Personalized Label Inference Attack in Federated Transfer Learning via Contrastive Meta Learning

Hanyu Zhao[1], Zijie Pan[2], Yajie Wang[1]*, Zuobin Ying[2], Lei Xu[1], Yu-an Tan[1],

[1]Beijing Institute of Technology, Beijing, China
[2]City University of Macau, Macau, China
hanyuzhao@bit.edu.cn, wangyajie0312@foxmail.com

Abstract

Federated Transfer Learning (FTL) is a popular approach to solve the problem of heterogeneous feature space and label distribution. Among the mainstream strategies for FTL, parameter decoupling, which balance the impact of a single global model and multiple personalized models under data heterogeneity, has attracted the attention of many researchers. However, few attacks have been proposed to evaluate the privacy risk of FTL. We find that the fine-tuned structures and the gradient update mechanisms of parameter decoupling would be more likely to leak personalized information for the server to infer private labels. Based on our findings, we propose the label inference attack that combines meta classifier with contrastive learning in FTL. Our experiments show that the proposed attack has ability to extract local personalized information from the differences before and after fine-tuning to improve the accuracy of the attack in the absence of a downstream model. Our research can reveal potential privacy risks in FTL and motivate more research on private and secure FTL.

Introduction

With the growing popularity of big data and the increasing complexity of various scenarios, federated transfer learning gains significant attention (Liu et al. 2020; Kevin et al. 2021; Saha and Ahmad 2021). Through FTL, knowledge is transferred from the multiple source domain to the target domain in a continuous interaction as the central server aggregates and participants interact. This allows local models obtained from a specific domain to be used by other participants, thus alleviating limitations such as data heterogeneity, system heterogeneity, incremental data and labeled data scarcity. FTL mainly includes two mainstreams: data-based strategies and model-based strategies (Guo et al. 2024). Data-based strategies emphasize knowledge transfer by modulating and transforming participants' data to accommodate the preservation or adaptation of space, distribution, and data attributes without exposing any original private data (Zhuang et al. 2020). Model-based strategies aim to improve the accuracy of client's prediction by modeling other participants. Among the various strategies employed in FTL, parameter decoupling gains significant attention due to its ability to balance the influence of global model with multiple personalized models, particularly in scenarios involving data heterogeneity (Guo et al. 2024). This approach allows for the creation of more tailored models for individual clients while maintaining the benefits of shared learning (Yu et al. 2020; Kang et al. 2019; Yosinski et al. 2014). A large number of research has demonstrated the simplicity and efficacy of this methodology (Collins et al. 2021; Oh, Kim, and Yun 2021; Arivazhagan et al. 2019; Liang et al. 2020; Pillutla et al. 2022; Yu et al. 2022; Liu et al. 2022; Jang et al. 2022).

Despite the growing popularity of FTL, there is a paucity of exploration of the potential privacy and security risks in this domain. Few research examine the potential for backdoor attacks on FTL (Ye et al. 2024; Lyu et al. 2024), and there is no discernible work on the examination of inference attacks. Admittedly, FTL is more challenging to attack than other scenarios within FL. On the one hand, transfer learning has the ability to mitigate privacy breaches, which may result in researchers failing to pursue further research. Techniques such as knowledge distillation and parameter decoupling facilitate the separation of the model into an upstream and a downstream component. Incomplete model increase the complexity of attacks to infer privacy knowledge from heterogeneous data. Furthermore, the content of personalized information in the global model after multi-source migration learning decrease and the shared model becomes more generalized. In order to adapt to different data distributions, the shared model may become less detailed in capturing individual data features. This can somewhat mitigate the risk of personalized information leakage and reduce the likelihood of an attacker obtaining specific individual data through reverse inference of the model.

Despite these challenges, we discover that the current design of FTL still presents significant privacy concerns that warrant further examination. In particular, the parameter decoupling strategy, which involves fine-tuning structures and gradient update mechanisms, may inadvertently reveal personalised information, thereby increasing the probability that a server could infer private labels from client data. This gap in the literature highlights the necessity for a comprehensive investigation into the potential privacy vulnerabilities that may emerge in FTL.

*Corresponding author

In response to fill this gap, we propose a novel label inference attack that leverages a combination of meta-classifiers and contrastive learning within the context of FTL, called the CML attack. Our approach is designed to exploit the local personalized information embedded in the differences observed during FTL fine-tuning process. We first show that label inference attacks already available in FL fail in FTL. Under parameter decoupling strategy, feature extraction by complementary models or by directly exploiting the difference in posterior outputs before and after local fine-tuning fails to infer a valid distribution of personalized labels. In order to fully utilize the downstream models uploaded by trained clients, we introduce a meta-classifier for high-dimensional feature extraction. Experiments show that the meta-classifier attack obtains an attack success rate of 77.51% on the CIFAR-10 image dataset with only 5% of auxiliary labeled samples and Dirichlet distribution parameter $\alpha = 0.1$. To enhance the attack model's capacity to represent personalized information and improve the attack performance, we further introduce contrastive learning (CL), which enables the attack model to distinguish between personalized and global generic information. Experimental results demonstrate that the personalized representations extracted through CL effectively enhance the attack performance. Subsequently, we integrate CL into meta-classifier to develop CML attack. Our experiments indicate that CML attack can achieve high attack success rate of 79.11% on the evaluated dataset with Dirichlet distribution parameter $\alpha = 0.1$.

