To Talk or to Work: Energy Efficient Federated Learning over Mobile Devices via the Weight Quantization and 5G Transmission Co-Design

Rui Chen
rchen19@uh.edu
University of Houston
Houston, Texas, USA

Liang Li
liliang_1127@outlook.com
Xidian University
Xi’an, China

Kaiping Xue
kpxue@ustc.edu.cn
University of Science and Technology of China
China

Chi Zhang
chizhang@ustc.edu.cn
University of Science and Technology of China
China

Lingjia Liu
ljliu@vt.edu
Virginia Tech
Blacksburg, Virginia, USA

Miao Pan
mpan2@uh.edu
University of Houston
Houston, Texas, USA

ABSTRACT

Federated learning (FL) is a new paradigm for large-scale learning tasks across mobile devices. However, practical FL deployment over resource constrained mobile devices confronts multiple challenges. For example, it is not clear how to establish an effective wireless network architecture to support FL over mobile devices. Besides, as modern machine learning models are more and more complex, the local on-device training/intermediate model update in FL is becoming too power hungry/radio resource intensive for mobile devices to afford. To address those challenges, in this paper, we try to bridge another recent surging technology, 5G, with FL, and develop a wireless transmission and weight quantization co-design for energy efficient FL over heterogeneous 5G mobile devices. Briefly, the 5G featured high data rate helps to relieve the severe communication concern, and the multi-access edge computing (MEC) in 5G provides a perfect network architecture to support FL. Under MEC architecture, we develop flexible weight quantization schemes to facilitate the on-device local training over heterogeneous 5G mobile devices. Observed the fact that the energy consumption of local computing is comparable to that of the model updates via 5G transmissions, we formulate the energy efficient FL problem into a mixed-integer programming problem to elaborate the quantization strategies and allocate the wireless bandwidth for heterogeneous 5G mobile devices. The goal is to minimize the overall energy consumption (computing + 5G transmissions) over 5G mobile devices while guaranteeing learning performance and training latency. Generalized Benders’ Decomposition is applied to develop feasible solutions and extensive simulations are conducted to verify the effectiveness of the proposed scheme.

CCS CONCEPTS

- Computing methodologies → Distributed artificial intelligence: Neural networks;
- Theory of computation → Mixed discrete-continuous optimization.

KEYWORDS

5G networks, federated learning, weight quantization, optimization

ACM Reference Format:
Rui Chen, Liang Li, Kaiping Xue, Chi Zhang, Lingjia Liu, and Miao Pan. 2020. To Talk or to Work: Energy Efficient Federated Learning over Mobile Devices via the Weight Quantization and 5G Transmission Co-Design. In Mobihoc ’21: ACM International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing, June 03–05, 2021, Shanghai, China. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnn.nnnnnn

1 INTRODUCTION

Due to the incredible surge of mobile data and the growing computing capabilities of mobile devices, it becomes a trend to apply deep learning (DL) on these devices to support fast responding and customized intelligent applications [7]. Recently, federated learning (FL) is expected as a promising DL solution to provide an efficient, flexible, and privacy-preserving learning framework on a large scale of mobile devices. Under the FL framework [20], each mobile device executes model training locally and then transmits the model updates instead of raw data to an FL server. The server would aggregate the intermediate results and broadcast the updated model to the participating devices. Its potential has prompted wide applications in various domains such as keyboard predictions [12], physical hazards detection in smart home [35], health event detection [3], etc. However, it faces significant challenges to deploy FL over mobile devices in practice. First, although mobile devices are gradually equipped with artificial intelligence (AI) computing capabilities, the limited resources (e.g., battery and storage capacity) restrain them from training deep and complicated learning models. Second, it is not clear how to establish an effective wireless network architecture to support FL over mobile devices. Last but not least, the power-hungry local computing and wireless communications...
over iterations in FL may be too much for the power-constrained mobile devices to afford.

The mismatch between the computational and storage requirements of DL models and the limited resources of mobile devices becomes more challenging due to the increasing complexity of the state-of-art DL models. To enable on-device learning, the most popular approach is to compress a trained network\cite{11, 13, 18, 22}. Han et al.\cite{11} successfully applied multiple compression methods, e.g., pruning and quantization, to several state-of-the-art large-scale neural networks (e.g., AlexNet and VGG-16). Those works help to reduce the model complexity by orders of magnitude and speed up the model inference on mobile devices. However, on-device training is less explored and more difficult than its inference counterpart. Some pioneering works\cite{30, 31} make efforts on quantizing the model parameters to make it possible to conduct computationally efficient on-device training. Nevertheless, most existing compressed learning frameworks and the associated convergence analysis for the potential on-device training only consider the case of a single mobile device. A few works\cite{14} simply extend to the distributed learning setting by assigning the same compression strategy across differently mobile devices, which is not appropriate in the FL setting. Since FL encompasses massively distributed mobile devices that are highly heterogeneous, it is in dire need to develop a compression scheme to handle the device heterogeneity and investigate its impacts on learning performance.

Besides the on-device training for local computing, the energy consumption for FL over mobile devices also includes the wireless communications for the intermediate model exchanges. Particularly, with the advance of computing hardware and wireless communication techniques, we have observed that the energy consumption for local computing in FL is comparable to that for wireless transmissions on 5G mobile devices. For instance, the energy consumption of local computing (e.g., 0.06J for one modern GPU of one training iteration for AlexNet with batch size of 128\cite{25}) is comparable to that of wireless communications (e.g., 0.028J for transmitting and receiving 240MB Alexnet model parameters at 10 Gbps data rate\cite{1}). Thus, a viable design of the energy efficient FL over mobile devices has to consider the energy consumption of both “working” (i.e., local computing) and “talking” (i.e., wireless communications). However, most existing works in wireless communities have mainly carried out the radio resource allocation under the FL convergence constraints\cite{4, 24, 28, 29}, while ignoring to essentially reduce the energy consumption of learning algorithm itself. Moreover, among the previous works, the targeted learning models are either relatively simple (i.e., with convex loss functions) or shallow networks\cite{30, 31}. That is inconsistent with the current trend of the over-parameterized DL models. On the other hand, most efforts in the machine learning communities have focused on communication efficient FL algorithm designs, such as compressing the model updates or reducing the update frequency during the training stage. The basic assumption is that the wireless transmission data rate is slow, which leads to the bottleneck to support complicated learning models over mobile devices. Therefore, the goal of those designs is to reduce the number of communications in model updates without considering any nature or advances of wireless transmissions.

Fortunately, the recent surging technology, 5G wireless transmissions, featured by very high data rate (peak rate of 10 Gbps and average rate of 1 Gbps in 5G standard\cite{1}) with ultra low latency, can be leveraged to relieve the communication bottleneck with proper design. Furthermore, the multi-access edge computing (MEC), an important ingredient of 5G, enhances the computing capabilities at the edge networks\cite{16}, and hence provides an ideal architecture to support viable FL. In MEC, the 5G edge server/gNodeB can serve as the FL server, and 5G mobile devices can serve as local FL agents.

