CA6001 Chapter 5
Computer Vision

Dr Zhang Jiehuang

College of Computing and Data Science

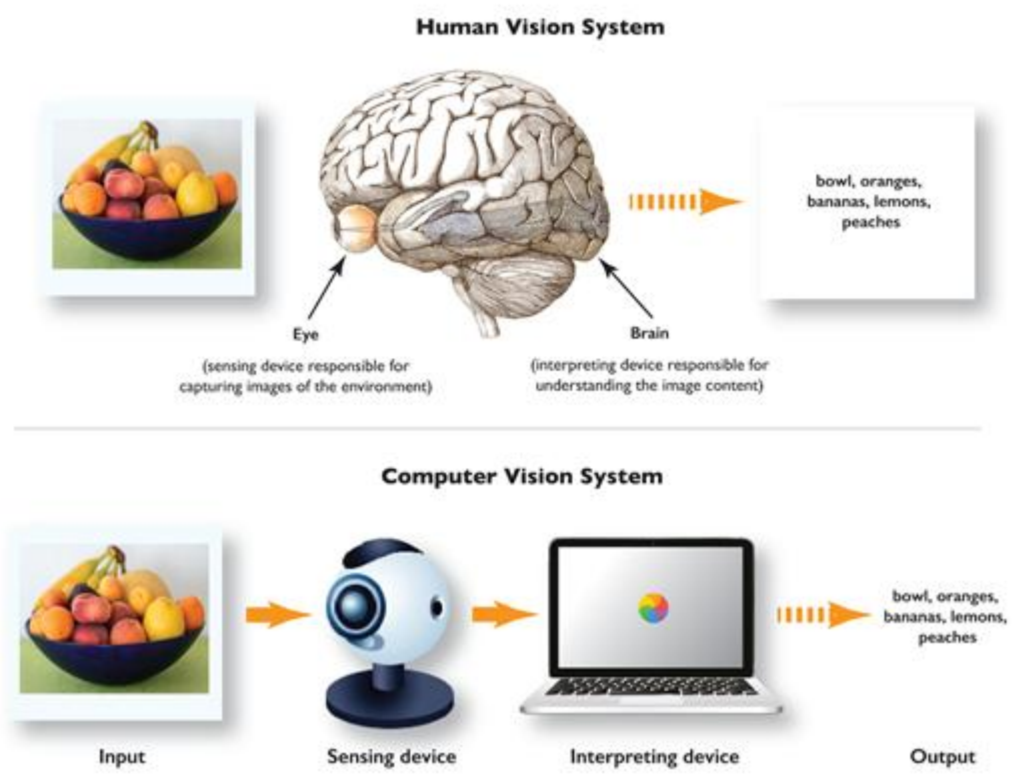
Nanyang Technological University

email: *jiehuang.zhang@ntu.edu.sg*

Chapter 11 – Computer Vision

1. What Is Computer vision and How do Computers see
2. Applications of Computer Vision
3. Traditional VS Modern Computer Vision
4. Convolutional Neural Networks
5. Activation Functions
6. Uses Cases of Computer Vision
7. Classic Architectures
8. Multimodal Models
9. Fun Activity

What is Computer Vision?



Computer Vision(CV) is the field of AI that allows computers to understand and process visual information

