DEEP LEARNING WITH GPUS
Maxim Milakov, Senior HPC DevTech Engineer, NVIDIA
TOPICS COVERED

Convolutional Networks
Deep Learning
Use Cases
GPUs
cuDNN
MACHINE LEARNING

- **Training**
  - Train the model from supervised data

- **Classification (inference)**
  - Run the new sample through the model to predict its class/function value
CONVOLUTIONAL NETWORKS

Local Receptive Fields

- Neurophysiologists David Hubel and Torsten Wiesel, 1962

CONVOLUTIONAL NETWORKS

Neocognitron: shared weights

- Kunihiko Fukushima, 1980

CONVOLUTIONAL NETWORKS

Training DNN with Backpropagation

- Yann LeCun et al, 1998

- MNIST: 0.7% error rate

"Gradient-Based Learning Applied to Document Recognition", Proceedings of the IEEE 1998
High need for computational resources
Low ConvNet adoption rate until ~2010
USE CASES

GTSRB: Traffic sign recognition

- The German Traffic Sign Recognition Benchmark, 2011

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team</th>
<th>Error rate</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IDSIA, Dan Ciresan</td>
<td>0.56%</td>
<td>CNNs, trained using GPUs</td>
</tr>
<tr>
<td>2</td>
<td>Human</td>
<td>1.16%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>NYU, Pierre Sermanet</td>
<td>1.69%</td>
<td>CNNs</td>
</tr>
<tr>
<td>4</td>
<td>CAOR, Fatin Zakhouta</td>
<td>3.86%</td>
<td>Random Forests</td>
</tr>
</tbody>
</table>

http://benchmark.ini.rub.de/?section=gtsrb
USE CASES

ImageNet: natural image classification

- Alex Krizhevsky et al, 2012
- 1.2M training images, 1000 classes
- Scored 15.3% Top-5 error rate with 26.2% for the second-best entry for classification task
- CNNs trained with GPUs

http://www.image-net.org/challenges/LSVRC/
USE CASES

ImageNet: results for 2010-2014

- % Teams using GPUs
- Top-5 error

- 2010: 28%
- 2011: 26%
- 2012: 15%
- 2013: 83%
- 2014: 95%

0% 5% 10% 15% 20% 25% 30%

0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

2010 2011 2012 2013 2014
USE CASES

Dogs vs. Cats: Transfer Learning

- Dogs vs. Cats, 2014
- Train model on one dataset - ImageNet
- Re-train the last layer only on a new dataset - Dogs and Cats

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team</th>
<th>Error rate</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pierre Sermanet</td>
<td>1.1%</td>
<td>CNNs, model transferred from ImageNet</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Maxim Milakov</td>
<td>1.9%</td>
<td>CNN, model trained on Dogs vs. Cat dataset only</td>
</tr>
</tbody>
</table>

https://www.kaggle.com/c/dogs-vs-cats
USE CASES

Speech recognition

- Acoustic model is DNN
  - Usually fully-connected layers
  - Some try using convolutional layers with spectrogram used as input
  - Both fit GPU perfectly
- Language model is weighted Finite State Transducer (wFST)
  - Beam search runs fast on GPU

It is all about supercomputing, right?
## GPU

### Tesla K40 and Tegra K1

<table>
<thead>
<tr>
<th></th>
<th>NVIDIA Tesla K40</th>
<th>NVIDIA Jetson TK1</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUDA cores</td>
<td>2880</td>
<td>192</td>
</tr>
<tr>
<td>Peak performance, SP</td>
<td>4.29 Tflops</td>
<td>326 Gflops</td>
</tr>
<tr>
<td>Peak power consumption</td>
<td>235 Wt</td>
<td>~10 Wt, for the whole board</td>
</tr>
<tr>
<td>Deep Learning tasks</td>
<td>Training, Inference</td>
<td>Inference, Online Training</td>
</tr>
</tbody>
</table>

USE CASES

Pedestrian detection+ on Jetson TK1

- Ikuro Sato, Hideki Niihara, R&D Group, Denso IT Laboratory, Inc.
- Real-time pedestrian detection with depth, height, and body orientation estimations

How do we run DNNs on GPUs?
CUdNN

cuDNN (and cuBLAS)

- Library for DNN toolkit developer and researchers
- Contains building blocks for DNN toolkits
  - Convolutions, pooling, activation functions e t.c.
- Best performance, easiest to deploy, future proofing
- Jetson TK1 support coming soon!

[developer.nvidia.com/cuDNN](https://developer.nvidia.com/cuDNN)

- cuBLAS (SGEMM for fully-connected layers) is part of CUDA toolkit,
cuDNN is already integrated in major open-source frameworks

- Caffe - [caffe.berkeleyvision.org](https://caffe.berkeleyvision.org)
- Torch - [torch.ch](https://torch.ch)
- Theano - [deplearning.net/software/theano/index.html](http://deplearning.net/software/theano/index.html), already has GPU support, cuDNN support coming soon!
REFERENCES

- HPC by NVIDIA: www.nvidia.com/tesla
- Jetson TK1 Development Kit: www.nvidia.com/jetson-tk1
- CUDA Zone: developer.nvidia.com/cuda-zone
- Parallel Forall blog: devblogs.nvidia.com/parallelforall
- Contact me: mmilakov@nvidia.com