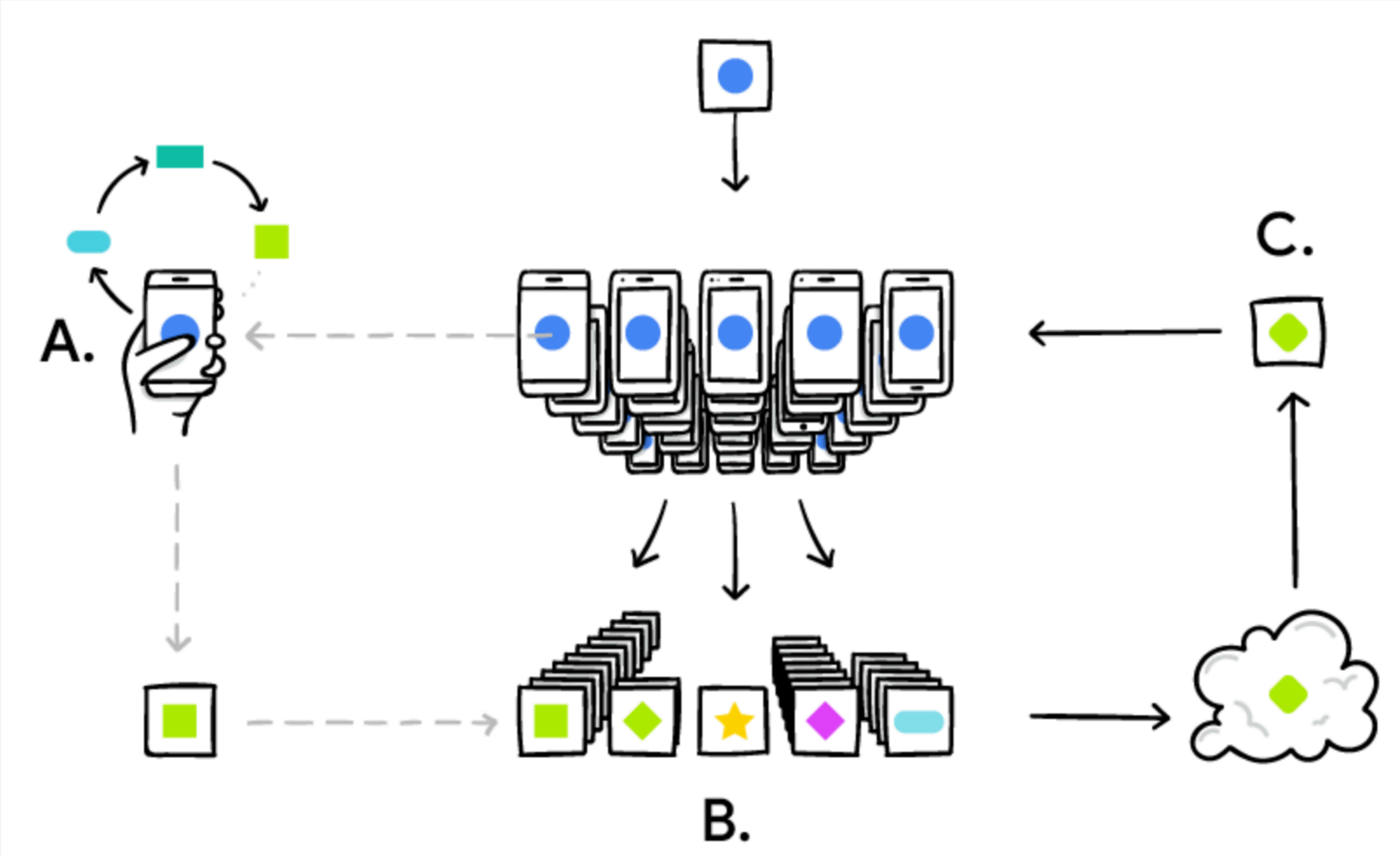
Towards Optimal Multi-Modal Federated Learning on Non-IID Data with Hierarchical Gradient Blending

Sijia Chen, Baochun Li
University of Toronto

Federated Learning

- *enables resource-constrained edge clients, such as mobile phones and IoT devices, to learn a shared global model for prediction, while keeping the training data local.*

