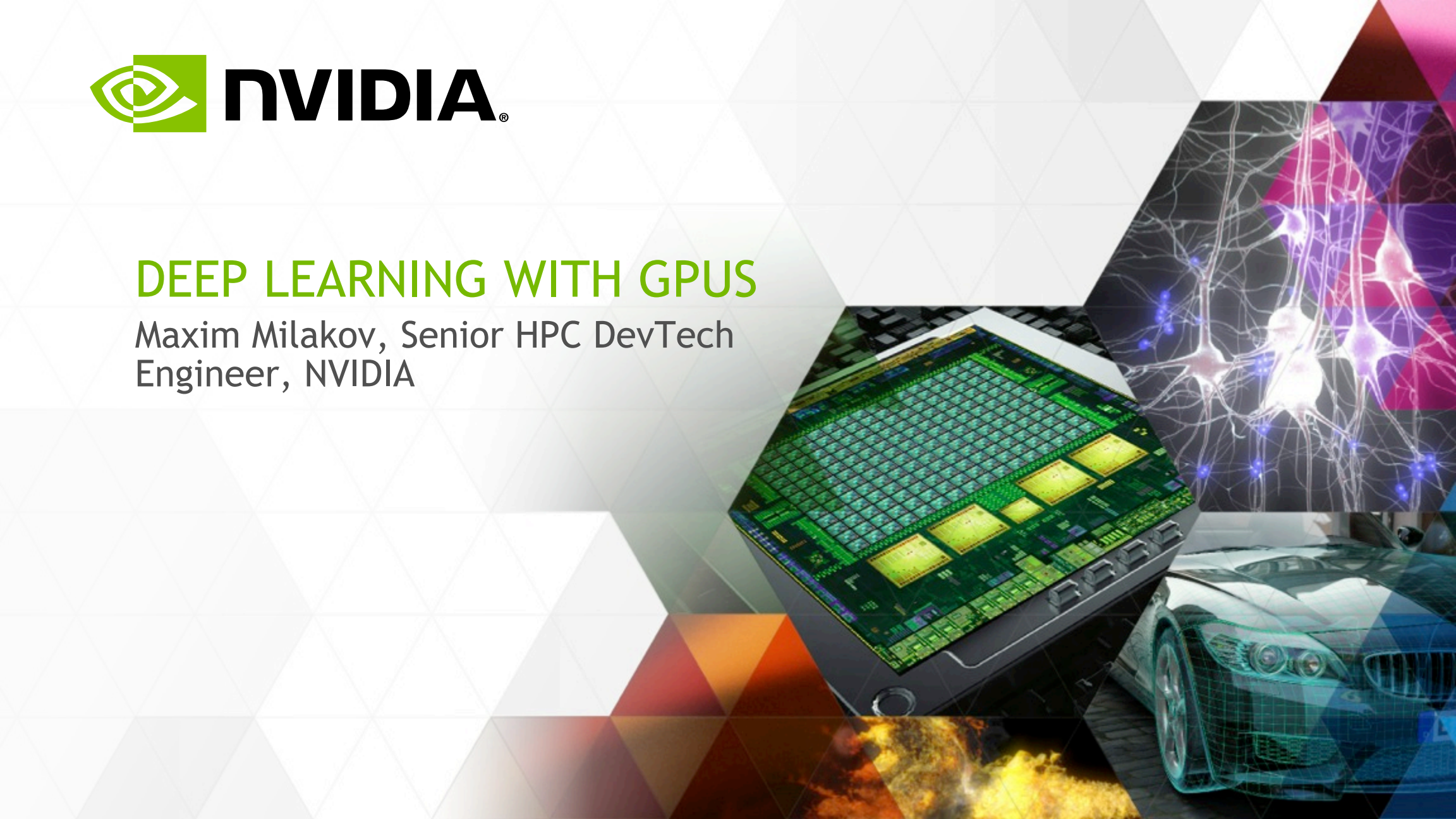


DEEP LEARNING WITH GPUS

Maxim Milakov, Senior HPC DevTech Engineer, NVIDIA

TOPICS COVERED

Convolutional Networks

Deep Learning

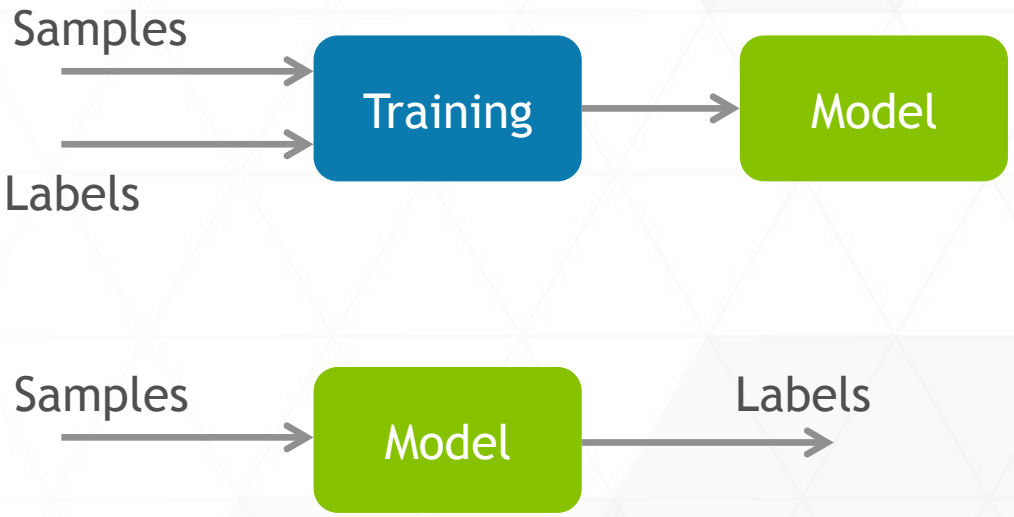
Use Cases

GPUs