Motivated by the remaining challenges (i.e., inefficient on-device training and large overall energy consumption of FL training), in this paper, we try to bridge the 5G and FL, and develop a wireless transmission and on-device weight quantization co-design for energy efficient FL over heterogeneous 5G mobile devices. We aim to 1) facilitate efficient on-device training on heterogeneous local devices via a flexible quantization scheme, and 2) minimize the overall energy consumption of the FL learning process considering the learning performance and training latency. Based on the derived convergence analysis, we formulate the energy minimization problem to determine optimal quantization strategies and bandwidth allocations. Our salient contributions are summarized as follows.

- We propose a novel efficient FL procedure over 5G mobile devices to reduce the overall energy consumption in communication and computation. Briefly, subject to their current storage capacities, the mobile devices are allowed to compress the model and compute the gradients of the compressed version of the models. Meanwhile, for a given training time threshold, the network resource allocation is optimized to minimize the total computing and communication energy cost for FL training.
- To facilitate on-device training for FL over 5G heterogeneous mobile devices, weight quantization is employed to best utilize the limited storage capacities by representing model parameters with different bit-widths. We further provide the theoretical analysis of the convergence rate of FL with quantization in non-convex cases and obtain a closed-form expression for the novel convergence bound to explore the relationship among the training data heterogeneity, quantization error, and the performance of the FL algorithm.
- Based on the theoretical bound above, the energy minimization problem of FL training is formulated as a mixed-integer nonlinear programming (MINLP) problem to jointly determine the bandwidth allocation and local compression strategies for different 5G mobile devices. The Generalized Benders’ Decomposition algorithm is applied to seek for the feasible solutions.
- We evaluate the performance of the proposed method via extensive simulations using various real-world datasets and models to verify the effectiveness of our proposed scheme. Compared with different schemes, our proposed method shows great superiority in terms of energy efficiency for FL over heterogeneous 5G mobile devices.

The rest of this paper is organized as follows: In Section 2, a detailed description of the system model is presented. The convergence analysis, our optimization problem formulation and quantization selection algorithm are presented in Sections 3 and 4, respectively. In Section 5, the feasible solutions from the real datasets are
analyzed. The related work is discussed in Section 6 and the paper is concluded in Section 7.

2 PRELIMINARY AND MODEL DESCRIPTION

2.1 Preliminaries on Weight Quantization

In this subsection, we introduce the related concepts about weight quantization. Quantization is one attractive solution to implement DL models on mobile devices efficiently. It represents model parameters, including the weights, feature maps, and even gradients, with smaller bit-widths, which leads to smaller memory requirements. As the intermediate data is stored in lower-precision numerical formats (e.g., int8), the computational demand of complex neural networks is reduced, and the frequency of memory access becomes less. Hence, energy efficiency can be greatly improved.

To train the shared model with low-precision weight, we define a quantization function $Q(\cdot)$ to convert a real-valued number $w$ into a quantized version $\hat{w} = Q(w)$. We use the same notation for quantizing vectors as $Q$ acts on each dimension of the vector independently in the same manner. Moreover, we employ stochastic rounding (SR) [10] in our proposed model and analyze its convergence properties. SR, also known as unbiased rounding, possesses the important property such that $\mathbb{E}[Q(w)] = w$. This property avoids the negative effect of quantization noise, which is useful for the theory of non-convex setting [19]. For each component $w_i$ of a vector $w$, the function $Q(\cdot)$ converts the data type from the 32-bit full precision into $q$-bit integer, defined as:

$$Q(w_i) = s \cdot \text{sgn}(w_i) \cdot \begin{cases} M_{k+1}, & \text{w.p. } \frac{|w_i|}{2^{q_k}} - M_{k+1}, \\ M_k, & \text{w.p. } M_{k+1} - \frac{|w_i|}{2^{q_k}}, \end{cases}$$

where $\text{sgn}(\cdot)$ represents sign function, $s = ||w||_\infty$, denotes scaling factor, the index $k$ satisfies $M_k \leq |w_i| \leq M_{k+1}$, quantization set $S_w = \{-M_K, \cdots, M_0, \cdots, M_K\}$ with $K = 2^{q-1} - 1$, $0 = M_0 \leq M_1 \leq \cdots \leq M_K$ are uniformly spaced, and $\Delta_q$ denotes the quantization resolution as $\Delta_q = M_{k+1} - M_k = 1/(2^q - 1)$. Smaller resolution leads to a smaller gap and keeps as much information as the original weight, while it has higher memory requirements. In practice, the bit-width for the weight quantization can be extremely small, like 1 bit without notable performance degradation. Other parameters, such as the weight gradient calculations and updates, are applied to capture accumulate small changes in SGD. In contrast, the quantization makes them insensitive to such information and may impede convergence performance during training. Therefore, we keep a higher precision for the gradients than the weights and inputs so that the 5G server aggregates the local gradients and updates the global model in full precision.

2.2 Model Overview

We consider a MEC architecture in 5G networks consisting of one 5G edge server and a set $N = \{1, 2, \cdots, N\}$ of distributed mobile devices, collaboratively training a DNN model through FL framework, which is depicted in Fig. 1. Each 5G mobile device $i$ is equipped with a single antenna and has its own dataset $D_i$ with data size $|D_i|$. The data is collected locally by the edge device $i$ itself (e.g., all the photos a user takes and daily usage). Generally, each learning model has a particular non-convex loss function $f_i(w)$ with the parameter vector $w$ for each data sample $j$. The loss function represents the difference of the model prediction and groundtruth of the training data.

$$F(w) = \frac{1}{N} \sum_{i=1}^{N} f_i(w),$$

where $d$ denotes the learning model size.

![Figure 1: Federated edge learning system with MEC frameworks.](image)

Considering the sensitive nature of the users’ data, mobile devices keep the data locally instead of uploading to the 5G edge server. A FL framework [20] is adopted to solve problem (2). It requires multiple global training iterations (i.e., communication rounds) to achieve accuracy loss $\epsilon$. Let $r$ denotes the $r$-global-iteration of FL. In each global round, there are interactions between the mobile devices and edge server: At the $r$-th global training iteration, each 5G mobile device $i \in N$ downloads the latest model and computes the SGD over a mini-batch size of $M$, i.e., $g_i^{r-1} = \frac{1}{M} \sum_{j=1}^{M} f_{ij}(w^{r-1})$, where $f_{ij}$’s are randomly sampled from $(f_1, \cdots, f_{ID})$; To transmit their gradient results to the edge server, each 5G mobile device is allocated with a sub-channel by deploying Orthogonal Frequency-Division Multiple Access (OFDMA) scheme to avoid severe interference. Followed by the synchronous global aggregation protocol, the edge server will update the model parameters after aggregating all the gradient results from the mobile devices $N$. Then, the edge server broadcasts the latest model parameters $w'$ to the set of participating mobile devices $i \in N$ to minimize their $F_i(w)$ and derive the gradients in the next global iteration.

