FedBABU: Toward Enhanced Representation for Federated Image Classification

Jaehoon Oh*, Sangmook Kim*, Se-Young Yun

KAIST

ICLR'22

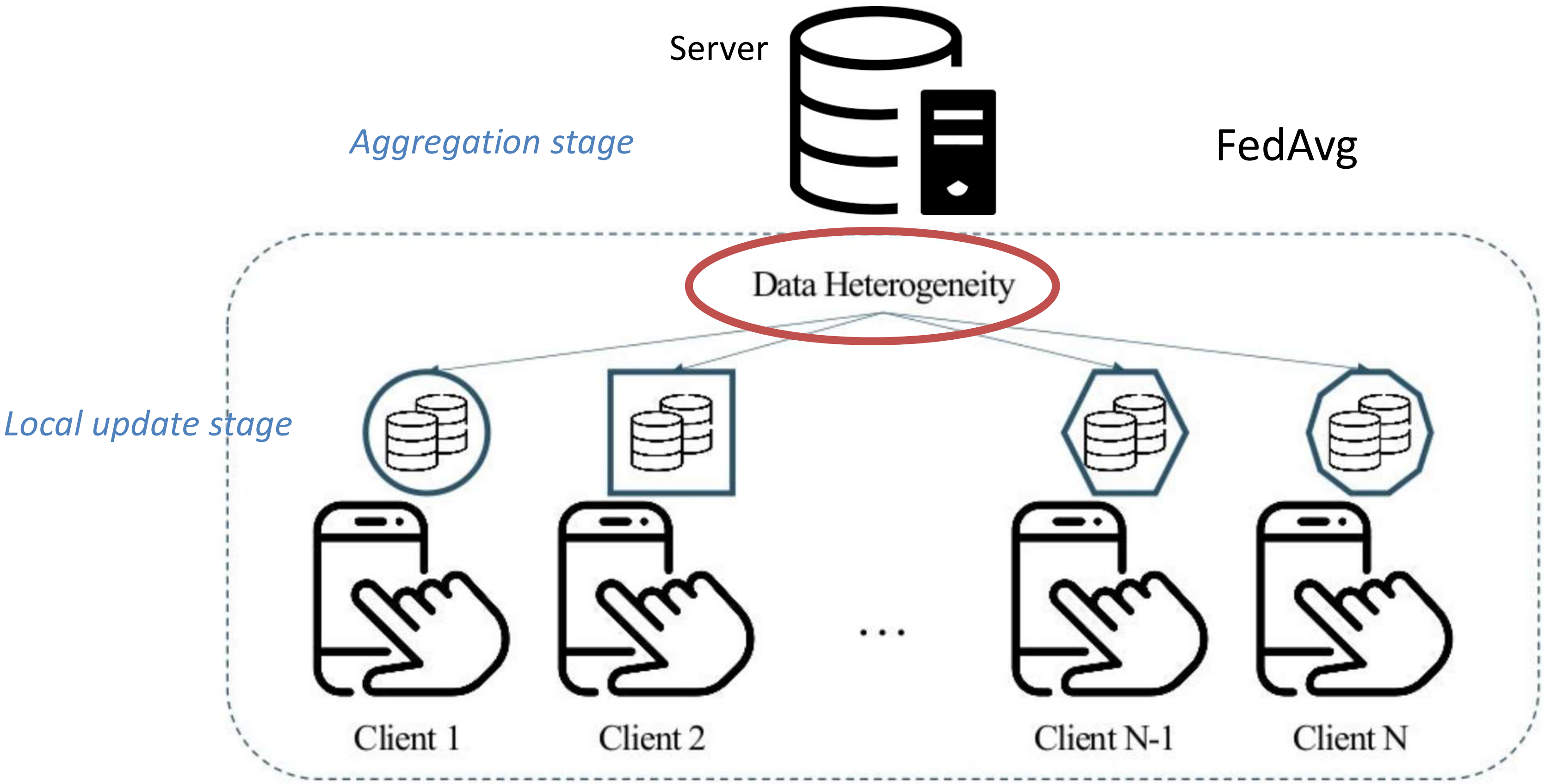
Presented by,

Pavana Prakash

Outline

- Motivation and Background
- FedBABU
- Experiments
- Conclusion

