Personalized Federated Learning through Local Memorization

Othmane Marfoq[1, 2] Giovanni Neglia[1] Laetitia Kameni[2] Richard Vidal[2]

[1]Inria, Université Côte d'Azur [2]Accenture Labs

Paper Code

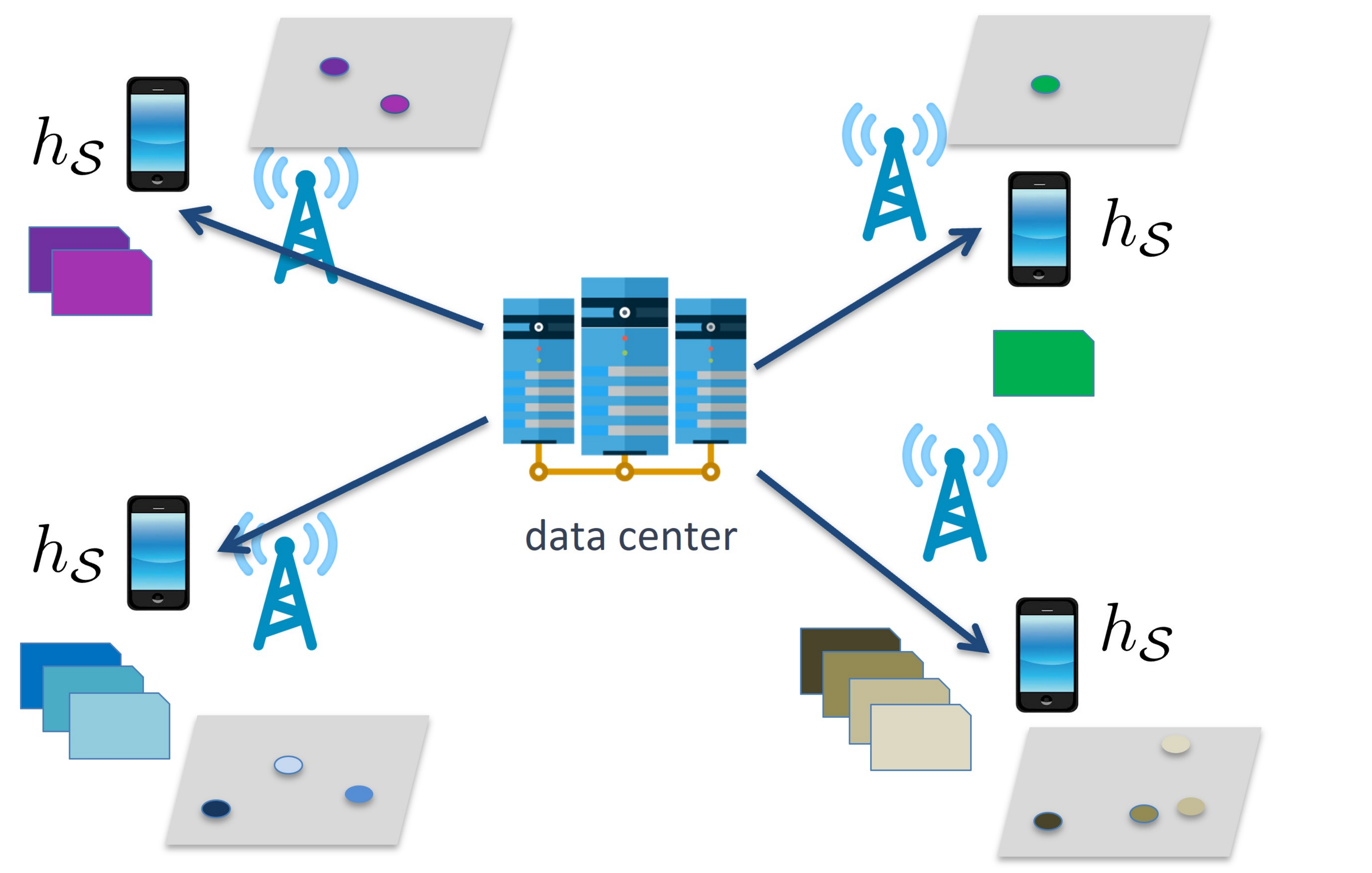
The problem

- We consider M classification (or regression) tasks, one for each client
- Data $\mathcal{S}_m = \{(\mathbf{x}_m^{(i)},\ y_m^{(i)})\}_{i=1}^{n_m}$ at client m is drawn from a local distribution $\mathcal{D}_m$ over $\mathcal{X} \times \mathcal{Y}$
- Client $m \in [M]$ wants to learn hypothesis $h_m \in \mathcal{H}_m$ mapping input $\mathbf{x} \in \mathcal{X}$ to a probability distribution over the set $\mathcal{Y}$:

