HOHAI UNIVERSITY

Joint Focal Loss and Dominant Gradient Correction for Gradient Conflict in Federated Learning

Jiajun Wang[b], Yingchi Mao[a,b], Zibo Wang[b], Jun Wu[b] and Jie Wu[c]

[a] Key Laboratory of Water Big Data Technology of Ministry of Water Resources, Hohai University, Nanjing, China
[b] School of Computer and Information, Hohai University, Nanjing, China
[c] Center for Networked Computing, Temple University, Philadelphia USA

IEEE MASS
The 20th IEEE International Conference on Mobile Ad-Hoc and Smart Systems

Content

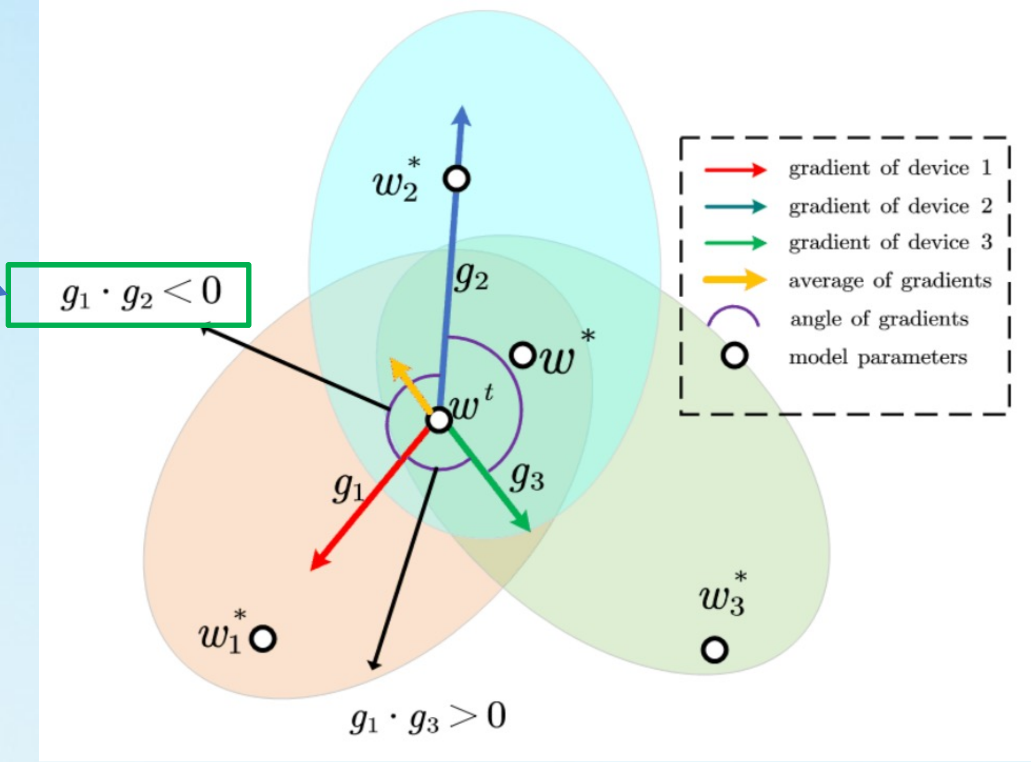
Introduction

■ Background

In federated learning, the data collected by IoT devices are typically heterogeneous .In this case, as the training advances the local model gradually converges to its local target optimum, resulting in a conflict among the upl oaded gradients in global aggregation phase.