cuDNN

MACHINE LEARNING

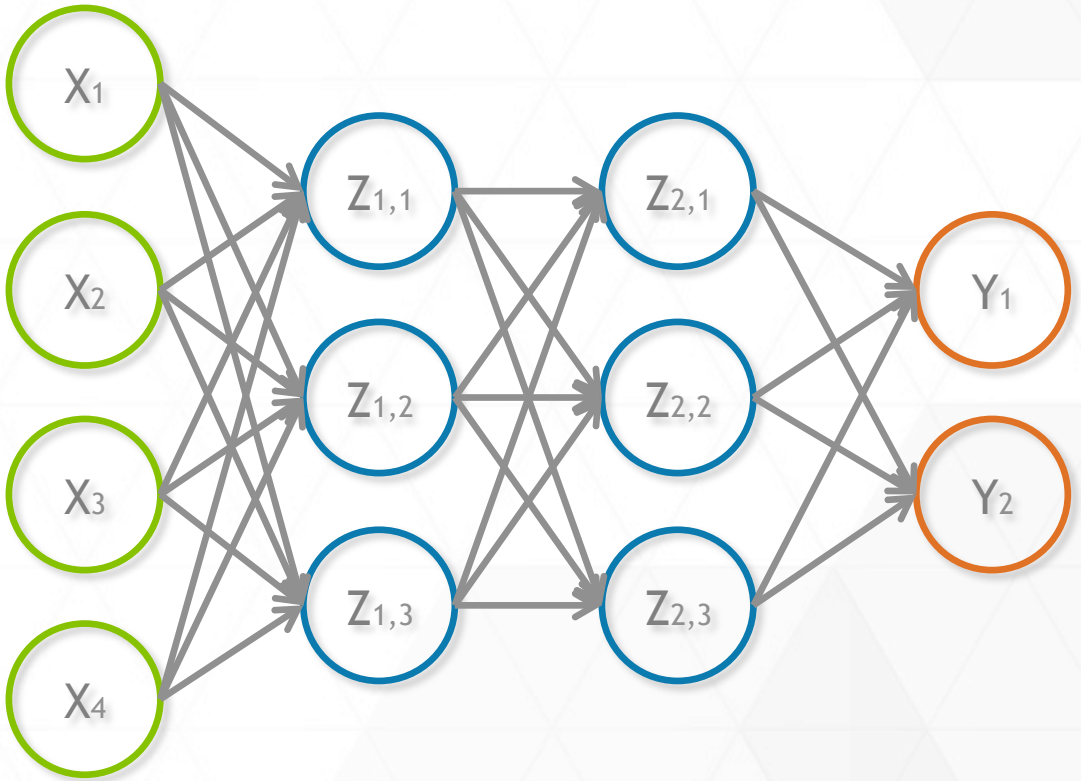
- Training
  - Train the model from supervised data
- Classification (inference)
  - Run the new sample through the model to predict its class/ function value



