Deep learning

6.6. Using GPUs

François Fleuret

https://fleuret.org/dlc/
The size of current state-of-the-art networks makes computation a critical issue, in particular for training and optimizing meta-parameters.
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 3090 has 10,500 computing cores clocked at 1.5GHz, and delivers a peak performance of $\approx 35$ 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.
A standard NVIDIA GTX 3090 has 10,500 computing cores clocked at 1.5GHz, and delivers a peak performance of $\approx 35$ 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.
A standard NVIDIA GTX 3090 has 10,500 computing cores clocked at 1.5GHz, and delivers a peak performance of $\approx 35$ 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.
A standard NVIDIA GTX 3090 has 10,500 computing cores clocked at 1.5GHz, and delivers a peak performance of \(\sim 35\) TFlops.
A standard NVIDIA GTX 3090 has 10,500 computing cores clocked at 1.5GHz, and delivers a peak performance of \( \sim 35 \) TFlops.
A standard NVIDIA GTX 3090 has 10,500 computing cores clocked at 1.5GHz, and delivers a peak performance of \(\sim 35\) TFlops.
A standard NVIDIA GTX 3090 has 10,500 computing cores clocked at 1.5GHz, and delivers a peak performance of \( \sim 35 \) TFlops.
A standard NVIDIA GTX 3090 has 10,500 computing cores clocked at 1.5GHz, and delivers a peak performance of $\sim 35$ 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.
### TABLE 7. COMPARATIVE EXPERIMENT RESULTS (TIME PER MINI-BATCH IN SECOND)

<table>
<thead>
<tr>
<th>Platform</th>
<th>Desktop CPU (Threads used)</th>
<th>Server CPU (Threads used)</th>
<th>Single GPU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>G980</td>
<td>G1080</td>
</tr>
<tr>
<td>FCN-S</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caffe</td>
<td>1.329 0.710 0.423</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNTK</td>
<td>1.227 0.660 0.435</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF</td>
<td>7.062 4.789 2.648</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MXNet</td>
<td>4.621 2.607 2.162</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Torch</td>
<td>4.645 2.429 1.424</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AlexNet-S</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caffe</td>
<td>1.606 0.999 0.719</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNTK</td>
<td>3.761 1.974 1.276</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF</td>
<td>6.525 2.936 1.749</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MXNet</td>
<td>2.977 2.340 2.250</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Torch</td>
<td>4.645 2.429 1.424</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RenNet-50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caffe</td>
<td>11.554 7.671 5.652</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNTK</td>
<td>23.905 16.435 10.206</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MXNet</td>
<td>48.000 46.154 44.444</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Torch</td>
<td>13.178 7.500 4.736</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FCN-R</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caffe</td>
<td>2.476 1.499 1.149</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNTK</td>
<td>1.845 0.970 0.661</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF</td>
<td>2.647 1.913 1.157</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MXNet</td>
<td>1.914 1.072 0.719</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Torch</td>
<td>1.670 0.926 0.565</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AlexNet-R</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caffe</td>
<td>3.558 2.587 2.157</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNTK</td>
<td>9.956 7.263 5.519</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF</td>
<td>4.355 3.225 1.911</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MXNet</td>
<td>13.401 12.305 12.278</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Torch</td>
<td>5.352 3.866 3.162</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RenNet-56</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caffe</td>
<td>6.741 5.451 4.989</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNTK</td>
<td>- - -</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF</td>
<td>- - -</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MXNet</td>
<td>34.409 31.255 30.069</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Torch</td>
<td>5.758 3.222 2.368</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caffe</td>
<td>- - -</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNTK</td>
<td>0.186 0.120 0.090</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF</td>
<td>4.662 3.385 1.935</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MXNet</td>
<td>- - -</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Torch</td>
<td>6.921 3.831 2.682</td>
<td></td>
<td></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.
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.
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.
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)
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)
```
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)
```
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')
```

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

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

```python
>>> 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.
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()`. 
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])
>>> 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')
```python
>>> 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]])

>>> 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'
```
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]])

>>> 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_()`.

Another exception to this rule are 0d tensors, which act as scalars and can be combined without device constraint.
The method \texttt{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 \texttt{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
device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')
```

and then have some `device = device` in some places, and/or

```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)
```
Multiple GPUs with `nn.DataParallel`
A very simple way to leverage multiple GPUs is to wrap the model in a `nn.DataParallel`. 
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`.

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

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


