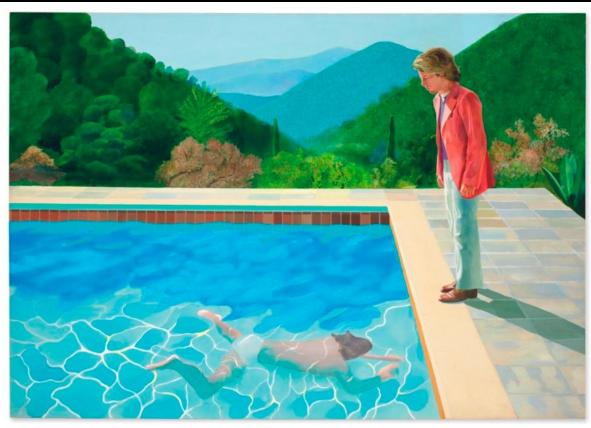
CS 444: Deep Learning for Computer Vision



D. Hockney, Pool with two figures, 1972

https://slazebni.cs.illinois.edu/spring24/

Overview

- Logistics
- Motivation: The statistical learning viewpoint
- A taxonomy of learning problems
- Topics to be covered in class

How can we build an agent to...

Play chess?



Recognize object categories?



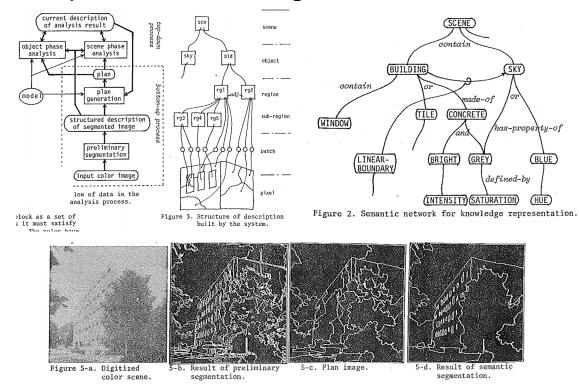
Translate between languages?



Fly a drone?

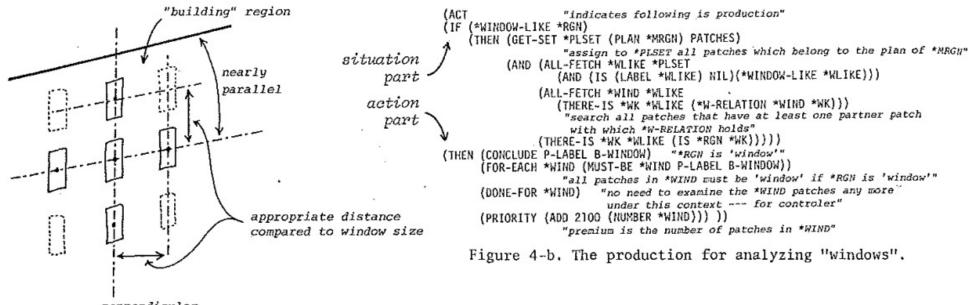


 Good old-fashioned AI (GOFAI) answer: Program expertise into the agent

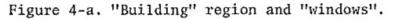


Y. Ohta, T. Kanade and T. Sakai. <u>An Analysis System for Scenes Containing objects with Substructures.</u> Proc. of the Fourth International Joint Conference on Pattern Recognition, pp. 752-754, 1978

• Good old-fashioned AI (GOFAI) answer: Program expertise into the agent



perpendicular



Good old-fashioned AI (GOFAI) answer: • Program expertise into the agent

Appendix-B Complete Listing of the Model

eSCENE knowledge-black-of-scene (DBJECTS (ASKY #TREE #BUILDING #ROAD #UNKNDUN) SUB-OBJECTS (AB-UINDON #CAR #C-SHADDU) KEY-PATCH-IS (IGREATERP (AREA #PCH) 3880 (#PCH)]

IF-PLAN-IS-MUDIFIED (IF-DDNE (

(F-Ru-E)-REGUINE OF-SOL (Vie-Ru-E)-REGUINE OF-SOL (Vie-Ru-E)-REGUINE P-SELECT (TO-00 (

rule-for-initial-start [GACT (AND (PROBABLY BUILDING APCH) (NDTFOUND BUILDING))

nali-for-tra-socialian Tucht (No Loader, 44/25), 44/25), 47/25 (1990) - 10,497 (1990) -

rula-for-trae-garbage [LACT PRODULLY TREE aPCH0 (THON COXRLUGE P-LAREL TREE) (SCORE-15 (ASK-VALUE TREE aPCH1)))(aPCH01))

P-LABEL (IF-DONE (

if-done-rule-to-be-activated-shan-kaypatch-is-labsled [IACT (NOT (IS (OF PLAN aPCH) NTL)) (THEN LEXEDUTE PLAN-EVALUATION))(ePCH)]))

ASKY knowledge-block-of-sky

0PR0PENTY-RLLES ((IGN NOT (42,04EP +0200)11,8, 8, 8,0)(4P200)1 (IGN (491NH06 +02001,1,8, 8, 22)(4P200)1 (IGN NOT (40,04E +0200 +0200)1,8, 8, 22)(4P200)1 (IGN NOT (401CH14A, 42001,14, 8, 8, 72)(4P200)1 (ISTR (TODENIDA 4P20149-5100)(8,7, 8, 22)(4P200)1)

