भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

IIT Hyderabad

Indian Institute of Technology Hyderabad

# FedALA :Adaptive Local Aggregation for Personalized Federated Learning


Presented by
T.A.

Debika Samanta
Phanindra Revanth

Under Guidance of

Prof. C. Krishna Mohan



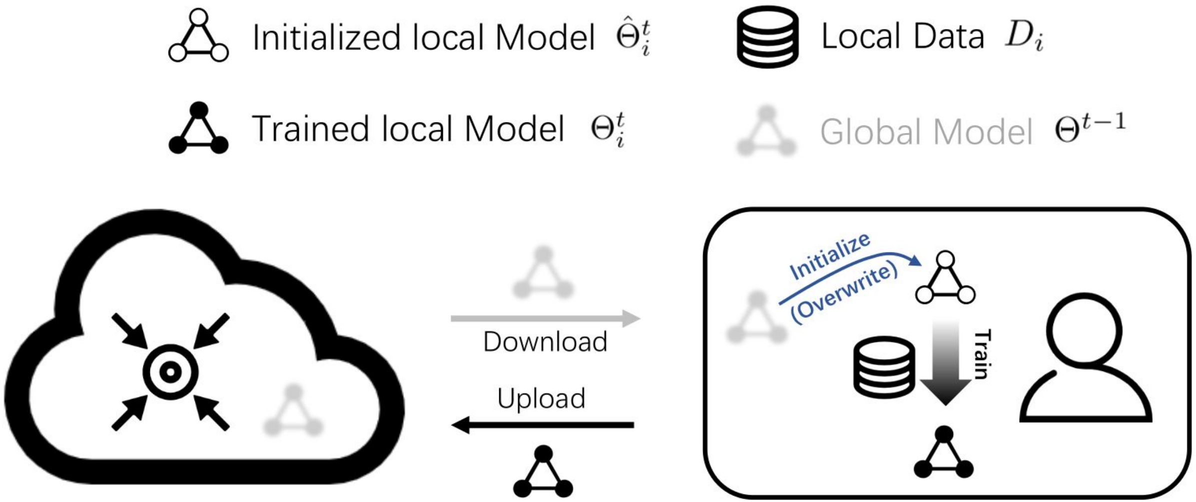
Zhang,Jianqing, Yang Hua,Hao Wang,Tao Song,Zhengui Xue,Ruhul Ma,and Haibing Guan, "Fedala: Adaptive local aggregation for personalized federated learning".
In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 9, pp. 11237-11244,2023

# Index

- Overview
- Issue in the current system
- Methodology used in the implementation
- Flow of the process
- Results of the implementation
- Novel idea
- Conclusion

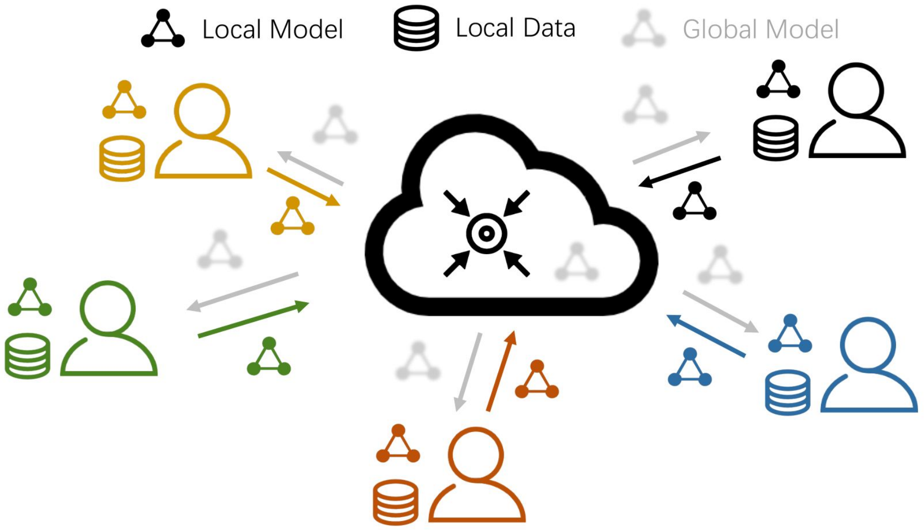
# Federated Learning

- ML technique in which model is trained across multiple devices.
- Protect privacy without uploading local data to central server.
- Learn AI model among clients by sharing model with server.
- Finally output single global model.



