ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton

Presented by Baotuan Nguyen and Markham Anderson
ECS 289G, 1.18.2018

*Slides adapted from Tugce Tasci and Kyunghee Kim for CS231B at Stanford University

ImageNet

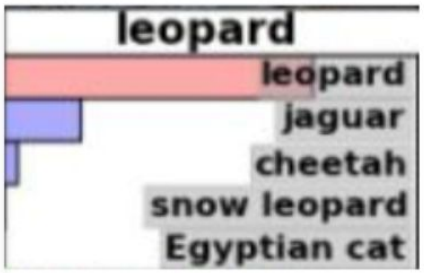
- Over 15M labeled high resolution images
- Roughly 22K categories
- Collected from web and labeled by Amazon Mechanical Turk

http://image-net.org/

ILSVRC

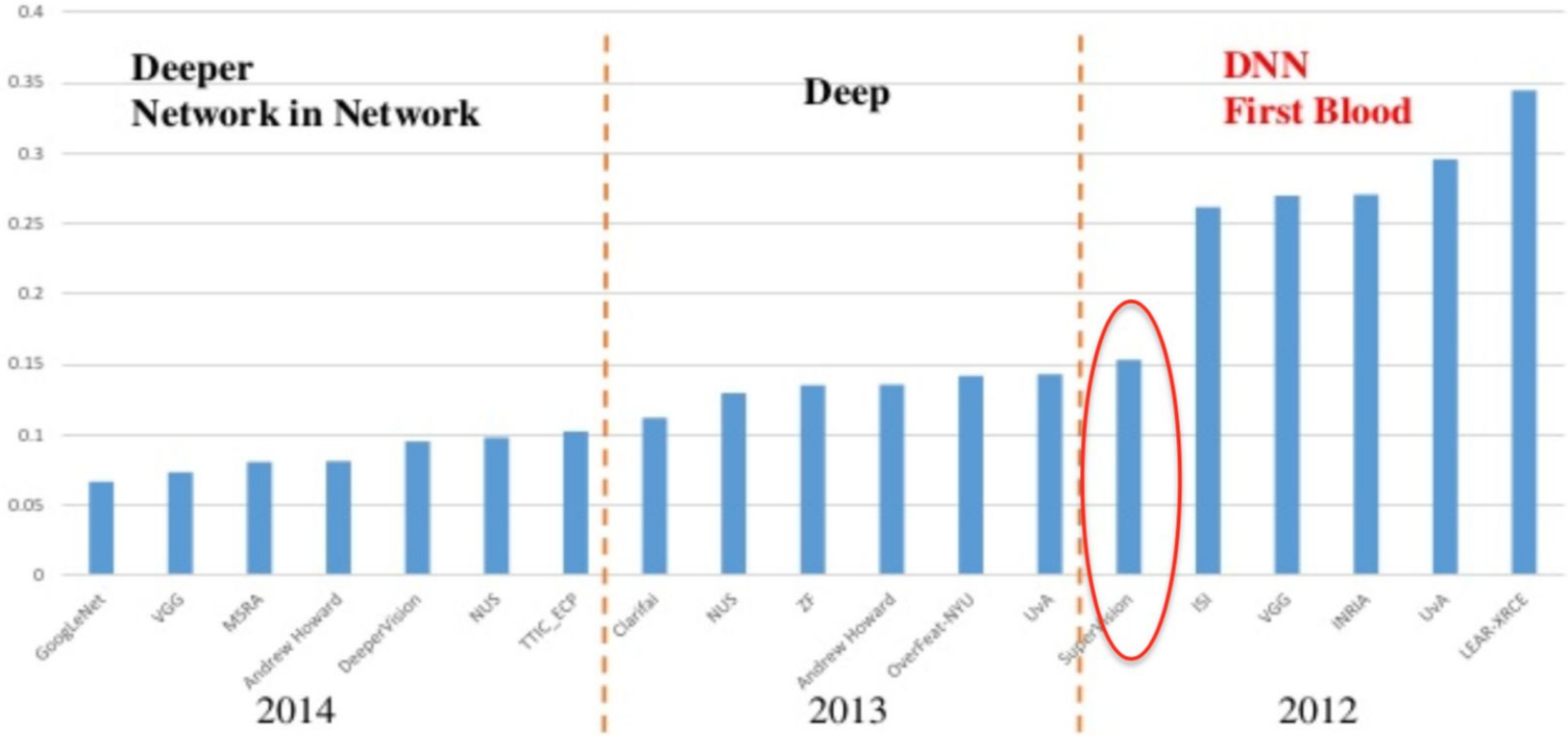
- Annual competition of image classification at large scale
- 1.2M images in 1K categories
- Classification: make 5 guesses about the image label

Classification



ILSVRC

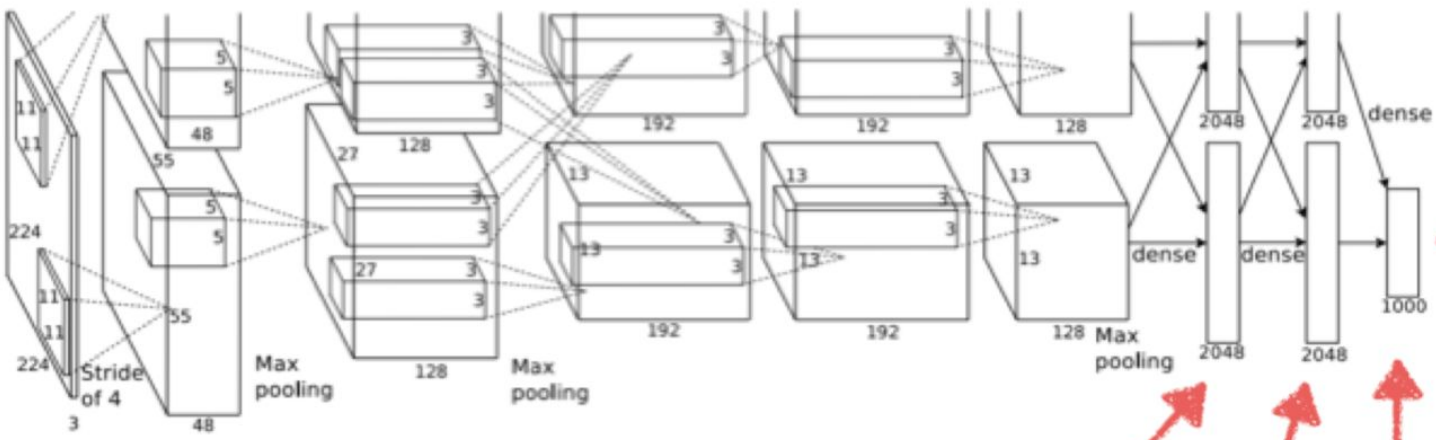
ImageNet Classification error throughout years and groups



Li Fei-Fei: ImageNet Large Scale Visual Recognition Challenge, 2014 http://image-net.org/

Architecture

5 Convolutional Layers



1000-way
softmax

3 Fully Connected Layers

SuperVision (SV)

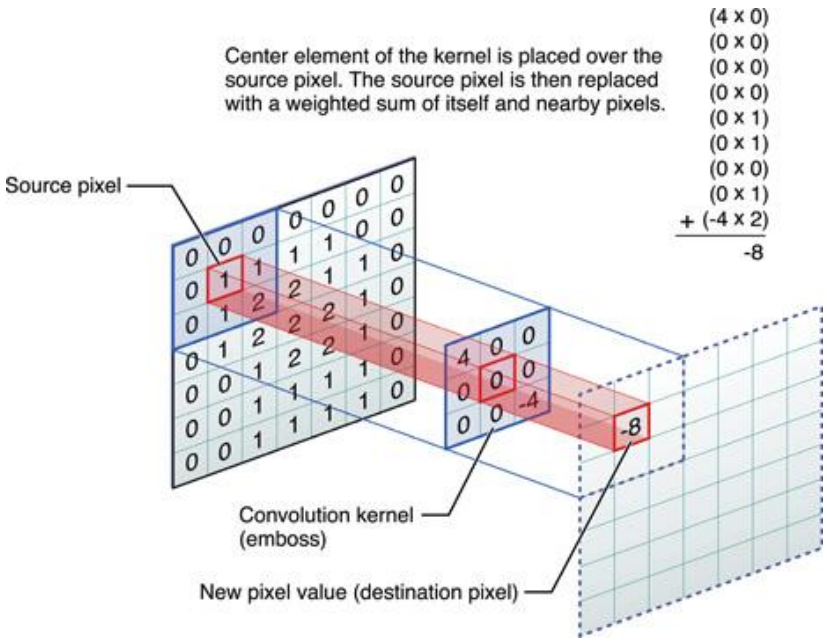
Image classification with deep convolutional neural networks

- 7 hidden “weight” layers
- 650K neurons
- 60M parameters
- 630M connections

http://image-net.org/challenges/LSVRC/2012/supervision.pdf

Why Convolutions for Images?

- Images are very large inputs, (i.e. 200x200 = 40K pixels) not scalable for fully connected input to a traditional NN.
- Convolutions are small filters (i.e. 3x3 = 9 parameters) that we can learn and apply over portions of an image.