(THEN ICONCLICE P-LARL ISOTORIO) ICONCLICE P-LARL ISOTORIDA:-ULLE aPOINTD:-Delta ICONCLICE P-LARL ISOT ICONCLIC

(THEN (CONCLUDE P-LABEL TREE) (SCORE-15 CADD 4.8 (CONFIDENCE-VALUE &PCH)))) (&PCH))

vije tura-algeneta al 10 - 000 (400)
100-01 (400)
100-01 (400)
100-01 (400)
100-01 (400)
100-01 (400)
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
100-01
1

(HeN (CORLINE CORLINE OF), (CORLINE OF), (CO

P-SELECT (10.40 (10.47 Biol.-82 Sirv. 4700 (Then ISODEL-15 LARD 2.8 (LARC-HULE Sirv. 47601)11) (47601) (LART LARD (154-LUK 2014-47800) LARD(147601) (THEN ISODEL-15 2.81) (47604) 476001 (LARD LARD(274-6016 LISODEL-15 2.857) (47601))

1E-DONE ([FACT aTe (THEN (DONDLUDE P-LABEL SKY) (DONDLUDE R-HERGE (MMSTER aPCH)))) (aPCH)])) APRIERI-VALUE-1S 0.1)

wTREE knowledge-block-of-tree (MADE-DF (eLEAVES)

PROPERTY-RULES ((IGEN: GANLOLE ARGN) (8.6 . 8.31) (ARGN)] LISTR: GAHEAVY-TEXTURE ARGN) (8.8 . 8.21) (ARGN)])

P-SELECT (

10-00 ((UGC 1947-86 1962 #PO-0 (INEN (60096-15 MAD 2.8 UASK-VALUE 1962 #PO(1)))) (#PO(1) [UACT MAD (15-PLAN #PO+ #POR0 001 (#SHIMMC #PO(1)) (INEN (50096-15 3.8))) (#POH #POR0)

IF-DONE ([(ACT aT# (THEN (CONCLIDE P-LABEL TREE) (CONCLIDE R-HERCE (MASTER aPDH011) (aPCH)1)) AFRIORI-VALUE-IS 8.2)

eBUILDING knowledge-block-of-building UNADE-OF (OR #CONCRETE #TILE #BRICK) SUB-OBJECTS (#B-HINDOW)

PROPERTY-RULES (LIERN (-MIDDLE -HEN) (8,6 , 8,31) (ARD/O] LIERT (-MWAYEDLE -HEN) (8,6 , 8,21) (ARD/O] LIERT (-MWAYEDLE -HEN) (8, 4 , 6,21) (-HEN) LIERN (-HEN) (8, 9 , 6,51) (-HEN) (-LIERN (-HEN) (8, 9 , 6,51) (-HEN) (-

1000 (1) 1000 (1000 (200) (2000 (200) (200) (2000 (200) (2000 (2000 (200) (2000 (2000 (2000 (2000 (2000 (200) (2000 (2000 (200) (2000 (200) (2000 (200) (2000 (200) (2000 (200) (2000 (200) (2000 (200) (2000 (200) (2000 (200) (2000 (200) (200) (2000 (200) (200

volume-resistantian
(uct) or well served and served served and serve

TWACT CARD 115-PCA APROX ISAPE-ZONE APCH APROXI (THEN ICONCLUE P-LABLE BUILDING) (CONCLUE R-AMDICE APROX (SECRE-15 2.481)) (APCH APROX)

0-NERSE (IF-ODNE (LIACT wTw (DESCRIBE-BUILDING (NEGION #PCHI))(#PCHI1 1)

O-OREATE (IF-DOME (LGAET #T# (THEN DEXTRUCT-BUILDING-GRAPE (HEDIIN #FCH)) (DESCRIBE-BUILDING (HEDIIN) (#FCH)) (DESCRIBE-BUILDING (HEDIIN) (#FCH)) APRIORI-VALUE-IS 8.2)

AB167 knowledge-block-of-road

CHADE-OF IOR WASPHALT &CONCRETE) SUB-OBLICITS WASPHALT &CONCRETE)

PROPERTY-FULES (IGEN (ALDEER #GON (8.8 . 8.4))(4000)) IGEN (ANDELDONTALLYLING 4000)(8.7 . 8.2))(4000) (619) (TOUCHUS 460) LEMER-SIDE (8.5 . 8.2))(4000)

- Good old-fashioned AI (GOFAI) answer: Program expertise into the agent
 - Never worked (in general)

- Good old-fashioned AI (GOFAI) answer: Program expertise into the agent
 - Never worked (in general)
 - Though not without exceptions...

ANNALS OF TECHNOLOGY

THE PASTRY A.I. THAT LEARNED TO FIGHT CANCER

In Japan, a system designed to distinguish croissants from bear claws has turned out to be capable of a whole lot more.

