Investigating the local-global accuracy trade-off in Federated Learning

Alexandru-Andrei Iacob
Homerton College

UNIVERSITY OF CAMBRIDGE

A dissertation submitted to the University of Cambridge in partial fulfilment of the requirements for the degree of Master of Philosophy in Advanced Computer Science

University of Cambridge
Computer Laboratory
William Gates Building
15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom

Email: aai30@cam.ac.uk

May 10, 2022

Declaration

I Alexandru-Andrei Iacob of Homerton College, being a candidate for the M.Phil in Advanced Computer Science, hereby declare that this report and the work described in it are my own work, unaided except as may be specified below, and that the report does not contain material that has already been used to any substantial extent for a comparable purpose.

Total word count: Add Here

add word-count

Signed:

Date:

Abstract (1 page)

Federated Learning is a form of distributed Machine Learning which leverages local data storage and training on edge-devices. It attempts to reduce communication costs by alternating the local training steps with a global aggregation phase which combines model parameters without ever directly sharing private data. While local-only training presents distinct advantages with regard to communication efficiency and client privacy over other distributed training paradigms, it introduces significant difficulties regarding data and system heterogeneity. Most relevant to this work, clients with highly unusual data distributions may receive a worse model following the global training process than they could have trained on their own. This may be exacerbated by recent techniques which attempt to restrict how far a local model could diverge from the global one in order to improve the federated process convergence. Two primary directions for addressing the trade-off between local and global accuracy have been explored in the academic literature. The first attempts to create a "fairer" global model, thus constraining how much the accuracy of the global model can vary on the local datasets of the clients. The second relies on local adaptation techniques to construct what is effectively a customized local model for each client alongside the global one. This work brings three contributions to the field of Federated Learning. First, it provides an experimental analysis of how the distance between the global and local models affects level of fairness or local adaptation necessary to provide a given client with better results than they could have obtained individually. Second, it provides the first direct comparison between Fair Federated Learning methods and local adaptation methods in terms of their ability to. Third, it proposes a new sub-category of Federated Learning which alternates normal training and aggregation steps with local adaptation ones. The experimental results show

add results show

Contents

Add word-count

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction (2 page)

As the amount of data produced and gathered globally has grown rapidly, Machine Learning (ML) methods capable of making practical use of it have become central to the functioning of large sections of the global economy. Deep Learning (DL) methods specifically have become the de-facto standard solution to core technical challenges faced by multi-billion-dollar industries, such as Natural Language Processing [4, 32] in translation or writing assistance or Image Processing [1, 26] in computer vision. Despite their success, such methods may require prohibitive amounts of computation or high-quality training data.

The increasingly large number of internet-connected consumer devices represents a pool of both computation power and training data for ML tasks. Unlike the computational units used in classical Distributed ML [25], such devices may not be wholly dedicated to the training task. As such, they may have privacy concerns with regard to sharing their training data, computational constraints on how much they can contribute to the training process at a given time, or data transfer limitations. Furthermore, heterogeneity in device characteristics and the local data they have gathered over time represent inherent challenges of the setting. Importantly, this heterogeneity means that the ideal scenario for each device would be an entirely custom model.

In order to tackle the peculiar challenges of the setting, McMahan et al. [23]

introduced Federated Learning (FL). This distributed ML paradigm shifted the focus towards communication-efficiency and privacy preservation, in accordance with the principles of focused collection and data minimization outlined in White House [30]. The algorithm they introduced, Federated Averaging (FedAvg), has been central to the development of a wide array of direct descendants [13, 18, 19] and the field as a whole. FedAvg organizes training into rounds, at the beginning of each round the server sends a global model to all clients which then proceed train it locally before sending it back for aggregation to create a new global model. The aggregation step consists off constructing a weighted average of the model parameters from all clients according to the share of the global training data that they hold. This simple aggregation mechanism proved empirically effective, however, it's limited focus to global model convergence irrespective of the local training process of each client has been a significant challenge for subsequent research.

add extra

The main issue of concern for this work is what will be referred to as the "local-global accuracy trade-off". This topic can be broadly defined as the tendency of the final trained global model to perform worse on the local data of highly heterogeneous clients despite good performance on the global distribution [19]. In some cases, it may be possible that the global model is worse than one they could have trained entirely locally in spite of the higher training time and data availability [33]. According to Yu et al. [33], such clients are severely disincentivized from participating in the training process. While the effect is well-documented when using FedAvg, no research assessing the impact of newer aggregation algorithms which further emphasizes global model convergence—such as FedProx [18], has been found.

