Computer Vision Basics

In the Context of Robot Perception

Prof. Yunzhu Li

Spring 2024

CS598YL: Deep Learning for Robotic Manipulation

* Course materials adapted from Prof. Yuke Zhu's CS391R at UT Austin.

Logistics

- Website is ready (https://robopil.github.io/cs598yl/24sp/)

CS598YL: Deep Learning for Robotic Manipulation

Spring 2024

Course Description

In the modern economy, specialized robots are revolutionizing industries by enhancing efficiency, safety, and product quality. However, their potential is curtailed by their inability to adapt to diverse, unstructured settings typical of everyday life. This course emphasizes robots' capacity to perceive, decide, and act in varied situations using their sensors and intelligent algorithms, to achieve various robotic manipulation tasks across different horizons, encountering objects made of diverse materials – from rigid to deformable, granular to cloth. It explores advanced subjects focused on 1) how robots interpret complex, unstructured environments using raw sensory data, moving beyond pre-programmed scenarios to dynamic real-world contexts? 2) how robots make informed choices based on their understanding of their surroundings? 3) how robots learn and adapt continuously, evolving their capabilities in response to the physical world. The course will be in a seminar style, including paper reviews, paper presentations, and hands-on projects.

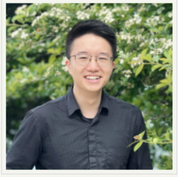
Course Time and Location

Lecture: 3:30 - 4:45 PM, Tuesdays and Thursdays
Location: Campus Instructional Facility | Room 1035

Online Platforms

Campuswire for announcements and discussions
Canvas for grade management

Instructor



Yunzhu Li
Office Hour: Fri 9:00-10:00 AM
Appointment: Calendly
Website

Teaching Assistant

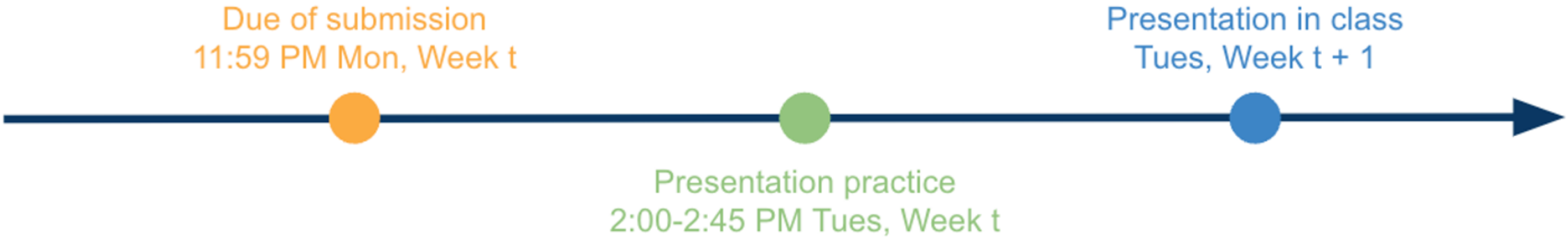
Mingtong Zhang
Office Hour: Mon 9:00-10:00 AM
Appointment: Calendly
Website

Logistics

- Website is ready (https://robopil.github.io/cs598yl/24sp/)
- Use Calendly to sign up office hour
- Use Campuswire for discussion
- Use Canvas to submit paper reviews, project proposal, milestones, etc.

Logistics

- Sign up presentation preference form
  - Deadline: Thur 11:59 PM, Jan 18 (Tonight!)
- The first week presentation submission deadline is Mon 11:59 PM Jan 22.
- Presentation practice: 2:00-2:45PM Tues and Thur with the instructor
  - Around 15 mins for each presentation
- Each student is required to present one paper



Today’s Agenda

- What is Robot Perception?
- Robot Vision vs. Computer Vision
- Landscape of Robot Perception
- Quick Review of Deep Learning for Computer Vision

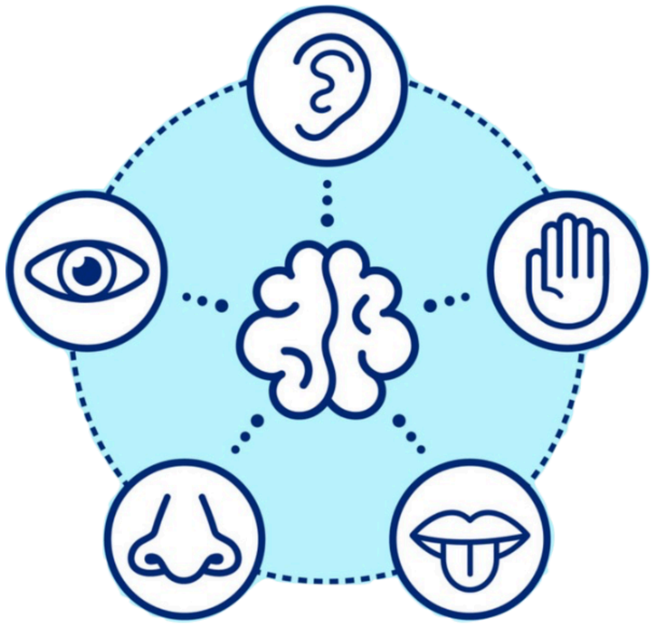
What is Robot Perception?

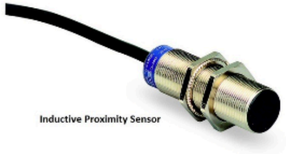
Making sense of the unstructured real world …

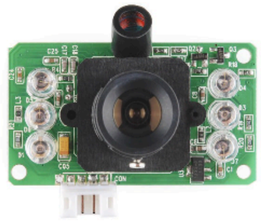
- Incomplete knowledge of objects and scenes
- Imperfect actions may lead to failure
- Environment dynamics and other agents

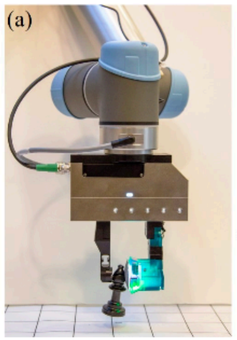
Sensors for Robotics

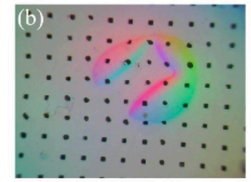
Understanding the interactions with world through multimodal senses

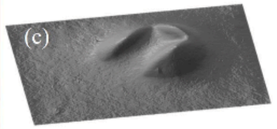






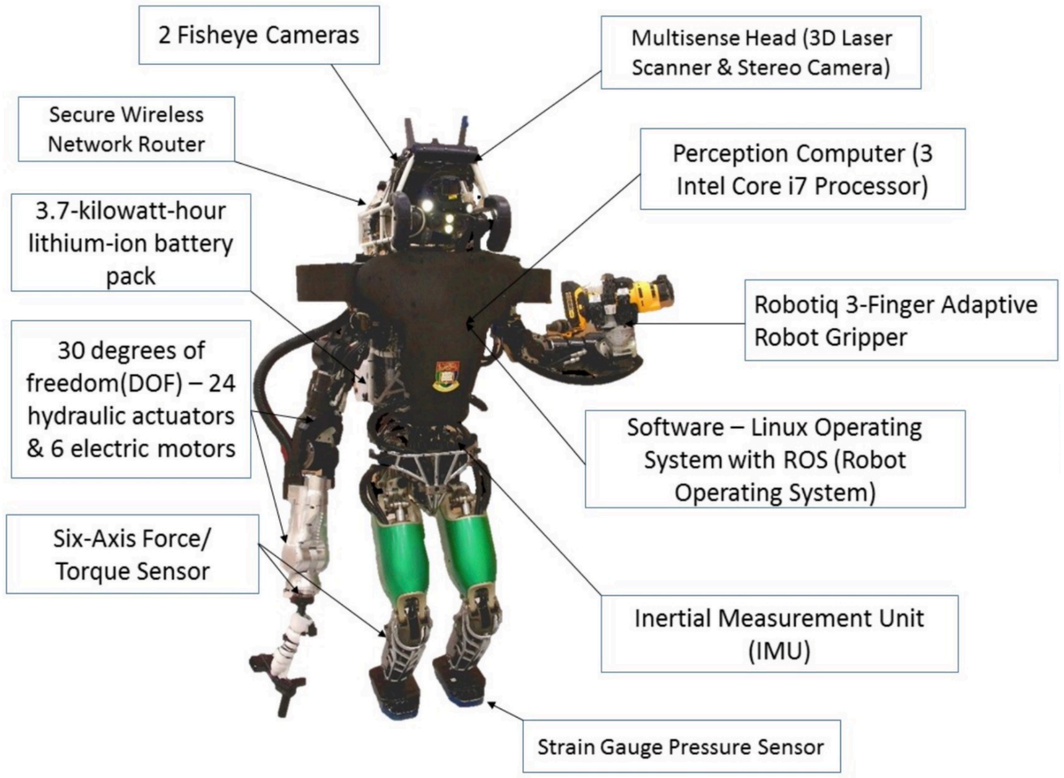


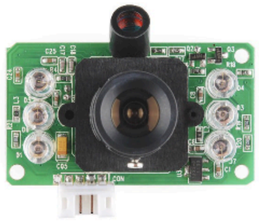


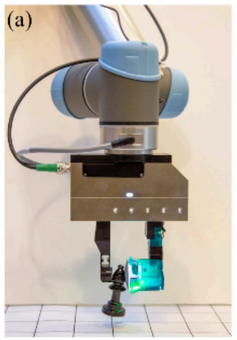


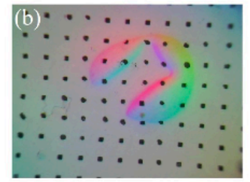
Sensors for Robotics

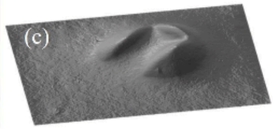
Understanding the interactions with world through multimodal senses







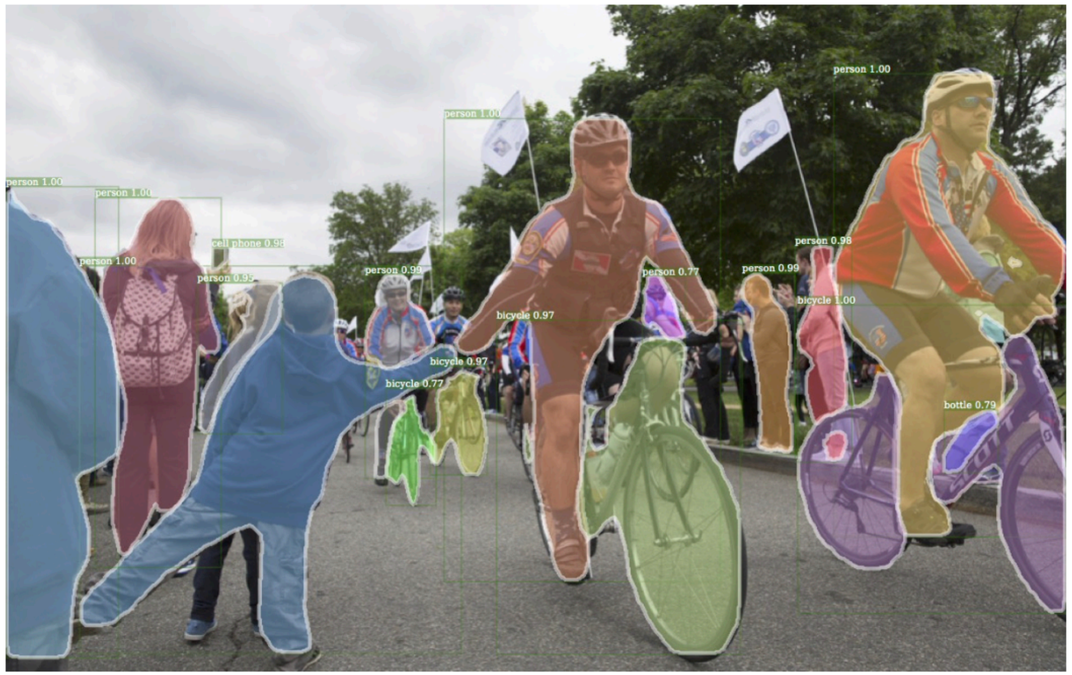




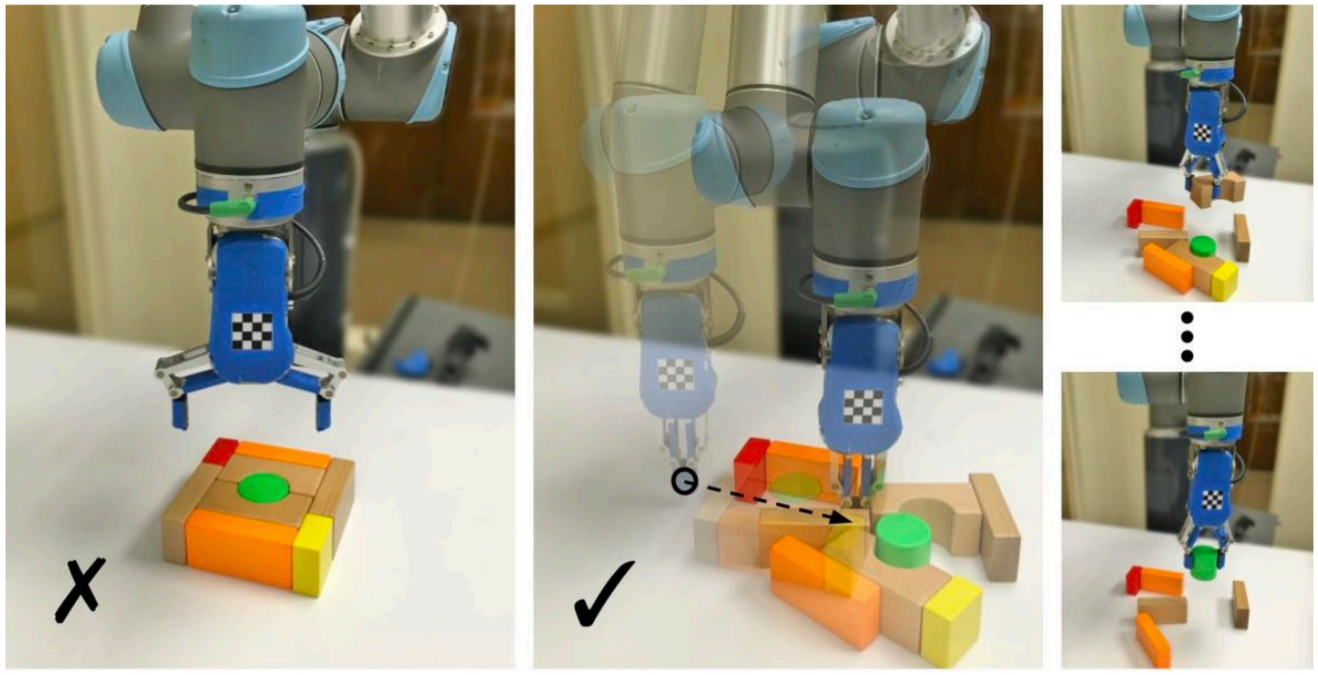
[Source: HKU Advanced Robotics Laboratory]

Robot Vision vs. Computer Vision

Robot vision is **embodied**, **active**, and **environmentally situated**.



[Detectron - Facebook AI Research]



[Zeng et al., IROS 2018]

Robot Vision vs. Computer Vision

Robot vision is **embodied**, **active**, and **environmentally situated**.

