Building High-level Features Using Large Scale Unsupervised Learning

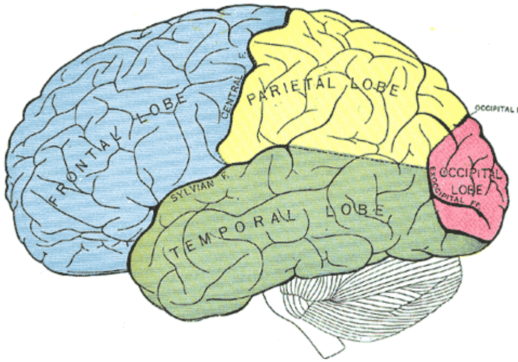
Machine Learning Seminar

Overview

- Motivation and objective
- Training set
- Algorithm
- Test set
- Performance
- Visualization
- Other high level concepts
- Summary

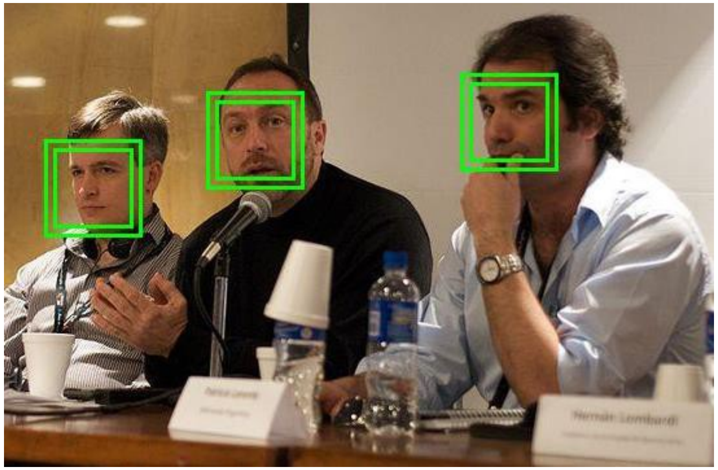
Motivation

- Neuroscientic conjecture:
  - highly class-specific neurons in the human brain
  - „grandmother neurons"
  - some neurons in the temporal cortex are highly selective for object categories such as faces or hands



Computer Vision

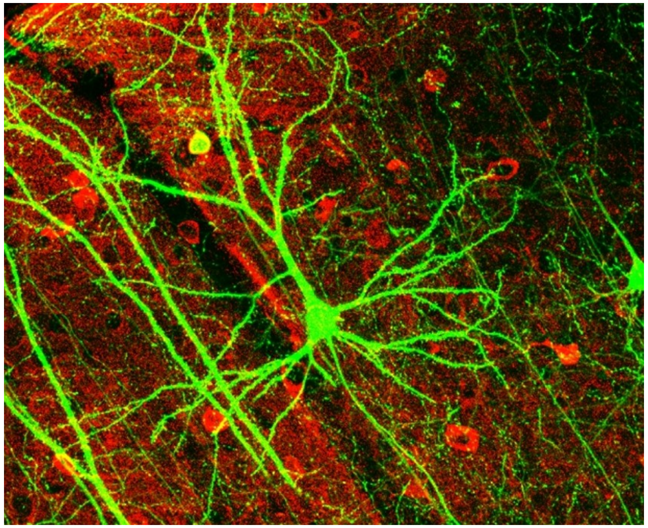
- Contemporary computer vision methodology typically:
  - emphasizes the role of labeled data
  - e.g. a large collection of labeled images to build a face detector
  - labeled data are rare for many problems



