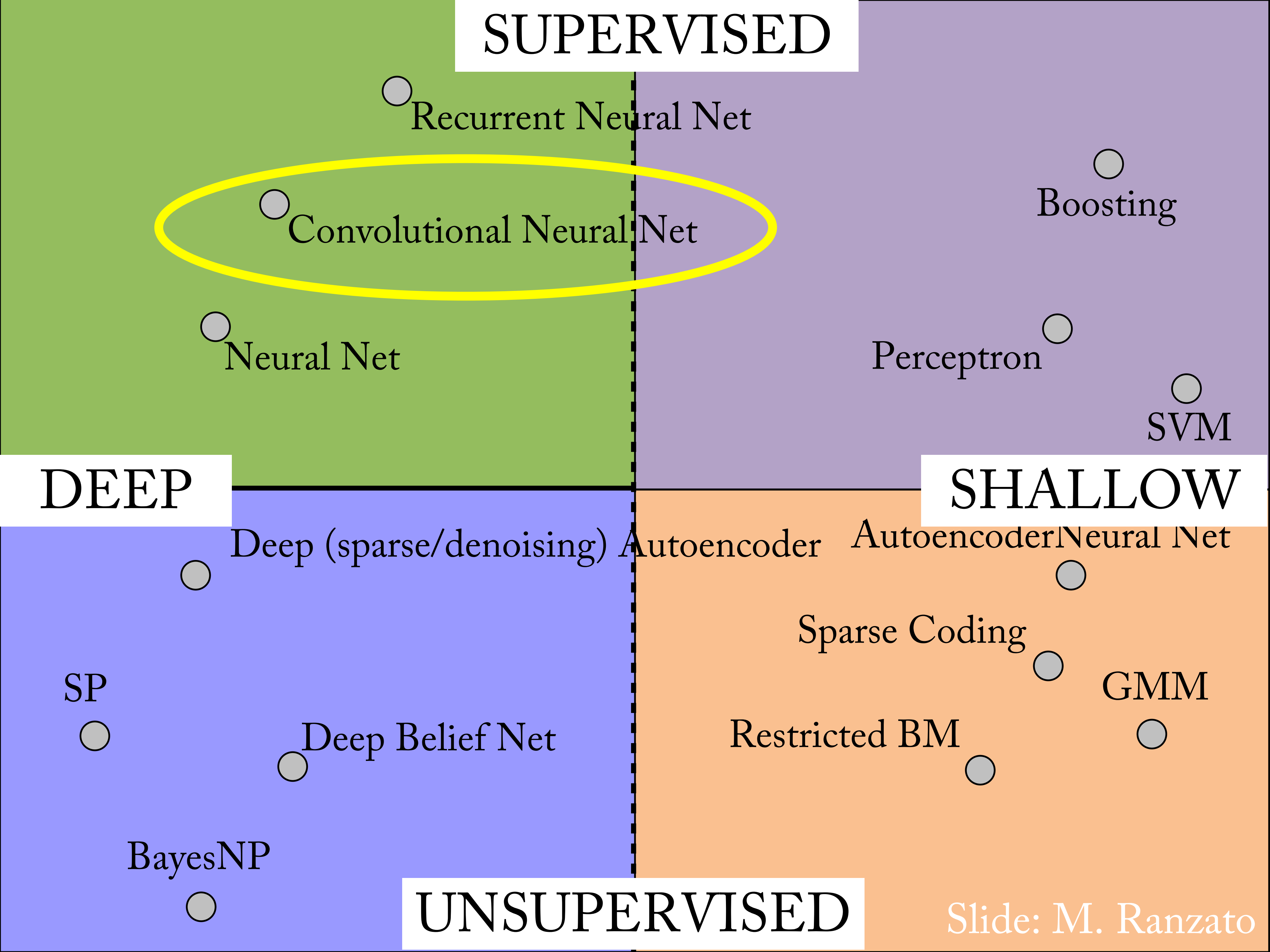
Introduction to Convolutional Networks

Lecture 7

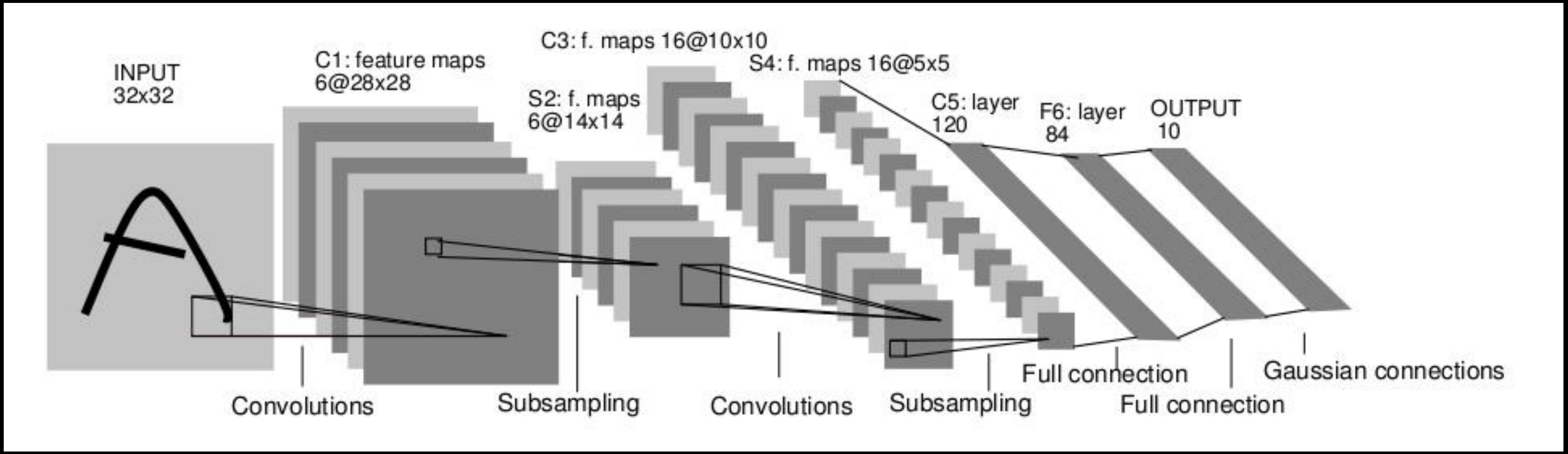
Rob Fergus

New York University


SUPERVISED
Recurrent Neural Net
Convolutional Neural Net
Neural Net
Boosting
Perceptron
SVM
DEEP
SHALLOW
Deep (sparse/denoising) Autoencoder
AutoencoderNeural Net
Sparse Coding
SP
GMM
Deep Belief Net
Restricted BM
BayesNP
UNSUPERVISED
Slide: M. Ranzato

Convolutional Neural Networks

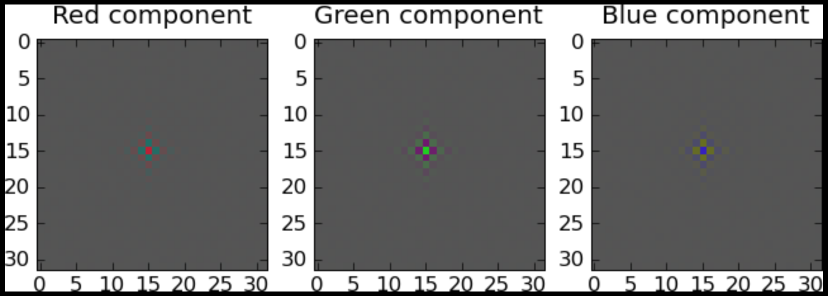
- LeCun et al. 1989
- Neural network with specialized connectivity structure



ConvNet Architecture

- Exploits two properties of images:

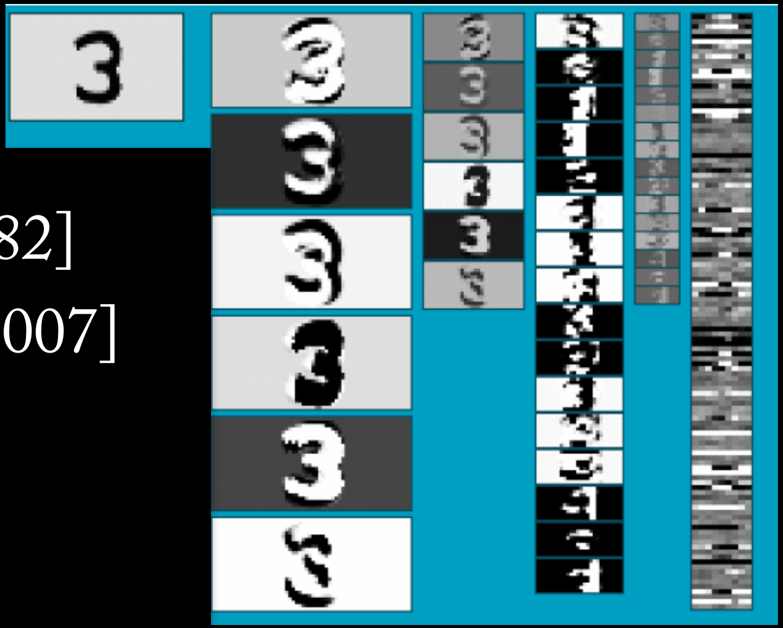
- 1. Dependencies are local
  - No need to have each unit connect to every pixel