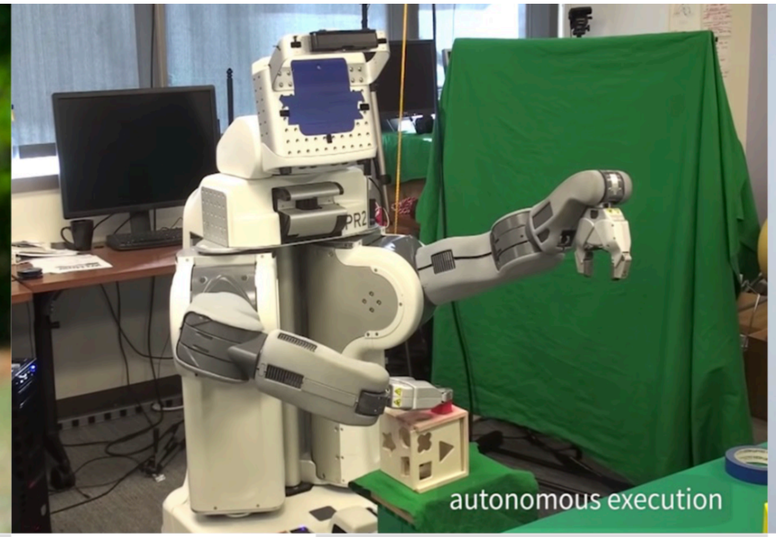
- **Embodied:** Robots have physical bodies and experience the world directly. Their actions are part of a dynamic with the world and have immediate feedback on their own sensation.

- **Active:** Robots are active perceivers. It knows why it wishes to sense, and chooses what to perceive, and determines how, when and where to achieve that perception.

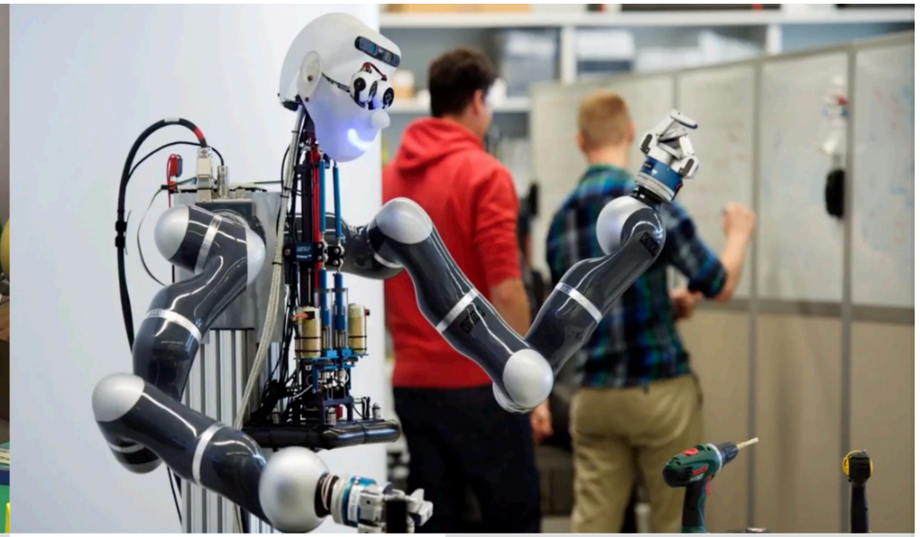
- **Situated:** Robots are situated in the world. They do not deal with abstract descriptions, but with the “here” and “now” of the world directly influencing the behavior of the system.

The Perception-Action Loop

[Sa et al. IROS 2014]

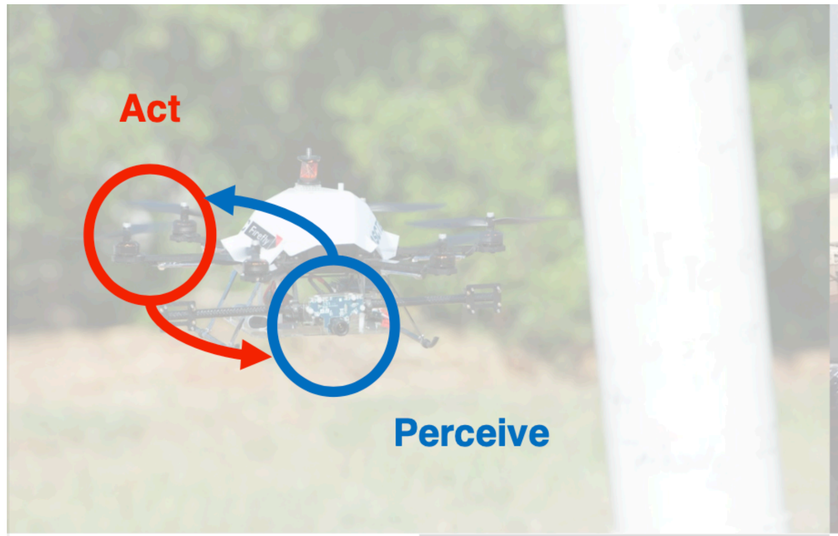


[Levine et al. JMLR 2016]

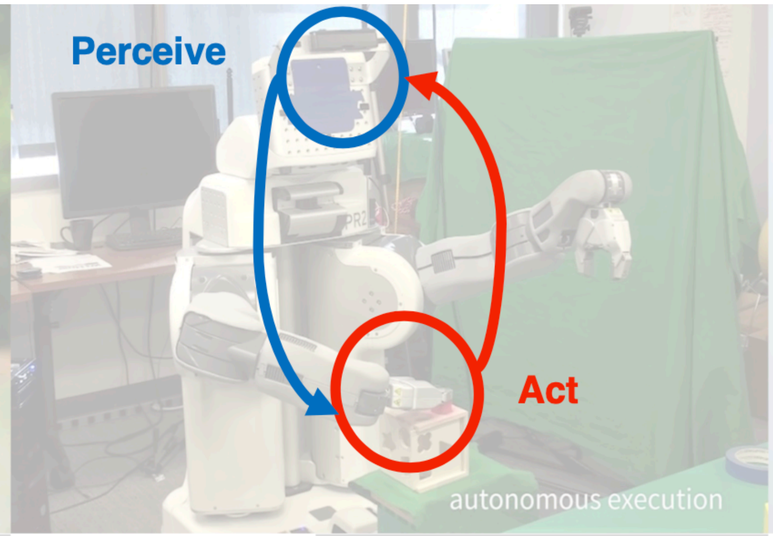


[Bohg et al. ICRA 2018]

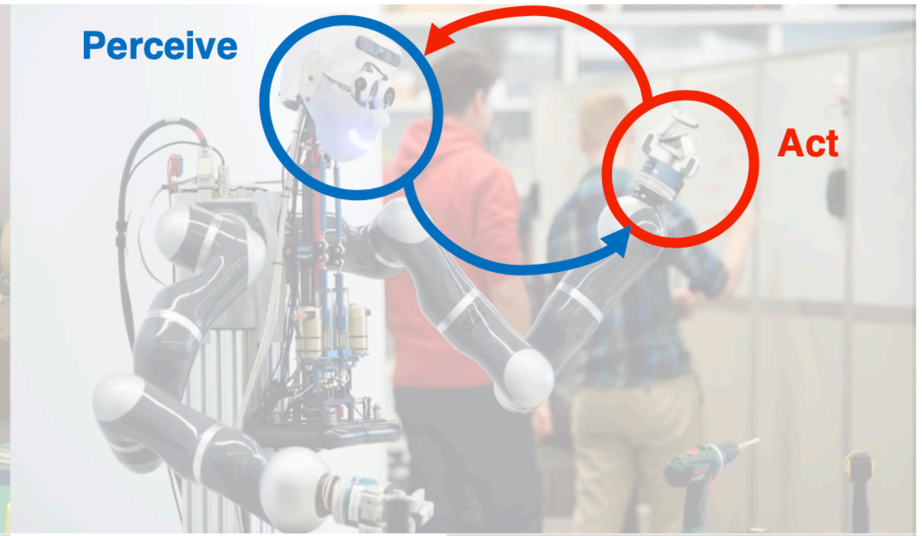
The Perception-Action Loop



[Sa et al. IROS 2014]



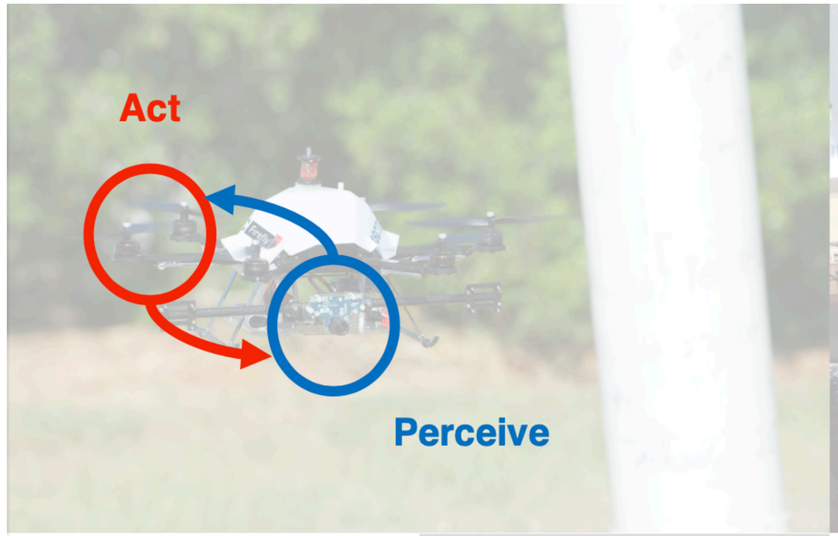
[Levine et al. JMLR 2016]



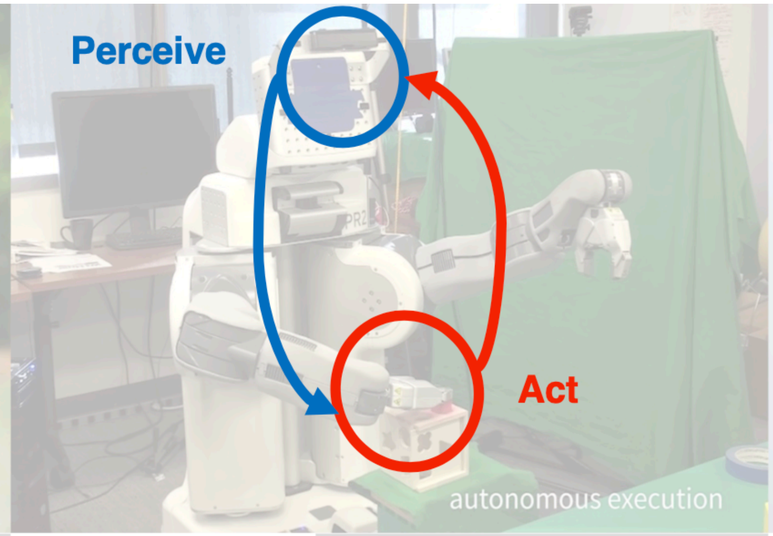
[Bohg et al. ICRA 2018]

The Perception-Action Loop

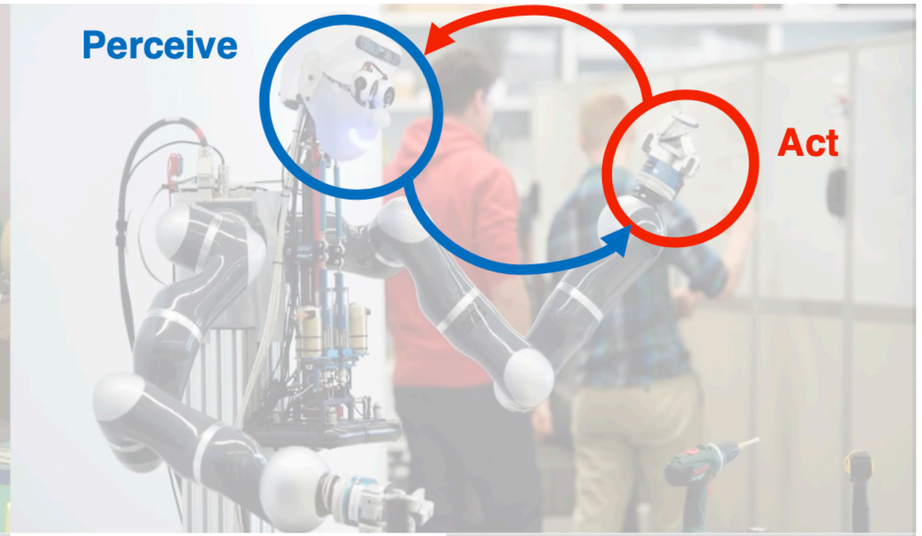
A key challenge in **Robot Learning** is to close the **perception**-**action** loop.



[Sa et al. IROS 2014]



[Levine et al. JMLR 2016]



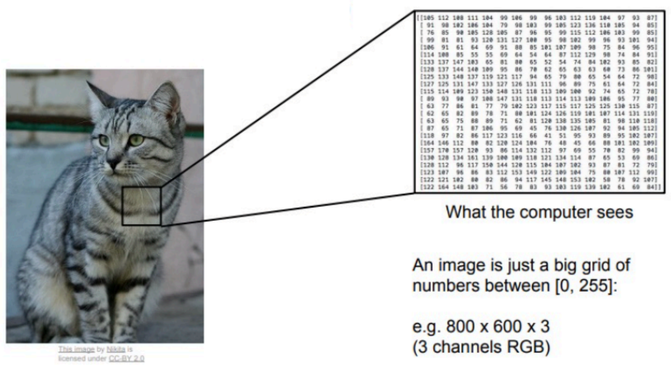
[Bohg et al. ICRA 2018]

Robot Perception: Landscape

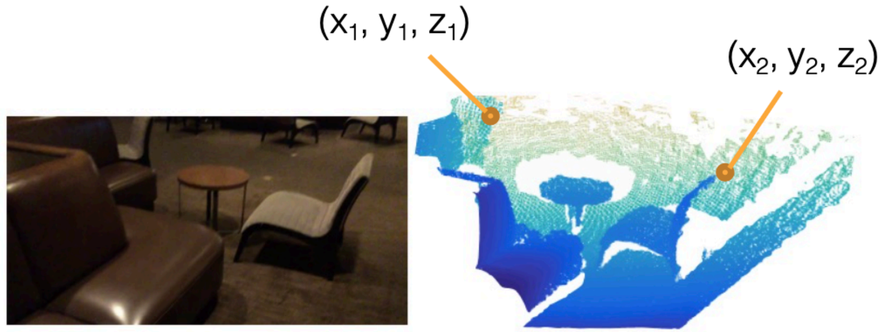
What you will learn in: Robot Perception

- **Modalities:** neural network architectures designed for different sensory modalities
- **Representations:** representation learning algorithms without strong supervision
- **Tasks:** state estimation tasks for robot navigation and manipulation
- **Cutting-Edge Topics:** embodied visual learning & synthetic data for visual AI

Robot Perception: Modalities

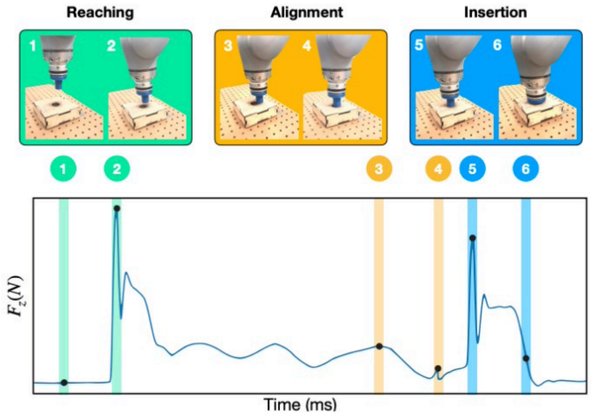


Pixels (from RGB cameras)