By James Somers March 18, 2021

https://www.newyorker.com/tech/annals-of-technology/the-pastryai-that-learned-to-fight-cancer



- Good old-fashioned AI (GOFAI) answer: Program expertise into the agent
 - Never worked (in general)
 - Though not without exceptions...

ANNALS OF TECHNOLOGY

THE PASTRY A.I. THAT LEARNED TO FIGHT CANCER

In Japan, a system designed to distinguish croissants from bear claws has turned out to be capable of a whole lot more.

By James Somers March 18, 2021

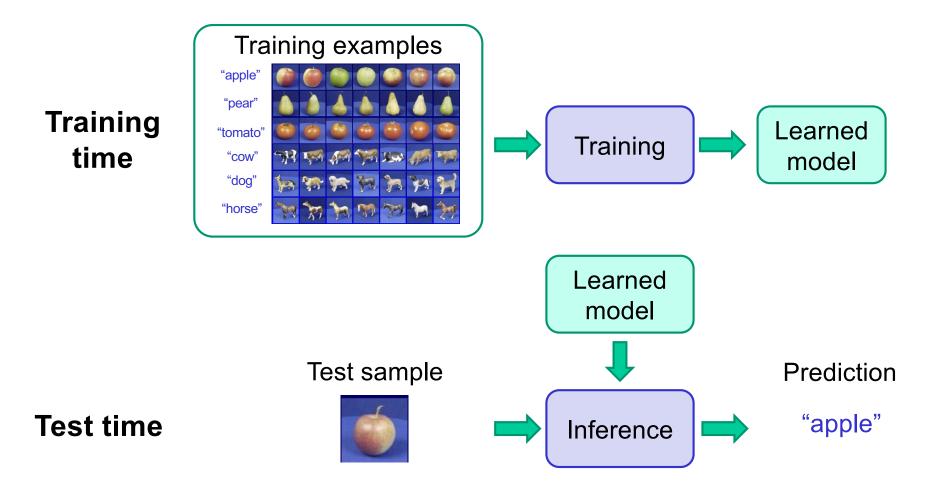
not well fed.) But this was all under carefully controlled conditions. In a real bakery, the lighting changes constantly, and BRAIN's software had to work no matter the season or the time of day. Items would often be placed on the device haphazardly: two pastries that touched looked like one big pastry. A subsystem was developed to handle this scenario. Another subsystem, called "Magnet," was made to address the opposite problem of a pastry that had been accidentally ripped apart.

be used elsewhere. Today, solving the pastry problem without deep learning would seem impossible; it's a wonder that, in 2007, when neural networks weren't a viable option, Kambe even took it on. The system that he and his team managed to build over the following fifteen years must surely be one of the more sophisticated achievements in "classical" computer vision—a fact obscured, perhaps, by its origin in baked goods.

https://www.newyorker.com/tech/annals-of-technology/the-pastryai-that-learned-to-fight-cancer

- Good old-fashioned AI (GOFAI) answer: Program expertise into the agent
- Modern answer: Program into the agent the ability to improve performance based on experience
 - **Performance** needs to be quantified using some score or metric (loss, reward, etc.)
 - Experience comes from training data or demonstrations
 - Improvement results from the *learning algorithm*
 - Leap of faith: agent that can achieve good performance on training data will *generalize* to never-before-seen inputs

The basic statistical learning framework



Overview

- Logistics
- Motivation: The statistical learning viewpoint
- A taxonomy of learning problems

Taxonomy of learning problems

- Type of output
 - Classification
 - Regression
 - Structured prediction
 - Dense prediction
 - Multi-modal prediction

Type of supervision

- Fully supervised
- Unsupervised
- Self-supervised or predictive learning

Training regime

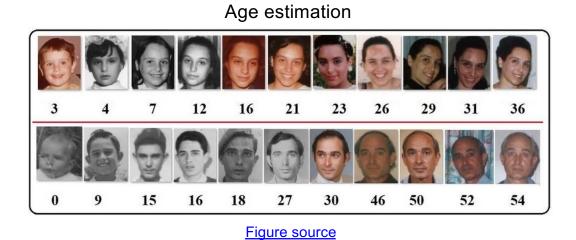
- Batch offline learning
- Online/continual learning
- Active learning
- Reinforcement learning

Type of output: Classification

ImageNet Large-Scale Visual Recognition Challenge (ILSVRC)

mite	container ship	motor scooter	leopard
mite	container ship		leopard
black widow	lifeboat	go-kart	jaguar
cockroach	amphibian	moped	cheetah
tick	fireboat	bumper car	snow leopard
starfish	drilling platform	golfcart	Egyptian cat
grille	mushroom	cherry	Madagascar cat
convertible	agaric	dalmatian	squirrel monkey
grille	mushroom	grape	spider monkey
pickup	jelly fungus	elderberry	titi
beach wagon	gill fungus	ffordshire bullterrier	indri
fire engine	dead-man's-fingers	currant	howler monkey

