POSTER P5285 | CONTACT NAME Min Lin: mavenlin@gmail.com

GPU TECHNOLOGY CONFERENCE

Purine: a bi-graph based deep learning framework

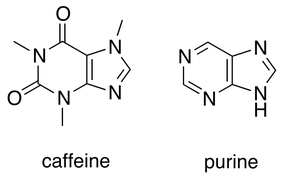
Min Lin[1, 2], Shuo Li[2], Xuan Luo[2], Shuicheng Yan[2]

[1] Graduate School for Integrative Sciences and Engineering
[2] Department of Electronic & Computer Engineering

NUS National University of Singapore

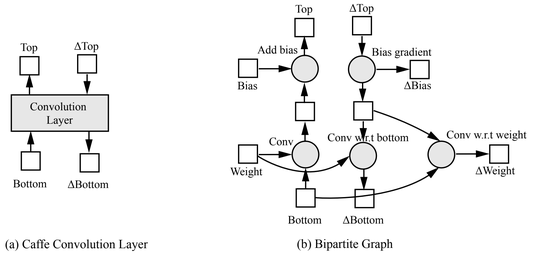
The name

1. We benefited from the open source deep learning framework Caffe.
2. The math functions and core computations are adapted from Caffe.
3. Similar molecular structure.



Bi-Graph abstraction

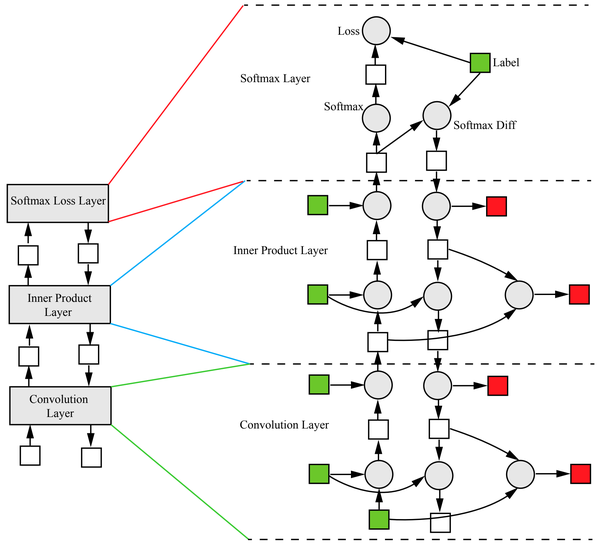
Comparison of Caffe layers and their bigraph representations.



(a) Caffe Convolution Layer (b) Bipartite Graph

Optimization

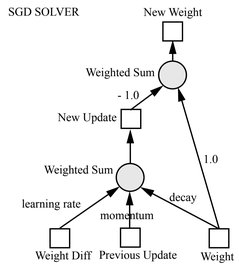
Example Network converted to Bi-graph representation.



1. Start from sources and end at sinks of the graph.
2. Prune unnecessary nodes.

Advantages of Bi-Graph Abstraction

1. Less hard coding and more reusability.
2. All concepts are consistently expressed with graph. (SGD solver, Forward & backward pass, etc.)



3. Flexible to implement various schemes of parallelization.

```
type: blob
name: weight
size: [96, 3, 11, 11]
location:
  ip: 127.0.0.1
  device: 0
```

Example Op defined in YAML

```
type: op
op_type: Conv
name: conv1
inputs: [ bottom, weight ]
outputs: [ top ]
location:
  ip: 127.0.0.1
  device: 0
thread: 1
other fields ...
```