ARTIFICIAL NEURAL NETWORKS

Deep networks

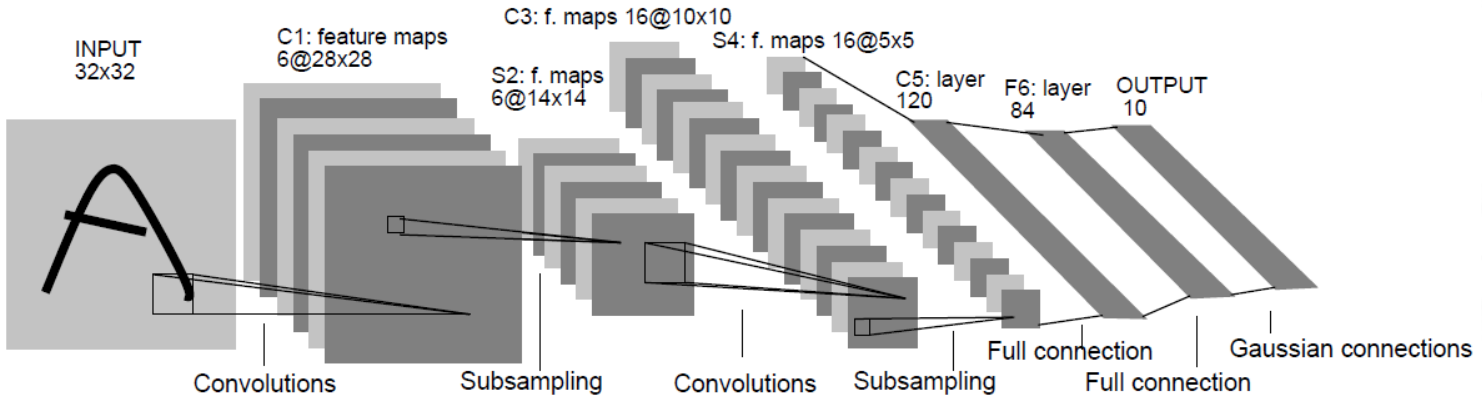
- Deep nets: with multiple hidden layers
- Trained usually with backpropagation

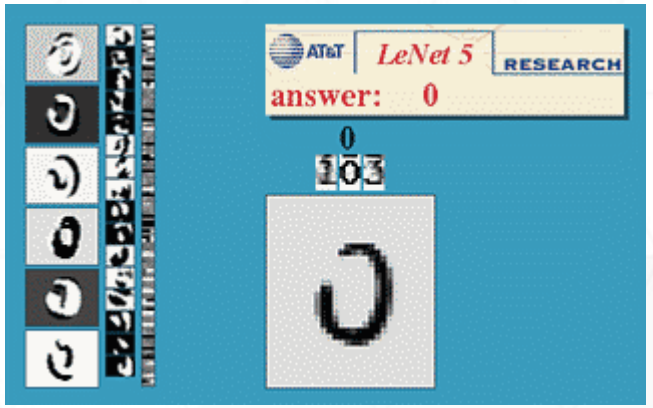


CONVOLUTIONAL NETWORKS

Local receptive field + weight sharing

- Yann LeCun et al, 1998





- MNIST: 0.7% error rate

"Gradient-Based Learning Applied to Document Recognition", Proceedings of the IEEE 1998, http://yann.lecun.com/exdb/lenet/index.html