Targeting energy-efficient on-device FL training in 5G networks, we propose a flexible weight quantization (FWQ) algorithm for heterogeneous mobile devices. After the mobile devices receive the full-precision model, they would first compress it to satisfy their current storage budget. Unlike the prior works that maintain the same compression strategy across all the mobile devices, FWQ considers device heterogeneity and allows the mobile devices to perform weight quantization with different bit-widths of “q”. Note that the weight and gradient at the server side remain in full precision operations to avoid further model performance degradation. A pseudo-code of our FWQ algorithm is presented in Algorithm 1.
Algorithm 1 Flexible Weight Quantized Federated Learning (FWQ)

Input: $\eta$ = learning rate; $Q(\cdot)$ = quantization function; initial $w^0$; mini-batch size $M$ 

Output: $w^R$

1: for each global iteration $r = 0, \cdots, R - 1$ do 
2: Edge server sends $w^r$ to the set of participating mobile devices $\mathcal{N}$
3: for each mobile device $i \in \mathcal{N}$ in parallel do 
4: $\bar{w}_i^r \leftarrow Q_i(w^r)$ [store the weight in the low precision]
5: Sample mini-batch data set $(x_m, y_m)^{M}_{m=1}$ from $\mathcal{D}_i$
6: $g_i^r = \frac{1}{M} \sum_{m=1}^{M} f_m(\bar{w}_i^r)$ [obtain the gradient in the high precision]
7: Send $g_i^r$ to the 5G edge server
8: end for
9: 5G edge server updates $w^{r+1}$ as follows
10: $G^r \leftarrow \frac{1}{N} \sum_{i=1}^{N} g_i^r$
11: $w^{r+1} \leftarrow w^r - \eta G^r$ [update the weight in the high precision]
12: end for

3 CONVERGENCE ANALYSIS

In this section, we discuss the convergence of Algorithm (1) and find an upper bound of $\|\nabla F(w^R)\|_2^2$.

3.1 Assumptions

For the purpose of deriving theoretical guarantees, we state the following assumptions to the loss function.

**Assumption 1.** All the loss functions $f_j$ are differentiable and their gradients are $L$-Lipschitz continuous in the sense of 2-norm: for all $x$ and $y \in \mathbb{R}^d$, $\|f_j(x) - f_j(y)\|_2 \leq L \|x - y\|_2$.

**Assumption 2.** Assume that $\bar{f}_j$ is randomly sampled from $i$-th mobile device local loss functions. The stochastic gradient is unbiased estimator and its variance: $E \|\nabla \bar{f}_j(w^r) - \nabla f_j(w^r)\|_2^2 \leq \tau_i^2 \ , \forall i = 1, \cdots, N$.

**Assumption 3.** Assume that $\bar{f}_i$ is randomly sampled from the participating mobile devices. The variance of gradient among mobile devices is bounded, that is, $E \|\nabla \bar{f}_i(w^r) - \nabla F(w^r)\|_2^2 \leq \varphi^2$.

The first and second assumptions are commonly used for non-convex analysis of SGD [17, 34]. Assumption 3 captures the impact of different data distributions on each mobile device. The key ideas of the proofs firstly bound the difference between the local model in full precision $w^r$ and the quantized model $Q(w^r)$ in term of the changes of weight and gradient calculation. It is valuable to bound the expected sum-of-squares of model updates to guarantee convergence performance. Then we bound the expected terms in a non-convex case with some sufficient conditions to derive the convergence rate.

3.2 Main Results

**Lemma 1.** Suppose the Assumption 1 holds, the global loss function $F(w)$ is $L$-smooth.

**Proof.** Straightforwardly form Assumption, the definition of $F(w)$ and triangle inequality.

**Lemma 2.** As Assumption 2, it follows that

$$E \|\nabla F_i (w^r) - \nabla F_i (w^r)\|_2^2 \leq \frac{4L^2}{4} \delta_i^2.$$  (5)

**Proof.** From Assumption 2, the variance of the stochastic gradients in mobile device $i$ is bounded by $\tau_i^2$ and $E \|\nabla \bar{f}_j (Q_i (w^r)) - \nabla F_i (Q_i (w^r))\|_2^2$.

$$= \frac{1}{M^2} \sum_{j=1}^{M} E \|\nabla \bar{f}_j (Q_i (w^r)) - \nabla F_i (Q_i (w^r))\|_2^2 \leq \frac{1}{M^2} \tau_i^2.$$  (4)

**Lemma 3.** For any weight vector $w^r \in \mathbb{R}^d$, it holds that

$$E \|\nabla F_i (Q_i (w^r)) - \nabla F_i (w^r)\|_2^2 \leq \frac{4L^2}{4} \delta_i^2.$$  (5)

where $\delta_i = \lambda A_i$ is the quantization noise.

**Proof.** In [6], the authors have shown that for any vector $w^r \in \mathbb{R}^d$, it holds

$$E \|\nabla F_i (Q_i (w^r)) - \nabla F_i (w^r)\|_2^2 \leq \frac{d}{4} \delta_i^2.$$  (6)

Under the Assumption of $L$-smooth of $f$, we easily obtain

$$E \|\nabla F_i (Q_i (w^r)) - \nabla F_i (w^r)\|_2^2 \leq L^2 E \|\nabla F_i (Q_i (w^r)) - \nabla F_i (w^r)\|_2^2 \leq L^2 \frac{d}{4} \delta_i^2.$$  (7)

**Theorem 1** (Non-Convex Objective Analysis). Let Assumptions 1-3 hold, the convergence rate of the proposed scheme satisfies, with the quantization series $\{q_i\}_{i=1}^{N}$.

$$\eta = \frac{2L^n + 8\eta^L L^n \sum_{i=1}^{N} q_i^2 + 2L^n \tau + \eta^2}{8N \sum_{i=1}^{N} q_i^2 + \frac{4L^n \tau}{\eta^L}} \leq \frac{F^* - F(w^0)}{R \cdot H},$$  (8)

where $H = \frac{8N \sum_{i=1}^{N} q_i^2 + \frac{4L^n \tau}{\eta^L}}{\eta^L}$ and $F^*$ is the global minimum of objective $F$.

