CD-SGD: Distributed Stochastic Gradient Descent with Compression and Delay Compensation

Speaker: Enda Yu

Author: Enda Yu, Dezun Dong, Yemao Xu, Shuo Ouyang, Xiangke Liao

From National University of Defense Technology
Background of Distributed Training

Calculation stage

Communication stage
S-SGD and A-SGD

- **S-SGD**
  - High accuracy
  - Slow speed

- **A-SGD**
  - Low accuracy
  - Fast speed
Several typical optimization methods

- **SSGD(a):** Synchronous Stochastic Gradient Descent
- **Local(b):** A local update method.
- **BIT-SGD(c):** 2-bit quantization method provided by MXNet
✓ The local update mechanism is introduced into quantization method.
✓ K-step correction method is used to periodically repair the loss of accuracy caused by quantization.
Compression overhead hiding

- Quantization always has an impact on the timing of the end of communication.

- The next FP/BP(calculation) in CD-SGD does not need to wait for the current communication, solving the speed impact caused by quantization.
**Convergence Accuracy**

CD-SGD can help BIT-SGD achieve accuracy close to or even better than S-SGD.

The upper limit of CD-SGD accuracy is affected by the quantification method used and the warm-up time.
Analysis about k-step correction

• Setting k equal to 5 is suitable for most situations.

The average epoch wall-clock time and communication time of ResNet-20 on CIFAR-10 (/sec).

<table>
<thead>
<tr>
<th>Time</th>
<th>SSGD</th>
<th>BIT–SGD</th>
<th>k2</th>
<th>k5</th>
<th>k10</th>
<th>k20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Te (M=4)</td>
<td>2.24</td>
<td>2.22</td>
<td>1.88</td>
<td>1.86</td>
<td>1.86</td>
<td>1.85</td>
</tr>
<tr>
<td>Te (M=2)</td>
<td>4.32</td>
<td>3.65</td>
<td>3.61</td>
<td>3.51</td>
<td>3.46</td>
<td>3.44</td>
</tr>
<tr>
<td>Tc (M=4)</td>
<td>2.19</td>
<td>2.26</td>
<td>2.14</td>
<td>2.08</td>
<td>2.05</td>
<td>2.11</td>
</tr>
<tr>
<td>Tc (M=2)</td>
<td>4.11</td>
<td>4.10</td>
<td>3.67</td>
<td>3.62</td>
<td>3.81</td>
<td>3.90</td>
</tr>
</tbody>
</table>
The performance improvement range of CD-SGD to S-SGD and BIT-SGD is 0 to 89% and 3% to 45%, respectively.

Example: the average epoch time of ResNet50 training with 4 workers batch size 32 per GPU on V100 (/sec).

<table>
<thead>
<tr>
<th></th>
<th>S-SGD</th>
<th>BIT-SGD</th>
<th>OD-SGD</th>
<th>CD-SGD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epoch-time</td>
<td>824.8</td>
<td>601.2</td>
<td>712.2</td>
<td>567.9</td>
</tr>
</tbody>
</table>
Conclusion

• Distributed training communication optimization is necessary

• Three challenges in combining compression methods with system-level methods: Additional compression overhead, Accuracy, Sufficient benefits

• The local update mechanism can cover the additional computational overhead of compression.

• K-step correction method can solve the accuracy reduction problem of the compression method.
CD-SGD: Distributed Stochastic Gradient Descent with Compression and Delay Compensation

THANKS