CS231A

Computer Vision: From 3D Reconstruction to Recognition

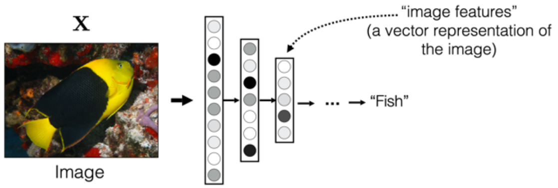
Representation & Representation Learning

How to reach me?

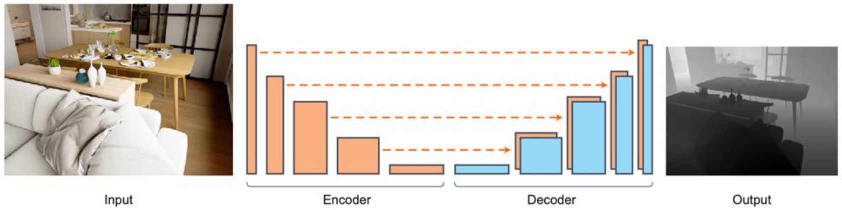
- Jeannette Bohg, CS, Assistant Professor in Robotics
- Office hours, Wednesdays 9am, Gates 244 or on zoom

Learning Goals for Upcoming Lectures

Representations & Representation Learning



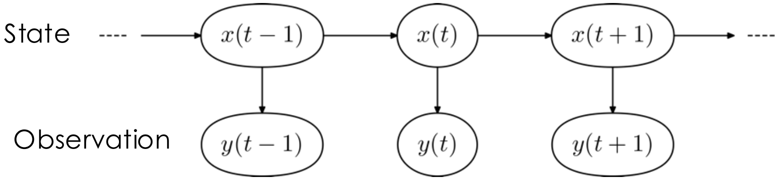
Monocular Depth Estimation, Feature Tracking



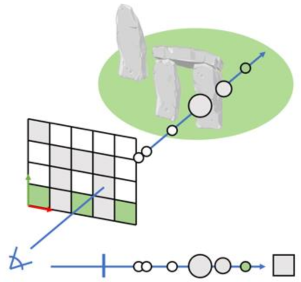
Optical & Scene Flow

A Database and Evaluation Methodology for Optical Flow.
Baker et al. IJCV. 2011

Optimal Estimation



Neural Radiance Fields



Exercise

- Use an to manipulate (pen, water bottle, a mask, …)
- What information do you need to solve a manipulation task, i.e., to make decision?
- How do you get this information?

Representations for Manipulation Tasks

0 surveys completed

0 surveys underway

How do you get this information? (Be concise)

Join by Web

PollEv.com/jeannetteboh707

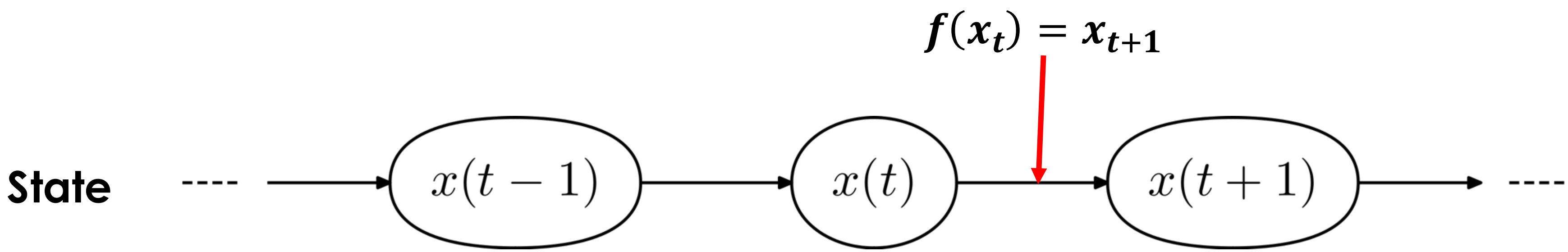
Join by QR code

Scan with your camera app

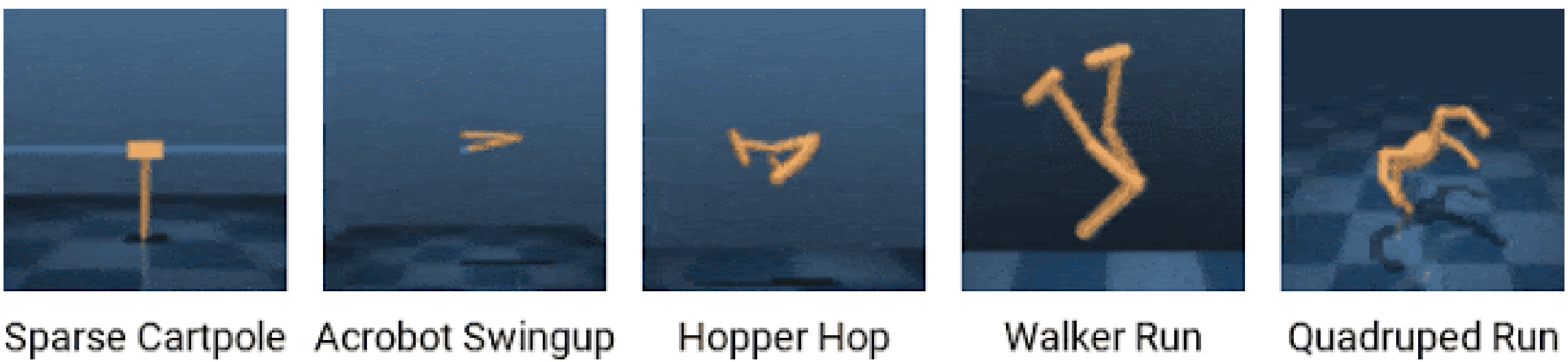
Outline of this lecture

- What is a state? What is a representation?
- What are the different kinds of representations?
- How can we extract state from raw sensory data?
- How can we learn good representations form data?

What is a state? What is a representation?

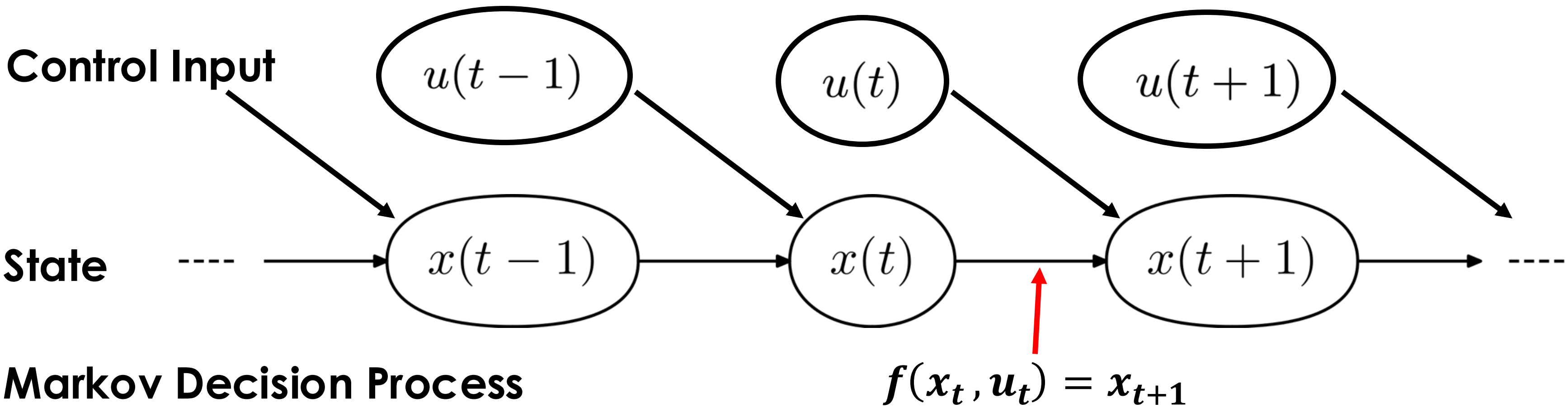


Markov Model



DeepMind Control Suite. Tassa et al. 2018

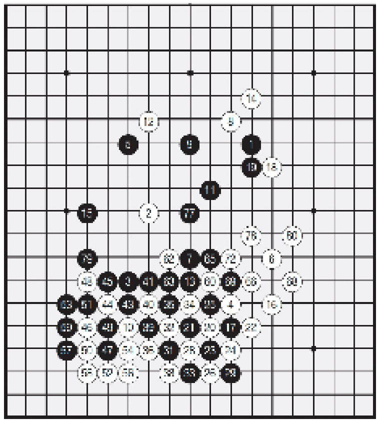
What is a state? What is a representation?



Sparse Cartpole Acrobot Swingup Hopper Hop Walker Run Quadruped Run

DeepMind Control Suite. Tassa et al. 2018

What is a state? What is a representation?



3^{361} states?

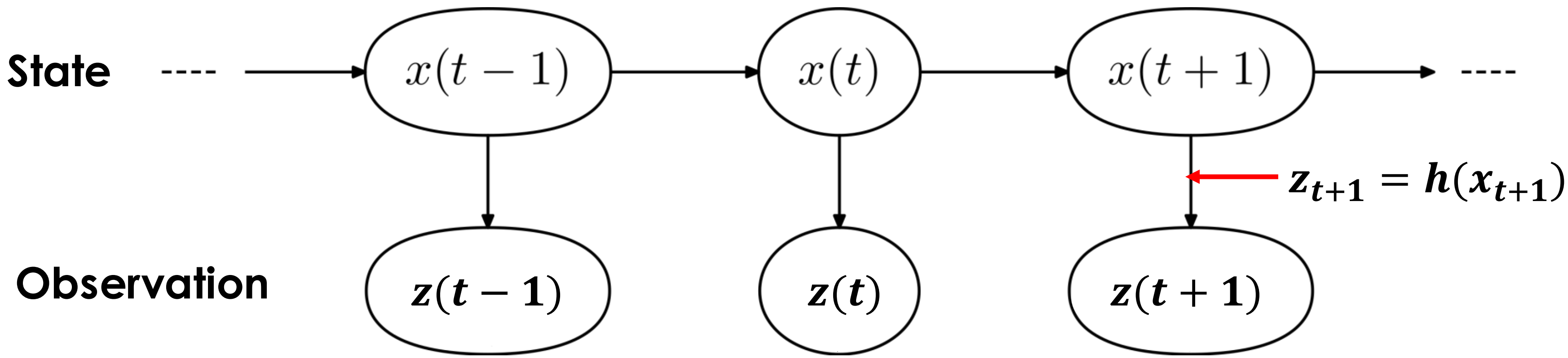
Game of Go
- an exponentially large number of states?
- infeasible to enumerate, memorize, or search

$256^{3x500x500}$?

Images
Image space has exponentially more states than Go.

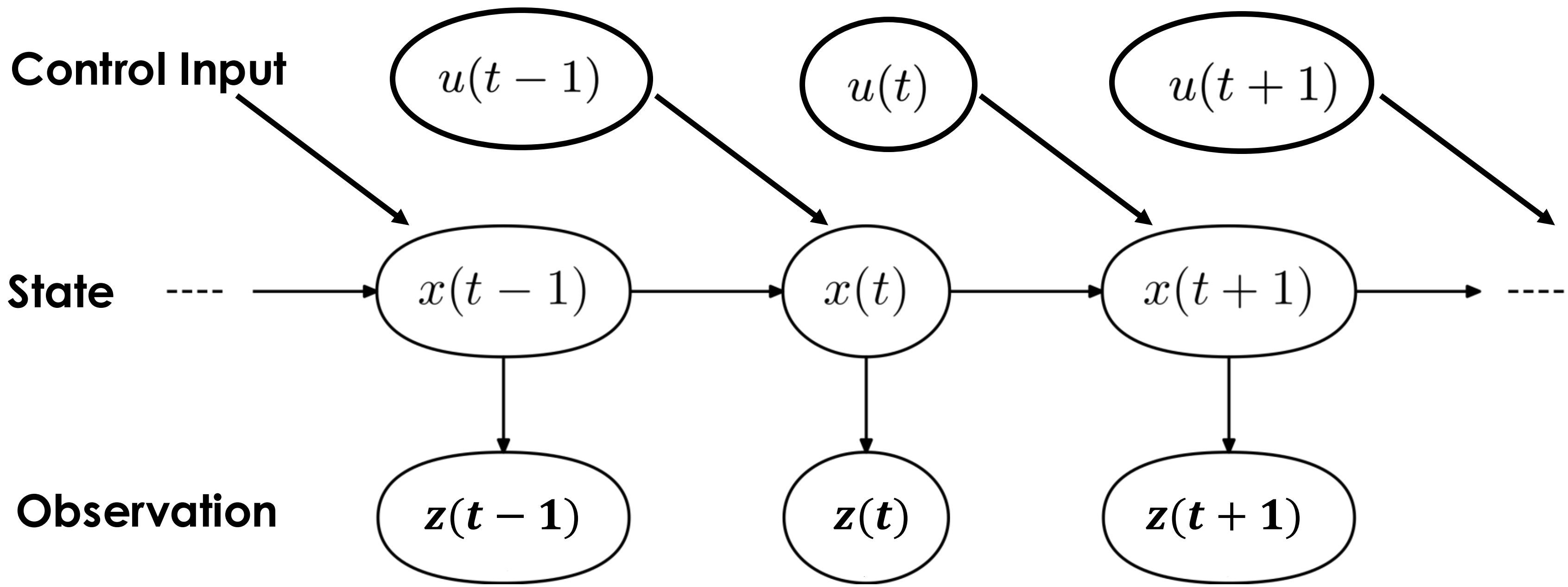
Examples from MIT - 6.8300/1 Advances in Computer Vision

What is a state? What is a representation?



Hidden Markov Model

What is a state? What is a representation?

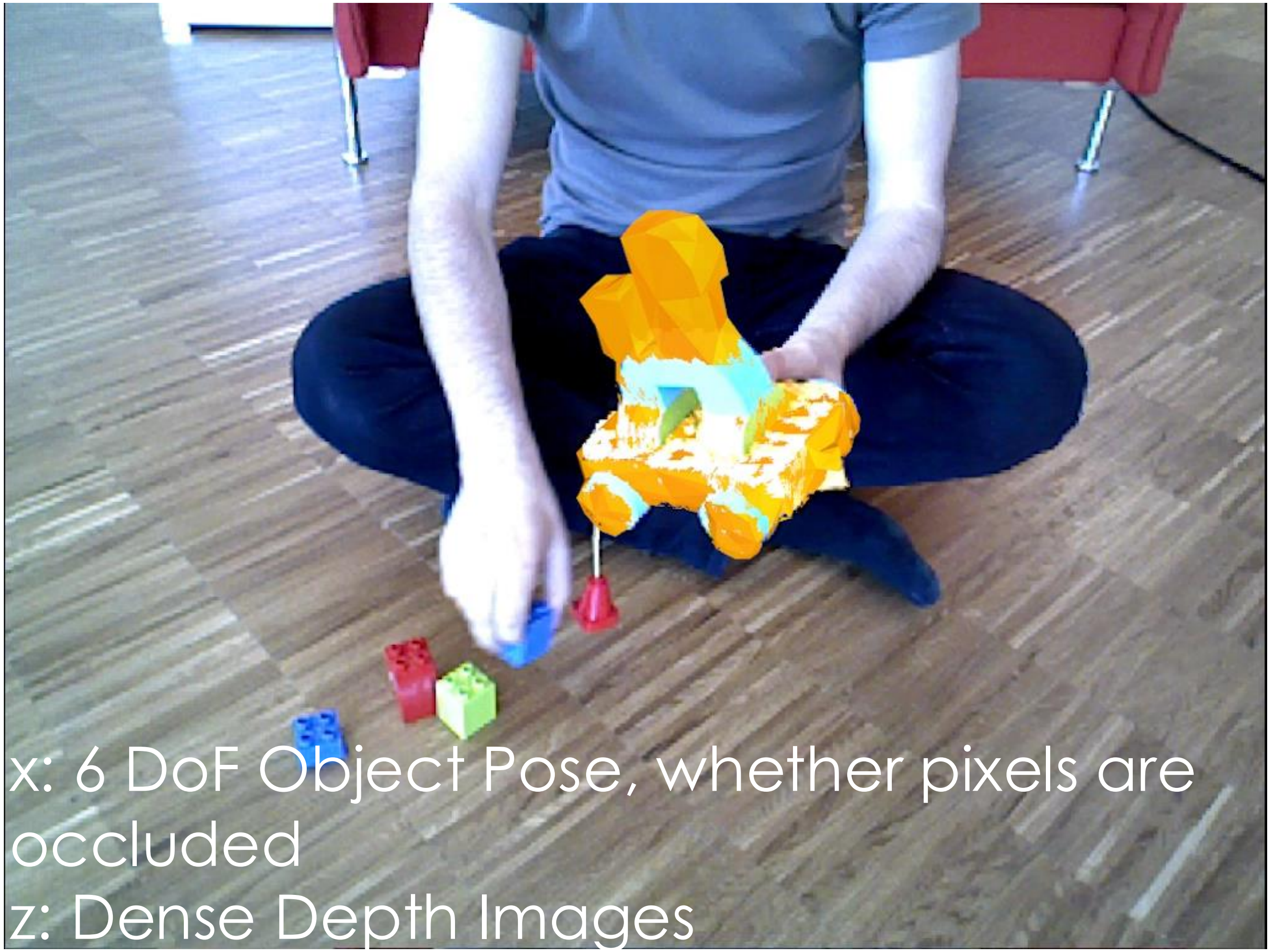


Partially Observable Markov Decision Process

Representations for Autonomous Driving

Image adapted from NuScenes by Motional. nuscenes.org

Representations for Manipulation



Manuel Wühtrich et al. "Probabilistic Object Tracking using a Depth Camera", IROS 2013