Gradient conflict

Impairs the global accuracy



Introduction

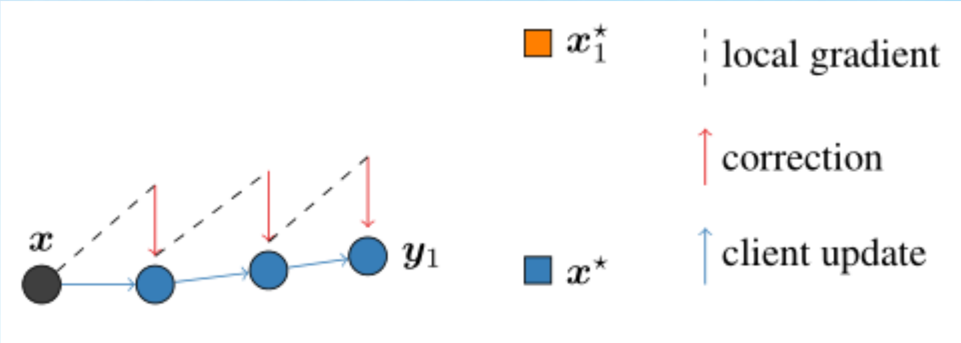
- Optimization object

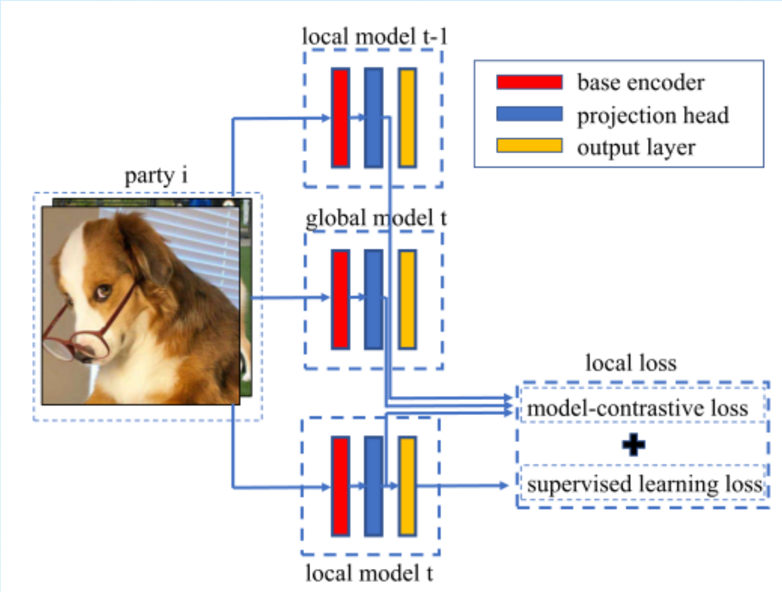
  Reduce the conflict among uploaded gradients.

- Goal

  Improve the accuracy of the global model.

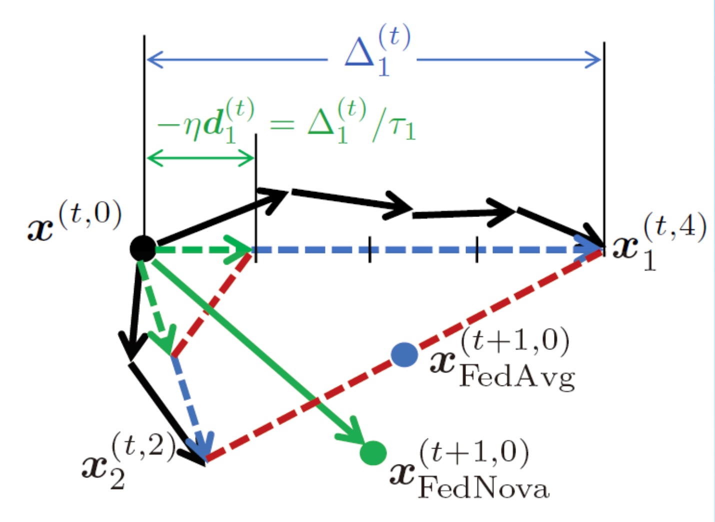
Local training optimization





- SCAFFOLD was proposed as a new stochastic control averaging algorithm
- Li et al corrected local updates of clients by injecting projection heads into the model with a model-level comparison learning method.
- Although local model divergence is suppressed, model preferences can still lead to gradient conflicts

