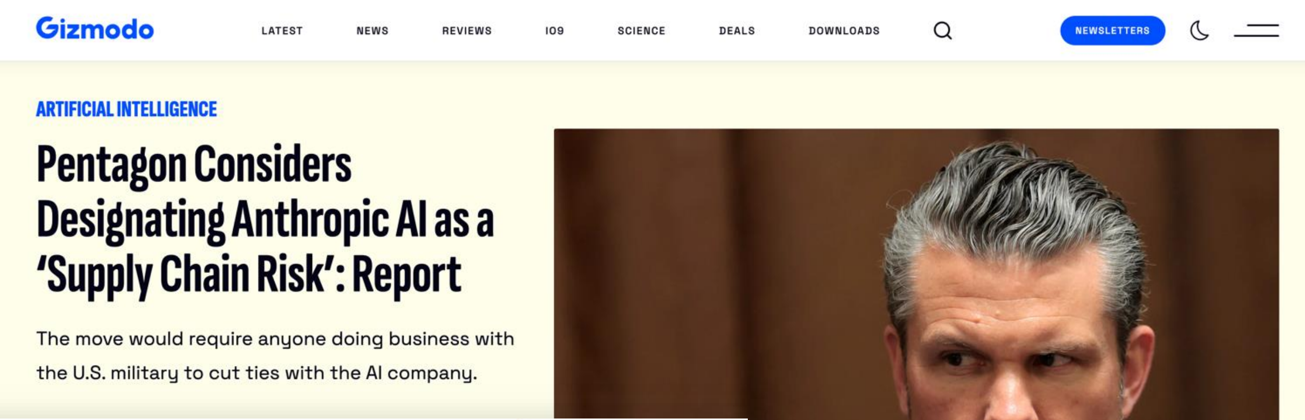
Computer Vision

© Gordon Burtch, 2026

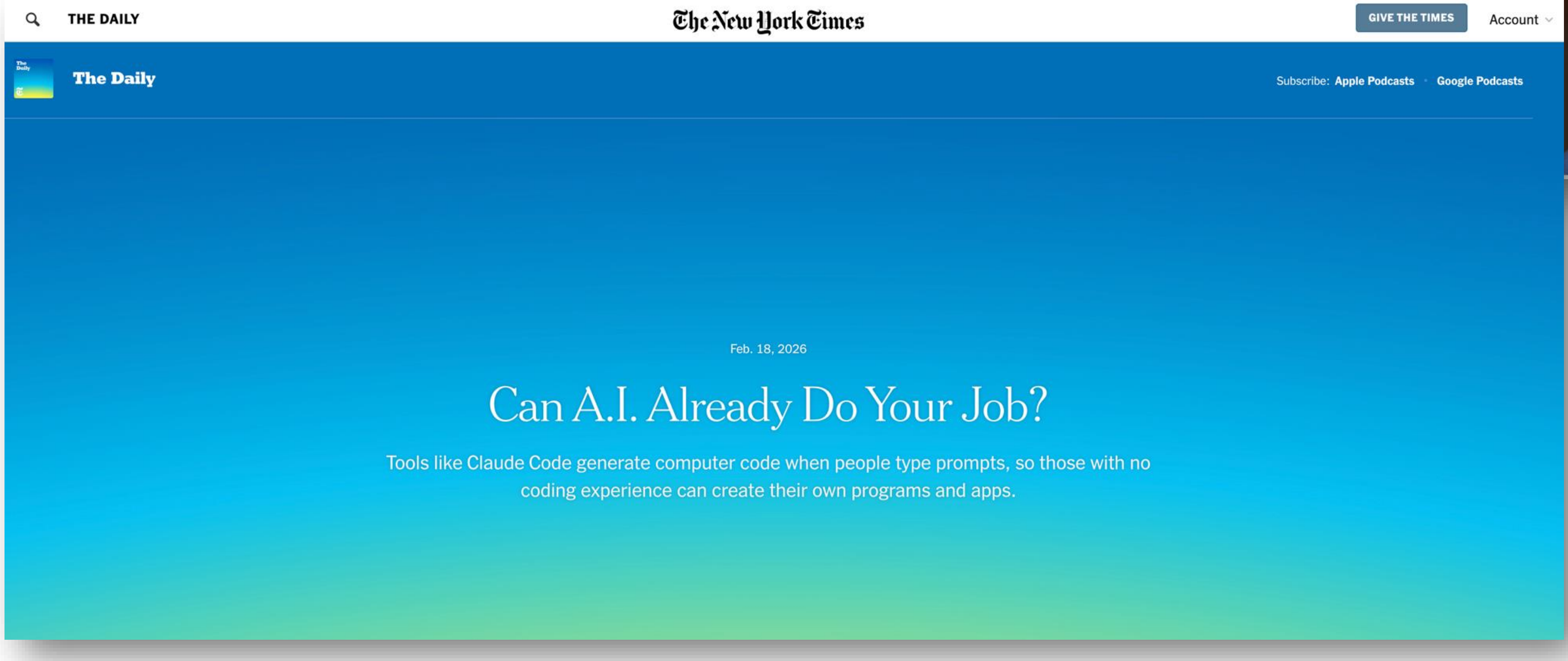
What's In the News?



ARTIFICIAL INTELLIGENCE

Pentagon Considers Designating Anthropic AI as a 'Supply Chain Risk': Report

The move would require anyone doing business with the U.S. military to cut ties with the AI company.



The Daily

Subscribe: Apple Podcasts · Google Podcasts

Feb. 18, 2026

Can A.I. Already Do Your Job?

Tools like Claude Code generate computer code when people type prompts, so those with no coding experience can create their own programs and apps.

This Week

Convolutional Neural Networks (CNNs)

- What CNNs try to accomplish
- What is a convolution?
    - Padding, strides, filters
- What is pooling?
    - Max, min, avg pooling.

Other Stuff

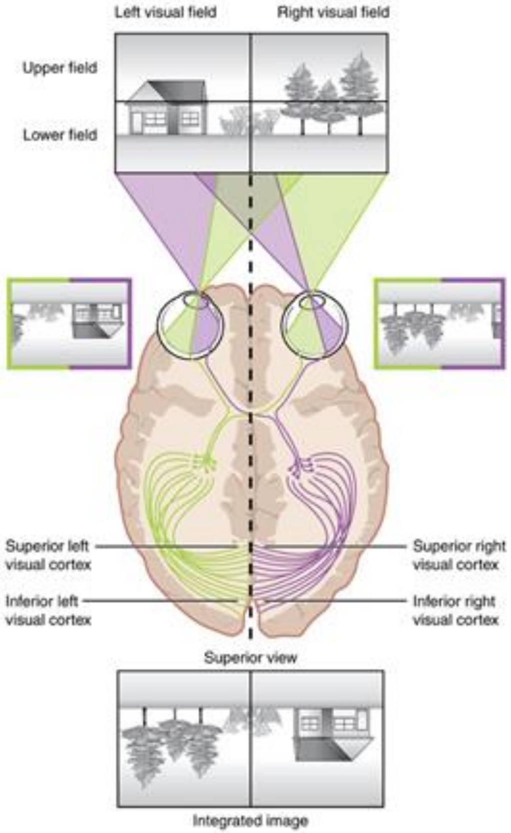
- CNN specific techniques to avoid overfitting (data augmentation).
- Extracting feature representations from your trained model.
- Adapting pre-trained models (transfer learning).

© Gordon Burtch, 2026

Inspiration for Convnets

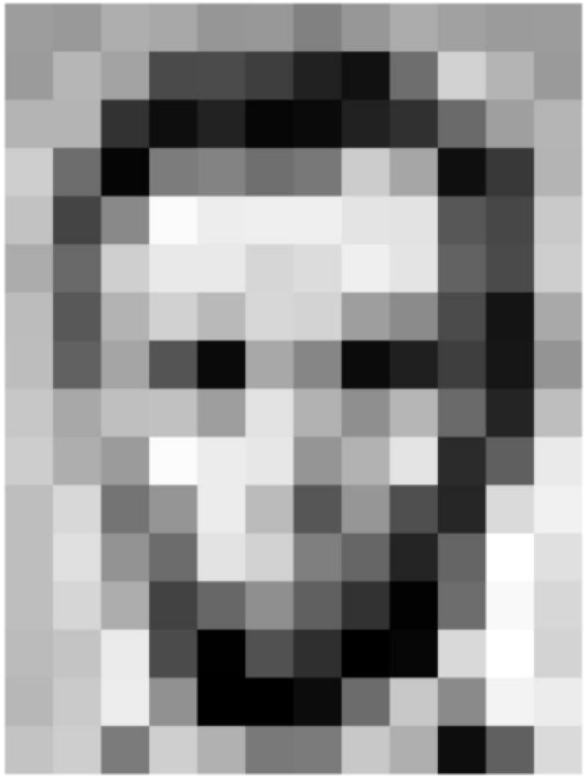
Our Visual System

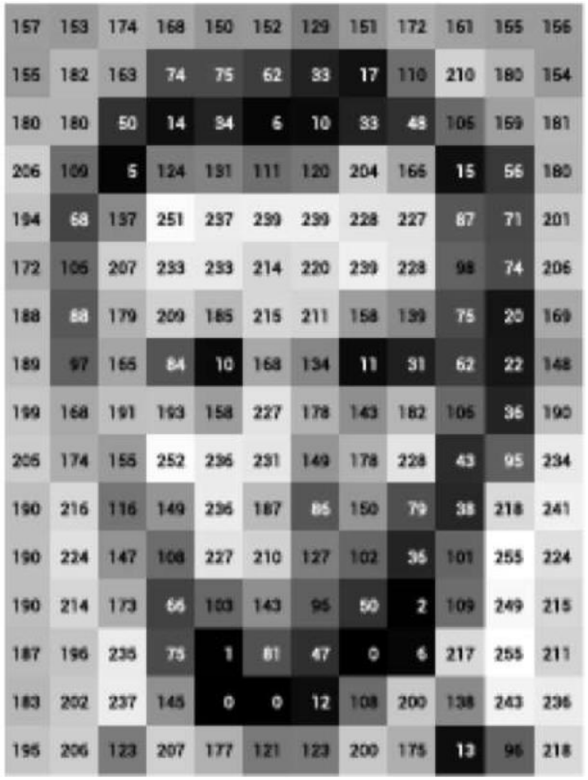
- Human eye is basically a 576-megapixel video camera.
- For comparison, the Pixel 6 camera is 50-megapixels.
- The human field of vision is not a square; something like a video camera that records individual image frames comprised of 24,000 x 24,000 pixels.


