The size of current state-of-the-art networks makes computation a critical issue, in particular for training and optimizing meta-parameters.

Although they were historically developed for mass-market real-time CGI, the highly parallel architecture of GPUs is extremely fitting to signal processing and high dimension linear algebra.

Their use is instrumental in the success of deep-learning (Raina et al., 2009; Ciresan et al., 2010; Krizhevsky et al., 2012; Shi et al., 2016).
A standard NVIDIA GTX 1080 has 2,560 single-precision computing cores clocked at 1.6GHz, and delivers a peak performance of \(\sim 9\) TFlops.

The precise structure of a GPU memory and how its cores communicate with it is a complicated topic that we will not cover here.

François Fleuret
Deep learning / 6.6. Using GPUs

<table>
<thead>
<tr>
<th>Network</th>
<th>Mini-batch size</th>
<th>CPU</th>
<th>Use</th>
<th>Runtime (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN-S</td>
<td>64</td>
<td>Caffe</td>
<td>Single threaded</td>
<td>0.120</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Caffe</td>
<td>Dual threaded</td>
<td>0.126</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MXNet</td>
<td>Single threaded</td>
<td>1.935</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MXNet</td>
<td>Dual threaded</td>
<td>1.940</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Torch</td>
<td>Single threaded</td>
<td>1.065</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Torch</td>
<td>Dual threaded</td>
<td>1.070</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RenNet-50</td>
<td>Single threaded</td>
<td>4.351</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RenNet-50</td>
<td>Dual threaded</td>
<td>4.356</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FCN-R</td>
<td>Single threaded</td>
<td>0.091</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FCN-R</td>
<td>Dual threaded</td>
<td>0.094</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MXNet</td>
<td>Single threaded</td>
<td>2.395</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MXNet</td>
<td>Dual threaded</td>
<td>2.400</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Torch</td>
<td>Single threaded</td>
<td>1.110</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Torch</td>
<td>Dual threaded</td>
<td>1.115</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LSTM</td>
<td>Single threaded</td>
<td>43.243</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LSTM</td>
<td>Dual threaded</td>
<td>55.172</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Torch</td>
<td>Single threaded</td>
<td>0.040</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Torch</td>
<td>Dual threaded</td>
<td>0.043</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G1080</td>
<td>Single threaded</td>
<td>0.060</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G1080</td>
<td>Dual threaded</td>
<td>0.063</td>
</tr>
<tr>
<td></td>
<td></td>
<td>K80</td>
<td>Single threaded</td>
<td>1.340</td>
</tr>
<tr>
<td></td>
<td></td>
<td>K80</td>
<td>Dual threaded</td>
<td>1.343</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Single GPU</td>
<td>Single threaded</td>
<td>4.351</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Single GPU</td>
<td>Dual threaded</td>
<td>4.356</td>
</tr>
<tr>
<td></td>
<td>1024</td>
<td>Caffe</td>
<td>Single threaded</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Caffe</td>
<td>Dual threaded</td>
<td>0.033</td>
</tr>
<tr>
<td></td>
<td>128</td>
<td>MXNet</td>
<td>Single threaded</td>
<td>2.977</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MXNet</td>
<td>Dual threaded</td>
<td>3.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Torch</td>
<td>Single threaded</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Torch</td>
<td>Dual threaded</td>
<td>0.017</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LSTM</td>
<td>Single threaded</td>
<td>1.567</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LSTM</td>
<td>Dual threaded</td>
<td>1.592</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Torch</td>
<td>Single threaded</td>
<td>0.137</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Torch</td>
<td>Dual threaded</td>
<td>0.152</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G1080</td>
<td>Single threaded</td>
<td>0.045</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G1080</td>
<td>Dual threaded</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td></td>
<td>K80</td>
<td>Single threaded</td>
<td>1.845</td>
</tr>
<tr>
<td></td>
<td></td>
<td>K80</td>
<td>Dual threaded</td>
<td>1.870</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Single GPU</td>
<td>Single threaded</td>
<td>2.222</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Single GPU</td>
<td>Dual threaded</td>
<td>2.250</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LSTM</td>
<td>Single threaded</td>
<td>1.329</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LSTM</td>
<td>Dual threaded</td>
<td>1.347</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Torch</td>
<td>Single threaded</td>
<td>0.090</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Torch</td>
<td>Dual threaded</td>
<td>0.093</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G1080</td>
<td>Single threaded</td>
<td>0.366</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G1080</td>
<td>Dual threaded</td>
<td>0.382</td>
</tr>
<tr>
<td></td>
<td></td>
<td>K80</td>
<td>Single threaded</td>
<td>0.702</td>
</tr>
<tr>
<td></td>
<td></td>
<td>K80</td>
<td>Dual threaded</td>
<td>0.729</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Single GPU</td>
<td>Single threaded</td>
<td>0.447</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Single GPU</td>
<td>Dual threaded</td>
<td>0.464</td>
</tr>
</tbody>
</table>