Global aggregation optimization



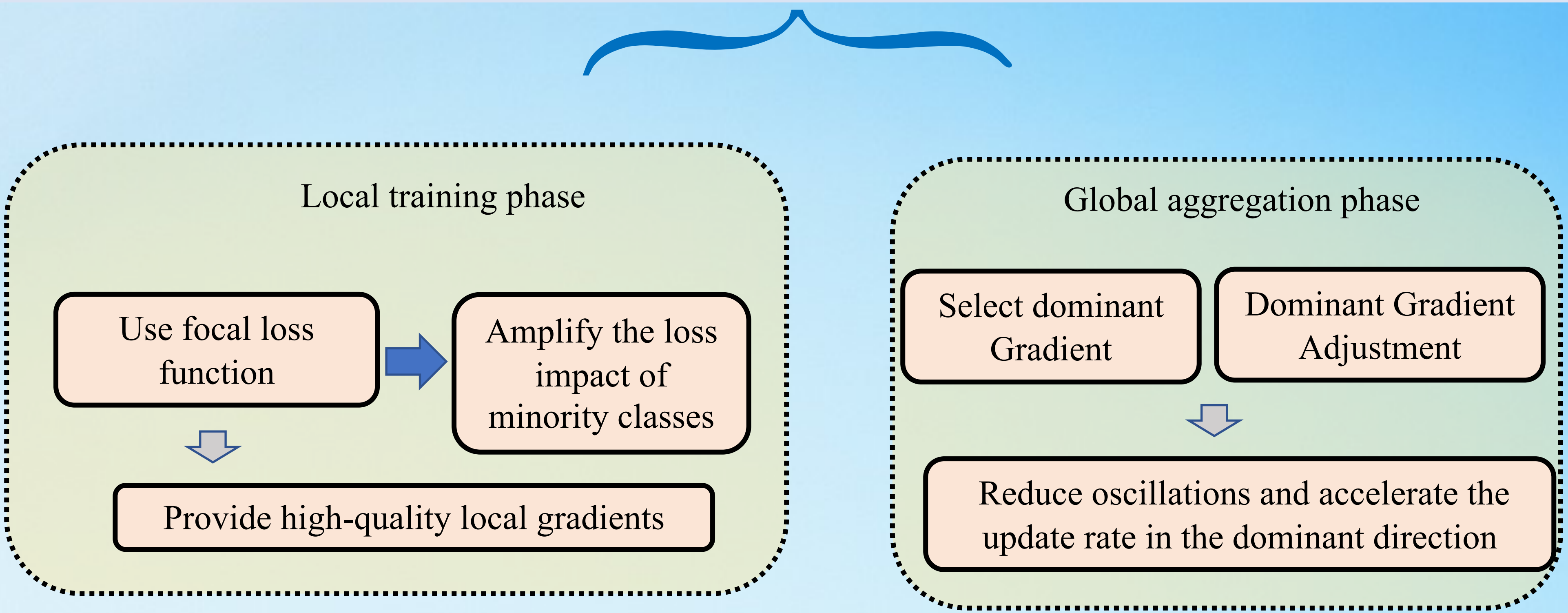
Algorithm 2 FedAdagrad, FedYogi, and FedAdam

Initialization: x_0, $v_{-1} \geq \tau^2$, decay parameters $\beta_1, \beta_2 \in [0,1)$
for $t = 0, \cdots, T-1$ **do**
  Sample subset $\mathcal{S}$ of clients
  $x_{i,0}^t = x_t$
  for each client $i \in \mathcal{S}$ **in parallel do**
    for $k = 0, \cdots, K-1$ **do**
      Compute an unbiased estimate $g_{i,k}^t$ of $\nabla F_i(x_{i,k}^t)$
      $x_{i,k+1}^t = x_{i,k}^t - \eta_l g_{i,k}^t$
    $\Delta_i^t = x_{i,K}^t - x_t$
  $\Delta_t = \beta_1 \Delta_{t-1} + (1-\beta_1)\left(\frac{1}{|\mathcal{S}|}\sum_{i \in \mathcal{S}} \Delta_i^t\right)$
  $v_t = v_{t-1} + \Delta_t^2$ (FedAdagrad)
  $v_t = v_{t-1} - (1-\beta_2)\Delta_t^2 \operatorname{sign}(v_{t-1} - \Delta_t^2)$ (FedYogi)
  $v_t = \beta_2 v_{t-1} + (1-\beta_2)\Delta_t^2$ (FedAdam)
  $x_{t+1} = x_t + \eta \frac{\Delta_t}{\sqrt{v_t}+\tau}$

- Wang et al. [10] presented FedNova, which standardized and adjusted local updates in accordance with the amount of local iterations.
- Adaptive schemes such as ADAM, YOGI, and ADAGRAD were introduced on the parameter server to improve the global model accuracy.
- Although the above methods make the global aggregation smoother, the gradient conflict problem is not properly solved.