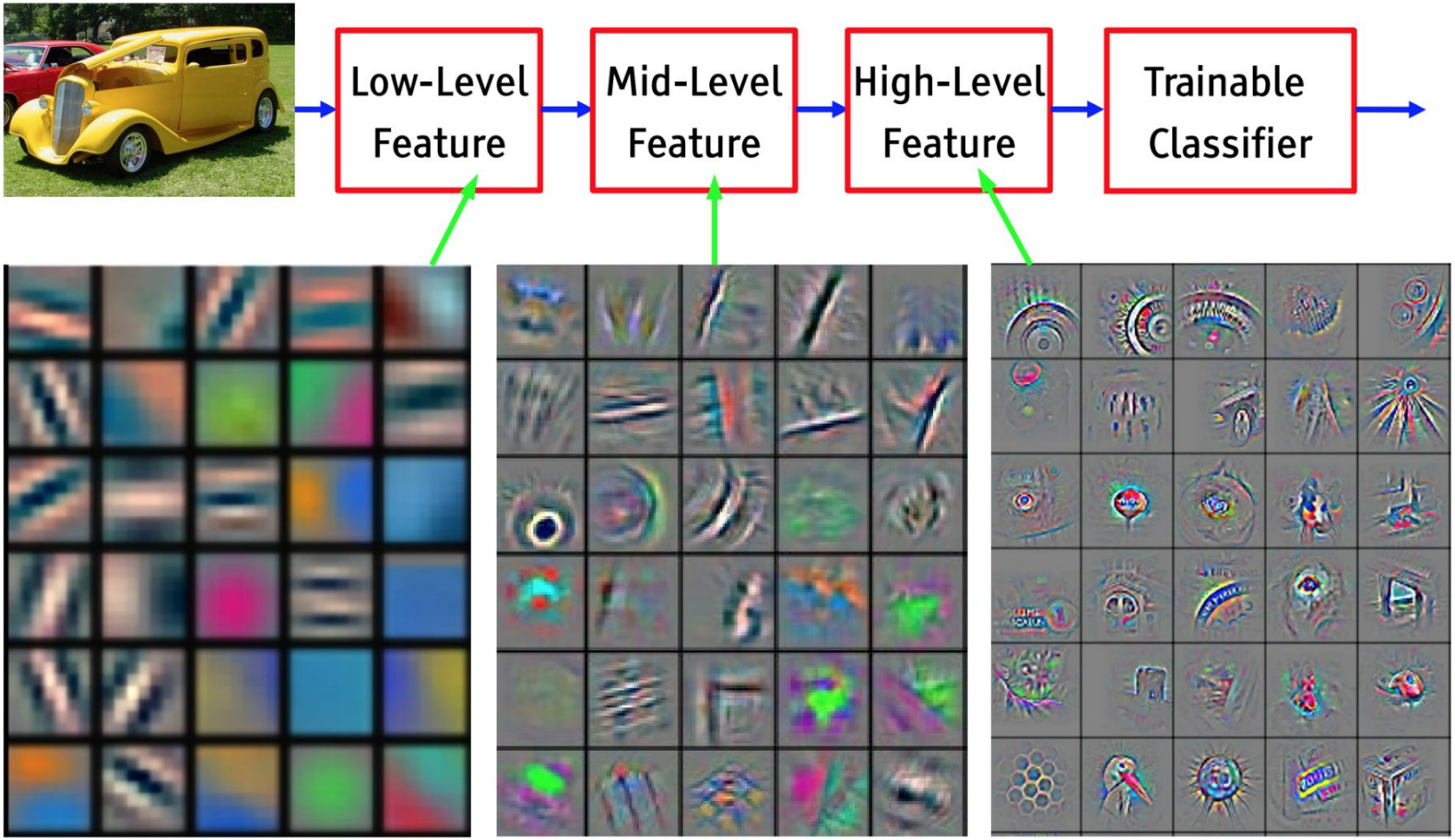
Input image

Convolution Kernel

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

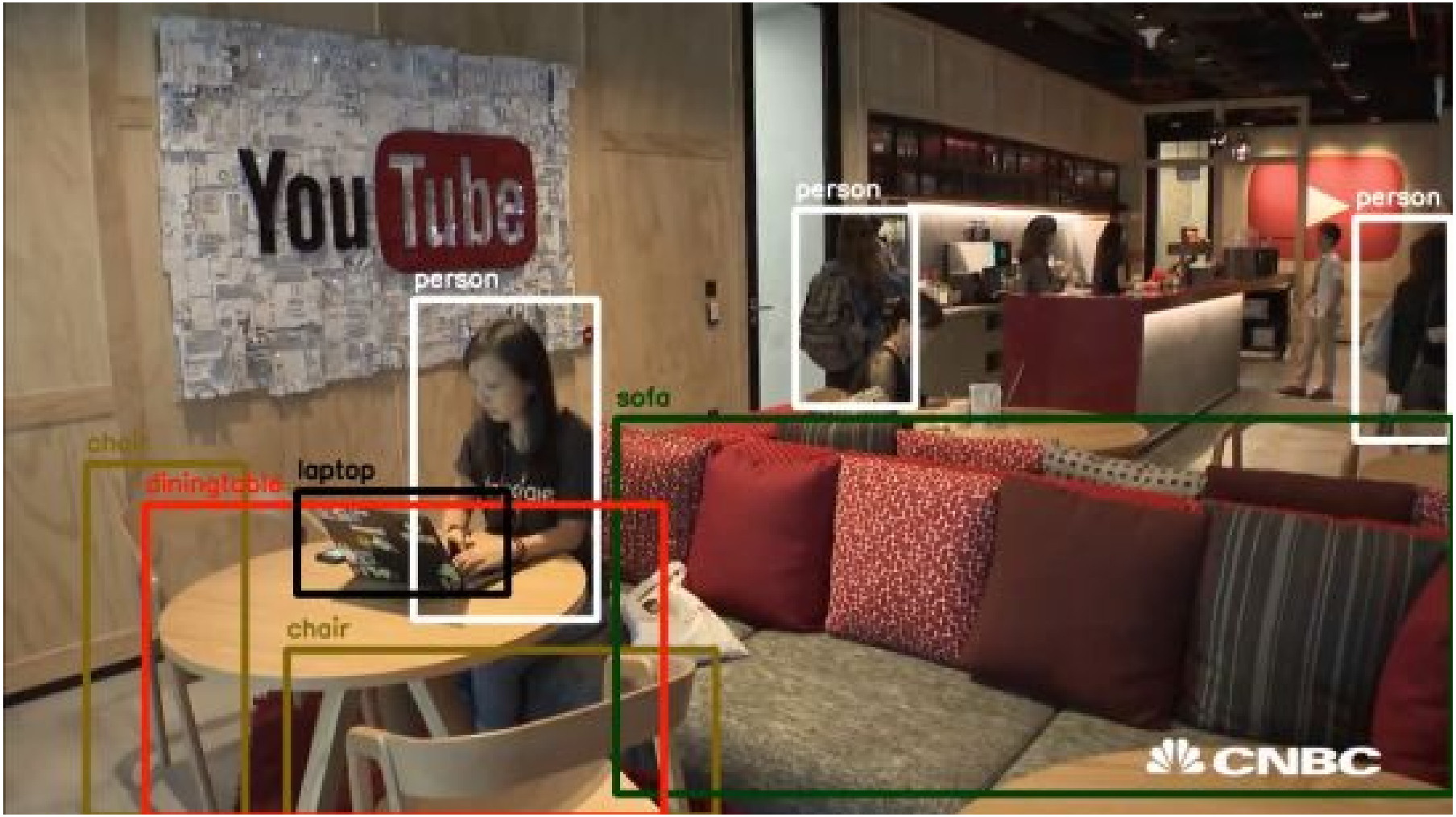
Feature map

Hierarchical representation



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Why Does This Work? Pixel Locality!



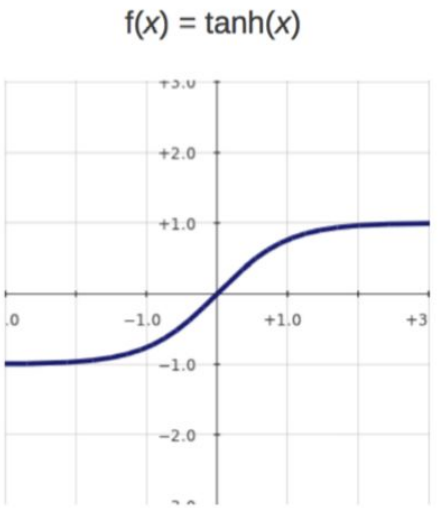
Architecture

RELU Nonlinearity

- Standard way to model a neuron

  $f(x) = tanh(x)$ or $f(x) = (1 + e^{-x})^{-1}$

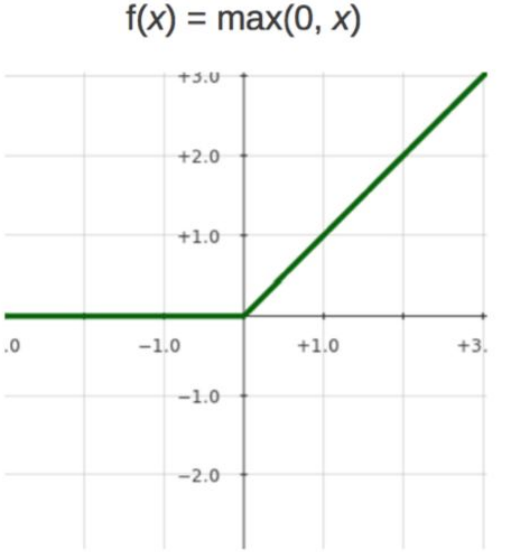
  Very slow to train



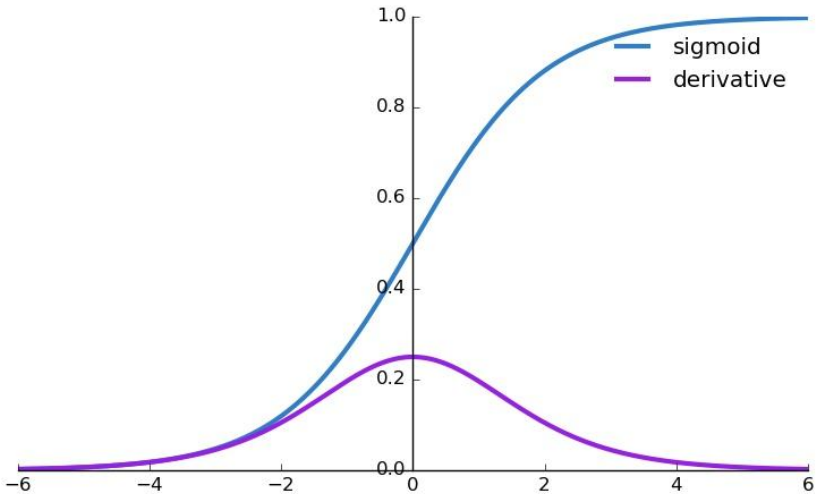
- Non-saturating nonlinearity (RELU)

