FedRACE: A Hierarchical and Statistical Framework for Robust Federated Learning

Wan Du

University of California, Merced

Gang Yan (University of California, Merced)

Sikai Yang (University of California, Merced)

The Thirty-Ninth Annual Conference on Neural Information Processing Systems

1 Challenges and Motivation

NEURAL INFORMATION PROCESSING SYSTEMS 2025

Challenges:

- **Static Representation Space**: Frozen backbones make all clients share the same latent space, allowing malicious clients to inject semantic backdoors that spread globally
- **Gradient-Based Defenses Fail**: Without gradient signals, traditional defenses (e.g., Krum, FLTrust) relying on update distances lose effectiveness
- **Non-IID Data Amplifies Confusion**: Heterogeneous client data causes natural drift, making it hard to distinguish benign deviation from malicious manipulation
- **Lack of Statistical Interpretability**: Existing methods rely on heuristics, with no quantitative or explainable measure of semantic inconsistency

Motivation:

Frozen-backbone FL improves efficiency but sacrifices robustness and transparency. We need a **new defense paradigm** that:

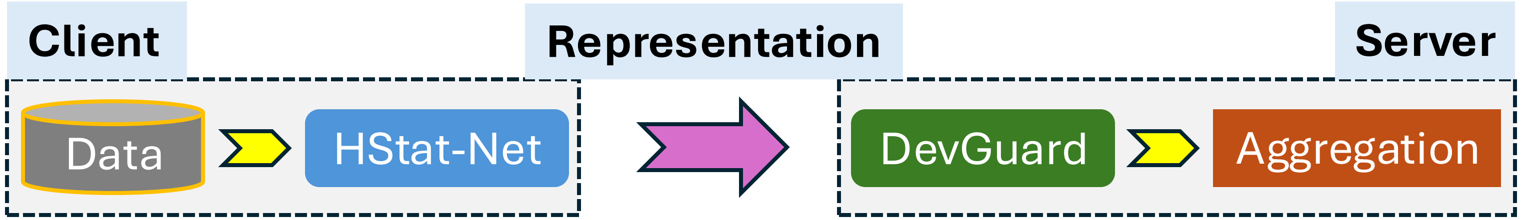
- Works without gradients
- Evaluates clients by semantic behavior
- And ensures statistical interpretability

→ This motivates ***FedRACE***, a framework combining hierarchical representation learning and statistical deviance analysis

Overview of FedRACE

❑ Goal:

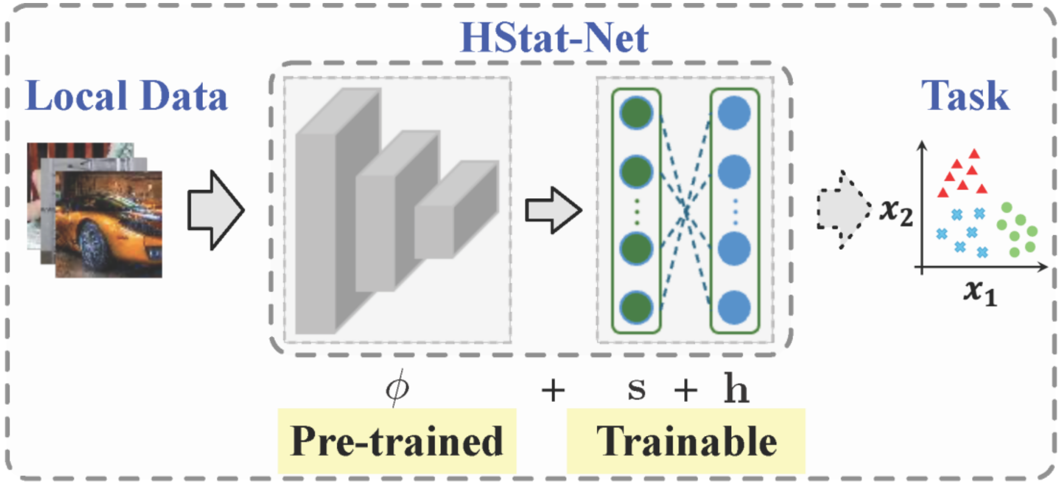
- Enable robust and interpretable federated learning under frozen-backbone settings
- Detect malicious clients and maintain global consistency without gradient information
- Bridge representation learning and statistical inference for explainable robustness