$$\underset{h_m \in \mathcal{H}}{\operatorname{minimize}}\ \mathcal{L}_{\mathcal{D}_m}(h_m) \triangleq \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_m}\left[l\left(h_m\left(\mathbf{x}\right), y\right)\right]$$

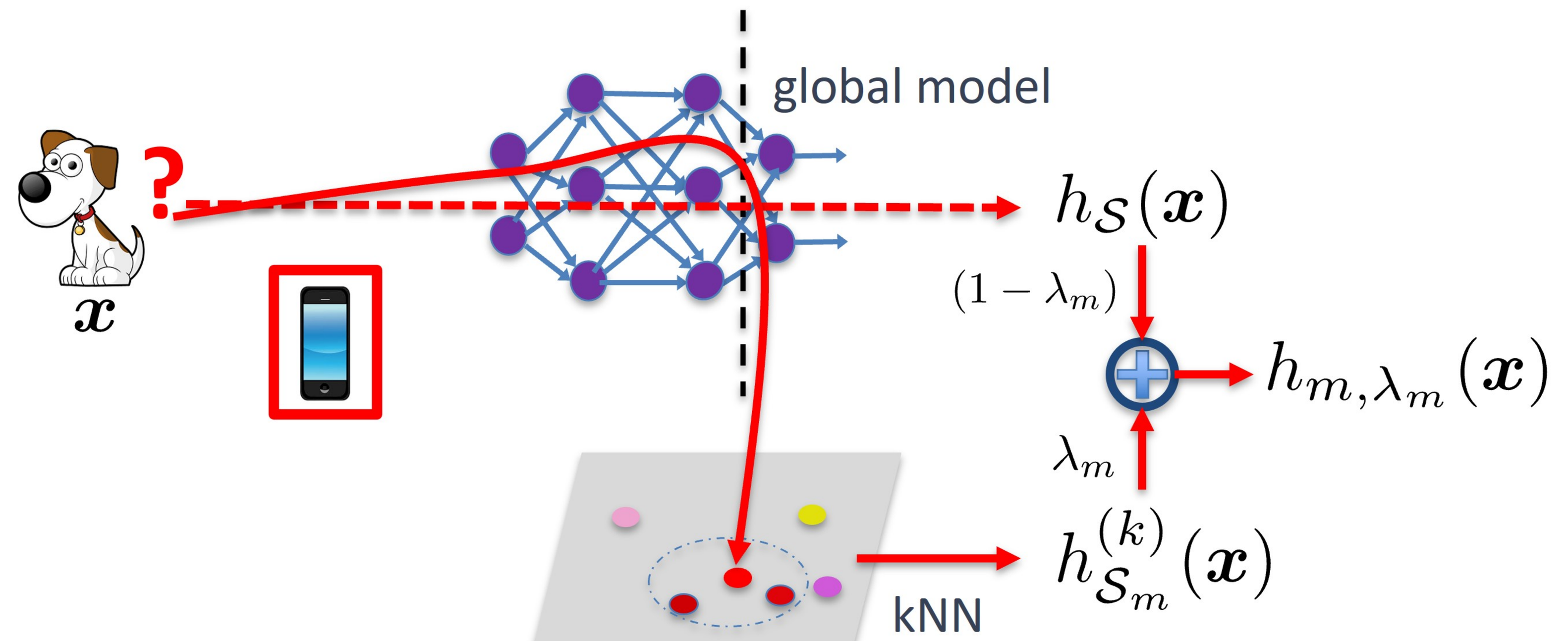
- `FedAvg` minimizes $\mathbb{E}_{(\mathbf{x},y)\sim\bar{\mathcal{D}}}\left[l\left(h(\mathbf{x}), y\right)\right]$, where $\bar{\mathcal{D}} = \sum_{m=1}^{M} \frac{n_m}{n} \cdot \mathcal{D}_t$ (asymptotically in the total number of samples)
- In many applications, e.g., language modeling, clients' local datasets differ both in size and distribution (*statistical heterogeneity*)
- Clients may differ in their storage and computational capabilities (*system heterogeneity*)

