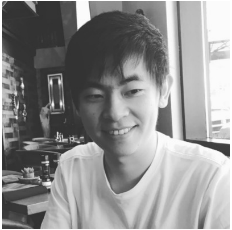
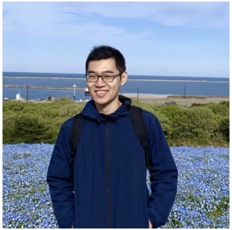
Tackling Data Heterogeneity in Federated Learning with Class Prototypes

Yutong Dai[1] Zeyuan Chen[2] Junnan Li[2]
Shelby Heinecke[2] Lichao Sun[1] Ran Xu[2]

Lehigh University[1] Salesforce Research[2]

Feburary, 2023

Washington, DC, USA

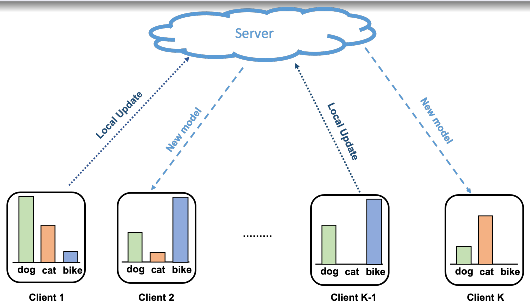





Outline

Outline

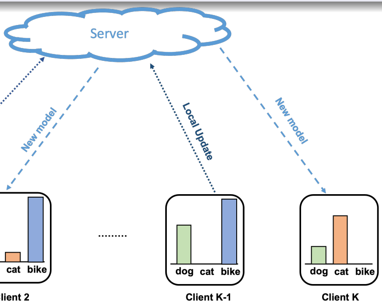
Introduction

Federated learning and analytics are a distributed approach for collaboratively learning models (or statistics) from **decentralized data**, motivated by and designed for **privacy** protection.



Introduction

Federated learning and analytics are a distributed approach for collaboratively learning models (or statistics) from **decentralized data**, motivated by and designed for **privacy** protection.



Address data heterogeneity with a particular focus on class imbalance.

Introduction (Cont'd)

Classic Problem Formulation

$$F^{\mathrm{ERM}}(\boldsymbol{x}) = \sum_{i=1}^{N} p_i F_i^{\mathrm{ERM}}(\boldsymbol{x}); \quad \text{where } F_i^{\mathrm{ERM}}(\boldsymbol{x}) = \frac{1}{|D_i|} \sum_{\in D_i} f_i(\boldsymbol{x}; \ ) \text{ and } \sum_{i=1}^{N} p_i = 1$$

- **Across clients**: Heterogeneous data distribution leads to inconsistent local objective functions, which imposes challenges into the optimization process.
- **Within a client**: Imbalanced data makes the local model likely to overfit **dominant classes**.

Introduction (Cont'd)

Classic Problem Formulation

$$F^{\mathrm{ERM}}(\boldsymbol{x}) = \sum_{i=1}^{N} p_i F_i^{\mathrm{ERM}}(\boldsymbol{x}); \quad \text{where } F_i^{\mathrm{ERM}}(\boldsymbol{x}) = \frac{1}{|D_i|} \sum_{\xi \in D_i} f_i(\boldsymbol{x}; \xi) \text{ and } \sum_{i=1}^{N} p_i = 1$$

- **Across clients**: Heterogeneous data distribution leads to inconsistent local objective functions, which imposes challenges into the optimization process.
- **Within a client**: Imbalanced data makes the local model likely to overfit **dominant classes**.

Personalized federated learning comes to the rescue.

Brief Literature Review

- Personalized federated learning
  - **goal**: tailor personalized models to client-speci c tasks;
  - **methods**: parameter decoupling, regularization, model interpolation, and more[1];
    - parameter decoupling: body (representation learning) + head (classification task).
- Class-imbalance learning
  - **data-level**: Over-sampling minority classes or under-sampling majority classes
  - **algorithm-level**:
    - sample-wise[2] or class-wise[3] class-balanced losses;
    - decoupling the training procedure into the representation learning and classification phases[4].

[1] Alysa Ziying Tan et al. "Towards personalized federated learning". In: *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[2] Tsung-Yi Lin et al. "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[3] Yin Cui et al. "Class-balanced loss based on effective number of samples". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 9268–9277.