Our contributions can be summarized in the following perspectives:

- To the best of our knowledge, this work represents a pioneering effort in evaluating the privacy risks associated with FTL. We offer insights into the causes of label leakage and expose how such leakage occurs in FTL through the lens of parameter decoupling strategies.
- We introduce a label inference attack that combines contrastive learning with meta-classifiers. Our attack demonstrates the feasibility of distinguishing personalized information from global information in FTL. As we known, we are the first to introduce contrastive learning in inference attack of the federated learning, which lead to a completely new attack perspective.
- We evaluate our attacks on various tasks under different heterogeneous data settings and achieve outstanding attack performance. We demonstrate that our proposed attack method is effective in extracting sensitive information, highlighting potential privacy risks that have been largely overlooked in related research. Our work underscore the importance of developing more secure and privacy-preserving techniques in FTL.

Related Work

Federated transfer learning strategies. Federated transfer learning is particularly useful in scenarios where there is minimal overlap in both features and samples between participants, such as a collaboration between banks and superstores in different regions (Li et al. 2020). Current research strategies for FTL mainly solve challenges of data shift and model shift. Instance augmentation enhance data homogeneity of various participants through techniques like oversampling (Younis and Fisichella 2022; Wu et al. 2020) and downsampling (Tsai et al. 2019). Feature clustering use model-related (Ouyang et al. 2022) and data-related (Duan et al. 2020) information to find a more abstract representation of original features to group similar data distributions together (Zhuang et al. 2020). Parameter decoupling decompose models of participants and perform global aggregation by sharing a homogeneous feature extractor (Collins et al. 2021; Oh, Kim, and Yun 2021) or sharing a homogeneous classifier (Liu et al. 2022; Jang et al. 2022). Knowledge distillation (Daliang and Junpu 2019) is used for dealing with model adaptive FTL induced by model heterogeneity. Other techniques including category relationship matrix (Liu et al. 2021) and domaindependent consistency regularization (Shen et al. 2020) are involved in semisupervised FTL scenario (Jeong et al. 2020) where some participants have fully labeled data while others have only unlabeled data.

Contrastive learning in federated learning. CL enables models to extract meaningful representations from unlabeled data. Extensive research focus on introducing CL in federated learning to mitigate the problem of model and data heterogeneity. MOON is to utilize the similarity between model representations to correct the local training of individual parties, conducting contrastive learning in model-level (Li, He, and Song 2021). FedPCL shares knowledge across clients through their class prototypes and builds client-specific representations in a prototype-wise contrastive manner (Tan et al. 2022).

Label inference attack. Ensuring the privacy of the private labels should be a fundamental guarantee provided by federated learning, as the labels might be the key asset of the participant or highly sensitive. Several works have demonstrated label leakage risks associated with FL across a variety of attack method. A passive label inference attack with model completion is presented with the direct label inference attack and the active label inference attack (Fu et al. 2022), revealing and shedding lights on the new label leakage issue of VFL. Under IoT seniors with edge computing, LDIA (Gu and Bai 2023) learns individual features of the output layer updates over different label distributions, and then perform inference from local models uploaded by users.

With privacy threats, more work begins to focus on defences against label inference attacks. DCAE (Zou et al. 2022) significantly boosts the main task accuracy than other known methods when defending various label inference attacks based on autoencoder and entropy regularization to disguise true labels. As for the differential privacy, label-DP limit the threat of LIAs below a certain level by showing the semantic protection and choosing property ε (Wu et al. 2022). Furthermore, a novel framework KDk (Arazzi, Nicolazzo, and Nocera 2024) combines knowledge distillation and k-anonymity to provide a defense mechanism against potential label inference attacks in a VFL scenario.

Other privacy-revealing attack. Privacy has always been a hot topic of concern for machine learning. In federated learning, training deep neural networks (Nasr, Shokri, and Houmansadr 2019) for white-box membership inference attac exploits the privacy vulnerabilities of the stochastic gradient descent algorithm. Recent work uses a GAN for data augmentation on limited prior federation data. The adversary then merges outputs from global and user models, leaking individual privacy. (Liu, Jiang, and Zhu 2023). In terms of transfer learning, the black-box meta-classifier with the "query tuning" technique (Xu et al. 2021) can conduct property inference attacks on a victim's tuned downstream model (Tian et al. 2023). In the context of online learning, an encoder-decoder formulation extract label leakage from the change in the output of a black-box ML model before and after being updated. The proposed CBM-GAN even can reconstruct accurate updated samples (Salem et al. 2020a). The scenarios of these tasks are similar to Federated Transfer Learning and can therefore serve as a reference by migrating to Federated Transfer Learning.

Methodology

In this section, we begin by introducing the parameter decoupling as typical strategy in FTL framework, and then we share the insights on why parameter decoupling is vulnerable to label inference attacks. Subsequently, we outline our label inference attacks based on the insights.

