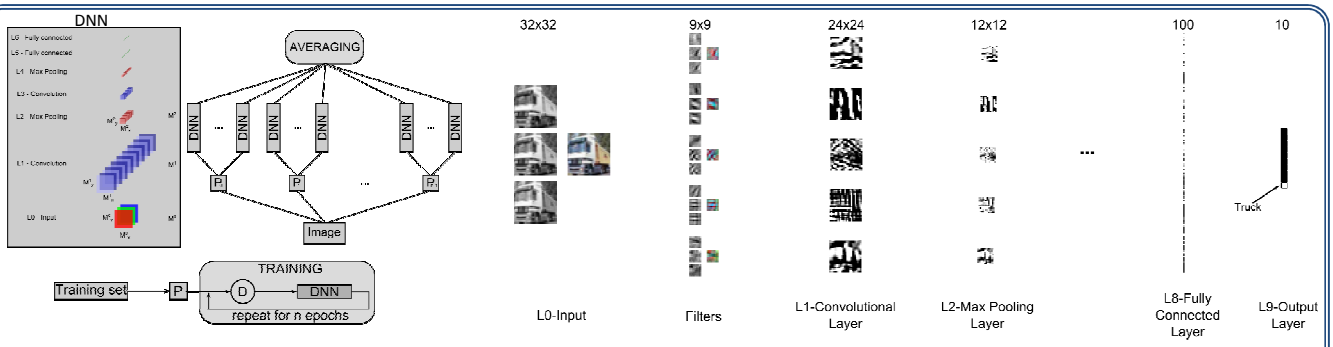
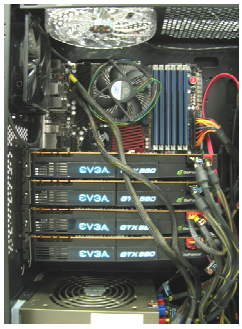
CVPR 2012, June 20th

Multi-column Deep Neural Networks for Image Classification

Dan C. Cireşan, Ueli Meier, Jürgen Schmidhuber
IDSIA, USI and SUPSI, Manno-Lugano, Switzerland
{dan,ueli,juergen}@idsia.ch



- 4 x GTX 580 1.5GB RAM
- >6 TFLOPS (maximum theoretical speed)
- 50-100x speed-up compared with a single threaded CPU version of the CNN program (one day on GPU instead of two months on CPU)

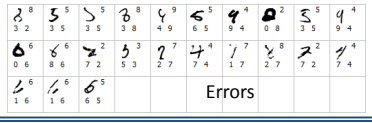


Deep Neural Networks (DNN)

- 6-10 layers (deep) of non-linear neurons, and have hundreds of maps per layer (large), loosely inspired from biology.
- Introduced by Fukushima (80) and refined by LeCun et al.(98), Riesenhuber et al.(99), Simard et al.(03), Behnke (03), Ciresan et al.(11).
- Fully supervised, with randomly initialized filters, trained minimizing the misclassification error.
- Hard/impossible to train with gradient descent two decades ago
- Current CPU are more than 60000 times faster
- Given sufficient labeled training data, neither supervision nor carefully prewired synapses are needed
- Architecture:
  - convolutional layer (extracts features), max-pooling layer (selects features)
  - one output neuron per class, normalized with soft-max activation function

MNIST

- Simard et al. (2003) – 0.40%, Ciresan et al. (2010) – 0.35%
- Distortions: translation, rotation, scaling, elastic
- 800 epochs of training for each DNN
- DNN architecture: 29x29-20C4-MP2-40C5-MP3-150N-10N
- 35column MCDNN: 6+1 width normalizations x 5 instances
- **0.23%** error rate (20 out of 23 digits have a correct second prediction)



NIST SD 19

- Collection of ~0.8M handwritten digits and Latin characters
- Same setup as for MNIST

Data (# classes)	MCDNN error [%]	Previous best result [%]
All (62)	**11.63**	-
Digits (10)	**0.77**	1.88
Letters (52)	**21.01**	30.91
Letters (26) - merged	**7.37**	13.00
Merged (37)	**7.99**	-
Uppercase (26)	**1.83**	6.44
Lowercase (26)	**7.47**	13.27