Objective

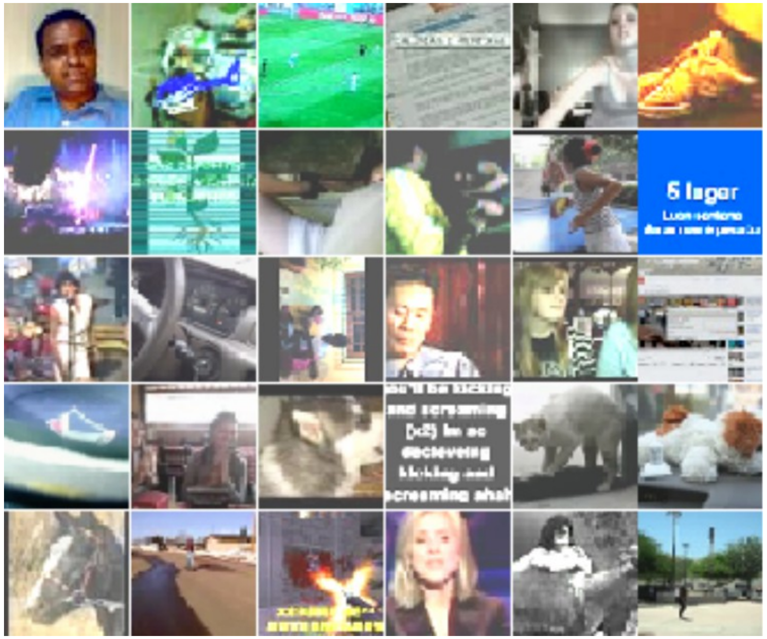
TECHNISCHE UNIVERSITÄT DARMSTADT

- This work investigates:
  - possibility to learn „grandmother neuron" from unlabeled data
  - feasibility of building high-level features from only unlabeled data



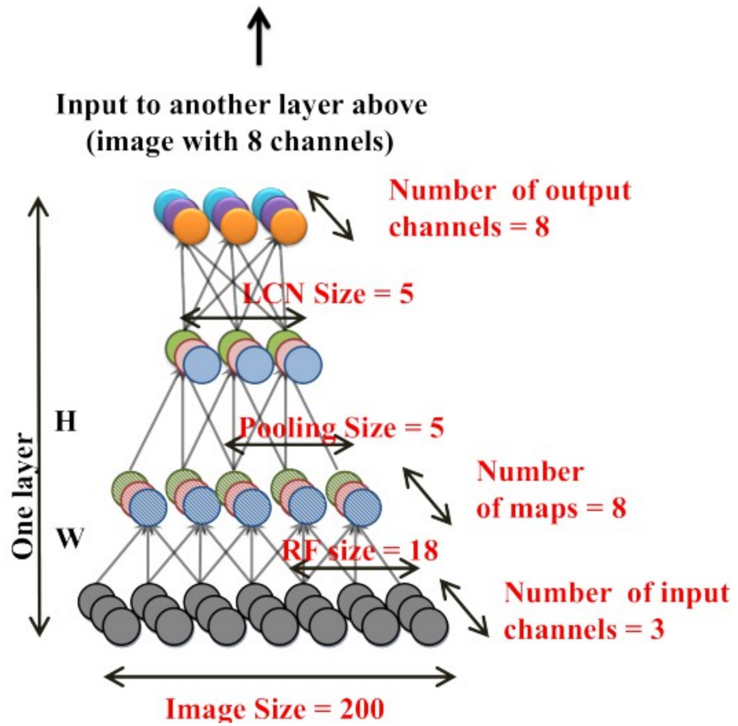
Training set

- Constructed by sampling frames from 10 million YouTube videos
- Each video contributes only one image to the dataset
- Each example is a color image with 200x200 pixels