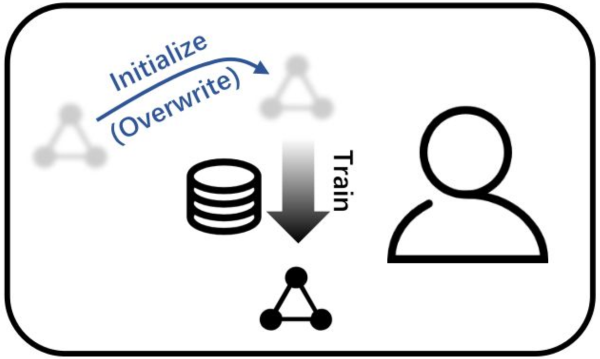
# Issues in Federated Learning

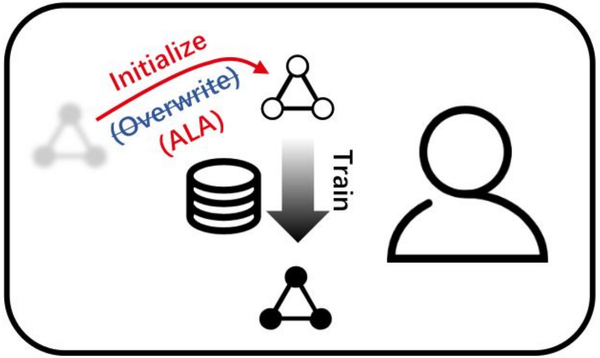
- Statistical heterogeneity
- Poor generalization ability of the global model on each client.



# Motivation of FedALA

- Almost all the other FL models overwrites local model with the entire global model for local initialization in each iteration.
- Only the desired information that improves the quality of the local model is beneficial for the client.
- Desired and undesired information exist in the global model resulting in poor generalization.

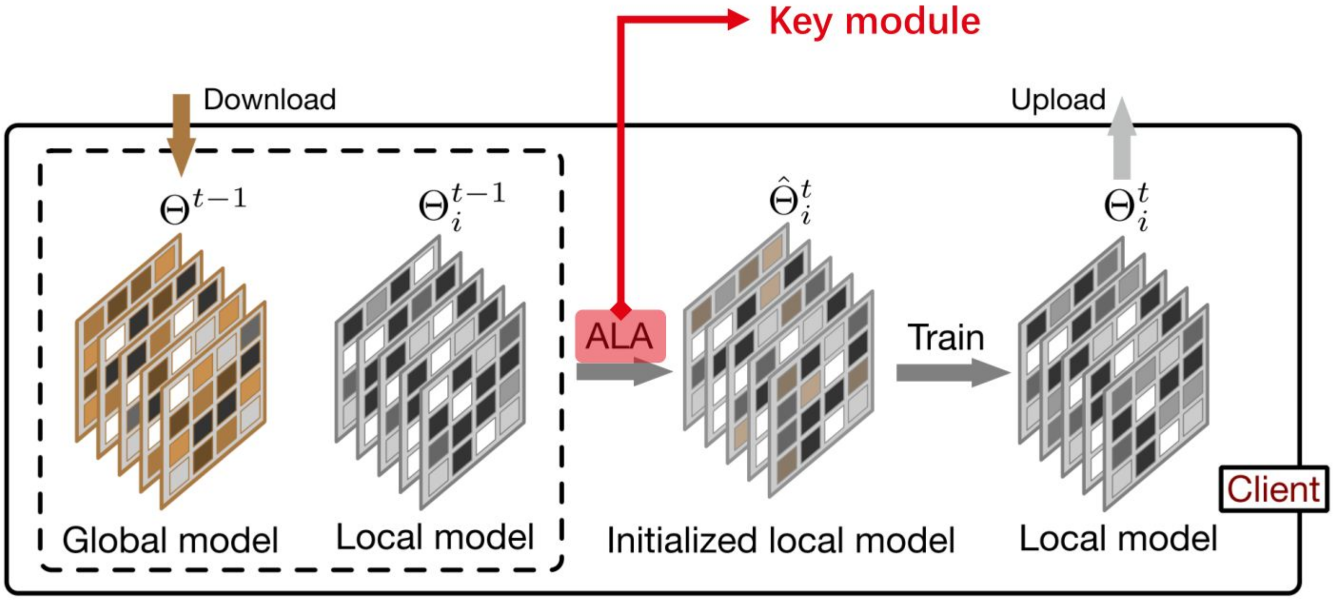


# FedALA

**ALA** : Adaptively aggregate the global model and local model for initialization.

**Train** : Train the local model on the local data.




Workflow on the client in one iteration

# FedALA : ALA module

- N clients
- Private training data $D_1, \ldots, D_N$, respectively.
- $D_1, \ldots, D_N$ are sampled from N distinct distributions.
- Individual local models $\hat{\Theta}_1, \ldots, \hat{\Theta}_N$
- Using $\{D_i\}_{i=1}^{N}$ for each client i.
- Objective:

Global loss $\sum_{i=1}^{N} k_i \mathcal{L}_i$

$$k_i = |D_i| / \sum_{j=1}^{N} |D_j|$$

$$\{\hat{\Theta}_1, \ldots, \hat{\Theta}_N\} = \arg\min \mathcal{G}(\mathcal{L}_1, \ldots, \mathcal{L}_N),$$

Total samples of client i

$$\mathcal{L}_i = \mathcal{L}(\hat{\Theta}_i, D_i; \Theta), \forall i \in [N]$$

Loss function

# FedALA : ALA module

- Element-wisely aggregate the global model and local model in an adaptive way

$$\hat{\Theta}_i^t := \Theta_i^{t-1} \odot \underbrace{W_{i,1}}_{\text{Trainable weights}} + \Theta^{t-1} \odot \underbrace{W_{i,2}}_{\text{Trainable weights}},$$

$$s.t. \quad \underbrace{w_1^q + w_2^q = 1, \forall \text{ valid } q}_{\text{Hard to learn weights with constraints}}$$

- Combine $W_{i,1}$ and $W_{i,2}$

$$\hat{\Theta}_i^t := \Theta_i^{t-1} + \underbrace{(\Theta^{t-1} - \Theta_i^{t-1})}_{\text{Called "update"}} \odot \underbrace{W_i}_{\text{Trainable weights}}$$

View the local aggregation as an **update process** for old local model

# FedALA : ALA module

- Element-wisely aggregate the global model and local model in an adaptive way

$$\hat{\Theta}_i^t := \Theta_i^{t-1} \odot W_{i,1} + \Theta^{t-1} \odot W_{i,2},$$
$$s.t. \quad w_1^q + w_2^q = 1, \forall \text{ valid } q$$

Trainable weights

Hard to learn weights with constraints

- Combine $W_{i,1}$ and $W_{i,2}$

$$\hat{\Theta}_i^t := \Theta_i^{t-1} + (\Theta^{t-1} - \Theta_i^{t-1}) \odot W_i$$

Trainable weights

Called "update"

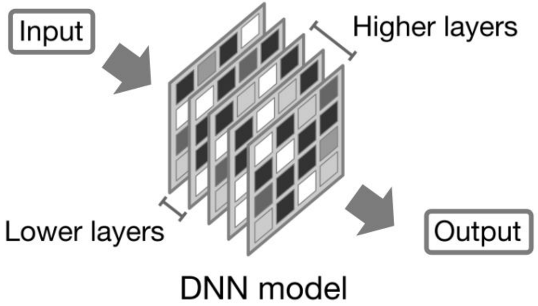
ALA covers the entire model

  - remove constraints
  - with weight clipping[1]

$$\boxed{\begin{aligned} \sigma(w) &= \max(0, \min(1, w)) \\ & w \in [0, 1], \forall w \in W_i \end{aligned}}$$

# FedALA : ALA module

- The lower layers in the DNN learn more general information than the higher layers[2]




- **Only apply ALA on $p$ higher layers**
- **Still overwrite the lower layers with global parameters**

$$\hat{\Theta}_i^t := \Theta_i^{t-1} + (\Theta^{t-1} - \Theta_i^{t-1}) \odot [\mathbf{1}^{|\Theta_i|-p}; W_i^p]$$

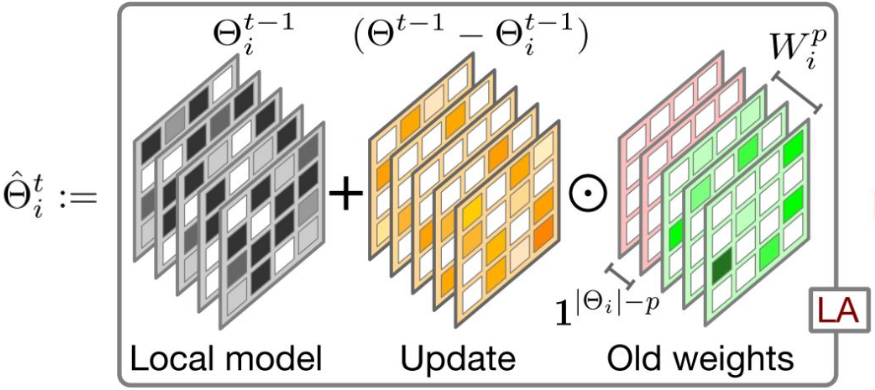
Fewer weights to train in ALA

Less computation overhead

# FedALA : ALA module

- Only apply ALA on $p$ higher layers
- Still overwrite the lower layers with global parameters