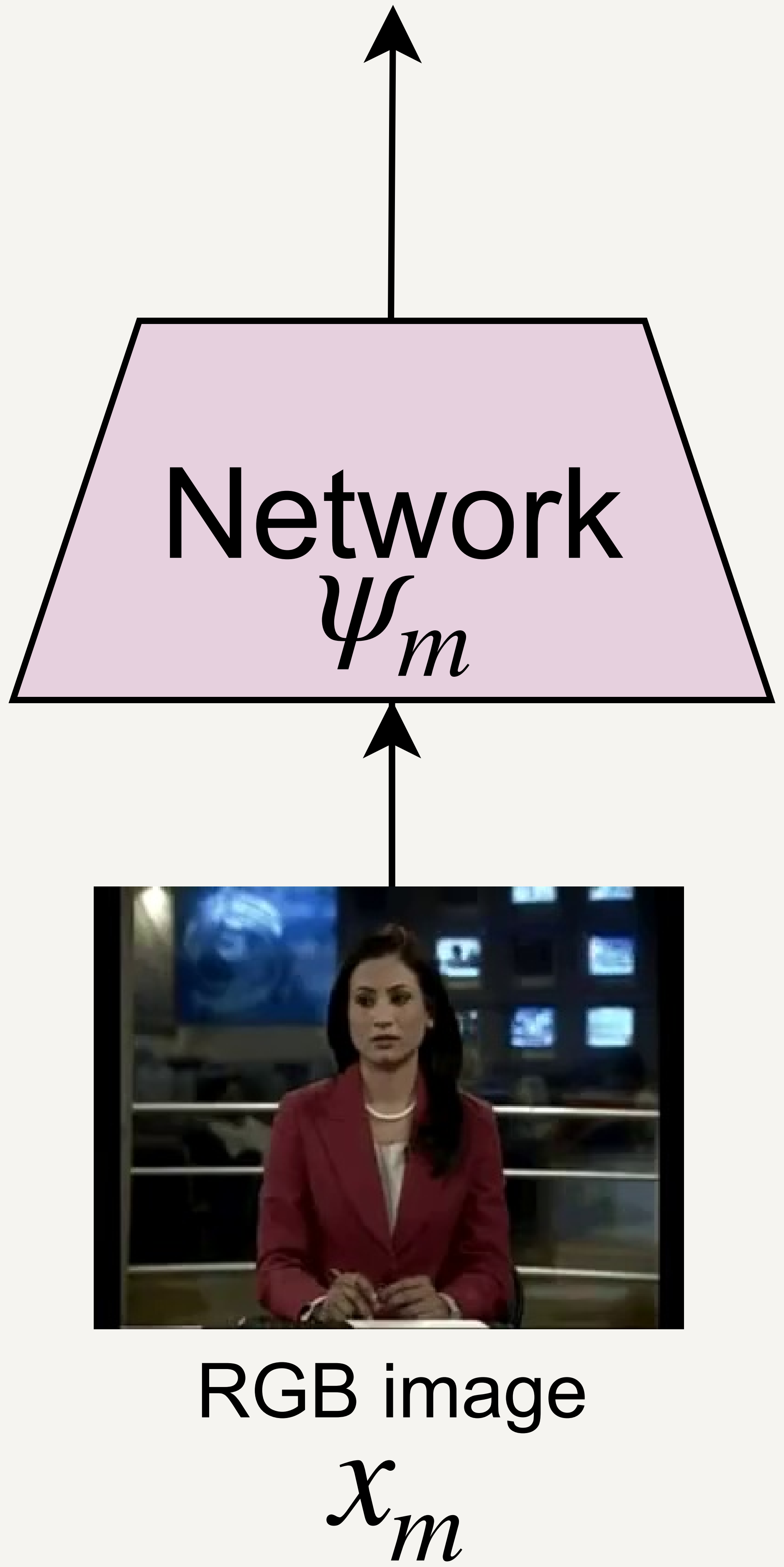


Uni-modal Federated Learning

- The global model receives one type of data modality as input.