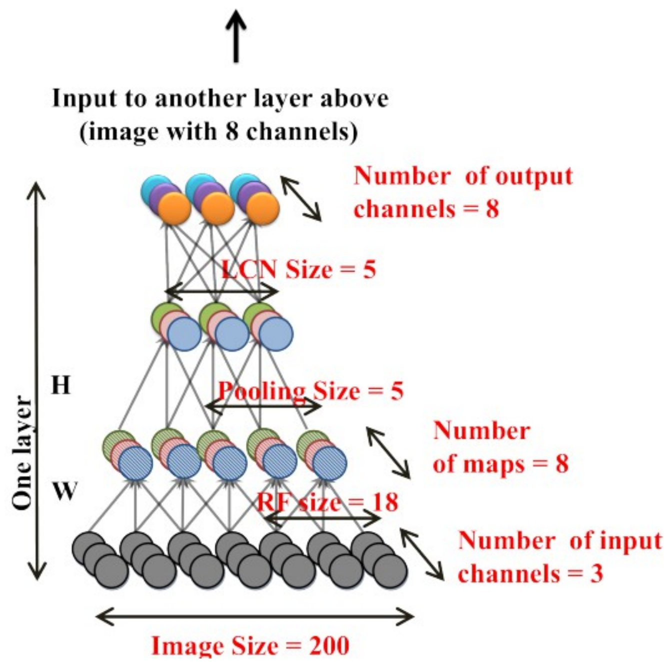
Algorithm

- Sparse deep autoencoder with three important ingredients:
  - local receptive fields
  - pooling
  - local contrast normalization

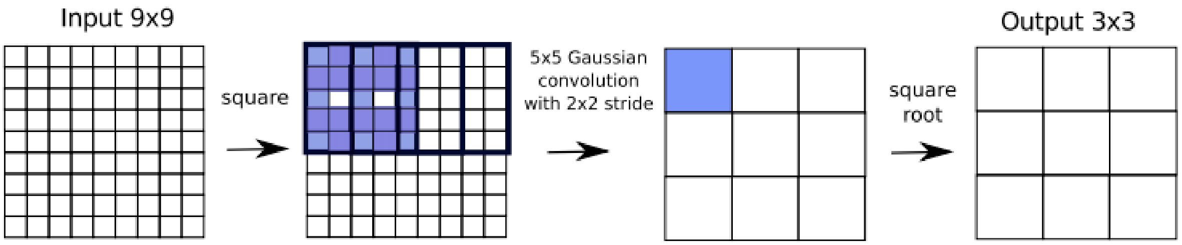


Local receptive fields

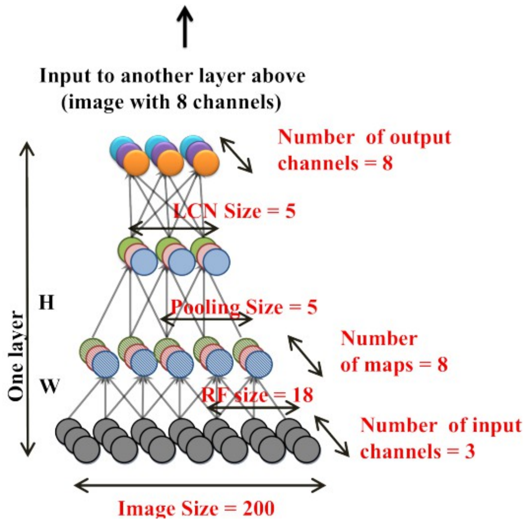
- Hubel and Wiesel's discovery of neurons in the cat's visual system (goes back to the early 60s)
- Neurons can learn to extract elementary visual features
- Different sets of units can be forced to have identical weight vectors
- Feature map: units in planes that share the same set of weights



Lp pooling



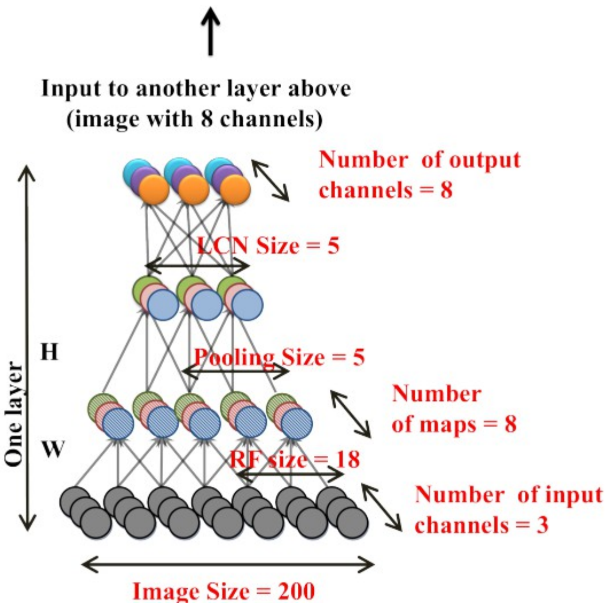
- Learning of invariant features
- $O = (\sum \sum I(i,j)^P \times G(i,j))^{1/P}$
- G: Gaussian kernel
  I: the input feature map
  O: the output feature map
- giving an increased weight to stronger features and suppressing weaker features