The existing body of work on the mitigation of this trade-off has focused on two primary means of improving performance on heterogeneous clients. Li et al. [19] propose a Fair Federated Learning (FFL) algorithm which tunes the federated objective function to focus on clients with large losses in an attempt to smooth-out the accuracy distribution of the final global model across the client data partitions. Alternatively, personalization (fine-tuning) methods have been proposed by Yu et al. [33] and Mansour et al. [22] as a means of quickly constructing effective local models for heterogeneous clients *after* training the global one using a standard FL algorithm. To date, there has not been an attempt to reconcile the two methods. Furthermore, little attention has been paid to the practical computational and engineering cost of either FFL or any of the relevant personalization algorithms.

This work makes three primary contributions to the field of Federated Learning

1. It extends the work of Yu et al. [33] by exploring the impact of personalization techniques on the FedProx [18] algorithm rather than FedAvg. FedProx purposefully limits how far a client model can diverge from the global one during one training round thus affecting highly heterogeneous clients disproportionately. Experimental results show that the local-global accuracy trade-off has a higher impact on FedProx than FedAvg and the effect scales based on the *degree* of the limitation that model divergence is subject to. edit
2. Second, it combines q-FedAvg[19] with personalization techniques to assess if fine-tuning is still beneficial in the context of a fairer accuracy distribution. This takes into account the computational cost of local adaptation as well as the disproportionate impact it has upon an ML model lifecycle. Experimental results show that models trained via q-FedAvg are in need of local adaptation less frequently. Additionally, when they do apply local adaptation it requires fewer rounds on average for the adapted model to exceed the performance of a purely local one. edit
3. Third, it proposes a new form of Federated Learning which interleaves rounds of local training, global aggregation and selective local adaptation, while keeping compatibility with pre-existing aggregation strategies. edit

Chapter 2

Background and related work (10 page)

Given the recent emergence of the field, a great deal of the published literature is concerned with the optimization of the fundamental Federated Learning process with the explicit goal of constructing a well-performing global model. Although several secondary research directions have been identified, the literature concerning them is largely exploratory in nature. As such, this chapter seeks to address the fundamentals of FL as a field broadly in the first section while taking a narrow and detailed view of research relevant to the local-global accuracy trade-off in the second.

2.1 Federated Learning

Previous sections have referenced the common cross-device client-server FL architecture, however, several alterations have been proposed in the literature. Kairouz et al. [13, sec 1] distinguishes between two versions of client-server FL. In cross-silo FL, the clients are organizations providing their siloed data. In cross-device FL, the clients are mobile or IoT devices. They have in common a large federated network with heterogeneous clients that cannot directly share data, although the cross-silo condition may be seen as somewhat closer to standard distributed ML given the higher degree of possible coordination and control over the training process. While the findings of this work will likely be relevant to both, future sections will refer to cross-device FL unless stated otherwise.

2.1.1 Foundational challenges

As previously mentioned, standard FL may can be conceptualized as an attempt to utilize latent data and computational resources on edge-devices while taking their non-committal nature into account. The data-transfer ability of clients has been established as the primary training performance bottleneck of this setting since the initial work of McMahan et al. [23]. As such, the design space of centralized client-server FL algorithms has been historically constrained by the need for communication-efficiency through the maximization of the on-device computation and minimization of data sharing. Data minimization also plays a role in the privacy-preserving aspects of FL—which are beyond the scope of this work. According to the surveys of Kairouz et al. [13] and Li et al. [20], this paradigmatic shift towards local computation and data storage lead to the following major challenges being shared across FL systems.

Look at zhang's survey

Statistical (data) heterogeneity Data generation and accrual naturally vary across devices in terms of both quantity and characteristics. Factors such as sensor capabilities, geographic location, time, or user behaviour may influence the precise deviations seen by a client [13, sec.3.1]. This heterogeneity results in data which cannot be assumed, as in traditional ML, to be Independent and Identically Distributed (IID)—as first noted by McMahan et al. [23]. Non-IID data has been shown to negatively impact both practical accuracy [11, 34] and theoretical convergence bounds [21]. A variety of contextually beneficial techniques have been used to restrict the impact of data heterogeneity by either targeting the global model [11, 13, 18, 19, 21, 34] or creating a personalized one for the final clients [2, 5, 12, 15, 22, 33].

