Towards Personalized Federated Learning

Alysa Ziying Tan[1,2,3], **Han Yu**[1*], **Lizhen Cui**[4,5] and **Qiang Yang**[6*]
[1]School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore
[2]Alibaba-NTU Singapore Joint Research Institute, NTU, Singapore
[3]Alibaba Group, Hangzhou, China
[4]School of Software, Shandong University (SDU), Jinan, China
[5]Joint SDU-NTU Centre for Artificial Intelligence Research (C-FAIR), SDU, Jinan, China
[6]Department of Computer Science and Engineering, Hong Kong University of Science and Technology
{S190109, han.yu}@ntu.edu.sg, clz@sdu.edu.cn, qyang@cse.ust.hk

Abstract

As artificial intelligence (AI)-empowered applications become widespread, there is growing awareness and concern for user privacy and data confidentiality. This has contributed to the popularity of federated learning (FL). FL applications often face data distribution and device capability heterogeneity across data owners. This has stimulated the rapid development of Personalized FL (PFL). In this paper, we complement existing surveys, which largely focus on the methods and applications of FL, with a review of recent advances in PFL. We discuss hurdles to PFL under the current FL settings, and present a unique taxonomy dividing PFL techniques into data-based and model-based approaches. We highlight their key ideas, and envision promising future trajectories of research towards new PFL architectural design, realistic PFL benchmarking, and trustworthy PFL approaches.

1 Introduction

The pervasiveness of edge devices in modern society, such as mobile phones and wearable devices, has led to the rapid growth of private data originating from distributed sources. While the wealth of such data offers vast opportunities for machine learning applications, societies are increasingly concerned about data privacy with the introduction of legislations such as the General Data Protection Regulation (GDPR) [Voigt and von dem Bussche, 2017]. This has contributed to the growing popularity of Federated Learning (FL) [Yang *et al.*, 2019], a learning paradigm that enables the development of a joint machine learning model on data silos in a collaborative and privacy-preserving manner. The key motivation for individual clients to participate in FL is to leverage the shared pool of knowledge from other clients in the federation. Individual clients often face data constraints such as data scarcity, low data quality and unseen classes that limit their capacity to train good performing local models.

The prevailing FL setting assumes a federation of clients (e.g., mobile devices, organizations) that collaboratively train a model under the orchestration of a central parameter server. The training data is stored locally and is not shared during the training process. Most of the existing training methods are variants of Federated Averaging (FedAvg), the pioneering FL algorithm introduced by [McMahan *et al.*, 2017]. The goal is to train a global model that performs well on most FL clients. Compared to local training, the globally-shared model trained through FL generalizes well to unseen data as it is trained on large amounts of data. However, these models are designed to fit the "average client". They might therefore not perform well in the presence of statistical data heterogeneity (i.e. if the local data distribution of a client deviates significantly from the global data distribution). Enabling FL to deal with heterogeneous data is important given the non-IID nature of data that originate from clients in the real world. Besides data heterogeneity, FL also needs to deal with heterogeneity in device capabilities in edge computing applications [Wu *et al.*, 2020].

In recent years, personalized federated learning (PFL) has attracted significant interest from researchers. PFL can be viewed as an intermediate paradigm between the server-based FL paradigm that produces a global model and the local model training paradigm [Mansour *et al.*, 2020]. The challenge is to strike a careful balance between local task-specific knowledge and shared knowledge useful for the generalization properties of FL models. There are several surveys on the general concepts, methods and applications of FL [Yang *et al.*, 2019; Li *et al.*, 2020a]. Others review FL from the perspectives of privacy [Mothukuri *et al.*, 2021] and robustness [Lyu *et al.*, 2020a]. There has been a short survey [Kulkarni *et al.*, 2020] that provides a brief overview of PFL. However, there is a lack of a comprehensive survey on PFL that provides a systematic perspective on this important topic.

In this paper, we bridge this gap by offering a unique data-based vs. model-based perspective for reviewing the PFL literature. We begin by analyzing the key hurdles to PFL in the current FL setting, and then present existing works following our hierarchical taxonomy while highlighting their main ideas. Additionally, we discuss common datasets for PFL benchmarking. With this perspective, we envision promising future trajectories of research towards new PFL architectural design, realistic PFL benchmarking, and trustworthy PFL approaches.

*Corresponding Authors

2 Hurdles to PFL

Limitations of the Prevailing FL Architecture: With the pioneering works in FL formulating the objective as training a single global model on data silos in a privacy-preserving manner, this has framed the prevailing FL model training under the central parameter server-based FL architecture. Recently, several works have questioned the suitability of this architecture in the presence of statistical data heterogeneity across FL clients. In [Zhao *et al.*, 2018], the authors analyzed the effect of non-IID data on FedAvg and observed a significant reduction in accuracy. They attributed this performance degradation to the phenomenon of weight divergence, as a result of the rounds of local training and synchronization on local data distributions that are non-IID. In [Li *et al.*, 2020d], the authors presented a theoretical analysis on the FedAvg algorithm and showed the slowing of convergence on non-IID data. They also highlighted the need for careful tuning of hyperparameters (e.g., learning rate decay) to improve learning stability. To build PFL models, alternatives to the central parameter server-based FL architecture are emerging.

