GradMA: A Gradient-Memory-based Accelerated Federated Learning with Alleviated Catastrophic Forgetting




Kangyang Luo, Xiang Li, Shiyun Lan, Ming Gao

East China Normal University, Shanghai, China

Paper tag: TUE-AM-353

➢ Conclusion

Federated Learning (FL) has emerged as a de facto machine learning area and received rapid increasing research interests from the community. However, catastrophic forgetting caused by data heterogeneity and partial participation poses distinctive challenges for FL, which are detrimental to the performance.

To tackle the problems, we propose a new FL approach, dubbed as GradMA (**Grad**ient-**M**emory-based **A**ccelerated Federated Learning), which takes inspiration from continual learning to both correct the server-side and worker-side update directions as well as take full advantage of server's rich computing and memory resources. Furthermore, we elaborate a memory reduction strategy to enable GradMA to accommodate FL with a large scale of workers.We then analyze convergence of GradMA theoretically under the smooth non-convex setting and show that its convergence rate achieves a linear speed up w.r.t the increasing number of sampled active workers. At last, our extensive experiments on various image classification tasks show that GradMA achieves significant performance gains in accuracy and communication efficiency compared to SOTA baselines.

➢ GradMA

- **Correcting gradient for the worker side**

Algorithm 2 Worker_Update(x', x, η_l, I)

1: Sets $x_{-1}^{(i)} = x'$, $x_0^{(i)} = x$.
2: **for** $\tau = 0, 1, \ldots, I-1$ **do**
3: $\quad g_\tau^{(i)} = \nabla f_i(x_\tau^{(i)})$,
4: $\quad G_\tau^{(i)} = [\nabla f_i(x_{\tau-1}^{(i)}), \nabla f_i(x), x_\tau^{(i)} - x_t]$,
5: $\quad \tilde{g}_\tau^{(i)} = \mathrm{QP}_l(g_\tau^{(i)}, G_\tau^{(i)})$,
6: $\quad x_{\tau+1}^{(i)} = x_\tau^{(i)} - \eta_l \tilde{g}_\tau^{(i)}$.
7: **end for**
8: **Output:** $x_I^{(i)}$.

$$\tilde{g}_\tau^{(i)} = G_\tau^{(i)} z_\tau^\star + g_\tau^{(i)}$$
$$= z_{\tau,1}^\star \nabla f_i(x_{\tau-1}^{(i)}) + z_{\tau,2}^\star \nabla f_i(x) + z_{\tau,3}^\star (x_\tau^{(i)} - x_t) + g_\tau^{(i)}$$

➢ GradMA

- **Correcting update direction for the server side**

Algorithm 3 Server_Update($[\boldsymbol{d}_{t+1}^{(i)}, i \in \mathcal{S}_t]$, $\tilde{\boldsymbol{m}}_t$, $\boldsymbol{D}$, η_g, β_1, β_2, buf, new_buf)

$\boldsymbol{d}_{t+1} = \frac{1}{S}\sum_{i\in\mathcal{S}_t} \boldsymbol{d}_{t+1}^{(i)}$, $\boldsymbol{m}_{t+1} = \beta_1 \tilde{\boldsymbol{m}}_t + \boldsymbol{d}_{t+1}$.
for $c(i) \in buf$ **do**
  if $i \in \mathcal{S}_t$ **then**
    $\boldsymbol{D}[i] \leftarrow \begin{cases} \beta_2 \boldsymbol{D}[i] + \boldsymbol{d}_{t+1}^{(i)}, c(i) \notin new\_buf \\ \boldsymbol{d}_{t+1}^{(i)}, c(i) \in new\_buf \end{cases}$.
  else if $i \notin \mathcal{S}_t$ **then**
    $\boldsymbol{D}[i] \leftarrow \beta_2 \boldsymbol{D}[i]$.
  end if
end for
$\tilde{\boldsymbol{m}}_{t+1} = \mathrm{QP}_g(\boldsymbol{m}_{t+1}, \boldsymbol{D})$, $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_g \tilde{\boldsymbol{m}}_{t+1}$.
Output: $\boldsymbol{D}$, $\boldsymbol{x}_{t+1}$, $\tilde{\boldsymbol{m}}_{t+1}$.

