Abstract
Federated Learning (FL) is very appealing for its privacy benefits: essentially, a global model is trained with updates computed on mobile devices while keeping the data of users local. Standard FL infrastructures are however designed to have no energy or performance impact on mobile devices, and are therefore not suitable for applications that require frequent (online) model updates, such as news recommenders.

This paper presents FLEET, the first Online FL system, acting as a middleware between the Android OS and the machine learning application. FLEET combines the privacy of Standard FL with the precision of online learning thanks to two core components: (i) I-Prof, a new lightweight profiler that predicts and controls the impact of learning tasks on mobile devices, and (ii) AnaSGD, a new adaptive learning algorithm that is resilient to delayed updates.

Our extensive evaluation shows that Online FL, as implemented by FLEET, can deliver a 2.3× quality boost compared to Standard FL, while only consuming 0.036% of the battery per day. I-Prof can accurately control the impact of learning tasks by improving the prediction accuracy up to 3.6× (computation time) and up to 19× (energy). AnaSGD outperforms alternative FL approaches by 18.4% in terms of convergence speed on heterogeneous data.

1 Introduction
The number of edge devices and the data produced by these devices have grown tremendously over the last 10 years. While in 2009, mobile phones only generated 0.7% of the worldwide data traffic, in 2018 this number exceeded 50% [74]. This exponential growth is raising challenges both in terms of scalability and privacy. As the volume of data produced by mobile devices explodes, users expose increasingly detailed and sensitive information, which in turn becomes more costly to store, process, and protect. This dual challenge of privacy and scalability is pervasive in machine learning (ML) applications such as recommenders, image-recognition apps, and personal assistants. These ML-based application often operate on highly personal and possibly sensitive content, including conversations, geolocation, or physical traits (faces, fingerprints), and typically require tremendous volumes of data for training their underlying ML models. For example, people in the USA of age 18-24, type on average around 900 words per day (128 messages per day [69] with an average of 7 words per message [17]). The Android next-word prediction service is trained on average with sequences of 4.1 words [36] which means that each user generates around 220 training samples daily. With tens of millions or even billions of user devices [8] scalability issues arise.

Federated Learning. To address this dual privacy and scalability challenge, large industrial players are now seeking to exploit the rising power of mobile devices to reduce the demand on their server infrastructures while, at the same time, protecting the privacy of their users. Federated Learning (FL) is a new computing paradigm (spearheaded among others by Google [11,41,72]) where a central server iteratively trains a global model (used by an ML-based application) without the need to centralize the data. The iterative training orchestrated by the server consists of the following synchronous steps for each update. First, the server selects the contributing mobile devices and sends them the latest version of the model. Each device then performs a learning task based on its local data and sends the result back to the server. The server aggregates a predefined number of results (typically a few hundreds [8]) and finally updates the model. The server drops any results received after the update. FL is "privacy-ready" and can provide formal privacy guarantees by using standard techniques such as secure aggregation and differential privacy [9].

The standard use of FL has so far been limited to a few lightweight and extremely privacy-sensitive services, such as next-word prediction [82], but its popularity is bound to grow. Privacy-related scandals continue to unfold [53,54], and new data protection regulations come into force [31,76]. The popularity of FL is clearly visible in two of the most popular ML frameworks (namely TensorFlow and PyTorch) [29,70], and also in the rise of startups such as S20.ai [71] or SNIPS (now part of Sonos) [73], which are betting on private decentralized learning.

Limitation of Standard FL. These are encouraging signs, but we argue in this paper that Standard FL [8] is unfortunately not effective for a large segment of ML-based...
applications, mainly due to its constraint for high device availability: the selected mobile devices need to be idle, charging and connected to an unmetered network. This constraint removes any impact perceived by users, but also limits the availability of devices for learning tasks. Google observed lower prediction accuracy during the day as few devices fulfill this policy and these generally represent a skewed population [82]. With most devices available at night the model is generally updated every 24 hours.

This constraint may be acceptable for some ML-based services but is problematic to what we call online learning systems, which underlie many popular applications such as news recommenders or interactive social networks (e.g., Facebook, Twitter, Linkedin). These systems involve large amounts of data with high temporality, that generally become obsolete in a matter of hours or even minutes [55]. To illustrate the limitation of Standard FL, consider two users, Alice and Bob, who belong to a population that trains the ML model underlying a news recommendation system (Figure 1). Bob wakes up earlier than Alice and clicks on some news articles. To deliver fresh and relevant recommendations, these clicks should be used to compute recommendations for Alice when she uses the app, slightly after Bob. In Standard FL (upper half Figure 1), the device of Bob would wait until much later (when idle, charging and connected to WiFi) to perform the learning task thus negating the value of the task results for Alice. In an online learning setup (lower half of Figure 1), the activity of Bob is rapidly incorporated into the model, thereby improving the experience of Alice.

**Challenges and contributions.** In this paper we address the aforementioned limitation and enable Online FL. We introduce FLeet, the first FL system that specifically targets online learning, acting as a middleware between the operating system of the mobile device and the ML-based application. FLeet addresses two major problems that arise after forfeiting the high device availability constraint.

First, learning tasks may have an energy impact on mobile devices now powered on a battery. Given that learning tasks are generally compute intensive, they can quickly discharge the device battery and thereby degrade user experience. To this end, FLeet includes I-Prof (§2.2), our new profiling tool which predicts and controls the computation time and the energy consumption of each learning task on mobile devices. The goal of I-Prof is not trivial given the high heterogeneity of the devices and the performance variability even for the same device over time [61] (as we show in §3).

Second, as mentioned above, synchronous training discards all late results arriving after the model is updated thus wasting the battery of the corresponding devices and their potentially useful data. Frequent model updates call for small synchronization windows that given the high performance variability, amplify this waste. We therefore replace the synchronous scheme of Standard FL with asynchronous updates. However, asynchronous updates introduce the challenge of staleness as multiple users are now free to perform learning tasks at arbitrary times. A stale result occurs when the learning task was computed on an outdated model version; meanwhile the global model has progressed to a new version. Stale results add noise to the training procedure, slow down or even prevent its convergence [39,85]. Therefore, FLeet includes AdaSGD (§2.3), our new Stochastic Gradient Descent (SGD) algorithm that tolerates staleness by dampening the impact of outdated results. This dampening depends on (a) the past observed staleness values and (b) the similarity with past learning tasks.

