Towards Privacy-Guaranteed Label Unlearning in Vertical Federated Learning: Few-Shot Forgetting Without Disclosure

Hanlin Gu[2*] Hong Xi Tae[1*] Lixin Fan[2] Chee Seng Chan[1]

Universiti Malaya[1] Webank AI Lab[2]

* *Equal* contribution

Paper Code Personal Page

Introduction

- **Vertical Federated Unlearning (VFU)** refers to the process of ***removing the influence of a passive party's data*** from a model trained under vertical federated learning.
- Previous works on VFU have focused on
  - *i. passive party unlearning,* and
  - *ii. lack a focus on privacy leakage during unlearning.*

Key Challenges

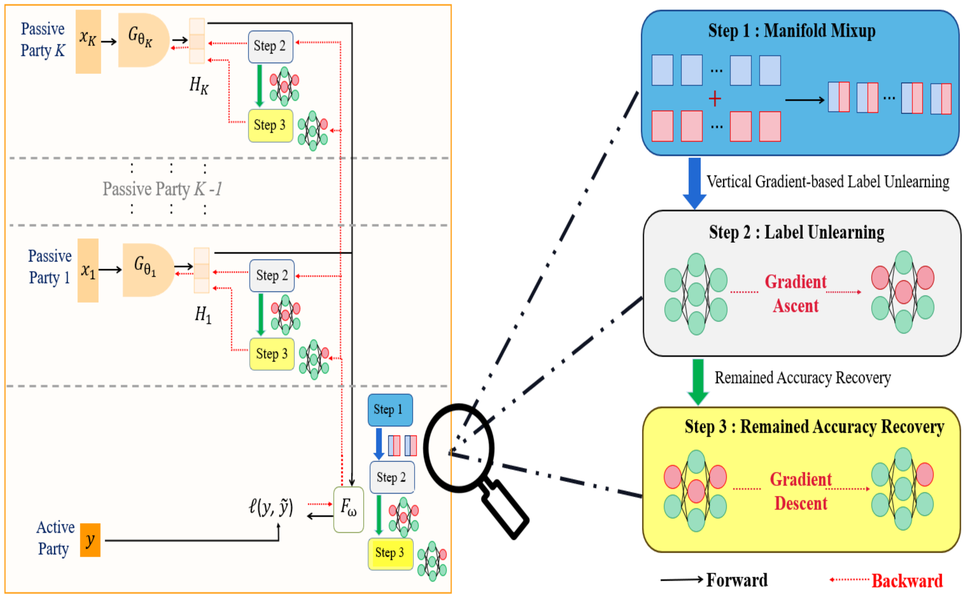
a) **Access to label gradients is required** to perform unlearning effectively.

b) The procedure **necessitates** active participation from **all involved parties**.

Contributions

1. To the best knowledge, this is the ***first work*** to address label unlearning in Vertical Federated Learning (VFL).
2. We propose a novel ***few-shot label unlearning*** method that removes labels from both active and passive models in VFL using only a limited amount of public data.
3. We conduct extensive experiments on diverse benchmark datasets, including
   - ***image-based:*** MNIST, CIFAR-10/100, ModelNet, Brain Tumor MRI, COVID19 Radiography, and
   - ***text-based:*** Yahoo Answers

Proposed Method



Our method consists of ***three*** key steps:

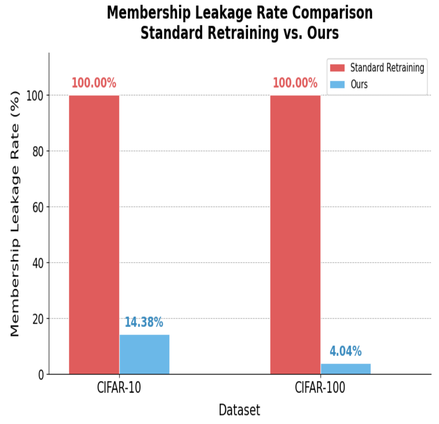
1. **Manifold mixup** to generate more synthetic samples from limited data.
2. **Gradient ascent** for efficient label unlearning.
3. **Gradient descent** to remain performance on retained data.

Conclusions

The proposed few-shot unlearning method leverages manifold mixup to reduce label privacy risks while enabling efficient unlearning for both active and passive models. Overall, it deepens the understanding of unlearning in vertical federated learning and offers a useful basis for future research in privacy-preserving collaborative machine learning.

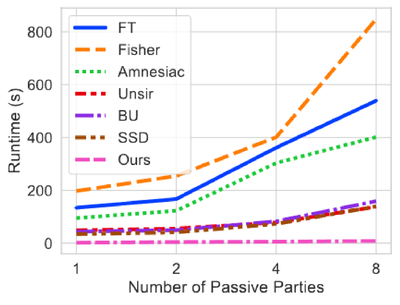
Experiment Results

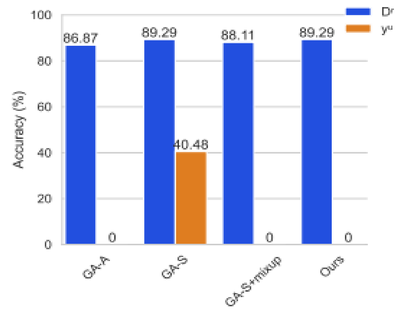
1. Standard retraining shows **100% membership leakage** on both datasets.
2. Our method reduces leakage to **14.38%** on CIFAR-10 and **4.04%** on CIFAR-100.
3. This indicates ***much stronger privacy protection*** against membership inference.



Selected Datasets	Metrics	Accuracy (%)		
		Baselines	Retrain	**Ours**
CIFAR10	D^r	90.61	91.26 ± 0.12	**89.29 ± 0.19**
	y^u	93.10	0.00 ± 0.00	**0.00 ± 0.00**
CIFAR100	D^r	71.43	71.03 ± 0.12	**69.96 ± 0.12**
	y^u	83.00	0.00 ± 0.00	**0.00 ± 0.00**

Our proposed method preserving the utility of ***retained data*** D^r, and effectively achieves strong unlearning performance in ***unlearn data*** y^u.





Our proposed method maintains ***high computational efficiency*** across varying participant counts *(left)* through a streamlined, module-optimized design *(right)*.