System (hardware) heterogeneity Devices within the federated network may differ from one-another in terms of characteristics such as computational ability, storage, network speed and reliability and data-gathering hardware. They may also differ from themselves at a different point in time as their battery power, network connection, or operational mode vary. Differences in data-generating hardware, such as sensors, are linked to data heterogeneity. However, the other aspects of system heterogeneity together with device unreliability create barriers to achieving fault and straggler-tolerant algorithm capable of accommodating different client training capabilities.

add citation and relevancy to current work

2.1.2 Federated learning objective

Together, the previously mentioned challenges require that client devices, their data, and the models they train be seen as distinctly relevant entities from the global model. Despite this fact, the research to date has tended to predominantly focus on a Federated Learning objective concerned *only* with global model performance.

The standard loss function of FL is formulated by Li et al. [20] as seen in Eq. (2.1)

$$\min_{w} f(w) = \sum_{k=1}^{m} p_k F_k(w) \tag{2.1}$$

where f is the federated global loss function, m is the total number of devices, w is the model at the beginning of a round, and F_k is the local loss function of client k weighted by the associated p_k. For a total number of training examples n, p_k is typically defined as either the proportion of total training examples held by the client $\frac{n_k}{n}$ or as the inverse of the total number of clients $\frac{1}{m}$. This formulation of f does not optimize for performance on the individual data partitions and may result in skewed models for clients with a disproportionately large fraction of the global data pool. Thus, tackling the local-global accuracy trade-off necessitates changing the FL objective function either explicitly in Fair Federated Learning (Section 2.2.2) or implicitly through personalization (Section 2.2.3)

2.1.3 FedAvg

The fundamental structure of Federated Averaging, seen in Alg (Algorithm 2), has been reused to varying extents in a majority of published FL algorithms, including those used in this work—FedProx [18] and q-FedAvg [19]. Since the publication of McMahan et al. [23], the convergence of the algorithms is known to be highly dependent on the degree of statistical heterogeneity, number of stragglers, and the number aggregation rounds relative to local training steps. Specifically, a higher level of heterogeneity or number of stragglers require increasing the frequency of aggregation steps or decreasing local training steps to avoid global model divergence. Since the global data distribution cannot be known prior to training, the number of necessary aggregation rounds is unpredictable. This is inherited by q-FedAvg and directly addressed by FedProx.

Algorithm 1 Federated Averaging, adapted from McMahan et al. [23]. Each client is assumed to handle their training parameters.

```
Input: $M, K, T$
$w_0 \leftarrow$ init()
for each round $t \leftarrow 1, \ldots T$ do
    $n$, $O_t \leftarrow 0$, $\varnothing$
    $S_t \leftarrow$ K selected clients out of M
    for for each client $k \in S_t$ do
        $w_k^{t+1} \leftarrow$ train(k, $w^t$)
        if $w_k^{t+1} \notin \varnothing$ then
            $O_t \leftarrow O_t \cup k$
            $n \leftarrow n + len(k.data)$
    $w^{t+1} \leftarrow \sum_{z \in O_t} len(z.data) \times w_z^{t+1} / n$
```

2.1.4 The local-global accuracy trade-off

Complete

2.1.5 Flower

2.2 Related work

2.2.1 Limiting global model divergence

In order to tackle both statistical and systems heterogeneity, Li et al. [18] introduced FedProx as a successor to FedAvg capable of tolerating partial work from clients while smoothing the training process. It achieved this by modifying the local client training process rather than the aggregation algorithm. The objective function of a local client F_k from Eq. (2.1) is combined with a "proximal term" (Eq. (2.2))

$$\min_{w} h_k(w, w_g) = F_k(w) + \tfrac{\mu}{2}||w_g - w||_2 \tag{2.2}$$

which is meant to limit the distance of its model from the global one in accordance to the hyperparameter μ.