We fully implemented the server side and the Android client of FLeet. We evaluate the potential of FLeet and show that it can increase the accuracy of a recommendation system (that employs Standard FL) by 2.3x on average, by performing the same number of updates but in a more timely (online) manner. Even though the learning tasks drain energy directly from the battery of the phone, they consume on average only 0.036% of the battery capacity of a modern smartphone per user per day. We also evaluate the components of FLeet on 40 commercial Android devices, by using popular benchmarks for image classification. Regarding I-Prof, we show that 90% of the learning tasks deviate from a fixed Service Level Objective (SLO) of 3 seconds by at most 0.75 seconds in comparison to 2.7 seconds for the competitor (the profiler of MAUI [18]). The energy deviation from an SLO of 0.075% battery drop is 0.01% for I-Prof and 0.19% for the competitor. We also show that our staleness-aware learning algorithm (AdaSGD) learns 18.4% faster than its competitor (DynSGD [39]) on heterogeneous data.

2 FLeet

FLeet incorporates two components we consider necessary in any system that has the ambition to provide both, the (a) privacy of FL and (b) the precision of online learning systems. The first component is I-Prof, a lightweight ML-based profiling mechanism that controls the computation
time and energy of the learning task by using ML-based estimators. The second component of FLEET is AdaSGD, a new adaptive learning algorithm that tolerates stale updates by automatically adjusting their weight.

2.1 Architectural Overview

Similar to the implementation of Standard FL [8], FLEET follows a client-server architecture (Figure 2) where each user hosts a worker and the service provider hosts the server (typically in the cloud). In FLEET, the worker is a library that can be used by any mobile ML-based application (e.g., a news articles application). The model training protocol of FLEET is the following (the numbers refer to Figure 2):

(1) The worker requests a learning task and sends information regarding the labels of the local data along with information about the state of the mobile device. We introduce the purpose of this information in Steps 2 and 3.

(2) I-Prof employs the device information to bound the workload size (i.e., set a mini-batch size bound) that will be allocated to this worker such that the computation time and energy consumption approximate an SLO set by the service provider or negotiated with the user (details in §2.2).

(3) AdaSGD computes a similarity for the requested learning task with past learning tasks in order to adapt to updates with new data (details in §2.3).

(4) In order to prevent the computation of learning tasks with low or no utility for the learning procedure, the controller checks if both the mini-batch size and the similarity value pass certain thresholds set by the service provider. If the check fails, the request of the worker is rejected, otherwise the controller sends the model parameters and the mini-batch size to the worker and the learning task execution begins (details about setting these thresholds in §2.4).

(5) Based on the mini-batch size returned by the server, the worker samples from its locally collected data, performs the learning task, i.e., computes the model gradient and sends it back to the server. On the server side, AdaSGD updates the model after dynamically adapting this gradient based on its staleness and on its similarity value (details in §2.3).

The above protocol maintains the key “privacy-readiness” of Standard FL: the user data never leave the device during the learning procedure.

2.2 Workload Bound via Profiling

In Online FL, a mobile device should be able to compute model updates at any time, not only during the night, when the mobile device is idle, charging and connected to WiFi. Therefore, FLEET drops the constraint of Standard FL for high device availability. Hence, the learning task now drains energy directly from the battery of the device. Controlling the impact of a learning task on the user application in terms of energy consumption and computation time becomes crucial. To this end, FLEET incorporates a profiling mechanism that determines the workload size (i.e., the mini-batch size) appropriate for each mobile device.

Best-effort solution. To highlight the need for a specific profiling tool, we first consider a naive solution in which users process data points until they reach the SLO either in terms of computation time or energy. At this point, a worker sends back the resulting “best-effort” gradient. The service provider cannot decide beforehand whether for a given device, the cost (in terms of energy, time and bandwidth) to download the model, compute and upload the gradient is worth the benefit to the model. Updates computed on very small mini-batch sizes (by weak devices) will perturb the convergence of the overall model, and might even negate the benefit of other workers.

To illustrate this point, consider the experiment of Figure 3. The figure charts the result of training a Convolutional Neural Network on CIFAR10 [43] under different combinations of “strong” and “weak” workers. The strong workers compute on a mini-batch size of 128 while the weak workers compute on a mini-batch size of 1. We observe that even 2 weak workers are enough to cancel the benefit of distributed learning, i.e., the performance with 10 strong + 2 weak workers is the same as training with a single strong worker.

One way to avoid this issue could be to drop all the gradients computed on a mini-batch size lower than a given bound or weigh them with a tiny factor according to the size of their underlying mini-batch. This way would however waste the energy required to obtain these gradients. A profiler tool that can estimate the maximum mini-batch size (workload bound) that a worker can compute is necessary for the controller to decide whether to reject the computation request of

![Figure 2](image-url)  
**Figure 2.** The architecture of FLEET.

![Figure 3](image-url)  
**Figure 3.** Motivation for lower bounding the mini-batch size. The noise introduced by weak workers (i.e., with small mini-batch sizes) may be detrimental to learning.
this worker, before the gradient computation. Unfortunately, existing profiling approaches \[10,13,14,18,35,44,83\] are not suitable because they are either relatively inaccurate (see §3.3) or they require privileged access (e.g., rooted Android devices) to low-level system performance counters.

**I-Prof.** Mobile devices have a significantly lower level of parallelism in comparison with cloud servers. For example, the graphical accelerators of mobile devices generally have 10-20 cores \[6,68\] while the GPUs on a server have thousands of cores \[62\]. Given this low level of parallelism, even a relatively small mini-batch size can fill the processing pipelines. Hence, any additional workload will linearly increase the computation time and the energy consumption. Based on this observation, we built I-Prof, a lightweight profiler specifically designed for Online FL systems. We design I-Prof with three goals in mind: (a) operate effectively with data from a wide range of device types, (b) do so in a lightweight manner, i.e., introduce only a negligible latency to the learning task and (c) rely only on the data available on a stock (non-rooted) Android device.

I-Prof employs an ML-based scheme to capture how the device features affect the computation time and energy consumption of the learning task. I-Prof predicts the largest mini-batch size a device can process while respecting both the time and the energy limits set by the SLO. To this aim, I-Prof uses two predictors, one for computation time and one for energy. Each predictor updates its state with data from the device information sent by the workers.

Designing such predictors is however tricky, as modern mobile phones exhibit a wide range of capabilities. For example, in a matrix multiplication benchmark, Galaxy S6 performs 7.11 Gflops whereas Galaxy S10 performs 51.4 Gflops \[46\]. Figure 4 illustrates this heterogeneity on three different mobile devices by executing successive learning tasks with decreasing mini-batch size \((n)\). For example, in a matrix multiplication benchmark, Galaxy S6 performs 7.11 Gflops whereas Galaxy S10 performs 51.4 Gflops \[46\]. Figure 4 illustrates this heterogeneity on three different mobile devices by executing successive learning tasks with decreasing mini-batch size \((n)\). For example, in a matrix multiplication benchmark, Galaxy S6 performs 7.11 Gflops whereas Galaxy S10 performs 51.4 Gflops \[46\].