Type of output: Regression



Date prediction



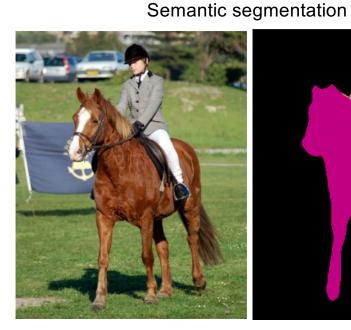
Vittayakorn et al. (2017)

Location prediction



Type of output: Dense prediction

6



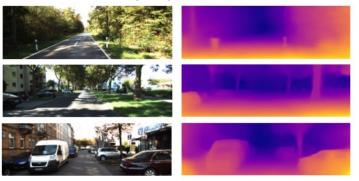
Long et al. (2016)

Image colorization



Zhang et al. (2016)

Depth prediction



Wang et al. (2017)

Type of output: Structured prediction

Object detection

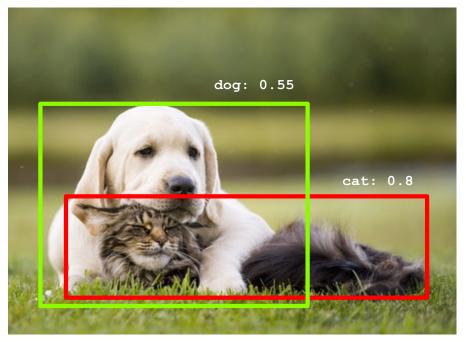


Image description



"man in black shirt is playing guitar."



"girl in pink dress is jumping in air."



"construction worker in orange

safety vest is working on road."

"black and white dog jumps over bar."



"two young girls are playing with lego toy."



"young girl in pink shirt is swinging on swing."

Karpathy & Fei-Fei (2015)

Dense + structured prediction

Object detection + instance segmentation

Keypoint detection



K. He, G. Gkioxari, P. Dollar, and R. Girshick, Mask R-CNN, ICCV 2017

Overview

- Logistics
- Motivation: The statistical learning viewpoint
- A taxonomy of learning problems
- Topics to be covered in class

Taxonomy of learning problems

- Type of output
 - Classification
 - Regression
 - Structured prediction
 - Dense prediction
 - Multi-modal prediction

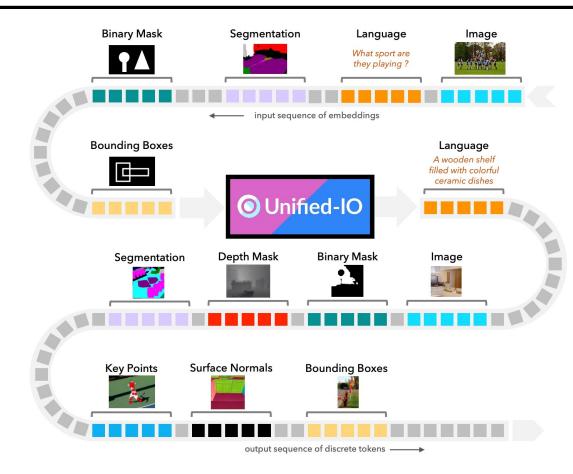
Type of supervision

- Fully supervised
- Unsupervised
- Self-supervised or predictive learning

Training regime

- Batch offline learning
- Online/continual learning
- Active learning
- Reinforcement learning

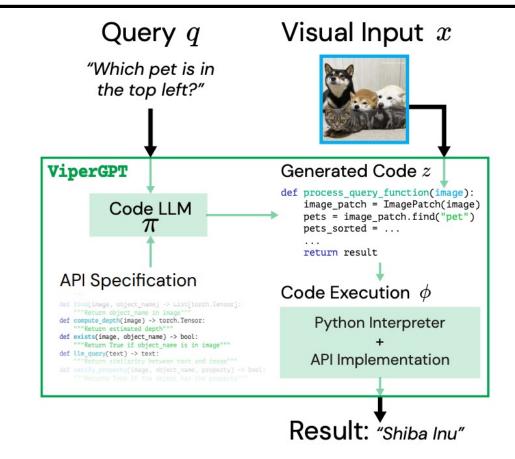
Multi-modal prediction



Tasks Image Classification **Object Detection** Semantic Segmentation **Depth Estimation** Surface Normal Estimation Segment-based Image Generation Image Inpainting **Pose Estimation Relationship Detection** Image Captioning Visual QA **Referring Expressions** Situation Recognition Text-based Image Generation Visual Commonsense Classification in context **Region Captioning GLUE Benchmark tasks** Reading comprehension Natural Language Inference

