Out-of-Distribution Generalization of Federated Learning via Implicit Invariant Relationships

Yaming Guo, Kai Guo, Xiaofeng Cao, Tieru Wu, Yi Chang

Jilin University

Introduction

- **Federated Learning (FL)** Multiple clients collaborate to solve machine learning problems under the coordination of a server, where each client's raw data is stored locally and is not exchanged or transferred. The federated networks are usually comprised of a large number of clients that generate and collect data in a non-identical distribution manner, most of which may never participate in training. The standard FL follows the ***E**mpirical **R**isk **M**inimization* (ERM) principle and is formalized as:

$$\min_f \mathop{\mathbb{E}}_{c\sim\mathcal{Q}_{\text{par}}}\left[\mathcal{R}_c(f)=\mathop{\mathbb{E}}_{X^c,Y^c}\ell(f(X^c),Y^c)\right].$$

- **Out-of-distribution (OOD) Generalization** Since the distribution shift probably exists between participating and non-participating (unseen) clients, models that follow ERM may perform poorly on the non-participating clients. In order to generalize the model appropriately to non-participating clients, we examine the problem of OOD generalization in FL, formally defined as:

$$\min_f \max_{c\in\mathcal{C}_{all}}\left[\mathcal{R}_c(f)=\mathop{\mathbb{E}}_{X^c,Y^c}\ell(f(X^c),Y^c)\right].$$

- **Invariant Relationships** A proven strategy in the OOD generalization literature is to learn the invariant relationships that are stable across distributions and build a model that works equally well over OOD. Intuitively, an invariant relationship is a statistical relationship between inputs and target variables that is maintained across all data distributions. This can be expressed by the following equation, which holds for all $c,c'\in\mathcal{C}_{\text{all}}$ and for all $z\in\text{supp}(\mathbb{P}(\Phi(X^c)))\cap\text{supp}(\mathbb{P}(\Phi(X^{c'})))$:

$$\mathop{\mathbb{E}}_{X^c,Y^c}[Y^c|\Phi(X^c)=z]=\mathop{\mathbb{E}}_{X^{c'},Y^{c'}}[Y^{c'}|\Phi(X^{c'})=z].$$

Remark. The relationship between representation $\Phi(X)$ and target Y is fixed across distributions in $\mathcal{C}_{\text{all}}$, i.e., using $\Phi(X)$ to predict Y is **invariant**.

Motivation

Question: *could the current techniques for learning invariant relationships adhere entirely to the federated principles of* ***privacy-preserving*** *and* ***limited communication****?*

An Explicit Perspective

Most existing work concentrates on learning invariant relationships explicitly from three angles: data, representation, and distribution.

- the data/representation-based methods: require a centralized setting where data or representation is shared across clients.
  ✗ **privacy-preserving**
- the distribution-based methods: assume the presence of only a small number of participating clients, most of which are involved in each round of communication.
  ✗ **limited communication**

⇓

A New Perspective: Implicit

Considering that the model parameter is usually the only interaction between the client and the server, we thus stand on a new perspective, i.e., **restrict the method to the parameter space for learning invariant relationships implicitly**.

- the implicit method doesn't need to communicate anything other than the parameter.
  ✓ **privacy-preserving**
- the implicit method can be analyzed in the stochastic optimization framework like standard federated techniques.
  ✓ **limited communication**

Method: FedIIR

- This paper proposes ***Fed**erated Learning with **I**mplicit **I**nvariant **R**elationships* (FedIIR), which implicitly learns invariant relationships for OOD generalization while adhering to the federated principles of privacy-preserving and limited communication.
  - quantify invariant relationships using prediction disagreement:

$$\mathcal{I}(\Phi,\mathcal{C})=\sup_{z\in\text{U}(\Phi,\mathcal{C})}\sup_{(c,c')\in\mathcal{C}^2}|w_c^*(z)-w_{c'}^*(z)|.$$

  - obtain surrogate objectives by parameterization:

$$\sup_{(c,c')\in\mathcal{C}^2}|w(z;\omega-\nabla_\omega\mathcal{R}_c(\theta))-w(z;\omega-\nabla_\omega\mathcal{R}_{c'}(\theta))|$$
$$\lesssim \sup_{(c,c')\in\mathcal{C}^2}\|\nabla_\omega w(z;\omega)\|\|\nabla_\omega\mathcal{R}_c(\theta)-\nabla_\omega\mathcal{R}_{c'}(\theta)\|.$$

Optimization Objective

The proposed FedIIR attempts to **minimize the risk** and **align the inter-client gradient** w.r.t. the classifier, which is formalized as:

$$\min_f \mathop{\mathbb{E}}_{c\sim\mathcal{Q}_{\text{par}}}\left[\mathcal{R}_c(f)+\frac{\gamma}{2}\|\nabla_\omega\mathcal{R}_c(f)-\nabla_\omega\mathcal{R}(f)\|^2\right],$$

where $\mathcal{R}(f)=\mathbb{E}_{c\sim\mathcal{Q}_{\text{par}}}\mathcal{R}_c(f)$ is the global risk.

