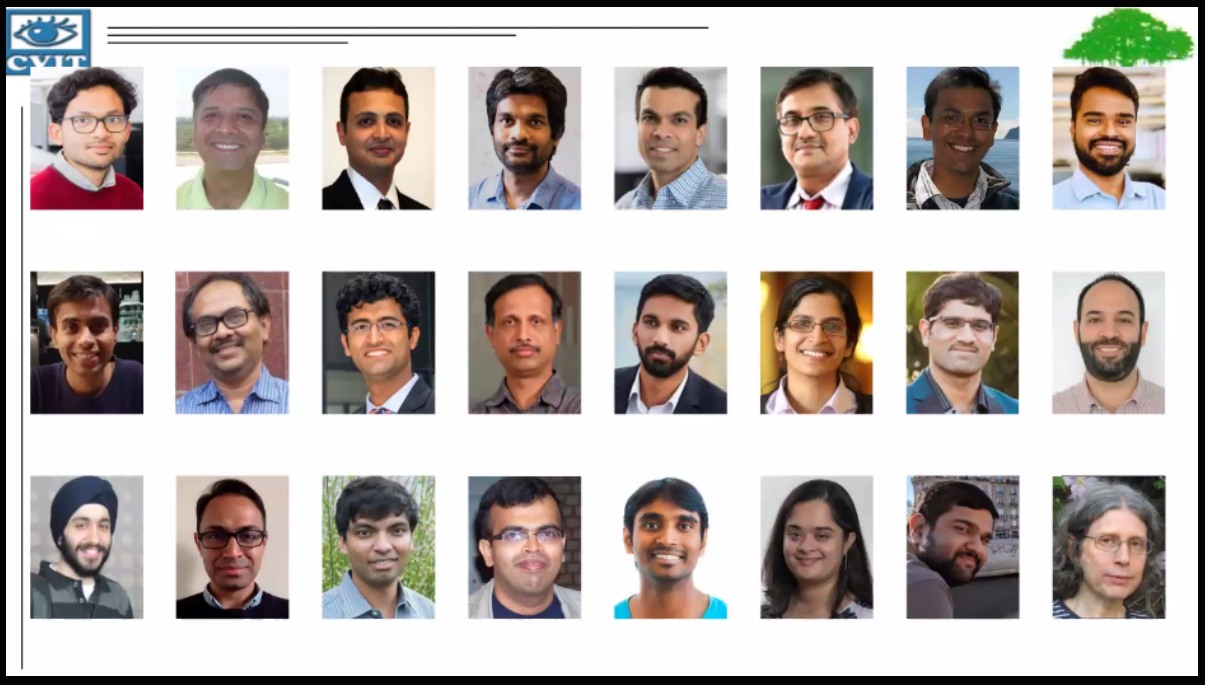
CVIT Summer School, 2021 - ONLINE Mode

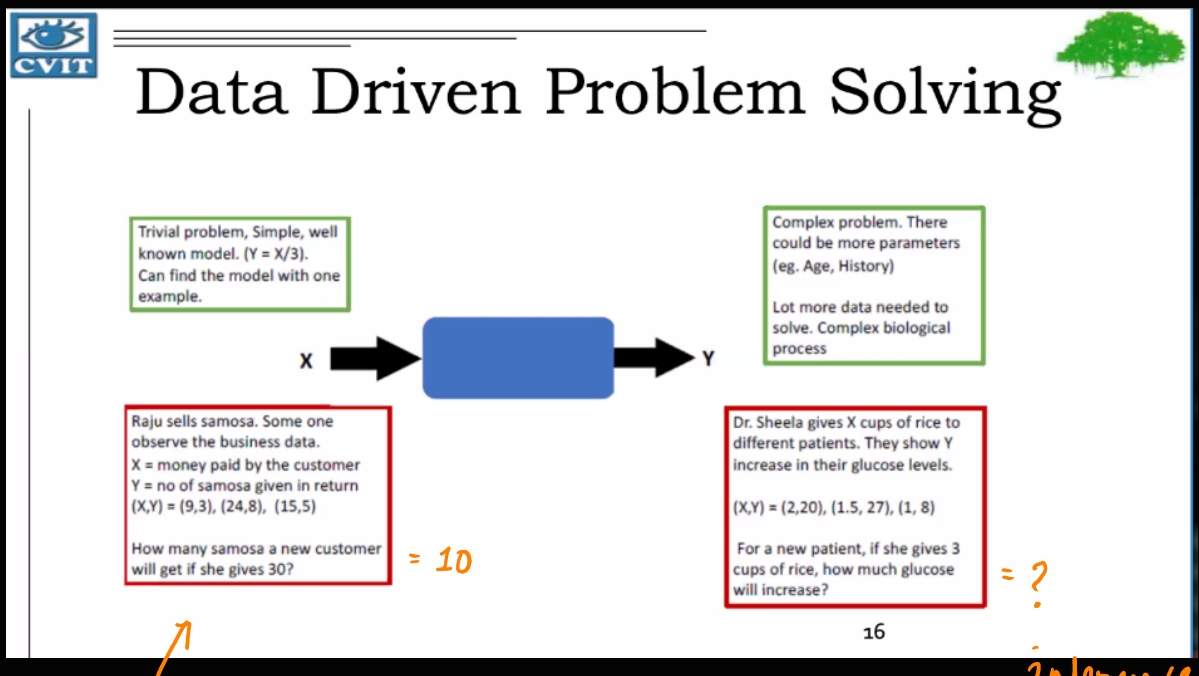
Day 1 : Prof. C.V. Jawahar

Topic: Welcome and Introductions

Time line: Glimse of Previous Summer Schools.

Speakers:





= Difficult to determine as the data seems insufficient for some inference. May be more parameters/data will make the work easier

This problem is easier as it follows a trend & can be computed by a fixed algorithm

More Examples

In this match , Rajasthan captain Ajinkya Rahane won the toss and decided to bowl first .

X

Y

General Strategy: Given many examples of (X,Y), learn an automated solution to predict Y given a new X.

इस मैच में राजस्थान के कप्तान अजिंक्य रहाणे ने टॉस जीतकर पहले गेंदबाजी का फैसला किया ।

"Black and White Dog Jumps over Bars"

If we give sufficient (x,y) pairs, the AI algo can predict y, given x.

* Paper to Read *

ILL Posed Problems: Why do they yield results?

X Y

Can Human(experts) do this?

How do they do?

Title: Biscuits

Ingredients:
Flour, butter, sugar, egg, milk, salt.

Instructions:
- Preheat oven to 450 degrees.
- Cream butter and sugar.
- Add egg and milk.
- Sift flour and salt together.
- Add to creamed mixture.
- Roll out on floured board to 1/4 inch thickness.
- Cut with biscuit cutter.
- Place on ungreased cookie sheet.
- Bake for 10 minutes.

