

Deep Learning

Lecture 1 - A history of deep learning

Are you in the right place?

Location: CSE2 G01

Lectures: Wednesdays @ 6:30pm-9:20pm

Recitations: Fridays

Canvas: <https://canvas.uw.edu/courses/1861735>

Gradescope: <https://www.gradescope.com/courses/1205206>
(Entry Code: YBR7Y6)

Website: <https://courses.cs.washington.edu/courses/csed502/26wi/>

EdStem: <https://edstem.org/us/courses/90173>

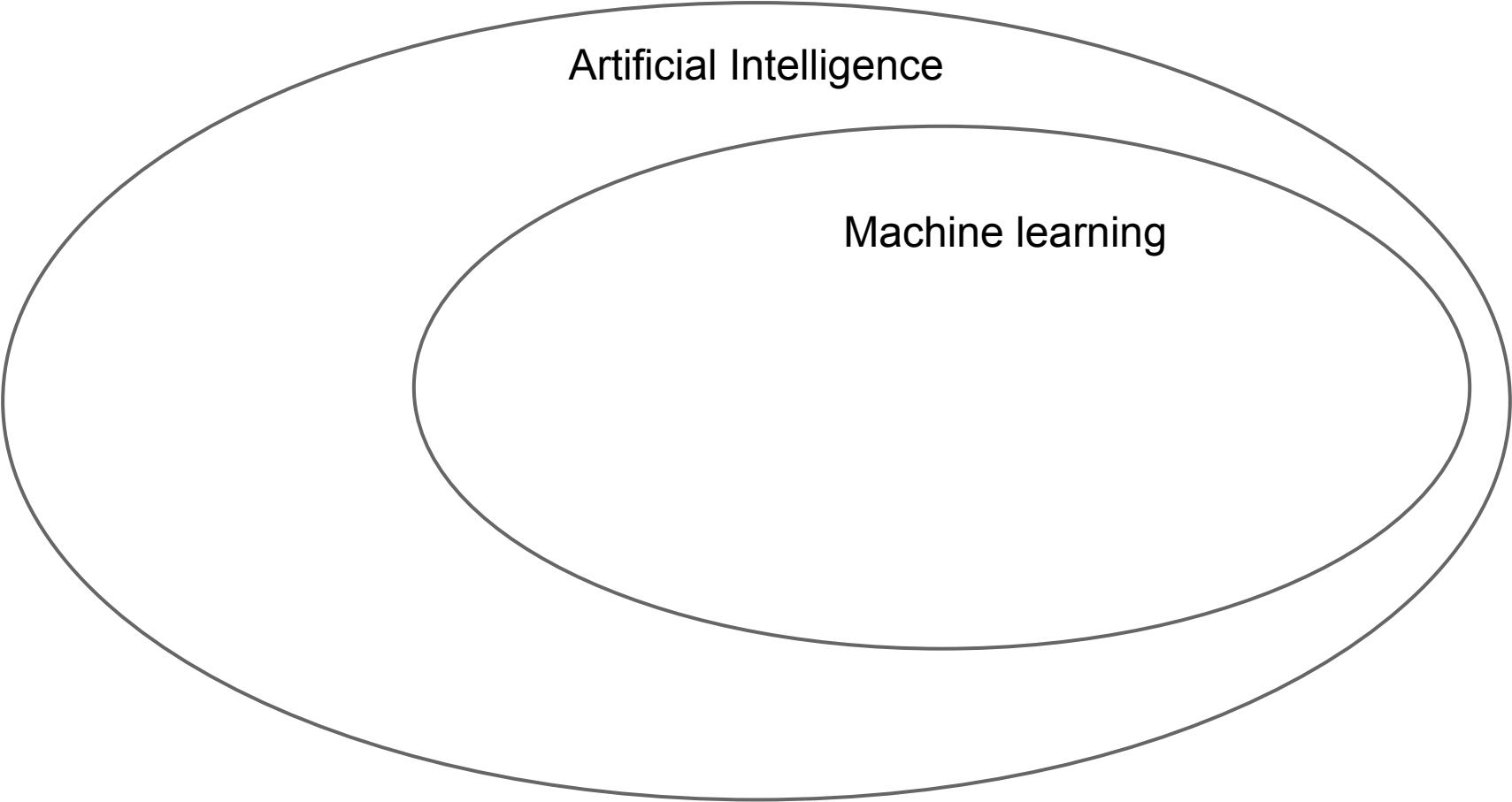
What is Deep Learning?

Building artificial systems that learn
from **data and experience**

What is Deep Learning?

Hierarchical systems with many “layers” of processing, which can learn from data and experience

Artificial Intelligence



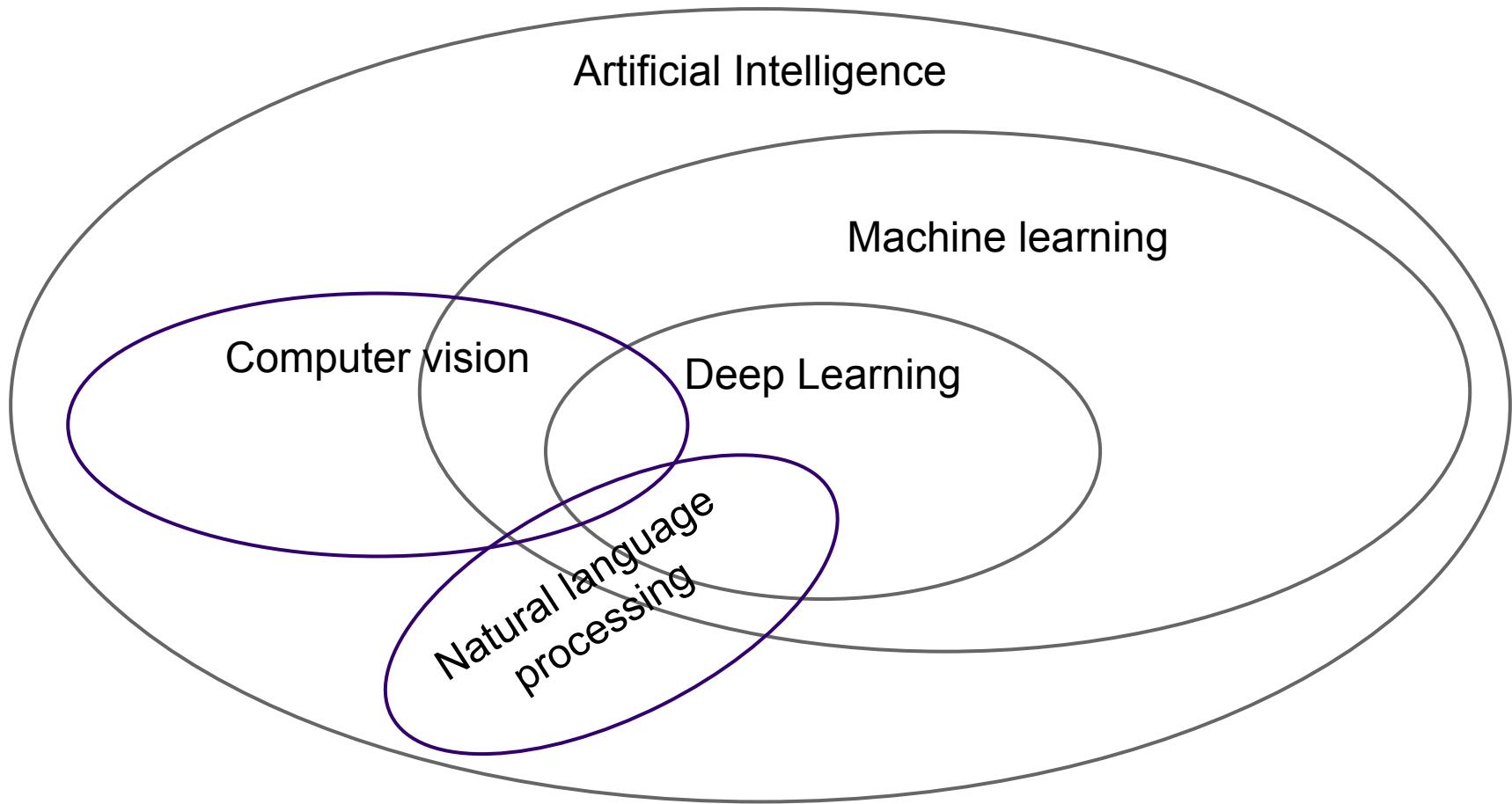
Artificial Intelligence

Machine learning

Artificial Intelligence

Machine learning

Deep Learning



Artificial Intelligence

Robotics

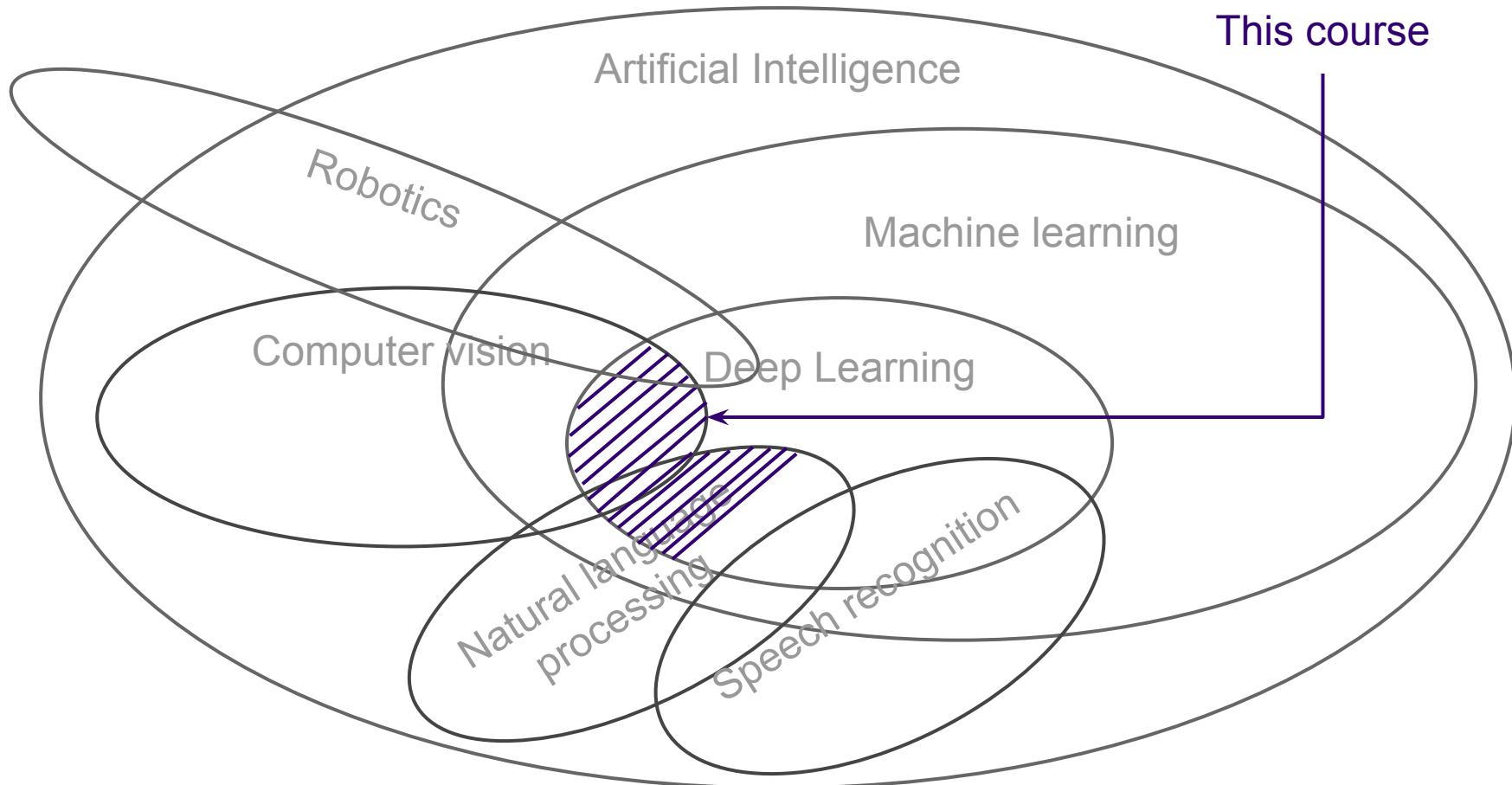
Machine learning

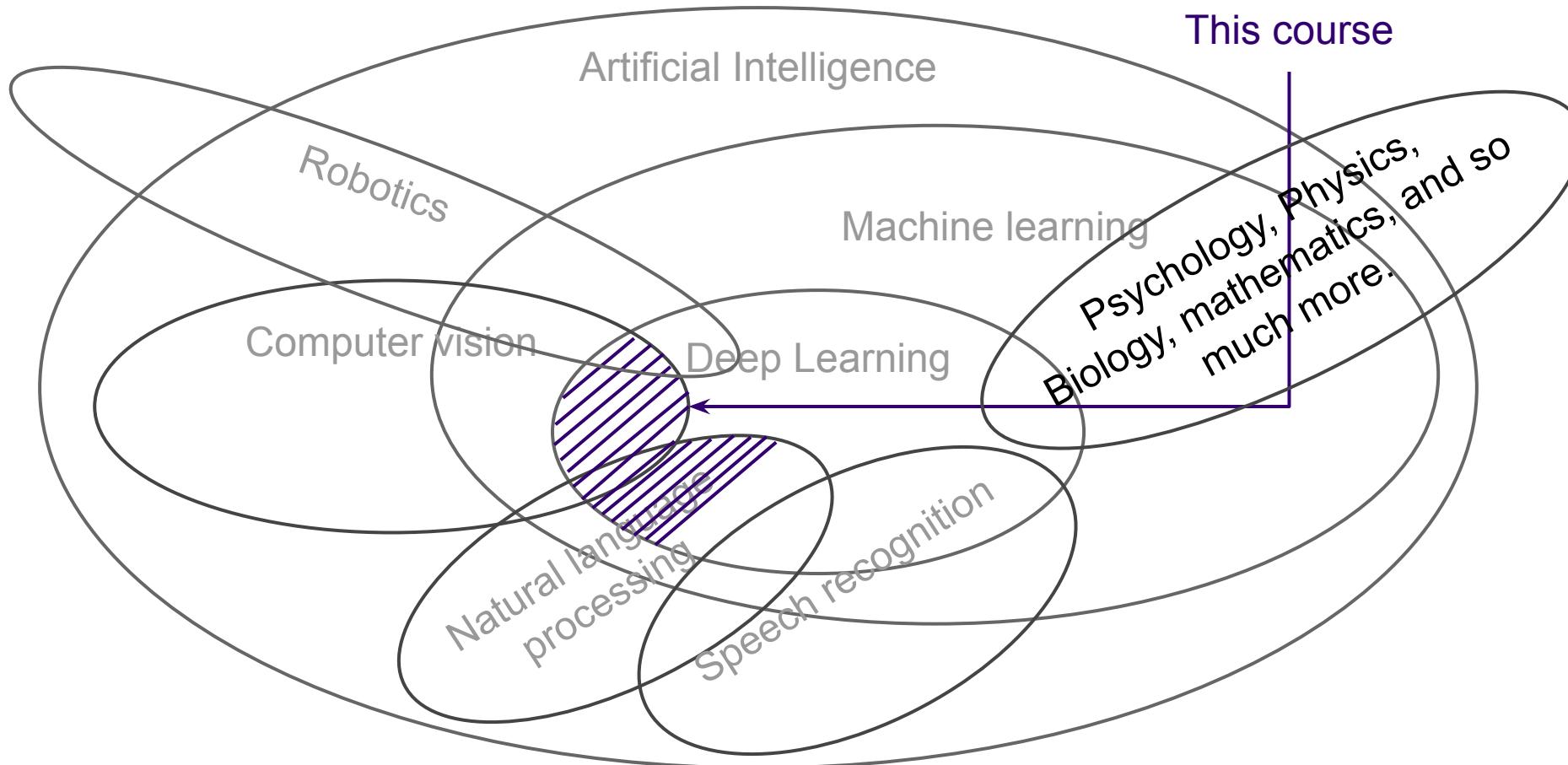
Computer vision

Deep Learning

Natural language
processing

Speech recognition





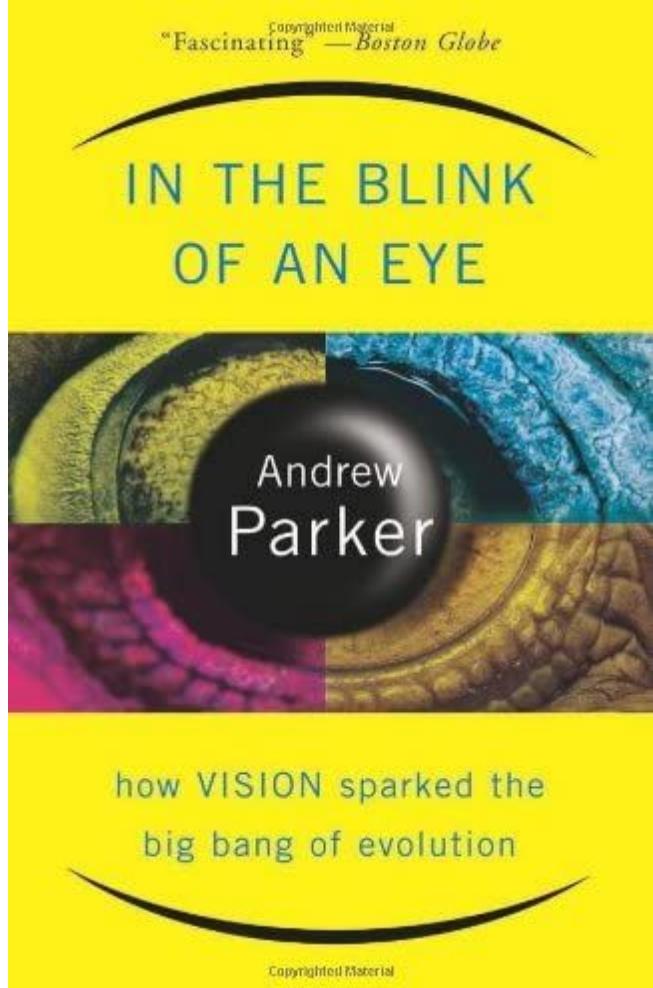
Today's agenda

- A brief history of deep learning
- CSE 493G1 overview



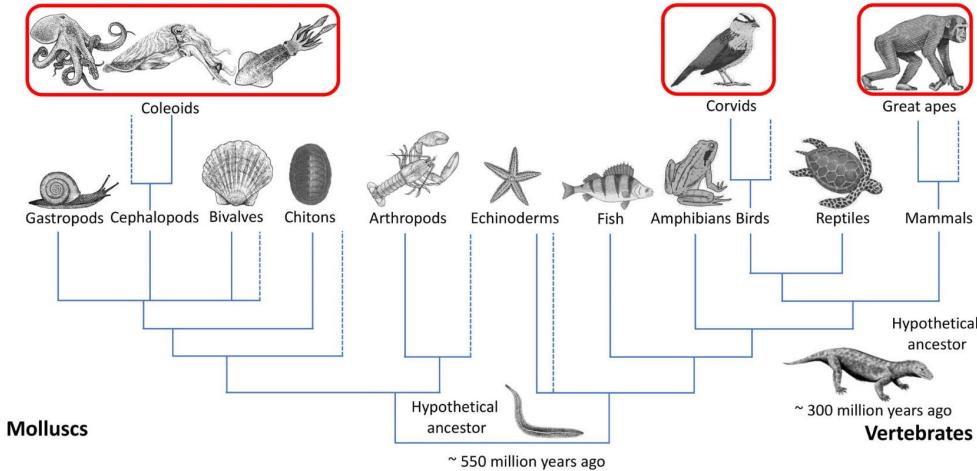
Vision is core to the evolution of intelligence

543 million years ago.



