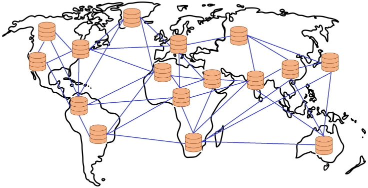
WeightGrad: Geo-Distributed Data Analysis Using Quantization for Faster Convergence and Better Accuracy

Syeda Nahida Akter

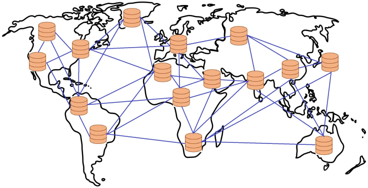
Muhammad Abdullah Adnan

Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology (BUET)
Dhaka, Bangladesh

Problem Overview



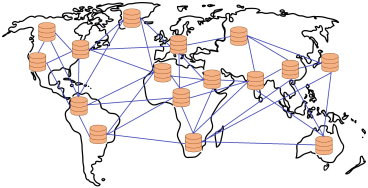
Problem Overview

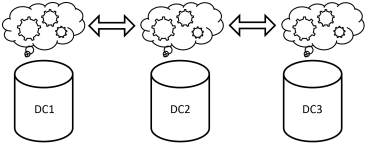


Problem

- Powerful machines.
- Huge amount of time.

Problem Overview





Problem

- Powerful machines.
- Huge amount of time.

Solution

- Distribute the DNN system across multiple data centers.

Problem Definition

For distributed setup, we define the problem as

Problem Definition

For distributed setup, we define the problem as

- How to efficiently utilize limited WAN b/w

Problem Definition

For distributed setup, we define the problem as

- How to efficiently utilize limited WAN b/w
- How to ensure faster convergence without loss of accuracy

Methodology

We propose **WeightGrad** that

- adapts both weight and gradient quantization to provide best speedup possible on WAN
- proposes a synchronous structure to prevent the loss in accuracy due to quantization