Our algorithm: `kNN-Per`



1. Clients train a global model $h_{\mathcal{S}}$ using a federated learning algorithm, e.g., `FedAvg`
2. Each client creates its local datastore for kNN inference (samples embedded through $h_{\mathcal{S}}$)
3. The global model and the local kNN are interpolated:

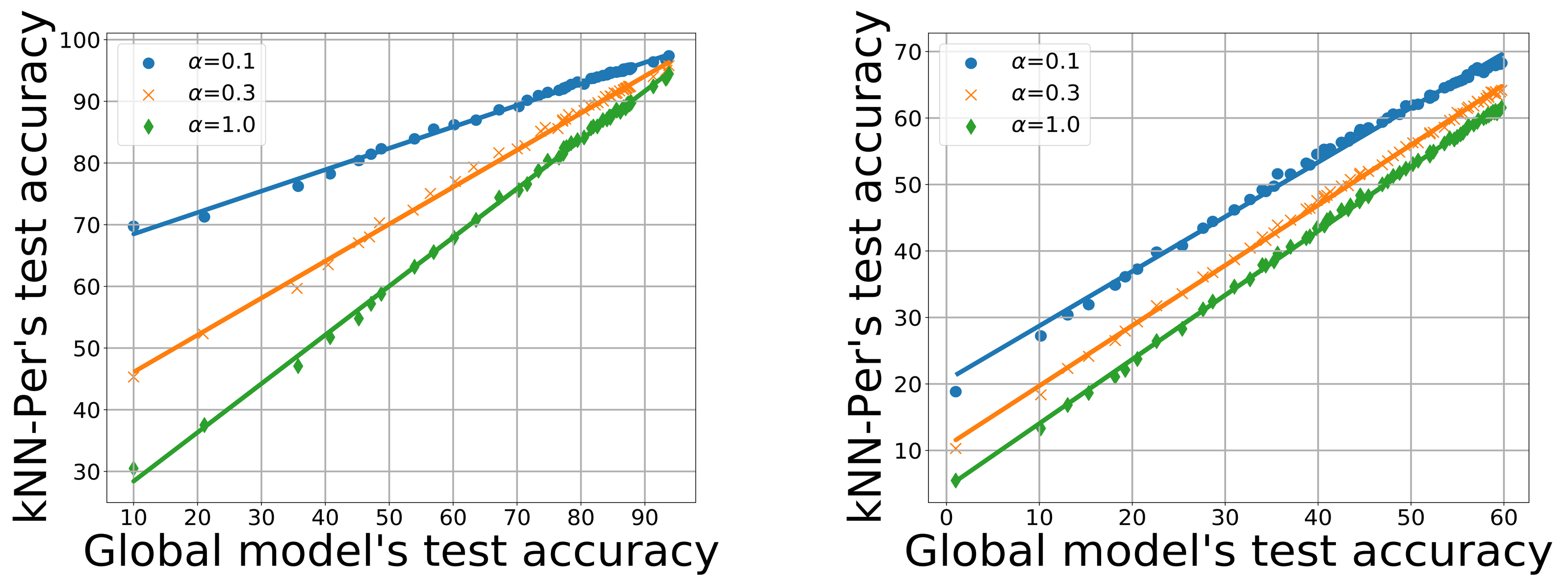
$$h_{m,\lambda_m}\left(\mathbf{x}\right) = \lambda_m \cdot h_{\mathcal{S}_m}^{(k)}\left(\mathbf{x}\right) + \left(1-\lambda_m\right) \cdot h_{\mathcal{S}}\left(\mathbf{x}\right)$$