Parameter Decoupling

Parameter decoupling is the process of sending part of the local model for aggregation by decomposing the participant's model into body (downstream model) and head (upstream model) parameters, which can improve the accuracy of the target domain while taking into account the accuracy of the global model. We take the typical structure FedRep (Collins et al. 2021) as an instance. The server and the client jointly learn $n-1$ layers of the model (body), and the client train the personalized classification layer (head) on its own. In each round, after the server's global model is broadcast, the selected client i freezes the global model to update the head h on f using one gradient descent in η steps: $\mathbf{h}_i^{t,s} = \text{GRD}\left(f_i\left(\mathbf{h}_i^{t,s-1}, \phi^t\right), \mathbf{h}_i^{t,s-1}, \eta\right)$, where t is global training round and s is local tuning round. After training τ steps to update h, client similarly freeze h to update body parameters ϕ: $\phi_i^{t,s} = \text{GRD}\left(f_i\left(\mathbf{h}_i^{t,\hat{h}}, \phi_i^{t,s-1}\right), \phi_i^{t,s-1}, \eta\right)$. The client completes the update and upload $\phi_i^{t,\tau}$ to server for aggregation, so that the next round global model becomes: $\phi^{t+1} = \sum_{i\in\mathcal{N}} \phi_i^{t,\tau}$.

Possible Privacy Leakage in Parameter Decoupling

In the training process of parameter decoupling, the participants and the server only exchange the parameters of the body. Although the fine-tuned classification layer is not accessible to the server, the difference between the parameters of the downstream model during fine-tuning still poses a risk of privacy leakage. In the following, we share findings on two components that can lead to label leakage: leakage from the upload downstream model and leakage from the difference during fine-tuning.

Leakage from upload downstream model. After each client completes the local fine-tuning of the head, they freeze it and proceed to train the body. Although personalized information is primarily encoded in the classification layer, the downstream model still learns a few latent features derived from the input features and the fine-tuning layer distribution. As global rounds progress and the model fit, the representation of the downstream model shared by the client increasingly aligns with the classification layer's distribution. Consequently, an honest and curious server could utilize the shadow dataset to analyze the uploaded downstream model, extract potential features from the posterior results, and ultimately predict the distribution of labels.

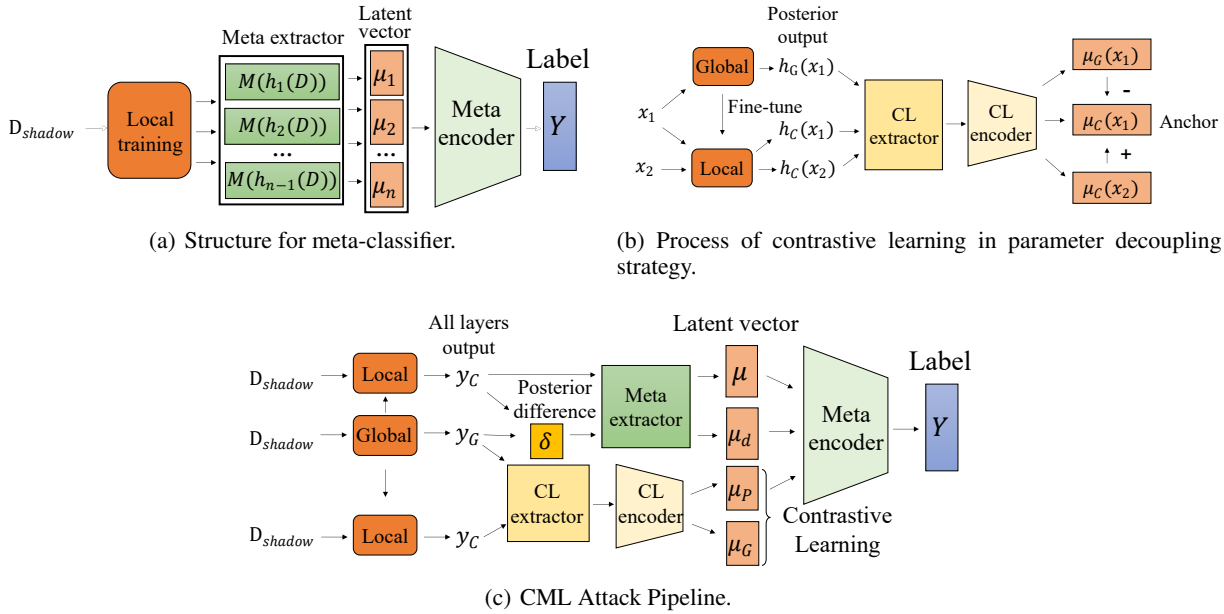
Leakage from model differences during fine-tuning. Previous work proposes that changes of a black-box model before and after fine-tuning with a small amount of data leaks the distribution of labels of the updated dataset (Salem et al. 2020b). Although the most important output layer is unavaliable in FTL, the rest part gradually converges to the distribution of the output layer, and it is reasonable to believe that even the difference in the downstream model leaks the information of the local dataset. Meanwhile, in parameter decoupling strategy, the aggregation of the global model strengthens the characterization of the common data information of each participant, while the fine-tuning of the local model enhanced the understanding of the personalized information. Therefore, the server can leverage the broadcast and accepted model difference degrees to extract and strengthen the personalized representations using deep learning with comparative learning to perform label inference attacks.

In this paper, we propose a label inference attack that combines downstream model extraction with fine-tuned differences by deep representation of personalized information of the target client.

Threat Model

In the proposed attack, we consider black-box access of the adversary to the target model. The attacker is honest and curious server and can only query the model with a set of data samples (i.e., shadow dataset) by obtaining body part of the uploaded model and obtaining the corresponding posterior. This is a reasonable and difficult attack setup for the server in FTL (Shokri et al. 2017).