Meaning in English

“the way that someone or something is shown or described:”

“a sign, picture, model, etc. of something”

- Cambridge Dictionary

Representations in Cognitive Science

Symbolic View

"[…] a hypothetical internal cognitive symbol that represents external reality" (Morgan '14)

"[…] a formal system for making explicit certain entities or types of information […]" (Marr '10)

"[…] intermediaries between the observing subject and the objects, processes or other entities observed in the external world. These intermediaries […] represent to the mind the objects of that world." (Wikipedia - Mental Representations - Representationalism)

Embodied View

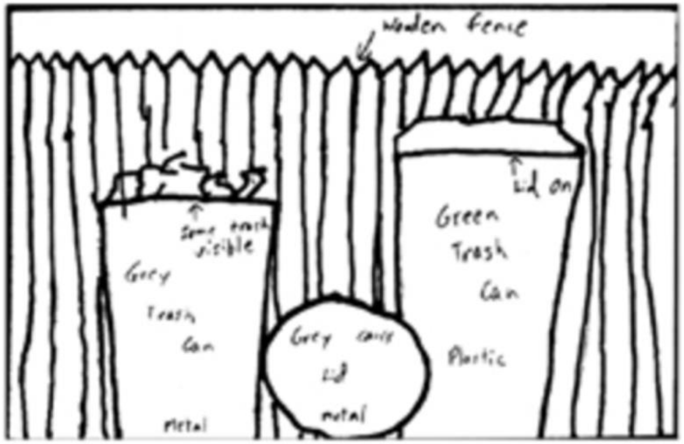
"… actions are directly triggered by stimuli in the environment without the need for internal representations" (Gibson '66, Zech '19 on Embodied Cognition)

"… actions are represented by their anticipated effect, that is, action representations essentially entail a mental model of a needed future environmental state" (Jeannerod '06, Zech '19)

Representations in Cognitive Science

Observed image

Drawn from memory



[Bartlett, 1932]
[Intraub & Richardson, 1989]

Example from Advances in Computer Vision – MIT – 6.869/6.819

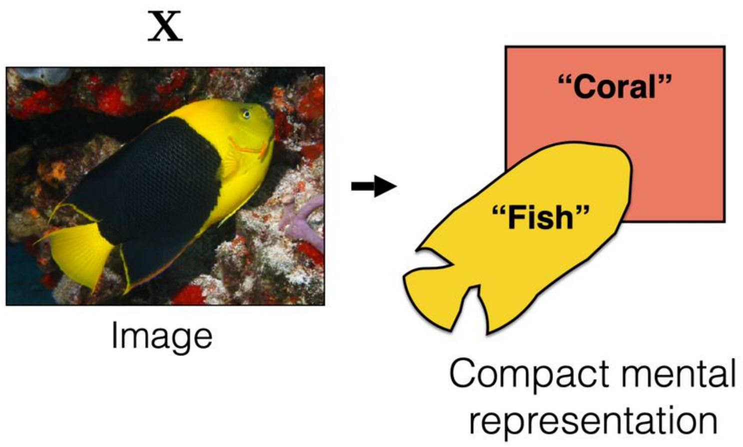
Representations in Machine Learning

"Features", "A good representation is also one that is useful as input to a supervised predictor." (Bengio '14)

"create a representation of the data to provide the model with a useful vantage point into the data's key qualities. [...] to train a model, you must choose the set of features that best represent the data." (Google Crash Course of ML Concepts)

" [...] world models, internal models of how the world works."; "(1) estimate missing information about the state of the world not provided by perception, (2) predict plausible future states of the world." (YLC '22)

Representations in Computer Vision

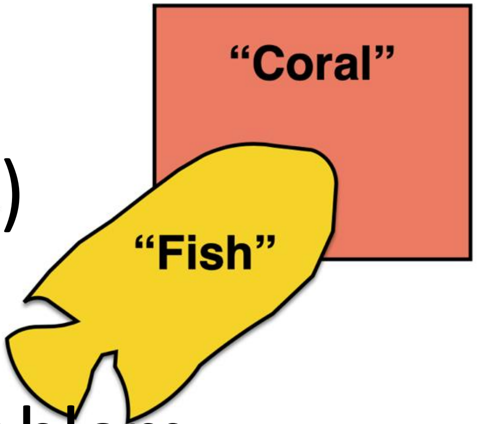


Input: grey scale, color, depth image, point cloud – Sensor measurements of the world
Output: Symbols, abstract shapes, 6D pose – Often Representation for Decision-Making
Intermediate Representations: Compact summary of the sensory information

Example from Advances in Computer Vision – MIT – 6.869/6.819

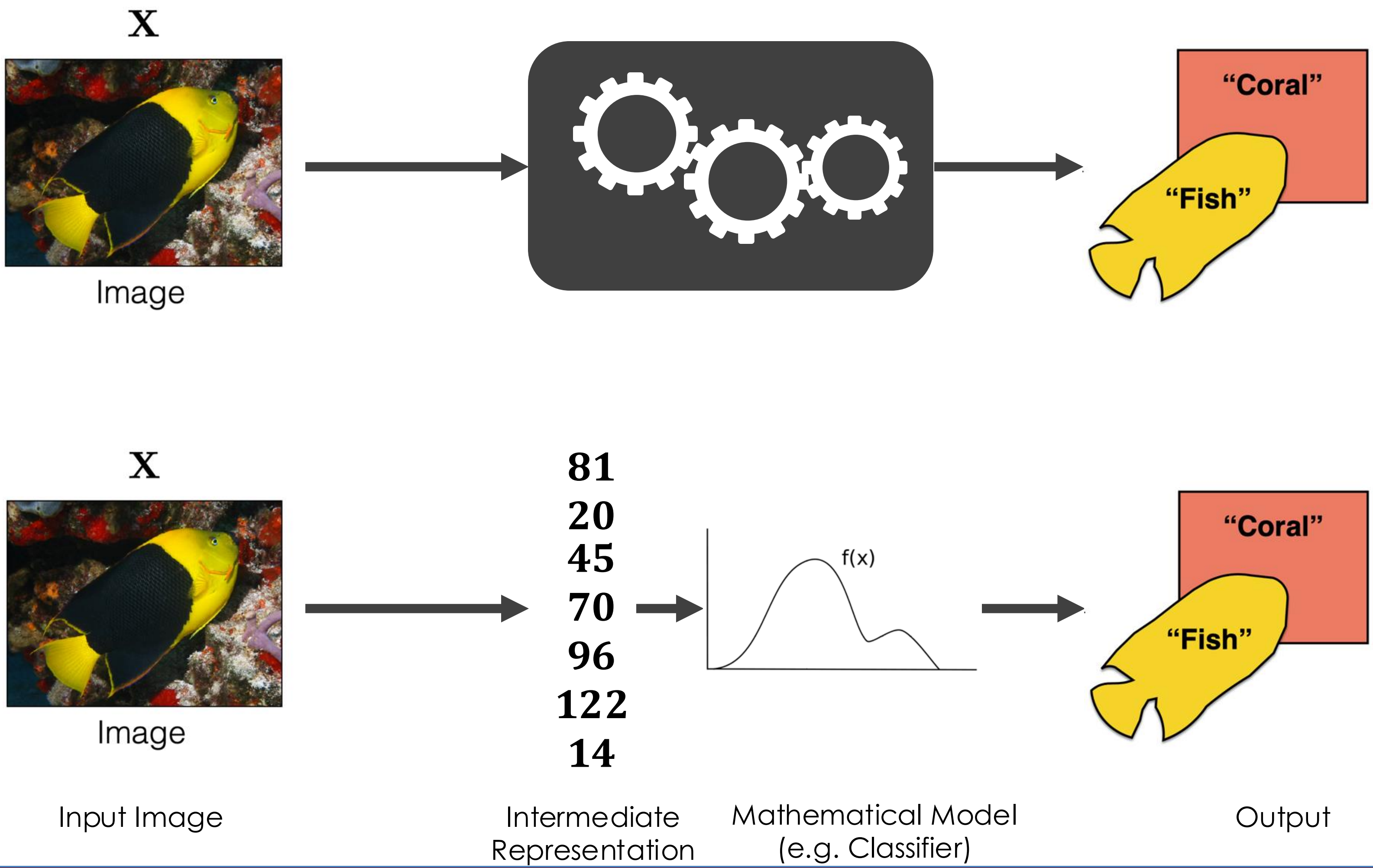
Requirements for Good Representations

- Compact (minimal)
- Explanatory (sufficient)
- Disentangled (independent factors)
- Hierarchical (feature reuse)
- Makes downstream perception problem easier
- Generalizes over many tasks



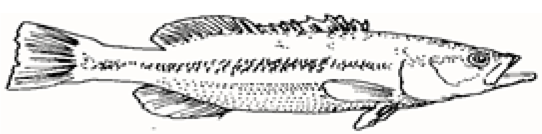
[See "Representation Learning", Bengio 2013, for more commentary]

Typical CV Pipeline

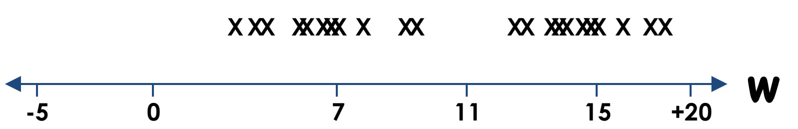


Example

~12 lbs



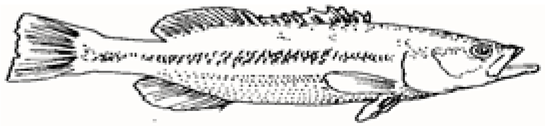
~8 lbs



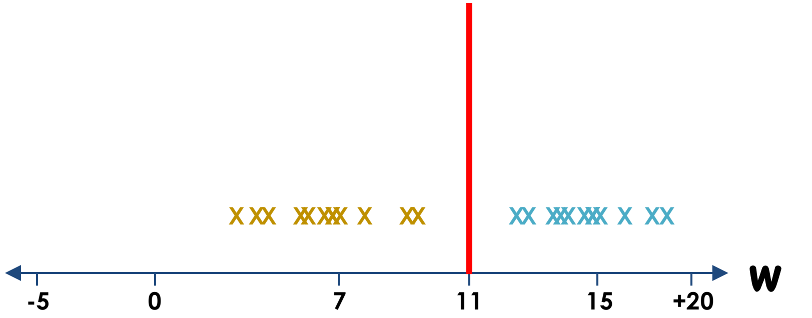
Example from CS331B: Representation Learning in Computer Vision

Example

~12 lbs



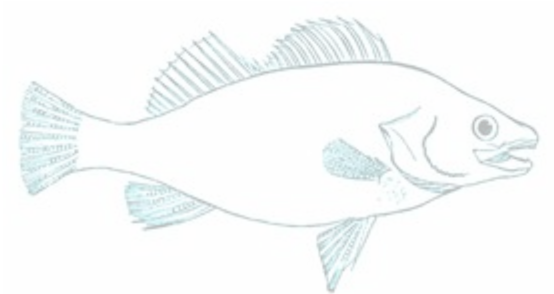
~8 lbs



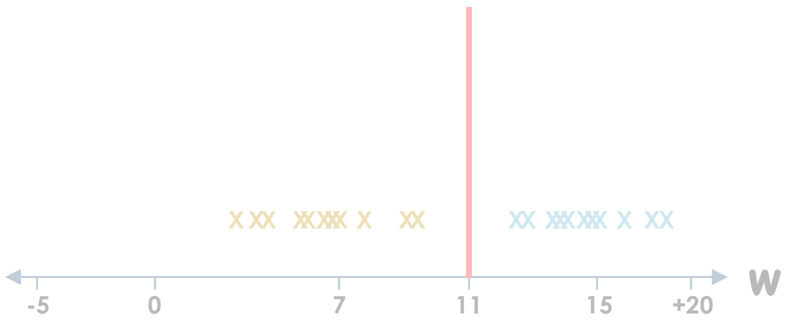
Example from CS331B: Representation Learning in Computer Vision

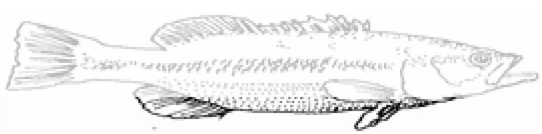
Example from CS331B: Representation Learning in Computer Vision

Example

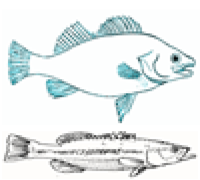


~12 lbs





~8 lbs



Representation

Weight (w)

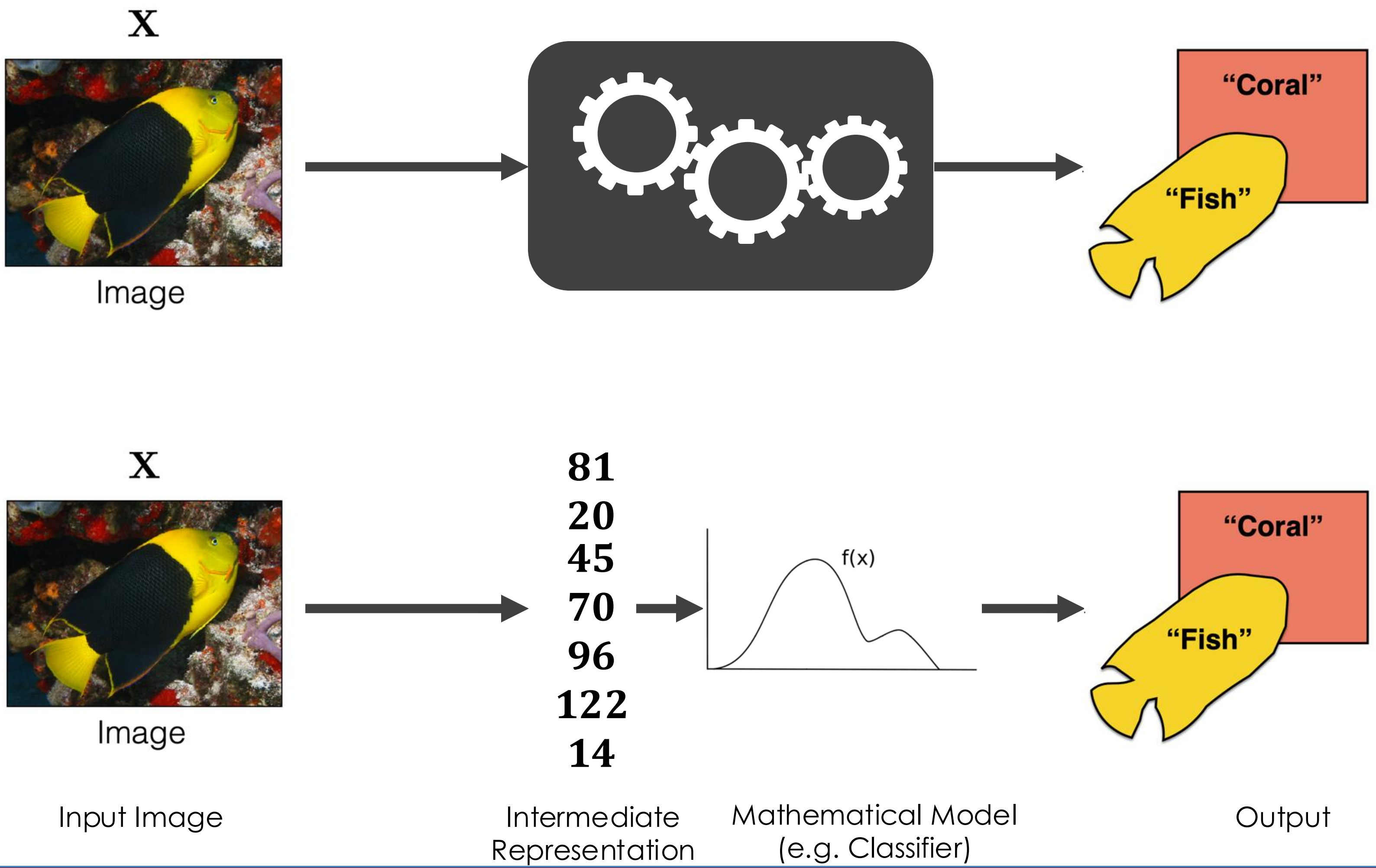
Mathematical Model (Classifier)

