migrating training tasks across different geographic locations according to time-dependent availability of sustainable energy.

FL architectures can be broadly divided into centralized and decentralized, each involving different energy models, as shown in Fig. 1. For example, vanilla Federated Averaging (FA) [2] resorts to the classical server-client architecture, where a Parameter Server (PS) coordinates the learning process, collects the local models (rather than raw data as in classical Machine Learning (ML)) and sends back the updated models (by averaging) to the devices. Both the PS and the federation of the devices contribute to the energy footprint. On the other hand, in decentralized FL tools, such as Consensus-driven FA (CFA), the local model parameters are shared and synchronized across multiple learners via mesh, or Device-to-Device (D2D) networking, without relying on the PS [3].

The problem of quantifying the energy and GHG emissions, namely the carbon footprint, of FL has been recently tackled (see e.g., [4], [14]) although the development of sustainable design and monitoring tools still remains partially unexplored.

Quantitative and qualitative evaluation of energy-efficient FL strategies has been also carried out by comparing different implementations [10], [5]. Quantization and sparsification approaches [15], [22] along with optimized strategies for selecting suitable devices or informative ML model parameters [11], [12] to be exchanged during the FL process are to be considered as critical for reducing FL emissions.

Contributions. This paper develops a *carbon tracking* framework aiming to provide: 1) a systematic reporting of carbon and energy footprints in FL processes over wireless networks; and 2) a toolkit to assess and compare different computationally efficient methods for ML model parameter selection, compression and quantization, adapted to both centralized (FA) and decentralized (CFA) processes, and taking into account their carbon and energy footprints. Compared to the previous works [4], [14], the developed tool quantifies explicitly the impact of model parameter sparsification and quantization, proposed in emerging FL designs [15], [16], on the energy balance. In particular, the proposed carbon tracking framework quantifies the carbon emissions on each FL round and accounts for the emissions from energy grids, as well as the energy outputs from CPU/GPUs of the individual devices. Results consider two increasingly complex image classification tasks, namely MNIST [8] and CIFAR10 [9], as well as ML model sizes to comprehensively evaluate the energy/carbon consumption of FA and CFA implementations. The goal is to study the optimal configuration of quantization and ML model sparsification to limit the FL process energy demands while maximizing learning performances. The analysis shows that decentralized FL (CFA) policies should be preferred when energy-inefficient communication protocols are adopted. On the other hand, vanilla FA solutions may be selected when more efficient uplink/downlink communications are available. Moreover, devices and PS carbon emissions need also to be well balanced to provide sustainable training processes.

The remainder of this paper is organized as follows. Sec. II introduces the carbon tracking framework, while Sec. III presents the quantization and sparsification strategies adopted. Sec. IV assesses the energy/carbon footprint of the proposed FL designs. Finally, Sec. V concludes the paper.

II. CARBON TRACKING FRAMEWORK

Energy and carbon footprint models in FL (see [4], [14] for a review) assume that the energy budget of the learning process can be broken down into computing and communication costs. Notice that both the PS (denoted with index $k = 0$), when used, and the learners/devices ($k > 0$) contribute to the energy balance, and their footprints should be evaluated separately. In what follows, we assume that a carbon tracking update is issued on every new FL round i that occurs at (discrete) time $t_i \geq t_0$, with t_0 being the starting time of the learning process. Let $\mathrm{C}_k(t_i)$ refer to the equivalent GHG emissions of device k observed up to time t_i, the proposed carbon tracking tool quantifies the GHG footprint through an iterative approach

$$\mathrm{C}_k(t_i) = E_{k,i} \cdot I_{k,i} + \mathrm{C}_k(t_{i-1}), \tag{1}$$

where $E_{k,i}$ is the energy cost to implement the new FL round i, including local model optimization and communication, and $I_{k,i} = I_k(t_i)$ is the Carbon Intensity (CI) of the electricity generation [21], measured in Kg CO2-equivalent emissions per kWh (KgCO2-eq/kWh) at time t_i. CI quantifies how much carbon emissions are produced per kilowatt hour of locally generated electricity. The $I_{k,i}$ terms are typically monitored on an hourly basis [7], [14] as they depend on the specific regional energy grid where the device k or PS is installed. Notice that grid differences can result in large variations in eq. CO2 emissions [1], as analyzed in Sec. IV.

In the following, we model the energy cost $E_{k,i}$ based on the specific FL method employed. The communication costs are quantified on average, in terms of the corresponding energy efficiencies (EE), standardized by the European Telecommunications Standards Institute (ETSI) [18]. Efficiency terms in downlink (DL) $\mathrm{EE_D}$, uplink (UL) $\mathrm{EE_U}$, or sidelink (S) $\mathrm{EE_S}$ transmissions are measured here in bit/Joule [bit/J] [19]. The efficiencies also include the power dissipated in the RF front-end, baseband processing, and transceiver stages.