Main assumption

Let $h_m^* \in \arg\min_{h\in\mathcal{H}} \mathcal{L}_{D_m}(h)$. There exist constants $\gamma_1, \gamma_2 > 0$, such that for any dataset $\mathcal{S}$ drawn from $\mathcal{X} \times \mathcal{Y}$ and any data points $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, we have

$$\underbrace{\left|\mathbb{P}\left[y=1|\mathbf{x}\right] - \mathbb{P}\left[y=1|\mathbf{x}'\right]\right|}_{\text{labels' distance}} \leq \underbrace{d\left(\phi_{h_{\mathcal{S}}}\left(\mathbf{x}\right), \phi_{h_{\mathcal{S}}}\left(\mathbf{x}'\right)\right)}_{\text{representations' distance}} \times \left(\gamma_1 + \gamma_2 \underbrace{\left(\mathcal{L}_{\mathcal{D}_m}\left(h_{\mathcal{S}}\right) - \mathcal{L}_{\mathcal{D}_m}\left(h_m^*\right)\right)}_{\text{global model's quality}}\right).$$



(a) CIFAR-10. (b) CIFAR-100.

Figure: Effect of the global model quality on the test accuracy of `kNN-Per` with λ tuned per client.

Generalization bound

Under proper assumptions, there exists $c \in \mathbb{R}$, such that

$$\underset{\mathcal{S}\sim\otimes_{m=1}^{M}\mathcal{D}_m^{n_m}}{\mathbb{E}}\left[\mathcal{L}_{\mathcal{D}_m}\left(h_{m,\lambda_m}\right)\right] \leq \left(1+\lambda_m\right)\mathcal{L}_{\mathcal{D}_m}\left(h_m^*\right) + c\left(1-\lambda_m\right)\operatorname{disc}_{\mathcal{H}}\left(\bar{\mathcal{D}}, \mathcal{D}_m\right) + \left(1-\lambda_m\right)\tilde{\mathcal{O}}\left(\sqrt{\frac{d_{\mathcal{H}}}{n}}\right)$$
$$+ \lambda_m\left(1+\operatorname{disc}_{\mathcal{H}}\left(\bar{\mathcal{D}}, \mathcal{D}_m\right)\right) \cdot \mathcal{O}\left(\frac{\sqrt{p}}{\sqrt[p+1]{n_m}}\right) + \lambda_m \cdot \tilde{\mathcal{O}}\left(\sqrt{\frac{d_{\mathcal{H}}}{n}} \cdot \frac{\sqrt{p}}{\sqrt[p+1]{n_m}}\right),$$

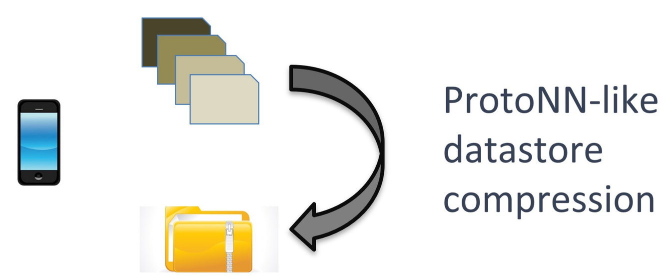
where $d_{\mathcal{H}}$ is the the VC dimension of the hypothesis class $\mathcal{H}$, $\bar{\mathcal{D}} = \sum_{m=1}^{M}\frac{n_m}{n} \cdot \mathcal{D}_m$ and $\operatorname{disc}_{\mathcal{H}}$ is the label discrepancy associated to the hypothesis class $\mathcal{H}$.