Privacy-Preservation Constraints: As the current personalization approaches in machine learning do not adequately address privacy concerns, they cannot be directly applied to achieve personalization in the heterogeneous FL setting. The study of personalization techniques under privacy constraints remains an ongoing challenge for the FL research community. Without explicit data sharing, it is challenging to learn personalized models as the full local data distributions are concealed from external access. It is also a challenge to understand the extent of heterogeneity among the clients' datasets without access to the raw data. This has motivated several works to relax the key privacy assumption in FL to allow some local data [Zhao *et al.*, 2018; Jeong *et al.*, 2018] or metadata [Duan *et al.*, 2021] to be shared with the parameter server. Other works assume the availability of a proxy dataset that is representative of the population distribution [Yang *et al.*, 2020b; Li and Wang, 2019]. These methods may not always be applicable, especially for privacy-sensitive applications or cases where the prior distributions of the datasets are unknown.

3 PFL Approaches

This section reviews existing PFL approaches. We organize them around the proposed taxonomy (Figure 1) that divides PFL methods into data-based and model-based approaches. Data-based approaches focus on smoothing the statistical heterogeneity among clients' datasets to improve FL model convergence, while model-based approaches enhance the performance of FL models under different levels of personalization.

3.1 Data-based Approaches

Data-based approaches aim to smooth the statistical heterogeneity of data residing at participating clients. This helps to mitigate the problem of weight divergence arising from multiple rounds of local training and weight synchronization on non-IID datasets during the FL training process.

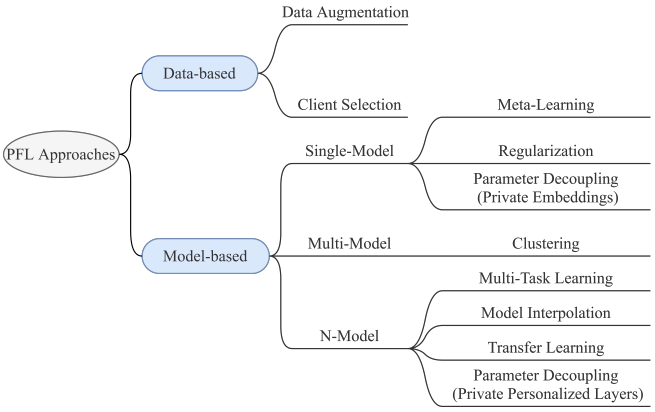

Figure 1: The proposed taxonomy of FL personalization approaches.

Data Augmentation

As the IID property of training data is a fundamental assumption in statistical learning theory, data augmentation methods to enhance the statistical homogeneity of the data have been extensively studied in the field of machine learning. Oversampling techniques involving synthetic data generation (e.g., SMOTE [Chawla *et al.*, 2002] and ADASYN [Haibo He *et al.*, 2008]), and under-sampling algorithms (e.g., Tomek links [Kubat and Matwin, 1997]) have been proposed to reduce data imbalance. These techniques, however, cannot be directly applied under the FL setting, where data residing at the clients in the federation are distributed and private.

Data augmentation in FL is highly challenging as it often requires some form of data sharing or assumes the availability of a proxy dataset that is representative of the overall data distribution. In [Zhao *et al.*, 2018], the authors proposed a data sharing strategy that distributes a small amount of global data balanced by classes to each client. Their experiments show that there is potential for significant accuracy gains ($\sim$30%) with the addition of a small amount of data. [Jeong *et al.*, 2018] proposed FAug, a federated augmentation approach that involves training a Generative Adversarial Network (GAN) in the server. Some data samples of the minority classes are uploaded to the parameter server to train the GAN model. The trained GAN model is then distributed to each client to generate additional data to augment its local data towards yielding an IID dataset. In [Duan *et al.*, 2021], the authors proposed Astraea, a self-balancing FL framework to handle class imbalance by using Z-score based data augmentation and down-sampling of local data. The FL server requires statistical information about clients' local data distributions (e.g., class sizes, mean and standard deviation).

The applicability of these data augmentation approaches under the FL setting is limited to some extent as the possibility of privacy leakage from data sharing has not been eliminated by design. Additionally, the assumption of a proxy dataset consisting of a representative population distribution is strong and may not always be achievable.

Client Selection

Another line of work focuses on designing client selection mechanisms to enable sampling from a more homogeneous

data distribution, with the aim of improving model generalization performance. In [Wang *et al.*, 2020a], the authors proposed FAVOR which selects a subset of participating clients for each training round in order to mitigate the bias introduced by non-IID data. A deep Q-learning formulation for client selection was designed with the objective of maximizing accuracy while minimizing the number of communication rounds. In a similar approach, a client selection algorithm based on the Multi-Armed Bandit formulation was proposed in [Yang *et al.*, 2020b] to select the subset of clients with minimal class imbalance. The local class distributions are estimated by comparing the similarity between the local gradient updates submitted to the parameter server with the gradients inferred from a balanced proxy dataset residing on the server.

In [Lyu *et al.*, 2020b], the authors proposed the Fair and Privacy-Preserving Deep Learning (FPPDL) approach to select clients based on the perceived value of their local datasets to each other's learning tasks in order to dynamically establish FL model training collaborations among data owners. A local credibility mutual evaluation mechanism is proposed to guarantee fairness, while a three-layer onion-style encryption scheme is proposed to achieve both model accuracy and data privacy. Different from the prevailing FL paradigm, FPPDL is a fully distributed federated learning paradigm in which data owners self-organize based on trust without the need for a dedicated parameter server. At the end of the FL model training process under FPPDL, each client receives a model with performance reflecting how much value the community of FL clients place on its local data.

