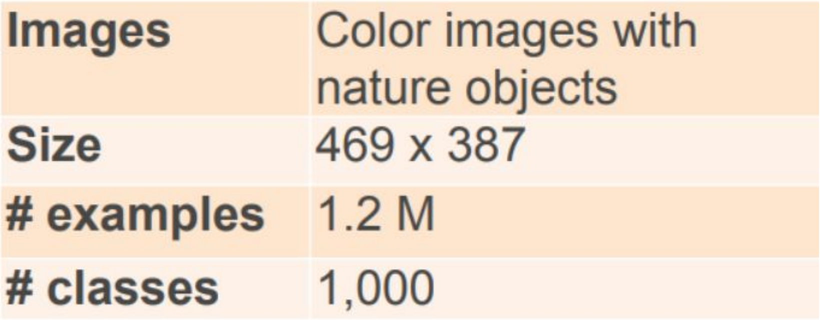
ImageNet (AlexNet) Classification with Deep Convolutional Neural Networks

Presenter: Shuozhe Li

08/30/2022

ImageNet Large Scale Visual Recognition Challenge (ILSVRC 2012)

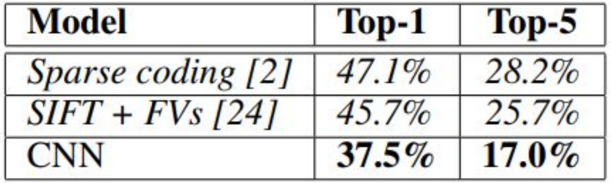
- **What do we do with AlexNet?**



Images	Color images with nature objects
Size	469 x 387
# examples	1.2 M
# classes	1,000

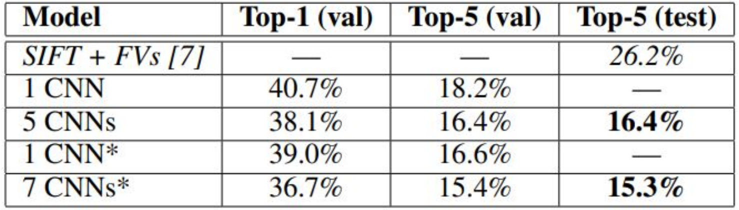
Experimental Results

- **ILSVRC-2010 Test Set Error Rate**



Model	Top-1	Top-5
Sparse coding [2]	*47.1%*	*28.2%*
SIFT + FVs [24]	*45.7%*	*25.7%*
CNN	**37.5%**	**17.0%**

- **Results on 2012 Validation and Testing Set**

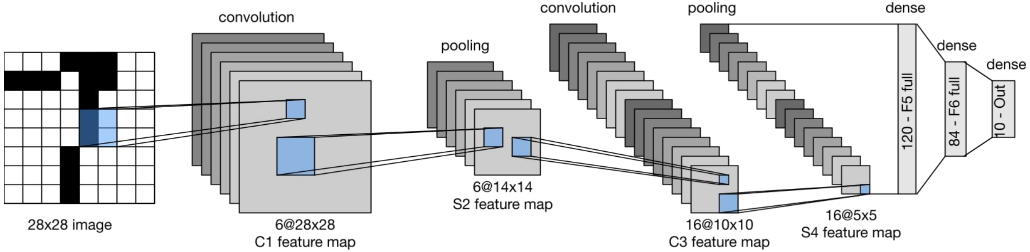


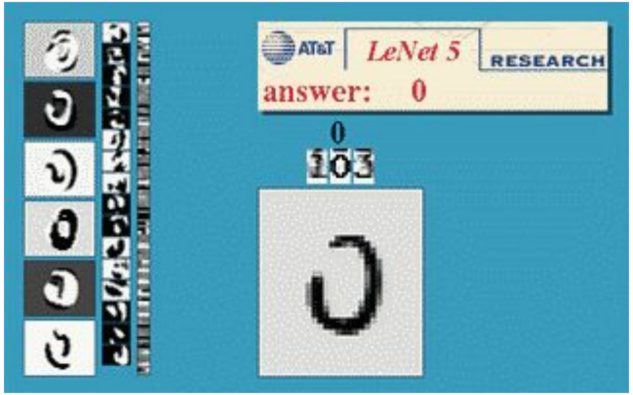
Model	Top-1 (val)	Top-5 (val)	Top-5 (test)
SIFT + FVs [7]	—	—	*26.2%*
1 CNN	40.7%	18.2%	—
5 CNNs	38.1%	16.4%	**16.4%**
1 CNN*	39.0%	16.6%	—
7 CNNs*	36.7%	15.4%	**15.3%**