Chinese characters – CASIA GB1

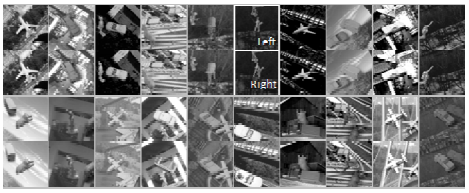
- 300 writers, 3755 classes, 1.2 Million characters, 3GB
- One week of training on GPU (corresponding to ~14 months on CPU)
- DNN: 48x48-100C3-MP2-200C2-MP2-300C2-MP2-400C2-MP2-500N-3755N
- First method which works directly on images
- Offline task: **6.5%** error vs. 10.01% (Liu et al. 2010)
- Online task: **5.61%** vs. 7.61% (Liu et al. 2010)
- **1th** place at ICDAR 2011 competition

Jittered-cluttered NORB

- 3D object recognition from stereo images
- Training set: 291600 48x48 stereo images
- 5 classes with 10 instances: 5 instances for training and 5 for testing
- Challenging dataset, only 5 instances/class, some instances from test set are completely different than the one from training set
- DNN: 2x48x48-100C5-MP2-100C5-MP2-100C4-MP2-300N-100N-6N
- Distortions: translation (max 15%), rotation (max 15°), scaling (max 15%)

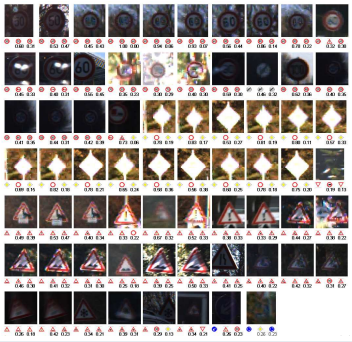
Training set size	Errors for 4 runs [%]				Mean [%]
First 2 folds	4.49	4.71	4.82	4.85	4.72±0.16
	4-net MCDNN error: 3.57%				
All 10 folds	3.32	3.18	3.73	3.36	3.40±0.23
	4-net MCDNN error: **2.70%**				

Previous lowest error: 5.00% (Coates et al. 2011)



Traffic signs - GTSRB

- 26640 training and 12569 testing images, 43 classes
- 25 column MCDNN: 4+1 normalization x 5 instances
- DNN: 3x48x48-100C7-MP2-150C4-150MP2-250C4-250MP2-300N-43N
- **1th** place at Traffic Sign competition at IJCNN 2011
- Error rate: **0.54%**; outperforms the second best algorithm by a factor of three

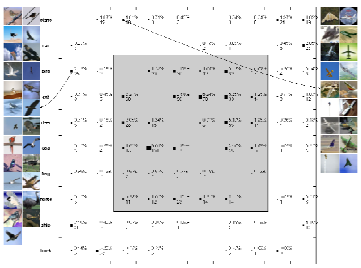


CIFAR10

- Small, 32x32 pixels color images, complex backgrounds
- DNN: 3x32x32-300C3-MP2-300C2-MP2-300C3-MP2-300C2-MP2-300N-100N-10N
- Affine distortions

preprocessing	Errors for 8 runs [%]				Mean [%]
Yes	16.47	19.20	19.72	20.31	18.93±1.69
No	15.63	15.85	16.13	16.05	15.91±0.22
	8-net average error: 17.42±1.96%				
	8-net MCDNN error: **11.21%**				

Previous lowest error: 18.50% (Coates et al. 2011)



Conclusions

- **Human-competitive results are reported on widely used computer vision benchmarks.**
- **MCDNN improves the state-of-the-art by 30-80%.**
- **Fully supervised, does not use any additional unlabeled data source.**
- **No need to extract handcrafted features, works on raw pixel images**
- **Robust (smallest error rates) and fast (10^3-10^5 images/s) for immediate industrial applications.**

Dataset	Best result of others [%]	MCDNN error [%]	Relative improvement [%]
MNIST	0.39	0.23	41
NIST SD 19			30-80
HWDB1.0 on.	7.61	5.61	26
HWDB1.0 off.	10.01	6.50	35
CIFAR 10	18.50	11.21	39
Traffic signs	1.69	0.54	72
NORB j.-c.	5.00	2.70	46

What is next?

- **Segmentation: neuronal structures in EM stacks (1th place at ISBI 2012 EM segmentation competition)**
- **More at www.idsia.ch/~ciresan**