Approach

Federated Learning Mitigating Gradient Conflict(FedMGC)



Approach

Handling of Class Imbalance

Cross Entropy loss function

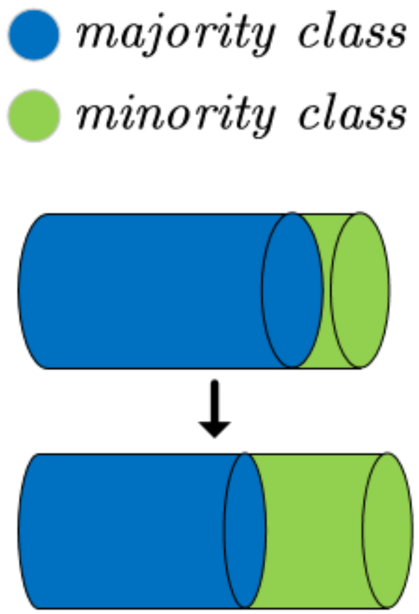
↓ Class imbalance

Focal Loss function

$$FL(p_t) = -\beta(1-p_t)^{\gamma}\log(p_t)$$

↓

Provide high-quality local gradients



Increase the proportion of minority class

Approach

Dominant Gradient Generation(DGG)

Algorithm 1 Dominant Gradient Generation

Input: Local gradients $g_t = \{g_1^t, ..., g_K^t\}$, loss values $l_t = \{l_1^t, ..., l_K^t\}$, dominant gradient selection ratio λ.

Output: Dominant gradient dg^t.

```
Initialize p^t = {}, z^t = {}
for g_i^t ∈ g_t do
    for g_j^t ∈ g_t do
        p_{i,j}^t = ( (g_i·g_j)/||g_j|| + (g_j g_i)/||g_i|| ) /2
    end for
end for
for i = 1, ..., K do
    add p_i^t = Σ_{j=1}^{K} p_{i,j}^t to p^t
end for
for i = 1, ..., K do
    add z_i^t = p_i^t / l_i^t to z^t.
end for
sort array z^t in descending order to achieve z_s = {z_{s_1}^t, ..., z_{s_K}^t}, then choose top ⌈λK⌉ gradients dg^t = {g_{s_1}^t, ..., g_{s_⌈λK⌉}^t} as dominant gradients which map to z_s.
return dg^t
```

$$p_{i,j}^t = \left(\frac{g_i \cdot g_j}{\|g_j\|} + \frac{g_j g_i}{\|g_i\|}\right)/2$$

$$p_i^t = \sum_{j=1}^{K} p_{i,j}^t, \qquad z_i^t = \frac{p_i^t}{l_i^t}$$

$$z_s = \{z_{s_1}^t, ..., z_{s_K}^t\}, \qquad dg^t = \{g_{s_1}^t, ..., g_{s_{\lceil \lambda K \rceil}}^t\}$$

Calculate the of mutual projection between two gradients, then calculate the gradient projection outlier of client i

By dividing the gradient projection outlier with loss value to get gradient outliers

Sort in descending order and select top [λK] gradients as the dominant gradients

Approach

Dominant Gradient Adjustment

Algorithm 2 Dominant Gradient Adjustment

Input: Local gradients $g_t = \{g_1^t, ..., g_K^t\}$, dominant gradients $dg^t = \left\{g_{s_1}^t, ..., g_{s_{\lceil \lambda K \rceil}}^t\right\}$.

Output: Corrected gradients g^t.

