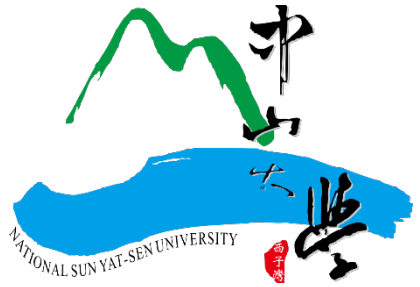
Deep Learning for Computer Vision



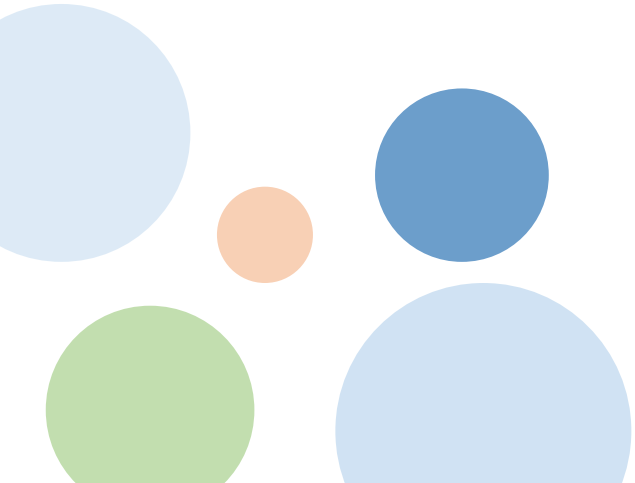
Outlines

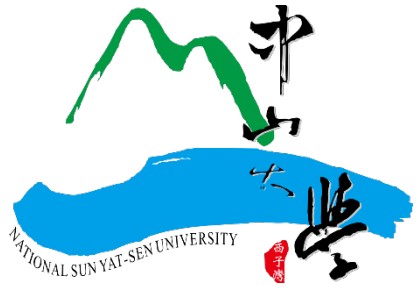
- Introduction
- Convolutional Neural Network (CNN)
  - Artificial Neural Network (ANN)
  - Convolutional Neural Network (CNN)
  - Training
- Benchmark CNN Models
  - CNN representative architectures
  - Recurrent Neural Network (RNN)
  - Generative Adversarial Network (GAN)
- Applications in Computer Vision
  - Object Detection
  - Segmentation
  - Image Captioning
  - Image-to-image Translation (style transfer, image enhancement, …)
  - Emotion Recognition
- Accelerators for CNN
  - Nvidia GPU (Graphics Processing Unit)
  - ASIC (Application Specific IC)

References and Credits

- Stanford CS231n, “Convolutional Neural Networks for Visual Recognition”
  - by Fei-Fei Li, Justin Johnson, and Serena Yeung
- MIT 6.S191, “Deep Learning”
  - by Alexander Amini, and Ava Soleimany
- UVA Deep Learning , Univ. of Amsterdam
  - by Efstratios Gavves
- CMSC 35264 Deep Learning, Univ. of Chicago
  - by Shubhendu Trivedi and Risi Kondor
- Deep Learning for Computer Vision
  - by 台大資工系 莊永裕 教授
- book: Deep Learning
  - by Ian Goodfellow, Yoshua Bengio, and Aaron Courville

Introduction

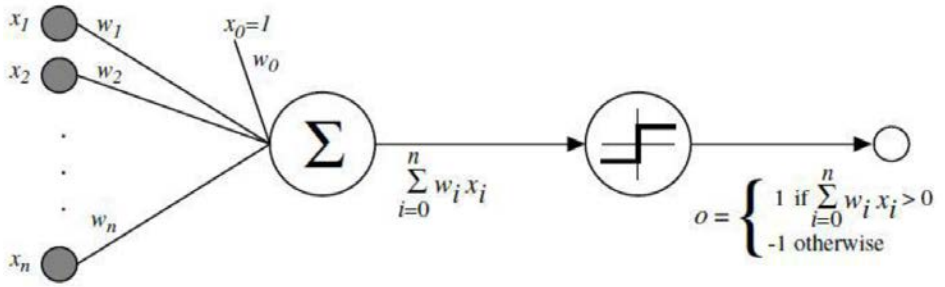




Outlines

- Evolution of AI
- Features in computer vision
- Deep Learning
- ILSVRC ImageNet image classification
- Applications
- Dataset

Evolution of AI



- first appeared in 1960s: perception for binary decision
- 1970~2000: backpropagation, recurrent nets with few layers
- 2005~: deep learning with many layers

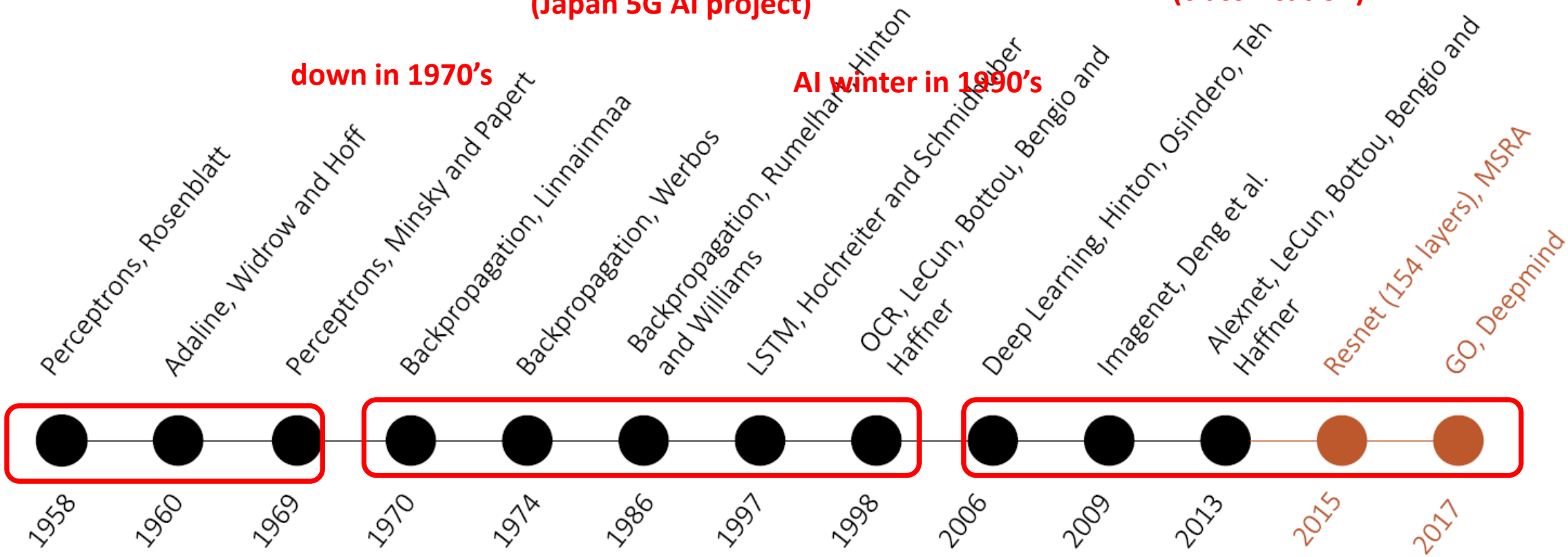
hot in 1960's

up in 1980's (Japan 5G AI project)

hot again in 2010's (classification)

down in 1970's

AI winter in 1990's



Three Waves of AI

- 1st wave: Handcrafted Knowledge (1980's)
  - create sets of **rules** to represent knowledge in well-defined domain (expert system)
  - no learning capability
- 2nd wave: Statistical Learnings (2010's)
  - create **statistical models** for specific problem domains and train them on big data
  - nuanced classification and prediction capability
  - limited explainable capability through visualization
  - no contextual capability
- 3rd wave: Contextual Adaptation (2020's)
  - construct **contextual explanatory** models
  - Artificial General Intelligence (AGI)
  - DARPA AI-Next projects in late 2018

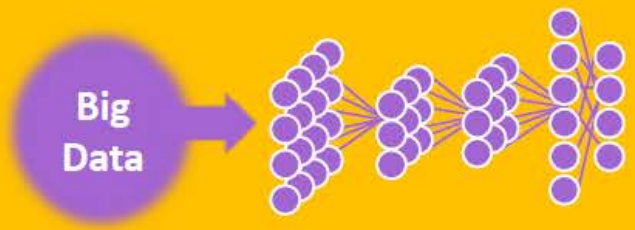
Three Waves of AI

Handcrafted Knowledge	Statistical Learning	Contextual Adaptation
First Wave	Second Wave	Third Wave
Expert Systems	Deep Learning, CNN, RNN, GAN, Explainable AI	Symbiosis AI, AGI*

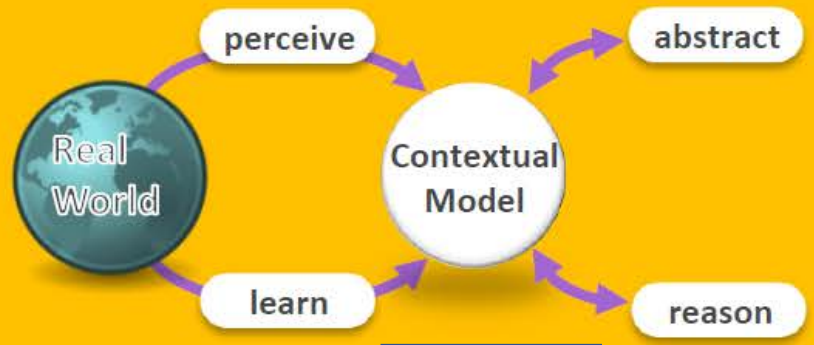
* Artificial General Intelligence

Human create
sets of rules
to represent knowledge
in well-defined domains

Human create
statistical models
for specific problem domains
and train them on big data