3.2 Model-based Approaches

Although data-based approaches improve solution convergence when learning the global model, they are limited in their capacity to personalize as they still involve training a single global model. Modifying the local data distributions may also result in the loss of valuable information associated with the inherent diversity of client behaviors that can be useful for personalizing the global model for each client.

In contrast, model-based personalization approaches aim to enable FL models to adapt to the diverse data distributions among clients. We divide them into Single-Model, Multi-Model and N-Model approaches. This naming convention indicates the number of models trained during the FL process. Techniques that follow the original single global model design are classified under the single-model category, while those that produce a personalized model for different subsets of clients fall under the multi-model category. Techniques that produce a personalized model for each of the N clients in the federation are classified under the N-model category.

Single-Model PFL Approaches

For Single-Model PFL approaches, a global FL model is trained and adapted on local data. Personalization is achieved through finding a more optimal initialization of the global model for local personalization, or by learning more desirable task-specific local representations. As these approaches are formulated based on a single global model design, they are best suited when the local data distributions do not deviate significantly from the global data distribution.

Meta-learning: Commonly known as "learning to learn", meta-learning aims to improve the learning algorithm itself through the exposure to a variety of tasks. This enables the model to learn a new task quickly and effectively. Optimization-based meta-learning algorithms, like Model-Agnostic Meta-Learning (MAML) [Finn *et al.*, 2017] and Reptile [Nichol *et al.*, 2018], are known for their good generalization and fast adaptation on new heterogeneous tasks. They are also model-agnostic and can be applied to any gradient descent-based approaches, enabling applications in problems such as supervised learning and reinforcement learning.

In [Jiang *et al.*, 2019], the authors drew parallels between meta-learning algorithms and FL. Meta-learning algorithms run in two phases: meta-training and meta-testing. They mapped the meta-training step in MAML to the FL global model training process, and meta-testing to FL personalization where a few steps of gradient descent are performed on local data during local adaptation. They also show that FedAvg is analogous to the Reptile algorithm, and are in fact equivalent when all clients possess equal amounts of local data. [Fallah *et al.*, 2020] proposed Per-FedAvg, a variant of FedAvg that builds on the MAML formulation. In contrast to FedAvg where the goal is to learn a global model that performs well on most participating clients, the new goal is transformed to learn a good initial global model that performs well on a new heterogeneous task. This problem formulation is suitable for learning on heterogeneous data distributions to provide an improved global model initialization for personalization. The authors in [Khodak *et al.*, 2019] proposed the ARUBA framework, an online learning algorithm for adaptive meta-learning. When applied to FedAvg, it improves model generalization performance and eliminates the need for hyperparameter optimization during personalization.

Regularization: Model regularization is a common strategy to prevent overfitting when training machine learning models. In FL, regularization techniques can be applied to achieve convergence stability and improve model generalization. Several studies employ regularization to tackle the weight divergence problem in FL settings involving non-IID local data. FedProx [Li *et al.*, 2020b] introduces a proximal term to the local subproblem that considers the dissimilarity of global and local models to limit the impact of local updates. Along with model dissimilarity, FedCurv [Shoham *et al.*, 2019] and FedCL [Yao and Sun, 2020] further consider parameter importance in the regularized local loss function. Using Elastic Weight Consolidation (EWC) [Kirkpatrick *et al.*, 2017] from the field of continual learning, parameter importance is estimated and penalization steps are carried out to preserve important parameters. This can prevent catastrophic forgetting when moving across learning tasks.

Besides improving convergence stability, regularization can also enforce desirable behaviors that can improve personalization. It has been used in [Huang *et al.*, 2021] to model an attentive collaboration mechanism to enforce stronger collaboration amongst FL clients with similar data distributions. In [Guo *et al.*, 2020], the authors proposed an approach that aligns model predictions with personal humour preferences through a local adapter at each FL client.

Parameter Decoupling (Private Embeddings): Another

line of work achieves personalization in FL by decoupling the local private model parameters from the global federated model parameters. Private parameters are trained locally and not shared with the parameter server. This enables task specific representations to be learnt for better personalization.

In [Bui *et al.*, 2019], a document classification model using a Bidirectional LSTM architecture is trained via FL by treating user embeddings as private parameters, and treating character embeddings, LSTM and MLP layers as federated parameters. In [Liang *et al.*, 2020], the authors proposed the Local Global Federated Averaging (LG-FedAvg) algorithm that combines local representation learning and global federated training. Learning lower dimensional local representations improves communication and computational efficiency for federated global model training. It also offers flexibility as specialized encoders can be designed based on the source data modality (e.g., image, text). The authors also demonstrated how fair and unbiased representations that are invariant to protected attributes (e.g., race, gender) can be learnt by incorporating adversarial learning into the FL model training.

Multi-Model PFL Approaches

For applications where there are inherent partitions among clients or data distributions that are significantly different, training a single global model is not effective. A multi-model approach where an FL model is trained for each homogeneous client cluster is more suitable.

