University of Glasgow

School of Computing Science
Knowledge & Data Engineering Systems

Viva

Qianyu Long, PhD Candidate, University of Glasgow

November 9, 2024

Introduction

Introduction

School *of* Computing Science
Knowledge & Data Engineering Systems

Thesis Title: Collaborative Distributed Machine Learning: From Knowledge Reuse to Sparsification in Federated Learning

Supervised by *Dr.Christos Anagnostopoulos* and *Dr. Fani Deligianni*

Introduction

School *of* Computing Science
Knowledge & Data Engineering Systems

- **Key Publications:**
  - Knowledge Reuse in Edge Computing Environments, *Journal of Network and Computer Applications*
    Qianyu Long, Kostas Kolomvatsos, Christos Anagnostopoulos
  - Model Reuse in Distributed Computing: A Multitask Learning Approach based on Partial Learning Curves, *Transactions on Emerging Topics in Computing*
    Qianyu Long, Kostas Kolomvatsos, Christos Anagnostopoulos
  - FedDIP: Federated Learning with Extreme Dynamic Pruning and Incremental Regularization, *International Conference on Data Mining 2023*
    Qianyu Long, Christos Anagnostopoulos, Shameem Puthiya Parambath, Daning Bi
  - Decentralized Personalized Federated Learning based on a Conditional Sparse-to-Sparser Scheme, *Under review in Transactions on Neural Networks and Systems*
    Qianyu Long, Qiyuan Wang, Christos Anagnostopoulos, Daning Bi
  - FedPhD: Federated Pruning with Hierarchical Learning of Diffusion Models, *In preparation for submission to the International Conference on Data Engineering 2025*
    Qianyu Long, Christos Anagnostopoulos

Overview

School of Computing Science
Knowledge & Data Engineering Systems

Content

- Background on Distributed Machine Learning
- Efficient Distributed Learning with Direct Reuse
- Efficient Distributed Learning with Enhanced Reusability
- Efficient Centralized Federated Learning with Pruning
- Efficient Decentralized Federated Learning with Pruning
- Conclusion

Background on Distributed ML

Definition and Motivation[4]

School of Computing Science
Knowledge & Data Engineering Systems

Definition: Distributed Machine Learning (DML) enables training machine learning models across multiple devices, addressing scalability and privacy challenges.

Motivation:

- Scalability
- Efficiency
- Data Location Constraints

Categories:

- Data Parallelism
- Model Parallelism
- Hybrid

Applications on DML

Applications

- Autonomous Vehicles
- Smart Grid
- IoT-Enabled Healthcare
- Digital Twin

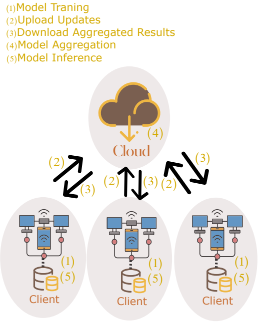


Figure 1: Example of Distributed ML (Server-Client)

Challenges in Distributed ML

School of Computing Science
Knowledge & Data Engineering Systems

- **System Heterogeneity**: Diverse hardware, network capacities, and computational power across devices.
- **Data Heterogeneity**: Non-uniform data distributions across nodes.
- **Communication Bottlenecks**: Limited bandwidth and latency issues in inter-node data exchange.
- **Data Privacy**: Maintaining data confidentiality across distributed training environments.
- **Computation Constraints**: Limited computational resources on edge devices.

School *of* Computing Science
Knowledge & Data Engineering Systems

Efficient Distributed Learning with Direct Reuse

Challenges in EC Environments

School of Computing Science
Knowledge & Data Engineering Systems

Edge Computing: Edge computing is a distributed computing framework where data processing occurs near the data source.

Motivation

- Limited computational resources and expensive communication.
- Data redundancy exists under certain situations (e.g., routine commuting, traffic cameras).

Figure 2: Home-to-School of Computing Science, University of Glasgow

Method: Knowledge Reuse

Target: Reduces costs while keeping model performance.

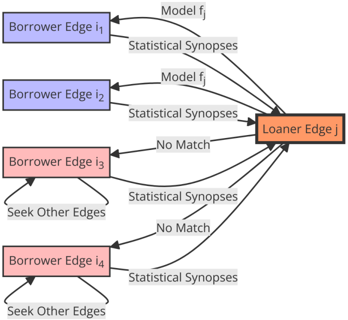


Figure 3: Mechanism of BLM

Solution: Reuse models from other nodes to avoid retraining.

- Borrower-Loaner-Match to decide **reusable** models.
- Model-Reusability-Monitoring to ensure model performance.