❑ Core Components：

- **Hierarchical Statistical Network (HStat-Net):** Transforms frozen features into structured, low-dimensional embeddings, enhancing class separability and enabling semantic-level comparison across clients
- **Deviance-based Guard Mechanism (DevGuard):** Models each client’s head as a GLM, measures semantic deviation from the global distribution via statistical deviance, and detects abnormal clients using an adaptive, theoretically grounded threshold

3 Hierarchical Statistical Network



☐ Architecture:

$$\emptyset(x) \rightarrow s(\emptyset(x)) \rightarrow h\Big(s(\emptyset(x))\Big)$$

- ∅: Frozen feature extractor (e.g., CLIP)
- $s\,(\cdot)$: Statistical projection layer
- $h\,(\cdot)$: Lightweight task head

☐ Two-phase Optimization:

- **Phase 1:** Fix $s\,(\cdot) \rightarrow$ train h with cross-entropy loss (task alignment)
- **Phase 2:** Fix $h\,(\cdot)$:$\rightarrow$ train s with triplet loss (structural compactness)

→ Builds a linearly separable, statistically stable representation space, enabling semantic-level comparison and robust aggregation

☐ Validation: The hierarchical features become linearly separable and semantically stable, enabling effective statistical evaluation in DevGuard design

Method	Raw	CLIP	HStat-Net
Fisher	0.149	0.480	**1.602**
MI	0.162	0.275	**0.556**

4 Deviance-based Guard Mechanism

☐ Core Idea:

- Model each client's head h_i as a Generalized Linear Model
- Compute deviance residuals Δ_i from predictions on global class representations
- Higher Δ_i indicates stronger semantic deviation from the global consensus

☐ Formulation:

$$\Delta_i = \sum_c (-2 \cdot \log \hat{\boldsymbol{y}}_i^c) \log(-2 \cdot \log \hat{\boldsymbol{y}}_i^c)$$

where $\hat{\boldsymbol{y}}_i^c$ is the predicted probability for class c. Clients are ranked by Δ_i; large values imply inconsistency.

☐ Thresholding & Voting:

- **Sort** residuals $\Delta_{[1]} \leq \Delta_{[2]} \leq \cdots \leq \Delta_{[n]}$
- For each candidate index p, **estimate** benign/malicious (μ_B, μ_M) and σ_p^2
- **Choose** $\hat{p}$ to minimize the upper bound of total misclassification rate
- **Repeat** for K random subsets; clients flagged in $> \frac{K}{2}$ steps are marked malicious

