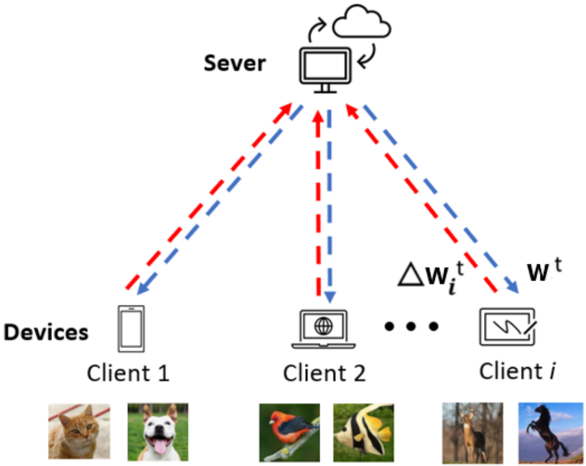
Heterogeneity-Guided Client Sampling: Towards Fast and Efficient Non-IID Federated Learning

Presenter: Huancheng Chen
University of Texas at Austin

Nov, 2024

Federated Learning

Federated learning (FL), a communication-efficient and privacy-preserving alternative to training on centrally aggregated data, relies on collaboration between clients devices.



Client Sampling

Prior Works DivFL (Balakrishnan et al., 2022), CS (Fraboni et al., 2021): aim to select clients such that the resulting model update is an unbiased estimate of the true update while minimizing the variance

$$\min_{\mathcal{S}^{(t)}} \left\| \frac{1}{N} \sum_{k=1}^{N} \nabla F_k(\mathbf{W}^{(t)}) - \frac{1}{K} \sum_{k \in \mathcal{S}^{(t)}} \nabla F_k(\mathbf{W}^{(t)}) \right\|_2^2 .$$

Assumption (Bounded Dissimilarity under Data Heterogeneity)

Gradient $\nabla F_k(\mathbf{W}^{(t)})$ of the k-th local model at global round t is such that

$$\left\| \nabla F_k(\mathbf{W}^{(t)}) - \nabla F(\mathbf{W}^{(t)}) \right\|^2 \leq \kappa - \rho e^{\beta\left(H(\mathcal{D}^{(k)}) - H(\mathcal{D}_0)\right)} = \sigma_k^2,$$

where $\mathcal{D}^{(k)}$ is the data label distribution of client k, $\mathcal{D}_0$ denotes uniform distribution, $H(\cdot)$ is Shannon's entropy of a stochastic vector, and $\beta > 0, \kappa > \rho > 0$.

Convergence Rate Relying on Selected Clients

With Assumptions:

- $F_k(\cdot)$ is L-smooth;
- $g_k(\mathbf{W}^{(t)})$ is unbiased and the variance is bounded by σ^2;
- Bounded Dissimilarity under Data Heterogeneity

Theorem

Let η and R be the learning rate and the number of local epochs, respectively. If the learning rate is such that $\eta \leq \frac{1}{8LR}$, $R \geq 2$, then

$$\min_{t \in [\mathcal{T}]} \left\| \nabla F(\mathbf{W}^{(t)}) \right\|^2 \leq \frac{1}{\mathcal{T}} \left(\frac{F(\mathbf{W}^{(0)}) - F(\mathbf{W}^*)}{\mathcal{A}_1} + \mathcal{A}_2 \sum_{t=0}^{\mathcal{T}-1} \sum_{k=1}^{N} \omega_k^t \sigma_k^2 \right) + \mathbf{\Phi},$$

where $\mathcal{A}_1$, $\mathcal{A}_2$, $\mathbf{\Phi}$ are positive constants, and ω_k^t is the probability of sampling client k at round t.

Estimating local data heterogeneity

We can obtain the correlation between $\Delta \mathbf{b}_j$ and the label distribution.

$$\mathbb{E}\left[\Delta \mathbf{b}_j\right] = \eta \left(D_j \sum_{c=1}^{C} \mathcal{S}_c - \mathcal{S}_j \right),$$

where D_j is the proportion of samples with label j in the training batch.

We define a proxy function to estimate the data heterogeneity of client k,

$$\hat{H}(\Delta \mathbf{b}^{(k)}) \triangleq H(\mathrm{softmax}(\Delta \mathbf{b}^{(k)}, \tau)),$$

where $H(\cdot)$ is Shannon's entropy and τ is a temperature parameters.

Heterogeneity-guided Clustering

Clustering: the server performs clustering algorithm to group clients with varying level of heterogeneity based on the distance

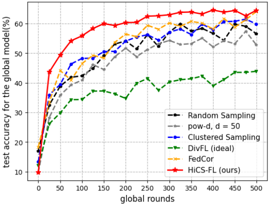
$$\textbf{Distance}(u,k) = \arccos\left(\frac{\Delta\mathbf{b}^{(u)} \cdot \Delta\mathbf{b}^{(k)}}{|\Delta\mathbf{b}^{(u)}| \cdot |\Delta\mathbf{b}^{(k)}|}\right) + \lambda \left|\hat{H}\Delta\mathbf{b}^{(u)}) - \hat{H}(\Delta\mathbf{b}^{(k)})\right|.$$

Hierarchical Sampling: sample clients from each cluster with probability based on the average value $\bar{H}_m$ across all clients in the cluster

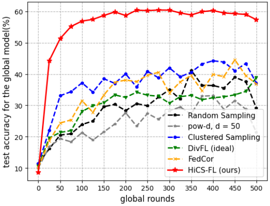
$$\pi^t = \left[\frac{\exp(\gamma^t \bar{H}_1^t)}{\sum_{m=1}^M \exp(\gamma^t \bar{H}_m^t)}, \ldots, \frac{\exp(\gamma^t \bar{H}_M^t)}{\sum_{m=1}^M \exp(\gamma^t \bar{H}_m^t)}\right], \tag{1}$$

where γ^t is a hyper-parameter.

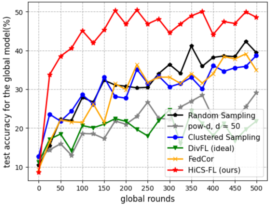
Part of the results on CIFAR100 dataset:



(a) mild heterogeneity



(b) medium heterogeneity



(c) severe heterogeneity

Observation:

- HiCS-FL (ours) improves the converged accuracy under medium and severe heterogeneity.

Experimental Results

Schemes	FMNIST		CIFAR10		Mini-ImageNet	
	acc $= 0.75$	speedup	acc $= 0.6$	speedup	acc $= 0.5$	speedup
Random	149	$1.0\times$	898	$1.0\times$	191	$1.0\times$
pow-d	79	$1.8\uparrow$	1037	$0.9\downarrow$	432	$0.4\downarrow$
CS	114	$1.3\uparrow$	748	$1.2\uparrow$	186	$1.0\times$
DivFL	478	$0.3\downarrow$	1417	$0.6\downarrow$	726	$0.3\downarrow$
FedCor	88	$1.7\uparrow$	711	$1.3\uparrow$	229	$0.8\uparrow$
HiCS-FL	**60**	**2.5**$\uparrow$	**123**	**7.3**$\uparrow$	**86**	**2.2**$\uparrow$

Observation:

- HiCS-FL (ours) can accelerate the convergence by at most 7.3 times.

References I

Balakrishnan, R., Li, T., et al. (2022). Diverse client selection for federated learning via submodular maximization. In *International Conference on Learning Representations*.

Fraboni, Y., Vidal, R., et al. (2021). Clustered sampling: Low-variance and improved representativity for clients selection in federated learning. In *International Conference on Machine Learning*.