A Survey of Federated Learning and Its Intersection with LoRA, Finetuning, Personalization, and Edge AI

Abstract

Federated Learning (FL) has emerged as a transformative paradigm in AI, enabling decentralized model training while preserving data privacy. This survey paper explores the integration of FL with key technologies such as LoRA, finetuning, personalization, distributed computing, and edge AI, highlighting their roles in enhancing model adaptability, efficiency, and personalization. FL addresses critical challenges in traditional centralized methods by maintaining data privacy and optimizing communication efficiency, particularly in heterogeneous environments. The paper delves into the architecture and principles of FL, emphasizing the importance of managing data heterogeneity and communication bottlenecks. The integration of LoRA and finetuning techniques offers solutions for efficient model adaptation and resource utilization. Personalization in FL is underscored as a crucial factor for improving model performance across diverse client data landscapes. The survey further examines the role of distributed computing and model optimization in achieving scalable and efficient FL systems. The convergence of edge AI with FL facilitates real-time processing and latency reduction, essential for applications requiring immediate responses. Hierarchical and decentralized architectures are explored for their potential to enhance scalability and efficiency. The paper concludes with insights into future research directions, emphasizing the need for adaptive mechanisms, enhanced privacy protection, and robust defense strategies to expand FL's applicability across various domains. These advancements position FL as a pivotal force in advancing AI technologies, ensuring secure and effective collaborative learning in real-world scenarios.

1 Introduction

1.1 Federated Learning in the AI Landscape

Federated Learning (FL) has emerged as a pivotal paradigm in artificial intelligence, enabling edge devices or clients to collaboratively train machine learning models while keeping their private data secure [1]. This capability is particularly significant in heterogeneous data environments where traditional centralized approaches struggle to accommodate diverse data distributions across clients [2]. By preserving local data privacy, FL facilitates collaborative training, effectively addressing challenges related to data heterogeneity and privacy concerns.

Typically, FL architecture involves a centralized server coordinating model training among distributed devices. This setup, however, faces challenges due to stringent communication and bandwidth constraints as the number of participating devices increases [3]. Nonetheless, FL reduces communication overhead by allowing local model training, sharing only model updates with the server, thus optimizing latency and enhancing system efficiency.

FL's impact spans various domains, including healthcare, where it enables secure data sharing among institutions while maintaining patient confidentiality [4]. In personalized services, FL allows for the fine-tuning of pre-trained models on local devices, customizing AI solutions to individual preferences

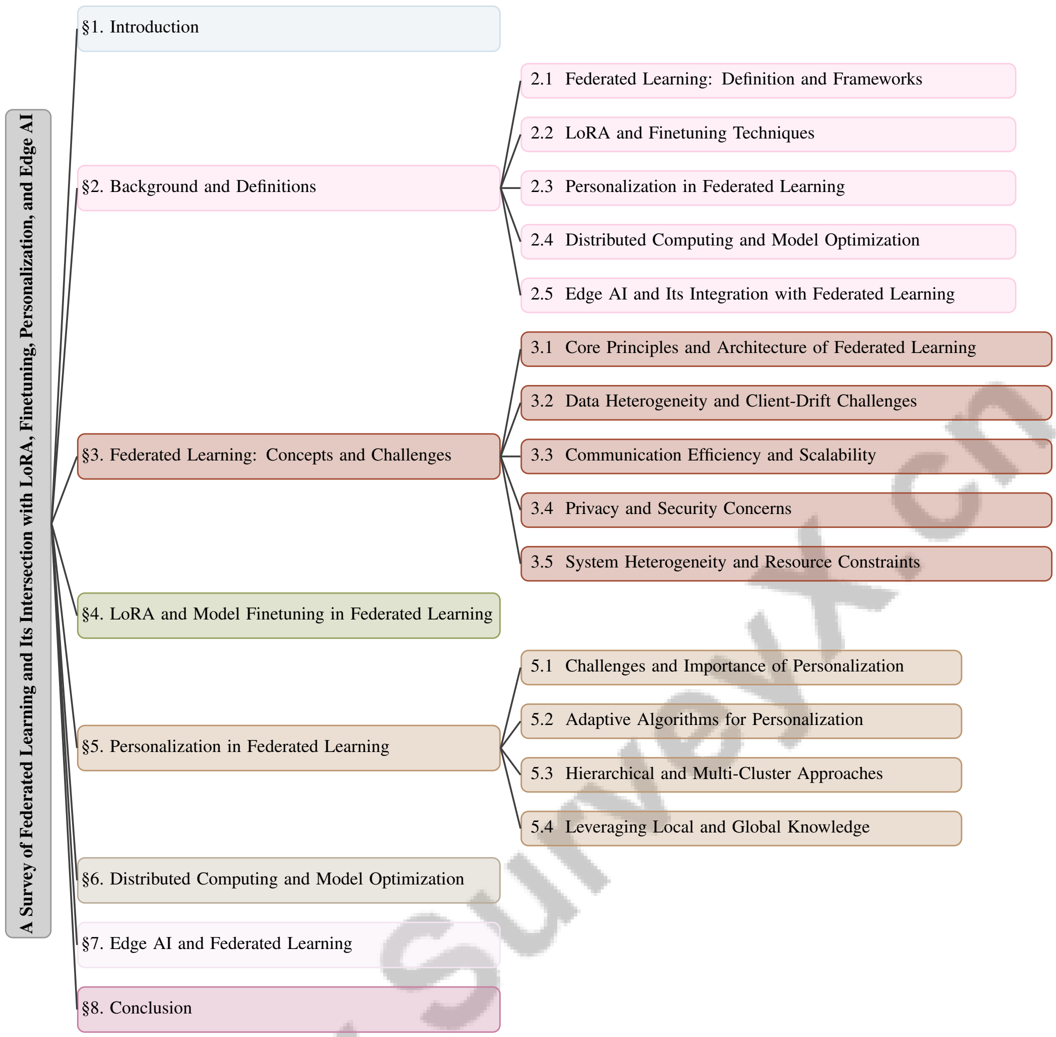

Figure 1: chapter structure

without compromising privacy [2]. As FL evolves, it enhances model adaptability and accessibility, establishing itself as a transformative force in AI with profound implications for future applications.

1.2 Challenges Addressed by Federated Learning

Federated Learning (FL) addresses critical challenges inherent in traditional centralized machine learning frameworks, focusing on privacy, data security, and computational efficiency. By decentralizing the learning process, FL ensures raw data remains on local devices, preserving privacy and enhancing data security, particularly in sensitive domains like healthcare where compliance with stringent regulations is essential [4].

A significant challenge that FL tackles is the inefficiency in resource utilization and communication costs during model training on edge devices with limited hardware resources [5]. The heterogeneity of data and devices affects computational capabilities and data distribution, leading to inefficiencies and fairness issues in model training [6]. This often results in local models diverging from the global model, reducing accuracy and generalization.

Communication bottlenecks pose another challenge, as frequent model updates between numerous edge devices and central servers can cause delays, especially in bandwidth-constrained environments [5]. The scalability of FL methods can be limited by their inability to fully exploit the parallel computing power of multiple clients while maintaining effective communication and data privacy.

AI-generated, for reference only.

The straggler effect, where devices with limited computational power lead to infrequent gradient updates, can hinder training performance.

Moreover, federated learning systems are vulnerable to model poisoning attacks, where malicious clients send corrupted updates that degrade the global model's performance. Traditional distributed machine learning systems, which rely on central servers for model updates, present significant security risks and compromise user confidentiality. In contrast, FL allows model training directly on edge devices without transmitting sensitive data to the cloud, preserving privacy while maintaining performance comparable to centralized models. This approach is particularly beneficial in fields like digital health, where user data protection is paramount [7, 8]. By addressing these multifaceted challenges, FL advances distributed machine learning, ensuring secure and efficient model training across diverse environments. The continuous evolution of FL methodologies promises to enhance scalability, security, and communication efficiency, broadening the applicability and impact of federated learning systems.

1.3 Privacy-Preserving Machine Learning

Privacy-preserving machine learning has become a critical focus in AI development due to increasing demands for data privacy and security, as evidenced by advancements in federated learning techniques that facilitate model training on edge devices without transferring sensitive data to the cloud. This approach is particularly beneficial in sectors like digital health, enabling effective collaboration across diverse datasets while minimizing the risk of data breaches [9, 10, 11, 7, 8]. FL addresses these concerns by enabling decentralized model training while ensuring raw data remains on client devices, thereby mitigating privacy risks and reducing communication overhead.

In the realms of the Internet of Things (IoT) and edge computing, FL is advantageous for upholding data privacy and security amidst bandwidth constraints [12]. By facilitating on-device model updates, FL supports privacy-preserving machine learning, crucial for mobile and edge applications that require adherence to legal privacy standards and protection against data poisoning attacks.