w>11

Type A

Type B

Typical CV Pipeline

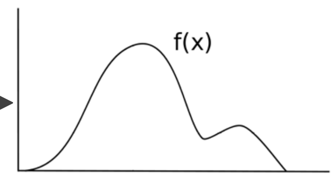


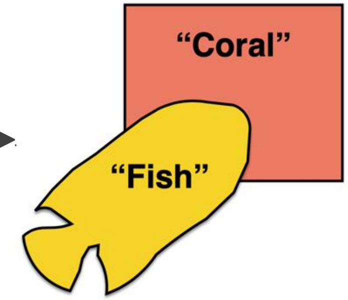
Traditional CV Pipeline

x

Image

81
20
45
70
96
122
14



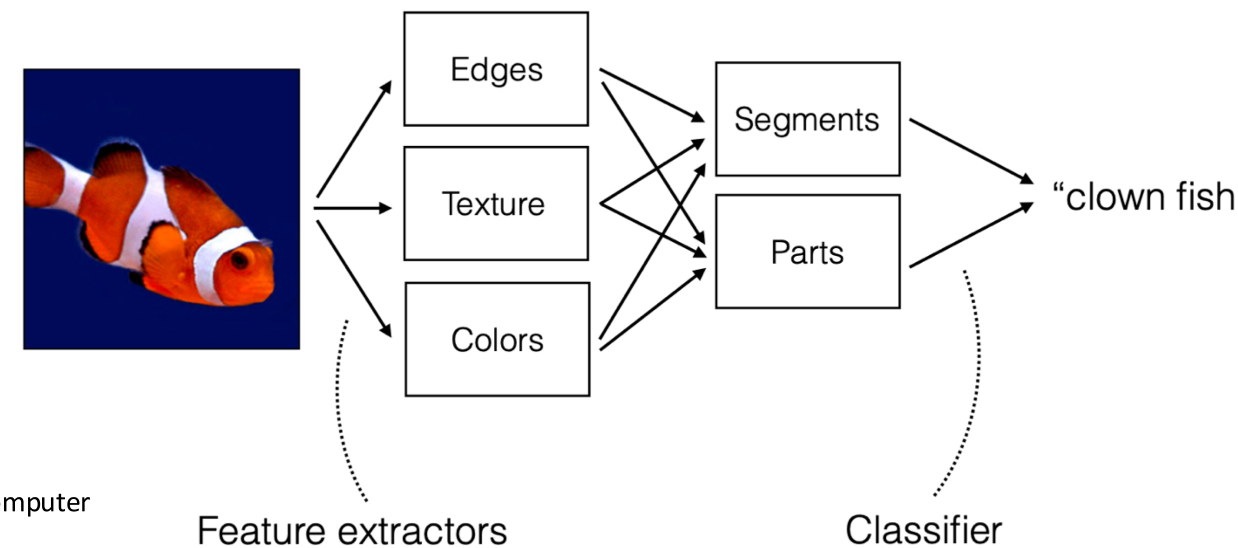


Input Image

Intermediate Representation

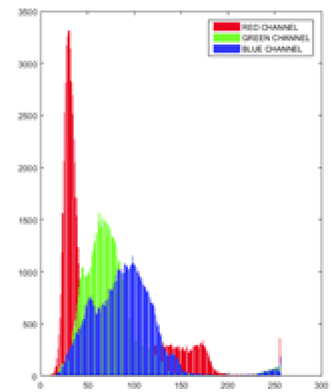
Mathematical Model (e.g. Classifier)

Output



Example from Advances in Computer Vision – MIT – 6.869/6.819

Represent these cats with a cat detector!

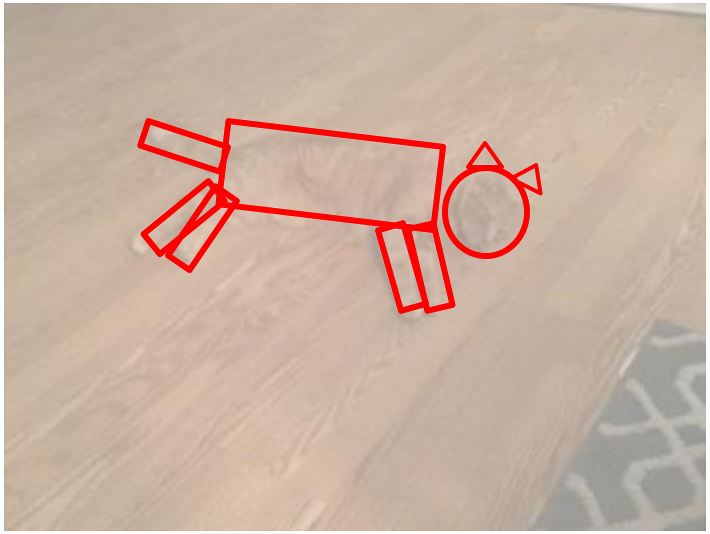


Example from CS331B: Representation Learning in Computer Vision

Represent these cats with a cat detector! (II)

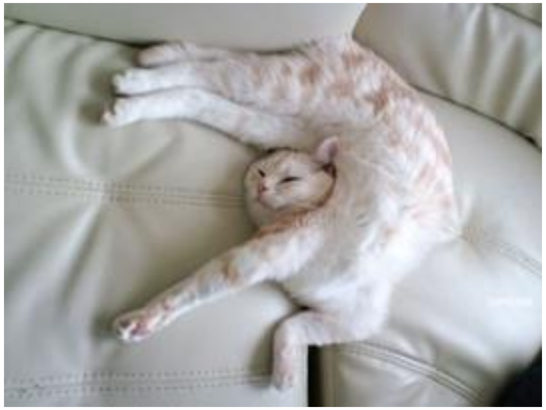
Example from CS331B: Representation Learning in Computer Vision

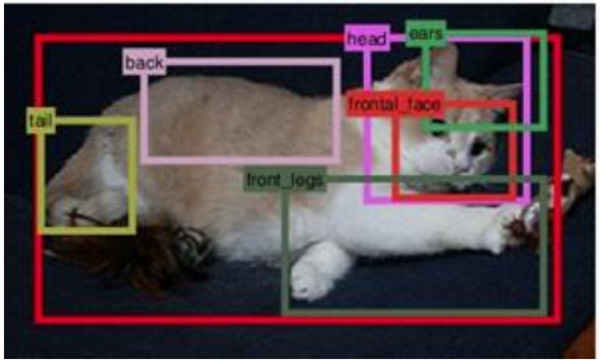
Represent these cats with a cat detector! (II)



Example from CS331B: Representation Learning in Computer Vision

Represent these cats with a cat detector! (III)



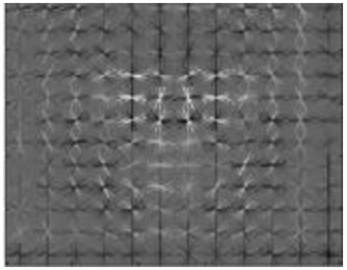



Example from CS331B: Representation Learning in Computer Vision

Represent these cats with a cat detector! (IV)



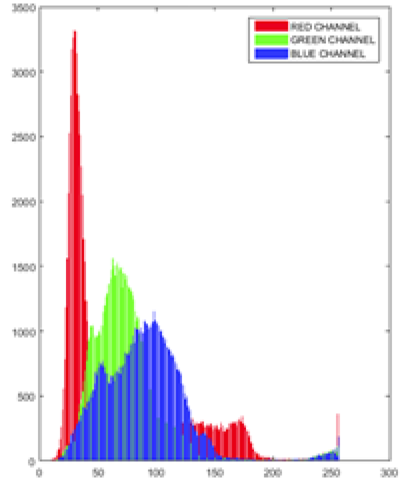


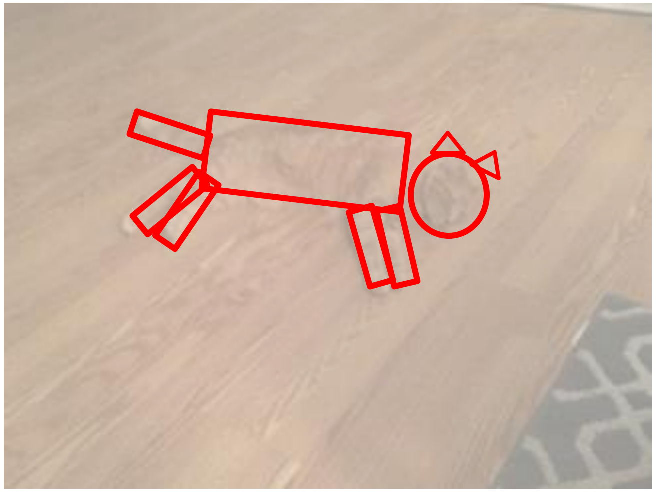


Example from CS331B: Representation Learning in Computer Vision

Summary of Traditional Components

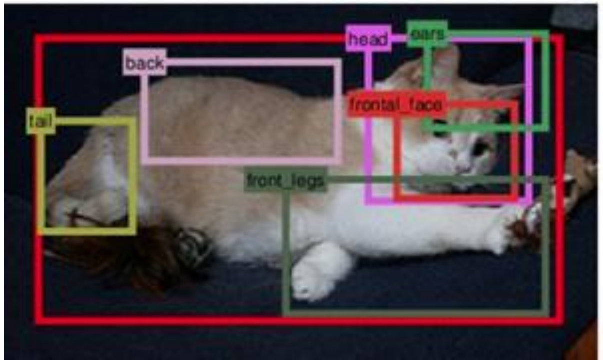
Color Histograms

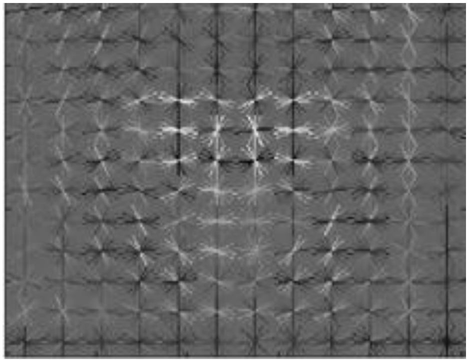




Model based Shapes

Deformable Part based Models (DPM)



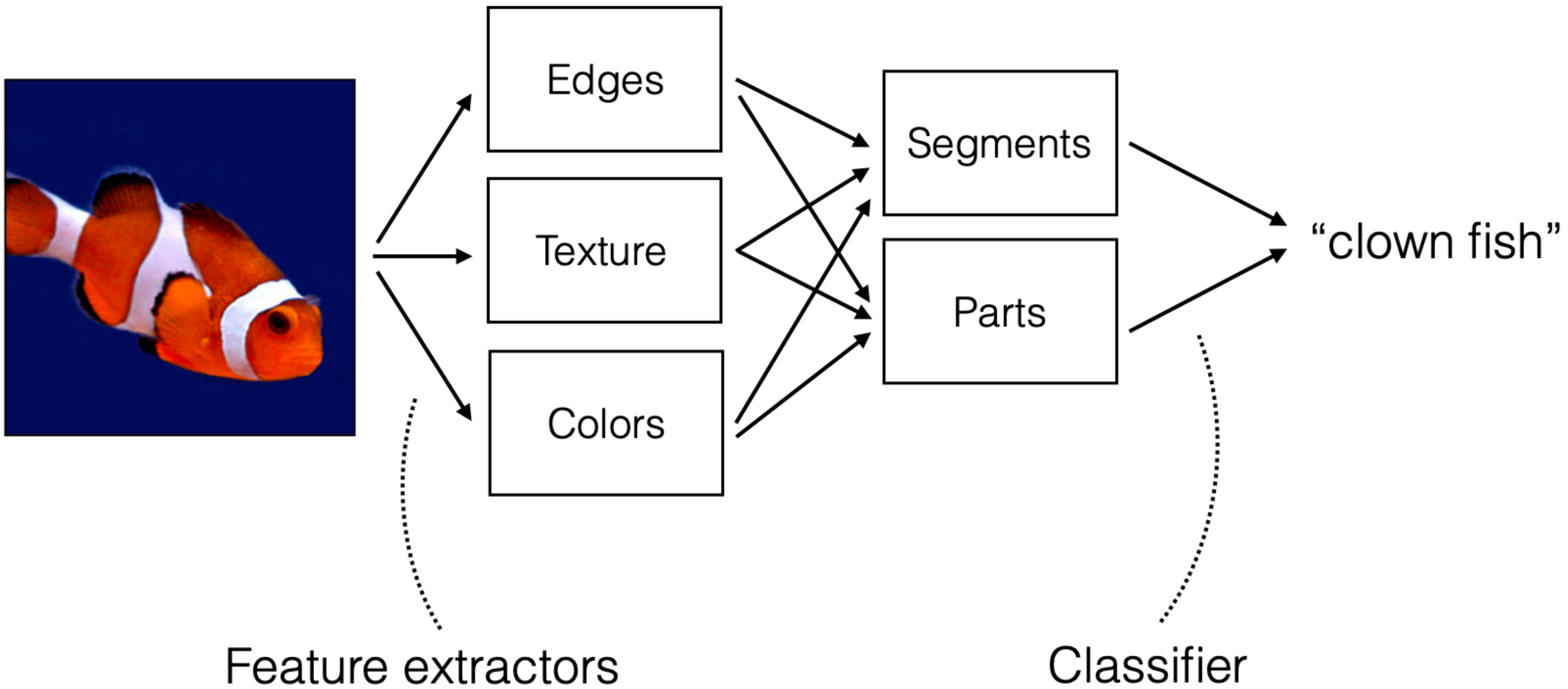


Histogram of Gradients (HOG)

Felzenszwalb et al. 2010.
Dalal and Triggs, 2005.
Beis and Lowe, 1997.

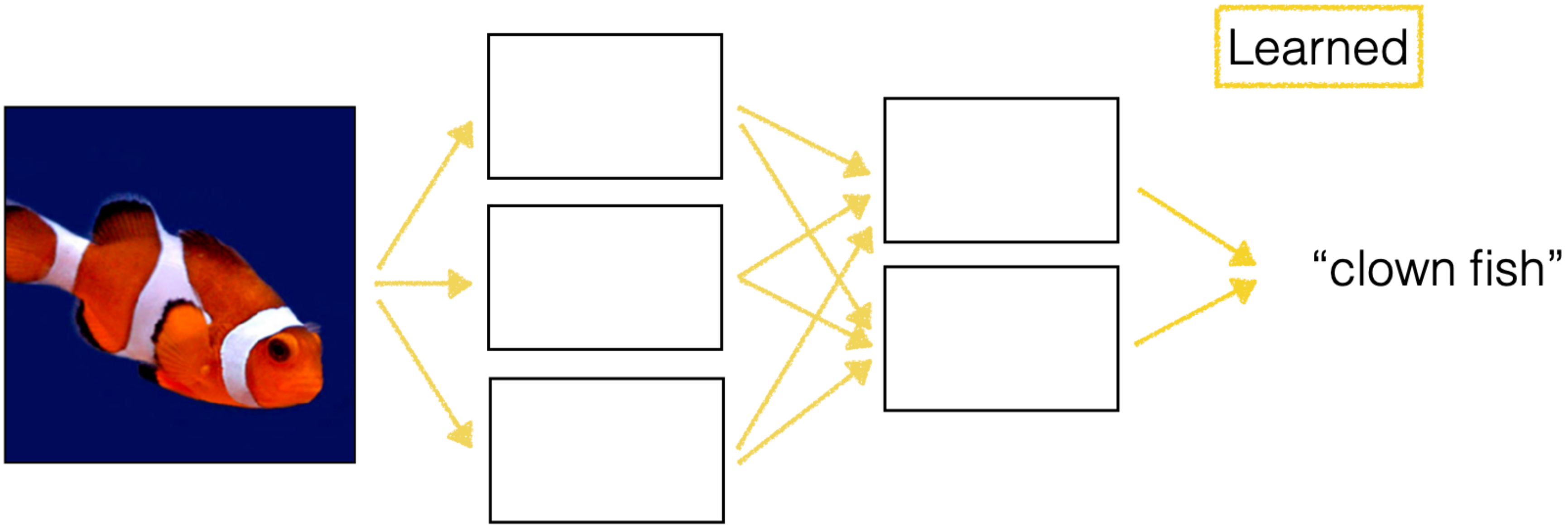
Example from CS331B: Representation Learning in Computer Vision

Traditional CV Pipeline



Example from Advances in Computer Vision – MIT – 6.869/6.819

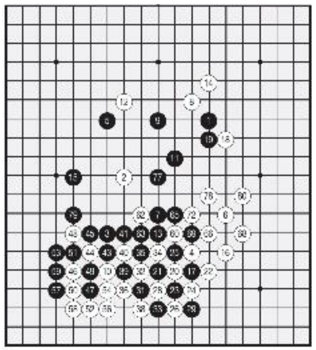
Learned CV Pipeline



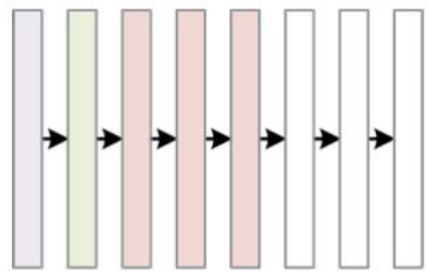
Example from Advances in Computer Vision – MIT – 6.869/6.819

Learned CV Pipeline

Go playing can be solved in representation space.