$$f_k = \frac{1}{|D^k|} \sum_{s \in D^k} l\left(\psi_m(x_m; v_m); y)\right)$$

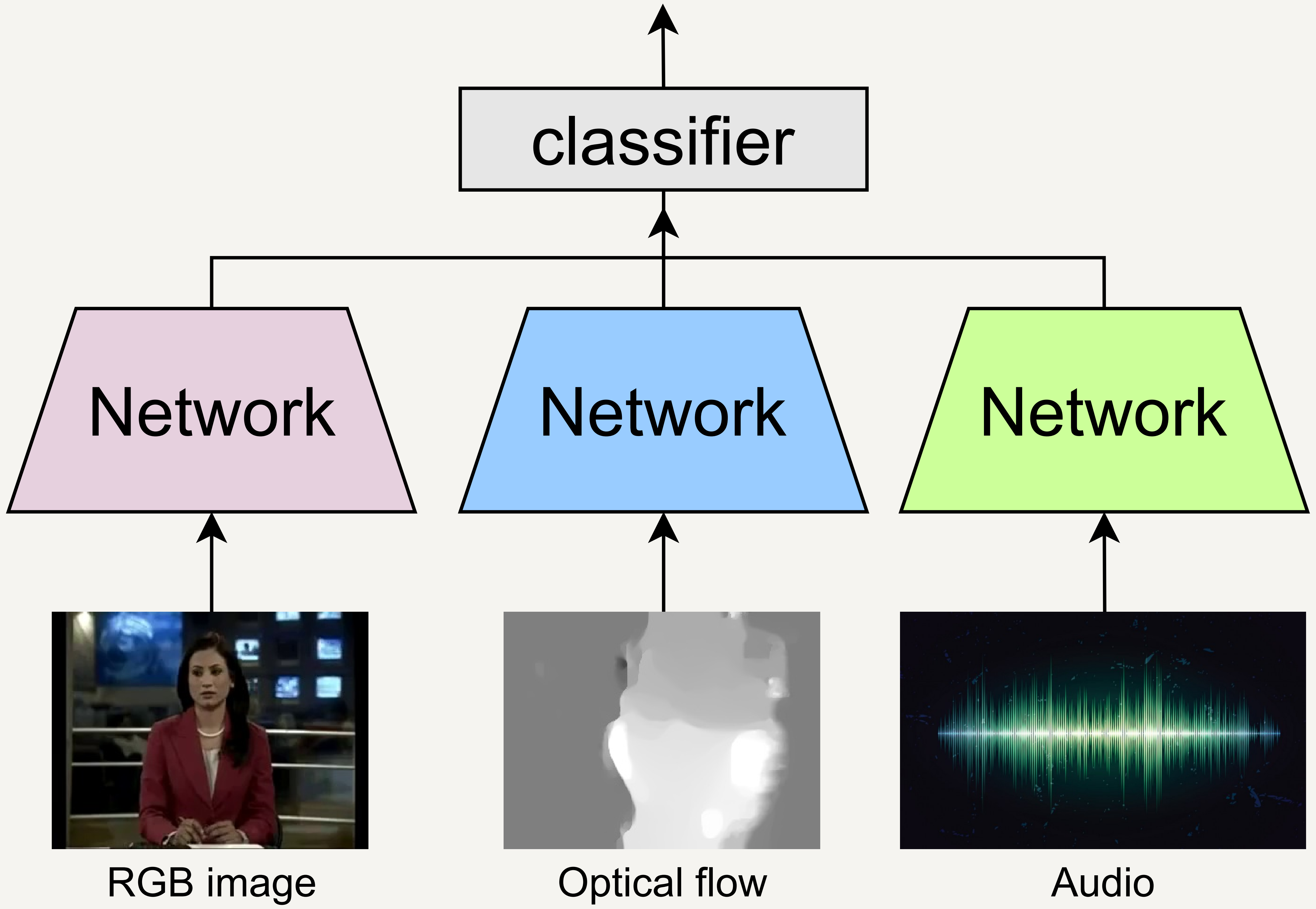
- The x_m denote samples extracted from single data modality such as RGB frames, audio, or optical flows.



An example of image classification

Multi-modal Machine Learning

- aims to build models that can process and relate information from multiple modalities.



Multi-modal Federated Learning

- The global multi-modal model is trained under the federated learning paradigm.

$$f_k = \frac{1}{|D^k|} \sum_{s \in D^k} l\left(\mathbb{C}(\psi_1(x_1), \ldots, \psi_M(x_M)); y)\right)$$

  - Each client contains samples from M modalities.
  - The global multi-modal model contains M sub-networks that are going to be jointly trained.