Innovative frameworks like FedMA enhance privacy-preserving capabilities by resolving technical challenges such as permutation invariance through layer-wise parameter matching [13]. Additionally, cooperative federated edge learning (CFEL) offers a scalable solution by utilizing a network of cooperative edge servers, minimizing reliance on central cloud servers [14].

FL's role in privacy-preserving machine learning is further exemplified in real-time model updates for computer vision, where frameworks like FedVision enable updates without sensitive data leaving local devices [15]. Despite challenges such as statistical heterogeneity and distribution shifts, FL continues to evolve, providing robust solutions to privacy challenges in distributed machine learning environments [16].

By addressing the inefficiencies of traditional AI algorithms due to centralized data collection [17], federated learning significantly contributes to privacy-preserving machine learning. It provides a framework that balances collaborative model training with stringent privacy requirements, facilitating AI deployment in privacy-sensitive domains.

1.4 Overview of Paper Structure

This survey delves into the multifaceted architecture of federated learning (FL) systems, examining their functional components, types of parallelism, aggregation algorithms, data manipulation techniques, and various FL frameworks [18]. The paper is structured to provide a comprehensive exploration of FL's intersection with key machine learning concepts and technologies.

We introduce federated learning, a groundbreaking approach in artificial intelligence that enables collaborative model training across multiple devices while preserving individual data confidentiality. Its critical importance lies in addressing significant challenges within the AI landscape, particularly those related to data privacy and security. As organizations face risks associated with centralized data storage, federated learning offers a viable solution, allowing for the sharing of insights without compromising sensitive information. We outline various forms of federated learning—such as horizontal, vertical, and transfer learning—and discuss inherent security vulnerabilities that must be navigated to foster trust and facilitate broader adoption in sensitive sectors like healthcare and finance [9, 10, 19]. Following this, we define and explore core concepts, including LoRA, finetuning,

AI-generated, for reference only.

personalization, distributed computing, model optimization, and edge AI, elucidating their roles and interrelations in the context of federated learning.

The survey transitions to a detailed discussion on the principles and challenges of federated learning, such as data heterogeneity, communication efficiency, and privacy concerns, providing a comparative assessment of federated learning against traditional centralized methods [20]. We examine the role of LoRA and model finetuning in enhancing federated learning's efficiency and personalization, exploring strategies and algorithms that tailor models to individual user preferences.

Subsequent sections analyze the integration of distributed computing and model optimization techniques, focusing on their contributions to improving performance and efficiency in federated learning environments. We also explore the intersection of edge AI and federated learning, emphasizing real-time processing, latency reduction, and the challenges of resource-constrained environments.

The survey culminates in a comprehensive synthesis of key findings and insights regarding federated learning, emphasizing its current state and integration with emerging technologies. It highlights the advantages of federated learning—such as enhanced data privacy and scalability—while addressing significant challenges, including statistical and system heterogeneity, communication bottlenecks, and security concerns. Furthermore, we outline potential future research directions aimed at overcoming these challenges and advancing the field, reflecting on the latest methodologies, frameworks, and applications shaping the landscape of federated learning today [21, 22, 23]. This structured approach ensures a thorough understanding of federated learning and its pivotal role in advancing AI applications across diverse domains.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Federated Learning: Definition and Frameworks

Federated Learning (FL) is a decentralized machine learning paradigm that allows model training across a network of devices while ensuring data remains localized, crucial for privacy in sensitive sectors such as healthcare and finance [5, 12]. FL is typically framed as regularized empirical risk minimization, accommodating the diverse capabilities and characteristics of local data [24].

FL systems are divided into cross-silo, involving a few powerful clients like institutions, and cross-device, characterized by numerous less powerful clients such as mobile devices [25]. A major challenge in FL is data heterogeneity, especially non-IID data among clients, which can degrade model performance.

To address these challenges, several frameworks have been developed. FedAvg is widely used for managing non-IID data by aggregating model updates at a central server to improve convergence [12]. CE-FedAvg enhances scalability and reduces latency by enabling collaborative training across multiple edge servers [14]. ConFederated Learning (CFL) connects multiple edge servers with individual devices, facilitating decentralized collaboration without overloading a central server [3].

Frameworks like Federated Continual Learning (FCL) tackle client drift in continual learning scenarios [26]. The Heroes framework optimizes resource use and training performance through advanced neural composition and adaptive local updates [27]. Agnostic Federated Learning (AFL) optimizes a centralized model for any target distribution formed by a mixture of client distributions [28].

Advancements in FL frameworks are addressing critical issues such as data heterogeneity, communication efficiency, and resource optimization. Innovations like personalized model training and robust aggregation techniques are enhancing training efficiency and accuracy, reinforcing FL's role in advancing AI technologies across various domains [29, 30, 31].

2.2 LoRA and Finetuning Techniques

Low-Rank Adaptation (LoRA) and model finetuning are pivotal in federated learning, enhancing model adaptability and efficiency in decentralized environments. LoRA employs low-rank matrices for fine-tuning large models, significantly reducing computational overhead while maintaining performance across diverse client devices [32]. This is particularly beneficial in federated settings with constrained computational resources and data heterogeneity challenges [33].

Frameworks like FL-TAC apply LoRA by using task-specific low-rank adapters for each client, facilitating efficient fine-tuning and model update aggregation [32]. This approach improves convergence and ensures personalized models reflect individual client data characteristics. By using specialized compression algorithms and focusing on salient parameters, LoRA supports creating resource-efficient sub-models tailored for federated scenarios [34].

Model finetuning adapts pre-trained models to new datasets, enhancing accuracy and relevance in specific applications. Strategies like PFIT and Personalized Federated Task Tuning (PFTT) focus on personalization while reducing communication overhead in federated environments [33]. Integrating finetuning techniques with federated frameworks, such as MAFL, allows collaborative training of local models while respecting data network structures, further improving accuracy and convergence [34]. These advancements underscore the importance of LoRA and finetuning techniques in optimizing federated learning systems for efficiency and adaptability.

2.3 Personalization in Federated Learning

Personalization in federated learning (FL) is essential for addressing the unique data distributions of individual clients, enhancing user experience and model performance. The primary challenge in FL is statistical heterogeneity and non-I.I.D. client data, necessitating customized models to leverage these unique characteristics effectively [35].

Personalization adapts global models to meet individual client needs, often unmet by traditional global model approaches [36]. Strategies to enhance personalization include client-specific model parameter adaptations, yielding more accurate outputs [37]. By empowering clients to maintain ownership of personalized models, pFL ensures effectiveness and efficiency across diverse environments [35].

Implementing pFL techniques is vital for overcoming challenges posed by device, data, and model heterogeneity. Clustering clients into groups and fitting personalized models for each cluster can manage data diversity and improve personalized model efficacy [37]. This strategy boosts performance and ensures scalability and adaptability to various client needs.

2.4 Distributed Computing and Model Optimization

Distributed computing is fundamental to federated learning (FL), enabling collaborative model training across numerous decentralized devices, enhancing computational efficiency and scalability [38]. This approach manages large data volumes and complex models typical in FL, allowing parallel execution across devices. Techniques such as data, model, and pipeline parallelism optimize resource allocation and improve training efficiency [38].

Communication inefficiency poses a significant challenge in distributed computing within FL, particularly as model sizes and communication rounds increase [39]. Techniques like FedAST use buffered asynchronous aggregation, reducing communication overhead from frequent updates [1]. The CFL framework employs random scheduling to select devices for updates, significantly cutting communication costs [3].

Model optimization in FL aims to enhance the performance and efficiency of decentralized models. Efficient Adaptive Federated Optimization (EAFO) dynamically balances communication, computation, and model precision trade-offs, improving efficiency [17]. Techniques like FedLPS use a shared encoder among multiple task-specific predictors, employing adaptive pruning to minimize model size and resource usage [40]. These strategies accommodate diverse client capabilities and ensure efficient training across heterogeneous environments.

Hybrid federated-centralized learning approaches, like HFCL, offer flexibility by allowing clients to send either model updates or local datasets to a parameter server based on their computational capabilities [41]. This adaptability optimizes resource utilization and enhances federated learning performance. The survey categorizes existing research by training phases—client selection, configuration, and reporting—providing a structured view of optimization strategies [42].

Advancements in distributed computing and model optimization are crucial for progressing federated learning, enabling efficient and scalable training across diverse, resource-constrained environments. These strategies address communication efficiency, resource optimization, and model performance challenges, expanding FL's applicability across various domains [43]. However, current studies often overlook FL's robustness against various attacks, and the complexity of implementing effective defenses remains a significant challenge [19].