- 5 CNNs: averaging 5 similar CNNs trained with ILSVRC-2010
- 1 CNN: six convolutional layers trained with ILSVRC-2011 and fine-tuning it on ILSVRC-2012
- 7 CNNs: averaging two 1 CNN trained with ILSVRC-2011 and 5 CNNs trained with ILSVRC-2010

Historical Background

- **LeNet (for image classification, LeCun et al. in 1989)**
  - Trained on MNIST (60,000, 28×28 handwritten number digit)
  - One of earliest largely deployed CNN in Post Service and Banks

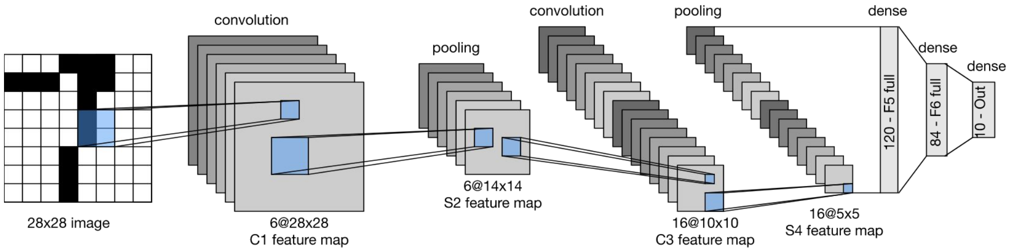




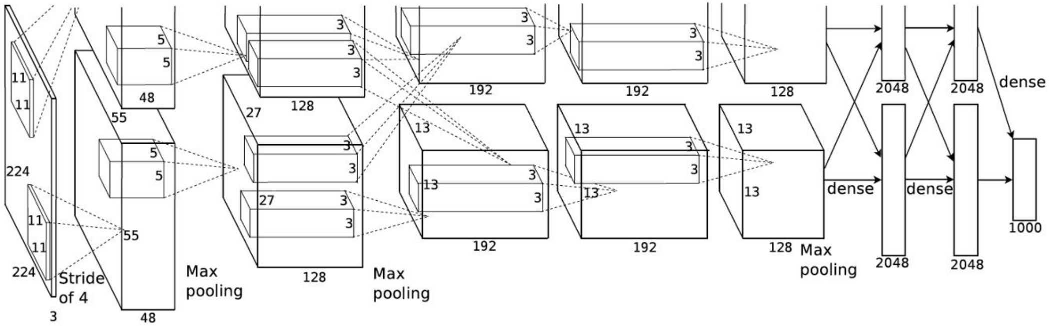
Historical Background

- **Why does it take so long for AlexNet to come?**

LeCun et al. in 1989