Left visual field
Right visual field
Upper field
Lower field
Superior left visual cortex
Inferior left visual cortex
Superior right visual cortex
Inferior right visual cortex
Superior view
Integrated image

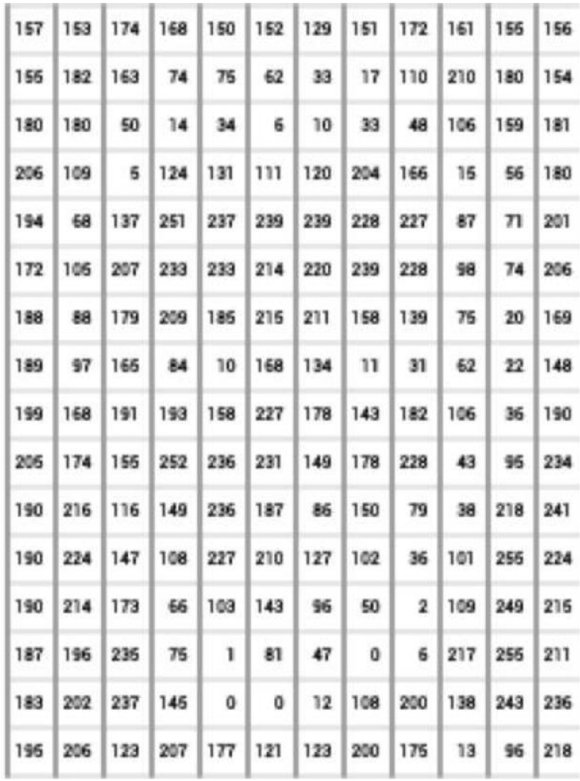
© Gordon Burtch, 2026

Images are Numbers





What the computer sees



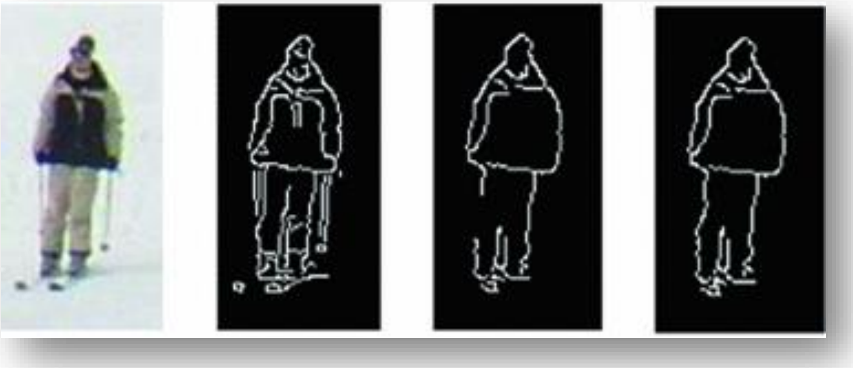
157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

An image is just a matrix of numbers [0,255]! i.e., 1080x1080x3 for an RGB image

We Don't Detect Features 'Manually'

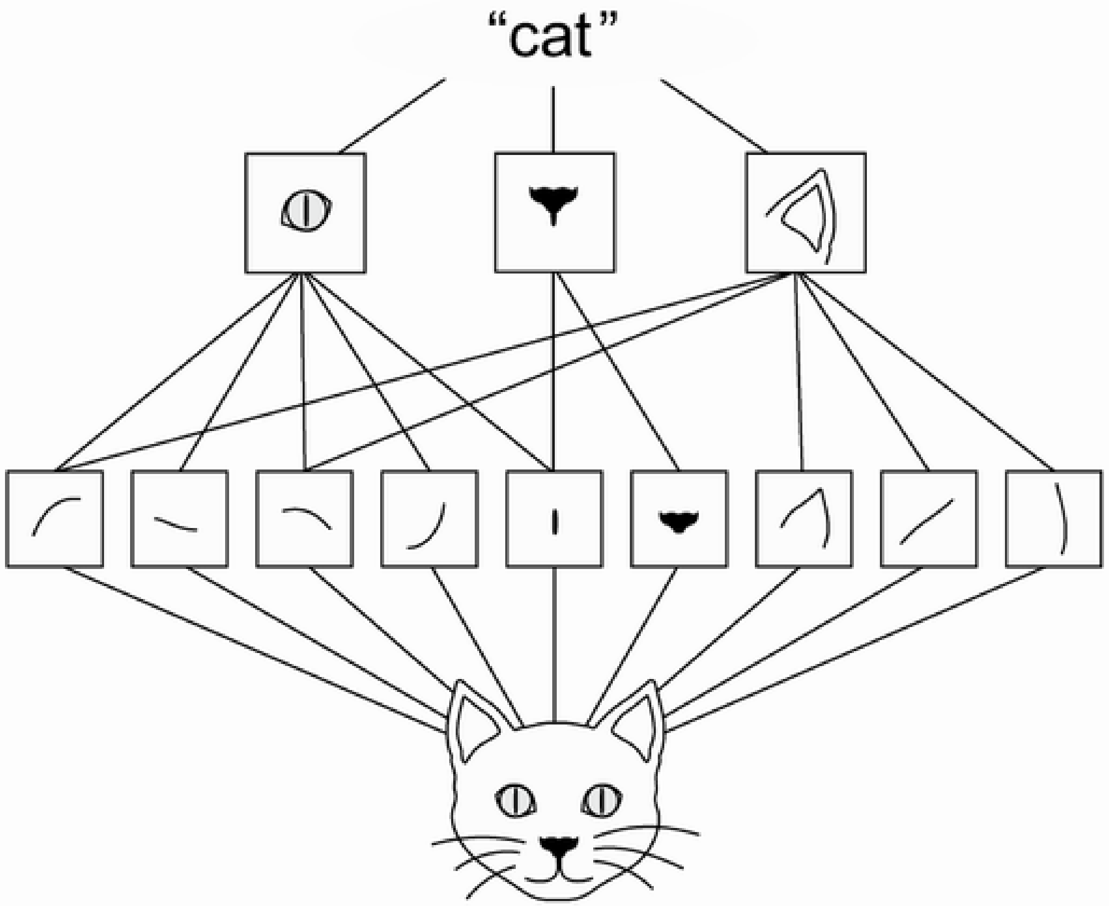
How Does Your Visual System Work?

- We think that your brain processes individual visual receptors in groups, identifies combinations of inputs in proximity to one another that imply something like an edge (edge detection), combines that with color and so on. These low-level features are then processed together to arrive at higher level objects (e.g., a nose, a mouth, an eye).
- Those higher-level features are then processed together to yield a face (perhaps someone we know or do not know). Hence why you might have a hard time recognizing someone who has a new haircut, or who is wearing a facemask!



© Gordon Burtch, 2026

Feature Detection / Aggregation

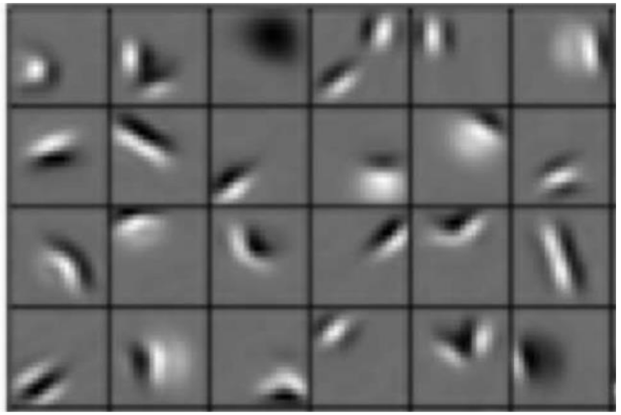


© Gordon Burtch, 2026

Learning Feature Representations

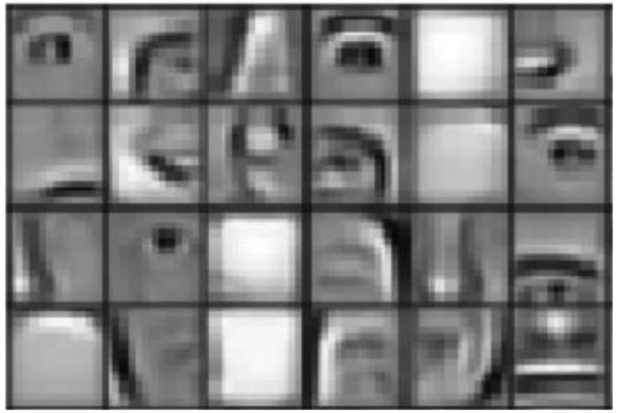
Can we learn a **hierarchy of features** directly from the data instead of hand engineering?

Low level features



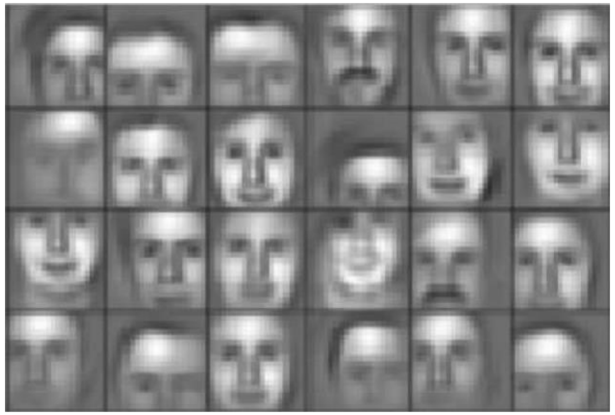
Edges, dark spots

Mid level features



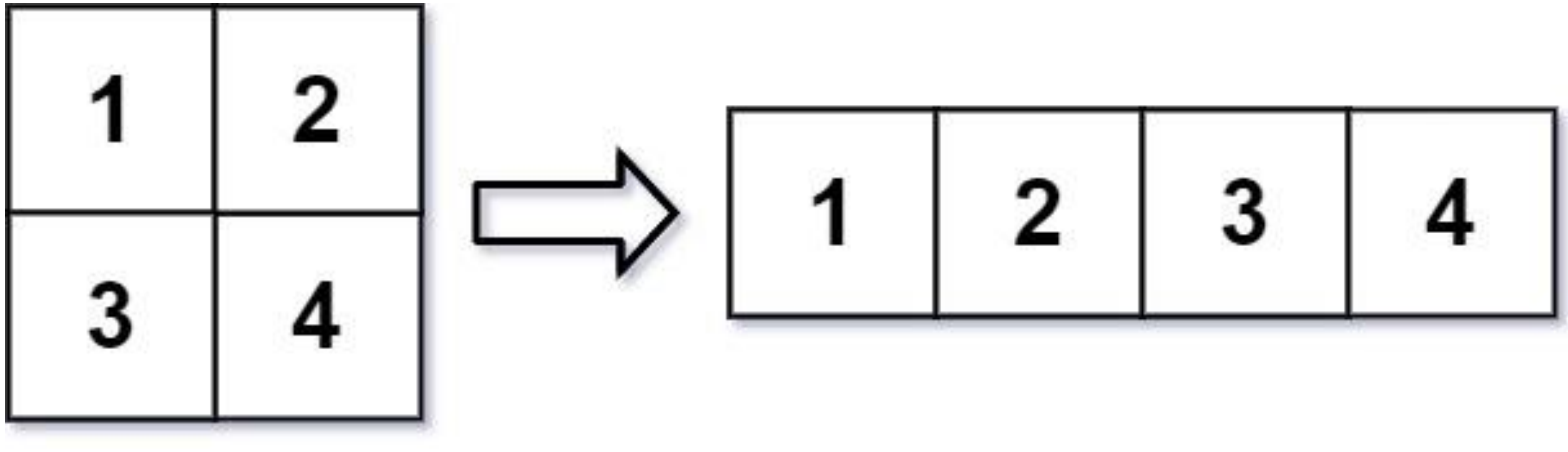
Eyes, ears,nose

High level features



Facial structure

Review: Why are Fully Connected Dense Layers (MLPs) Not Good for This?