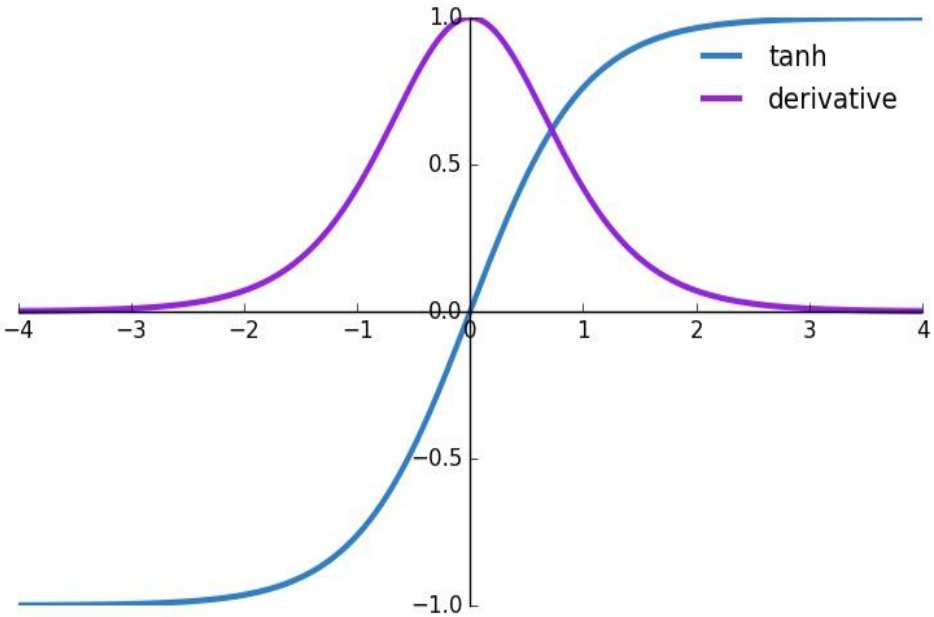
  $f(x) = max(0, x)$

  Quick to train

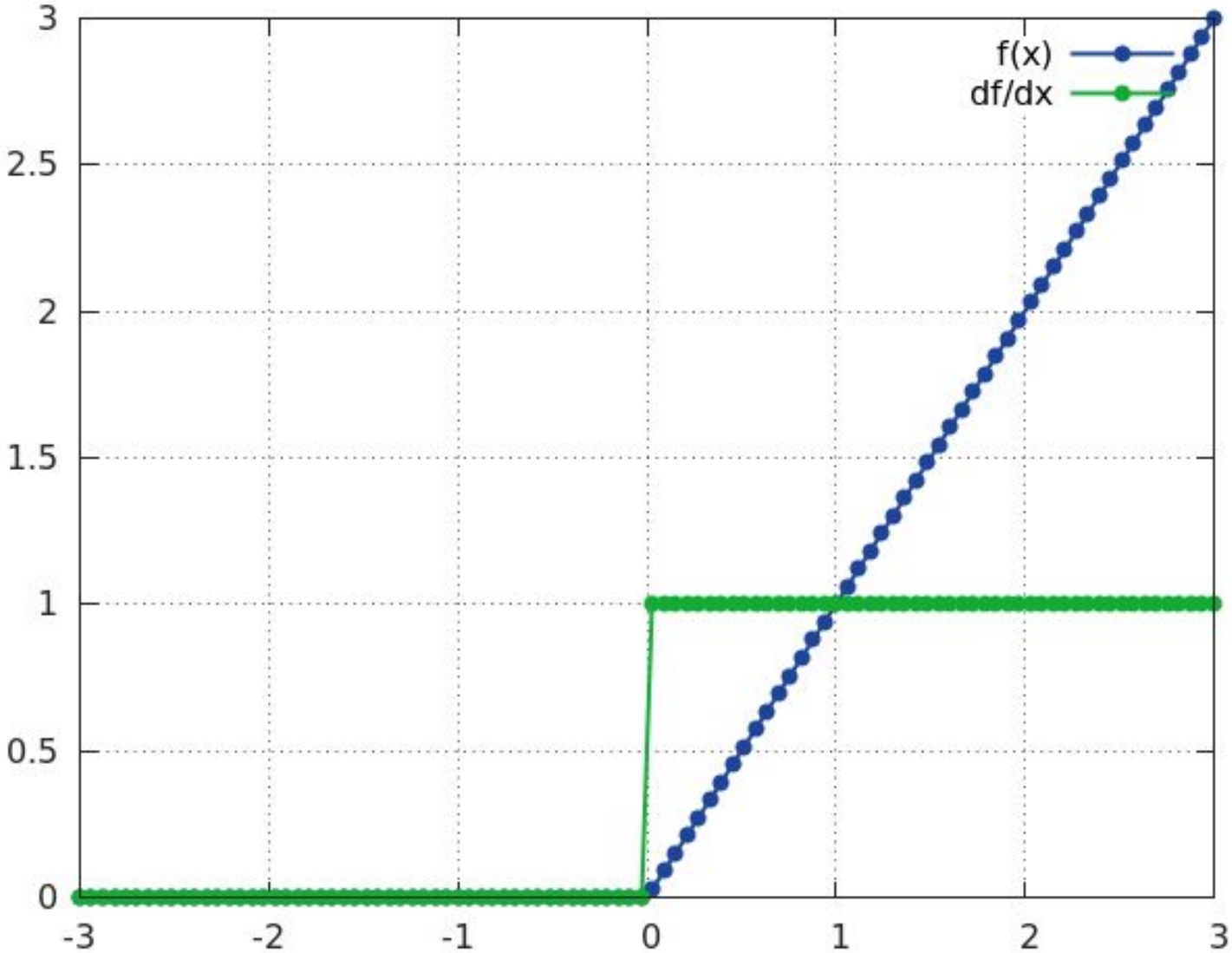


The Vanishing Gradient Problem



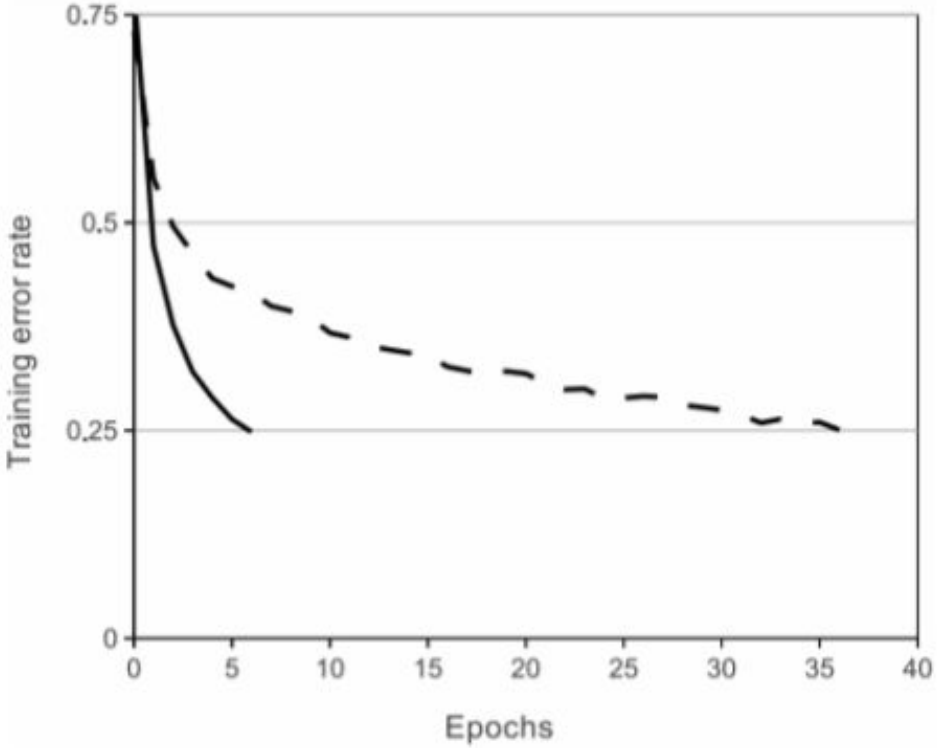


RELU to the Rescue



Architecture

RELU Nonlinearity



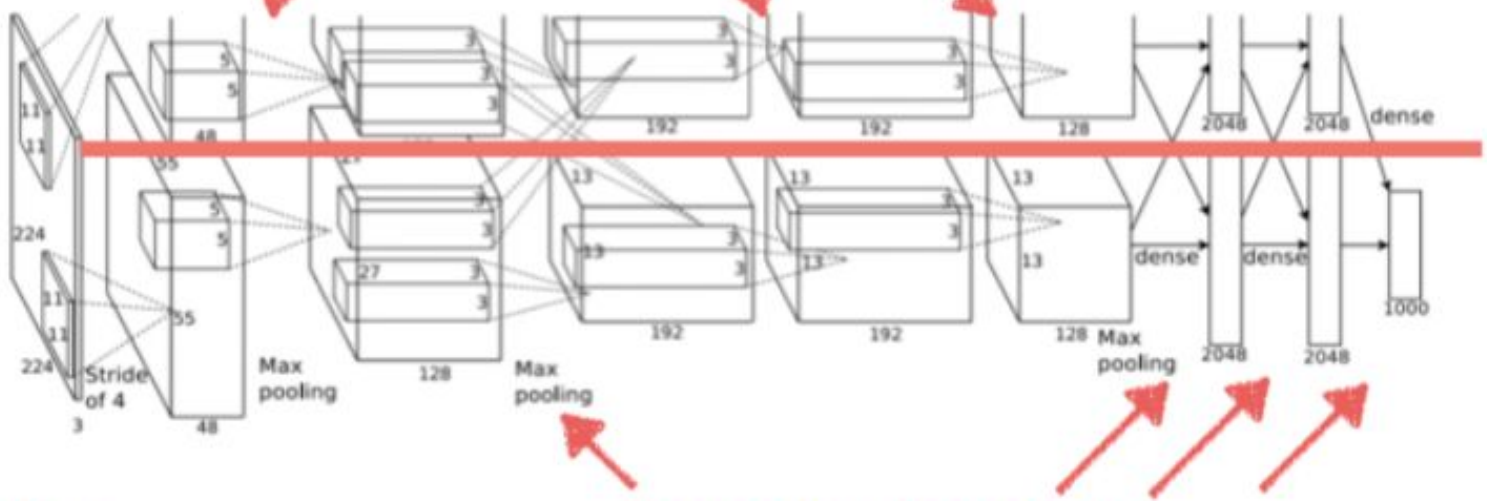
A 4 layer CNN with ReLUs (solid line) converges six times faster than an equivalent network with tanh neurons (dashed line) on CIFAR-10 dataset