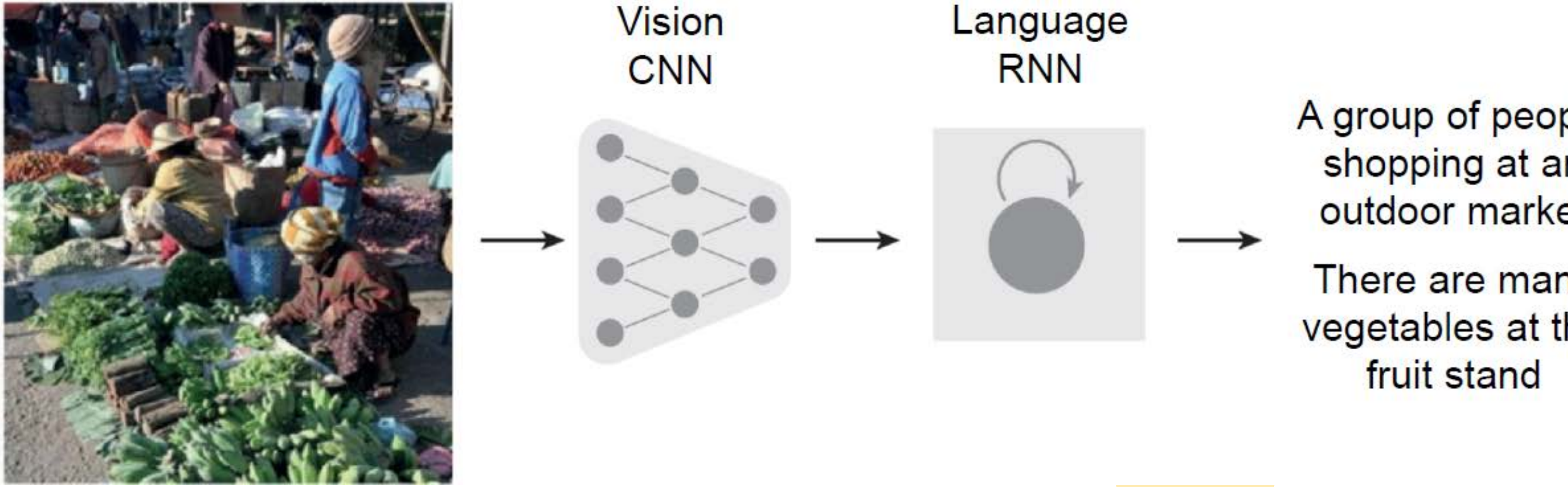
Systems construct
contextual explanatory models
for classes of
real world phenomena



Here we are!

Image Captioning Example

- a deep convolution neural net (CNN) produces a set of “words”
- a language-generating recurrent neural net (RNN) “translate” the words into captions



Challenges with 2nd Wave AI

- Generative Adversarial Network (GAN) 生成對抗網路

- image captioning and age estimation

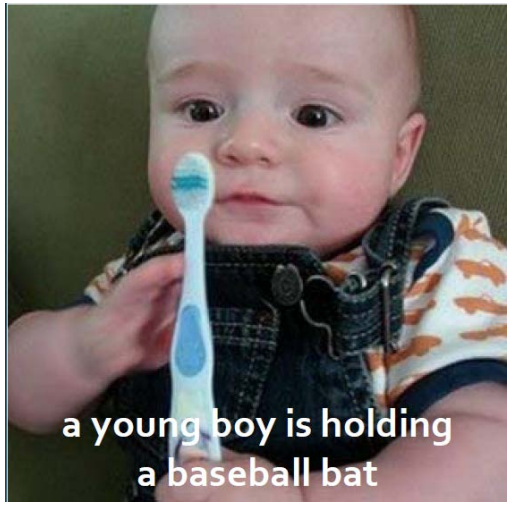
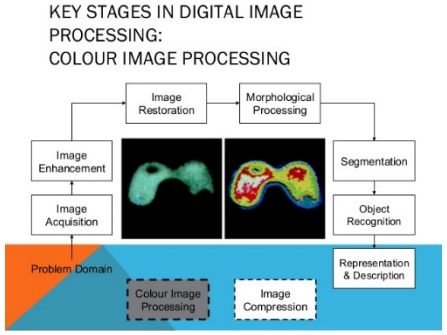
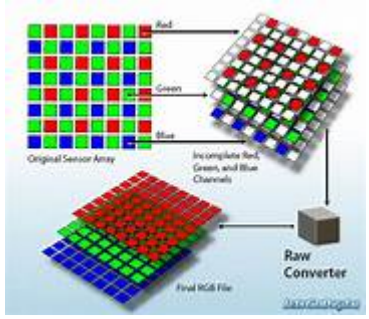
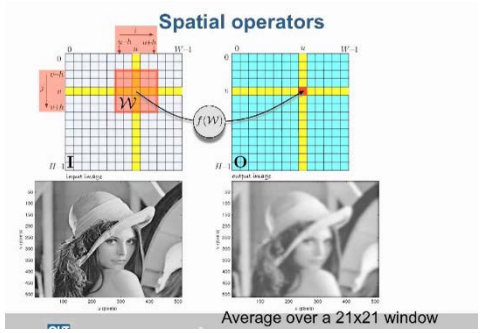


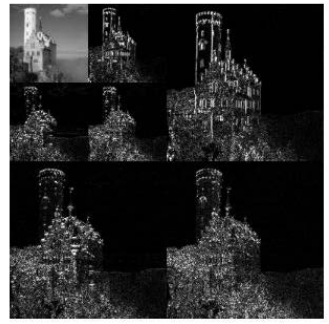
Image Processing vs. Computer Vision

- Image Processing
  - use of computer algorithms to perform image processing on digitalimages
  - e.g., filtering, sharpening, contrast, compression, image editing, restoration…

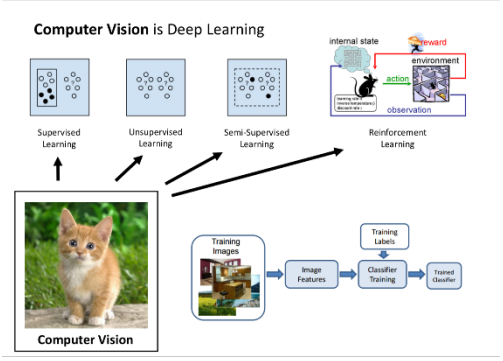


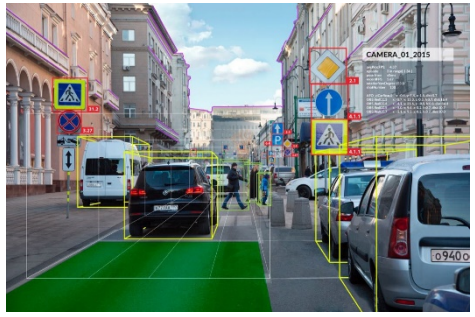


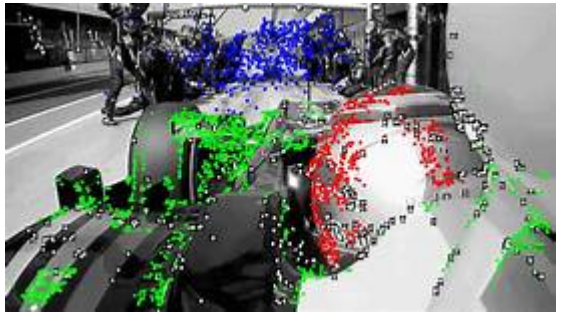


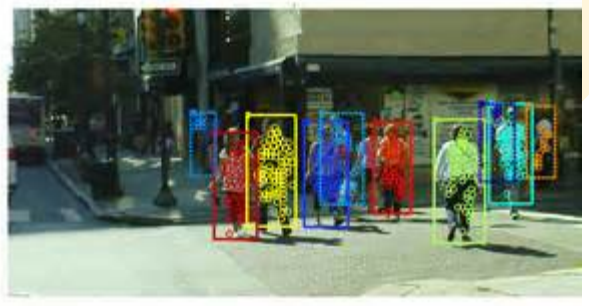


- Computer Vision
  - use of computer algorithms to gain high-level understanding from images or videos
  - e.g., classification, pattern recognition, image understanding, …



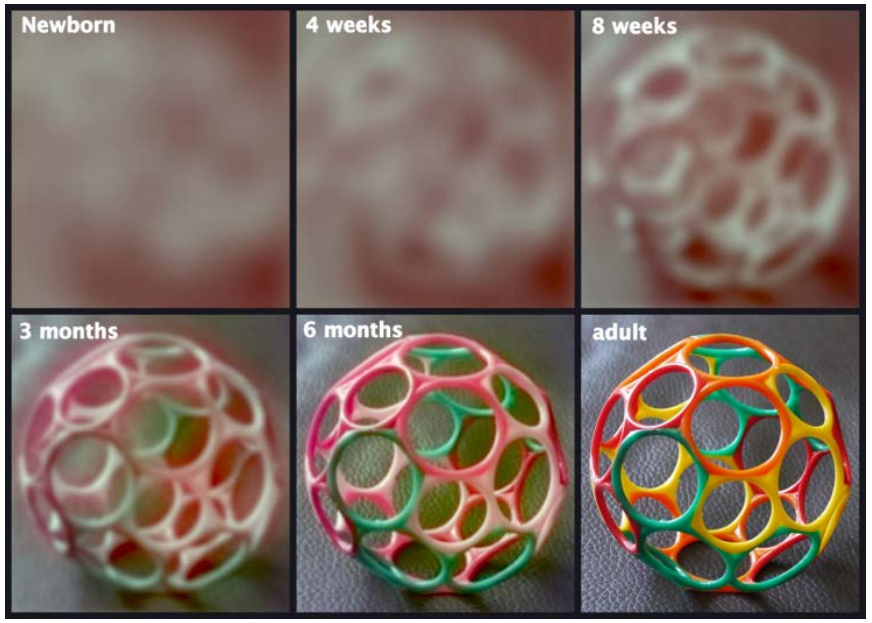






Human vs. Computer Vision

- Human vision
  - How vision develops in the first six months of life?
  - we see in context