Extensive use of Prior Knowledge.

Composition of parts seen in the past.

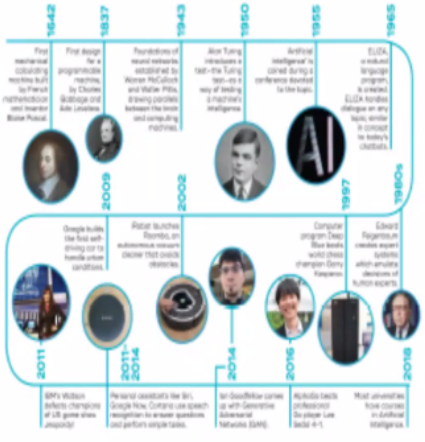
Inverse Cooking (CVPR 2019)

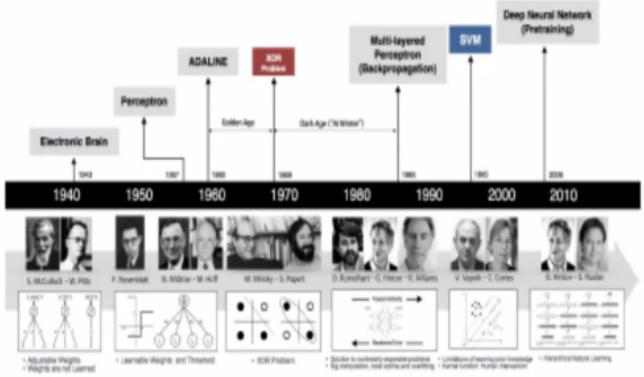
Human's yield results and predict the ingredients and recipe owing to their expertise.

i.e seeing more data (such recipe, ingredients) in the past.

History:

Is this all really new?

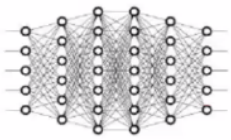




9 Why AI research is booming today?

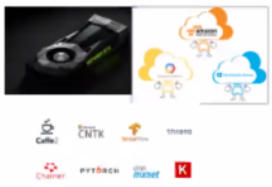
Why AI started to work now?

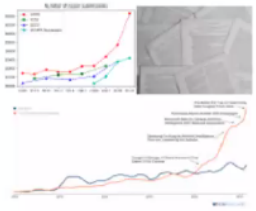
Data;
Internet;
Connectivity



Algorithms,
Deep Learning

Compute,
Cloud, APIs,
Libraries





More People,
Papers, Results,
Funding, People.
Positive Feedback.

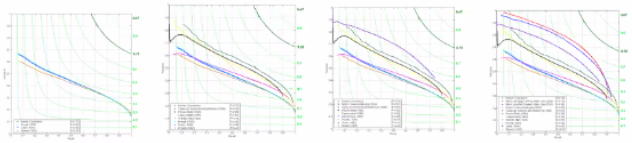
20

Introduction to Computer Vision. (Timeline)

History from Low/Mid-Level Vision

Malik, BISD, ICCV 2001

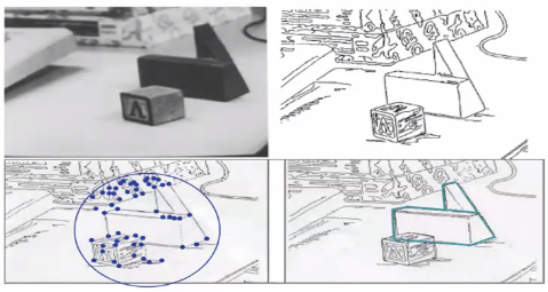
Sobel (1968,0.48), Canny(1986,0.54),
Martin(2004,0.63), Marie(2008,0.70),
Human(0.79)



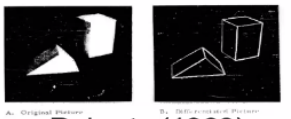
1970s 1990s 2004 2008

Maire, Arberaez, et. al., IEEE PAMI 2011

Edges and Corners



Huttenlocher & Ullman (1990)



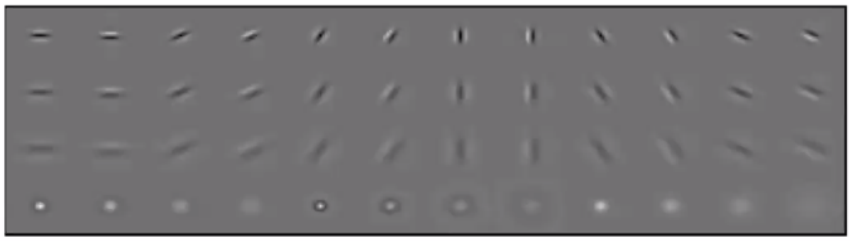
Roberts (1963)

Brooks & Binford (1981)

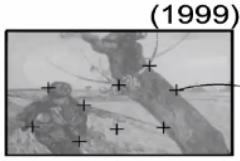
Rothwell, Zisserman, Forsyth & Mundy (1995)

From Jitenrda Malik

Filters



Malik & Perona (1990), Leung & Malik (1999) for texture analysis

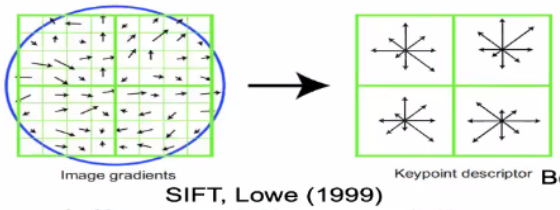


Schmid & Mohr (1997) for image retrieval

Viola & Jones (2001) for face detectior

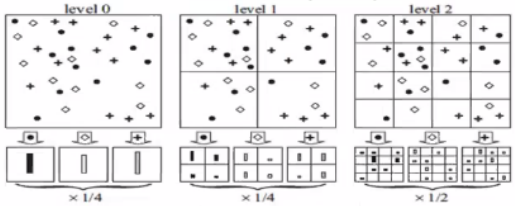
From Jitendra Malik

Histograms

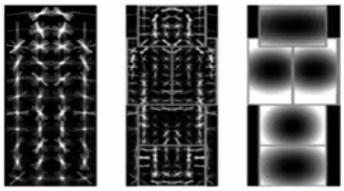


SIFT, Lowe (1999)

Shape Contexts, Belongie, Malik & Puzicha(2001)



Spatial Pyramid Matching, Lazebnik, Schmid & Ponce (2006)

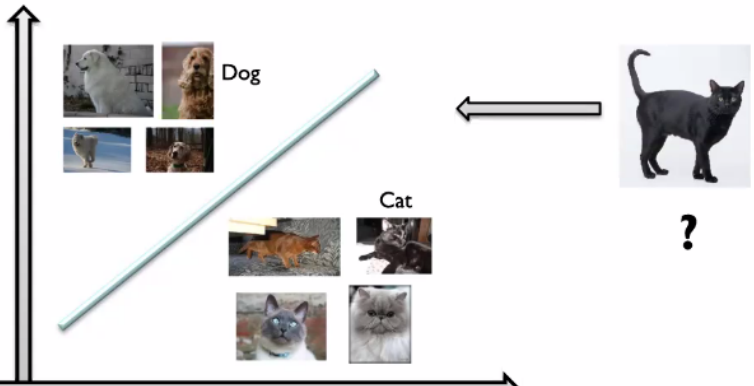


DPM based on HOG, Felzenswalb et al (2010)

From Jitendra Malik

Category Classification

Given an image, classify it as a cat or a dog.