In the following, we describe how I-Prof predicts the mini-batch size \((n)\) given a computation time SLO \((T_{SLO})\). The computation time linearly increases with the workload size, i.e., \(t_{comp} = \alpha \cdot n\), where \(\alpha\) depends on the device and its state. Considering the goal (i.e., \(t_{comp} \rightarrow T_{SLO}\)), the optimal mini-batch size is predicted as:

\[
\hat{n} = \max \left(1, \frac{T_{SLO}}{\hat{\alpha}} \right)
\]  

(1)

I-Prof estimates the slope \(\hat{\alpha}\) from the device characteristics and operational conditions using a method that combines linear regression and online passive-aggressive learning \[16\].

The input to this method is a set of device features based on measurements available through the Android API, namely available memory, total memory, temperature and sum of the maximum frequency over all the CPU cores. However, these features only encode the computing power of a device. For the prediction based on the energy SLO, I-Prof also needs a feature that encodes the energy efficiency of each device. We choose this additional feature as the energy consumption per non-idle CPU time \[2\]. We show in our evaluation (§3.3) that these features achieve our three design goals. Given a vector of device features \((x)\), and a vector of model parameters \((\theta_{prof})\), the slope \(\hat{\alpha}\) is estimated as \(\hat{\alpha} = x^T \theta_{prof}\).

I-Prof uses a cold-start linear regression model for the first request of each user device. We pre-train the cold-start model using ordinary least squares with an offline dataset. This dataset consists of data collected by executing requests from a set of training devices with a mini-batch size increasing from 1 till a value such that the computation time reaches twice the SLO. I-Prof periodically re-trains the cold-start model after appending new data (device features).

Furthermore, I-Prof creates a personalized model for every new device model (e.g., Galaxy S7) and employs it for every following request coming from this particular model. I-Prof bootstraps the new model with the first request (for which the cold-start model is used to estimate the computation time). For all the following learning tasks that result in pairs of \((x^{(k)}, \alpha^{(k)})\), I-Prof incrementally updates a Passive-Aggressive (PA) model \[16\] as:

\[
\theta_{prof}^{(k+1)} = \theta_{prof}^{(k)} + \frac{f^{(k)} - u^{(k)}}{||x^{(k)}||^2} u^{(k)}
\]

where \(u^{(k)} = \text{sign}(\alpha^{(k)} - x^{(k)}^T \theta_{prof}^{(k)})x^{(k)}\) denotes the update

1The prediction method given an energy SLO is the same.

2CPU time spent by processes executing in user or kernel mode.
direction, and \( f \) the loss function:

\[
f(\theta_{\text{prof}}, x, \alpha) = \begin{cases} 
0 & \text{if } |x^T \theta_{\text{prof}} - \alpha| \leq \epsilon \\
|x^T \theta_{\text{prof}} - \alpha| - \epsilon & \text{otherwise.}
\end{cases}
\]  

(2)

The parameter \( \epsilon \) controls the sensitivity to prediction error and thereby the aggressiveness of the regression algorithm, i.e., the smaller the value of \( \epsilon \) the larger the update for each new data instance (more aggressive).

I-Prior focuses solely on the time and energy spent during an SGD computation. Despite network costs (in particular when transferring models) having also an important impact, they fall outside the scope of this work as one can rely on prior work [4,51,66] to estimate the time and energy of network transfers within FLEET.

### 2.3 Adaptive Stochastic Gradient Descent

The server-driven synchronous training of Standard FL is not suitable for Online FL, as the latter requires frequent updates and needs to exploit contributions from all workers, including slow ones (§1). Therefore, we introduce AdaSGD, an asynchronous learning algorithm that is robust to stale updates. AdaSGD is responsible for aggregating the gradients sent by the workers and updating the application model (\( \theta_{\text{app}} \))^3. Each update takes place after AdaSGD receives \( K \) gradients. The aggregation parameter \( K \) can be either fixed or based on a time window (e.g., update the model every 1 hour). The model update is:

\[
\theta_{\text{app}}^{(t+1)} = \theta_{\text{app}}^{(t)} - \gamma_t \sum_{i=1}^{K} \min \left( 1, A(t_i) \cdot \frac{1}{\text{sim}(x_i)} \right) \cdot G(\theta_{\text{app}}^{(t)}, \xi_t)
\]  

(3)

where \( \gamma_t \) is the learning rate, \( t \in \mathbb{N} \) denotes the global logical clock (or step) of the model at the server (i.e., the number of past model updates) and \( t_i \leq t \) denotes the logical clock of the model that the worker receives from the server. \( G(\theta_{\text{app}}^{(t)}, \xi_t) \) is the gradient computed by the client w.r.t the model parameters \( \theta_{\text{app}}^{(t)} \) on the mini-batch \( \xi_t \) drawn uniformly from the local dataset \( x_t \).

The workers send gradients asynchronously that can result in stale updates. The staleness of the gradient (\( t_i := t - t_i \)) shows the number of model updates between the model pull and gradient push of worker \( i \). One option is to directly apply this gradient, at the risk of slowing down or even completely preventing convergence [39,85]. The Standard FL algorithm (FedAvg [52]) simply drops stale gradients. However, even if computed on a stale model, the gradient may incorporate potentially valuable information. Moreover, in FLEET, the gradient computation may drain energy directly from the battery of the phone, thus making the result even more valuable. Therefore, AdaSGD utilizes even stale gradients without jeopardizing the learning process, by multiplying each gradient with an additional weight to the learning rate. This weight consists of (a) a dampening factor based on the staleness (\( \Lambda(t_i) \)) and (b) a boosting factor based on the user’s data novelty (\( \frac{1}{\text{sim}(x_i)} \)), that we describe in the following.

**Staleness-based dampening.** AdaSGD builds on prior work on staleness-aware learning that has shown promising results [39,85]. In order to accelerate learning, AdaSGD relies on a system parameter: the expected percentage of non-stragglers (denoted by \( s \)). We highlight that this value is not a hyperparameter that needs tuning for each ML application but a system parameter that solely depends on the computing and networking characteristics of the workers, while it can be adapted dynamically [64,65]. We define the staleness-aware damping factor \( \Lambda(t) = e^{-\beta t} \), with \( \beta \) chosen s.t. \( \frac{1}{\text{thres}} = e^{-\beta \text{thres}} \) (i.e., the inverse dampening function [39]) intersects with our exponential dampening function in \( \frac{\text{thres}}{2} \), where \( \text{thres} \) is the \( s \)-th percentile of past staleness values. Figure 5 shows the dampening factor of AdaSGD compared to the inverse dampening function (employed by DynSGD [39]). Our hypothesis is that the perturbation to the learning process introduced by stale gradients, increases exponentially and not linearly with the staleness. We empirically verify the superior performance of our exponential dampening function compared to the inverse in §3.2.