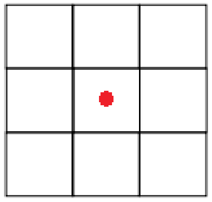


Local contrast normalization

- For each unit in the 3rd sublaeyr:
  - subtract the mean of the unit values in a fixed window (3x3 units, centered on the unit)
  - if euclidean norm of the resulting 9-dimensional vector greater than 1
    - divide this value by euclidean norm: $\|\mathbf{p}\| = \sqrt{p_1^2 + p_2^2 + \cdots + p_n^2}$

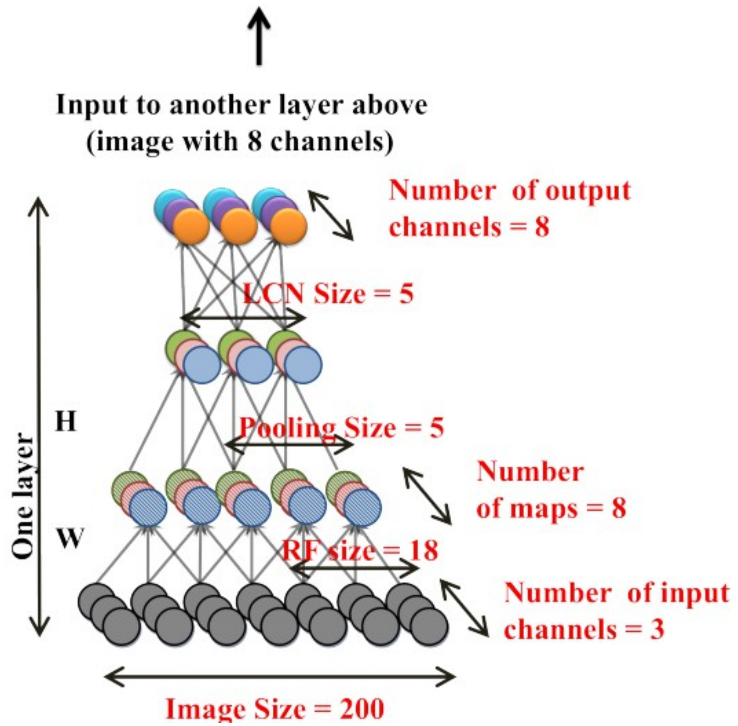
- Normalization can reduce responses, but not enhance them





Algorithm

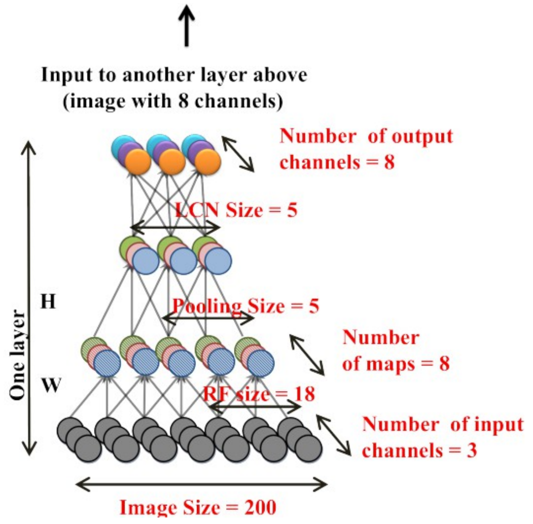
- Replicating three times the same stage composed of:
  - local receptive fields
  - local pooling
  - local contrast normalization
- The output of one stage is the input to the next one
- The overall model can be interpreted as a nine-layered network



Optimization

- The parameters of the second sublayers (H) fixed to uniform weights
- Encoding weights W_1 and decoding weights W_2 are adjusted