Architecture

Training on Multiple GPUs

GPU #1

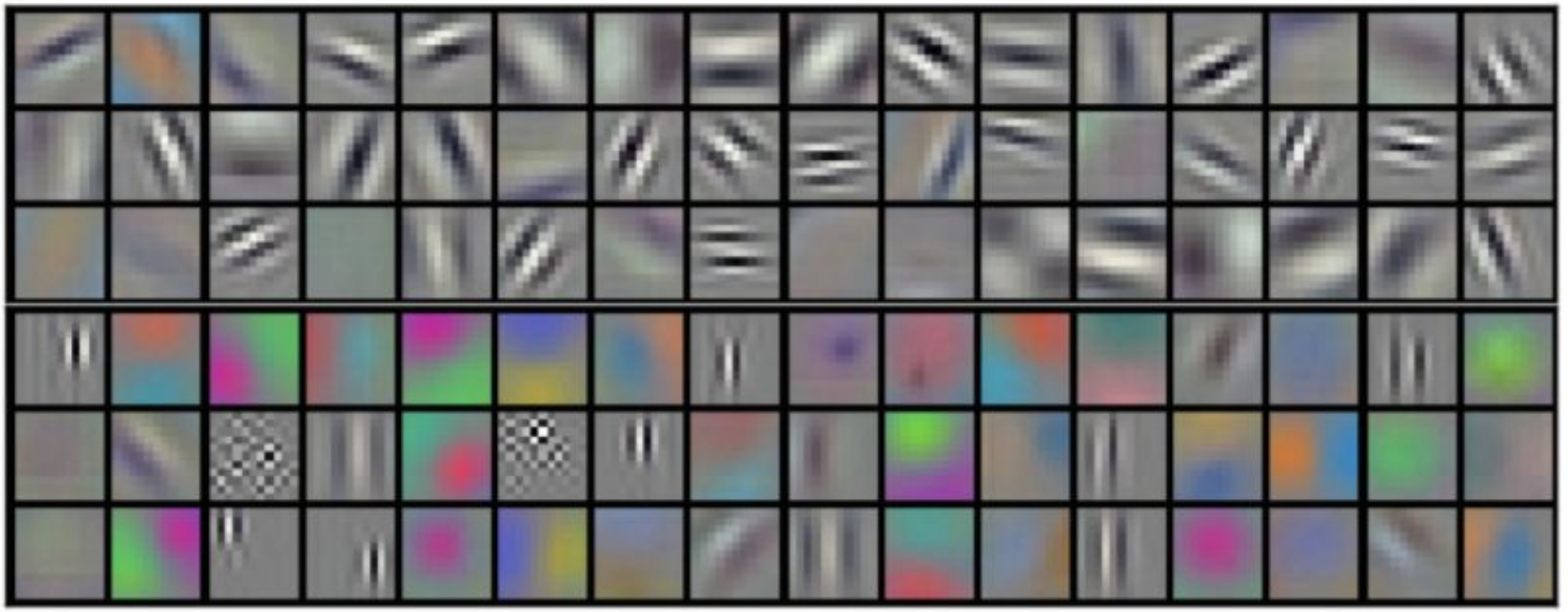
intra-GPU connections



GPU #2

inter-GPU connections

96 Convolutional Kernels

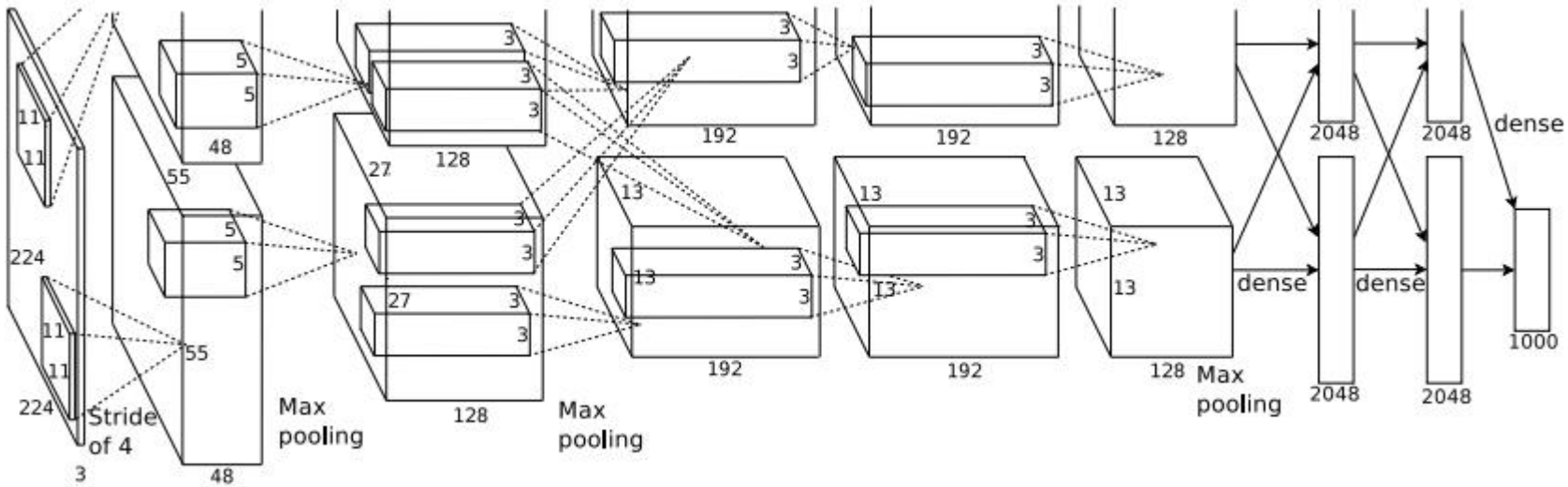


- 11 x 11 x 3 size kernels.
- top 48 kernels on GPU 1 : color-agnostic
- bottom 48 kernels on GPU 2 : color-specific.

Why?

Artificially Enlarging the Dataset

- 60 million parameters
- 1.2 million training images



Data Augmentation: Altering RGB Intensities

"Object identity is invariant to changes in intensity and color of the illumination."

p = eigenvector
ƛ = eigenvalue
α ~ N(0,0.1)

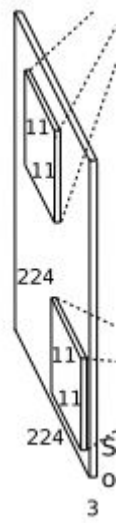
$$[I^R_{xy}, I^G_{xy}, I^B_{xy}] + [p_1, p_2, p_3][\alpha_1 ƛ_1, \alpha_2 ƛ_2, \alpha_3 ƛ_3]$$

PC: https://photodune.net/item/grand-canal-at-night-venice/4823526

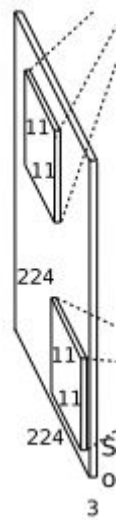
Data Augmentation: Multiple Patch Extraction

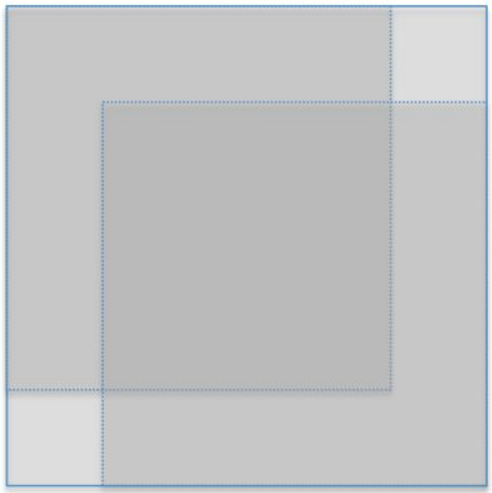
Downsample and crop to 256x256

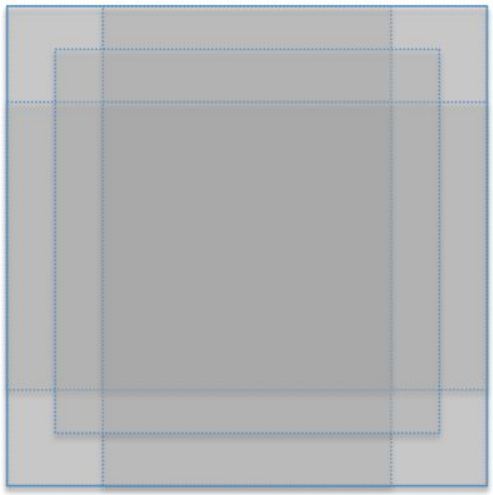
But input layer is 224x224



Data Augmentation: Multiple Patch Extraction



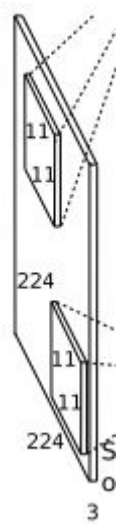




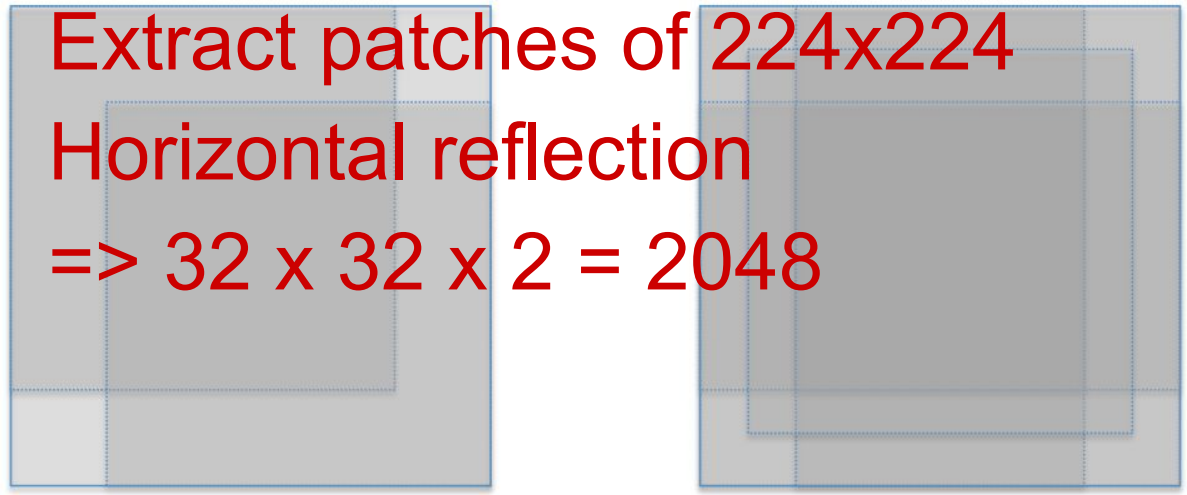
PC: Presentation by Tugce Tasci, Kyunghee Kim 05/18/2015

Data Augmentation: Multiple Patch Extraction

Downsample and crop to 256x256



- Extract patches of 224x224
- Horizontal reflection
- => 32 x 32 x 2 = 2048



60 million parameters vs 108 million images

PC: Presentation by Tugce Tasci, Kyunghee Kim 05/18/2015

Data Augmentation: Other Label-Preserving Transformations

- rotation
- scaling
- translation

D. Cireşan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. Arxiv preprint arXiv:1202.2745, 2012.

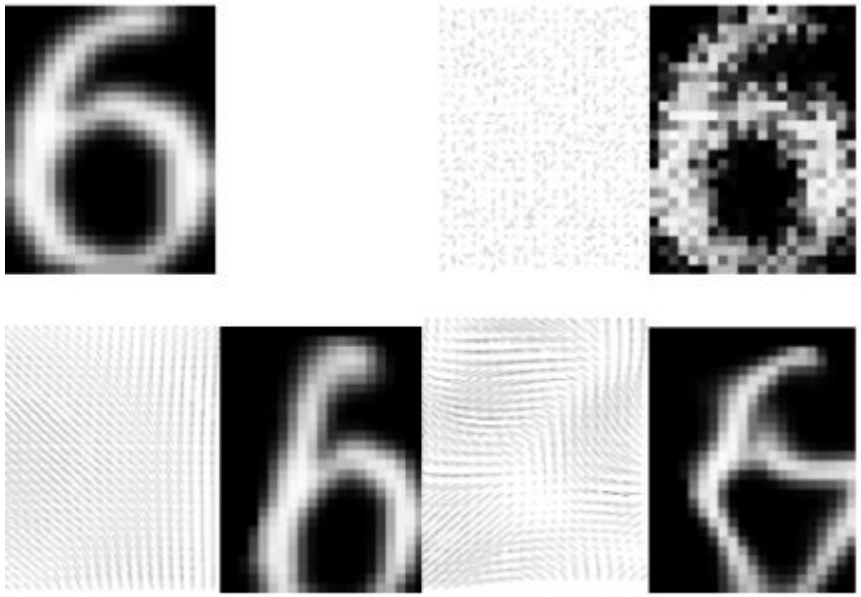
- horizontal shearing
- elastic deformations

D.C. Cireşan, U. Meier, J. Masci, L.M. Gambardella, and J. Schmidhuber. High-performance neural networks for visual object classification. Arxiv preprint arXiv:1102.0183, 2011.

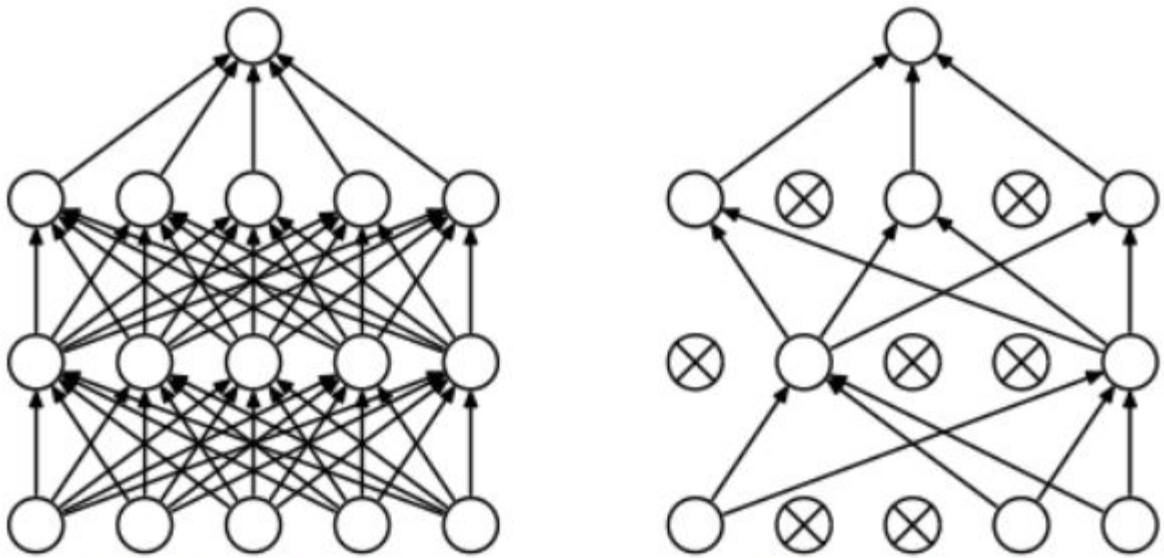
Data Augmentation: Other Label-Preserving Transformations

Displacement fields with smoothing (bilinear interpolation)

P.Y. Simard, D. Steinkraus, and J.C. Platt. Best practices for convolutional neural networks applied to visual document analysis. In Proceedings of the Seventh International Conference on Document Analysis and Recognition, volume 2, pages 958–962, 2003.