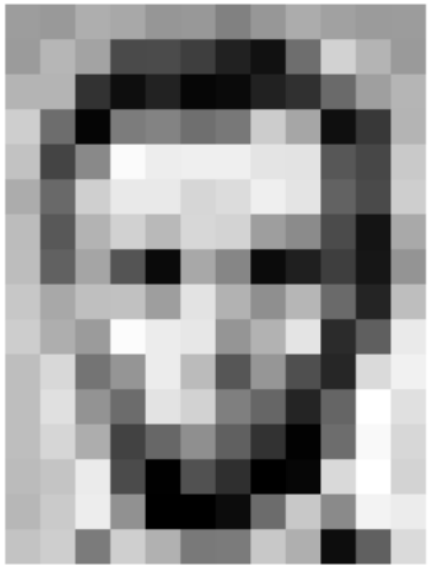
- Computer Vision
  - eigenface: express a particular face as a “sum” of notional faces through a machine learning process

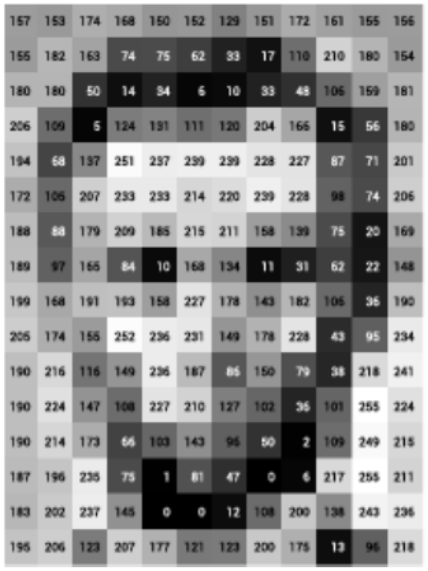
I see face

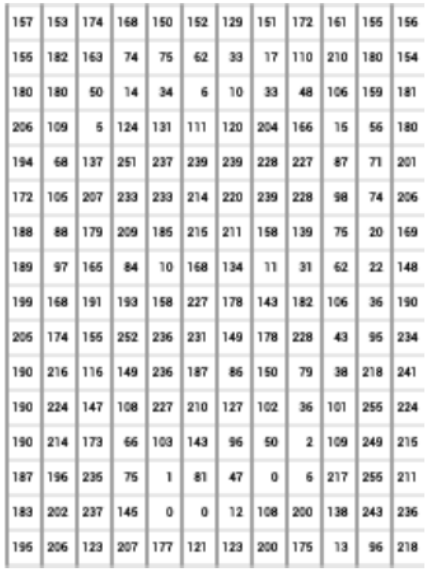
Eigenfaces look like something out of a scary movie

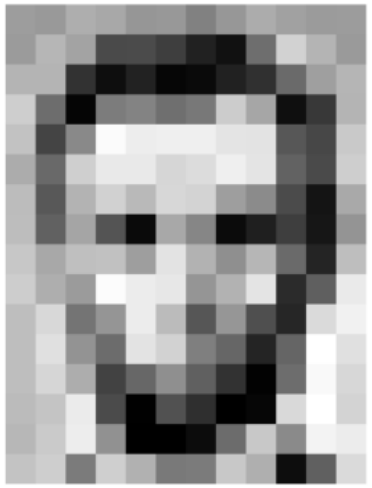
Image as Numbers

- image size (image resolution)
  - i.e., 1080x1080
- image pixel values
  - e.g., 8-bit in [0, 255]
- gray
  - 1 channel: Y
- color
  - 3 channels: R, G, B
- classification
  - based on 2D matrix of pixel values

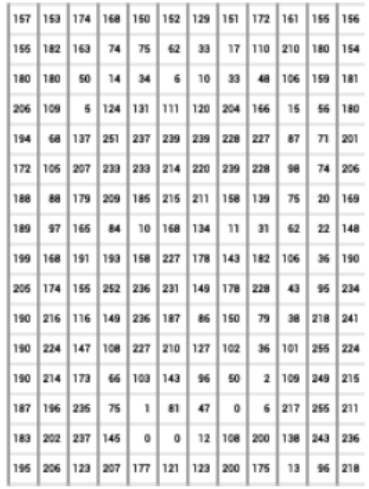








Input Image



Pixel Representation

classification

Lincoln	0.8
Washington	0.1
Jefferson	0.05
Obama	0.05

Convolution

- 2D filter kernel on images
  - different filter weights capture different features
- smoothing

$$G = \frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$$

  - Gaussian filter
  - box filter

$$B = \frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

- sharpening
- feature extraction using handcrafted algorithms
  - edge
  - corner

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

input image

*

1	0	-1
1	0	-1
1	0	-1

kernel

=

-5	-4	0	8
-10	-2	2	3
0	-2	-4	-7
-3	-2	-3	-16

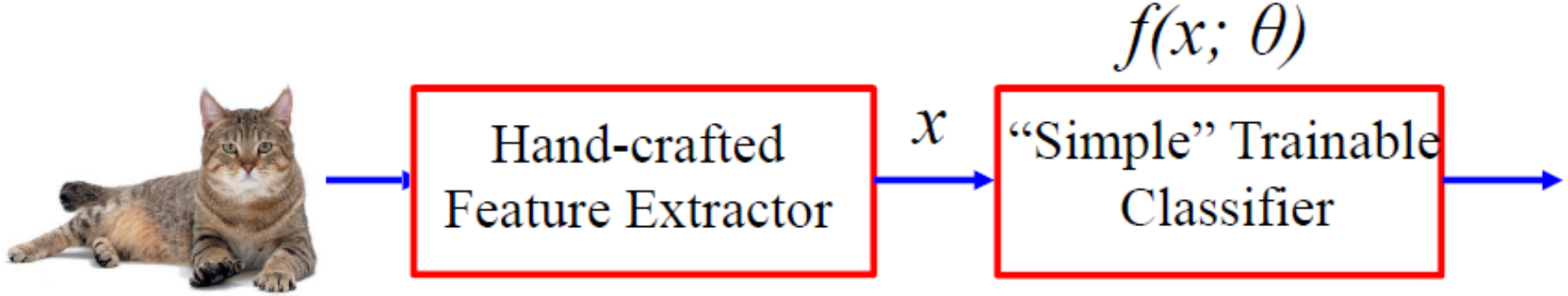
output image

3^{1}	0^{0}	1^{-1}	2	7	4
1^{1}	5^{0}	8^{-1}	9	3	1
2^{1}	7^{0}	2^{-1}	5	1	3
0	1	3	1	7	8
4	2	1	6	2	8
2	4	5	2	3	9

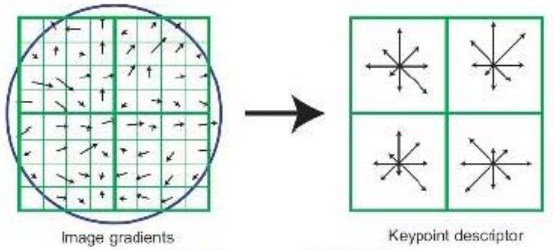
3	0^{1}	1^{0}	2^{-1}	7	4
1	5^{1}	8^{0}	9^{-1}	3	1
2	7^{1}	2^{0}	5^{-1}	1	3
0	1	3	1	7	8
4	2	1	6	2	8
2	4	5	2	3	9

3	0	1^{1}	2^{0}	7^{-1}	4
1	5	8^{1}	9^{0}	3^{-1}	1
2	7	2^{1}	5^{0}	1^{-1}	3
0	1	3	1	7	8
4	2	1	6	2	8
2	4	5	2	3	9

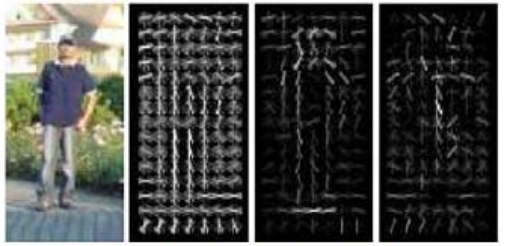
Conventional Image Classification



- handcraft features
  - Edges, SIFT, SURF, HOG, Optical Flow, …
  - Different objects might have different features
  - many feature algorithms are proposed in the past decade in computer vision

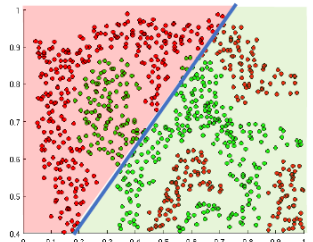


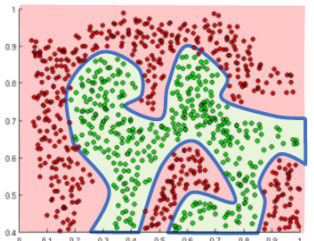
SIFT [Lowe, IJCV'04]
Citations: 43465



HoG [Dalal & Triggs, CVPR'05]
Citations: 20174

- classifier
  - SVM dominates classifier in conventional image classification





Feature Extraction with Convolution

- different filters extract different features from images
  - filter coefficients (weights) are determined during training in machine learning (instead of handcrafted feature extraction)
- below are some handcrafted filters for feature extraction in conventional computer vision

Original

Sharpen

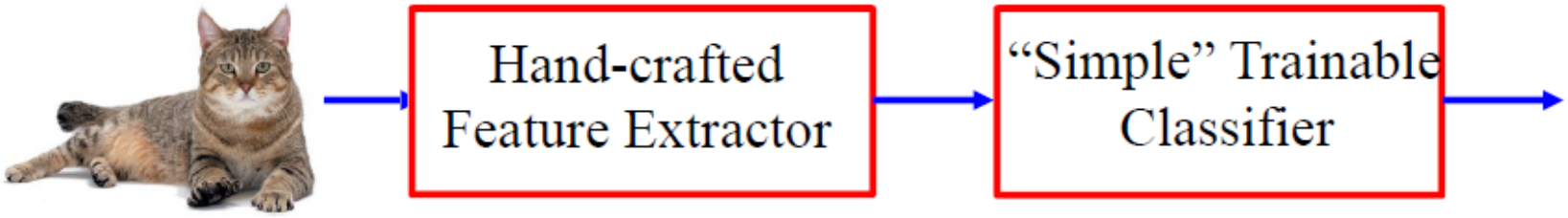
Edge Detect

"Strong" Edge Detect

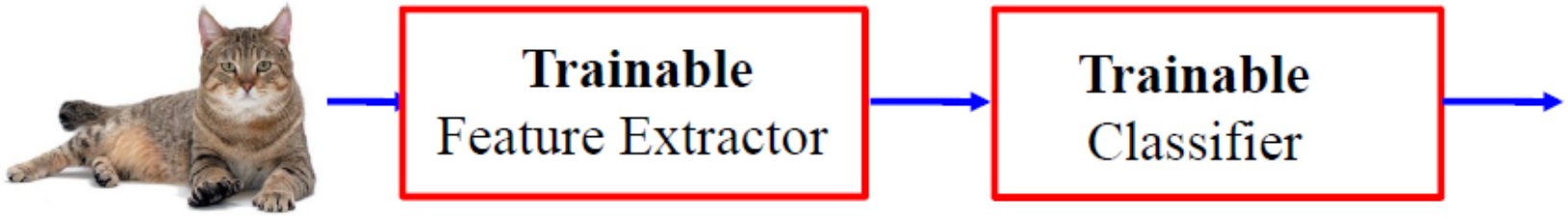
Classification (Conventional vs. Deep Learning)

