Thesis Progress report

Ching-Yuan, Tsai

2018,1,17
Presented by Ching-Yuan, Tsai

Outline

1. Introduction
2. Prune
3. Gradients request scheduling
4. Timetable

Motivation

- In large module, distributed training may be slower than single-machine training.
- Parameter server uses a lot of time on synchronous stage.
- Network is bottleneck in our system.

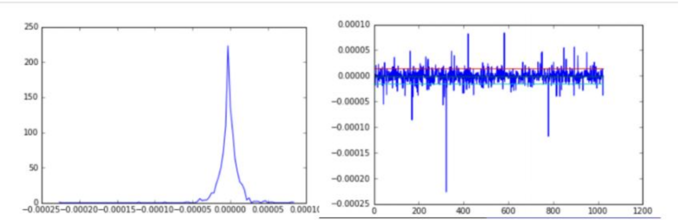
Objective

- Use software method to reduce network load.
  - Prune gradients by threshold.
    - static threshold
    - standard deviation threshold
    - dynamic threshold
  - Gradients request scheduling.

Outline

1. Introduction
2. Prune
3. Gradients request scheduling
4. Timetable

Observation



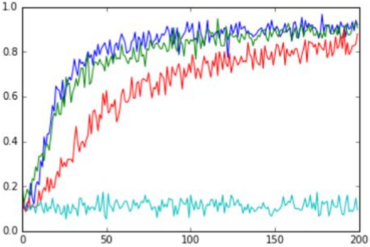
Prune

- Prune gradients: Send gradients which are absolutely greater than threshold.
  - Static threshold: 10%, 1%, 0.1%.
  - Standard deviation threshold: 1, 2, 3.
  - Dynamic threshold: mean of the gradients which are greater than last threshold.

Standard deviation threshold

- Compute standard deviation on gpu.
- Faster 30 times than cpu.
- Use 25% computation time.

experiment 1



experiment 2

- To do: accuracy for each time.

Static threshold

- Compute on gpu.
- Selection algorithm.

experiment 1

- To do: accuracy for each interation.

experiment 2

- To do: accuracy for each time.

Dynamic threshold

- Compute on gpu.
- Next threshold = mean of the gradients which are greater than current threshold

experiment 1

- To do: accuracy for each interation.

experiment 2

- To do: accuracy for each time.

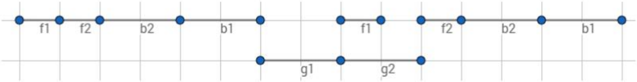
Outline

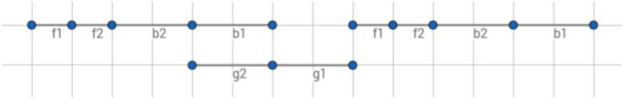
1. Introduction
2. Prune
3. Gradients request scheduling
4. Timetable

Introduction

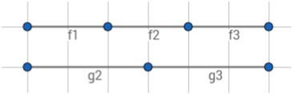
- Client changes gradient request order to get some benefits.

Reduce suspend





Reduce staleness



Input

- B = staleness bound.
- L = number of layers.
- I = number of iterations.
- W = number of workers.
- Cp = computation time : f1, b1, f2, b2, ...
- Cm = communication time : t1, t2, ...

Intermediate data

- Server maintains a set of variables indicating the minimum iteration of each layer.
- c1, c2, ... , cL

Objective

- Minimize training time and staleness.

Request pool

- Pull request.
- Push request.

Scheduling

- First in first out.
- First in last out.
  - Starvation.

Optimal solution

- Pull
    - Pull the stalest layer.
- Push
    - Push the layer which can update intermediate data.

Outline

1 Introduction

2 Prune

3 Gradients request scheduling

4 Timetable

Timetable

- -3/31 : Finish all experiment.
- -4/30 : Write paper.
- 6/15- : Oral.