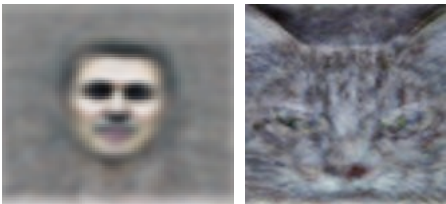


# Neural Nets

## Neural Nets



ML watches YouTube for three straight days!  
(and learns to recognize cats)

<http://www.npr.org/2012/06/26/155792609/a-massive-google-network-learns-to-identify>  
**Building High-level Features Using Large Scale Unsupervised Learning**  
 Quoc V. Le, Marc Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeffrey Dean, and Andrew Y. Ng

## State-of-the-Art Image Classification

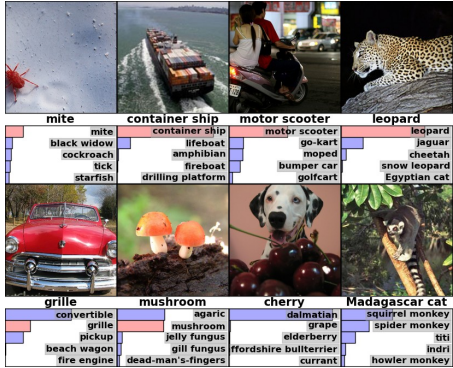
---

### ImageNet Classification with Deep Convolutional Neural Networks

---

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
---	--	--

## State-of-the-art Image Classification



## Digit Recognition

Handwritten digit recognition

0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9

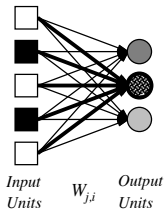
3-nearest-neighbor = 2.4% error  
 400-300-10 unit MLP = 1.6% error  
 LeNet: 768-192-30-10 unit MLP = 0.9% error

## What is a Neural Network?

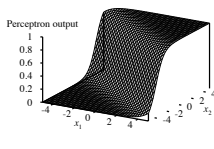
- Biological view: models neurons in the brain
- Mathematical view
  - Flexible parametric class of non-linear functions
  - Compose many linear/logistic regression models
  - Easy to compute
    - $h(x)$ : “feed-forward”
    - Partial derivatives: “backward propagation”
- **Develop on board**

### Hypothesis Class

**Single-layer perceptrons**



Input Units      $W_{ji}$      Output Units

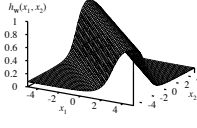


Perceptron output

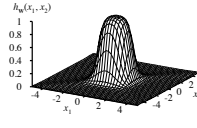
### Hypothesis Class

**Expressiveness of MLPs**

All continuous functions w/ 2 layers, all functions w/ 3 layers




2 layers




3 layers

### Learning in Neural Networks

- “Backprop” + “Stochastic Gradient Descent”



Neural net specific



Generally useful!

### Stochastic Gradient Descent

- Like gradient descent, but training examples treated one-by-one
- **Develop on board**

### SGD Discussion

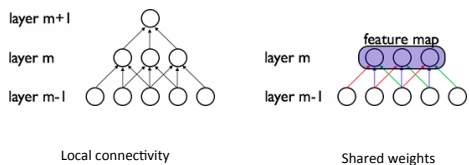
- Simple
- Memory efficient
  - Applies to huge data sets
- Online
  - Trivial to add new examples, or get rid of old ones
- Can solve huge fraction of ML problems

Pillar of **large-scale** machine learning

### Backprop

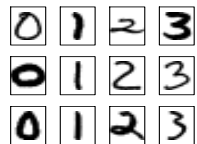
- Backprop = chain rule for partial derivatives
  - Works out nicely for “feed-forward” neural nets
  - Compute function: forward pass through network
  - Compute derivatives: backward pass
- **Develop on board**

### Architectures

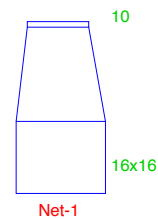


<http://www.deeplearning.net/tutorial/>

### Digits Example

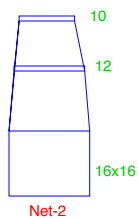


From [Hastie, Tibshirani, Friedman]



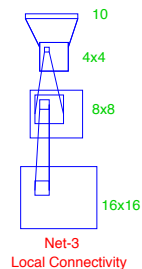
256 inputs, 10 outputs:  
multiclass logistic regression

### Digits Example: Net 2



Add hidden layer with 12 units

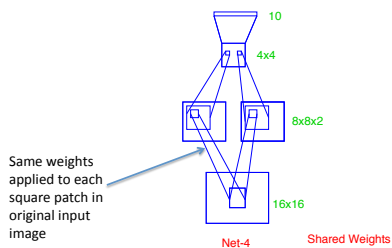
### Digits Example: Net 3



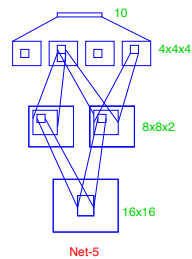
Local Connectivity

Two hidden layers, each hidden unit summarizes a small square patch of input image

### Digits Example: Net 4

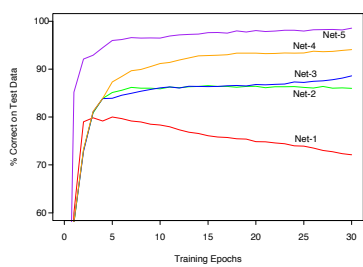


### Digits Example: Net 5



Even fancier shared weights

## Digits Example



320 train / 120 test