- in modern end-to-end deep learning (DL) classification
  - Trained automatic feature extractor using Convolutional (Conv) layers
  - Trained classifier using Fully Connected (FC) layers

- Classical CV



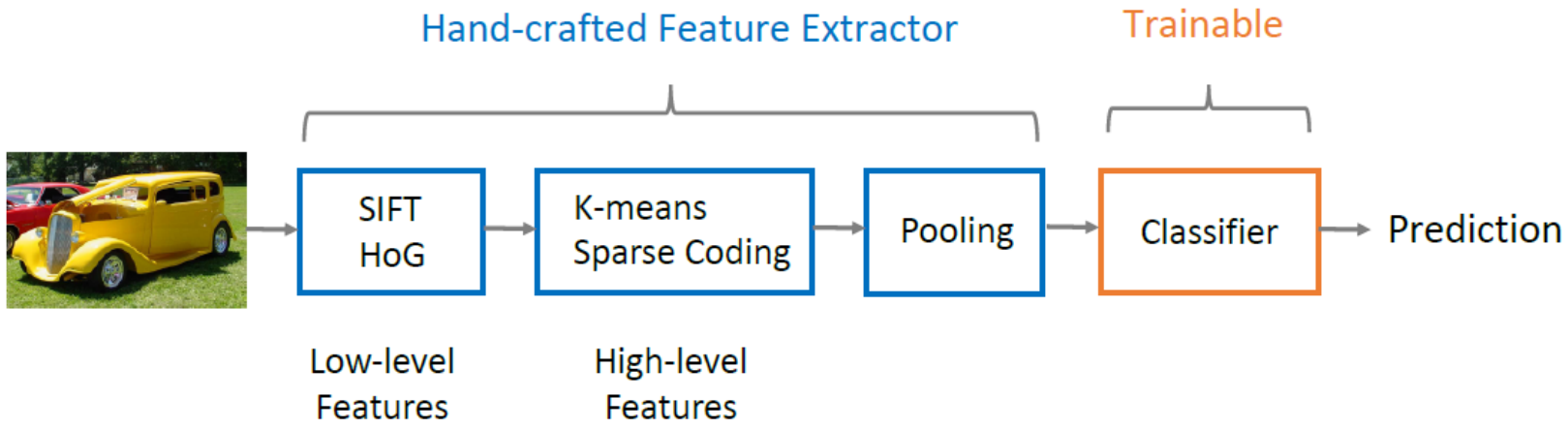
- Deep learning



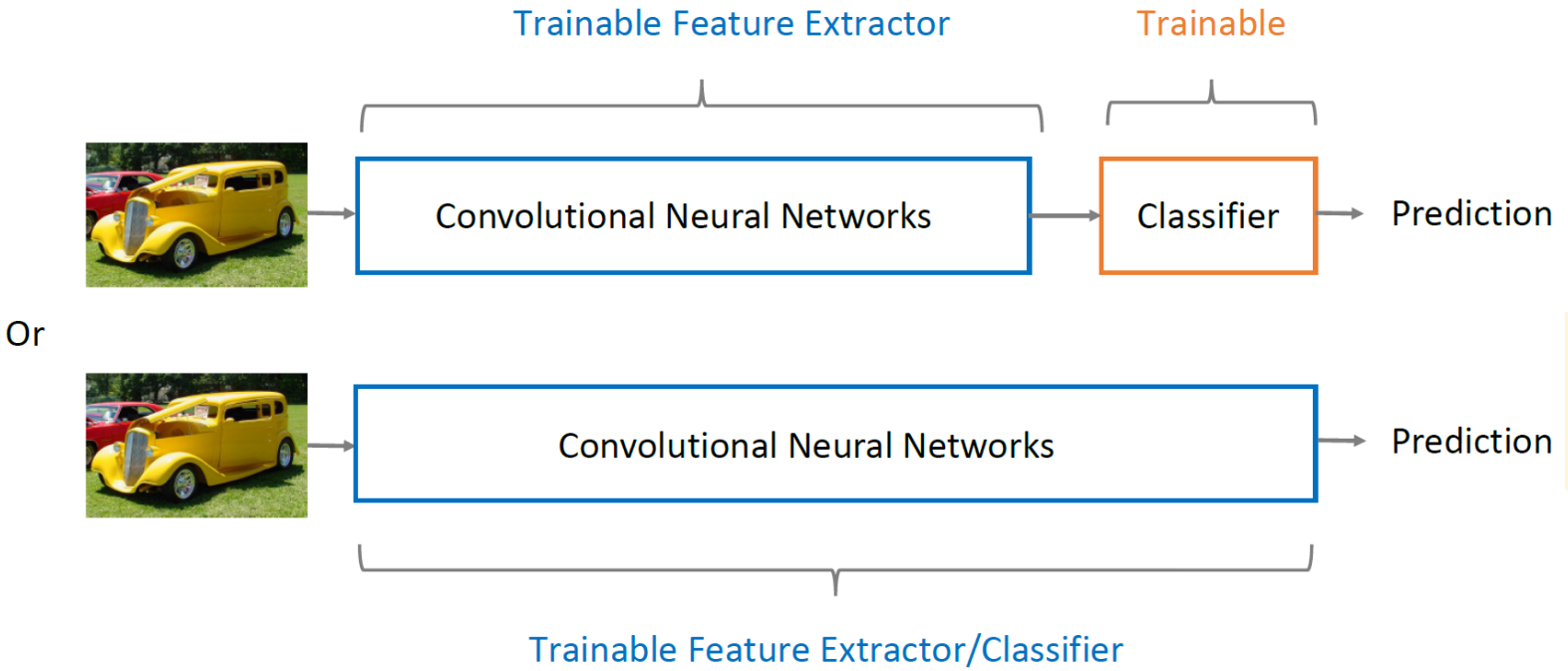
End-to-End Learning

Conventional vs. CNN classification

- conventional classification methods (before 2012)

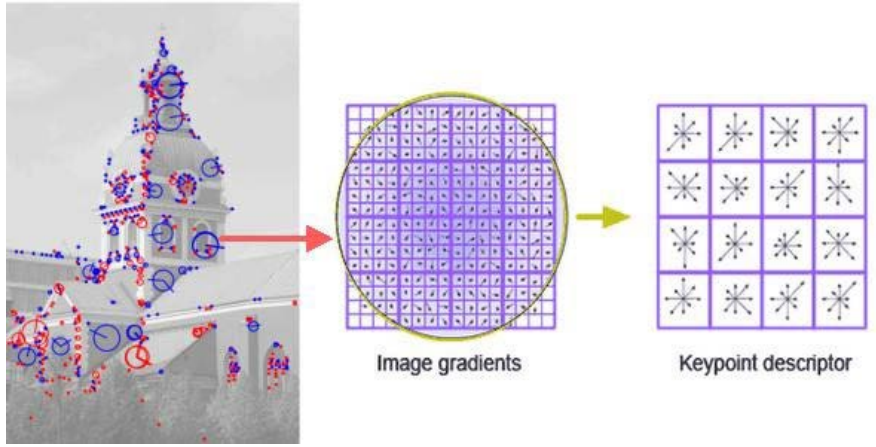


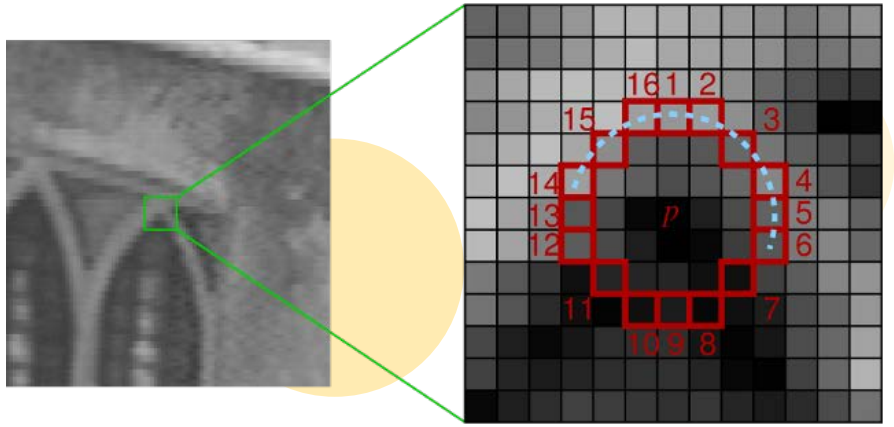
- Convolutional Neural Network (CNN) methods (after 2012)



Handcrafted Features

- Edges: Sorbel, Canny
- Corners: Harris corner
- SIFT (Shift-Invariant Feature Transform)
- SURF (Speeded Up Robust Features)
- FAST (Feature from an Accelerated Segment Test)
- ORB (Oriented Robust Binary features)
- BRIEF (Binary Robust Independent Features)
- HOG (Histogram Of Gradients)
- Optical Flow
- …