- 2. Spatially stationary statistics
  - Translation invariant dependencies
  - Only approximately true

Multistage Hubel-Wiesel Architecture

- Stack multiple stages of simple cells / complex cells layers
- Higher stages compute more global, more invariant features
- Classification layer on top



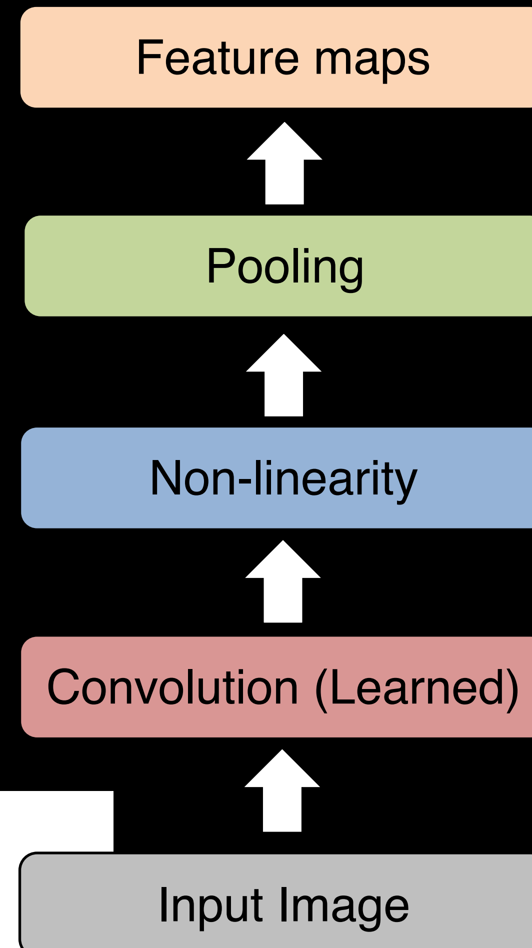
History:

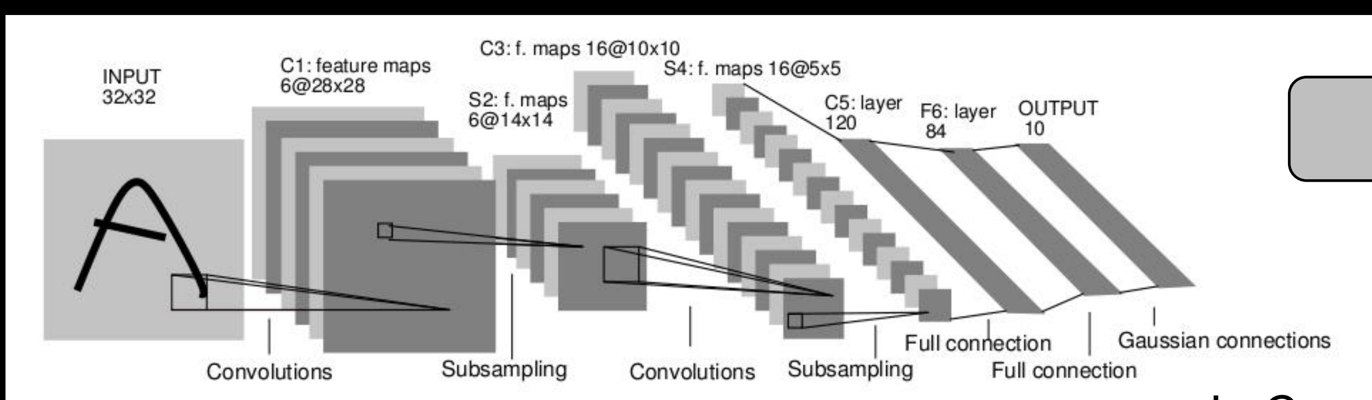
- Neocognitron [Fukushima 1971-1982]
- Convolutional Nets [LeCun 1988-2007]
- HMAX [Poggio 2002-2006]
- Many others….

Slide: Y.LeCun

Overview of Convnets

- Feed-forward:
  - Convolve input
  - Non-linearity (rectified linear)
  - Pooling (local max)
- Supervised
- Train convolutional filters by back-propagating classification error

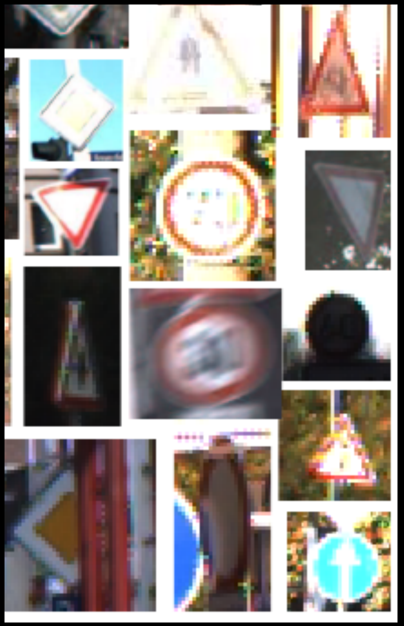




Convnet Successes

- Handwritten text/digits
  - MNIST (0.17% error [Ciresan et al. 2011])
  - Arabic & Chinese [Ciresan et al. 2012]

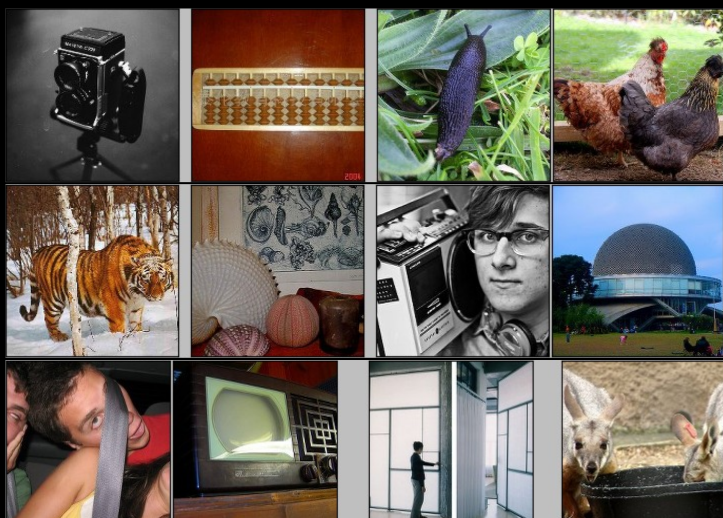
- Simpler recognition benchmarks
  - CIFAR-10 (9.3% error [Wan et al. 2013])
  - Traffic sign recognition
    - 0.56% error vs 1.16% for humans [Ciresan et al. 2011]



- But less good at more complex datasets
  - E.g. Caltech-101/256 (few training examples)

Application to ImageNet

IMAGENET



[Deng et al. CVPR 2009]

- ~14 million labeled images, 20k classes
- Images gathered from Internet
- Human labels via Amazon Turk

ImageNet Classification with Deep Convolutional Neural Networks [NIPS 2012]