Dropout



Standard Neural Net

After applying dropout.

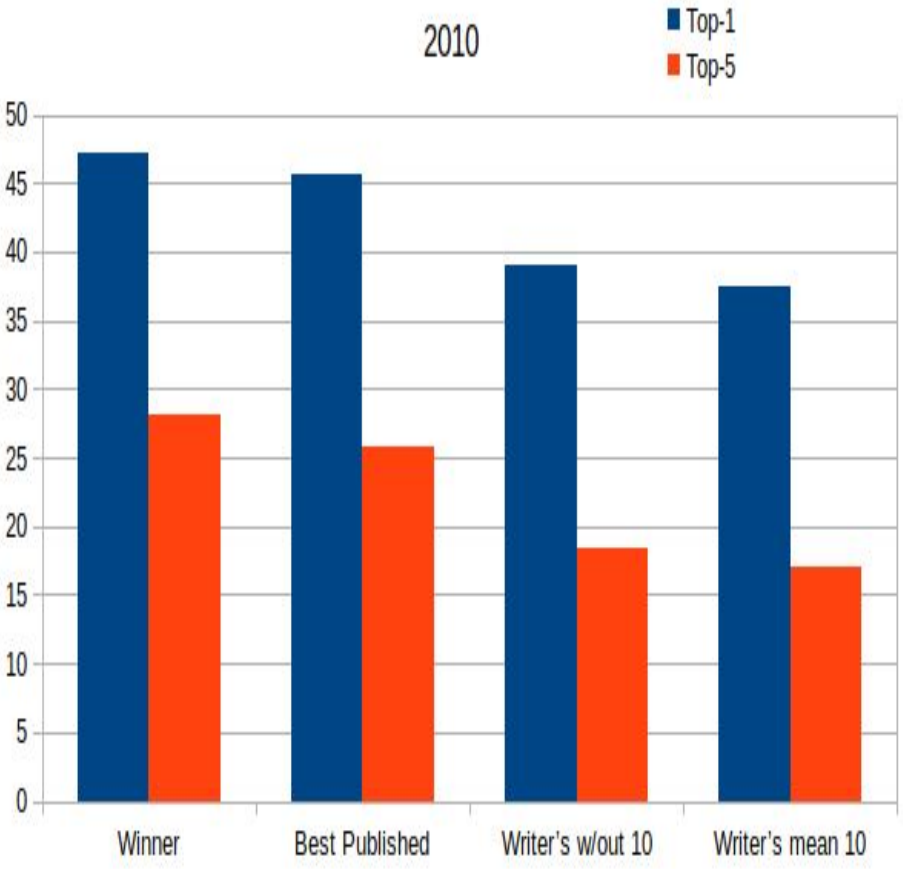
PC: [Srivastava et al.]

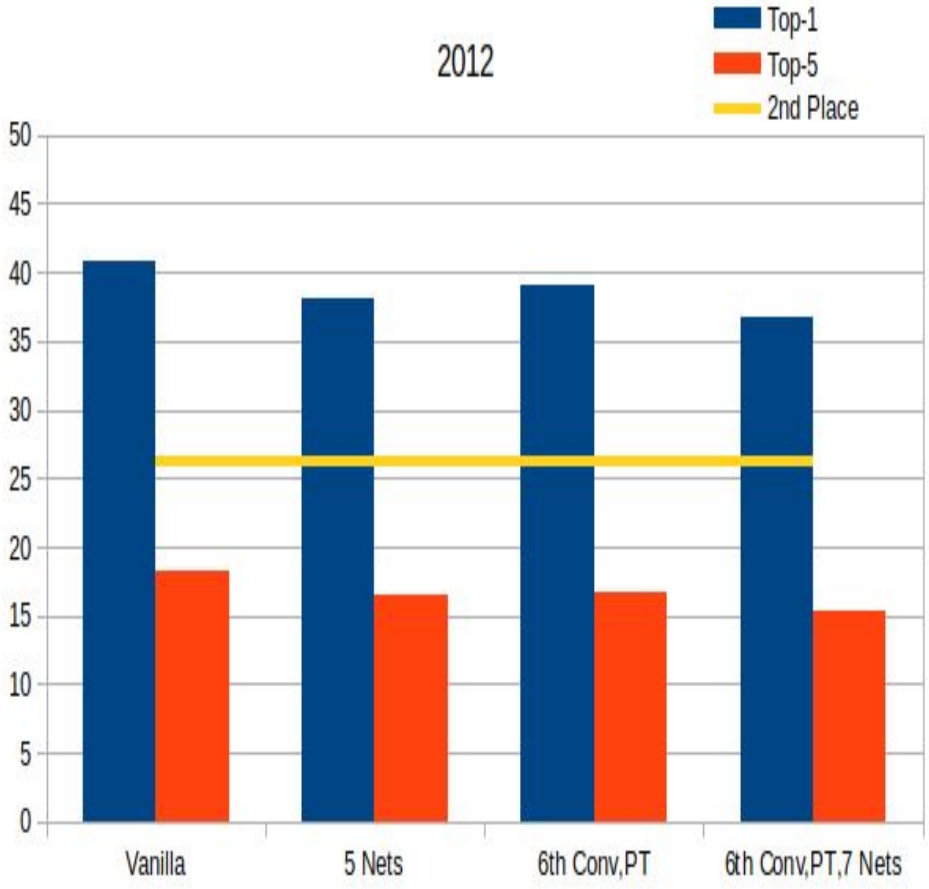
Datasets

	images	categories
ImageNet 2009	8.9 million	22000
†ImageNet 2011	15.0 million	10184
*ILSVRC 2010	1.2 million	1000
‡ILSVRC 2012	“”	“”

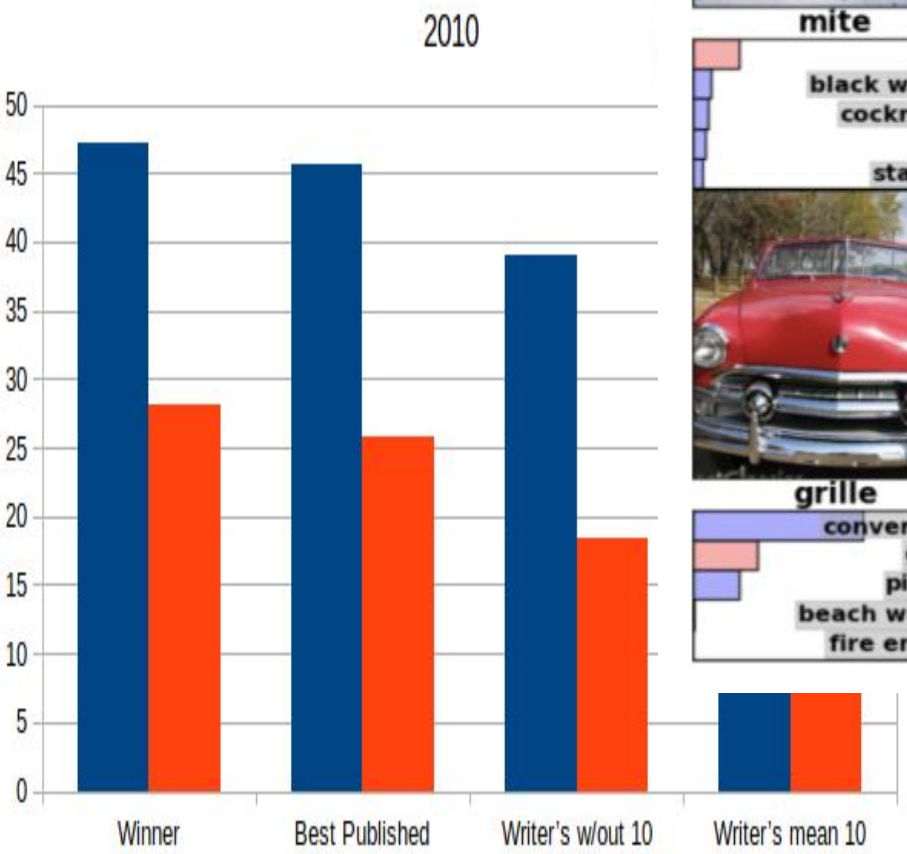
*primary reporting
†pretraining
‡competition entry