Note: The mini-batch sizes for FCN-S, AlexNet-S, ResNet-50, FCN-R, AlexNet-R, ResNet-56 and LSTM are 64, 16, 16, 1024, 1024, 128 and 128 respectively.

(Shi et al., 2016)
The current standard to program a GPU is through the CUDA (“Compute Unified Device Architecture”) model, defined by NVIDIA.

Alternatives are OpenCL, backed by several CPU/DSP manufacturers, and more recently AMD’s HIP/ROCm.

Google developed its own line of processors for deep learning dubbed TPU (“Tensor Processing Unit”) which offer excellent flops/watt performance.

In practice, as of today (23.03.2020), NVIDIA hardware remains the default choice for deep learning, and CUDA is the reference framework in use.

From a practical perspective, libraries interface the framework (e.g. PyTorch) with the “computational backend” (e.g. CPU or GPU)

- BLAS (“Basic Linear Algebra Subprograms”): vector/matrix products, and the cuBLAS implementation for NVIDIA GPUs,
- LAPACK (“Linear Algebra Package”): linear system solving, Eigen-decomposition, etc.
- cuDNN (“NVIDIA CUDA Deep Neural Network library”) computations specific to deep-learning on NVIDIA GPUs.
Using GPUs in PyTorch

The use of the GPUs in PyTorch is done by creating or copying tensors into their memory.

Operations on tensors in a device’s memory are done by the said device.
As for the type, the device can be specified to the creation operations as a device, or as a string that will implicitly be converted to a device.

```python
>>> x = torch.zeros(10, 10)
>>> x.device
device(type='cpu')
>>> x = torch.zeros(10, 10, device = torch.device('cuda'))
>>> x.device
device(type='cuda', index=0)
>>> x = torch.zeros(10, 10, device = torch.device('cuda:1'))
>>> x.device
device(type='cuda', index=1)
>>> x = torch.zeros(10, 10, device = 'cuda:0')
>>> x.device
device(type='cuda', index=0)
```

The `torch.Tensor.to(device)` returns a clone on the specified device if the tensor is not already there or returns the tensor itself if it was already there.

The argument device can be either a string, or a device.

Alternatives are `torch.Tensor.cuda([gpu_id])` and `torch.Tensor.cpu()`.

⚠️ Moving data between the CPU and the GPU memories is far slower than moving it inside the GPU memory.
>>> u = torch.tensor([1, 2, 3])
>>> u.device
device(type='cpu')
>>> v = u.to('cuda') # copy of u
>>> v
tensor([1, 2, 3], device='cuda:0')
>>> v[0] = 5
>>> u
tensor([1, 2, 3])
>>> w = u.to('cpu') # this is u itself
>>> w
tensor([1, 2, 3])
>>> w[0] = 5
>>> u
tensor([5, 2, 3])

>>> m = torch.empty(10, 10).normal_()
>>> m.device
device(type='cpu')
>>> x = torch.empty(10, 100).normal_()
>>> q = m@x
>>> q.device
device(type='cpu')
>>> m = m.to('cuda')
>>> x = x.to('cuda')
>>> q = m@x # This is done on GPU (#0)
>>> q.device
device(type='cuda', index=0)
Since operations maintain the types and devices of the tensors, you generally do not need to worry about making your code generic regarding these aspects.