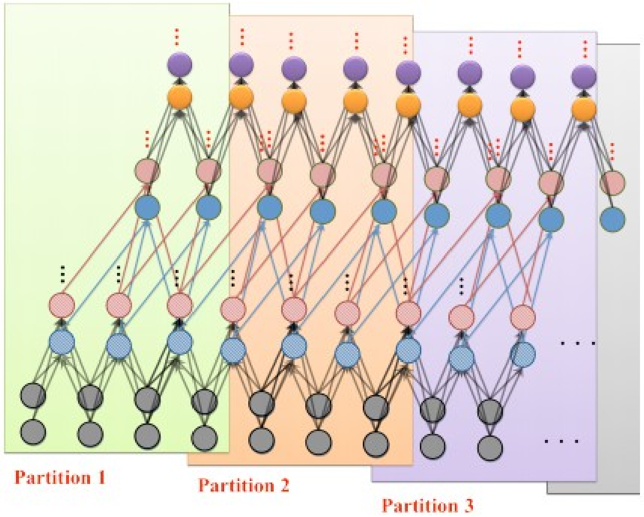
- m, k are the number of examples and pooling units in a layer respectively
- H_j is the vector of weights of the j-th pooling unit
- $\lambda = 0.1$



$$\underset{W_1,W_2}{\text{minimize}} \quad \sum_{i=1}^{m} \Big(\|W_2 W_1^T x^{(i)} - x^{(i)}\|_2^2 + \lambda \sum_{j=1}^{k} \sqrt{\epsilon + H_j (W_1^T x^{(i)})^2} \Big).$$

Model Training

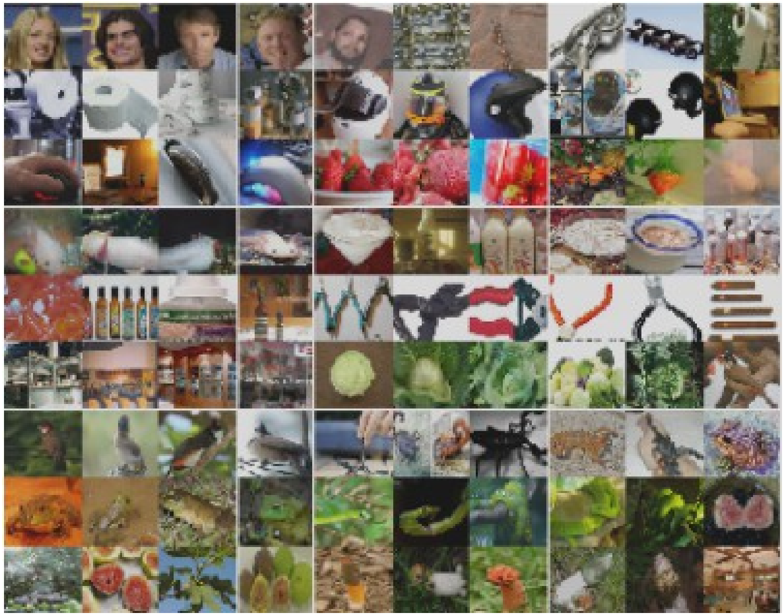
- Model parallelism by distributing the local weights W_1, W_2 and H to different machines
- Weights are divided according to the locality of the image and stored on different machines
- The network trained on a cluster with 1,000 machines for three days



$$\underset{W_1, W_2}{\text{minimize}} \quad \sum_{i=1}^{m} \left(\left\| W_2 W_1^T x^{(i)} - x^{(i)} \right\|_2^2 + \lambda \sum_{j=1}^{k} \sqrt{\epsilon + H_j (W_1^T x^{(i)})^2} \right).$$