Initialize $n = |dg| = \lceil \lambda K \rceil$, $m = |g_t|$, $g_i^{cur} \leftarrow g_i^t$

for $i < m$ **do**

  for $j < b$ **do**

    if $g_i^{cur} \cdot g_{s_j} < 0$ and $i \neq s_j$ **then**

      $g_i^{cur} = g_i^{cur} - \frac{g_i^{cur} \cdot g_{s_j}^t}{\|g_{s_j}^t\|^2} g_{s_j}^t$

    end if

  end for

end for

$g^t = \frac{1}{m} \sum_{i=1}^{m} g_i^{cur}$

return g^t

Detects gradient conflict

Correct the gradient

Aggregate the corrected gradients

Experiment

Dataset & Model

Dataset	Model
MNIST	Two convolutional layers and two fully connected layers.
CIFAR-10	Two convolutional layers and three fully connected layers
CIFAR-100	Two convolutional layers and three fully connected layers

Dirichlet distribution: q ~Dir(α). A smaller value of α indicates stronger data heterogeneity

Parameter Settings

Parameter	Value
Batch size	128
Local epoch	10
Local learning rate	0.001

Baselines

FedAvg, FedProx, SCAFFOLD, FedNova

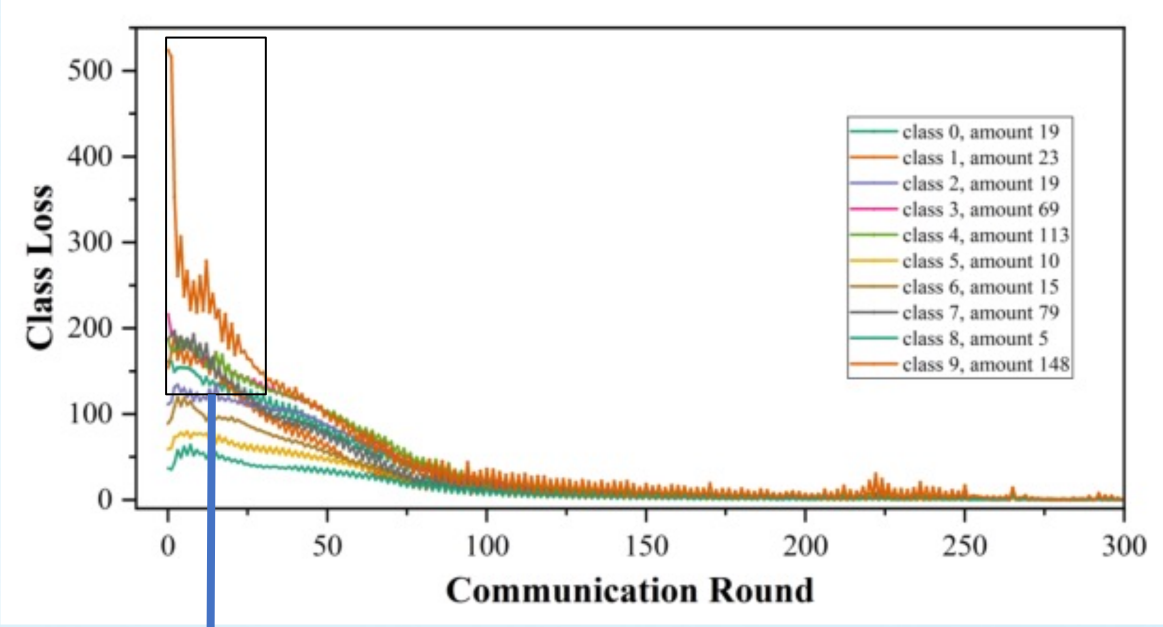
Indicators

Test accuracy and class loss

Experiment

Ablation Experiments

Settings: α = 1, the number of clients is 100, the client participation rate is 10%, CIFAR-10 dataset.