2.5 Edge AI and Its Integration with Federated Learning

Edge AI involves deploying artificial intelligence models on devices at the network's edge, enabling real-time data processing and decision-making with minimal latency. This approach is beneficial in scenarios requiring immediate responses, such as autonomous vehicles, industrial automation, and smart city applications [44]. Integrating Edge AI with federated learning (FL) creates a powerful paradigm that combines decentralized learning benefits with edge device computational efficiency.

The E-Tree Learning framework illustrates this integration, enhancing model accuracy and convergence through hierarchical learning structures [45]. This framework shows how edge devices can collaboratively train models, improving performance without transferring large data volumes to a central server. By processing data locally, FL minimizes communication overhead and preserves data privacy, crucial for applications involving sensitive information [8].

The Edge-Cloud Collaborative Training (ECCT) framework emphasizes the synergy between edge devices and cloud resources, enabling real-time, privacy-sensitive feature extraction at the edge while leveraging the cloud's computational power for complex model training tasks [46]. This collaboration enhances personalization and model accuracy by allowing adaptive learning based on immediate context and user-specific data at the edge.

Platforms like FedVision support the integration of Edge AI with FL, facilitating online visual object detection and safety monitoring in smart city projects [44]. These applications highlight Edge AI's potential for real-time analytics and decision-making in distributed environments where latency and privacy are paramount.

Additionally, tools such as the Python Testbed for Federated Learning Algorithms (PTB-FLA) offer a runtime environment for developing and testing federated learning algorithms, simulating edge scenarios and evaluating FL system performance in resource-constrained environments [47].

3 Federated Learning: Concepts and Challenges

Understanding the core principles and architecture of federated learning (FL) is crucial for addressing the challenges inherent in decentralized model training. FL enables collaborative learning while preserving data privacy, facilitating its application across various domains. As depicted in Figure 2, this figure illustrates the primary concepts and challenges in federated learning, categorizing them into core principles and architecture, data heterogeneity and client-drift, communication efficiency and scalability, privacy and security concerns, and system heterogeneity and resource constraints. Each category is further divided into specific challenges and strategies, highlighting the complex interplay of factors affecting federated learning systems. This section explores the foundational concepts and architectural frameworks of FL, emphasizing decentralized data processing and innovative strategies that enhance model performance and resource utilization, setting the stage for deeper discussions in subsequent sections.

3.1 Core Principles and Architecture of Federated Learning

Federated Learning (FL) is a paradigm that enables distributed machine learning across decentralized devices while ensuring data privacy and minimizing communication overhead. FL typically involves edge devices conducting local training and relaying updates to a central server, crucial for resource efficiency and scalability [48]. This decentralized approach is effective in industries like healthcare, smart cities, and autonomous driving, where data privacy is paramount [12].

A key principle of FL is managing model heterogeneity, which requires flexible aggregation strategies to accommodate diverse client architectures [43]. Decentralized methods, such as cooperative edge servers, enhance scalability and reduce latency, addressing limitations of traditional centralized approaches [14]. The CFL framework exemplifies this by employing decentralized architecture and random scheduling, minimizing communication costs while maintaining model accuracy through local computations [3].

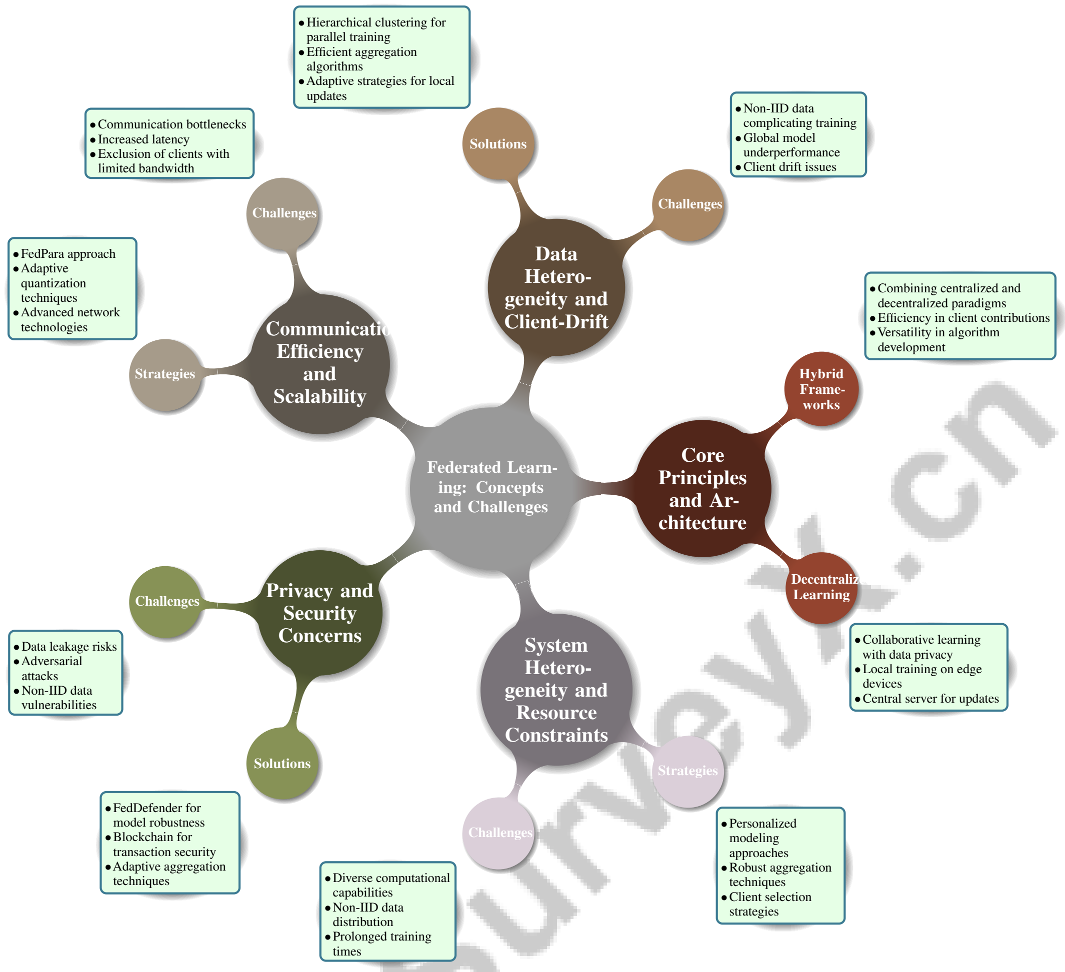

Figure 2: This figure illustrates the primary concepts and challenges in federated learning, categorizing them into core principles and architecture, data heterogeneity and client-drift, communication efficiency and scalability, privacy and security concerns, and system heterogeneity and resource constraints. Each category is further divided into specific challenges and strategies, highlighting the complex interplay of factors affecting federated learning systems.

Hybrid frameworks like HFCL augment FL effectiveness by combining centralized and decentralized paradigms, enabling efficient client contributions regardless of computational capabilities [41]. Similarly, PTB-FLA supports diverse learning paradigms, providing a versatile platform for developing and testing FL algorithms [47].

Communication efficiency remains a significant challenge, particularly with the costs of transmitting large model updates. Strategies utilizing cooperative edge servers for aggregation and innovative fine-tuning methods like FL-TAC, which employs low-rank task-specific adapters, are crucial for reducing these costs while enhancing model convergence and personalization [32].

Figure 3 illustrates the core principles, architectures, and challenges of Federated Learning (FL), highlighting key concepts such as data privacy, decentralized methods, and scalability issues. This visual representation underscores the intricate balance required to address these challenges while leveraging the advantages of FL in various applications.

3.2 Data Heterogeneity and Client-Drift Challenges

Data heterogeneity and client drift are significant challenges in federated learning (FL) due to the non-IID nature of client data, complicating training and impairing model convergence [26]. This heterogeneity can lead to a global model underperforming compared to locally trained models, as it

AI-generated, for reference only.

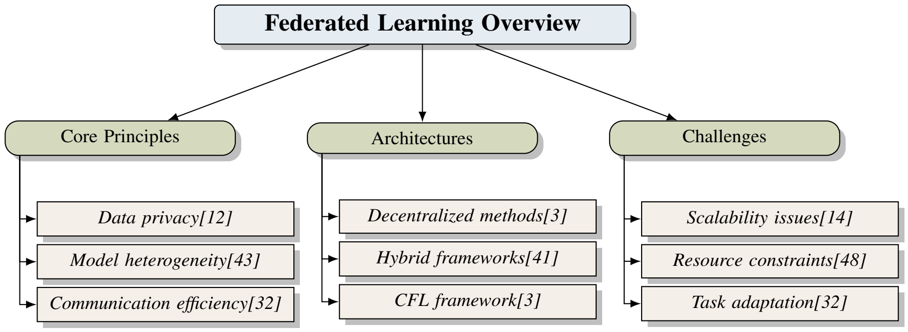

Figure 3: This figure illustrates the core principles, architectures, and challenges of Federated Learning (FL), highlighting key concepts such as data privacy, decentralized methods, and scalability issues.

may not accurately represent diverse client data distributions [6]. Statistical heterogeneity can cause model updates from one client to overwrite learned parameters from others, resulting in client drift and suboptimal global performance [26].