**Proof.** Please refer to the detailed proof in the Github.

Here, the average expected squared gradient norm is utilized to characterize the convergence rate due to the non-convex objective in modern learning models. Given the generic result in Theorem 1, we obtain the novel convergence rate by appropriately choosing the learning rate $\eta$.

https://github.com/ansdfla1/blob/master/proof.pdf
Corollary 1. If we choose the learning rate \( \eta \) as
\[
\eta = \frac{1}{4L + \sqrt{\frac{4R}{MN} + \varphi \sqrt{R}}},
\]
we have the following convergence rate,
\[
1 \sum_{r=0}^{R-1} \mathbb{E}[\|F(w^r)\|^2_2] 
\leq \frac{4LK}{R} + \frac{9dL^2}{N} \sum_{i=1}^{N} \delta_i^2 + \frac{(K + 4L)\sqrt{\varphi}}{\sqrt{MN}R} + \frac{(K + 8L)\varphi}{R},
\]
where \( K = 4\mathbb{E}[F(w^0)] - \mathbb{E}[F^*] \).

Proof. Given the choice of \( \eta = \frac{1}{4L + \sqrt{\frac{4R}{MN} + \varphi \sqrt{R}}} \), we have \( \frac{1}{2}\eta \leq \eta(1 - 2\eta L) \leq \eta \). Also, from Theorem 1, we obtain
\[
\left( \frac{\eta}{2} - 2\eta L \right) \sum_{r=0}^{R-1} \mathbb{E}[\|F(w^r)\|^2_2] \leq F(w^0) - F^* + R \cdot H,
\]
followed by
\[
1 \sum_{r=0}^{R-1} \mathbb{E}[\|F(w^r)\|^2_2] 
\leq \frac{2\mathbb{E}[F(w^0)] - \mathbb{E}[F^*]}{\eta(1 - 2\eta L)R} + \frac{2nL}{(1 - 2\eta L)} \left( \frac{\tau}{MN} + 2\varphi^2 \right) 
+ \frac{(1 + 8nL)dL^2}{2N(1 - 2\eta L)} \sum_{i=1}^{N} \delta_i^2 
\leq \frac{4\mathbb{E}[F(w^0)] - \mathbb{E}[F^*]}{\eta R} + 4\eta L \left( \frac{\tau}{MN} + 2\varphi^2 \right) + \frac{9dL^2}{N} \sum_{i=1}^{N} \delta_i^2 
\leq \frac{4\mathbb{E}[F(w^0)] - \mathbb{E}[F^*]}{R} \left( 4L + \sqrt{\frac{4R}{MN} + \varphi \sqrt{R}} \right) 
+ \frac{4L \sqrt{\tau}}{\sqrt{MN}} + \frac{8L\varphi}{\sqrt{R}} + \frac{9dL^2}{N} \sum_{i=1}^{N} \delta_i^2 
\leq \frac{4dL^2}{N} \sum_{i=1}^{N} \delta_i^2 + 16L \mathbb{E}[F(w^0)] - \mathbb{E}[F^*] 
+ \frac{(4\mathbb{E}[F(w^0)] - \mathbb{E}[F^*]) + 4L \sqrt{\tau}}{\sqrt{MN}R} 
+ \frac{(4\mathbb{E}[F(w^0)] - \mathbb{E}[F^*]) + 8L \varphi}{\sqrt{R}}.
\]

4 OPTIMIZATION FOR ENERGY EFFICIENT FL TRAINING

Motivated by the above discussion, the quantization strategies \( (q_i)_{i=1}^{N} \) act as key design parameters to balance the trade-off between per round resource consumption and training performance. In this section, our goal is to minimize the energy consumption during the overall training process with model performance and training delay guarantee. We start by discussing the computation and communication energy model. Later, problem formulation and solution are presented.

4.1 Energy Model

4.1.1 Computation Model. GPUs have become one key attributor for building high performance deep learning models [15]. Its powerful computing capacities and massively parallel architecture can efficiently handle compute-intensive matrix multiplication and other operations that significantly accelerate the training process. Despite that, GPUs are power hungry for computing/memory intensive tasks. Thus, the quantization is employed on mobile GPU in the training phase to save energy consumption. Noted that the local computation of mobile device \( i \) involves the data fetching in GPU memory modules and the arithmetic in GPU core modules, where the voltage and frequency of each module are independent and configurable.

1) GPU runtime power model of 5G mobile device \( i \) is modeled as a function of the core/mem voltage/frequency [21],
\[
F_i^\text{comp} = F_i^\text{core} + \gamma_i^\text{mem} \cdot w_i + \gamma_i^\text{core}(\nu_i^\text{core})^2 w_i^\text{core},
\]
where $P^G_{t^0}$ is the summation of the power consumption unrelated to the GPU voltage/frequency scaling; $V^\text{core}_{i}, f^\text{core}_{i}, f^\text{mem}_{i}$ denote the GPU core voltage, GPU core frequency, and GPU memory frequency respectively; $q^\text{mem}_{i}$ and $q^\text{core}_{i}$ are the constant coefficient that depends on the hardware and arithmetic for one training iteration, respectively.

2) **GPU execution time model of 5G mobile device $i$ with quantization level $q_{i}$** is formulated as,

$$T^\text{comp}_{i}(q_{i}) = t^0_{i} + \frac{c_{1}(q_{i})q^\text{mem}_{i}}{f^\text{mem}_{i}} + \frac{c_{2}(q_{i})q^\text{core}_{i}}{f^\text{core}_{i}},$$  \hspace{1cm} (17)

where $t^0_{i}$ represents the other component unrelated to training task; $q^\text{mem}_{i}$ and $q^\text{core}_{i}$ denote the number of cycles for 5G mobile device $i$ to fetch data and to compute one mini-batch size of data samples, respectively, which is measured on a platform-based experiment in this work. Due to the weight quantization, the number of cycles for data fetching and computing are reduced with scaling $c_{1}(q_{i})$ and $c_{2}(q_{i})$, respectively. For simplicity, we assume that the number of cycles for data fetching and computing scales $c_{1}(q_{i})$ and $c_{2}(q_{i})$ are the linear functions of data bit-width $q_{i}$, respectively. This is reasonable since the quantization reduces the bit-widths, and the data size scales linearly to the bit representation [32].

With the above GPU power and performance model, the local energy consumed to pass a single mini-batch SGD with quantization $q_{i}$ of $i$-th 5G mobile device is the product of the runtime power and the execution time, such that,

$$E^\text{comp}_{i}(q_{i}) = P^\text{comp}_{i} \cdot T^\text{comp}_{i}(q_{i}).$$  \hspace{1cm} (18)

