Fundamentals for Distributed Machine Learning

Luo Mai

University of Edinburgh
Motivation

Gap to be filled by Distributed ML Systems
Divide-and-Conquer

(a) Single-worker Execution

(b) Distributed Multi-worker Execution
Why distributed ML systems?

• Performance
  • Reducing the time to complete a data epoch

• Memory wall

• Economy
  • Multiple commodity servers, instead of a single expensive high-end server

• Hardware failure tolerance
Basic execution model

Workers must have sufficient memory to store data, weights, activations & gradients
• Otherwise, you get Out-Of-Memory (OOM) exception
Questions?
## Parallel training methods

<table>
<thead>
<tr>
<th></th>
<th>Single Data</th>
<th>Multiple Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single Program</strong></td>
<td>Single-Program-Single-Data (SPSD)</td>
<td>Single-Program-Multiple-Data (SPMD)</td>
</tr>
<tr>
<td></td>
<td><strong>Single Worker Training</strong></td>
<td><strong>Data Parallel Training</strong></td>
</tr>
<tr>
<td><strong>Multiple Program</strong></td>
<td>Multiple-Program-Single-Data (MPSD)</td>
<td>Multiple-Program-Multiple-Data (MPMD)</td>
</tr>
<tr>
<td></td>
<td><strong>Model Parallel Training</strong></td>
<td><strong>Hybrid Parallel Training</strong></td>
</tr>
</tbody>
</table>

Data parallel training

Worker 1
Partition 1
Data
Op1
Gradient 1
Op2
Gradient 2
Op3
Gradient 3
Worker 2
Partition 2
Data
Op1
Gradient 1
Op2
Gradient 2
Op3
Gradient 3

Remove computation bottleneck

Allreduce (Averaged gradients)
Model parallel training: Intra-operator

Program Partition 1

Worker 1

Data

Op1

Partition 1

Op2

Combine (Gather)

Broadcast

Program Partition 2

Worker 2

Op1

Partition 2

Remove memory bottleneck
Model parallel training: Inter-operator

Worker 1
Data

Program Partition 1
Op1

Send

Program Partition 2
Op2

Worker 2

Remove memory bottleneck

Receive
Hybrid parallel training

Worker 1

Partition 1 -> Op1
Data

Gradient 1

Worker 2

Op2 -> Gradient 2

Replicated Program Partition 1

Allreduce (Averaged gradients)

Worker 3

Data

Op1

Worker 4

Partition 2 -> Op2

Replicated Program Partition 2

Data

Op1

Remove memory & computation bottlenecks
Questions?
How to choose parallelism methods?

- **Empirical parallelism**
  - TensorFlow Mesh

- **Semi-automatic parallelism**
  - Manually partition a few upstream operators and propagate the partitioning to downstream operators

- **Automatic parallelism**
  - Build a cost model for evaluating different parallel methods
  - Search for the best methods that incur minimal costs
Pipeline parallelism

Creating pipeline by dividing a data partition into micro-batches

Computing the **averaged gradients** for micro-batches

Model Partitions

- Worker 0
- Worker 1
- Worker 2
- Worker 3

Creating pipeline by dividing a data partition into **micro-batches**
Optimising micro-batch size

- **Small micro-batch** reduces bubble size; but incur large micro-batch scheduling overheads
- **Large micro-batch** incurs large bubble; but come with small micro-batch scheduling overheads
- Optimal micro-batch size must **balance bubble size and scheduling overheads**
Design aspects of distributed ML systems

• Cluster elasticity
  • Reserving a large number of GPU servers is prohibitively expensive

• Device roles in ML
  • CPUs for data processing, GPUs for training (PyTorch & TensorFlow)
  • This is causing problems in graph learning and reinforcement learning

• Mixed precision training
  • FP8, FP16, FP32, FP64
Summary

• Distributed ML systems are keys to tackle “End of Moore Law”
• Performance and economy benefits
• Spatial parallelism: Data-parallel, model-parallel and hybrid-parallel
• Temporal parallelism: Pipeline parallelism
Reading

• Optional reading
  • PyTorch distributed overview
  • Google GPipe paper
Questions?
Large-Scale System Software Group

- ML system projects (capstone project / final-year projects)
- Open-source project: the 1st open-sourced textbook for ML systems

Personal website: https://luomai.github.io
Email: luo.mai@ed.ac.uk

If interested, please send me your CV, transcript, and a description of your interest, and we can arrange a follow-up meeting.