Challenges are intensified by asynchronous device participation, leading to stale updates and inconsistent global model aggregation [42]. Methods like FederatedAveraging (FedAvg) do not guarantee convergence under heterogeneous data distributions, leading to inefficiencies in communication and model training [24]. The lack of adaptive strategies for balancing local updates and parameter compression further complicates FL system optimization [17].

Device heterogeneity presents additional challenges, as disparities in processing power and training times hinder optimization of training round durations [49]. Frameworks like ConFederated Learning (CFL) may impose higher computational demands on resource-limited devices [3].

Strategies such as hierarchical clustering enable independent and parallel model training for different client clusters, improving performance in non-IID settings [42]. Efficient aggregation algorithms are essential for coordinating model updates while maintaining privacy and security, enhancing federated learning systems' robustness and scalability. Addressing these challenges is crucial for improving FL applicability and effectiveness across various domains.

3.3 Communication Efficiency and Scalability

Communication efficiency and scalability are pivotal challenges in federated learning (FL), as the distributed nature of data and computation requires effective coordination among numerous decentralized devices. Communication bottlenecks arise from the need for clients to transmit full model updates to the server in each training round, causing delays and potentially excluding clients with limited bandwidth [50]. This issue is exacerbated by the requirement for multiple communication rounds in existing FL methods, leading to increased latency and inefficiency in aggregating updates from distributed devices [51].

To enhance communication efficiency and ensure scalable model training, several strategies have been developed. The FedPara approach achieves comparable accuracy with significantly reduced communication costs, demonstrating compatibility with other federated learning methods [52]. Adaptive quantization techniques, such as those in Hier-Local-QSGD, improve communication efficiency and convergence speed, further optimizing the FL process [39].

Advanced network technologies, including 6G networks, offer promising solutions for optimizing communication efficiency in FL. By leveraging computing and network convergence, FL systems can achieve more efficient and robust model training across diverse environments [53]. This is particularly vital in scenarios with heterogeneous devices and variable channel conditions, where reconfigurable intelligent surfaces can enhance communication efficiency and learning accuracy [54].

The FedPAQ method addresses communication bottlenecks by reducing the frequency of communication rounds and increasing local computation, thus minimizing delays and improving overall efficiency [5]. Furthermore, optimizing task assignments in heterogeneous environments, as proposed

AI-generated, for reference only.

by Pilla et al., contributes to minimizing the duration of FL rounds by controlling the data used for training by each device [49].

3.4 Privacy and Security Concerns

Privacy and security concerns are paramount in federated learning (FL) systems, given the decentralized nature of model training, which introduces risks of data leakage and adversarial attacks. Protecting sensitive information during training is crucial, especially when data is shared among various providers or outsourced to cloud services [33]. Unverified local model submissions can compromise the global model, underscoring the necessity for robust security measures [11].

FL inherently addresses some privacy issues by enabling collaborative model training without transferring raw data, thus complying with regulations like GDPR [55]. However, the non-IID nature of client data can lead to performance degradation and increased vulnerability to adversarial attacks [56]. Techniques such as FedDefender have been developed to bolster the robustness of FL models against model poisoning, significantly enhancing performance when integrated with existing server-side defense strategies [57].

The complexity of decentralized evaluation methods and data privacy issues complicates the assessment of client contributions [58]. Current research often falls short in adequately addressing these privacy concerns and the computational limitations inherent in FL systems [2]. Moreover, reliance on public datasets for knowledge distillation raises privacy concerns and limits the applicability of these methods in real-world scenarios [43].

Decentralized architectures, including those leveraging blockchain technology, have been proposed to enhance robustness against model poisoning attacks by securing transactions and ensuring integrity [59]. Despite these advancements, the interpretability of models and the robustness of FL systems against adversarial attacks remain underexplored [38]. Addressing these multifaceted challenges necessitates a comprehensive approach that balances data privacy, communication efficiency, and model robustness, ensuring secure and effective collaborative learning across decentralized environments [60]. By employing adaptive aggregation techniques and enhancing the utilization of diverse data, FL systems can achieve improved accuracy and resilience against privacy threats.

3.5 System Heterogeneity and Resource Constraints

System heterogeneity and resource constraints are significant challenges in federated learning (FL), primarily due to the diverse computational capabilities and data distributions across client devices. The variability among systems complicates efficient and consistent model training, as clients differ widely in processing power, memory, and network connectivity [30]. This disparity can lead to inefficiencies and prolonged training times, particularly when high-performance clients are scarce, as existing methods like neural composition do not fully optimize all parameters [27].

The non-IID nature of client data exacerbates these challenges, resulting in skewed model updates that hinder convergence and degrade global model performance [61]. Addressing these issues necessitates robust aggregation techniques and personalized modeling approaches that can accommodate the statistical heterogeneity inherent in FL environments [30]. Client selection strategies are crucial for mitigating resource constraints by prioritizing clients based on computational capabilities and data quality, thereby optimizing the overall training process.

The interplay between system heterogeneity—characterized by variations in data distribution, quality, and computational capabilities—and resource constraints in FL environments highlights the urgent need for innovative approaches. Solutions must effectively navigate the trade-offs between computational efficiency and model accuracy, especially in decentralized contexts where data privacy is critical, and diverse client capabilities can lead to performance bottlenecks. Strategies such as personalized models, robust aggregation techniques, and adaptive training methods are essential for optimizing model performance while ensuring fairness and efficiency across various operational environments [30, 43, 62, 63, 64]. By exploring these interdependencies and developing adaptive approaches, FL systems can enhance scalability and effectiveness, ensuring robust model training across diverse and resource-constrained environments.

4 LoRA and Model Finetuning in Federated Learning

4.1 Introduction to LoRA in Federated Learning

Low-Rank Adaptation (LoRA) enhances federated learning (FL) by improving model finetuning and resource efficiency in decentralized settings. By incorporating low-rank matrices, LoRA reduces computational and communication burdens, aligning with FL's objectives of optimizing resource usage while preserving data privacy [5]. This approach is particularly advantageous for devices with limited resources, enabling efficient training on decentralized data with minimal data transmission.

LoRA's implementation in FL is exemplified by frameworks utilizing adaptive optimization techniques to stabilize model updates and mitigate client drift, thereby enhancing model adaptability and performance. For instance, the Heroes framework improves neural composition and modulates local update frequencies to address resource limitations and client heterogeneity. Additionally, Layer Selective Training in Federated Learning (LST-FL) enhances resource efficiency by allowing edge devices to randomly select a subset of model layers for training in each communication round. This selective strategy reduces server load and communication costs, cutting data transmission by approximately 75

Architectural innovations, such as low-rank parameterization methods, further decrease communication costs without compromising model performance. The CFEL framework demonstrates scalability and high accuracy through multiple cooperative edge servers, underscoring LoRA's effectiveness in enhancing federated learning efficiency. Moreover, algorithms like FedPAQ improve convergence speed in FL systems by balancing communication overhead, computational demands, and model precision, illustrating the practical benefits of LoRA techniques in optimizing FL frameworks, particularly in scenarios with client data heterogeneity and resource constraints [65, 66, 67, 31, 68].

4.2 Benefits of LoRA for Model Finetuning

Incorporating Low-Rank Adaptation (LoRA) into federated learning (FL) frameworks offers significant advantages for model finetuning, especially in environments characterized by data heterogeneity and resource constraints. The use of low-rank matrices effectively reduces computational and communication overhead, enhancing the efficiency and scalability of FL systems [3]. This reduction is crucial for achieving faster convergence rates and improved performance, particularly in contexts with non-IID data distributions [6].

LoRA's adaptability allows for dynamic adjustments in model update contributions, resulting in enhanced convergence speed and accuracy compared to traditional FL methods [24]. This flexibility enables LoRA to efficiently manage diverse client capabilities, minimizing the impact of stragglers and optimizing training performance. Techniques such as hybrid pruning identify smaller subnetworks for each client, facilitating personalized parameter averaging that improves performance under non-IID conditions [6].

Furthermore, LoRA enhances training efficiency in bandwidth-limited environments through model sparsification and linear compression techniques [5]. The OLAR framework exemplifies the ability to produce optimal task assignments while accommodating heterogeneous device constraints, significantly boosting training efficiency [49].