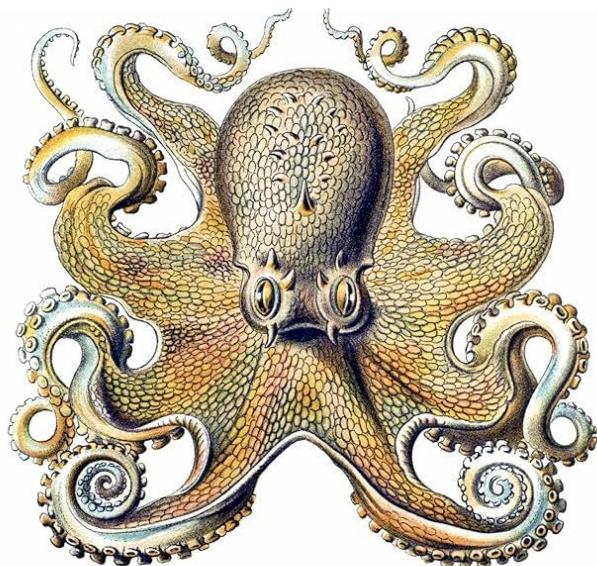
Octopus evolved to have the same eyes as we do

They split from us before eyes evolved.



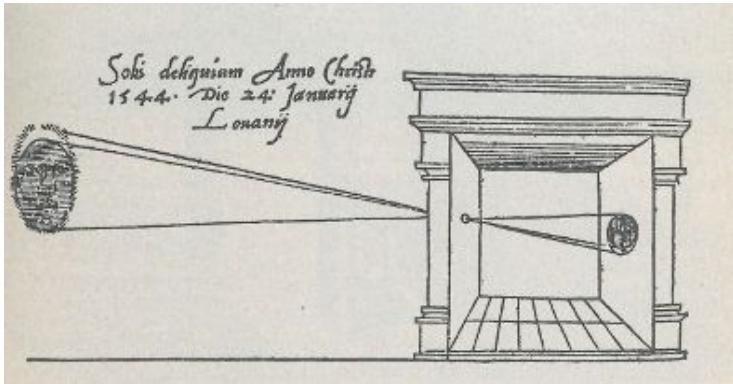
OTHER MINDS

THE OCTOPUS,
THE SEA, AND
THE DEEP ORIGINS
of CONSCIOUSNESS



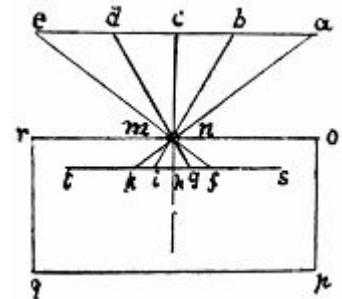
PETER GODFREY-SMITH

The first attempts at capturing the visual world

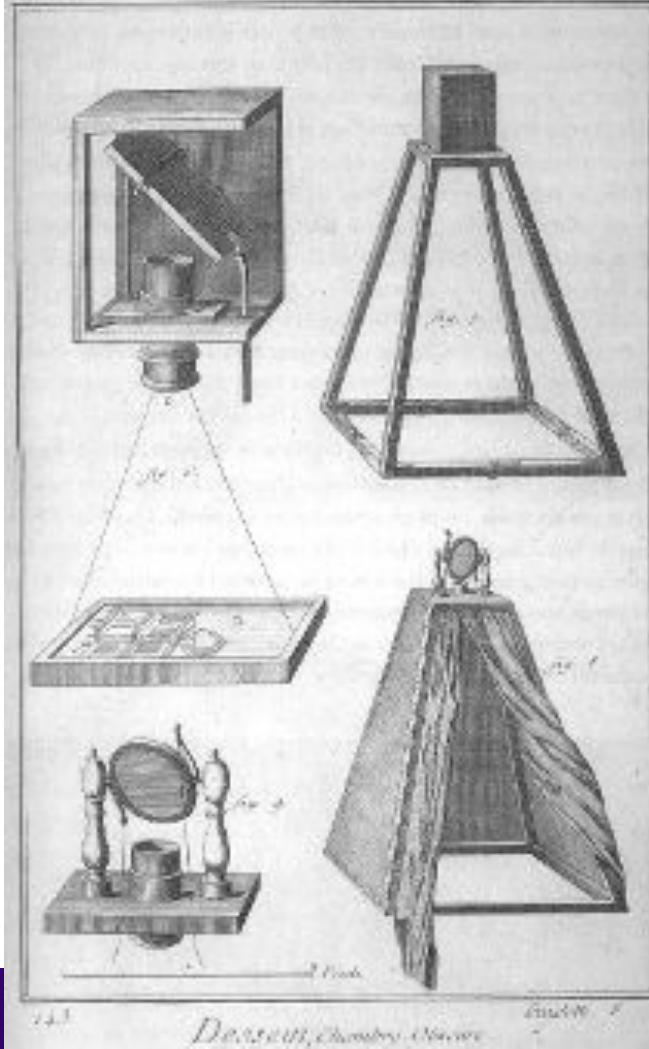


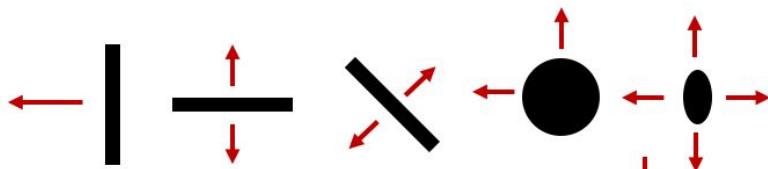
Camera obscura by Gemma Frisius, 1545

Inspired Leonardo da Vinci,
16th Century AD



Examples from 18th century Encyclopedia
Lecture 1 -





Hubel & Wiesel, 1959

How does animal vision work?

Won Nobel Prize in 1981

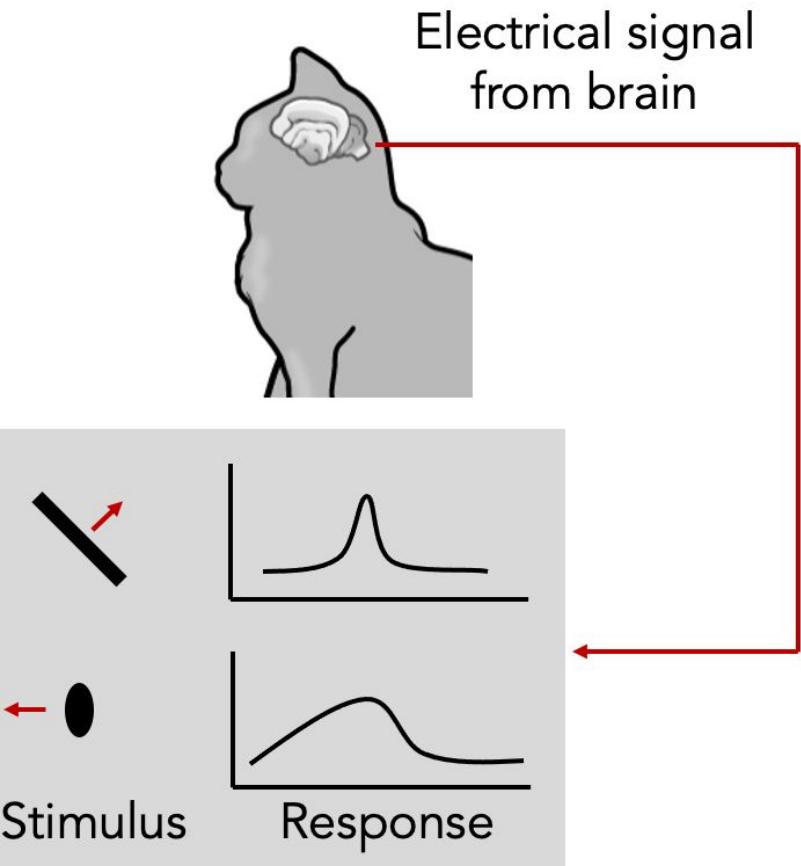
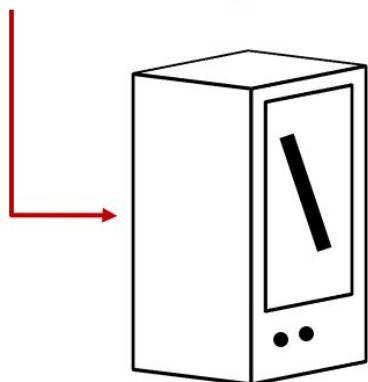
Visual processing is hierarchical, involving recognizing simpler structures, edges, etc.



No response



Response (end point)



The summer vision project

Organized by
Seymour Papert

Computer vision was
meant to be just a
simple summer
intern project

Artificial Intelligence Group
Vision Memo. No. 100.

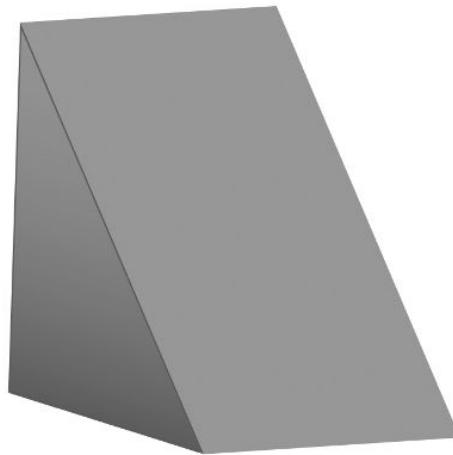
July 7, 1966

THE SUMMER VISION PROJECT

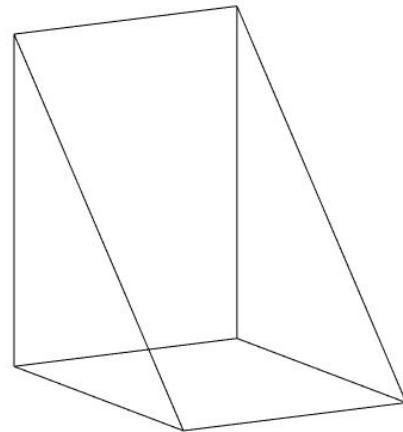
Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

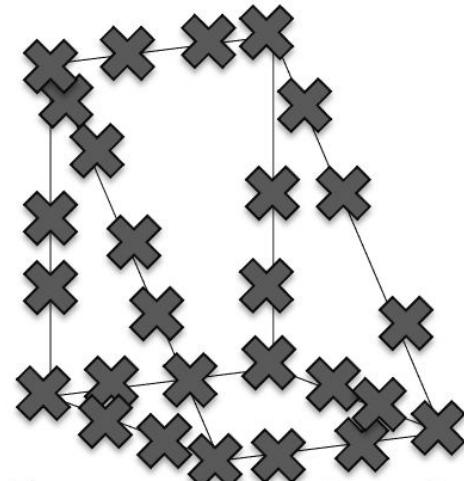
Larry Roberts - Father of computer vision



(a) Original picture



(b) Differentiated picture



(c) Feature points selected

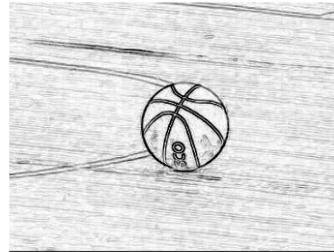
Synthetic images, building up the visual world from simpler structures

Input image

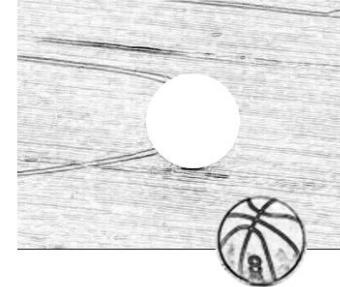


[This image is CC0 1.0 public domain](#)

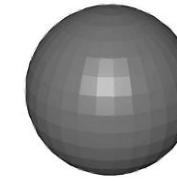
Edge image



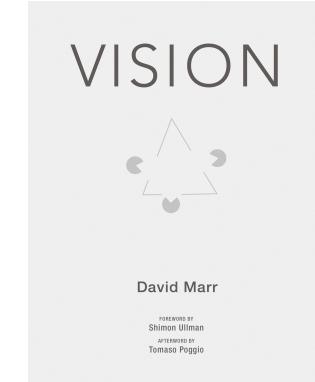
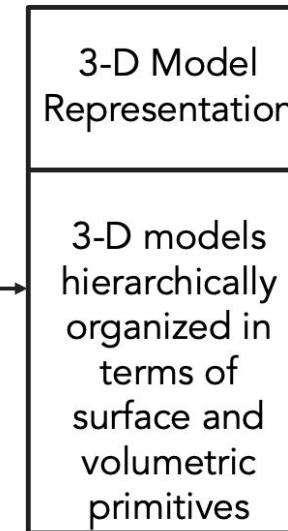
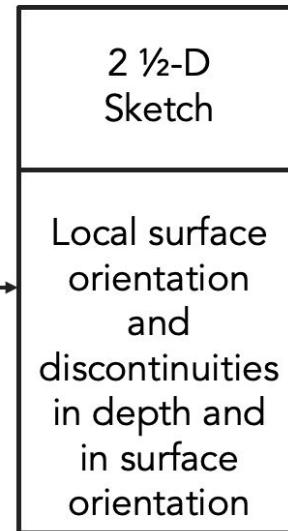
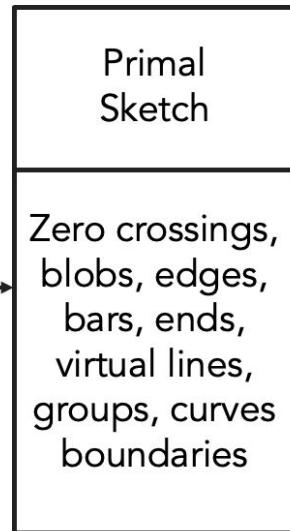
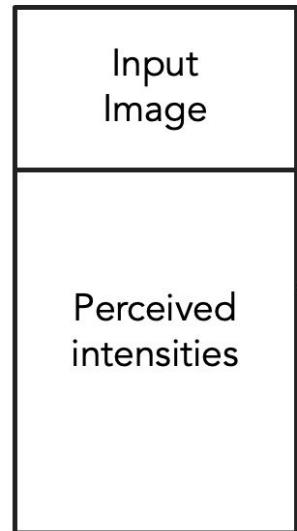
2 ½-D sketch



3-D model

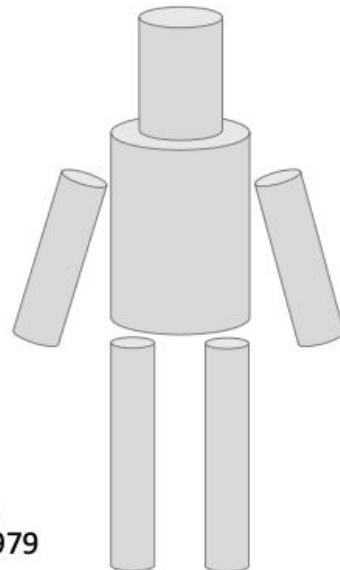


[This image is CC0 1.0 public domain](#)

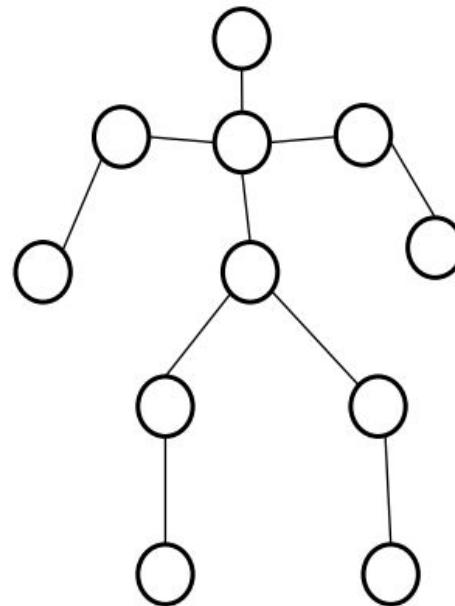


Book published in 1970

Recognition via parts (1970s)



Generalized Cylinders,
Brooks and Binford, 1979



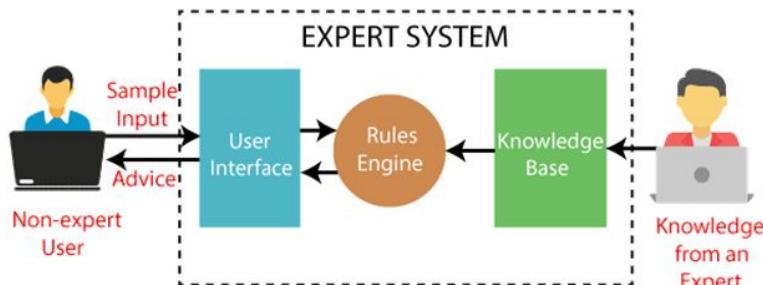
Pictorial Structures,
Fischler and Elshlager, 1973

Recognition via edge detection (1980s)



John Canny, 1986 David Lowe, 1987

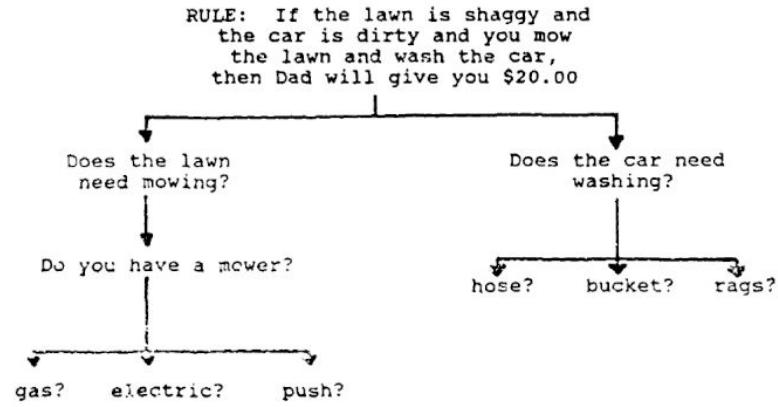
1980s caused one of the larger AI winters (the second AI winter)



Originally called heuristic programming project.