A. Energy tracking for Parameter Server based FL (FA)

In vanilla FL methods (FA), the PS collects the local models produced by K learners and produces a global model of size $b_{\mathbf{W}}$ bits, which is fed back to active devices on each round. Learners are powered on as they run the local optimizer (LO) and decode the updated global model $\mathbf{W}_{i-1}$ obtained from the PS at round $i-1$. During a FL round, the energy cost $E_{k,i} = E_{k,i}^{(\mathrm{FA})}$ [Joule] of a device ($k > 0$) can be written as

$$E_{k,i}^{(\mathrm{FA})} = E_{k,i}^{(\mathrm{C})} + \frac{b_{\mathbf{W}}}{\mathrm{EE_D}} + \frac{\mathbb{Q}_{k,i}[b_{\mathbf{W}}]}{\mathrm{EE_U}} + E_{k,\mathbb{Q}}^{(\mathrm{C})} + E_{k,\mathrm{sleep}}^{(\mathrm{C})}, \tag{2}$$

namely, the superposition of the energy spent for receiving the global model from the PS $\frac{b_{\mathbf{W}}}{\mathrm{EE_D}}$, the cost for the LO at round i $E_{k,i}^{(\mathrm{C})}$ required for SGD computation, and the UL communication $\frac{\mathbb{Q}_k[b_{\mathbf{W}}]}{\mathrm{EE_U}}$ of the selected local model parameters according to the model selection and quantization policy $\mathbb{Q}_k[\cdot]$ (see Sec. III). Implementation of policy $\mathbb{Q}_k[\cdot]$ has a cost which is quantified here as $E_{k,\mathbb{Q}}^{(\mathrm{C})}$. Finally, $E_{k,\mathrm{sleep}}^{(\mathrm{C})}$ is the energy cost in sleep mode, which is required by devices while waiting for the global model to be produced by the PS.

The PS ($k = 0$) energy per round is

$$E_{0,i}^{(\mathrm{PS})} = K \cdot E_{0,\mathrm{global}}^{(\mathrm{C})} + \frac{b_{\mathbf{W}}}{\mathrm{EE_U}} + \sum_{k=1}^{K} \frac{\mathbb{Q}_{k,i}[b_{\mathbf{W}}]}{\mathrm{EE_D}} + E_{0,\mathrm{sleep}}^{(\mathrm{C})}, \tag{3}$$

which accounts for the cost for K global model updates each with cost $E_{0,\mathrm{global}}^{(\mathrm{C})}$, global model DL publication $\frac{b_{\mathbf{W}}}{\mathrm{EE_U}}$, and collection of local model parameters $\frac{\mathbb{Q}_{k,i}[b_{\mathbf{W}}]}{\mathrm{EE_D}}$ from the K learners. The total carbon emissions produced by FA can be then evaluated as

$$\mathrm{C}_{\mathrm{tot}}^{(\mathrm{FA})} = \sum_i \sum_{k=1}^{K} \mathrm{C}_k^{(\mathrm{FA})}(t_i) + \sum_i \mathrm{C}_0^{(\mathrm{PS})}(t_i) \tag{4}$$

where $\mathrm{C}_k^{(\mathrm{FA})}(t_i) = E_{k,i}^{(\mathrm{FA})} \cdot I_{k,i} + \mathrm{C}_k^{(\mathrm{FA})}(t_{i-1})$ is the carbon footprint of device k at round i while $\mathrm{C}_0^{(\mathrm{PS})}(t_i) = E_{0,i}^{(\mathrm{PS})} \cdot I_{0,i} + \mathrm{C}_0^{(\mathrm{PS})}(t_{i-1})$ is the carbon footprint of the PS at round i.

B. Energy tracking for decentralized CFA

Decentralized CFA techniques [2], [3] do not employ the PS as the devices mutually exchange their local model parameters, i.e., through publish-subscribe operations [13]. Each learner is responsible for producing a global model representation which is the result of a consensus over the received local model parameters obtained from neighbor devices. CFA adopts a distributed weighted averaging approach [3], [2] used to combine the received neighbor models.