Additionally, LoRA contributes to reduced communication costs and increased robustness against device failures. Methods like FedPAQ enhance communication efficiency by combining periodic averaging and quantization techniques, significantly lowering the communication load compared to traditional federated learning approaches [5].

Figure 4 illustrates the hierarchical benefits of Low-Rank Adaptation (LoRA) in model finetuning, highlighting key areas of efficiency improvements, communication cost reduction, and training optimization within federated learning frameworks. This visual representation complements the aforementioned advantages, providing a clear overview of how LoRA facilitates enhanced performance in diverse FL environments.

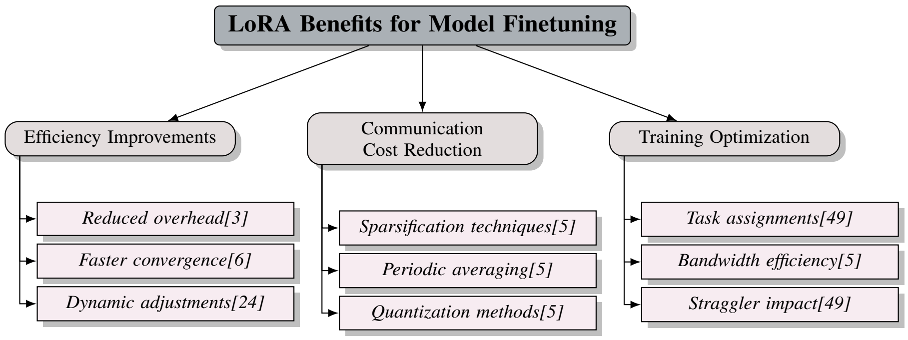

Figure 4: This figure illustrates the hierarchical benefits of Low-Rank Adaptation (LoRA) in model finetuning, highlighting key areas of efficiency improvements, communication cost reduction, and training optimization within federated learning frameworks.

5 Personalization in Federated Learning

5.1 Challenges and Importance of Personalization

Personalization in federated learning (FL) is essential for effectively managing data heterogeneity and distribution shifts in decentralized environments. Achieving a balance between personalization and generalization is crucial, as demonstrated by PERADA, which enhances model performance across diverse conditions [69]. Adapting to concept drift is necessary to improve user experience and prevent increased error rates due to rapid data distribution changes [70].

Figure 5 illustrates the challenges and solutions related to personalization in federated learning, highlighting the importance of balancing personalization with generalization, techniques to enhance personalization, and methods to handle resource constraints. A significant challenge in personalization involves developing models that learn from individual client data while maintaining shared learning benefits. This is particularly important for personalized large language models (LLMs) that must utilize diverse client data without compromising privacy [33]. Techniques like the stateless FedFOR, which achieves faster convergence, are well-suited for real-world FL applications with numerous clients [71].

Personalization enhances model performance through knowledge distillation, allowing for the management of diverse data and labels while preserving privacy [56]. Adaptive client selection strategies, such as those in SRPFL, improve efficiency and robustness in personalized federated learning, particularly in heterogeneous settings [16].

Resource constraints pose another challenge, as low-memory devices may struggle to participate effectively in FL. Selective model layer training is crucial for reducing communication costs and enhancing personalization [48]. Additionally, server-side defenses like FedDefender improve robustness against model poisoning, ensuring that personalized models remain secure and effective [57].

5.2 Adaptive Algorithms for Personalization

Adaptive algorithms are crucial for enhancing personalization in FL by dynamically adjusting to individual clients' unique data distributions and computational capabilities. These algorithms address challenges related to data heterogeneity and the non-IID nature of client data. Strategies such as personalized model training, robust aggregation techniques, and client selection methods enhance federated learning model performance and user satisfaction across diverse environments [72, 30, 43, 73, 6].

Personalization can be achieved through adaptive learning rates and model parameters fine-tuned to each client's data characteristics. Techniques like hierarchical clustering and adaptive client selection group clients with similar data distributions, enabling more accurate and efficient model training [42]. These strategies optimize resource allocation and improve convergence rates, even in highly diverse environments [35].

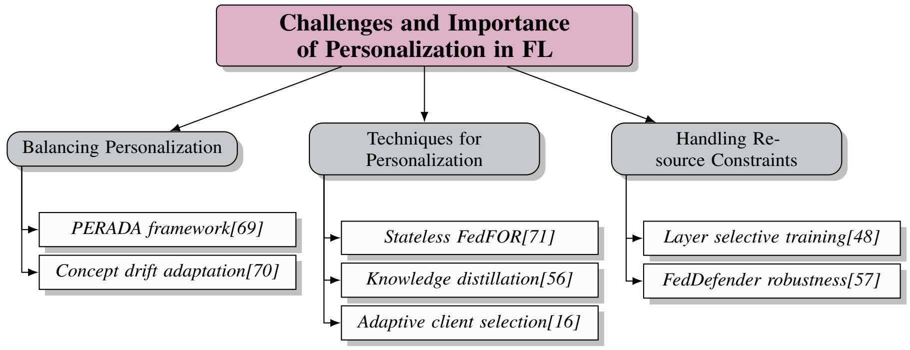

Figure 5: This figure illustrates the challenges and solutions related to personalization in federated learning, highlighting the importance of balancing personalization with generalization, techniques to enhance personalization, and methods to handle resource constraints.

Implementing model-agnostic meta-learning (MAML) within FL frameworks allows rapid adaptation of global models to local client data, enhancing personalization without sacrificing collaborative learning benefits [33]. This approach fosters the development of efficient and effective personalized models that can adapt to evolving client needs.

Integrating adaptive algorithms with advanced communication techniques, such as quantization and compression, reduces communication overhead and enhances FL system scalability [39]. These methods facilitate efficient model updates and improved personalization, even in resource-constrained settings.

5.3 Hierarchical and Multi-Cluster Approaches

Hierarchical and multi-cluster approaches in FL significantly enhance personalization by managing diverse data distributions and client computational capabilities. These strategies leverage natural client clustering based on data characteristics, allowing for tailored model training and improved efficiency [42]. Organizing clients into hierarchical structures or clusters optimizes resource allocation and reduces communication overhead, thereby enhancing model performance and scalability [39].

Hierarchical federated learning frameworks, such as Hier-Local-QSGD, utilize quantization and clustering techniques to tackle data heterogeneity and communication inefficiencies [39]. By grouping clients with similar data distributions, these frameworks enable more accurate model updates and faster convergence, even in diverse data environments [42]. This structure facilitates efficient aggregation of model updates at various levels, alleviating the communication burden on the central server and enhancing overall system performance.

Multi-cluster approaches further improve personalization by allowing the development of specialized models for each cluster, reflecting the unique data characteristics of clients within that cluster [42]. By concentrating on these localized models, FL systems achieve higher accuracy and relevance, ensuring models are well-suited to individual client needs. This approach enhances model performance while ensuring that personalization efforts remain scalable and adaptable.

5.4 Leveraging Local and Global Knowledge

Leveraging both local and global knowledge is a critical strategy in FL for enhancing personalization, ensuring models are accurate and adaptable to individual client needs. Integrating local and global knowledge facilitates model development that is finely tuned to unique client data distributions while benefiting from insights gained from aggregated global data [37].

One effective strategy is employing transfer learning and meta-learning techniques, allowing models to adapt quickly to new tasks or data distributions by utilizing pre-learned knowledge from global models. This ensures efficient incorporation of new information while maintaining prior learning benefits, thus enhancing personalization [37].

Additionally, multi-task learning frameworks enable simultaneous training of multiple related tasks, allowing models to leverage shared representations to improve performance and personalization [37]. Focusing on commonalities between tasks enhances the model's ability to generalize and adapt to diverse client data, ensuring personalized models are robust and effective.

Knowledge distillation is another vital technique that transfers knowledge from a global model to local models, facilitating the development of lightweight and efficient personalized models. This process distills knowledge from a complex global model into simpler local models, which can be fine-tuned to meet individual client needs [37]. Consequently, FL systems achieve high levels of personalization without compromising model performance or efficiency.

Finally, the mixture of local and global models offers a balanced approach to personalization, combining local adaptations with global insights. By integrating local model updates with global model parameters, FL systems develop personalized models that are well-suited to the unique data landscapes of individual clients while benefiting from the broader context provided by global models [37].