Performance Degradation of Classical Methods

- The classical federated learning method, FedAvg, presents performance degradation when training the multi-modal global model.

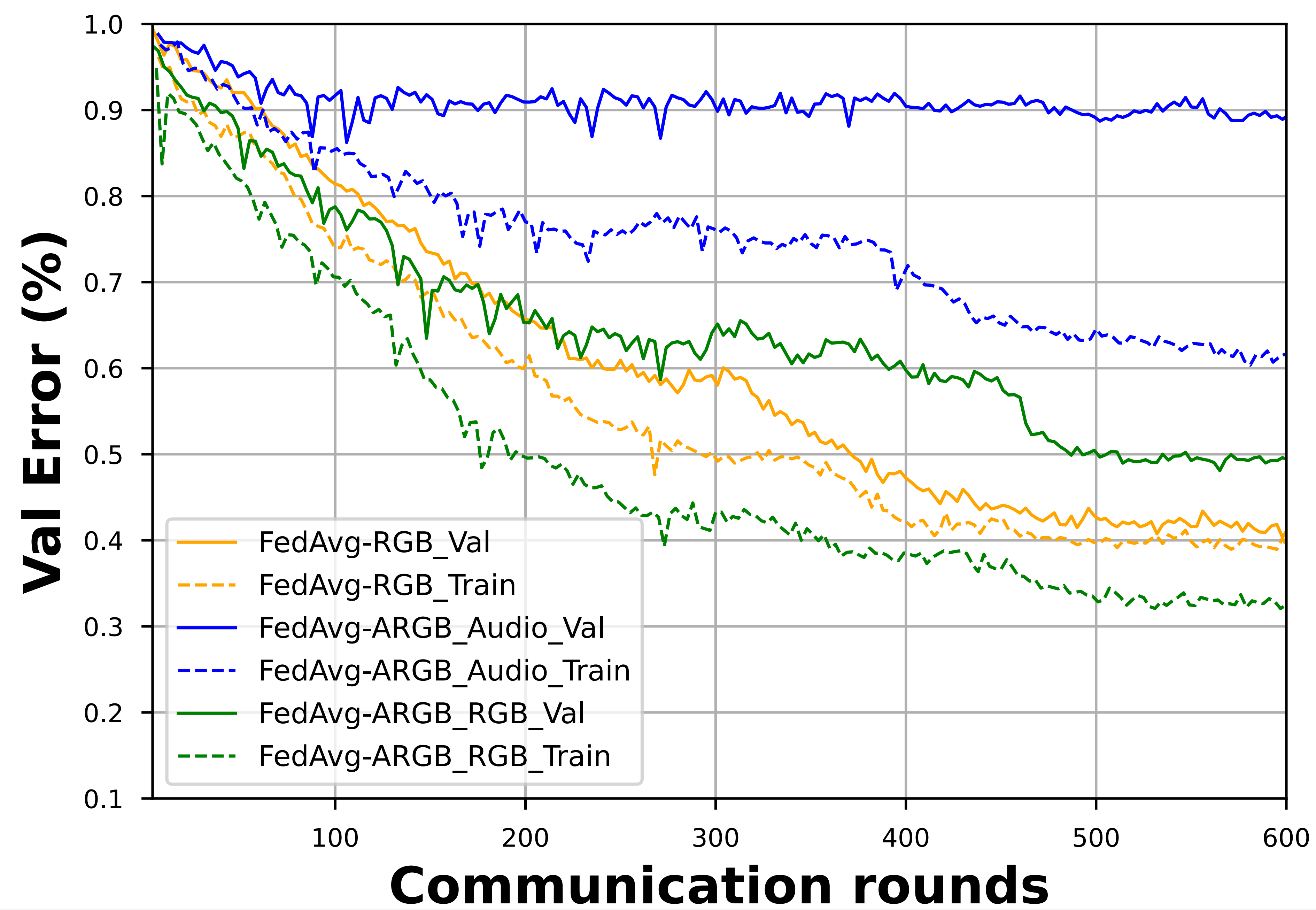


TABLE I: The performance comparison of the FedAvg method on uni-modal FL and the multi-modal FL under the non-IID data settings of three modalities.