We also assume that the server has a small number of shadow data samples, which come from random probes and does not necessarily contain the complete and balanced label distribution. In addition, we assume that we are able to know a small amount of random data in the target client through purchase and collection. The adversary can optionally exploit the received model to conduct attack in the training stage. Additionally, the adversary knows the corresponding client of each uploaded model to get the difference between the locally trained models of the target client. With the help of a small amount of auxiliary labeled samples, the adversary can further train a model for label inference based



(a) Structure for meta-classifier.

(b) Process of contrastive learning in parameter decoupling strategy.

(c) CML Attack Pipeline.

Figure 1: Attack ASR vs attacked FedRep rounds

on the upload downstream model and the difference during fine-tuning.

Meta Classifier for Label Inference Attack

To obtain the information directly in the output layer of the downstream model, a complemented passive label inference attack (Fu et al. 2022) obtain the ideal attack effect by complementing the classifier layer. However, we verified that this complementary is limit in FTL. Parameter decoupling strategy separates the representation layer from the classification layer, causing features from the representation layer to be less applicable to local classifiers, particularly when client data distributions vary significantly. This poor coupling leads to difficulties for the supplementary classification layer to utilize the information in the representation layer (body) for effective classification, thus affecting the effectiveness of inference attacks. We therefore suppose to enhance the accuracy of the attack by extracting features that contribute to the classification through a more comprehensive latent feature extraction dependence on the downstream model.

Following the inspiration of the comprehensive framework for white-box membership inference attack (Nasr, Shokri, and Houmansadr 2019), the output of each layer after the model is feature extracted to train meta-classifiers. The shadow samples are fed into the downstream model uploaded by the target client to get the output of each layer of the model: $h_1(x),\ h_2(x),\ \ldots,\ h_{n-1}(x)$ where $\mathcal{D}$ represents the shadow dataset and $x \in \mathcal{D}$. After corresponding feature extraction module of each layer learning personalized information, generated embedding will be concatenated together and brought into the encoder to calculate loss with the label of the samples. This structure help to maximize the extraction of the potential feature space in the model, shown in Fig. 1(a).

After the meta-classifier training is completed, the adversary is given a complete representation of the downstream model. In contrast to utilizing only the last layer of a posterior information, the meta-classifier extracts more high-dimensional information, extracting fully privacy leaked from the model. With an efficient encoder, this information is possible to predict the label of each data of the adversary.

Label Inference Attack via Contrastive Meta Learning

For non-IID sample distributions, each client's sample contains a distribution of individualized features and labels that are unique, and a distribution of global features and labels that are common to every client else. This is a common data characteristic for federated migration learning, which provides us with new perspectives that separating these two distributions as much as possible with the model can yield more accurate personalization information.

As our mentioned in the previous section about the causes of privacy revelation, we found that in parameter decoupling strategy, the global model averages the parameters of the fine-tuned model for each individualized client, whose training represents homogeneous information in the data distribution of all clients, while the local fine-tuned model implies the personalized information unique to that client. For the same sample, the global model maps to a homogeneous feature space, while the local model maps to a personalized label distribution. We hope to expand the difference between this two spaces through CL to extract a more effective representation of personalized information that maximally distinguishes it from the global information for label inference.

Additionally, the label distribution of the shadow dataset differs from the label distribution of the target client in the real scenarios. CL allows unsupervised information extraction rather than relying on labels. However, there are two difficulties with method.

The first problem is whether the posterior output of the model is valid as a sample for CL. Sever do not have access to the data, therefore it cannot directly perform representational learning on the data to extract personalized information. However, the posterior output of the model can be substituted as a valid input sample. For federated learning, the posterior outputs of the model are labeled distributions, reflecting the probability that the samples belong to each category. Using these distributions as the input of CL ensures that comparisons of samples are consistent. Furthermore, the posterior output of the model shows a high-dimensional representation of the data, and contrastive learning in this case means comparing the learned knowledge of the model, which is consistent with our learning objective.

The contrastive loss function for posterior output as a sample is:

$$\mathcal{L}_{\text{contrastive}} = \sum_{x \in \mathcal{D}} \ell\left(f(h_c(x), f(h_g(x))\right) \tag{1}$$

where h_c is the client upload model and and h_g is the broadcast global model. f is a deep neutral network and ℓ is the contrastive loss function. We use the body model as part of contrastive learning, integrated with the DNN for feature extraction. For the distance calculation of similarity, we choose the Euclidean distance, which reflects the difference in directions as well as values of vectors. Since the heterogeneous data of customers are Diriclet distributed, the spatial distance between global and personalized information will not be too far, so that the distance between the two embedding vectors needs to be controlled within a reasonable range. We set a threshold to control the maximum distance. In addition, the distance of personalized vector between each sample is supposed to be small. This reminds us of the triplet loss function with construction of anchors, positive samples and negative samples. The output of the model before fine-tuning, which as the model broadcasted by the server, is taken as the negative sample. The output from model uploaded by the client serves as the anchor sample. They form the negative sample pairs: $\{h_c(x_i), h_g(x_i)\}$. The output of the fine-tuned model within the batch is formed a positive sample pair with the anchor samples: $\{h_c(x_i), h_c(x_{i+1})\}$. For deep and potential information, we also use output from all layers mentioned in previous method. The loss function turns out to following equation:

$$\begin{aligned} \mathcal{L}_{\text{contrastive}} \quad = \quad & (\max(0, P\_distance \\ & -N\_distance + margin)) \end{aligned} \tag{2}$$

where

$$P\_distance = \|f(h_C(x_i)) - f(h_C(x_{i+1}))\|_2$$
$$N\_distance = \|f(h_C(x_i)) - f(h_G(x_i))\|_2$$
$$h_C(x_i) = [h_{c_{\text{layer1}}}(x_i), \ldots, h_{c_{\text{layern}}}(x_i)]$$
$$h_G(x_i) = [h_{g_{\text{layer1}}}(x_i), \ldots, h_{g_{\text{layern}}}(x_i)]$$

and $margin$ is threshold to avoid non-difference of the distances between positive and negative sample pairs. f is similar with the previous meta classifier, extract all layers of the model and generate a embedding vector. The process of contrastive learning shown as Fig. 1(b).