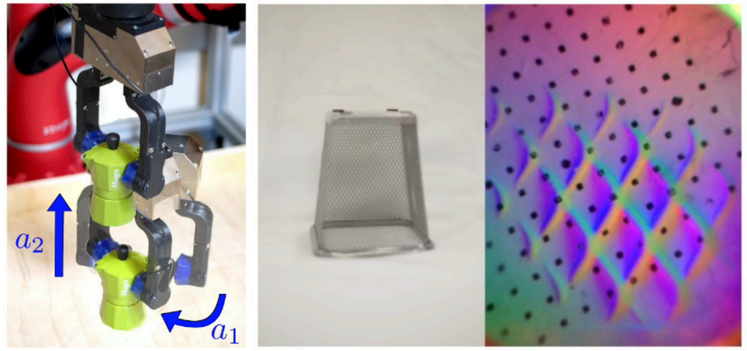
[Source: PointNet++; Qi et al. 2016]

Point cloud (from structure sensors)



[Source: Lee*, Zhu*, et al. 2018]

Time series (from F/T sensors)

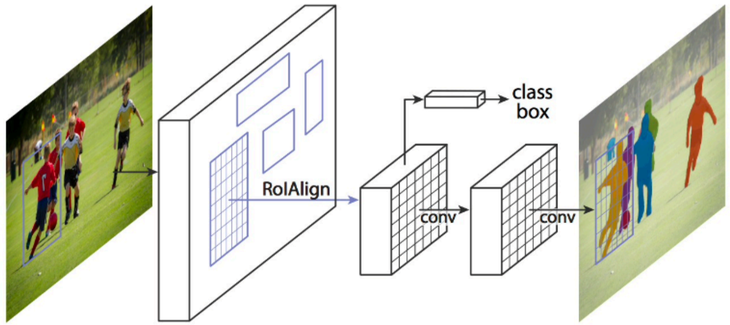


[Source: Calandra et al. 2018]

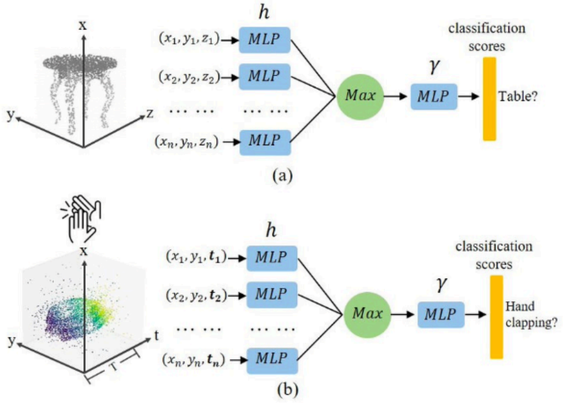
Tactile data (from the GelSights sensors)

Robot Perception: Modalities

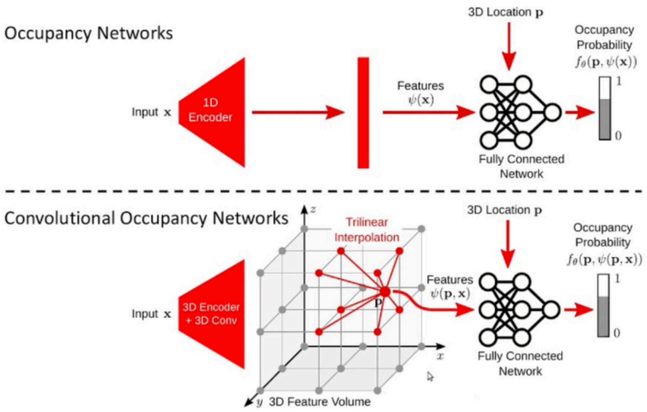
How can we design the **neural network architectures** that can effectively process raw sensory data in vastly different forms?



2D Perception



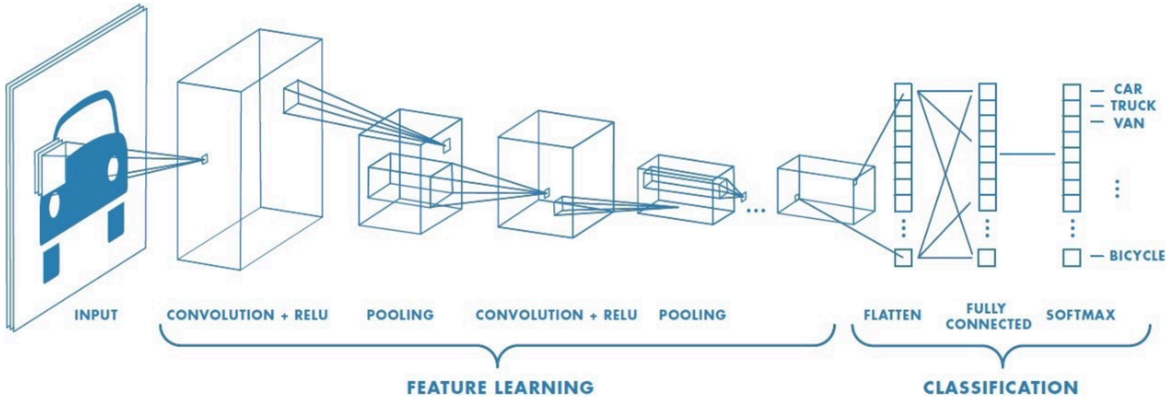
3D Perception



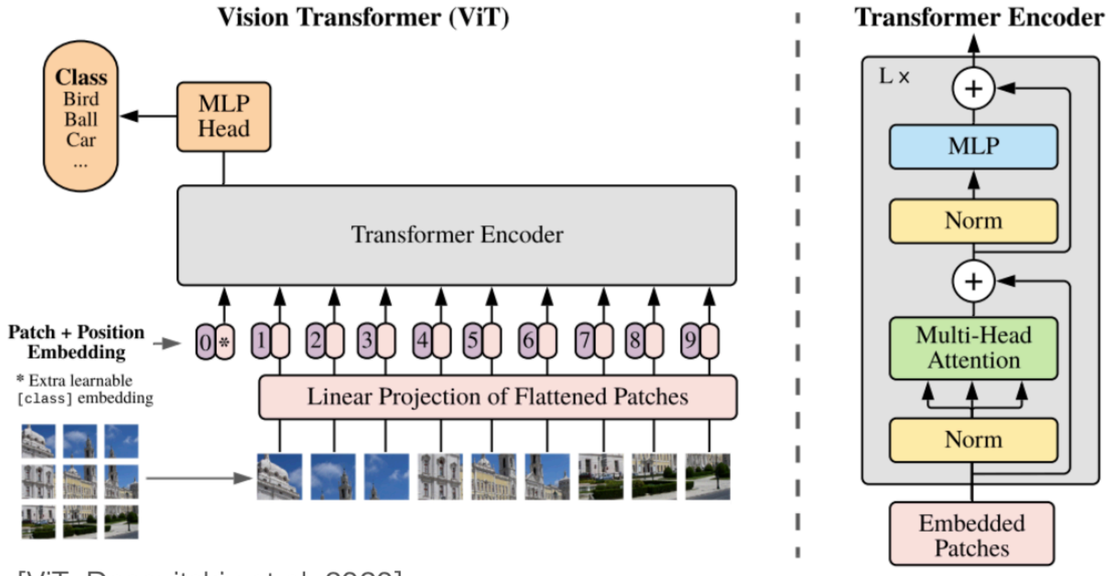
Neural Fields

Robot Perception: Modalities

How can we design the **neural network architectures** that can effectively process raw sensory data in vastly different forms?



[Credit: Sumit Saha]



[ViT, Dosovitskiy et al. 2020]

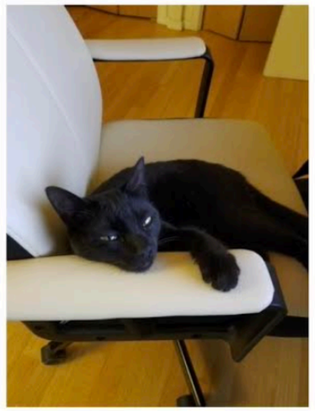
Convolution Architectures

Attention Architectures

Robot Perception: Representations

A fundamental problem in robot perception is to learn the proper **representations** of the unstructured world.

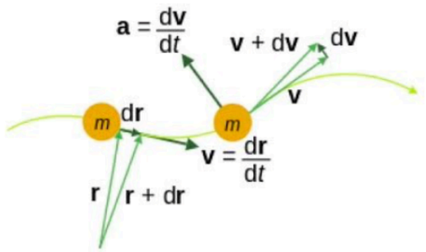
Things...

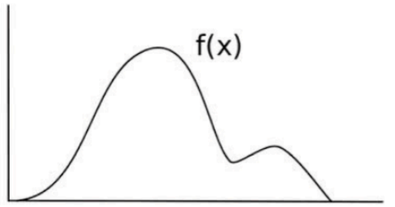


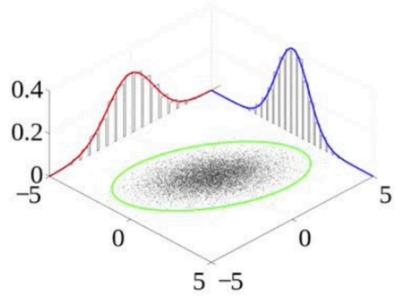
My heart beats as if the world is dropping, you may not feel the love but i do its a heart breaking moment of your life. enjoy the times that we have, it might not sound good but one thing it rhymes it might not be romantic but i think it is great,the best rhyme i've ever heard.

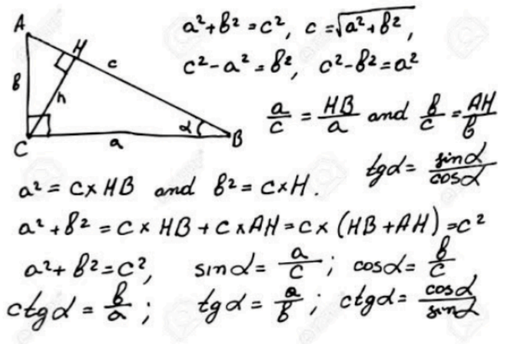
Representation

Engineering Knowledge...









[Source: Stanford CS331b]

Robot Perception: Representations

"Solving a problem simply means representing it so as to make the solution transparent."

Herbert A. Simon, Sciences of the Artificial

Our secret weapon? **Learning**

Robot Perception: Representations

"Solving a problem simply means representing it so as to make the solution transparent."

Herbert A. Simon, Sciences of the Artificial

What representation to learn? How to learn them?

Google Scholar

Top publications

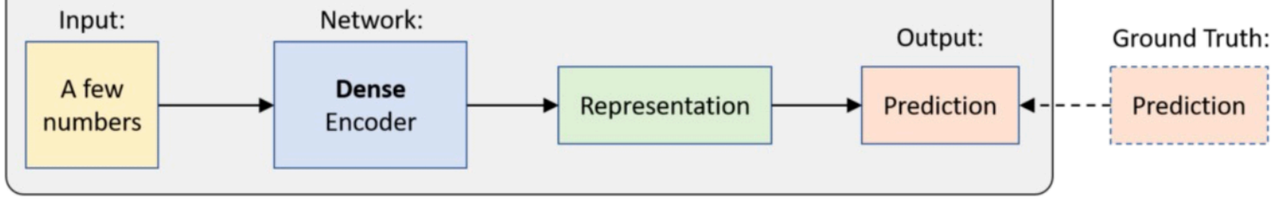
Categories ▾ English ▾

	Publication	h5-index	h5-median
1.	Nature	467	707
2.	The New England Journal of Medicine	439	876
3.	Science	424	665
4.	IEEE/CVF Conference on Computer Vision and Pattern Recognition	422	681
5.	The Lancet	368	688
6.	Nature Communications	349	456
7.	Advanced Materials	326	415
8.	Cell	316	503
9.	Neural Information Processing Systems	309	503
10.	International Conference on Learning Representations	303	563

International Conference on Learning Representation (ICLR)
2013-present

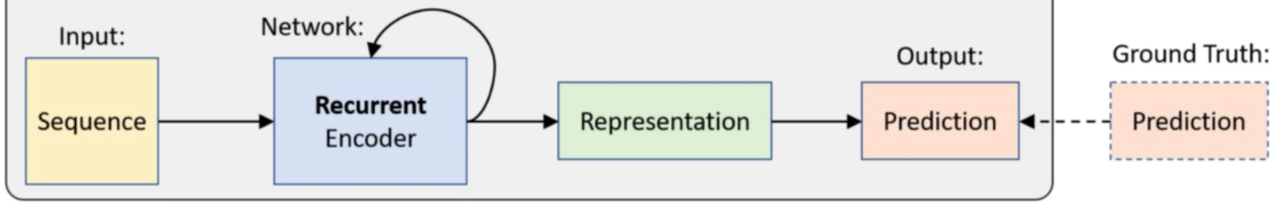
Supervised Learning

1. Feed Forward Neural Networks

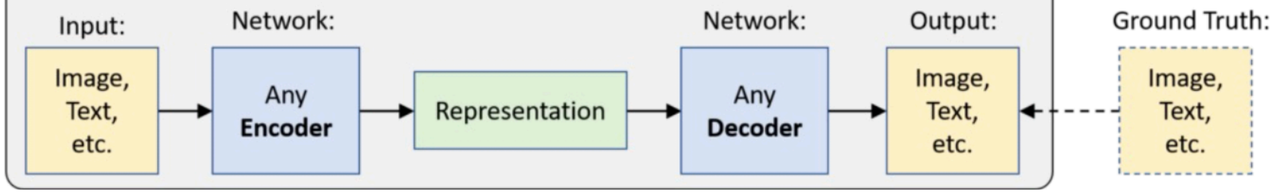


2. Convolutional Neural Networks

3. Recurrent Neural Networks

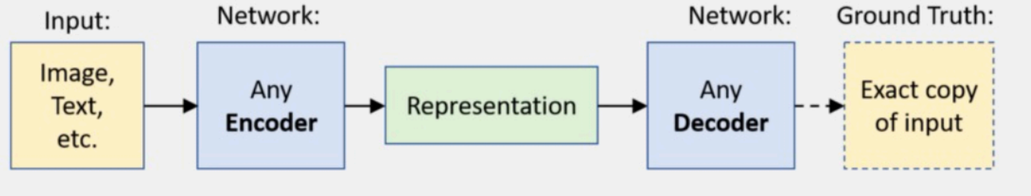


4. Encoder-Decoder Architectures

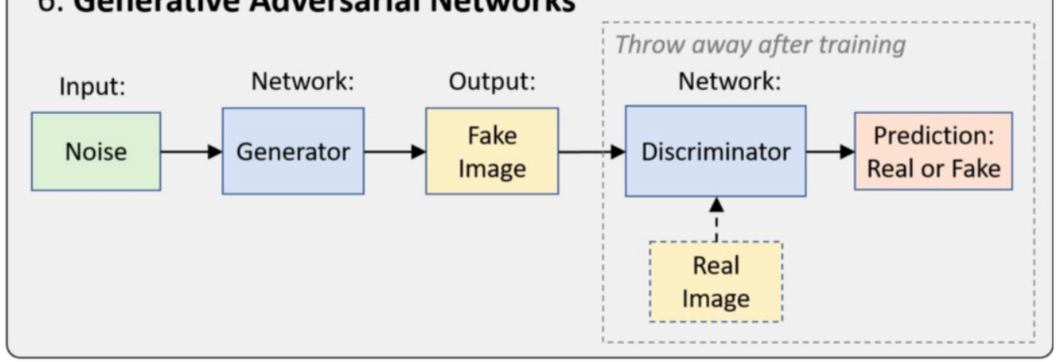


Unsupervised Learning

5. Autoencoder

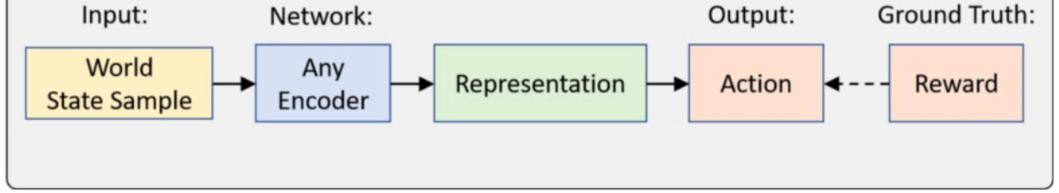


6. Generative Adversarial Networks



Reinforcement Learning

7. Networks for Actions, Values, Policies, and Models



[Source: 6.S094, MIT]

Robot Perception: Representations

How can we learn **representations of the world** with limited supervision?

