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

Jiajun Wang\textsuperscript{b}, Yingchi Mao\textsuperscript{a,b}, Zibo Wang\textsuperscript{b}, Jun Wu\textsuperscript{b} and Jie Wu\textsuperscript{c}

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

IEEE MASS
The 20th IEEE International Conference on Mobile Ad-Hoc and Smart Systems
1 Introduction
2 Related Work
3 Approach
4 Experiment
5 Conclusion
**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 uploaded gradients in the global aggregation phase.

Gradient conflict

\[ g_1 \cdot g_2 < 0 \]

Impairs the global accuracy
Optimization object

Reduce the conflict among uploaded gradients.

Goal

Improve the accuracy of the global model.
Related Work

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
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.
Federated Learning Mitigating Gradient Conflict (FedMGC)

**Local training phase**
- Use focal loss function
- Amplify the loss impact of minority classes
- Provide high-quality local gradients

**Global aggregation phase**
- Select dominant Gradient
- Dominant Gradient Adjustment
- Reduce oscillations and accelerate the update rate in the dominant direction

Approach
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 $g_t = \{g_1^t, \ldots, g_K^t\}$, loss values $l_t = \{l_1^t, \ldots, l_K^t\}$, dominant gradient selection ratio $\lambda$.

**Output:** Dominant gradient $dg^t$.

1. Initialize $p^t = \{\}$, $z^t = \{\}$
2. for $g_i^t \in g_t$ do
3.   for $g_j^t \in g_t$ do
4.     $p_{i,j}^t = \frac{(g_i^t \cdot g_j^t) + \|g_i^t\| \cdot \|g_j^t\|}{2}$
5.   end for
6. end for
7. for $i = 1, \ldots, K$ do
8.    add $p_i^t = \sum_{j=1}^K p_{i,j}^t$ to $p^t$
9. end for
10. for $i = 1, \ldots, K$ do
11.    add $z_i^t = \frac{p_i^t}{l_i^t}$ to $z^t$
12. end for
13. sort array $z^t$ in descending order to achieve $z_s = \{z_{s_1}^t, \ldots, z_{s_K}^t\}$, then choose top $\lceil\lambda K\rceil$ gradients $dg^t = \{g_{s_1}^t, \ldots, g_{s_{\lceil\lambda K\rceil}}^t\}$ as dominant gradients which map to $z_s$.
14. return $dg^t$

- 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 $\lceil\lambda K\rceil$ gradients as the dominant gradients
### Algorithm 2 Dominant Gradient Adjustment

**Input:** Local gradients $g_t = \{g_t^1, ..., g_t^K\}$, dominant gradients

$dg_t = \{g_s^t, ..., g_s^{(\lambda K)}\}$.

**Output:** Corrected gradients $g_t^i$.

1. **Initialize** $n = |dg| = |\lambda K|$, $m = |g_t|$, $g_{i, cur} \leftarrow g_i^t$
2. **for** $i < m$ **do**
3.  **for** $j < b$ **do**
4.   **if** $\langle g_{i, cur}^t \cdot g_{s,j}^t \rangle < 0$ and $i \neq s_j$ **then**
5.     $g_{i, cur}^t = g_{i, cur}^t - \frac{g_{i, cur}^t \cdot g_{s,j}^t}{\|g_{s,j}^t\|^2} g_{s,j}^t$
6.   **end if**
7. **end for**
8. **end for**
9. $g_i^t = \frac{1}{m} \sum_{i=1}^{m} g_{i, cur}^t$
10. **return** $g_i^t$

---

**Detected gradient conflict**

**Correct the gradient**

**Aggregate the corrected gradients**
**Experiment**

### Dataset & Model

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>Two convolutional layers and two fully connected layers.</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>Two convolutional layers and three fully connected layers</td>
</tr>
<tr>
<td>CIFAR-100</td>
<td>Two convolutional layers and three fully connected layers</td>
</tr>
</tbody>
</table>

**Dirichlet distribution:** \( q \sim \text{Dir}(\alpha) \). A smaller value of \( \alpha \) indicates stronger data heterogeneity.

### Parameter Settings

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Batch size</td>
<td>128</td>
</tr>
<tr>
<td>Local epoch</td>
<td>10</td>
</tr>
<tr>
<td>Local learning rate</td>
<td>0.001</td>
</tr>
</tbody>
</table>

### Baselines


### Indicators

Test accuracy and class loss
Ablation Experiments

Settings: $\alpha = 1$, 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.

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

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

Focal loss and DGC can effectively alleviate the gradient conflict.
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.
Analysis of test accuracy while client partial participation

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

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

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

The gradient conflict seriously affects the global accuracy, and FedMGC can effectively mitigate the gradient conflict.
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.
Thanks for your attention!