4.1.2 **Communication Model.** We consider OFDMA protocol for mobile devices to upload their local results to the edge server. The total channel bandwidth $B_{\text{max}}$ of edge server is divided into sub-channel of $B_{i,r}$ each, and each sub-channel is allocated to user terminals in slots. As a result, the achievable transmission rate (bit/s) of 5G mobile device $i$ can be calculated as,

$$y_{i,r} = B_{i,r} \ln \left(1 + \frac{h_{i,r} P^\text{comm}_{i}}{\sigma^2}\right),$$  \hspace{1cm} (19)

where $B_{i,r}$ is allocated bandwidth, $\sigma^2$ represents the noise power, $P^\text{comm}_{i}$ is the transmission power. Here, $h_{i,r}$ denotes the channel gain of the mobile device $i$ and would change in each global iteration $r$. Such information can be estimated in prior, and channel estimation in OFDMA systems has been thoroughly studied in the literature, which is out of our paper scope. The dimension of the gradient vector $g_{i}^r$ is fixed for a given model so that the overall data size to transmit the gradient vector is the same for all the 5G mobile devices, which is denoted by $D_{g}$. Then, the communication time to transmit $D_{g}$ for mobile device $i$ is

$$T^\text{comm}_{i}(B_{i,r}) = \frac{D_{g}}{y_{i,r}} = \frac{D_{g}}{B_{i,r} \ln \left(1 + \frac{h_{i,r} P^\text{comm}_{i}}{\sigma^2}\right)}.$$  \hspace{1cm} (20)

Thus, the communication energy consumption of 5G mobile device $i$ can be derived,

$$E^\text{comm}_{i}(B_{i,r}) = P^\text{comm}_{i} \cdot T^\text{comm}_{i}(B_{i,r}) = \frac{D_{g} P^\text{comm}_{i} B_{i,r}}{\ln \left(1 + \frac{h_{i,r} P^\text{comm}_{i}}{\sigma^2}\right)}.$$  \hspace{1cm} (21)

4.2 **Problem Formulation**

Considering the computing capabilities of different mobile devices varies, we formulate the optimization problem to minimize the total energy consumption during the training process as,

$$\min_{q,B} \sum_{r=1}^{R} \sum_{i=1}^{N} E^\text{comm}_{i}(B_{i,r}) + E^\text{comp}_{i}(q_{i})$$  \hspace{1cm} (22)

s.t. \hspace{1cm} (23) - (29)

where $R$ is the total number of global iterations, $U_{i}$ and $C_{i}$ represent the learning model size (MB) stored in full precision and memory capacity in the 5G mobile device $i$, respectively. $c_{s}(q_{i})$ is the ratio of the bit-width to full precision. $q = \{q_{1}, \ldots, q_{N}\}$ represent the quantization strategies of mobile devices and $B_{r} = \{B_{1,r}, \ldots, B_{N,r}\}$ are the resource allocation at $r$-th global iteration and long-term allocation strategies denote as $B = \{B_{r}\}_{r=1}^{R}$. In constraint (24), the bandwidth allocation to the mobile devices must not exceed the channel bandwidth available to the edge server in any global iteration. Constraint (25) states the model size stored on the mobile device $i$ does not exceed the capacity of the device $i$. The constraints in (27) and (26) ensure the entire training time can be completed within a predefined deadline $T_{\text{max}}$. Constraints (28) and (29) indicate that variables take the values from a set of non-negative numbers. Bit representation is set to be a power of 2, ranging from 8 to 32 bits, which is a standard setting and hardware friendly [27]. The number of global iterations $R$ is determined by the FL model convergence. We set it as a large constant number. The necessary number of iterations to converge is the same with and without weight quantization due to the derivation in the previous section. However, Corollary 1 illustrates that the model with weight quantization cannot converge to the original optimal model $w^{\star}$ and the unavoidable error degrades the final performance. Hence, the constraint (23) controls such negative impact to be smaller than a given tolerance level. $\delta_{i} = \frac{s_{i}}{2^{q_{i}}} \text{ and } \epsilon_{2}$ is constant used to approximate the big-$O$ in (10).

For convenience, we denote the $a^1_{i,r} = \frac{D_{g} P^\text{comm}_{i} B_{i,r}}{\ln \left(1 + \frac{h_{i,r} P^\text{comm}_{i}}{\sigma^2}\right)}$ and $a^2_{i,r} = \frac{D_{g}}{\ln \left(1 + \frac{h_{i,r} P^\text{comm}_{i}}{\sigma^2}\right)}$, and reformulate the problem (22)-(29) as:

$$\min_{q,B} \sum_{r=1}^{R} \sum_{i=1}^{N} \frac{a^1_{i,r}}{B_{i,r}} + \frac{P^\text{comp}_{i} \cdot T^\text{comp}_{i}(q_{i})}{\ln \left(1 + \frac{h_{i,r} P^\text{comm}_{i}}{\sigma^2}\right)}$$  \hspace{1cm} (30)

s.t. \hspace{1cm} (23) - (29).  \hspace{1cm} (31)
4.3 Generalized Benders’ Decomposition for Solutions

The optimization problem in (30)-(31) is a MINLP. Furthermore, the continuous decision variables, $B_{i,r}$, are coupled with the integer decision variables $q_i$ in constraint (26). Therefore, we utilize the generalized Benders’ decomposition (GBD) algorithm [9] to solve the proposed problem. It decomposes the optimization problem (30)-(31) into a master problem and a primal problem. Specifically, the primal problem is a convex problem involving variables $B_{i,r}$ and its solution for given $q$ yields an upper bound for the optimal values of (30)-(31). The master problem is a mix integer problem and the corresponding solution provides a lower bound for the optimal values of (30)-(31). Accordingly, the master and primal problems are solved iteratively until their solutions converge. In the following, we first describe such two problems for a single round, and then we describe the interaction of them between two consecutive rounds. For convenience, we further simplify the description of GPU time model as $T_{i}^{\text{comp}}(q_i) = \beta_2 q_i + \beta_1$, where $\beta_1 = t_i^g$ and $\beta_2 = \omega_1^{\text{mem}}/t_i^g + \omega_2^{\text{core}}/t_i^g$.

1) Primal Problem: At $(z)$-th round, for the given integer parameters $\bar{q} = q^{(z)}$, we relax constraint (26) and introduce a new variable, $T_r$, indicating per global iteration latency, to get the following primal problem of communication cost:

$$
\nu(\bar{q}) = \min_{B,T} \sum_{i=1}^{N} \sum_{r=1}^{R} \frac{\alpha_{i,r}^1}{B_{i,r}} + p_i^{\text{comp}} \cdot (\beta_2 q_i + \beta_1) \quad (32)
$$

s.t. (24), (28),

$$
T_r - \frac{\alpha_{i,r}^2}{B_{i,r}} \geq (\beta_2 q_i + \beta_1), \forall i, r,
$$

$$
\sum_{r=1}^{R} T_r \leq T_{max}. \quad (34)
$$

The primal problem is a convex optimization problem [2] concerning the optimization variables and can be solved by the standard convex optimization algorithm, e.g., interior point method [23]. The Lagrangian duality $L_1$ of the primal problem is formulated as follow:

$$
L_1(B, T, \bar{q}, \omega_1^1, \omega_2^1, \omega_3)
= \sum_{i=1}^{N} \sum_{r=1}^{R} \frac{\alpha_{i,r}^1}{B_{i,r}} + p_i^{\text{comp}} \cdot (\beta_2 q_i + \beta_1)
+ \sum_{r=1}^{R} \frac{\alpha_{i,r}^2}{B_{i,r}} (T_r - \beta_2 q_i - \beta_1) + \omega_3 \left( T_{max} - \sum_{i=1}^{N} B_{i,r} \right)
\quad (35)
$$

where $\omega_1^1$, $\omega_2^1$ and $\omega_3$ represent the Lagrangian multipliers associated with constraints (24), (33) and (34), respectively. If the primal problem (32)-(34) is feasible, the optimal solutions of $T = T^{(z)^*}$, $B = B^{(z)^*}$ and the Lagrange multipliers of the primal problem be $\omega_1^{(z)^*}$, $\omega_2^{(z)^*}$ and $\omega_3^{(z)^*}$ at the $z$-th round. The strong duality holds due to convexity of the primal problem, and the solution of the dual problem is the solution of the problem (32)-(34). The optimal objective value of the primal problem is the valid upper bound of problem (30)-(31). The corresponding optimal cut is added as a new constraint into the master problem for the next round.