J. Lu et al. Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks. ICLR 2023

Multi-modal prediction



S. Menon et al. ViperGPT: Visual Inference via Python Execution for Reasoning. ICCV 2023

Multi-modal prediction

Query: How many muffins can each kid have for it to be fair? Execution kid_patches = image patch.find("kid") **Generated** Code muffin_patches = image_patch.find("muffin") def execute command(image): image_patch = ImagePatch(image) muffin patches = image patch.find("muffin") Len(muffin_patches)=8 kid_patches = image_patch.find("kid") len(kid_patches)=2 return str(len(muffin_patches) // len(kid_patches)) ▶8//2 = 4 Result:4 Query: Drink with zero alcohol drink_patches= drink_name = 'tullamore dew' alcoholic = 'yes' def execute_command(image): image_patch = ImagePatch(image) Result: drink name = 'bacardi' drink patches = image patch.find("drink") alcoholic = 'yes' for drink_patch in drink_patches: drink_name = drink_patch.simple_query("What is this?") alcoholic = llm_query(f"Does the {drink_name} have alcohol?") bdrink name = 'gin' if alcoholic == "no": return drink_patch >alcoholic = 'yes' return None drink_name = 'dr pepper' alcoholic = 'no' Query: What would the founder of the brand of the car on the left say to the founder of the brand of the car on the right? car patches = def execute_command(image): image_patch = ImagePatch(image) image_patch.find("car") car_patches = image_patch.find("car") > left_car_brand='lamborghini' car_patches.sort(key=lambda car: car.horizontal_center) right_car_brand='ferrari' left_car = car_patches[0]

right car = car patches[-1] left_car_brand = left_car.simple_query("What is the brand of this car?") right_car_brand = right_car.simple_query("What is the brand of this car?") left_car_founder = llm_query(f"Who is the founder of {left_car_brand}?") right_car_founder = llm_query(f"Who is the founder of {right_car_brand}?") return llm_query(f"What would {left_car_founder} say to {right_car_founder}?")



>left_car_founder='Ferruccio Lamborghini' ▶ right_car_founder='Enzo Ferrari'

Result: "Ferruccio Lamborghini might say, 'It's been an honor to be a rival of yours for so many years, Enzo. May our cars continue to push each other to be better and faster!"

S. Menon et al. ViperGPT: Visual Inference via Python Execution for Reasoning. ICCV 2023

Taxonomy of learning problems

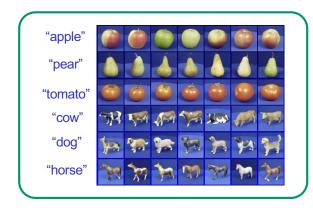
- Type of output
 - Classification
 - Regression
 - Structured prediction
 - Dense prediction
 - Multi-modal prediction

• Type of supervision

- Fully supervised
- Unsupervised
- Self-supervised or predictive learning

Type of supervision

• Traditional (over-simplified) dichotomy





Supervised learning:

clean, complete training labels for the task of interest

Unsupervised learning: no labels

- Given: large collection of unlabeled data
- Goal: ???



Image source

Clustering ullet

Discover groups of "similar" data points •

cute rabbit bunny animal cheerleader football girls



abandoned decay old

urban rust industrial

factory jail rusty

basketball girls dance university sports college



music concert rock live city urban manhattan new home design office house building downtown night festival band scientists dance drum



underwater fish diving scuba coral sea ocean reef dive







nature macro flower closeup green insect bravo red yellow





snow winter ice cold nature trees mountains white mountain



Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A Multi-View Embedding Space for Modeling Internet Images, Tags, and Their Semantics. IJCV 2014

baby adorable pet

interior kitchen fashion

work room

bird birds nature wildlife

animal booby eagle

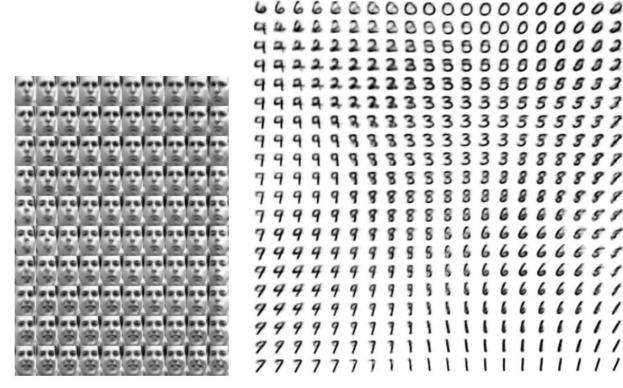
hawk flight



portrait face self girl woman eyes smile

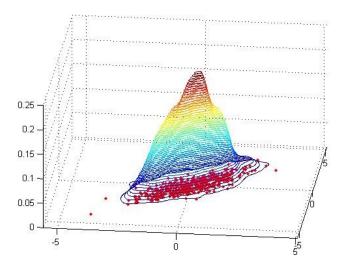
Dimensionality reduction, manifold learning

• Discover a lower-dimensional surface on which the data lives

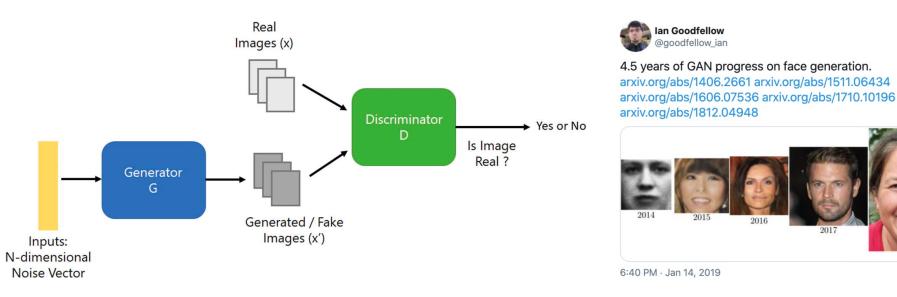


D. Kingma and M. Welling, Auto-Encoding Variational Bayes, ICLR 2014

- Learning the data distribution
 - **Density estimation**: Find a function that approximates the probability density of the data (i.e., value of the function is high for "typical" points and low for "atypical" points)
 - An extremely hard problem for high-dimensional data...