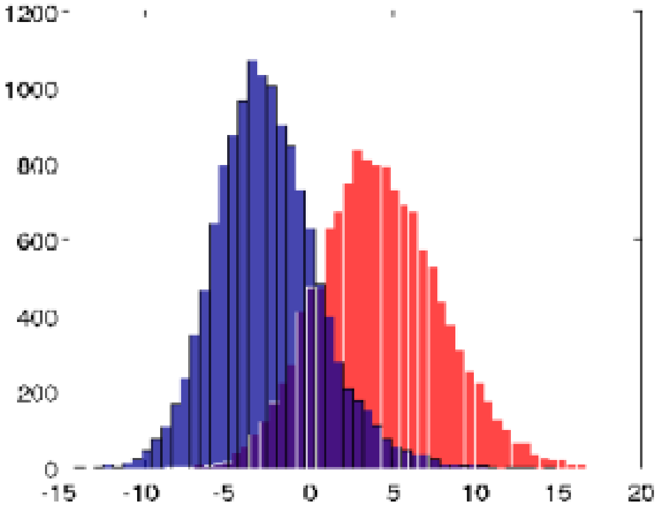
Test set

- 37,000 images sampled from two datasets:
  - Labeled Faces In the Wild
  - ImageNet
- 13,026 faces sampled from non-aligned Labeled Faces in The Wild
- The rest are distractor objects randomly sampled from ImageNet



Measure the performance

- For each neuron:
  - find its maximum and minimum activation values
  - pick 20 equally spaced thresholds in between
  - take the best classication accuracy among 20 thresholds
- The best neuron achieves 81.7% accuracy in detecting faces



Histograms of faces (red) vs. no faces (blue)

Visualization

- Two visualization techniques to verify if the optimal stimulus of the neuron is indeed a face:
  - visualizing the most responsive stimuli in the test set
  - perform numerical optimization

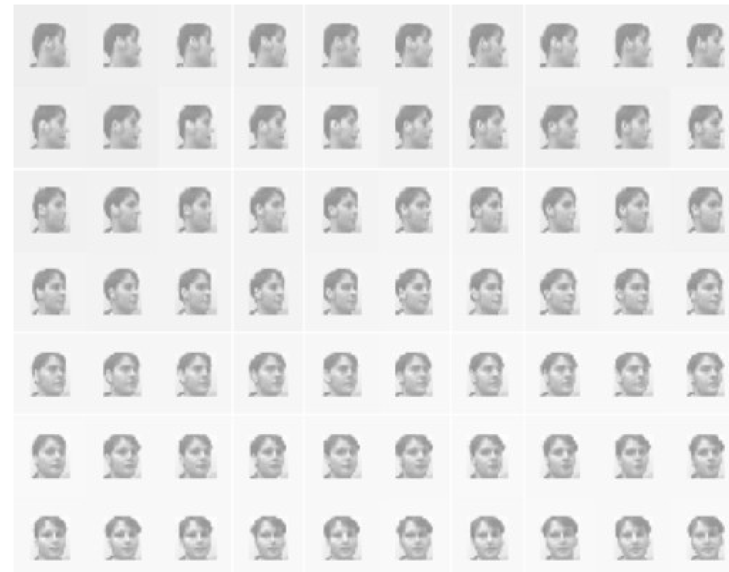
- Top: top 48 stimuli of the best neuron from the test set

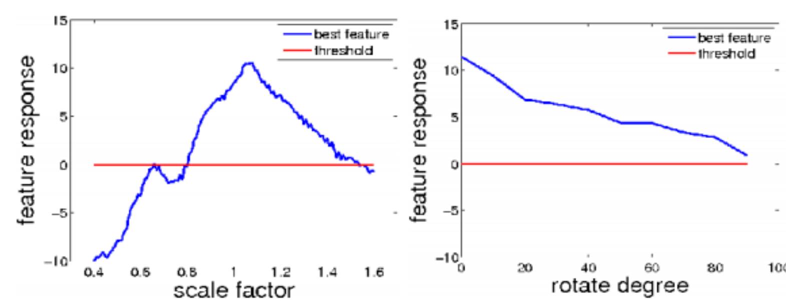
- Bottom: the optimal stimulus according to numerical constraint optimization

Robustness

- Robustness of the face detector against common object transformations:
  - scaling
  - out-of-plane
  - translation

- Results show that the neuron is robust against complex and difficult transformations





Comparison with state-of-the-art baselines

Dataset version	2009 (~9M images, ~10K categories)	2011 (~14M images, ~22K categories)
State-of-the-art	16.7% (Sanchez & Perronnin, 2011)	9.3% (Weston et al., 2011)
Our method	16.1% (without unsupervised pretraining) **19.2%** (with unsupervised pretraining)	13.6% (without unsupervised pretraining) **15.8%** (with unsupervised pretraining)