	Centralized	FedAvg	
Modalities	V@1	V@1	#Rounds
RGB	71.52	58.43	480
A+RGB	72.17	49.91	519
OF+RGB	72.3	52.57	504
A+OF+RGB	73.62	38.62	557

Non-IID Multi-modal Data Challenge

- Multi-modal weights divergence:

$$\| w_{t_p}^{(f)} - w_{t_p}^{(c)} \| \leq \| w_{t_{p-1}}^{(f)} - w_{t_{p-1}}^{(c)} \| + \sum_{k=1}^{K} p_k \sum_{j=1}^{E} (d_{local} + d_{local\_global})$$

  - Local divergence d_{local}, $\| \nabla f_k(v_j^k, \xi^k) - \nabla f_k(w_j, \xi^k) \|$.
  - Local-global divergence d_{local_global}, $\| \nabla f_k(v_j^k, \xi^k) - \nabla f(w_j, \xi_j) \|$.

Non-IID Multi-modal Data Challenge

- Local divergence $\| \nabla f_k(v_j^k, \xi^k) - \nabla f_k(w_j, \xi^k) \|$:

$$\sum_{m=1}^{M} z_m^k g_{max}\left(w_{m_{j-1}}\right) \sum_{i \in \mathcal{Y}} B_{mi}^k \frac{\triangle d_m^k}{B_m^k} \left(\left(\eta B_m^k + 1\right)^{j-1-t_{p-1}} - 1 \right)$$

  - Gradient divergence from M sub-networks.

$$g_{max}\left(w_{m_{j-1}}\right) = \max_{i \in \mathcal{Y}} \| \nabla \psi_{(i)}\left(x_m; w_{m_{j-1}}\right) \|$$

Non-IID Multi-modal Data Challenge

- Local-global divergence $\| \nabla f_k(v_j^k, \xi^k) - \nabla f(w_j, \xi_j) \|$:

$$\sum_m z_m^k g_{max}\left(w_{m_j}\right) \sum_{i \in \mathcal{Y}} \left(\underline{p_m^k(y=i) - p_m(y=i)}\right)$$

  - Gradient divergence from participating clients.
  - Data distribution distance of modality m between local data k and the global data.

Hierarchical gradient blending

- The high-level idea is to update the model to reduce the training loss while achieving low evaluation loss.

  - HGB directly minimizes the overfitting-to-generalization ratio (OGR).

$$\min_{\{z_m\}_{m=1}^{M}, \{p_k\}_{k=1}^{K}} \left(\frac{[L^T(w_{t_{p-1}} - L^T(w_{t_p})] - [L^*(w_{t_{p-1}}) - L^*(w_{t_p})]}{L^*(w_{t_{p-1}}) - L^*(w_{t_p})} \right)^2$$

  - Achieve the best OGR for adjacent global parameters $w_{t_{p-1}}$ and w_{t_p} obtained by aggregating local models from K clients.

Optimal hierarchical gradient blending

- Computes the optimal $\left\{z_m^k\right\}_{m=1}^{M}, k \in [1,K]$ in the local updates.

$$z_m^{k*} = \frac{1}{Q}\frac{\langle \nabla l_k^*, g_m^k \rangle}{\sigma_m^2}, Q = \frac{\sum_{m=1}^{M} \frac{\langle \nabla l_k^*, g_m^k \rangle}{\sigma_m^2}}{2}$$

  - To achieve the minimum overfitting-to-generalization ratio (OGR) when jointly training M sub-networks in the local update.

Optimal hierarchical gradient blending

- Computes the optimal $\{p_k\}_{k=1}^{K}$ in the global aggregation.

$$p_k^* = \frac{1}{M} \frac{\triangle G^k(t_{p-1}, t_p)}{2\left(\triangle O^k(t_{p-1}, t_p)\right)^2}, M = \sum_{k=1}^{K} \frac{\triangle G^k(t_{p-1}, t_p)}{2\left(\triangle O^k(t_{p-1}, t_p)\right)^2}$$

- To achieve the minimum overfitting-to-generalization ratio (OGR) when aggregating gradients from K participating clients.