The second obstacle is that when FTL training, the differences in local model before-and-after fine-tuning gradually diminish, which leads to convergence issue of contrastive learning. Therefore, we attack the model in the front rounds during FTL training for contrastive learning. To integrate with meta-classifier, the personalized embedding vector obtained from the inference of the CL model μ_p concatenate the latent vector of meta-extractor μ. We also put the posterior output differences of the models during the fine-tuning directly into the meta-extractor and join latent vector μ_d together with μ_p and μ. The final integrated vector is embedded in the meta-encoder.The total CML attack pipeline is shown in Fig. 1(c).

Experimental Evaluation

Datasets, models, and hyperparameters. We evaluated the CML label inference attack on 2 large-scale benchmark datasets: CIFAR-10 and CIFAR-100 (Krizhevsky, Hinton et al. 2009). CIFAR-10 is selected for the convenience of later visualized analysis. CIFAR-100 is selected to prove the feasibility of our attacks on datasets with many class labels. On each dataset, we adopt the setting of the Diriclet non-IID data distribution. The Dirichlet distribution is a family of continuous multivariate probability distributions parameterized by a vector of positive real numbers α, whose elements sum to 1. This parameter vector determines the shape and properties of the Dirichlet distribution. When α equal to 1, the Dirichlet distribution becomes a uniform distribution. The closer the value of α is to 1, the more the Dirichlet distribution resembles a uniform distribution. Conversely, the smaller the value of α, the more heterogeneous the distribution becomes. We take $\alpha = 0.5, 0.3, and\ 0.1$ in our experiment to illustrate the effect of heterogeneity on attacks. There were 20 clients in involved. AlexNet is chosen for the training model in FedRep, whose global model contains 7 layers, which means that the classification layer is excluded. 50% clients are randomly chosen to participate training per round, total training finish at the 30-th round. Each client train 8 rounds, which 4 rounds to fine-tune the head and freeze body, and 4 rounds to freeze the head and fine-tune the body. All algorithms were implemented using PyTorch v2.2.0 and executed on an NVIDIA V100 GPU with 32 GB of memory.

Baseline attacks. Since our work is pioneering in the study of label inference attacks in FTL, there are limited existing methods for comparison. Our initial baseline extends the update-based label inference attack method (Salem et al. 2020a) to the context of federated learning (We call it ULIA). This approach utilizes the differences in local learning updates as inputs to the encoder, whose structure shown in the Appendix. Furthermore, we involve the Complementary Label Inference Attack (CLIA) (Fu et al. 2022) as our

second baseline, which perfectly fits parameter decoupling strategy.

Dataset	Method	l_r	Batch	Epoch
CIFAR-10	ULIA	0.001	64	600
CIFAR-10	CLIA	0.01	256	600
CIFAR-10	CML(Ours)	0.001	256	600
CIFAR-100	All methods	0.001	256	4000

Table 1: Hyperparameter description for attack model.

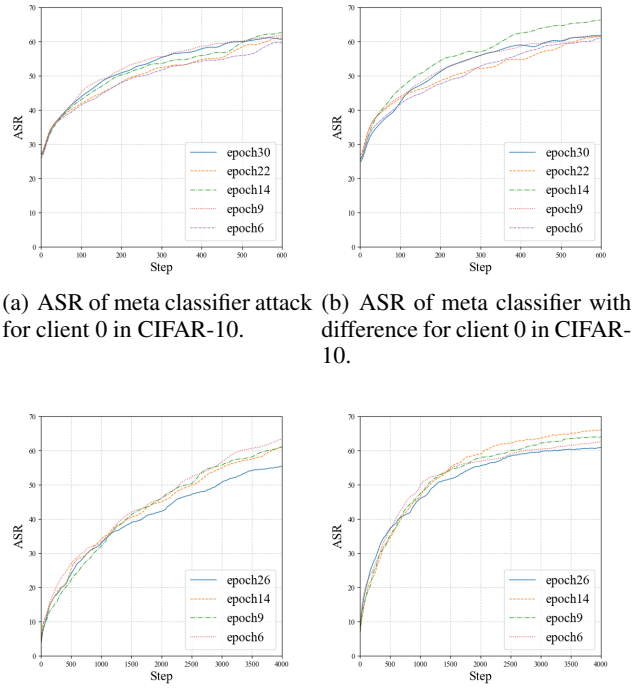
Attack settings. In the label inference attack, the number of samples per client is 3,000, and our auxiliary data is 500 random label samples, including 150 random samples from the target client. For the meta-classifier extractor of the Alexnet, convolutional module and the linear module are used for feature extraction on the output of 5 convolutional layers and the output of 2 linear layer respectively. The encoder of the meta-classifier is a 4-layer MLP. The structure of the extractor and encoder shown in Appendix B. The structure of the extractor in the CML is the same with meta-classifier, while the encoder map to a 16-dimensional embedding vector. 10 rounds training of contrastive learning are performed with the maximum threshold of Euclidean distance set to 10 and margin set to 1. The hyperparameters of the attack model are demonstrated in Table 1. 2 clients (client 0 and client 10) are attacked in our experiment to test attacking robustness. The adversary attack client 0 for the 6-th, 9-th, 14-th, 22-th, 30-th training iterations of FedRep, and attack client 10 for the 6-th, 9-th, 14-th, 26-th training iterations. Each attack effect was tested 3 times taking the average attack success rates (ASR).

