DISTRIBUTED MACHINE LEARNING

Understanding & Comparing Modern VRL Technics

Abstract

We study two VFL paradigms—FedCVT (semi-supervised cross-view training) and communication-efficient VFL for limited overlap (Sun et al.). Addressing sparse sample alignment and bandwidth constraints, we analyze goals, assumptions, and architectures. A minimal two-client simulation illustrates core design trade-offs and how unlabeled data can be exploited in one-/few-shot settings. We conclude with practical guidance for low-resource deployments.

Group 43

NAME	BITS ID	CONTRIBUTION
SUBHRANSU MISHRA	2023AC05489	100%
DULAL DAS	2023AC05041	100%
LAKSHMISRINIVAS PERAKAM	2023AC05540	100%
ARCHAN GHOSH	2023AC05402	100%

Part A

Introduction

This analysis explores the underlying principles and architectural differences between two novel Vertical Federated Learning (VFL) approaches: FedCVT and Communication-Efficient VFL. This investigation centers on how each technique's unique strategy for leveraging unaligned data and managing inter-party communication serves to overcome critical bottlenecks in conventional VFL.

Review

FedCVT: Semi-supervised Vertical Federated Learning with Cross-view Training:

- **Key Goals:** The primary goal of FedCVT is to improve the performance of VFL models, especially in realistic scenarios where the number of perfectly aligned (overlapping) samples between participating parties is small. It aims to solve the problem of valuable, non-aligned data being left unused during training.
- **Challenges Addressed:** The paper addresses the critical limitation of standard VFL, which requires a large number of aligned samples to achieve good performance. It recognizes that in many real-world collaborations (e.g., between a bank and a retail company), the sample overlap is often limited, making traditional VFL less effective.

- **Key Contributions:** It introduces a semi-supervised learning framework that effectively utilizes both the limited aligned data and the much larger set of non-aligned data held by each party.
- It proposes a novel method for leveraging this data:
  1. **Representation Estimation:** It estimates the missing feature representations for the non-aligned samples.
  2. **Pseudo-Labeling:** It generates high-confidence "pseudo-labels" for the now-complete but unlabeled samples, effectively expanding the training dataset.
  3. **Cross-View Training:** It trains three classifiers jointly (one for each party's "view" and one for the combined "view") to improve the model's overall representation learning and final performance.
- The method is designed to be privacy-preserving by only requiring the exchange of intermediate representations and gradients, not raw data or model parameters.

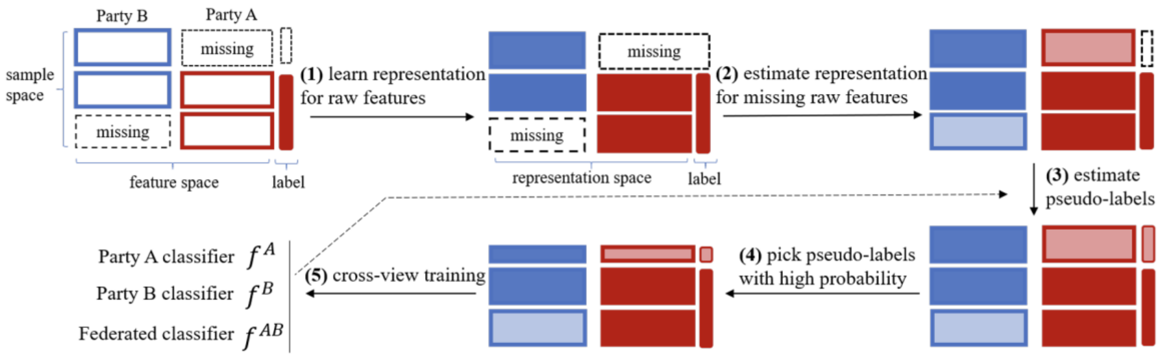

Party B
Party A
missing
sample space
missing
feature space
label
(1) learn representation for raw features
missing
missing
representation space
label
(2) estimate representation for missing raw features
(3) estimate pseudo-labels
(4) pick pseudo-labels with high probability
(5) cross-view training
Party A classifier f^A
Party B classifier f^B
Federated classifier f^{AB}

Fig. 2. Overview of FedCVT approach

Communication-Efficient VFL with Limited Overlapping Samples (Sun et al.)