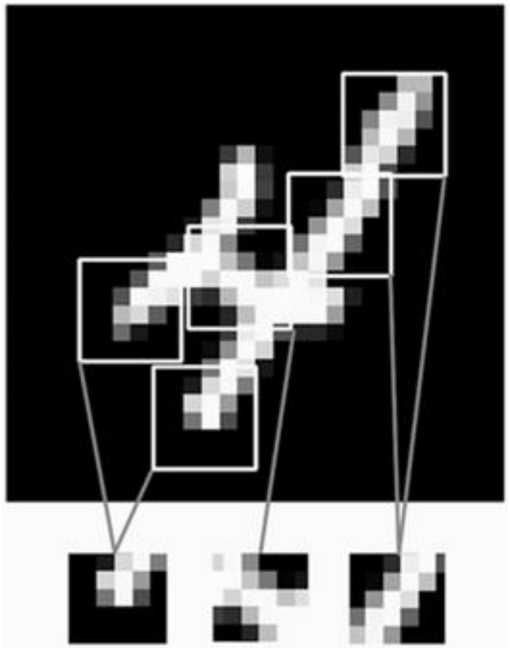


© Gordon Burtch, 2026

We Need to Preserve Spatial Arrangements

Hone-in On Sub-sections of the Image

- So, if we have a 28x28 image, we might separately consider 3x3 pixel subsection of that image. Each subsection (they can be overlapping) is represented by its own node in the first hidden layer.
- That local input matrix (subfield) is considered in tandem with a 'filter' a matrix of weights. A filter might be something like

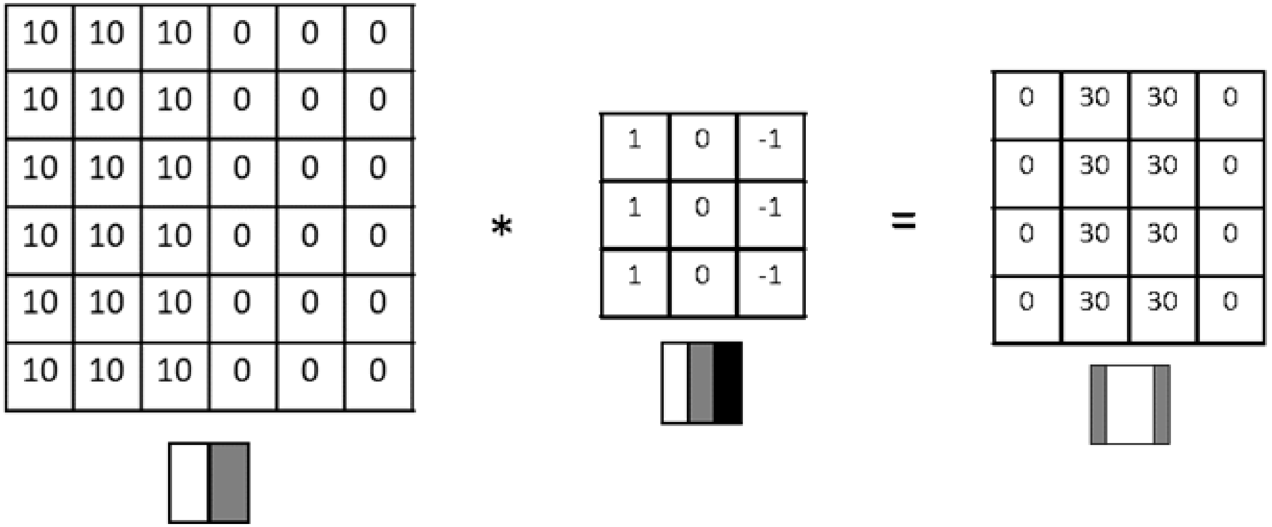


© Gordon Burtch, 2026

The Convolution Operation

Consider in Matrix Representation

- We have the raw image data (a.k.a. input feature map), the filter, and the result of passing our filter over our image (a.k.a. output feature map). We will have one output feature map for a given image, per filter (each filter is intended to detect a different type of feature).
- The filter elements are just weights for the Conv layer; we learn the filter values as part of the backpropagation process. So, the CNN will figure out what features to look for to predict the label (probably what a baby does when its first board and first learning how to process visual information).



$$\begin{bmatrix} 10 & 10 & 10 & 0 & 0 & 0 \\ 10 & 10 & 10 & 0 & 0 & 0 \\ 10 & 10 & 10 & 0 & 0 & 0 \\ 10 & 10 & 10 & 0 & 0 & 0 \\ 10 & 10 & 10 & 0 & 0 & 0 \\ 10 & 10 & 10 & 0 & 0 & 0 \end{bmatrix} * \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 30 & 30 & 0 \\ 0 & 30 & 30 & 0 \\ 0 & 30 & 30 & 0 \\ 0 & 30 & 30 & 0 \end{bmatrix}$$

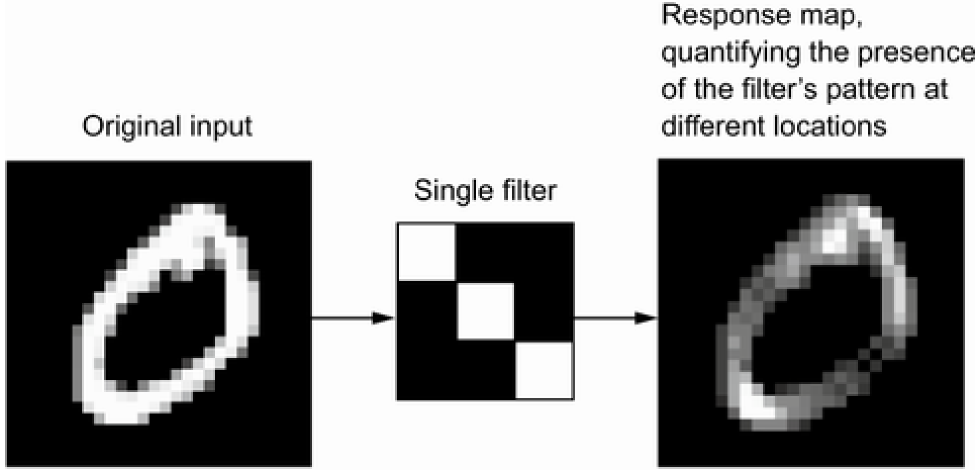
© Gordon Burtch, 2026

The Convolution Operation

Consider in Matrix Representation

- We have the raw image data (a.k.a. input feature map), the filter, and the result of passing our filter over our image (a.k.a. output feature map). We will have one output feature map for a given image, per filter (each filter is intended to detect a different type of feature).

- The filter elements are just weights for the Conv layer; we learn the filter values as part of the backpropagation process. So, the CNN will figure out what features to look for to predict the label (probably what a baby does when its first board and first learning how to process visual information).

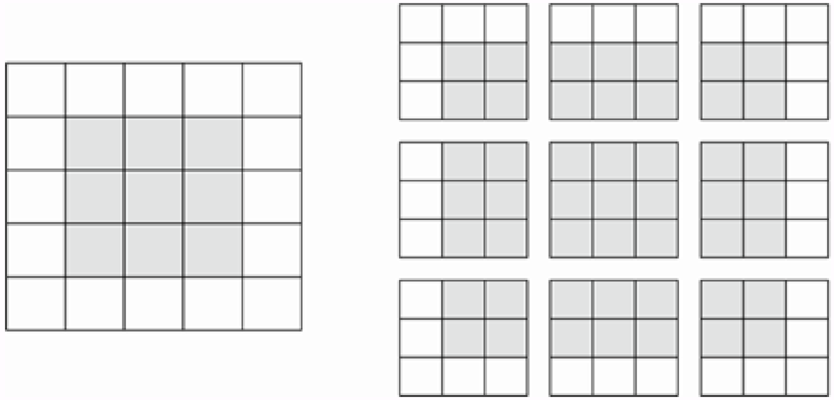

Original input
Single filter
Response map, quantifying the presence of the filter's pattern at different locations

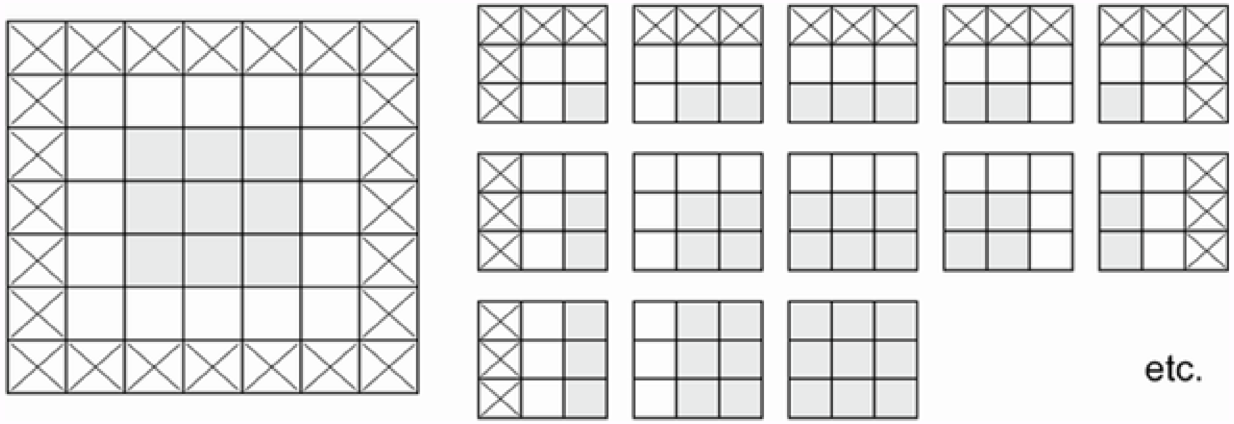
© Gordon Burtch, 2026

Padding

Padding

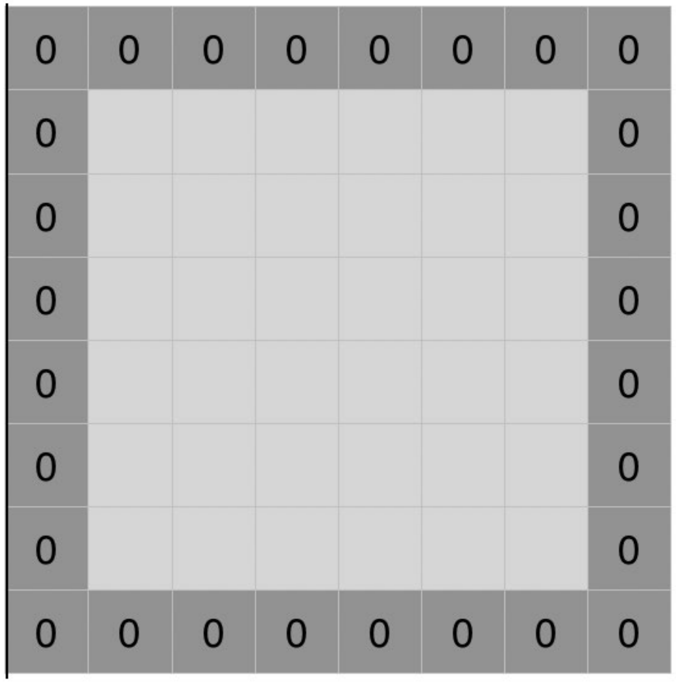
- To prevent the transformation from down-sampling (reducing the size of the matrix during convolution to output), we can pad the edges of the image with 0’s.