6 Distributed Computing and Model Optimization

6.1 Model Optimization and Resource Management

Method Name	Optimization Techniques	Communication Efficiency	Data Heterogeneity Management
CE-FedAvg[14]	Decentralized Approach	Gossip Protocol	Computational Heterogeneity
EAFO[17]	Adaptive Balancing Coefficients	Parameter Compression	Local Update Coefficients
SFA[6]	Hybrid Pruning Technique	Reduced Communication Costs	Personalized Subnetworks

Table 1: Comparison of Federated Learning Optimization Techniques: This table outlines various methods employed to enhance federated learning systems by detailing their optimization techniques, communication efficiency strategies, and approaches for managing data heterogeneity. The methods include CE-FedAvg, EAFO, and SFA, each offering unique solutions to address the challenges of decentralized model training.

Efficient model optimization and resource management are critical for federated learning (FL) systems operating under diverse computational and network constraints. Effective strategies enhance scalability and performance across client devices, especially in wireless environments where communication efficiency is crucial. Challenges such as data distribution heterogeneity, varying device capabilities, and participant availability can hinder model convergence and introduce bias. Techniques like intelligent participant selection, adaptive subnetwork scheduling, and overlapping local computation with communication mitigate these issues, reducing training latency and resource consumption while maintaining model accuracy. By optimizing resource use at edge nodes and minimizing communication costs, these methods enable FL to function effectively even on mobile devices and unreliable networks [74, 75, 76, 77, 48]. Table 1 provides a comprehensive comparison of different federated learning methods, highlighting their optimization techniques, communication efficiency, and strategies for managing data heterogeneity, which are critical for improving model performance and resource management.

The CE-FedAvg framework exemplifies collaborative model training among devices within each edge server's coverage, facilitating inter-server cooperation to learn a shared global model, thus minimizing communication overhead and enhancing model convergence [14]. Efficient Adaptive Federated Optimization (EAFO) enhances performance by optimizing local updates and parameter compression, dynamically balancing communication, computation, and precision [17]. Sub-FedAvg tackles data heterogeneity by combining pruning techniques to create personalized subnetworks, improving model performance tailored to individual data characteristics [6]. Resource-efficient frameworks optimize computational resource use while ensuring high model quality through intelligent participant selection, dynamic server updates, and adaptive model pruning, enhancing training efficiency and reducing communication costs [78, 74, 79, 31, 48].

6.2 Integration with Edge and Cloud Resources

Integrating federated learning with edge and cloud resources advances distributed learning by enabling bi-directional knowledge transfer between edge devices and cloud servers. This integration enhances personalization and model heterogeneity while addressing computational resource constraints and network bottlenecks. By allowing cloud servers to leverage historical and interactive features while edge nodes perform local updates, scalability and performance improve, reducing resource utilization and communication costs, thus facilitating efficient global model convergence [48, 46]. The FedASM framework, validated through multiple models and datasets, exemplifies efficient asynchronous updates that reduce communication overhead and enable real-time processing on edge devices [80].

Centralized federated learning frameworks allow more devices to participate in training without overwhelming the central server, facilitating seamless edge-cloud integration [81]. Bonawitz et al. demonstrate scalability, accommodating approximately 10 million devices [82]. Frameworks like FedCL enhance initial performance through continual local training, minimizing communication overhead [83]. By efficiently utilizing edge and cloud resources, these systems improve the adaptability and responsiveness of FL, making them suitable for real-world applications where efficiency and performance are critical.

6.3 Enhancing Data Privacy and Processing Efficiency

Enhancing data privacy and processing efficiency in federated learning (FL) is vital for deployment in environments marked by data heterogeneity and resource constraints. Decentralized architectures, as implemented in frameworks like Flotta, ensure secure instruction execution across federated data spaces, maintaining data integrity and preventing unauthorized access [25]. Federated Feature Maps (FFMs) enhance privacy and reduce communication overhead by effectively utilizing decentralized data, facilitating efficient data handling and model training [84]. The ASO-Fed framework showcases the benefits of asynchronous updates, improving computational efficiency and reducing model aggregation wait times [85].

FeDeRA demonstrates the integration of low-rank adaptations to reduce communication and computation costs, employing low-rank adaptations of weight matrices for efficient resource use, especially in bandwidth-limited environments [66]. Adaptive Federated Dropout (AFD) minimizes communication burdens by exchanging only necessary model parameters, maintaining model quality during training [50]. FedPAQ addresses communication and scalability challenges, optimizing strategies to enhance both privacy and processing efficiency [5].

Iterative pruning methods retain personalized parameters, optimizing resource allocation and ensuring efficient training across diverse environments [6]. Despite advancements, challenges in addressing data space heterogeneity persist. Future research should explore implications within intricate network architectures and investigate regularization techniques to enhance privacy, critical for widespread adoption. By examining these aspects, researchers can mitigate risks like communication bottlenecks and potential attacks, fostering a secure environment for FL applications, especially in sensitive fields like digital health where privacy is paramount [8, 10]. Innovative methods and optimized communication strategies can achieve robust privacy protections and efficient processing, ensuring applicability in diverse real-world scenarios.

7 Edge AI and Federated Learning

The integration of edge AI with federated learning (FL) represents a significant advancement in distributed systems, enhancing processing efficiency and addressing challenges related to latency and real-time data handling. This section investigates the role of FL in improving data access for model fine-tuning under resource constraints at the network edge, emphasizing real-time processing and latency reduction. Time-sensitive FL frameworks are analyzed to optimize model training efficiency and reduce runtime, crucial for applications like autonomous vehicles and industrial automation [86, 87].

7.1 Real-Time Processing and Latency Reduction

The synergy between edge AI and federated learning significantly enhances real-time processing and reduces latency, essential for applications requiring immediate responses, such as autonomous vehicles and smart cities. FL facilitates local data processing on edge devices, minimizing data transmission to central servers and thus reducing latency [44]. This capability is critical for timely decision-making [45].

Frameworks like FedVision exemplify this integration by enabling online visual object detection and safety monitoring in smart city projects, leveraging the reduced latency and enhanced processing capabilities of edge devices for immediate insights and actions [44]. The E-Tree Learning framework demonstrates how hierarchical learning structures can optimize model accuracy and convergence, enhancing real-time processing without extensive data transfer [45].

The Edge-Cloud Collaborative Training (ECCT) framework further illustrates the effective management of data and computational demands by combining edge and cloud resources. It conducts privacy-sensitive feature extraction at the edge while utilizing cloud resources for complex model training, maintaining low latency [46].

7.2 Challenges in Resource-Constrained Environments

Federated learning in resource-constrained environments faces significant challenges due to limited computational power, memory, and network bandwidth of edge devices. These constraints impede the execution of complex machine learning tasks, necessitating adaptive strategies for optimizing resource utilization while maintaining model performance. Simulations of various edge server configurations highlight these challenges [86].

Balancing computation and communication is a primary challenge, as frequent model updates can overwhelm limited network capacities, increasing latency. This issue is particularly pronounced in heterogeneous environments, where disparities in device capabilities can lead to uneven contributions to the global model, affecting performance [86].

Efficient resource allocation is crucial to accommodate diverse client device capabilities. Strategies like selective layer training and adaptive pruning are vital for optimizing resource usage, enabling resource-limited devices to participate in federated learning without sacrificing model quality [86].

Additionally, the non-IID nature of client data complicates model training, potentially skewing updates and hindering convergence. Robust aggregation techniques and personalized modeling approaches are required to adapt to the statistical heterogeneity inherent in FL environments [86].

7.3 Hierarchical and Decentralized Architectures

The integration of edge AI with federated learning through hierarchical and decentralized architectures offers a promising strategy for enhancing scalability, efficiency, and model performance in distributed learning systems. Hierarchical architectures utilize multi-layered structures to facilitate efficient data processing and model training across diverse edge devices, optimizing resource allocation and reducing communication overhead [39]. By organizing clients into hierarchical clusters based on data characteristics and computational capabilities, these architectures enable accurate and efficient aggregation of model updates, improving convergence rates and overall system performance [42].

Decentralized architectures further enhance this integration by distributing the learning process across multiple cooperative edge servers, minimizing reliance on a central server and reducing communication delays [14]. This approach allows for independent local computations, enabling real-time data processing and decision-making while preserving data privacy and security [45]. Decentralized architectures are particularly beneficial in scenarios where latency and bandwidth constraints are significant, facilitating efficient model updates and improved scalability.

Frameworks like E-Tree Learning exemplify the implementation of these architectures by employing hierarchical learning structures to optimize model accuracy and convergence without extensive data transfer [45]. Additionally, frameworks like ECCT demonstrate the potential of combining edge and cloud resources to enhance real-time processing and latency reduction, enabling federated learning systems to efficiently manage data and computational demands while maintaining low latency [46].