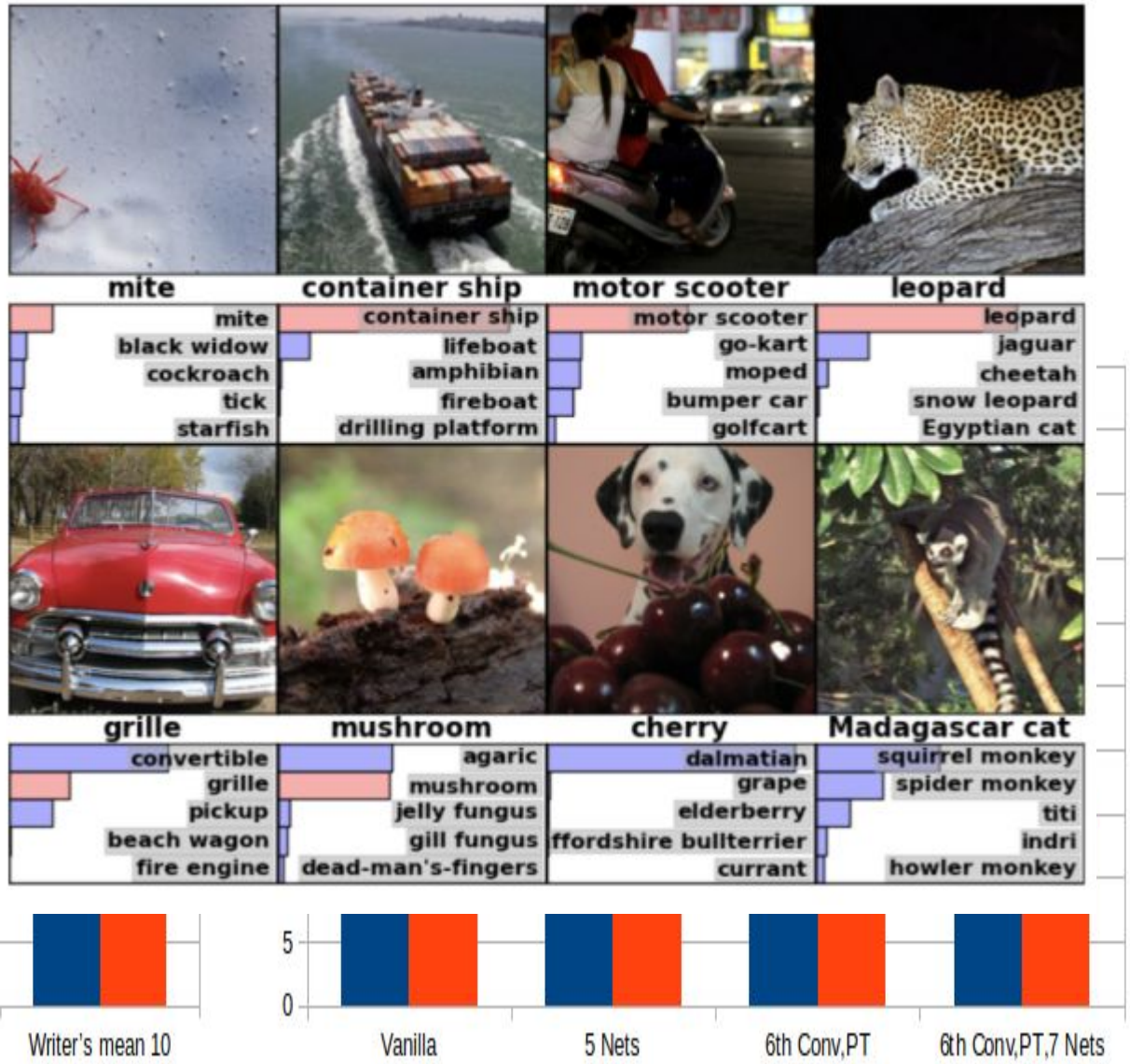
Evaluating Performance: Error



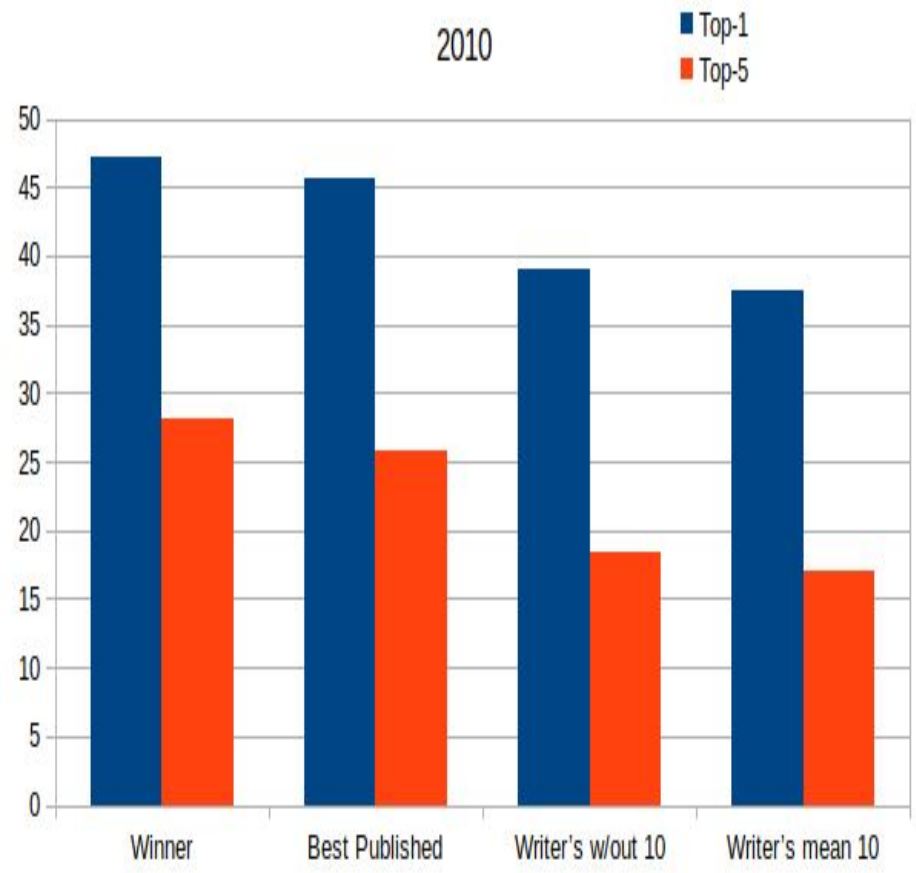


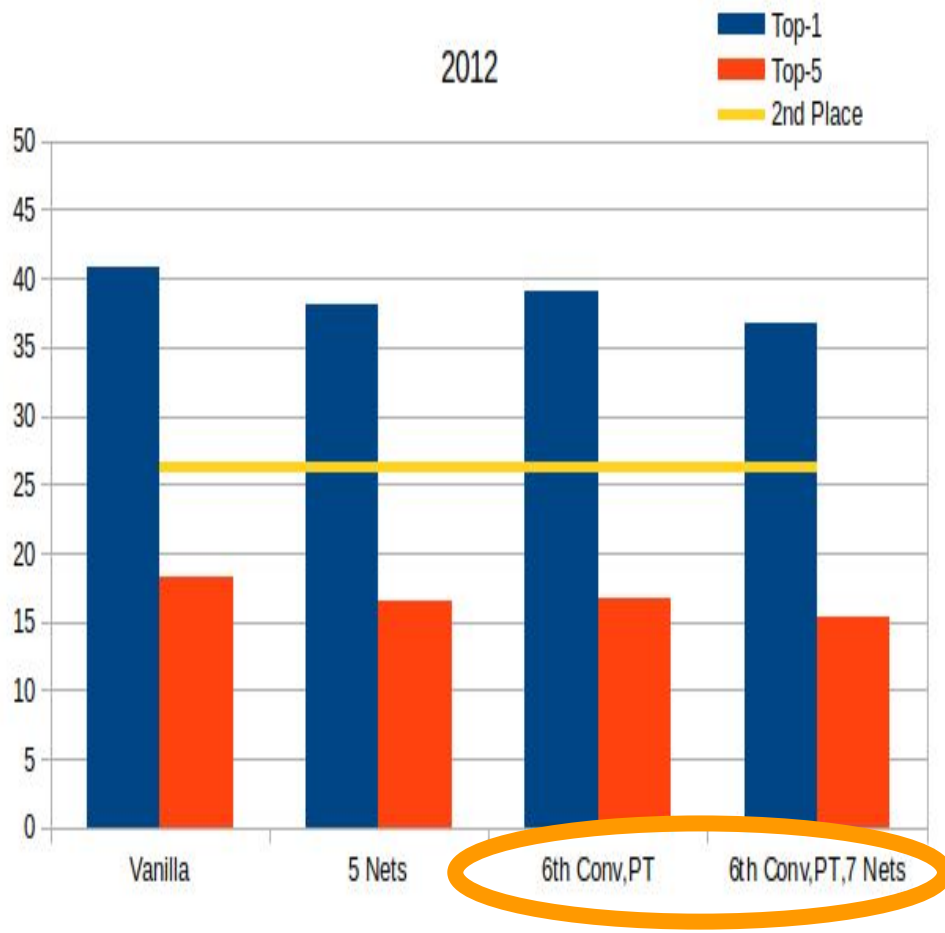
Evaluating Perforr



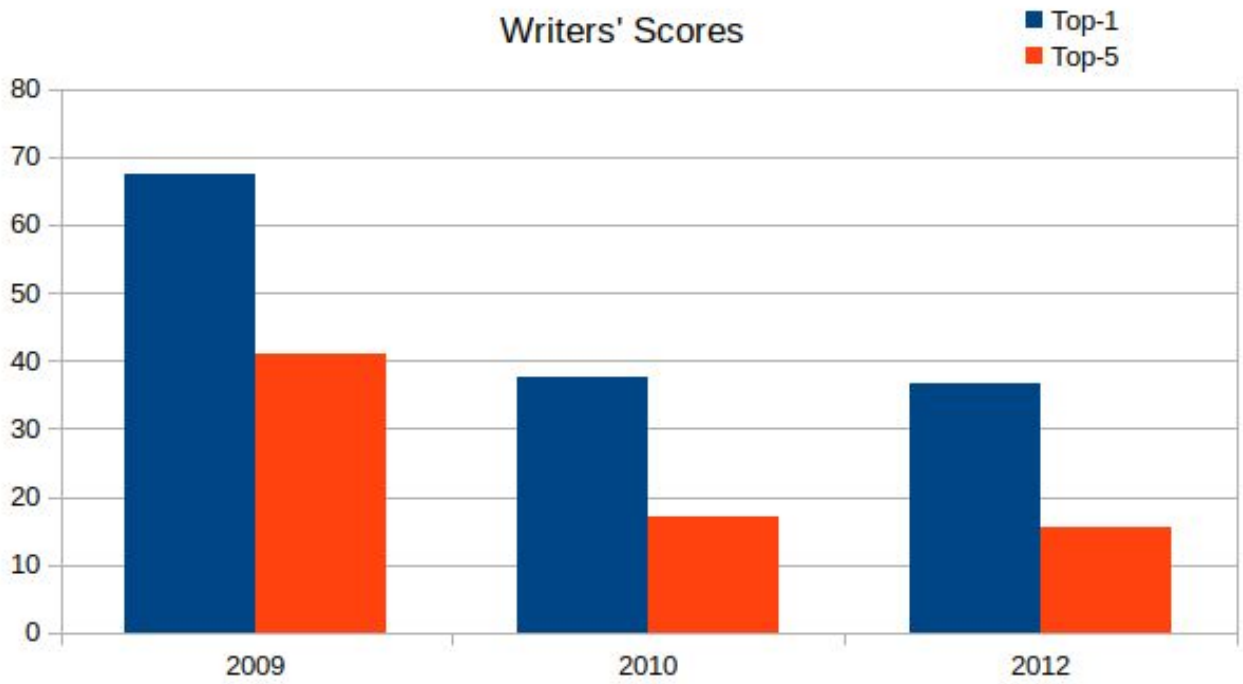


Evaluating Performance: Error