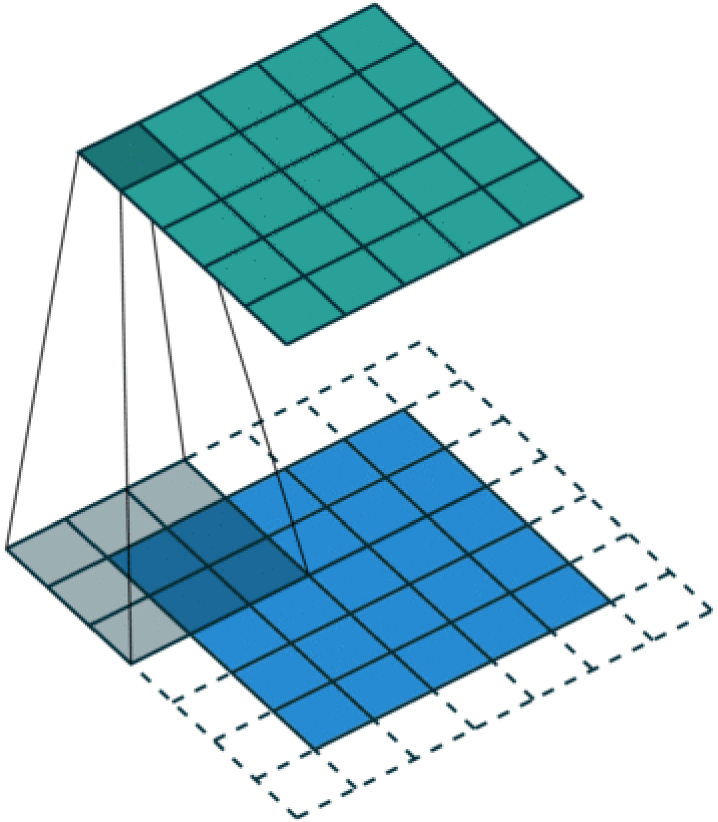


© Gordon Burtch, 2026

Padding Technique



Zero-padding added to image



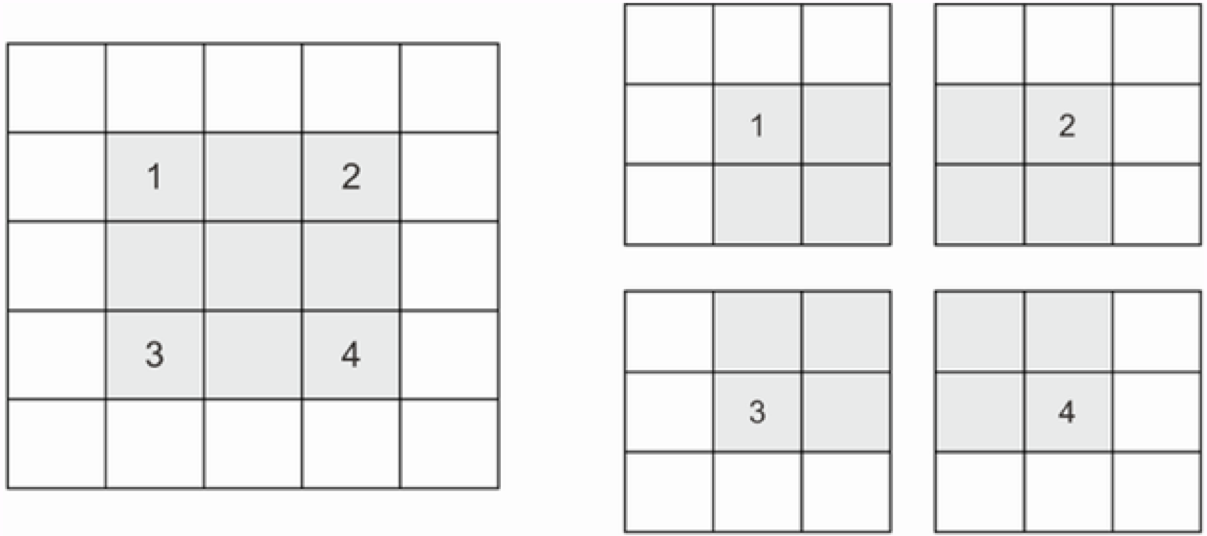
Add appropriate padding to keep the dimension after filtering

$m\times m$input image with $ff\times ff$filters
add $pp= (ff-1)/2$ padding keep the same dimension of input image

Strides

Strides

- Often, we will pass the filter over every pixel cell, but we don’t have to; we might pass over every other cell. This is what strides refers to (skipping).



© Gordon Burtch, 2026

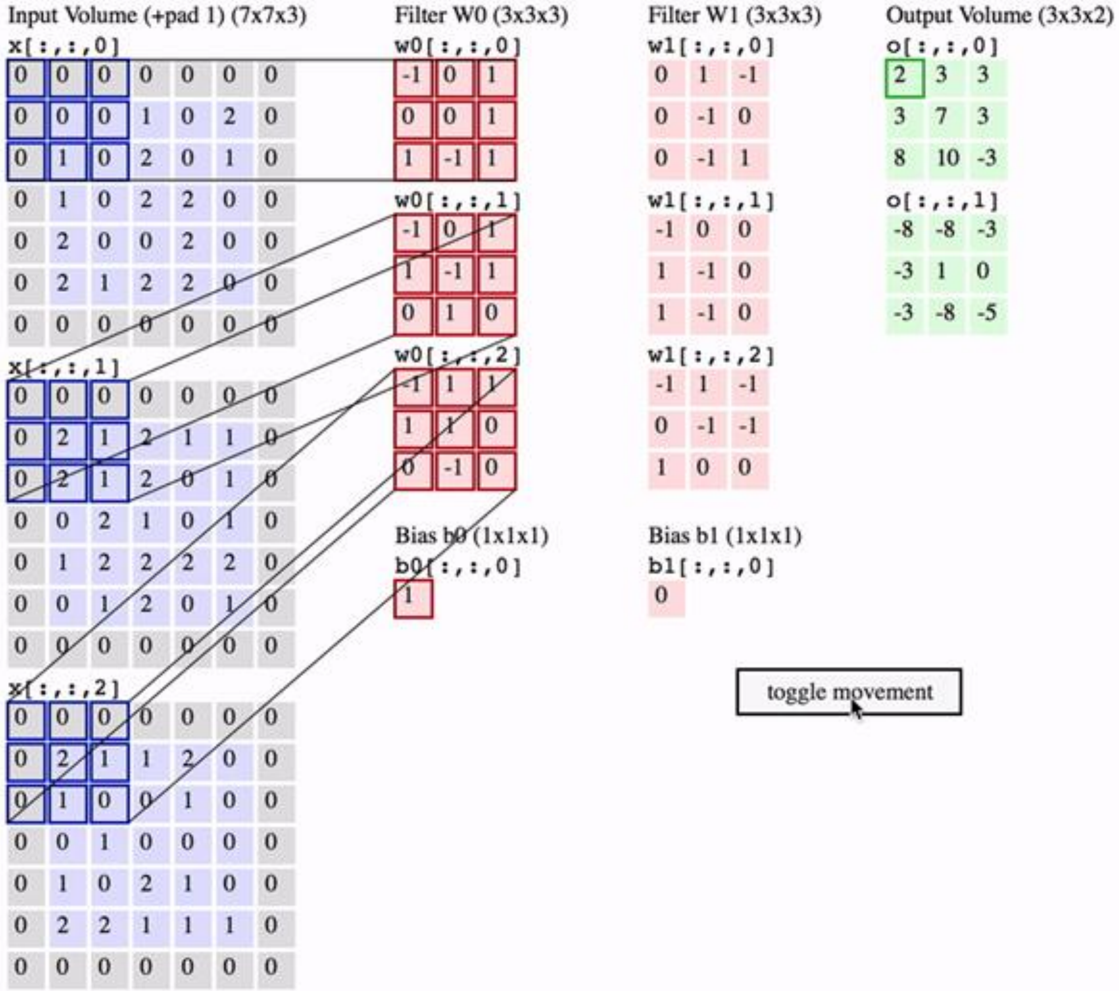
Padding and Strides

Padding

- To prevent the transformation from down-sampling (reducing the size of the matrix during convolution to output), we can pad the edges of the image with 0's.

Strides

- Often, we will pass the filter over every pixel cell, but we don't have to; we might pass over every other cell.
- This is what strides refers to (skipping).



Input Volume (+pad 1) (7x7x3)

x[:,:,0]

0	0	0	0	0	0	0
0	0	0	1	0	2	0
0	1	0	2	0	1	0
0	1	0	2	2	0	0
0	2	0	0	2	0	0
0	2	1	2	2	0	0
0	0	0	0	0	0	0

x[:,:,1]

0	0	0	0	0	0	0
0	2	1	2	1	1	0
0	2	1	2	0	1	0
0	0	2	1	0	1	0
0	1	2	2	2	2	0
0	0	1	2	0	1	0
0	0	0	0	0	0	0

x[:,:,2]

0	0	0	0	0	0	0
0	2	1	1	2	0	0
0	1	0	0	1	0	0
0	0	1	0	0	0	0
0	1	0	2	1	0	0
0	2	2	1	1	1	0
0	0	0	0	0	0	0

Filter W0 (3x3x3)

w0[:,:,0]

-1	0	1
0	0	1
1	-1	1

w0[:,:,1]

-1	0	1
1	-1	1
0	1	0

w0[:,:,2]

-1	1	1
1	1	0
0	-1	0

Bias b0 (1x1x1)

b0[:,:,0]

1

Filter W1 (3x3x3)

w1[:,:,0]

0	1	-1
0	-1	0
0	-1	1

w1[:,:,1]

-1	0	0
1	-1	0
1	-1	0

w1[:,:,2]

-1	1	-1
0	-1	-1
1	0	0

Bias b1 (1x1x1)

b1[:,:,0]

0

Output Volume (3x3x2)

o[:,:,0]

2	3	3
3	7	3
8	10	-3

o[:,:,1]

-8	-8	-3
-3	1	0
-3	-8	-5

© Gordon Burtch, 2026

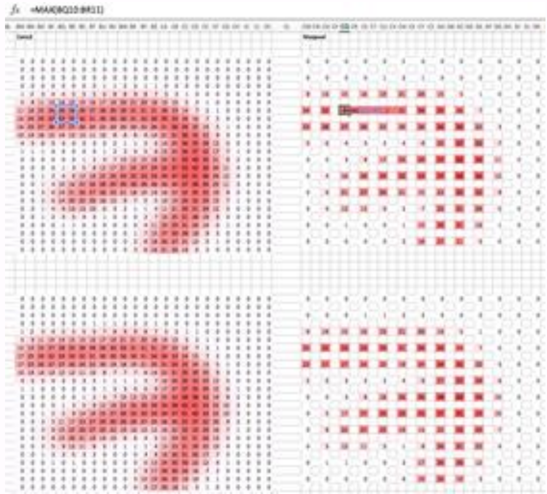
What is Pooling?