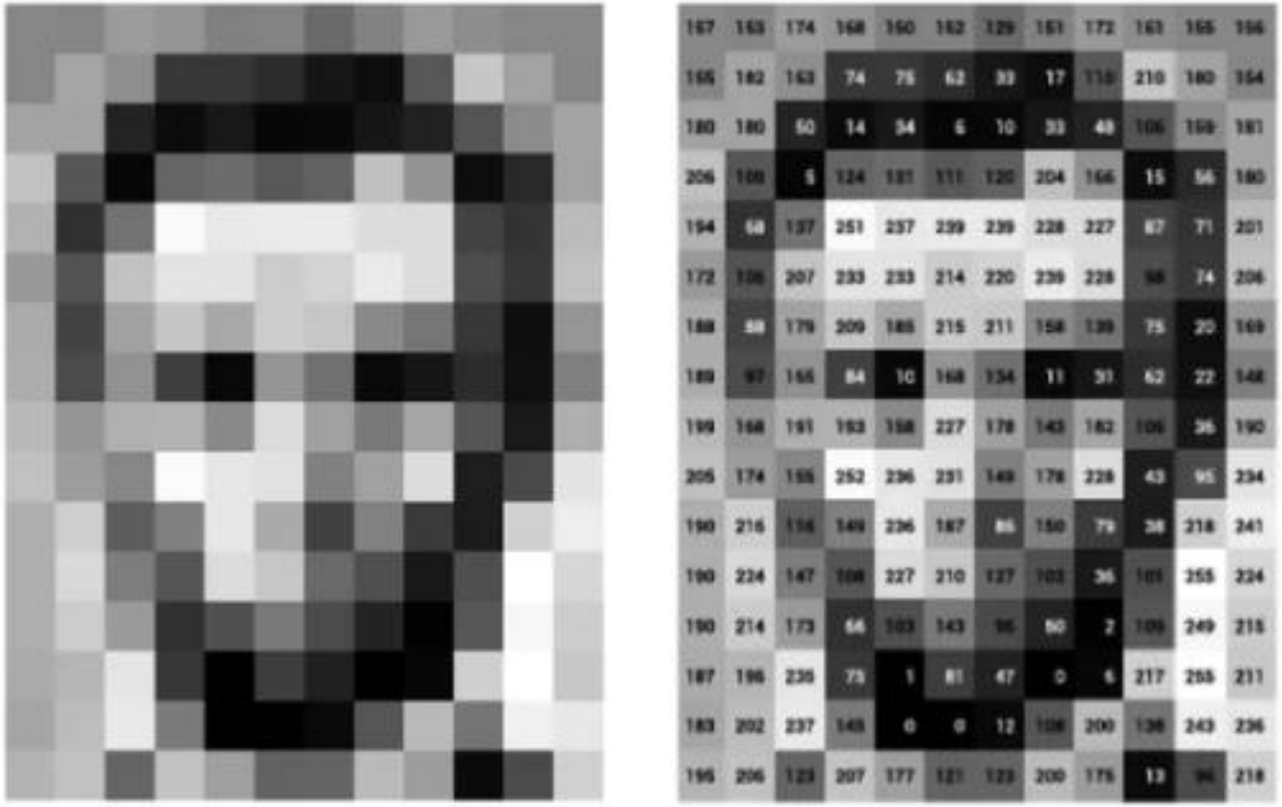
“teach a computer to see like humans”

While NLP allows computers to “speak”, Computer Vision allows computers to “see”

Deep Learning based Computer Vision requires massive amounts of data, for the model to learn to see edges, shapes, and other features

Convolutional Neural Networks(CNNs) drive recent advances in CV

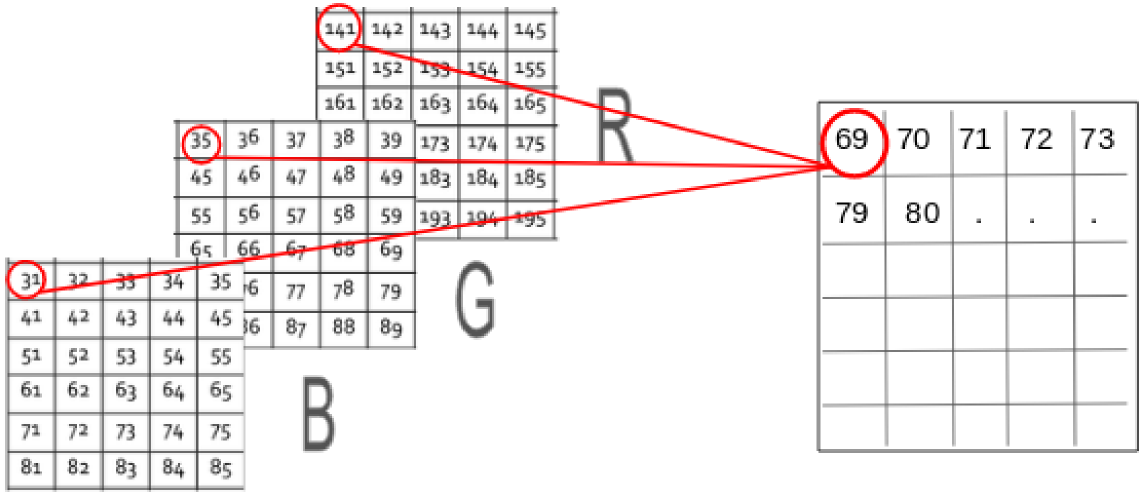
How do Computers See?