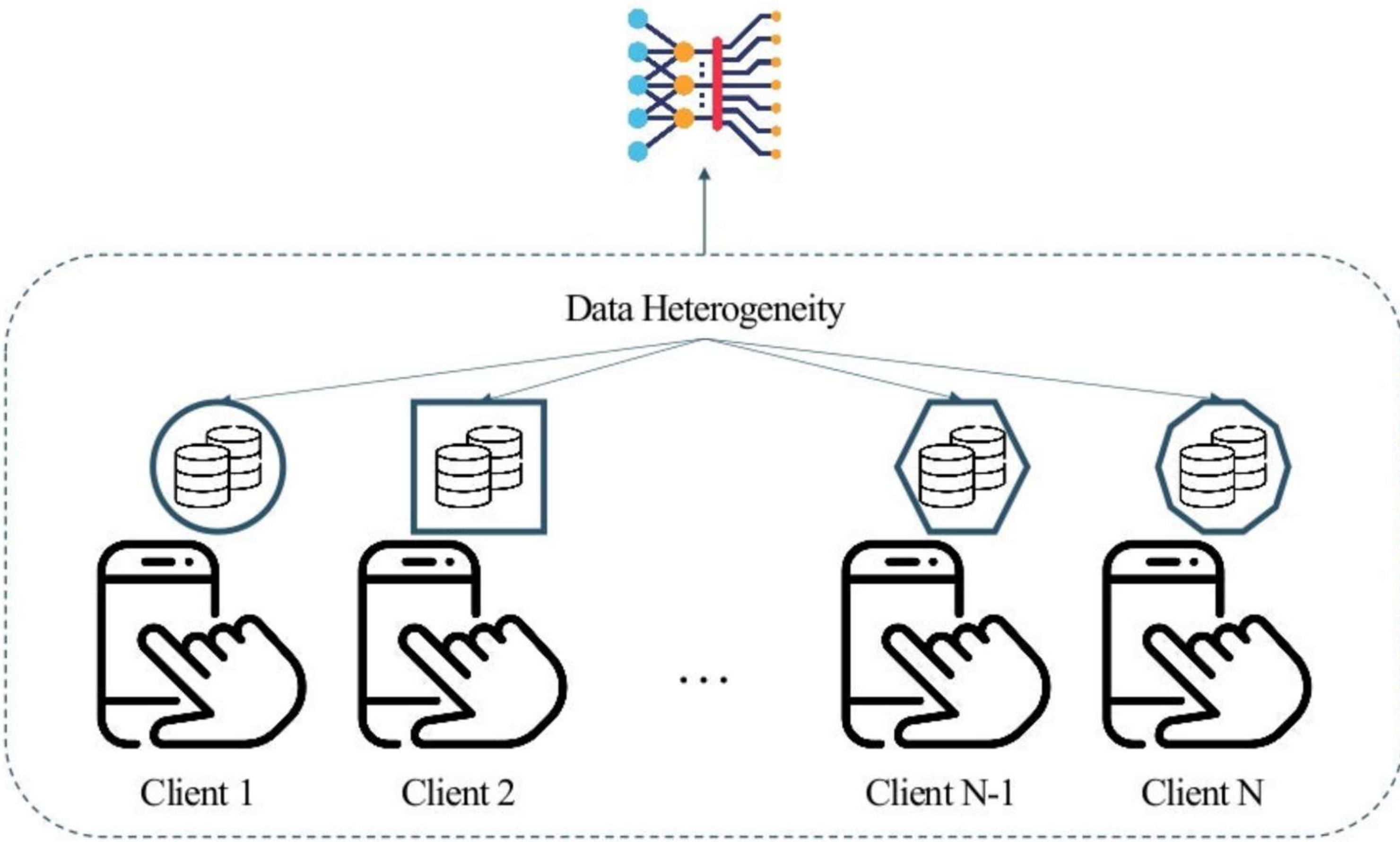
Federated Learning (FL)



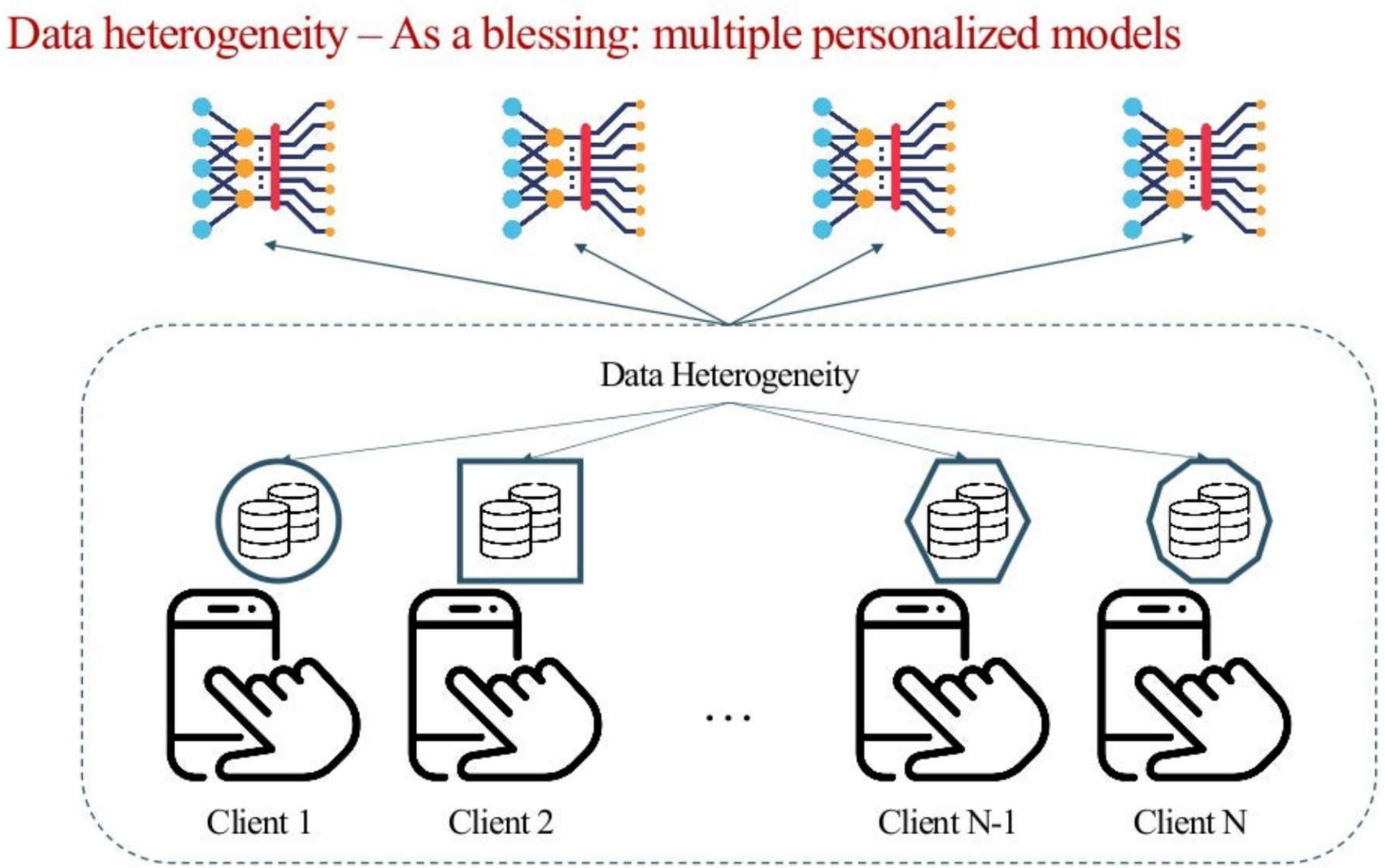
[1] McMahan, Brendan, et al. "Communication-Efficient Learning of Deep Networks from Decentralized Data." Artificial intelligence and statistics. PMLR, 2017.