BACKWARD CHAINING

GOAL: Make \$20.00



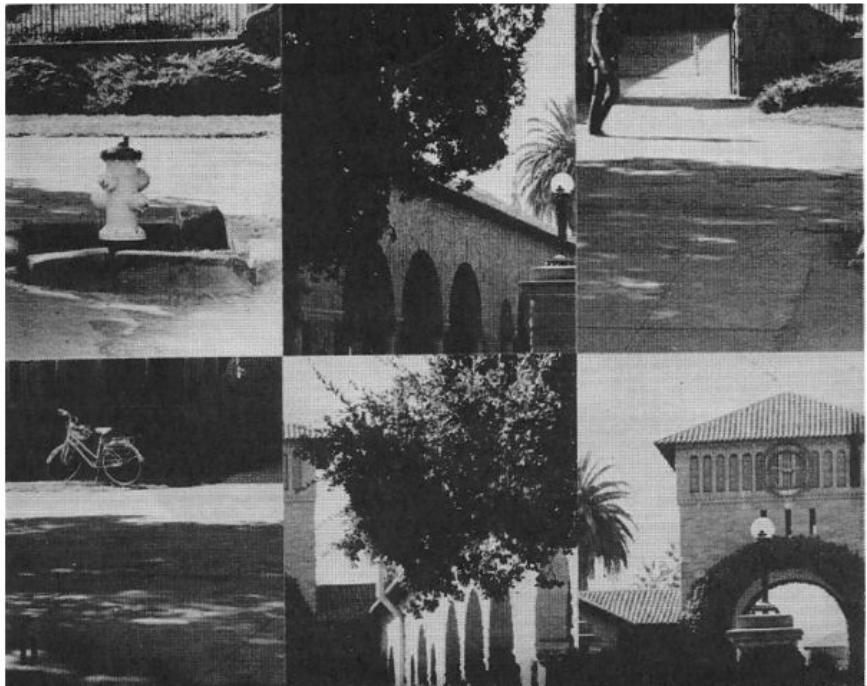
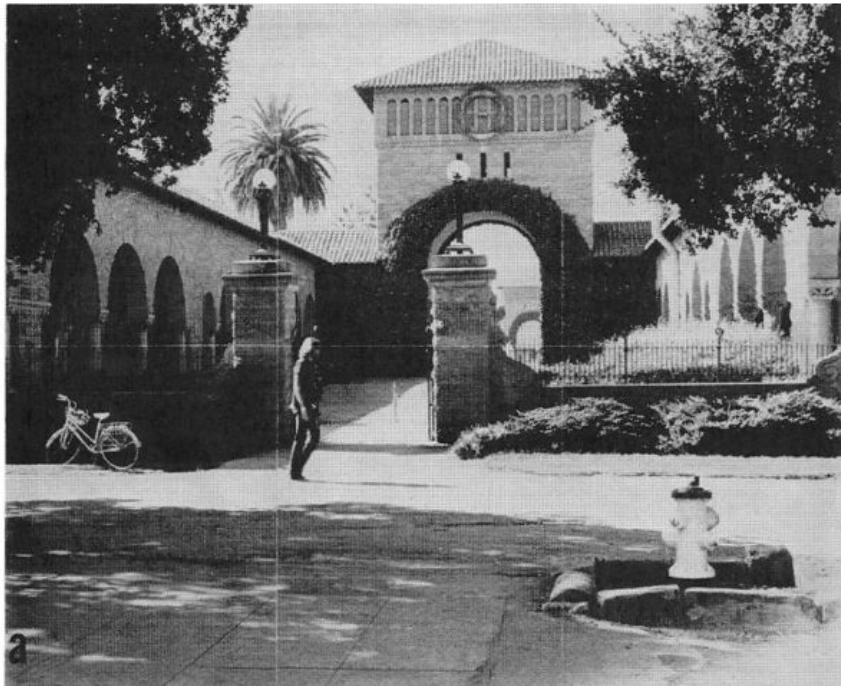
*** The inference engine will test each rule or ask the user for additional information.

- Enthusiasm (and funding!) for AI research dwindled
- “Expert Systems” failed to deliver on their promises
- But subfields of AI continued to grow
 - Computer vision, NLP, robotics, compbio, etc.

In the meantime...seminal work
in cognitive and neuroscience

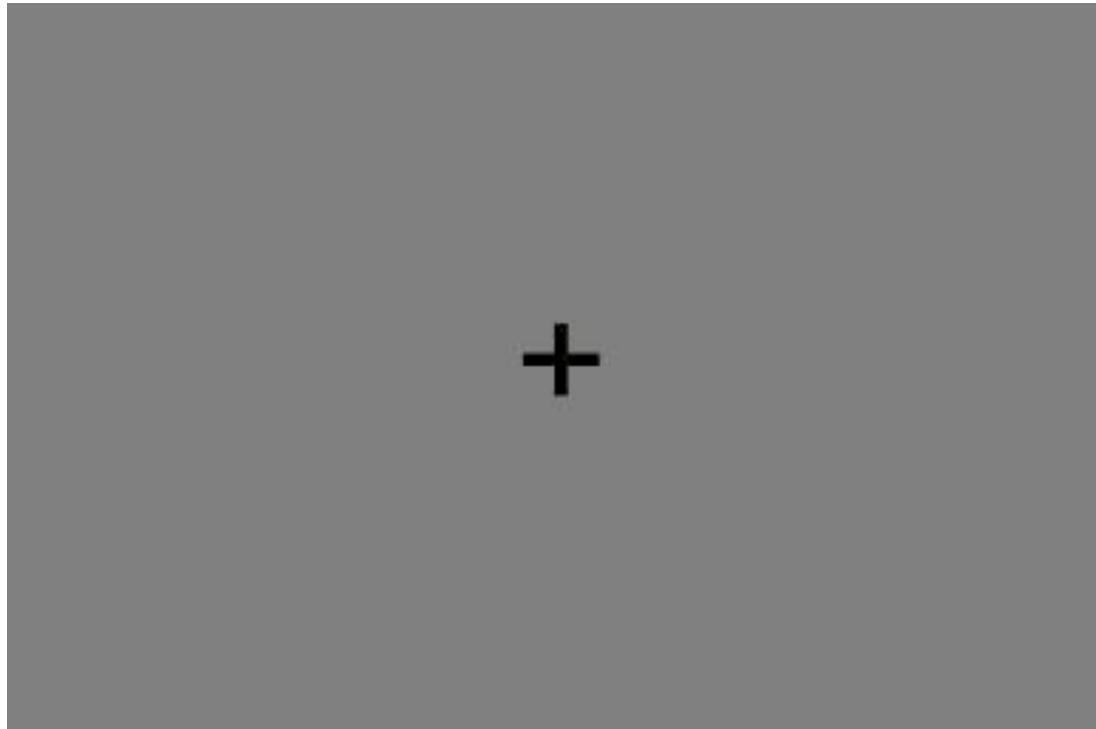
Perceiving real-world scenes

Irving Biederman

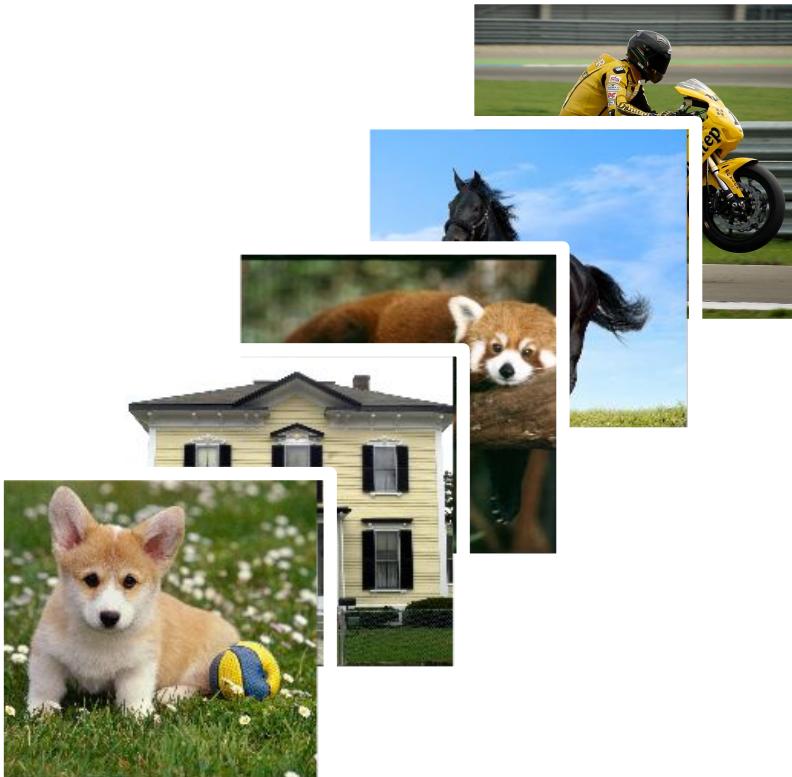


I. Biederman, *Science*, 1972

Rapid Serial Visual Perception (RSVP)



Potter, etc. 1970s



RSVP: Rapid Serial Visual Presentation

Krishna et al. Embracing Error with Rapid Crowdsourcing. CHI 2015









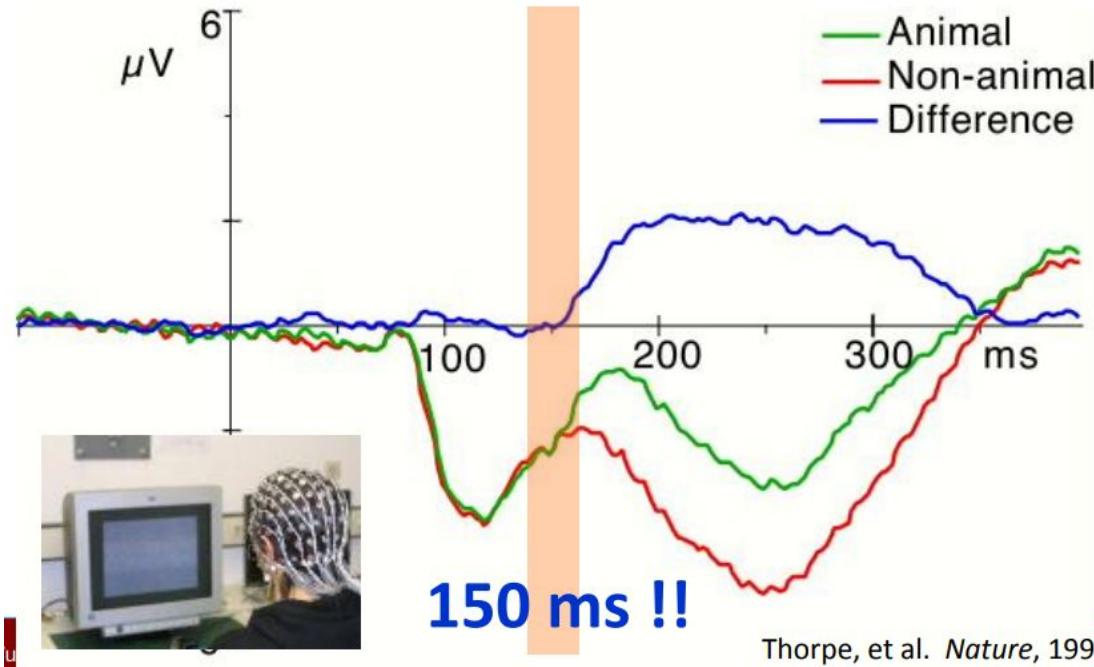






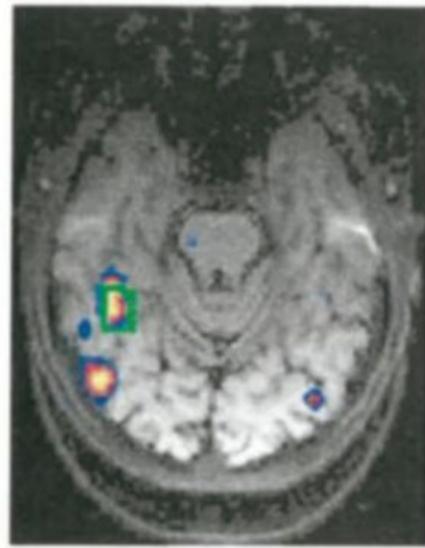


Speed of processing in the human visual system (Thorpe et al. *Nature* 1996)



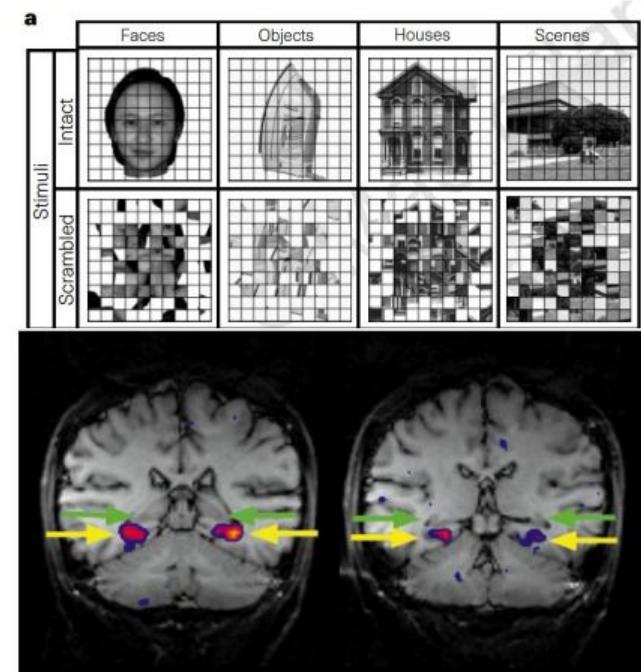
Neural correlates of object & scene recognition

Faces > Houses



% signal change

Kanwisher et al. J. Neuro. 1997



Epstein & Kanwisher, Nature, 1998

Visual recognition is a fundamental to
intelligence [[link](#)]

Searching for Computer Vision North Stars

AUTHORS: Fei-Fei Li and Ranjay Krishna

Until the 90s,
computer vision was not
broadly applied to real world
images

The focus was on algorithms! Recognition via Grouping (1990s)



Shi & Malik,
Normalized Cut, 1997

Recognition via Matching (2000s)



[Image](#) is public domain



[Image](#) is public domain

SIFT, David Lowe, 1999

First **commercial** success of computer vision

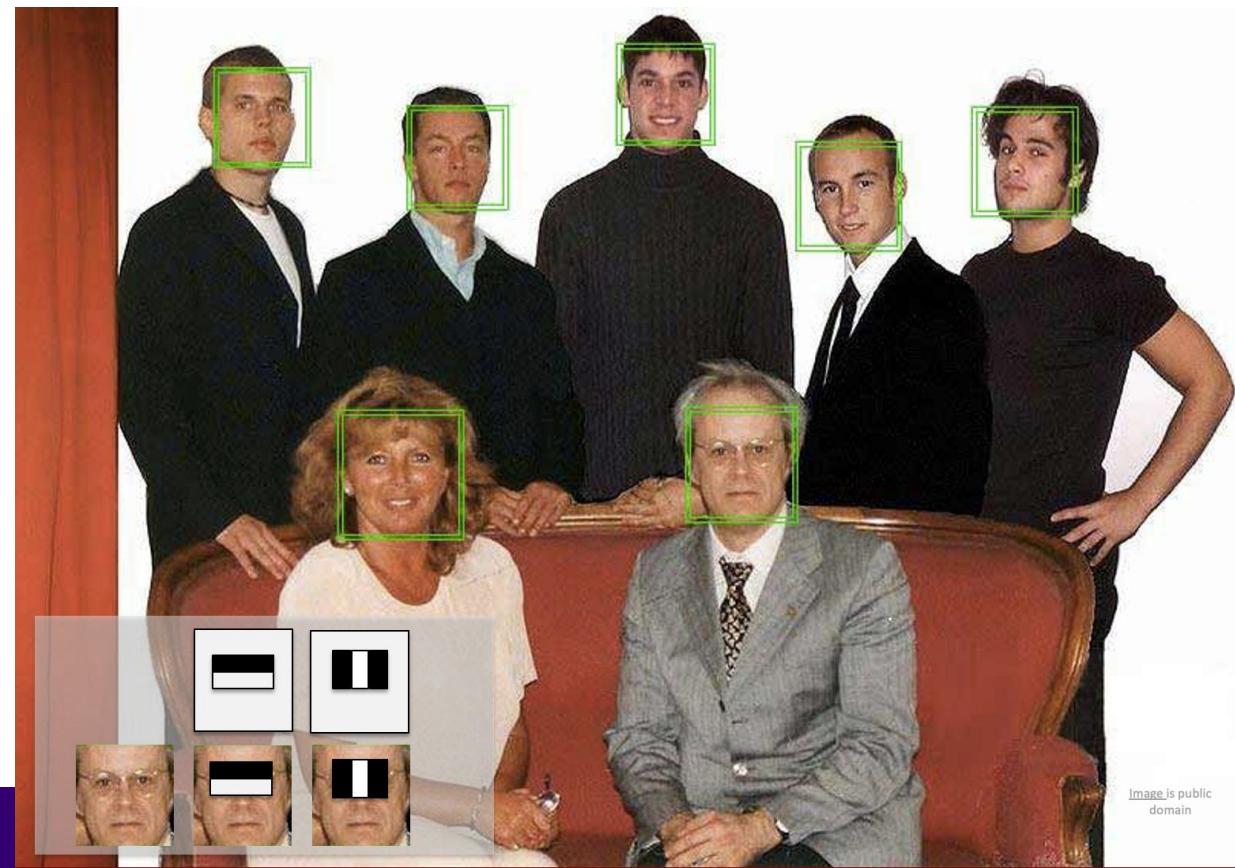
It came from embracing machine learning in 2001.

Does anyone know what it was?