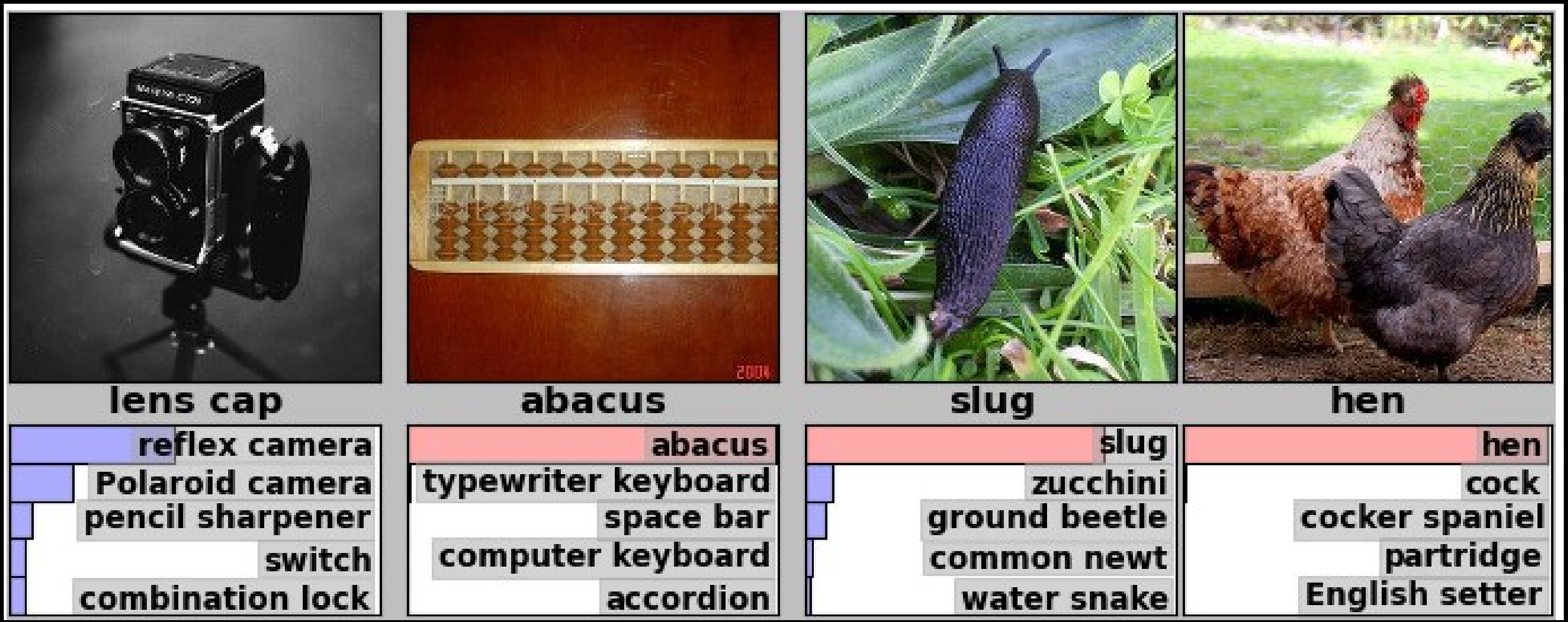
Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Goal

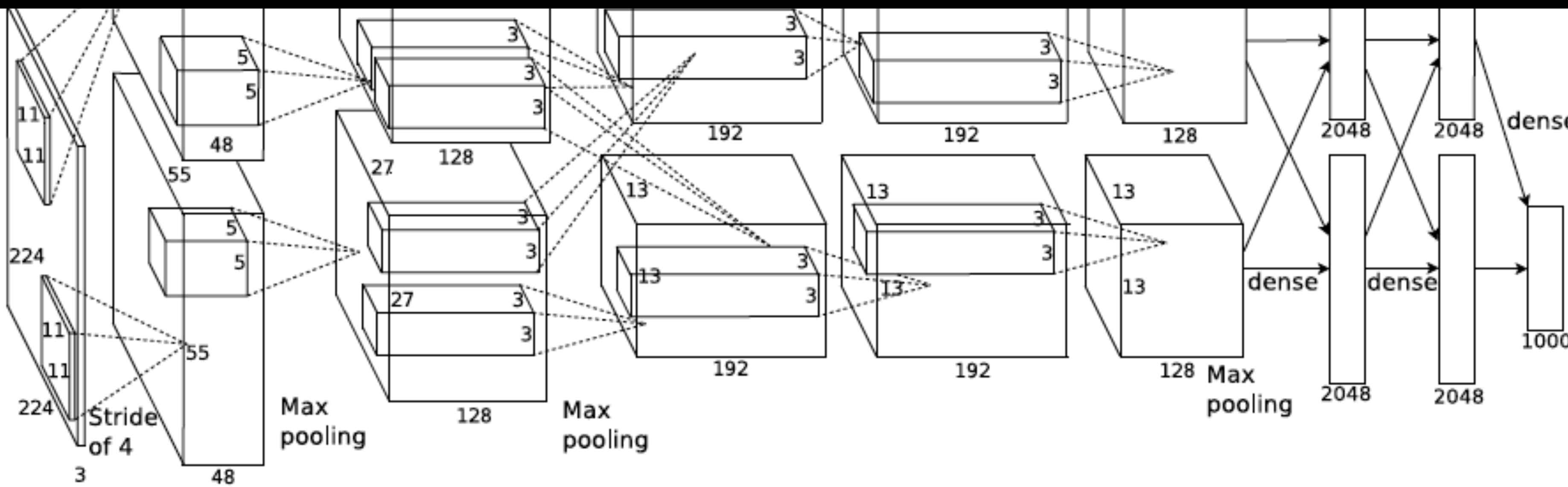
- Image Recognition
  - Pixels → Class Label



[Krizhevsky et al. NIPS 2012]

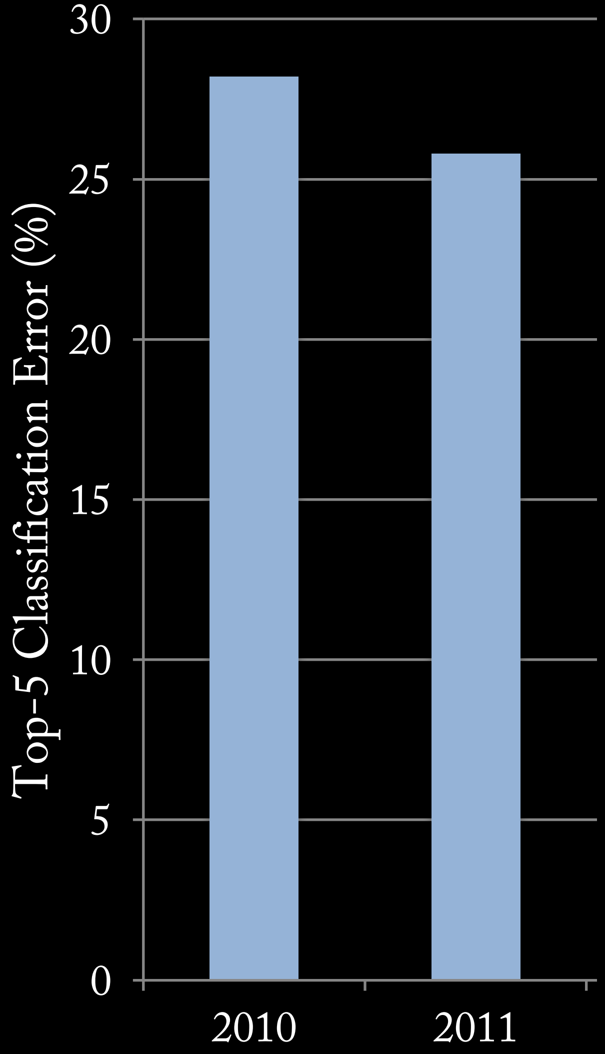
Krizhevsky et al. [NIPS2012]

- Same model as LeCun'98 but:
  - Bigger model (8 layers)
  - More data (10^6 vs 10^3 images)
  - GPU implementation (50x speedup over CPU)
  - Better regularization (DropOut)



- 7 hidden layers, 650,000 neurons, 60,000,000 parameters
- Trained on 2 GPUs for a week

ImageNet Classification (2010 – 2015)



Examples

- From Clarifai.com

Predicted Tags:

food	(16.00%)
dinner	(3.10%)
bbq	(2.90%)
market	(2.50%)
meal	(1.40%)
turkey	(1.40%)
grill	(1.30%)
pizza	(1.30%)
eat	(1.10%)
holiday	(1.00%)

Stats:

Size: 247.24 KB

Time: 110 ms

Examples

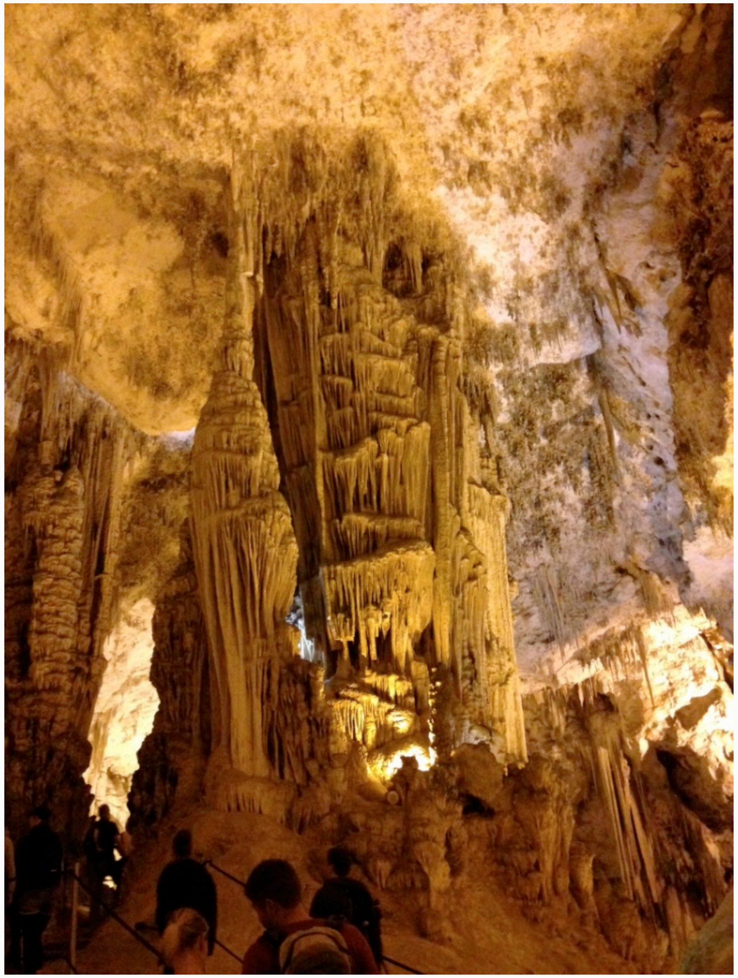
- From Clarifai.com

Predicted Tags:

ship	(2.30%)
helsinki	(1.80%)
fish	(1.40%)
port	(1.10%)
istanbul	(1.10%)
beach	(1.00%)
denmark	(1.00%)
copenhagen	(0.90%)
sea	(0.80%)
boat	(0.80%)