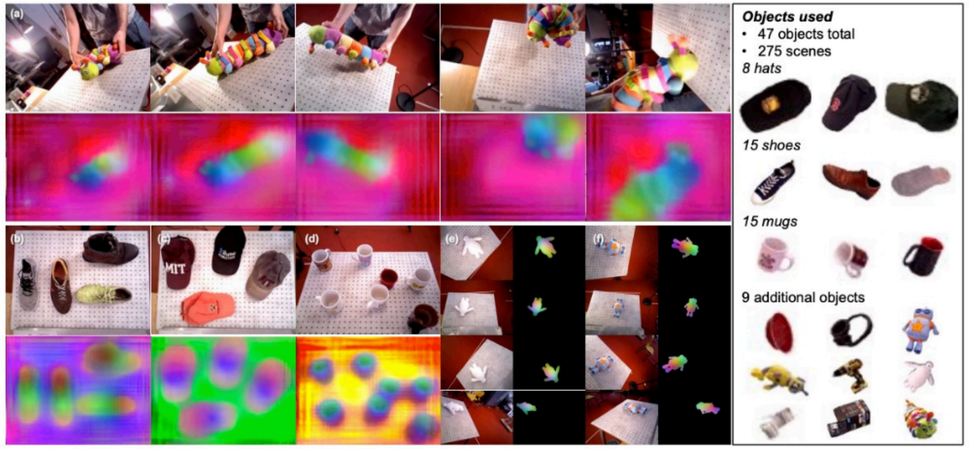
"Self-supervised learning"

Supervision comes from the unlabeled data themselves

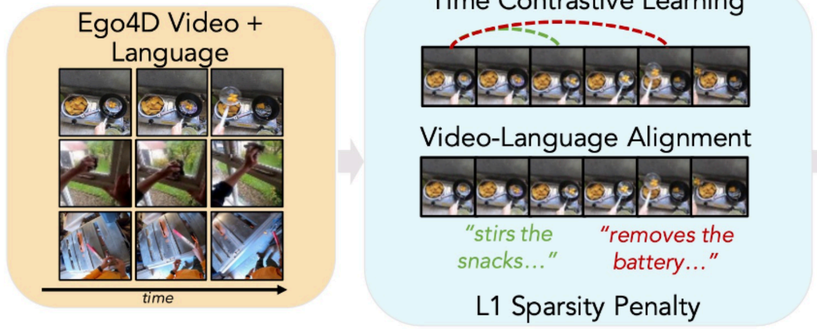
Babies learning by playing

Robot Perception: Representations

How can we learn **representations of the world** with limited supervision?



[Dense Object Nets, Florence et al. 2018]



[R3M, Nair et al. 2022]

Representation Learning for Robotics

Robot Perception: Representations

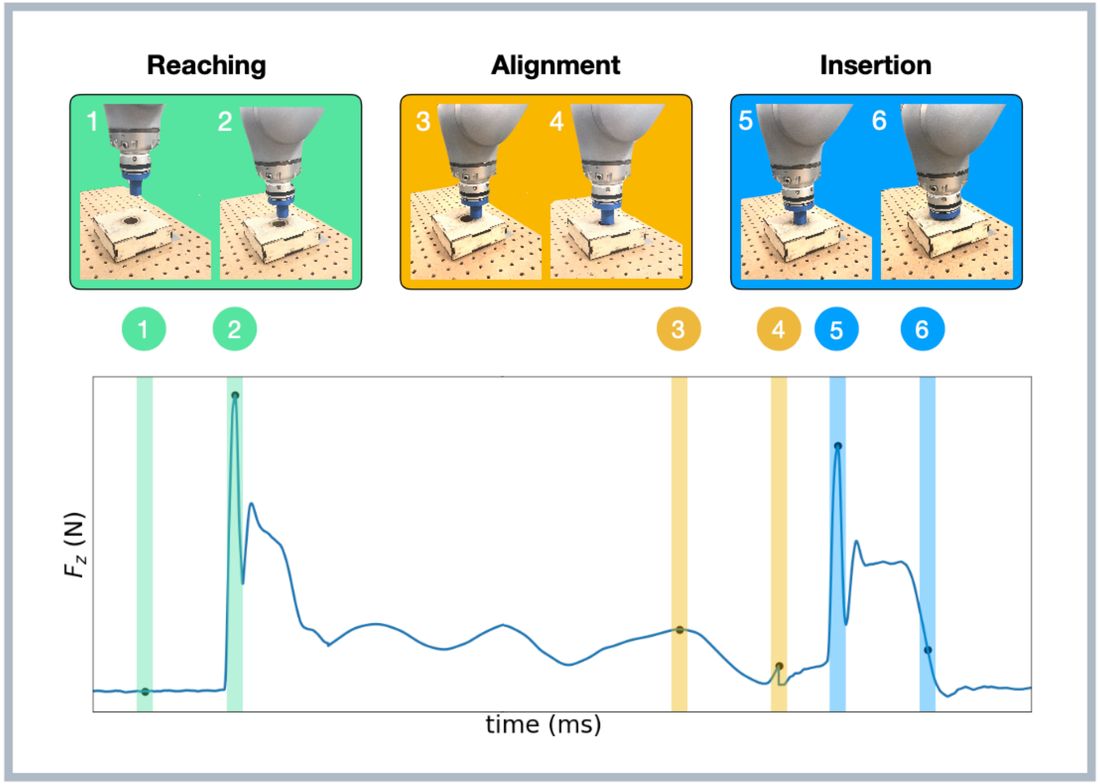
How can we learn representations that **fuse multiple sensory** modalities together?

Is seeing believing?

[The McGurk Effect, BBC] https://www.youtube.com/watch?v=2k8fHR9jKVM

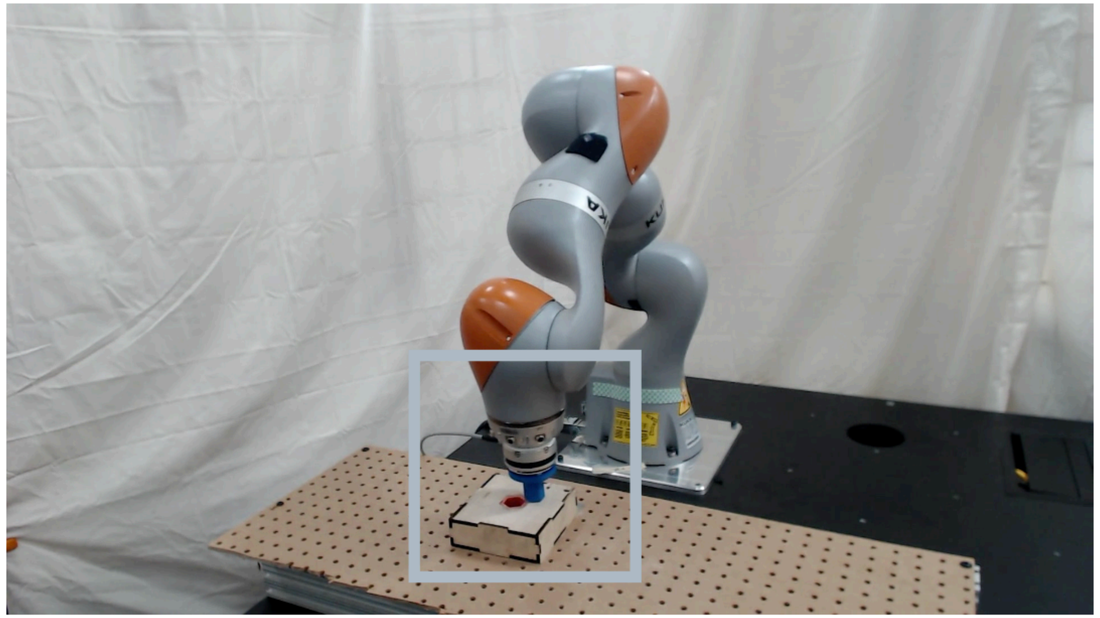
Robot Perception: Representations

How can we learn representations that **fuse multiple sensory** modalities together?



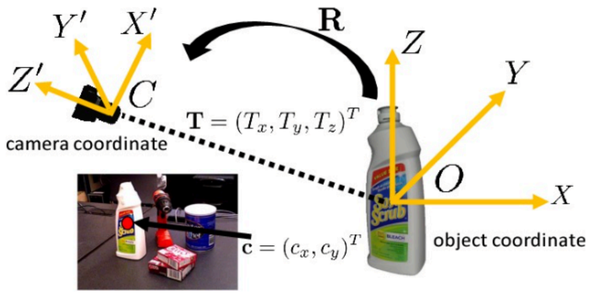
Combining vision and force for manipulation

Multi-sensory fusion



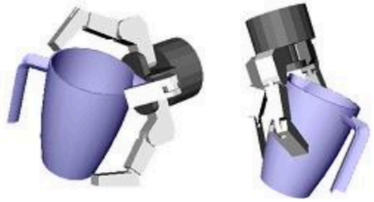
[Lee*, Zhu*, et al. 2018]