WeightGrad System

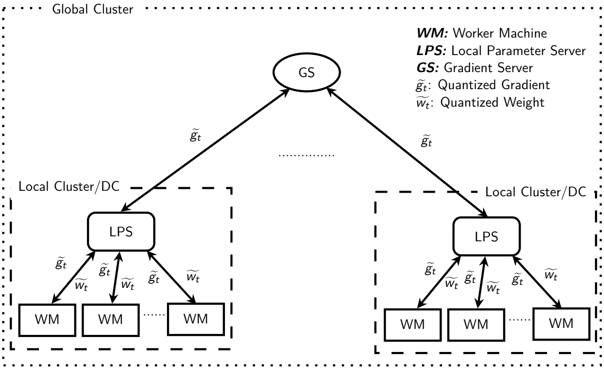


Figure: WeightGrad Tree Structure

Two Level Structure

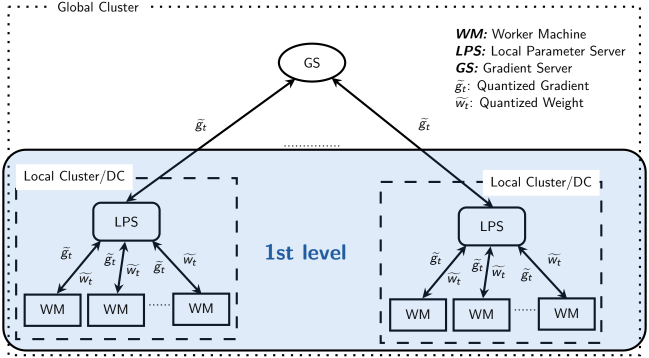


Figure: WeightGrad Tree Structure

Two Level Structure

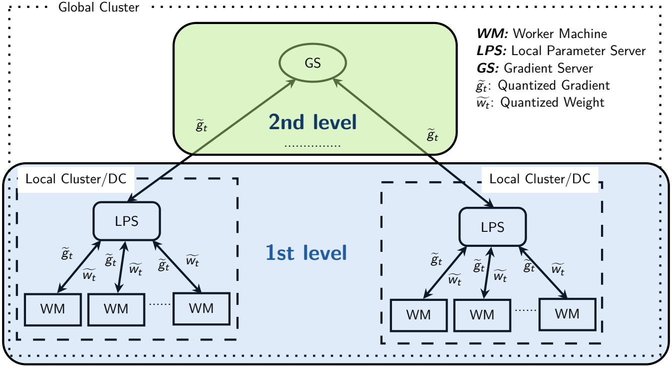


Figure: WeightGrad Tree Structure

WeightGrad System

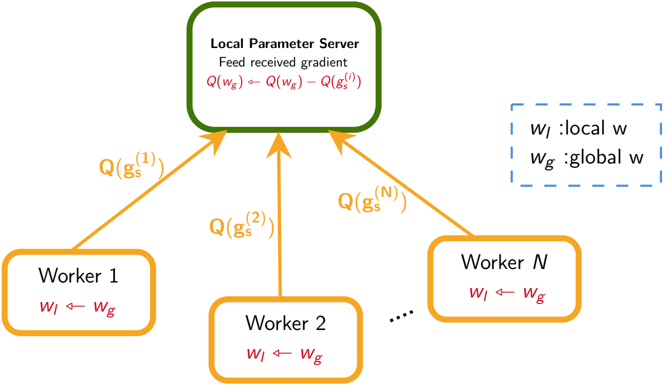


Figure: WeightGrad: Local Cluster

WeightGrad System

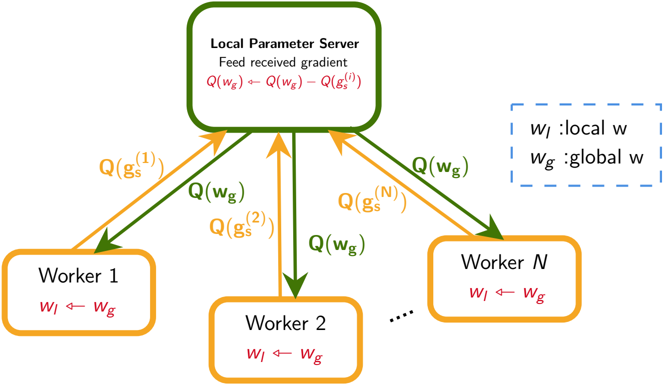


Figure: WeightGrad: Local Cluster

LPS with Gradient Synchronizer

We introduce Gradient Synchronizer which performs two functions:

LPS with Gradient Synchronizer

We introduce Gradient Synchronizer which performs two functions:

Dynamic Threshold

- Checks significance of aggregated quantized gradients by comparing it with the threshold.

LPS with Gradient Synchronizer

We introduce Gradient Synchronizer which performs two functions:

Dynamic Threshold

- Checks significance of aggregated quantized gradients by comparing it with the threshold.

Fixed Interval

- Maintains a fixed interval T, within which LPS receives aggregated gradient values from the GS.

Amazon EC-2

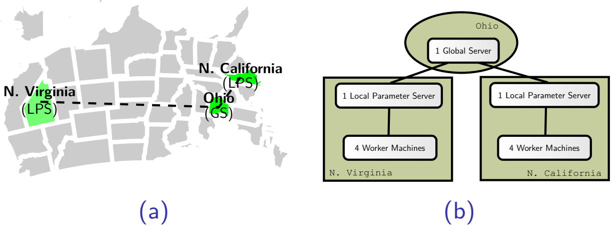
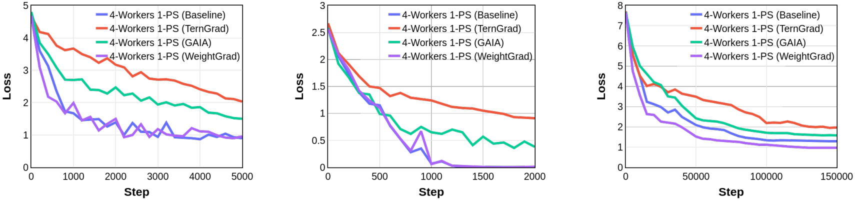