Examples

- From Clarifai.com



Predicted Tags:

barcelona	(6.50%)
street	(3.00%)
cave	(2.20%)
sagrada	(1.90%)
old	(1.80%)
night	(1.40%)
familia	(1.40%)
jerusalem	(1.40%)
guanajuato	(1.10%)
alley	(1.00%)

Stats:

Size: 278.96 KB

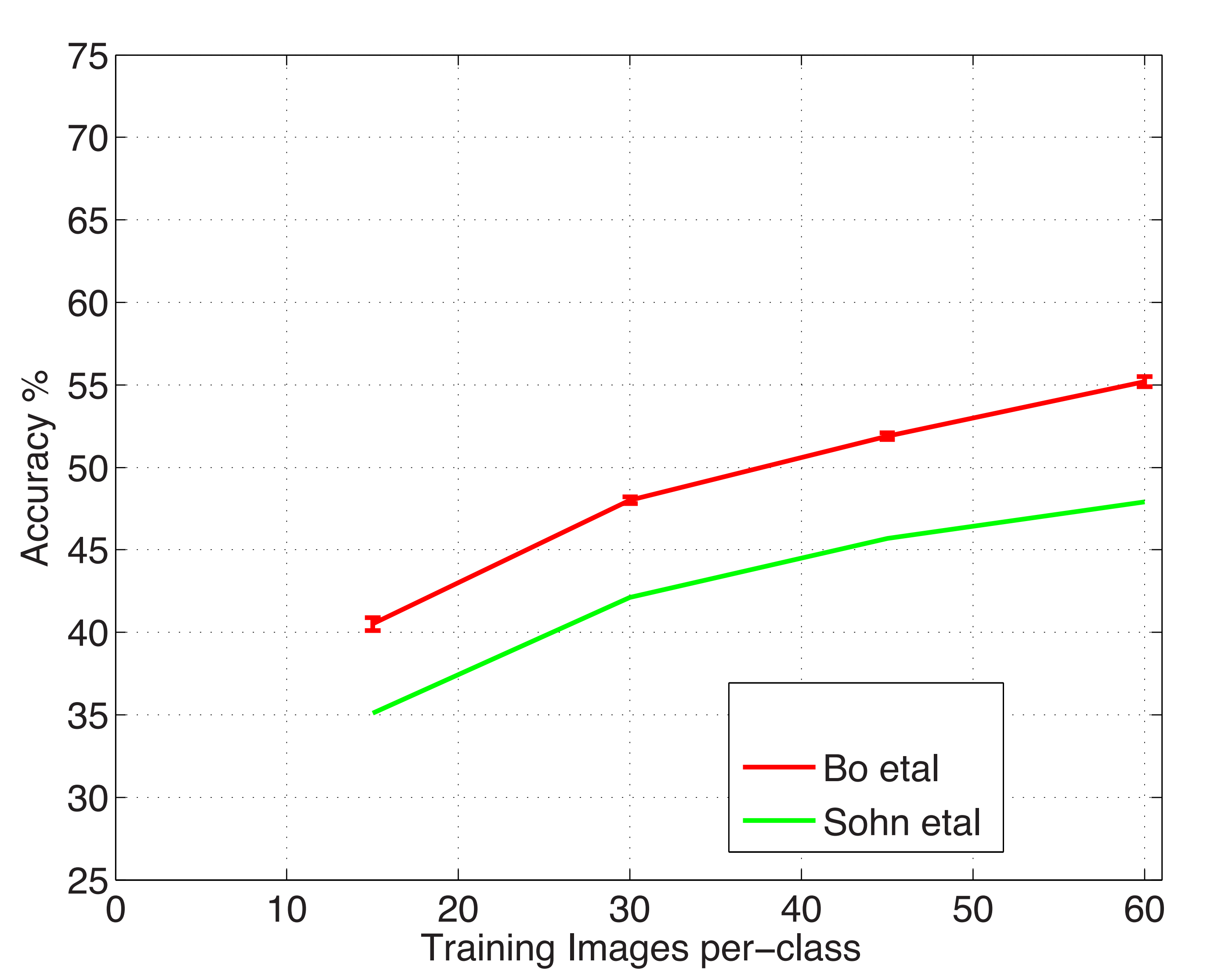
Time: 113 ms

Using Features on Other Datasets

- Train model on ImageNet 2012 training set
- Re-train classifier on new dataset
  - Just the top layer (softmax)
- Classify test set of new dataset

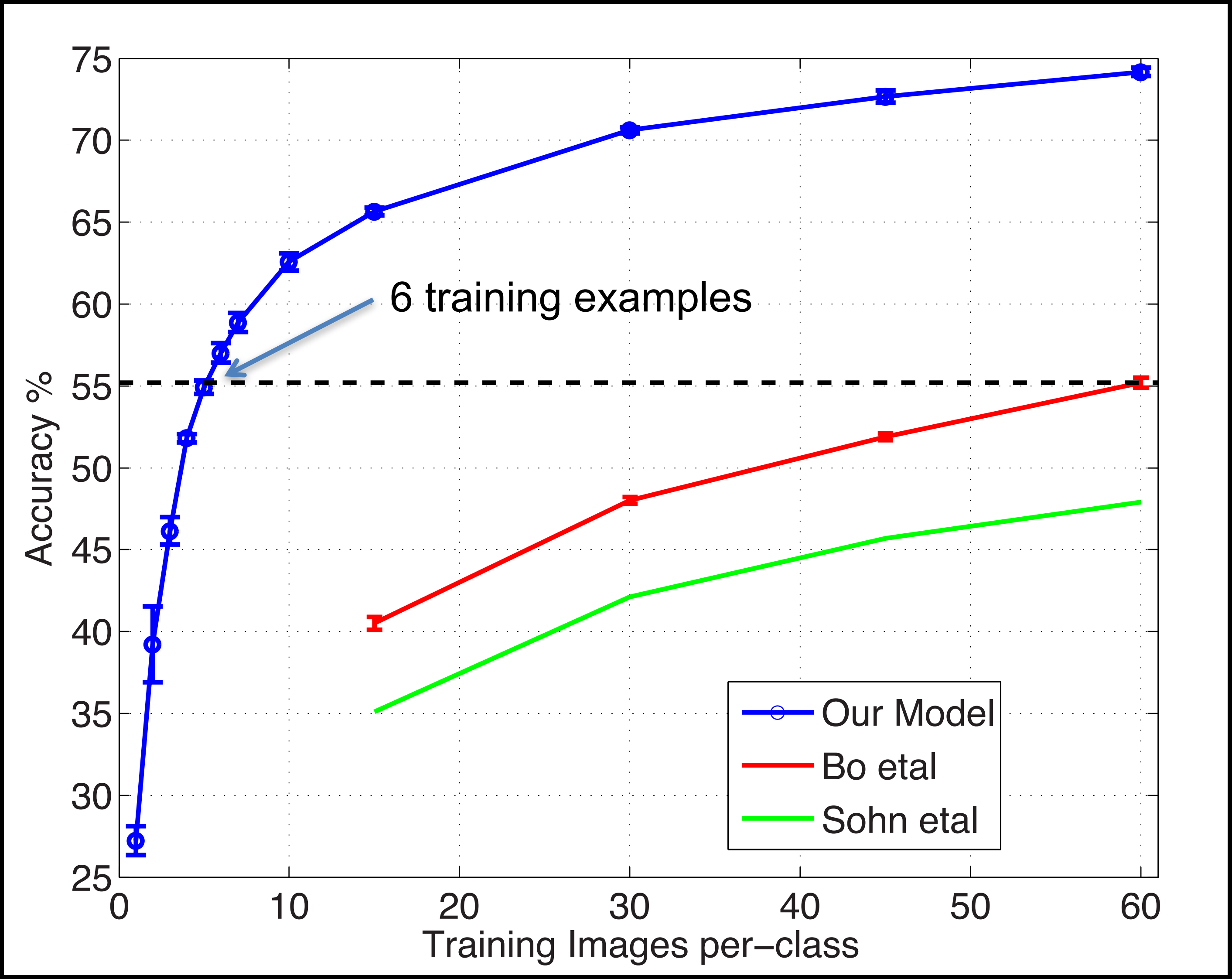
Caltech 256

Zeiler & Fergus, *Visualizing and Understanding Convolutional Networks*, arXiv 1311.2901, 2013