Robot Perception: Tasks



Noisy Sensory Data

Physical State

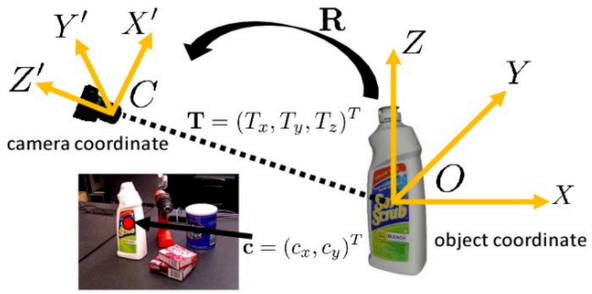


Perception &
Computer Vision

Robot Control &
Decision Making

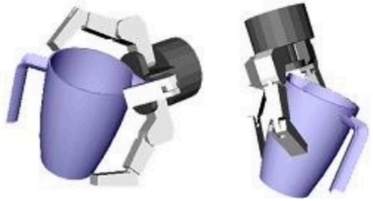
Robot Perception: Tasks

- Localization
- Pose Estimation
- Visual Tracking



Noisy Sensory Data

Physical State

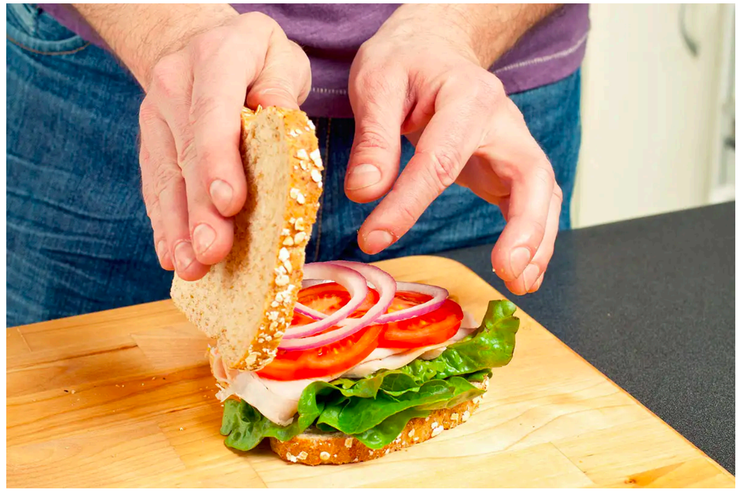


Perception & Computer Vision

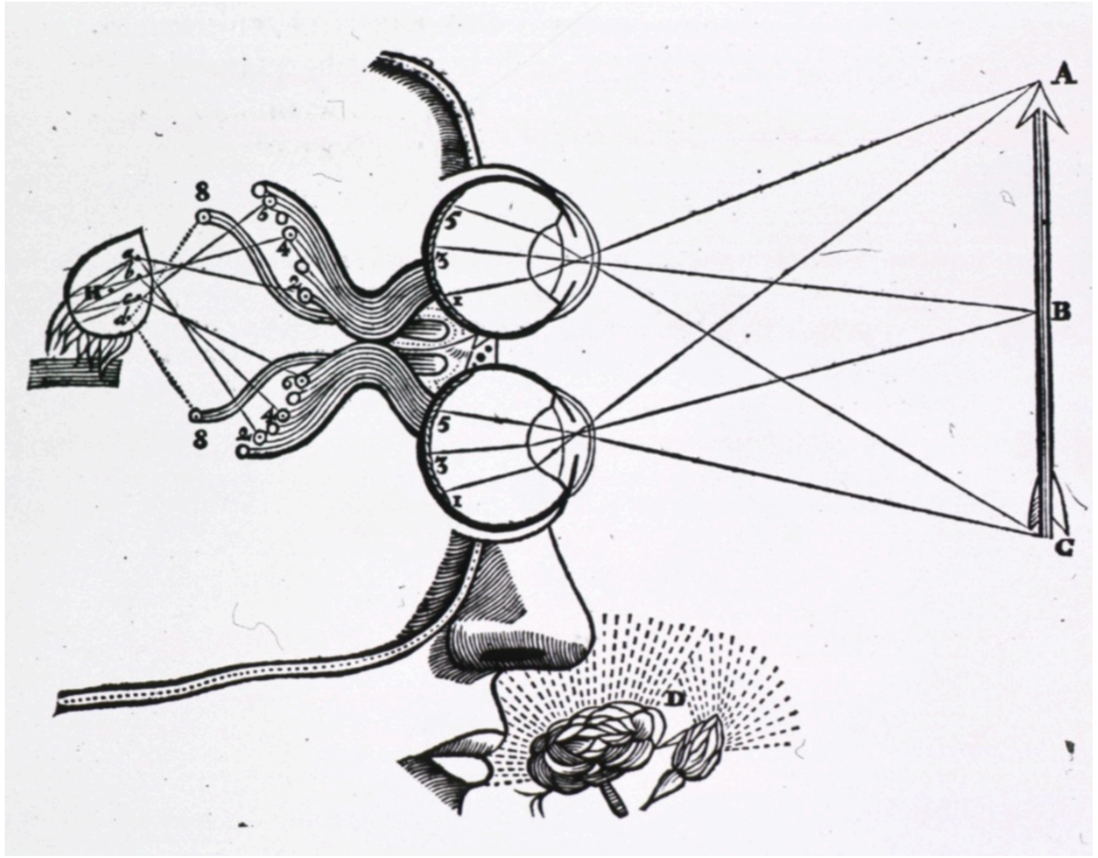
Robot Control & Decision Making

Robot Perception: Tasks

Robot Perception: Tasks



Robot Perception: Embodiment



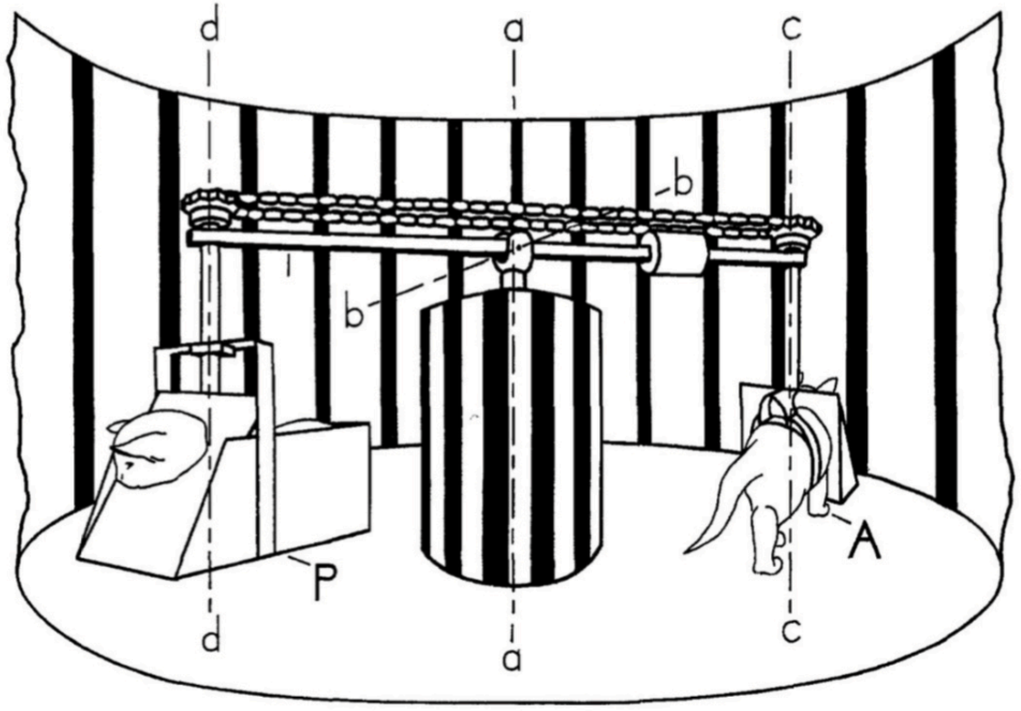
Input-Output Picture (Susan Hurley, 1998)

Conventional View of Perception

- Perception is the process of building an internal representation of the environment
- Perception is input from world to mind, and action is output from mind to world, thought is the mediating process.

[Action in Perception, Alva Noë 2004]

Robot Perception: Embodiment

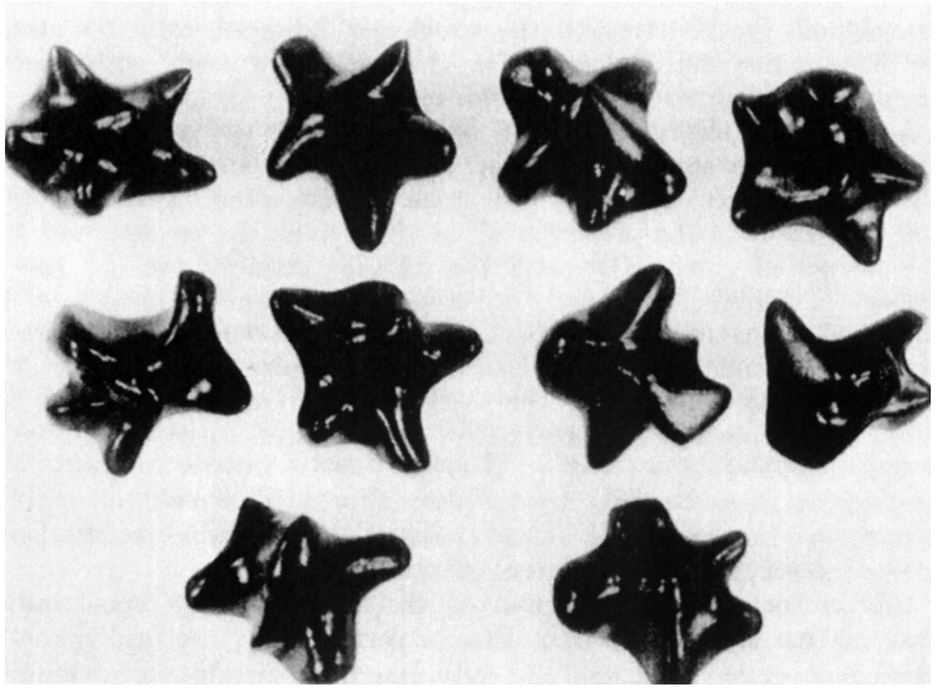


Kitten Carousel (Held and Hein, 1963)

Embodied View of Perception

- As the active cat (A) walks, the other cat (P) moves and perceives the environment passively.
- Only the active cat develops normal perception through self-actuated movement.
- The passive cat suffers from perception problems, such as 1) not blinking when objects approach, and 2) hitting the walls.

Robot Perception: Embodiment



Pebbles (James J. Gibson 1966)

Embodied View of Perception

- Subjects asked to find a reference object among a set of irregularly-shaped objects

- Three groups
  - Passive observers of one static image (49%)
  - Observers of moving shapes (72%)
  - Interactive observers (99%)

- The ability to condition input signals with actions is crucial to perception.

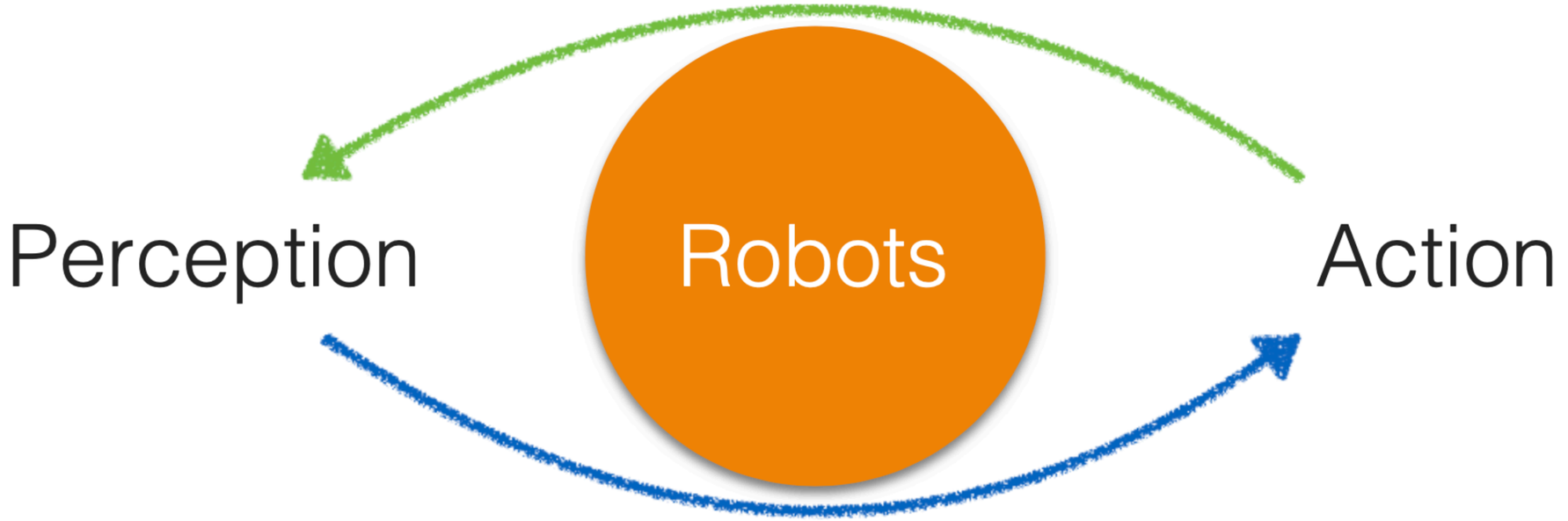
Robot Perception: Embodiment

Take-home messages

- Perceptual experiences do not present the sense in the way that a photograph does.
- Perception is developed by an embodied agent through actively exploring in the physical world.

- “We see in order to move; we move in order to see.” – William Gibson

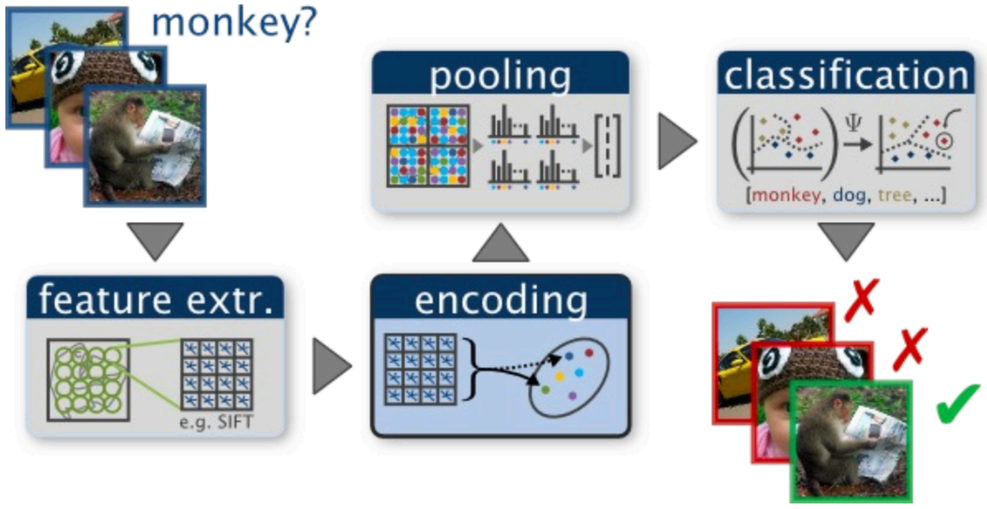
Closing the Perception-Action Loop



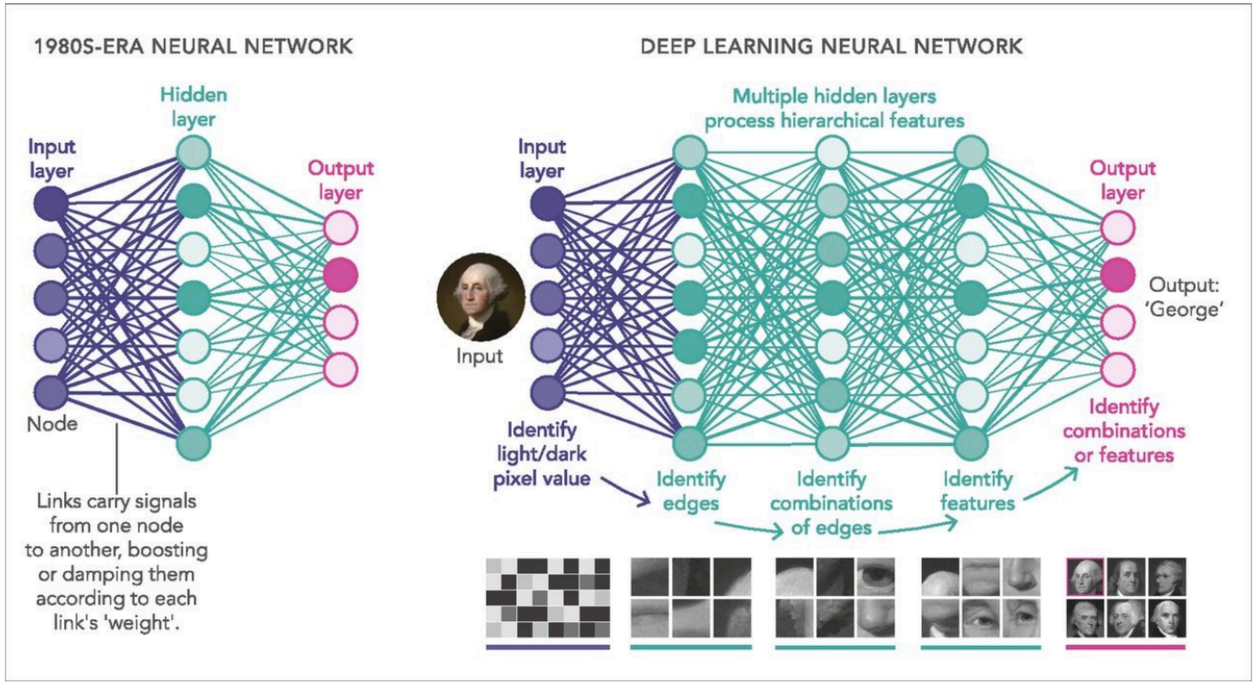
- How robots develop better perception from embodied sensorimotor experiences
- How robots' intelligent behaviors are guided by their interactive perception

Visual Processing Methods

What is new since 1980s?