If the primal problem (32)-(34) is infeasible for given integer parameters $\bar{q}$, we formulate the corresponding feasibility problem as $l_1$-minimization problem:

$$
\min_{B_i,r, \nu_1^1, \nu_2^1, \nu_3^1, T_r} \sum_{r=1}^{R} \nu_1^1 + \sum_{i=1}^{N} \nu_2^2 + \nu_3^3 \quad (36)
$$

s.t. $\sum_{r=1}^{R} B_{i,r} - \nu_1^1 \leq B_{max}, \forall r,$

$$
\frac{\alpha_{i,r}^2}{B_{i,r}} - \nu_2^2 \leq T_r - \beta_2 q_i - \beta_1, \forall i, r \quad (38)
$$

$$
\sum_{r=1}^{R} T_r - \nu_3^3 \leq T_{max}. \quad (39)
$$

$B_{i,r}, 0, T_r \geq 0, \nu_1^1, \nu_2^2, \nu_3^3 \geq 0, \forall r, i. \quad (40)$

where $\nu_1^1$, $\nu_2^2$, $\nu_3^3$ are auxiliary variables to quantify the constraint violation. To solve problem (36)-(40), the solutions of $\bar{B}$ and $\bar{T}$, and the corresponding Lagrange multipliers $\nu_1^1, \nu_2^2, \nu_3^3$ are generated as the following feasible cut in the master problem for the next iteration,

$$
L_2(\bar{B}, \bar{T}, \bar{q}, \nu_1^1, \nu_2^2, \nu_3^3) \leq 0, \quad (41)
$$

where

$$
L_2(\bar{B}, \bar{T}, \bar{q}, \nu_1^1, \nu_2^2, \nu_3^3)
= \sum_{r=1}^{R} \nu_1^1 \sum_{i=1}^{N} B_{i,r} - B_{max} + \nu_3^3 \left( \sum_{r=1}^{R} T_r - T_{max} \right)
+ \sum_{r=1}^{R} \sum_{i=1}^{N} \nu_2^2 \left( \frac{\alpha_{i,r}^2}{B_{i,r}} + \beta_2 q_i + \beta_1 - T_r \right). \quad (42)
$$

2) Master Problem: Based on the optimal cut and the feasible cut, the master problem to solve the binary variables in the $z$-th iteration is formulated as

$$
\min_{\phi} \phi \quad (43)
$$

s.t. $\phi \geq L_1(\bar{B}^{(z)}, \bar{T}^{(z)}, \bar{q}, \omega_1^{(z)}, \omega_2^{(z)}, \omega_3^{(z)}), \forall \nu \in OC, \quad (44)

\geq L_2(\bar{B}^{(z)}, \bar{T}^{(z)}, \bar{q}, \nu_1^{(z)}, \nu_2^{(z)}, \nu_3^{(z)}), \forall \nu \in FC, \quad (45)

(25), (29), (23), \quad (46)

where $OC = \{\nu | \nu < z \text{ and primal problem is infeasible} \}$ and $OC = \{\nu | \nu < z \text{ and primal problem is feasible} \}$ are defined as the sets of all the round indices for feasible cut and optimal cut, respectively. As the master problem is a small-scale mixed integer programming problem, the problem (43)-(46) can be solved optimally by branch and bound methods. At each iteration, the master problem has one additional constraint compared to those defined in the previous iteration. Thus, the newly obtained optimum of the master problem is always less than or equal to the previous one and therefore this upper bound is always non-increasing. The proposed algorithm is guaranteed to converge to the feasible solution. The overall algorithm can be summarized in Algorithm 2.
5 PERFORMANCE EVALUATION

In this section, we evaluate the performance of the proposed energy efficient FL via simulations.

5.1 Data and settings

1) Learning Model and Dataset: To test the model performance, we choose two commonly-used deep learning models: ResNet-34 and Mobilenet. The well-known datasets, CIFAR-10 and CIFAR-100, are used to train FL models for image classification tasks. The training data is distributed over different mobile devices in the non-i.i.d setting.

2) Communication and Computing Models: For communication model, we assume the noise power is \( N_0 = -174 \text{ dBm} \). The maximum transmitting power of each user ranges from 2 dBm to 20 dBm [33]. For the GPU computing model, the scaling factors of energy consumption are measured by Nvidia profiling tools on RTX8000.

3) Several Comparing Schemes: In order to show the advantages of our proposed FWQ algorithm, we compare it with other existing algorithms: 1) Full Precision: All mobile devices compute their local models with the full precision operation, i.e., without quantization, and upload the local results using the allocated bandwidth. 2) Unified Q: All the devices are set to use the same quantization strategy regardless of various budgets of different mobile devices. 3) Rand Q: All mobile devices choose a quantization level randomly without considering the learning performance. The resource allocations strategies are optimized by solving a simplified version of the problem (32).

5.2 Convergence analysis

First, convergence performance is studied. We implement the above learning model and use a large batch-size in FL training. Here, we assign the bit-width as 16 bits in "Unified Q" scheme. Fig. 2(a) and Fig. 2(c) show the model performance in terms of testing accuracy. We observe that the models trained by "FWQ", "Unified Q" and "Rand Q" are inferior to the "Full Precision" scheme, and the "Rand Q" has the worst performance. That is consistent with our convergence analysis that the discretization error induced by the quantization is unavoidable. This error is accumulated by all the participating mobile devices, which indicates some mobile devices take aggressive quantization strategies (e.g., 8 bit) due to their resource limitation. As for the proposed "FWQ", since it considers this error in the quantization selection, the degradation is well controlled and relatively small. The corresponding energy consumption of the FL training is presented in Fig. 2(b) and Fig. 2(d). It demonstrates the effectiveness of the proposed scheme. Since the quantization reduces the memory costs, the training energy is smaller than that in "Full Precision". Besides, "FWQ" enhances the energy efficiency and consumes x2 - x100 less energy than the other three schemes in the FL training process, which is essential for battery-limited mobile devices.