- Learning the data distribution
 - Learning to sample: Produce samples from a data distribution that mimics the training set



Generative adversarial networks (GANs)

Image source

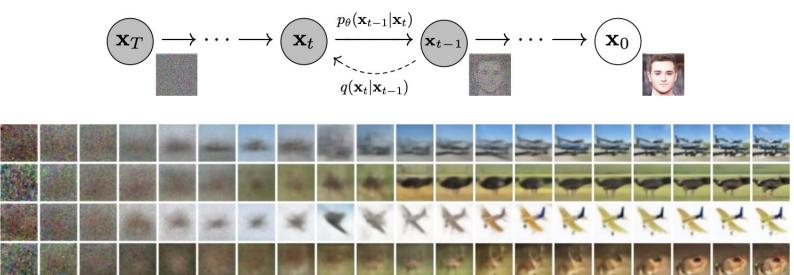
y

2018

(i)

- Learning the data distribution
 - Learning to sample: Produce samples from a data distribution that mimics the training set

Denoising diffusion probabilistic models (DDPMs)



- Learning the data distribution ۲
 - Learning to sample: Produce samples from a data distribution that mimics the training set

Denoising diffusion probabilistic models (DDPMs)



Salvador Dalí with a robotic half face

a close up of a handpalm with leaves growing from







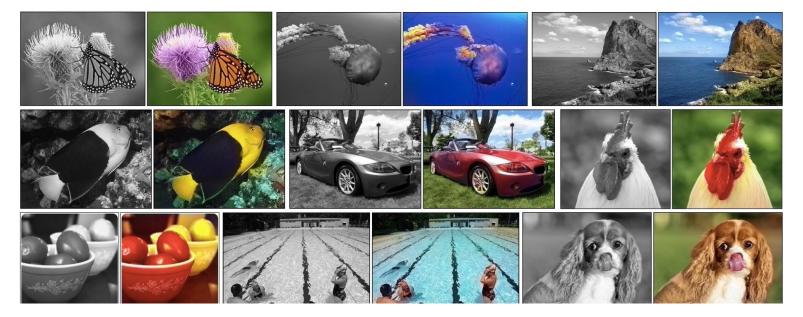
Source: DALL-E 2

n espresso machine that makes coffee from human souls, artstation

panda mad scientist mixing sparkling chemicals, artstation

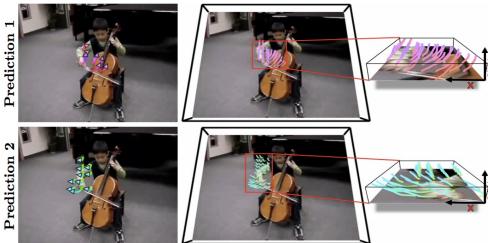
a corgi's head depicted as an explosion of a nebula

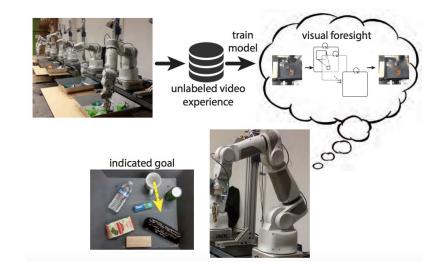
- Use part of the data to predict other parts of the data
 - Example: Image colorization



R. Zhang et al., Colorful Image Colorization, ECCV 2016

- Use part of the data to predict other parts of the data
 - Example: Future prediction





J. Walker et al. An Uncertain Future: Forecasting from Static Images Using Variational Autoencoders. ECCV 2016 C. Finn and S. Levine. Deep Visual Foresight for Planning Robot Motion. ICRA 2017. YouTube video

- Use part of the data to predict other parts of the data
 - Example: Grasp prediction



L. Pinto and A. Gupta. Supersizing self-supervision: Learning to grasp from 50K tries and 700 robot hours. ICRA 2016

- Use part of the data to predict other parts of the data
 - Example: Next/masked word prediction

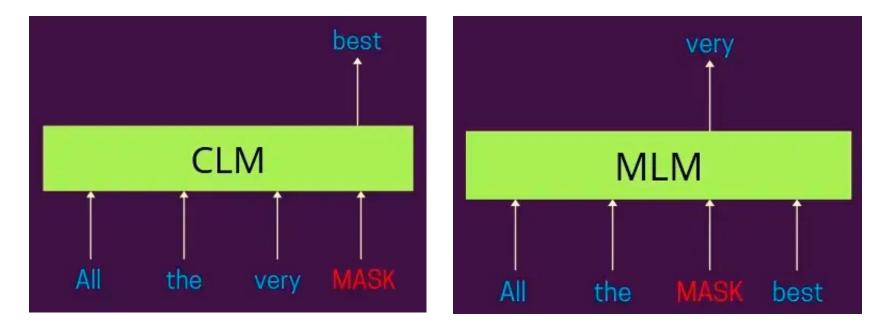
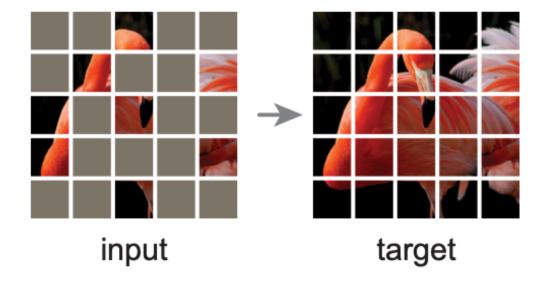


