Reading Assignment 1

Authors: Group 1, Chuan Su, Diego Alonso Guillen Rosaperez

1 Motivation

Deep Convolutional Neural Networks (CNN) have proven in the last years to be general and effective across many tasks such as image classification, face recognition, text classification, and game playing. Therefore, computational requirements for training them have growth recently.
A current approache to face this issue is to parallelize training onto multiple devices by applying a single parallelization strategy to all layers in a network. On one side, data parallelism replicates the entire network on each device and assigns a subset of the data to each of them, which is inefficient for layers with large numbers of network parameters. This strategy suits more optimally a convolutional layer to eliminate data transfers from the previous layer. On the other side, model parallelism divides the network parameters into disjoints subsets and trains each of them on a dedicated device. This works better on densely-connected layers since it reduces the communication cost for synchronizing parameters.
Nevertheless, this approach results in a suboptimal runtime performance in large scale distributed training since different layers in a network may prefer different parallelization strategies.

2 Contributions

The authors propose a layer-wise parallelism technique, which allows different layers in a network to use individual parallelization configurations.

Also, the authors defined the search space of possible parallelization configurations for a layer and presented a cost model to quantitively evaluate the runtime performance of training a network. This includes an algorithm to jointly find a global optimal parallelization strategy.

Additionally, the authors provided an implementation that supports layer-wise parallelism and show that this technique can increase training throughput and reduce communication costs over state-of-the-art approachers, while the scalability to multiple GPUs is improved. This outperforms state-of-the-art approaches by increasing training throughput, reducing communication costs, achieving better scalability to multiple GPUs, while maintaining original network accuracy.

3 Solution

The author propose a layer-wise parallelism that allows each layer in a network to use an individual parallelization strategy. It performs the same computation for each layer as it is defined in the original network, thus it keeps the same accuracy. Then, the goal is to find the optimal individual parallelization stategy for each layer by solving a graph search problem.

The author started by defining a cost model to quantitively evaluate the runtime of different parallelization strategies, and to use a dynamic programming based graph search algorithm to find an optimal parallelization strategy out of it. This model related the time to process each layer including forward and back propagation; the time it takes to transfer the input tensors between the devices considering their sizes and communication bandwidth; and the time it requires to synchronize each parameter after back

propagation. This equation expresses the problem of finding an optimal parallelization strategy.

In practice in CNNs, each layer is only connected to a few layers with similar depths in a computation graph. Therefore, it is feasible to simplify computation on nodes and edges. In case an node is only connected by one edge to its input and with another one to its output, it is possible to eliminate this node and use a new edge to connect the previous layer with the next one. Also, in case a node is connected with two edges to another layer, it is posible to replace both edges by a single one. This algorithm has been tested and it achieved a lower time complexity and reduced the execution time by orders of magnitude over the baseline (*depth-first search* algorithm).

4 Strong Points

- **Clear structure and concise text**. The paper has a very clear structure, where each section has purposeful content with a suitable heading. The text is concise and easily read.

- **Good related work investigation**. The authors performed a good related work investigation that clearly pointed out the contributions and limitations of previous work, which enables readers to have even better understanding of authors' motivation to this research.

- **Clear Experiment setup description**. The authors did well in presenting their experiment setup in detail such as software, hardware, dataset, which significantly increases the reproducibility of their research result..

5 Weak Points

- **Inadequate cost model evaluation** The experiment result or execution time was measured by the cost model introduced by the authors. We think the authors should also evaluate their research with other cost models or motivate their cost model comparing to others.

- **Inadequate of experiment result discussion** The authors provided a brief analysis of their experiment result that showed the performance was improved. However the analysis or discussion was not referring to the details of the methods the developed and did not discuss why their approach improves the performance.

- **Not implemented in mainstream DL framwork** The authors implemented thier framework in Legion on the ground of limited interfaces in TensorFlow, PyTorch and Caffe2, which implies negative impacts on the adoption of the proposed approach.

- **Lack of Future work suggestion** The authors did not propose any future works or improvements to their work.