High need for computational resources
Low ConvNet adoption rate until ~2010

TRAFFIC SIGN RECOGNITION

GTSRB

- The German Traffic Sign Recognition Benchmark, 2011

Rank	Team	Error rate	Model
1	IDSIA, Dan Ciresan	0.56%	CNNs, trained using **GPUs**
2	Human	1.16%	
3	NYU, Pierre Sermanet	1.69%	CNNs
4	CAOR, Fatin Zaklouta	3.86%	Random Forests

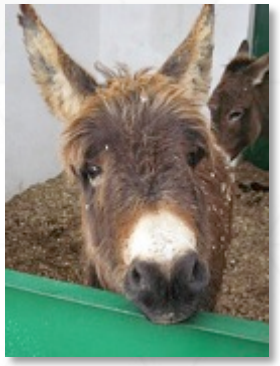
http://benchmark.ini.rub.de/?section=gtsrb

NATURAL IMAGE CLASSIFICATION

ImageNet

- 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**

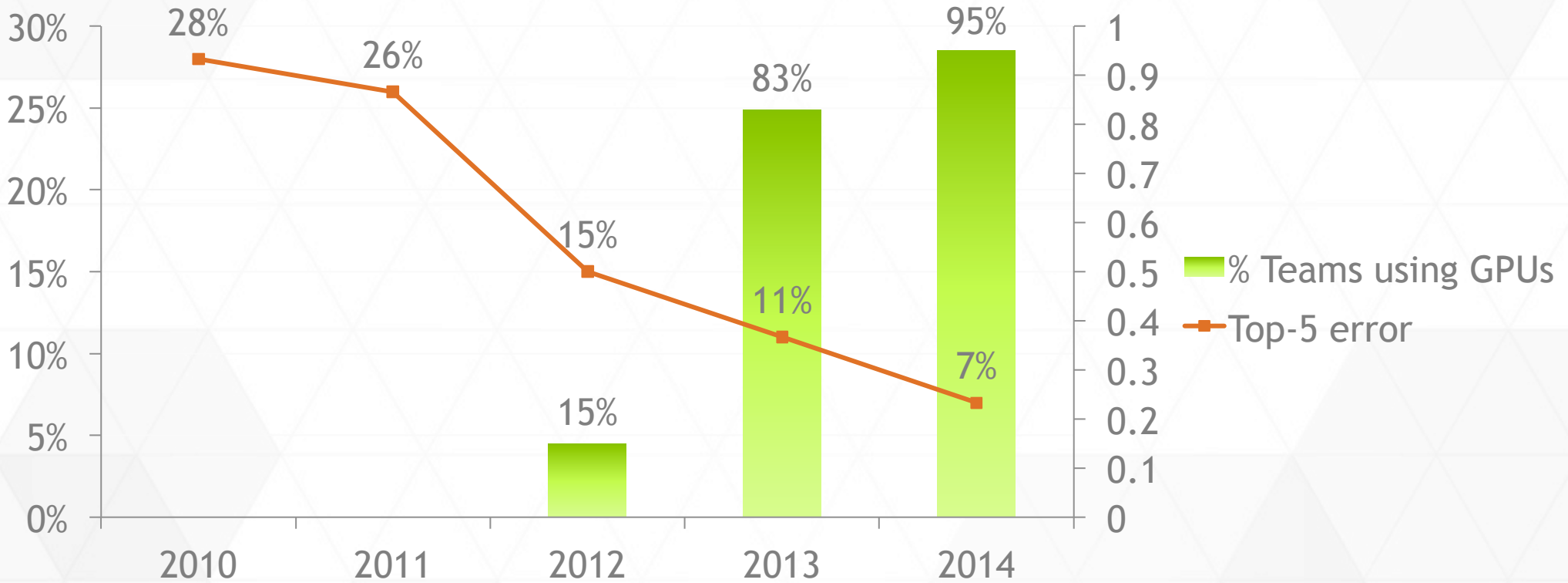





http://www.image-net.org/challenges/LSVRC/

NATURAL IMAGE CLASSIFICATION

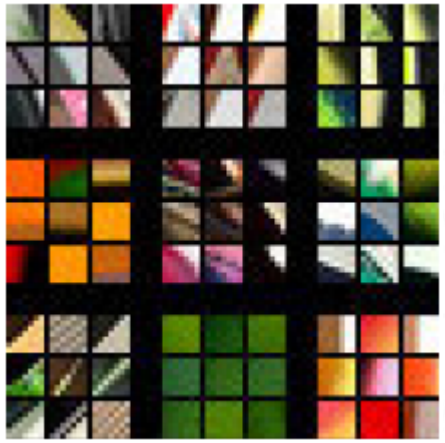
ImageNet: results for 2010-2014



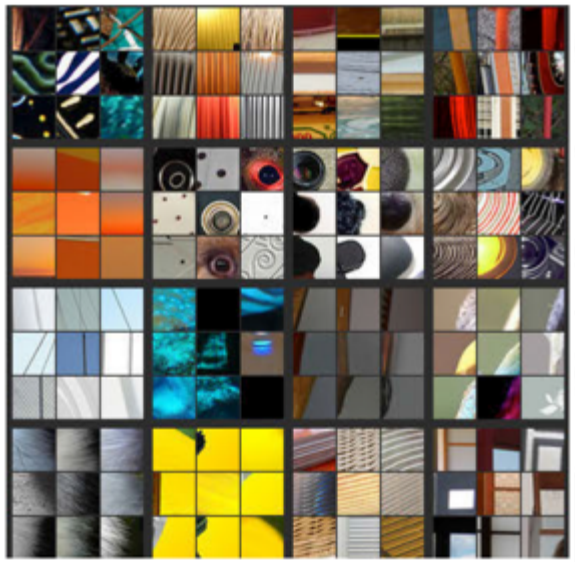
MODEL VISUALIZATION

- Matthew D. Zeiler, Rob Fergus