As a quantile, \( \text{thres} \) is estimated from the staleness distribution. In practice, for the past staleness values to be representative of the actual distribution, an initial bootstrapping phase can employ the dampening factor of DynSGD. After this phase, the service provider can set \( s \) and deploy AdaSGD. An underestimate of \( s \) will slow down convergence, whereas an overestimate may lead to divergence. As we empirically observe (§3.1), the staleness distribution often has a long tail. In such cases, the best choice of \( s \) is the one that sets \( \text{thres} \) at the beginning of the tail.

**Similarity-based boosting.** In the presence of stragglers with large delays (comparing to the mean latency), staleness can grow and drive \( \Lambda(t) \) close to 0, i.e., almost neglect the

---

3Not to be confused with the model of the profiler (\( \theta_{\text{prof}} \)).
We highlight that worker, (iii) an encoding scheme for transferring C++ objects works in C++ on top of (DL4J [22]) to deployed on an HTTP server) and the worker as an Android I-Prof interface (JNI) for the server, (ii) the Android NDK for the back-end running app. Therefore, allowing the worker to be killed by the operating system to free resources for background applications have low priority so their access to resources is heavily restricted and they are likely to be killed by the operating system to free resources for the foreground running app. Therefore, allowing the worker to run in the background would make its performance very unpredictable and thus impact the predictions of I-ProF.

We build our main library for Convolutional Neural Networks in C++ on top of FLeet. We employ (i) the Java Native Interface (JNI) for the server, (ii) the Android NDK for the worker, (iii) an encoding scheme for transferring C++ objects through the java streams, and (iv) a thread-based parallelization scheme for the independent gradient computations of the worker. On recent mobile devices that support NEON [5], FLeet accelerates the gradient computations by using SIMD instructions. We also port a popular deep learning library (DL4J [22]) to FLeet, to benefit from its rich ecosystem of ML algorithms. However, as DL4J is implemented in Java, we do not have full control over the resource allocation.

FLeet relies on the developer of the overlying ML application to ensure the performance isolation between the running application and the worker runtime. The worker can execute in a window of low user activity (e.g., while the user is reading an article) to minimize the impact of the overlying ML application on the predictive power of I-ProF.

**Resource allocation.** Allocating system resources is a very challenging task given the latency and energy constraints of mobile devices [24,56]. Our choice of employing only stock Android without root access means we can only control which cores execute the workload on the worker, with no access, for instance, to low-level advanced tuning. Given this limited control and the inherent mobile device heterogeneity, we opt for a simple yet effective scheme for allocating resources.

This scheme schedules the execution only on the “big” cores for ARM big.LITTLE architectures and on all the cores otherwise. In the case of computationally intensive tasks (such as the learning tasks of FLeet), big cores are more energy efficient than LITTLE cores because they finish the computation much faster [32]. Regarding ARMv7 symmetric architectures with 2 and 4 cores that equip older mobile devices, the energy consumption per workload is constant regardless of the number of cores: a higher level of parallelism will consume more energy but the workload will execute faster. For this reason, our allocation policy relies on all the available cores so that we can take advantage of the embarrassingly parallel nature of the gradient computation tasks. For such tasks, we empirically show (§3.4) that this scheme outperforms more complex alternatives [56].

**Controller thresholds.** In practice, the service provider can adopt various approaches to define the size and similarity thresholds of the controller (Figure 2). One option is A/B testing along with the gradual increase of the thresholds. In particular, the system initializes the thresholds to zero and divides the users into two groups. The first group tests the impact of the mini-batch size and the second the impact of the label similarity. Both groups gradually increase the thresholds until the impact on the service quality is considered acceptable. The server can execute this A/B testing procedure periodically, i.e., reset the thresholds after a time interval. We empirically evaluate the impact of these thresholds on prediction quality in §3.5.

3 Evaluation

Our evaluation consists of two main parts. First, in §3.1, we evaluate the claim that Online FL holds the potential to deliver better ML performance than Standard FL [8] for applications that employ data with high temporality (§1). Second, we evaluate in more detail the internal mechanisms
of FLeet, namely AdaSGD (§3.2), I-ProF (§3.3), the resource allocation scheme (§3.4) and the controller (§3.5).

We deploy the server of FLeet on a machine with an Intel Xeon X3440 with four CPU cores, 16 GiB RAM and 1 Gb Ethernet, on Grid5000 [33]. The workers are deployed on a total of 40 different mobile phones that we either personally own or belong to the AWS Device Farm [7] (Oregon, USA). In §3.1, we deploy the worker on a Raspberry Pi 4 as our hashtag recommender is implemented on TensorFlow that does not yet support training on Android devices.

3.1 Online VS Standard Federated Learning

We compare Online with Standard FL on a Twitter hashtag recommender. Tweepy [75] enables us to collect around 2.6 million tweets spanning across 13 successive days and located in the west coast of the USA. We preprocess these tweets (e.g., remove automatically generated tweets, remove special symbols) based on [23]. We then divide the data into shards, each spanning 2 days, and divide each shard into chunks of 1 hour. We finally group the data into mini-batches based on the user id.

Our training and evaluation procedure follows an Online FL setup. Our model is a basic Recurrent Neural Network implemented on TensorFlow with 123,330 parameters [30], that predicts the 5 hashtags with the largest values on the output layer. The model training consists of successive gradient-descent operations, with each gradient derived from a single mini-batch (i.e., sent by a single user). For the Online FL setup, the model is updated every 1 hour. Training uses the data of the previous hour and testing uses the data of the next hour. For the Standard FL setup, the model is updated every day. Training uses the data of the previous day and testing uses the data of the next day. We highlight that under this setup, the two approaches employ the same number of gradient computations and the difference lies only in the time they perform the model updates. We also compare against a baseline model that always predicts the most popular hashtags [42,63]. We evaluate the model on the data of each chunk and reset the model at the end of each shard.

Figure 6. Online FL boosts Twitter hashtag recommendations by an average of 2.3× comparing to Standard FL.

Quality boost. For assessing the quality of the hashtag recommender, we employ the F1-score @ top-5 [28,42] to capture how many recommendations were used as hashtags (precision) and how many of the used hashtags were recommended (recall). In particular, for each tweet in the evaluation set, we compare the output of the recommender (top-5 hashtags) with the actual hashtags of the tweet, and derive the F1-score. Figure 6 shows that Online FL outperforms Standard FL in terms of F1-score, with an average boost of 2.3×. Online FL updates the model in a more timely manner, i.e., soon after the data generation time, and can thus better predict (higher F1-score) the new hashtags than Standard FL. The performance of the baseline model is quite low as the nature of the data is highly temporal [45].