First commercial success of computer vision

Real time face detection
using using an algorithm
by Viola and Jones,
2001

- Fujifilm face detection in cameras
- [HP patent](#) immediately

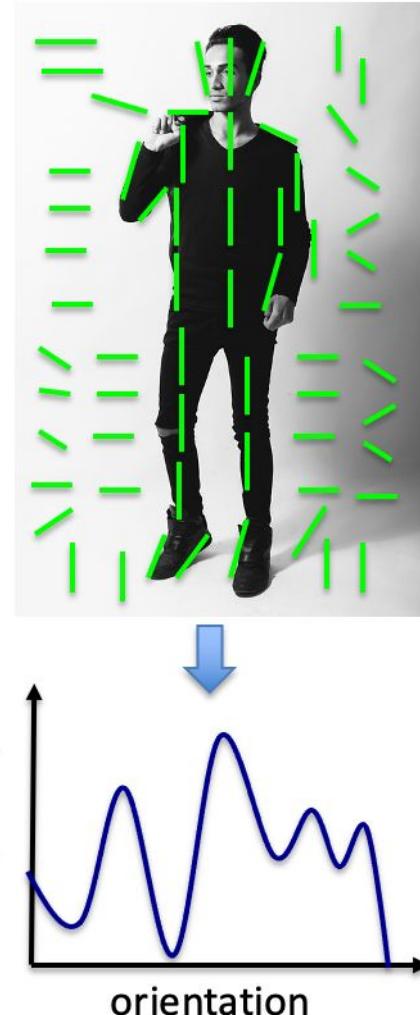


Designing better feature extraction became the focus

HoG features

- Histogram of oriented gradients
- Handcrafted

[Dalal & Triggs, HoG.
2005]



Caltech 101 images



PASCAL Visual Object Challenge

Image is CC0 1.0 public domain



Image is CC0 1.0 public domain

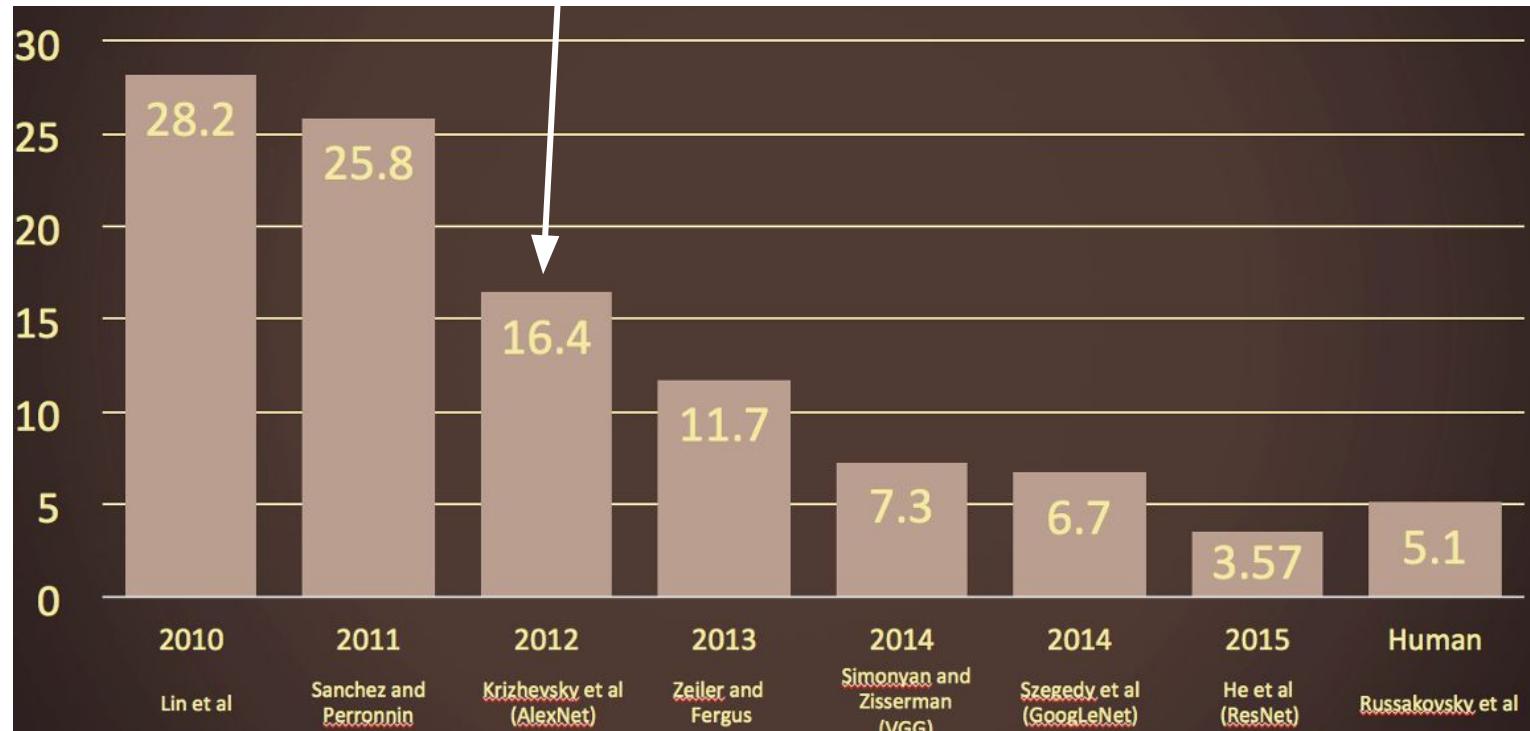
22K categories and **14M** images

- Animals
 - Bird
 - Fish
 - Mammal
 - Invertebrate
- Plants
 - Tree
 - Flower
 - Food
 - Materials
- Structures
 - Artifact
 - Tools
 - Appliances
 - Structures
- Person
- Scenes
 - Indoor
 - Geological Formations
 - Sport Activities

Hypothesis behind ImageNet

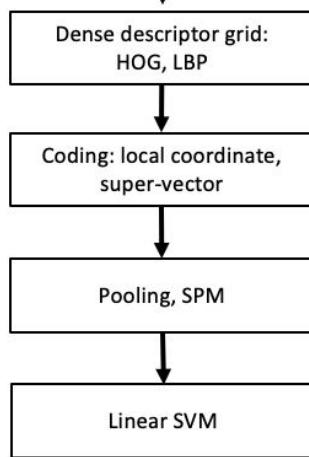
- A child sees nearly 3K unique objects by the age of 6
- Calculated by Irving Biederman
 - [Biederman. Recognition-by-components: a theory of human image understanding. 1983]
- But computer vision algorithms are trained on a handful of objects.

Object recognition error drops by half in 2012 (Enter deep learning)



Year 2010

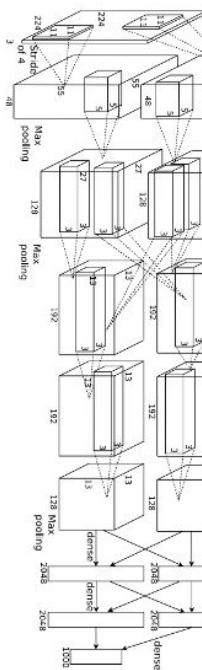
NEC-UIUC



[Lin CVPR 2011]

Year 2012

SuperVision



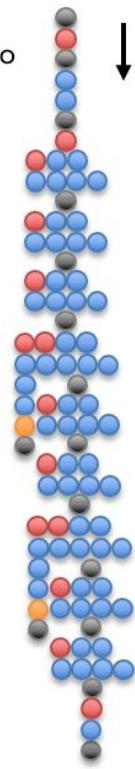
[Krizhevsky NIPS 2012]

Year 2014

GoogLeNet

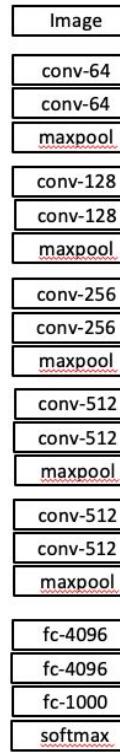
Legend:

- Pooling (Red circle)
- Convolution (Blue circle)
- n (Orange circle)
- Softmax (Black circle)
- Other (Grey circle)



[Szegedy arxiv 2014]

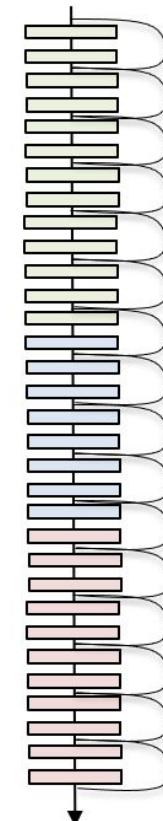
VGG



[Simonyan arxiv 2014]

Year 2015

MSRA



[He ICCV 2015]

AlexNet goes mainstream across computer vision

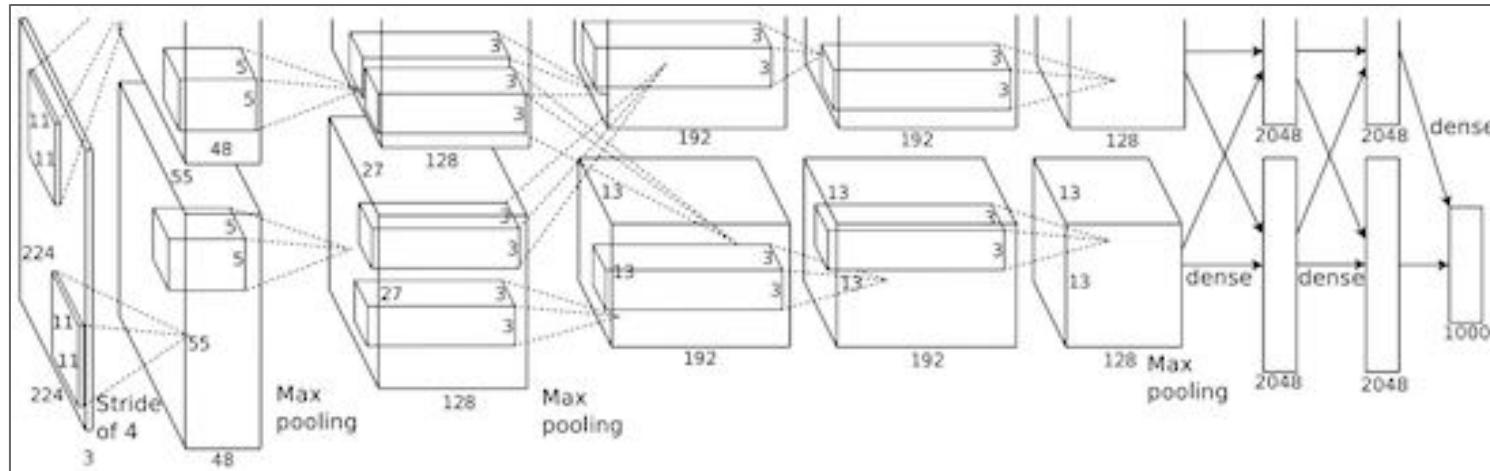


Figure copyright Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, 2012. Reproduced with permission.

“AlexNet”

Core ideas go back many decades!

The **Mark I Perceptron** machine was the first implementation of the perceptron algorithm.

The machine was connected to a camera that used 20×20 cadmium sulfide photocells to produce a 400-pixel image.

recognized
letters of the alphabet

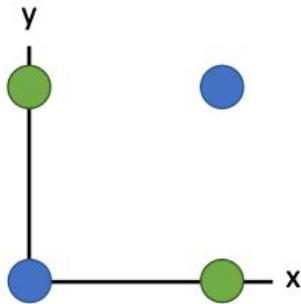
Frank Rosenblatt, ~1957: Perceptron



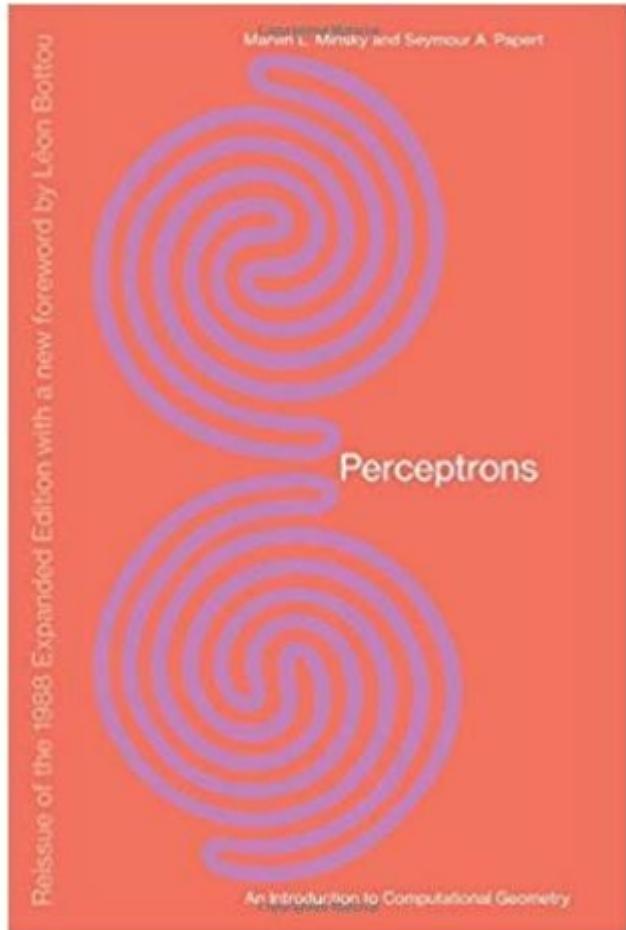
[This image](#) by Rocky Acosta is licensed under [CC-BY 3.0](#)

Minsky and Papert, 1969

X	Y	$F(x,y)$
0	0	0
0	1	1
1	0	1
1	1	0



Showed that Perceptrons could not learn the XOR function
Caused a lot of disillusionment in the field

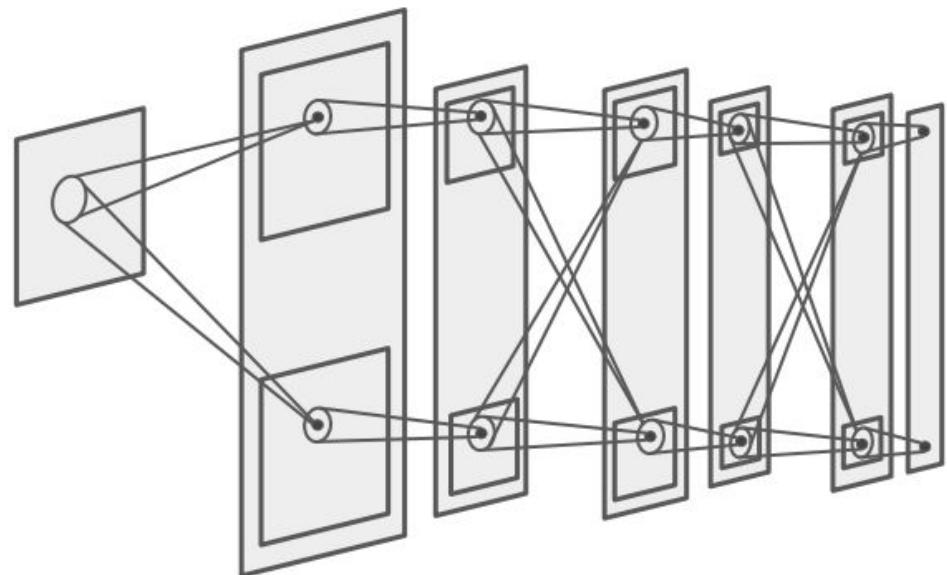


Neocognitron: Fukushima, 1980

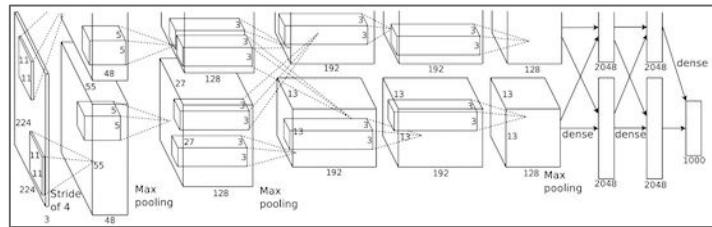
Computational model the visual system, directly inspired by Hubel and Wiesel's hierarchy of complex and simple cells

Interleaved simple cells (convolution) and complex cells (pooling)

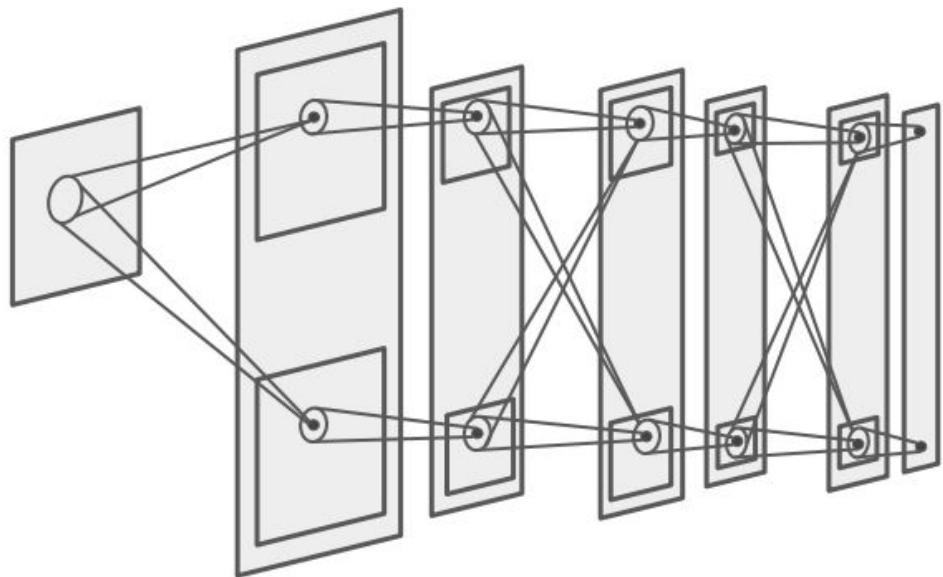
No practical training algorithm



A lot like AlexNet today



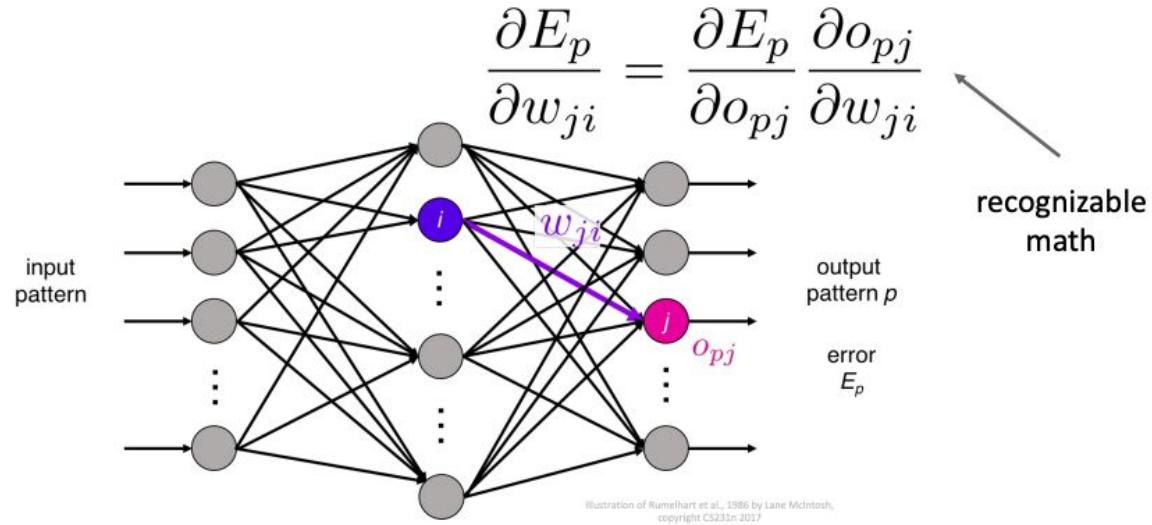
“AlexNet”