Data Heterogeneity in FL

Data heterogeneity – As a curse: a single global model

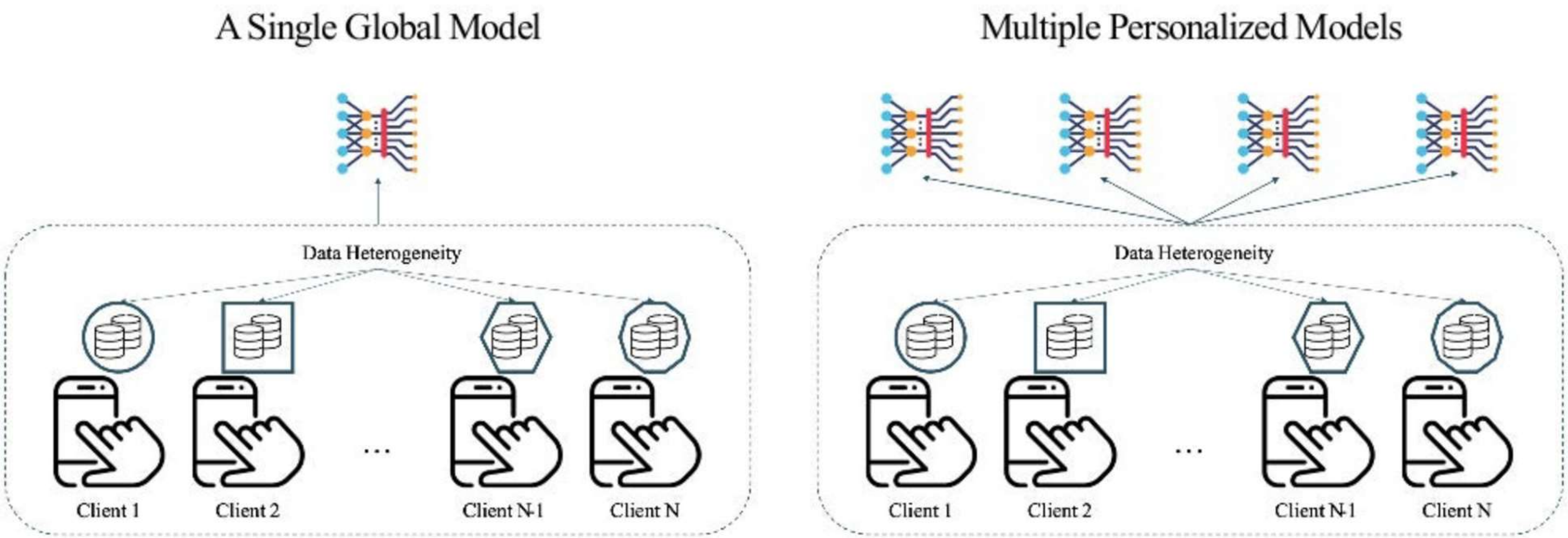


Data Heterogeneity in FL



Introduction

Problem statement: How to train a global model that can personalize



Personalization of a Single Global Model

Observation 1: Effectiveness of a global single model (FedAvg)

- The trained global model can be personalized with a few epochs, particularly under large data heterogeneity.