Parkhi et al, "Cats and Dogs", CVPR 2012

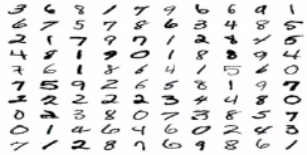
Many Challenges
- Appearance
- Occlusion
- Pose
- Etc.

Classifiacation
- Many #Classes
- Distractors

* Binary Classification and its challenges

History: Early object categorization

- Turk and Pentland, 1991
- Belhumeur, Hespanha, & Kriegman, 1997
- Schneiderman & Kanade 2004
- Viola and Jones, 2000

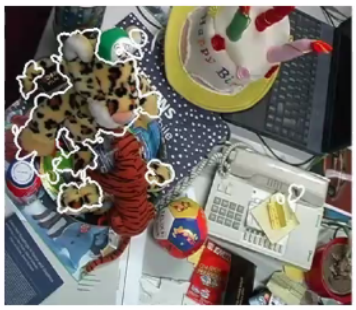


- Amit and Geman, 1999
- LeCun et al. 1998
- Belongie and Malik, 2002

- Schneiderman & Kanade, 2004
- Argawal and Roth, 2002
- Poggio et al. 1993

* Data was not collected from the wild
↓
Not much occlusions
∴ Poor Inference Results.

Instance (vs Category) Recognition



Slide credit Fei-Fei, Fergus, Torralba CVPR07 Short Course

* Two types of Classification
→ Instance &
→ Category

Towards Finer Understanding

Classification → Recognition.

Is there a dog in this image?

If yes, where is the dog?

← Localization.

Semantic → Segmentation

Which pixels exactly?

What breed?

→ Finer Classification

Variations in Problems

- Binary Classification $\mathbf{x} \longrightarrow l \in L = \{l_1, l_2\}$
- Multi Class Classification $\mathbf{x} \longrightarrow l \in L = \{l_1, l_2, ..., l_{\|C\|}\}$
- Multi Label Classification

$$\mathbf{x} \longrightarrow l^X = \{l_1^X, l_2^X, ...\} \quad l_i^X \in L = \{l_1, l_2, ..., l_{\|C\|}\}$$

- Structured Output Prediction

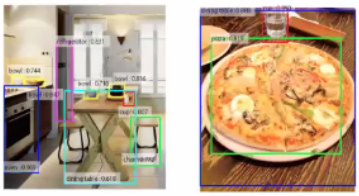
$$\mathbf{x} \longrightarrow y \in \mathcal{Y} = \{1, ..., k\}^l$$

- $\mathcal{Y}$ are complex (structured outputs)
  - Images, text, audio, folds of protein

* 'Classification' is a broad problem
  ← Many types

Problems Getting Solved

Segmentation and Object Detection



Object Detection

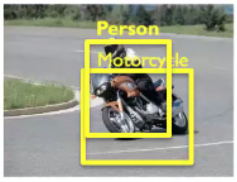
Describing Images in context

Human Keypoints Detection

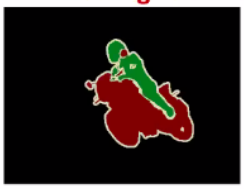
Evolution of the Space

Classification

Detection

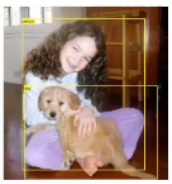


Semantic Segmentation



Generation

Caltech (2003)
Simple isolated objects

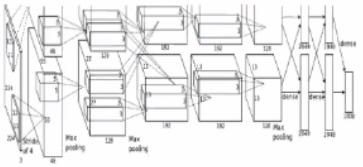


PASCAL (2005-2012)
20 classes.

1000 classes
> 1M images

IMAGENET

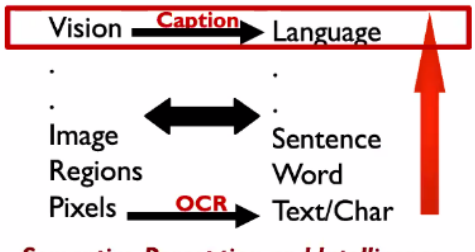
2010 - ?



CNNs, RNNs, Deep Learning

"Understanding" Images

Visual Question-Answering: Types



Semantics, Perception and Intelligence

SEMANTIC GAP IN IMAGE UNDERSTANDING ?

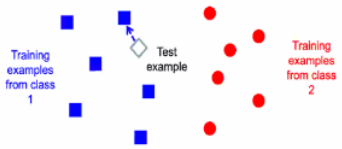
Evolution →

CV:What enabled this success?

- Modern Features
  - Invariant to popular transformations
  - Capable of capturing local and global (shape, colour, texture) characteristics reliably
  - Features than can be learnt
- Machine Learning
  - Learn from examples rather than handcoding
  - New algorithms: effective, efficient
  - Efficient algorithms to solve complex optimization tasks
- Realistic Data
  - Huge amount; partly annotated
  - Regular competitions
  - Challenging problem statements. Evaluation Metrics
- Advances in Computational Resources
  - GPUs
  - Industrial scale clusters
  - Well established libraries, community

Introduction to Machine Learning

Classification

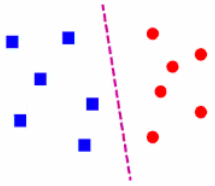


f(x) = label of the training example nearest to **x**

- All we need is a distance function for our inputs
- No training required!