Backprop: Rumelhart, Hinton, and Williams, 1986

Introduced backpropagation for computing gradients in neural networks

Successfully trained perceptrons with multiple layers

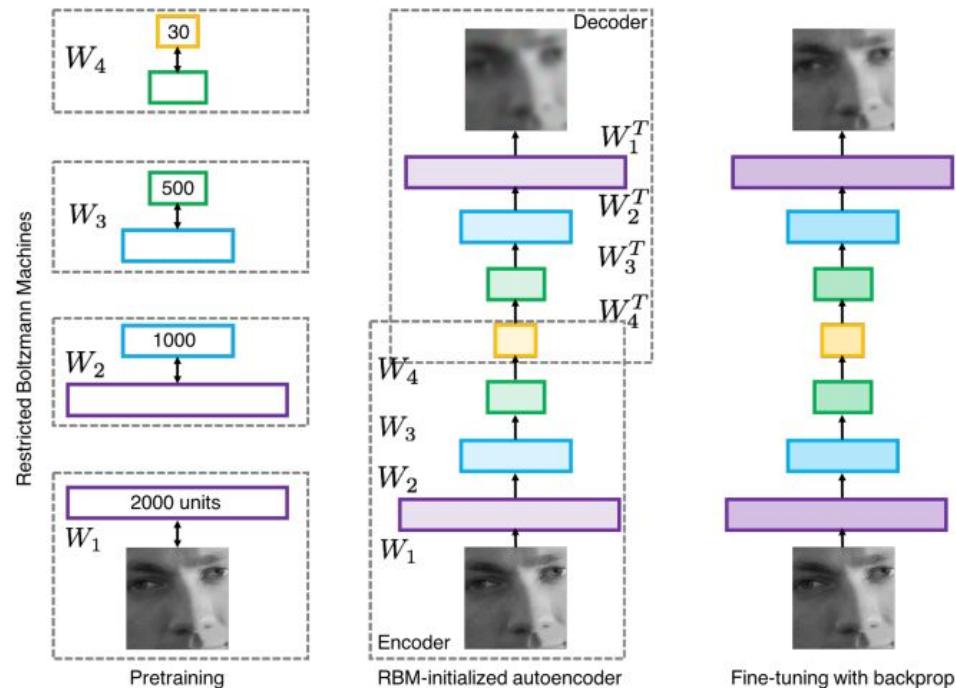


2000s: “Deep Learning”

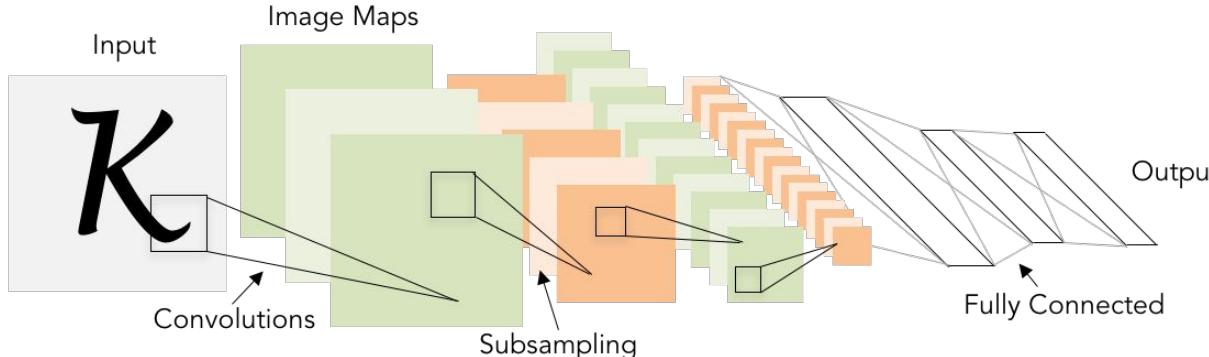
People tried to train neural networks that were deeper and deeper

Not a mainstream research topic at this time

Hinton and Salakhutdinov, 2006
Bengio et al, 2007 Lee et al, 2009
Glorot and Bengio, 2010



1998 LeCun et al.

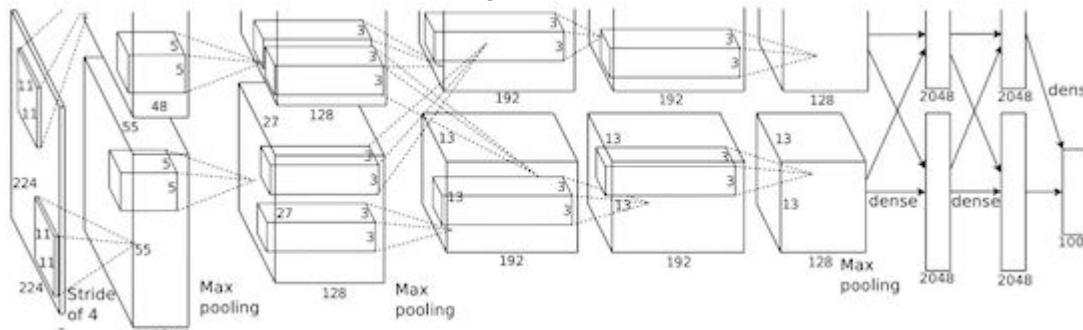


of transistors



10^6

2012 Krizhevsky et al.



of transistors



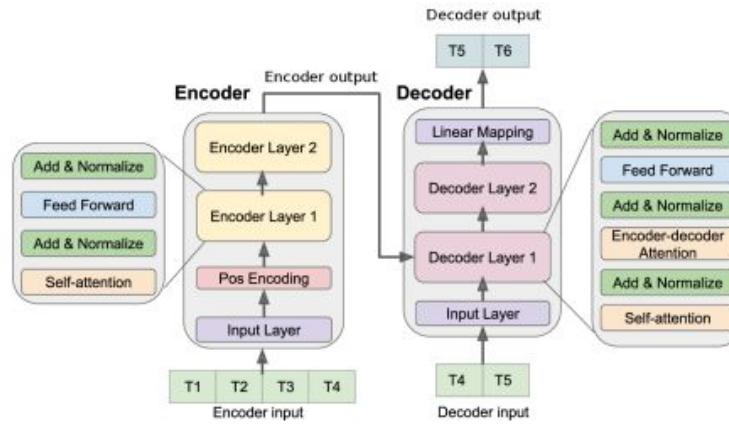
10^9

of pixels used to train:
 10^{14} **IMAGENET**

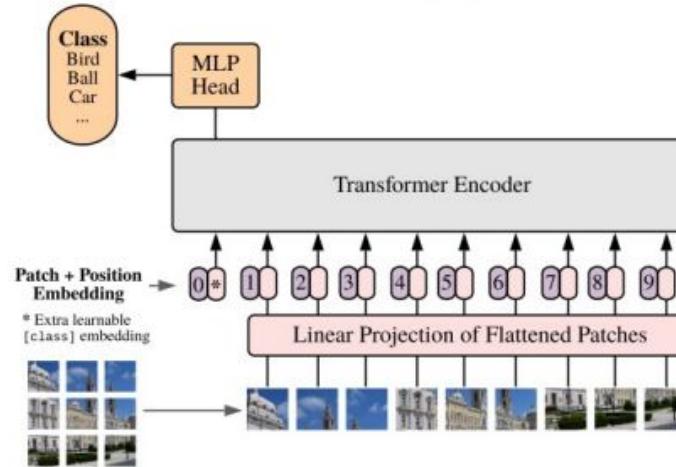
Figure copyright Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, 2012. Reproduced with permission.

Today: Homogenization of Deep Learning

Same models for GPT-4 and image recognition



Transformer Models
originally designed for NLP



Almost identical model (Visual
Transformers) can be applied to
Computer Vision tasks

2012 to present: deep learning is everywhere

Image Classification

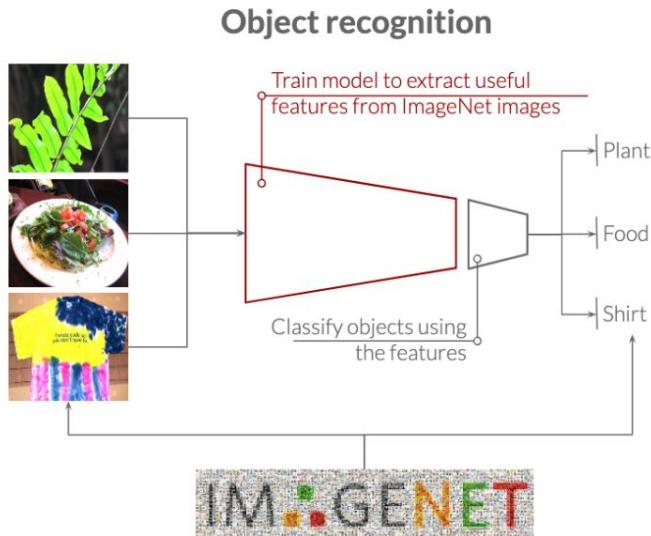


Image Retrieval

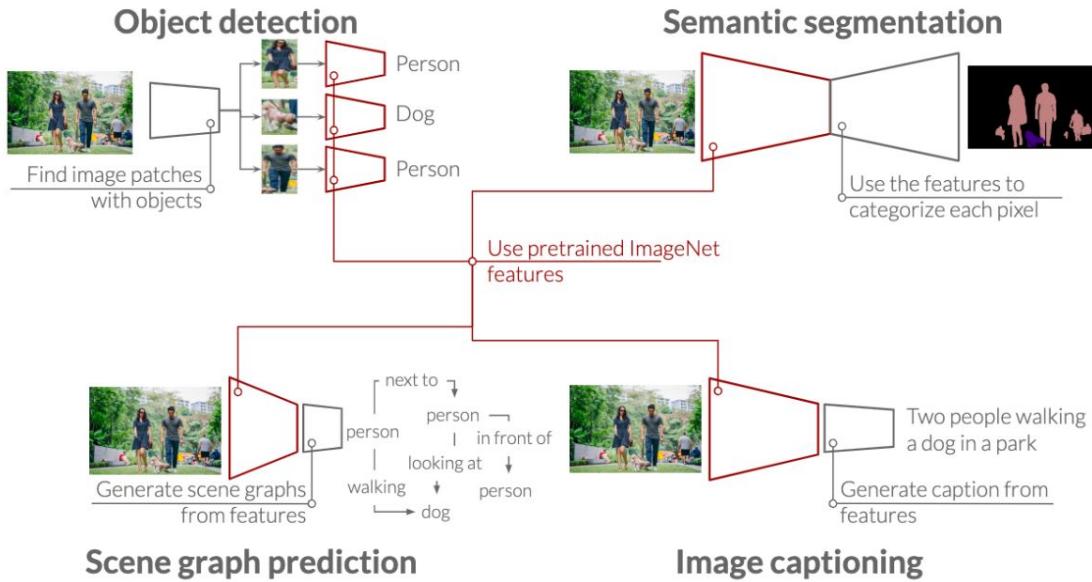


Data hungry machine learning models are now everywhere

Pretraining on ImageNet for object classification

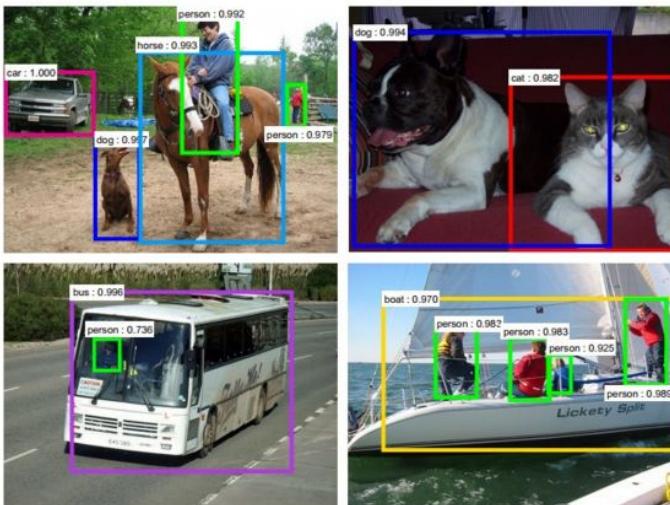


Transfer ImageNet features for many other tasks:



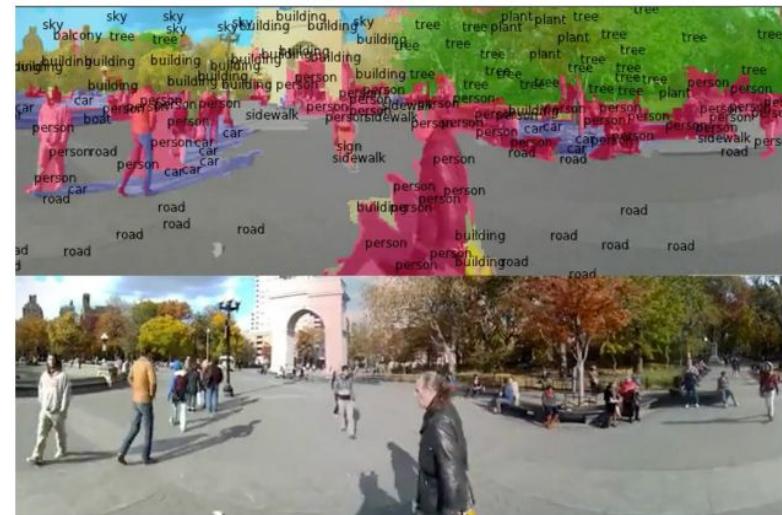
They are used for predicting more than 1 label

Object Detection



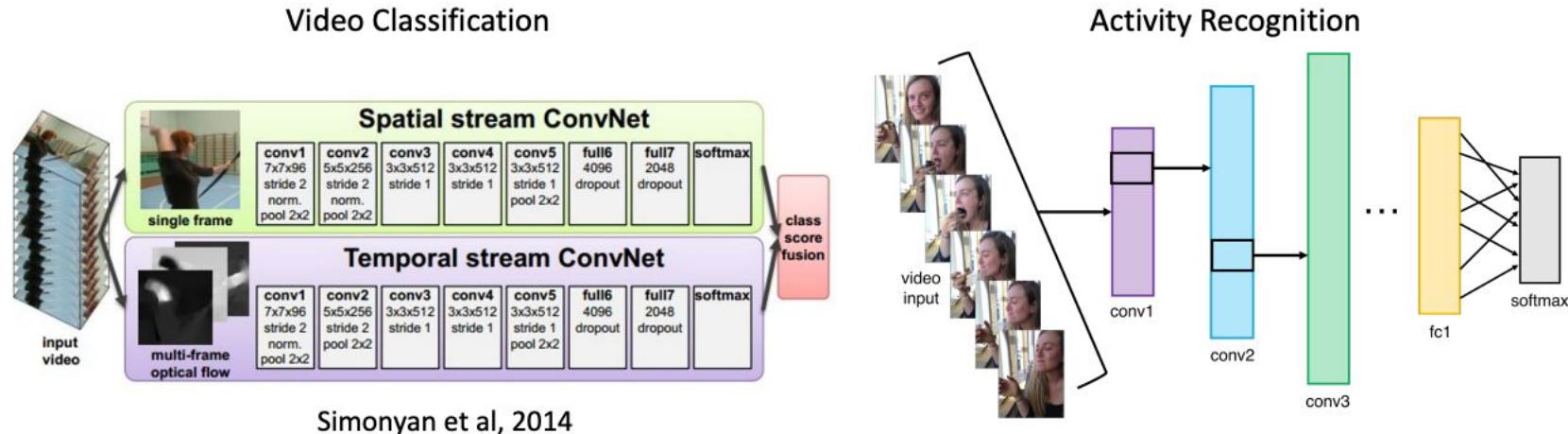
Ren, He, Girshick, and Sun, 2015

Image Segmentation



Fabaret et al, 2012

For accepting not just images but also videos of varying length

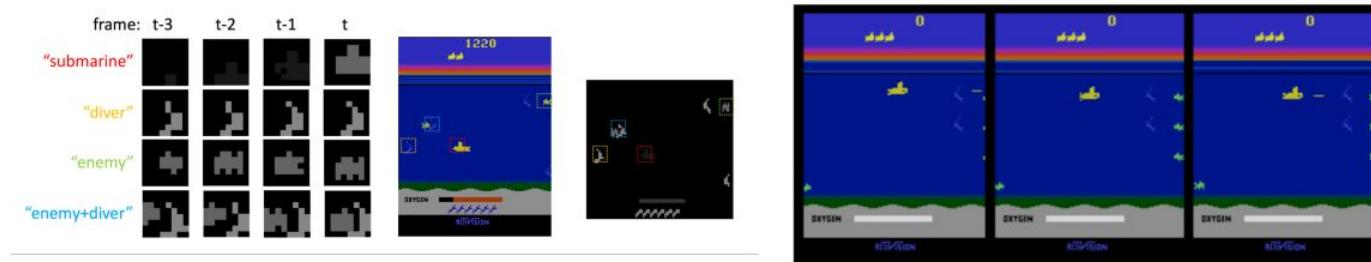


They can be used to track people and their bodies, even play video games

Pose Recognition (Toshev and Szegedy, 2014)

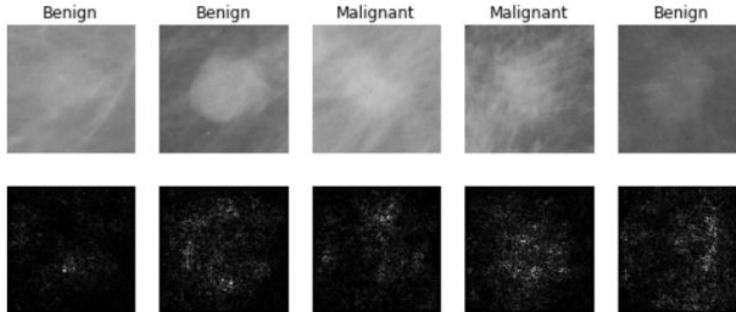


Playing Atari games (Guo et al, 2014)



They can be adapted to new domains/applications

Medical Imaging



Levy et al, 2016

Figure reproduced with permission

Galaxy Classification



Dieleman et al, 2014

From left to right: public domain by NASA, usage permitted by
ESA/Hubble, public domain by NASA, and public domain

Whale recognition



[Kaggle Challenge](#)

This image by Christin Khan is in the public domain and originally came from the U.S. NOAA.