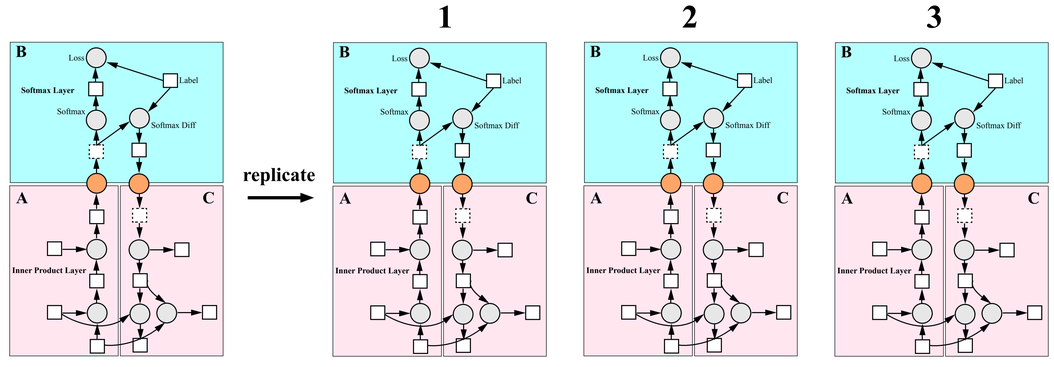
Location:
The location that the blob/op resides on, including:

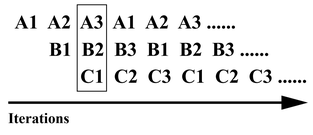
- ip address of the target machine
- what device it is on (CPU/GPU)

Thread:
Thread is needed for op because both CPU and GPU can be multiple threaded (Streams in terms of NVIDIA GPU).

Parallelization

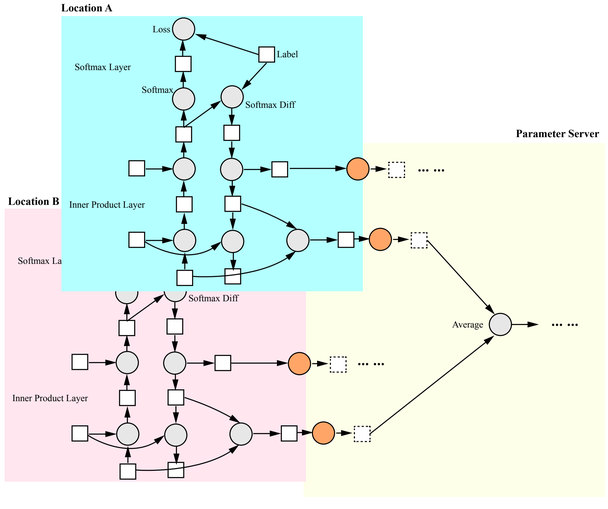
PIPELINING

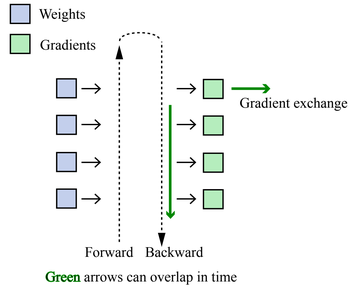




DATA PARALELLISM

1. Asynchronous update hides communication latency.
2. Synchronous (all reduce) is possible by overlapping data transfer with computation.





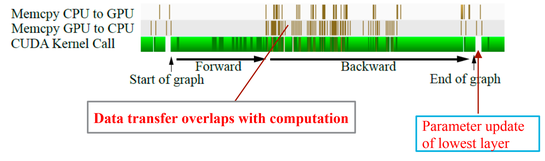
Green arrows can overlap in time

Higher layer gradients are computed earlier than lower layers.

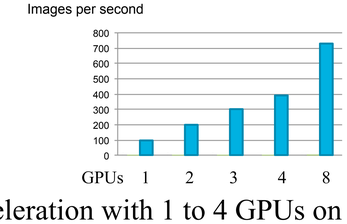
Higher layer can send gradients to parameter server and get them back while the lower layers are doing their computation.

Especially true for very deep networks

Profiling result with nvprofiler.



Acceleration ratio with different number of GPUs.



Linear acceleration with 1 to 4 GPUs on the same machine. Approximate linear acceleration with 8 GPUs on two machines interconnected with 1 gigabit ethernet.