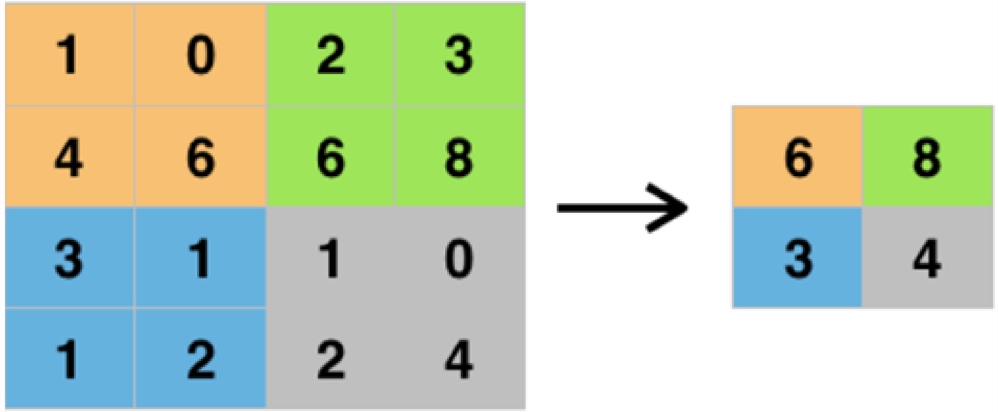
Down-sampling Detected Features

- The idea is to compress the resulting data down into a coarser representation, to reduce model complexity, and to also force attention toward a broader section of the original image (helps reduce overfitting).

Forcing Attention to Larger Blocks of the Original Image

- Because we typically use stride = pool width, the pooling output is aggregating over segments of the input.





1	0	2	3
4	6	6	8
3	1	1	0
1	2	2	4

→

6	8
3	4

© Gordon Burtch, 2026

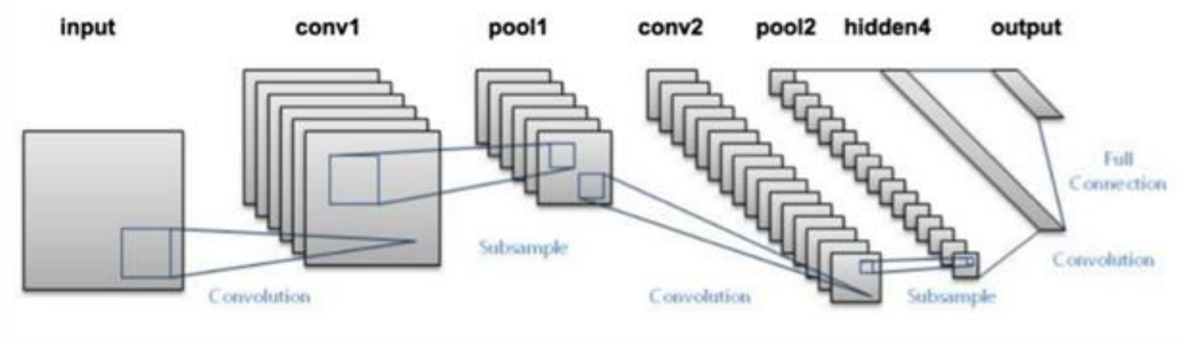
A Basic Image Labeling Architecture

Convolution Layers Apply (Multiple) Filters

- Each filter is a matrix of trainable weights.
- As you move through the network, the number of features typically rises exponentially.
- More filters as you move along means it allows more permutations / combinations

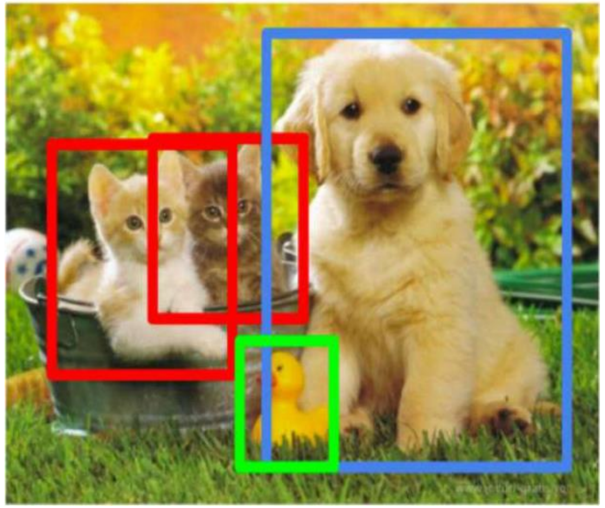
Progressively Smaller Filter Maps

- Smaller filter map arises from the pooling steps, which means that each element of the final map distills features (high level features, derived from low level features, derived from raw pixels) derived from a larger segment of the original picture.

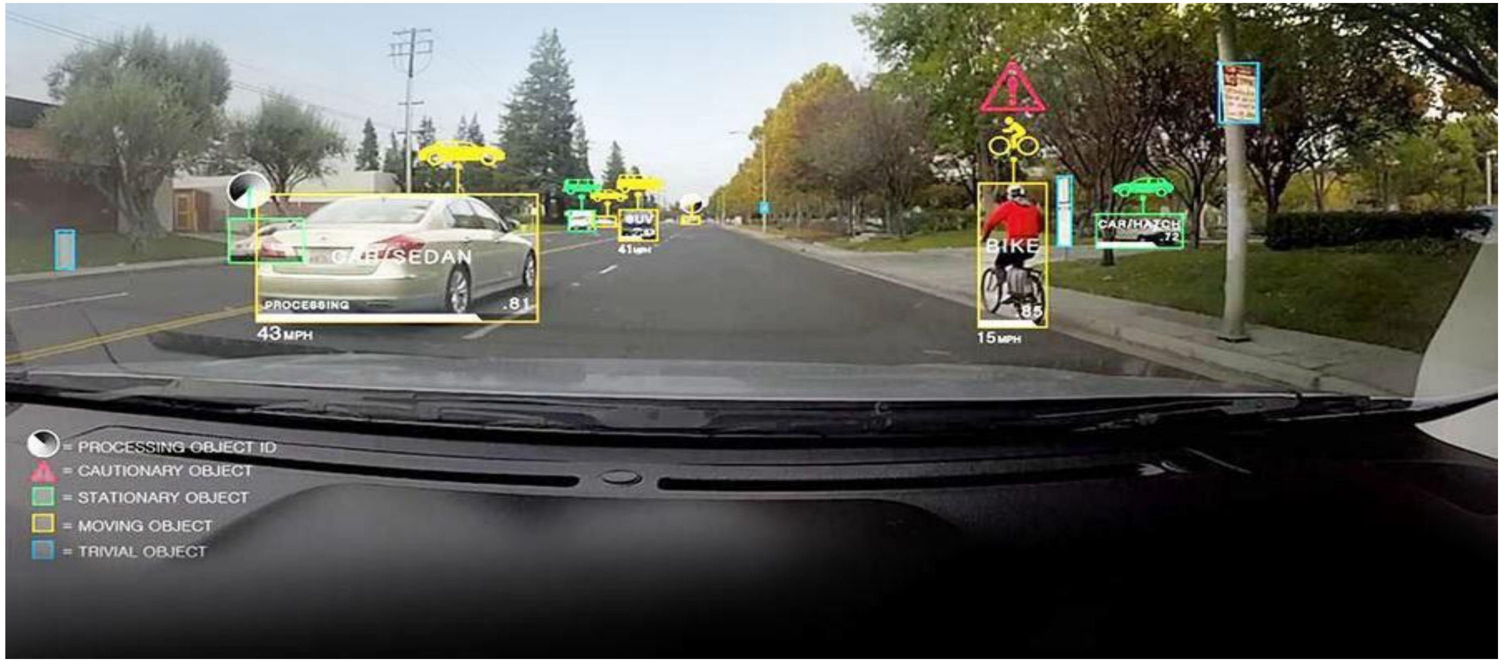


© Gordon Burtch, 2026

Object Recognition Tasks



CAT, DOG, DUCK

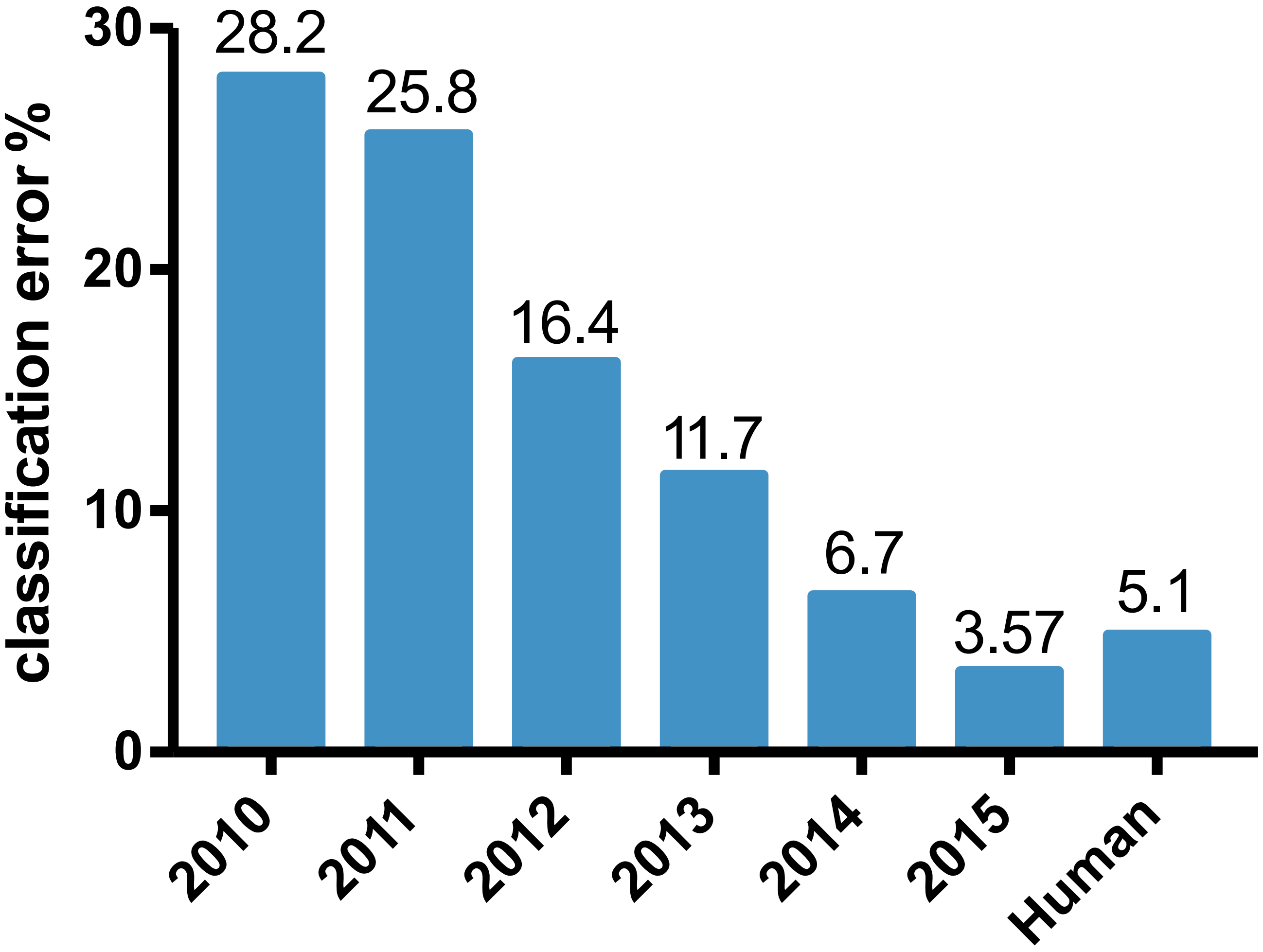


© Gordon Burtch, 2026

ImageNet Challenge

Classification task: "Top 5 error among 1000 categories": rate at which the model does not output correct label in top 5 predictions

ImageNet Challenge: Classification Task



2012:AlexNet. First CNN to win.

