Contents

Big Data Computing

First Lecture

Classroom: phdlyrl

Second Lecture

What's Big Data?

It is sometimes used as a buzzword, but it describes an actual phenomenon.

5 V's:

1. Value: Extracting knowledge from data.
2. Volume: The amount of data. (Measuring in terabytes, petabytes, exabytes)
3. Variety: Different formats, structured (SQL tables) and Unstructured (text, images, videos).
4. Velocity: The speed at which data is generated. (usually in real-time, you can't parse it all)
5. Veracity: The quality of the data. (Is it reliable?)

Scaling

When scaling, you have two options:

- Scaling *up*, you buy better hardware for your monolithic machine, however, moore's law is slowing down, and you will get diminishing returns, in general, you will not be able to keep up with increases in Volume/Velocity.
- Scaling *out*, you buy more machines, and distribute the workload, this is the way to go for big data, since this allows for exponential growth.

Scaling *out*, or *horizontal scaling*, also allows you for more flexibility, as once you deal with the orchestration of the workload, you can dynamically increase/reduce the number of allocated machines depending on a rolling cost/benefit analysis.

Network Bottlenecks

A bottleneck during a request is generated when:

- There is a skew in the ratio of communication/computation
- There is a non-overlappable part of communication/computation

Third Lecture

In this lecture, we will take a look at a use case of Big Data, *Distributed Deep Learning*.

Naturally, Deep Learning performs better with more data, this is actually an empirically derived law that has come up in recent OpenAI research: exponentially more data leads to a linear increase in model performance, irrespective of the model architecture (assuming a reasonable architecture).

How does a DNN work

To put it simply, we want to build a function estimator over a dataset D:

$$\hat{f} : D \to Y$$

Such that:

$$\hat{f}(x) \approx y$$

Where x is the input data, and y is the output data.

To do this, we use iterative optimization, where we start with a random function, and we iteratively update it to minimize the error between the output of the function and the actual output. We use the data from the dataset to ***fit*** our model, under an ***inductive*** bias, which is an assumption that the ***test*** data is similarly distributed (so that our fit will generalize).

Forward and Backward passes

Our iterative fitting process can be split in two passes:

1. Forward pass, where we compute the output of the model given the input.
2. Backward pass, where we compute the gradient of the error with respect to the model parameters.

We use the obtained gradient to refine the model parameters, and we repeat the process until we reach a stopping criterion (which indicates that the model has converged).

In general, most of these operations are matrix operations, which can be parallelized on the GPU in an efficient manner.

Parallelism Models

There are three orthogonal parallelism directions in which we can distribute the workload:

1. Data Parallelism
2. Pipeline Parallelism
3. Model Parallelism

The last two together are also known as *Model Parallelism*.

Data Parallelism

The simplest one is *Data Parallelism*. We split the data in n subsets, and we train n models in parallel, each on a different subset. We then perform *gradient aggregation* across these sets, and we update the model parameters.

In this way all the models are updated in parallel, and the only bottleneck is the gradient aggregation.

Gradient aggregation is *quite* costly, it contributes to a significant portion (20% to 50%) of total training time, with the portion increasing as the available bandwidth decreases.

Naturally, this means that this operation has been the subject of heavy optimization, algorithmically and infrastructurally.

Parameter Servers

An initial solution was to use a *Parameter Server*, which is a server that holds the model parameters, and is responsible for the aggregation of the gradients. This does *not* scale well, as it introduces a single point of failure, and a bottleneck in the network.

Splitting the Aggregation

We can split the aggregation into K servers, which will each aggregate a subset of the gradients, and then we will aggregate the results of these servers. This is a more scalable solution, but it still has a bottleneck in the aggregation.

Naturally, as long as K is small, this is a good solution, but as K grows, the aggregation time will grow linearly, and the bottleneck will be reached.

Cost Model

A cost model on a K-PS (Parameter Server) system is:

$$\max\left(\frac{n}{\beta}, \frac{pn}{k\beta}\right)$$

Where n is the number of bytes in the gradient vector, p is the number of GPUs/workers, k is the number of servers, and β is the bandwidth of the network.

Data Parallelism Recap

In data parallelism, each worker must keep a copy of the model, therefore you can also encounter out-of-memory errors, as the model grows in size.

Data parallelism is simple and easy to implement, but has limited potential to be scalable.

Replication of the model is also a waste of resources, as we are just keeping copies around, and we are not using them. It should only be used as a last resort if we have more resources than we can use after applying the other parallelism models, and we still want idle time to be minimized.

Model Parallelism

In order to train models that do *not* fit into a single GPU, we can use *Model Parallelism.* in essence, model parallelism is the process of splitting the model into k parts, and training each part on a different GPU.

Now, the bottleneck is the GPU-to-GPU communication, since we have a sequential dependency between the parts of the model (each part depends on the output of the previous part). This happens in both the forward and backward passes.

Pipelining

Naturally, you can reduce the *bubble overhead* by pipelining the computation, so that the forward pass of the next part can start before the backward pass of the previous part has finished. The same concept can be applied on subsections of the forward/backward passes that use layers that are distributed across a network.