Energy impact. We measure the energy impact of the gradient computation on the Raspberry Pi worker. The Raspberry Pi has no screen; nevertheless recent trends in mobile/embedded processor design show that the processor is dominating the energy consumption, especially for compute intensive workloads such as the gradient computation [34]. We measure the power consumption of every update of Online FL by executing the corresponding gradient computation 10 times and by taking the median energy consumption. We observe that the power depends on the batch size and increases from 1.9 Watts (idle) to 2.1 Watts (batch size of 1) and to 2.3 Watts (batch size of 100). The computation latency is 5.6 seconds for batch size of 1 and 8.4 for batch size of 100. Across all the updates of Online FL (that employ various batch sizes and result in the quality boost shown in Figure 6), we measure the average, median, 99th percentile and maximum values of the daily energy consumption as 4, 3.3, 13.4 and 44 mWh respectively. Given that most modern smartphones have battery capacities over 11000 mWh, we argue that Online FL imposes a minor energy consumption overhead for boosting the prediction quality.

Staleness distribution. We study the staleness distribution of the updates on our collected tweets, in order to set our experimental setup for evaluating AdaSGD (§3.2). We assume that the round-trip latency per model update (gradient computation time plus network latency) follows an exponential distribution (as commonly done in the distributed
learning literature \cite{3,25,48,58}). The network latency for downloading the model (123,330 parameters) and uploading the gradients is estimated to 1.1 second for 4G LTE and 3.8 seconds for 3G HSPA+ \cite{37}. We then estimate the average computation latency to be 6 seconds, based on our latency measurements on the Raspberry Pi. Therefore, we choose the exponential distribution with a minimum of $6 + 1.1 = 7.1$ seconds and a mean of $\frac{(6+1.1)+6+3.8}{2} = 8.45$ seconds. Given the exponential distribution for the round-trip latency and the timestamps of the tweets, we observe (in Figure \ref{fig:7}) that the staleness follows a Gaussian distribution with a long tail (as assumed in \cite{85}). The long tail is due the presence of certain peak times with hundreds of tweets per second.

### 3.2 AdaSGD Performance

We now dissect the performance of AdaSGD via an image classification application that involves Convolutional Neural Networks (CNNs). We choose this benchmark due to its popularity for the evaluation of SGD-based approaches \cite{2,12,40,52,85,86}. We employ multiple scenarios involving various staleness distributions, data distributions, and a noise-based differentially private mechanism.

**Image classification setup.** We implement the models shown in Table \ref{table:1} in Fleet\textsuperscript{4} to classify handwritten characters and colored images. We use three publicly available datasets: MNIST \cite{47}, E-MNIST \cite{15} and CIFAR-100 \cite{43}. MNIST consists of 70,000 examples of handwritten digits (10 classes) while E-MNIST consists of 814,255 examples of handwritten characters and digits (62 classes). CIFAR-100 consists of 60,000 colour images in 100 classes, with 600 images per class. We perform min-max scaling as a pre-processing step for the input features.

We split each dataset into training / test sets: 60,000 / 10,000 for MNIST, 69,932 / 116,323 for E-MNIST and 50,000 / 10,000 for CIFAR-100. Unless stated otherwise, we set the aggregation parameter $K$ (§2.3) to 1 (for maximum update frequency), the mini-batch size to 100 examples \cite{60}, the $\epsilon$ (the Passive-Aggressive parameter) to 0.1 and the learning rate to $15 \times 10^{-4}$ for CIFAR-100, $8 \times 10^{-4}$ for E-MNIST, and $5 \times 10^{-4}$ for MNIST.

Since the training data present on mobile devices are typically collected by the users based on their local environment and usage, both the size and the distribution of the training data will typically heavily vary among users. Given the terminology of statistics, this means that the data are not Independent and Identically Distributed (\textit{non-IID}). Following recent work on FL \cite{77,78,84,86}, we employ a non-IID version of MNIST. Based on the standard data decentralization scheme \cite{52}, we sort the data by the label, divide them into shards of size equal to $\frac{\text{number of users}}{69000}$, and assign 2 shards to each user. Therefore, each user will contain examples for only a few labels.

\textsuperscript{4}We implement the CNN for E-MNIST on DL4J and the rest on our default CNN library.

| Dataset   | Parameters | Input              | Conv1 | Pool1 | Conv2 | Pool2 | FC1  | FC2  | FC3  |
|-----------|------------|--------------------|-------|-------|-------|-------|------|------|------|
| MNIST     | Kernel size Strides | 28x28x1 5x5x8 1x1 9x1 | 3x3 | 3x3 | 5x5x48 2x2 9x1 | 2x2 | 3x3 | 2x2 | 10   | --   |
| E-MNIST   | Kernel size Strides | 28x28x1 5x5x10 1x1 2x2 5x5x10 2x2 9x1 | 2x2 | 2x2 | 5x5x10 2x2 9x1 | 2x2 | 5   | 42   | --   |
| CIFAR-100 | Kernel size Strides | 32x32x3 5x5x16 1x1 2x2 5x5x64 1x1 4x4 | 2x2 | 2x2 | 5x5x32 4x4 16x1 | 8   | 384 | 192  | 100  |

#### Staleness awareness setup.

To be able to precisely compare AdaSGD with its competitors, we control the staleness of the updates produced by the workers of Fleet. Based on \cite{85} and the shape of the staleness distribution shown in Figure \ref{fig:7}, we employ Gaussian distributions for the staleness with two setups: $D_1 := \mathcal{N}(\mu = 6, \sigma = 2)$ and $D_2 := \mathcal{N}(\mu = 12, \sigma = 4)$, to measure the impact of increasing the staleness. We set the expected percentage of non-stragglers ($\%$) to 99.7%, i.e., $\tau_{\text{thres}} = \mu + 3\sigma$. We evaluate the SGD algorithms on Fleet by using commercial Android devices from AWS.

We evaluate the performance of AdaSGD against three learning algorithms: (i) DynSGD \cite{39}, a staleness-aware SGD algorithm employing an inverse dampening function ($A(\tau) = \frac{1}{1+\tau^2}$), that AdaSGD builds upon (§2.3), (ii) the standard SGD algorithm with synchronous updates (SSGD) that represents the ideal (staleness-free) convergence behaviour, and (iii) FedAvg \cite{52}, the standard staleness-unaware SGD algorithm that is based on gradient averaging.

**Staleness-based dampening.** Figure \ref{fig:8} depicts that AdaSGD outperforms the alternative learning schemes for the non-IID version of MNIST. As expected, the staleness-free scenario (SSGD) delivers the fastest (ideal) convergence, whereas the staleness-unaware FedAvg diverges. The comparison between the two staleness-aware algorithms (DynSGD and AdaSGD) shows that our solution (AdaSGD) better adapts the dampening factor to the noise introduced by stale gradients (§2.3). AdaSGD achieves 80% accuracy 14.4% faster than DynSGD for $D_1$ and 18.4% for $D_2$. Figure \ref{fig:8} also depicts the impact of staleness on DynSGD and AdaSGD. We observe that the larger the staleness, the slower the convergence of both algorithms. The advantage of AdaSGD over DynSGD grows with the amount of staleness as the larger amount of noise gives more leeway to AdaSGD to benefit from its superior dampening scheme.