Staged Visual Recognition Pipeline



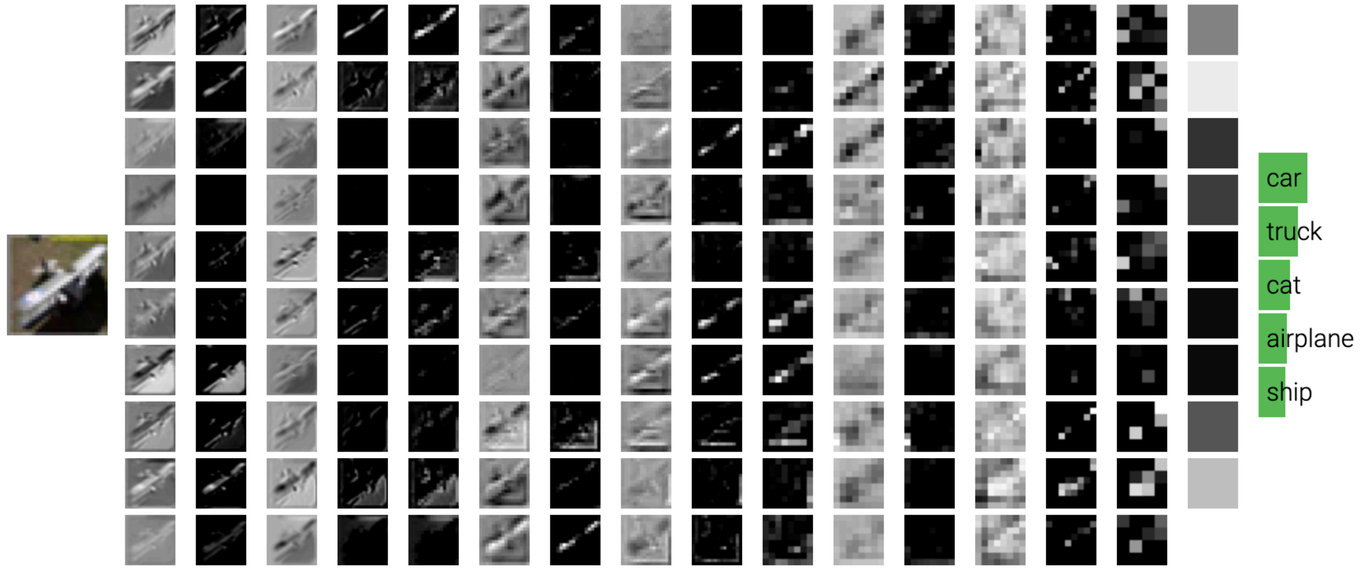
End-to-end Deep Learning

Quick Review of Deep Learning for Computer Vision

CS231n Home | Course Notes | Schedule | Project | Office Hours | Lecture Videos | Ed

CS231n: Deep Learning for Computer Vision

Stanford - Spring 2023



*This network is running live in your browser

Course Instructors

Fei-Fei Li

Andrej Karpathy

Justin Johnson

Quick Review of Deep Learning for Computer Vision

Instructors

Fei-Fei Li

Yunzhu Li

Ruohan Gao

Teaching Assistants

Ziang Liu
(Head TA)

Tanmay Agarwal

Samir Agarwala

Samuel Clarke

Zane Durante

Yuan Gao

Jeff He

Minjune Hwang

Hao Li

Eva Prakash

Manasi Sharma

Bokui (William) Shen

Haochen Shi

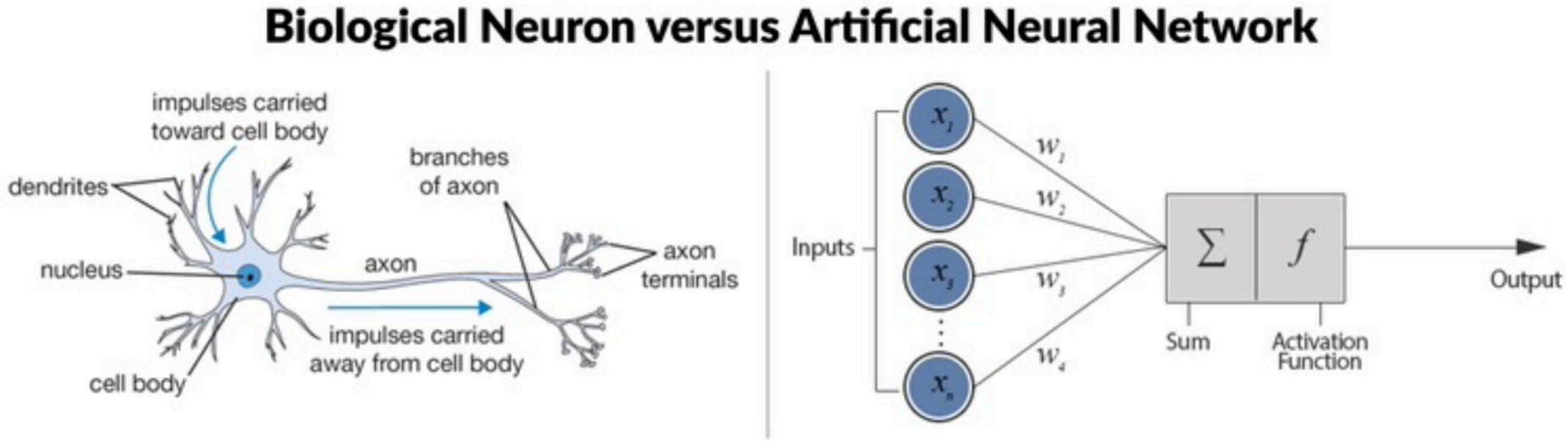
Manuka Stratta

Tiange Xiang

Course Manager

Amelie Byun

Deep Learning for Computer Vision: Artificial Neurons



Biological Neuron

Computational building block for the brain

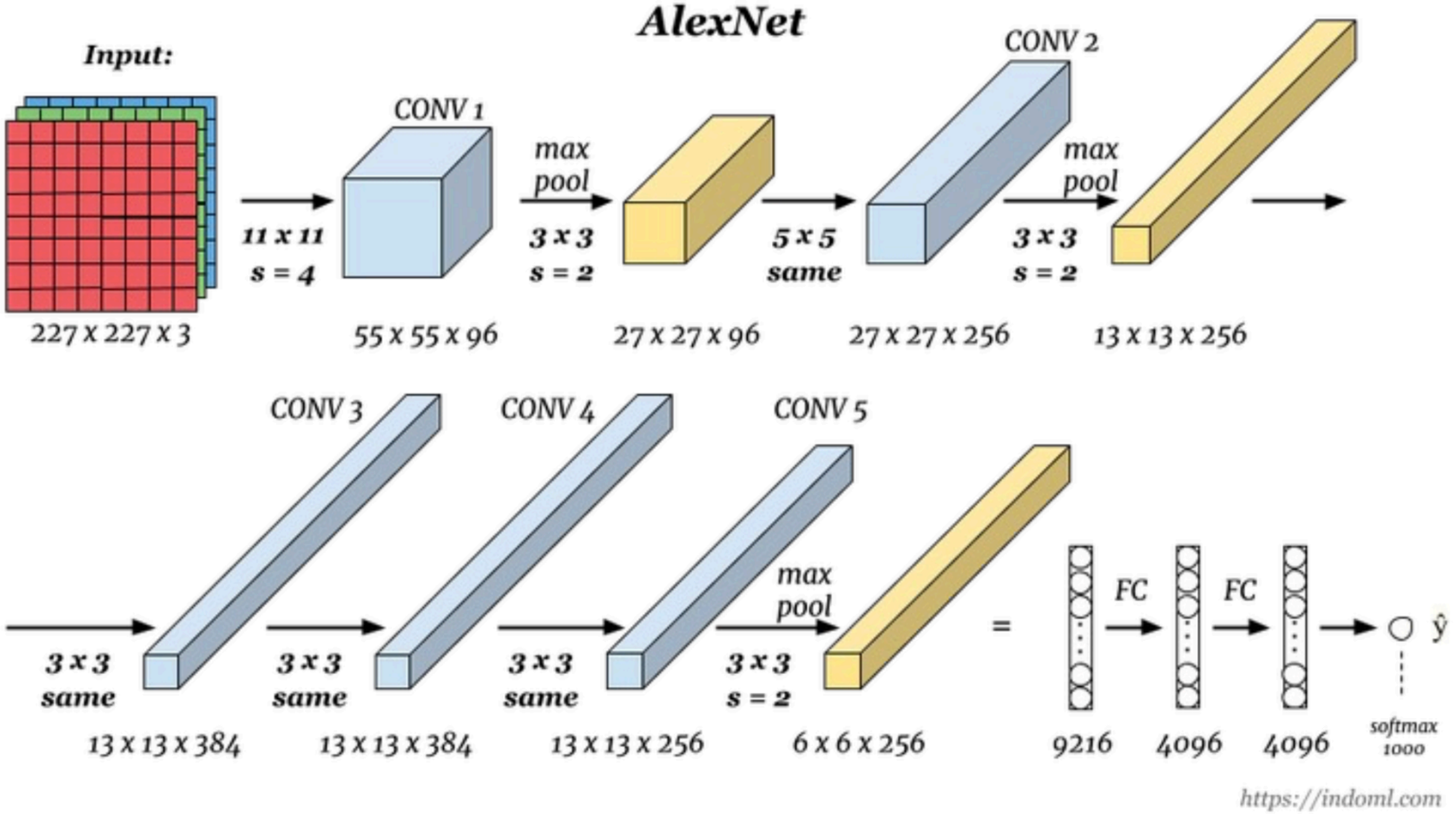
Artificial Neuron

Computational building block for the neural network

Note: Many differences exist – be careful with the brain analogies!

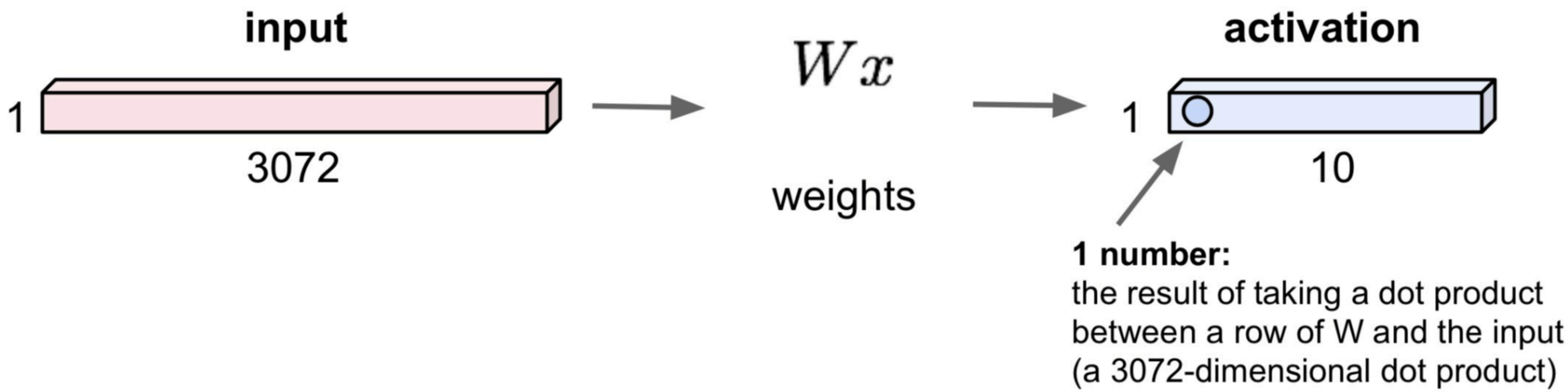
[Dendritic Computation, Michael London and Michael Hausser 2015]

Deep Learning for Computer Vision: Convolutional Networks



Deep Learning for Computer Vision: Fully-Connected Layers

32x32x3 image -> stretch to 3072 x 1

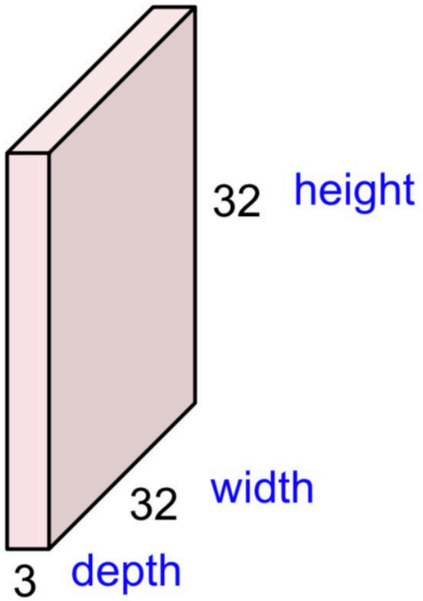


What is the dimension of W?

[Source: Stanford CS231N]

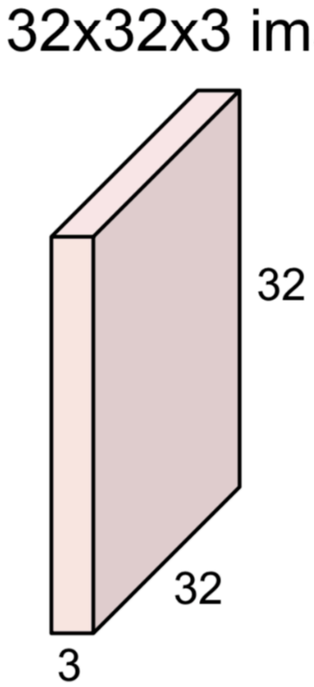
Deep Learning for Computer Vision: Convolutional Layers

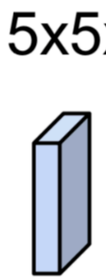
32x32x3 image -> preserve spatial structure



[Source: Stanford CS231N]

Deep Learning for Computer Vision: Convolutional Layers

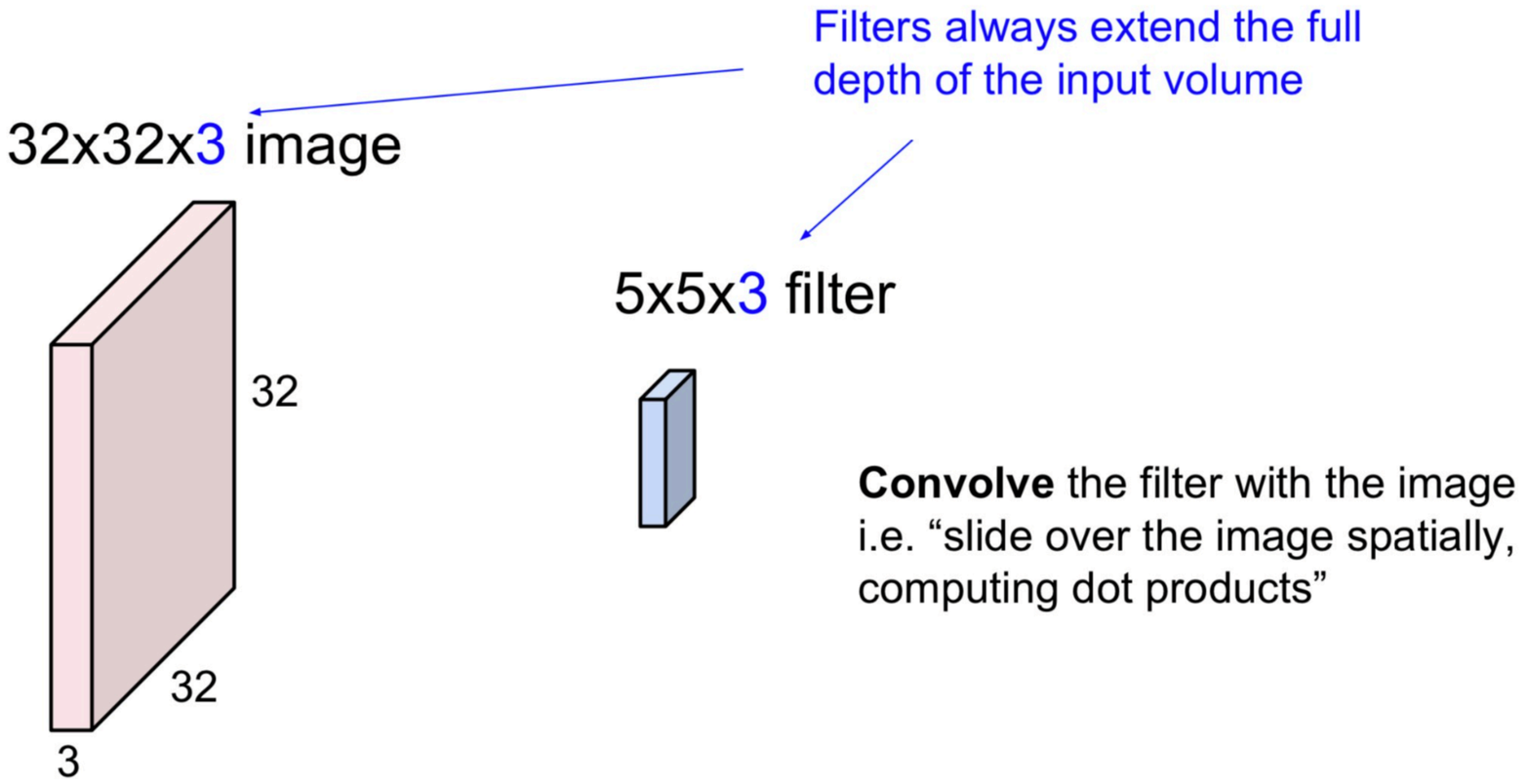




Convolve the filter with the image
i.e. “slide over the image spatially,
computing dot products”