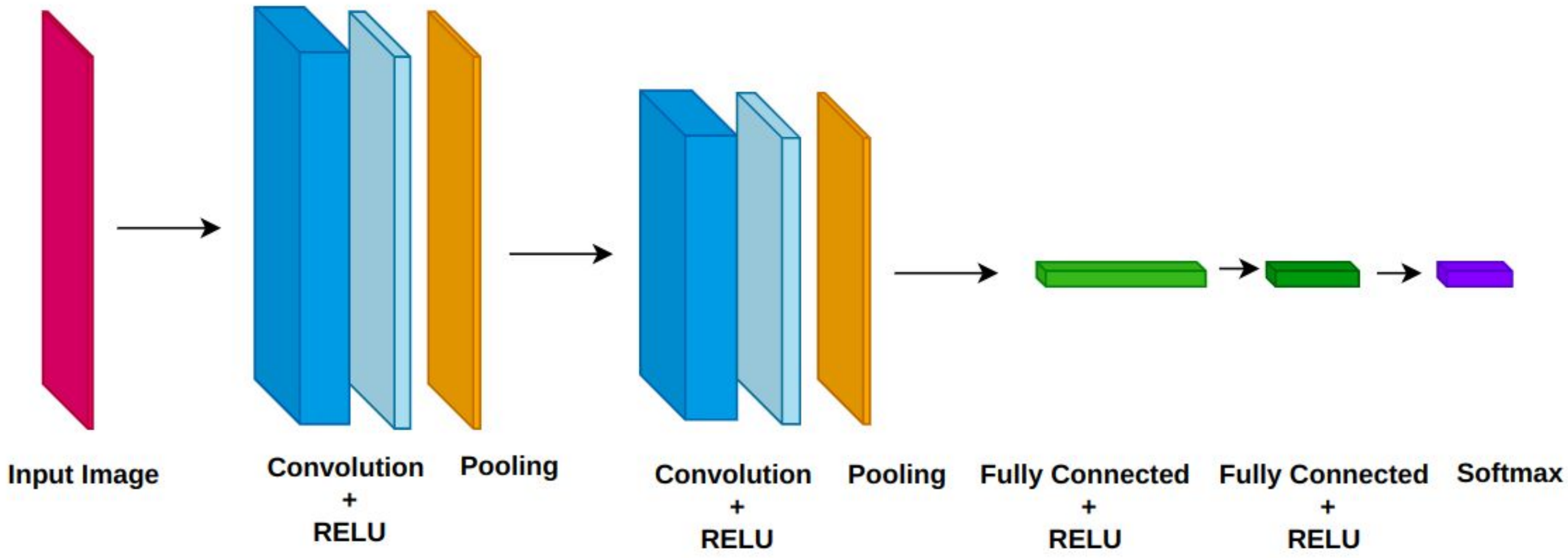
$$\hat{\Theta}_i^t := \Theta_i^{t-1} + (\Theta^{t-1} - \Theta_i^{t-1}) \odot [\mathbf{1}^{|\Theta_i|-p}; W_i^p]$$




How to train weights?

Local aggregation (LA)

# FedAvgCNN



Input Image
Convolution
+
RELU
Pooling
Convolution
+
RELU
Pooling
Fully Connected
+
RELU
Fully Connected
+
RELU
Softmax