3^{361} states?



Good representation

$256^{3x500x500}$?

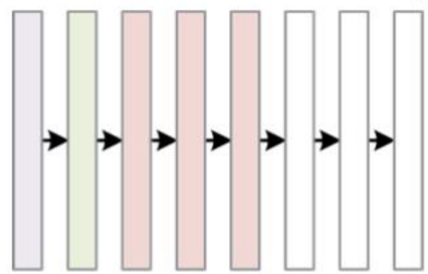
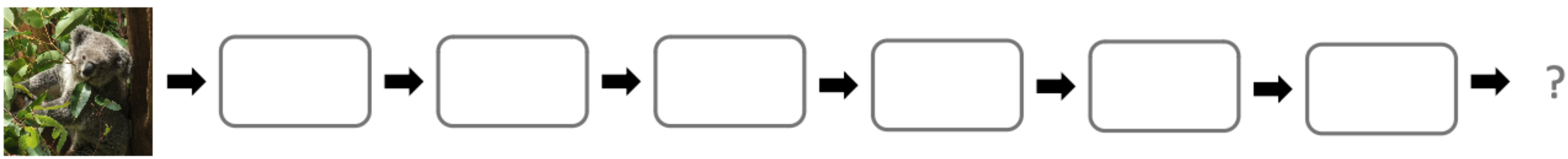


Image recognition is solved in representation space.

Examples from MIT - 6.8300/1 Advances in Computer Vision

Learned CV Pipeline

general modules (instead of specialized features)



compose simple modules into complex functions

- build multiple levels of abstractions
- learn by back-prop
- learn from data
- reduce domain knowledge and feature engineering

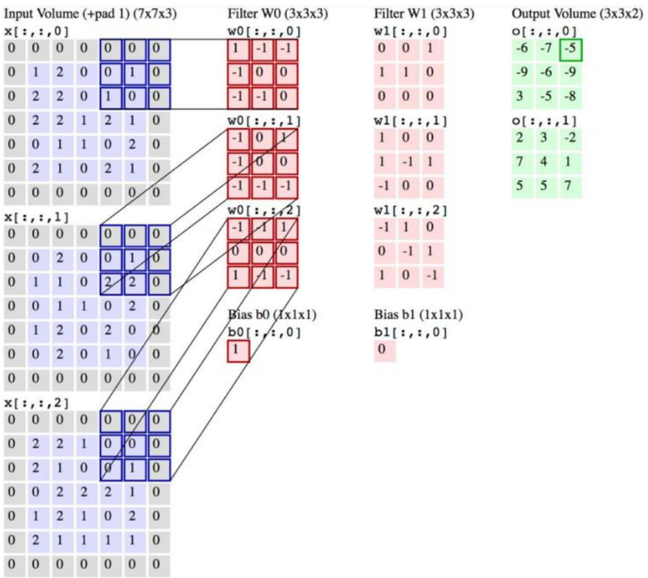
Examples from MIT - 6.8300/1 Advances in Computer Vision

Introduction to Neural Networks and CNNs

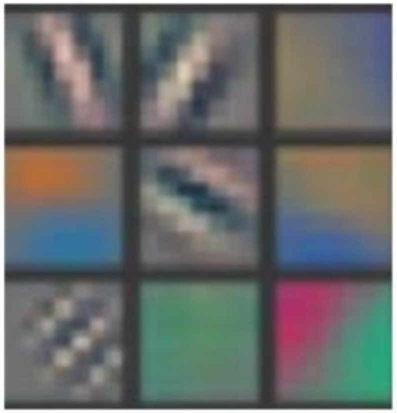
- Check Friday's CA Session (5/9)

Components

- Each convolutional "layer" is represented by a 3D tensor of shape $[h \times w \times n_{channels}]$
- Between two convolutional layers, the weights are of the shape [`relative x-position, relative y-position, input channels, output channels`]
- "Convolve" operation consists of 4 hyperparameters:
  - Number of filters, or *depth* (each channel also called an "activation map")
  - *Spatial extent*, or *receptive field*
  - The stride
  - Amount of zero-padding
- With this, the shape of layer convolved from layer − 1 is:
  - $[\,(W - F + 2P)/S + 1, (H - F + 2P)/S + 1, K\,]$



Multiple Levels of Representations



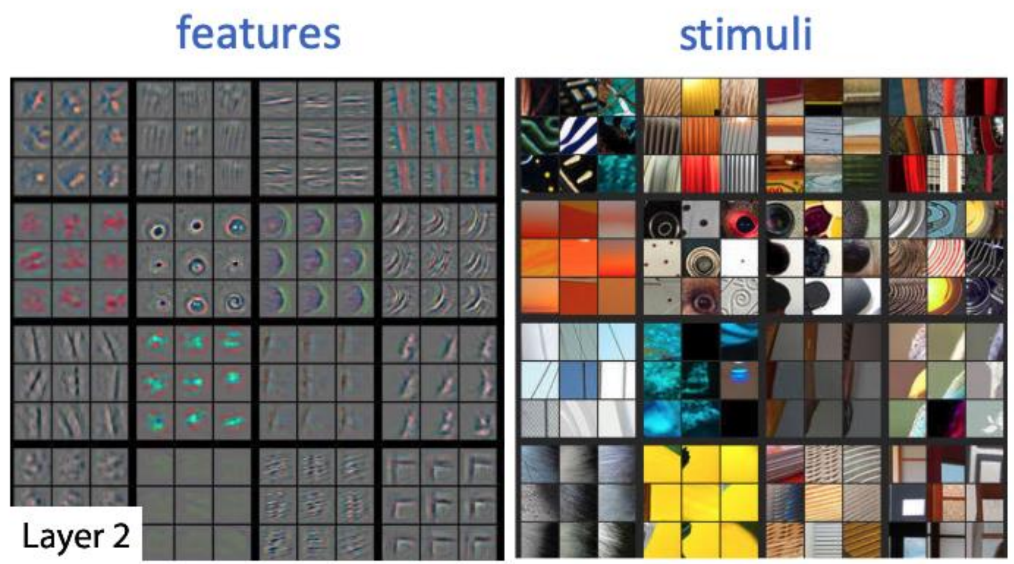
features

stimuli
(patches with the highest
1-hot activations)

"Visualizing and Understanding Convolutional Networks", Zeiler & Fergus. ECCV 2014

Examples from MIT - 6.8300/1 Advances in Computer Vision

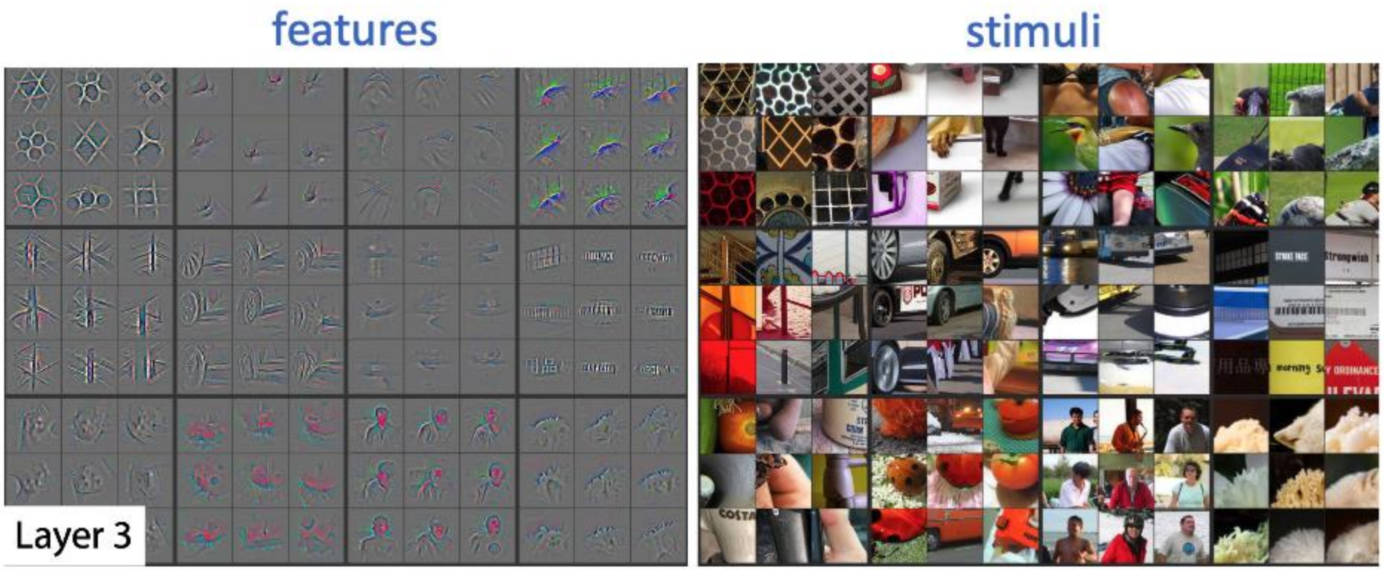
Multiple Levels of Representations



"Visualizing and Understanding Convolutional Networks", Zeiler & Fergus. ECCV 2014

Examples from MIT - 6.8300/1 Advances in Computer Vision

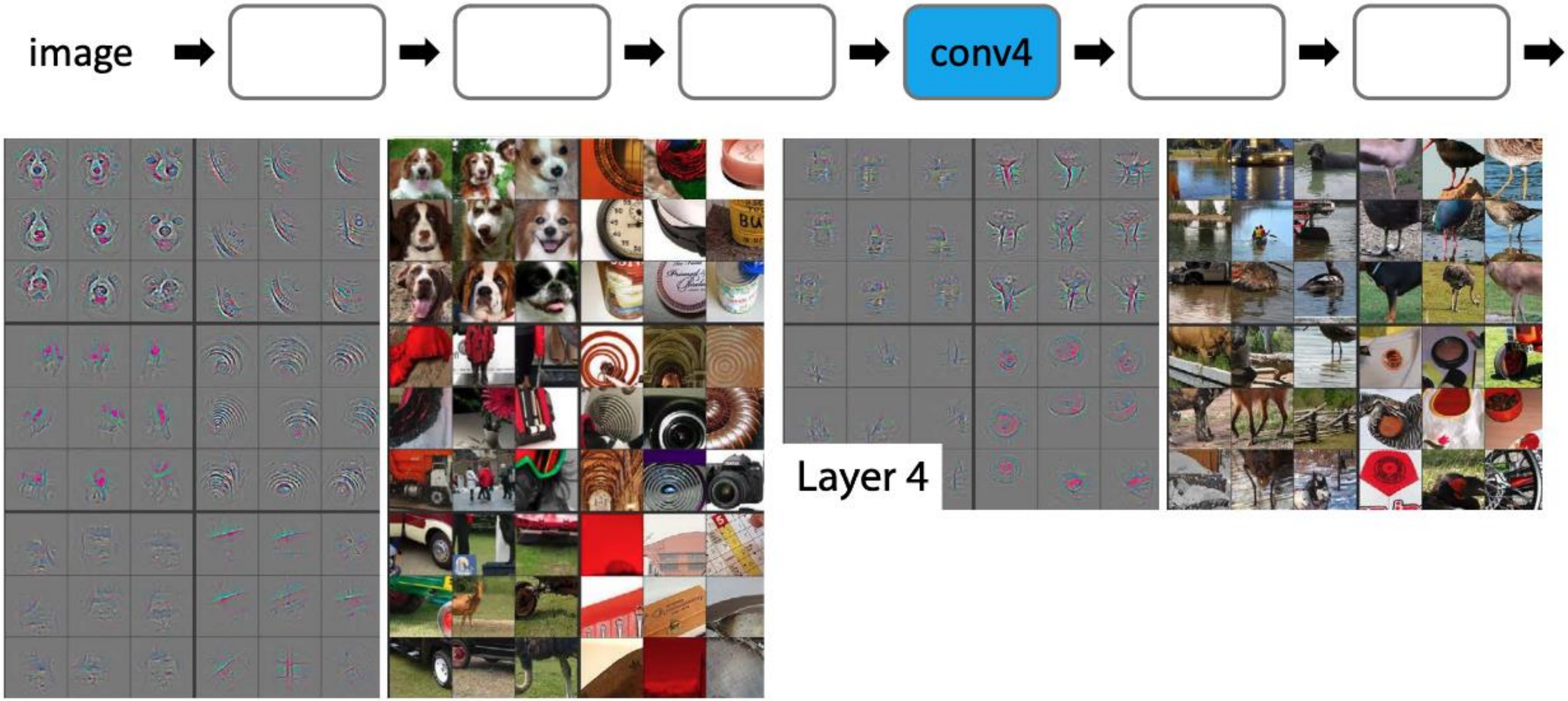
Multiple Levels of Representations



“Visualizing and Understanding Convolutional Networks”, Zeiler & Fergus. ECCV 2014

Examples from MIT - 6.8300/1 Advances in Computer Vision

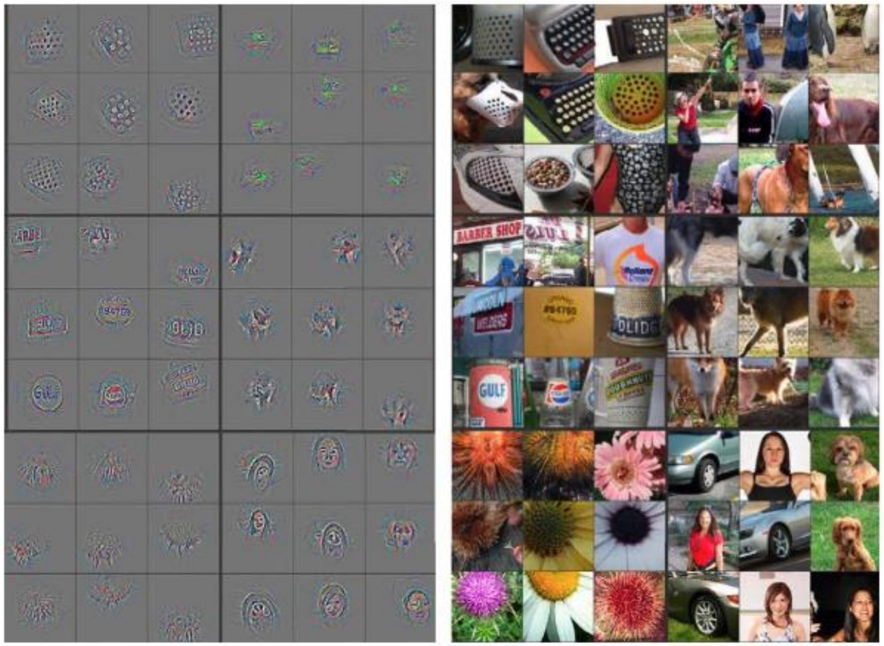
Multiple Levels of Representations

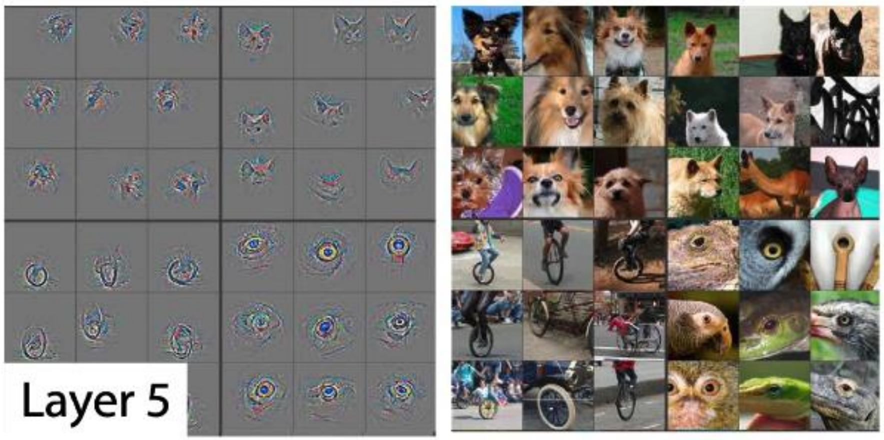


"Visualizing and Understanding Convolutional Networks", Zeiler & Fergus. ECCV 2014