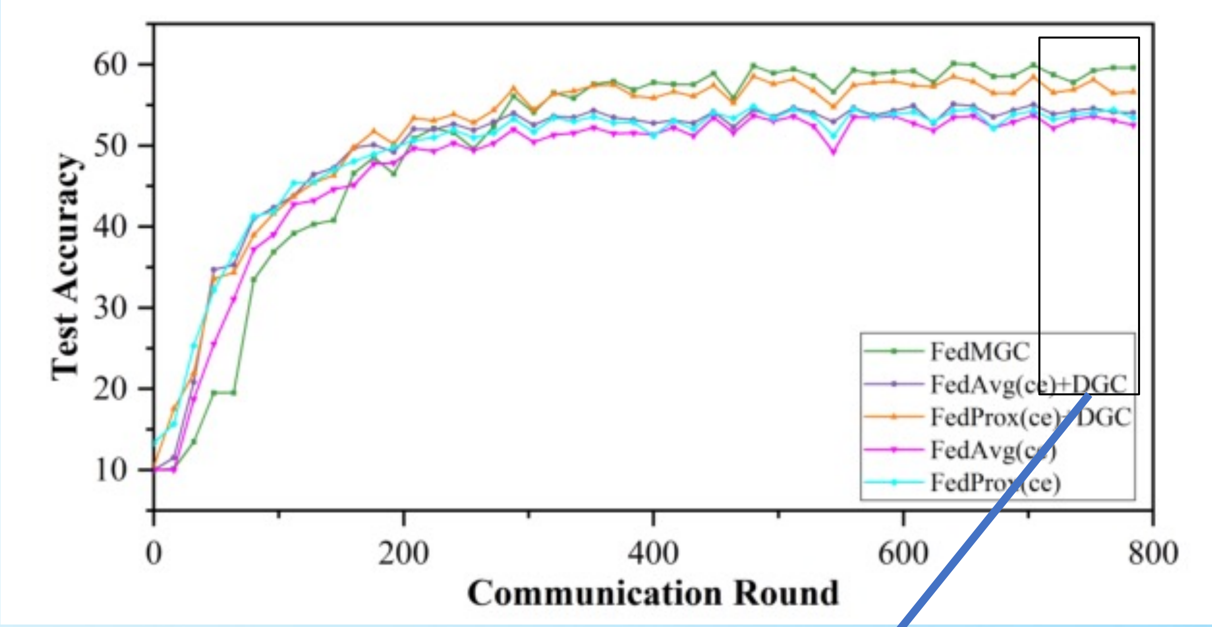


Fig. 1. Test accuracy of FedMGC and related ablation methods.

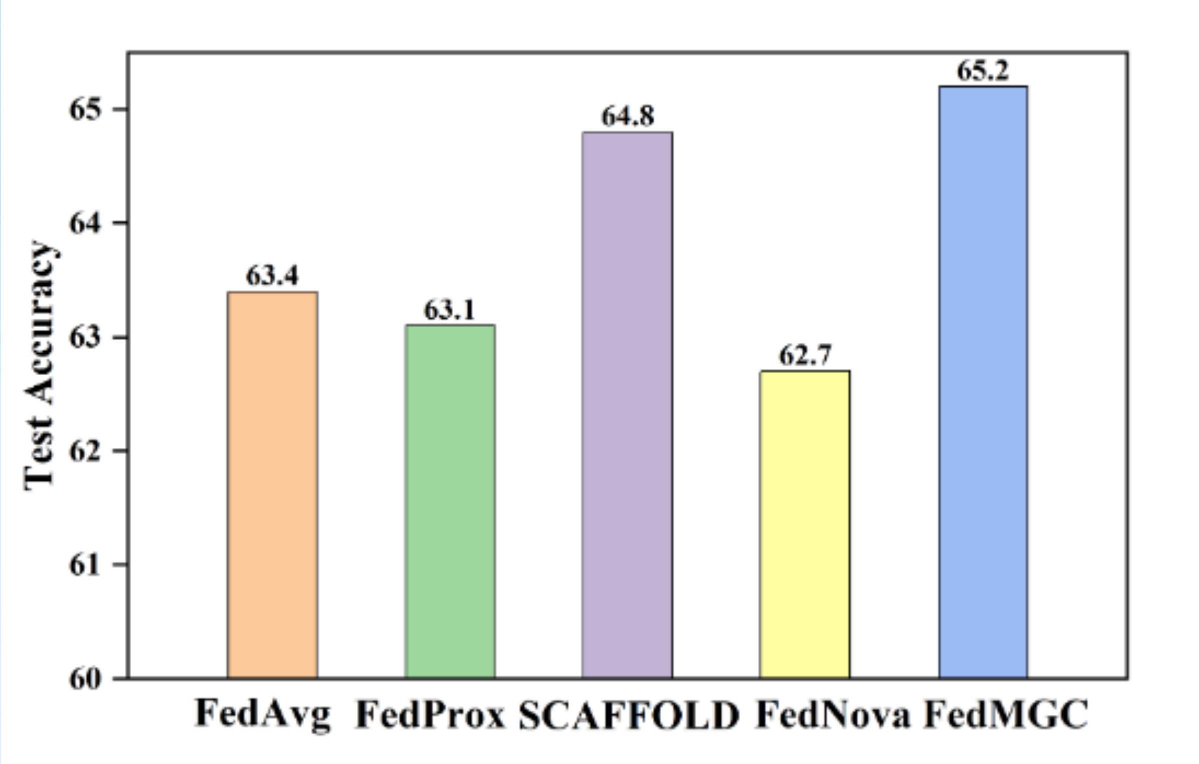
Fig. 2. Class loss variation in a client with class imbalance.