Nearest Neighbour

Linear Classifier

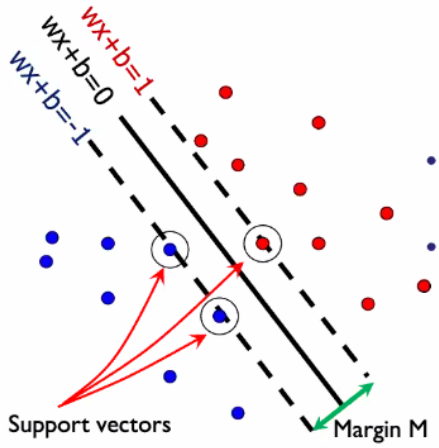


- Find a *linear function* to separate the classes:

f(x) = sgn(**w** · **x** + b)

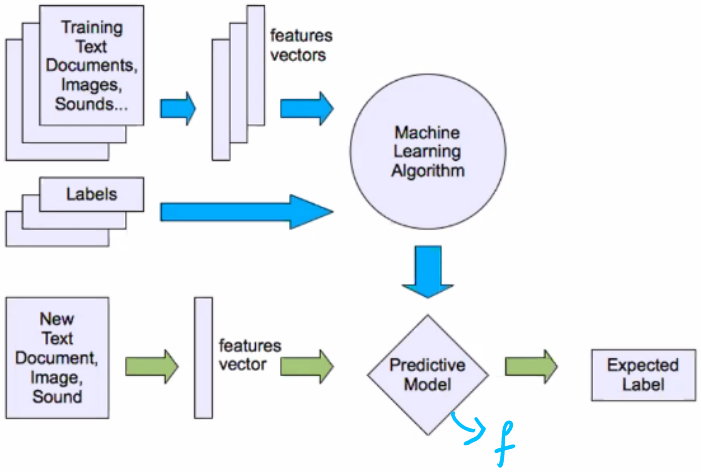
→ Linear Regression

Support Vector Machines



- SVM: maximizes the margin.
  - Convex Optimization
- Other Algorithms and Tricks
  - Random Forests
  - Ensembling

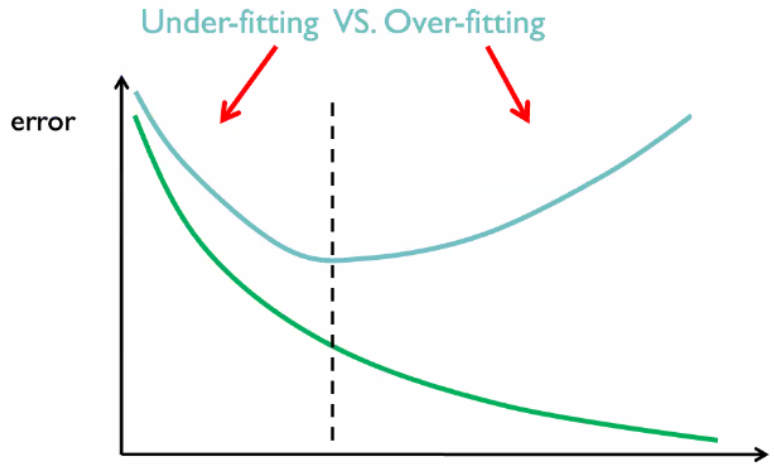
Supervised Learning



Some Key Words

- ***Training:*** Find f and W
- ***Testing:*** Evaluate f on a specific example
- Training, Testing and Validation splits of the data
- ***Generalization:*** Goal is to do well on "unseen data"
- Error, ***Loss,*** Objective Functions
- Complexity of the solution (eg. Number of free parameters)
- Generative classifiers try to model the data. Discriminative classifiers try to predict the label.

Worry



* Most crucial — Hyperparam-eters tuning

(Dis)Advantage of the Data Driven Solutions

Good/ Appropriate **Examples** → **Machine Learning Algorithms** → Data Driven **Solutions**

Train

Learn from (large) realistic examples

Validation

Evaluate on realistic data

Test

Statistics on real data tells us how good is the solution

Challenges

- Your solution is only "as good as" your data
  - Bias, Generalization, (Fairness?)
- Need of many many examples
  - Few shot learning, Transfer learning, incremental
- Is Supervised Learning everything?
  - Unsupervised, RL, Self-Supervised

* limited data or single distribution of data.

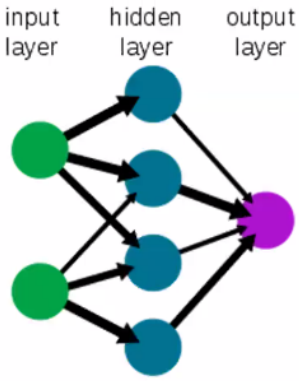
* If few or examples available.

→ other types of learning.

Neural Networks and Deep Learning

Neural Networks

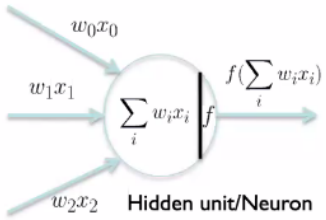
A simple neural network



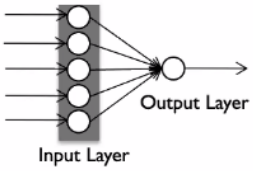
- Biologically inspired networks.
- Complex function approximation through composition of functions.
- Can learn arbitrary Nonlinear decision boundary

Biological N/w's Resemblence to Neural Networks → Loose Defination.