$$\boldsymbol{d}_{t+1} = \frac{1}{S}\sum_{i\in\mathcal{S}_t} \boldsymbol{d}_{t+1}^{(i)}, \boldsymbol{m}_{t+1} = \beta_1 \tilde{\boldsymbol{m}}_t + \boldsymbol{d}_{t+1},$$

$$\boldsymbol{D}[i] \leftarrow \begin{cases} \beta_2 \boldsymbol{D}[i] + \boldsymbol{d}_{t+1}^{(i)}, i \in \mathcal{S}_t \\ \beta_2 \boldsymbol{D}[i], i \notin \mathcal{S}_t \end{cases},$$

$$\tilde{\boldsymbol{m}}_{t+1} = \mathrm{QP}_g(\boldsymbol{m}_{t+1}, \boldsymbol{D}), \boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_g \tilde{\boldsymbol{m}}_{t+1}.$$

➢ GradMA

- **A Practical Memory Reduction Strategy**

Algorithm 4 mem_red$(m, \mathcal{S}, c, \boldsymbol{D}, buf, new_buf)$

for $i \in \mathcal{S}$ **do**
  if $c(i) \in buf$ **then**
    $c(i) \leftarrow c(i) + 1$.
  else if $c(i) \notin buf$ **then**
    if $Length(buf) = m$ **then**
      $old\_buf = \{\}$.
      for $k \in buf$ **do**
        if $k \notin \mathcal{S}$ **then**
          $old\_buf \leftarrow old\_buf \cup \{c(k)\}$.
        end if
      end for
      Discarding $c(i')$ with the smallest value from old_buf and set $c(i') = 0$.
      Discarding $\boldsymbol{D}[i']$ from memory state $\boldsymbol{D}$.
    end if
    $c(i) \leftarrow c(i) + 1$.
    $buf \leftarrow buf \cup \{c(i)\}$.
    $new\_buf \leftarrow new\_buf \cup \{c(i)\}$.
  end if
end for
Output: $c, \boldsymbol{D}, buf, new\_buf$

➢ Convergence Results for GradMA

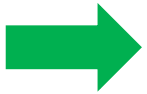
Assumption 1 (Global function below bounds). *Set* $f^* = \inf_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x})$ *and* $f^* > -\infty$.

Assumption 2 (L-smooth). $\forall i \in [N]$, *the local functions* f_i *are differentiable, and there exist constant* $L > 0$ *such that for any* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$.

Assumption 3 (Bounded data heterogeneity). *The degree of heterogeneity of the data distribution across workers can be quantified as* $\|\nabla f_i(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2 \leq \rho^2$, *for any* $i \in [N]$ *and some constant* $\rho \geq 0$.

Assumption 4 (Bounded optimal solution error for QP_l). *Given* $\boldsymbol{g}^{(i)} = \nabla f_i(\boldsymbol{x})$ *(see Alg. 2), then there exists* $\varepsilon_l > 0$ *such that* $\|\boldsymbol{g}^{(i)} - \tilde{\boldsymbol{g}}^{(i)}\|^2 \leq \varepsilon_l^2$.

Assumption 5 (Bounded optimal solution error for QP_g). *Given* $\beta_2 \in [0, 1)$ *and* $\boldsymbol{m}$ *(see Alg. 3), then there exists* $\varepsilon_g > 0$ *such that* $\|\boldsymbol{m} - \tilde{\boldsymbol{m}}\|^2 \leq \frac{\varepsilon_g^2}{1-\beta_2}$.



Theorem 1 *Assume Assumptions 1-5 exist. Let* $\eta_l \leq \frac{1}{160^{0.5}LI}$, $\eta_g\eta_l \leq \frac{(1-\beta_1)^2 S(N-1)}{IL(\beta_1 S(N-1)+4N(S-1))}$ *and* $320I^2\eta_l^2L^2 + \frac{64I\eta_g\eta_l L(1+40I^2\eta_l^2L^2)}{(1-\beta_1)^2}\frac{N-S}{S(N-1)} \leq 1$. *For all* $t \in [0, \cdots, T-1]$, *the following relationship generated by Alg. 1 holds:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] \leq \frac{8(1-\beta_1)(f(\boldsymbol{x}_0)-f^\star)}{I\eta_g\eta_l T} + C_1\varepsilon_l^2 + C_2\varepsilon_g^2 + C_3\rho^2,$$

where the expectation $\mathbb{E}$ *is w.r.t the sampled active workers per communication round, and* $C_1 = 8 + 320I^2\eta_l^2L^2 + \frac{64I\eta_g\eta_l L(1+40I^2\eta_l^2L^2)}{(1-\beta_1)^2}\frac{N-S}{S(N-1)}$, $C_2 = \frac{20\eta_g L}{(1-\beta_1)^2(1-\beta_2)I\eta_l} + \frac{8}{(1-\beta_2)I^2\eta_l^2}$, $C_3 = C_1 - 8$.