Deep learning techniques work across images with language



Image Captioning

Vinyals et al, 2015
Karpathy and Fei-Fei, 2015



A man riding a wave on top of a surfboard



A cat sitting on a suitcase on the floor

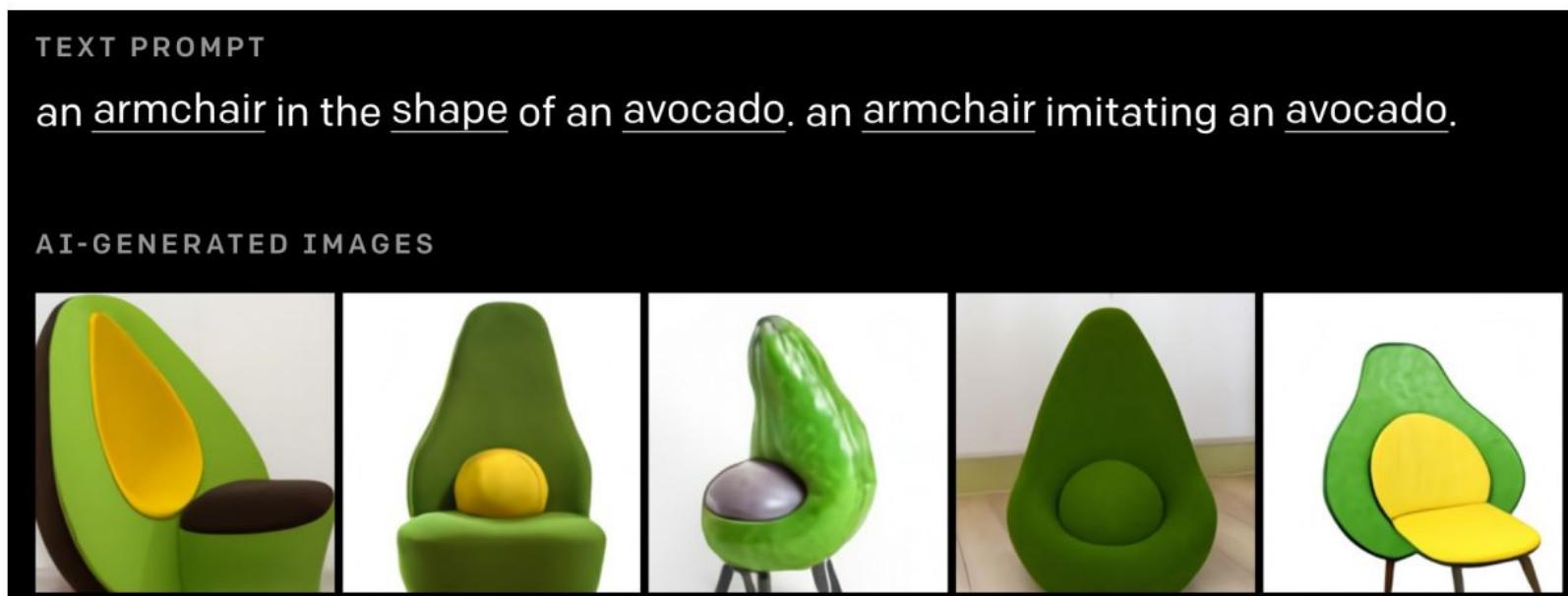


A woman standing on a beach holding a surfboard

All images are CC0 Public domain:
<https://pixabay.com/en/luggage-antique-cat-1643010/>
<https://pixabay.com/en/teddy-plush-bears-cute-teddy-bear-1623436/>
<https://pixabay.com/en/woman-female-model-portrait-adult-983715/>
<https://pixabay.com/en/handsstand-lake-meditation-490008/>
<https://pixabay.com/en/baseball-player-shortstop-infield-1045263/>

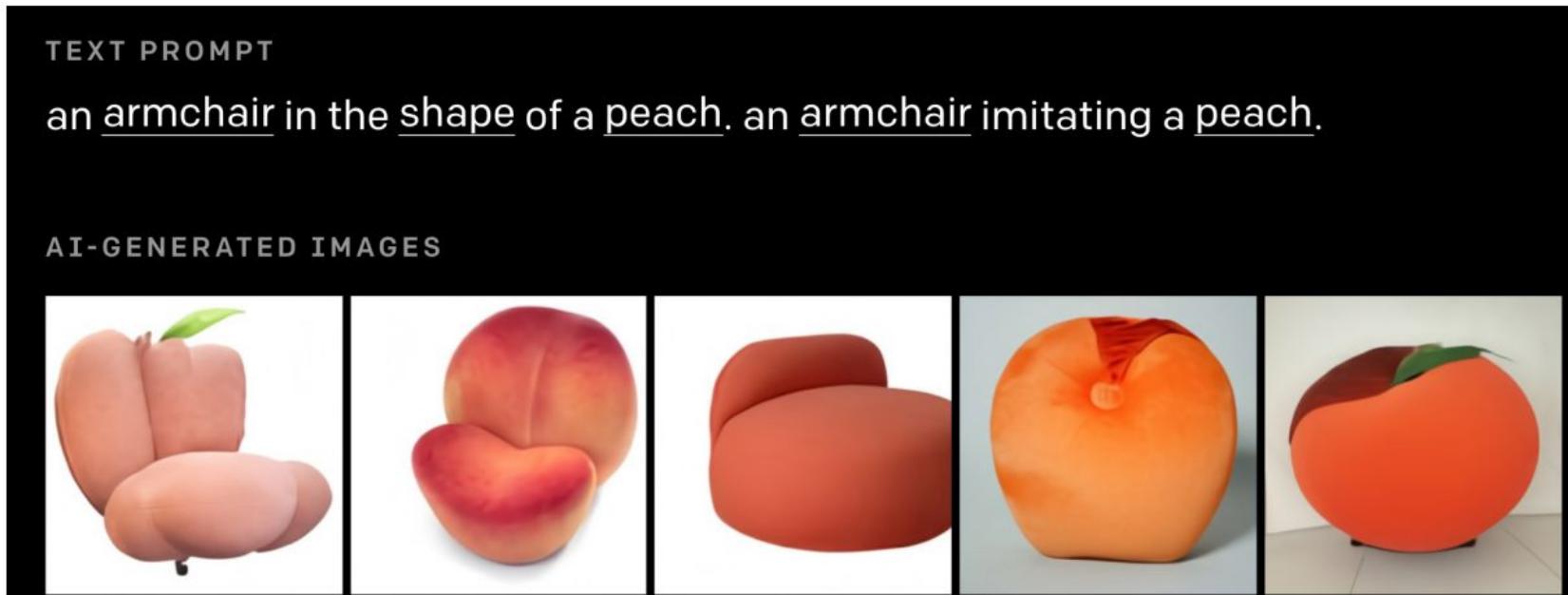
Captions generated by Justin Johnson using Neuraltalk2

Deep learning can generate images



Ramesh et al, “DALL·E: Creating Images from Text”, 2021.
<https://openai.com/blog/dall-e/>

Generations can be controlled by users



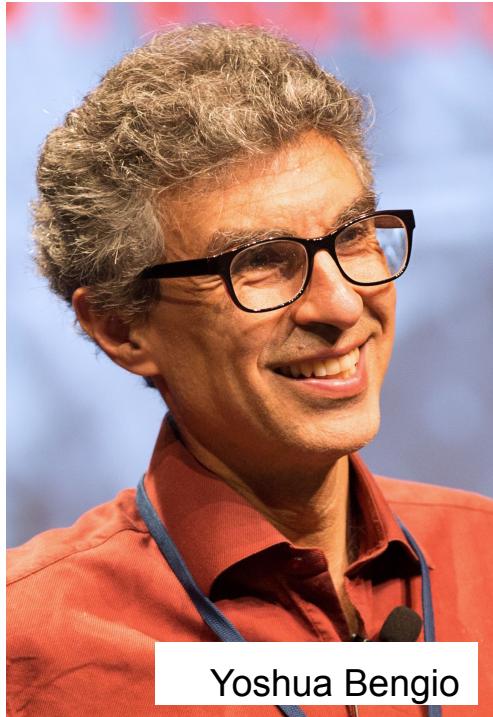
Ramesh et al, "DALL·E: Creating Images from Text", 2021.
<https://openai.com/blog/dall-e/>

2018 Turing Award for deep learning models

most prestigious technical award, is given for major contributions of lasting importance to computing.



Jeffrey Hinton



Yoshua Bengio



Yann LeCun

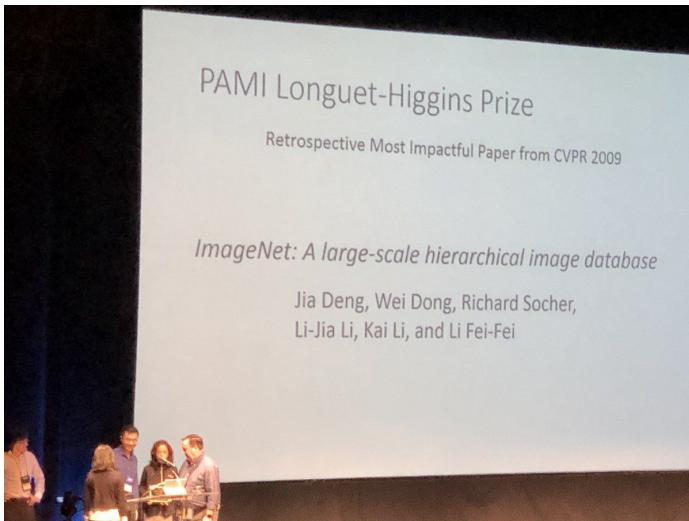
[This image is CC0 public domain](#)

[This image is CC0 public domain](#)

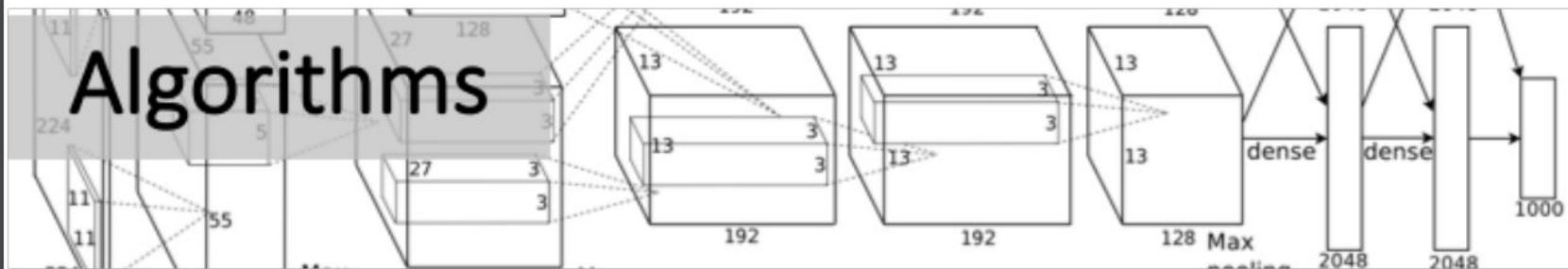
[This image is CC0 public domain](#)

IEEE PAMI Longuet-Higgins Prize awarded in 2019 to ImageNet (published in 2009)

Award recognizes ONE Computer Vision paper from **ten years ago** with **significant impact on computer vision** research.



Algorithms



Data



Computation



Queen Elizabeth Prize for Engineering in 2025

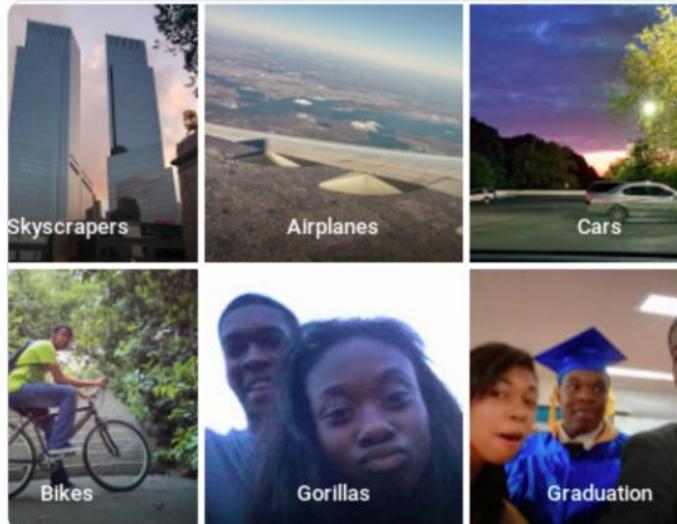


NVIDIA provided the hardware (GPUs)



Despite progress, deep learning can be harmful

Harmful Stereotypes



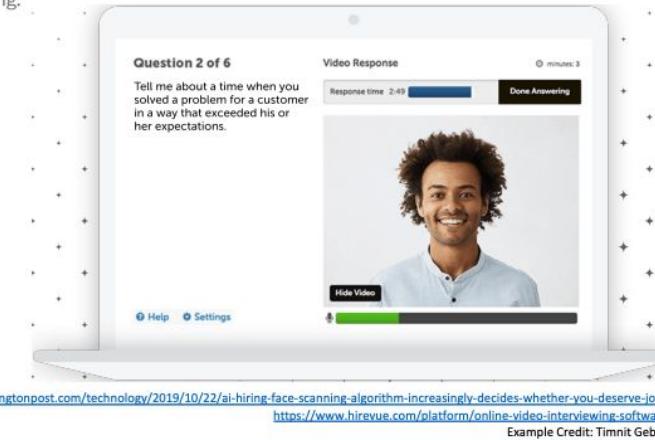
Barocas et al, "The Problem With Bias: Allocative Versus Representational Harms in Machine Learning", SIGKDD 2017
Kate Crawford, "The Trouble with Bias", NeurIPS 2017 Keynote
Source: <https://twitter.com/jackyalcine/status/615329515909156865> (2015)

Affect people's lives

Technology

A face-scanning algorithm increasingly decides whether you deserve the job

HireVue claims it uses artificial intelligence to decide who's best for a job. Outside experts call it 'profoundly disturbing.'



Source: <https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/>
<https://www.hirevue.com/platform/online-video-interviewing-software>

Example Credit: Timnit Gebru

In this course, we will study these
algorithms and architectures
starting from a grounding in
Visual Recognition

A fundamental and general problem in Computer Vision, that has roots in
Cognitive Science

Image Classification: A core task in Computer Vision



cat

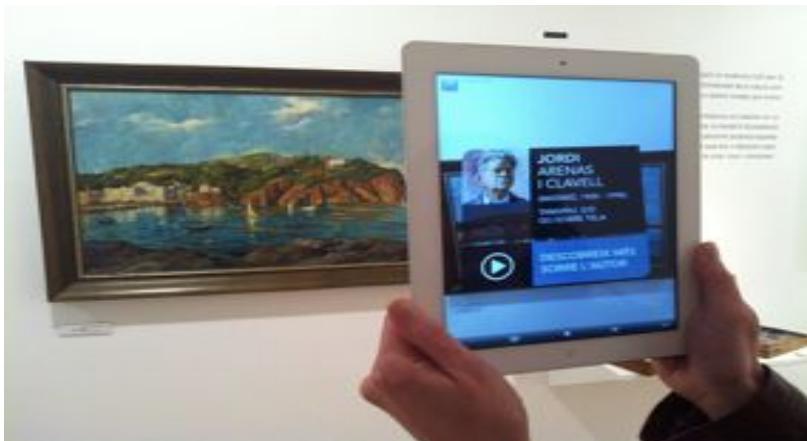
This image by [Nikita](#) is
licensed under [CC-BY 2.0](#)



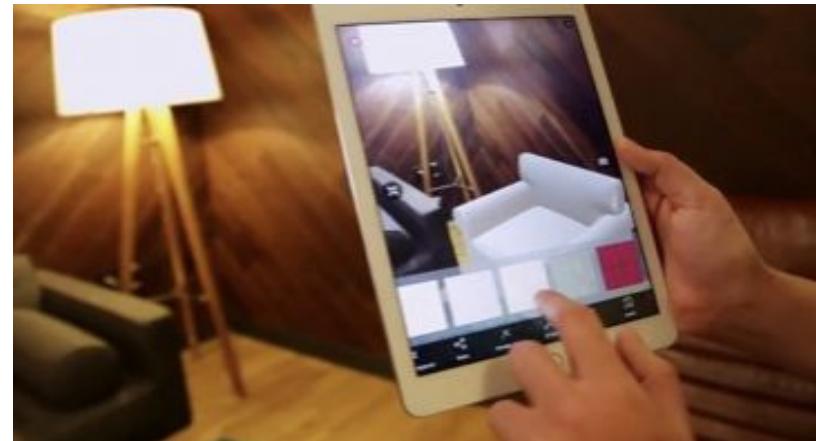
[Image](#) by [US Army](#) is licensed under [CC BY 2.0](#)



[Image](#) is [CC0 1.0](#) public domain



[Image](#) by [Kippelboy](#) is licensed under [CC BY-SA 3.0](#)



[Image](#) by Christina C. is licensed under [CC BY-SA 4.0](#)

Object detection
car



[This image](#) is licensed under [CC BY-NC-SA 2.0](#);
changes made

Action recognition
bicycling



[This image](#) is licensed under [CC BY-SA 3.0](#);
changes made

Scene graph prediction
<person - holding - hammer>

Captioning:
a person holding a hammer



[This image](#) is licensed under [CC BY-SA 3.0](#);
changes made

Beyond recognition: Segmentation, 2D/3D Generation



[This image](#) is CC0 public domain



Progressive GAN, Karras 2018.