8 Conclusion

8.1 Future Directions and Research Opportunities

The trajectory of federated learning (FL) is poised to address pivotal challenges that will enhance its adaptability and efficacy across diverse applications. A crucial area for future exploration is the refinement of adaptive mechanisms within frameworks like EAFO, which aim to optimize the balance between communication frequency and model accuracy in complex IoT environments. Extending methodologies such as FedPAQ to more sophisticated FL contexts remains a priority.

Addressing computational heterogeneity and enhancing privacy measures within frameworks like CFEL are essential for future FL advancements. Research should focus on developing lightweight, privacy-preserving algorithms and fortifying defenses against malicious activities, with blockchain technology as a potential tool for secure model aggregation. Another promising direction involves integrating prior knowledge into the AFL framework to create personalized models that leverage both centralized and client-specific data.

Advancements in FL algorithms through platforms like PTB-FLA, particularly for operations within local area networks and edge systems, are critical. Additionally, enhancing client selection strategies and pioneering model fusion techniques for effective management of multi-modal data are necessary steps forward.

Further research will aim to optimize pruning strategies, bolster robustness against diverse data distributions, and assess applicability in various FL scenarios. Adapting the OLAR framework for real-world FL platforms, with a focus on energy consumption optimization and integration with other algorithms to improve execution efficiency, presents promising avenues for future inquiry.

AI-generated, for reference only.

References

[1] Baris Askin, Pranay Sharma, Carlee Joe-Wong, and Gauri Joshi. Fedast: Federated asynchronous simultaneous training, 2024.

[2] Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE transactions on neural networks and learning systems*, 34(12):9587–9603, 2022.

[3] Bin Wang, Jun Fang, Hongbin Li, Xiaojun Yuan, and Qing Ling. Confederated learning: Federated learning with decentralized edge servers, 2022.

[4] Latif U Khan, Walid Saad, Zhu Han, Ekram Hossain, and Choong Seon Hong. Federated learning for internet of things: Recent advances, taxonomy, and open challenges. *IEEE Communications Surveys & Tutorials*, 23(3):1759–1799, 2021.

[5] Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization, 2020.

[6] Saeed Vahidian, Mahdi Morafah, and Bill Lin. Personalized federated learning by structured and unstructured pruning under data heterogeneity, 2021.

[7] Margarita Kirienko, Martina Sollini, Gaia Ninatti, Daniele Loiacono, Edoardo Giacomello, Noemi Gozzi, Francesco Amigoni, Luca Mainardi, Pier Luca Lanzi, and Arturo Chiti. Distributed learning: a reliable privacy-preserving strategy to change multicenter collaborations using ai. *European Journal of Nuclear Medicine and Molecular Imaging*, 48:3791–3804, 2021.

[8] Anirban Das and Thomas Brunschwiler. Privacy is what we care about: Experimental investigation of federated learning on edge devices, 2019.

[9] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications, 2019.

[10] Ghazaleh Shirvani, Saeid Ghasemshirazi, and Behzad Beigzadeh. Federated learning: Attacks, defenses, opportunities, and challenges, 2024.

[11] Aditya Pribadi Kalapaaking, Ibrahim Khalil, and Mohammed Atiquzzaman. Smart policy control for securing federated learning management system, 2023.

[12] Tuo Zhang, Lei Gao, Chaoyang He, Mi Zhang, Bhaskar Krishnamachari, and Salman Avestimehr. Federated learning for internet of things: Applications, challenges, and opportunities, 2022.

[13] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. *arXiv preprint arXiv:2002.06440*, 2020.

[14] Zhenxiao Zhang, Zhidong Gao, Yuanxiong Guo, and Yanmin Gong. Scalable and low-latency federated learning with cooperative mobile edge networking, 2022.

[15] Jie Wen, Zhixia Zhang, Yang Lan, Zhihua Cui, Jianghui Cai, and Wensheng Zhang. A survey on federated learning: challenges and applications. *International Journal of Machine Learning and Cybernetics*, 14(2):513–535, 2023.

[16] Isidoros Tziotis, Zebang Shen, Ramtin Pedarsani, Hamed Hassani, and Aryan Mokhtari. Straggler-resilient personalized federated learning, 2022.

[17] Zunming Chen, Hongyan Cui, Ensen Wu, and Yu Xi. Efficient adaptive federated optimization of federated learning for iot, 2022.

[18] Ji Liu, Jizhou Huang, Yang Zhou, Xuhong Li, Shilei Ji, Haoyi Xiong, and Dejing Dou. From distributed machine learning to federated learning: A survey, 2022.

[19] Priyanka Mary Mammen. Federated learning: Opportunities and challenges. *arXiv preprint arXiv:2101.05428*, 2021.

[20] Ibrahim Abdul Majeed, Sagar Kaushik, Aniruddha Bardhan, Venkata Siva Kumar Tadi, Hwang-Ki Min, Karthikeyan Kumaraguru, and Rajasekhara Duvvuru Muni. Comparative assessment of federated and centralized machine learning, 2022.

[21] Bingyan Liu, Nuoyan Lv, Yuanchun Guo, and Yawen Li. Recent advances on federated learning: A systematic survey, 2023.

[22] Taki Hasan Rafi, Faiza Anan Noor, Tahmid Hussain, Dong-Kyu Chae, and Zhaohui Yang. A generalized look at federated learning: Survey and perspectives, 2023.

[23] 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.

[24] Langming Liu and Dingxuan Zhou. Analysis of regularized federated learning, 2024.

[25] Claudio Bonesana, Daniele Malpetti, Sandra Mitrović, Francesca Mangili, and Laura Azzimonti. Flotta: a secure and flexible spark-inspired federated learning framework, 2024.

[26] Yeshwanth Venkatesha, Youngeun Kim, Hyoungseob Park, Yuhang Li, and Priyadarshini Panda. Addressing client drift in federated continual learning with adaptive optimization, 2022.

[27] Jiaming Yan, Jianchun Liu, Shilong Wang, Hongli Xu, Haifeng Liu, and Jianhua Zhou. Heroes: Lightweight federated learning with neural composition and adaptive local update in heterogeneous edge networks, 2023.

[28] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning, 2019.

[29] Hong Zhang, Ji Liu, Juncheng Jia, Yang Zhou, Huaiyu Dai, and Dejing Dou. Fedduap: Federated learning with dynamic update and adaptive pruning using shared data on the server, 2022.

[30] Tatjana Legler, Vinit Hegiste, Ahmed Anwar, and Martin Ruskowski. Addressing heterogeneity in federated learning: Challenges and solutions for a shared production environment, 2024.

[31] Ji Liu, Juncheng Jia, Hong Zhang, Yuhui Yun, Leye Wang, Yang Zhou, Huaiyu Dai, and Dejing Dou. Efficient federated learning using dynamic update and adaptive pruning with momentum on shared server data, 2024.

[32] Siqi Ping, Yuzhu Mao, Yang Liu, Xiao-Ping Zhang, and Wenbo Ding. Fl-tac: Enhanced fine-tuning in federated learning via low-rank, task-specific adapter clustering, 2024.

[33] Feibo Jiang, Li Dong, Siwei Tu, Yubo Peng, Kezhi Wang, Kun Yang, Cunhua Pan, and Dusit Niyato. Personalized wireless federated learning for large language models, 2024.

[34] A. Jung, S. Abdurakhmanova, O. Kuznetsova, and Y. SarcheshmehPour. Towards model-agnostic federated learning over networks, 2023.

[35] Jiahao Liu, Jiang Wu, Jinyu Chen, Miao Hu, Yipeng Zhou, and Di Wu. Feddwa: Personalized federated learning with dynamic weight adjustment, 2023.

[36] Sotirios Nikoloutsopoulos, Iordanis Koutsopoulos, and Michalis K. Titsias. Personalized federated learning with exact stochastic gradient descent, 2022.

[37] 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. IEEE, 2020.

[38] Ji Liu, Jizhou Huang, Yang Zhou, Xuhong Li, Shilei Ji, Haoyi Xiong, and Dejing Dou. From distributed machine learning to federated learning: A survey. *Knowledge and Information Systems*, 64(4):885–917, 2022.

[39] Lumin Liu, Jun Zhang, Shenghui Song, and Khaled B. Letaief. Hierarchical federated learning with quantization: Convergence analysis and system design, 2023.

[40] Yongzhe Jia, Xuyun Zhang, Amin Beheshti, and Wanchun Dou. Fedlps: Heterogeneous federated learning for multiple tasks with local parameter sharing, 2024.

[41] Ahmet M. Elbir, Sinem Coleri, and Kumar Vijay Mishra. Hybrid federated and centralized learning, 2021.

[42] Zhifeng Jiang, Wei Wang, Bo Li, and Qiang Yang. Towards efficient synchronous federated training: A survey on system optimization strategies, 2022.