➢ Convergence Results for GradMA

Theorem 1 *Assume Assumptions 1-5 exist. Let* $\eta_l \leq \frac{1}{160^{0.5}LI}$, $\eta_g\eta_l \leq \frac{(1-\beta_1)^2S(N-1)}{IL(\beta_1 S(N-1)+4N(S-1))}$ *and* $320I^2\eta_l^2L^2 + \frac{64I\eta_g\eta_l L(1+40I^2\eta_l^2L^2)}{(1-\beta_1)^2}\frac{N-S}{S(N-1)} \leq 1$. *For all* $t \in [0, \cdots, T-1]$, *the following relationship generated by Alg. 1 holds:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] \leq \frac{8(1-\beta_1)(f(\boldsymbol{x}_0)-f^\star)}{I\eta_g\eta_l T} + C_1\varepsilon_l^2 + C_2\varepsilon_g^2 + C_3\rho^2,$$

where the expectation $\mathbb{E}$ *is w.r.t the sampled active workers per communication round, and* $C_1 = 8 + 320I^2\eta_l^2L^2 + \frac{64I\eta_g\eta_l L(1+40I^2\eta_l^2L^2)}{(1-\beta_1)^2}\frac{N-S}{S(N-1)}$, $C_2 = \frac{20\eta_g L}{(1-\beta_1)^2(1-\beta_2)I\eta_l} + \frac{8}{(1-\beta_2)I^2\eta_l^2}$, $C_3 = C_1 - 8$.

Corollary 1 *Assume Assumptions 1-5 exist. We set* $\eta_l = \frac{1}{T^{0.5}LI}$, $\eta_g = \frac{S^{0.5}}{I^{0.5}}$, $\varepsilon_l = \frac{1}{T^{0.5}}$ *and* $\varepsilon_g = \frac{I^{0.25}}{T^{0.75}S^{0.25}L}$. *For* $T \geq \max\left\{160, \frac{(\beta_1 S(N-1)+4N(S-1))^2}{I^2(1-\beta_1)^4S(N-1)^2}, \frac{(b+(b^2+1280)^{0.5})^2}{4}\right\}$ *where* $b = \frac{128(N-S)}{(1-\beta_1)^2I^{0.5}S^{0.5}(N-1)}$ *in Theorem 1, we have:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] = \mathcal{O}\left(\frac{I^{0.5}}{S^{0.5}T^{0.5}} + \frac{1}{T}\right).$$

➢ Empirical Study

- **Datasets**

  MNIST, CIFAR-10, CIFAR-100 and Tiny-ImageNet

- **Our methods**

  GradMA, GradMA-W (for Worker_Update()), GradMA-S (for Server_Update())

- **Baselines**

  FedAvg;

  FedAvg’s improvements on worker side (FedProx, MOON, FedMLB, Scaffold);