As shown in Fig. 1, the devices mutually exchange their local model parameters with an assigned number $N < K$ of neighbors. Let $\mathcal{N}_{k,i}$ be the set that contains the N chosen neighbors of node k at round i, the energy cost $E_{k,i} = E_{k,i}^{(\mathrm{CFA})}$ of an individual learner ($k > 0$) can be written as

$$E_{k,i}^{(\mathrm{CFA})} = E_{k,i}^{(\mathrm{C})} + N \cdot E_{k,\mathrm{global}}^{(\mathrm{C})} + \sum_{h \in \mathcal{N}_{k,i}} \frac{\mathbb{Q}_{h,i}[b_{\mathbf{W}}]}{\mathrm{EE_S}} + \frac{\mathbb{Q}_{k,i}[b_{\mathbf{W}}]}{\mathrm{EE_S}} + E_{k,\mathbb{Q}}^{(\mathrm{C})} + E_{k,\mathrm{sleep}}^{(\mathrm{C})}, \tag{5}$$

where now each learner runs the LO (with cost $E_{k,i}^{(\mathrm{C})}$), produces a global model representation via N weighted averaging steps $E_{k,\mathrm{global}}^{(\mathrm{C})}$, distributes the local (selected) parameters and obtains the neighbor ones using sidelink communications.

The total carbon consumption under CFA can be quantified as

$$\mathrm{C}_{\mathrm{tot}}^{(\mathrm{CFA})} = \sum_i \sum_{k=1}^{K} \mathrm{C}_k^{(\mathrm{CFA})}(t_i), \tag{6}$$

with $C_k^{(\mathrm{CFA})}(t_i) = E_{k,i}^{(\mathrm{CFA})} \cdot I_{k,i} + C_k^{(\mathrm{CFA})}(t_{i-1})$ being the carbon footprint of device k at round i.

III. QUANTIZATION AND PARAMETER SELECTION IN FL

This section presents the compression strategies employed for evaluating the impact of the communication on the energy/carbon footprint for the proposed FL setups. To guarantee a fair comparison, we assume that both centralized and decentralized FL tools rely on the same compression operators. More specifically, we consider two widely-adopted mechanisms, namely top-t sparsification [23] and probabilistic quantization [22], which are applied independently by each device member of the federation.

We consider a generic input vector $\mathbf{w}$, of N_P elements w_n, $n = 1, ..., N_P$, that collects the entries of the model parameters $\mathbf{W}_{k,i}$ or some surrogate quantities, such as the corresponding gradients or updates. The compression policy aims at obtaining a lower bit representation of $\mathbf{w}$ by successively applying sparsification and quantization as

$$\bar{\mathbf{w}} = \mathcal{Q}(\mathbf{w}) = f_q(f_s(\mathbf{w})), \tag{7}$$

where $\mathcal{Q}(.)$ denotes the overall compression function, while $f_s(.)$ and $f_q(.)$ indicate the operators required for top-t sparsification and probabilistic quantization, respectively. The sparsification operator $f_s(.)$ outputs a new representation $\widetilde{\mathbf{w}} = f_s(\mathbf{w})$ that selects the t largest absolute values of $\mathbf{w}$, and sets all other entries to 0. Quantization $f_q(.)$ encodes the sparsification output $\widetilde{\mathbf{w}}$ via a randomized rounding operation. By setting the output quantization bits to $N_b \leq N_{bc}$, with (typical) $N_{bc} = 32$ bits, the element $\bar{w}_n$ of $\bar{\mathbf{w}}$ is defined as [22]

$$\bar{w}_n = \|\widetilde{\mathbf{w}}\|_2 \cdot \mathrm{sign}(\widetilde{w}_n) \cdot \xi_n(\widetilde{\mathbf{w}}, 2^{N_b}), \tag{8}$$

where $\mathrm{sign}(.)$ denotes the sign operator, while $\xi_n(\widetilde{\mathbf{v}}, 2^{N_b})$ is defined as in [22].

Compression output $\bar{\mathbf{w}} = \mathcal{Q}(\mathbf{w})$ in (7) corresponds to the vector $\bar{\mathbf{w}} = [\bar{w}_1 \cdots \bar{w}_{N_P}]^{\mathrm{T}}$, and can be used as a lower bit representation of $\mathbf{w}$. The number $\mathbb{Q}_{k,i}[b_{\mathbf{W}}]$ of bits that are sent by device k on round i according to carbon tracking models (2)-(5), can be quantified as

$$\mathbb{Q}_{k,i}[b_{\mathbf{W}}] = \delta \cdot \frac{N_b}{N_{bc}} \cdot b_{\mathbf{W}} = t \cdot N_b, \tag{9}$$

where $\delta = \frac{t}{N_P}$ represents the fraction of the model parameters selected by top-t sparsification function $f_s(.)$, while $\frac{N_b}{N_{bc}}$ sets the quantization level according to $f_q(.)$ for each parameter.