[4] Bingyi Kang et al. "Decoupling Representation and Classifier for Long-Tailed Recognition". In: *International Conference on Learning Representations*. 2019.

Brief Literature Review

- Class-imbalance learning + FL
  - **CReFF**[1] adapts the idea of[2] into FL setting while observing the privacy.
  - **FedROD**[3] designs a two-head-one-body architecture, where one head is trained with class-balanced loss while the other head is trained with empirical loss.

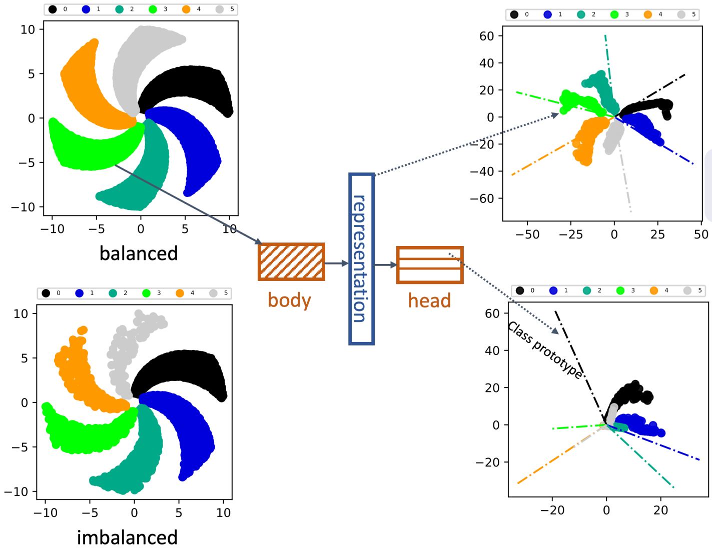
[1] Xinyi Shang et al. "Federated Learning on Heterogeneous and Long-Tailed Data via Classifier Re-Training with Federated Features". In: *arXiv preprint arXiv:2204.13399* (2022).

[2] Bingyi Kang et al. "Decoupling Representation and Classifier for Long-Tailed Recognition". In: *International Conference on Learning Representations*. 2019.

[3] Hong-You Chen and Wei-Lun Chao. "On Bridging Generic and Personalized Federated Learning for Image Classification". In: *International Conference on Learning Representations*. 2021.

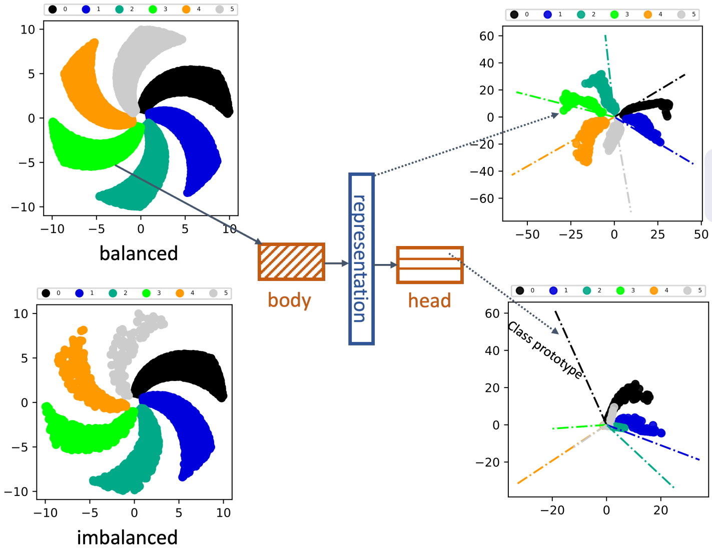
Outline

Centralized Training on A Toy Dataset



Visualization is over the same **balanced** testing dataset for a fully trained MLP.