Clustering: Several recent works focus on clustering for FL personalization. The underlying assumption of FL clustering approaches is the existence of a natural grouping of clients' local datasets. In [Sattler *et al.*, 2019], the authors proposed integrating hierarchical clustering into FL as a post-processing step. An optimal bi-partitioning algorithm based on cosine similarity is used divide the FL clients into clusters. Another hierarchical clustering framework for FL has been proposed in [Briggs *et al.*, 2020]. The approach is designed for a wider range of non-IID settings and allows training on a subset of clients during each round of FL model training. The formulation reduces clustering to a single step to lower computation and communication loads. In [Huang *et al.*, 2019], the authors proposed a community-based FL algorithm to predict patient hospitalization time and mortality. They train a denoising autoencoder and cluster patients based on the encoded features of their private data. An FL model is then trained for each cluster. In [Xie *et al.*, 2020], the authors proposed a multi-center formulation that learns multiple global models. Expectation Maximization is used to derive the optimal matching of clients to each cluster center.

Federated clustering frameworks often incur high computation and communication costs, which may limit their feasibility for practical applications. Additional architectural components for the management and deployment of the clustering mechanism (e.g., cluster allocation) are also required.

N-Model PFL Approaches

For N-Model PFL approaches, a model is learnt for each individual client in the federation. Broadly, these approaches are well-suited for applications with diverse data distributions, or when the heterogeneity of the underlying data distributions is unclear. Fine-tuning of hyperparameters may be required to improve model convergence and learning stability.

Multi-task Learning (MTL): The goal of MTL is to train a model that jointly learns several related tasks. This improves generalization by leveraging domain-specific knowledge across the learning tasks. In [Smith *et al.*, 2017], the authors observed that by treating each FL client as a task in MTL, there is potential to learn and capture relationships in heterogeneous data across the clients. They proposed the MOCHA algorithm which extends distributed MTL to FL using a primal-dual optimization method. The algorithm addresses communication and system challenges prevalent in FL which are not considered in the field of MTL. Unlike the conventional FL design which learns a single global model, MOCHA learns a personalized model for each individual FL client. While MOCHA improves personalization, it is not suitable for cross-device FL applications as all clients are required to participate in every round of FL model training. Another drawback of MOCHA is that it is only applicable to convex models and is thus unsuitable for deep learning implementations. This motivated [Corinzia and Buhmann, 2019] to propose the VIRTUAL federated MTL algorithm designed for non-convex models. Recently, [Dinh *et al.*, 2021] proposed FedU, a framework for federated MTL with Laplacian regularization. It has achieved improved performance over FedAvg, MOCHA and personalized meta-learning approaches like Per-FedAvg [Fallah *et al.*, 2020].

In contrast to single-model approaches in which the model is trained by aggregating model updates from the local clients, the MTL formulation improves model generalization by learning the underlying collaboration relationships among heterogeneous datasets. Such approaches implicitly assume that there exists inherent partitioning of data and inter-task relationships to be learnt.

Model Interpolation: To address the limitations of a global shared model in the prevailing FL paradigm, [Hanzely and Richtárik, 2020] proposed a new formulation that learns a mixture of global and local models. In this formulation, each FL client learns an individual local model. A penalty parameter λ is used to discourage the local models from being too dissimilar from the mean model. Pure local model learning occurs when λ is set to zero. As λ increases, mixed model learning occurs and the local models become increasingly similar to each other. The setting approximates the global shared FL model formulation when λ approaches infinity. In this way, the degree of personalization can be controlled.

In a related line of work, [Deng *et al.*, 2020] proposed the APFL algorithm with the goal of finding the optimal combination of global and local models in a communication-efficient manner. They introduced a mixing parameter α that is adaptively learnt during the FL training process to control the balance between the global and local models. This enables the optimal degree of personalization for each client to be learnt. The weighting factor on a particular local model is expected to be larger if the local and global data distributions are not well-aligned, and vice versa. A similar formulation involving the joint optimization of local and global models to determine the optimal interpolation weight has been proposed in [Mansour *et al.*, 2020].

Recently, [Diao *et al.*, 2021] proposed the HeteroFL framework which trains local models of varying computational complexities, based on a single global model. By adaptively allocating local models of varying complexity levels according to the computation and communication capabilities of each client, it achieves PFL relating to client capability heterogeneity in edge computing scenarios.

Transfer Learning: Transfer learning has been used for model personalization in non-federated settings. It aims to transfer knowledge from a source to a target domain, where the domains are often different but related. There have been a number of studies of federated transfer learning (FTL) in the healthcare domain to improve model personalization (e.g., FedHealth [Chen *et al.*, 2020] and FedSteg [Yang *et al.*, 2020a]). The approach generally involves: (i) learning local models by fine-tuning a pre-trained model on local data; (ii) training a global model via FL; and (iii) training personalized models by integrating the global and local models via transfer learning. A correlation alignment layer [Sun *et al.*, 2016] is added before the softmax layer for adaptation between the source and target domains. In [Li and Wang, 2019], the authors proposed FedMD, an FL framework based on transfer learning and knowledge distillation which allows clients to design independent models using their own private data. Before the FL training phase, transfer learning is first carried out by each individual client by training a model to convergence on a public dataset which is then fine-tuned on local data.

Parameter Decoupling (Private Personalized Layers): In parameter decoupling, the classification of private and federated parameters is an architectural design decision. In [Arivazhagan *et al.*, 2019], the authors proposed a base + personalized layers design for deep feed-forward neural networks. Personalized layers are kept private at the clients for local training, while the base layers are shared with the FL server.