In what follows, we review the specific operations required for implementing the considered compression strategies for the FA (Sec. III-A) and CFA (Sec. III-B) schemes.

A. Parameter Server based FL (FA)

In FA, the devices send to the PS the model updates being more suited for compression [2]. Given the model parameters $\mathbf{W}_{k,i}$, the model updates are evaluated as

$$\mathbf{P}_{k,i} = \mathbf{W}_{k,i+1/2} - \mathbf{W}_{k,i}, \tag{10}$$

where $\mathbf{W}_{k,i+1/2}$ denotes the parameters obtained after applying the LO for device k at round i. Then, $\mathbf{P}_{k,i}$ is compressed as

$$\bar{\mathbf{P}}_{k,i} = \mathcal{Q}(\mathbf{P}_{k,i}), \tag{11}$$

with $\mathcal{Q}(.)$ as in (7) and transmitted to the PS. The PS updates its global model by aggregating the received contributions as

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \sum_{k=1}^{N} \sigma_k \bar{\mathbf{P}}_{k,i}, \tag{12}$$

where σ_k is a weighting factor chosen as in [14]. Note that the PS forwards back the updated global model uncompressed as in Sec. II.

B. CFA based on CHOCO-SGD framework

The CFA process considered here relies on the CHOCO-SGD algorithm proposed in [16]. This scheme employs two additional variables stored at each device for preserving the average of the model iterates across consecutive rounds and for controlling the noise introduced by the compression [16]. In the same spirit as FA, each device k compresses the model updates in a manner similar to (10)-(11) leading to

$$\bar{\mathbf{P}}_{k,i} = \mathcal{Q}(\mathbf{P}_{k,i}) = \mathcal{Q}(\mathbf{W}_{k,i+1/2} - \mathbf{X}_{k,i}), \tag{13}$$

TABLE I
MODEL AND ENERGY PARAMETERS FOR FA AND CFA SCHEMES

	MNIST	CIFAR10
N_P	59500	28146954
$b_{\mathbf{W}}$	0.24 MB	112.59 MB
δ	10%, 50%, 100%	10%, 50%, 100%
N_b	8, 16, 32	16, 24, 32
$E_{k,i}^{(C)}$	3.51 Joule	5.53 KJoule
$E_{k,\mathbb{Q}}^{(C)}$	0.04 - 0.14 Joule	18.9 - 66.2 Joule
$E_{k,sleep}^{(C)}$	0.12 Joule	59.92 Joule
$E_{k,global}^{(C)}$	0.06 Joule	29.96 Joule
I_k	0.449 KgC02-eq/kWh	0.449 KgC02-eq/kWh
$E_{0,sleep}^{(C)}$	0.70 Joule	1.10 KJoule
$E_{0,global}^{(C)}$	0.24 Joule	114.02 Joule
I_0	0.449 KgC02-eq/kWh	0.449 KgC02-eq/kWh

where $\mathcal{Q}(.)$ is defined as in (7) and $\mathbf{X}_{k,i}$ is a local variable, with $\mathbf{X}_{k,0} = \mathbf{0}$ for round $i = 0$. The compressed representation $\bar{\mathbf{P}}_{k,i}$ is then exchanged over the network.

Upon receiving the contributions from its neighbors, device k updates $\mathbf{X}_{k,i}$ as

$$\mathbf{X}_{k,i+1} = \mathbf{X}_{k,i} + \bar{\mathbf{P}}_{k,i} \tag{14}$$

and then uses the compressed models received from the neighboring devices to update an additional local variable $\mathbf{S}_{k,i}$, with $\mathbf{S}_{k,0} = \mathbf{0}$ at round $i = 0$, as

$$\mathbf{S}_{k,i+1} = \mathbf{S}_{k,i} + \sum_{j \in \mathcal{N}_{k,i}} \omega_{k,j} \bar{\mathbf{P}}_{j,i}, \tag{15}$$

where $\omega_{k,j}$ is the (k,j)-th entry of a symmetric doubly stochastic matrix $\mathbf{\Omega}$. Finally, each device k updates its local model using the following update rule

$$\mathbf{W}_{k,i+1} = \mathbf{W}_{k,i+1/2} + \gamma(\mathbf{S}_{k,i+1} - \mathbf{X}_{k,i+1}), \tag{16}$$

where $0 < \gamma \leq 1$ is the consensus step-size.