[43] Boyu Fan, Siyang Jiang, Xiang Su, Sasu Tarkoma, and Pan Hui. A survey on model-heterogeneous federated learning: Problems, methods, and prospects, 2024.

[44] Yang Liu, Anbu Huang, Yun Luo, He Huang, Youzhi Liu, Yuanyuan Chen, Lican Feng, Tianjian Chen, Han Yu, and Qiang Yang. Fedvision: An online visual object detection platform powered by federated learning, 2020.

[45] Lei Yang, Yanyan Lu, Jiannong Cao, Jiaming Huang, and Mingjin Zhang. E-tree learning: A novel decentralized model learning framework for edge ai, 2021.

[46] Zexi Li, Qunwei Li, Yi Zhou, Wenliang Zhong, Guannan Zhang, and Chao Wu. Edge-cloud collaborative learning with federated and centralized features, 2023.

[47] Miroslav Popovic, Marko Popovic, Ivan Kastelan, Miodrag Djukic, and Silvia Ghilezan. A simple python testbed for federated learning algorithms, 2023.

[48] Sadi Alawadi, Addi Ait-Mlouk, Salman Toor, and Andreas Hellander. Toward efficient resource utilization at edge nodes in federated learning, 2024.

[49] Laércio Lima Pilla. Optimal task assignment to heterogeneous federated learning devices, 2020.

[50] Nader Bouacida, Jiahui Hou, Hui Zang, and Xin Liu. Adaptive federated dropout: Improving communication efficiency and generalization for federated learning, 2020.

[51] Peng Yang, Yuning Jiang, Ting Wang, Yong Zhou, Yuanming Shi, and Colin N. Jones. Over-the-air federated learning via second-order optimization, 2022.

[52] Nam Hyeon-Woo, Moon Ye-Bin, and Tae-Hyun Oh. Fedpara: Low-rank hadamard product for communication-efficient federated learning, 2023.

[53] Yizhuo Cai, Bo Lei, Qianying Zhao, Jing Peng, Min Wei, Yushun Zhang, and Xing Zhang. Communication efficiency optimization of federated learning for computing and network convergence of 6g networks, 2023.

[54] Hang Liu, Xiaojun Yuan, and Ying-Jun Angela Zhang. Reconfigurable intelligent surface enabled federated learning: A unified communication-learning design approach, 2021.

[55] Muhammad Ammad ud din, Elena Ivannikova, Suleiman A. Khan, Were Oyomno, Qiang Fu, Kuan Eeik Tan, and Adrian Flanagan. Federated collaborative filtering for privacy-preserving personalized recommendation system, 2019.

[56] Hussain Ahmad Madni, Rao Muhammad Umer, and Gian Luca Foresti. Federated learning for data and model heterogeneity in medical imaging, 2023.

[57] Sungwon Park, Sungwon Han, Fangzhao Wu, Sundong Kim, Bin Zhu, Xing Xie, and Meeyoung Cha. Feddefender: Client-side attack-tolerant federated learning, 2023.

[58] Behnaz Soltani, Yipeng Zhou, Venus Haghighi, and John C. S. Lui. A survey of federated evaluation in federated learning, 2023.

[59] Joost Verbraeken, Martijn de Vos, and Johan Pouwelse. Bristle: Decentralized federated learning in byzantine, non-i.i.d. environments, 2021.

[60] Michael Ben Ali, Omar El-Rifai, Imen Megdiche, André Peninou, and Olivier Teste. Comparative evaluation of clustered federated learning methods, 2024.

[61] Daniel M. Jimenez G., David Solans, Mikko Heikkila, Andrea Vitaletti, Nicolas Kourtellis, Aris Anagnostopoulos, and Ioannis Chatzigiannakis. Non-iid data in federated learning: A survey with taxonomy, metrics, methods, frameworks and future directions, 2024.

[62] Ha Min Son, Moon Hyun Kim, and Tai-Myoung Chung. Compare where it matters: Using layer-wise regularization to improve federated learning on heterogeneous data, 2021.

[63] Basem Suleiman, Muhammad Johan Alibasa, Rizka Widyarini Purwanto, Lewis Jeffries, Ali Anaissi, and Jacky Song. Optimisation of federated learning settings under statistical heterogeneity variations, 2024.

[64] Bart Cox, Lydia Y. Chen, and Jérémie Decouchant. Aergia: Leveraging heterogeneity in federated learning systems, 2022.

[65] Aritra Mitra, Rayana Jaafar, George J. Pappas, and Hamed Hassani. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients, 2021.

[66] Yuxuan Yan, Qianqian Yang, Shunpu Tang, and Zhiguo Shi. Federa:efficient fine-tuning of language models in federated learning leveraging weight decomposition, 2024.

[67] Pengxin Guo, Shuang Zeng, Yanran Wang, Huijie Fan, Feifei Wang, and Liangqiong Qu. Selective aggregation for low-rank adaptation in federated learning, 2024.

[68] Changxin Xu, Yuxin Qiao, Zhanxin Zhou, Fanghao Ni, and Jize Xiong. Enhancing convergence in federated learning: A contribution-aware asynchronous approach, 2024.

[69] Chulin Xie, De-An Huang, Wenda Chu, Daguang Xu, Chaowei Xiao, Bo Li, and Anima Anandkumar. Perada: Parameter-efficient federated learning personalization with generalization guarantees, 2024.

[70] Amir Hossein Estiri and Muthucumaru Maheswaran. Attentive federated learning for concept drift in distributed 5g edge networks, 2021.

[71] Junjiao Tian, James Seale Smith, and Zsolt Kira. Fedfor: Stateless heterogeneous federated learning with first-order regularization, 2022.

[72] Hongrui Shi, Valentin Radu, and Po Yang. Closing the gap between client and global model performance in heterogeneous federated learning, 2022.

[73] Dashan Gao, Xin Yao, and Qiang Yang. A survey on heterogeneous federated learning, 2022.

[74] Ahmed M. Abdelmoniem, Atal Narayan Sahu, Marco Canini, and Suhaib A. Fahmy. Resource-efficient federated learning, 2022.

[75] Jiaxiang Geng, Boyu Li, Xiaoqi Qin, Yixuan Li, Liang Li, Yanzhao Hou, and Miao Pan. Fedex: Expediting federated learning over heterogeneous mobile devices by overlapping and participant selection, 2024.

[76] Pavana Prakash, Jiahao Ding, Maoqiang Wu, Minglei Shu, Rong Yu, and Miao Pan. To talk or to work: Delay efficient federated learning over mobile edge devices, 2021.

[77] Huai an Su, Jiaxiang Geng, Liang Li, Xiaoqi Qin, Yanzhao Hou, Hao Wang, Xin Fu, and Miao Pan. Whale-fl: Wireless and heterogeneity aware latency efficient federated learning over mobile devices via adaptive subnetwork scheduling, 2024.

[78] Ruirui Zhang, Xingze Wu, Yifei Zou, Zhenzhen Xie, Peng Li, Xiuzhen Cheng, and Dongxiao Yu. A resource-adaptive approach for federated learning under resource-constrained environments, 2024.

[79] Mohamed Yassine Boukhari, Akash Dhasade, Anne-Marie Kermarrec, Rafael Pires, Othmane Safsafi, and Rishi Sharma. Boosting federated learning in resource-constrained networks, 2023.

[80] Ji Liu, Juncheng Jia, Tianshi Che, Chao Huo, Jiaxiang Ren, Yang Zhou, Huaiyu Dai, and Dejing Dou. Fedasmu: Efficient asynchronous federated learning with dynamic staleness-aware model update, 2023.

[81] Li Chou, Zichang Liu, Zhuang Wang, and Anshumali Shrivastava. Efficient and less centralized federated learning, 2021.

[82] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečný, Stefano Mazzocchi, H. Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roselander. Towards federated learning at scale: System design, 2019.

[83] Xin Yao and Lifeng Sun. Continual local training for better initialization of federated models, 2020.

[84] Sixing Yu, J Pablo Muñoz, and Ali Jannesari. Federated foundation models: Privacy-preserving and collaborative learning for large models. *arXiv preprint arXiv:2305.11414*, 2023.

[85] Yujing Chen, Yue Ning, Martin Slawski, and Huzefa Rangwala. Asynchronous online federated learning for edge devices with non-iid data, 2020.

[86] Yong Xiao, Xiaohan Zhang, Guangming Shi, Marwan Krunz, Diep N. Nguyen, and Dinh Thai Hoang. Time-sensitive learning for heterogeneous federated edge intelligence, 2023.

[87] Herbert Woisetschläger, Alexander Isenko, Shiqiang Wang, Ruben Mayer, and Hans-Arno Jacobsen. Federated fine-tuning of llms on the very edge: The good, the bad, the ugly, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.

AI-generated, for reference only.