Average performance and fairness of personalized model

Dataset	Local	FedAvg	FedAvg+	ClusteredFL	Ditto	FedRep	APFL	kNN-Per (Ours)
FEMNIST	71.0 / 57.5	83.4 / 68.9	84.3 / 69.4	83.7 / 69.4	84.3 / 71.3	85.3 / 72.7	84.1 / 69.4	**88.2** / **78.8**
CIFAR-10	57.6 / 41.1	72.8 / 59.6	75.2 / 62.3	73.3 / 61.5	80.0 / 66.5	77.7 / 65.2	78.9 / 68.1	**83.0** / **71.4**
CIFAR-100	31.5 / 19.8	47.4 / 36.0	51.4 / 41.1	47.2 / 36.2	52.0 / 41.4	53.2 / 41.7	51.7 / 41.1	**55.0** / **43.6**
Shakespeare	32.0 / 16.0	48.1 / 43.1	47.0 / 42.2	46.7 / 41.4	47.9 / 42.6	47.2 / 42.3	45.9 / 42.4	**51.4** / **45.4**

Table: Test accuracy: average across clients / bottom decile.

Adding compression techniques



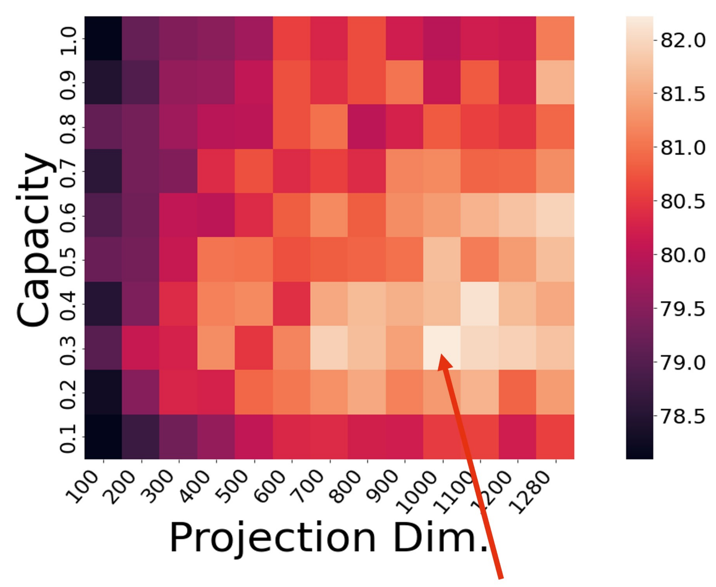
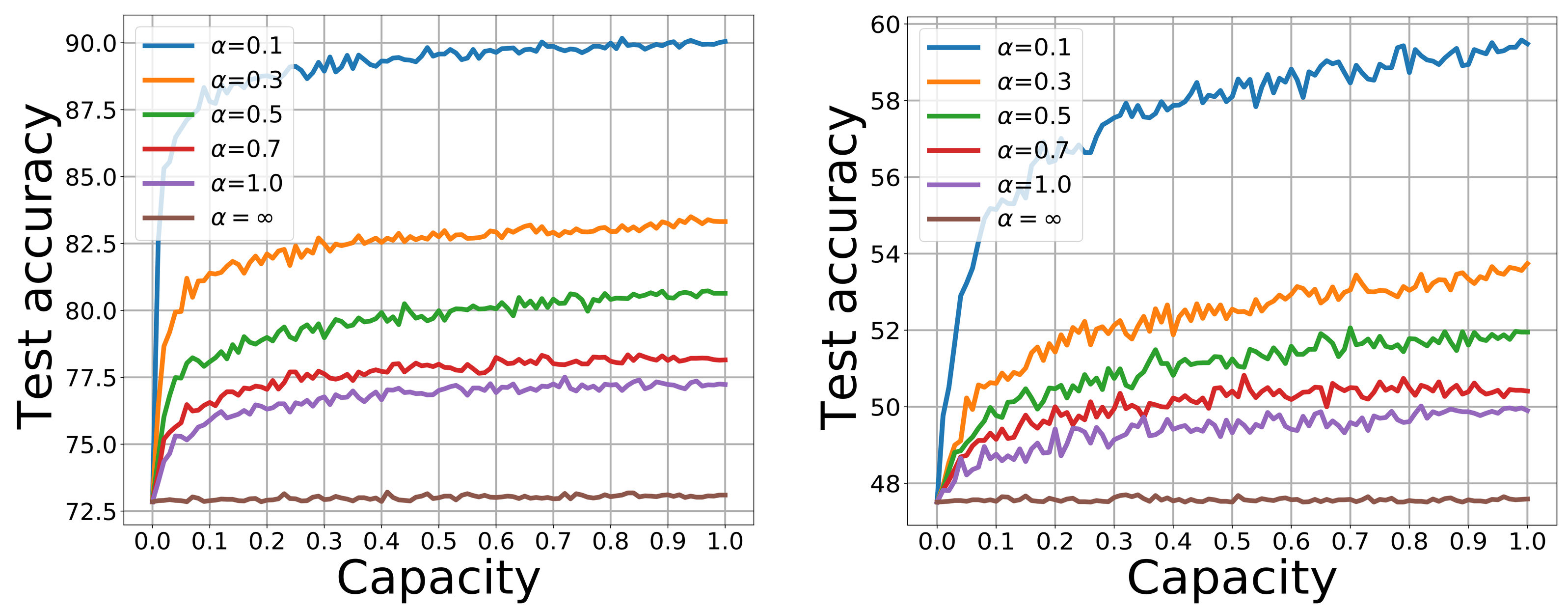