α	FedAvg	FedRep		
		Train ACC	ULIA	CLIA
0.5	46.14	52.79	33.76	37.49
0.3	45.11	66.34	32.19	37.53
0.1	30.68	87.76	26.71	37.53

Table 2: Training and attacking performance for FedAvg and FedRep on CIFAR-10.

Attack Performance

To confirm the motivation of our work, we compare the attack effect of the two baselines under FedAvg and FedRep in different heterogeneous settings. As shown in Table 2, we verified that the two methods of baseline have limitations in the FTL. The learning effect of FedRep is significantly better than that of FedAvg in each data settings. For labeling inference attacks, the sever in FedAvg has complete model to infer the label, whose ASR can regard as training accuracy. However, the effect of the existing attack methods against FedRep is disappointing. For ULIA, the a posterior output dimension of the upload Alexnet model is 4096 dimensions, and the difference between the local before and after training values is relatively small, which does not allow for the extraction of valid information by direct model learning. As for CLIA, although the model is complemented, the personalized training of FedRep makes each client's classifier significantly different from the others, which further increases the complexity of the challenge.Even if some information is obtained from the representation layers, it is not enough to reliably infer the parameters or label distribution of the classification layer. As a result, the attack is not as effective as expected.



(a) ASR of meta classifier attack for client 0 in CIFAR-10.

(b) ASR of meta classifier with difference for client 0 in CIFAR-10.

(c) ASR of meta classifier attack for client 10 in CIFAR-100.

(d) ASR of meta classifier with difference for client 10 in CIFAR-100.

Figure 2: Attack ASR vs attacked FedRep rounds

We compare the CML attack with 2 baselines. Since the previous section mentioned the need to select the forward training rounds for the CML attack to ensure the convergence of the comparison learning, we attack round 6th in FedRep training for the comparison experiments. Due to the limited length of the article, the results for CIFAR-10 are shown in this section. For the results of CIFAR-100, please refer to Appendix. In the following tables, we use bold fonts to highlight the highest ASR values obtained among all attack methods. As shown in Table 3, with the same heterogeneous setup, CML attack can obtain significant higher ASR value than the other baseline attack methods. This demonstrates the effectiveness of the CML attack, particularly under extreme heterogeneous data conditions in FTL.

We conduct ablation experiments to compare the separate effects of relying only on the meta-classifier and adding model difference directly to the meta-extractor during fine-tuning to show the significance of each part in our attack. Compared to utilizing only the last layer of the upload

LIA attack	$\alpha = 0.5$		$\alpha = 0.3$		$\alpha = 0.1$	
	client0	**client10**	**client0**	**client10**	**client0**	**client10**
ULIA	33.76	33.15	32.19	35.86	26.71	38.31
CLIA	37.49	34.431	37.53	39.30	37.53	40.22
Meta classifier	74.28	74.15	62.84	77.02	61.23	73.99
Meta classifier + diff	74.30	74.06	63.03	76.72	61.91	77.26
CML(Ours)	**77.15**	**76.94**	**68.55**	**78.70**	**67.86**	**79.11**

Table 3: ASR of different label inference attacks on CIFAR-10 in round 6th (%).

	$\alpha = 0.1$					$\alpha = 0.3$				
epoch	**ULIA**	**CLIA**	**Meta**	**Meta+diff**	**CML(Ours)**	**ULIA**	**CLIA**	**Meta**	**Meta+diff**	**CML(Ours)**
6	**26.71**	**37.53**	61.23	61.91		**32.19**	**37.53**	62.84	63.03	
9	22.79	36.71	62.20	62.34		29.49	36.71	62.48	62.88	
14	20.71	36.21	**63.47**	**67.44**	**67.86**	30.24	36.21	**67.15**	**67.65**	**68.55**
22	10.37	36.35	61.94	62.34		10.23	36.35	61.72	63.02	
30	9.27	37.16	62.07	62.48		9.04	37.16	61.86	61.86	

Table 4: ASR of different label inference attacks in any attacked epoch on CIFAR-10 in client 0 (%).

	FedRep ACC		**CML ASR (Ours)**	
α	0.1	0.5	0.1	0.5
epsilon=0	87.76	52.79	67.86	77.15
epsilon=0.5	72.23	35.11	62.11	71.86
epsilon=1.2	68.41	33.24	59.31	68.22

Table 5: Impact of different privacy dudgets on model ACC and ASR of CML attack with CIFAR-10 in client 0 (%).

model, the meta-classifier extracts deeper into each layer of the network, reducing the dimension while extracting the privacy information leaked from the model itself, which is the foundation of the CML attack. Capturing subtle variations between models enhances the attack's precision, making it more effective than simply relying on the meta classifier. CML provides a further expansion of the personalized information to global information distance and extracts more accurate latent representations.