4 PFL Benchmarking

Realistic datasets are important for the development of the PFL research field. LEAF [Caldas *et al.*, 2019] is one of the earliest and most popular benchmarking frameworks proposed for FL. At the time of writing, it provides 6 FL datasets covering a range of machine learning tasks including image classification, language modeling and sentiment analysis under both IID and non-IID settings. Examples of datasets include the Extended MNIST dataset split according to the writers of the character digits, and the Shakespeare dataset split according to the characters in the play. A set of accuracy and communication metrics, along with implementation references for well-known approaches such as FedAvg, SGD and MOCHA are also provided. As LEAF extends existing public datasets from traditional machine learning settings, it does not fully reflect the data heterogeneity in FL scenarios. Although there are a few real-world federated datasets, such as a street image dataset for object detection [Luo *et al.*, 2019] and a species dataset for image classification [Hsu *et al.*, 2020], they are often limited in size. To facilitate PFL research, datasets that include more modalities like audio, video and sensor signals, and involve a broader range of machine learning tasks from real-world applications are required.

5 Promising Future Research Directions

Based on the above review of current PFL techniques, we envision promising future trajectories of research towards new PFL architectural design, realistic benchmarking, and trustworthy PFL approaches.

5.1 Opportunities for PFL Architectural Design

FL Client Data Distribution Analytics: The inherent heterogeneity of data among FL clients is a key consideration when assessing the type of PFL required. For example, a multi-model approach is preferred for applications where there are inherent partitions or data distributions that are significantly different. In order to facilitate experimentation on non-IDD data, recent works in PFL have proposed metrics like Total Variation, 1-Wasserstein [Fallah *et al.*, 2020] and Earth Mover's Distance (EMD) [Zhao *et al.*, 2018] to quantify the statistical heterogeneity of data distributions. However, these metrics can only be calculated with direct access to raw data. The problem of how to perform heterogeneity diagnostics in a privacy-preserving manner remains open.

Aggregation Procedure: In more complex PFL scenarios, performing a simple average on the local updates of each client may not be an ideal approach in handling data heterogeneity. Model averaging is adopted in most prevailing FL architectures, and its effectiveness as an aggregation method has not been well-studied from a theoretical perspective [Xiao *et al.*, 2020]. Recently, [Wang *et al.*, 2020b] proposed a layer-wise matched averaging formulation for CNN and LSTM architectures. The design of specialized aggregation procedures for PFL architectures remains to be explored.

PFL Architecture Search: In the presence of statistical heterogeneity, federated neural architectures are highly sensitive to hyperparameter choices and may therefore experience poor learning performance if they are not tuned carefully [Li *et al.*, 2020d]. The choice of the FL model architecture may also not be optimal, leading to poor fitting of the underlying non-IID distribution. Neural Architecture Search (NAS)-based [Zhu *et al.*, 2021] PFL is a promising research direction to reduce manual design effort to optimize the model architecture with the given PFL scenario.

Spatial Adaptability refers to the ability of the FL system to handle variations across client datasets as a result of (i) the addition of new clients, and/or (ii) dropouts and stragglers. These are practical issues prevalent in complex edge computing environments, where there is significant variability in hardware capabilities in terms of computation, memory, power and network connectivity [Wu *et al.*, 2020].

(i) Existing FL personalization approaches commonly assume a fixed client pool at the start of an FL training cycle, and that new clients cannot join the training process midway [Jeong *et al.*, 2018; Briggs *et al.*, 2020]. Other approaches involve a pre-training step [Li and Wang, 2019] that require time for local computation. Besides meta-learning approaches [Fallah *et al.*, 2020] that encourage fast learning on a new client, there is limited work addressing the cold-start problem in PFL. Current deep FL techniques are also prone to catastrophic forgetting of previously learnt knowledge when new clients join, due to the stability-plasticity dilemma in

neural networks [Kemker *et al.*, 2018]. As a result, existing clients may experience a degradation in performance. A promising direction is to incorporate continual learning [Delange *et al.*, 2021] into FL to mitigate catastrophic forgetting.

(ii) With the prevalence of dropouts and stragglers in large-scale federated systems due to network, communication and computation constraints, it is necessary to design for robustness in FL systems. Developing communication-efficient algorithms to mitigate the problem of stragglers is an ongoing research direction, where gradient compression [Lin *et al.*, 2018] and asynchronous model updates [Chen *et al.*, 2019] are common strategies for addressing FL communication bottlenecks. These issues require further study in PFL to formalize the trade-offs between overhead and performance.

Temporal Adaptability refers to the ability of the FL system to learn on non-stationary data. In dynamic real-world systems, we may expect temporal changes in the underlying data distributions. This phenomenon is known as concept drift. Concept drift learning is commonly classified into three key components: (i) drift detection (whether drift has occurred); (ii) drift understanding (when, how and where the drift occurs); and (iii) drift adaptation (response to drift) [Lu *et al.*, 2018]. [Casado *et al.*, 2021] is one of the few works that study the problem of concept drift in FL. It extends FedAvg with the Change Detection Technique (CDT) for drift detection. It remains an open direction to leverage existing drift detection and adaptation algorithms to improve learning on dynamic real-world data in federated systems.