5 Experimental Results

NEURAL INFORMATION PROCESSING SYSTEMS 2025

Dataset	Defense	Untargeted		Targeted					
		Min-Max	IPMA	TLFA		ECBA		DBA	
		ACC	ACC	ASR	ACC	BA	ACC	BA	ACC
CIFAR-100	Multi-krum	$72.59_{0.27}$	$76.16_{0.32}$	$1.52_{0.10}$	$75.93_{0.28}$	$20.05_{0.11}$	$76.03_{0.31}$	$23.20_{0.28}$	$75.68_{0.27}$
	Trimmed-mean	$75.15_{0.35}$	$76.43_{0.27}$	$1.79_{0.25}$	$75.83_{0.24}$	$10.34_{0.26}$	$76.53_{0.26}$	$12.16_{0.29}$	$76.65_{0.26}$
	FLAIR	$73.07_{0.29}$	$75.74_{0.27}$	$0.61_{0.16}$	$74.49_{0.30}$	$1.30_{0.23}$	$76.21_{0.32}$	$0.96_{0.17}$	$75.65_{0.28}$
	FedRoLA	$76.05_{0.33}$	$76.84_{0.28}$	$11.92_{0.28}$	$74.88_{0.29}$	$39.28_{0.28}$	$76.47_{0.30}$	$2.89_{0.28}$	$77.04_{0.27}$
	FLShield	$76.86_{0.24}$	$76.66_{0.25}$	$2.27_{0.29}$	$75.63_{0.28}$	$1.67_{0.28}$	$76.81_{0.27}$	$1.46_{0.27}$	$76.99_{0.31}$
	FEDRACE	$\mathbf{76.69}_{0.32}$	$\mathbf{76.99}_{0.32}$	$\mathbf{0.07}_{0.10}$	$\mathbf{77.02}_{0.33}$	$\mathbf{0.06}_{0.11}$	$\mathbf{76.98}_{0.31}$	$\mathbf{0.36}_{0.23}$	$\mathbf{77.21}_{0.31}$
Food-101	Multi-krum	$52.31_{0.33}$	$55.70_{0.27}$	$2.07_{0.13}$	$55.85_{0.27}$	$20.22_{0.13}$	$55.87_{0.28}$	$49.13_{0.30}$	$55.23_{0.29}$
	Trimmed-mean	$54.37_{0.31}$	$56.37_{0.31}$	$2.34_{0.26}$	$56.08_{0.28}$	$27.58_{0.29}$	$56.22_{0.32}$	$30.84_{0.29}$	$56.54_{0.29}$
	FLAIR	$53.16_{0.30}$	$54.27_{0.30}$	$0.43_{0.15}$	$52.09_{0.29}$	$5.67_{0.30}$	$55.24_{0.29}$	$1.48_{0.25}$	$53.33_{0.29}$
	FedRoLA	$56.40_{0.29}$	$55.59_{0.29}$	$12.74_{0.29}$	$54.10_{0.29}$	$45.27_{0.26}$	$56.16_{0.31}$	$8.14_{0.28}$	$56.51_{0.28}$
	FLShield	$56.24_{0.29}$	$56.07_{0.31}$	$14.02_{0.32}$	$54.76_{0.30}$	$6.36_{0.29}$	$56.25_{0.31}$	$1.44_{0.28}$	$56.65_{0.27}$
	FEDRACE	$\mathbf{56.38}_{0.27}$	$\mathbf{56.76}_{0.26}$	$\mathbf{0.27}_{0.16}$	$\mathbf{56.68}_{0.27}$	$\mathbf{0.31}_{0.16}$	$\mathbf{56.70}_{0.26}$	$\mathbf{1.01}_{0.31}$	$\mathbf{56.72}_{0.27}$
Tiny ImageNet	Multi-krum	$71.04_{0.32}$	$72.38_{0.28}$	$0.63_{0.10}$	$72.70_{0.27}$	$19.27_{0.12}$	$72.85_{0.27}$	$45.71_{0.29}$	$72.05_{0.28}$
	Trimmed-mean	$71.95_{0.28}$	$72.44_{0.29}$	$0.95_{0.22}$	$72.74_{0.28}$	$33.06_{0.28}$	$72.33_{0.30}$	$35.09_{0.23}$	$72.67_{0.25}$
	FLAIR	$71.23_{0.35}$	$72.59_{0.28}$	$0.28_{0.19}$	$70.58_{0.28}$	$4.43_{0.28}$	$71.89_{0.28}$	$0.24_{0.15}$	$70.91_{0.30}$
	FedRoLA	$73.36_{0.21}$	$72.78_{0.29}$	$4.87_{0.27}$	$71.92_{0.29}$	$47.14_{0.28}$	$72.73_{0.25}$	$4.75_{0.28}$	$73.13_{0.21}$
	FLShield	$73.29_{0.24}$	$73.19_{0.32}$	$9.85_{0.28}$	$71.84_{0.29}$	$5.84_{0.28}$	$73.11_{0.28}$	$0.53_{0.19}$	$73.21_{0.32}$
	FEDRACE	$\mathbf{73.06}_{0.29}$	$\mathbf{73.40}_{0.29}$	$\mathbf{0.07}_{0.10}$	$\mathbf{73.24}_{0.31}$	$\mathbf{0.08}_{0.10}$	$\mathbf{73.44}_{0.29}$	$\mathbf{0.13}_{0.13}$	$\mathbf{73.42}_{0.29}$

- **Across all datasets and attack types, FEDRACE achieves the best performance:**
  - Highest clean accuracy (e.g., 76–77% on CIFAR-100, 56–57% on Food-101)
  - Lowest attack success rates, even under severe targeted attacks
- **Competing methods show clear weaknesses:**
  - FLAIR and FedRoLA exhibit residual backdoor effects (ASR up to 40 %)
  - FLShield performs better but still trails FEDRACE, especially on complex datasets

→ FedRACE's advantage is **consistent** across both untargeted and targeted settings, indicating robust global model stability under frozen-backbone constraints.

Conclusions

NEURAL INFORMATION PROCESSING SYSTEMS 2025

FedRACE introduces a new defense paradigm for FL:

- ❑ Learns hierarchical statistical representations for semantic alignment
- ❑ Performs statistical deviance evaluation for reliable client assessment

- ❖ Works without gradient information and achieves **reliable robustness** across diverse datasets and attacks
- ❖ Offers **theoretical guarantees** on detection mechanism and demonstrates scalability in large federated systems

Thank You