Evaluating Performance: Error



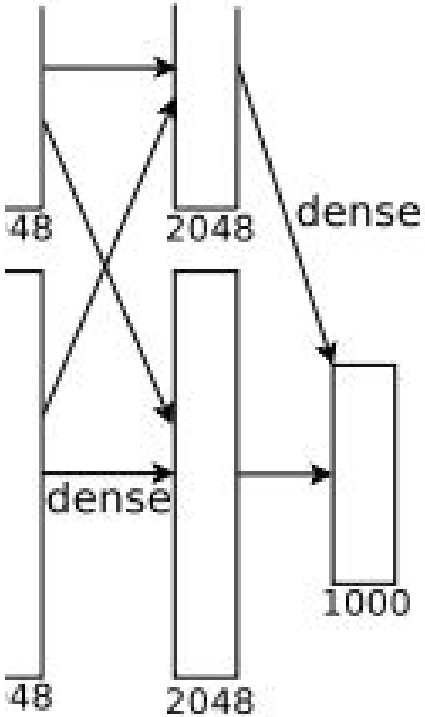
2010, 2012

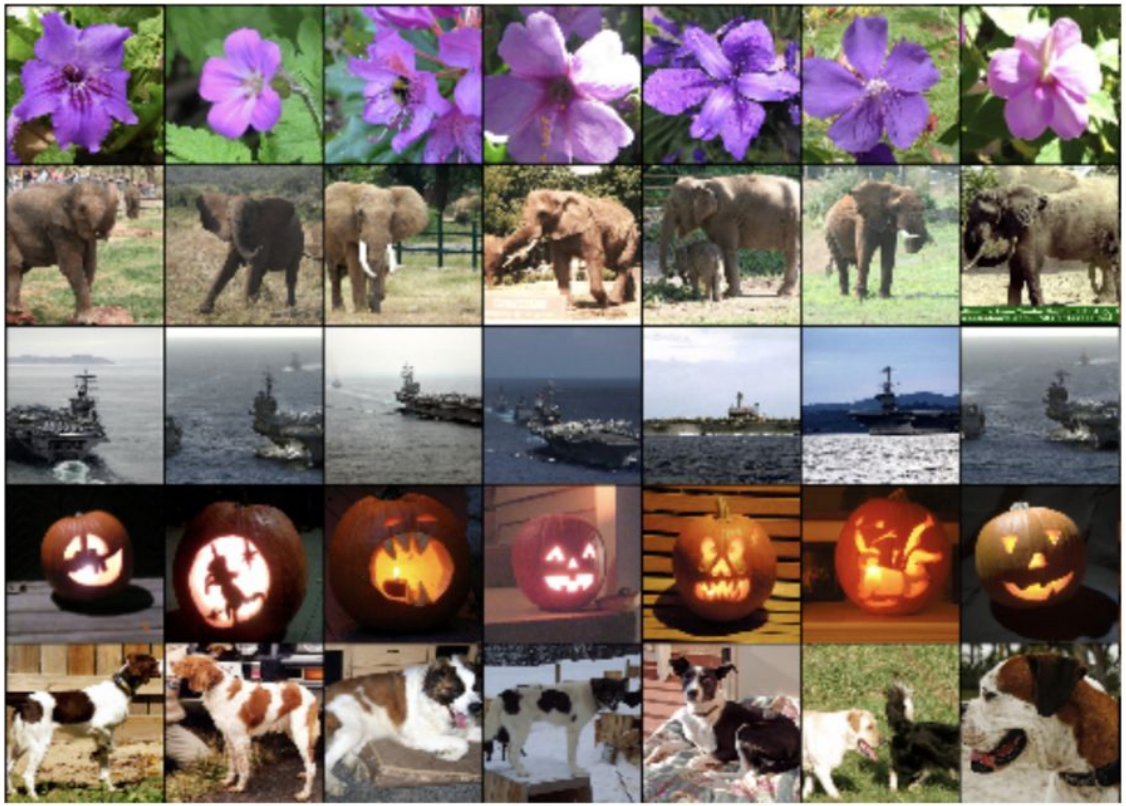
- 1000 categories
- 1.8 million images

2009

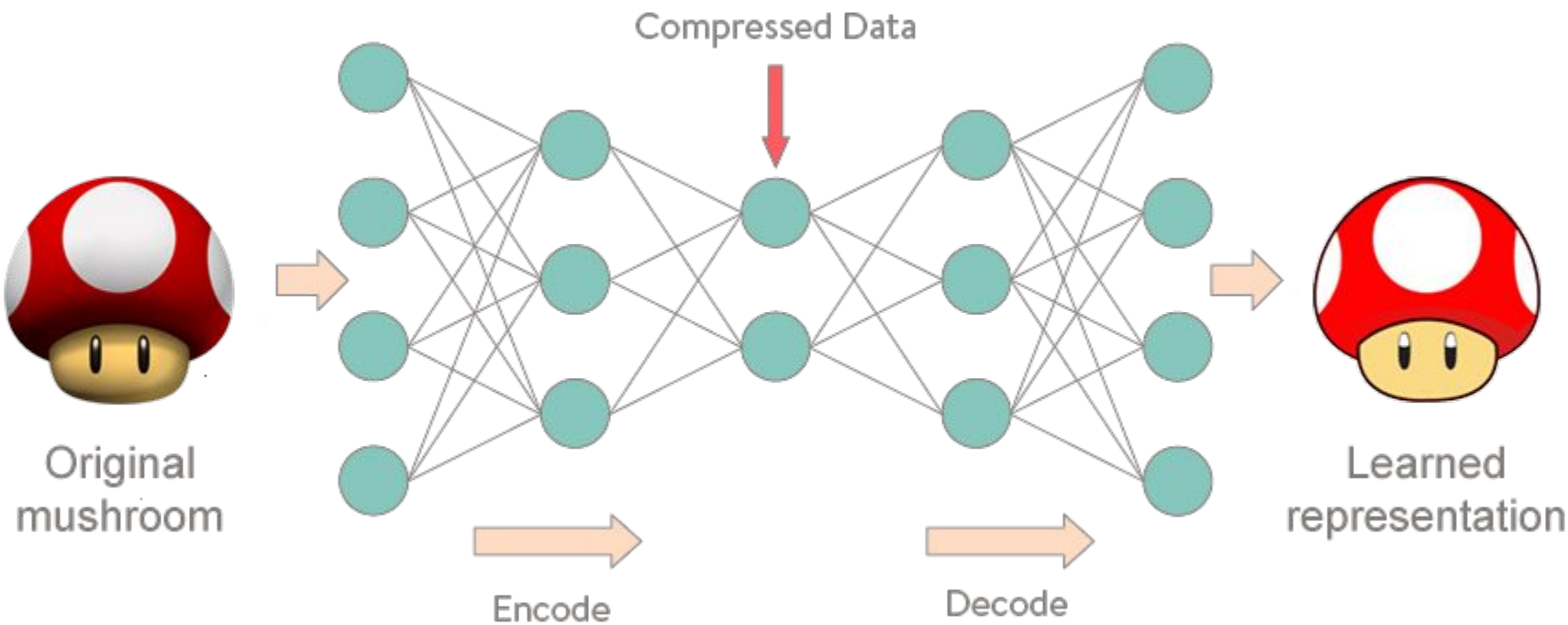
- 10,184 categories
- 8.9 million images