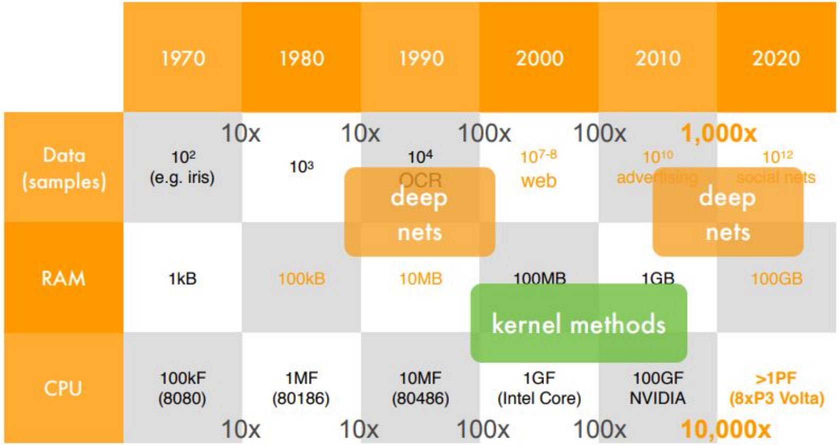
Krizhevsky et al. in 2012



Historical Background

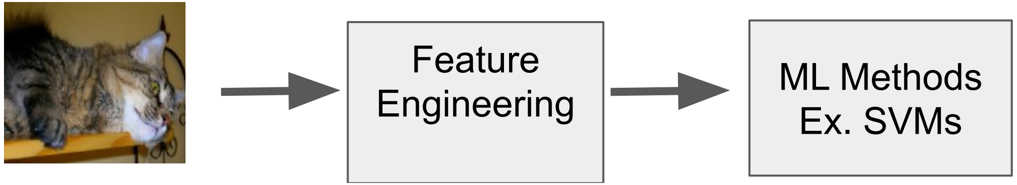
- Why does it take so long for AlexNet to come?

Images	Color images with nature objects	Gray image for hand-written digits
Size	469 x 387	28 x 28
# examples	1.2 M	60 K
# classes	1,000	10

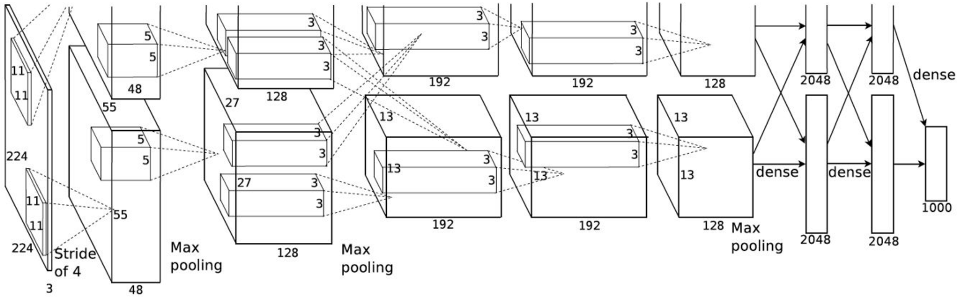


	1970	1980	1990	2000	2010	2020
	10x	10x	100x	100x	1,000x	
Data (samples)	10^2 (e.g. iris)	10^3	10^4 OCR	$10^{7\text{-}8}$ web	10^{10} advertising	10^{12} social nets
RAM	1kB	100kB	10MB	100MB	1GB	100GB
CPU	100kF (8080)	1MF (80186)	10MF (80486)	1GF (Intel Core)	100GF NVIDIA	>1PF (8xP3 Volta)
	10x	10x	100x	100x	10,000x	