# Resnet 18

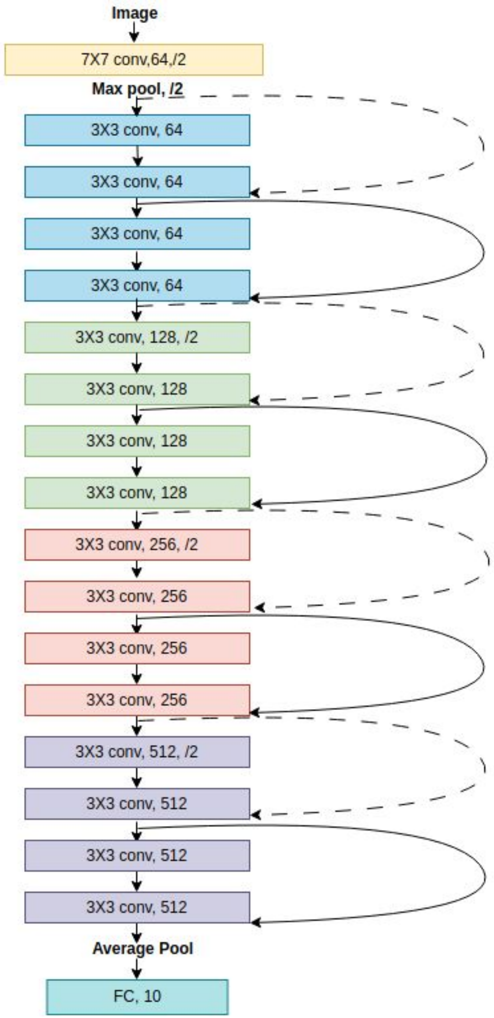
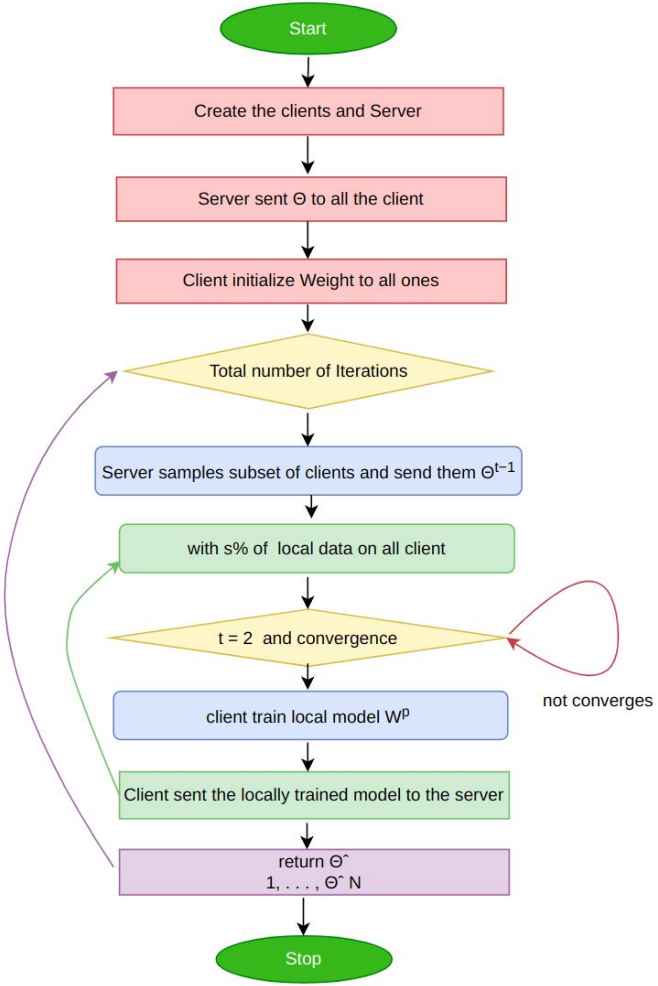


Image
7X7 conv,64,/2
Max pool, /2
3X3 conv, 64
3X3 conv, 64
3X3 conv, 64
3X3 conv, 64
3X3 conv, 128, /2
3X3 conv, 128
3X3 conv, 128
3X3 conv, 128
3X3 conv, 256, /2
3X3 conv, 256
3X3 conv, 256
3X3 conv, 256
3X3 conv, 512, /2
3X3 conv, 512
3X3 conv, 512
3X3 conv, 512
Average Pool
FC, 10

# Flow chart of FedALA



# FedALA: analysis

- Two main equations (omitting $p$):

$$\hat{\Theta}_i^t := \Theta_i^{t-1} + (\Theta^{t-1} - \Theta_i^{t-1}) \odot W_i$$

$$W_i^p \leftarrow W_i^p - \eta \nabla_{W_i^p} \mathcal{L}(\hat{\Theta}_i^t, D_i^{s,t}; \Theta^{t-1})$$

Denote $\mathcal{L}(\hat{\Theta}_i^t, D_i^{s,t}; \Theta^{t-1})$ as $\mathcal{L}_i^t$

- Rewrite the gradient term as $\nabla_{W_i} \mathcal{L}_i^t = \eta(\Theta^{t-1} - \Theta_i^{t-1}) \odot \nabla_{\hat{\Theta}_i} \mathcal{L}_i^t$

- We view updating $W_i$ as updating $\hat{\Theta}_i^t$

$$\hat{\Theta}_i^t \leftarrow \hat{\Theta}_i^t - \eta(\Theta^{t-1} - \Theta_i^{t-1}) \odot (\Theta^{t-1} - \Theta_i^{t-1}) \odot \nabla_{\hat{\Theta}_i} \mathcal{L}_i^t$$