- If the inter-client gradient is aligned, the model's local learning on one client will also improve its performance on other clients. This indicates that the model implicitly learns invariant relationships that work equally for all clients.

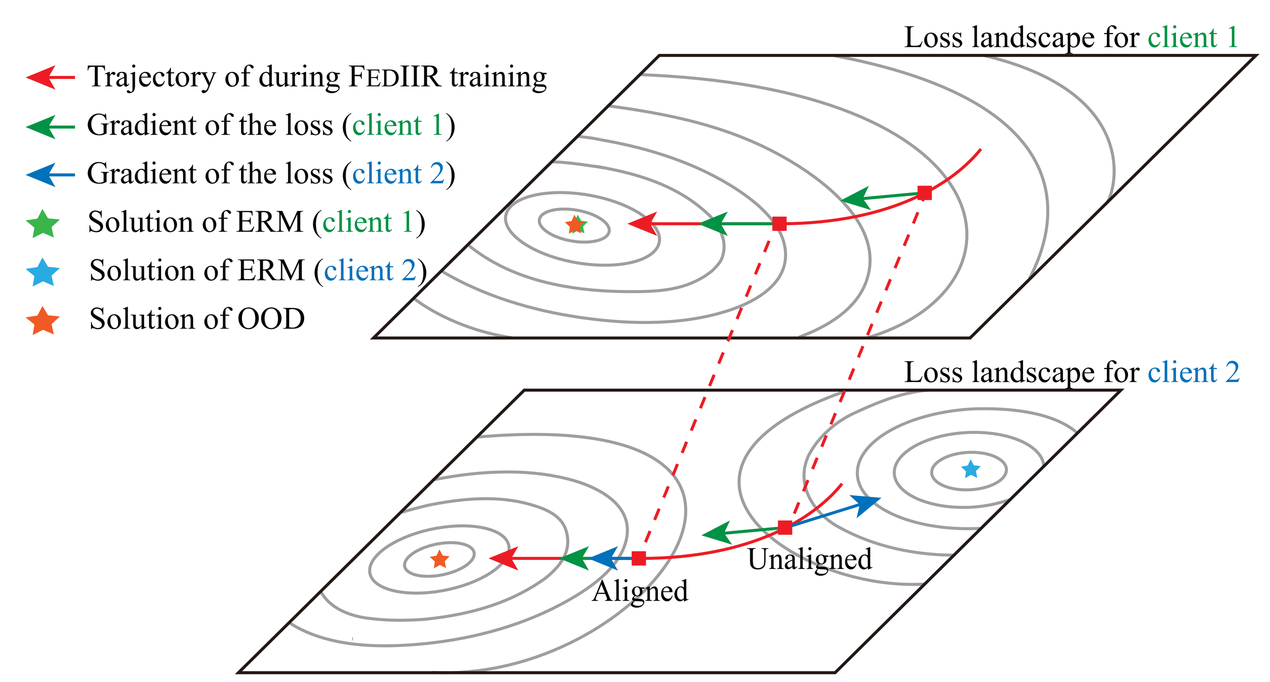

Figure 1. Illustration of inter-client gradient alignment with two clients.

Generalization Analysis

When the number of participating clients is finite in practice, what is the range of non-participating clients to which FedIIR is expected to generalize?

Theorem 3 of the paper

Given the collection $\mathcal{C}_{\text{par}}$ of clients, let's assume that $\ell(\cdot,\cdot)\le M$. Then for all $f=w\circ\Phi\in\mathcal{F}$, we have the following risk bound for the affine combination of participating clients:

$$\sup_{\lambda\in\Lambda_\nu}\mathcal{R}_\lambda(f)\le\mathcal{R}(f)+\widetilde{M}\mathcal{I}(\Phi,\mathcal{C}_{\text{par}})+\widetilde{M}\sup_{(c,c')\in\mathcal{C}^2_{\text{par}}}\rho(\mathbb{P}_c(X),\mathbb{P}_{c'}(X)),$$

where $\widetilde{M}=(1+|\mathcal{C}_{\text{par}}|\nu)M$ is monotonic in ν, and $\rho(\mathbb{P}_c(X),\mathbb{P}_{c'}(X))=\sup_X|\mathbb{P}_c(X)-\mathbb{P}_{c'}(X)|$ is the total variation distance.

If the global risk, invariance constraint and covariate shift are sufficiently small, FedIIR promises to generalize to non-participating clients included in the affine combination of participating clients.

Convergence Analysis

How does the convergence speed of FedIIR fare in the scenario where clients are massively distributed with limited communication?