Edge

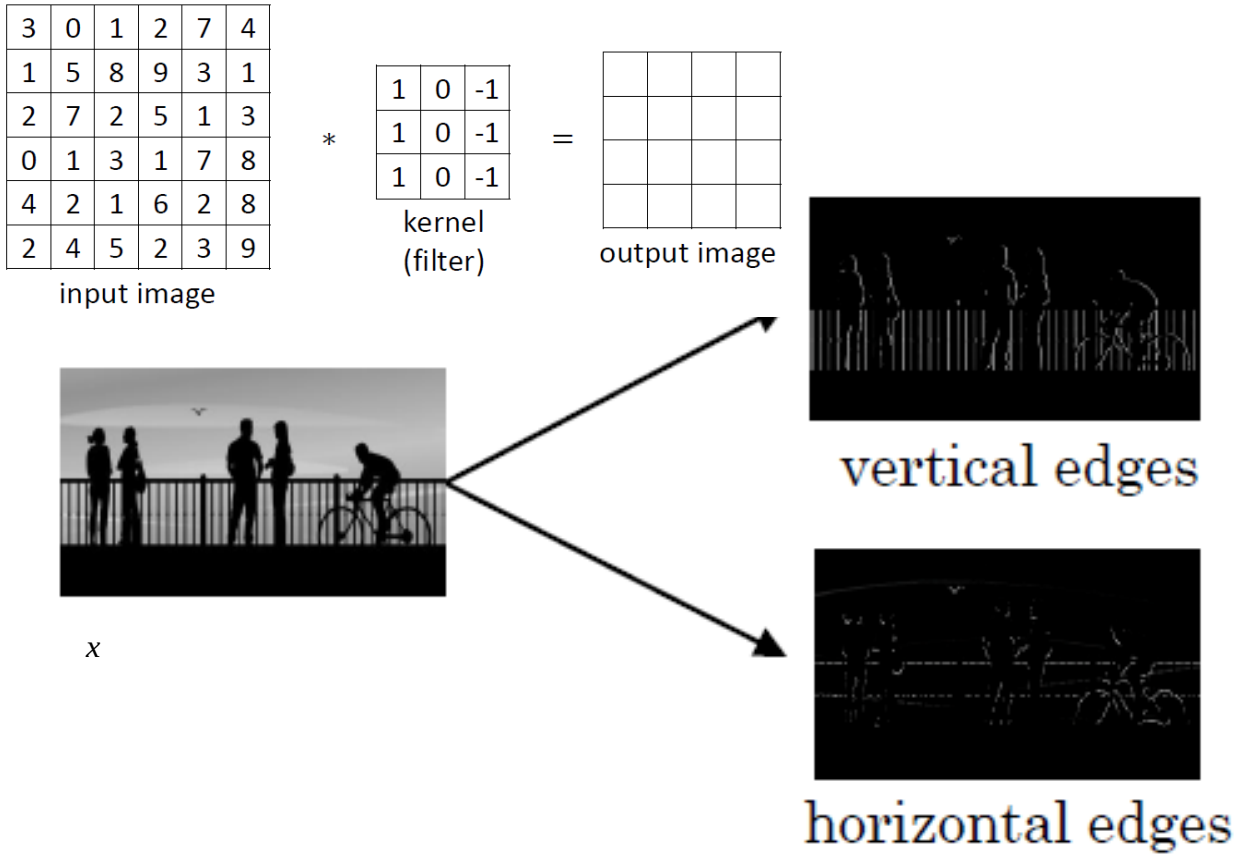
- Sorbel edge detector
  - convolution with simple Sorbel vertical and horizontal operators

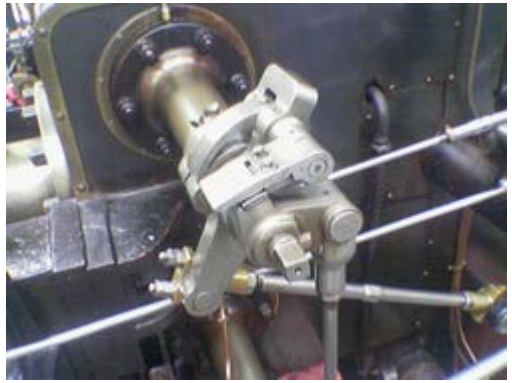
$$S_x = \begin{bmatrix} -1 & 2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}, S_y = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$$

- Laplacian edge detector

$$L_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & \end{bmatrix}, L_2 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -8 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

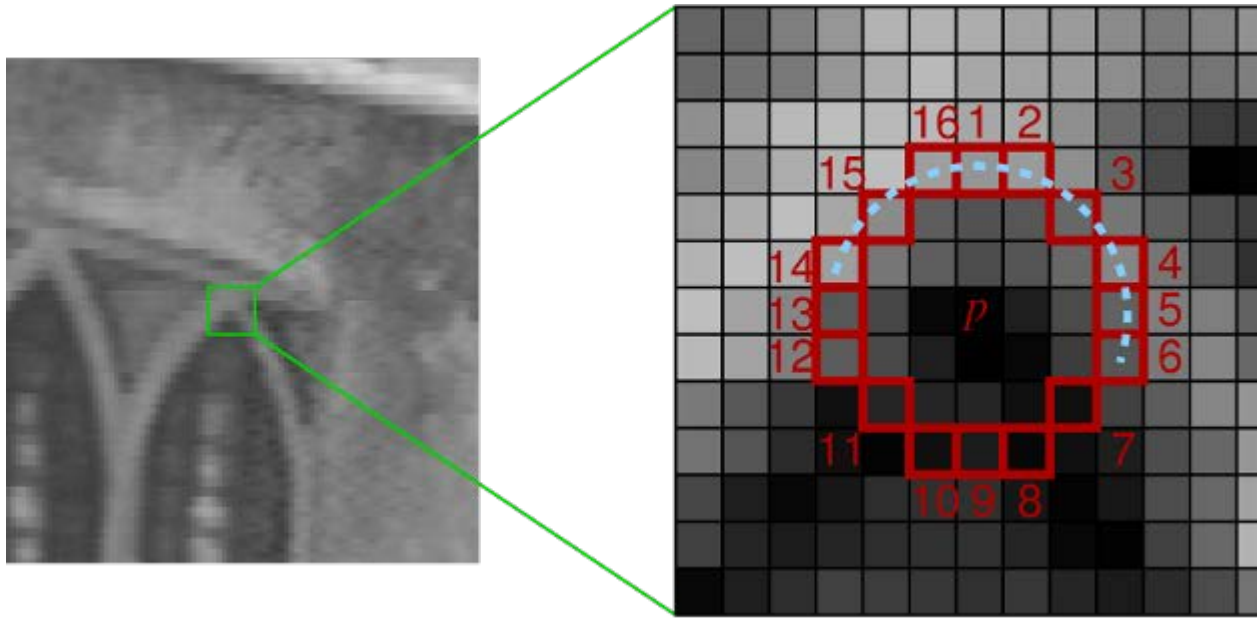
- Canny edge detector
  - more complicate algorithm, but more accurate results

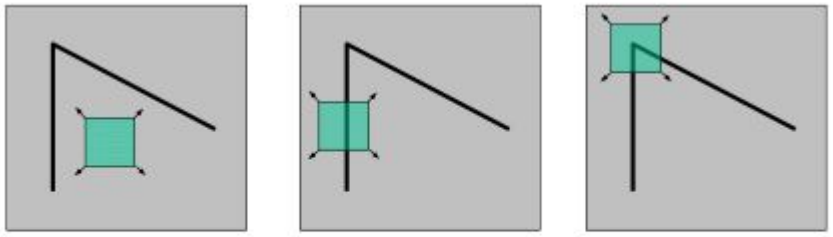




Corner

- a corner is where two edges intersect
- Harris corner detection





"flat" region:
no change in all directions

"edge":
no change along the edge direction

"corner":
significant change in all directions

Manual Feature Extraction

Domain knowledge → Define features → Detect features to classify

Viewpoint variation

Scale variation

Deformation

Occlusion

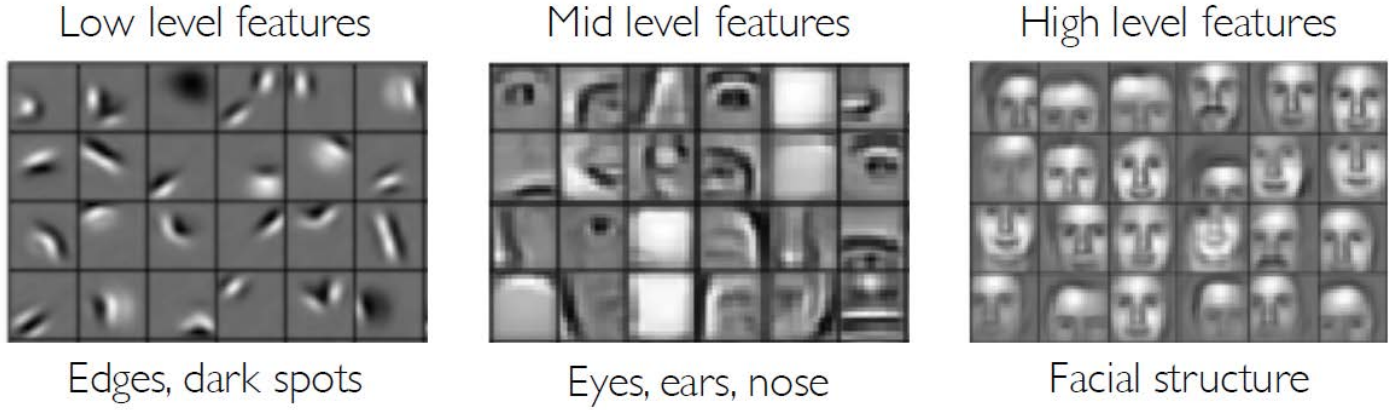
Illumination conditions

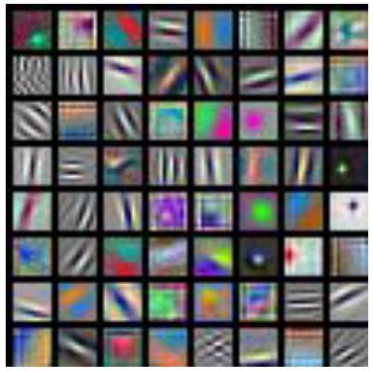
Background clutter

Intra-class variation

Feature Extraction by Machine Learning

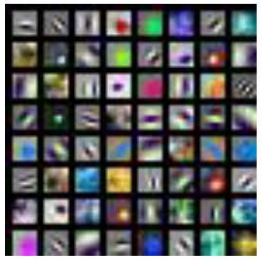
- extract features directly from raw data (instead of handcrafted)
  - using many convolution layers to extract various levels of features
  - low-level features in first (bottom) layers
  - high-level features in later (top) layers