Figure source

Self-supervised or predictive learning

- Use part of the data to predict other parts of the data
 - Example: Masked patch prediction



K. He et al. Masked autoencoders are scalable vision learners. CVPR 2022

Taxonomy of learning problems

- Type of output
 - Classification
 - Regression
 - Structured prediction
 - Dense prediction
 - Multi-modal prediction

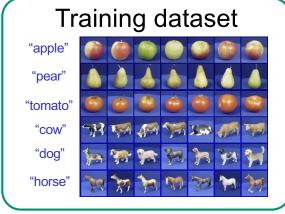
Type of supervision

- Fully supervised
- Unsupervised
- Self-supervised or predictive learning

Training regime

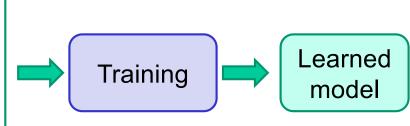
- Batch offline learning
- Online/continual learning
- Active learning
- Reinforcement learning

Offline learning

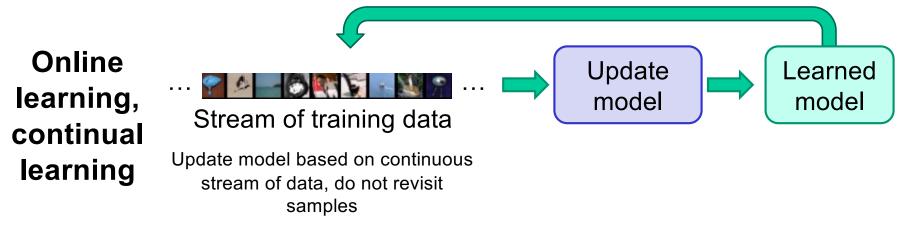


Process the entire training set (typically in multiple passes)

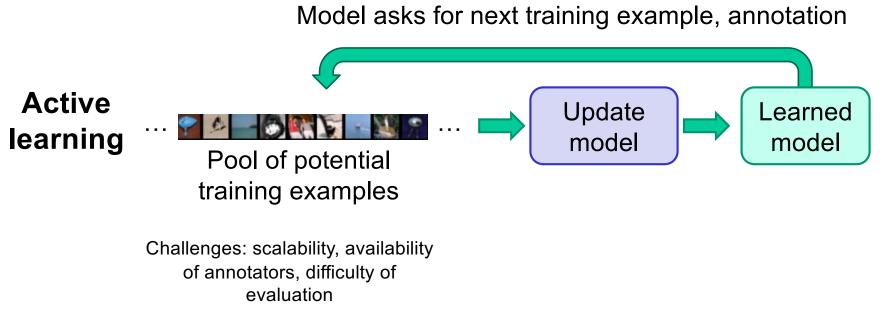
Challenges: static dataset, high storage, memory requirements



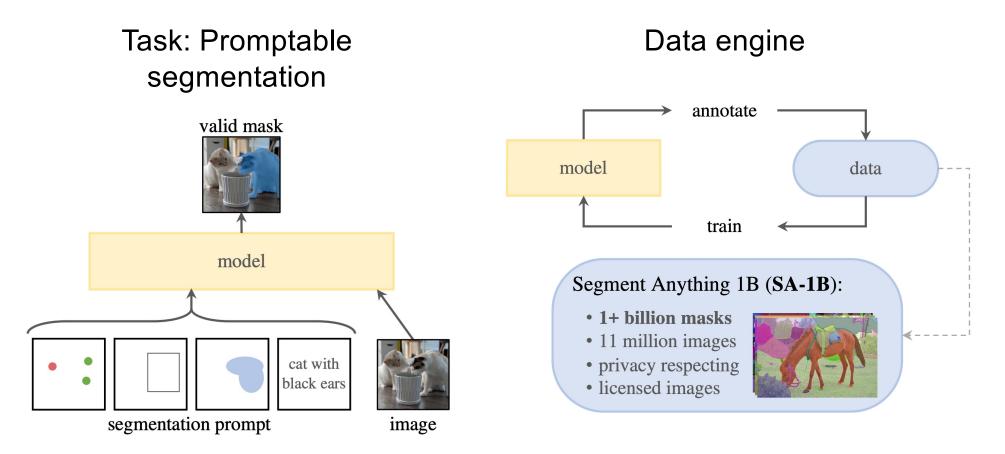
Pre-training dataset Transfer **Pre-trained** Training learning model New dataset ellow worble Fine-**Fine-tuned** tuning model tree swallow black tern scarlet tanage green violetear hooded merganser florida jay