Assumptions

- **Smoothness** For all clients c, we assume that $R_c(\omega)$ is L-smoothness and Moral-smoothness.
- **Bounded Statistical Heterogeneity** For all clients c, we assume that when there is no perturbation, the variance of the local gradient w.r.t. the global gradient is bounded by G.
- **Bounded Intra-client Variance** For all clients c, we assume that $\nabla R_c(\omega;\zeta)$, $\nabla^2R_c(\omega;\zeta)$, and $\nabla^2R_c(\omega;\zeta)\nabla R_c(\omega;\zeta)$ are unbiased estimates of $\nabla R_c(\omega)$, $\nabla^2R_c(\omega)$, and $\nabla^2R_c(\omega)\nabla R_c(\omega)$, respectively, with variances bounded by σ^2.
- **μ-PL Inequality** We assume that $R(\omega)$ satisfies the μ-PL inequality with $\mu>0$.

Theorem 4 of the paper

Let aforementioned assumptions hold and FedIIR updates with constant local and global step-size such that $\eta_l\le\frac{1}{4KL\sqrt{1+\gamma^2}}$, $\tilde{\eta}=K\eta_g\eta_l<\frac{1}{2\alpha\mu}$. Then, the sequence of iterates generated by FedIIR satisfies

$$\mathbb{E}[R(\omega^t)-R^*]\le(1-2\alpha\mu\tilde{\eta})^t[R(\omega^0)-R^*]+\eta_l\frac{\beta_1G^2+\beta_2\gamma^2\sigma^2+\beta_3\gamma^2G^2\sigma^2}{2\alpha\mu},$$

where $\alpha>0$ is a constant, and β_1,β_2,β_3 are the polynomials in η_l.

For the μ-PL inequality case, FedIIR has a linear convergence rate up to a solution that is proportional to η_l, where the penalty factor γ affects the suboptimality of the solution.

Experiments

- **Results on a small number of clients scenario.**

Algorithm	RotatedMNIST	VLCS	PACS	OfficeHome	Average
	ConvNet	ResNet-18	ResNet-18	ResNet-50	
FedAvg	94.5$_{\pm0.1}$	76.3$_{\pm0.4}$	83.1$_{\pm0.0}$	68.5$_{\pm0.1}$	80.6
FedAdg	94.7$_{\pm0.0}$	**77.1**$_{\pm0.1}$	83.1$_{\pm0.2}$	68.4$_{\pm0.2}$	80.8
FedSR	94.7$_{\pm0.1}$	75.8$_{\pm0.4}$	83.4$_{\pm0.3}$	69.1$_{\pm0.2}$	80.8
FedIIR	**95.0**$_{\pm0.2}$	76.6$_{\pm0.6}$	**83.7**$_{\pm0.3}$	**69.2**$_{\pm0.0}$	**81.1**

Table 1. Average test accuracy (%) using leave-one-out domain validation in the scenario with a small number of clients. Each training domain is treated as a separate participating client, and all participating clients are sampled in each round of communication.

- **Results on a large number of clients scenario (limited communication).**

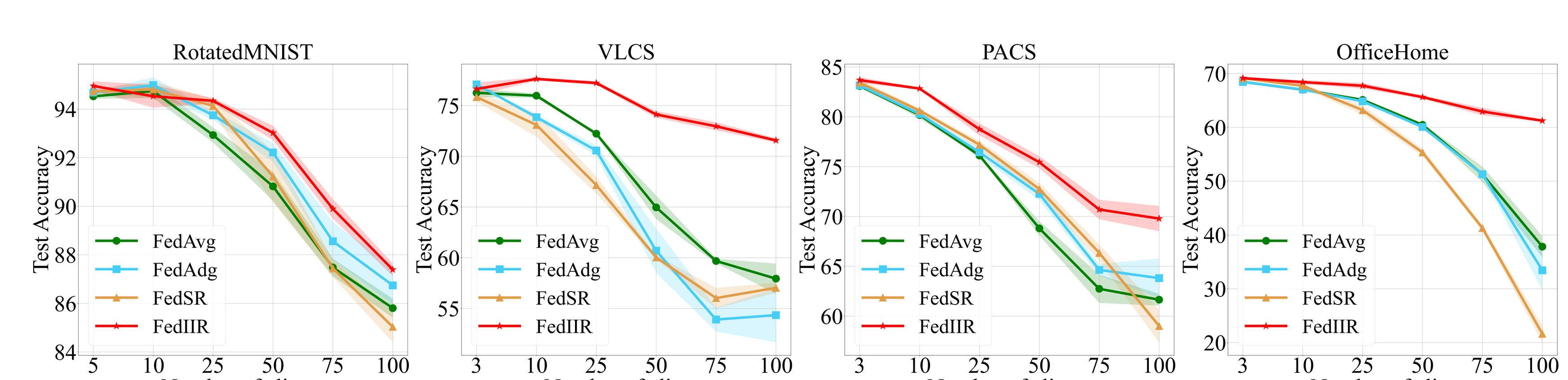

Figure 2. Average test accuracy (%) versus the total number of participating clients, with the number of sampled clients in one communication round matches the number of training domains.