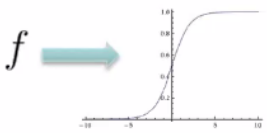
Neuron, Perceptron and MLP



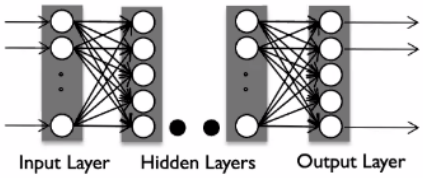
Hidden unit/Neuron



Perceptron

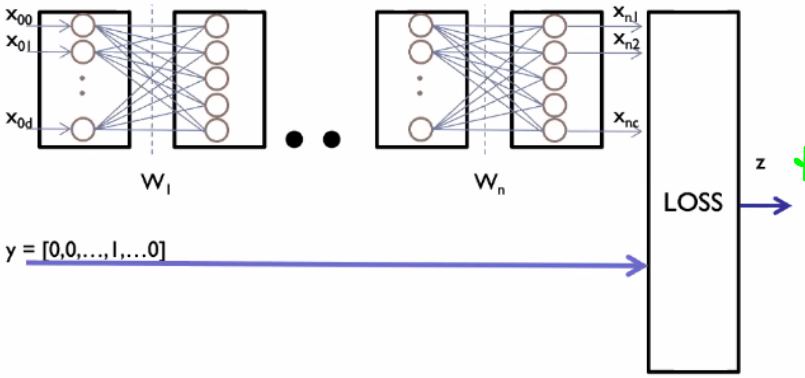


E.g. Sigmoid Activation Function



Multi Layer Perceptron

Neural Networks and Learning

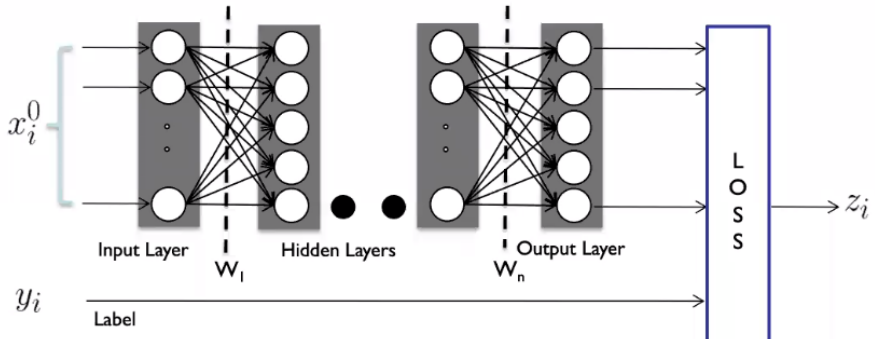


* We have to minimize this loss.

Loss = $(x_{ni} - y_i)$ {say}

* Weight formula.

Loss or Objective

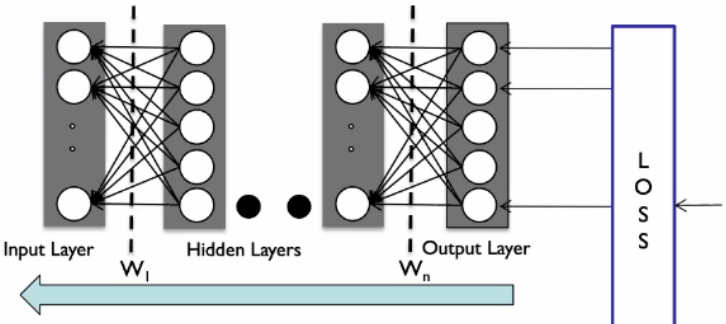


Objective: Find out the best parameters which will minimizes the loss.

$$W^* = arg\min_W \sum_{i=1}^{N} L(x_i^n, y_i; W)$$ → Weight Vector

$$z_i = \frac{1}{2} \parallel x_i^n - y_i \parallel_2^2$$ E.g. Squared Loss

Back propagation

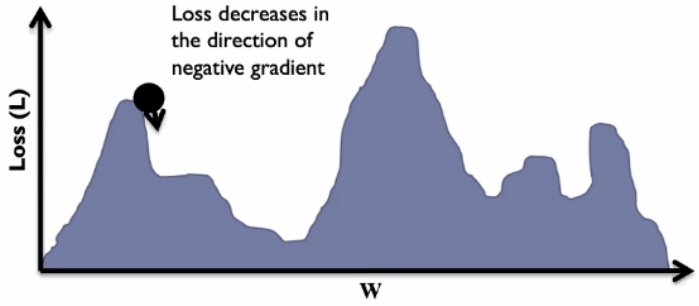


Solution: Iteratively update W along the direction where loss decreases.

Each layer weights are updated based on the derivative of its output w.r.t. input and weights

Gradient Descent

- Visualization of loss function



$$W_{i+1} = W_i - \eta \times \frac{\partial L}{\partial W}$$ Parameter update

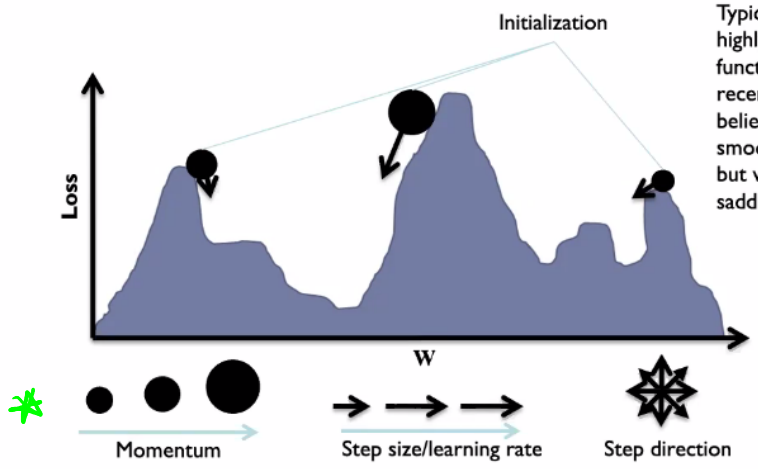
* Initialization
* Momentum
* Speed etc.

are imp. factors

* More parameters will give a more complex mapping.

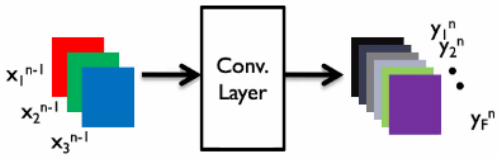
Training

- Visualization of loss function



Typically viewed as highly non-convex function but more recently it's believed to have smoother surfaces but with many saddle regions !

Convolutional Layer



$$y_j^n = f(\sum_{k=1}^{F} x_k^{n-1} * w_{kj}^n)$$

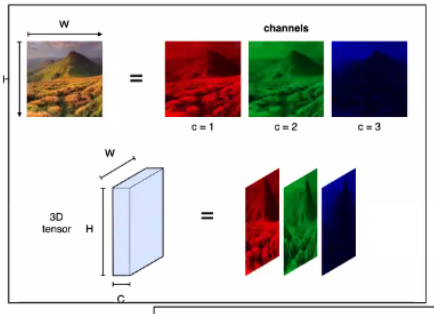
Here "f" is a non-linear activation function.
F= no. of feature maps
n= layer index
"*" represents element-by-element multiplication

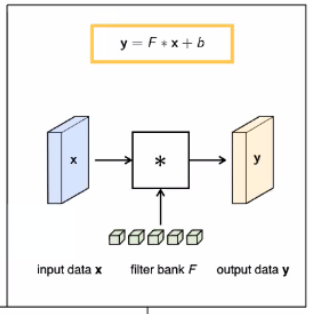
→ FC N/w invites a lot of parameters making the process complex ≈ ∴ CNN

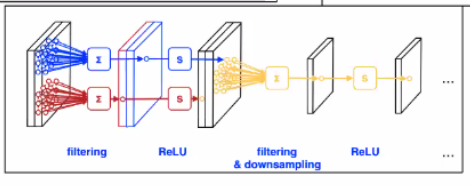
Fully Connected

Specially the high dimensional data

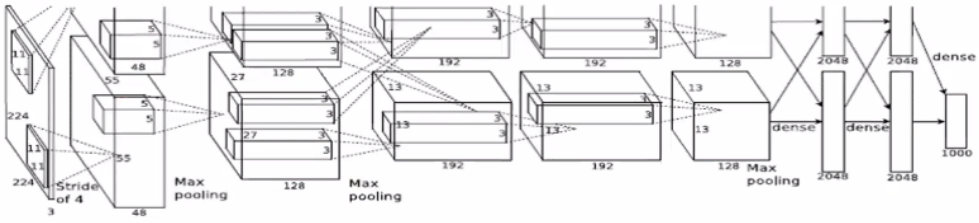
Convolutional Neural Networks







AlexNet (NIPS 2012)



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

ImageNet Classification Task:

Previous Best : ~25% (CVPR-2011)
AlexNet : ~15 % (NIPS-2012)

Breakthrough in CNN.