deep nets
kernel methods
deep nets

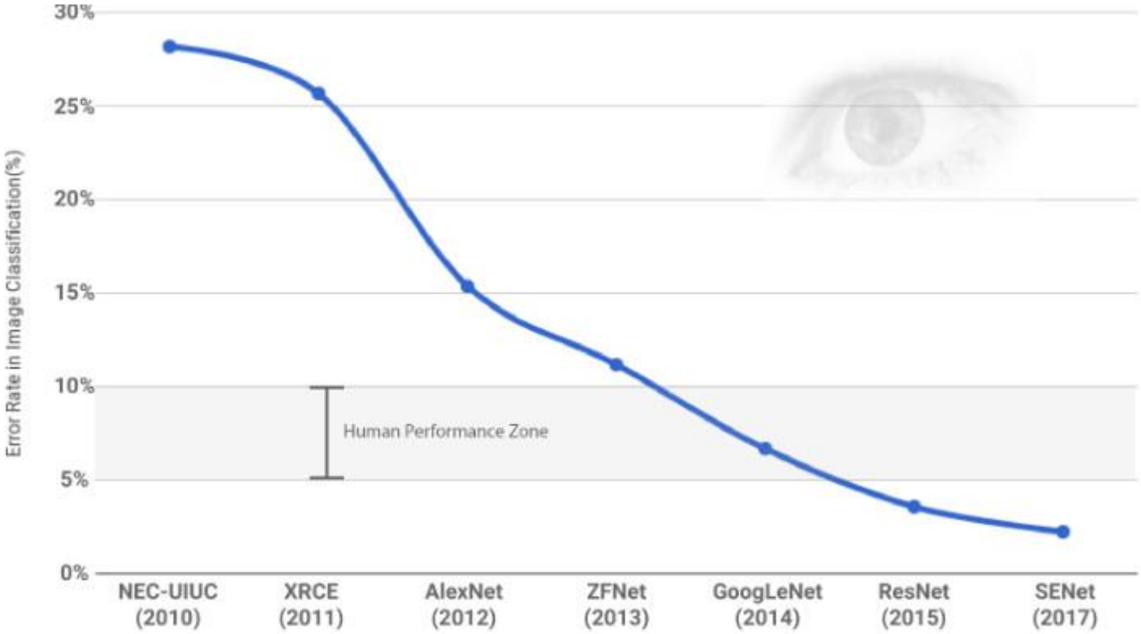

Feature Engineering

ML Methods Ex. SVMs



Historical Background

- **AlexNet, the Game Changer**



Implementation Details

- **Pre-Process:** Rescaled to 256x256
- **Data Augmentation:**

During Training:

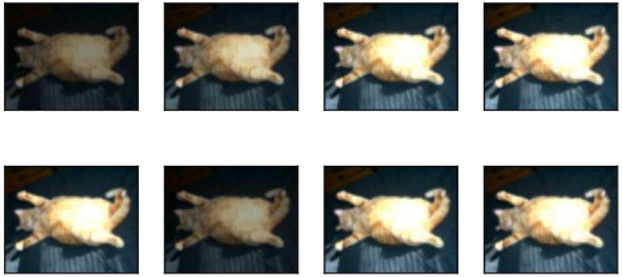
1. Random cropping to 224x224 with Horizontal Reflections (enlarged by a scale factor of 2048)
2. Random changing to the intensity and color of the illumination

$$I_{xy} = [I_{xy}^R, I_{xy}^G, I_{xy}^B]^T + [\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3][\alpha_1\lambda_1, \alpha_2\lambda_2, \alpha_3\lambda_3]^T$$

I_{xy} :RGB image pixel

$\mathbf{p}_i, \lambda_i$:ith eigenvector and eigenvalue by PCA

α_i :Drawn from *N*(avg=0, *∂*=0.1)

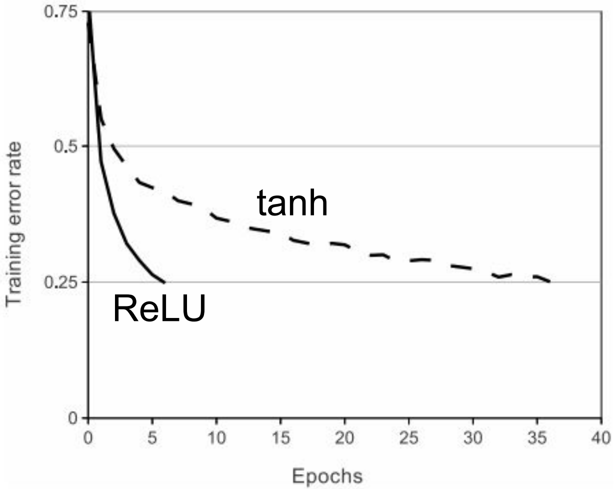


During Testing:

Apply Horizontal Reflection and 4 corner and 1 center

Cropping, averaging the predictions of ten patches from softmax

Network Architecture



- **ReLU Nonlinearity:**
  - shorter training time with gradient descent than tanh() or sigmoid
- **Local Response Normalization:**
  - constrain ReLU output within a bounded range
  - lateral inhibition: carry out local contrast enhancement for the next layers