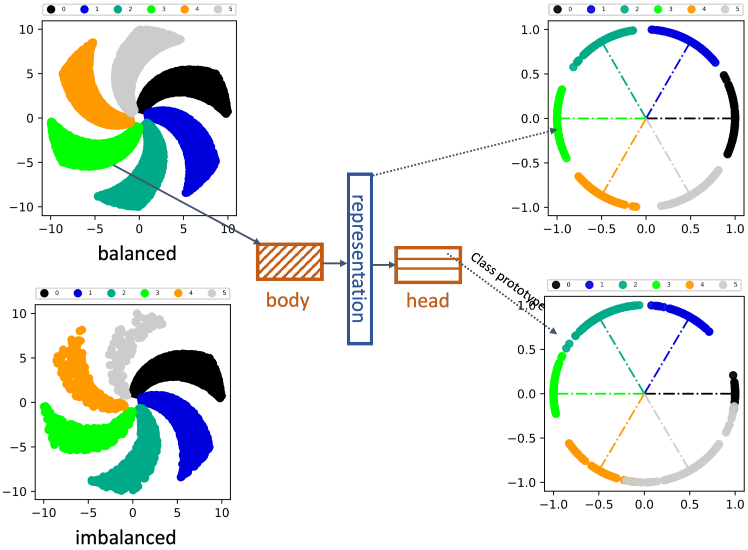
Centralized Training on A Toy Dataset



Visualization is over the same **balanced** testing dataset for a fully trained MLP.

- Balanced Dataset
  - uniformly distributed class prototypes
  - separated representations
- Imbalanced Dataset
  - crowded class prototypes
  - overlapped representations

Centralized Training on A Toy Dataset (Cont'd)



- no harm on the balanced dataset
- rescue the minority classes

FL with A Toy Dataset

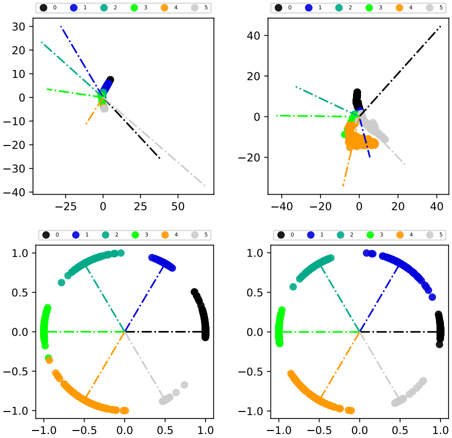


Figure: Top: Visualization of representations and prototypes on two of FedAvg clients. Bottom: Visualization of representations and prototypes (xed at prior and never updated) on the same two of FedAvg clients.

Uniformity in Class Prototypes

Initialization Strategy

$$\max_{\{w_1, \cdots, w_{|C|}, M\}} M^2 \quad (1)$$

$$\text{s.t. } \|w_i - w_j\|^2 \geq M^2, \|w_i\|^2 = 1 \text{ for all } i \in [|C|], i \neq j.$$

- Can be solved with any constrained optimization solver, e.g., interior point method.
- Need only to be solved once.
- The solution can be approximated with an orthonormal base, which is similar to FedBABU[4].

[4] Jaehoon Oh, SangMook Kim, and Se-Young Yun. "FedBABU: Toward Enhanced Representation for Federated Image Classification". In: *International Conference on Learning Representations*. 2021.

Uniformity in Class Prototypes

Initialization Strategy

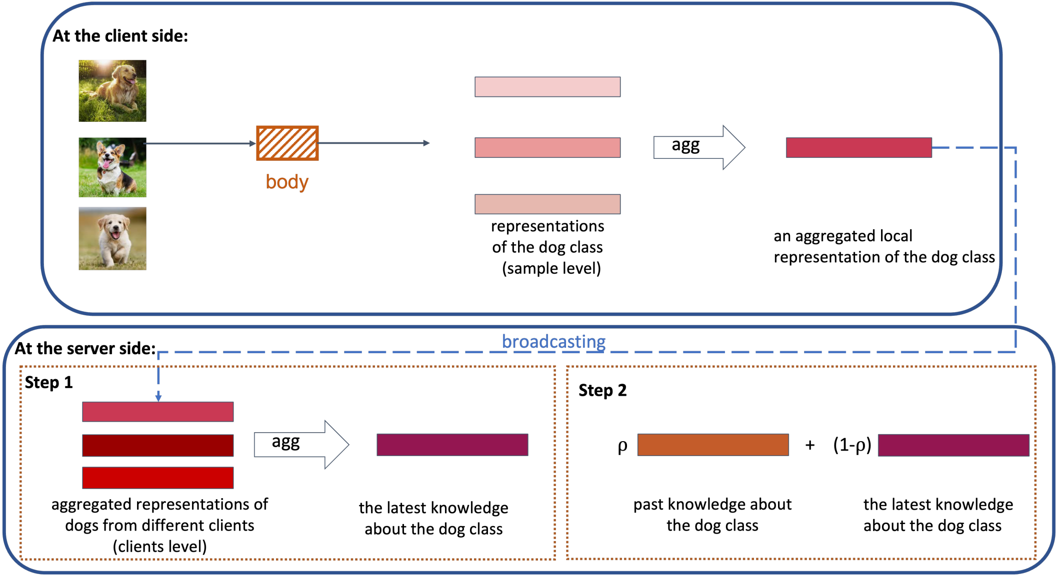
$$\max_{\{w_1,\dots,w_{|C|},M\}} M^2 \tag{1}$$
$$\text{s.t. } \|w_i - w_j\|^2 \geq M^2, \|w_i\|^2 = 1 \text{ for all } i \in [|C|], i \neq j.$$

