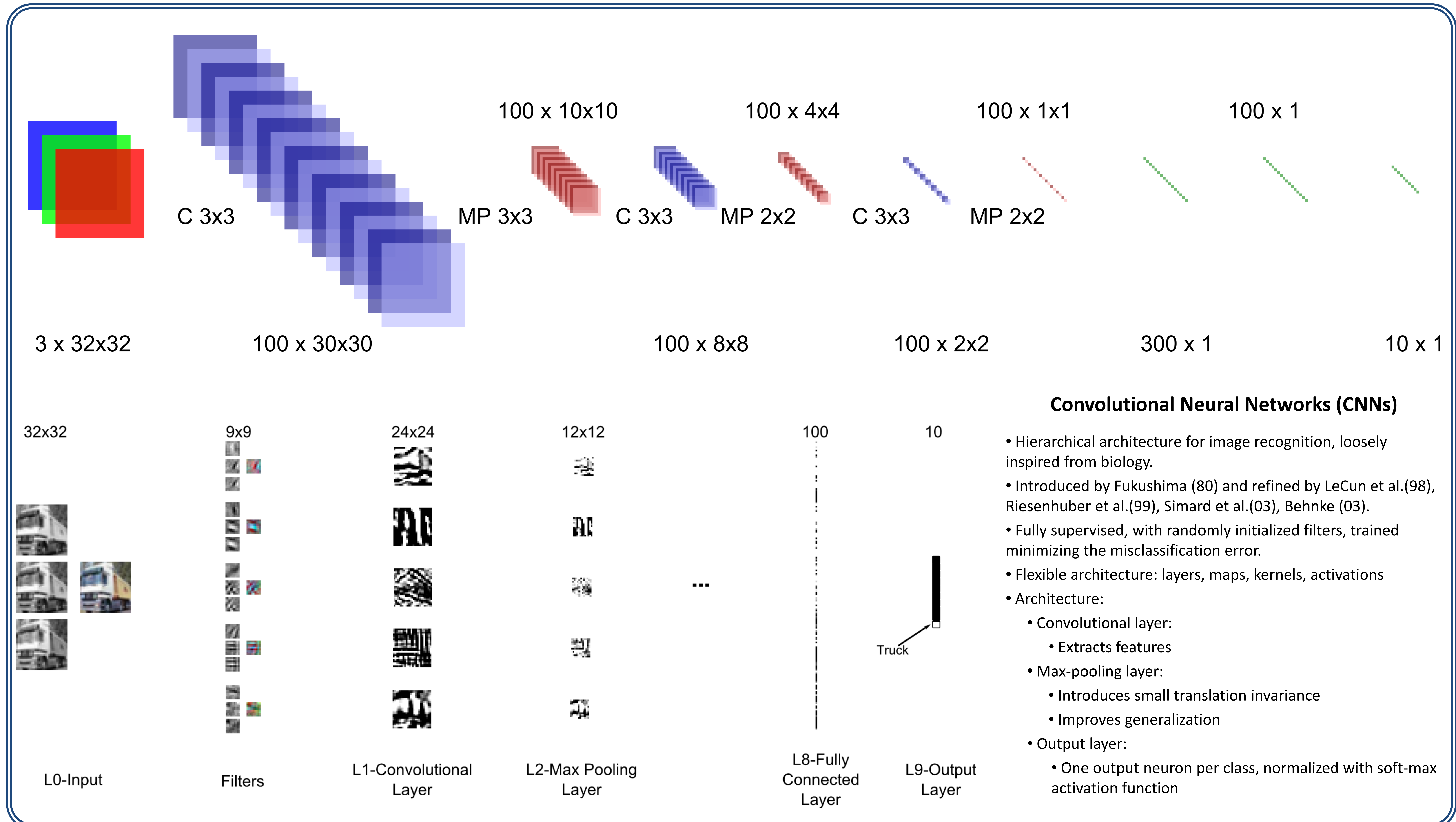


Flexible, High Performance Convolutional Neural Networks for Image Classification