Formulation and Analysis

Math Form:

- $$\mathrm{MMD}(i,j) = \|\mu_i - \mu_j\|_{\mathcal{H}}$$
- $$\mathrm{CD}(i,j) = 1 - \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\|\|\mathbf{v}_j\|}$$
- $$S_t = \alpha Y_t + (1-\alpha)(S_{t-1} + b_{t-1}), \quad b_t = \beta(S_t - S_{t-1}) + (1-\beta) b_{t-1}$$

- **Reusability:** Over 67% success in reusability thresholds, with minimal false positives.
- **Similarity Metrics:** MMD and CD effectively guide borrower-loaner matches.
- **Accuracy:** High predictive accuracy maintained. (With only around 2% drop in accuracy)
- **Monitoring:** Holt-Winters detects data drift, ensuring model relevancy.

Efficient Distributed Learning with Enhanced Reusability

Reusability in Distributed Learning

School *of* Computing Science
Knowledge & Data Engineering Systems

Motivation

- Distributed systems (IoT, edge devices) generate redundant data. Hence, training separate models for each task is cost-inefficient.
- Pre-existing reusable models might be unavailable.

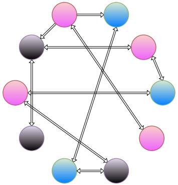


Figure 4: **Data Distribution Across Nodes**: For nodes with similar data distributions (same color), models trained on one node are reusable for some of others.

Method: PLC-based DMtL Framework

School of Computing Science
Knowledge & Data Engineering Systems

Target: Minimize resource consumption by optimizing model reuse in distributed training.

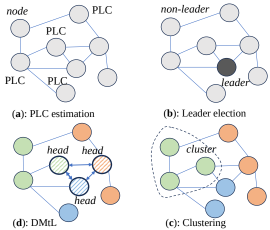


Figure 5: Two-Phase DMtL Process based on Partial Learning Curves

Method: Efficient knowledge reuse across tasks through model sharing.

- Partial Learning Curves computation with bootstrapping method, to estimate task similarity.
- PLC-based clustering and leader election.
- Distributed Multitask Learning across Leaders based on similarity information.

Problem Formulation and Analysis

School *of* Computing Science
Knowledge & Data Engineering Systems

Problem Formulation

- **Partial Learning Curves (PLC):** $V_i = \left[V_i^{(S_1)}, V_i^{(S_2)}, \ldots, V_i^{(S_p)}\right]^T$
- **Task Relationship Matrix:** $\Omega_{i,j}^{-1} = \frac{2}{m} \cdot \frac{1}{1+\exp(\epsilon \cdot d_{i,j})}$
- **Optimization Objective:** $J(W) = \sum_{k=1}^{K} \sum_{t=1}^{n_k} L_k(w_k^T x_k^t, y_k^t) + \frac{\lambda_1}{2} \mathrm{tr}(W\Omega^{-1}W^T) + \frac{\lambda_2}{2} \|W\|_F^2$

Key Results

- More than 80% communication computation reduction via clustering and head selection.
- **Sørensen-Dice Coefficient:** $\mu_{DC} > 0.9$ for efficient clustering.
- Minimal loss with reused models ($\xi \approx 0.05$).
- Improved Model performance, compared with SOTA baselines with 0.8% to 2% across CIFAR10 and Sentiment Datasets.

Efficient Centralized Federated Learning with Pruning

Definition and Motivation

School of Computing Science
Knowledge & Data Engineering Systems

Definition: Federated Learning (FL) is a decentralized machine learning approach that enables training across multiple client devices without sharing raw data.[3]

Motivation:

- Resource constraints on edge devices.
- Communication bottleneck on the central server.
- Suboptimal sparsity in SOTA methods

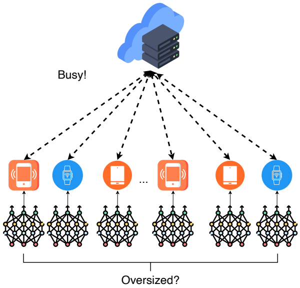


Figure 6: Illustration of Challenges

FedDIP Framework Overview

- Introduces dynamic pruning with error feedback into FL: (**DPF**[2]):

$$\omega_{t+1} = \omega_t - \eta_t \nabla f(\omega_t \odot \mathbf{m}_t) \qquad (1)$$
$$= \omega_t - \eta_t \nabla f(\omega_t + \mathbf{e}_t) \qquad (2)$$

- Inspired by **GReg**[5], we combine incremental regularization to achieve extreme sparsity.

$$\lambda_t = \begin{cases} 0 & \text{if } 0 \le t < \frac{T}{Q} \\ \vdots & \vdots \\ \frac{\lambda_{\max}(Q-1)}{Q} & \text{if } \frac{(Q-1)T}{Q} \le t \le T \end{cases} \qquad (3)$$