Caltech 256

Zeiler & Fergus, *Visualizing and Understanding Convolutional Networks*, arXiv 1311.2901, 2013



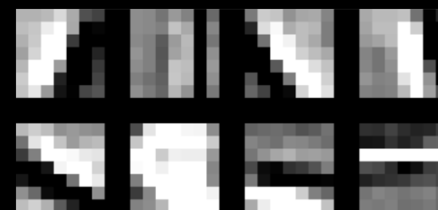
The Details

- Operations in each layer
- Architecture
- Training
- Results

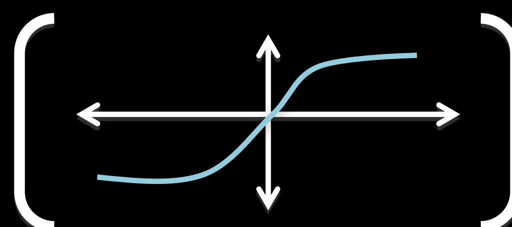
Components of Each Layer

Pixels / Features

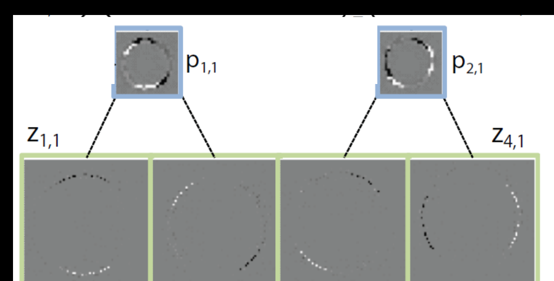
Filter with learned dictionary



Non-linearity



Spatial local max pooling

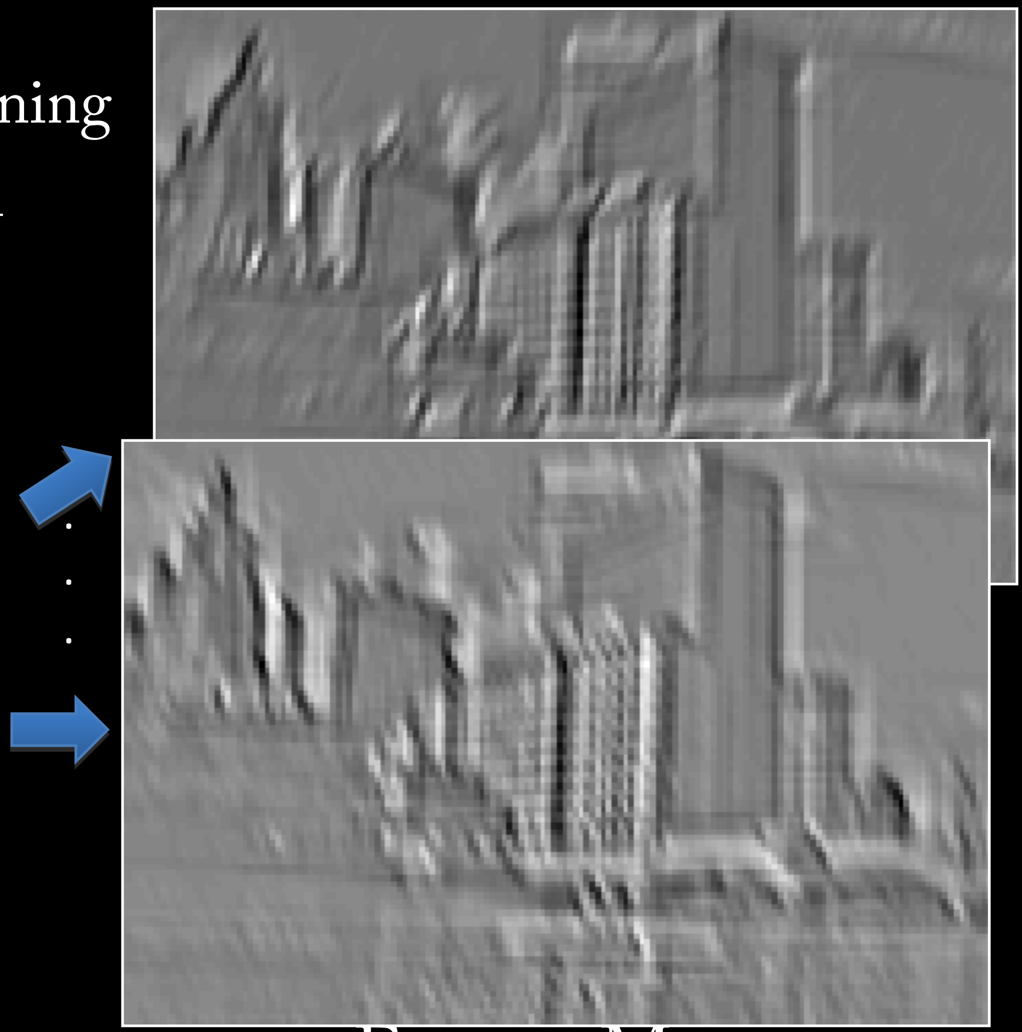


Output Features

Filtering

- Convolution
  - Filter is learned during training
  - Same filter at each location