Layer 1



Layer 2



Layer 5

- Critique by Christian Szegedy et al

Visualizing and Understanding Convolutional Networks, http://arxiv.org/abs/1311.2901
Intriguing properties of neural networks, http://arxiv.org/abs/1312.6199

TRANSFER LEARNING

Dogs vs. Cats

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

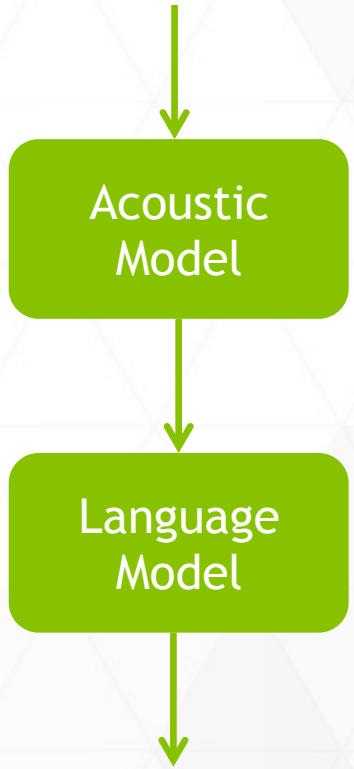
Rank	Team	Error rate	Model
1	Pierre Sermanet	1.1%	CNNs, model transferred from ImageNet
	…		
5	Maxim Milakov	1.9%	CNN, model trained on Dogs vs. Cat dataset only

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

SPEECH RECOGNITION

Acoustic model

- 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



http://devblogs.nvidia.com/parallelforall/cuda-spotlight-gpu-accelerated-speech-recognition/

It is all about supercomputing, right?