5.3 Impact of Users, Computing, and Communication Capacities

To show how the number of users affects the total energy consumption, we vary the number of users from 2 to 35. Fig. 3 shows that the average energy consumption decreases with more mobile devices participating in FL. The costs are unchanging when the number of users exceeds a certain level in all the schemes. The reason is that more users can enrich the training dataset and help to speed up the model convergence and thus reduce energy consumption. Once the model obtains sufficient information from a large enough number of mobile devices, the training iteration number becomes a constant to maintain the learning accuracy. Besides, the proposed "FWQ" outperforms the others because it reduces more local computing energy costs than "Full Precision" and "Rand Q". Moreover, "FWQ" is more flexible than the "Unified Q" since it enables participating mobile devices with different resource budgets to choose their own suitable quantization levels to further reduce the computing energy consumption.

We evaluate the impact of device heterogeneity concerning computing capability. Here, we keep the number of mobile devices as 10 and divide them into 4 groups. Fixing the minimum capacity as 1400MHz, we set different capacities into 4 groups such that \( C = 2000 \text{ MHz}, C + 500 \text{ MHz}, C + 1500 \text{ MHz}, \) and \( C + 2000 \text{ MHz} \), respectively. The values of \( L \) range from 0 to 10. Large \( L \) means high heterogeneity among mobile devices and implies that the optimized quantization strategy has more diverse values. From Fig. 4, we observe that the energy costs grow with the device heterogeneity, indicating a negative impact of heterogeneous mobile devices on FL training. As expected, the proposed methods achieve much lower energy consumption than the other three schemes since we consider the differences among mobile devices and optimize the quantization selection.
To Talk or to Work: Energy Efficient Federated Learning over Mobile Devices

Mobihoc ’21, June 03–05, 2021, Shanghai, China

Figure 2: Convergence Analysis. (a)-(b): Resnet-34 on CIFAR-100. (c)-(d): Mobilenet on CIFAR-10

Figure 3: Energy vs numbers of devices. Figure 4: Energy vs device heterogeneity. Figure 5: Quantization vs bandwidth.

Figure 5 shows the impacts of the wireless conditions on the optimal quantization selection. We vary the total available bandwidth from 20 MHz to 38 MHz and divide the mobile devices into 4 different groups, denoted as \( g_1, g_2, g_3, g_4 \), where \( h(g_1) \leq h(g_2) \leq h(g_3) \leq h(g_4) \). From Fig. 5, we see that, as the overall bandwidth becomes small, the ratio of the communication energy consumption to the overall energy consumption grows, which means wireless communications have a larger impact on the total energy consumption than local computing. As a result, the mobile devices in group 1, with small channel gain, become the stragglers in FL training and could slow down the gradient update time for one iteration. To avoid prolonging the update time for the next iteration and reduce the overall energy consumption, they have to take aggressive actions to compress their local models into the smallest bits. In contrast, it results in large discretization noise and degrades the performance, as stated in Corollary 1. To compensate for that, those who have better channel gain need to “work” more by using a higher precision model to perform local training. Similarly, when the available bandwidth increases, the computing contributes more to the overall energy consumption. Those mobile devices with smaller local computing capacities choose to compress their models more to save computing energy.

6 RELATED WORK

Several research efforts are carried out on the FL resource consumption optimization in wireless networks in both computing and communication [26, 28–31]. Particularly, to address the straggler problem, Shi et al. in [26] proposed a device scheduling scheme to balance the trade-off between the training rounds and per round delay, including computation and communication. In [29], Vu et al. leveraged cell-free massive MIMO to support FL and minimize the training time. Tran et al. in [28] studied the trade-offs between the FL training latency and energy cost in wireless communities and determine the optimal computing and wireless resource allocation strategies considering the heterogeneity of device capacities. However, most of the works omit to remove the redundant resource consumption from the learning algorithm itself.

Various works have been developed for on-device learning to reduce the model complexities via low precision operation and storage requirements. In the extreme case, the weights and activations are represented in one bit, called Binary Neural Networks (BNN) [5], while the performance degrades significantly in large DNNs. For weight quantization, the prior work such as “LQ-Net” in Zhang et al. [36] quantized weights and activations such that the inner products can be computed efficiently with bit-wise operations, performing in the case of single machine computation. Similar to our work, Fu et al. [8] considered the weight quantization for local devices in the distributed learning setting and proposed to quantize activations via estimating Weibull distributions. However, they assigned the same quantization level across different participating devices and limits the performance when facing the challenges of device heterogeneity. Unlike these existing works, in our proposed model, a mobile-compatible FL algorithm with flexible weight quantization strategies is introduced. By jointly considering the device and wireless heterogeneity, we formulate an optimization problem to obtain the optimal strategies and minimize the overall computing and communication energy consumption.
7 CONCLUSION
In this paper, we have studied the energy efficiency of FL training via 5G transmission and quantization. We have jointly exploited the flexible weight quantization selection and the bandwidth allocation to develop an energy efficient FL training optimization over 5G heterogeneous mobile devices, constrained by the training delay and learning performance. The weight quantization approach has been leveraged to deal with the mismatch between the high model computing complexity and the limited computing capacities of mobile devices. The convergence bound for FL with local quantization has been analyzed in non-convex cases. Guided by the derived theoretical bound, the energy efficient FL training problem is formulated as a MI-NLP, and the GBD algorithm has been applied to obtain the feasible solutions. By comparing with different compression strategies through extensive simulations, we have demonstrated the effectiveness of our proposed scheme in handling device heterogeneity and reducing overall energy consumption of FL over heterogeneous 5G mobile devices.