**Similarity-based boosting.** We evaluate the effectiveness of the similarity-based boosting property of AdaSGD (§2.3) in the case of long tail staleness (Figure \ref{fig:7}). We employ the non-IID MNIST dataset, $D_1$ (thus $\tau_{\text{thres}} = 12$) and set the staleness to $4 \cdot \tau_{\text{thres}} = 48$ for all the gradients computed on data with class 0. This setup essentially captures the case where a particular label is only present in stragglers. Figure \ref{fig:9}(a) shows that AdaSGD incorporates the knowledge from class 0 much faster than DynSGD.
Figure 8. Impact of staleness on learning.

Figure 9(b) shows the CDF for the damping values used to weight the gradients of Figure 9(a). We mark the two points of interest regarding the $\tau_{\text{thres}}$ by vertical lines (as also shown in Figure 5). If AdaSGD had no similarity-based boosting, all updates related to class 0 would almost not be taken into account, as they would be nullified by the exponential damping function, therefore leading to a model with poor predictions for this class. Given the low class similarity of the learning tasks involving class 0, AdaSGD boosts their damping value. The second vertical line denotes the staleness value ($\tau_{\text{thres}} = 6$) for which AdaSGD and DynSGD give the same damping value (0.14). The slope of each curve at this point indicates that the damping values for DynSGD are more concentrated whereas the ones for AdaSGD are more spread around this value.

Figure 9. Impact of long tail staleness on learning.

**IID data.** Although data are more likely to be non-IID in an FL environment, the data collected on mobile devices might in some cases be IID. We thus benchmark AdaSGD under two additional datasets (E-MNIST and CIFAR-100) with the staleness following D2. Figure 10 shows that our observations from Figure 8 hold also with IID data. As with non-IID data, FedAvg diverges also in the IID setting, and AdaSGD performs better than DynSGD on both datasets.

**Differential privacy.** Differential privacy [26] is a popular technique for privacy-preserving FL with formal guarantees [9]. We thus compare AdaSGD against DynSGD in a differentially private setup by perturbing the gradients as in [2]. We keep the previous setup (IID data with D2) and employ the MNIST dataset. Based on [80], we fix the probability $\delta = 1/N^2 = \frac{1}{60000}$ and measure the privacy loss ($\epsilon$) with the moments accountant approach [2] given the sampling ratio ($\text{mini-batch size} = \frac{100}{60000}$), the noise amplitude, and the total number of iterations.

Figure 11 demonstrates that the advantage of AdaSGD over DynSGD also holds in the differentially private setup. A better privacy guarantee (i.e., smaller $\epsilon$) slows down the convergence for both staleness-aware learning schemes.

3.3 I-Prof Performance

We compare I-Prof against the profiler of MAUI [18], a mobile device profiler aiming to identify the most energy-consuming parts of the code and offload them to the cloud. MAUI predicts the energy by using a linear regression model (similar to the global model of I-Prof) on the number of CPU cycles ($\hat{E} = \theta_0 \cdot n$), to essentially capture how the size of the workload affects the energy (as in [59]). We adapt the profiler of MAUI to our setup by replacing the CPU cycles with the mini-batch size for two main reasons. First, our workload has a static code path so the number of CPU cycles on a particular mobile device is directly proportional to the mini-batch size. Second, measuring the number of executed CPU cycles requires root access that is not available on AWS.

We bootstrap the global model of I-Prof and the model of MAUI by pre-training on a training dataset. To this end, we use 15 mobile devices in AWS (that are different from the ones used for the rest of the experiments), assign them learning tasks with increasing mini-batch size until the computation time becomes 2 times the SLO, and collect their device information for each task. We rely on the same methodology to
evaluate energy consumption but use only 3 mobile devices in our lab, as AWS prohibits energy measurements.

For testing, we use a different set of 20 commercial mobile devices in AWS, each performing requests for the image classification application (on MNIST), starting at different timestamps (log-in events) as shown in Figure 12(a). In order to ensure a precise comparison with MAUI, we add a round-robin dispatcher to the profiler component which alternates the requests from a given device between I-Prof and MAUI.

Computation time SLO. Figure 12(b) shows that I-Prof largely outperforms MAUI in terms of deviation from the computation time SLO. 90% of approximately 280 learning tasks deviate from an SLO of 3 seconds by at most 0.75 seconds with I-Prof and 2.7 seconds with MAUI. This is the direct outcome of our design decisions. First, I-Prof adds dynamic features (e.g., the temperature of the device) to train its global model (§2.2). As a result, the predictions are more accurate for the first request of each user. Second, I-Prof uses a personalized model for each device that reduces the error (deviation from the SLO) with every subsequent request (Figure 12(c)). Figure 12(d) shows that the personalized models of I-Prof are able to output a wider range of mini-batch sizes that better match the capabilities of individual devices.

On the contrary, MAUI relies on a simple linear regression model which has acceptable accuracy for its use-case but is inefficient when profiling heterogeneous mobile devices.

Energy SLO. To assess the ability of I-Prof to also target the energy SLO, we use the same setup as for the computation time, except on 5 mobile devices. We configure I-Prof with a significantly smaller error margin, \( \epsilon = 6 \times 10^{-5} \) (Equation 2), because the linear relation (capture by \( \alpha \) as defined in §2.2) is significantly smaller for the energy than for the computation time (as shown in Figure 4).

Figure 13 shows that I-Prof significantly outperforms MAUI in terms of deviation from the energy SLO. 90% of

36 learning tasks deviate from an SLO of 0.075% battery drop by at most 0.01% for I-Prof and 0.19% for MAUI. The observation that I-Prof is able to closely match the latency SLO, while MAUI suffers from huge deviations, holds for the energy too. The PA personalized models are able to quickly adapt to the state of the device as opposed to the linear model of MAUI that provides biased predictions.