Examples from MIT - 6.8300/1 Advances in Computer Vision

Multiple Levels of Representations



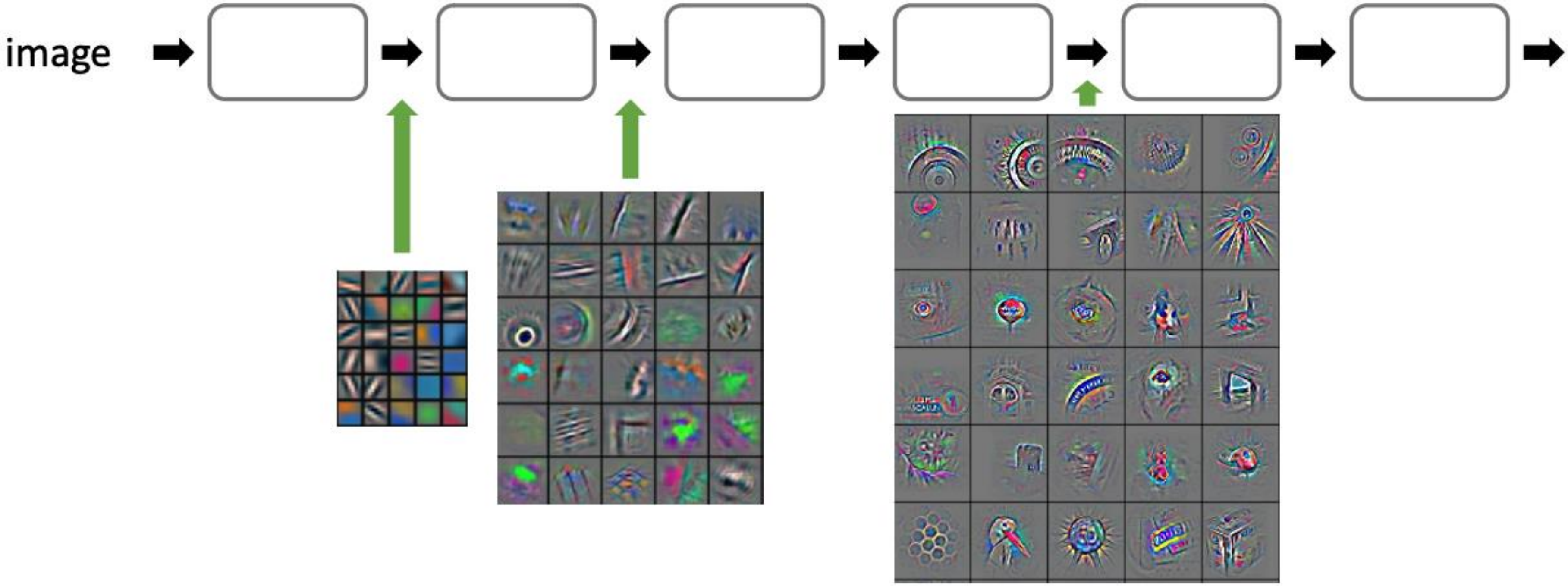


Deeper layers have “higher-level” features.

“Visualizing and Understanding Convolutional Networks”, Zeiler & Fergus. ECCV 2014

Examples from MIT - 6.8300/1 Advances in Computer Vision

Multiple Levels of Representations



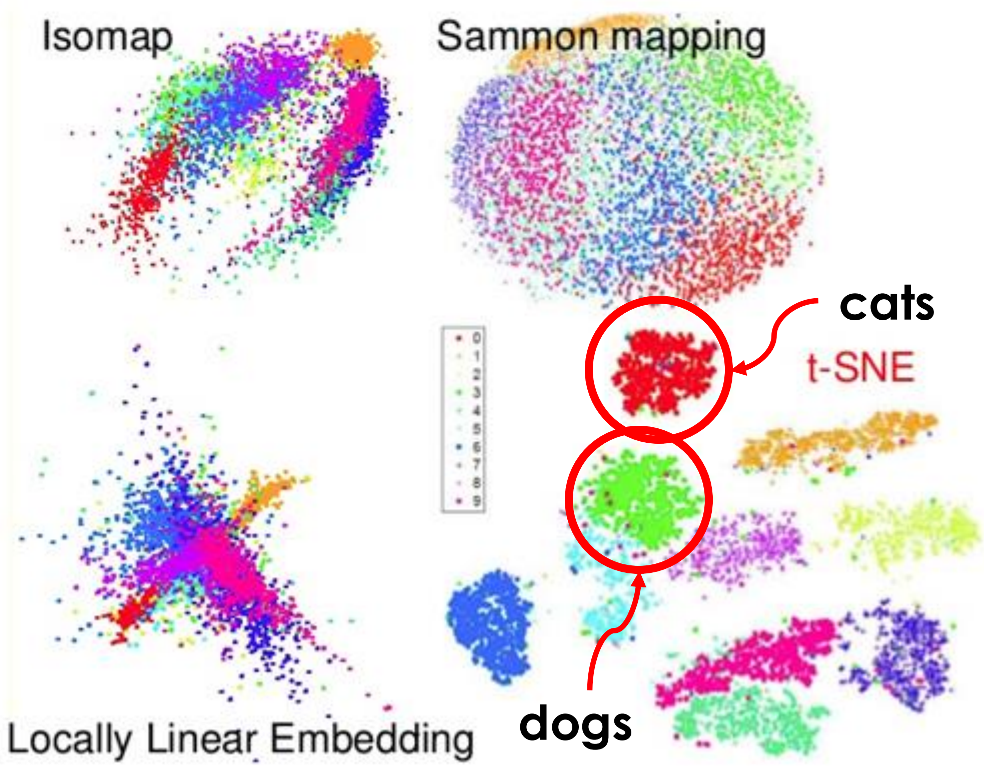
Deeper layers have “higher-level” features.

“Visualizing and Understanding Convolutional Networks”, Zeiler & Fergus. ECCV 2014

Examples from MIT - 6.8300/1 Advances in Computer Vision

Understanding representations through low-dimensional embeddings

- 6000 MNIST Digit
  - tSNE
  - Isomap
  - Sammon M
  - LLE

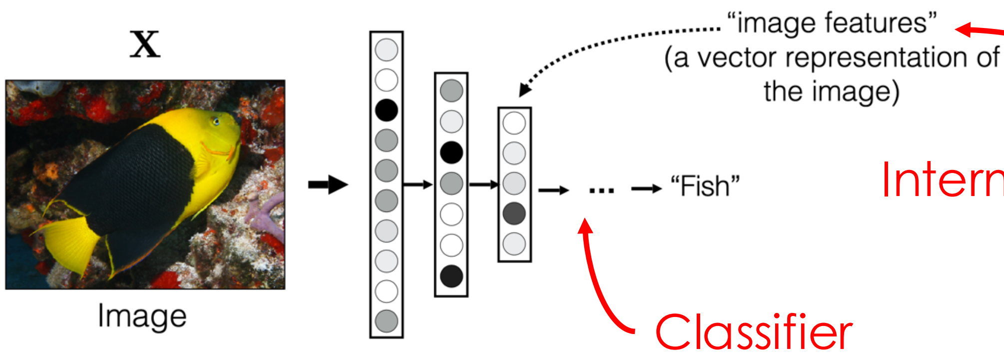


Understanding representations through low-dimensional embeddings

- tSNE

Van der Maaten & Hinton. 2008

How do you learn a representation?



Intermediate representation

Classifier

Training data

Input images $\{x_1, y_1\}$ Labels

$\{x_2, y_2\}$ $\longrightarrow$ Learner $\longrightarrow$ $f : X \rightarrow Y$

$\{x_3, y_3\}$

$\ldots$

$$f^* = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{N} \mathcal{L}(f(x_i), y_i)$$

Loss function/Cost

Learned Representations are Transferable

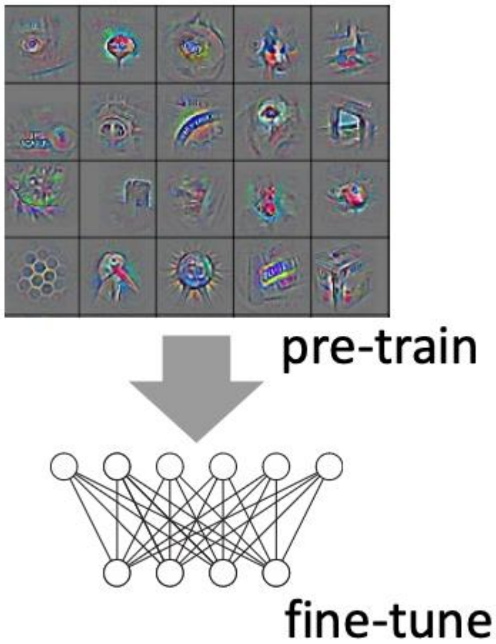
The single most important discovery in DL revolution

Transfer learning:

- pre-train on large-scale data
- fine-tune on small-scale data

- enable DL for small datasets
- revolutionize computer vision

- data: engine for general representation
- GPT: a similar principle



“DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition”, Donahue et al. arXiv 2013
“Visualizing and Understanding Convolutional Networks”, Zeiler & Fergus. arXiv 2013
“CNN Features off-the-shelf: an Astounding Baseline for Recognition”, Razavian. arXiv 2014

Slide from MIT - 6.8300/1 Advances in Computer Vision

Transfer learning: Pre-training & Fine-tuning

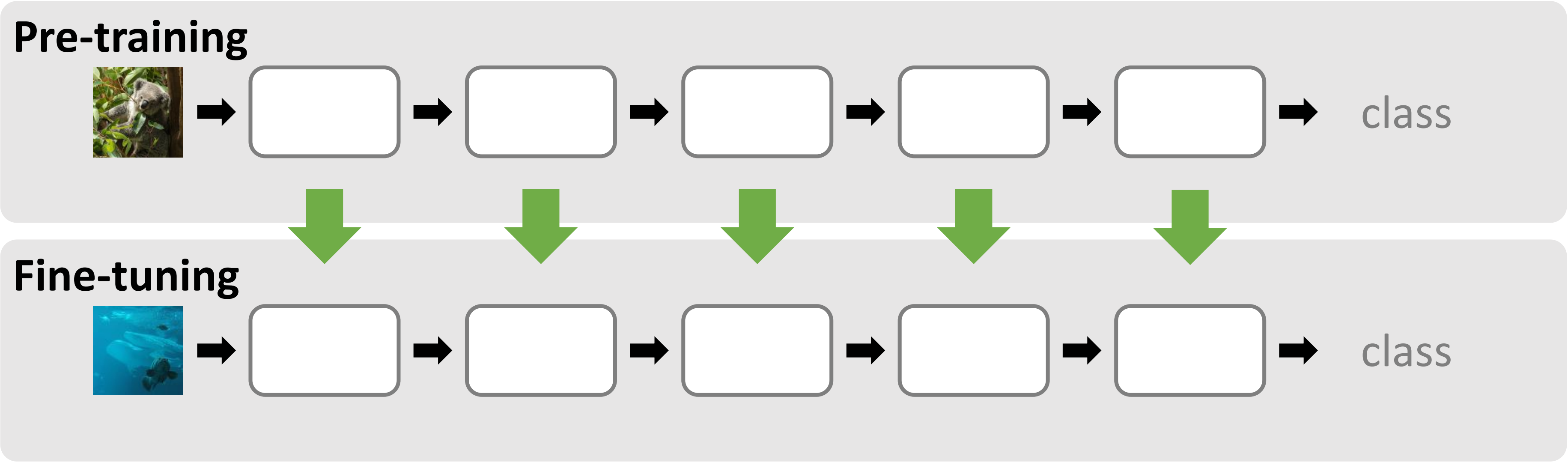
Pre-training

Pre-training:

- to learn **general** representations
- on **large-scale** data
- train for a **long** time
- with **large** models

Slide from MIT - 6.8300/1 Advances in Computer Vision

Transfer learning: Pre-training & Fine-tuning

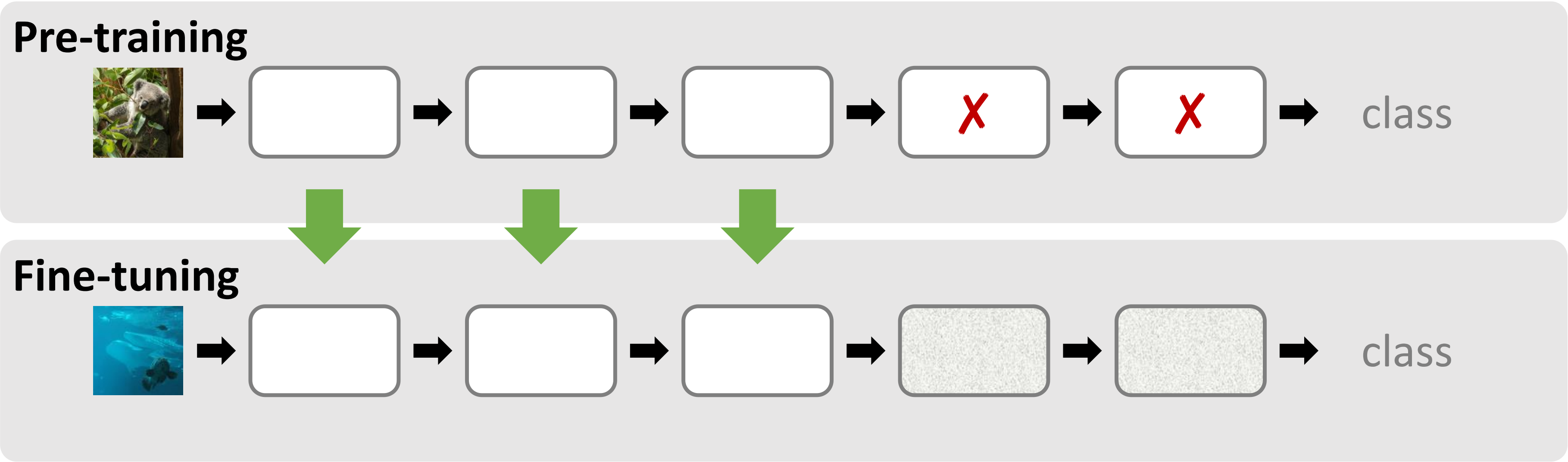


Fine-tuning:

- transfer weights to **specific** tasks
- on **small-scale** data
- train for a **short** time, **lower** learning rate
- enable **large** models with lower risk of overfitting