Discussion The experimental analysis conducted by Li et al. [18] prove that higher weighing of the proximal term allow FedProx to better handle a certain degree of statistical heterogeneity or percentage of stragglers. Of interest to this examination is the interaction between the proximal term and the local-global accuracy trade-off. Since highly heterogeneous clients would diverge more from the global model during training, larger weighing of the proximal term should impact their contribution disproportionately. This potential effect was beyond the scope of the early FL work conducted by Li et al. [18].

2.2.2 Fair Federated Learning

The canonical loss function of FL presented in 2.1 trains the sole global model without regards to the distribution of client loss values. Although no global model can fit the exact distribution of each client, this standard formulation solely emphasizes performance on the average case with regard to the global data distribution. As such, rather than incurring the training cost of constructing an additional local model per client, Li et al. [19] propose a form of FL, Fair Federated Learning

(FFL), which imposes a different distribution of model performance.

The authors construct the objective function for q-FFL, a specific version of FFL, as seen in Eq. (2.3)

$$\min_{w} f(w) = \sum_{k=1}^{m} \tfrac{p_k}{q+1} F_k^q(w) \tag{2.3}$$

where the new parameter q controls the degree of desired fairness. A value of $q = 0$ corresponds to standard FL, while larger values impose an increasingly fairer distribution. As q grows $\lim_{q \to \infty} q$, the objective function approaches optimizing solely for the client with the largest loss.

Discussion The exact formulation of Eq. (2.3) was proposed by Li et al. [19] in order to allow for tuning the degree of fairness through a single parameter. The authors drew inspiration from fair resource allocation in wireless networks [17]. Unlike other potential definitions amounting to a linear or geometric re-weighing of p_k, such as those based on generalized Gini Social-evaluation Functions introduced by Weymark [29], the exponential function of q-FFL has wide implications for the entire training process. First, Li et al. [19] show that choosing a q value is highly domain specific with optimal q-values ranging from $q = 0.001$ to $q = 5$ across evaluated tasks. Second, the exponential scaling of the loss function heavily impacts the convergence rate of the global model.

Q-FedAvg

Li et al. [19] show that their adaptation of the FedAvg algorithm to FFL, q-FedAvg (Algorithm 2), is capable of reducing the variance in global model performance across clients in a wide variety of task. By carefully tuning the value of q, the authors have shown that this is feasible without incurring a significant decrease in the test-performance of the global model or the convergence speed of the training process.

To fit the convergence rate of the new q-based exponential loss function, q-FedAvg adjusts the step-size of the client training process in relation to q. Rather than directly tuning the step size for each client, the authors choose to derive it based on

Algorithm 2 Q-FedAvg, adapted from Li et al. [19].

```
Input: $M, K, T$
$w_0 \leftarrow$ init()
for each round $t \leftarrow 1, \ldots T$ do
    $n$, $O_t \leftarrow 0$, $\varnothing$
    $S_t \leftarrow$ K selected clients out of M
    for for each client $k \in S_t$ do
        $w_k^{t+1} \leftarrow$ train(k, $w^t$)
        if $w_k^{t+1} \notin \varnothing$ then
            $\Delta w_k^t = L(w^t - w_k^{t+1})$
            $\Delta_k^t = F_k^q(w^t)\Delta w_k^t$
            $h_k^t = qF_k^{q-1}(w^t)\|w_k^t\|^2 + LF_k^q(w^t)$
            $O_t \leftarrow O_t \cup k$
    $w^{t+1} \leftarrow w^t - \sum_{z \in O_t} \Delta_k^t \,/\, \sum_{z \in O_t} h_k^t$
```

an estimation the Lipschitz constant—to which the optimal step-size is inversely related—of the functions gradient for $q = 0$.

Discussion The estimation procedure of the Lipschitz constant consists of tuning a step-size for $q = 0$ and then taking the inverse of the optimum. Despite its empirical success, this method creates difficulties in assessing the impact of modifying q-FedAvg in manners which affect the convergence rate of the algorithm. To date, no attempts to extend q-FFL beyond q-FedAvg have been made and no investigation into its interaction with personalization techniques has been found.

2.2.3 Personalization techniques

On the other end of the spectrum from FFL, personalized local models can be used to handle client heterogeneity and the local-global accuracy trade-off instead of tuning the global model. The existing literature on model personalization is extensive[13, 33], as surveyed by Kulkarni et al. [15], when compared to limited number of publications on fairness. However, the primary concern of this paper is in extending the experimental evaluation conducted by Yu et al. [33] which

add citations

add citations

covered a majority of preponderant techniques.

