Tensorflow

Guangyi Tao
June/09/2017
Introduction

• Tensorflow is a deep-learning framework published by Google.
• Available at https://www.tensorflow.org/ (open-source)
• Works both for GPU and CPU
Single and Multiple Process Worker
Node Placement

• Simulate activity on each device in system as it progresses.
• Make sure each node has its set of feasible devices due to different operations.
• Use a greedy heuristic examine or measure execution time of operation.
The Greedy Heuristic

- Starts from the source of the graph
- The heuristic includes execution time and communication time
- Use the device that takes least time
- Then calculate the downstream nodes
Some constraints

• Sometimes we can collocate some nodes to some others (i.e., The ReLU)
• Once the colocation is determined, then we need to consider these two nodes to be placed on one device
• So we need to use union-find to find the intersection of feasible device
• Edges are cut. We add node ‘send’ and ‘recv’ on the edge to isolate the communication.
Let’s say if there is a device who failed, and we need one more iteration. We can let the device only run by itself.

The feed is for the input and the fetch takes out the output, both of them are specially-initialized.

Each node is Connected to a ‘Save’ node, Saves the data of this node in case it failed or need further usage.
Data parallel Training

• Used for SGD

• If we have mini-batch size of 1000 elements we can use 10 replicas of the model to each compute the gradient for ‘the’ 100 elements on local device and then update the integration by a client
Parallel computing for single device
Small Tricks

• Scheduling
  – Scheduling can help save the memory of intermediate results from being kept too long.
  – Add control edges to tell the device when to start

• Asynchronous Kernels
  – Gain benefits for I/O, use a queue to list all the tasks, no need to wait for all the threads to finish compared to synchronous kernel

• Use less precise data
  – Change the 32bits float to 16bits float may help the data communication
Other Framework

• By facebook:
  – Caffe2
  – Torch/PyTorch(newest)

• By Google:
  – Tensorflow(come from Theano)

• Other:
  – Caffe
  – Theano