Delay Gradient Averaging: Tolerate the Communication Latency in Federated Learning

Ligeng Zhu¹, Hangzhou Lin², Yao Lu³, Yujun Lin¹, Song Han¹

¹MIT ²Amazon ³Google

delaygradientaveraging.com

Federated Learning Allows Training without Sharing

• Security: Data never leaves devices thus promises security and regularization.
• Customization: Models continually adapt to new data from the sensors.

Challenge: Network Communication Bottleneck

Bandwidth can be always improved by

• Customization
• Physical limits: Shanghai to Boston, even with the speed of light, still takes 162ms.

Federated Learning Allows Training without Sharing

User's data
Intelligent Edge Devices
Cloud Server

• Wired or wireless network
• Bandwidth as high as 100Gb/s, latency as low as 1us
• WiFi or cellular network: Bandwidth up to 1Gb/s, latency ~200ms

Existing Methods Improves the Bandwidth, but not the Latency

• Bandwidth can be always improved by
  • Hardware upgrade
  • Gradient compression and quantization

• Latency is hard to improve because
  • Physical limits: Shanghai to Boston, even with speed of light, still takes 162ms.
  • Signal congestion: Urban office and home creates a lot of signal contention.

Delay Gradient Averaging [Ours]

1. Sample and calculate \( \nabla W_{(i,j)} \)
2. If \( i \mod K = 0 \)
   1. Send fresh \( \nabla W_{(i,j)} \) to other nodes
   Delay \( D \) steps
3. If \( i \mod K = D \)
   1. Recv stale \( \nabla W_{(i-D,j)} \) from other nodes
   2. \( \nabla W_{(i,j)} = \frac{1}{D} \sum_{j=1}^{D} \nabla W_{(i-D,j)} \)
4. \( w_{(i,j)} = w_{(i,j)} - \eta ( \nabla W_{(i,j)} - \nabla W_{(i-D,j)} + \nabla W_{(i-D+1,j)} ) \)

With delay: All the local machines are blocked to wait for the synchronization to finish.

Send and recv params

W/o delay: All the local machines are blocked to wait for the synchronization to finish.

Send and recv params

Discussion of the Compensation Term

\[
\begin{align*}
  w_{(i,j)} &= w_{(i,j)} - \eta ( \nabla W_{(i,j)} - \nabla W_{(i-D,j)} + \nabla W_{(i-D+1,j)}) \\
  w_{(i,j)} &= w_{(i,j)} - \eta ( \nabla W_{(i-D,j)} - \nabla W_{(i-D+1,j)} + \cdots + \nabla W_{(i-D+2,j)} + \cdots + \nabla W_{(i-D+D-1,j)}) \\
  w_{(i,j)} &= w_{(i,j)} - \eta ( \nabla W_{(i-D,j)} - \nabla W_{(i-D+1,j)} + \cdots + \nabla W_{(i-D+2,j)} + \cdots + \nabla W_{(i-D+D-1,j)}) \\
\end{align*}
\]

Consider the 3rd iteration with \( D = 2 \)

\[
\begin{align*}
  w_{(i,j)} &= w_{(i,j)} - \eta ( \nabla W_{(i,j)} - \nabla W_{(i-D,j)} + \nabla W_{(i-D+1,j)}) \\
\end{align*}
\]

Replacing oldest local gradients with global averaged ones.

\[
\begin{align*}
  \nabla W_{(i,j)} &= \nabla W_{(i,j)} - \nabla W_{(i-D,j)} + \nabla W_{(i-D+1,j)} + \cdots + \nabla W_{(i-D+D-1,j)} \\
\end{align*}
\]

The divergence is bounded.

\[
\begin{align*}
  \nabla W_{(i,j)} &= \nabla W_{(i,j)} - \nabla W_{(i-D,j)} + \nabla W_{(i-D+1,j)} + \cdots + \nabla W_{(i-D+D-1,j)} \\
\end{align*}
\]

Table 1: Ablation studies about our gradient correction term. Without our correction term, using pure

Experimental Results

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Partition</th>
<th>FedAvg (K=5)</th>
<th>FedAvg (K=10)</th>
<th>FedAvg (K=20)</th>
<th>DGA (K=5, D=20)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>Speedup</td>
<td>Acc</td>
<td>Speedup</td>
<td>Acc</td>
</tr>
<tr>
<td>CIFAR</td>
<td>i.i.d</td>
<td>88.7</td>
<td>1</td>
<td>88.5</td>
<td>88.1</td>
</tr>
<tr>
<td></td>
<td>non-i.i.d</td>
<td>48.2</td>
<td></td>
<td>47.2</td>
<td>43.9</td>
</tr>
<tr>
<td>ImageNet</td>
<td>i.i.d</td>
<td>55.4</td>
<td>1</td>
<td>76.5</td>
<td>76.2</td>
</tr>
<tr>
<td></td>
<td>non-i.i.d</td>
<td>36.9</td>
<td>1</td>
<td>47.3</td>
<td>46.8</td>
</tr>
<tr>
<td>Shakespeare</td>
<td>i.i.d</td>
<td>47.6</td>
<td>1</td>
<td>34.3</td>
<td>47.4</td>
</tr>
<tr>
<td></td>
<td>non-i.i.d</td>
<td>36.9</td>
<td>1</td>
<td>34.3</td>
<td>30.1</td>
</tr>
</tbody>
</table>

DGA demonstrates consistent training speedup, but also maintains the accuracy, on both i.i.d and non-i.i.d partition.

On a realistic federated learning scenarios, a raspberry-pi cluster consists of 16 devices, DGA is robust against network stragglers and shows improvements.