IV. CARBON FOOTPRINT EVALUATION

This section discusses the main factors that are expected to steer the choice between vanilla FA and CFA paradigms towards sustainable designs. The goal is to provide a first look into the impact of quantization and sparsification of model parameters on carbon emissions and learning accuracy, for varying communication energy efficiencies typically found in wireless communication systems. We consider two scenarios with increasing model size and training dataset complexities. The first case (Sec. IV-A) focuses on the MNIST classification task [8], while the second (Sec. IV-B) concentrates on the CIFAR10 learning problem [9]. Energy and carbon footprints are influenced by the PS and device hardware configurations. Table I summarizes the main parameters used by the carbon tracking tool under the two scenarios. For the PS (FA), we used a CPU (Intel i7 8700K, 3.7 GHz, 64 GB, GPU not used). A realistic pool of $K = 10$ resource-constrained FL learners is adopted, namely Jetson Nano boards based on a low-power CPU (ARM-Cortex-A57 and GPU 128-core Maxwell). Their energy expenditure parameters when implementing the FA and CFA schemes reviewed in Sec. III are summarized in Table I. Each device is designed to track its carbon emissions independently as a function of the estimated CI $I_{k,i} = I_k$ which is assumed as constant for all FL rounds.

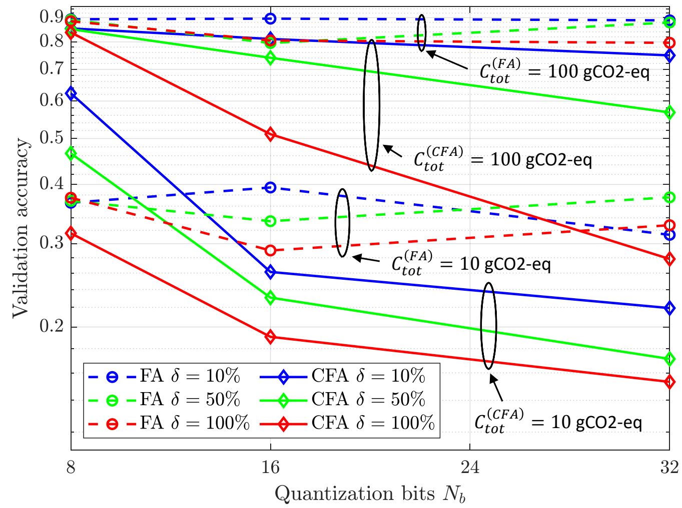

Fig. 2. Analysis of the validation accuracy achieved by FA and CFA schemes under different levels of quantization and sparsification. Rings group together curves related to the same carbon footprint.

In what follows, rather than choosing a specific communication or carbon emission setting, we consider a what-if analysis approach as proposed in [4], [14]. We thus quantify the achievable loss/accuracy of the proposed FL designs under the assumption of different DL/UL and SL efficiencies (setting $\mathrm{EE_{COM}} = \mathrm{EE_D} = \mathrm{EE_U} = \mathrm{EE_S}$) and CI I_k (see https://app.electricitymaps.com/map for reference values). To guarantee a fair comparison, we also assume that the PS collects the (compressed) model updates from all K devices. Similarly, CFA uses (all) $N = K - 1$ neighbors. Each device uses an SGD LO with learning rate 0.01, momentum 0.9, and batch size of 64 examples chosen in accordance with the computational capabilities of the chosen learners. For CFA, we set $\gamma = 0.01$ in all experiments. Note that the actual emissions may be larger than the estimated ones depending on the specific devices and implementations. Therefore, in the following, we will highlight relative comparisons.

A. Impact of sparsification, energy and carbon efficiencies

In the example, each device has access to 300 observations randomly drawn from the MNIST database [8]. The devices employ a Lenet-5 model [8], with parameters in Table I.

Fig. 2 analyzes the quantization and sparsification impact on the learning accuracy by enforcing a max. total carbon emission (carbon budget) of $\mathrm{C_{tot}}$. We consider two carbon emission targets, namely $\mathrm{C_{tot}} = \mathrm{C_{tot}^{(FA)}} = \mathrm{C_{tot}^{(CFA)}} = \{10, 100\}$ gCO2-eq, modeling low to medium carbon consumptions. The quantization bits are in the range $N_b = 8 - 32$ and the percentage of parameters shared $\delta = 10\% - 100\%$. $\mathrm{EE_{COM}} = 10$ Kbit/Joule, which roughly corresponds to an LTE design for macro-cell delivery [19], [21]. The devices/PS are