- 8 layers, 61 million parameters

2013:ZFNet

- 8 layers, more filters

2014:VGG

- 19 layers

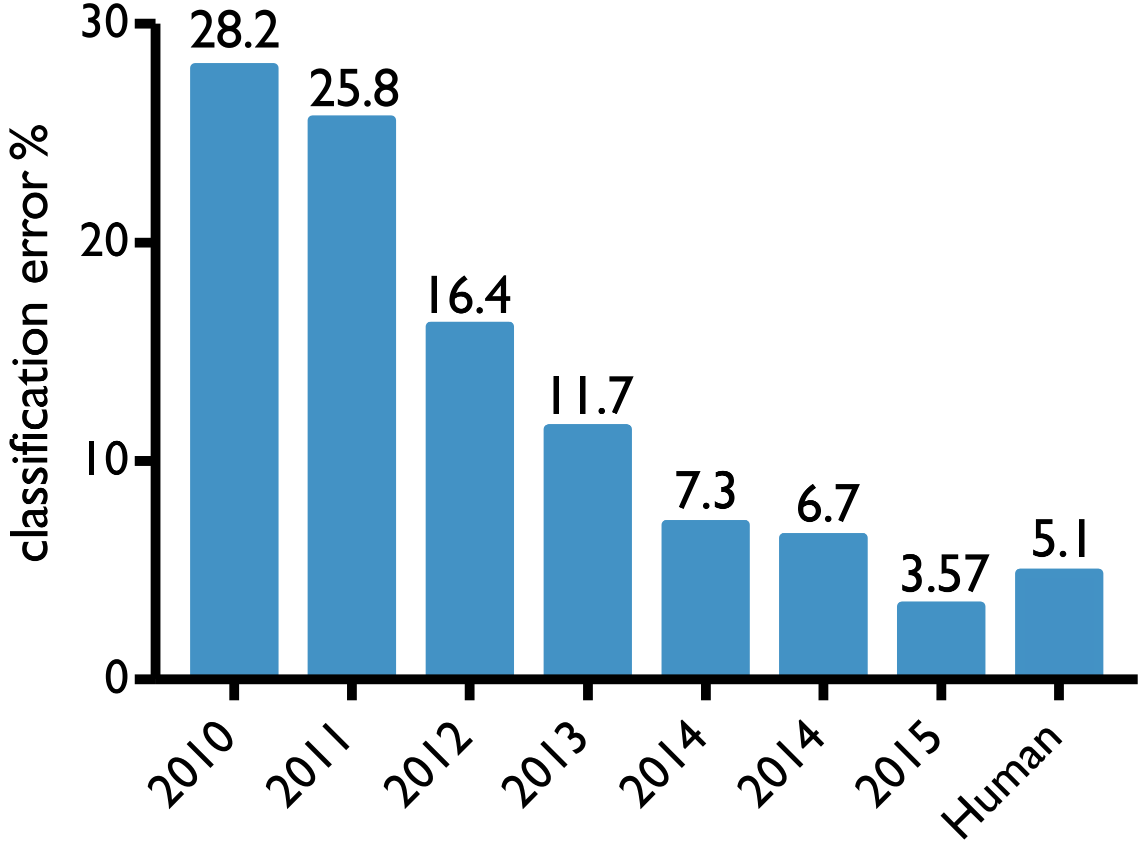
2014:GoogLeNet

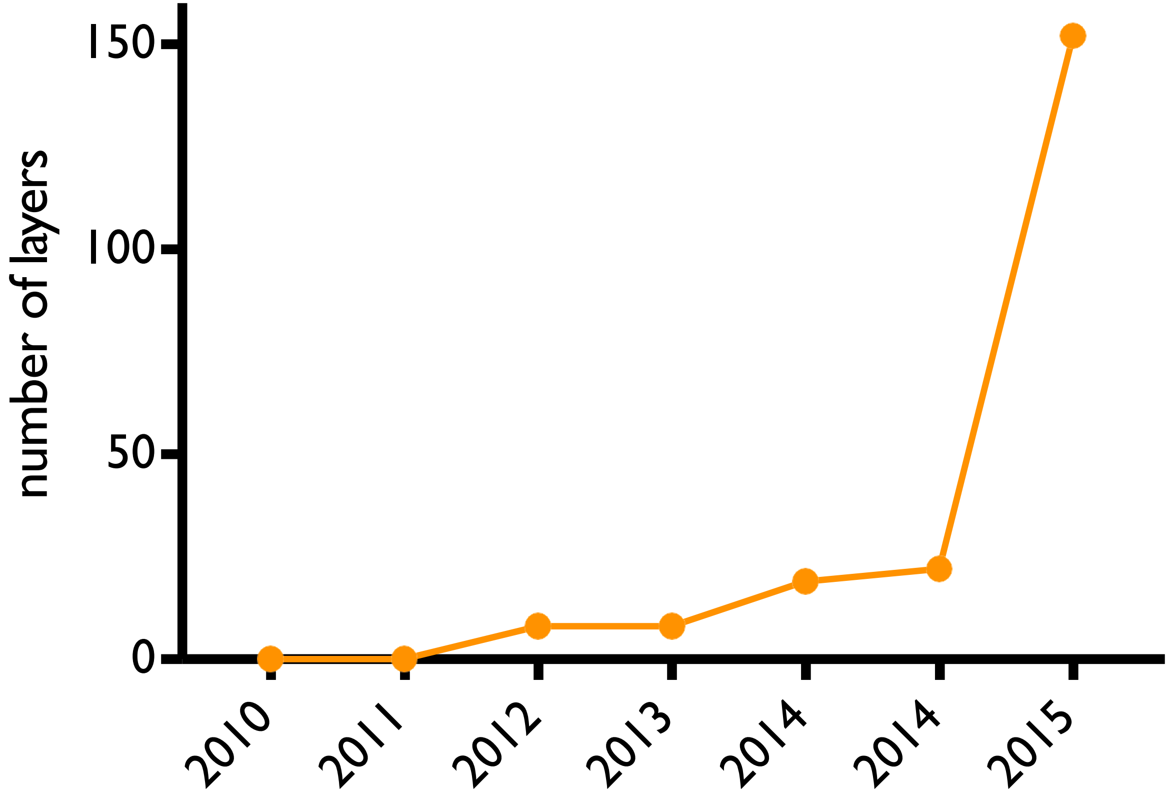
- "Inception" modules
- 22 layers, 5million parameters

2015:ResNet

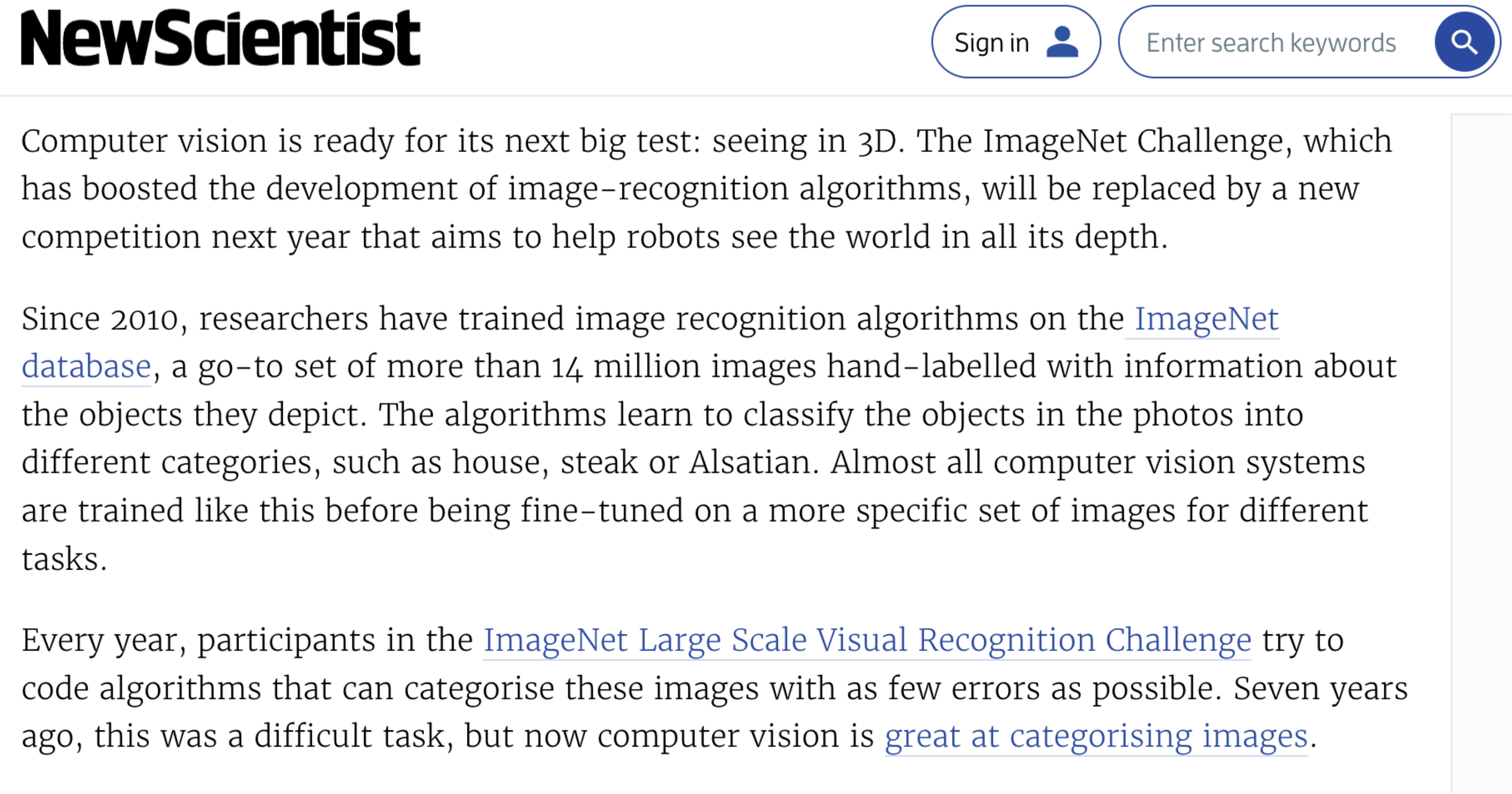
- 152 layers

ImageNet Challenge: Classification Task





ImageNet Challenge Was Ended After 2017



Computer vision is ready for its next big test: seeing in 3D. The ImageNet Challenge, which has boosted the development of image-recognition algorithms, will be replaced by a new competition next year that aims to help robots see the world in all its depth.