# FedALA: analysis

- Two main equations (omitting $p$):

$$\hat{\Theta}_i^t := \Theta_i^{t-1} + (\Theta^{t-1} - \Theta_i^{t-1}) \odot W_i$$

$$W_i^p \leftarrow W_i^p - \eta \nabla_{W_i^p} \mathcal{L}(\hat{\Theta}_i^t, D_i^{s,t}; \Theta^{t-1})$$

Denote $\mathcal{L}(\hat{\Theta}_i^t, D_i^{s,t}; \Theta^{t-1})$ as $\mathcal{L}_i^t$

- Rewrite the gradient term as $\nabla_{W_i} \mathcal{L}_i^t = \eta(\Theta^{t-1} - \Theta_i^{t-1}) \odot \nabla_{\hat{\Theta}_i} \mathcal{L}_i^t$

- We view updating $W_i$ as updating $\hat{\Theta}_i^t$

$$\hat{\Theta}_i^t \leftarrow \hat{\Theta}_i^t - \eta \underbrace{(\Theta^{t-1} - \Theta_i^{t-1}) \odot (\Theta^{t-1} - \Theta_i^{t-1})}_{\text{* Dynamic generic information}} \odot \nabla_{\hat{\Theta}_i} \mathcal{L}_i^t$$

# Hyperparameters

- **p: the range of ALA**

  To reduce computation overhead, we introduce a hyperparameter p to control the range of ALA by applying it on p higher layers and overwriting the parameters in the lower layers.

- **s%: the percent of local data in ALA**

  To further reduce computation overhead, we randomly sample s% of Di in iteration t for each client.

# FedALA: results for computation reduction

Reduce computation overhead with small p (applying ALA on p higher layers)

The test accuracy (%) and the number of trainable parameters (in millions) of FedALA on Tiny-ImageNet using ResNet-18 ($s = 80$)

| | $p=6$ | $p=5$ | $p=4$ | $p=3$ | $p=2$ | $p=1$ |
|---|---|---|---|---|---|---|
| Acc. | 41.71 | 41.54 | 41.62 | 41.86 | **42.47** | 41.94 |
| Param. | 11.182 | 11.172 | 11.024 | 10.499 | 8.399 | 0.005 |

**Accuracy hardly changes with different *p***

**Parameter amount decreases greatly with small *p***

# FedALA: results for computation reduction

Reduce computation overhead with small p (applying ALA on p higher layers)

The test accuracy (%) and the number of trainable parameters (in millions) of FedALA on Tiny-ImageNet using ResNet-18 ($s = 80$)

| | $p=6$ | $p=5$ | $p=4$ | $p=3$ | $p=2$ | $p=1$ |
|---|---|---|---|---|---|---|
| Acc. | 41.71 | 41.54 | 41.62 | 41.86 | **42.47** | 41.94 |
| Param. | 11.182 | 11.172 | 11.024 | 10.499 | 8.399 | 0.005 |

**Accuracy hardly changes with different $p$**

**Parameter amount decreases greatly with small $p$**

**Set $p = 1$ to greatly reduce computation overhead**

# FedALA: results for computation reduction

The test accuracy (%) of FedALA on Tiny-ImageNet using ResNet-18 ($p = 1$)

| | $s=5$ | $s=10$ | $s=20$ | $s=40$ | $s=60$ | $s=80$ | $s=100$ |
|---|---|---|---|---|---|---|---|
| Acc. | 39.53 | 40.62 | 40.02 | 40.23 | 41.11 | 41.94 | **42.11** |

Accuracy decreases with smaller $s$

# FedALA: results for computation reduction

Reduce computation overhead with small s (training weights with s% local data)

The test accuracy (%) of FedALA on Tiny-ImageNet using ResNet-18 ($p = 1$)

| | $s = 5$ | $s = 10$ | $s = 20$ | $s = 40$ | $s = 60$ | $s = 80$ | $s = 100$ |
|---|---|---|---|---|---|---|---|
| Acc. | 39.53 | 40.62 | 40.02 | 40.23 | 41.11 | 41.94 | **42.11** |

**Accuracy decreases with smaller $s$**

**Set $s = 80$ to reduce computation overhead**

**FedALA performs well with only 5% local data for ALA**

# Heterogeneity Setting:

- **Practical Heterogeneity Setting :**

  Clients are separated into groups, which is based on the similarity among clients.

- **Pathological Heterogeneity Setting :**

  controlled by dirichlet distribution denoted by Dir(**β**).

  The smaller the **β** is the more heterogeneous the setting is.