Figure: Test accuracy on CIFAR-10 dataset when the kNN mechanism is implemented through `ProtoNN` for different values of projection dimension and number of prototypes (expressed as a fraction of the local dataset).

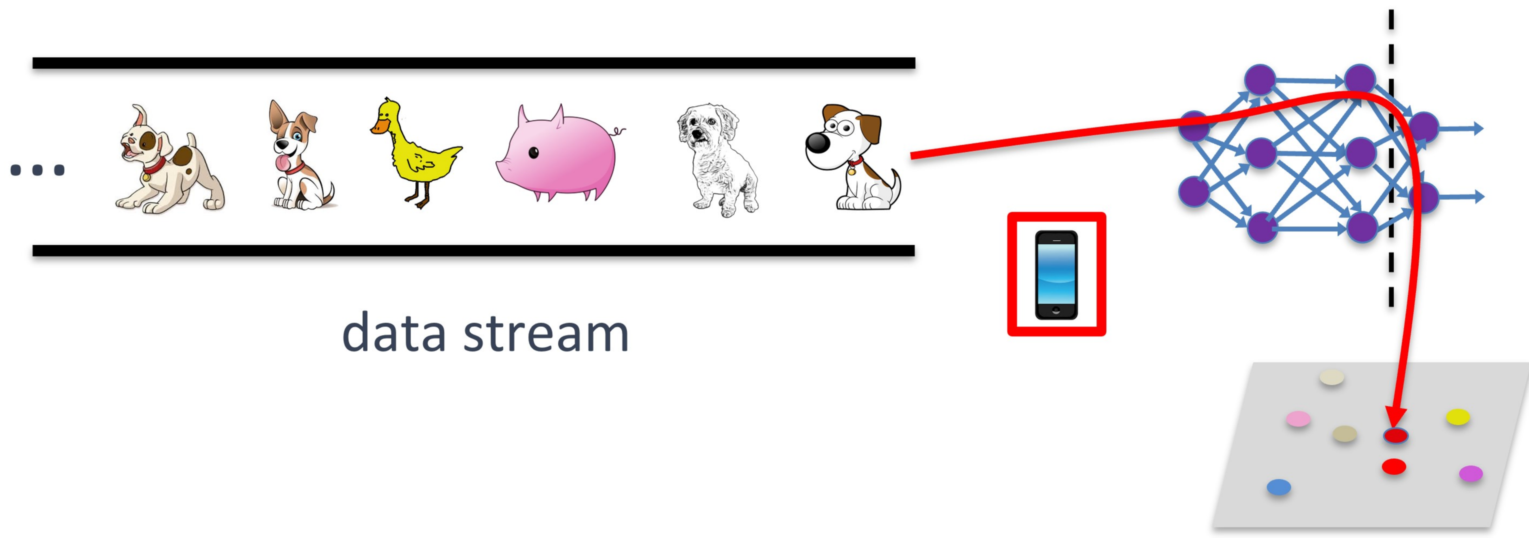
Effect of local datastore size and data heterogeneity