Evaluations

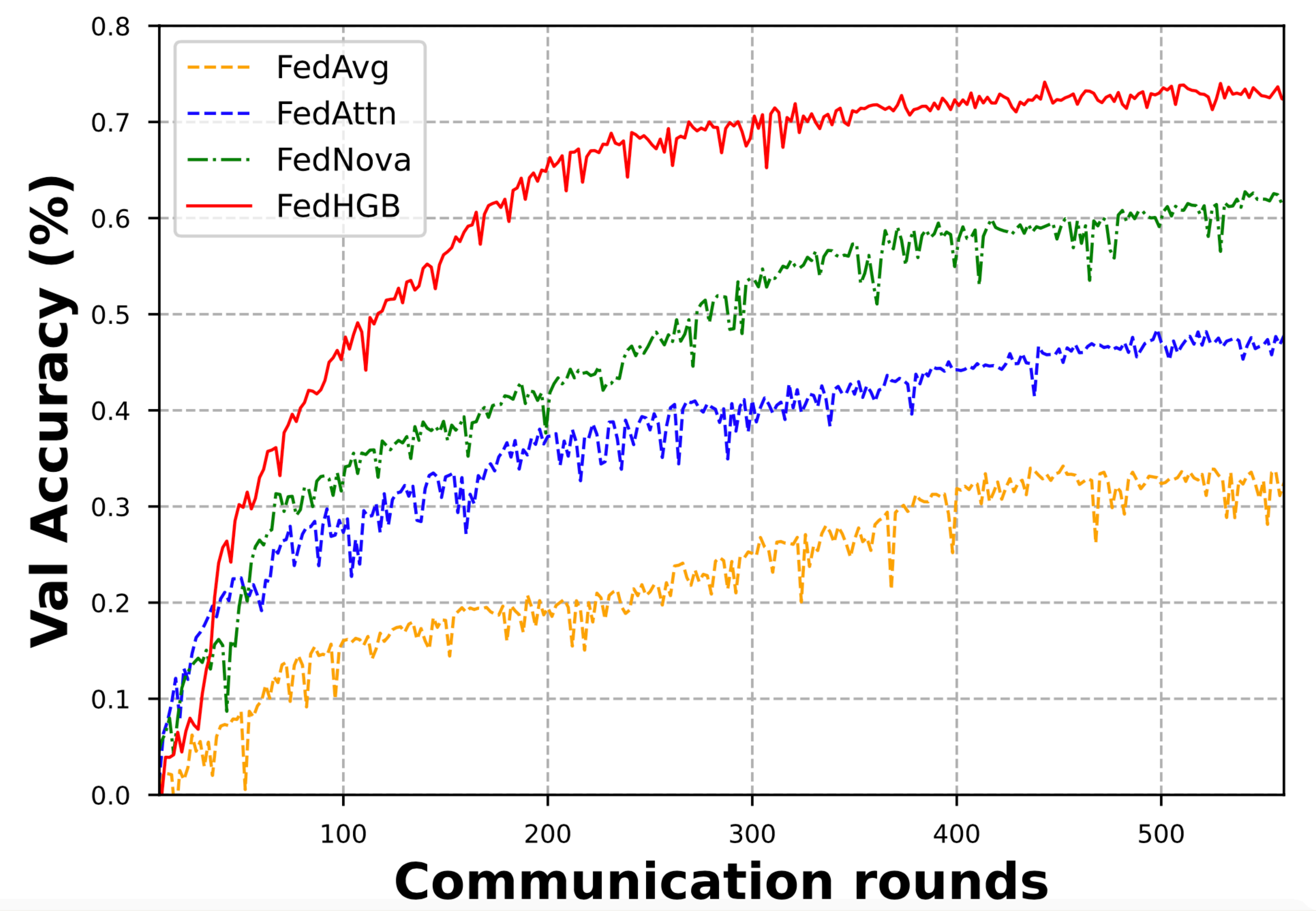
- Targeting the video recognition task
  - Kinetics
  - FineGym
- Designing the non-IID multi-modal data as:
  - In case A, each client contains all modalities.
  - In case B, each client can only contain subset modalities.
  - Case C is built on case B but adds the sample skewness among modalities.

Performance

- Our method outperforms alternative leading methods, including FedAttn [1] and FedNova [2], in terms of classification accuracy and convergence speed.

TABLE II: The performance comparison of methods in case B with two modality non-IID types (i.e., Mixed-B and 2M-B). The evaluation metric is the top-1 accuracy and the communication rounds distance ($\triangle CR$) between FedHGB and the fastest method.