Evaluating Performance: Nearest Neighbours



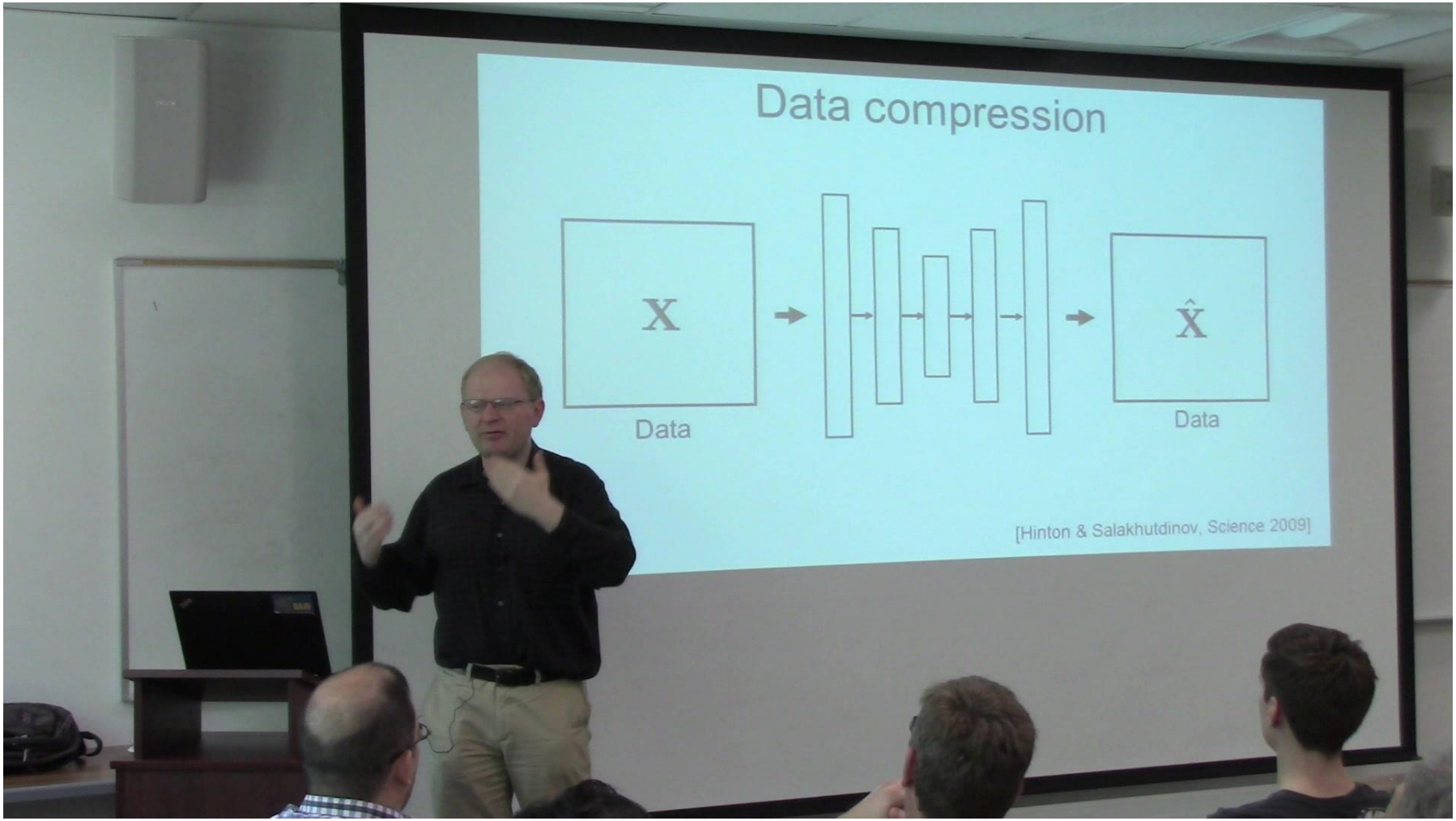


Autoencoders



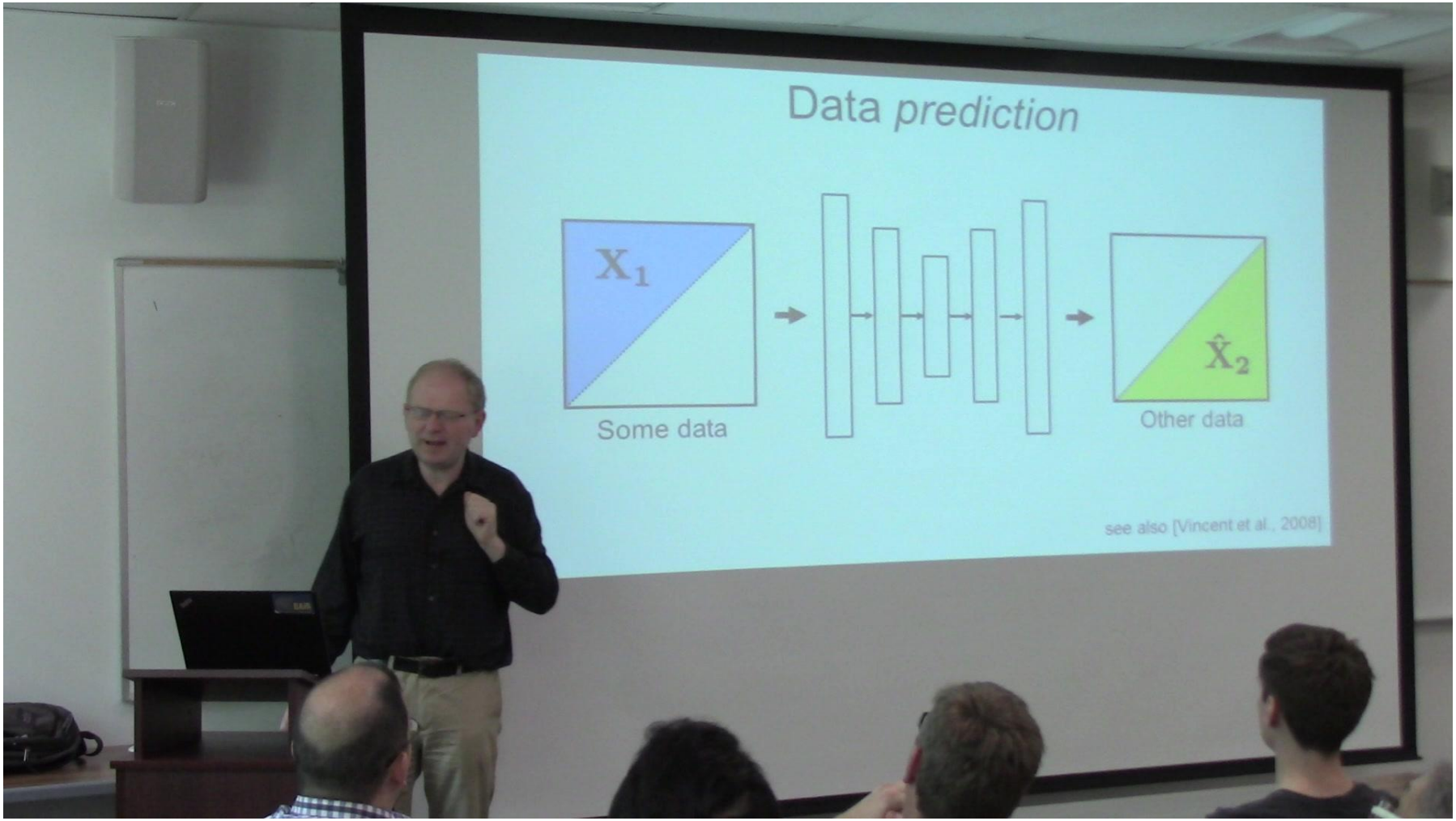
PC: https://medium.com/@curiousily/credit-card-fraud-detection-using-autoencoders-in-keras-tensorflow-for-hackers-part-vii-20e0c85301bd

Unsupervised Learning Without Reconstruction



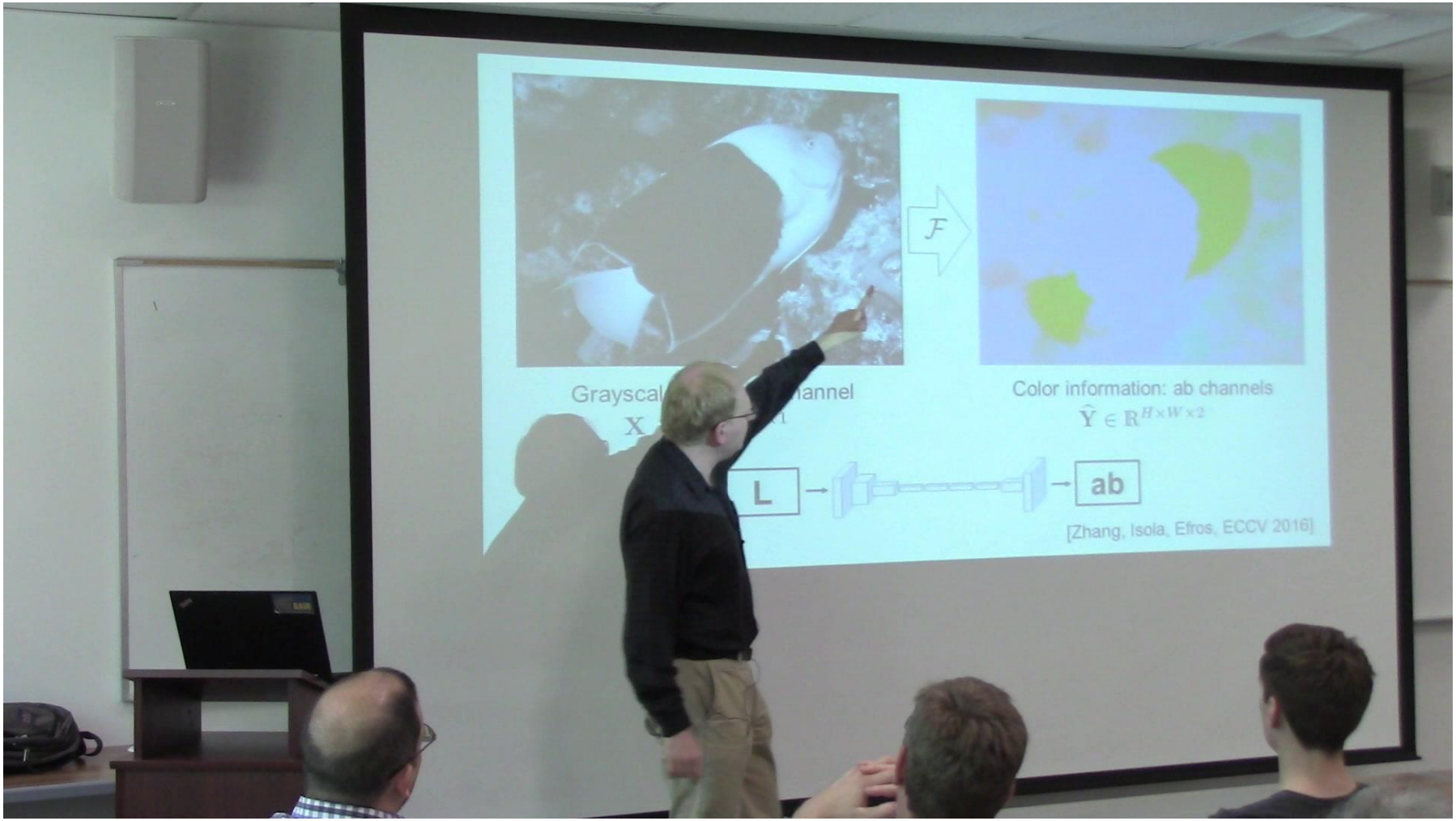
PC: ttps://ucdavis.box.com/s/xf5wvae9wxxh159pkqfs0bvr065umh47

Unsupervised Learning Without Reconstruction



PC: ttps://ucdavis.box.com/s/xf5wvae9wxxh159pkqfs0bvr065umh47

Unsupervised Learning Without Reconstruction



PC: ttps://ucdavis.box.com/s/xf5wvae9wxxh159pkqfs0bvr065umh47

Strengths

- First attempt at deep learning for ILSVRC.
- Paved the path for state of the art in computer vision.

Weaknesses

- Several parts were not well explained, making it difficult for first timers.
- Ex. Local Response Normalization

$$b^i_{x,y} = a^i_{x,y} / \left(k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a^j_{x,y})^2 \right)^\beta$$

Discussion

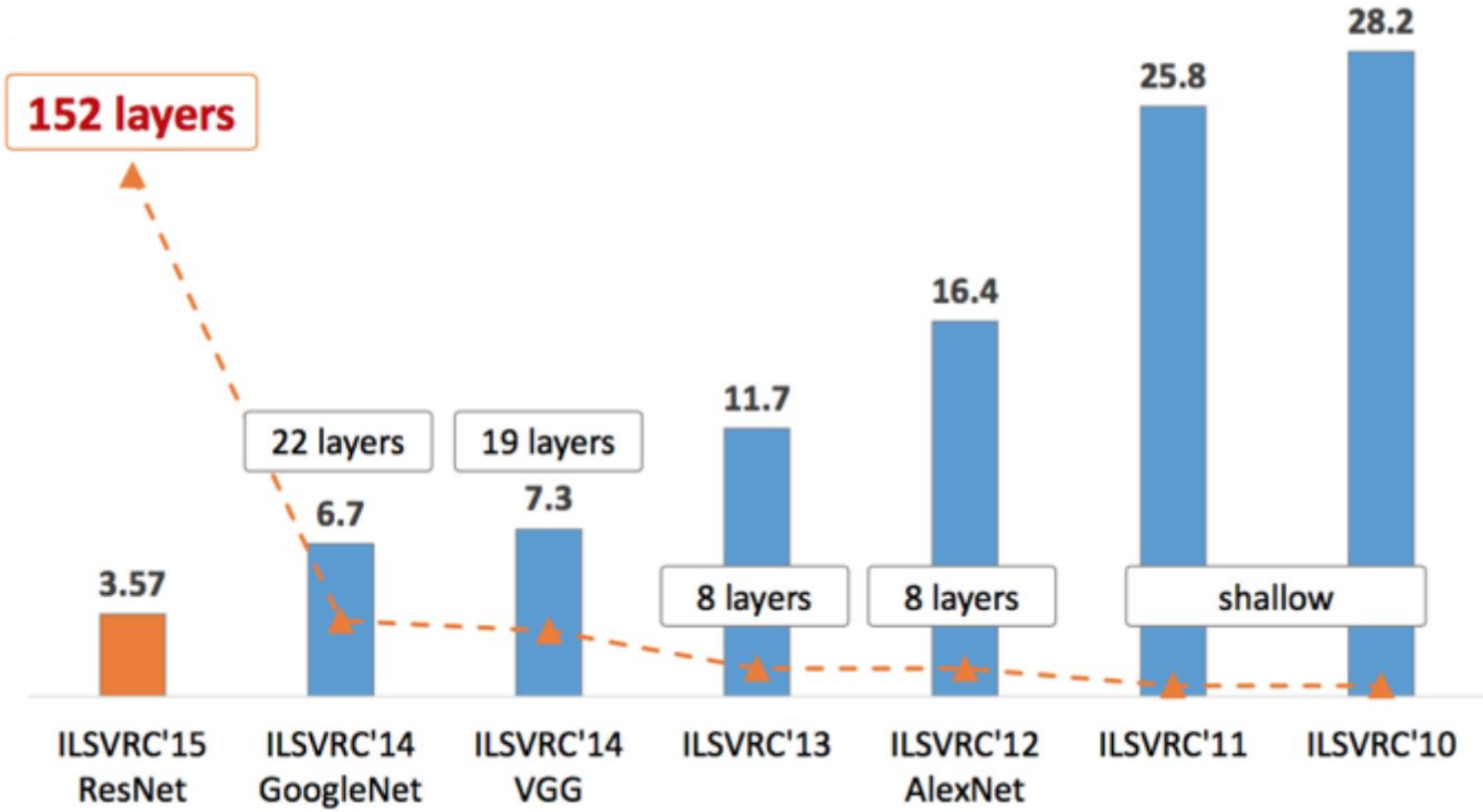
- Depth is really important.

  removing a single convolutional layer degrades the performance.

  K. Simonyan, A. Zisserman.
  Very Deep Convolutional Networks for Large-Scale Image Recognition. Technical report, 2014.

  → 16-layer model, 19-layer model. 7.3% top-5 test error on ILSVRC-2012

Going deep and deeper…….



Questions?