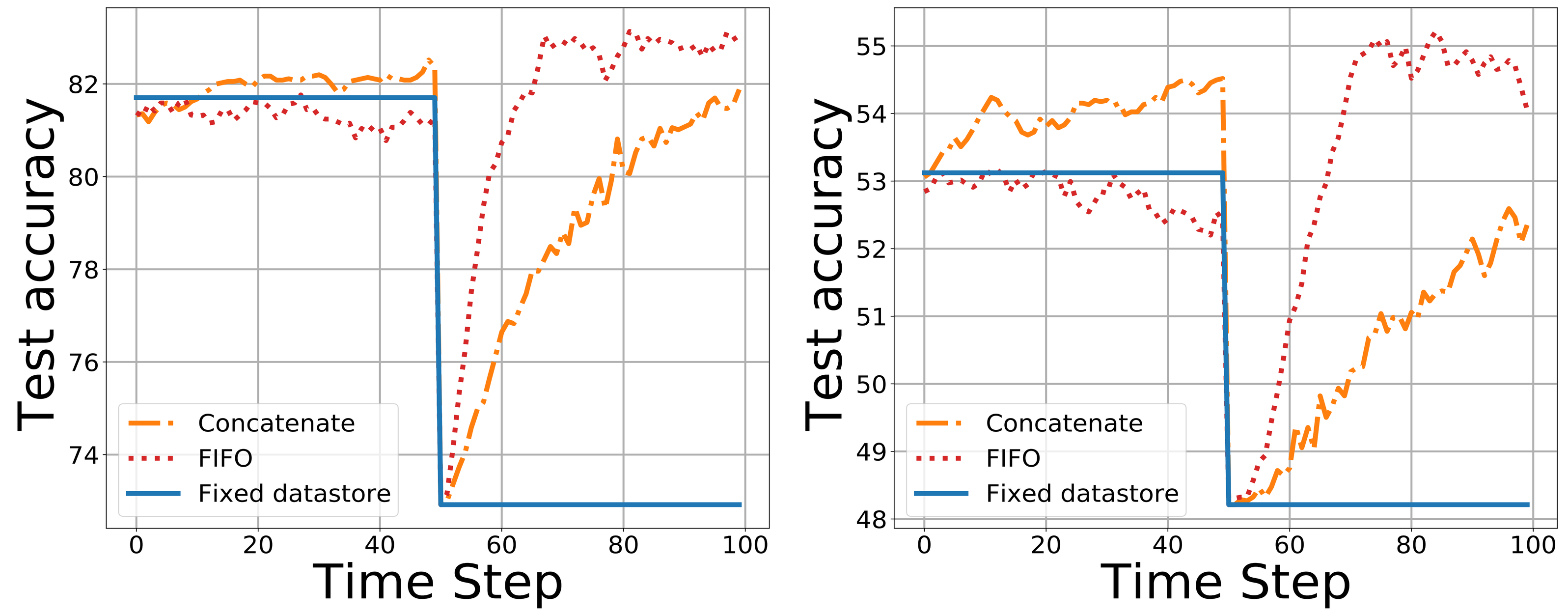


(a) CIFAR-10 (b) CIFAR-100

Figure: Accuracy vs capacity (local datastore size). The capacity is normalized with respect to the initial size of the client's dataset partition. Smaller values of α correspond to more heterogeneous data distributions across clients.

Robustness to distribution shift





(a) CIFAR-10 (b) CIFAR-100

Figure: Test accuracy when a distribution shift happens at time step $t_0 = 50$ for different datastore management strategies.

Conclusions

- `kNN-Per` offers a simple and effective way to address statistical heterogeneity in FL
- `kNN-Per` has a limited leakage of private information and can be easily combined with differential privacy techniques
- `kNN-Per` partially addresses system heterogeneity as data-store's size and approximate kNN's choice can be adapted to client's capabilities
- `kNN-Per` adapts to data distribution shifts over time by updating the local datastore