Since 2010, researchers have trained image recognition algorithms on the ImageNet database, a go-to set of more than 14 million images hand-labelled with information about the objects they depict. The algorithms learn to classify the objects in the photos into different categories, such as house, steak or Alsatian. Almost all computer vision systems are trained like this before being fine-tuned on a more specific set of images for different tasks.

Every year, participants in the ImageNet Large Scale Visual Recognition Challenge try to code algorithms that can categorise these images with as few errors as possible. Seven years ago, this was a difficult task, but now computer vision is great at categorising images.

Shifted Focus to Harder Problems

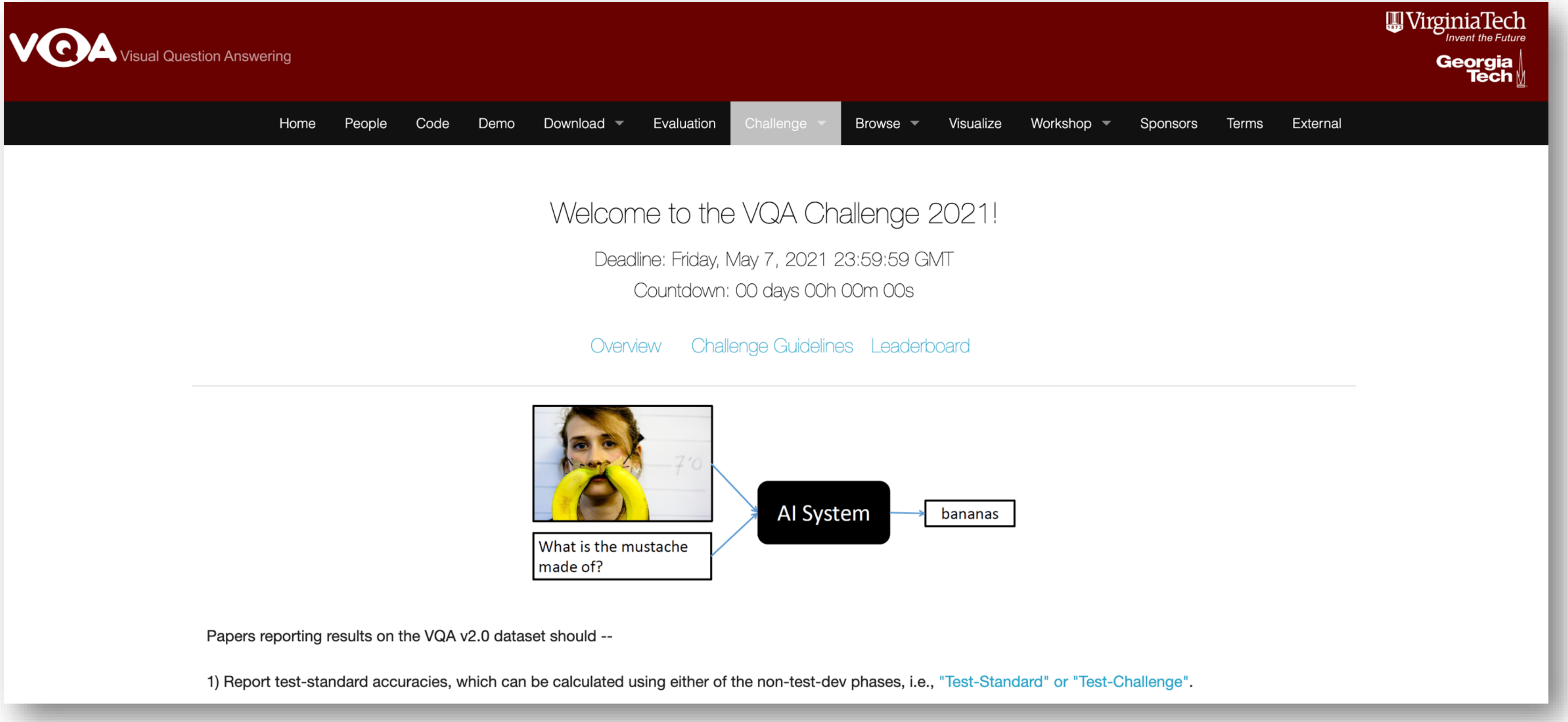
ACTIVITYNET
Large Scale Activity Recognition Challenge
HOME
PEOPLE
CHALLENGE
PROGRAM
DATES
EVALUATION
CONTACT
CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA
Travel Information

International Challenge on Activity Recognition (ActivityNet)

CVPR 2022 Workshop

© Gordon Burtch, 2026

Shifted Focus to Harder Problems



Welcome to the VQA Challenge 2021!

Deadline: Friday, May 7, 2021 23:59:59 GMT

Countdown: 00 days 00h 00m 00s

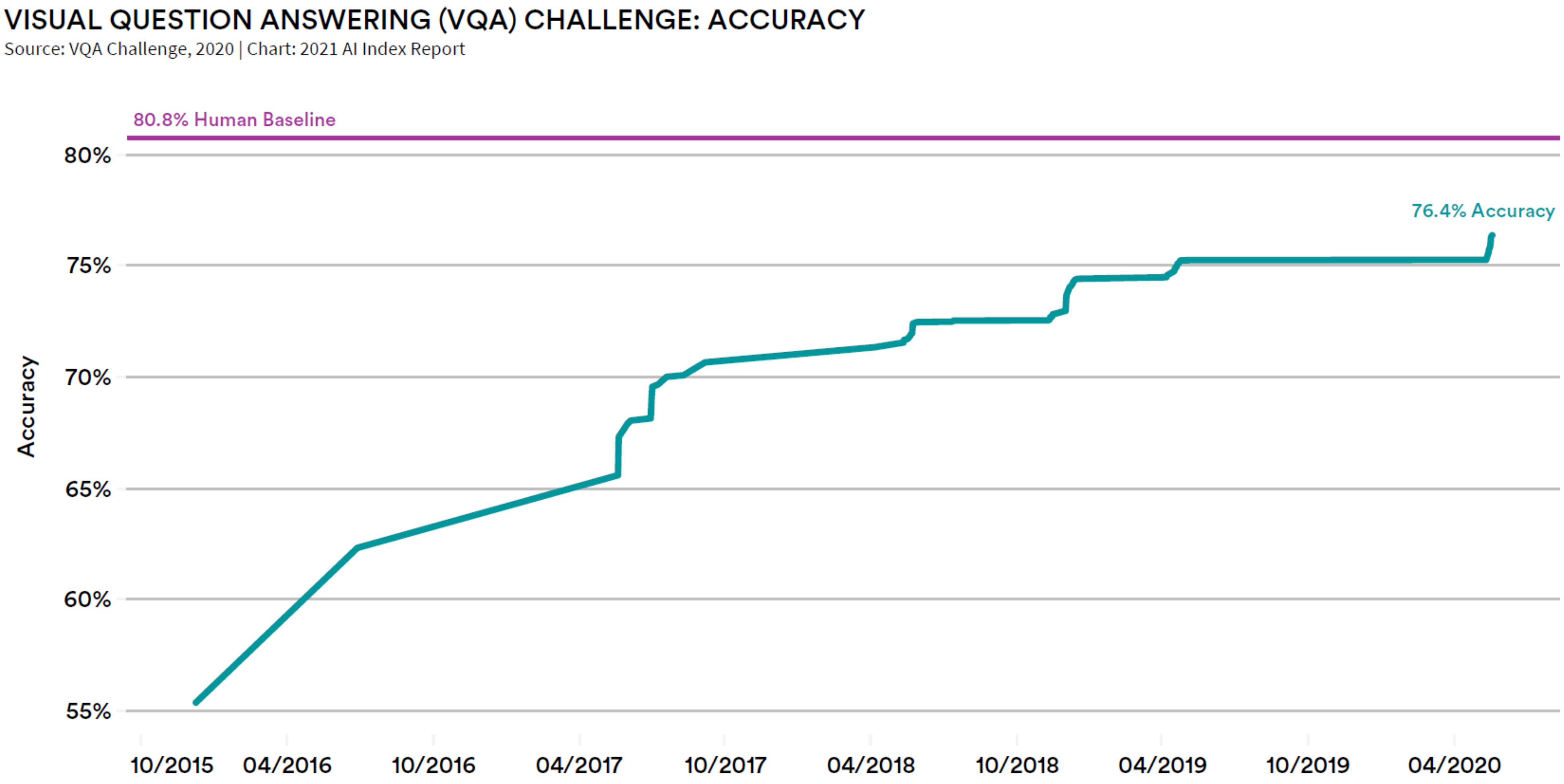
Overview Challenge Guidelines Leaderboard

Papers reporting results on the VQA v2.0 dataset should --

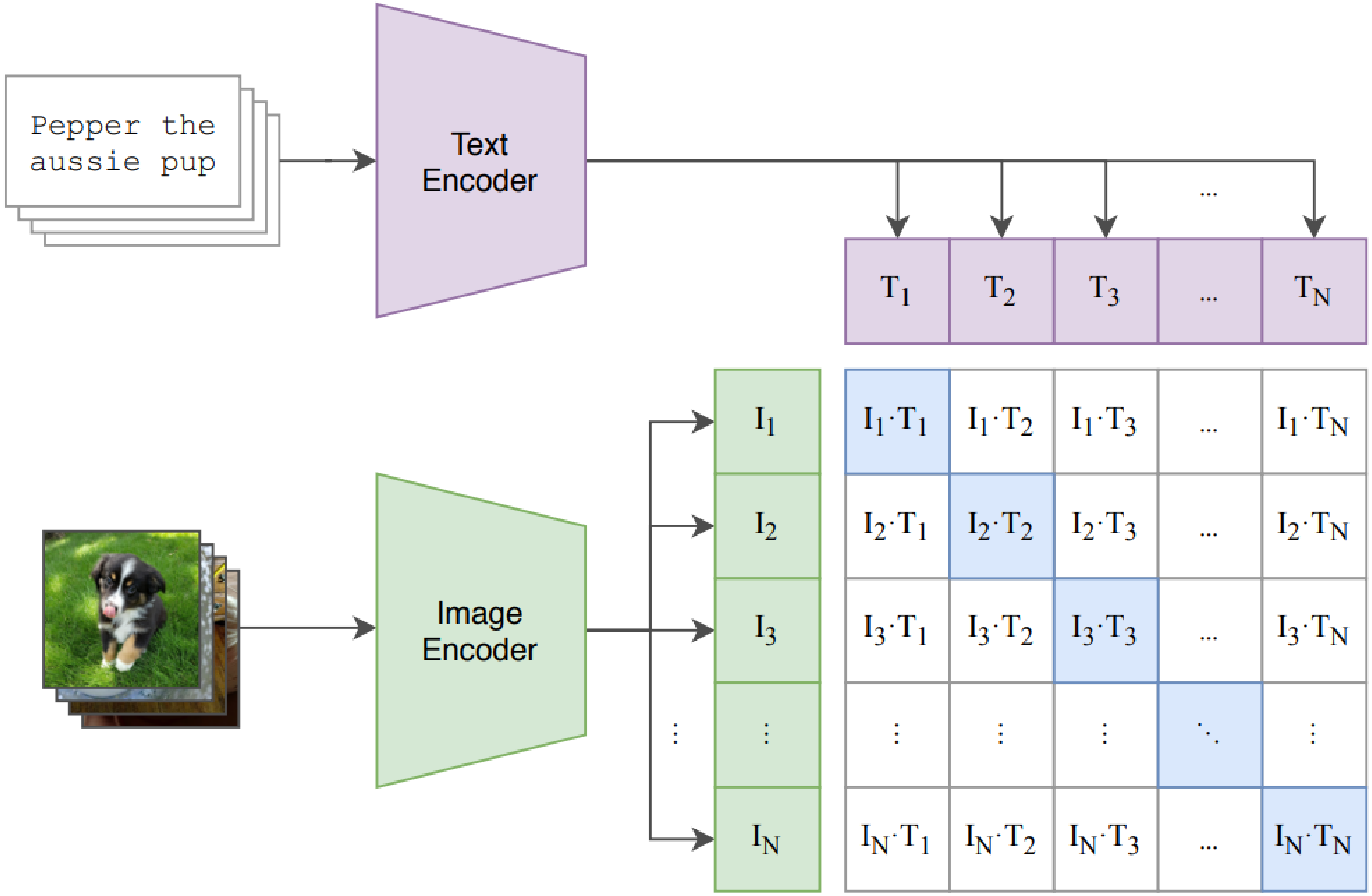
1) Report test-standard accuracies, which can be calculated using either of the non-test-dev phases, i.e., "Test-Standard" or "Test-Challenge".