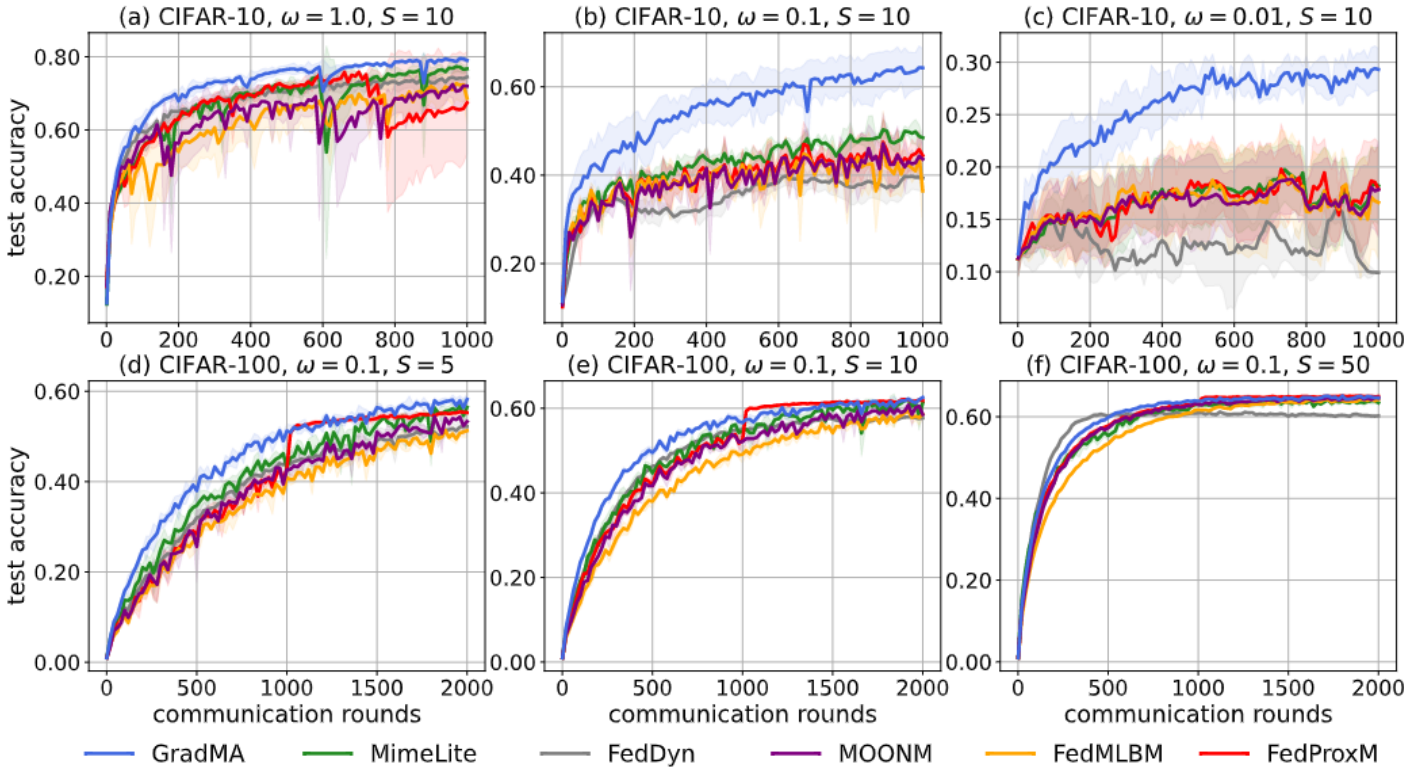
  FedAvg’s improvements on server side (FedAvgM, MIFA, MIFAM);

  FedAvg’s improvements on server side and woker side (FedProxM, FedMLBM, MOONM, Feddyn, MimeLite).