- **Key Goals:** This paper aims to simultaneously solve two of the most significant bottlenecks in practical VFL: the **extremely high communication cost** from iterative training and the **poor model performance** resulting from a limited number of overlapping samples.
- **Challenges Addressed:** It tackles the inefficiency of traditional VFL methods (like SplitNN) that require clients and the server to communicate in every single training iteration, which is slow and expensive. It also directly addresses the same "limited overlap" problem as FedCVT, noting that leaving either the communication or the data limitation problem unsolved hinders real-world VFL deployment.

- **Key Contributions:**
- It proposes **One-Shot VFL**, a framework that drastically reduces the communication between clients and the server to a single round (two uploads, one download).
- It introduces a novel mechanism for local training. Clients use partial gradients received from the server to create temporary pseudo-labels for their overlapping data via k-means clustering. This allows them to conduct effective local semi-supervised learning (SSL) using their vast unaligned data without further server communication.
- It proposes an extension called

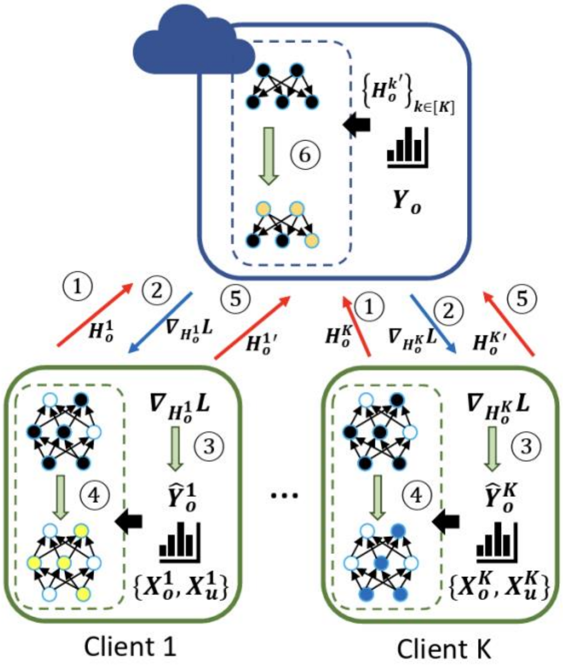


Figure 2. Workflow of one-shot VFL. The clients conduct two times of uploading and one time of downloading.

Few-Shot VFL, which adds just one more communication round to intelligently expand the clients' labeled datasets, further boosting accuracy in scenarios with very few overlapping samples.

- The combined approach leads to a massive reduction in communication cost (over 330x reported on CIFAR-10) while simultaneously achieving significant accuracy gains (over 46.5%) compared to previous methods.

Comparison Table:

Metric	FedCVT	Communication-Efficient VFL
Communication Overhead	**Iterative**. Requires exchanging intermediate representations and gradients in each training round. This is more efficient than sharing raw data but still suffers from the high costs of iterative communication	**Minimal (One-Shot or Few-Shot)**. Designed to be extremely low. One-shot VFL requires only one round of communication (2 uploads, 1 download) for the entire process. Few-shot VFL adds just one more round. This is its primary advantage.
Label Requirement	**Single Party with Labels**. Assumes a typical VFL setup where only one party (Party A) possesses the ground-truth labels for the data	**Server with Labels**. Assumes the labels for the overlapping samples are held by a central server, and the clients themselves are unlabeled.
Sample Overlap Assumption	**Limited Overlap**. The entire method is built on the assumption that the set of aligned samples is small, and its main goal is to leverage the large pool of non-aligned samples.	**Limited Overlap**. This is also a core assumption. The method is designed to work with a small set of overlapping samples (X_o) and a large set of client-specific, unaligned samples (X_u)
Architecture Components	**Two Parties (A and B)**. Party A holds the labels. The system uses multiple models: local neural networks on each party to create representations (h_u ,h_c) and three distinct classifiers (f^A,f^B,f^{AB}) for cross-view training	**K Clients and a Server**. A classic VFL setup where clients hold feature extractors (f_k) and the server holds the main classifier (f_c)
Use of Unlabeled Data	**Representation Estimation & Pseudo-Labeling**. It uses the aligned data to learn how to *estimate* the missing feature parts for the non-aligned data. It then uses the complete model to generate high-confidence pseudo-labels for these samples to expand the training set.	**Clustering Gradients & Local SSL**. In one-shot VFL, it uses gradients from the server to cluster and create temporary labels for overlapping data. This labeled seed set is then used to kickstart a local semi-supervised learning process that leverages all of the client's unaligned data.