MobileNet on CIFAR100 with 100 clients

s - Shard per client

Client fraction

Local epochs

FL settings		s=100 (heterogeneity ↓)		s=50		s=10 (heterogeneity ↑)	
f	τ	Initial	Personalized	Initial	Personalized	Initial	Personalized
1.0	1	46.93±5.47	51.93±5.19	45.68±5.50	57.84±5.08	37.27±6.97	77.46±5.78
	4	37.44±4.98	42.66±5.09	36.05±4.04	47.17±4.26	24.17±5.50	70.41±6.83
	10	29.58±4.87	34.62±4.97	29.57±4.29	40.59±5.23	17.85±7.38	63.51±7.38
0.1	1	39.07±5.22	43.92±5.55	38.20±5.73	49.55±5.36	29.12±7.11	71.24±7.82
	4	35.39±4.58	39.67±5.21	33.49±4.72	43.63±4.77	21.14±6.86	67.14±6.72
	10	28.18±4.83	33.13±5.22	27.34±4.96	38.09±5.17	14.40±5.64	62.67±6.52

mean±standard deviation of the accuracies across all clients

The initial and personalized accuracy indicate the evaluated performance without fine-tuning and after five fine-tuning epochs for each client, respectively.

ANTS LAB

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Personalization of a Single Global Model

Observation 2: Motivation for our study: Data sharing

Server has a small portion of p of the non-private client data of the clients

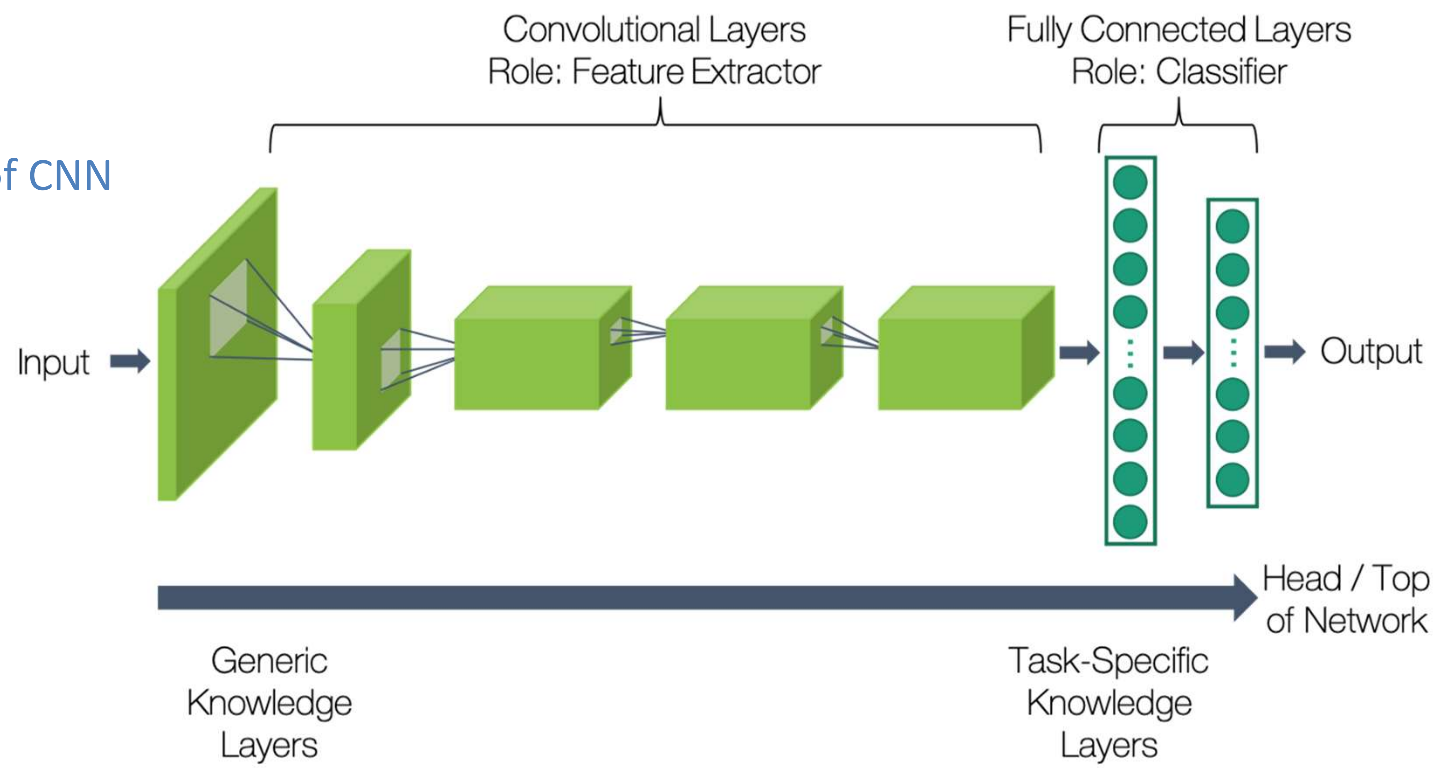
p	s=100 (heterogeneity ↓)		s=50		s=10 (heterogeneity ↑)	
	Initial	Personalized	Initial	Personalized	Initial	Personalized
0.00	28.18±4.83	33.13±5.22	27.34±4.96	38.09±5.17	14.40±5.64	62.67±6.52
0.05 (Full)	29.23±5.03	32.59±5.24	27.13±4.34	34.34±4.78	18.22±5.64	54.68±6.77
0.05 (Body)	28.50±4.93	33.03±5.36	27.96±4.86	39.10±5.55	14.78±5.59	60.19±6.46
0.10 (Full)	30.59±4.93	33.34±5.30	29.62±4.27	35.50±4.84	19.24±5.15	49.62±7.48
0.10 (Body)	32.90±4.77	36.82±4.66	32.81±4.97	40.80±5.62	18.35±6.75	60.94±7.30