AlexNet:
64 x 3 x 11 x 11

ResNet-18:
64 x 3 x 7 x 7



ResNet-101:
64 x 3 x 7 x 7

DenseNet-121:
64 x 3 x 7 x 7

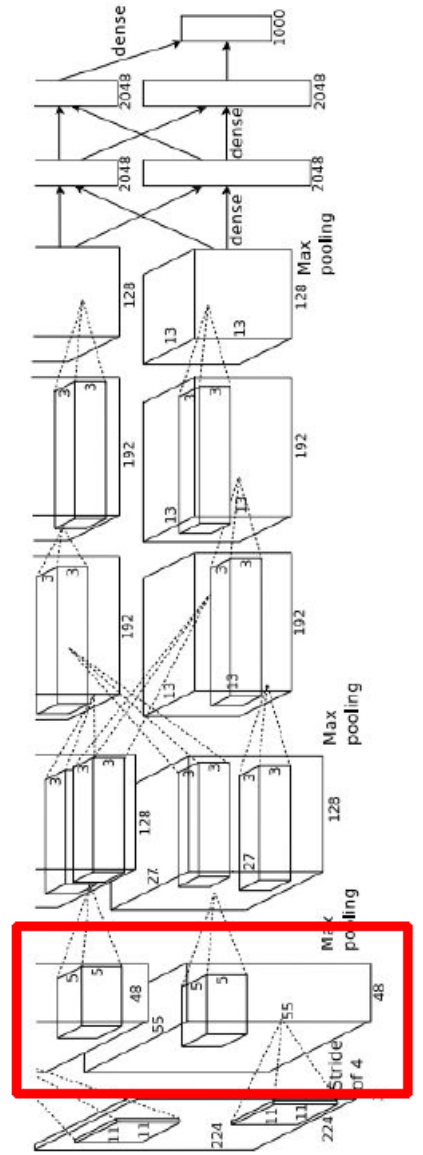
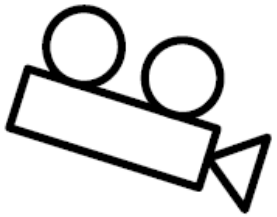
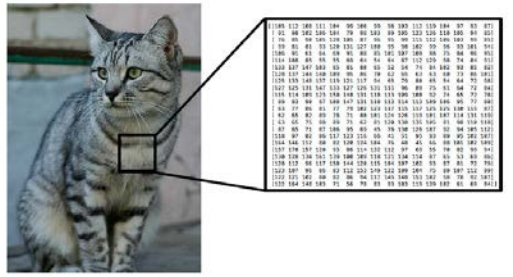


Image Classification Challenges

- viewpoint variation
  - all pixel values change when camera moves





- illumination

- deformation

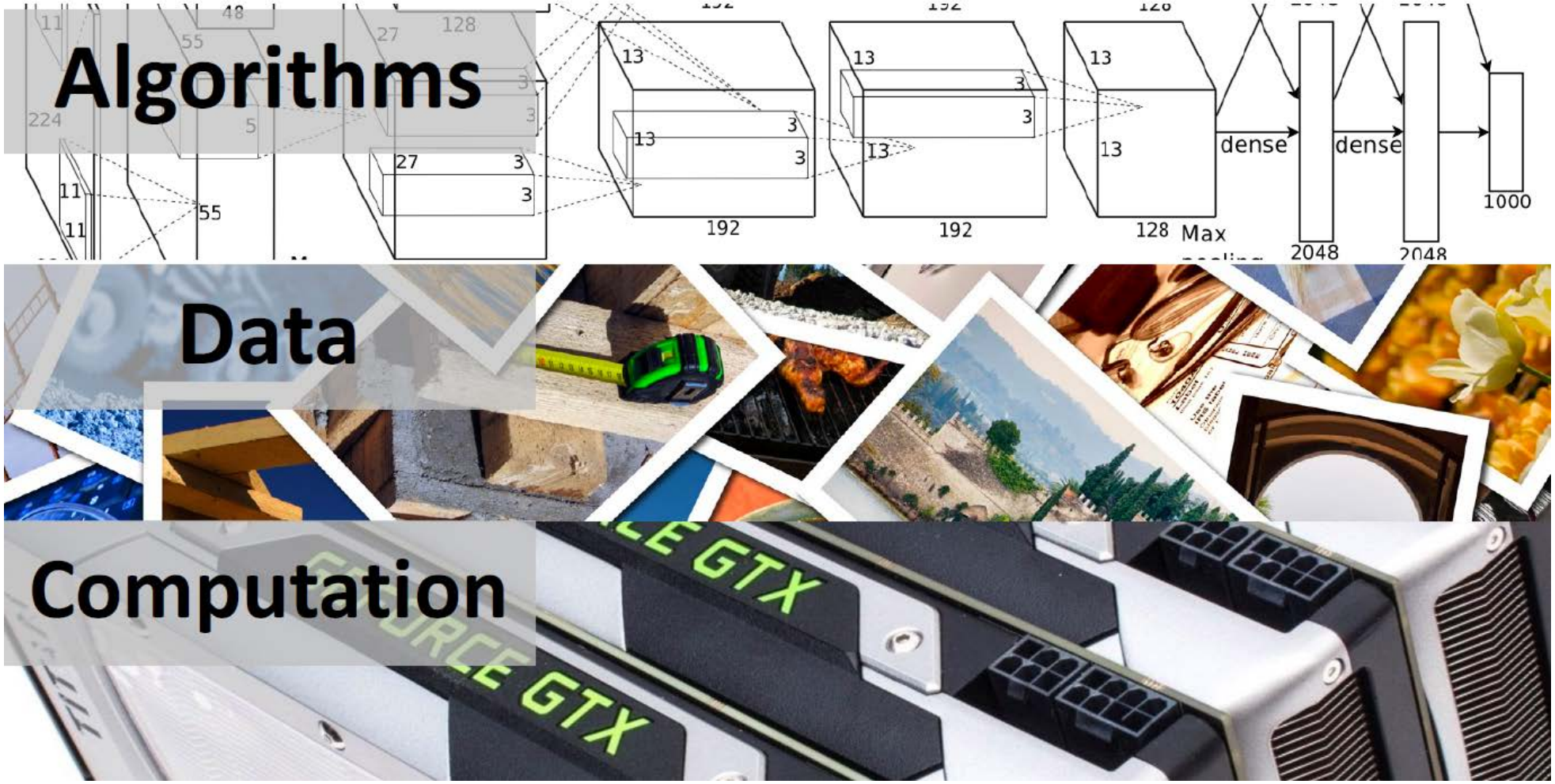
- occlusion

- background clutter

- intraclass variation

Ingredients in Deep Learning

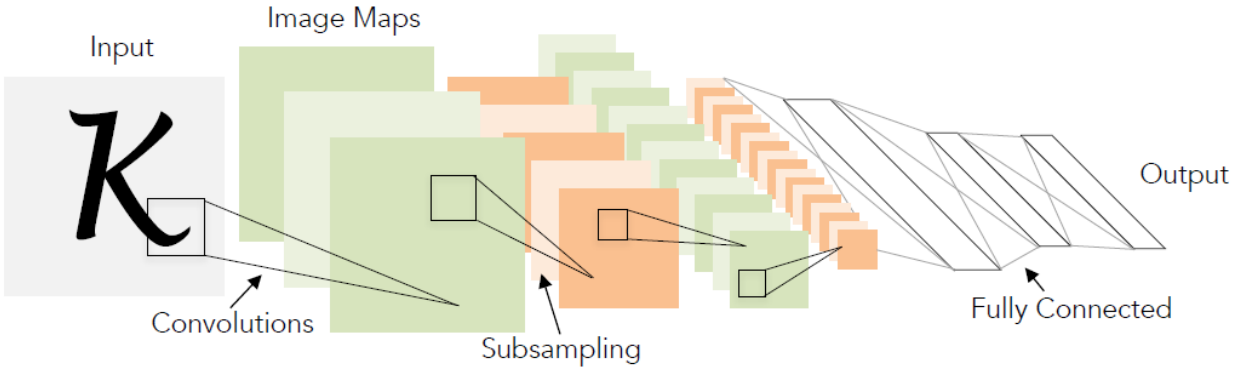
- Algorithm (Model) + Training Data + Computation
  - CNN models + Big Data + Accelerators



Evolution of CNN Image Classification

- LeNet (1998)
  - digit classification (0~9)
  - Convolutional (Conv) layers + Fully Connected (FC) layers
  - trained using MNIST dataset (60,000 samples)

1998
LeCun et al.



# of transistors

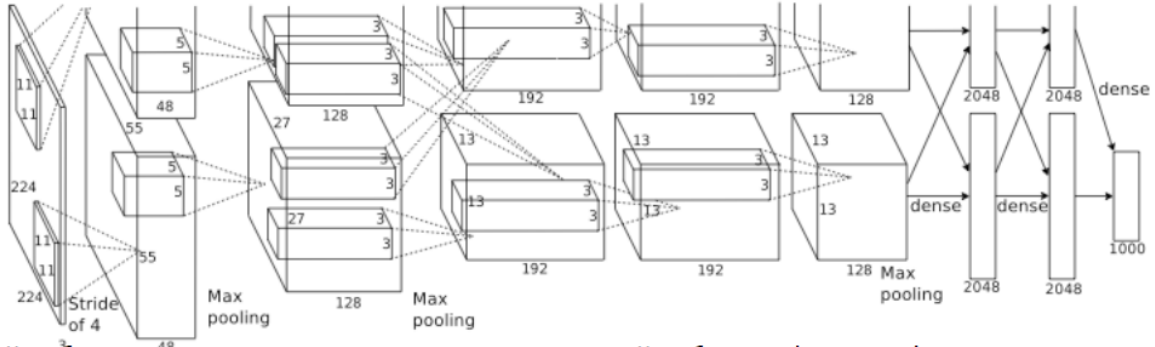
10^6

# of pixels used in training

10^7 NIST

- AlexNet (2012)
  - 1000-category classification
  - more Conv layers
  - trained using ImageNet dataset (1,200,000 samples)