Slide from MIT - 6.8300/1 Advances in Computer Vision

Transfer learning: Pre-training & Fine-tuning

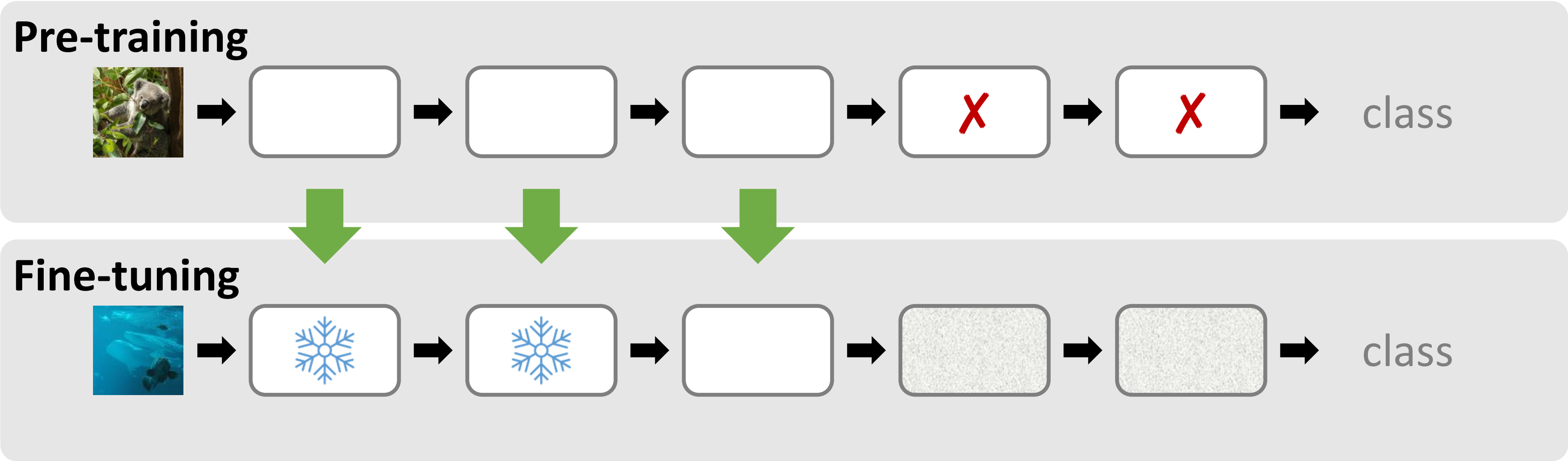


Partial transfer

- pre-train and target domains may differ
- highest-level features are too adapted to pre-training
- randomly initialize new layers

Slide from MIT - 6.8300/1 Advances in Computer Vision

Transfer learning: Pre-training & Fine-tuning

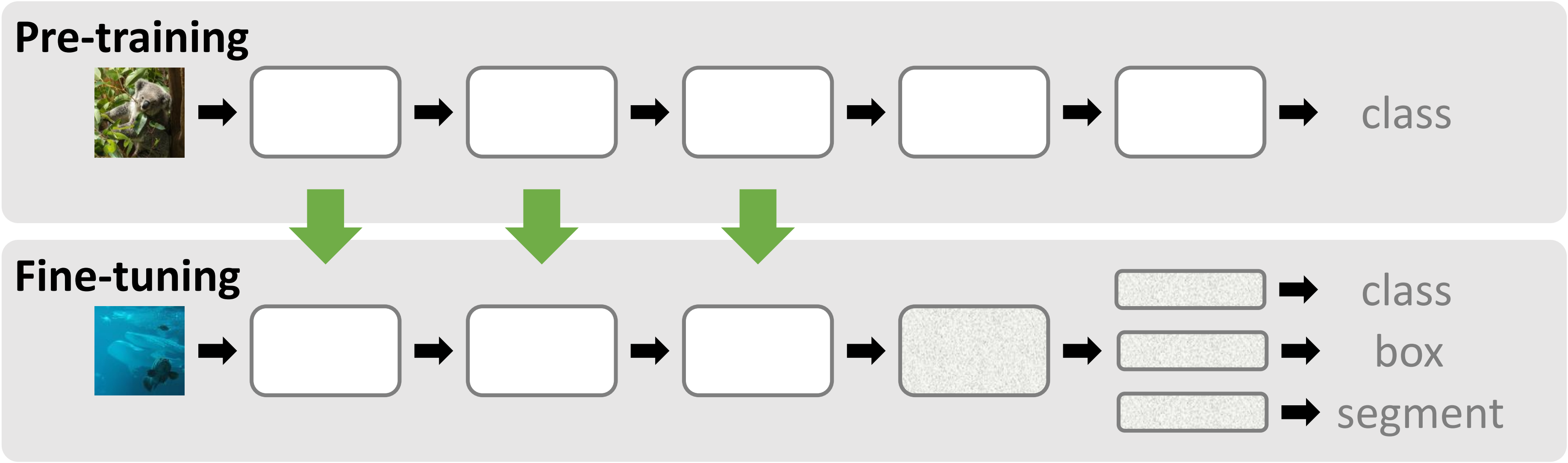


Frozen weights

- freeze some/all pre-trained weights
- reduce overfitting if data is too little
- save memory, speed up training

Slide from MIT - 6.8300/1 Advances in Computer Vision

Transfer learning: Pre-training & Fine-tuning



Network surgery

- re-purpose the model for other tasks (detect, segment)
- general features + task-specific predictions

Slide from MIT - 6.8300/1 Advances in Computer Vision

How can we (pre-)train good representations?

- **Model** (network architecture)
  - scaling: deep, wide, large
  - inductive bias: convolution, recurrency, attention
- **Data**
  - scaling
  - curating, cleaning, filtering, ...
- **Learning objective**
  - supervised
  - unsupervised
  - self-supervised

Slide from MIT - 6.8300/1 Advances in Computer Vision

Supervised Object Recognition

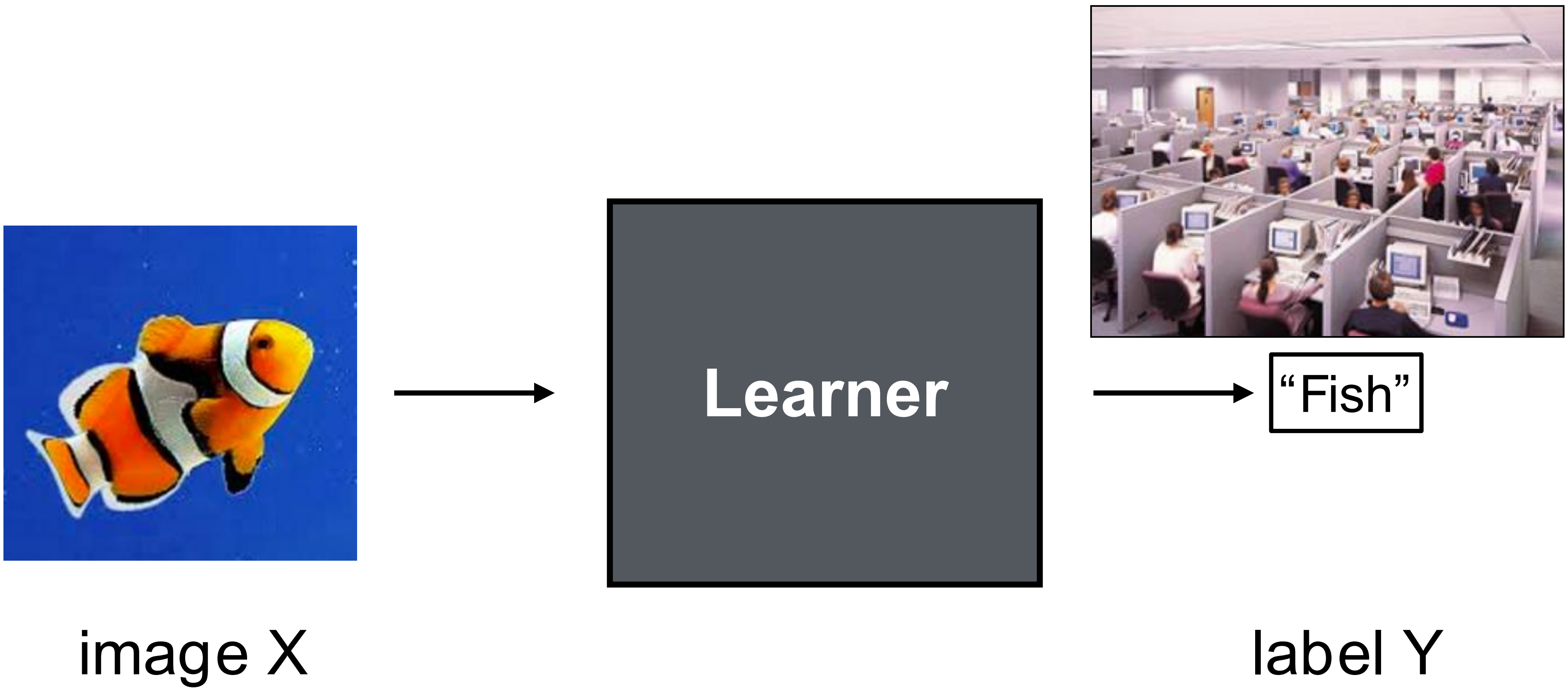
Learner

"Fish"

image X

label Y

Supervised Object Recognition



Supervised Object Recognition

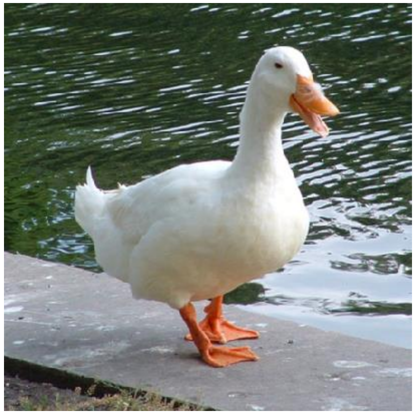
Learner

"Fish"

image X

label Y

Supervised Object Recognition



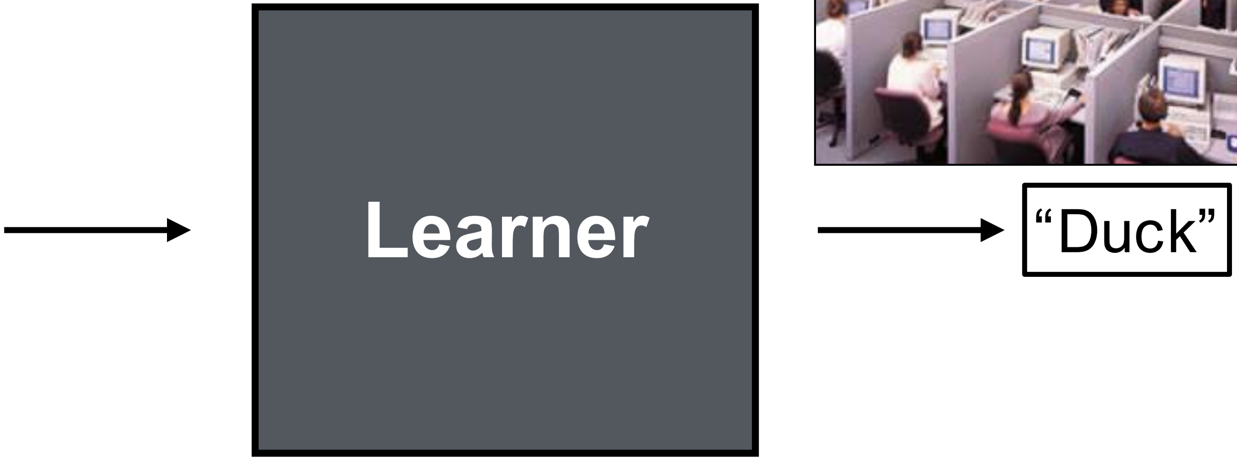


image X

label Y

Learning in the wild

Time lapse of a baby playing with toys. Francis Vachon. YouTube

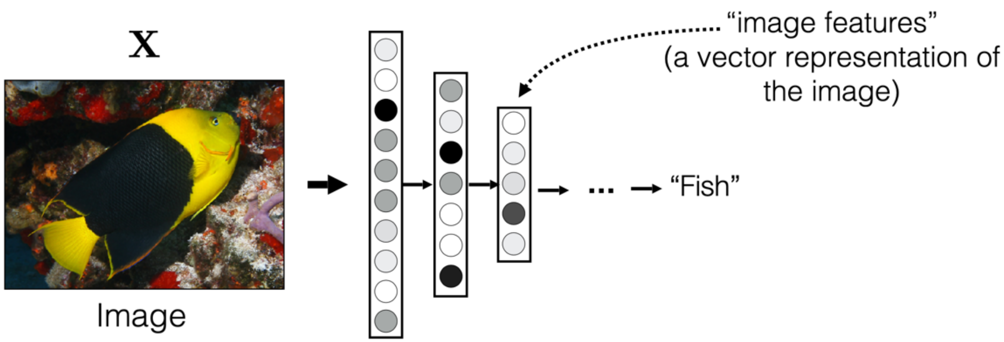
Supervised Computer Vision

- Informative
- Expensive
- Limited to teacher knowledge

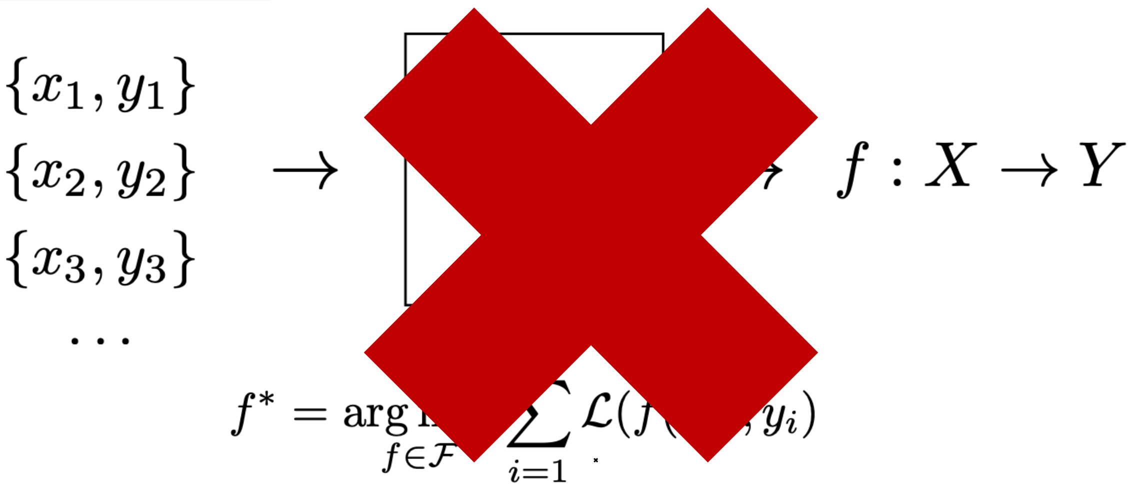
Vision in Nature

- Cheap
- Noisy
- Harder to interpret

Learning without Labels



Training data

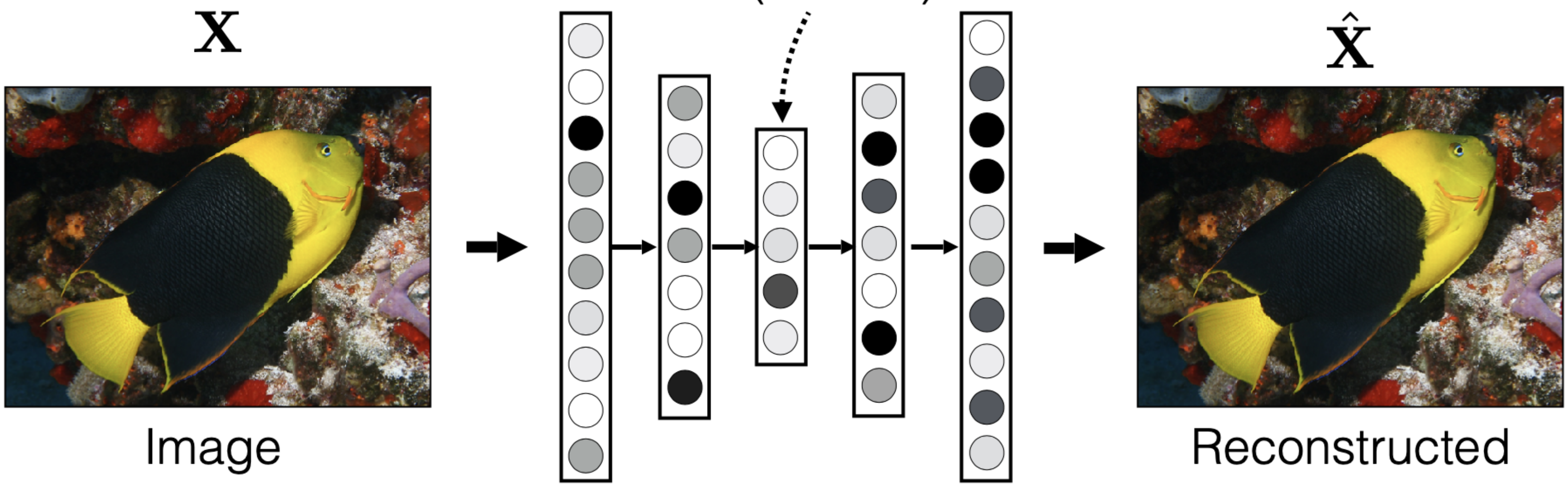


Unsupervised Representation Learning

No category or symbolic label. Instead: learn to reconstruct.