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

$\alpha^i_{x,y}$:activity of a neuron

N :number of channel

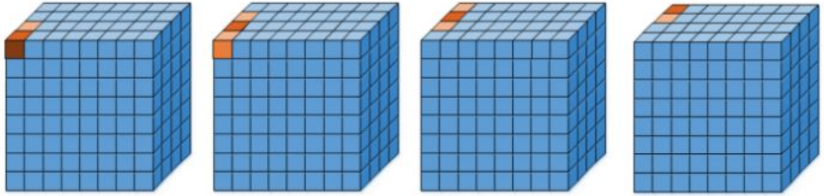
n :neighborhood length

i :index of center neuron

k :avoid division by zero

α:normalization constant

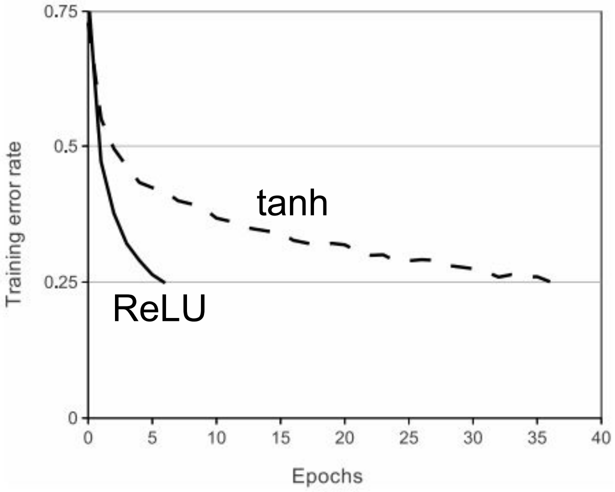
β :contrasting constant



a) Inter-Channel LRN (n=2)

Network Architecture

- **ReLU Nonlinearity:**
  - shorter training time with gradient descent than tanh() or sigmoid
- **Local Response Normalization:**
  - ~~constrain ReLU output within a bounded range~~
  - ~~lateral inhibition: carry out local contrast enhancement for the next layers~~

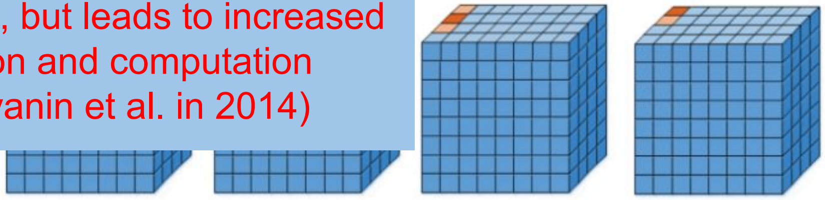


$$b^i_{x,y} = a^i_{x,y} / \left(k + \alpha \sum_{j=\max(0,i}^{\min(N-1,i} \right.$$

~~$\alpha^i_{x,y}$:activity of a neuron~~
~~N :number of channel~~
~~n :neighborhood length~~

~~i :inde~~
~~k :avoid division by zero~~
~~α:normalization constant~~
~~β :contrasting constant~~

LRN "does not improve the performance on the ILSVRC dataset, but leads to increased memory consumption and computation time" (VGG, Simonyanin et al. in 2014)



a) Inter-Channel LRN (n=2)

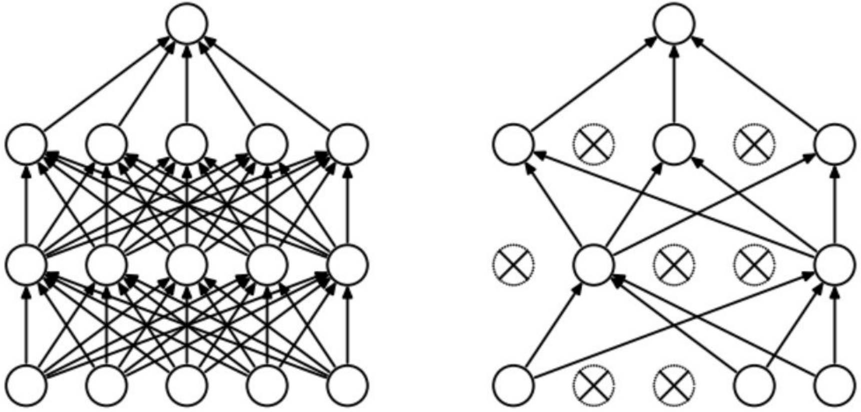
Network Architecture

- **Overlapping MaxPooling (LeNet: non-overlapping average pooling):**
  - introducing overlap maxpoling to prevent overfitting
- **Dropout:**
  - combining the predictions of many different models

$$x = \begin{cases} 0 & \textit{with probability } 0.5 \\ 0.5x & \textit{otherise} \end{cases}$$