To explicitly create new tensors you can use a tensor’s new_*() methods.

```python
>>> u = torch.empty(3, 5, dtype = torch.float64).normal_()
>>> v = u.new_zeros(1, 2)
>>> v
tensor([[0., 0.]], dtype=torch.float64)
>>> w = torch.empty(3, 5, dtype = torch.float16, device = 'cuda:1').fill_(1.0)
>>> w.new_full((2, 3), 1.4)
tensor([[1.4004, 1.4004, 1.4004],
        [1.4004, 1.4004, 1.4004]], device='cuda:1', dtype=torch.float16)
```

Apart from copy_(), operations cannot mix different tensor types or devices:

```python
>>> import torch
>>> x = torch.empty(3, 5).normal_()
>>> y = torch.empty(3, 5).normal_.to('cuda')
>>> x.copy_(y)
tensor([[ 0.4071, 0.7589, -0.5321, 0.9103, -1.4985],
        [-0.1059, 2.1554, -0.0774, -0.4520, 1.5123],
        [ 0.1322, 0.1002, -0.4071, 1.8927, -0.5800]])
```

```python
>>> x + y
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
RuntimeError: Expected object of type torch.FloatTensor but found type torch.cuda.FloatTensor for argument #3 'other'
```

Similarly if multiple GPUs are available, cross-GPUs operations are not allowed by default, with the exception of copy_().
The method `torch.Module.to(device)` moves all the parameters and buffers of the module (and registered sub-modules recursively) to the specified device.

⚠ Although they do not have a “_” in their names, these `Module` operations make changes in-place.

The method `torch.cuda.is_available()` returns a Boolean value indicating if a GPU is available, so a typical GPU-friendly code would start with

```python
if torch.cuda.is_available():
    device = torch.device('cuda')
else:
    device = torch.device('cpu')
```

and then have some

```python
model.to(device)
criterion.to(device)
train_input, train_target = train_input.to(device), train_target.to(device)
test_input, test_target = test_input.to(device), test_target.to(device)
```
A very simple way to leverage multiple GPUs is to wrap the model in a `nn.DataParallel`.

The forward of `nn.DataParallel(my_module)` will

1. split the input mini-batch along the first dimension in as many mini-batches as there are GPUs,
2. send them to the forwards of clones of `my_module` located on each GPU,
3. concatenate the results.

And it is (of course!) autograd-compliant.
If we define a simple module to printout the calls to forward.

```python
class Dummy(nn.Module):
    def __init__(self, m):
        super(Dummy, self).__init__()
        self.m = m

    def forward(self, x):
        print('Dummy.forward', x.size(), x.device)
        return self.m(x)
```

```python
x = torch.empty(50, 10).normal_()
model = Dummy(nn.Linear(10, 5))

print('On CPU')
y = model(x)

x = x.to('cuda')
model.to('cuda')

print('On GPU w/o nn.DataParallel')
y = model(x)

print('On GPU w/ nn.DataParallel')
parallel_model = nn.DataParallel(model)
y = parallel_model(x)
```

will print, on a machine with two GPUs:

```
On CPU
Dummy.forward torch.Size([50, 10]) cpu
On GPU w/o nn.DataParallel
Dummy.forward torch.Size([50, 10]) cuda:0
On GPU w/ nn.DataParallel
Dummy.forward torch.Size([25, 10]) cuda:0
Dummy.forward torch.Size([25, 10]) cuda:1
```
References


