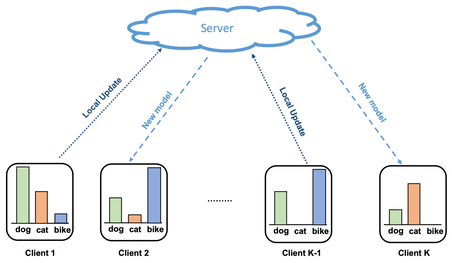
Tackling Data Heterogeneity in Federated Learning with Class Prototypes

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

Lehigh University[1] Salesforce Research[2]

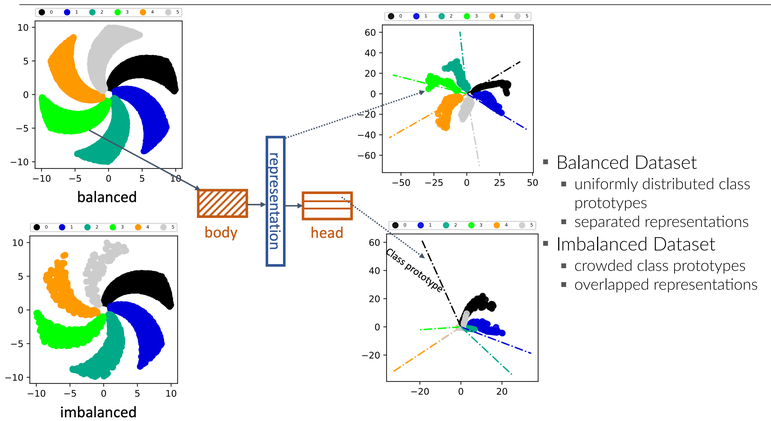
arXiv

Introduction



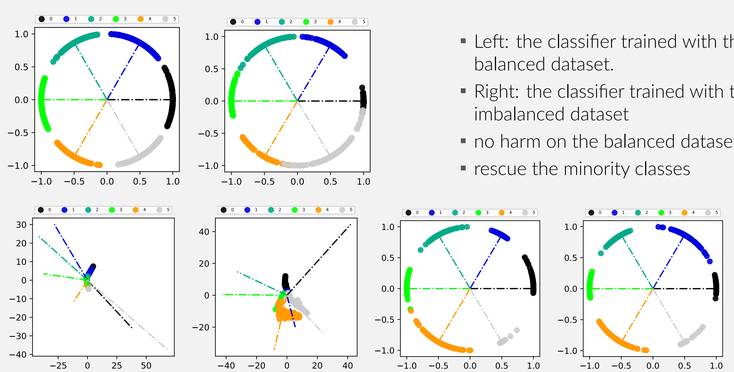
- Among clients: different distributions;
- Within a client: class imbalance;
- Personalized models have biased performances on dominant classes;
- Few works address both date heterogeneity and class imbalance without requiring auxiliary datasets or potential privacy leakage.

A Motivating Example



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

What if we enforce uniformity in the class prototypes?



- Left: the classifier trained with the balanced dataset.
- Right: the classifier trained with the imbalanced dataset
- no harm on the balanced dataset
- rescue the minority classes

Top: FedAvg clients. Bottom: FedAvg with a fixed head. Learned representations are consistent for different clients for FedAvg with the uniform head compared with vanilla FedAvg.

Methodology

Uniformity in class prototypes: an initialization strategy

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

Infuse class semantics

Algorithm 1 FedNH - Skeleton

Initialization: the body θ; the head $W \in \mathbb{R}^{|\mathcal{C}| \times d}$;
for $t = 0, \ldots, R-1$ communication rounds **do**
  Select a subset of clients S^t.
  ...
  for each selected client $k \in S^t$ **in parallel do**
    $\left(\theta_k^{t+1}, \mu_k^{t+1}\right) \leftarrow \text{ClientUpdate}(\theta^t, W^t)$.
  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 $c \in \mathcal{C}$.
  ...
  Body Update: $\theta^{t+1} = \frac{1}{|S^t|} \sum_{k \in S^t} \theta_k^t$.
  ...
end for

A Convergence Result

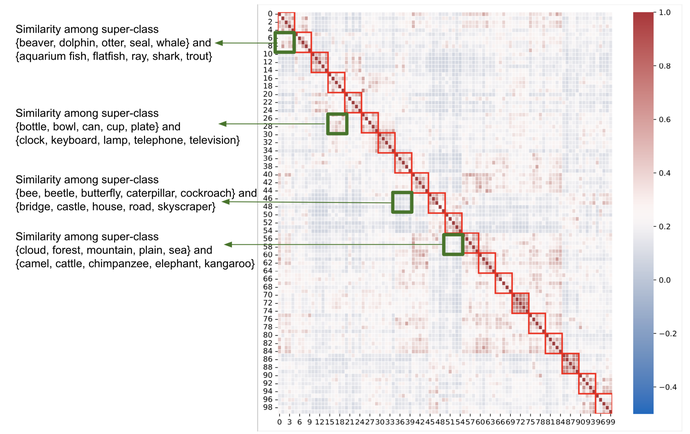
Let the kth client uniformly at random returns an element from $\{\theta_k^{t,j}\}$ as the solution, denoted as θ_k^*. Further, let W^* share the same round index as θ_k^*. Then for any $\epsilon > 0$, set $\rho \in (\nu_1(\epsilon, M_G, M_f), 1)$ and $\eta \in \left(0, \nu_2(\epsilon, L_g, \sigma^2, \rho, M_G, M_f)\right)$, if $R > \mathcal{O}(\epsilon^{-1})$, one gets

$$\mathbb{E}\left[\left\|\nabla_\theta F_k(\theta_k^*; W^*)\right\|^2\right] \leq \epsilon,$$

where $\nu_1(\epsilon, M_G, M_f)$, $\nu_2(\epsilon, L_g, \sigma^2, \rho, M_G, M_f)$, M_G, and M_f are some positive constants.

Experiments

Learned class semantics: Visualize the pair-wise cosine similarity of class prototypes on Cifar100. Similar plots for other methods are to be found in our paper.



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

$$\text{acc}_i = \frac{\sum_{(x_j,y_j)\sim\mathcal{D}^{\text{test}}} \alpha_i\left(y_j\right) \mathbf{1}\left(y_j = \hat{y}_j\right)}{\sum_{(x_j,y_j)\sim\mathcal{D}^{\text{test}}} \alpha_i\left(y_j\right)}.$$

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

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**

References

[1] H.-Y. Chen and W.-L. Chao, "On bridging generic and personalized federated learning for image classification," in *International Conference on Learning Representations*, 2021.

[2] J. Oh, S. Kim, and S.-Y. Yun, "Fedbabu: Toward enhanced representation for federated image classification," in *International Conference on Learning Representations*, 2021.

[3] X. Shang, Y. Lu, G. Huang, and H. Wang, "Federated learning on heterogeneous and long-tailed data via classifier re-training with federated features," *arXiv preprint arXiv:2204.13399*, 2022.

[4] T. Li, P. Cao, Y. Yuan, L. Fan, Y. Yang, R. S. Feris, P. Indyk, and D. Katabi, "Targeted supervised contrastive learning for long-tailed recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6918–6928, 2022.