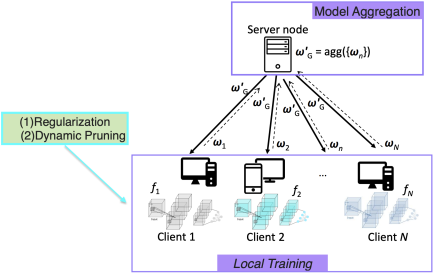
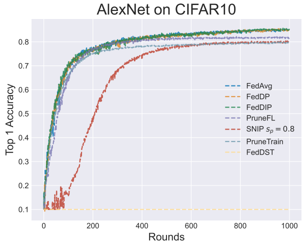


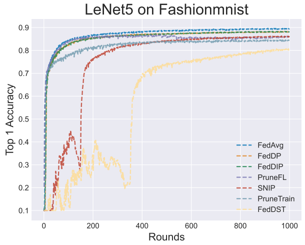
Figure 7: Illustration of FedDIP

School of Computing Science
Knowledge & Data Engineering Systems

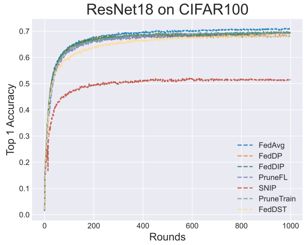
Analysis



(a) AlexNet on CIFAR10: Top-1 Accuracy over 1000 rounds for various FL methods.



(b) LeNet5 on FashionMNIST: Top-1 Accuracy over 1000 rounds for different FL methods.



(c) ResNet18 on CIFAR100: Top-1 Accuracy over 1000 rounds for multiple FL methods.

Figure 8: Top-1 accuracy comparison of AlexNet, LeNet5, and ResNet18 across 1000 communication rounds using various federated learning methods.

Analysis

Summary of Contributions

- Accelerated training times
- Reduced memory usage
- Lower download costs

Detailed Results

- Enables extreme sparsity pruning while preserving accuracy: FedDIP achieved up to **90%** sparsity with only **1.25%** accuracy loss.
- Demonstrates efficiency across various model architectures in experiments with Fashion-MNIST, CIFAR10, and CIFAR100 datasets.
- Offers theoretical convergence guarantees for FedDIP.

Efficient Decentralized Federated Learning with Pruning

Definition and Motivation

From Centralized Federated Learning to Decentralized Federated Learning

Definition: Decentralized Federated Learning (DFL) is a variation of Federated Learning where devices collaboratively train a model by communicating directly with each other, eliminating the need for a central server.

Motivation:

- Higher Communication
- Higher Computation
- Higher Maintenance
- Faster Convergence

Method

- Dynamic Aggregation: Clients reuse models within the same communication round, splitting neighbors into prior $N_k^{(a)}$ and posterior $N_k^{(b)}$ subsets. - Aggregated model for client k at time t:

$$\tilde{\omega}_k^t = \left(\sum_{j \in G_k^t} \omega_j^t + \omega_k^t \right) \odot m_k^t$$

- Sparsity-Driven Pruning: - Utilizes PQ Index (PQI)[1] for layer-wise compressibility assessment:

$$\mathrm{PQI}(\omega_{k,l}^t) = 1 - \left(\frac{1}{d_l^t} \right)^{\frac{1}{q} - \frac{1}{p}} \cdot \frac{\|\omega_{k,l}^t\|_p}{\|\omega_{k,l}^t\|_q}$$

- Pruning occurs based on the threshold δ_{pr} to control pruning frequency:

$$\frac{|\Delta_0^t - \Delta_0^{t-1}|}{|\Delta_0^1|} < \delta_{pr}$$

DA-DPFL Diagram

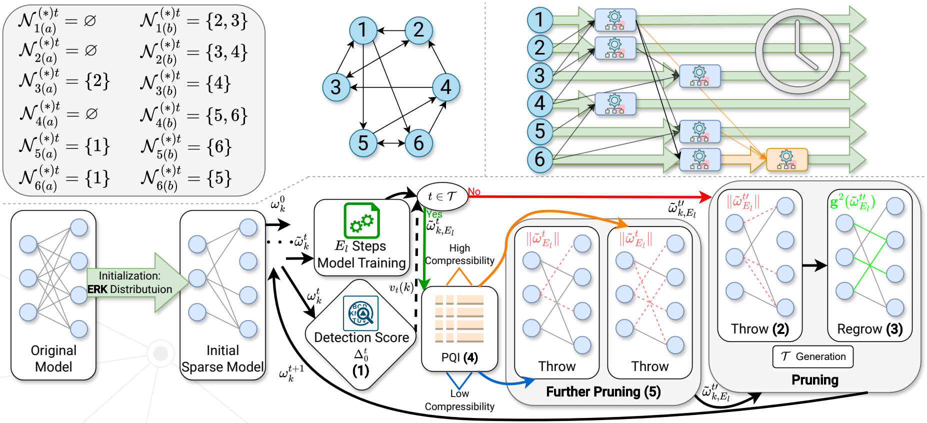


Figure 9: Iullstration of DA-DPFL

Analysis

School *of* Computing Science
Knowledge & Data Engineering Systems

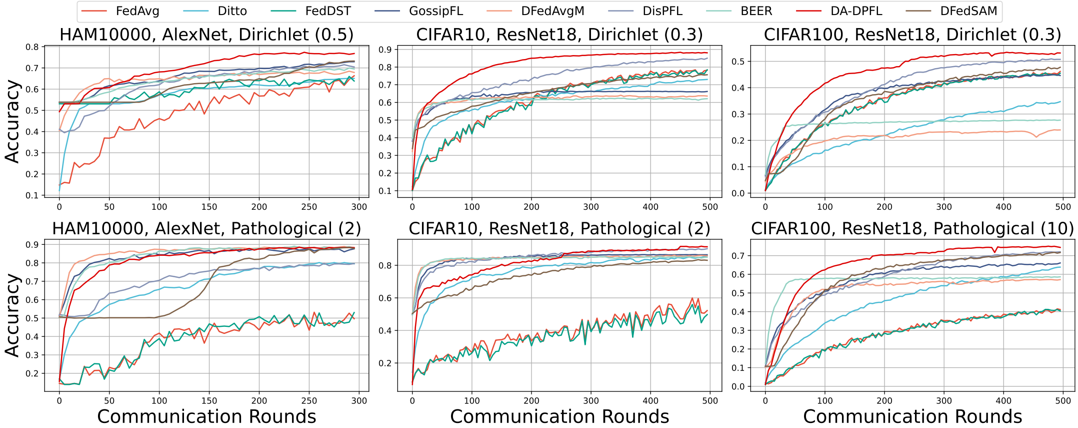


Figure 10: Performance comparison on multiple datasets and models.

Analysis

- **Model Accuracy**: DA-DPFL consistently outperforms baselines, achieving top-1 accuracy across datasets, with up to 3.2% higher accuracy on CIFAR10 (ResNet18), 2.6% on HAM10000 (AlexNet), and 2.4% on CIFAR100 (VGG11) under various data partitioning schemes.
- **Energy and Communication Efficiency**:
  - Reduces busiest communication cost by 5x, thanks to sparsity-driven pruning and dynamic aggregation.
  - Achieves high model sparsity (**up to 80%**) with minimal/no accuracy loss.
- **Convergence Efficiency**: Fewer communication rounds are needed to reach target accuracy, outperforming DisPFL and other baselines, due to adaptive pruning and dynamic client scheduling.

Conclusion & Research Contributions

Conclusion & Research Contributions

School *of* Computing Science
Knowledge & Data Engineering Systems

Core Advances:

- Developed efficient distributed learning frameworks for **knowledge reuse and sparsification**, addressing resource constraints in edge and federated systems.
- Introduced **pruning and dynamic aggregation** methods (FedDIP and DA-DPFL) to reduce communication costs and improve convergence with minimal accuracy loss.

Empirical Validation:

- Demonstrated efficiency across diverse datasets and architectures, **outperforming state-of-the-art** baselines in accuracy, communication, and computation.

Future Directions:

- Extending model reusability frameworks to more **heterogeneous environments**.
- Exploring further **sparsification methods** for lightweight, real-time federated systems in mobile settings.

Reference

References I

[1] Enmao Diao et al. “Pruning Deep Neural Networks from a Sparsity Perspective”. In: *The Eleventh International Conference on Learning Representations*. 2022.

[2] Tao Lin et al. “Dynamic Model Pruning with Feedback”. In: *International Conference on Learning Representations (ICLR)*. 2019.

[3] Brendan McMahan et al. “Communication-efficient learning of deep networks from decentralized data”. In: *Artificial intelligence and statistics*. PMLR. 2017, pp. 1273–1282.

[4] Joost Verbraeken et al. “A survey on distributed machine learning”. In: *Acm computing surveys (csur)* 53.2 (2020), pp. 1–33.

References II

School *of* Computing Science
Knowledge & Data Engineering Systems

[5] Huan Wang et al. "Neural Pruning via Growing Regularization". In: *International Conference on Learning Representations (ICLR)*. 2021.

Thanks