Boosting a single global model can hurt personalization

- Data sharing increases initial accuracy but decreases personalized accuracy.
- By narrowing the update parts, the personalization degradation problem is remedied significantly. It implies that training the head can deteriorate the personalization performance since head is biased.

Neural Network Layers

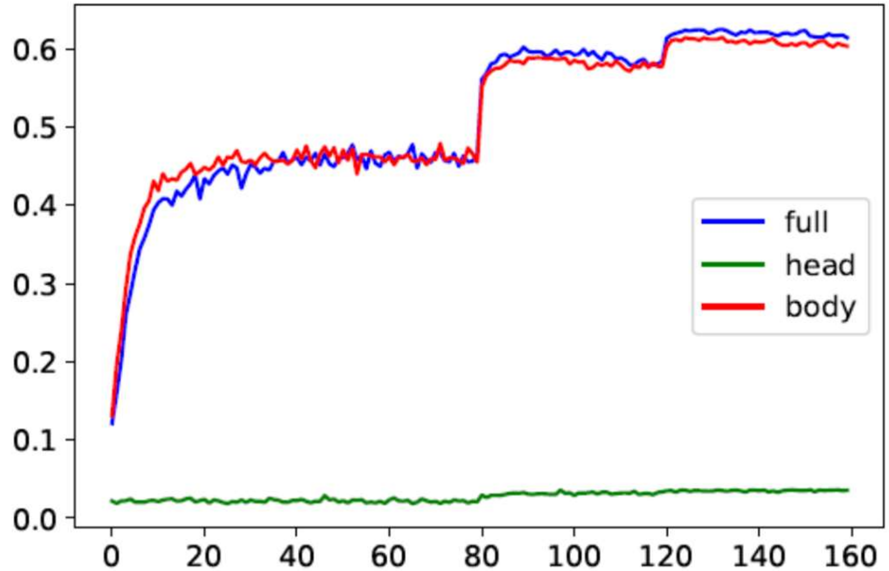
Overview of CNN



- Popular networks such as ResNet, MobileNet have only one linear layer at the end of the model
- This linear layer : Head → linear decision boundary learning
- All of the layers except the head : Body → representation learning

Frozen Head in the Centralized Setting

Observation 3



Test accuracy curves

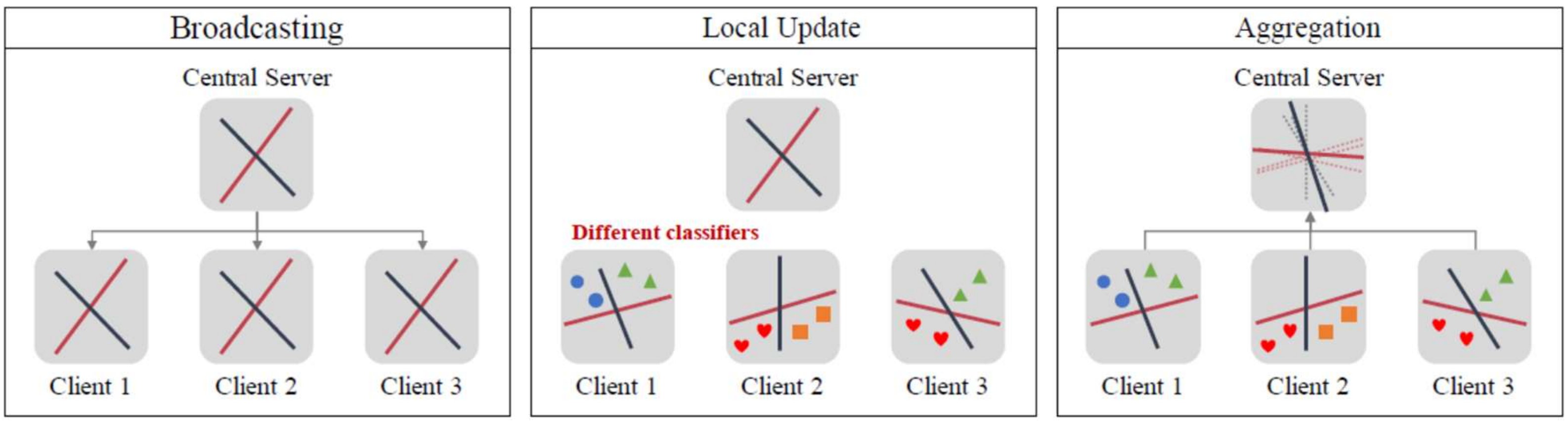
- MobileNet on CIFAR100 trained in centralized setting for 160 total epochs
- Full: accuracy when all layers are trained
- Body: accuracy when only the body of the model is trained
- Head: accuracy when only the head of the model is trained
- *Model with the initialized and fixed head has comparable performance to a model that jointly learns the body and the head!*