We compared the effect of the selected FedRep rounds on our attacks. Fig. 2 shows the meta classifier trains a better model in the later rounds as the FedRep training is fitted under Diriclet $\alpha = 0.1$. The meta classifier with difference attack method reach the best ASR in the middle rounds by incorporating differences from earlier training rounds into the attack model. The method capitalizes on the local training dynamics of each client in early-round, allowing the attack to exploit inconsistencies between clients' updates.

Table 4 further demonstrates that the CML attack consistently achieves optimal ASR across all rounds, outperforming other attack methods. This is largely due to its ability to effectively separate and extract personalized information from the global model updates. The extraction model's understanding of the distribution of personalized data in the target is enhanced through comparative learning, and the obtained personalized representation is integrated with the features extracted by the meta-classifier through a gating layer, which combines two privacy leakage pattern. As a result, the CML attack can accurately infer labels even in highly heterogeneous environments. Moreover, this is the first attempt to use contrastive learning in an attack, demonstrating the potential of unsupervised methods in attack scenarios.

Denfence Performance

We conducted additional validation experiments on existing defense mechanisms to highlight the advantages of our attacks over the most prominent defenses against label inference attacks . All our attacks leverage the posterior difference as input. Therefore, differential privacy is a comprehensive privacy-enhancing strategy against our attacks. In each iteration, the clients add Gaussian-distributed noise upstream of the model to protect the parameters of the target uploaded model. Shown as Table 5, our experiments reveal that in scenarios of extreme data heterogeneity, the utility of the model can drop significantly under a small privacy budget. However, the effectiveness of our attack drops by only 5%. We leave an in-depth exploration of effective defense mechanisms against our attacks for future work.

Conclusion

Most previous research on FTL focuses on performance and effectiveness, with little attention, however, paid to the issue of FTL privacy and security risks. To bridge this gap, we provide a first investigation of privacy risks in FTL. Taking the commonly used parameter decoupling structure as an example, our study explicitly reveals the key causes of privacy leakage in federated transfer learning tasks. Instantiating the theoretical discussion, we propose an effective integration attack method, CML attack, to mine local personalized information from the differences during fine-tuning under extreme heterogeneous data distribution. Extensive experimental results show that the CML attack is an effective privacy threat to FTL. This should motivate better defense efforts in the future.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant No.2023YFF0905300), the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 62402040), the Postdoctoral Fellowship Program of CPSF (Grant No.GZB20230938), the China Postdoctoral Science Foundation (Grant No.2024T171132 and No.2023M740246).

References

Arazzi, M.; Nicolazzo, S.; and Nocera, A. 2024. KDk: A Defense Mechanism Against Label Inference Attacks in Vertical Federated Learning. *arXiv preprint arXiv:2404.12369*.

Arivazhagan, M. G.; Aggarwal, V.; Singh, A. K.; and Choudhary, S. 2019. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*.

Collins, L.; Hassani, H.; Mokhtari, A.; and Shakkottai, S. 2021. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, 2089–2099. PMLR.

Daliang, L.; and Junpu, W. 2019. FedMD: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*, 2.

Duan, M.; Liu, D.; Chen, X.; Liu, R.; Tan, Y.; and Liang, L. 2020. Self-balancing federated learning with global imbalanced data in mobile systems. *IEEE Transactions on Parallel and Distributed Systems*, 32(1): 59–71.

Fu, C.; Zhang, X.; Ji, S.; Chen, J.; Wu, J.; Guo, S.; Zhou, J.; Liu, A. X.; and Wang, T. 2022. Label inference attacks against vertical federated learning. In *31st USENIX security symposium (USENIX Security 22)*, 1397–1414.

Gu, Y.; and Bai, Y. 2023. LDIA: Label distribution inference attack against federated learning in edge computing. *Journal of Information Security and Applications*, 74: 103475.

Guo, W.; Zhuang, F.; Zhang, X.; Tong, Y.; and Dong, J. 2024. A comprehensive survey of federated transfer learning: challenges, methods and applications. *Frontiers of Computer Science*, 18(6): 1–34.

Jang, J.; Ha, H.; Jung, D.; and Yoon, S. 2022. Fedclassavg: Local representation learning for personalized federated learning on heterogeneous neural networks. In *Proceedings of the 51st International Conference on Parallel Processing*, 1–10.

Jeong, W.; Yoon, J.; Yang, E.; and Hwang, S. J. 2020. Federated semi-supervised learning with inter-client consistency & disjoint learning. *arXiv preprint arXiv:2006.12097*.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2019. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*.

Kevin, I.; Wang, K.; Zhou, X.; Liang, W.; Yan, Z.; and She, J. 2021. Federated transfer learning based cross-domain prediction for smart manufacturing. *IEEE Transactions on Industrial Informatics*, 18(6): 4088–4096.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Li, L.; Fan, Y.; Tse, M.; and Lin, K.-Y. 2020. A review of applications in federated learning. *Computers & Industrial Engineering*, 149: 106854.

Li, Q.; He, B.; and Song, D. 2021. Model-Contrastive Federated Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10713–10722.

Liang, P. P.; Liu, T.; Ziyin, L.; Allen, N. B.; Auerbach, R. P.; Brent, D.; Salakhutdinov, R.; and Morency, L.-P. 2020. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*.

Liu, C.; Yang, Y.; Cai, X.; Ding, Y.; and Lu, H. 2022. Completely heterogeneous federated learning. *arXiv preprint arXiv:2210.15865*.