What changed over time?

	LeNet(1989)	LeNet(1998)	AlexNet(2012)
Task	Digit	Digit	Objects
# Classes	10	10	1000
image size	16 × 16	28 × 28	256 × 256 × 3
# examples	7291	60,000	1.2 M
units	1256	8084	658,000
parameters	9760	60K	60 M
connections	65K	344K	652M
Operations	11 billion	412 billion	200 quadrillion

* NN Based

Indeed, Many Changes

Regularization
- DropOut, DropConnect, Batch Normalization, Data Augmentation, Noise in Data/Label/Gradient

Weight Initialization
- Xavier's initialization, He's initialization

Choosing Gradient Descent Parameters
- Adagrad, RMSProp, Adam, Momentum, Nesterov Momentum

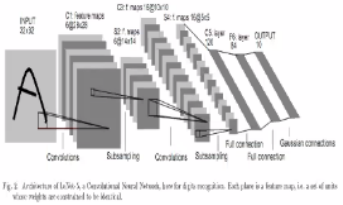
Activation Functions
- ReLU, PReLU, Leaky ReLU, ELU

Loss Functions
- Cross-Entropy, Embedding Loss, Mean-Squared Error, Absolute Error, KLDivergence, Max-Margin Loss

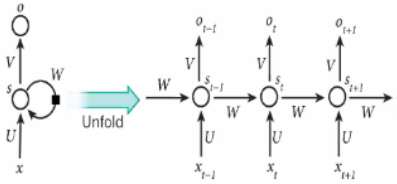
Little Pieces that have made the Whole

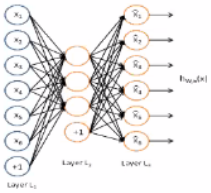
Popular DL Architectures

CNN



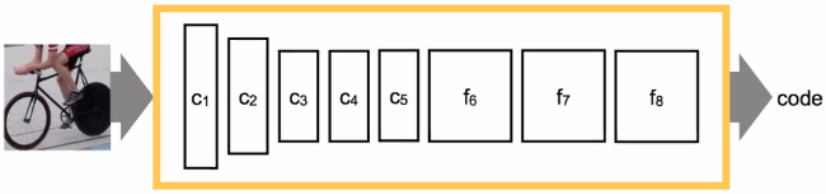
RNN





Auto Encoder

CNN Features are Generic



CNN Features can be used for wider applications:

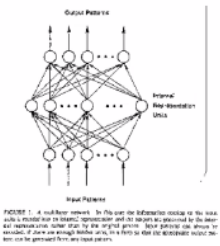
1. **Train the CNN (deep network) on a very large database such as imagenet.**
2. **Reuse CNN to solve smaller problems**
   1. **Remove the last layer (classification layer)**
   2. **Output is the code/feature representation**

Representation Learning

NN was always learning features!

Learning Internal Representations by Error Propagation

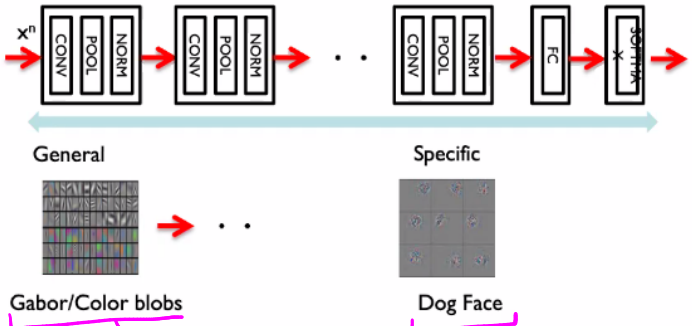
D. E. RUMELHART, G. E. HINTON, and R. J. WILLIAMS



Rumelhart's classical paper on error backpropagation, 1986

Transfer Learning

- A key observation that we noticed in visualization:-



Yosinski J, Clune J, Bengio Y, and Lipson H. How transferable are features in deep neural networks? NIPS '14

* Used when we have fewer examples in our problem set.

Fine features

Easier features (more abstract)

Unsupervised Learning

Unsupervised Learning

- "We expect unsupervised learning to become far more important in the longer term. Human and animal learning is largely unsupervised: we discover the structure of the world by observing it, not by being told the name of every object."
  - LeCun, Bengio, Hinton, Nature 2015

- As I've said in previous statements: most of human and animal learning is unsupervised learning. If intelligence was a cake, unsupervised learning would be the cake, supervised learning would be the icing on the cake, and reinforcement learning would be the cherry on the cake. We know how to make the icing and the cherry, but we don't know how to make the cake.
  - Yann LeCun, March 14, 2016 (Facebook)

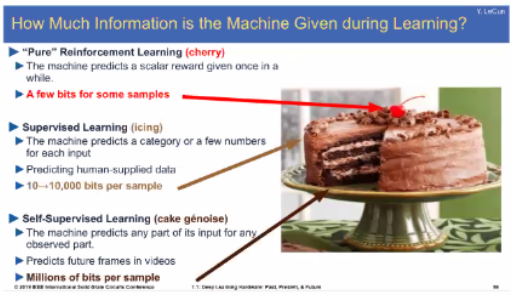
Old and New

2016

- "Pure" Reinforcement Learning (cherry)
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- Supervised Learning (icing)
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- Unsupervised/Predictive Learning (cake)
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

(Yes, I know, this picture is slightly offensive to RL folks. But I'll make it up)