2012
Krizhevsky et al.



# of transistors

10^9

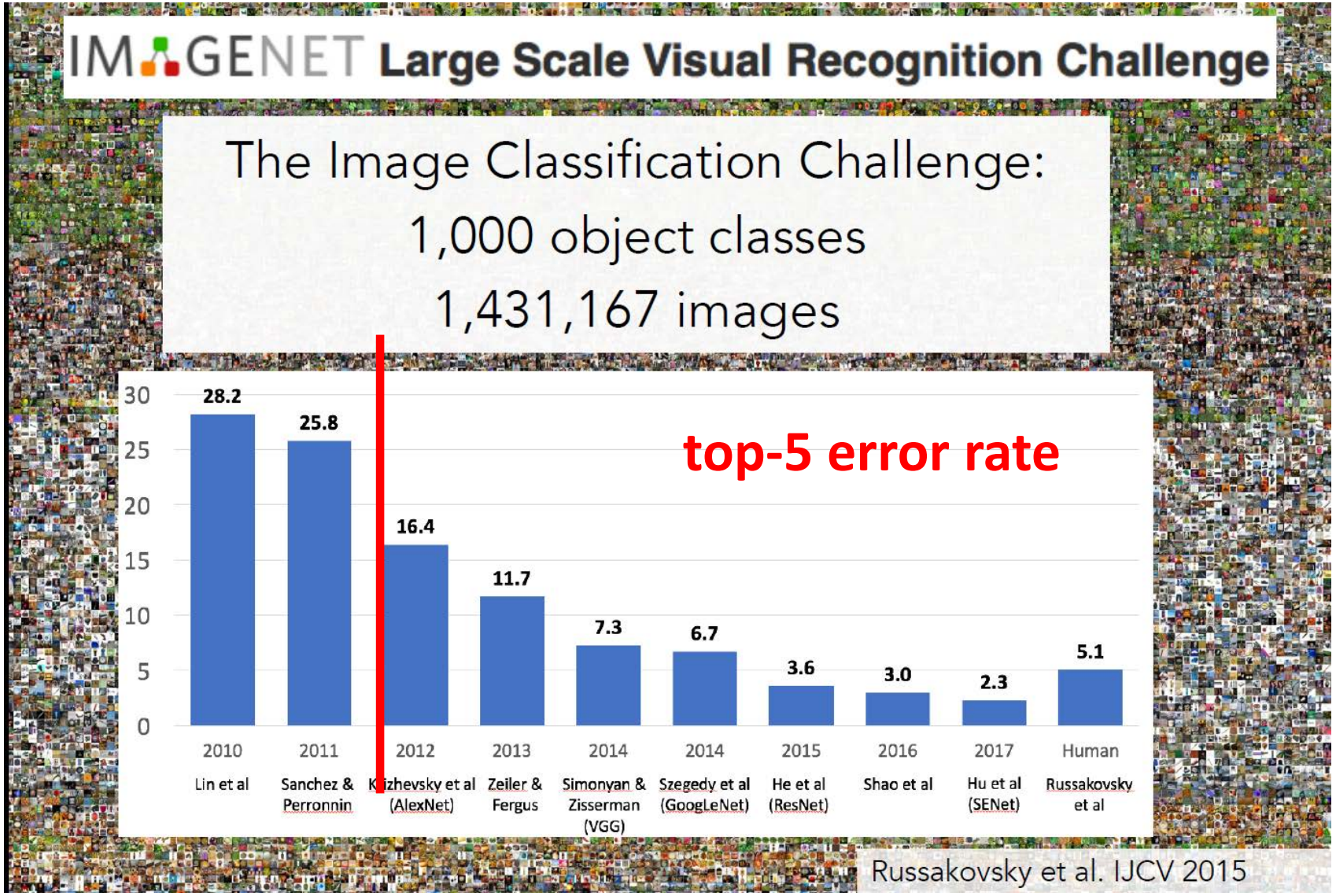
GPUs

# of pixels used in training

10^{14} IMAGENET

ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

- Deep learning model starts in 2012 (AlexNet)
  - significant accuracy improvement (**16.4%**)
- 2014 VGG model
  - 16, 19 layers (**7.3%**)
- 2014 GoogLeNet
  - 22 layers (**6.7%**)
- 2015 ResNet
  - >150 layers (**3.6% !!!**)
- human capability
  - top-5 error rate: **5%**



ILSVRC Winner CNN Models

Year 2010

NEC-UIUC

Dense descriptor grid: HOG, LBP

Coding: local coordinate, super-vector

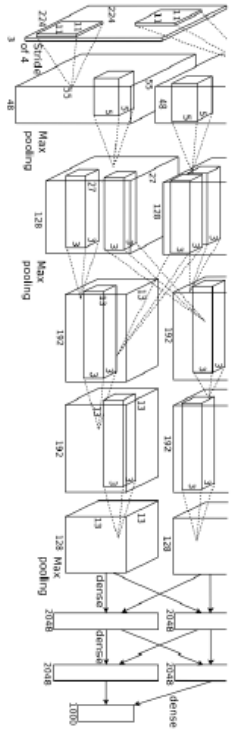
Pooling, SPM

Linear SVM

[Lin CVPR 2011]

Year 2012

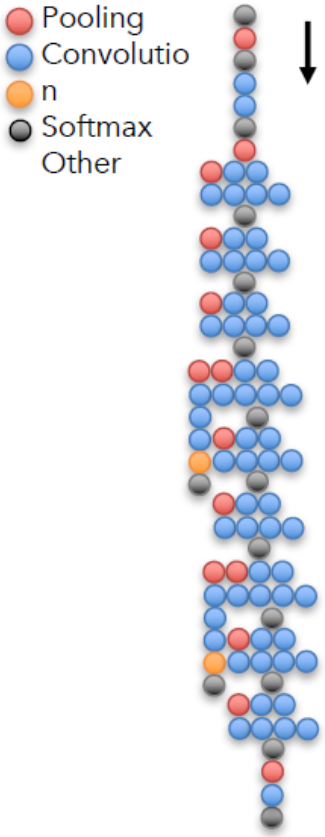
AlexNet



[Krizhevsky NIPS 2012]

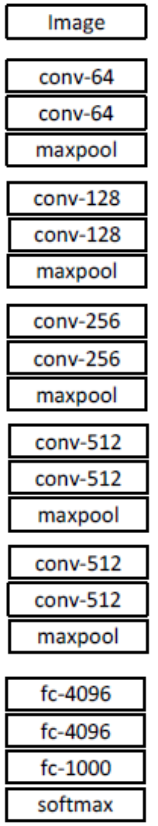
Year 2014

GoogLeNet



[Szegedy arxiv 2014]

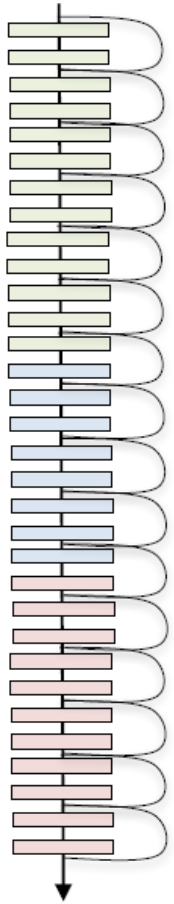
VGG



[Simonyan arxiv 2014]

Year 2015

ResNet



[He ICCV 2015]

Image Classification

- output probability of most likely one or five categories (top-1, top-5)
  - softmax appended at the last classification layer

$$\text{softmax}(y_i) = \frac{e^{y_i}}{\sum_j e^{y_j}}$$

CV Application Examples

- classification
- classification + localization
- object detection
- segmentation
- style transfer
- image captioning
- image enhancement