- „Unsupervised pretraining“:
  - learn features using described techniques
  - add one-versus-all logistic classifiers on top
- 70% relative improvement over the highest other result on ImageNet
- Random guess achieves less than 0.005% accuracy on ImageNet (22K categories)

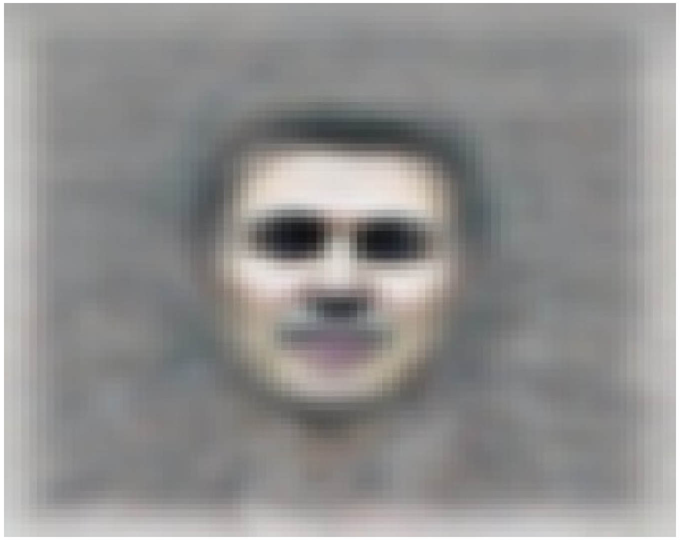
Cat and human body detectors

Concept	Our network	Deep autoencoders 3 layers	Deep autoencoders 6 layers	K-means on 40x40 images
Faces	**81.7%**	72.3%	70.9%	72.5%
Human bodies	**76.7%**	71.2%	69.8%	69.3%
Cats	**74.8%**	67.5%	68.3%	68.5%

- Is the network able to detect other high-level concepts
- Cats and body parts are quite common in YouTube
- Construct two new datasets:
  - human bodies against random backgrounds
  - cat faces against other random distractors

Summary

- We have seen that it is possible:
  - to learn „grandmother neuron" from unlabeled data
  - to build high-level features from only unlabeled data



Thanks

For your good

investigated time! :-)

References

- Quoc V. Le, Marc' A. Ranzato, R. Monga, M. Devin, K. Chen, Greg S. Corrado, J. Dean, Andrew Y. Ng. 2012
  Building High-level Features Using Large Scale Unsupervised Learning
- R. Quian Quiroga, L. Reddy, C. Koch, and I. Fried. 2007
  Decoding Visual Inputs From Multiple Neurons in the Human Temporal Lobe
- I. Arel, Derek C. Rose, and Thomas P. Karnowski. November 2010
  Deep Machine Learning – A New Frontier in Artificial Intelligence Research
- G. Hinton. October 26, 2006
  To Recognize Shapes, First Learn to Generate Images
- Bruno A. Olshausen & David J. Field. June 1996
  Emergence of simple-cell receptive field properties by learning a sparse code for natural images
- Andrew Ng (http://www.stanford.edu/class/cs294a/sparseAutoencoder_2011new.pdf)
  CS294A Lecture notes. Sparse autoencoder
- Y. LeCun, P. Haffner, L. Bottou and Y. Bengio.
  Object Recognition with Gradient-Based Learning
- Aapo Hyvärinen. 2009
  Statistical Models of Natural Images and Cortical Visual Representation
- P. Sermanet, S. Chintala and Y. LeCun. 2012
  Convolutional Neural Networks Applied to House Numbers Digit Classification
- Nicolas Pinto, David D. Cox, James J. DiCarlo. 2008
  Why is real-world visual object recognition hard?

Sparseness

- A random variable takes very small absolute values and very large values
- More often than a Gaussian random variable of the same variance
- To compensate: it takes values in between relatively more rarely
- The random variable is “activated” (significantly non-zero) only rarely