Figure: (a)Deployment Regions in AWS(b)Instance Hierarchy

Instances	Instance Type	RAM	vCPU	GPU	B/W
11	g3s.xlarge, 64-bit Ubuntu Server 16.04 LTS	30.5 GiB	4	NVIDIA Tesla M60 GPU	10 Gbps

Training Loss Analysis



(a) Training Loss for CifarNet

(b) Training Loss for VGGNet

(c) Training Loss for ImageNet

Figure: (a) Training loss for CifarNet model on CIFAR-10 dataset (b) Training loss for VGGNet model on CIFAR-10 dataset (c) Training loss for AlexNet on ImageNet dataset

SpeedUp Analysis

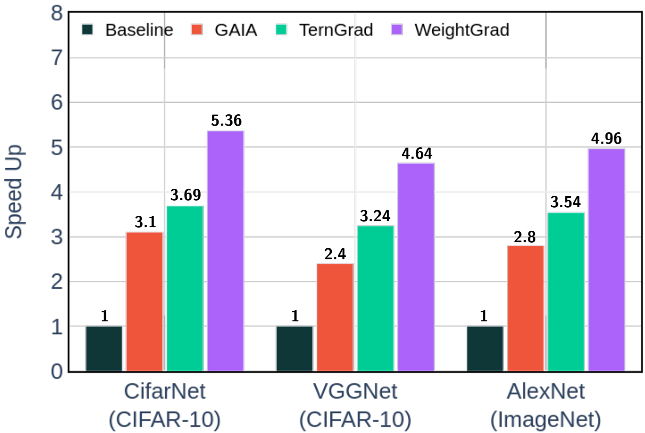


Figure: Training Speed Comparison

Accuracy Comparison

Model	SGD	Base LR	Total mini-batch size	Steps	Gradients	Workers	Accuracy
CifarNet	GD	0.1	128	50k	Baseline	4	84.56%
					Gaia	4	83.48%(-1.08%)
					TernGrad	4	82.41%(-2.15%)
					WeightGrad	**4**	**84.56%(-0.00%)**
	GD	0.1	512	50k	Baseline	8	83.19%
					Gaia	8	83.04%(-0.13%)
					TernGrad	8	81.40%(-1.79%)
					WeightGrad	**8**	**83.21%(+0.03%)**
VGG-Net	GD	0.1	512	50k	Baseline	8	88.14%
					Gaia	8	87.19%(-0.95%)
					TernGrad	8	86.3%(-1.84%)
					WeightGrad	**8**	**88.13%(-0.01%)**

(a)

Table: Comparison of training methods on (a) Cifar-10 data and (b) ImageNet

Accuracy Comparison

Model	SGD	Base LR	Total mini-batch size	Steps	Gradients	Workers	Accuracy
CifarNet	GD	0.1	128	50k	Baseline	4	84.56%
					Gaia	4	83.48%(-1.08%)
					TernGrad	4	82.41%(-2.15%)
					WeightGrad	**4**	**84.56%(-0.00%)**
	GD	0.1	512	50k	Baseline	8	83.19%
					Gaia	8	83.04%(-0.13%)
					TernGrad	8	81.40%(-1.79%)
					WeightGrad	**8**	**83.21%(+0.03%)**
VGG-Net	GD	0.1	512	50k	Baseline	8	88.14%
					Gaia	8	87.19%(-0.95%)
					TernGrad	8	86.3%(-1.84%)
					WeightGrad	**8**	**88.13%(-0.01%)**

(a)

Model	Steps	Training Method	Top-1 Accuracy	Top-5 Accuracy
AlexNet	185k	Baseline	58.17%	80.19%
		Gaia	58.02%(-0.15%)	80.20%(+0.01%)
		TernGrad	57.32%(-0.85%)	80.18%(-0.01%)
		Deep Gradient Compression	58.20%(+0.03%)	80.20%(+0.01%)
		WeightGrad	**59.28%(+1.06%)**	**80.25%(+0.06)**

(b)

Table: Comparison of training methods on (a) Cifar-10 data and (b) ImageNet

THANK YOU!