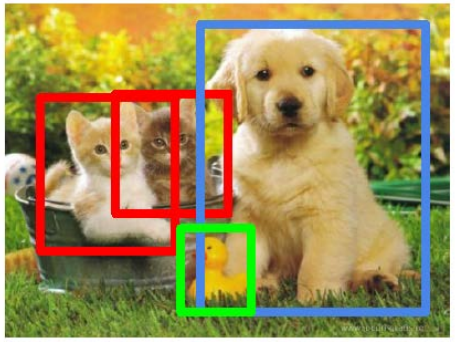
Classification

CAT

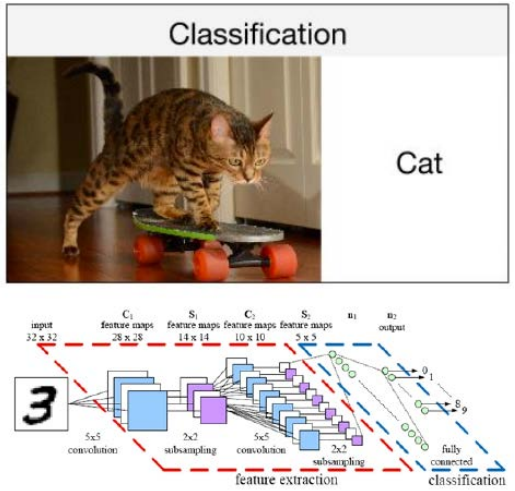
Classification + Localization

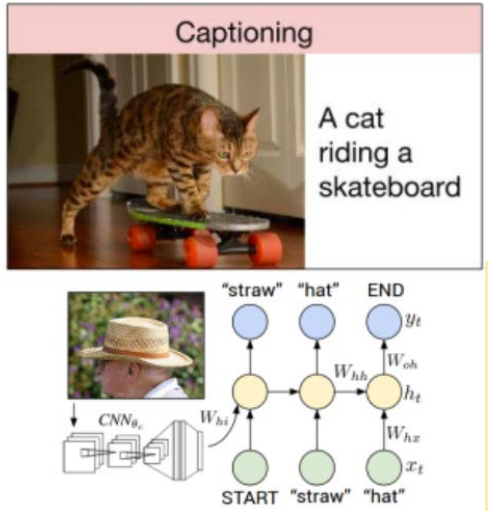
CAT

Object Detection



CAT, DOG, DUCK



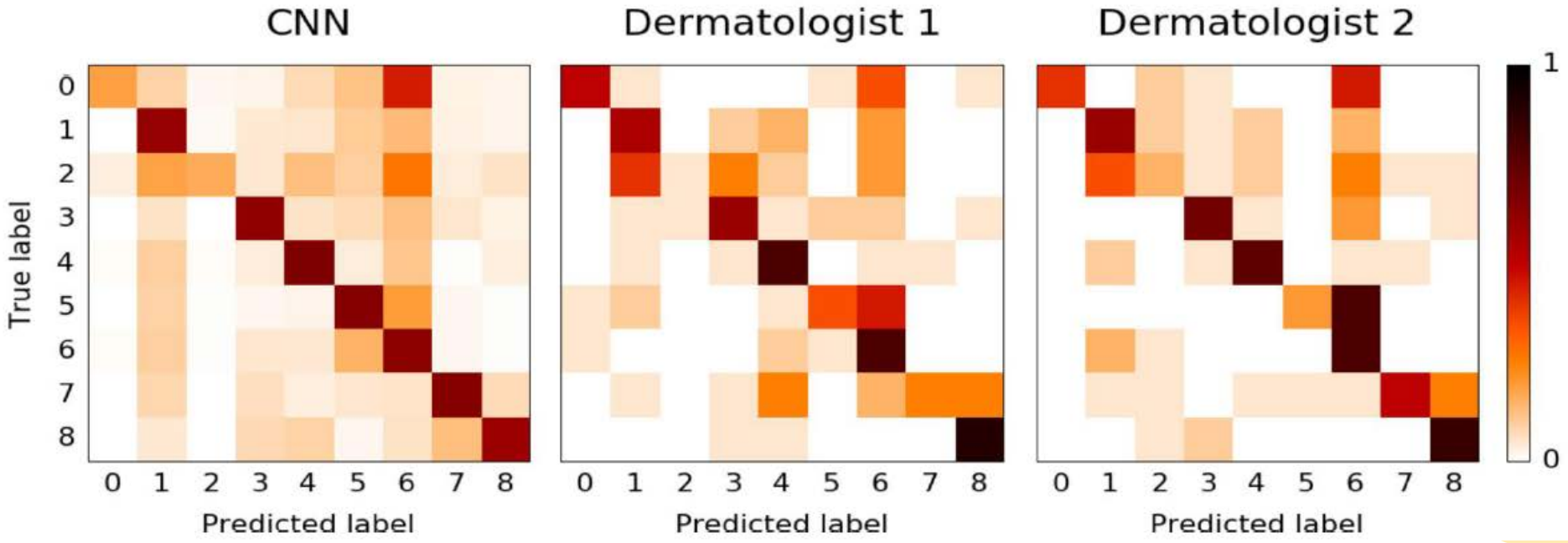


Segmentation

CAT, DOG, DUCK

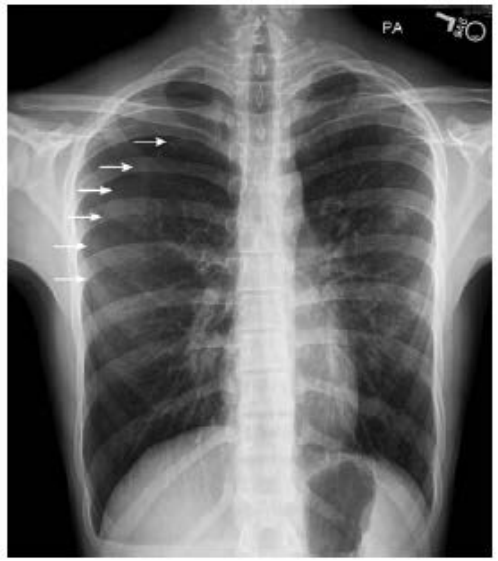
Medical Application Examples

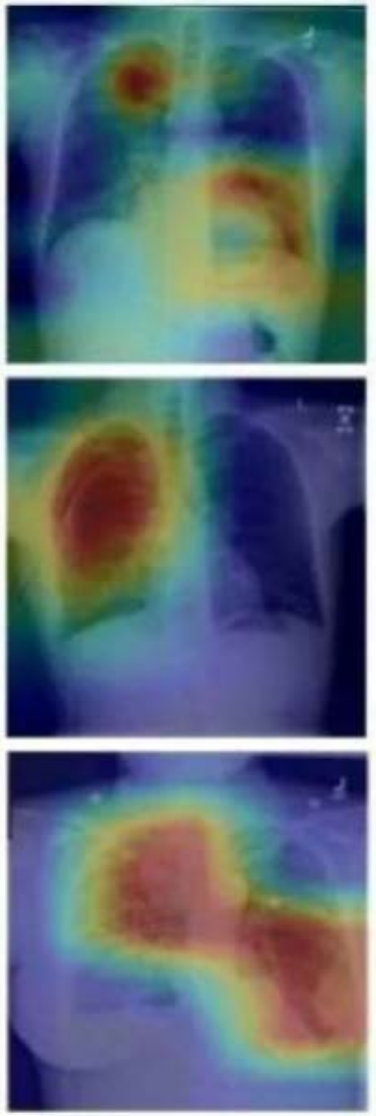
- Dermatology-level classification of 9 types of skin cancers with deep neural networks (2017/2 Nature)
- confusion matrices (more accurate with more diagonally concentrated)
- CNN outperforms dermatologists in categories 5 and 7, but loses in categories 6 and 8



Medical Application Examples

- detect pneumothorax in X-Ray scans
  - could output intensity map to highlight the critical area
- some critics about the accuracy of ground truth data labeling
  - data collection and labelling are important in **supervised learning** 監督式學習

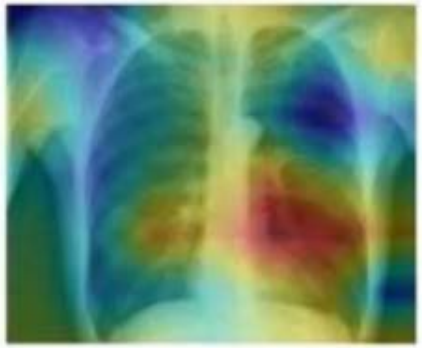




Input
Chest X-Ray Image

CheXNet
121-layer CNN

Output
Pneumonia Positive (85%)



ChexNet : 121 層的卷積神經網路，
匯入胸部透視圖，匯出患病機率。
同時定位圖中最有可能患病的位置。

Datasets

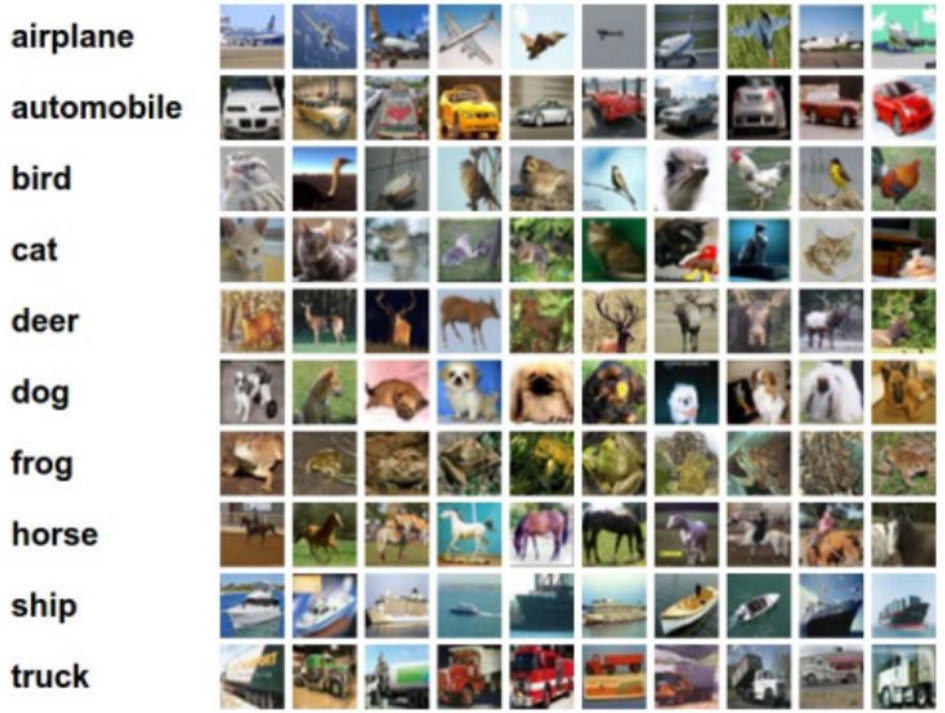
- Datasets are used for training deep neural network models
- Popular datasets
  - MNIST
    - digit 0~9
  - CIFAR-10/100
    - 10/100 categories
  - ImageNet
    - >1000 categories
  - PASCAL VOC
    - 20 categories
  - COCO
    - 80 categories
  - …

dataset	image size	# of samples	# of categories	image types	training set	test set
MNIST	28x28 gray	70K	10	digits	60K	10K
ImageNet	256x256 color	> 15 M	> 22,000	vision	1.2M	150K
Caltech-101	300x200 color	>9K	101	vision	5K	3K
CIFAR-10/100	32x32 color	6 K	10, 100	vision	60K	10K
PASCAL VOC	500*334 color	>10K	21	vision	10K	5K
LabelMe (MIT)	-	>2K	7	vision	2K	1K
SVHN	32x32 color	>70K	10	vision	600K	26K
COCO	-	330K	80	vision	10K	5K
CompCars	-	136K	-	cars	-	-
Stanford Cars	-	16K	196 car classes	cars	8K	8K
KITTI	64*32 color	>15K	5	pedestrian	>7K	>7K
INRIA	64*128 color	>0.8K	1	pedestrian	614	288

MNIST, CIFAR-10, ImageNet Datasets

- MNIST: 10 categories, 60,000 training images of 28x28x1, 10,000 test images
- CIFAR-10: 10 categories, 60,000 training images of 32x32x3, 10,000 test images
- ImageNet: 1000 categories, 1,200,000 training images of 256x256x3, 150,000 test images

MNIST: handwritten digits



ImageNet:
22K categories. 14M images.