The experimental analysis of Yu et al. [33] established that not only does the global model perform worse on highly heterogeneous clients, as previous investigations had already noted [13, 19], it may produce worse results than training a model locally. The relevance of this result becomes clear when taking into account the significantly higher amount of data and computation used during the FL training process.

add more

Yu et al. [33] apply three distinct personalization methods which operate entirely locally and do not require server involvement beyond providing a global model.

Fine-tuning (FT) When a client receives a global model after the FL process, it can apply Fine-tuning (see Paulik et al. [24], Wang et al. [27] and Mansour et al. [22, Section D.2]) to retrain the model parameters on its own local data. To avoid potential Catastrophic forgetting [6, 8, 14], Yu et al. [33] also opt to use Freezebase (FB) as an additional variant of FT which retrains only the top layer.

As noted by Mansour et al. [22, Section 5], the performance of FT in general has only been demonstrated empirically, the technique lacks any theoretical bounds on its potential for Catastrophic forgetting. Furthermore, the last layer of a Neural Network performance has been shown to contribute disproportionately to performance in the sparsity literature [3, 7, 9], thus FB may also be susceptible to forgetting.

Knowledge Distillation (KD) As an alternative to FT, Knowledge Distillation (see Hinton et al. [10]) uses the global model as a teacher for a student client model. While KD technically allows for the two to differ in structure, Yu et al. [33] opt to maintain the same architecture across both and to initialize the student with the teachers parameters—making it possible to re-introduce the client model into FL later. For the pure logit outputs of the global model $G(x)$ and client model $C(x)$, the client minimizes Eq. (2.4)

$$l(C,x) = \alpha K^2 L(C,x) + (1-\alpha) K_L(\sigma(G(x) / K), \sigma(C(x) / K)) \tag{2.4}$$

where L is the client loss function, K_l is the Kullback-Leibler [16] divergence, σ is the activation function for the final output, α is the weighing of the clients loss and K is the temperature.

Equation (2.4) states the optimization objective as a mixture of minimizing the loss on the local client data and the KL-dstance between the temperature-adjusted outputs of G and C. It is worth noting the practical similarity of this objective to that of FedProx [18] from Eq. (2.2) as they both intend to maintain a level of similarity between the global and local models.

Multi-task Learning (MTL) The task of the global model is to perform well on the distributions of all clients while the local model must perform on the distribution of a single client. Similarly to the other techniques, the global model must be used to create a client model optimized for the local task. Framing this as a Multi-task Learning (MTL) problem in Eq. (2.5)

$$l(C,x) = L(C,x) + \sum_i \tfrac{\lambda}{2} F[i](C[i] - G[i])^2 \tag{2.5}$$

where λ determines the weighing between the two tasks, F is the Fisher information matrix and i indexes model parameters. The loss L optimizes for the client model while the second term represents the task of improving performance over all participants—i.e., using the pre-trained global model to improve performance on the local task. Multiplying the squared distance of parameters by the i-th entry of F serves the role of mitigating Catastrophic forgetting via the Elastic Weight Consolidation technique introduced by Kirkpatrick et al. [14]. The $F[i]$ term acts as a surrogate for the second derivative near minimums and serves the purpose of maintaining weights which are particularly important to the global task loss close to their initial values.

Discussion The findings of Yu et al. [33] established the strong version of the local-global accuracy trade-off and serve as foundation for the present investigation. Their experimental results established the benefits of personalization techniques as relevant for two categories of clients. Clients with generally inaccurate models obtain the largest accuracy boost both from FL generally and local adaptation particularly. Those with local models more accurate than the global one gain some benefit from participating in FL by receiving an adapted model. This paper is interested in clients falling within both categories.

Yu et al. [33] undertook a relatively narrow research scope with regard to the investigated family of aggregation algorithms. The authors experimented using FedAvg (Algorithm 2) with two variations in the form of differential privacy [28] and robust aggregation [31]—both of which have a practical effect of reducing training effectiveness. Furthermore, the paper makes no attempt to explore potential effects of re-starting the FL process following personalization.