3.4 Resource Allocation

We evaluate our resource allocation scheme (§2.4) and compare it against CALOREE [56] which is a state of the art resource manager for mobile devices. The goal of CALOREE is to optimize resource allocation in order for the workload execution to meet its predefined deadline while minimizing the energy consumption. To this end, CALOREE profiles the target device by running the workload with different resource configurations (i.e., number of cores, core frequency). Since FLEET executes on non-rooted mobile devices, we can only adapt the number of big/little cores (but not their frequencies). By varying the number of cores allocated to our workload (i.e., gradient computation), we obtain the energy consumption of each possible configuration. From these configurations, CALOREE only selects those with the optimal energy consumption (the lower convex hull) which are packed in the so called performance hash table (PHT).
Table 2. Performance of CALOREE [56] on new devices.

| Running device | Deadline error (%) |
|----------------|--------------------|
| Galaxy S7      | 1.4                |
| Galaxy S8      | 9                  |
| Honor 9        | 46                 |
| Honor 10       | 255                |

**CALOREE on new devices.** In their thorough evaluation, the authors of CALOREE used the same device for training and running the workloads. Therefore, we first benchmark the performance of CALOREE when running on new devices. We employ Galaxy S7 to collect the PHT and set the mini-batch size that I-Prof gives for a latency SLO of 3 seconds (§3.3). We then run this workload with CALOREE on different mobile devices, as shown in Table 2.

The performance of CALOREE degrades significantly when running on a different device than the one used for training. The first line of Table 2 shows the baseline error when running on the same device. The error increases more than 6× for a device with similar architecture and the same vendor (Galaxy S8) and more than 32× for a device of similar architecture but different vendor (Honor 9 and 10). This significant increase for the error is due to the heterogeneity of the mobile devices which make PHTs not applicable across different device models.

**CALOREE vs FLEET.** We evaluate the resource allocation scheme of FLEET by comparing it to the ideal environment for CALOREE, i.e., training and running on the same device (a setup nevertheless difficult to achieve in FL with millions of devices). Following the setup used for the energy SLO evaluation (§3.3), we employ 5 devices and fix the size of the workload (mini-batch size) based on the output of I-Prof. In particular we set the mini-batch size to 280, 4320, 6720, 5280, 1200 for the devices shown in Figure 14 respectively. We set the deadline of CALOREE either equal or double than the computation latency of FLEET. We take 10 measurements and report on the median, 10th and 90th percentile.

Figure 14 shows the fact that in the ideal environment for CALOREE and even with double the time budget (giving more flexibility to CALOREE), FLEET has comparable energy consumption. Since gradient computation is a compute intensive task with high temporal and spacial cache locality, the configuration changes performed by CALOREE negatively impact the execution time and cancel any energy saved by its advanced resource allocation scheme. Additionally, the fewer configuration knobs available on non-rooted Android devices limit the full potential of CALOREE.

### 3.5 Learning Task Assignment Control

The controller of FLEET employs a threshold to prune learning tasks and thus control the trade-off between the cost of the gradient computations and the model prediction quality. This threshold can be based either on the mini-batch size or on the similarity values (Figure 2). To evaluate this trade-off, we employ non-IID MNIST with the mini-batch size following a Gaussian distribution $N(\mu = 100, \sigma = 33)$ (based on the distribution of the output of I-Prof shown in Figure 12(d)), and set the threshold to the $n$ – th percentile of the past values. Figure 15 illustrates that a threshold on the mini-batch size is more effective in pruning the less useful gradient computations than a threshold on the similarity. Figure 15(a) shows that even dropping up to 39.2% of the gradients (with the smallest mini-batch size) has a negligible impact on the accuracy (less than 2.2%). Figure 15(b) shows that one can drop 17% of the most similar gradients with an accuracy impact of 4.8%.

### 4 Related Work

**Distributed ML.** Adam [12] and TensorFlow [1] adopt the parameter server architecture [49] for scaling out on
high-end machines, and typically require cross-worker communication. FLEET also follows the parameter server architecture, by maintaining a global model on the server. However, FLEET avoids cross-worker communication, which is impractical for mobile workers due to the device churn.

A common approach for large-scale ML is to control the amount of staleness for boosting convergence [19,67]. In Online FL, staleness cannot be controlled as this would impact the model update frequency. The workers perform learning tasks asynchronously with end-to-end latencies that can differ significantly (due to device heterogeneity and network variability) or even become infinite (user disconnects).

Petuum [81] and TensorFlow handle faults (worker crashes) by checkpointing and repartitioning the model across the workers whenever failures are detected. In a setting with mobile devices, such failures may appear very often, thus increasing the overhead for checkpointing and repartitioning. FLEET does not require any fault-tolerance mechanism for its workers, as from a global learning perspective, they can be viewed as stateless.

**Federated learning.** In order to minimize the impact on mobile devices, Standard FL algorithms [8,38,52,72] require the learning task to be executed only when the devices are idle, plugged in, and on a free wireless connection. However, in §3.1, we have shown that these requirements may drastically impact the performance of some applications. Noteworthy, techniques for reducing the communication overhead [38] or increasing the robustness against adversarial users [20,21], are orthogonal to the online characteristic so they can be adapted for AdaSGD, and plugged into FLEET.

**Performance prediction for mobile devices.** Estimating the computation time or energy consumption of an application running on a mobile device is a very broad area of research. Existing approaches [10,13,14,35,44,83] target multiple applications generally executing on a single device. They typically benchmark the device or monitor hardware and OS-level counters that require root access. In contrast, FLEET targets a single application executing in the same way across a large range of devices. I-Prof poses a negligible overhead, as it employs features only from the standard Android API to enable Online FL, and requires no benchmarking of new devices. I-Prof is designed to make predictions for unknown devices.

Neurosurgeon [40] is a scheduler that minimizes the end-to-end computation time of inference tasks (whereas FLEET focuses on training tasks), by choosing the optimal partition for a neural network and offloading computations to the cloud. The profiler of Neurosurgeon only uses workload-specific features (e.g., number of filters or neurons) to estimate computation time and energy, and ignores device-specific features. By contrast, mobile phones, as targeted by I-Prof, exhibit a wide range of device-specific characteristics that significantly impact their latency and energy consumption (Figure 4).

Systems such as CALOREE [56] and LEO [57] profile mobile devices under different system configurations and train an ML model to determine the ones that minimize the energy consumption. They rely on a control loop to switch between these configurations such that the application does not miss the preset deadline. Due to the restrictions of the standard Android API, the available knobs are limited in our setup. For our application (i.e., gradient computation), we show that a simple resource allocation scheme (§2.4) is preferable even in comparison with an ideal execution model.

**5 Concluding Remarks**

This paper presented FLEET, the first system that enables online ML at the edge. FLEET employs I-Prof, a new ML-based profiler which determines the ML workload that each device can perform within predefined energy and computation time SLOs. FLEET also makes use of AdaSGD, a new staleness-aware learning algorithm that is optimized for Online FL. We showed the performance of I-Prof and AdaSGD on commercial Android devices with popular benchmarks. In our performance evaluation we do not focus on network and scalability aspects that are orthogonal to our work and addressed in existing literature. We also highlight that transferring the label and device information (Figure 2) poses a negligible network overhead compared to transferring the relatively large FL learning models.