Datasets	Kinetics		Gym	
Case B settings	Mixed-B	2M-B	Mixed-B	2M-B
FedAvg	38.04	44.32	42.33	51.42
FedAttn	51.79	56.91	58.07	64.52
FedNova	55.12	58.76	63.92	68.3
FedHGB	62.97	64.39	71.66	73.34
$\triangle CR$	34	15	51	20
Uni-RGB	62.33		70.52	



The validation curve on Gym dataset under Case C.

[1]. S. Ji, et. Al, Learning private neural language modeling with attentive aggregation, International Joint Conference on Neural Networks (IJCNN 2019).

[2]. Jianyu Wang, et. Al, Tackling the Objective Inconsistency Problem, Neural Information Processing Systems (NeurIPS 2020).

Ablation Study

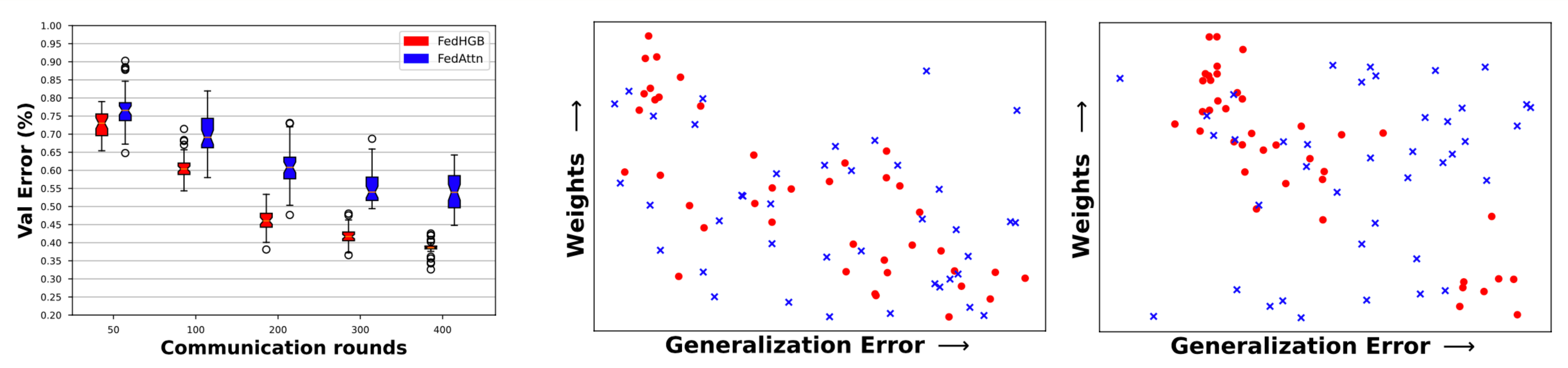
- M-GB computes optimal blending of sub-networks.
- C-GB computes optimal blending of clients' gradients.

TABLE III: The performance comparison between ablation methods of HGB in two datasets with all non-IID settings. The evaluation metric is the top-1 accuracy (%) and the communication rounds $\triangle CR$ that is computed as CR of M-GB minus the CR of C-GB .

Datasets	methods	CaseA	mixed-B	2M-B	1M-B	Case C
Kinetics	M-GB	65.92	56.55	60.47	58.73	57.66
	C-GB	63.83	55.91	57.02	58.64	52.81
	$\triangle CR$	25	48	66	95	67
Gym	M-GB	71.36	66.13	69.92	67.34	65.17
	C-GB	70.03	64.4	66.1	66.38	58.93
	$\triangle CR$	41	36	69	87	46

- M-GB performs well on the accuracy metric.
- C-GB performs well on the convergence speed.

Qualitative Analysis



Comparison of quantitative results on the Kinetics dataset in the non-IID Case C.

- The first column shows the generalization distribution of clients before aggregation in different communication rounds.
- The other two columns show the relationship between generalization error and the computed weight p_k^* for participating clients.

Conclusion Remarks

- Hierarchical Gradient Blending for Optimal Multi-Modal Federated Learning on Non-IID Data
  - Train multi-modal global model to consistently outperform uni-modal model.
  - Maintain high performance (i.e., accuracy and convergence speed) under different challenging non-IID multi-modal data.
  - Outperform alternative leading methods.
- Future Work:
  - Explore more complicated multi-modal federated learning tasks, such as visual grounding federated learning.

Contact: sjia.chen@mail.utoronto.ca