The literature review presented in this chapter points to several potentially conflicting directions within the field of Federated Learning, caused primarily by divergent methods of handling heterogeneity Section 2.1.1. These methods reflect prioritizing either global model convergence and performance (Section 2.2.1) or its performance on individual client datasets. Prioritizing client performance may take the form of creating a global model with a performance level more resilient to their heterogeneity (Section 2.2.2) or by adapting it specifically for each client (Section 2.2.3). While the two priorities are likely irreconcilable, no evidence for the incompatibility of the two methods of improving performance on client data distributions has been found.

Read all the new personalization references

Chapter 3

Methods (10 page)

3.1 Hypotheses

1. *Allowing a higher degree of divergence from the global model during training reduces the need for local adaptation.*
2. *Using a fairer global model reduces the need for local adaptation.*
3. *Targeted local adaptation has a non-negligible cost in terms of both the number of clients adapted and total training.*
4. *Using a fair aggregation algorithm or clustering are both more efficient overall than targeted local adaptation.*

3.2 Experimental Setup

The specific set of experiments is designed as to isolate the impact of statistical heterogeneity on the global and client model accuracies. As such, they are divided according to the specific goal.

All experiments will share the following characteristics.

3.2.1 Proximal Term Tuning

The FedAvg algorithm investigated by Yu et al. [33] is unpredictable in terms of how far the local client model diverges from the global one received at the start of the round. Since this divergence is potentially very large, the entire training process may fail to converge if sufficiently heterogeneous clients are present.

The implications of this upon the accuracy of the global model on local client data is unclear. The expectation expressed in H.1 is that the heterogeneous clients will have large changes in the global model and thus impact the overall training process more resulting in less of a need for local adaptation. However, multiple client models can diverge in highly contradictory directions thus resulting in a global model that does not perform particularly well on any one of them.

To tackle this issue, the first set of experiments attempts to answer H.11 by extending the work of Yu et al. [33]. This is achieved by incorporating the proximal term from FedProx [18]. Values of $\mu \in \{0, 0.5, 1.0\}$ will be tested, $\mu = 0$ corresponds to standard FedAvg.

Importantly, the proximal term of FedProx [18] can be incorporated into the local objective function of a client regardless of the specific aggregation algorithm. As such, it can be used in conjunction with q-FedAvg[19].

3.2.2 Fairness Tuning

A modified version of q-FedAvg[19] containing the proximal term from FedProx [18] is used for this set of experiments. The overall goal is to obtain a baseline for how much accuracy can and needs to be recovered by using local adaptation techniques after the usage of a fair aggregation algorithm.

The major drawback of q-FedAvg is the need to tune the fairness parameter q. Given that Li et al. [19] do not experiment with or provide q-values for the specific datasets used in this work, new q-values must be chosen for the experimental design. The procedure for doing so will consist of a hyperparameter search over potential q-values $q \in [0, 20]$. Based on final accuracy and accuracy variance, alongside $q = 0$ two other q-values will be selected representing a moderate or high degree of fairness.

Following the q-value choice for all datasets, the general set of experiments will be ran with local adaptation techniques afterwards to again establish a baseline in terms of the potential benefits of local adaptation when using Fair Federated Learning.

3.2.3 Targeted Local Adaptation

Chapter 4

Results (10 page)

Chapter 5

Discussion (10 page)

Chapter 6

Summary and Conclusions (3-4)

Bibliography

[1] Md. Zahangir Alom, Tarek M. Taha, Chris Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C. Van Essen, Abdul A. S. Awwal, and Vijayan K. Asari. The history began from alexnet: A comprehensive survey on deep learning approaches. *ArXiv*, abs/1803.01164, 2018.

[2] Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *CoRR*, abs/1912.00818, 2019. URL `http://arxiv.org/abs/1912.00818`.

[3] Guillaume Bellec, David Kappel, Wolfgang Maass, and Robert A. Legenstein. Deep rewiring: Training very sparse deep networks. *ArXiv*, abs/1711.05136, 2018.

[4] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. Natural language processing (almost) from scratch. *CoRR*, abs/1103.0398, 2011. URL `http://arxiv.org/abs/1103.0398`.

[5] Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *CoRR*, abs/2003.13461, 2020. URL `https://arxiv.org/abs/2003.13461`.

[6] Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999. ISSN 1364-6613. doi: https://doi.org/10.1016/S1364-6613(99)01294-2. URL `https://www.sciencedirect.com/science/article/pii/S1364661399012942`.