5.2 Opportunities for PFL Benchmarking

The establishment of systematic evaluation methodologies and metrics is also important for PFL research. In most existing studies, the evaluation of PFL algorithms is limited to a single type of non-IID setting such as quantity skew [Zhao *et al.*, 2018], feature distribution skew [Chen *et al.*, 2020] or label distribution skew [Li and Wang, 2019]. The experiments are performed by either leveraging an existing pre-partitioned public dataset (e.g., LEAF) or prepared by partitioning a public dataset to fit the target non-IID setting. For more effective research and fairer comparison, it is imperative for the research community to develop a deeper understanding of the different non-IID settings in real-world federated learning in order to simulate diverse realistic non-IID settings. The broad categorization of non-IID settings into: (i) Feature distribution skew; (ii) Label distribution skew; (iii) Quantity skew; (iv) Same label, different features; and (v) Same features, different labels [Kairouz *et al.*, 2019] can be a starting point for further research. Such an effort requires wider collaboration among researchers and industry practitioners, and will be beneficial for building a healthy PFL research ecosystem.

5.3 Opportunities for Trustworthy PFL

Fairness: As machine learning technologies become more widely adopted by businesses to support decision-making, there has been a growing interest in developing methods to ensure fairness in order to avoid undesirable ethical and social implications [Mehrabi *et al.*, 2019; Holstein *et al.*, 2019].

Current approaches do not adequately address the unique set of fairness related challenges presented in FL. This includes new sources of bias introduced by the diversity of participating FL clients due to unequal local data sizes, activity patterns, location, and connection quality, etc. [Kairouz *et al.*, 2019]. The study of fairness in FL is still in its infancy and the framing of fairness in FL has not yet been well-defined. A common notion of fairness in the literature is the satisfaction of accuracy parity. As it is not practical to enforce an equal accuracy for all clients where there is significant variability in large-scale FL settings, the authors in [Mohri *et al.*, 2019] introduced the notion of good-intent fairness. Its goal is to ensure that the model does not overfit to any client at the expense of others, which is achieved by maximizing the performance of the worst performing FL client. In [Li *et al.*, 2020c], fairness is defined in terms of minimizing the variability of error rate distributions. [Zhang *et al.*, 2020] uses client contribution to FL model training to measure fairness.

The study of fairness in FL is mostly focused on the prevailing server-based FL paradigm, although new work on fairness in alternative FL paradigms is emerging [Lyu *et al.*, 2020b]. As FL approaches maturity, advances in improving fairness for PFL in particular will become increasingly important in order for FL to be adopted at scale.

Explainability: Explainable Artificial Intelligence (XAI) is an active research area that has attracted significant interest recently, driven by pressure from government agencies and the general public for interpretable models [Arrieta *et al.*, 2020]. It is important for models in high stake applications such as healthcare to be explainable, where there is a strong need to justify decisions made [Tonekaboni *et al.*, 2019].

Explainability has not yet been systematically explored in the FL literature. There are complex challenges unique to achieving explainability in PFL due to the scale and heterogeneity of distributed datasets. Striving for FL model explainability may also be associated with potential privacy risks from inadvertent data leakage, as demonstrated in [Shokri *et al.*, 2020] where certain gradient-based explanation methods are prone to privacy leakage. There is limited work addressing both explainability and privacy objectives simultaneously. In [Harder *et al.*, 2020], the authors proposed using differentially private locally linear maps (LLMs) to approximate the mapping from the model to the input data. However, the model formulation is limited to a non-federated setting and has only been evaluated on simple datasets.

Developing an FL framework that balances the trade-off between explainability and privacy is an important future research direction. One possible approach to achieve this trade-off is to incorporate explainability into the global FL model but not the personalization component of the FL model.

Robustness: Although FL offers better privacy protection compared to traditional centralized model training approaches, recent research has exposed vulnerabilities of FL that could potentially compromise data privacy [Lyu *et al.*, 2020a]. It is therefore of paramount importance to study FL attack methods and develop defensive strategies to counteract these attacks in order to ensure robustness of the FL system. With more complex protocols and architectures developed for PFL, more work is needed to study related forms of attacks and defenses to enable robust PFL approaches to emerge.

References

[Arivazhagan *et al.*, 2019] Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated Learning with Personalization Layers. *arXiv:1912.00818*, 2019.

[Arrieta *et al.*, 2020] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, and Javier Del Ser et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[Briggs *et al.*, 2020] C. Briggs, Z. Fan, and P. Andras. Federated learning with hierarchical clustering of local updates to improve training on non-IID data. In *IJCNN*, pages 1–9, 2020.

[Bui *et al.*, 2019] Duc Bui, Kshitiz Malik, Jack Goetz, Honglei Liu, Seungwhan Moon, Anuj Kumar, and Kang G. Shin. Federated User Representation Learning. *arXiv:1909.12535*, 2019.

[Caldas *et al.*, 2019] Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. LEAF: A Benchmark for Federated Settings. *arXiv:1812.01097*, 2019.

[Casado *et al.*, 2021] Fernando E Casado, Dylan Lema, Roberto Iglesias, Carlos V Regueiro, and Senén Barro. Concept drift detection and adaptation for robotics and mobile devices in federated and continual settings. In *Advances in Physical Agents II*, pages 79–93. Springer, 2021.

[Chawla *et al.*, 2002] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *JAIR*, 16:321–357, 2002.

[Chen *et al.*, 2019] Yang Chen, Xiaoyan Sun, and Yaochu Jin. Communication-efficient federated deep learning with layerwise asynchronous model update and temporally weighted aggregation. *IEEE TNNLS*, 31(10):4229–4238, 2019.

[Chen *et al.*, 2020] Yiqiang Chen, Xin Qin, Jindong Wang, Chaohui Yu, and Wen Gao. Fedhealth: A federated transfer learning framework for wearable healthcare. *IEEE Intell. Syst.*, 35(4):83–93, 2020.