↓

$$x = \begin{cases} 0 & \textit{with probability } p \\ \dfrac{x}{1-p} & \textit{otherise} \end{cases}$$



(a) Standard Neural Net

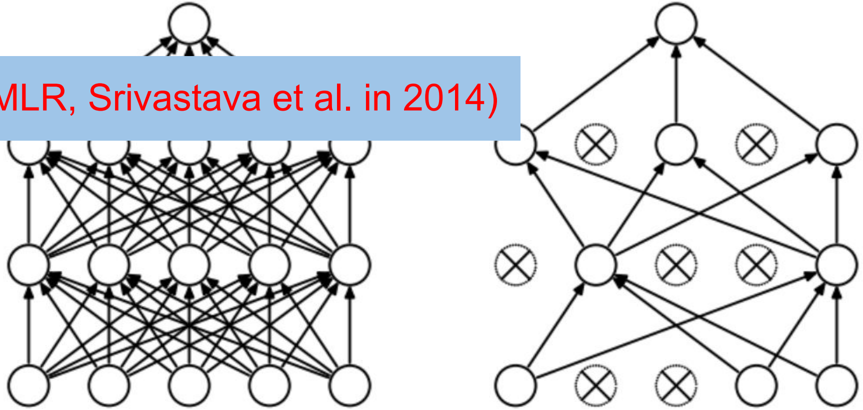
(b) After applying dropout.

Network Architecture

- **Overlapping Pooling:**
  - introducing overlap maxpoling to prevent overfit
- **Dropout:**
  - ~~combining the predictions of many different models~~

Dropout ”is a modified form of L2 regularization” (JMLR, Srivastava et al. in 2014)

$$x = \begin{cases} 0 & \textit{with probability } p \\ \dfrac{x}{1-p} & \textit{otherise} \end{cases}$$



(a) Standard Neural Net

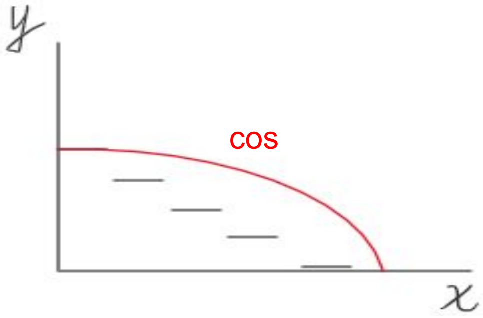
(b) After applying dropout.

Details of Training

- Stochastic Gradient Descent with batch size of 128, Momentum of 0.9, Weight Decay of 0.0005

$$
\begin{aligned}
v_{i+1} &:= 0.9 \cdot v_i - 0.0005 \cdot \epsilon \cdot w_i - \epsilon \cdot \left\langle \frac{\partial L}{\partial w}\Big|_{w_i} \right\rangle_{D_i} \\
w_{i+1} &:= w_i + v_{i+1}
\end{aligned}
$$

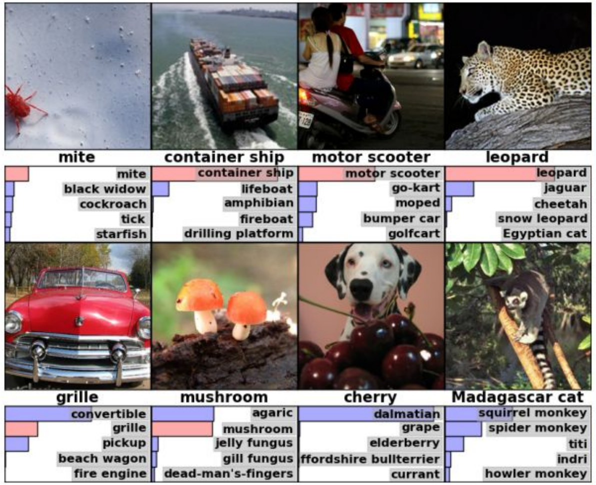
- Learning Rate initialized at 0.01 and divide by 10 when the validation error rate stopped improving
  - more techniques on adjusting learning rate
    Ex. Cosine or more specific Learning Rate Scheduler