[7] Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks. *CoRR*, abs/1902.09574, 2019. URL `http://arxiv.org/abs/1902.09574`.

[8] Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks, 2013. URL `https://arxiv.org/abs/1312.6211`.

[9] Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both weights and connections for efficient neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1135–1143, Cambridge, MA, USA, 2015. MIT Press.

[10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL `https://arxiv.org/abs/1503.02531`.

[11] Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-IID data quagmire of decentralized machine learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4387–4398. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/hsieh20a.html`.

[12] Yihan Jiang, Jakub Konečný, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *CoRR*, abs/1909.12488, 2019. URL `http://arxiv.org/abs/1909.12488`.

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

[14] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526, 2017.

[15] Viraj Kulkarni, Milind Kulkarni, and Aniruddha Pant. Survey of personalization techniques for federated learning. In *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, pages 794–797, 2020. doi: 10.1109/WorldS450073.2020.9210355.

[16] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951. doi: 10.1214/aoms/1177729694. URL `https://doi.org/10.1214/aoms/1177729694`.

[17] Tian Lan, David Kao, Mung Chiang, and Ashutosh Sabharwal. An axiomatic theory of fairness in network resource allocation. In *2010 Proceedings IEEE INFOCOM*, pages 1–9, 2010. doi: 10.1109/INFCOM.2010.5461911.

[18] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar,

and Virginia Smith. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018.

[19] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019.

[20] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[21] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

[22] Y. Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *ArXiv*, abs/2002.10619, 2020.

[23] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[24] Matthias Paulik, Matt Seigel, Henry Mason, Dominic Telaar, Joris Kluivers, Rogier van Dalen, Chi Wai Lau, Luke Carlson, Filip Granqvist, Chris Vandevelde, Sudeep Agarwal, Julien Freudiger, Andrew Byde, Abhishek Bhowmick, Gaurav Kapoor, Si Beaumont, Áine Cahill, Dominic Hughes, Omid Javidbakht, Fei Dong, Rehan Rishi, and Stanley Hung. Federated evaluation and tuning for on-device personalization: System design & applications, 2022. URL `https://arxiv.org/pdf/2102.08503.pdf`.

[25] Joost Verbraeken, Matthijs Wolting, Jonathan Katzy, Jeroen Kloppenburg, Tim Verbelen, and Jan S. Rellermeyer. A survey on distributed machine learning. *ACM Comput. Surv.*, 53(2), mar 2020. ISSN 0360-0300. doi: 10.1145/3377454. URL `https://doi.org/10.1145/3377454`.

[26] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, Eftychios Protopapadakis, and Diego Andina. Deep learning for computer vision: A brief review. *Intell. Neuroscience*, 2018, jan 2018. ISSN 1687-5265. doi: 10.1155/2018/7068349. URL `https://doi.org/10.1155/2018/7068349`.

[27] Kangkang Wang, Rajiv Mathews, Chloé Kiddon, Hubert Eichner, Françoise Beaufays, and Daniel Ramage. Federated evaluation of on-device person-

alization. *CoRR*, abs/1910.10252, 2019. URL http://arxiv.org/abs/1910.10252.

[28] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H. Yang, Farhad Farokhi, Shi Jin, Tony Q. S. Quek, and H. Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020. doi: 10.1109/TIFS.2020.2988575.

[29] John A Weymark. Generalized gini inequality indices. *Mathematical Social Sciences*, 1(4):409–430, 1981.

[30] White House. Consumer data privacy in a networked world: A framework for protecting privacy and promoting innovation in the global digital economy. *Journal of Privacy and Confidentiality*, 4(2), Mar. 2013. doi: 10.29012/jpc.v4i2.623. URL https://journalprivacyconfidentiality.org/index.php/jpc/article/view/623.

[31] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pages 5650–5659. PMLR, 2018.

[32] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing [review article]. *IEEE Computational Intelligence Magazine*, 13(3):55–75, 2018. doi: 10.1109/MCI.2018.2840738.

[33] Tao Yu, Eugene Bagdasaryan, and Vitaly Shmatikov. Salvaging federated learning by local adaptation. *arXiv preprint arXiv:2002.04758*, 2020.

[34] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data, 2018.