[Corinzia and Buhmann, 2019] Luca Corinzia and Joachim M. Buhmann. Variational Federated Multi-Task Learning. *arXiv:1906.06268*, 2019.

[Delange *et al.*, 2021] Matthias Delange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE TPAMI*, 2021.

[Deng *et al.*, 2020] Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive Personalized Federated Learning. *arXiv:2003.13461*, 2020.

[Diao *et al.*, 2021] Enmao Diao, Jie Ding, and Vahid Tarokh. Heterofl: Computation and communication efficient federated learning for heterogeneous clients. *ICLR*, 2021.

[Dinh *et al.*, 2021] Canh T Dinh, Tung T Vu, Nguyen H Tran, Minh N Dao, and Hongyu Zhang. FedU: A unified framework for federated multi-task learning with laplacian regularization. *arXiv preprint arXiv:2102.07148*, 2021.

[Duan *et al.*, 2021] Moming Duan, Duo Liu, Xianzhang Chen, Renping Liu, Yujuan Tan, and Liang Liang. Self-Balancing Federated Learning With Global Imbalanced Data in Mobile Systems. *IEEE T. Parallel Dist. Syst.*, 32(1):59–71, 2021.

[Fallah *et al.*, 2020] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *NeurIPS*, volume 33, pages 3557–3568, 2020.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135, 2017.

[Guo *et al.*, 2020] Xu Guo, Pengwei Xing, Siwei Feng, Boyang Li, and Chunyan Miao. Federated Learning for Personalized Humor Recognition. *arXiv:2012.01675*, 2020.

[Haibo He *et al.*, 2008] Haibo He, Yang Bai, E. A. Garcia, and Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *IJCNN*, pages 1322–1328, 2008.

[Hanzely and Richtárik, 2020] Filip Hanzely and Peter Richtárik. Federated Learning of a Mixture of Global and Local Models. *arXiv:2002.05516*, 2020.

[Harder *et al.*, 2020] Frederik Harder, Matthias Bauer, and Mijung Park. Interpretable and differentially private predictions. In *AAAI*, pages 4083–4090, 2020.

[Holstein *et al.*, 2019] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudík, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *CHI*, pages 1–16, 2019.

[Hsu *et al.*, 2020] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Federated visual classification with real-world data distribution. In *ECCV*, pages 76–92, 2020.

[Huang *et al.*, 2019] Li Huang, Andrew L Shea, Huining Qian, Aditya Masurkar, Hao Deng, and Dianbo Liu. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *Journal of Biomedical Informatics*, 99, 2019.

[Huang *et al.*, 2021] Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-iid data. In *AAAI*, 2021.

[Jeong *et al.*, 2018] Eunjeong Jeong, Seungeun Oh, Hyesung Kim, Jihong Park, Mehdi Bennis, and Seong-Lyun Kim. Communication-Efficient On-Device Machine Learning: Federated Distillation and Augmentation under Non-IID Private Data. In *NeurIPS*, 2018.

[Jiang *et al.*, 2019] Yihan Jiang, Jakub Konečný, Keith Rush, and Sreeram Kannan. Improving Federated Learning Personalization via Model Agnostic Meta Learning. *arXiv:1909.12488*, 2019.

[Kairouz *et al.*, 2019] Peter Kairouz, H. Brendan McMahan, and Brendan Avent et al. Advances and Open Problems in Federated Learning. *arXiv:1912.04977*, 2019.

[Kemker *et al.*, 2018] Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. Measuring catastrophic forgetting in neural networks. In *AAAI*, pages 3390–3398, 2018.

[Khodak *et al.*, 2019] Mikhail Khodak, Maria-Florina Balcan, and Ameet Talwalkar. Adaptive Gradient-Based Meta-Learning Methods. In *NeurIPS*, pages 5917–5928, 2019.

[Kirkpatrick *et al.*, 2017] James Kirkpatrick, Razvan Pascanu, and Neil Rabinowitz et al. Overcoming catastrophic forgetting in neural networks. *PNAS*, 114(13):3521–3526, 2017.

[Kubat and Matwin, 1997] Miroslav Kubat and Stan Matwin. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In *ICML*, pages 179–186, 1997.

[Kulkarni *et al.*, 2020] V. Kulkarni, M. Kulkarni, and A. Pant. Survey of Personalization Techniques for Federated Learning. In *WorldS4*, pages 794–797, 2020.

[Li and Wang, 2019] Daliang Li and Junpu Wang. FedMD: Heterogenous Federated Learning via Model Distillation. *FL-NeurIPS*, 2019.

[Li *et al.*, 2020a] 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.

[Li *et al.*, 2020b] Tian Li, Anit Kumar Sahu, Manzil Zaheer, and et al. Federated Optimization in Heterogeneous Networks. *MLSys*, 2:429–450, 2020.

[Li *et al.*, 2020c] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *ICLR*, 2020.

[Li *et al.*, 2020d] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the Convergence of FedAvg on Non-IID Data. In *ICLR*, 2020.

[Liang *et al.*, 2020] Paul Pu Liang, Terrance Liu, and Liu Ziyin et al. Think Locally, Act Globally: Federated Learning with Local and Global Representations. *FL-NeurIPS*, 2020.

[Lin *et al.*, 2018] Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William Dally. Deep Gradient Compression: Reducing the communication bandwidth for distributed training. In *ICLR*, 2018.

[Lu *et al.*, 2018] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang. Learning under concept drift: A review. *IEEE TKDE*, 31(12):2346–2363, 2018.