Dan C. Cireşan, Ueli Meier, Jonathan Masci, Luca M. Gambardella, Jürgen Schmidhuber
 IDSIA, USI and SUPSI, Manno-Lugano, Switzerland
 {dan,ueli,jonathan,luca,juergen}@idsia.ch



Graphics processing units (GPUs)

- 8 x GTX 480/580 1.5GB RAM
- >12 TFLOPS (theoretical speed)
- 40-80x speed-up compared with a single threaded CPU version of the CNN program (one day on GPU instead of two months on CPU)

Back-propagation of deltas

Uses pooling of deltas.

$$\max\left(\left\lfloor \frac{i-K_x+1}{S_x+1} \right\rfloor, 0\right) \leq x \leq \min\left(\left\lfloor \frac{i}{S_x+1} \right\rfloor, M_x-1\right)$$

$$\max\left(\left\lfloor \frac{i-K_y+1}{S_y+1} \right\rfloor, 0\right) \leq y \leq \min\left(\left\lfloor \frac{i}{S_y+1} \right\rfloor, M_y-1\right)$$

MNIST

- 28x28 grayscale images
- 60000 for training and 10000 for testing
- Simard et al. (2003) – 0.40%, Cireşan et al. (2010) – 0.35%
- 30 out of 35 digits have a correct second prediction
- deformations: translation, rotation, scaling, elastic

#M, #N in Hidden Layers	test error [%]
20M-60M	1.02
20M-60M-150N	0.55
20M-60M-100M-150N	0.38
20M-40M-60M-80M-100M-120M-150N	0.35

Small NORB

- 48600 96x96 stereo images
- 5 classes with 10 instances: 5 instances for training and 5 for testing
- bad/challenging dataset, only 5 instances/class, some instances from test set are completely different than the one from training set
- IP maps (Mexican hat) are needed only for this data set
- deformations: only translation (random for both axes, maximum 5%)

	trans. [%]	IP	Tfbv [%]	runs	time[s]
0	no	7.86 ± 0.55	50	1143	
5	no	4.71 ± 0.57	50	1563	
0	yes	3.94 ± 0.48	50	1658	
5	yes	2.53 ± 0.40	100	2080	

CIFAR10

- small, 32x32 pixels color images
- complex backgrounds
- 10 classes
- 50000 training images and 10000 test images
- deformations: only translation (random for both axes, maximum 5%)
- border effects

trans. [%]	IP	Test error [%]	Runs	time/epoch [s]
0; 100M	no	28.87 ± 0.37	11	93
0; 100M	edge	29.11 ± 0.36	15	104
5; 100M	no	20.26 ± 0.21	11	111
5; 100M	edge	21.87 ± 0.57	5	120
5; 100M	hat	21.44 ± 0.44	4	136
5; 200M	no	19.90 ± 0.16	5	248
5; 300M	no	19.51 ± 0.18	5	532
5; 400M	no	19.54 ± 0.16	5	875

Conclusions

- Our big deep nets combining CNN and other ideas are now state of the art for many image classification tasks.
- No need to extract handcrafted features.
- Supervised training with simple gradient descent training is best. No need for unsupervised pre-training (e.g. autoencoders) in case of sufficient training samples.
- Distorting the training set improves recognition rate on unseen data.
- CPUs are not enough anymore, use GPUs which are 2 orders of magnitude faster.
- Robust (smallest error rates) and fast enough (10³-10⁴ images/s) for immediate industrial applications.

What is next?

- Results on all benchmarks already improved 20-50% compared with this paper.
- Test the CNNs on different datasets:
 - already done:
 - Chinese characters: 3755 classes, >1M characters, 6.5% error rate, first place at ICDAR 2011 competition
 - Traffic signs: 43 classes, <1% error rate, first place at IJCNN 2011 competition
 - All Latin alphabet (NIST SD 19, >0.8M characters): state of the art results (ICDAR 2011)
 - next:
 - CALTECH 101 & 256, ImageNet, cluttered NORB, medical images
- Use CNNs for general scene understanding (segmentation).
- Robot vision applications (detect type, position and orientation of objects in a scene).
- Computer vision applications (de-blurring, segmentation, similarity check etc.).
- Try to reach and surpass human image classification performance.
- More at www.idsia.ch/~cireşan