Wang et al, "Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images", ECCV 2018

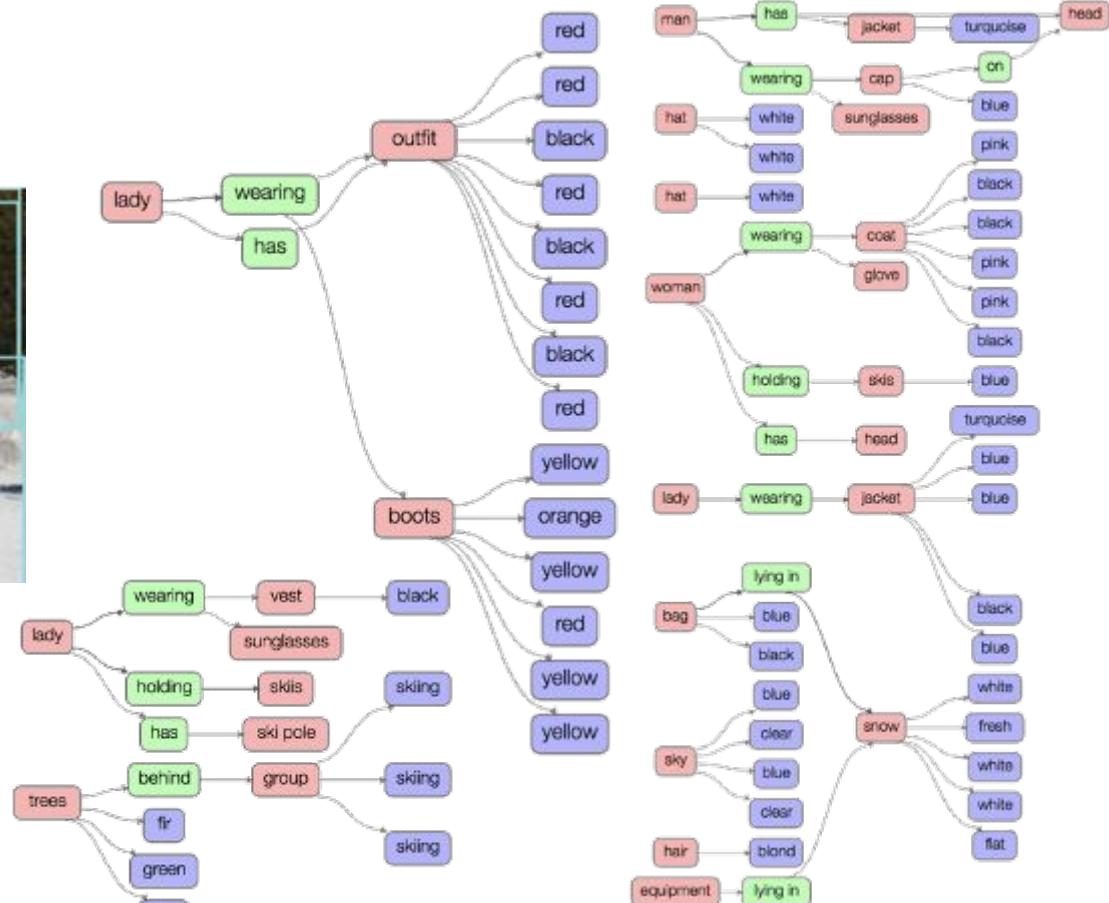
Scene Graphs



This image is CC0 public domain

Three Ways Computer Vision Is Transforming Marketing

- Forbes Technology Council



Krishna et al., Visual Genome: Connecting Vision and Language using Crowdsourced Image Annotations, IJCV 2017

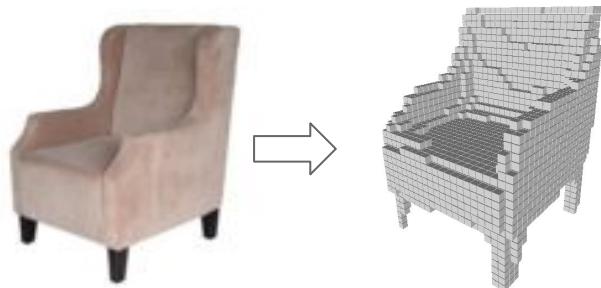
Spatio-temporal scene graphs

Action Genome: Actions as Spatio-Temporal Scene Graphs



Ji, Krishna et al., Action Genome: Actions as Composition of Spatio-temporal Scene Graphs, CVPR 2020

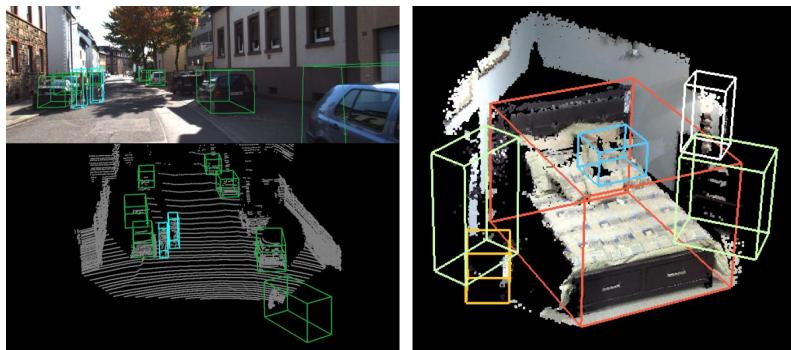
3D Vision & Robotic Vision



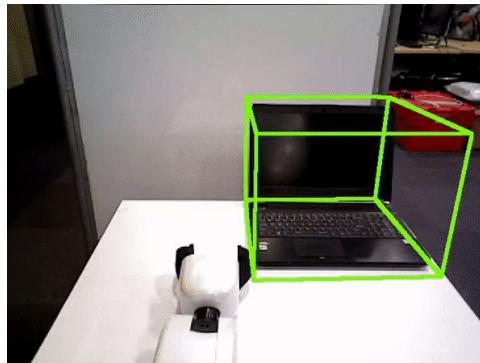
Choy et al., 3D-R2N2: Recurrent Reconstruction Neural Network (2016)



Mandlekar and Xu et al., Learning to Generalize Across Long-Horizon Tasks from Human Demonstrations (2020)



Xu et al., PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation (2018)



Wang et al., 6-PACK: Category-level 6D Pose Tracker with Anchor-Based Keypoints (2020)

Human vision

PT = 500ms



[Image](#) is licensed under [CC BY-SA 3.0](#); changes made

Some kind of game or fight. Two groups of two men? The man on the left is throwing something. Outdoors seemed like because i have an impression of grass and maybe lines on the grass? That would be why I think perhaps a game, rough game though, more like rugby than football because they pairs weren't in pads and helmets, though I did get the impression of similar clothing. maybe some trees? in the background.

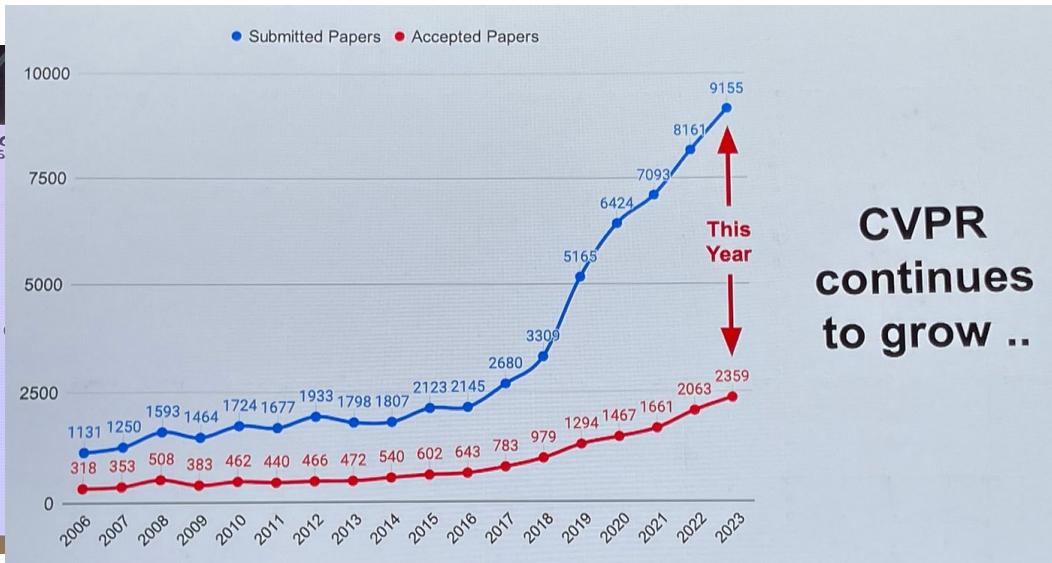
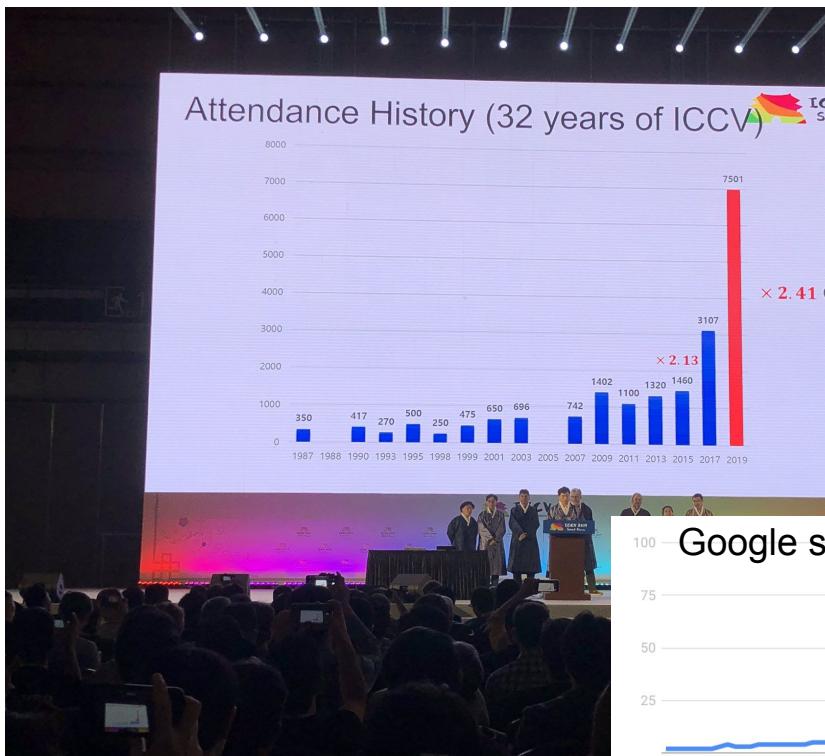
Fei-Fei, Iyer, Koch, Perona, *JoV*, 2007

And there is a lot we don't know how to do



https://fedandfit.com/wp-content/uploads/2020/06/summer-activities-for-kids_optimized-scaled.jpeg

What has deep learning done for computer vision?

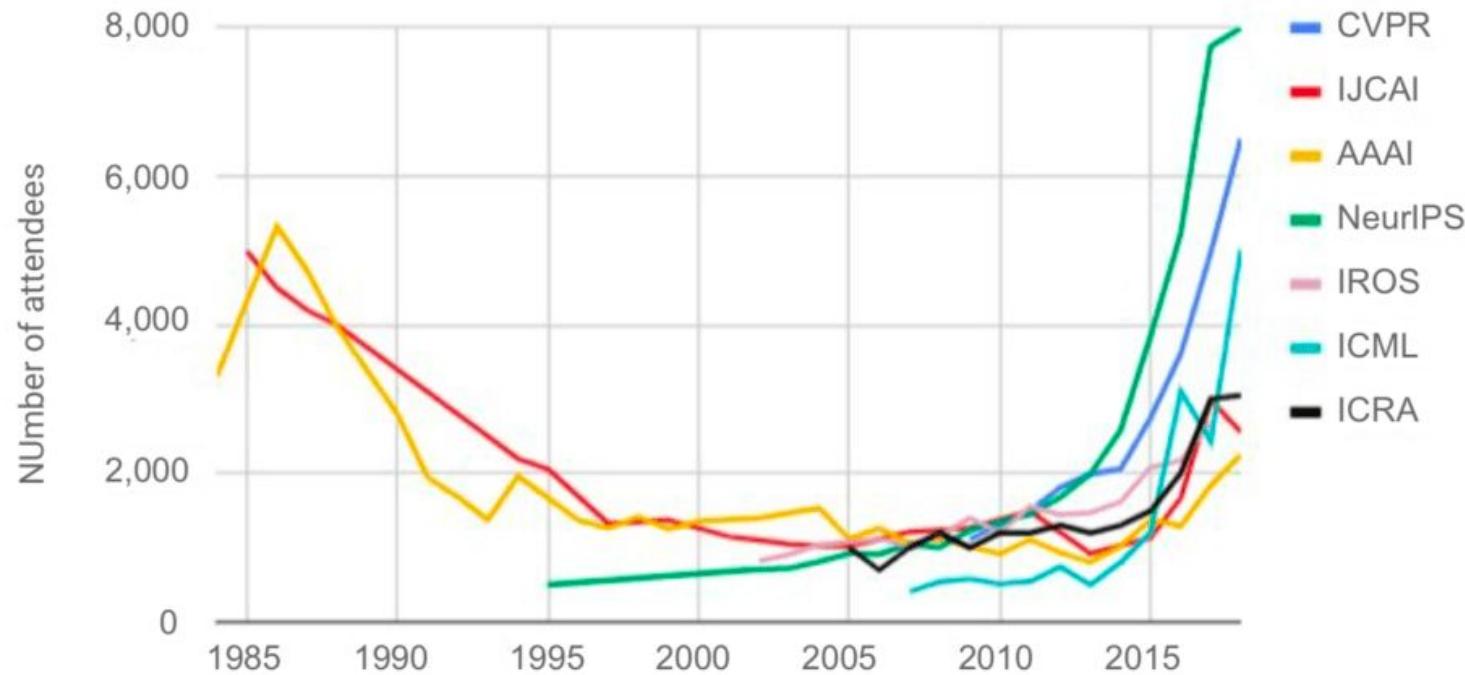


CVPR
continues
to grow ..



Attendance at large conferences (1984–2018)

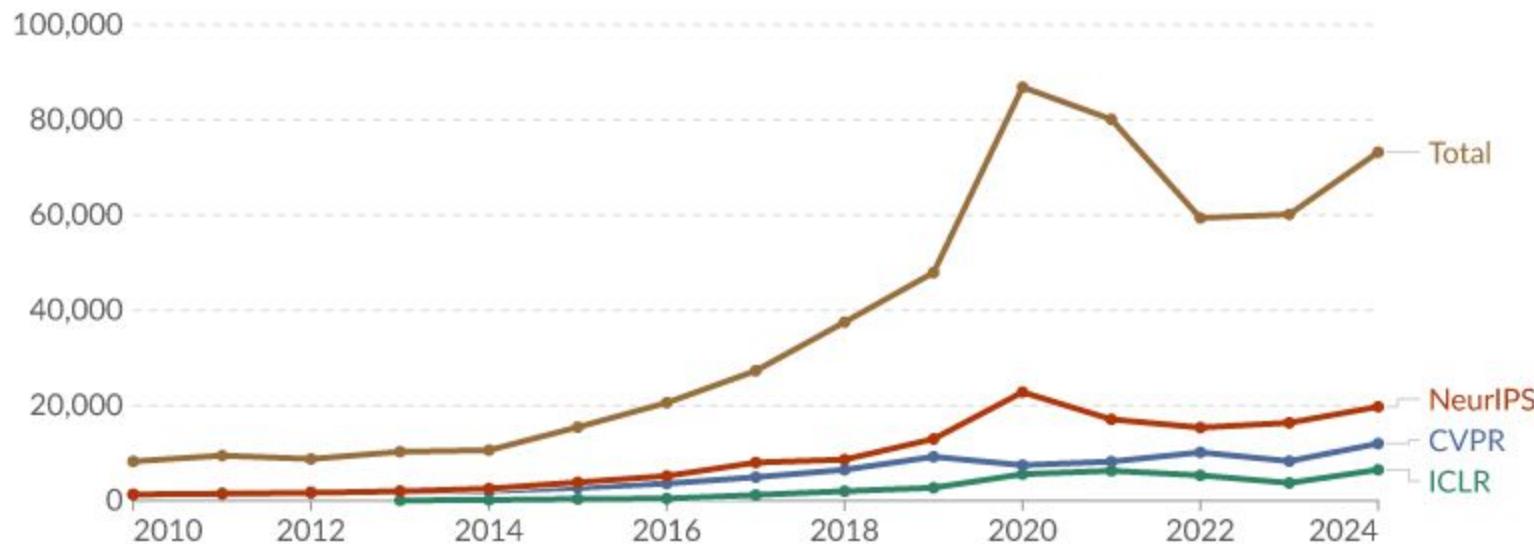
Source: Conference provided data



Annual attendance at major artificial intelligence conferences

Our World
in Data

Thirteen major conferences are included.



Data source: AI Index Report (2025)

CC BY

Note: Most conferences in 2020–2021 were held virtually due to the COVID-19 pandemic.

Today's agenda

- A brief history of computer vision
- CSED 502 overview

Survey - A show of hands

Undergrad?

M.S.?

Ph.D.?

CSE / EE?

Other Engineering?

Math / Natural Science?

Others?

Instructor**Ranjay Krishna**

Hours: Tue,
TBD

CSE2 304

ranjay@cs
.washington.edu

Teaching Assistant**Lindsey Li**

Hours: Wed,
4:30 - 6:00 PM

Forstall & Bahn TA Office - Gates Center 151 or [Zoom](https://zoom.us)

linjli@cs
.washington.edu

Who is Ranjay?

Ranjay Krishna (Assistant Professor at UW CSE)

- PhD from Stanford
- I worked with Fei-Fei Li (**AI**)
- And with Michael Bernstein (**HCI**)



Other courses:

- **UW CSE 455** [2024, 2025]: Computer vision
- **UW CSE 599H** [2023]: Artificial intelligence vs intelligence augmentation
- **Stanford CS 231N** [2020, 2021]: Convolutional neural networks for computer vision
- **Stanford CS 131** [2017, 2018, 2019]: Computer vision fundamentals and applications

What I do aside from teaching? I co-direct the RAIVN lab at UW



<https://raivn.cs.washington.edu/>

What I do aside from teaching?

I lead the computer vision & embodied AI team
At Ai2 (Allen Institute for Artificial Intelligence)



<https://prior.allenai.org/>

Syllabus

Deep learning Fundamentals

Data-driven approaches
Linear classification & kNN
Loss functions
Optimization
Backpropagation
Multi-layer perceptrons
Neural Networks
Convolutions
RNNs / LSTMs
Transformers

Practical training skills

Activation functions
Batch normalization
Transfer learning
Data augmentation
Momentum / RMSProp / Adam
Architecture design
LoRA
RLHF
DPO
Scaling Laws

Applications

Image captioning
Interpreting machine learning
Generative AI
Fairness & ethics
Data-centric AI
Deep reinforcement learning
Self-supervised learning
Diffusion
LLMs



Lectures

In person in Gates building: CSE2 G01

- Recordings will be shared via canvas:
- **Wednesdays between 6:30pm to 9:20pm**
 - To watch the lectures later, you must login to canvas. We highly recommend coming in person
- Slides posted to our website:
 - <https://courses.cs.washington.edu/courses/csed502/26wi/>

Friday recitation sections

Fridays

- Two recitation sections:
 - 9:30-10:20am (MGH 241)
 - 12:30-1:20pm (ECE 125)

Hands-on concepts, some tutorials, more practical details than tuesday/thursday lectures

Check the [syllabus page](#) for more information on what is going to be covered when.