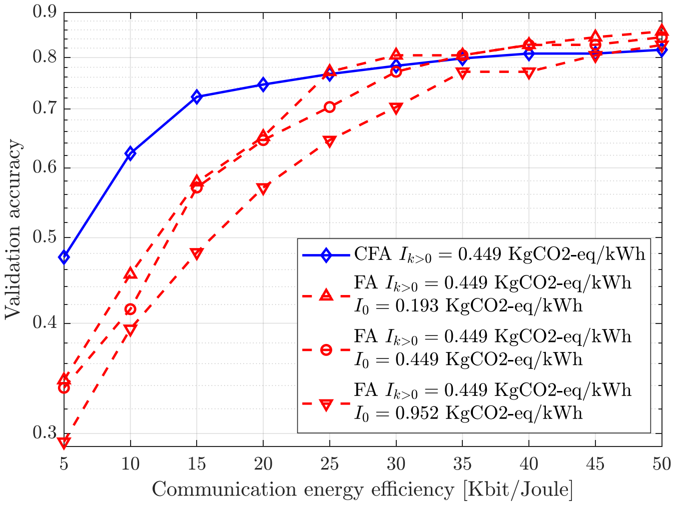

Fig. 3. Analysis of the validation accuracy under carbon constraints with different communication (energy) efficiencies as well as carbon intensities.

here located in the same region, i.e., Italy, with a corresponding CI $I_k = 0.449$ KgCO2-eq/kWh for $k = 0, \dots, N$.

Fig. 2 reports the validation accuracy obtained by the CFA and FA schemes under the two carbon budgets. For stringent emission constraints, i.e., $C_{\text{tot}} = 10$ gCO2-eq, CFA provides better performances when $\delta = \{10\%, 50\%\}$ and $N_b = 8$ bits, while for $C_{\text{tot}} = 100$ gCO2-eq FA achieves higher accuracy regardless of the number of parameters shared and the quantization bits employed. ML model compression is more critical in CFA rather than in FA since it reduces the sidelink channel use. FA is not that much affected by the specific choice of the compression parameters as long as the carbon budget is high enough, i.e., $C_{\text{tot}} = 100$ gCO2-eq. On the other hand, CFA schemes must generally employ a more aggressive compression, i.e., $\delta = 10\%$ and $N_b = 8$ bits, to guarantee a reasonable target accuracy. For FA, a good choice for the compression parameters is $\delta = 10\%$ and $N_b = 16$ bits as achieving the highest accuracy under all carbon budgets.

Fig. 3 analyzes the impact of the communication energy efficiency as well as the CI on the learning accuracy under the same carbon budget $\text{C}_{\text{tot}} = 10$ gCO2-eq. The energy efficiency EE_{COM} varies from 5 Kbit/Joule up to 50 Kbit/Joule, with the last case corresponding to 5G micro-cell delivery and WiFi. The results consider $\delta = 10\%$ and $N_b = 16$ bits for FA, while $\delta = 10\%$ and $N_b = 8$ bits for CFA following the previous analysis. In the same figure, we also study how the PS location and its CI I_0 affect the carbon consumption. Three CI values are considered, namely $I_0 = \{0.193, 0.449, 0.952\}$ KgCO2-eq/kWh, modeling different energy grid efficiencies. These values correspond to the geographical regions of Finland, Italy, and Poland, respectively. Comparing the results, CFA is shown to outperform FA for all CI terms, provided that the energy efficiency is below 25 Kbit/Joule. On the other hand, when the communication is more efficient, centralized FA should be preferred. High carbon intensities, i.e., $I_0 = 0.952$ KgCO2-eq/kWh, make FA tools more susceptible to accuracy losses due to the high energy costs. Interestingly, even when the PS is located in the same region of the devices, CFA strategies are to be preferred as more accurate compared to PS-based solutions.

B. CIFAR10 analysis: impact of large model size

To finalize the analysis, we here consider a more challenging image recognition task based on the CIFAR10 dataset and a much larger ML model. The goal is to evaluate the energy/carbon footprint of FA and CFA in a complex and potentially energy-hungry ML setup. Each learner is assigned 500 examples for each one of the 10 classes and employs a VGG11 architecture [24] with parameters defined in Table I.