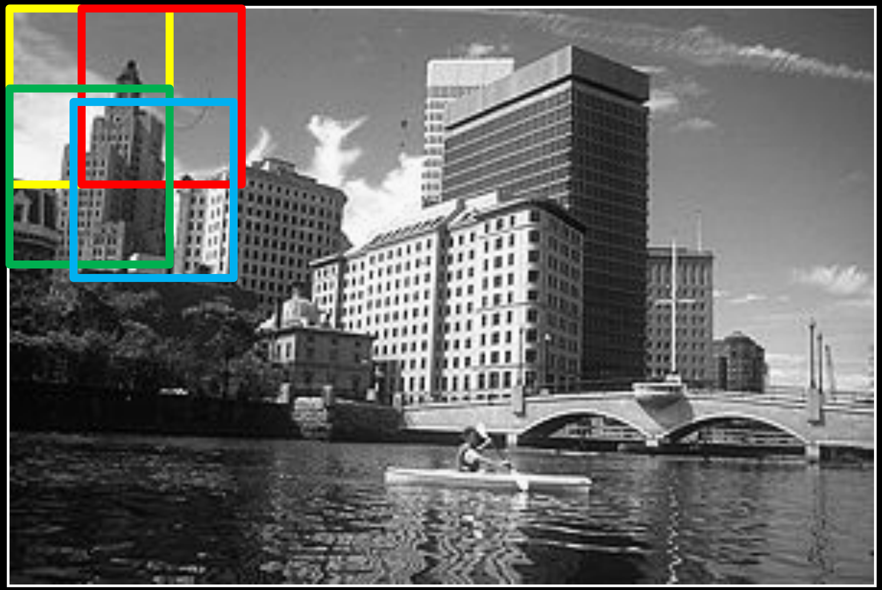
Input



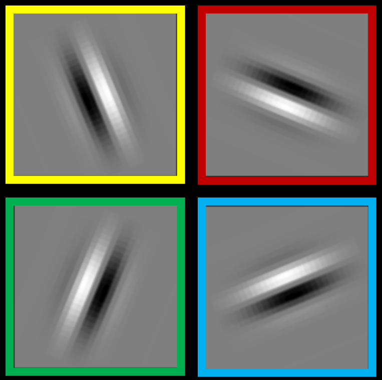
Feature Map

Filtering

- Local
  - Each unit layer above look at local window
  - But no weight tying

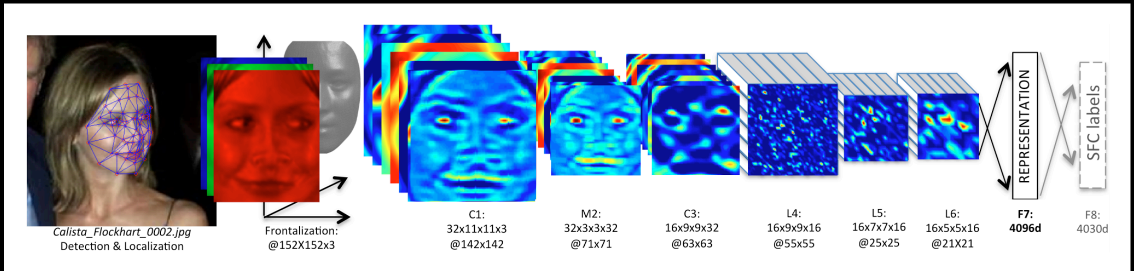


Input



Filters

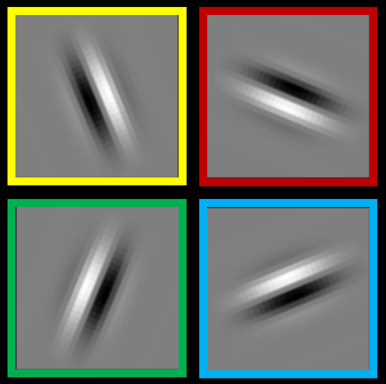
- E.g. face recognition



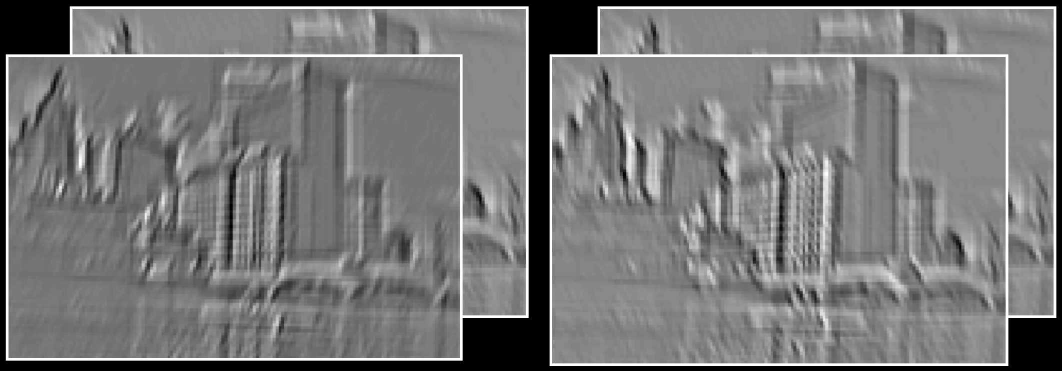
Filtering

- Tiled
  - Filters repeat every n
  - More filters than convolution for given # features

Input



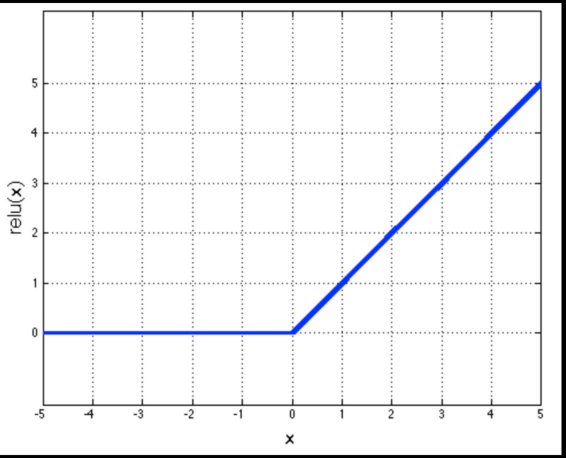
Filters



Feature maps

Non-Linearity

- Rectified linear function
  - Applied per-pixel
  - output = max(0,input)

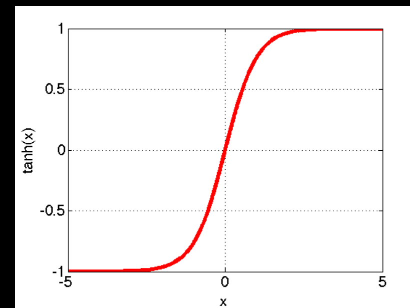


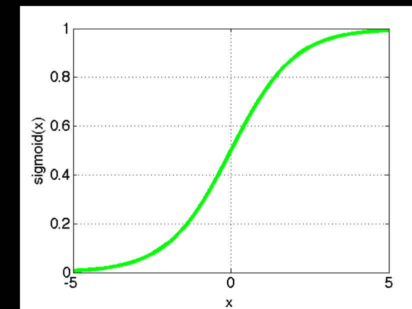
Input feature map

Output feature map

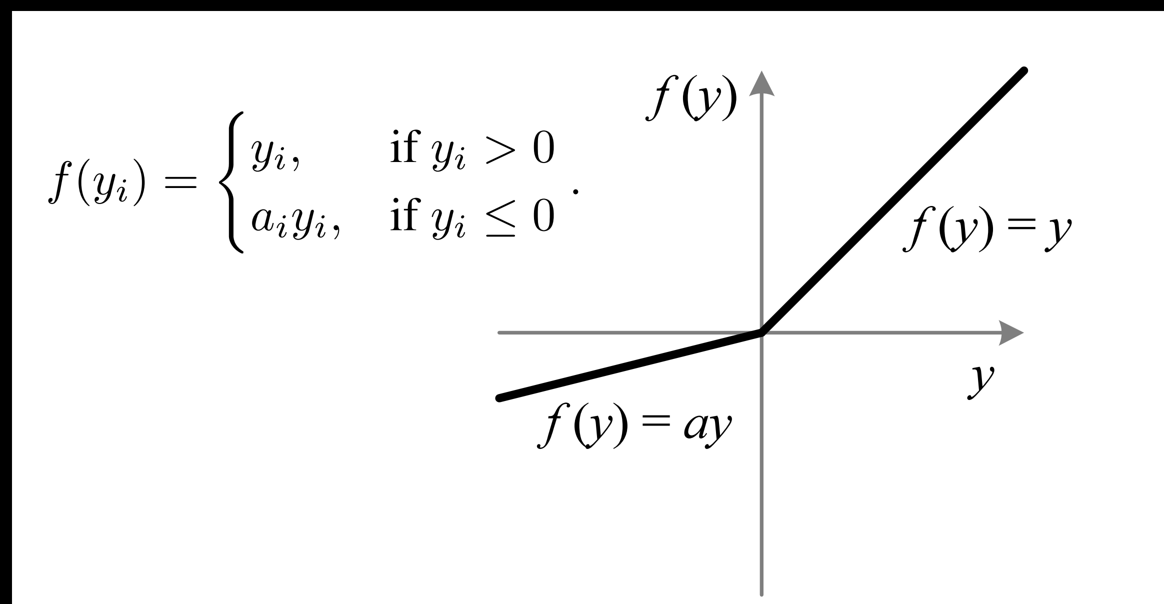
Non-Linearity

- Other choices:
  - Tanh
  - Sigmoid: 1/(1+exp(-x))
  - PReLU



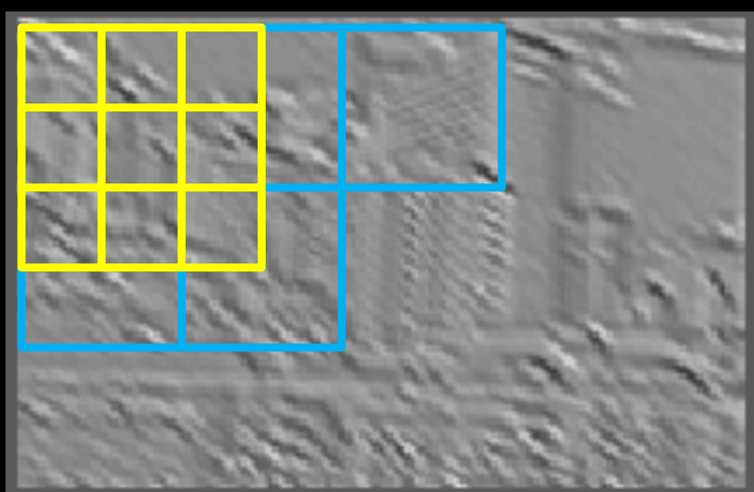


[Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, Kaiming He et al. arXiv:1502.01852v1.pdf, Feb 2015]

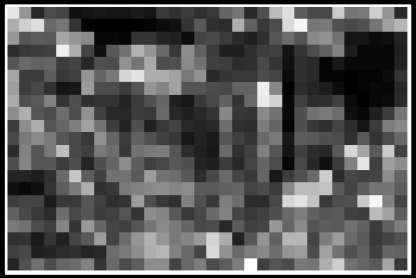


Pooling

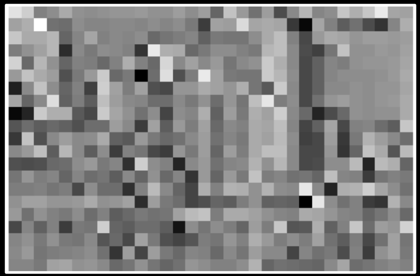
- Spatial Pooling
  - Non-overlapping / overlapping regions
  - Sum or max
  - Boureau et al. ICML'10 for theoretical analysis