FedBABU

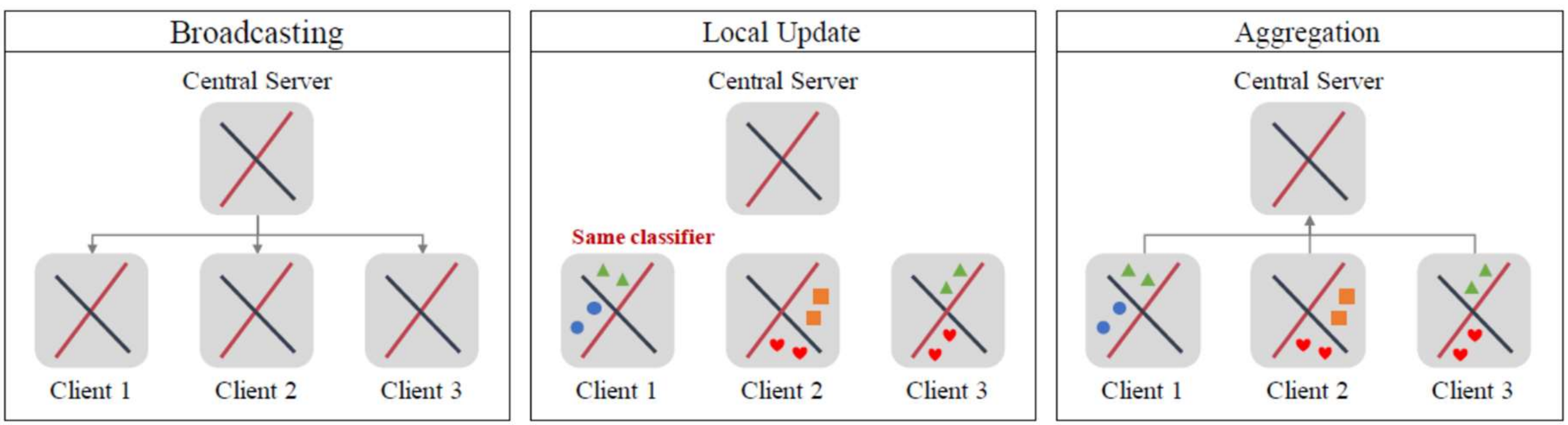
- Federated Averaging with Body Aggregation and Body Update
- Decouple the entire network into the body and the head
- Body (extractor), is trained for generalization → related to universality
- Head (classifier), is then trained for specialization → related to personalization
- Federated learning:
  - never train the head in the federated training phase (i.e., develop a single global model)
  - no need to aggregate the head
  - fine-tune the head for personalization (in the evaluation process)
  - Same fixed head on all clients serves as the same criteria on learning representations across all clients

FedBABU

FedBABU based on decoupling parameters



(a) FedAvg.



(b) FedBABU.

Algorithm 1 Training procedure of FedBABU.

initialize $\theta_G^0 = \{\theta_{G,ext}^0, \theta_{G,cls}^0\}$ initialized global parameter
for each round $k = 1, \cdots, \mathrm{K}$ **do**
  $m \leftarrow max(\lfloor Nf \rfloor, 1)$
  $C^k \leftarrow$ random subset of m clients
  for each client $C_i^k \in C^k$ in parallel **do**
    $\theta_i^k(0) \leftarrow \theta_G^{k-1} = \{\theta_{G,ext}^{k-1}, \theta_{G,cls}^0\}$
    $\theta_{i,ext}^k(\tau I_i^k) \leftarrow$ **ClientBodyUpdate**$(\theta_i^k(0), \tau)$
  end for
  $\theta_{G,ext}^k \leftarrow \sum_{i=1}^m \frac{n_{C_i^k}}{n} \theta_{i,ext}^k(\tau I_i^k), n = \sum_{i=1}^m n_{C_i^k}$ global body parameter
end for
return $\theta_G^{\mathrm{K}} = \{\theta_{G,ext}^{\mathrm{K}}, \theta_{G,cls}^0\}$ final global parameter

function ClientBodyUpdate(θ_i^k, τ)
  $I_i^k \leftarrow \lceil \frac{n_{C_i^k}}{B} \rceil$
  for each local epoch $1, \cdots, \tau$ **do**
    for each iteration $1, \cdots, I_i^k$ **do**
      $\theta_{i,ext}^k \leftarrow SGD(\theta_{i,ext}^k, \theta_{G,cls}^0)$ local body parameter
    end for
    Same fixed head parameter
  end for
  return $\theta_{i,ext}^k$
end function