Although we believe FLEET to represent a significant advance for online learning at the edge, there is still room for improvement. First, for the energy prediction, I-Prof requires access to the CPU usage that is considered as a security flaw on some Android builds and thus not exposed to all applications. In this case, I-Prof requires a set of additional permissions that belong to services from Android Runtime. Second, the transfer of the label distribution from the worker to the server introduces a potential privacy leakage. However, we highlight that the server has access only to the indices of the labels and not their values. In this paper, we focus on the protection of the input features and mention the possibility to deactivate the similarity-based boosting feature of AdaSGD in the case that this leakage is detrimental. We plan to investigate noise addition techniques for bounding this leakage [26] in our future work. Finally, theoretically proving the convergence of AdaSGD is non-trivial due to the unbounded staleness and the non Independent and Identically Distributed (non-IID) datasets among the workers. In this respect, a dissimilarity assumption similar to [50] may facilitate the derivation of the proof.

---

In their in-depth experimental evaluation the authors of [40] consider a single hardware platform and not Android mobile devices.
for smartphone applications. In USENIX ATC. 297–308.

[45] Su Mon Kywe, Tuan-Anh Hoang, Ee-Peng Lim, and Feida Zhu. 2012. On recommending hashtags in twitter networks. In International Conference on Social Informatics. Springer, 357–350.

[46] Primate Labs. 2020. Matrix multiplication benchmark. https://browser.geekbench.com.

[47] Yann LeCun. 1998. MNIST dataset. http://yann.lecun.com/exdb/mnist/.

[48] Kangwook Lee, Maximilian Lam, Ramtin Pedarsani, Dimitris Papailiou, and K. Ramchandran. 2017. Speeding up distributed machine learning using codes. IEEE Transactions on Information Theory 64, 3 (2017), 1514–1529.

[49] Mu Li, David G Andersen, Jun Woo Park, Alexander J Smola, Adam Ahmed, Vanja Josifovski, James Long, Eugene J Shekita, and Bor-Yiing Su. 2014. Scaling distributed machine learning with the parameter server. In OSDI. 583–598.

[50] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2018. Federated optimization for heterogeneous networks. arXiv preprint arXiv:1812.06127 (2018).

[51] Yan Liu and Jack YB Lee. 2015. An empirical study of throughput prediction in mobile data networks. In GLOBECOM. IEEE, 1–6.

[52] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2018. Federated optimization for heterogeneous networks. In OSDI. 583–598.

[53] Behnam Neyshabur, Ruslan R Salakhutdinov, and Nati Srebro. 2015. Path-sgd: Path-normalized optimization in deep neural networks from Decentralized Data. In AISTATS. 1273–1282.

[54] Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. 2019. Bayesian Nonparametric Federated Learning: Improving google keyboard query suggestions. In USENIX ATC. Vol. 12. 1–14.

[55] Wei Zhang, Suyog Gupta, Xiangru Lian, and Ji Liu. 2016. Staleness-aware asycn-sgd for distributed deep learning. In IJCAI. 321–334.

[56] Aurick Qiao, Abutalib Aghayev, Weiren Yu, Hoayong Chen, Qirong Ho, Garth A Gibson, and Eric P Xing. 2018. Litz: Elastic framework for high-performance distributed machine learning. In USENIX ATC. 631–644.

[57] Qualcomm. 2020. Adreno™ Graphics Processing Units. https://developer.qualcomm.com/software/adreno-gpu-sdk/gpu.

[58] European Union. 2016. Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (GDPR). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

[59] JLG Mishne, Jeff Dalton, Zhenghua Li, Aneesha Sharma, and Jimmy Lin. 2013. Fast data in the era of big data: Twitter’s real-time related query suggestion architecture. In SIGMOD. ACM, 1147–1158.

[60] Nikita Mishra, Connor Imes, John D Lafferty, and Henry Hoffmann. 2018. Path-sgd: Path-normalized optimization in deep neural networks from Decentralized Data. In AISTATS. 1273–1282.

[61] Joannis Miligiakas, Ce Zhang, Stefan Hadjis, and Christopher Ré. 2016. Asynchrony begets momentum, with an application to deep learning. In Annual Allerton Conference on Communication, Control, and Computing. IEEE, 997–1004.

[62] Radhika Mittal, Aman Kansal, and Ranveer Chandra. 2012. Empowering developers to estimate app energy consumption. In MobiCom. ACM, 317–328.

[63] Behnam Neyshabur, Ruslan R Salakhutdinov, and Nati Srebro. 2015. Path-sgd: Path-normalized optimization in deep neural networks. In NIPS. 2422–2430.

[64] Takayuki Nishio and Ryo Yonetani. 2019. Client selection for federated learning with heterogeneous resources in mobile edge. In MobileSoft. ACM, 330–333.

[65] Theo Ryffel, Morten Dahl, Bobby Wagner, Jason Mancuso, Daniel Rueckert, and Jonathan Passerat-Palmbach. 2018. A generic framework for privacy preserving deep learning. arXiv preprint arXiv:1811.04017 (2018).

[66] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet Talwalkar. 2017. Federated multi-task learning. In NIPS. 4424–4434.

[67] Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K Leung, Christoph Madaya, 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.

[68] Zhibo Wang, Mengkai Song, Zhiwei Zhang, Yang Song, Qian Wang, and Hairong Qi. 2019. Beyond inferring class representatives: User-level privacy leakage from federated learning. In INFOCOM. IEEE, 2512–2520.

[69] Wikipedia. 2019. Bhattacharyya coefficient. https://en.wikipedia.org/wiki/Bhattacharyya_distance.

[70] Xi Wu, Fengan Li, Arun Kumar, Kamalika Chaudhuri, Somesh Jha, and Jeffrey Naughton. 2017. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In SIGMOD. 1307–1322.

[71] Eric P Xing, Qirong Ho, Wei Dai, Jin-Kyu Kim, Seunghak Lee, Xun Zheng, Pengtao Xie, Abhimani Kumar, and Yaoliang Yu. 2015. Petuum: A New Platform for Distributed Machine Learning on Big Data. In KDD. 1335–1344.

[72] Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Cong, Daniel Ramage, and Françoise Beaufays. 2018. Applied federated learning: Improving google keyboard query suggestions. arXiv preprint arXiv:1812.02903 (2018).

[73] Chaomin Yoon, Dongwon Kim, Wonwoo Jung, Chulkang Koo, and Hojung Cha. 2012. AppScope: Application Energy Metering Framework for Android Smartphone Using Kernel Activity Monitoring. In USENIX ATC. Vol. 12. 1–14.

[74] Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. 2019. Bayesian Nonparametric Federated Learning of Neural Networks. In ICLR. 7252–7261.

[75] Wei Zhang, Suyog Gupta, Xiangru Lian, and Ji Liu. 2016. Staleness-aware asycn-sgd for distributed deep learning. In IJCAI. 2350–2356.

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