# Performance Comparison

FedALA outperforms 13 traditional FL and pFL methods

The test accuracy (%) in the pathological heterogeneous setting and practical heterogeneous setting.

| Settings | Pathological heterogeneous setting | | | Practical heterogeneous setting | | | | |
|---|---|---|---|---|---|---|---|---|
| Methods | MNIST | Cifar10 | Cifar100 | Cifar10 | Cifar100 | TINY | TINY* | AG News |
| FedAvg | 97.93±0.05 | 55.09±0.83 | 25.98±0.13 | 59.16±0.47 | 31.89±0.47 | 19.46±0.20 | 19.45±0.13 | 79.57±0.17 |
| FedProx | 98.01±0.09 | 55.06±0.75 | 25.94±0.16 | 59.21±0.40 | 31.99±0.41 | 19.37±0.22 | 19.27±0.23 | 79.35±0.23 |
| FedAvg-C | 99.79±0.00 | 92.13±0.03 | 66.17±0.03 | 90.34±0.01 | 51.80±0.02 | 30.67±0.08 | 36.94±0.10 | 95.89±0.25 |
| FedProx-C | 99.80±0.04 | 92.12±0.03 | 66.07±0.08 | 90.33±0.01 | 51.84±0.07 | 30.77±0.13 | 38.78±0.52 | 96.10±0.22 |
| Per-FedAvg | 99.63±0.02 | 89.63±0.23 | 56.80±0.26 | 87.74±0.19 | 44.28±0.33 | 25.07±0.07 | 21.81±0.54 | 93.27±0.25 |
| FedRep | 99.77±0.03 | 91.93±0.14 | 67.56±0.31 | 90.40±0.24 | 52.39±0.35 | 37.27±0.20 | 39.95±0.61 | 96.28±0.14 |
| pFedMe | 99.75±0.02 | 90.11±0.10 | 58.20±0.14 | 88.09±0.32 | 47.34±0.46 | 26.93±0.19 | 33.44±0.33 | 91.41±0.22 |
| Ditto | 99.81±0.00 | 92.39±0.06 | 67.23±0.07 | 90.59±0.01 | 52.87±0.64 | 32.15±0.04 | 35.92±0.43 | 95.45±0.17 |
| FedAMP | 99.76±0.02 | 90.79±0.16 | 64.34±0.37 | 88.70±0.18 | 47.69±0.49 | 27.99±0.11 | 29.11±0.15 | 94.18±0.09 |
| FedPHP | 99.73±0.00 | 90.01±0.00 | 63.09±0.04 | 88.92±0.02 | 50.52±0.16 | 35.69±3.26 | 29.90±0.51 | 94.38±0.12 |
| FedFomo | 99.83±0.00 | 91.85±0.02 | 62.49±0.22 | 88.06±0.02 | 45.39±0.45 | 26.33±0.22 | 26.84±0.11 | 95.84±0.15 |
| APPLE | 99.75±0.01 | 90.97±0.05 | 65.80±0.08 | 89.37±0.11 | 53.22±0.20 | 35.04±0.47 | 39.93±0.52 | 95.63±0.21 |
| PartialFed | 99.86±0.01 | 89.60±0.13 | 61.39±0.12 | 87.38±0.08 | 48.81±0.20 | 35.26±0.18 | 37.50±0.16 | 85.20±0.16 |
| FedALA | **99.88±0.01** | **92.44±0.02** | **67.83±0.06** | **90.67±0.03** | **55.92±0.03** | **40.54±0.02** | **41.94±0.05** | **96.52±0.08** |

# Performance Comparison

The test accuracy (%) in the pathological heterogeneous setting and practical heterogeneous setting.