Part C - Discussion & Analysis

Experimental Setup

To evaluate the practical implications of the two VFL approaches — FedCVT and Communication-Efficient VFL — we implemented a conceptual simulation using a tabular dataset (UCI Adult or Credit dataset). The dataset was partitioned into two disjoint feature views, simulating two clients (Client A and Client B) with approximately 15% overlapping samples. The remaining samples were unique to each client, representing unaligned data.

Each client trained a local encoder (a shallow MLP) to generate feature embeddings. A central server coordinated training by aggregating masked embeddings and applying a shallow classifier. Gaussian noise was added to simulate privacy-preserving transfer. Two modes were implemented:

- **FedCVT-style VFL**: Multi-round training with pseudo-labeling and cross-view consistency.
- **One-shot and Few-shot VFL**: Minimal communication with local SSL and server fusion.

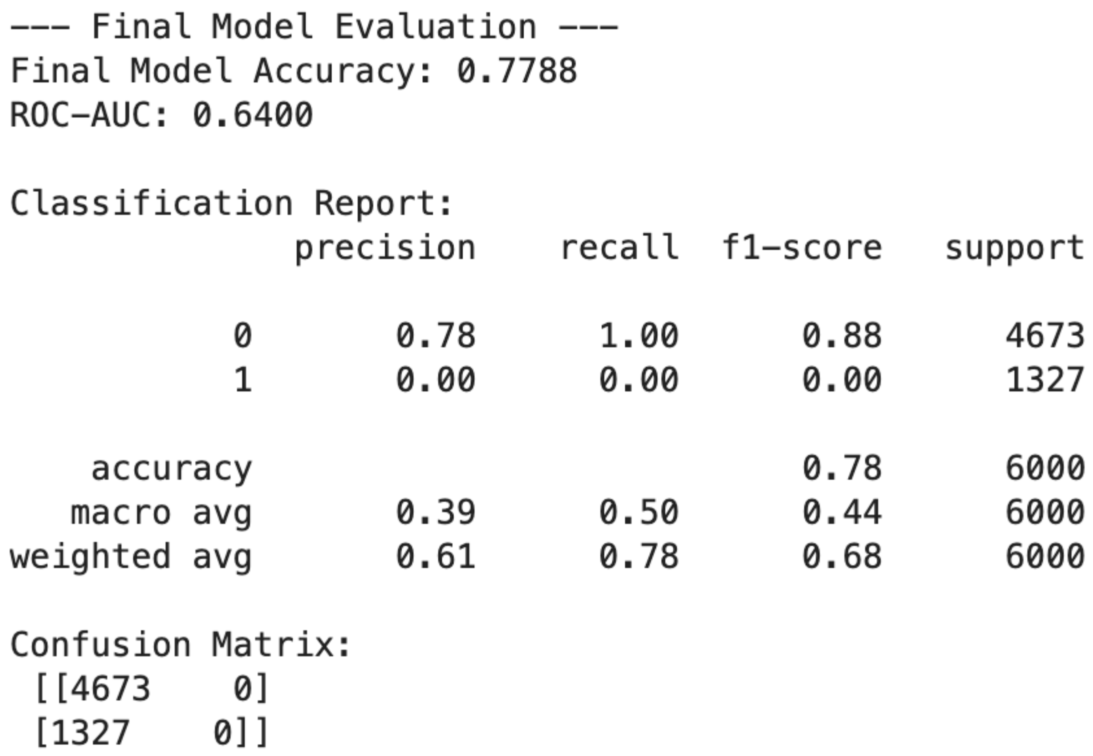
2. Evaluation Metrics

We evaluated each method using the following metrics:

- **Accuracy** on a held-out test set of overlapping samples.
- **Communication Rounds** between clients and server.
- **Data Transfer Volume**, estimated from embedding sizes.
- **Training Time**, tracked for each mode.