- Can be solved with any constrained optimization solver, e.g., interior point method.
- Need only to be solved once.
- The solution can be approximated with an orthonormal base, which is similar to FedBABU[4].

However, this is not enough...

[4] Jaehoon Oh, SangMook Kim, and Se-Young Yun. "FedBABU: Toward Enhanced Representation for Federated Image Classification". In: *International Conference on Learning Representations*. 2021.

Infuse Class Semantics



Algorithm

Algorithm FedNH - Skeleton

Initialization: the body θ; the head $W \in \mathbb{R}^{|C| \times d}$;
for $t = 0, \dots, R-1$ communication rounds **do**
    Select a subset of clients S^t.
    ...
    for each selected client $k \in S^t$ **in parallel do**
        $\theta_k^{t+1}, \mu_k^{t+1} \leftarrow$ ClientUpdate(θ^t, W^t). [μ_k^{t+1} is aggregated representations of classes of the local training dataset.]
    end for
    Head Update: $W_c^{t+1} \leftarrow \rho W_c^t + (1-\rho) \sum_{k \in S^t} \alpha_k^{t+1} \mu_{k,c}^{t+1}$ for all $c \in C$.
    ...
    Body Update: $\theta^{t+1} = \frac{1}{|S^t|} \sum_{k \in S^t} \theta_k^t$.
    ...
end for

Convergence Result

Theorem 1 (Informal)

Let the kth client uniformly at random returns an element from $f\ {}^{t;j}_{k} g$ *as the solution, denoted as* $_k$. *Further, let* W *share the same round index as* $_k$. *Then for any* > 0, *set* $2\ (\ _1(\ ; M_G; M_f); 1)$ *and* $2\ \ 0;\ _2(\ ; L_g;\ ^2;\ ; M_G; M_f)$, *if* $R > O(\ ^1)$, *one gets*

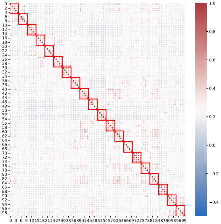
$$\mathrm{E}\ \overset{\mathrm{h}}{}\ k r\ \ F_k(\ _k; W\ ) k^2 \overset{\mathrm{i}}{}\ \ \ ;$$

where $_1(\ ; M_G; M_f)$, $_2(\ ; L_g;\ ^2;\ ; M_G; M_f)$, M_G, *and* M_f *are some positive constants.*

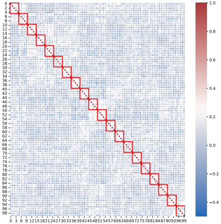
Outline

Learned Class Semantics

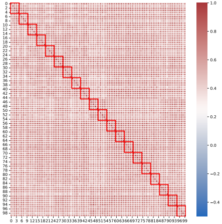
Visualize the pair-wise cosine similarity of class prototypes on Cifar100.



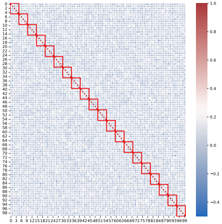
(a) Our method



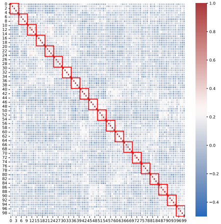
(b) FedAvg



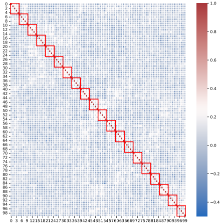
(c) FedProto



(d) FedBABU



(e) FedRod



(f) CReFF

- Cifar100 has 20 super-classes.
- Each block along the diagonal contains 5 sub-classes within one super-class.