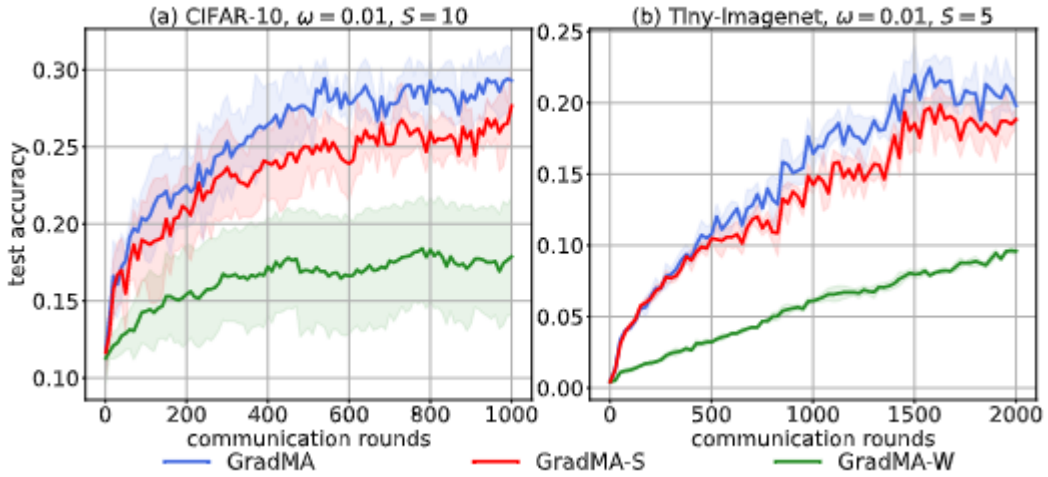
➢ Comparative Experiments

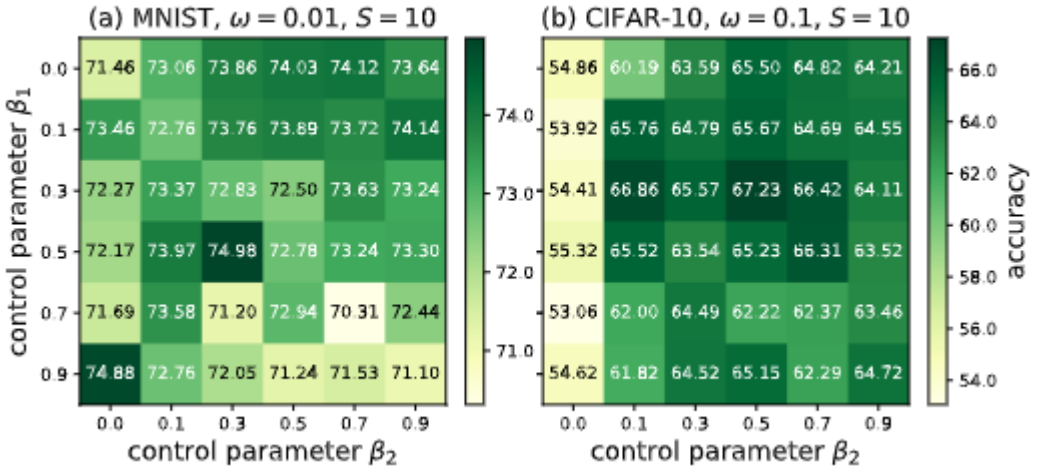
Table 1. Top test accuracy (%) overview given different FL scenarios.