[Luo *et al.*, 2019] Jiahuan Luo, Xueyang Wu, Yun Luo, Anbu Huang, Yunfeng Huang, Yang Liu, and Qiang Yang. Real-World Image Datasets for Federated Learning. *FL-NeurIPS*, 2019.

[Lyu *et al.*, 2020a] Lingjuan Lyu, Han Yu, and Qiang Yang. Threats to Federated Learning: A Survey. *FL-IJCAI*, 2020.

[Lyu *et al.*, 2020b] Lingjuan Lyu, Jiangshan Yu, Karthik Nandakumar, Yitong Li, Xingjun Ma, Jiong Jin, Han Yu, and Kee Siong Ng. Towards fair and privacy-preserving federated deep models. *IEEE TPDS*, 31(11):2524–2541, 2020.

[Mansour *et al.*, 2020] Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three Approaches for Personalization with Applications to Federated Learning. *arXiv:2002.10619*, 2020.

[McMahan *et al.*, 2017] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *AISTATS*, pages 1273–1282, 2017.

[Mehrabi *et al.*, 2019] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. *arXiv:1908.09635*, 2019.

[Mohri *et al.*, 2019] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic Federated Learning. In *ICML*, pages 4615–4625, 2019.

[Mothukuri *et al.*, 2021] Viraaji Mothukuri, Reza M. Parizi, Seyedamin Pouriyeh, Yan Huang, Ali Dehghantanha, and Gautam Srivastava. A survey on security and privacy of federated learning. *FGCS*, 115:619–640, 2021.

[Nichol *et al.*, 2018] Alex Nichol, Joshua Achiam, and John Schulman. On First-Order Meta-Learning Algorithms. *arXiv:1803.02999*, 2018.

[Sattler *et al.*, 2019] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered Federated Learning: Model-Agnostic Distributed Multi-Task Optimization under Privacy Constraints. *arXiv:1910.01991*, 2019.

[Shoham *et al.*, 2019] Neta Shoham, Tomer Avidor, Aviv Keren, Nadav Israel, Daniel Benditkis, Liron Mor-Yosef, and Itai Zeitak. Overcoming Forgetting in Federated Learning on Non-IID Data. *FL-NeurIPS*, 2019.

[Shokri *et al.*, 2020] Reza Shokri, Martin Strobel, and Yair Zick. On the privacy risks of model explanations. In *WHI-ICML*, 2020.

[Smith *et al.*, 2017] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet Talwalkar. Federated Multi-Task Learning. In *NeurIPS*, pages 4427–4437, 2017.

[Sun *et al.*, 2016] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, pages 2058–2065, 2016.

[Tonekaboni *et al.*, 2019] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. What clinicians want: contextualizing explainable machine learning for clinical end use. In *MLHC*, pages 359–380, 2019.

[Voigt and von dem Bussche, 2017] Paul Voigt and Axel von dem Bussche. *The EU General Data Protection Regulation (GDPR)*. Springer International Publishing, 2017.

[Wang *et al.*, 2020a] Hao Wang, Zakhary Kaplan, Di Niu, and Baochun Li. Optimizing Federated Learning on Non-IID Data with Reinforcement Learning. In *IEEE INFOCOM*, pages 1698–1707, 2020.

[Wang *et al.*, 2020b] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. In *ICLR*, 2020.

[Wu *et al.*, 2020] Q. Wu, K. He, and X. Chen. Personalized federated learning for intelligent iot applications: A cloud-edge based framework. *IEEE OJ-CS*, 1:35–44, 2020.

[Xiao *et al.*, 2020] Peng Xiao, Samuel Cheng, Vladimir Stankovic, and Dejan Vukobratovic. Averaging is probably not the optimum way of aggregating parameters in federated learning. *Entropy*, 22(3), 2020.

[Xie *et al.*, 2020] Ming Xie, Guodong Long, Tao Shen, Tianyi Zhou, Xianzhi Wang, and Jing Jiang. Multi-Center Federated Learning. *arXiv:2005.01026*, 2020.

[Yang *et al.*, 2019] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated Machine Learning: Concept and Applications. *ACM TIST*, 10(2):1–19, 2019.

[Yang *et al.*, 2020a] Hongwei Yang, Hui He, Weizhe Zhang, and Xiaochun Cao. FedSteg: A Federated Transfer Learning Framework for Secure Image Steganalysis. *IEEE TNSE*, 2020.

[Yang *et al.*, 2020b] Miao Yang, Akitanoshou Wong, Hongbin Zhu, Haifeng Wang, and Hua Qian. Federated learning with class imbalance reduction. *arXiv:2011.11266*, 2020.

[Yao and Sun, 2020] Xin Yao and Lifeng Sun. Continual Local Training for Better Initialization of Federated Models. In *IEEE ICIP*, pages 1736–1740, 2020.

[Zhang *et al.*, 2020] Jingfeng Zhang, Cheng Li, Antonio Robles-Kelly, and Mohan Kankanhalli. Hierarchically Fair Federated Learning. *arXiv:2004.10386*, 2020.

[Zhao *et al.*, 2018] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated Learning with Non-IID Data. *arXiv:1806.00582*, 2018.

[Zhu *et al.*, 2021] Hangyu Zhu, Haoyu Zhang, and Yaochu Jin. From federated learning to federated neural architecture search: a survey. *Complex & Intelligent Systems*, pages 1–19, 2021.