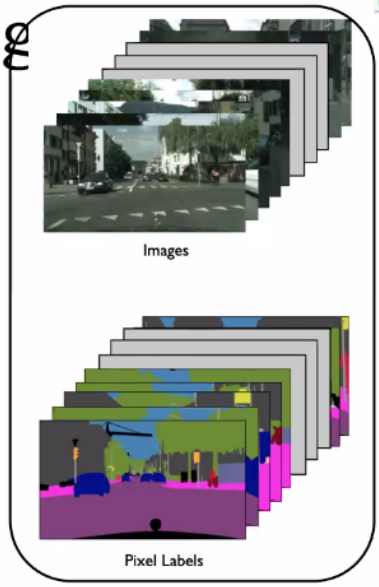
2019



Source: Y. LeCun at NIPS 2016

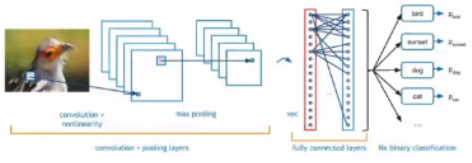
Supervised Learning

- Input space: $\mathcal{X}$
- Label space: $\mathcal{Y}$
- Input set: $X = (x_1, x_2, \cdots, x_n)$
- Label set: $Y = (y_1, y_2, \cdots, y_n)$
- Function: $f : \mathcal{X} \longrightarrow \mathcal{Y}$
- Loss function: $l : \underbrace{\mathcal{Y}}_{\text{Predicted Label Space}} \times \underbrace{\mathcal{Y}}_{\text{True Label Space}} \longrightarrow \underbrace{\mathbb{R}^{\geq 0}}_{\text{Loss Value}}$
  - Associates the value of predicted and true label with a cost
  - Used to estimate parameters of the model

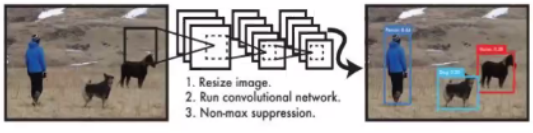


* Not enough → As we might not have sufficient examples.

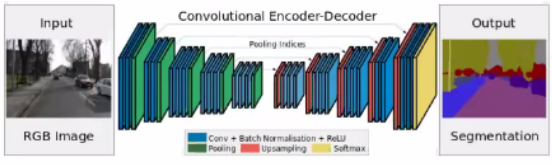
Supervised Models



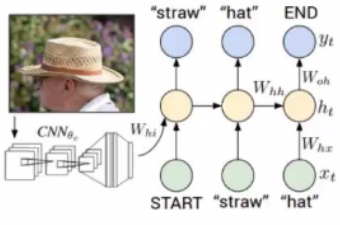
(a) Image Classification



(b) Object Detection



(c) Semantic Segmentation

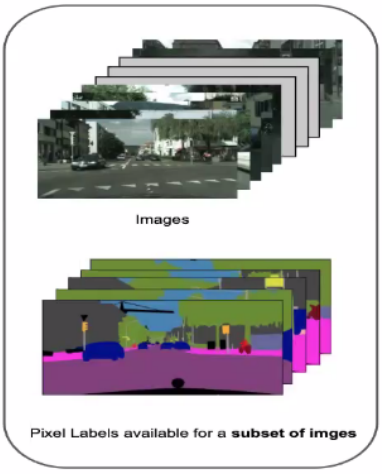


(d) Image Captioning

→ known labels.

Semi-Supervised Learning

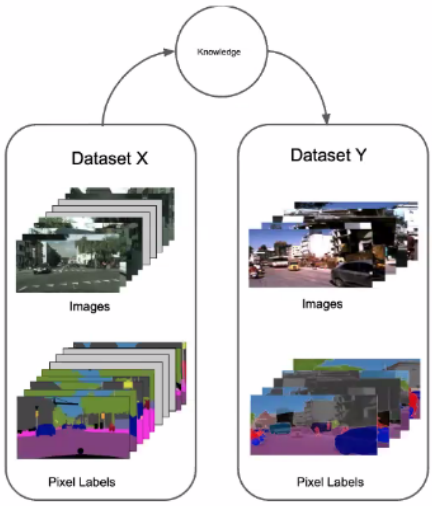
- Input set: $X = (x_1, x_2, \cdots, x_n)$
- Label set: $Y = (y_1, y_2, \cdots, y_m)$
- In semi-supervised setting ***m<<n***
- Self-training:
  - Produce proxy labels on unlabelled data
  - Use these labels along with labelled data
  - Disdavantage: Model unable to corrent it's mistakes
- Co-training:
  - Two models learn from each other's mistakes



Are only human annotated data useful? Should we discard others? NO!!

Transfer Learning

- Source domain space: $\mathcal{X}_s$
- Target domain space: $\mathcal{X}_t$
- Source model: $f_s : \mathcal{X}_s \longrightarrow \mathcal{Y}$
- Target model: $f_t : \mathcal{X}_t \longrightarrow \mathcal{Y}$
- Transfer of knowledge
  - Domain transfer
    - European roads to Indian roads
  - Task transfer
    - Classification to object detection
- Approach
  - Train a network for f_s
  - Reuse the model to initialize f_t
  - Fine-tune f_t