Table II summarizes the results of all the tests. In line with the previous analysis, we consider two main scenarios: 1) the learning process is constrained to a maximum carbon emission of $\text{C}_{\text{tot}}^{(\text{FA})} = \text{C}_{\text{tot}}^{(\text{CFA})} = 1.5$ KgCO2-eq and 2) the same process must achieve a 70% target accuracy with no constraints on carbon emissions. In particular, the first three rows refer to scenario 1 and show the validation accuracy obtained by FA and CFA tools for three quantization and sparsification cases, namely $(\delta, N_b) = (10\%, 16)$, $(\delta, N_b) = (50\%, 24)$ and $(\delta, N_b) = (100\%, 32)$, modeling moderate to no compression. We consider two communication efficiencies $\text{EE}_{\text{COM}} = 10$ Kbit/Joule and $\text{EE}_{\text{COM}} = 100$ Kbit/Joule. The following three rows consider scenario 2 with same values for compression and energy efficiencies, and show now the carbon emissions required to reach 70% accuracy. Comparing the results, in accordance with the analysis in Fig. 3, CFA provides a higher accuracy w.r.t. to FA schemes when the communication is inefficient, i.e., $\text{EE}_{\text{COM}} = 10$ Kbit/Joule. FA is more effective when the learning process is implemented over more communication-efficient networks. Focusing now on the carbon emissions needed to reach a 70% target accuracy, CFA requires a much larger footprint (i.e., $3\times$) compared to FA for all cases considered. Nevertheless, optimizing the compression parameters is beneficial for reducing carbon consumption with respect to uncompressed communications: for FA selecting $\delta = 10\%$ and $N_b = 16$ bits allows to reduce the carbon consumption by roughly 50% while for CFA $\delta = 50\%$ and $N_b = 24$ bits leads to a 42% carbon reduction.

The last four rows of Table II analyze the impact of the regional CI on the performances and energy consumption with same configurations of the PS as described in the previous section. The results confirm the findings of Fig. 3: FA tools are advantageous compared to CFA policies when the PS is located in a region with CI comparable to or better than the one of the devices. Indeed, optimizing the PS location allows to further reduce the carbon consumption by 10%.

V. Conclusions

This paper proposed a carbon tracking framework for monitoring the energy/carbon footprints of FL policies. The proposed framework enables to assess the impact of ML model quantization and sparsification on the carbon/energy consumption for both centralized (FA) and consensus-driven (CFA) FL implementations. The developed tool keeps track

TABLE II
CIFAR10 ANALYSIS CONSIDERING DIFFERENT CARBON CONSUMPTION TARGETS, CI VALUES AND COMMUNICATION ENERGY EFFICIENCIES.

Scenario	δ [%]	N_b [bits]	$\mathrm{EE_{COM}}$ [Kbit/Joule]	I_0 [KgCO2-eq/kWh]	$I_{k>0}$ [KgCO2-eq/kWh]	Validation accuracy FA [%]	Validation accuracy CFA [%]	Carbon $\mathrm{C_{tot}^{(FA)}}$ [KgCO2-eq]	Carbon $\mathrm{C_{tot}^{(CFA)}}$ [KgCO2-eq]
1	10	16	10 - 100	0.449	0.449	45.8 - 74.8	54.8 - 64.9	1.5	1.5
	50	24	10 - 100	0.449	0.449	26.1 - 74.5	59.8 - 69.8	1.5	1.5
	100	32	10 - 100	0.449	0.449	17.4 - 73.1	52.4 - 66.3	1.5	1.5
2	10	16	10 - 100	0.449	0.449	70	70	3.6 - 0.6	15.6 - 4.2
	50	24	10 - 100	0.449	0.449	70	70	5.1 - 0.7	14.8 - 1.8
	100	32	10 - 100	0.449	0.449	70	70	8.6 - 1.1	27.3 - 2.9
1	10	16	10	0.449	0.193 - 0.952	60.8 - 45.8	54.8 - 54.8	1.5	1.5
	50	24	10	0.449	0.193 - 0.952	51.4 - 26.1	59.8 - 59.8	1.5	1.5
2	10	16	10	0.449	0.193 - 0.952	70	70	3.3 - 4.1	15.6 - 15.6
	50	24	10	0.449	0.193 - 0.952	70	70	4.3 - 6.8	14.8 - 14.8

of the energy/carbon demands in an iterative manner and accounts for computing/communication energy consumption as well as emissions arising from energy grids, allowing to identify the optimal operating conditions of FL processes and their energy expenditure. Experimental results are based on two increasingly complex image classification tasks and ML model sizes that serve as reference examples to study the optimal configuration of the compression parameters.

The analysis shows that CFA tools are more suited under energy-demanding communication protocols (i.e., when the communication energy efficiency is below 25 Kbit/Joule), while FA policies should be preferred under higher energy efficiency regimes. The optimization of quantization and sparsification operations in FA and CFA tools allows to further reduce the carbon footprint on average by roughly 50% and 42%, respectively, with respect to sending the ML model uncompressed. Besides, FA strategies have been shown to provide significant energy savings (up to 10%) when the PS is located in a region with a carbon efficiency (intensity) that is at least 2 times larger than that experienced by the devices.