compressed image code (vector **z**)

Intermediate representatio



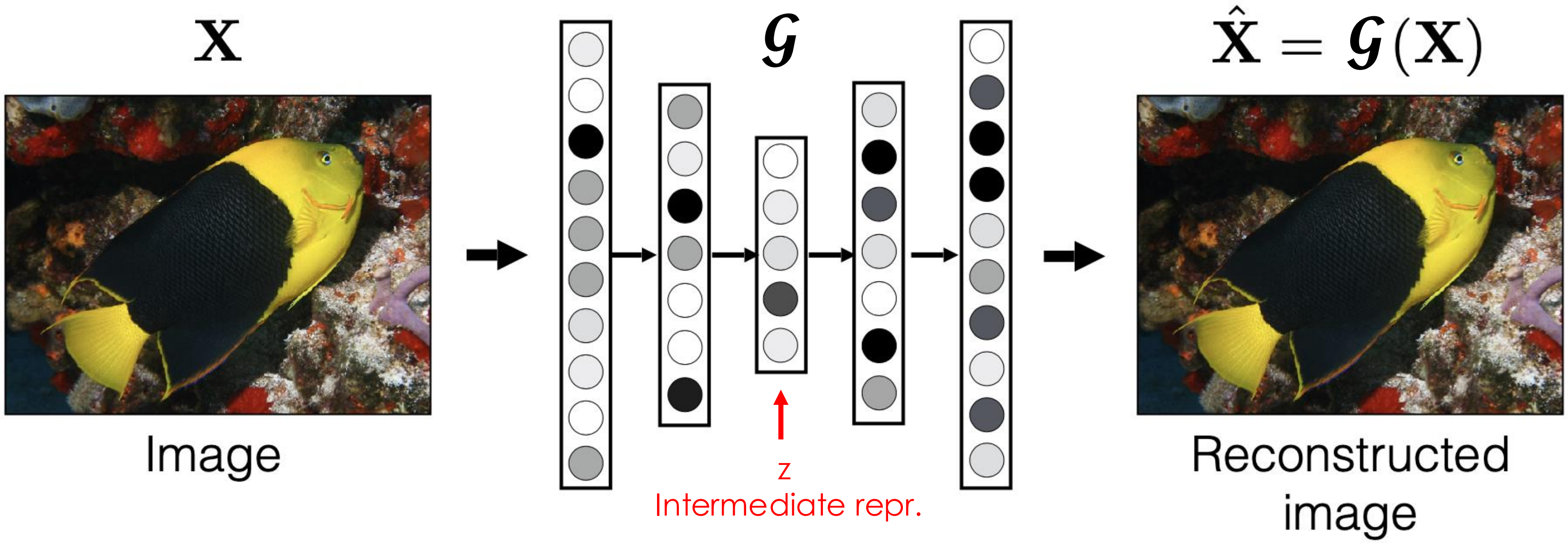
Image

Reconstructed image

One kind of unsupervised model: "Autoencoder"

[e.g., Hinton & Salakhutdinov, Science 2006]

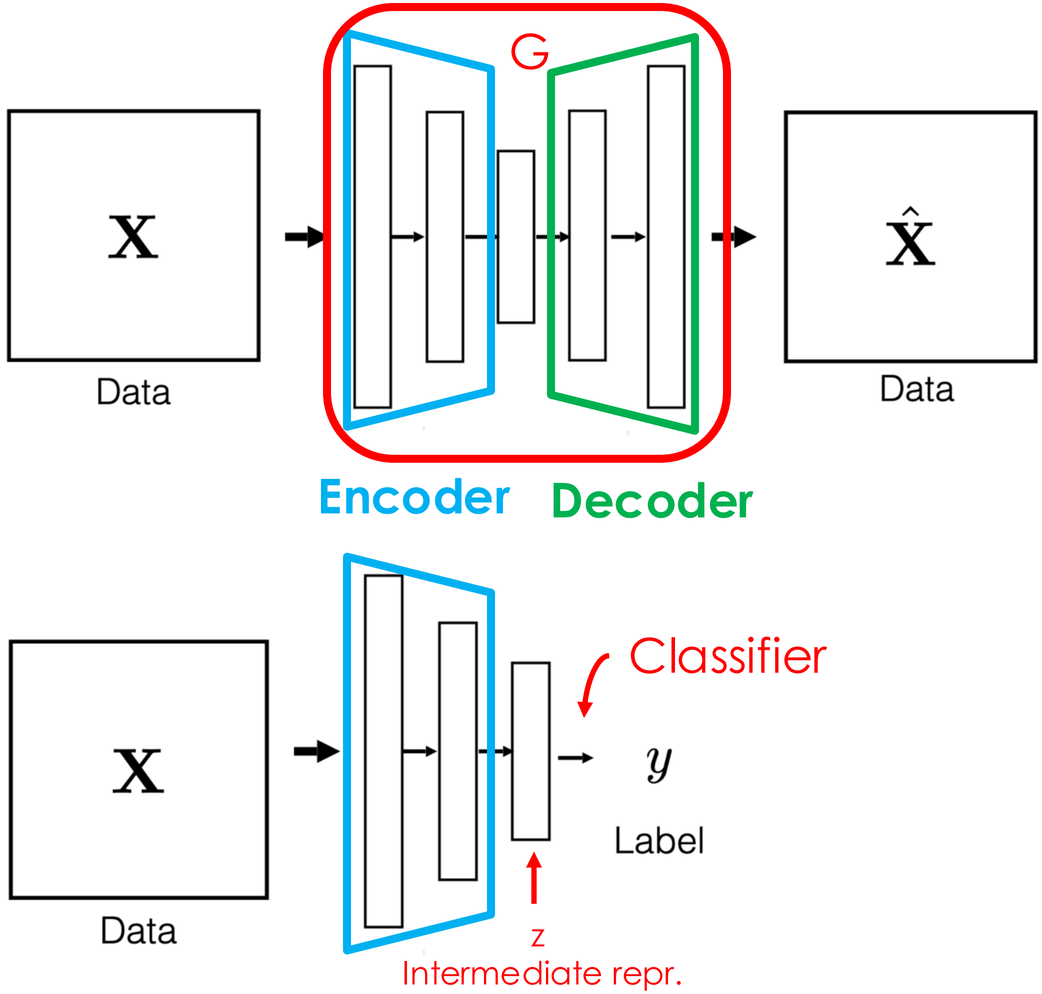
Autoencoder



$$\arg\min_{\mathcal{F}} \boxed{\mathbb{E}_{\mathbf{X}}[||\mathcal{G}(\mathbf{X}) - \mathbf{X}||]}$$

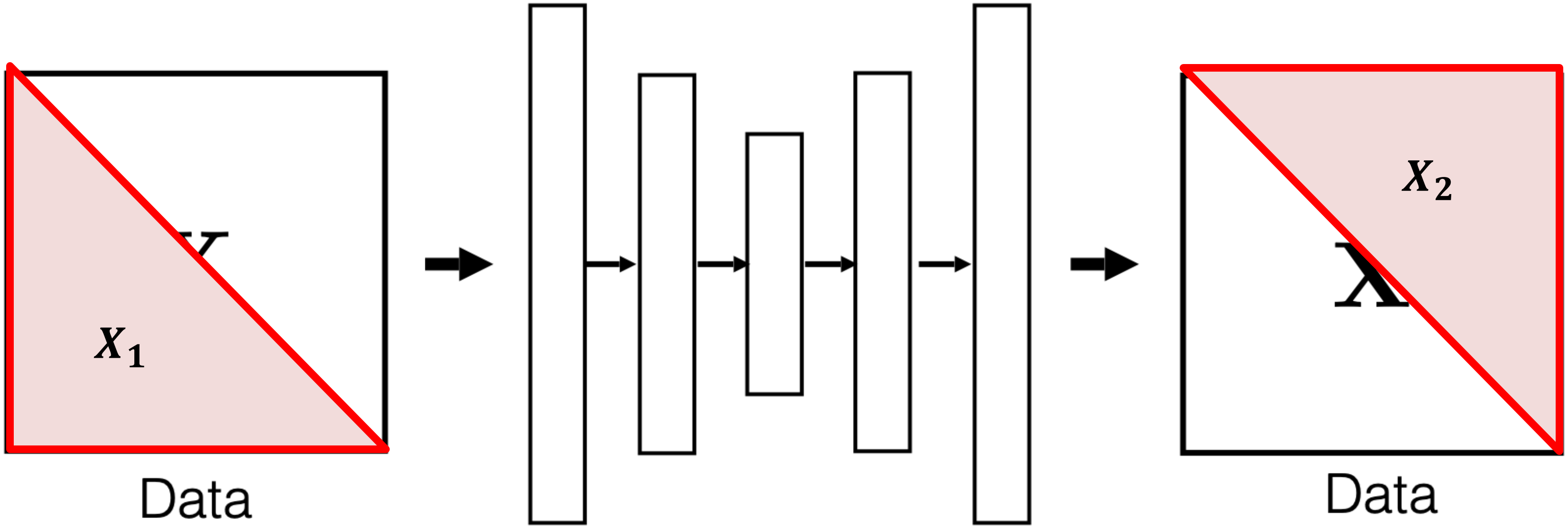
Reconstruction loss to minimize by finding optimal G

Data Compression & Task Transfer



e.g., image classification

Self-Supervision



$$G\left(X\right) = \hat{X}$$
$$G(X_1) = \hat{X}_2$$

Predictive Learning: Language Models

Next word prediction (GPT)

- Predict the next word (token) given a prefix

"The students opened their __ → model → books

Radford, et al., "Improving Language Understanding by Generative Pre-Training", 2018

Slide from MIT - 6.8300/1 Advances in Computer Vision

Predictive Learning: Language Models

Masked language modeling (BERT)

- Predict the masked words (tokens) in a text

The __ opened their __ and began to __ → model → students .. books .. read

Devlin, et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, 2018

Slide from MIT - 6.8300/1 Advances in Computer Vision

Predictive Learning: Computer Vision

Masked image modeling (Context Encoders)

- Predict the masked regions using ConvNets

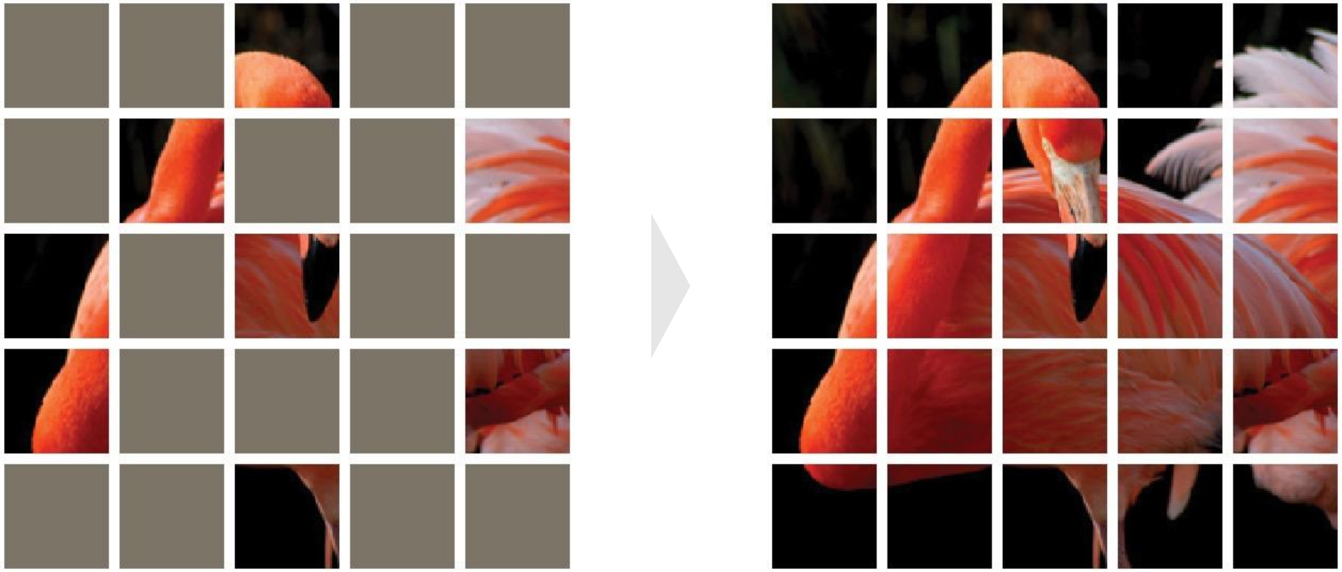
Pathak, et al., “Context Encoders: Feature Learning by Inpainting”, CVPR 2016

Slide from MIT - 6.8300/1 Advances in Computer Vision

Predictive Learning: Computer Vision

Masked image modeling (Masked Autoencoder)

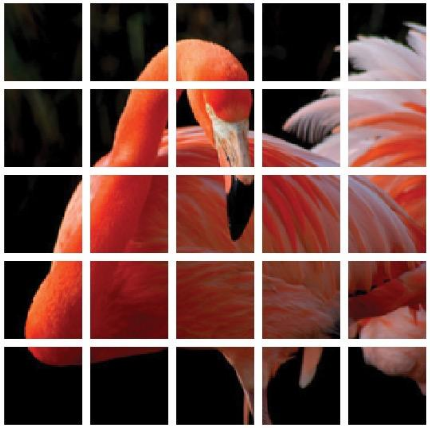
- Predict the masked patches using Transformers



He, et al., “Masked Autoencoders Are Scalable Vision Learners”, CVPR 2022

Slide from MIT - 6.8300/1 Advances in Computer Vision

Masked Autoencoder (MAE)

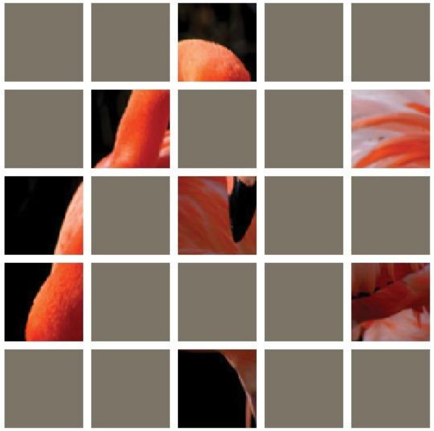


patches as visual tokens
(Vision Transformer)

He, et al., "Masked Autoencoders Are Scalable Vision Learners", CVPR 2022

Slide from MIT - 6.8300/1 Advances in Computer Vision

Masked Autoencoder (MAE)

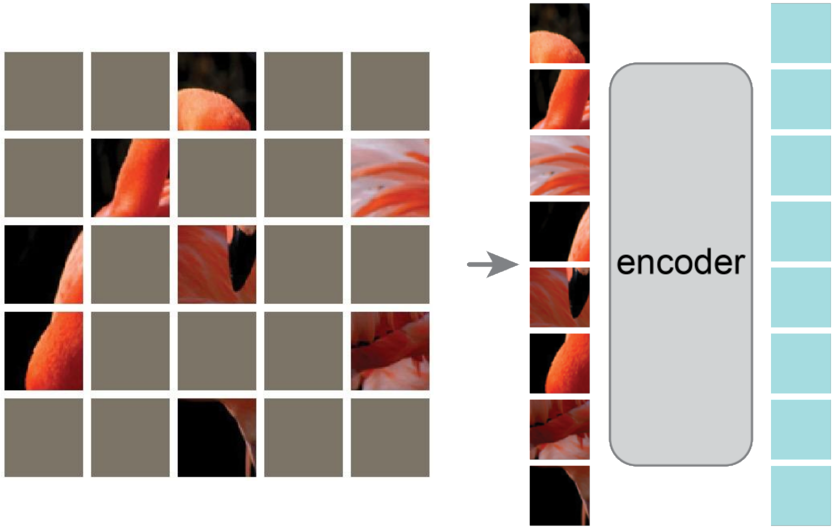


random masking

He, et al., “Masked Autoencoders Are Scalable Vision Learners”, CVPR 2022

Slide from MIT - 6.8300/1 Advances in Computer Vision

Masked Autoencoder (MAE)