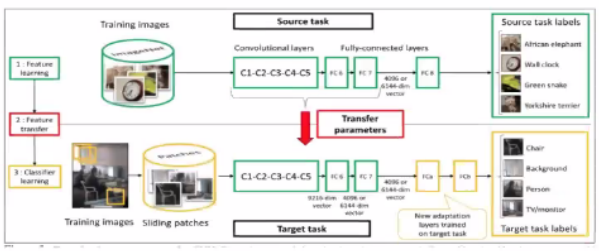


When we have insufficient data? ↑

Transfer Learning Models

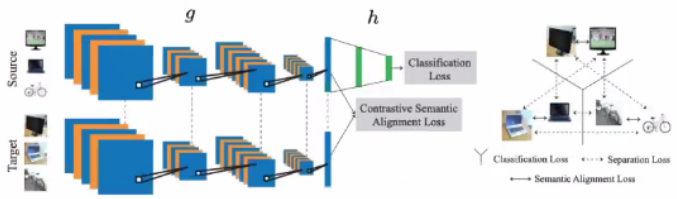
Different label space

$\mathcal{Y}_s \neq \mathcal{Y}_t$



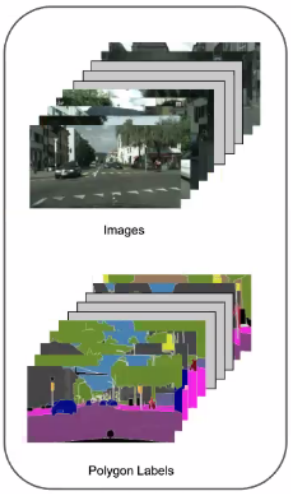
Different input space

$\mathcal{X}_s \neq \mathcal{X}_t$

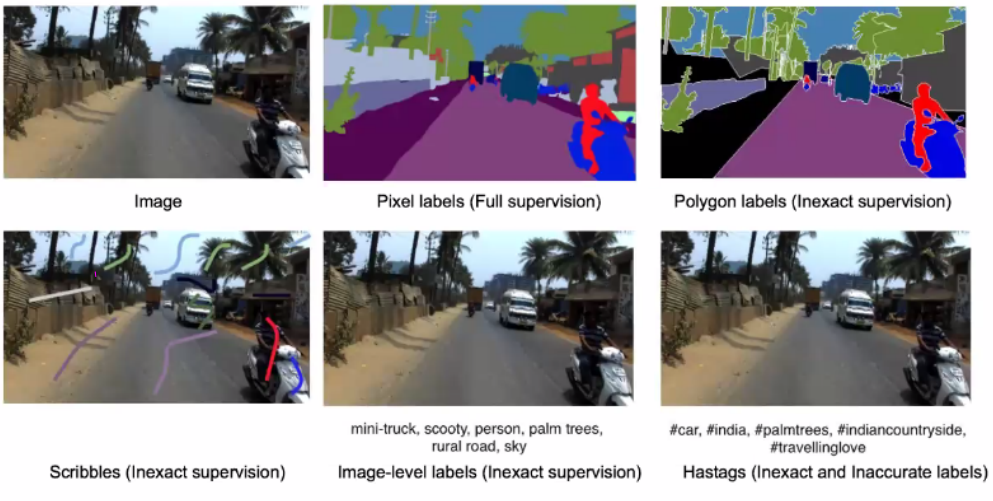


Weak Supervision

- Labels
  - Inaccurate/Noisy
    - Webly-supervised
  - Inexact
    - Heuristics
    - Distant supervision
  - Incomplete
    - A small subset of labels
- Multiple Instance Learning (MIL)
  - Bags: Images
  - Instances: Windows



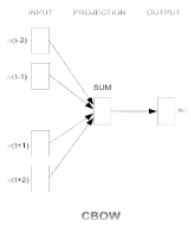
Weakly Supervised Models

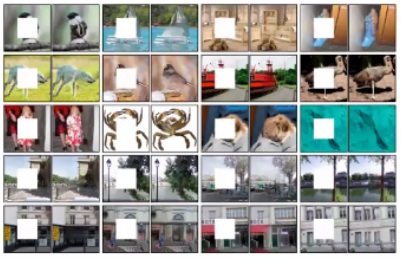


* The amount, style, design, labelling style are the variations of data available

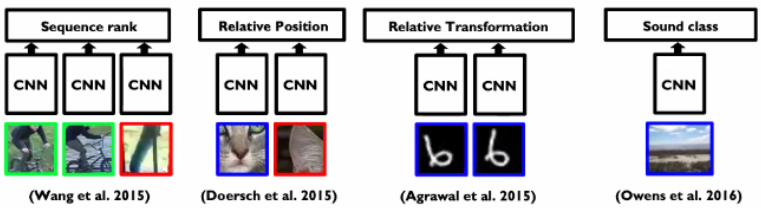
Self Supervised Learning

Word2Vec
Mikolov 2013



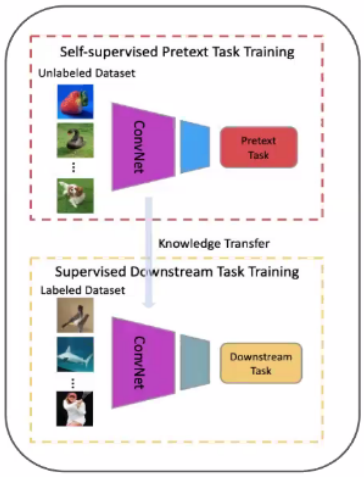


Pathak et al, 2016



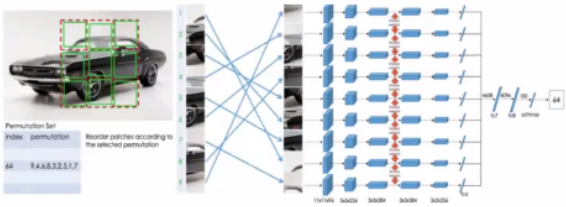
Self Supervised Learning

- Utilize naturally occuring “free” information contained within the data
- No need of explicit labels pertinent to the task
- Pretext tasks:
  - Color
  - Spatial order
  - Temporal order
  - Sound
  - Motion

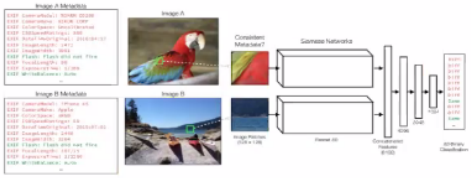


Self Supervised Models

- Spatial Context as a pretext task for feature learning
  - Learn a function to solve jigsaw puzzle
- EXIF information as a pretext task to detect image tampering
  - Each image has EXIF information
    - Camera model, ISO, shutter speed etc
  - Patches from same image → Similar EXIF
  - Patches from different images → Dissimilar EXIF
  - Training conducted using patches from untampered images
  - At test time, tampered image patches → Dissimilar EXIF values



(a) Spatial Context Prediction as proxy task for feature learning

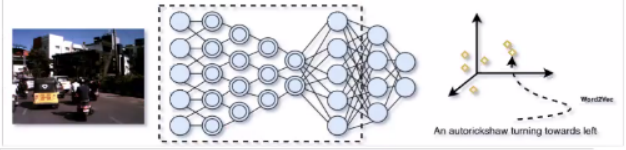


(b) EXIF information consistency as a proxy task for image tamparing

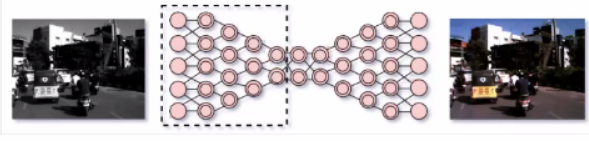
Comparison: Example

Full Supervision
(Bounding box)

Weak Supervision
(Caption as label)



Self-supervision
(Colorization as pretext task)



An autorickshaw turning towards left.

Check the keywords:

Summary: Harder and Harder Tasks

- Classification (2, 10s, 100s, 1000s, Hierarchical)
- Detection (Face, Rigid, Deformable Objects)
- Pixel-level Segmentation (two, multi, labels)
- Detection, Body parts, Human Pose, Human Actions
- Semantics: Nouns, Verbs, Semantic structure
- Annotations (Tags, phrases, captions)
- Visual Question Answering
- Summarization and Synthesis
- Predicting intend
- ....

Summary: DL

- CNNs and Image Classification
- Efficient Optimization
- CNN Features and Fine Tuning
- RNNs and Sequence to Sequence
- Vision + Language
- FCNs and Semantic Segmentation, Image to Image Models
- GANs and VAEs, Advserserial Training
- Transfer and Incremental Learning
- Few Shot Learning
- Deep Probabilistic Models
- Beyond Supervised Learning, Self and Unsupervised Learning
- ??