Liu, Q.; Yang, H.; Dou, Q.; and Heng, P.-A. 2021. Federated semi-supervised medical image classification via inter-client relation matching. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, 325–335. Springer.

Liu, Y.; Jiang, P.; and Zhu, L. 2023. Subject-Level Membership Inference Attack via Data Augmentation and Model Discrepancy. *IEEE Transactions on Information Forensics and Security*, 18: 5848–5859.

Liu, Y.; Kang, Y.; Xing, C.; Chen, T.; and Yang, Q. 2020. A secure federated transfer learning framework. *IEEE Intelligent Systems*, 35(4): 70–82.

Lyu, X.; Han, Y.; Wang, W.; Liu, J.; Zhu, Y.; Xu, G.; Liu, J.; and Zhang, X. 2024. Lurking in the shadows: Unveiling Stealthy Backdoor Attacks against Personalized Federated Learning. *arXiv preprint arXiv:2406.06207*.

Nasr, M.; Shokri, R.; and Houmansadr, A. 2019. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, 739–753. IEEE.

Oh, J.; Kim, S.; and Yun, S.-Y. 2021. Fedbabu: Towards enhanced representation for federated image classification. *arXiv preprint arXiv:2106.06042*.

Ouyang, X.; Xie, Z.; Zhou, J.; Xing, G.; and Huang, J. 2022. Clusterfl: A clustering-based federated learning system for human activity recognition. *ACM Transactions on Sensor Networks*, 19(1): 1–32.

Pillutla, K.; Malik, K.; Mohamed, A.-R.; Rabbat, M.; Sanjabi, M.; and Xiao, L. 2022. Federated learning with partial model personalization. In *International Conference on Machine Learning*, 17716–17758. PMLR.

Saha, S.; and Ahmad, T. 2021. Federated transfer learning: concept and applications. *Intelligenza Artificiale*, 15(1): 35–44.

Salem, A.; Bhattacharya, A.; Backes, M.; Fritz, M.; and Zhang, Y. 2020a. Updates-Leak: Data Set Inference and Reconstruction Attacks in Online Learning. In *29th USENIX*

Security Symposium (USENIX Security 20), 1291–1308. USENIX Association. ISBN 978-1-939133-17-5.

Salem, A.; Bhattacharya, A.; Backes, M.; Fritz, M.; and Zhang, Y. 2020b. {Updates-Leak}: Data set inference and reconstruction attacks in online learning. In *29th USENIX security symposium (USENIX Security 20)*, 1291–1308.

Shen, T.; Zhang, J.; Jia, X.; Zhang, F.; Huang, G.; Zhou, P.; Kuang, K.; Wu, F.; and Wu, C. 2020. Federated mutual learning. *arXiv preprint arXiv:2006.16765*.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, 3–18. IEEE.

Tan, Y.; Long, G.; Ma, J.; LIU, L.; Zhou, T.; and Jiang, J. 2022. Federated Learning from Pre-Trained Models: A Contrastive Learning Approach. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 19332–19344. Curran Associates, Inc.

Tian, Y.; Suya, F.; Suri, A.; Xu, F.; and Evans, D. 2023. Manipulating transfer learning for property inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15975–15984.

Tsai, C.-F.; Lin, W.-C.; Hu, Y.-H.; and Yao, G.-T. 2019. Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. *Information Sciences*, 477: 47–54.

Wu, Q.; Chen, X.; Zhou, Z.; and Zhang, J. 2020. Fedhome: Cloud-edge based personalized federated learning for in-home health monitoring. *IEEE Transactions on Mobile Computing*, 21(8): 2818–2832.

Wu, R.; Zhou, J. P.; Weinberger, K. Q.; and Guo, C. 2022. Does Label Differential Privacy Prevent Label Inference Attacks? *arXiv preprint arXiv:2202.12968*.

Xu, X.; Wang, Q.; Li, H.; Borisov, N.; Gunter, C. A.; and Li, B. 2021. Detecting ai trojans using meta neural analysis. In *2021 IEEE Symposium on Security and Privacy (SP)*, 103–120. IEEE.

Ye, T.; Chen, C.; Wang, Y.; Li, X.; and Gao, M. 2024. BapFL: You can Backdoor Personalized Federated Learning. *ACM Transactions on Knowledge Discovery from Data*.

Yosinski, J.; Clune, J.; Bengio, Y.; and Lipson, H. 2014. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27.

Younis, R.; and Fisichella, M. 2022. FLY-SMOTE: Rebalancing the non-IID iot edge devices data in federated learning system. *IEEE Access*, 10: 65092–65102.

Yu, H.; Zhang, N.; Deng, S.; Yuan, Z.; Jia, Y.; and Chen, H. 2020. The devil is the classifier: Investigating long tail relation classification with decoupling analysis. *arXiv preprint arXiv:2009.07022*.

Yu, S.; Nguyen, P.; Abebe, W.; Qian, W.; Anwar, A.; and Jannesari, A. 2022. Spatl: Salient parameter aggregation and transfer learning for heterogeneous federated learning. In *SC22: International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–14. IEEE.

Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; and He, Q. 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1): 43–76.

Zou, T.; Liu, Y.; Kang, Y.; Liu, W.; He, Y.; Yi, Z.; Yang, Q.; and Zhang, Y.-Q. 2022. Defending Batch-Level Label Inference and Replacement Attacks in Vertical Federated Learning. *IEEE Transactions on Big Data*, 1–12.