Delay Gradient Averaging: Tolerate the Communication Latency in Federated Learning

Ligeng Zhu[1], Hangzhou Lin[2], Yao Lu[3], Yujun Lin[1], Song Han[1]
[1]MIT [2]Amazon [3]Google

Hardware, AI and Neural-nets

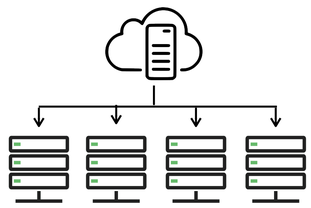
NEURAL INFORMATION PROCESSING SYSTEMS

dga.hanlab.ai

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

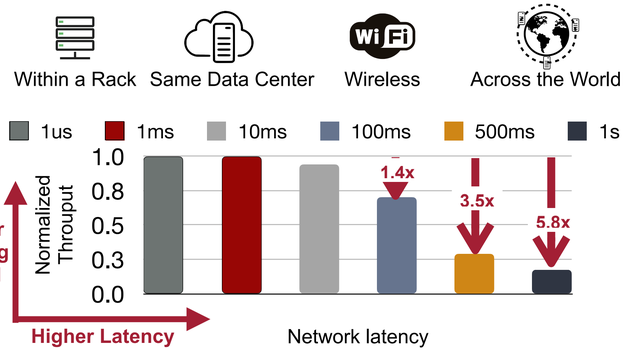


Wired ethernet or infinity band
Bandwidth as high as **100Gb/s**,
Latency as low as **1us**

WiFi or Cellular network
Bandwidth up to **1Gb/s**,
Latency **~200ms**.

Existing Methods Improves the Bandwidth, but not the Latency

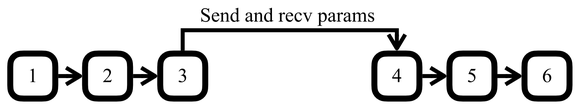
- Bandwidth can be always improved by
  - Hardware upgrade
  - Gradient compression and quantization
- Latency **is hard to improve** because
  - Physical limits: Shanghai to Boston, even with the speed of light, still takes 162ms.
  - Signal congestion: Urban office and home creates a lot of signal contention.



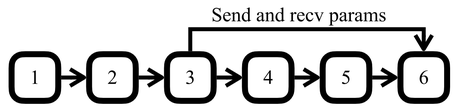
Delay Gradient Averaging [Ours]

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

Delay D steps



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



With delay: Worker keep performing local updates while the parameters are in transmission.

Discussion of the Compensation Term

$$w_{(i,j)} = w_{(i,j)} - \eta(\nabla w_{(i,j)} - \nabla w_{(i-D,j)} + \overline{\nabla w_{(i-D)}})$$

Consider the **3rd iteration** with $D = 2$

$$w_{(3,j)} = w_{(1,j)} - \eta(\nabla w_{(1,j)} + \nabla w_{(2,j)} + \nabla w_{(3,j)})$$

$\nabla w_{(1)}$

Consider the **4th iteration** with $D = 2$

$$w_{(4,j)} = w_{(1,j)} - \eta(\overline{\nabla w_{(1)}} + \nabla w_{(2,j)} + \nabla w_{(3,j)} + \nabla w_{(4,j)})$$

$\overline{\nabla w_{(2)}}$

Replacing **oldest local gradients** with **global averaged ones.**

Our DGA:

$$w_{(i,j)} = w_{(1,j)} - \eta(\overline{\nabla w_{(1)}} + \ldots + \overline{\nabla w_{(i-D,j)}} + \nabla w_{(i-D+1,j)} + \ldots + \nabla w_{(i,j)})$$

The divergence is bounded.

Vanilla Distributed SGD:

$$w_{(i,j)} = w_{(1,j)} - \eta(\overline{\nabla w_{(1)}} + \ldots + \overline{\nabla w_{(i-D,j)}} + \overline{\nabla w_{(i-D+1)}} + \ldots + \overline{\nabla w_{(i)}})$$

Theoretical Results

- Assumption 1: the loss function $F(w; x, y)$ is **Lipchitz smooth**

$$\nabla f_j(x) - \nabla f_j(y)\,|| \le L\,||x - y\,||. \quad \forall x, y \in \mathbb{R}^d$$

- Assumption 2: **Bounded gradients and variances**

$$\mathbb{E}_{\zeta_j}\,||\nabla F_j(w;\zeta_i)\,||^2 \le G^2, \forall w, \forall j,\ \mathbb{E}_{\zeta_j}\,||\nabla F_j(w;\zeta_j) - \nabla f_j(w)\,||^2 \le \sigma^2, \forall w, \forall j.$$

The convergence rate of DGA is $O(\frac{\Delta + \sigma^2}{\sqrt{JN}} + \frac{Jd^2}{N})$ (details in paper)

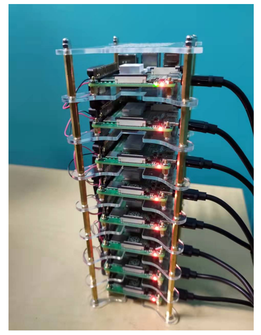
When $D < O(N^{\frac{1}{4}}J^{-\frac{3}{4}})$, **DGA converges as fast as original SGD** which is $O(\frac{\Delta + \sigma^2}{\sqrt{JN}})$.

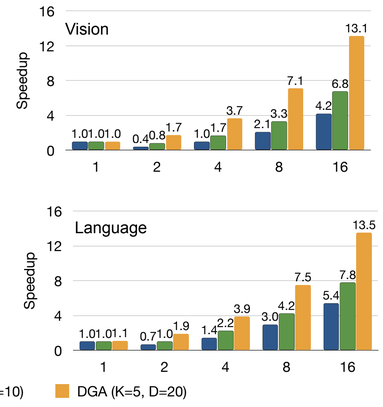
N: The number of iterations; J: The number of workers / clients.

Experimental Results

Datasets	Partition	FedAvg (K=5)		FedAvg (K=10)		FedAvg (K=20)		DGA (K=5, D=20)	
		Acc	Speedup	Acc	Speedup	Acc	Speedup	Acc	Speedup
CIFAR	i.i.d	88.7	1×	88.5	1.51×	88.1	2.05×	88.6	3.16×
	non-i.i.d	48.2		47.2		43.9		48.0	
ImageNet	i.i.d	76.6	1×	76.5	1.43×	76.2	1.81×	76.4	2.55×
	non-i.i.d	55.4		52.5		48.6		54.9	
Shakespeare	i.i.d	47.6	1×	47.3	1.66×	47.4	2.51×	47.1	4.07×
	non-i.i.d	36.9		34.3		30.1		36.3	

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.**