Experiments

- MobileNet on CIFAR100
- Number of clients: 100
- Each client has 500 training data and 100 test data
- Shards for heterogeneity: sort the data by labels and divide the data into the same-sized shards (dataset size/(total number of clients * number of shards per user))
- Hyperparameters: client fraction ratio f, local epochs τ , and shards per user s
- Initial accuracy: the learned global model is broadcast to all clients and is then evaluated on the test data set of each client
- Personalized accuracy: the learned global model is personalized using the training data set of each client by fine-tuning with the fine-tuning epochs of τf ; the personalized models are then evaluated on the test data set of each client
- TITAN RTX

ANTS LAB

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Evaluation

- Exp 1 . Representation power of a single global model

Initial accuracy of FedAvg and FedBABU under various settings

FL settings			FedAvg		FedBABU	
s	f	τ	w/ head	w/o head	w/ head	w/o head
100	1.0	1	46.93±5.47	46.23±4.53	48.61±4.75	49.97±4.69
		4	37.44±4.98	33.48±5.09	37.32±4.39	37.20±4.35
		10	29.58±4.87	25.11±4.60	26.69±4.50	27.70±4.51
	0.1	1	39.07±5.22	36.69±5.82	41.02±4.99	41.19±4.96
		4	35.39±4.58	32.58±4.37	36.77±4.47	36.61±4.64
		10	28.18±4.83	24.34±4.58	29.38±4.74	29.36±4.46
50	1.0	1	45.68±5.50	53.87±5.39	47.19±4.77	55.70±5.48
		4	36.05±4.04	42.65±4.76	37.27±5.25	45.25±5.45
		10	29.57±4.29	34.13±4.44	28.43±4.72	36.19±4.93
	0.1	1	38.20±5.73	44.57±5.34	41.33±5.10	49.18±5.73
		4	33.49±4.72	40.01±5.49	34.68±4.58	42.43±5.11
		10	27.34±4.96	33.10±5.08	27.91±5.27	36.49±5.37
10	1.0	1	37.27±6.97	67.18±7.27	45.32±8.52	71.23±6.71
		4	24.17±5.50	58.70±6.74	32.91±7.07	64.41±7.44
		10	17.85±7.38	51.72±7.65	22.15±5.72	55.63±7.24
	0.1	1	29.12±7.11	60.42±7.89	35.05±7.63	65.98±6.43
		4	21.14±6.86	54.91±6.72	25.67±7.31	59.44±6.43
		10	14.40±5.64	50.25±6.27	18.50±7.82	54.93±7.85

Improved representation power under large data heterogeneity

Experimental Evaluation

- Exp 2. Personalization of FedBABU

FL settings		Update part for personalization		
s	τ	Body	Head	Full
100	1	44.26±5.12	49.76±5.03	49.67±4.92
	4	39.61±4.68	44.74±5.02	44.74±5.10
	10	32.45±5.42	36.48±5.04	35.94±5.06
50	1	48.54±5.23	56.76±5.68	56.69±5.16
	4	41.27±5.04	49.45±5.41	49.55±5.58
	10	35.42±5.60	42.55±5.70	42.63±5.59
10	1	72.81±7.32	75.97±6.29	76.02±6.29
	4	69.12±6.70	70.74±6.47	71.00±6.63
	10	64.77±7.14	66.28±6.77	66.32±7.02

(a) FedBABU.

FL settings		Update part for personalization		
s	τ	Body	Head	Full
100	1	41.00±5.35	43.18±5.34	43.92±5.55
	4	37.43±4.98	38.29±4.96	39.67±5.21
	10	30.62±4.95	31.92±5.04	33.13±5.22
50	1	43.61±5.54	47.51±5.61	49.55±5.36
	4	37.99±4.68	41.48±4.61	43.63±4.77
	10	32.20±4.92	36.06±5.31	38.09±5.17
10	1	56.00±8.66	69.70±7.88	71.24±7.82
	4	34.49±7.92	65.32±6.81	67.14±6.72
	10	27.94±6.96	60.24±6.16	62.67±6.52

(b) FedAvg.