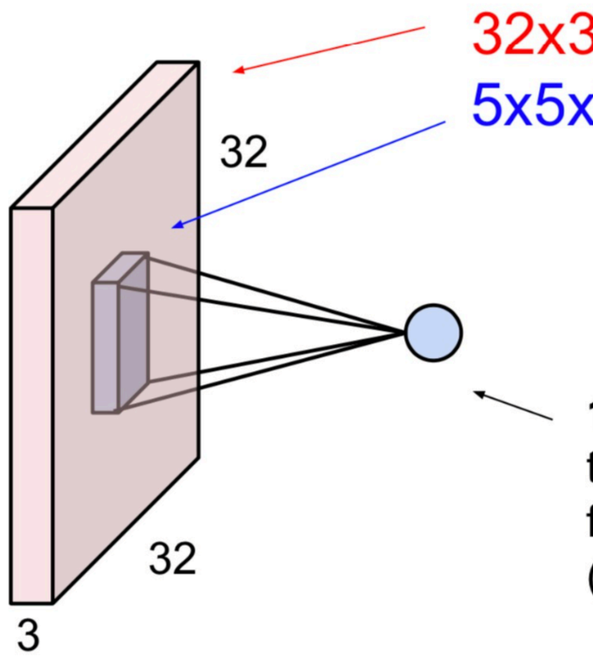
[Source: Stanford CS231N]

Deep Learning for Computer Vision: Convolutional Layers



[Source: Stanford CS231N]

Deep Learning for Computer Vision: Convolutional Layers

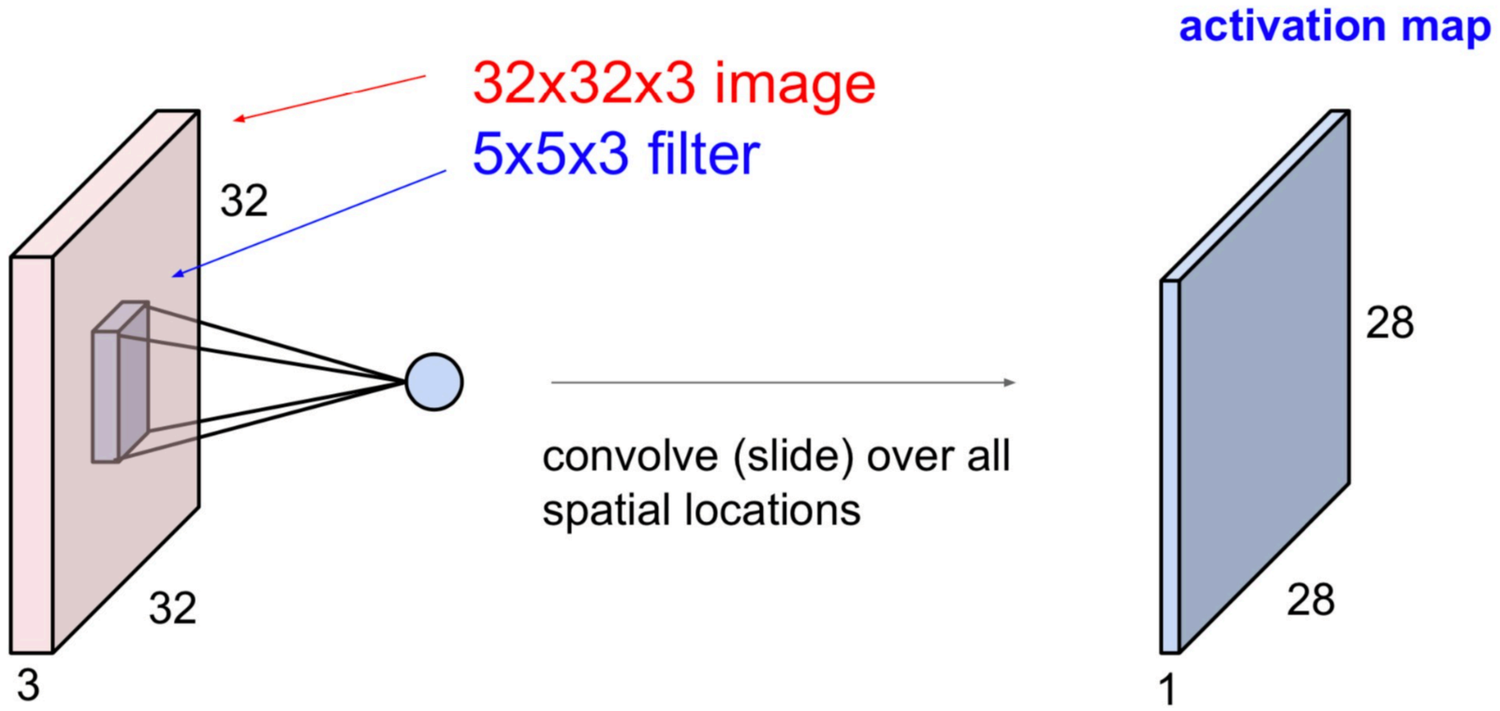


1 number:
the result of taking a dot product between the filter and a small 5x5x3 chunk of the image
(i.e. 5*5*3 = 75-dimensional dot product + bias)

$$w^T x + b$$

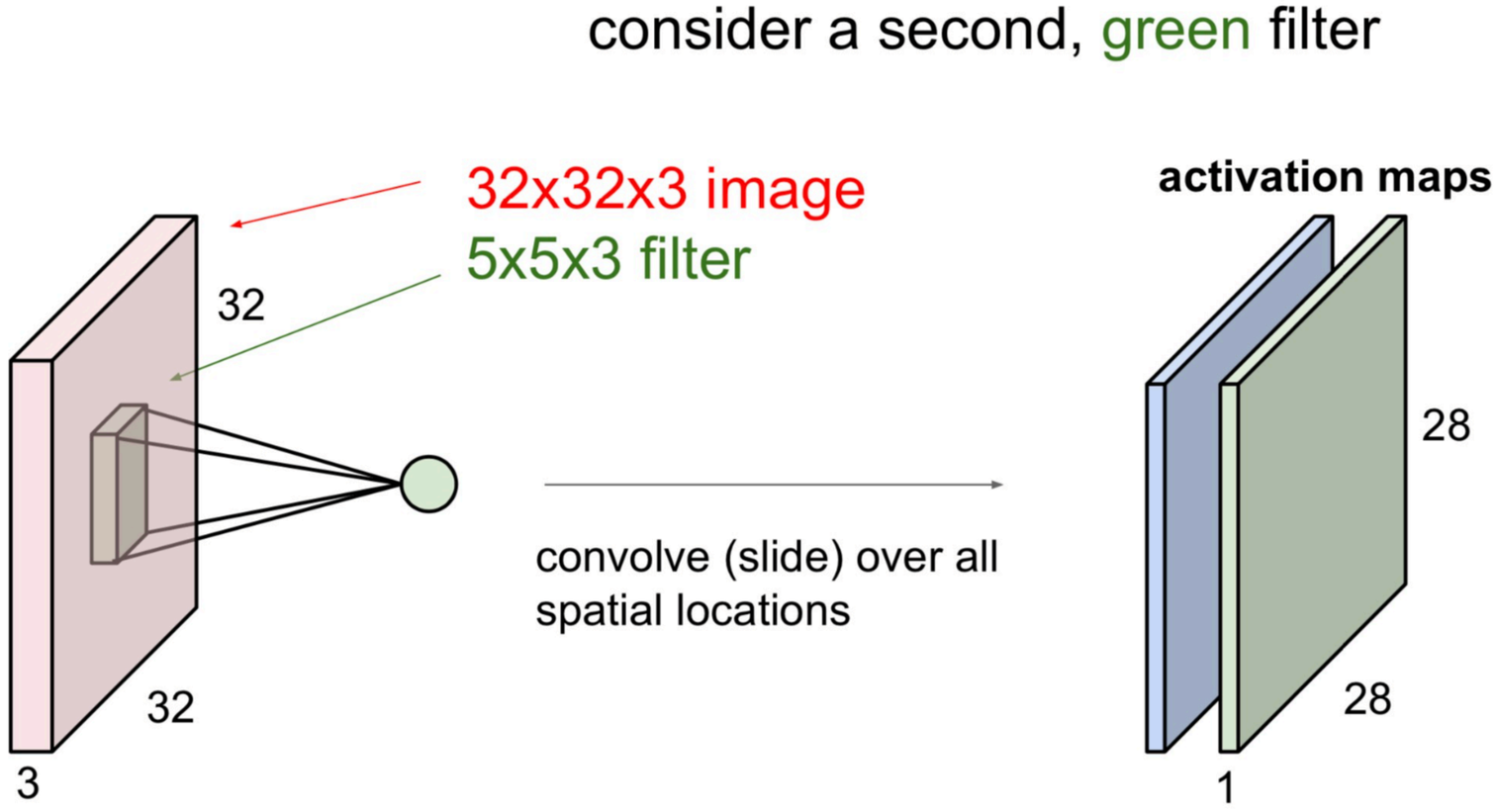
[Source: Stanford CS231N]

Deep Learning for Computer Vision: Convolutional Layers



[Source: Stanford CS231N]

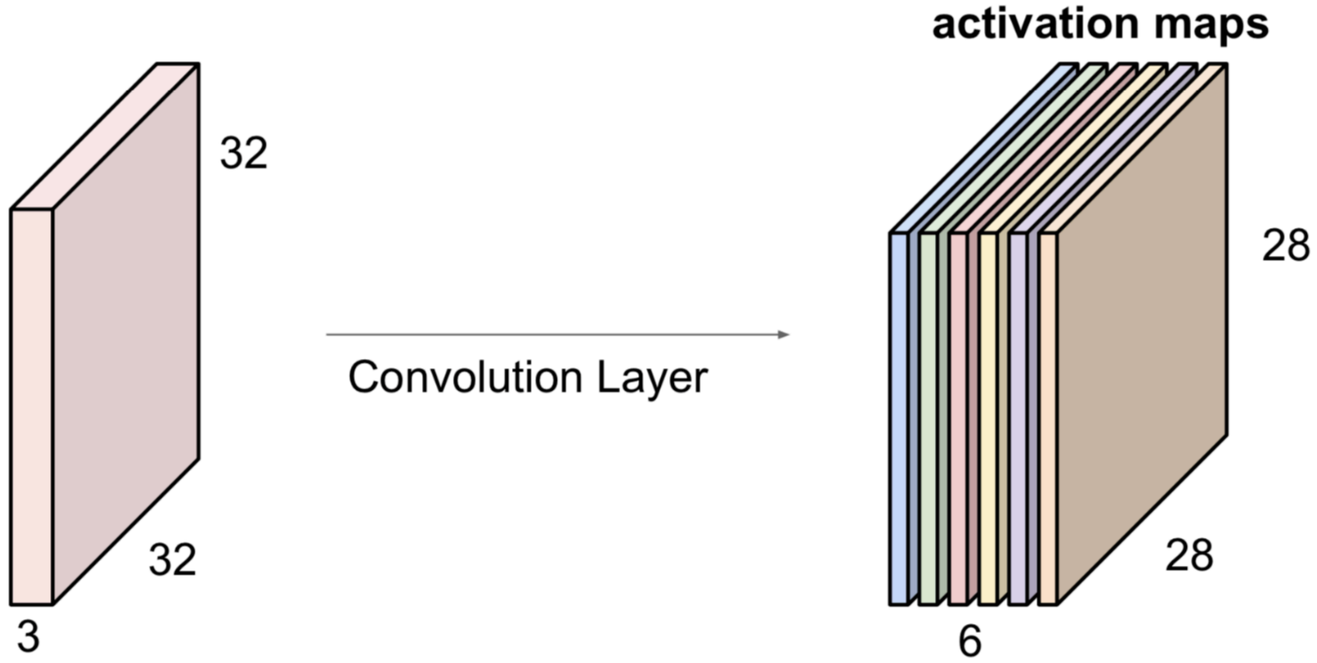
Deep Learning for Computer Vision: Convolutional Layers



[Source: Stanford CS231N]

Deep Learning for Computer Vision: Convolutional Layers

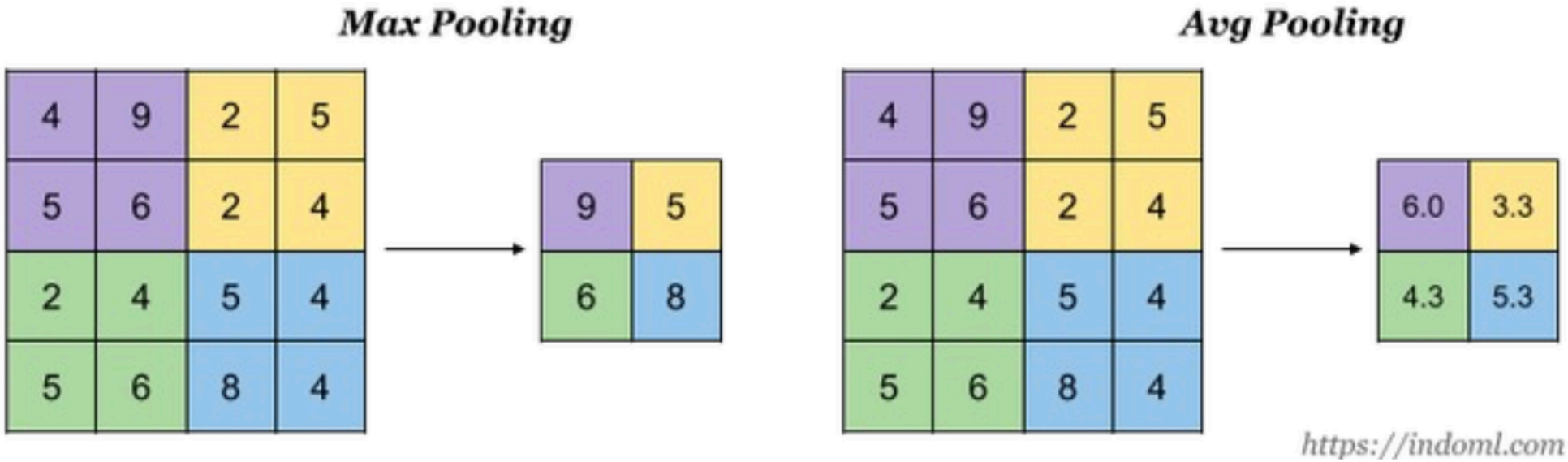
For example, if we had 6 5x5 filters, we'll get 6 separate activation maps:



We stack these up to get a "new image" of size 28x28x6!