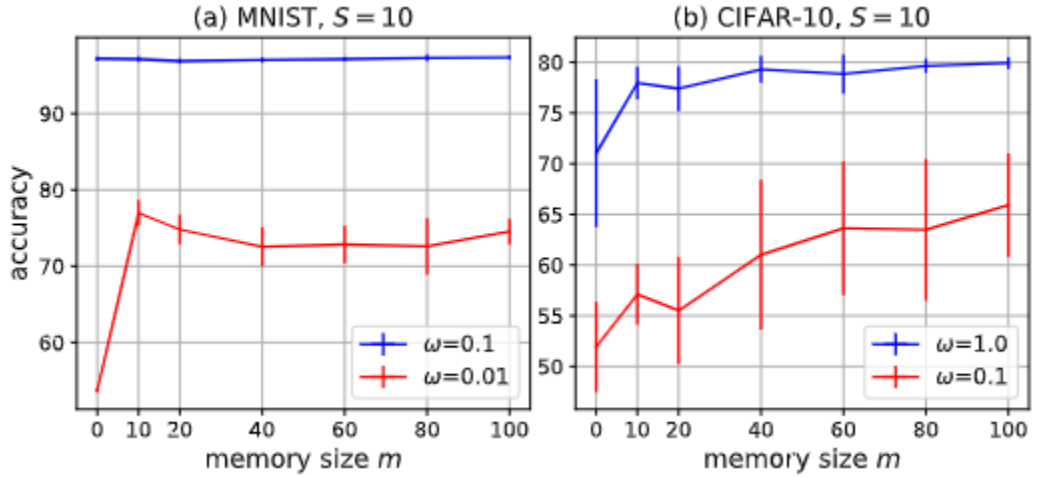
Alg.s	MNIST+NN, $S=10$			MNIST+NN, $\omega=0.01$		CIFAR-10+Lenet-5, $S=10$			CIFAR-100+VGG-11, $\omega=0.1$			Tiny-Imagenet+Resnet20, (ω, S)		
	$\omega=1.0$	$\omega=0.1$	$\omega=0.01$	$S=5$	$S=50$	$\omega=1.0$	$\omega=0.1$	$\omega=0.01$	$S=5$	$S=10$	$S=50$	$(0.01,5)$	$(1.0,5)$	$(1.0,10)$
FedAvg	98.22±0.05	97.11±0.39	46.19±1.29	49.65±3.88	68.32±4.16	69.48±8.28	47.86±5.26	**20.97±3.73**	56.02±0.37	61.22±0.16	64.78±0.43	7.50±0.32	41.80±0.55	42.90±0.12
FedProx	98.16±0.09	97.19±0.31	46.82±0.96	49.89±3.67	67.97±4.27	71.58±4.66	48.63±4.92	20.40±3.85	55.94±0.71	61.25±0.09	64.69±0.27	7.51±0.46	41.82±0.29	42.58±0.69
FedMLB	**98.31±0.06**	**97.26±0.40**	54.53±0.39	57.22±3.07	68.44±3.26	69.28±6.56	48.99±4.94	20.81±3.26	53.80±0.16	59.20±0.27	64.06±0.30	7.98±0.34	42.83±0.13	43.59±0.80
MOON	98.18±0.12	97.11±0.31	46.26±1.35	50.39±5.16	**68.75±4.50**	71.11±7.94	48.84±5.16	19.39±3.99	55.37±0.34	60.58±0.60	64.48±0.42	7.70±0.38	41.68±0.22	42.80±0.54
Scaffold	97.63±0.37	93.94±1.18	50.86±7.46	39.97±4.88	49.54±2.28	53.33±6.63	35.91±2.14	15.55±1.33	32.22±0.92	34.72±0.80	45.70±0.76	7.20±0.33	40.96±0.23	43.02±0.30
GradMA-W	98.15±0.10	97.01±0.23	**63.34±3.75**	**65.39±0.96**	65.13±2.54	**72.33±3.84**	**50.25±3.94**	18.99±4.06	**56.43±0.51**	**61.38±0.11**	**64.96±0.36**	**9.98±0.22**	**43.68±0.23**	**44.57±0.45**
FedAvgM	98.29±0.18	97.20±0.30	53.77±0.32	57.87±3.64	67.80±5.58	71.04±7.29	51.91±4.46	21.02±3.52	55.85±0.28	61.32±0.29	64.88±0.25	16.96±1.08	41.91±0.23	42.57±0.14
MIFA	98.02±0.12	96.88±0.56	66.92±2.53	56.04±3.92	52.84±4.89	71.41±5.81	50.60±11.87	23.78±2.04	50.37±1.02	58.74±0.42	64.71±0.31	8.88±0.33	41.42±0.22	42.83±0.13
MIFAM	98.02±0.15	96.90±0.44	67.15±2.23	55.28±6.05	53.35±6.84	73.48±1.37	52.13±9.71	24.17±1.24	49.30±0.86	58.91±0.24	64.61±0.33	12.01±0.32	41.94±0.06	43.17±0.09
GradMA-S	**98.38±0.09**	**97.35±0.28**	**74.52±1.71**	**75.93±0.97**	**69.09±3.83**	**78.76±1.96**	**64.60±5.87**	**28.41±2.43**	**59.08±0.43**	**63.23±0.22**	**65.63±0.35**	**20.93±1.49**	**48.83±1.06**	**49.65±0.72**
FedProxM	98.26±0.08	97.13±0.34	54.50±0.79	58.59±4.58	69.00±4.42	78.00±1.61	51.22±5.14	21.80±3.72	55.63±0.31	63.15±0.12	64.78±0.11	18.30±0.79	37.98±0.10	45.27±0.19
FedMLBM	98.26±0.16	**97.35±0.30**	61.12±1.48	64.12±4.17	68.78±3.28	73.70±4.62	49.90±5.82	21.53±2.93	53.91±0.78	60.44±0.34	64.85±0.18	17.32±0.82	44.62±0.32	45.18±0.27
MOONM	98.21±0.13	97.04±0.42	62.34±8.91	57.98±5.51	68.82±4.43	73.96±4.11	50.06±6.14	20.19±3.10	56.01±0.25	62.06±0.19	65.37±0.17	16.78±0.95	42.43±0.39	42.78±0.46
Feddyn	97.92±0.12	96.03±0.46	59.39±2.29	65.36±5.20	57.68±4.30	74.94±2.48	41.93±3.22	17.94±3.52	52.95±1.63	58.48±0.18	61.71±0.25	17.89±0.95	44.37±0.57	44.86±0.15
MimeLite	98.19±0.07	97.10±0.31	54.86±13.36	51.04±4.15	**69.41±4.15**	77.98±1.48	53.27±1.69	20.73±3.33	58.00±0.51	63.29±0.49	64.68±0.33	8.29±0.29	41.05±0.21	41.56±0.18
GradMA	**98.39±0.04**	97.34±0.35	**77.97±1.28**	**75.51±1.94**	66.68±3.03	**79.92±0.59**	**65.91±5.10**	**30.81±1.78**	**59.47±0.58**	**63.49±0.47**	**65.68±0.25**	**23.52±1.32**	**49.29±0.86**	**50.54±0.56**



➢ Ablation Study







Thank You!