| Settings | Pathological heterogeneous setting | | | Practical heterogeneous setting | | | | |
|---|---|---|---|---|---|---|---|---|
| Methods | MNIST | Cifar10 | Cifar100 | Cifar10 | Cifar100 | TINY | TINY* | AG News |
| FedAvg | 97.93±0.05 | 55.09±0.83 | 25.98±0.13 | 59.16±0.47 | 31.89±0.47 | 19.46±0.20 | 19.45±0.13 | 79.57±0.17 |
| FedProx | 98.01±0.09 | 55.06±0.75 | 25.94±0.16 | 59.21±0.40 | 31.99±0.41 | 19.37±0.22 | 19.27±0.23 | 79.35±0.23 |
| FedAvg-C | 99.79±0.00 | 92.13±0.03 | 66.17±0.03 | 90.34±0.01 | 51.80±0.02 | 30.67±0.08 | 36.94±0.10 | 95.89±0.25 |
| FedProx-C | 99.80±0.04 | 92.12±0.03 | 66.07±0.08 | 90.33±0.01 | 51.84±0.07 | 30.77±0.13 | 38.78±0.52 | 96.10±0.22 |
| Per-FedAvg | 99.63±0.02 | 89.63±0.23 | 56.80±0.26 | 87.74±0.19 | 44.28±0.33 | 25.07±0.07 | 21.81±0.54 | 93.27±0.25 |
| FedRep | 99.77±0.03 | 91.93±0.14 | 67.56±0.31 | 90.40±0.24 | 52.39±0.35 | 37.27±0.20 | 39.95±0.61 | 96.28±0.14 |
| pFedMe | 99.75±0.02 | 90.11±0.10 | 58.20±0.14 | 88.09±0.32 | 47.34±0.46 | 26.93±0.19 | 33.44±0.33 | 91.41±0.22 |
| Ditto | 99.81±0.00 | 92.39±0.06 | 67.23±0.07 | 90.59±0.01 | 52.87±0.64 | 32.15±0.04 | 35.92±0.43 | 95.45±0.17 |
| FedAMP | 99.76±0.02 | 90.79±0.16 | 64.34±0.37 | 88.70±0.18 | 47.69±0.49 | 27.99±0.11 | 29.11±0.15 | 94.18±0.09 |
| FedPHP | 99.73±0.00 | 90.01±0.00 | 63.09±0.04 | 88.92±0.02 | 50.52±0.16 | 35.69±3.26 | 29.90±0.51 | 94.38±0.12 |
| FedFomo | 99.83±0.00 | 91.85±0.02 | 62.49±0.22 | 88.06±0.02 | 45.39±0.45 | 26.33±0.22 | 26.84±0.11 | 95.84±0.15 |
| APPLE | 99.75±0.01 | 90.97±0.05 | 65.80±0.08 | 89.37±0.11 | 53.22±0.20 | 35.04±0.47 | 39.93±0.52 | 95.63±0.21 |
| PartialFed | 99.86±0.01 | 89.60±0.13 | 61.39±0.12 | 87.38±0.08 | 48.81±0.20 | 35.26±0.18 | 37.50±0.16 | 85.20±0.16 |
| FedALA | **99.88±0.01** | **92.44±0.02** | **67.83±0.06** | **90.67±0.03** | **55.92±0.03** | **40.54±0.02** | **41.94±0.05** | **96.52±0.08** |

# Performance Comparison

| Sample setting | Dataset | Paper Results | Recreation Results |
|---|---|---|---|
| Pathological heterogeneous | MNIST | 99.88 | 99.63 |
| | Cifar10 | 92.44 | 91.40 |
| | Cifar100 | 67.83 | 52.93 |
| Practical heterogeneous | Cifar10 | 90.67 | 90.64 |
| | Cifar100 | 55.92 | 55.88 |
| | TINY(CNN) | 40.54 | 37.66 (136) |
| | TINY(Resnet-18) | 41.94 | 31.42 (12) |
| | AG News | 96.52 | 96.40 (696) |

# Novelty:

- Parallelizing the local client initialization on each client with the help of multithreading.
- This will be in contrast to the sequential initialization of client (using for loop in the current code).

# Conclusion

- Contributions of FedALA:
    - **Adaptively aggregates** the global model and local model towards the local objective to **capture the desired information** from the global model.
    - Outperforms **11 SOTA** methods by up to 3.2% in test accuracy **without additional communication overhead** in each iteration.
    - The ALA module in FedALA can be **directly applied to existing FL methods** to enhance their performance by up to 24.19%.

# References


- A Downsampled Variant of Imagenet as an Alternative to the Cifar Datasets. arXiv preprint arXiv:1707.08819. Collins, L.; Hassani, H.; Mokhtari, A.; and Shakkottai, S. 2021.
- Exploiting Shared Representations for Personalized Federated Learning. In ICML. Courbariaux, M.; Hubara, I.; Soudry, D.; El-Yaniv, R.; and Bengio, Y. 2016.
- Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1. arXiv preprint arXiv:1602.02830. Fallah, A.; Mokhtari, A.; and Ozdaglar, A. 2020.
- Personalized Federated Learning with Theoretical Guarantees: A Model-Agnostic Meta-Learning Approach. In NeurIPS. Finn, C.; Abbeel, P.; and Levine, S. 2017.
- Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In ICML. He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016.
- Deep Residual Learning for Image Recognition. In CVPR. Huang, Y.; Chu, L.; Zhou, Z.; Wang, L.; Liu, J.; Pei, J.; and Zhang, Y. 2021.
- Personalized Cross-Silo Federated Learning on Non-IID Data. In AAAI. Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2017.

Thankyou ….