Further research activities may target the integration of adaptive compression frameworks together with heterogeneous datasets characterized by multi-modal inputs, i.e., images, time series, or text data, to fully characterize the energy demands of FL processes and identify the optimal operating conditions.

REFERENCES

[1] P. Henderson, et al. "Towards the systematic reporting of the energy and carbon footprints of machine learning," Journal of Machine Learning Research, no. 21, pp. 1-43, 2020.
[2] P. Kairouz et al., "Advances and open problems in federated learning," Foundations and Trends in Machine Learning, Now Publishers, 2021. [Online]. Available: https://arxiv.org/abs/1912.04977.
[3] S. Savazzi, et al., "Opportunities of Federated Learning in Connected, Cooperative and Automated Industrial Systems," IEEE Communications Magazine, vol. 59, no. 2, pp. 16–21, 2021.
[4] X. Qiu, et al., "A first look into the carbon footprint of federated learning," 2021. [Online]. Available: https://arxiv.org/abs/2102.07627
[5] Z. Yang, et al. "Energy Efficient Federated Learning Over Wireless Communication Networks," IEEE Trans. on Wireless Communications, vol. 20, no. 3, pp. 1935-1949, 2021.
[6] L.F.W. Anthony, et al., "Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models," Proc of ICML Workshop on Challenges in Deploying and monitoring ML syst., 2020.
[7] K. Stechemesser "Carbon accounting: a systematic literature review," Journal of Cleaner Production, 36:17–38, 2012.
[8] Y. Lecun, et al., "Gradient-based learning applied to document recognition," in Proc. of the IEEE, vol. 86, no. 11, pp. 2278-2324, 1998.
[9] A. Krizhevsky, et al., "Cifar-10 (canadian institute for advanced research)." [Online]. Available: https://www.cs.toronto.edu/~kriz/cifar.html.
[10] W. Liu, et al., "Decentralized federated learning: Balancing communication and computing costs," IEEE Trans. on Signal and Information Processing over Networks, vol. 8, pp. 131–143, 2022.
[11] L. Barbieri, et al., "Decentralized federated learning for extended sensing in 6G connected vehicles," Vehicular Communications, vol. 33, 100396, 2022.
[12] L. Barbieri, et al., "A Layer Selection Optimizer for Communication-Efficient Decentralized Federated Deep Learning," IEEE Access, vol. 11, pp. 22155-22173, 2023.
[13] B. Camajori Tedeschini et al., "Decentralized Federated Learning for Healthcare Networks: A Case Study on Tumor Segmentation," in IEEE Access, vol. 10, pp. 8693-8708, 2022.
[14] S. Savazzi, et al., "An Energy and Carbon Footprint Analysis of Distributed and Federated Learning," in IEEE Trans. on Green Communications and Networking, vol. 7, no. 1, pp. 248-264, 2023.
[15] G. Yan, et al., "AC-SGD: Adaptively Compressed SGD for Communication-Efficient Distributed Learning," IEEE Journal on Selected Areas in Communications, vol. 40, no. 9, pp. 2678-2693, 2022.
[16] A. Koloskova, et al. "Decentralized Stochastic Optimization and Gossip Algorithms with Compressed Communication." Proc. of the 36th International Conference on Machine Learning, 97:3478-3487, 2019.
[17] European Environment Agency, Data and maps: "Greenhouse gas emission intensity of electricity generation," 2020. [Online]. Available: https://tinyurl.com/36l5v5ht
[18] ETSI TC EE, "ES 203 228, Environmental Engineering (EE); Assessment of mobile network energy efficiency," V1.3.1, 2020.
[19] E. Björnson et al. "How Energy-Efficient Can a Wireless Communication System Become?," Proc. 52nd Asilomar Conf. on Sig., Syst., and Comp., Pacific Grove, CA, USA, pp. 1252–1256, 2018.
[20] A. Lacoste, et al., "Quantifying the Carbon Emissions of Machine Learning," [Online]. Available: http://arxiv.org/abs/1910.09700.
[21] J. Huang, et al., "A close examination of performance and power characteristics of 4G LTE networks," Proc. of the 10th international conference on Mobile Systems, Applications, and Services, pp. 225–238, 2012.
[22] D. Alistarh et al., "QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding," in Advances in Neural Information Processing Systems, vol. 30, 2017.
[23] S. Shi et al., "Understanding Top-k Sparsification in Distributed Deep Learning," [Online]. Available: https://arxiv.org/abs/1911.08772.
[24] K. Simonyan, and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in International Conference on Learning Representations (ICLR), 2015.