Max

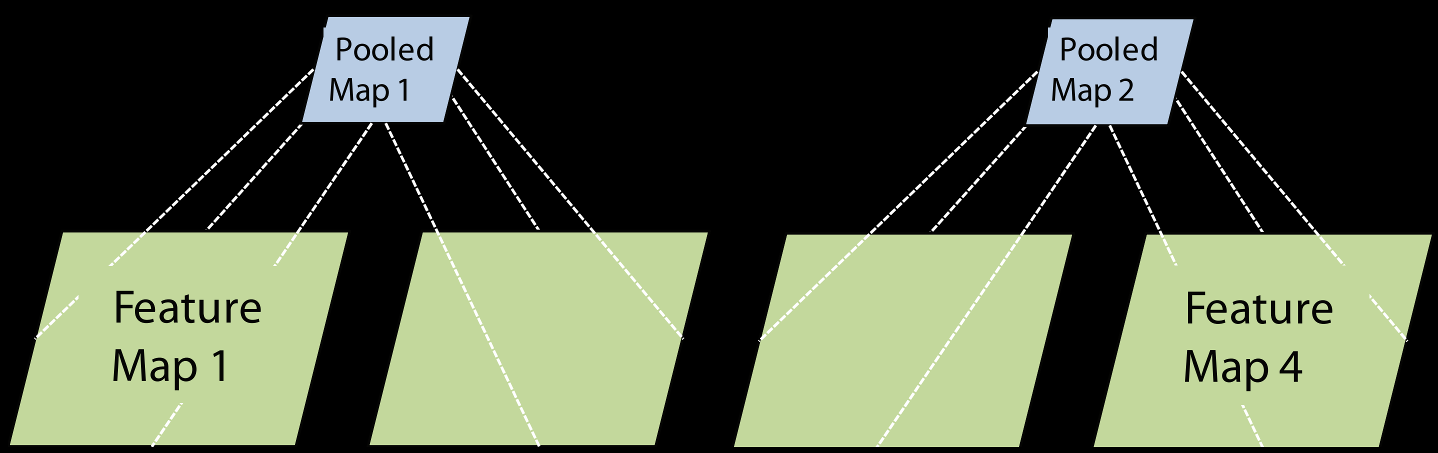


Sum



Pooling

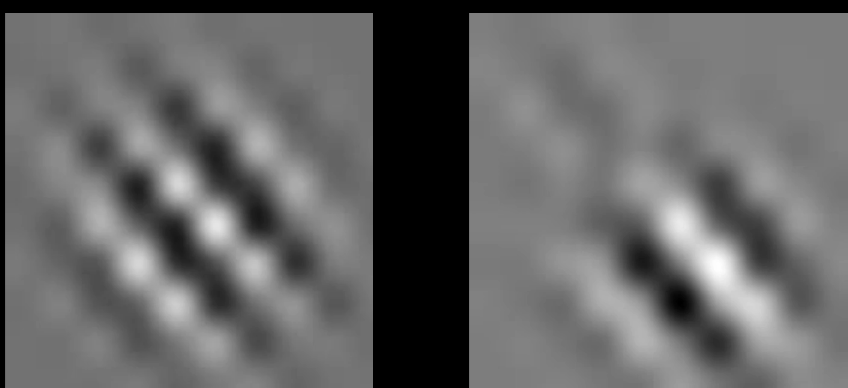
- Pooling across feature groups
  - Additional form of inter-feature competition
  - MaxOut Networks [Goodfellow et al. ICML 2013]



Role of Pooling

- Spatial pooling
  - Invariance to small transformations
  - Larger receptive fields (see more of input)

Visualization technique from [Le et al. NIPS'10]:



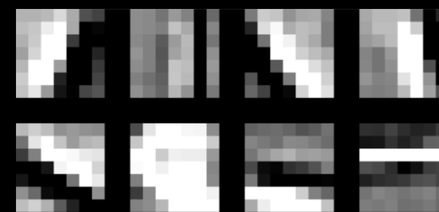
Zeiler, Fergus [arXiv 2013]

Videos from: http://ai.stanford.edu/~quocle/TCNNweb

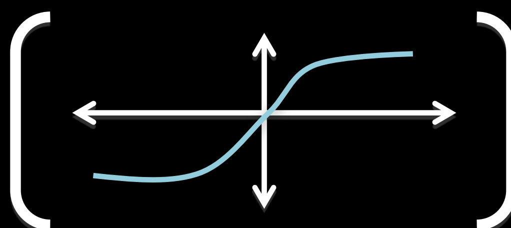
Components of Each Layer

Pixels / Features

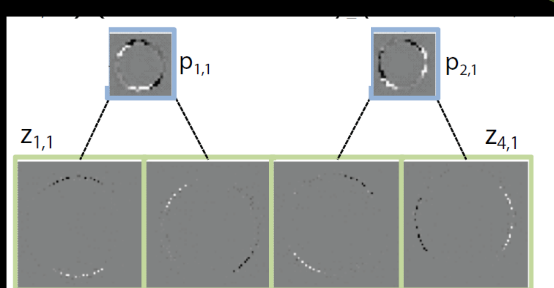
Filter with learned dictionary



Non-linearity



Spatial local max pooling



[Optional] Normalization across data/features

Output Features

Normalization

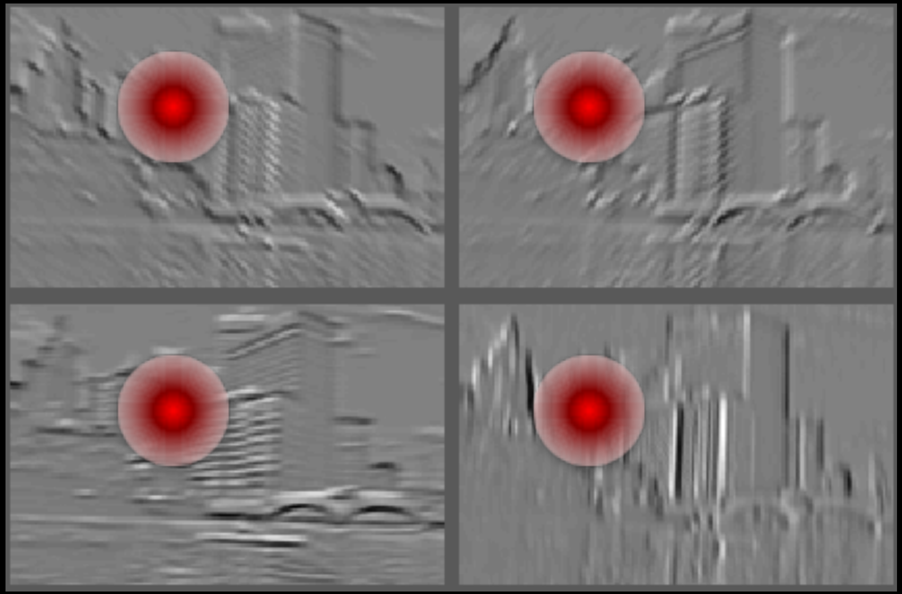
- Contrast normalization across features
  - See Divisive Normalization in Neuroscience

Input

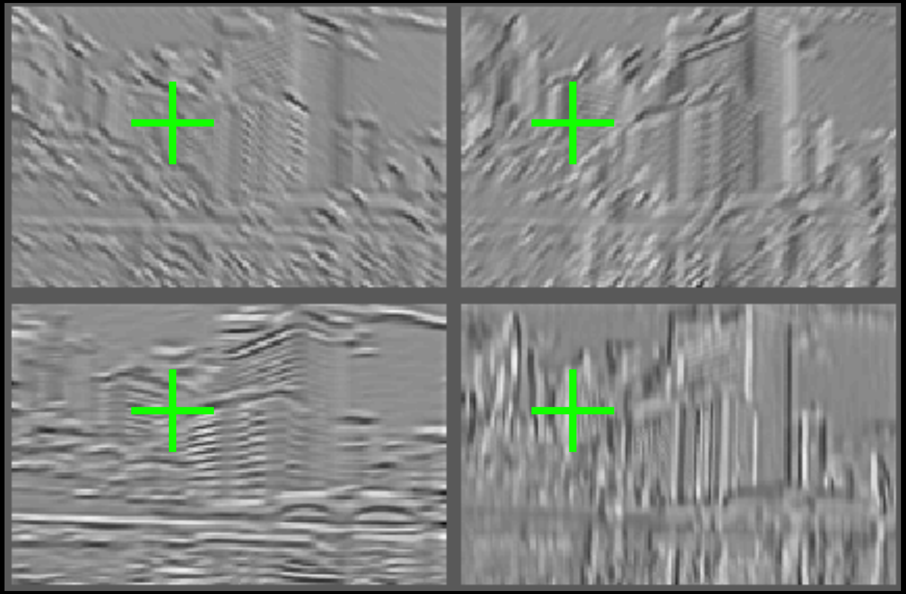
Filters

Normalization

- Contrast normalization (across feature maps)
  - Local mean = 0, local std. = 1, “Local” → 7x7 Gaussian
  - Equalizes the features maps



Feature Maps



Feature Maps
After Contrast Normalization

Role of Feature Normalization

- Introduces local competition between features
  - "Explaining away" in graphical models
  - Just like top-down models
  - But more local mechanism

- Also helps to scale activations at each layer better for learning
  - Makes energy surface more isotropic
  - So each gradient step makes more progress

- Empirically, seems to help a bit (1-2%) on ImageNet
- Most recent models don't seem to have use though

Normalization across Data

- Batch Normalization

[Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, Sergey Ioffe, Christian Szegedy, arXiv:1502.03167]

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1\ldots m}\}$;
Parameters to be learned: γ, β
Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_\mathcal{B} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma_\mathcal{B}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_\mathcal{B})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_\mathcal{B}}{\sqrt{\sigma_\mathcal{B}^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

Algorithm 1: Batch Normalizing Transform, applied to activation x over a mini-batch.