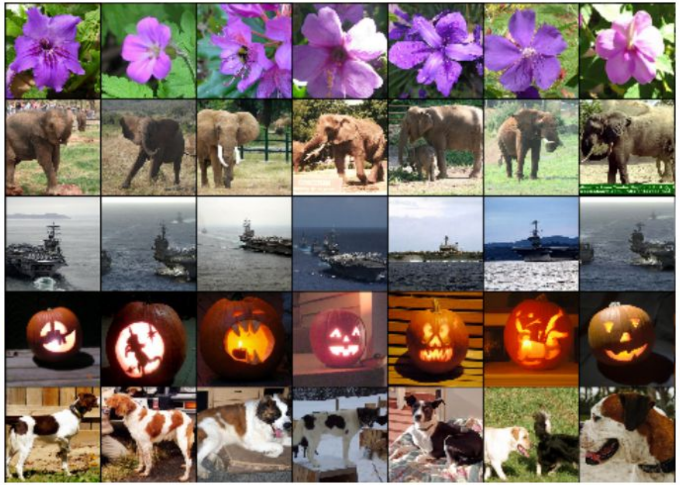


- Weights Initialization: set bias to 1 accelerates the early stage of learning (ReLU)
  - nowadays, does not really matter

Discussion of Results

- **Test Images from ILSVRC-2010**





Some test images and the five labels considered most probable by AlexNet

when comparing the output from the last fully connected layer, from right to the left, these are the six images that have small Euclidean separation to the first image

Summary

- **A variant from LeNet with techniques:** Dropout, ReLU, MaxPooling, Local Response Normalization
- **AlexNet won ImageNet classification challenge in 2012**
- **Changes the game of Computer Vision**

What Will Come After?

- **Can we build even deeper and wider CNNs to perform better？YES!**

  VGG: K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *CoRR*, vol. abs/1409.1556, 2014.

- **Can we use a similar network architecture on object localization and detection in image? YES!**

  (OverFeat) P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, “OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks.” 24-2014.

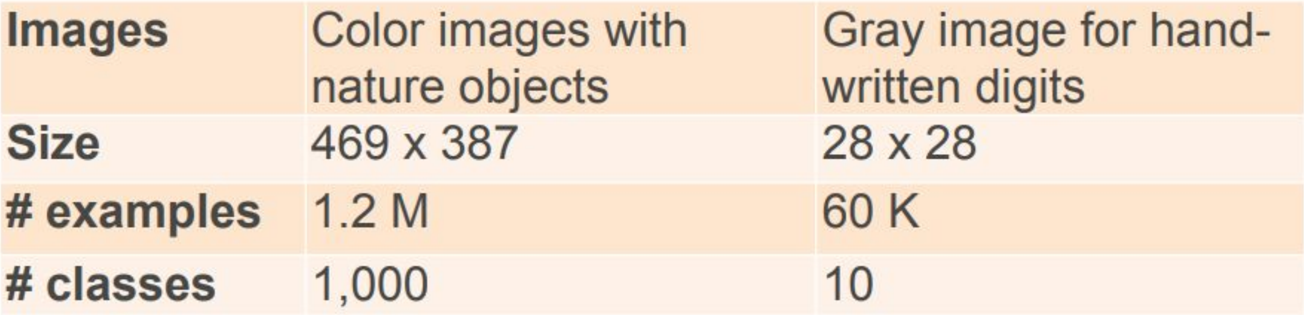
Thank You!

Shuozhe Li

ImageNet Large Scale Visual Recognition Challenge (ILSVRC 2012)

- **What do we do with AlexNet?**

MNIST



Images	Color images with nature objects	Gray image for hand-written digits
Size	469 x 387	28 x 28
# examples	1.2 M	60 K
# classes	1,000	10