The loss of majority classes rapidly decline, increasing the proportion of minority classes.

Focal loss and DGC can effectively alleviate the gradient conflict

Experiment

Analysis of test accuracy while full client participation



Settings: $\alpha = 0.5$, the number of client is 10, client participation rate is 100% , CIFAR-10 dataset.

FedMGC achieves higher test accuracy among all methods, which indicates that the global accuracy can be improved by mitigating the gradient conflict.

Fig. 3. Test accuracy of FedMGC and baselines with full client participation.

Experiment

Analysis of test accuracy while client partial participation

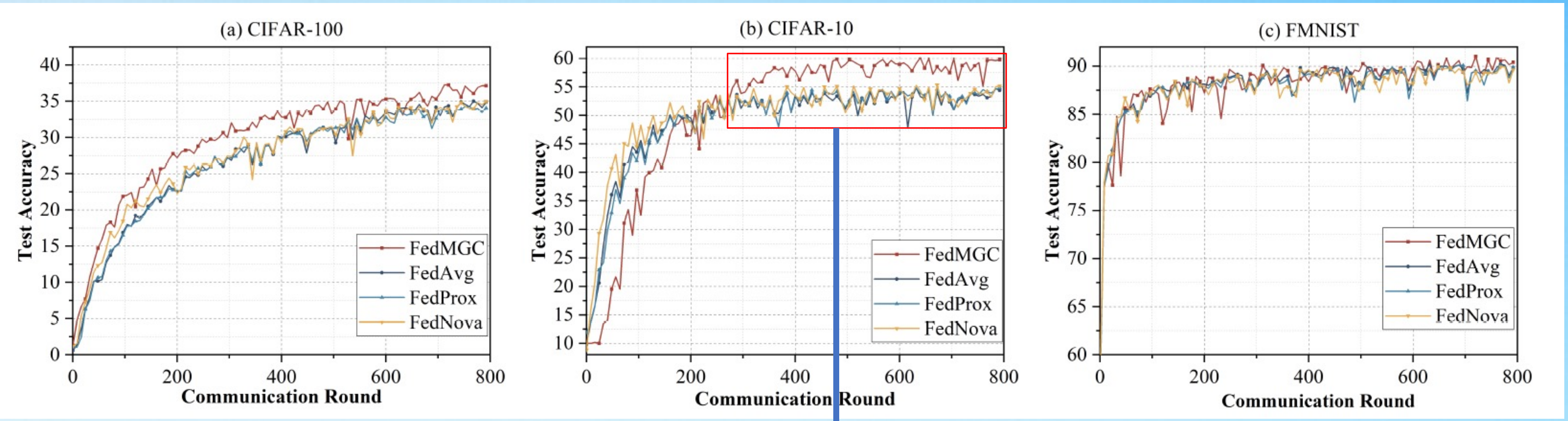


Fig. 4. Test Accuracy of FedMGC with baselines for α = 0.5 on CIFAR-100, CIFAR-10 and FMNIST datasets.

Settings: α = 0.5, the number of clients is 100, the client participation rate is 10%, on CIFAR-100,CIFAR-10and FMNIST dataset.

6.5%, 5.5%, and 5.2% higher FedAvg, FedProx, and FedNova.

The gradient conflict seriously affects the global accu racy, and FedMGC can effectively mitigate the gradient conflict

Conclusion

Conclusion

- FedMGC increases the loss contribution of minority classes and corrects gradients with conflict.
- FedMGC is able to achieve higher test accuracy on various heterogeneity of data.

Future Work

- Convergence analysis of FedMGC.
- Further optimize the FL function to reduce the tuning of the hyperparameter.

At Last

Thanks for your attention!