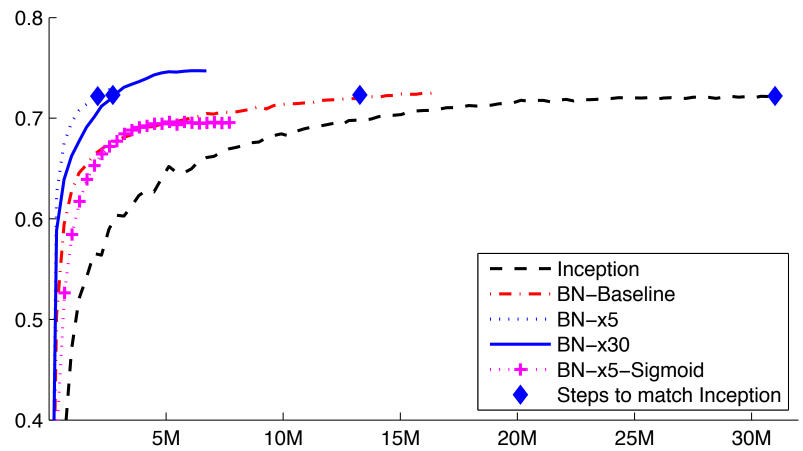


Figure 2: *Single crop validation accuracy of Inception and its batch-normalized variants, vs. the number of training steps.*

References

- [Slide 5]
- P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object Detection with Discriminatively Trained Part Based Models,IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, No. 9, September 2010
- Zheng Song*, Qiang Chen*, Zhongyang Huang, Yang Hua, and Shuicheng Yan. Contextualizing Object Detection and Classification. In CVPR'11. (* indicates equal contribution) [No. 1 performance in VOC'10 classification task]
- [Slide 6]
- Finding the Weakest Link in Person Detectors, D. Parikh, and C. L. Zitnick, CVPR, 2011.
- [Slide 7]
- Gehler and Nowozin, On Feature Combination for Multiclass Object Classification, ICCV'09
- [Slide 8]
- http://www.amazon.com/Vision-David-Marr/dp/0716712849
- [Slide 10]
- Yoshua Bengio and Yann LeCun: Scaling learning algorithms towards AI, in Bottou, L. and Chapelle, O. and DeCoste, D. and Weston, J. (Eds), Large-Scale Kernel Machines, MIT Press, 2007

References

- [Slide 11]
- S. Lazebnik, C. Schmid, and J. Ponce, Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, CVPR 2006
- [Slide 12]
- Christoph H. Lampert, Hannes Nickisch, Stefan Harmeling: "Learning To Detect Unseen Object Classes by Between-Class Attribute Transfer", IEEE Computer Vision and Pattern Recognition (CVPR), Miami, FL, 2009
- [Slide 14] Riesenhuber, M. & Poggio, T. (1999). Hierarchical Models of Object Recognition in Cortex. Nature Neuroscience 2: 1019-1025.
- http://www.scholarpedia.org/article/Neocognitron
- K. Fukushima: "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position", Biological Cybernetics, 36[4], pp. 193-202 (April 1980).
- Y. LeCun, L. Bottou, Y. Bengio and P. Haffner: Gradient-Based Learning Applied to Document Recognition, Proceedings of the IEEE, 86(11):2278-2324, November 1998

References

- [Slide 30]
- Y-Lan Boureau, Jean Ponce, and Yann LeCun, A theoretical analysis of feature pooling in vision algorithms, Proc. International Conference on Machine learning (ICML'10), 2010
- [Slide 31]
- Q.V. Le, J. Ngiam, Z. Chen, D. Chia, P. Koh, A.Y. Ng , Tiled Convolutional Neural Networks. NIPS, 2010
- http://ai.stanford.edu/~quocle/TCNNweb
- Matthew D. Zeiler, Graham W. Taylor, and Rob Fergus, Adaptive Deconvolutional Networks for Mid and High Level Feature Learning, International Conference on Computer Vision(November 6-13, 2011)
- [Slide 32]
- Yuanhao Chen, Long Zhu, Chenxi Lin, Alan Yuille, Hongjiang Zhang. Rapid Inference on a Novel AND/OR graph for Object Detection, Segmentation and Parsing. NIPS 2007.

References

- [Slide 35]
- P. Smolensky, Parallel Distributed Processing: Volume 1: Foundations, D. E. Rumelhart, J. L. McClelland, Eds. (MIT Press, Cambridge, 1986), pp. 194–281.
- G. E. Hinton, Neural Comput. 14, 1711 (2002).
- [Slide 36]
- M. Ranzato, Y. Boureau, Y. LeCun. "Sparse Feature Learning for Deep Belief Networks". Advances in Neural Information Processing Systems 20 (NIPS 2007).
- [Slide 39]
- Hinton, G. E. and Salakhutdinov, R. R., Reducing the dimensionality of data with neural networks. Science, Vol. 313. no. 5786, pp. 504 - 507, 28 July 2006.
- [Slide 41]
- A. Torralba, K. P. Murphy and W. T. Freeman, Contextual Models for Object Detection using Boosted Random Fields, Adv. in Neural Information Processing Systems 17 (NIPS), pp. 1401-1408, 2005.

References

- [Slide 42]
- Ruslan Salakhutdinov and Geoffrey Hinton, Deep Boltzmann Machines, 12th International Conference on Artificial Intelligence and Statistics (2009).
- [Slide 44]
- P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object Detection with Discriminatively Trained Part Based Models,IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, No. 9, September 2010
- Long Zhu, Yuanhao Chen, Alan Yuille, William Freeman. Latent Hierarchical Structural Learning for Object Detection. CVPR 2010.
- [Slide 45]
- Matthew D. Zeiler, Graham W. Taylor, and Rob Fergus, Adaptive Deconvolutional Networks for Mid and High Level Feature Learning, International Conference on Computer Vision(November 6-13, 2011)

References

- [Slide 48]
- S.C. Zhu and D. Mumford, A Stochastic Grammar of Images, Foundations and Trends in Computer Graphics and Vision, Vol.2, No.4, pp 259-362, 2006.
- [Slide 49]
- R. Girshick, P. Felzenszwalb, D. McAllester, Object Detection with Grammar Models, NIPS 2011
- [Slide 50]
- P. Felzenszwalb, D. Huttenlocher, Pictorial Structures for Object Recognition, International Journal of Computer Vision, Vol. 61, No. 1, January 2005
- M. Fischler and R. Elschlager. The Representation and Matching of Pictoral Structures. (1973)
- [Slide 51]
- S. Fidler, M. Boben, A. Leonardis. A coarse-to-fine Taxonomy of Constellations for Fast Multi-class Object Detection. ECCV 2010.
- S. Fidler and A. Leonardis. Towards Scalable Representations of Object Categories: Learning a Hierarchy of Parts. CVPR 2007.

References

- [Slide 52]
- Long Zhu, Chenxi Lin, Haoda Huang, Yuanhao Chen, Alan Yuille. Unsupervised Structure Learning: Hierarchical Recursive Composition, Suspicious Coincidence and Competitive Exclusion. ECCV 2008.
- [Slide 53]
- Hinton, G. E., Krizhevsky, A. and Wang, S, Transforming Auto-encoders. ICANN-11: International Conference on Artificial Neural Networks, 2011
- Matthew D. Zeiler, Graham W. Taylor, and Rob Fergus, Adaptive Deconvolutional Networks for Mid and High Level Feature Learning, International Conference on Computer Vision(November 6-13, 2011)
- [Slide 54]
- Q.V. Le, M.A. Ranzato, R. Monga, M. Devin, K. Chen, G.S. Corrado, J. Dean, A.Y. Ng., Building high-level features using large scale unsupervised learning. ICML, 2012.
- [Slide 55]
- Ruslan Salakhutdinov and Geoffrey Hinton, Deep Boltzmann Machines, 12th International Conference on Artificial Intelligence and Statistics (2009).

References

- [Slide 56]
- http://www.image-net.org/challenges/LSVRC/2010/
- Q.V. Le, M.A. Ranzato, R. Monga, M. Devin, K. Chen, G.S. Corrado, J. Dean, A.Y. Ng., Building high-level features using large scale unsupervised learning. ICML, 2012.
- Q.V. Le, W.Y. Zou, S.Y. Yeung, A.Y. Ng., Learning hierarchical spatio-temporal features for action recognition with independent subspace analysis, CVPR 2011