Operator Parallelism

In the case of *Operator Parallelism*, we split the model into operators, and we distribute the operators across the network. This is a more fine-grained approach, and it is usually used in conjunction with the other two parallelism models.

Here, the bottleneck is the communication between the operators, and the synchronization of the operators, however, we reduce the idle time of GPUs drastically.

Hierarchical Parallelism

Naturally, *Operator Parallelism* is applied on GPUs that are on the same node, so that the available bandwidth is maximized. Whereas *Model Parallelism* is applied on GPUs that are on different nodes, where communication is more costly.

This allows us to maximize the utilization of the available resources, and minimize the communication overhead.

Hybrid Parallelism

Hybrid parallelism (also called 3D-Parallelism), combines all three techniques, and offers maximum scalability and potential for resource utilization.

In general, one must be working with a sufficient quantity of data to *justify* the use of these techniques, as the overhead of the parallelism models can be significant, and can outweigh the benefits of the parallelism.

Challenges

Compute-Communication Overlap

When employing hybrid-parallelism, we perform computations and communications in a very fine-grained manner. The reason is that we want to perform gradient aggregation after each layer, so that we minimize the *bubbling* in our pipeline (we don't want to keep anyone waiting).

We want to maximize the overlap in windows of computation and communication, so that we can keep the GPUs and network as busy as possible.

Gradient Compression

Since DL is a stochastic process, one can afford to have some noise in the gradients, and still converge to a good solution. This allows us to compress the gradients, and reduce the communication overhead.

We have various techniques:

1. Quantization: Reducing the precision of the gradients.
2. Rank Decomposition: Decomposing the gradients into a low-rank matrix.
3. Sparsification: Sample the top k gradients, and send only those.

In sparsification, we can use various techniques to select the top k gradients, such as:

1. Top-k: Select the top k gradients by magnitude.
2. Thresholding: Select the gradients that are above a certain threshold λ.

There are theoretical proofs that show that asymptotically, model convergence is not affected by gradient sparsification, as long as the sparsification is done in a *smart* way.

Compression Tradeoff By compressing gradients, we lower the cost of a single iteration, but we increase the number of iterations needed to converge. This is a tradeoff that must be taken into account when designing a distributed deep learning system.

We have a real speed up if:

$$T_{comp} \cdot I_{comp} < T_{orig} \cdot I_{orig}$$

Aggregation Consistency

We observe that gradient accumulation acts as a *barrier*, which means that if a single GPU is, for some reason, considerably slower than the others, it will slow down the whole process. The probability of this happening is significant in large-scale systems, and it is a problem that must be addressed.

Possible solutions include: * Gradient Accumulation: Accumulate the gradients over multiple iterations, and then send them. * Ignoring Slow Workers: Ignore the gradients from slow workers, and only aggregate the gradients that you have currently available after a certain time.

Reduce Operations

We can use smart reduction techniques common in IPC (Inter-Process Communication) to reduce the number of operations needed to aggregate the gradients.

We keep a table of operations and total volume of communication:

Operation	Volume	Steps
Ideal	n	1
Parameter Server	$max(n, \frac{pn}{k})$	1
Näive AllReduce	$(p-1)\cdot n$	1
Ring AllReduce	$2\cdot N$	$2(p-1)$
Bandwidth-optimal recursive doubling	$2\cdot N$	$2\log_2(p)$

We explain each:

- Ideal: The ideal case, where we have a perfect system that can aggregate the gradients in a single operation.
- Parameter Server: The case where we have a parameter server, and we have to send the gradients to the server, and then the server sends the aggregated gradients back.
- Näive AllReduce: The case where we have a simple all-reduce operation, where each worker sends the gradients to all the other workers, and then each worker aggregates the gradients.
- AllReduce (ReduceScatter-AllGather): The case where we have a more sophisticated all-reduce operation, we scatter p subsets of the gradients to p workers, each worker aggregates the subset, and then we gather the results across all workers.

Ring AllReduce

We go in depth into the AllReduce operation. First, we build a logical ring topology of N workers, such that each one is connected to the next one, and the last one is connected to the first one.

We take our vector g, we divide it into N parts, and we send each part to the next worker. Each worker then aggregates the received part with its own part, and then sends the result to the next worker.

This results in a full aggregation of the gradients in $(p-1)$ steps, another full rotation is then performed to gather the results, and we have our aggregated gradients.

The communication volume is:

$$2 \cdot \frac{pn}{k} \approx 2n$$

Which is a 2-approximation of the ideal case.

We use a Ring AllReduce instead of a Tree AllReduce, since the Ring AllReduce is capable of only sending parts of the gradients, and not the full gradients, which can be problematic in the case of large models.

Bandwidth-optimal recursive doubling

A variation of the Ring AllReduce is the Bandwidth-optimal recursive doubling, this allows us to perform the aggregation in a logarithmic number of steps, and it is optimal in terms of bandwidth.

Latency-optimal recursive doubling

Using the same binary-tree like topology as the Bandwidth-optimal recursive doubling, we can build a latency-optimal recursive doubling, which is optimal in terms of latency, by sending the entire gradient vector in a single step.