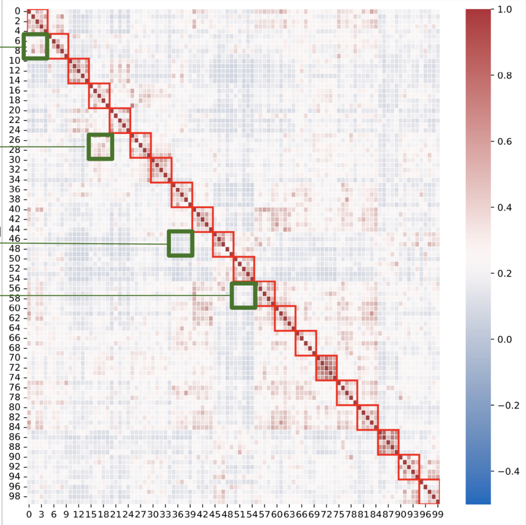
Learned Class Semantics (Cont'd)

Similarity among super-class
{beaver, dolphin, otter, seal, whale} and
{aquarium fish, flatfish, ray, shark, trout}

Similarity among super-class
{bottle, bowl, can, cup, plate} and
{clock, keyboard, lamp, telephone, television}

Similarity among super-class
{bee, beetle, butterfly, caterpillar, cockroach} and
{bridge, castle, house, road, skyscraper}

Similarity among super-class
{cloud, forest, mountain, plain, sea} and
{camel, cattle, chimpanzee, elephant, kangaroo}



Classification Accuracy

Dataset	Method	Dir(0.3)			Dir(1.0)		
		GM	PM(V)	PM(L)	GM	PM(V)	PM(L)
Cifar100	Local	—	13.63 ± 2.45	30.89 ± 1.82	—	9.44 ± 1.27	16.71 ± 1.03
	FedAvg	35.14 ± 0.48	31.85 ± 1.33	50.77 ± 2.31	36.07 ± 0.41	28.86 ± 1.23	38.35 ± 2.11
	FedPer	15.04 ± 0.06	16.15 ± 2.34	33.10 ± 1.50	14.69 ± 0.03	11.61 ± 2.17	19.08 ± 1.36
	Ditto	35.14 ± 0.48	26.19 ± 1.11	45.91 ± 2.17	36.07 ± 0.41	22.92 ± 1.77	32.81 ± 2.16
	FedRep	5.42 ± 0.03	13.59 ± 2.31	29.45 ± 2.45	6.37 ± 0.04	9.47 ± 2.27	16.07 ± 1.27
	FedProto	—	10.64 ± 1.02	19.11 ± 1.75	—	9.24 ± 1.33	12.61 ± 1.78
	CReFF	22.90 ± 0.30	31.85 ± 1.33	50.77 ± 2.31	22.21 ± 0.15	28.86 ± 1.23	38.35 ± 2.11
	FedBABU	32.41 ± 0.40	28.96 ± 2.16	47.86 ± 1.03	32.34 ± 0.49	25.84 ± 1.44	34.85 ± 1.80
	FedROD	33.83 ± 0.25	28.53 ± 1.27	42.93 ± 1.03	35.20 ± 0.19	27.58 ± 1.98	33.44 ± 1.76
	FedNH	**41.34 ± 0.25**	**38.25 ± 1.23**	**55.21 ± 2.11**	**43.19 ± 0.24**	**36.88 ± 1.15**	**45.46 ± 2.14**

Metric: the accuracy of the ith personalized model is computed as

$$\mathrm{acc}_i = \frac{\sum_{(x_j,y_j)\; D^{\text{test}}} {}_i(y_j)\,\mathbf{1}(y_j = \hat{y}_j)}{\sum_{(x_j,y_j)\; D^{\text{test}}} {}_i(y_j)}.$$

- D^{test} is a unified and **balanced** dataset.
- **PM(L)**: ${}_i(y)$ to 1 if the class y appears in ith client's training dataset and 0 otherwise.
- **PM(V)**: ${}_i(y) = \mathrm{P}_i(y = c)$, the probability of the sample y is from class c in the ith client.
- $\hat{y}$ is the predicted value and $\mathbf{1}(\ )$ is the indicator function.

Summary

- We proposed FedNH to address the data heterogeneity with class imbalance. FedNH combines the uniformity and semantics of class prototypes to learn high-quality representations for classification.

- Our idea currently only applies to the classification task, and the inductive bias from uniformity and semantics of class prototypes can only be imposed on the head of neural network architecture.

Q&A

Thank you and Questions?

Contact: yud319@lehigh.edu

arXiv