GPU

Tesla K40 and Tegra K1

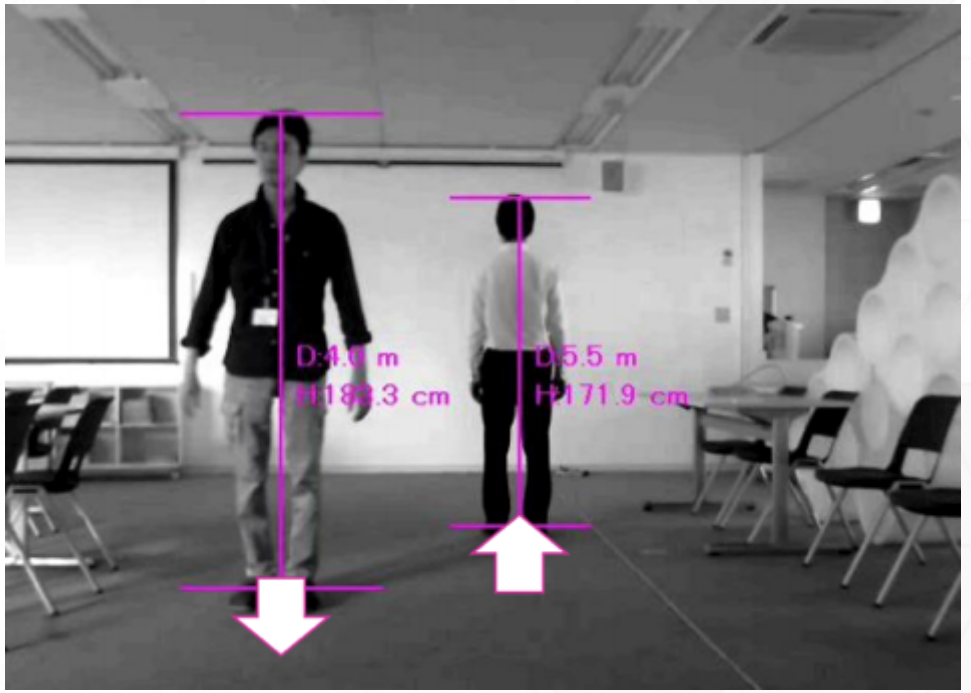
	NVIDIA Tesla K40	NVIDIA Jetson TK1
CUDA cores	2880	192
Peak performance, SP	4.29 Tflops	326 Gflops
Peak power consumption	235 Wt	~10 Wt, for the whole board
Deep Learning tasks	Training, Inference	Inference, Online Training
CUDA	Yes	Yes

http://www.nvidia.com/tesla http://www.nvidia.com/jetson-tk1 http://www.nvidia.com/object/jetson-automotive-development-platform.html

PEDESTRIAN + GAZE DETECTION

Jetson TK1

- Ikuro Sato, Hideki Niihara, R&D Group, Denso IT Laboratory, Inc.
- Real-time pedestrian detection with depth, height, and body orientation estimations
- http://www.youtube.com/watch?v=9Y7yzi_w8qo



http://on-demand.gputechconf.com/gtc/2014/presentations/S4621-deep-neural-networks-automotive-safety.pdf

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
- cuBLAS (SGEMM for fully-connected layers) is part of CUDA toolkit, developer.nvidia.com/cuda-toolkit

CUDNN

Frameworks

- cuDNN is already integrated in major open-source frameworks

- Caffe - caffe.berkeleyvision.org
- Torch - torch.ch
- Theano - deeplearning.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
- Jetson Pro: www.nvidia.com/object/jetson-automotive-development-platform.html
- CUDA Zone: developer.nvidia.com/cuda-zone
- Parallel Forall blog: devblogs.nvidia.com/parallelforall
- Contact me: mmilakov@nvidia.com

TECHNOLOGY INNOVATION ROUNDTABLE ISRAEL

NOV 5, 2014

Government of Israel
Economic Mission to the West Coast
Foreign Trade Administration Israel

HP
(Platinum Sponsor)

Lenovo
(Gold Sponsor)

DELL
(Gold sponsor)