[Source: Stanford CS231N]

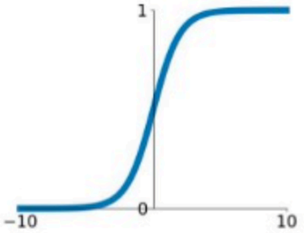
Deep Learning for Computer Vision: Pooling Operations



Deep Learning for Computer Vision: Activation Functions

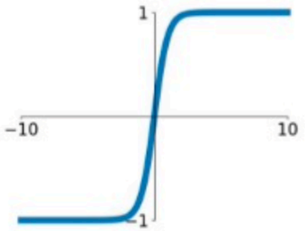
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



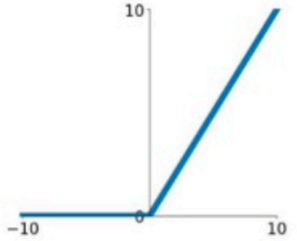
tanh

$$\tanh(x)$$



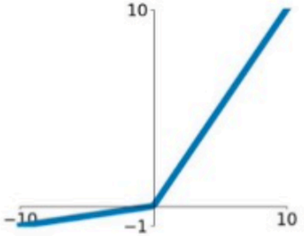
ReLU

$$\max(0, x)$$



Leaky ReLU

$$\max(0.1x, x)$$

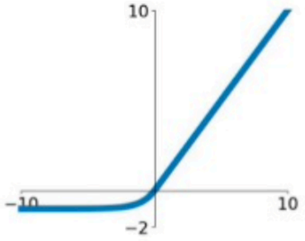


Maxout

$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

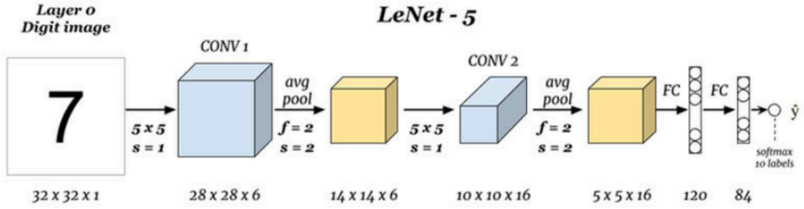
ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$

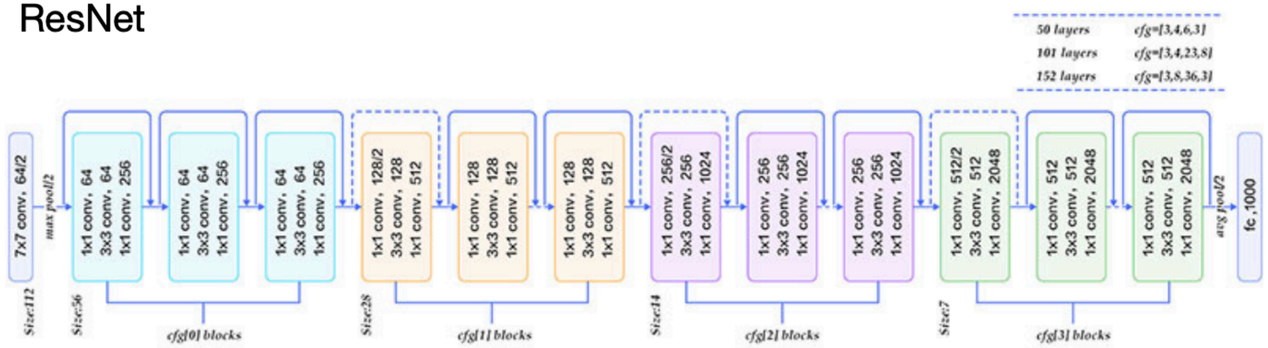


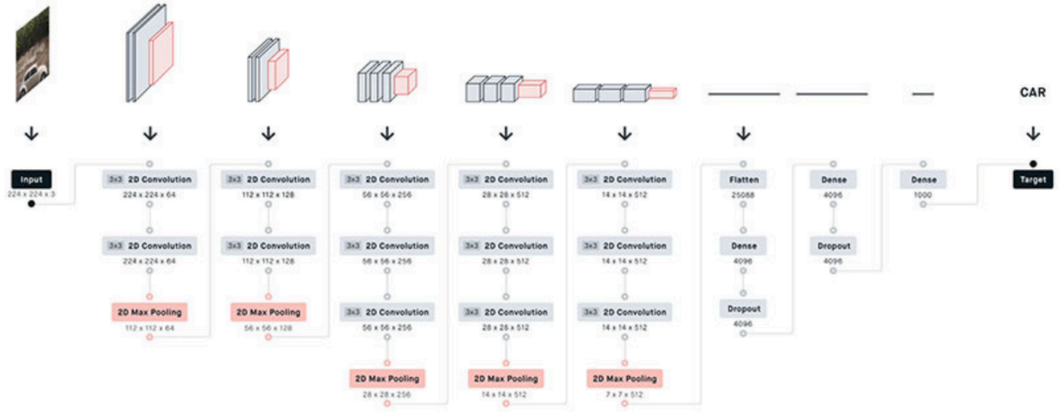
Deep Learning for Computer Vision: CNN Architectures

LeNet

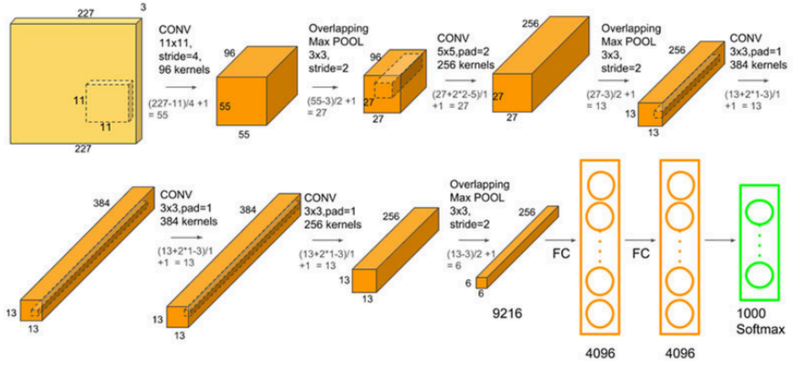


ResNet



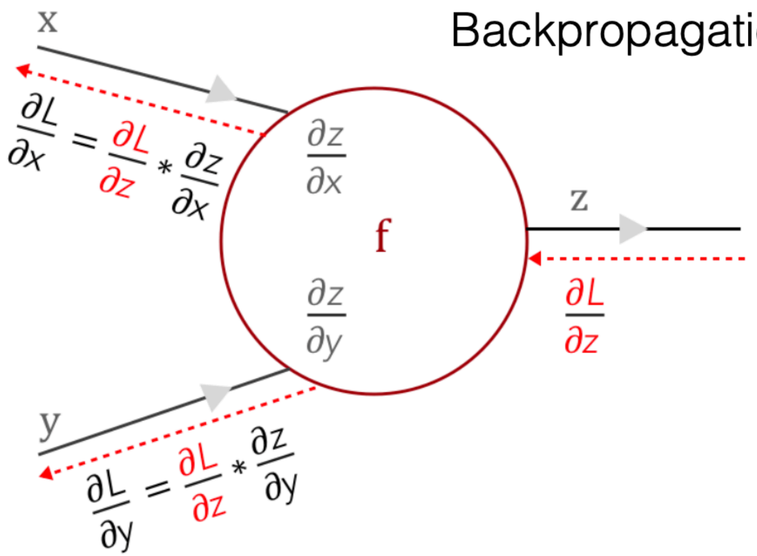


VGG-16



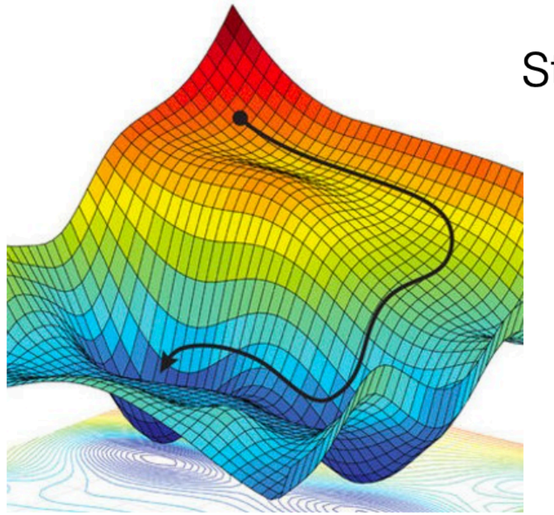
AlexNet

Deep Learning for Computer Vision: Optimization



$\frac{\partial z}{\partial x}$ & $\frac{\partial z}{\partial y}$ *are local gradients*

$\frac{\partial L}{\partial z}$ *is the loss from the previous layer which has to be backpropagated to other layers*



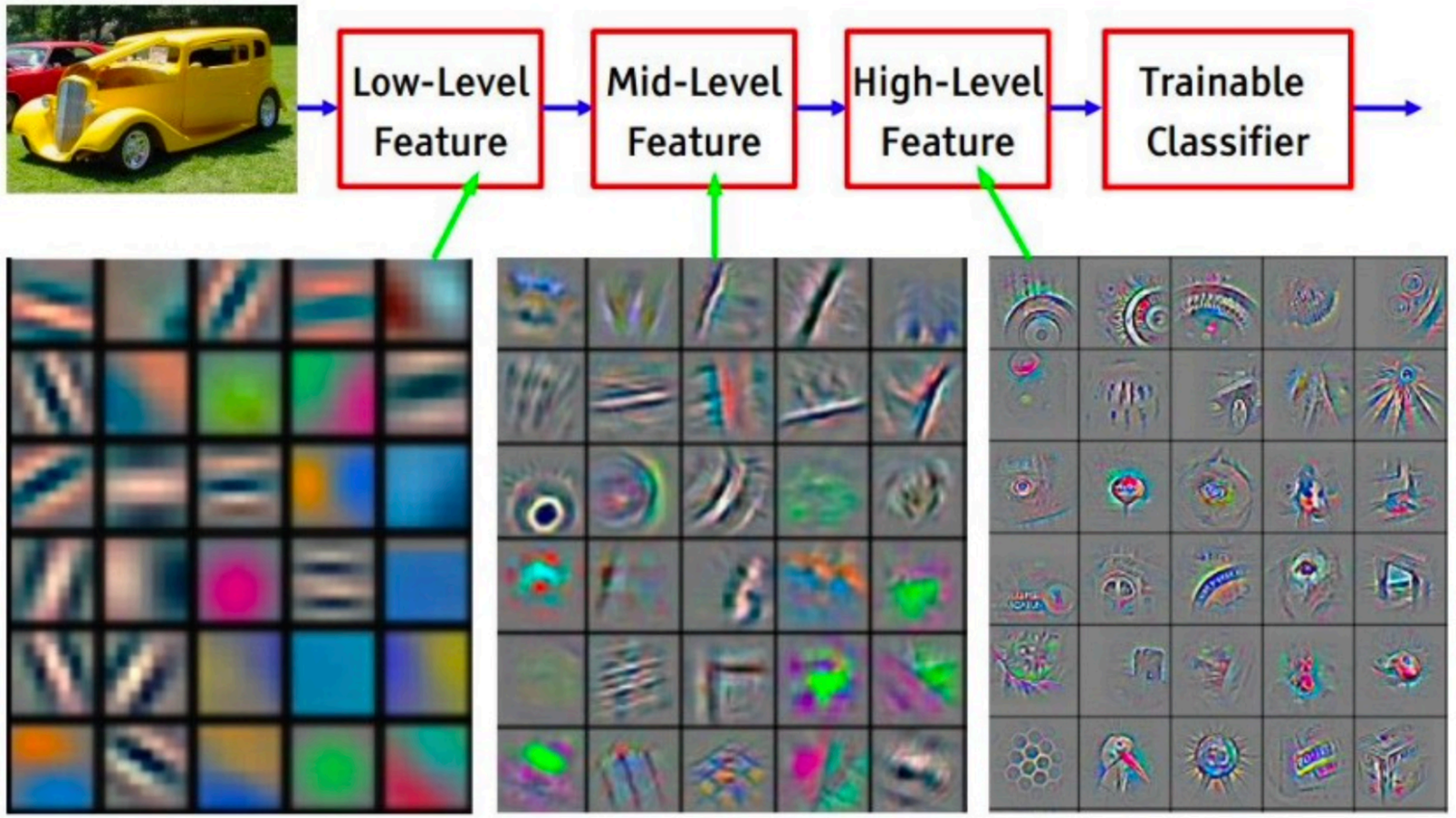
Stochastic Gradient Descent (SGD)

learning rate

$$\theta = \theta - \eta \nabla_\theta J(\theta; x^{(i)}; y^{(i)})$$

weights input label

Deep Learning for Computer Vision: Features



[Source: Stanford CS231N]

Deep Learning for Computer Vision: Implementation

PyTorch

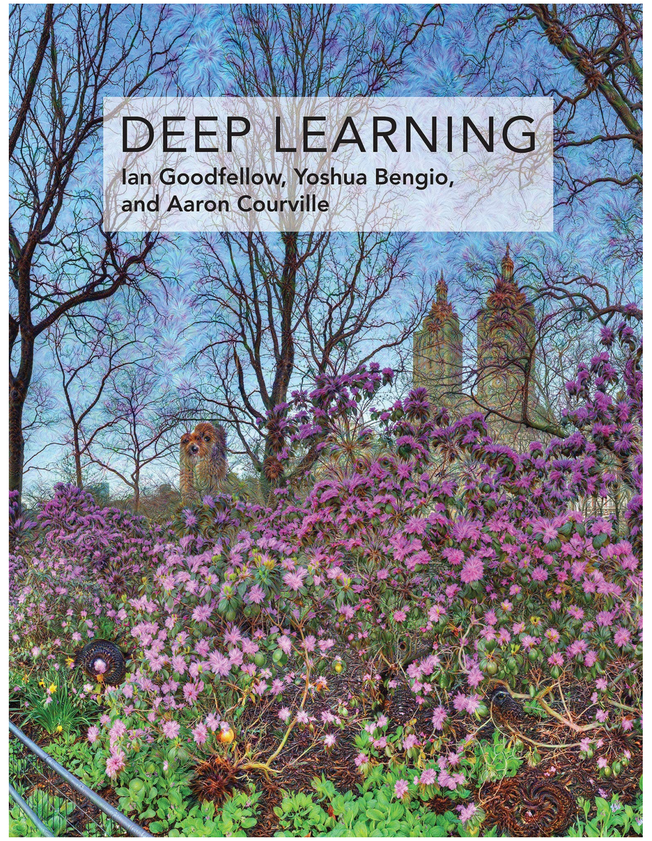
TensorFlow

```
[ ]  import torch
     from torch import nn

     class MNISTClassifier(nn.Module):

       def __init__(self):
         super(MNISTClassifier, self).__init__()

         # mnist images are (1, 28, 28) (channels, width, heig
         self.layer_1 = torch.nn.Linear(28 * 28, 128)
         self.layer_2 = torch.nn.Linear(128, 256)
         self.layer_3 = torch.nn.Linear(256, 10)

       def forward(self, x):
         batch_size, channels, width, height = x.size()

         # (b, 1, 28, 28) -> (b, 1*28*28)
         x = x.view(batch_size, -1)

         # layer 1
         x = self.layer_1(x)
         x = torch.relu(x)

         # layer 2
         x = self.layer_2(x)
         x = torch.relu(x)

         # layer 3
         x = self.layer_3(x)

         # probability distribution over labels
         x = torch.log_softmax(x, dim=1)

         return x
```

Deep Learning for Computer Vision: Resources

- Online Courses
  - CS231N: Deep Learning for Computer Vision
    - http://cs231n.stanford.edu/
  - MIT 6.S191: Introduction to Deep Learning
    - http://introtodeeplearning.com

- Textbooks:
  - Deep Learning
    - Ian Goodfellow, Yoshua Bengio, Aaron Courville
    - http://www.deeplearningbook.org/



Logistics

- Sign up presentation preference form
  - Deadline: Thur 11:59 PM, Jan 18 (Tonight!)
- The first week presentation submission deadline is Mon 11:59 PM Jan 22.
- Presentation practice: 2:00-2:45PM Tues and Thur with the instructor
  - Around 15 mins for each presentation
- Each student is required to present one paper