This Friday: Broadcasting & Matrix Calculus (Presenter: Tanush)

2 Midterms

Goal: Evaluate individual understanding of concepts from assignments and lecture

Will consist of multiple choice, T/F, and short answer questions and will take place in lecture (check syllabus page).

It will cover all concepts covered up till the lecture before each exam.

EdStem discussions

For questions about assignments, midterm, projects, logistics, etc, use [EdStem](#)!

SCPD students: Use your @uw.edu address to register for EdStem;

Online Office Hours

See course webpage for schedule.

- Add your name to a queue when you arrive for a particular office hours
- TAs will usually conduct 1-1 conversations in front of the whole group unless otherwise requested for a private conversation.

Optional textbook resources

- [Deep Learning](#)
 - by Goodfellow, Bengio, and Courville
 - Here is a [free version](#)
- Mathematics of deep learning
 - Chapters 5, 6 7 are useful to understand vector calculus and continuous optimization
 - [Free online version](#)
- Dive into deep learning
 - An interactive deep learning book with code, math, and discussions, based on the NumPy interface.
 - [Free online version](#)

Grading

All assignments, coding and written portions, will be submitted via [Gradescope](#).
(Entry Code: YBR7Y6)

We use an **auto-grading system**

- A consistent grading scheme,
- Public tests:
 - Students see results of public tests immediately
- Private tests
 - Generalizations of the public tests to thoroughly test your implementation

Grading

5 Assignments (A1-A5): **10%** each = **50%**

A0 is worth 0%

2 exams in lecture: **24%** each = **48%**

Participation (course evals, asking questions in class): **2%**

Extra Credit in lectures and recitation: up to **5%**

Grading

Late policy

- 5 free late days
- Can use at most 2 per assignment (or proposal or milestone)
- Afterwards, 25% penalty per day late
- No late days for project report
- Weekends count as 1 day.
 - So using 1 late day for a Friday 11:59pm deadline means you can submit by Sunday 11:59pm

Overview on communication

All content will be up to date on the **course website**:

- Syllabus, lecture slides, links to assignment downloads, etc

EdStem:

- Use this for most communication with course staff
- Ask questions about assignments, grading, logistics, etc
- Use private questions if you want to post code

Gradescope:

- For turning in homework and receiving grades

Canvas:

- For watching lecture videos

Assignments

All assignments will be completed using **Google Colab**

- We have a tutorial for how to use Google Colab on the website
- Must use CSE email for Colab, not UW email (non-cse students should already have received CSE email account)

Assignment 0 IS OUT!!!, due 1/13 by 11:59pm

- Easy assignment
- Hardest part is learning how to use colab and how to submit on gradescope
- Worth **0%** of your grade
- Used to evaluate how prepared you are to take this course

Assignments

Assignment 1 will be released this weekend!!!, due 1/23 by 11:59pm

- K-Nearest Neighbor
- Linear classifiers: SVM, Softmax

Final project

- Groups of up to 3
- You can form groups yourselves
 - For students looking for groups, we will help assign you
- Anything related to deep learning or computer vision

I will have project ideas posted

Example final project

Detecting AI-Generated Face Images: A Deep Learning Approach for Combating Disinformation

Yuan Tian, Kefan Ping, Ruijin Ye

Introduction

- Generative models in deep learning have achieved remarkable advancements, producing images that are indistinguishable from real images.
- However, there are concerns about the potential misuse of AI-generated images, such as creating deepfake videos to spread disinformation.
- Our goal is to develop deep neural networks that can automatically and accurately identify AI-generated human face images to prevent illegal activities enabled by AI.

Dataset

• 103,463 Real Faces:

- FFHQ: 70,000 high-quality face images with a resolution of 1024x1024 pixels, created by NVIDIA.
- CelebA-HQ: 30,000 high-quality celebrity face images with various poses and expressions, created by the Multimedia Laboratory at the Chinese University of Hong Kong.
- Quintic AI: 30,000 real face images cropped from the COCO training set and the Labeled Faces in the Wild dataset.

• 63,646 Generated Faces:

- Generated.photos: 10,000 high-quality generated faces that exhibit high variability produced by generated.photos.
- StyleGAN1: portion of the 100,000 generated face images by StyleGAN
- StyleGAN2: portion of the 100,000 generated face images by StyleGAN2
- Quintic AI: 15,076 generated face image: 8,505 by Stable Diffusion, 6,350 by Midjourney, 6,76 by DALL-E 2



Example of generated faces.

Methods

• Fully Connected Networks (Logistic Regression):

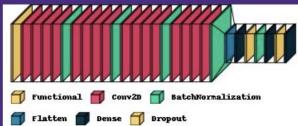
- Our baseline model consists of a single-layer Fully Connected (FC) network, which can be understood as a logistic regression model from a theoretical standpoint.

• Two Layer CNN:

- Another baseline model we have is a two-layer Conv Network. It consists of Conv-Conv-MaxPooling 2 . The resulting output is then flattened and passed through a FC layer of size 1000 and another FC layer.

• Residual Networks + CNN:

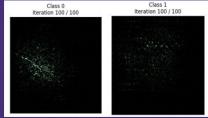
- The improved model incorporates a ResNet50 (pre-trained on ImageNet) on the top. It is followed by a sequence of Conv layers, specifically Conv-Cov-Cov-Cov-BatchNorm 4 . Subsequently, the output is flattened and passed through FC layers, with dropout and batch normalization applied in between. Finally, there is another FC layer with dropout, followed by a final FC layer. The model architecture is shown below (the scaling of the visualization may obscure the true complexity of a layer).



W

Analysis

- CNN features visualization
- Class Activation Maps

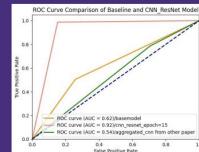


CNN feature visualization. (Class 0: Real; Class 1: Generated)



Result

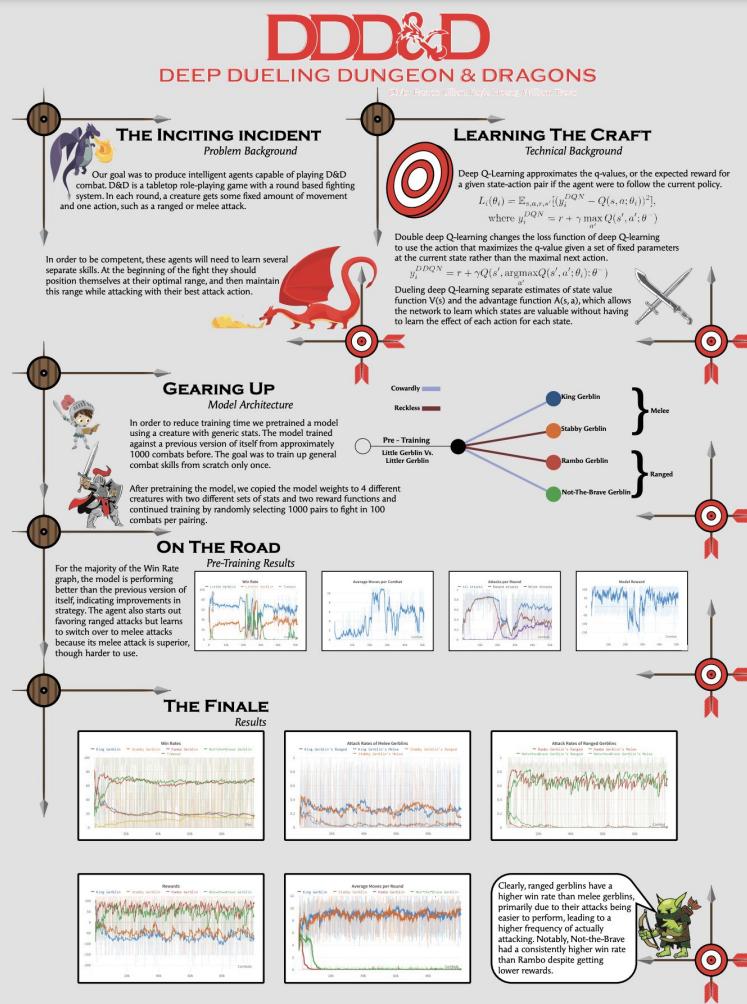
Model	Test Performance (AUCs)
Baseline(Ours)	0.62
Aggregated NNs	0.54
CNN-ResNet50(Ours)	0.92



Conclusion

Clearly, our fine-tuned CNN using the training data performed better than the other two methods in our study. However, while the aggregated CNN model from Mandelli et al.'s paper achieved remarkable accuracy (99%), it still failed to predict some test samples. This raises concerns about the robustness of these models, as they may eventually fail when faced with unseen synthetic images generated by unknown models.

Example final project



Example final project

Learning Codebooks of Discrete Representations for Embodied-AI

W PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

Department of STATISTICS

Ainaz Eftekhari and Peter Sushko

Object Navigation: Current Approaches & Limitations

- SLAMs: limited generalizability unless recalibrated for each new environment
- RNNs (LSTMs): fixed-size state vector leads to loss of information from early observations, optimization difficulties over long sequences
- Memory via maps: domain-based biases; requires estimate of agents position and costly annotated data
- Memory via transformers: limited efficiency and scalability especially for long episodes
- Our idea:** use an external/global memory (codebook)

EmbCLIP Framework (RNN-based)

- CLIP visual encoder + RNN (to provide memory)
- Agent's belief is the hidden state of the RNN

Memory-Augmented Architecture

- Global Codebook: 1024 memory cells, each entry is 512 dimensions
- Agent's belief is a convex combination memory cells indexed by the RNN

Codebook Collapse

Issue: Only a few entries receive non zero probability

Solutions:

- Random restarts (Dhariwal et al. 2022)
- Geometric approach (normalization) (Li et al. 2022)
- Linde-Buzo-Gray splitting algorithm (Dhariwal et al. 2022) ✓
 - Split the most frequent embedding into 2 and replace an unused one
- Code dropout (Dhariwal et al. 2022) ✓
 - Randomly drop 10% of the codebook entries

Experiment Design

ProcTHOR-10K dataset with 10,000 generated houses

- Trained using AllenAct framework
- PPO Algorithm
- 200M steps
- Batch sizes of 32,64,128
- Adam optimizer
- Learning rate decay
- Gradient clipping

AllenAct
A project for learning embodied intelligence

Comparison to SOTA

Validation results on ProcTHOR ObjectNav (1k houses)

Major Improvement upon all three baseline metrics

Method	Object Navigation		
	SR(%)	EL	SPL(%)
EmbCLIP (1k) (RNN-Based)	60.06	140.36	46.65
Global Memory (with codebook collapse)	48.0	99.18	33.65
Global Memory (Linde-Buzo-Gray splitting)	63.7	99.70	50.56
Global Memory (Code dropout)	66.5	83.28	51.81

Test Success by Object Type

What does the codebook encode?

Each cell encodes useful representation for object navigation such as information about semantics, affordances (possible actions), etc.

Memory spikes are sparse episode-wise and dense training-wise

Optional W Credit

- Undergrads can receive 4 “Additional Writing” credits towards their general education requirements
- Must complete 3 low-stakes writing assignments throughout the quarter
 - Should take a few hours each
- Must submit a longer final report a week early
 - Must then revise report and re-submit at normal deadline based on TA feedback
- More information [here](#)
 - Complete [interest form](#) by April 11th

Pre-requisites

Proficiency in Python

- All class assignments will be in Python (and use numpy)
- Later in the class, you will be using Pytorch and TensorFlow
- We will go over a Python tutorial on this Friday's recitation.

You need to know:

- **College Calculus,**
- **Linear Algebra,**
- **experience with Python**

No longer need Machine Learning as a prerequisite

Collaboration policy

Please follow [UW student code of conduct](#) – read it!

Here are our course specific rules:

- **Rule 1:** Don't look at solutions or code that are not your own; everything you submit should be your own work. We have automatic tools that detect plagiarism.
- **Rule 2:** Don't share your solution code with others; however discussing ideas or general strategies is fine and encouraged.
- **Rule 3:** Indicate in your submissions anyone you worked with.

Turning in something late / incomplete is better than violating the code

Plagiarism and Collaboration

We will run all assignments through plagiarism software.

Additionally, you may use online resources to understand concepts, but not to complete the coding portion of your assignments. This includes Stack Overflow and ChatGPT.

We will compare all student solutions to ChatGPT generated solutions. If we detect plagiarism in your assignments, you will get a 0 on the assignment and we will have no choice but to report to the university.

**** It is much better to turn in an incomplete assignment than to turn in code that is not your own! ****

Learning objectives

Formalize deep learning applications into tasks

- Formalize inputs and outputs for vision-related problems
- Understand what data and computational requirements you need to train a model

Develop and train deep learning models

- Learn to code, debug, and train convolutional neural networks.
- Learn how to use software frameworks like TensorFlow and PyTorch

Gain an understanding of where the field is and where it is headed

- What new research has come out in the last 0-9 years
- What are open research challenges?
- What ethical and societal considerations should we consider before deployment?

What you should expect from us

Fun: We will discuss fun applications like image captioning, GPT, generative AI



What we expect from you

Patience.

- Things will break; we will experience technical difficulties
- Bear with us and trust us to listen to you

Contribute

- Build a community with your peers
- Help one another - discuss topics you enjoy
- Give us anonymous feedback on EdStem

Why should you take this class?

Become a deep learning researcher (an incomplete list of conferences)

- Get involved with [research at UW](#): apply [using this form](#).

Conferences:

- [CVPR 2023](#), [ACL 2023](#), [NeurIPS 2023](#), [ICML 2023](#)

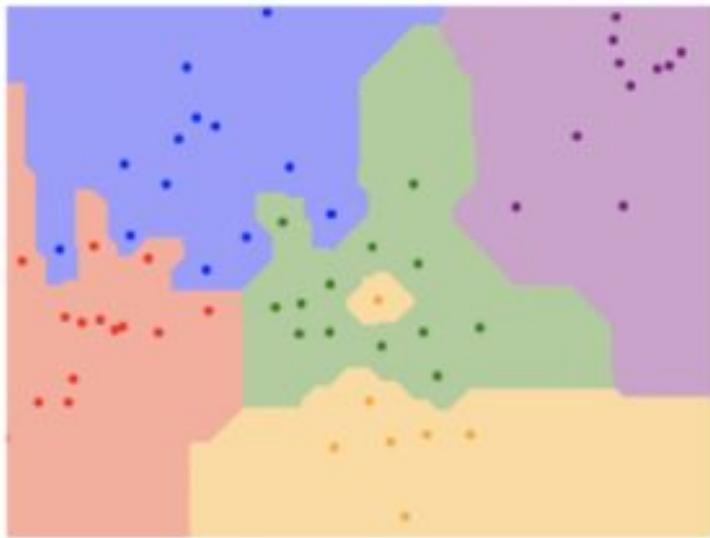
Become a deep learning engineer in industry (an incomplete list of industry teams)

- [Brain team at Google AI](#)
- [OpenAI](#)
- [Meta's Fundamental AI research team](#)
- [Microsoft's AI research team](#)

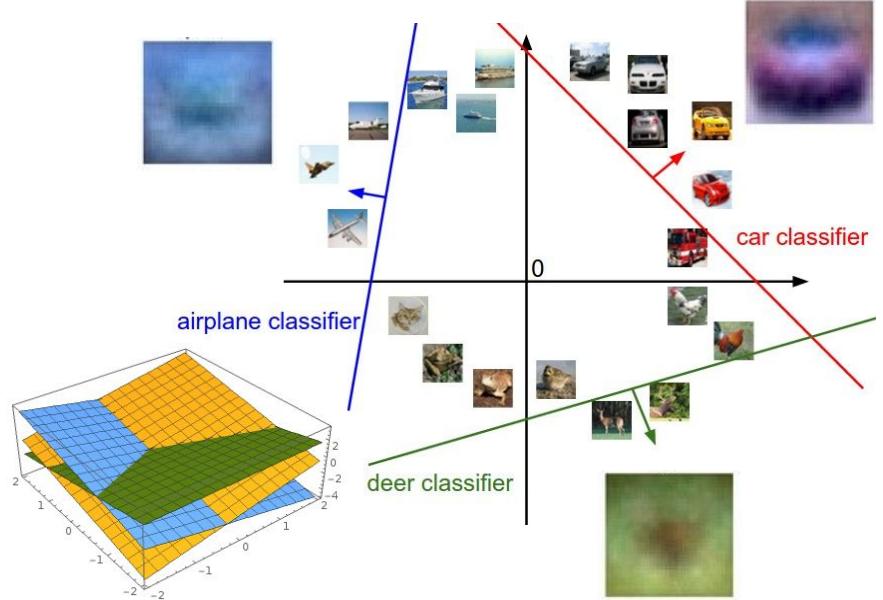
General interest

Next: Image classification

k- nearest neighbor



Linear classification



Plot created using [Wolfram Cloud](#)

References

- Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 1. IEEE, 2005. [\[PDF\]](#)
- Felzenszwalb, Pedro, David McAllester, and Deva Ramanan. "A discriminatively trained, multiscale, deformable part model." Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008 [\[PDF\]](#)
- Everingham, Mark, et al. "The pascal visual object classes (VOC) challenge." International Journal of Computer Vision 88.2 (2010): 303-338. [\[PDF\]](#)
- Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009. [\[PDF\]](#)
- Russakovsky, Olga, et al. "Imagenet Large Scale Visual Recognition Challenge." arXiv:1409.0575. [\[PDF\]](#)
- Lin, Yuanqing, et al. "Large-scale image classification: fast feature extraction and SVM training." Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011. [\[PDF\]](#)
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012. [\[PDF\]](#)
- Szegedy, Christian, et al. "Going deeper with convolutions." arXiv preprint arXiv:1409.4842 (2014). [\[PDF\]](#)
- Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014). [\[PDF\]](#)
- He, Kaiming, et al. "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition." arXiv preprint arXiv:1406.4729 (2014). [\[PDF\]](#)
- LeCun, Yann, et al. "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86.11 (1998): 2278-2324. [\[PDF\]](#)
- Fei-Fei, Li, et al. "What do we perceive in a glance of a real-world scene?." Journal of vision 7.1 (2007): 10. [\[PDF\]](#)