© Gordon Burtch, 2026

Visual Question and Answer (VQA) challenge



Contrastive Language-Image Pre-training (CLIP)



	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

© Gordon Burtch, 2026

Contrastive Loss

The goal here is *not* prediction. The goal is to train a model that can encode inputs into numeric vector representations in a shared embedding space, such that ‘similar’ vectors are conceptually related, and ‘dissimilar’ vectors are conceptually unrelated.

We have a special loss function that we minimize for this purpose, where it is once again a 2-part loss, depending on whether the ‘pair’ of inputs is positive (related) or negative (unrelated).

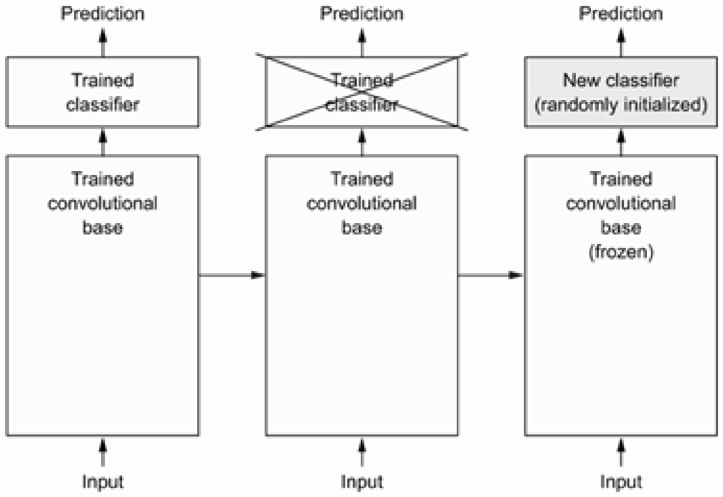
$$L = y \cdot D^2 + (1 - y) \cdot \max(0, m - D)^2$$

Pre-Trained Models: Feature Extraction

Take the convolutional base layers from someone else's model, then…

Two Options

- *Feed Data Through Model Base:* feed your images through convolutional base, take the outputs, and then use those as your predictors, feeding them into a network of dense layers.
- *Freeze Model Base and Include in Network:* Take the convolutional base layers from someone else's model and freeze them (make parameters non-trainable), then stack your (trainable) Dense layers onto the end. This lets you add data-augmentation to the front of the model.


Prediction
Trained classifier
Trained convolutional base
Input
Prediction
~~Trained classifier~~
Trained convolutional base
Input
Prediction
New classifier (randomly initialized)
Trained convolutional base (frozen)
Input

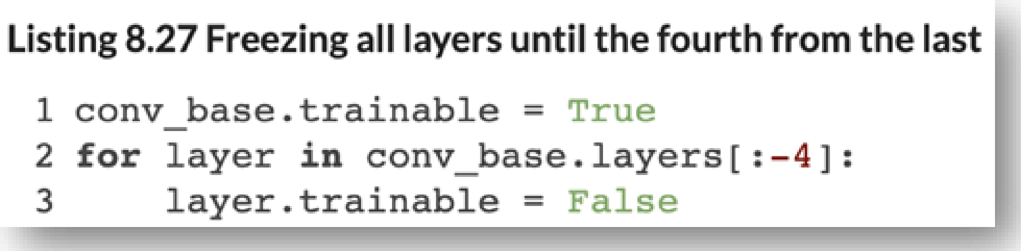
© Gordon Burtch, 2026

Pre-Trained Models: Fine Tuning

Take the convolutional base layers from someone else’s model, then…

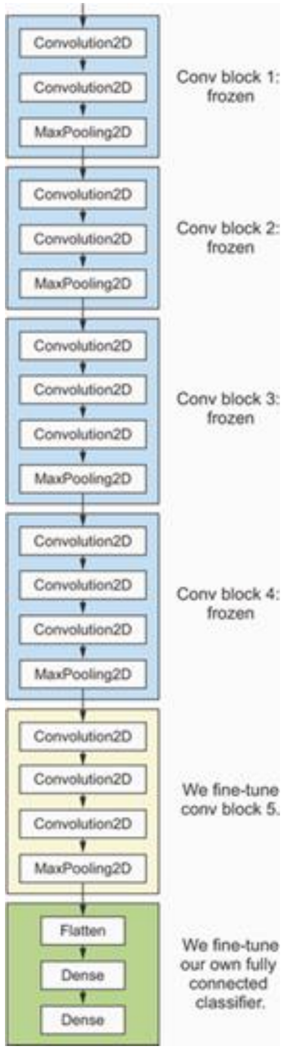
Freeze Only the First Several Layers

- Allow your network to modify / update the last few convolutional base layers as part of training, along with your own Dense layers…
- Iterate over the layers in the network and set the last few to be trainable.



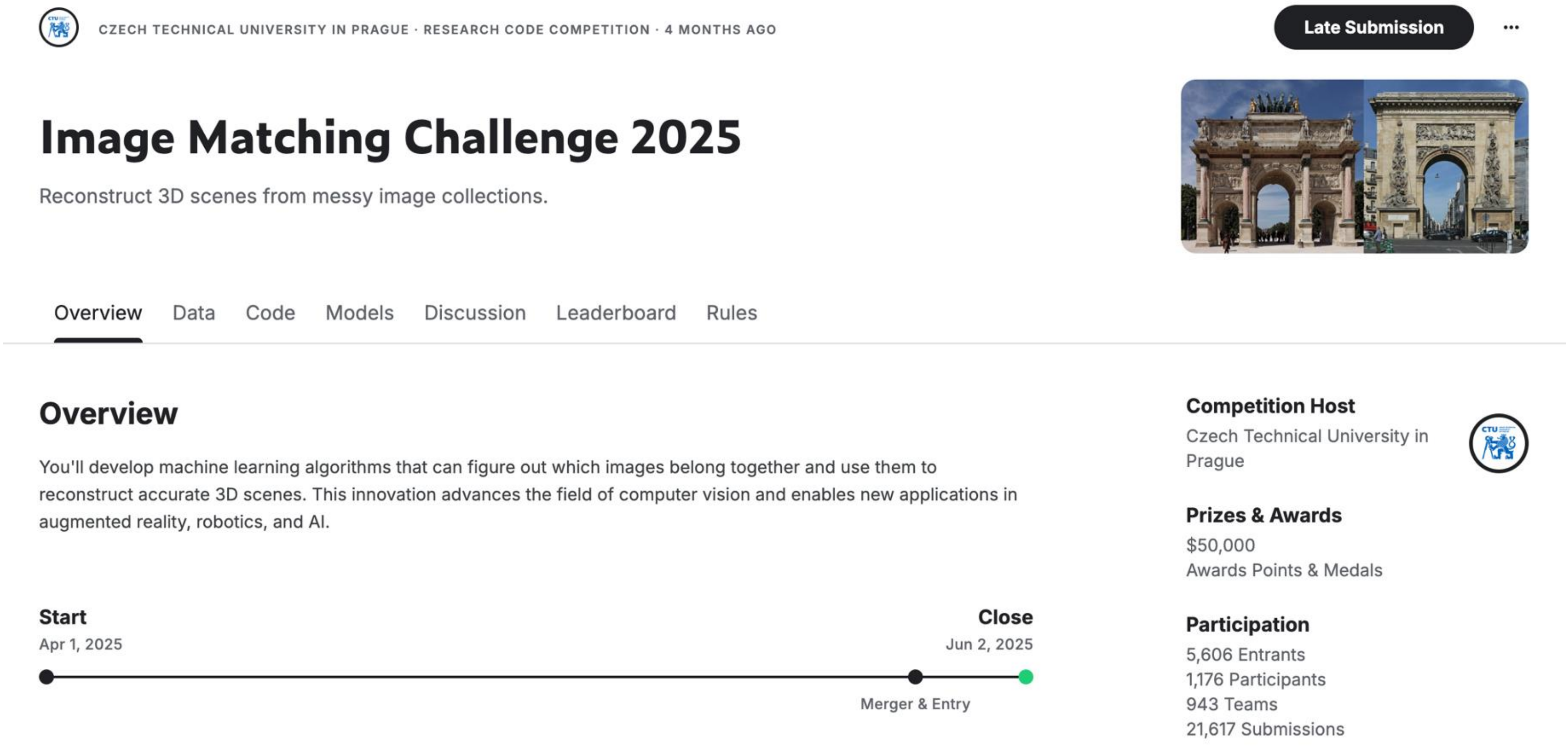
Listing 8.27 Freezing all layers until the fourth from the last

```python
conv_base.trainable = True
for layer in conv_base.layers[:-4]:
    layer.trainable = False
```



© Gordon Burtch, 2026

What Are People Tackling Now?
3D Scene Reconstruction



CZECH TECHNICAL UNIVERSITY IN PRAGUE · RESEARCH CODE COMPETITION · 4 MONTHS AGO

Image Matching Challenge 2025

Reconstruct 3D scenes from messy image collections.

Overview

You'll develop machine learning algorithms that can figure out which images belong together and use them to reconstruct accurate 3D scenes. This innovation advances the field of computer vision and enables new applications in augmented reality, robotics, and AI.

Start
Apr 1, 2025

Close
Jun 2, 2025

Merger & Entry

Competition Host
Czech Technical University in Prague

Prizes & Awards
$50,000
Awards Points & Medals

Participation
5,606 Entrants
1,176 Participants
943 Teams
21,617 Submissions

Self-Evaluation & Peer Call-outs

© Gordon Burtch, 2026

Questions?

© Gordon Burtch, 2026