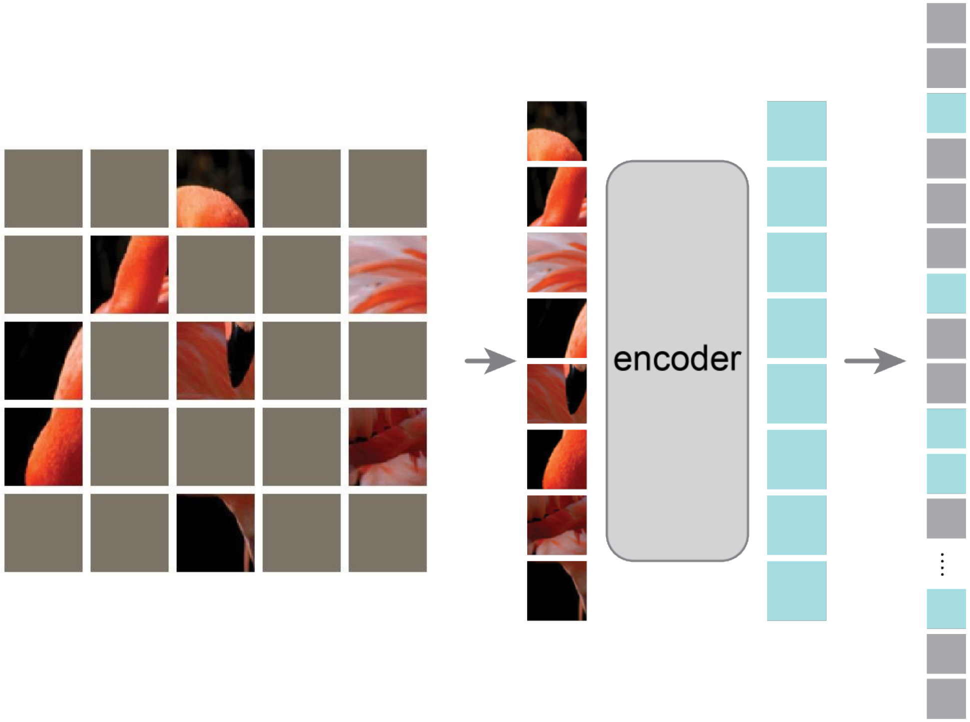


encode visible patches
w/ Transformer

He, et al., "Masked Autoencoders Are Scalable Vision Learners", CVPR 2022

Slide from MIT - 6.8300/1 Advances in Computer Vision

Masked Autoencoder (MAE)

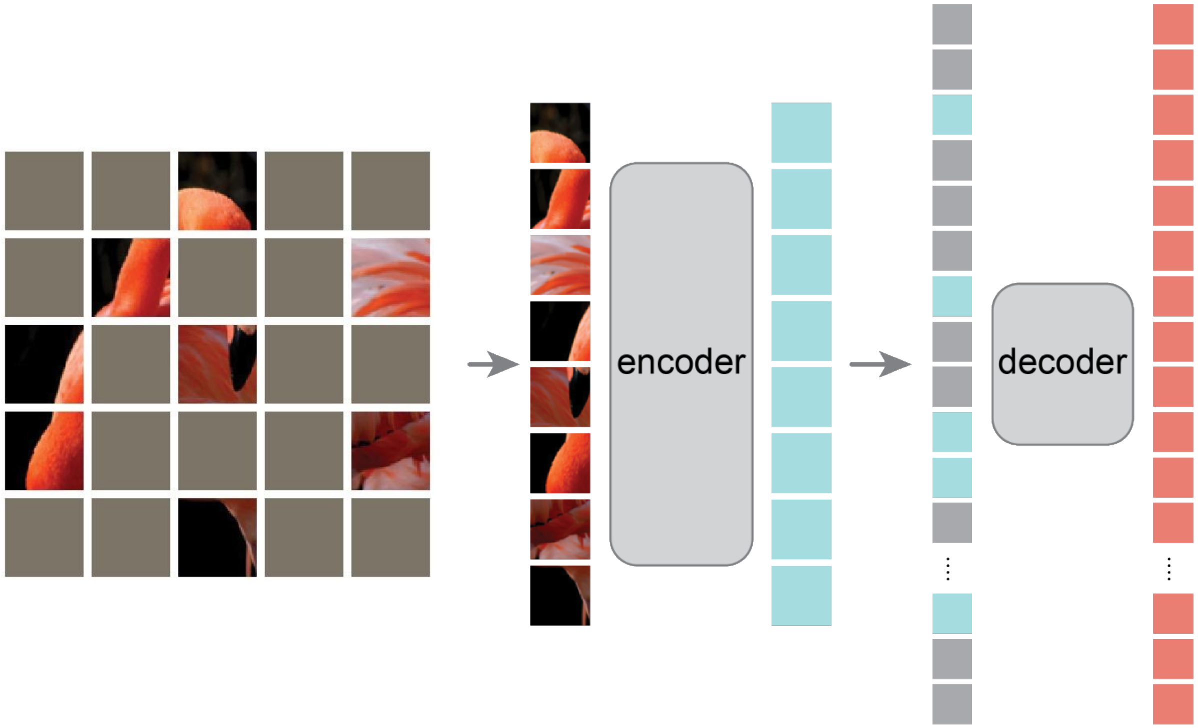


expand with “mask tokens”
(specify where to predict)

He, et al., “Masked Autoencoders Are Scalable Vision Learners”, CVPR 2022

Slide from MIT - 6.8300/1 Advances in Computer Vision

Masked Autoencoder (MAE)

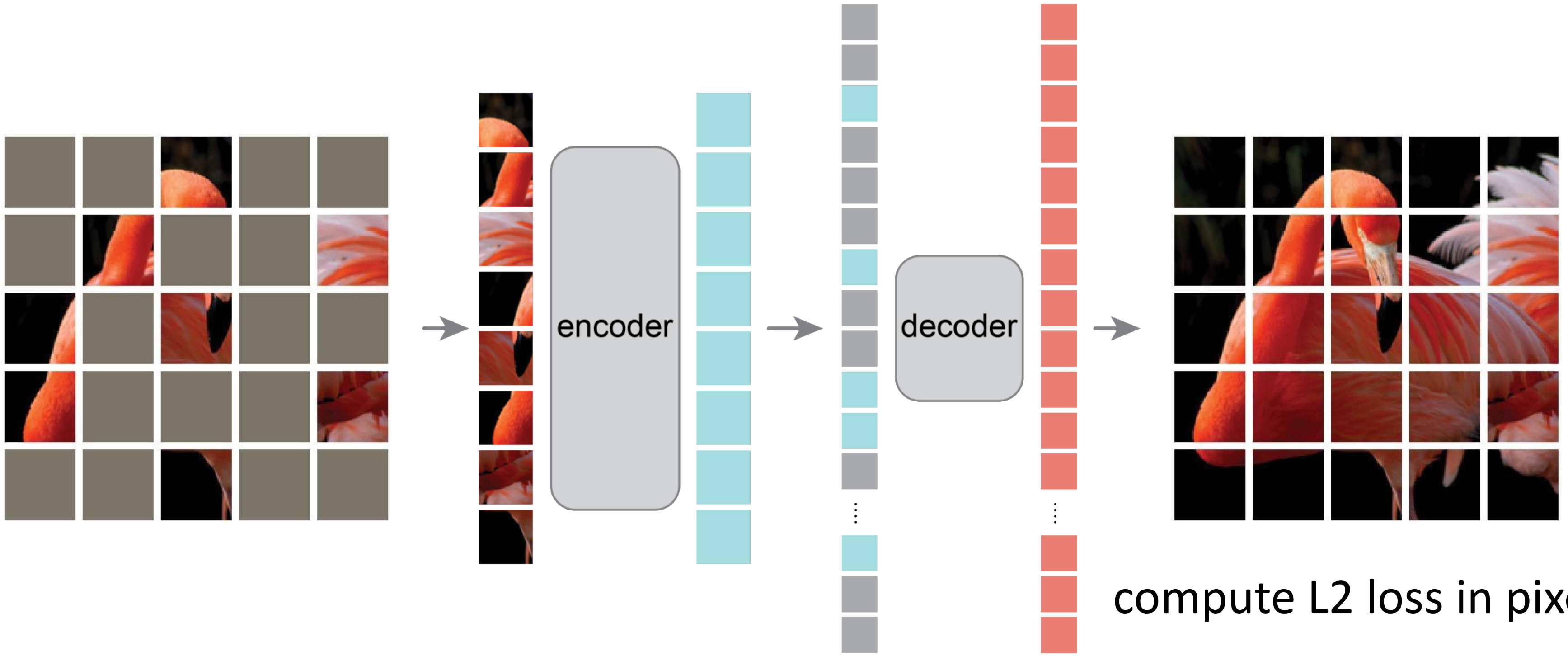


predict the unknown

He, et al., "Masked Autoencoders Are Scalable Vision Learners", CVPR 2022

Slide from MIT - 6.8300/1 Advances in Computer Vision

Masked Autoencoder (MAE)

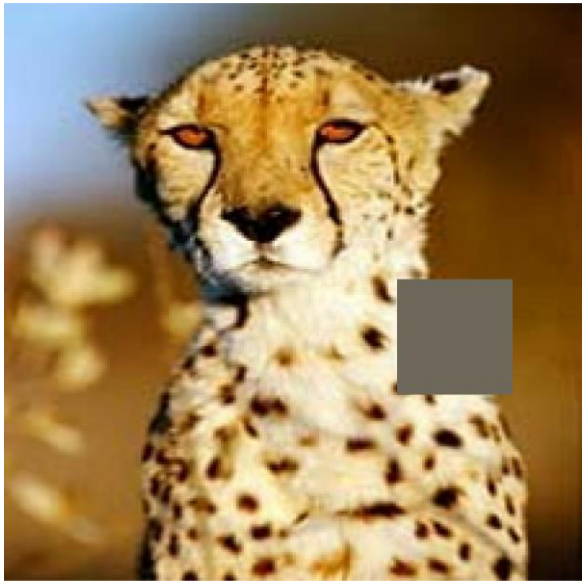


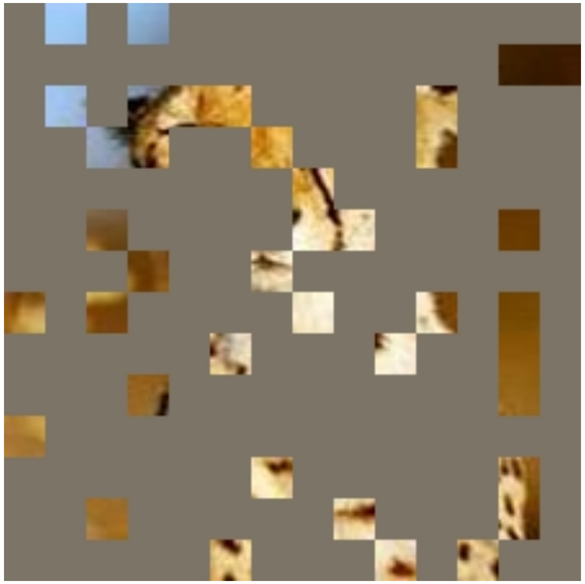
compute L2 loss in pixels

He, et al., “Masked Autoencoders Are Scalable Vision Learners”, CVPR 2022

Slide from MIT - 6.8300/1 Advances in Computer Vision

How to learn good representations by predicting?



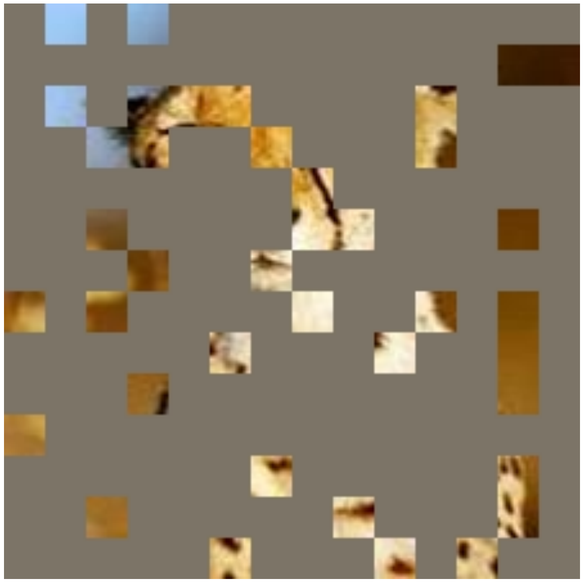


- predicting a <u>small portion</u> may not require high-level understanding
- predicting a <u>large portion</u> of unknown patches encourages to learn semantic features

He, et al., "Masked Autoencoders Are Scalable Vision Learners", CVPR 2022

Slide from MIT - 6.8300/1 Advances in Computer Vision

How to learn good representations by predicting?



input

MAE prediction

original

He, et al., “Masked Autoencoders Are Scalable Vision Learners”, CVPR 2022

Slide from MIT - 6.8300/1 Advances in Computer Vision

Representation Learning

Reinforcement Learning (Cherry)
Predicting a scalar reward given once in a while
A few bits for some samples

Supervised Learning (Chocolate Coat)
Predicting category or vector of scalars per input as provided by human labels.
10-10k bits per sample

Unsupervised / Self-Supervised Learning (Cake)
Predicting parts of observed input or predicting future observations or events
Millions of bits per sample

Visualisation Idea by Yann LeCun
Photo by Kristina Paukshtite from Pexels

Summary of what you learned today

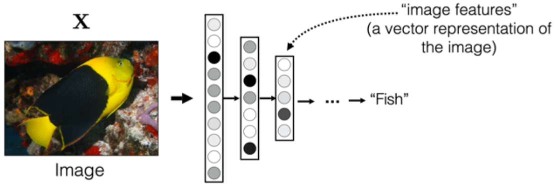
- **State:** Quantity that describes the most important aspect of a dynamical system at time t
- **Representation:** data format of input or output including a low-dimensional representation of sensor data

Summary of what you learned today

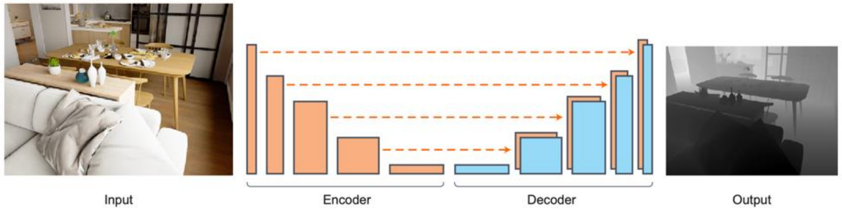
- Learned versus interpretable representations
- Visualize learned representations
- How to learn representations?
  - Supervised
  - Unsupervised
  - Self-supervised

Next Lectures

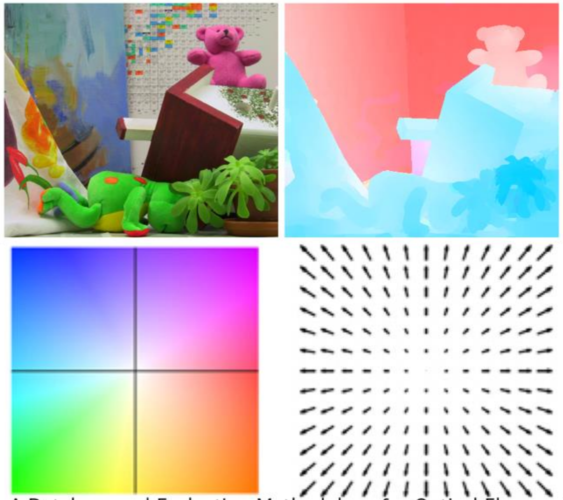
Representations & Representation Learning



Monocular Depth Estimation, Feature Tracking

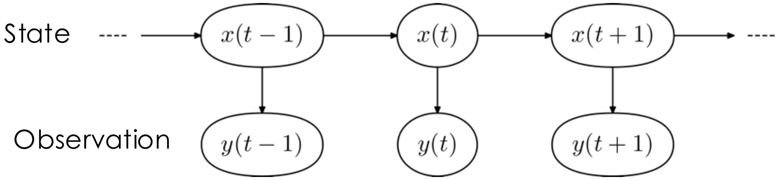


Optical & Scene Flow

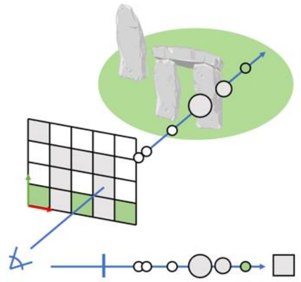


A Database and Evaluation Methodology for Optical Flow.
Baker et al. IJCV. 2011

Optimal Estimation



Neural Radiance Fields



CS231

Computer Vision:

From 3D Reconstruction to Recognition

Next lectures:

Midterm (Monday)

Monocular Depth Estimation & Feature Tracking (Wednesday)