REFERENCES
[1] 3GPP. 2019. Technical Specification Group Services and System Aspects; Release 15 Description; Summary of Rel-15 Work Items. Technical Specification (TS) 21.915. 3rd Generation Partnership Project (3GPP). https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3389 Version 2.0.0.
[2] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. 2004. Convex optimization. Cambridge university press.
[3] Theodora S Brismi, Rudi Chen, Theofanie Mela, Al Oleshvsky, Ioaniss Ch Paschalidis, and Wei Shi. 2018. Federated learning of predictive models from federated electronic health records. International journal of medical informatics 112 (2018), 59–67.
[4] Mingzhe Chen, Zhaohui Yang, Walid Saad, Changchuan Yin, H Vincent Poor, and Shuguang Cui. 2019. A joint learning and communications framework for federated learning over wireless networks. arXiv preprint arXiv:1909.07972 (2019).
[5] Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. 2016. Binarized neural networks: Training deep neural networks with weights and activations constrained to+1 or −1. arXiv preprint arXiv:1602.02830 (2016).
[6] Christopher De Sa, Megan Lezeczyński, Jian Zhang, Alana Marzoev, Christopher R Abberger, Kuneil Okotokin, and Christopher Ré. 2018. High-accuracy low-precision training. arXiv preprint arXiv:1803.03383 (2018).
[7] Jiashao Ding, Xinyue Zhang, Xiaohuan Li, Jinyi Wang, Rong Yu, and Miao Pan. 2020. Differentially private and fair classification via calibrated functional mechanism. In Proceedings of the AAAI Conference on Artificial Intelligence. New York, NY, 626–629.
[8] Fangcheng Fu, Yuzheng Hu, Yihan He, Jiawei Jiang, Yingxia Shao, Ce Zhang, and Bin Cui. 2020. Don’t waste your bits! squeeze activations and gradients for deep neural networks via TINYSQI. In International Conference on Machine Learning. PMLR, virtual, 3304–3314.
[9] Arthur M Geoffrion. 1972. Generalized benders decomposition. Journal of optimization theory and applications 10, 4 (October 1972), 237–260.
[10] Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint arXiv:1510.0049 (2015).
[11] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sen Augustin, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. 2018. Federated learning for mobile keyboard prediction. arXiv preprint arXiv:1813.03604 (2018).
[12] Zhiwu He, Xiangyu Zhang, and Jin Sun. 2017. Channel pruning for accelerating very deep neural networks. In Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 1389–1397.
[13] Lu Hou, Ruiliang Zhang, and James T Kwok. 2018. Analysis of quantized models. In International Conference on Learning Representations. Vancouver, Canada.
[14] Loc N Huynh, Youngki Lee, and Rajesh Krishna Balan. 2017. Deepmop: Mobile gpu-based deep learning framework for continuous vision applications. In Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services. Niagara Falls, NY, 82–95.
[15] Liang Li, Dian Shi, Ronghui Hou, Hui Li, Miao Pan, and Zhu Han. 2021. To Talk or to Work: Flexible Communication Compression for Energy Efficient Federated Learning over Heterogeneous Mobile Edge Devices. In Proc. IEEE Conference on Computer Communications (INFOCOM). Virtual Conference.
[16] Xiang Li, Kaixuan Huang, Wenhao Yang, Shunen Wang, and Zhihua Zhang. 2019. On the Convergence of FedAvg on Non-IID Data. In International Conference on Learning Representations (ICLR). New Orleans, Louisiana.
[17] Yuhang Li, Xin Dong, and Wei Wang. 2020. Additive-Powers-of-Two Quantization: An Efficient Non-uniform Discretization for Neural Networks. In International Conference on Learning Representations (ICLR). Virtual Conference.
[18] Zheng Li and Christopher M De Sa. 2019. Dimension-Free Bounds for Low-Precision Training. In Advances in Neural Information Processing Systems (NIPS). Vancouver, BC, Canada.
[19] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguerre y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In Artificial Intelligence and Statistics (AISTATS). Ft. Lauderdale, FL.
[20] Xinxin Mei, Xiaowen Chu, Hai Liu, Yu-Wing Leung, and Zongpeng Li. 2017. Energy efficient real-time task scheduling on CPU-GPU hybrid clusters. In IEEE Conference on Computer Communications (INFOCOM). Atlanta, GA, USA.
[21] Bert Moons, Koen Goetschalckx, Nick Van Berckelaer, and Mariam Verhelst. [n.d.]. Minimum energy quantized neural networks. In 2017 51st Asilomar Conference on Signals, Systems, and Computers. IEEE, 1921–1925.
[22] Yurii Nesterov and Arkadii Nemirovskii. 1994. Interior-point polynomial algorithms in convex programming. SIAM.
[23] Takayuki Nishio and Ryo Isonetani. 2019. Client selection for federated learning with heterogeneous resources in mobile edge. In IEEE International Conference on Communications (ICC). Shanghai, China.
[24] Hang Qi, Evan R, Sparks, and Ameet Talwalkar. 2017. Paleo: A Performance Model for Deep Neural Networks. In Proceedings of the International Conference on Learning Representations (ICLR). Toulon, France.
[25] Renqi Shi, Sheng Zhou, Zhisheng Niu, Miao Jiang, and Lu Geng. 2020. Joint device scheduling and resource allocation for latency constrained wireless federated learning. IEEE Transactions on Wireless Communications (2020).
[26] Emanuele Torti, Alessandro Fontanella, Miro Mucci, Nicola Blago, Danilo Pau, Francesco Leporati, and Marco Piastra. 2018. Embedded real-time fall detection with deep learning on wearable devices. In 2018 21st euromicro conference on digital system design (DSD). IEEE, Prague, Czech Republic, 405–412.
[27] Nguyen H Tran, Wei Bai, Albert Zomaya, Nguyen Minh NL, and Chong Seon Hong. 2019. Federated learning over wireless networks: Optimization model design and analysis. In IEEE Conference on Computer Communications (INFOCOM). Paris, France.
[28] Tung T Vu, Duy T Ngo, Nguyen H Tran, Hien Quoc Ngo, Minh N Dao, and Richard H Middleton. 2020. Cell-free massive MIMO for wireless federated learning. IEEE Transactions on Wireless Communications 19, 10 (2020), 6577 – 6592.
[29] Shiqiang Wang, Tiffany Tuoer, Theodoros Saloudis, Kim K Leung, Christian Makaya, Ting He, and Kevin Chan. 2019. Adaptive federated learning in resource-constrained edge computing systems. IEEE Journal on Selected Areas in Communications 37, 6 (2019), 1205–1221.
[30] Kai Yang, Tao Jiang, Yuanming Shi, and Zhi Ding. 2020. Federated learning via over-the-air computation. IEEE Transactions on Wireless Communications 19, 3 (2020), 2022–2035.
[31] Tian-Ju Yang, Yu-Hsin Chen, Joel Eren, and Vivienne Sze. 2017. A method to estimate the energy consumption of deep neural networks. In 2017 51st asilomar conference on signals, systems, and computers. IEEE, Pacific Grove, CA.
[32] Zhaohui Yang, Mingzhe Chen, Walid Saad, Chong Seon Hong, and Mohammad Shiloh-Bahari. 2020. Energy efficient federated learning over wireless communication networks. IEEE Transactions on Wireless Communications (Early Access) (2020).
[33] Hao Yu, Sen Yang, and Shenghou Zhu. 2019. Parallel restarted SGD with faster convergence and less communication. Demystifying why model averaging works for deep learning. In Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA.
[34] Tianlong Yu, Tian Li, Yuqiong Sun, Susanda Nanda, Virginia Smith, Vyas Sekar, and Srinivasan Seshan. 2020. Learning-Centric-Aware Policies from Multiple Smart Homes via Federated Multi-Task Learning. In Proc. of the IEEE/ACM Fifth International Conference on Internet-of-Things Design and Implementation (IoTDI). Sydney, Australia.
[35] Longqing Zhang, Jiaolong Yang, Dongqiang Ye, and Gang Hua. 2018. Lq-nets: Learned quantization for highly accurate and compact deep neural networks. In Proceedings of the European conference on computer vision (ECCV). Munich, Germany.