Fine-tuning epochs is 5 and f is 0.1

Experimental Evaluation

- Exp 3. Personalization performance comparison

FL settings			Personalized accuracy							
s	f	τ	FedBABU (Ours)	FedAvg (2017)	FedPer (2019)	LG-FedAvg (2020)	FedRep (2021)	Per-FedAvg (2020)	Ditto (2021)	Local-only
100	1.0	1	**55.79**±4.57	51.93±5.19	51.95±5.30	53.01±5.26	18.29±3.59	47.09±7.35	39.36±4.78	20.60±3.15
		4	**44.49**±4.91	42.66±5.09	40.87±5.05	43.09±4.74	15.32±3.79	39.07±7.59	31.58±5.00	
		10	33.20±4.54	34.62±4.97	32.91±4.97	**34.64**±5.03	13.45±3.26	30.22±6.59	21.18±4.54	
	0.1	1	**49.67**±4.92	43.92±5.55	45.17±4.70	40.91±5.50	23.84±3.92	48.10±7.42	45.46±4.73	
		4	**44.74**±5.10	39.67±5.21	39.30±4.92	37.87±4.99	16.01±3.48	33.70±7.04	32.46±5.42	
		10	**35.94**±5.06	33.13±5.22	32.08±4.97	30.08±5.34	11.11±3.13	25.82±5.83	23.96±4.64	
50	1.0	1	**61.09**±4.91	57.84±5.08	57.16±5.26	58.44±5.53	24.75±5.02	43.75±7.94	42.70±5.46	28.02±4.01
		4	**51.56**±5.04	47.17±4.26	48.89±5.40	47.78±4.72	21.55±4.36	37.59±7.87	36.57±5.11	
		10	**42.09**±5.12	40.59±5.23	39.90±5.54	40.32±4.70	19.92±4.50	28.75±6.40	27.27±5.04	
	0.1	1	**56.69**±5.16	49.55±5.36	51.63±5.27	42.64±5.55	32.88±5.09	43.96±7.40	43.22±5.82	
		4	**49.55**±5.58	43.63±4.77	46.31±5.63	38.54±4.71	21.13±3.96	28.67±6.98	31.65±5.08	
		10	**42.63**±5.59	38.09±5.17	39.81±4.88	30.79±6.12	15.15±4.01	21.64±6.16	22.16±4.67	
10	1.0	1	**79.17**±6.51	77.46±5.78	74.71±6.35	77.49±5.60	61.28±8.27	36.59±8.98	65.33±7.49	61.52±7.22
		4	**74.60**±6.69	70.41±6.83	65.61±7.13	69.97±6.42	50.59±7.94	18.31±10.57	64.47±7.45	
		10	**66.64**±6.84	63.51±7.38	59.71±7.35	61.50±7.28	42.13±7.53	11.54±8.87	51.68±7.44	
	0.1	1	**76.02**±6.29	71.24±7.82	69.36±6.77	51.75±9.32	60.13±7.72	31.21±11.66	31.91±15.10	
		4	**71.00**±6.63	67.14±6.72	62.62±7.63	35.80±10.55	45.91±7.68	14.34±9.51	23.70±15.84	
		10	**66.32**±7.02	62.67±6.52	59.50±7.33	25.04±12.02	34.30±7.84	9.17±6.95	14.24±15.67	

Personalized accuracy comparison

Experimental Evaluation

- Exp 4. Personalization speed of FedAvg and FedBABU

Performance according to the fine-tune epochs (FL setting: f=0.1, and τ =10).

s	Algorithm	Fine-tune epochs (τ_f)								
		0 (Initial)	1	2	3	4	5	10	15	20
50	FedAvg	27.34±4.96	29.17±5.01	32.39±4.77	34.97±5.13	36.78±5.13	38.09±5.17	40.56±5.43	41.20±5.51	40.86±5.13
	FedBABU	27.91±5.27	35.20±5.58	40.60±5.47	42.12±5.61	42.74±5.60	42.63±5.59	41.94±5.68	41.19±5.52	40.61±5.28
10	FedAvg	14.40±5.64	27.43±6.46	48.63±7.30	58.08±6.11	61.27±6.15	62.67±6.52	63.91±6.49	64.56±6.45	64.89±6.53
	FedBABU	18.50±7.82	63.29±7.55	66.05±6.93	66.10±6.54	66.40±7.24	66.32±7.02	66.07±7.57	66.24±7.67	66.32±7.71

FedBABU achieves better accuracy with a small number of epochs → can personalize accurately and rapidly, especially when fine-tuning is either costly or restricted.

Experimental Evaluation

- Exp 5. Body aggregation and body update on the FedProx

Algorithm	τ	s=100 (heterogeneity ↓)		s=50		s=10 (heterogeneity ↑)	
		Initial	Personalized	Initial	Personalized	Initial	Personalized
FedProx	1	46.52±4.56	50.95±4.65	42.20±4.90	51.29±5.20	28.16±9.00	66.39±7.79
	4	36.54±4.74	39.83±4.71	33.59±4.80	40.17±5.11	18.20±7.62	41.56±9.34
	10	28.63±4.40	31.90±4.16	26.88±4.59	32.92±5.00	13.62±7.73	43.48±9.32
FedProx +BABU	1	48.53±5.15	57.44±4.72	46.25±5.31	63.12±5.25	33.13±8.11	78.86±5.70
	4	37.17±4.41	45.26±4.76	33.86±5.44	50.18±5.14	22.94±9.90	75.71±5.33
	10	27.79±3.95	35.68±4.34	27.48±5.22	42.37±6.10	15.66±8.29	67.15±7.10

Initial and personalized accuracy of FedProx and FedProx+BABU with μ=0.01

FedProx+BABU performs better than the personalization of FedAvg!

ANTS LAB

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Conclusion

- Investigate the connection between a single global model and fine-tuned personalized models by analyzing the FedAvg algorithm at the client level and show that training the head using centralized data has a negative impact on personalization.
- Demonstrate that a fixed random classifier can have comparable performance to a learned classifier.
- Propose a novel algorithm, FedBABU, that reduces the update and aggregation parts from the entire model to the body of the model during federated training.
- Show that FedBABU is efficient, particularly under more significant data heterogeneity.
- Adapt the body update and body aggregation idea to the regularization-based federated learning algorithm (such as FedProx).

THANK YOU

Pavana Prakash

Department of Electrical and Computer Engineering
University of Houston
Houston, TX

UNIVERSITY of HOUSTON | ENGINEERING