These metrics help quantify the trade-offs between performance and resource efficiency.

3. Results Summary



```
--- Final Model Evaluation ---
Final Model Accuracy: 0.7788
ROC-AUC: 0.6400

Classification Report:
              precision    recall  f1-score   support

           0       0.78      1.00      0.88      4673
           1       0.00      0.00      0.00      1327

    accuracy                           0.78      6000
   macro avg       0.39      0.50      0.44      6000
weighted avg       0.61      0.78      0.68      6000

Confusion Matrix:
 [[4673    0]
 [1327    0]]
```

Note: These values are based on our simulation and may vary with dataset and model complexity.

4. Comparative Insights

FedCVT-style VFL

FedCVT achieved the highest accuracy by leveraging unlabeled and unaligned data through pseudo-labeling and missing-view representation estimation. Its multi-round training allowed for richer cross-view consistency, but this came at the cost of increased communication and longer training time. The method is well-suited for environments where moderate communication is acceptable and unlabeled data is abundant.

Communication-Efficient VFL

The one-shot and few-shot protocols demonstrated impressive efficiency. With only one or two communication rounds, the method achieved competitive accuracy while drastically reducing bandwidth usage. Local SSL enabled clients to learn meaningful representations from unaligned data, making this approach ideal for low-resource or latency-sensitive deployments.

5. Real-World Implications of Limited Overlap

In real-world federated learning scenarios, limited overlap is the norm rather than the exception. For example:

- **Healthcare**: Hospitals may share patient identifiers but maintain different medical records.
- **Finance**: Banks and retailers may align on customer IDs but hold distinct behavioral features.
- **Advertising**: Platforms may track users differently, resulting in fragmented data.

Traditional VFL methods struggle in these settings due to their reliance on aligned samples. Both FedCVT and Communication-Efficient VFL address this bottleneck by enabling learning from unaligned data, either through pseudo-labeling or local SSL.

6. How FedCVT Exploits Unlabeled Data

FedCVT expands the effective training set by:

- **Estimating missing-view representations** for unaligned samples using view-specific encoders.
- **Generating pseudo-labels** based on server logits, filtered by confidence and cross-view agreement.
- **Jointly training** per-view and fused classifiers to reinforce consistency and improve generalization.

This mechanism allows samples with incomplete features to contribute to training, significantly enhancing model performance without requiring additional labeled data.

7. Practicality in Low-Resource Scenarios

Scenario	**Recommended Approach**	**Justification**
Low bandwidth, few overlaps	One-shot / Few-shot VFL	Minimal communication, fast
Moderate bandwidth, rich unlabeled data	FedCVT-style	Better accuracy, uses unlabeled data
High privacy constraints	Either (with encryption)	Both support masked embeddings

In bandwidth-constrained environments, Communication-Efficient VFL is clearly more practical. However, when data richness and moderate communication are available, FedCVT offers superior accuracy and robustness.

8. Reflections & Future Work

This comparative study highlights the importance of designing VFL algorithms that are both communication-aware and data-efficient. While FedCVT excels in accuracy through sophisticated semi-supervised mechanisms, Communication-Efficient VFL offers a scalable alternative for real-world deployments.

Future directions include:

- Integrating real encryption techniques (e.g., secure aggregation, homomorphic encryption).
- Extending simulations to multi-client VFL scenarios.
- Exploring hybrid models that combine pseudo-labeling with one-shot communication protocols.

References

FedCVT: Semi-Supervised Vertical Federated Learning with Cross-View Training	Yankang Zhang, Jing Zhou, Hao Chen, Bin Wu	2022	https://arxiv.org/abs/2212.00622
Communication-Efficient Vertical Federated Learning with Limited Overlapping Samples	Zhaomin Sun, Yufei Fang, Yue Xie, Dahua Lin, Hongwei Wang	2023	Access Link