Challenges: changes in the data distribution, *catastrophic forgetting*



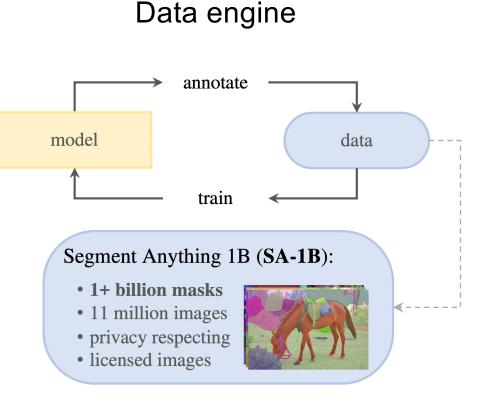
Today's trend: Data engines



A. Kirillov et al. Segment anything. ICCV 2023

Today's trend: Data engines

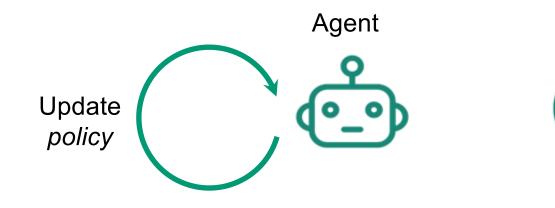
- Data engine steps:
 - 1. Pre-training using public datasets
 - 2. Assisted manual stage: interactive segmentation with SAM assisting annotators
 - 3. Semi-automatic stage: SAM generates confident masks, annotators add masks to improve diversity
 - 4. Fully automatic stage: SAM generates ~100 masks per image starting with a grid of points



A. Kirillov et al. Segment anything. ICCV 2023

Reinforcement learning

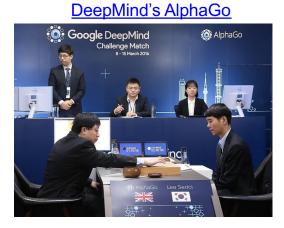
 Learning for an agent that can affect the world through actions



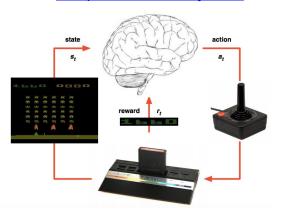


World

Reinforcement learning: Examples



DeepMind's Atari system



Sensorimotor learning



Sensorimotor learning

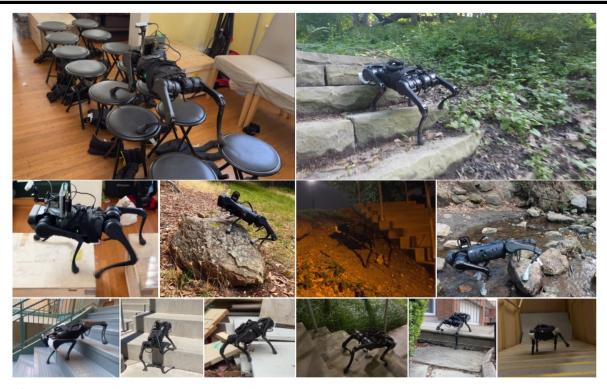


Figure 1: Our robot can traverse a variety of challenging terrain in indoor and outdoor environments, urban and natural settings during day and night using a single front-facing depth camera. The robot can traverse curbs, stairs and moderately rocky terrain. Despite being much smaller than other commonly used legged robots, it is able to climb stairs and curbs of a similar height. Videos at https://vision-locomotion.github.io

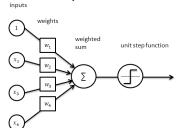
A. Agarwal, A. Kumar, J. Malik, and D. Pathak. Legged Locomotion in Challenging Terrains using Egocentric Vision. CoRL 2022

Overview

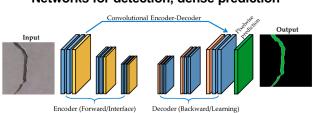
- Logistics
- Motivation: The statistical learning viewpoint
- A taxonomy of learning problems
- Topics to be covered in class

Topics to be covered in class

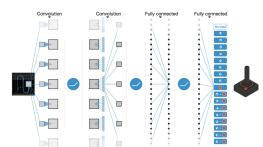
ML basics, linear classifiers



Networks for detection, dense prediction



Deep reinforcement learning



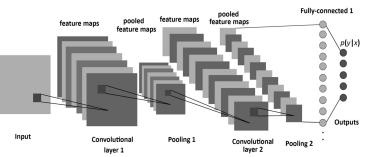
Multilayer neural networks, backpropagation

Outputs Input Layer Hidden Layers Output Layer

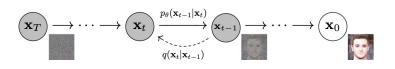
Self-supervised learning



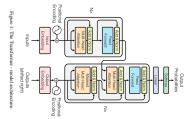
Convolutional networks for classification



Generative models: GANs, image-to-image translation, diffusion models



Transformers, large language models, transformers for vision





Models for sequence data