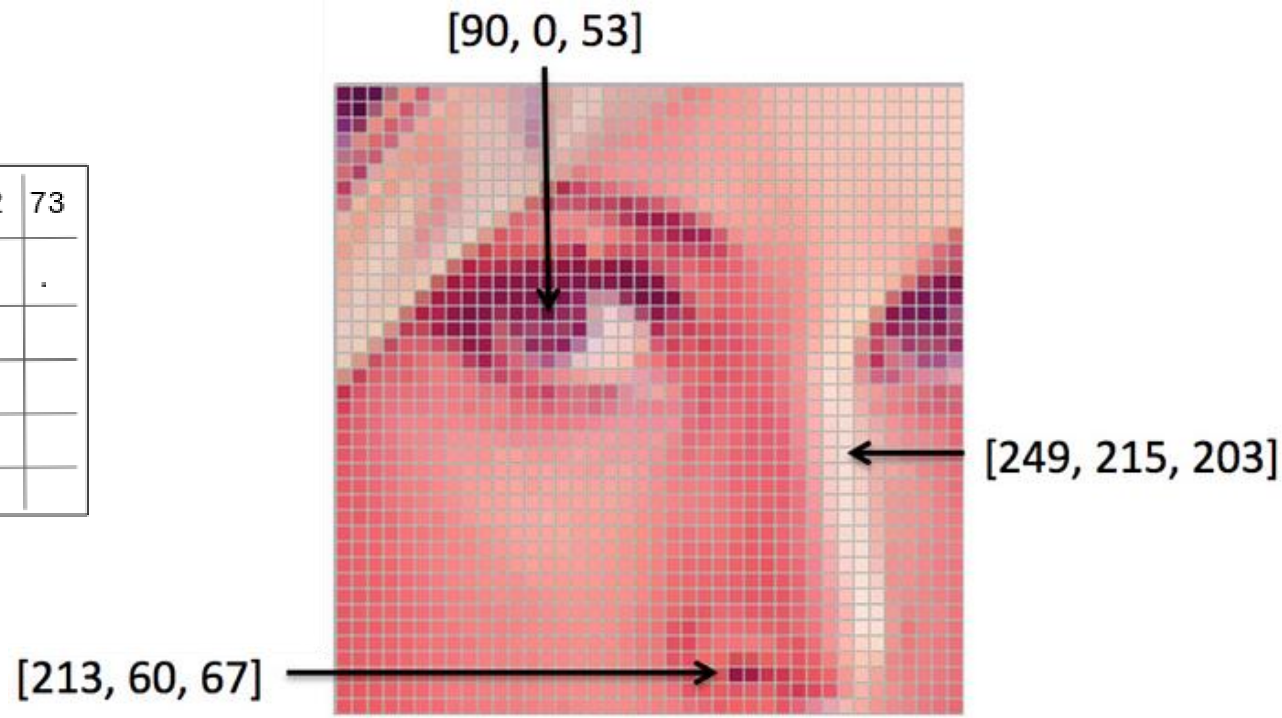


157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	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

A computer sees an image as a matrix of pixels, where the number assigned to a pixel indicates the intensity of the color

How do Computers See?





For a full color picture, each of the base colors (Red Green Blue) is a matrix of numbers.

Hence there is a 3 channel each consisting of a matrix of numbers indicating the intensity of the color

How do Computers See: Sliding Windows

Sliding windows: instead of analyzing an entire image at once, a fixed-size window is moved (or "slid") across different regions of the image to detect objects or extract features at various locations.

The direction of the sliding may not be as shown in the example, and size can be adjusted accordingly for bigger and smaller objects

https://blog.hireterra.com/machine-learning-in-computer-vision-484cbd84cabf

How do Computers See: Kernels/Filters

Pixel Values

0	0	0	0	0	0	0	0
0	1	0	4	2	125	67	0
0	8	2	5	4	34	12	0
0	20	13	25	15	240	2	0
0	76	8	6	6	100	76	0
0	34	66	134	223	201	3	0
0	255	123	89	55	32	2	0
0	0	0	0	0	0	0	0

Kernel 3 x 3 Pixels

1	2	1
2	4	2
1	2	1

Convoluted Image

22	27	36	313	722	576
91	110	120	522	984	576
284	257	198	755	1360	798
507	567	687	1312	1689	955
1061	1288	1496	1911	1659	702
1400	1480	1269	1249	870	279

Sliding windows with kernels: a kernel is a overlay used to slide across the input image to find patterns consistent with the kernel's template

There are kernels used to detect horizontal edges, vertical edges, round shapes etc. Their pixel values will reflect the type of kernel

Applications of Computer Vision

CV has enabled many use cases where previously it was impossible

Self driving cars are a major development in the automotive industry, where a paradigm shift is happening to change how people commute

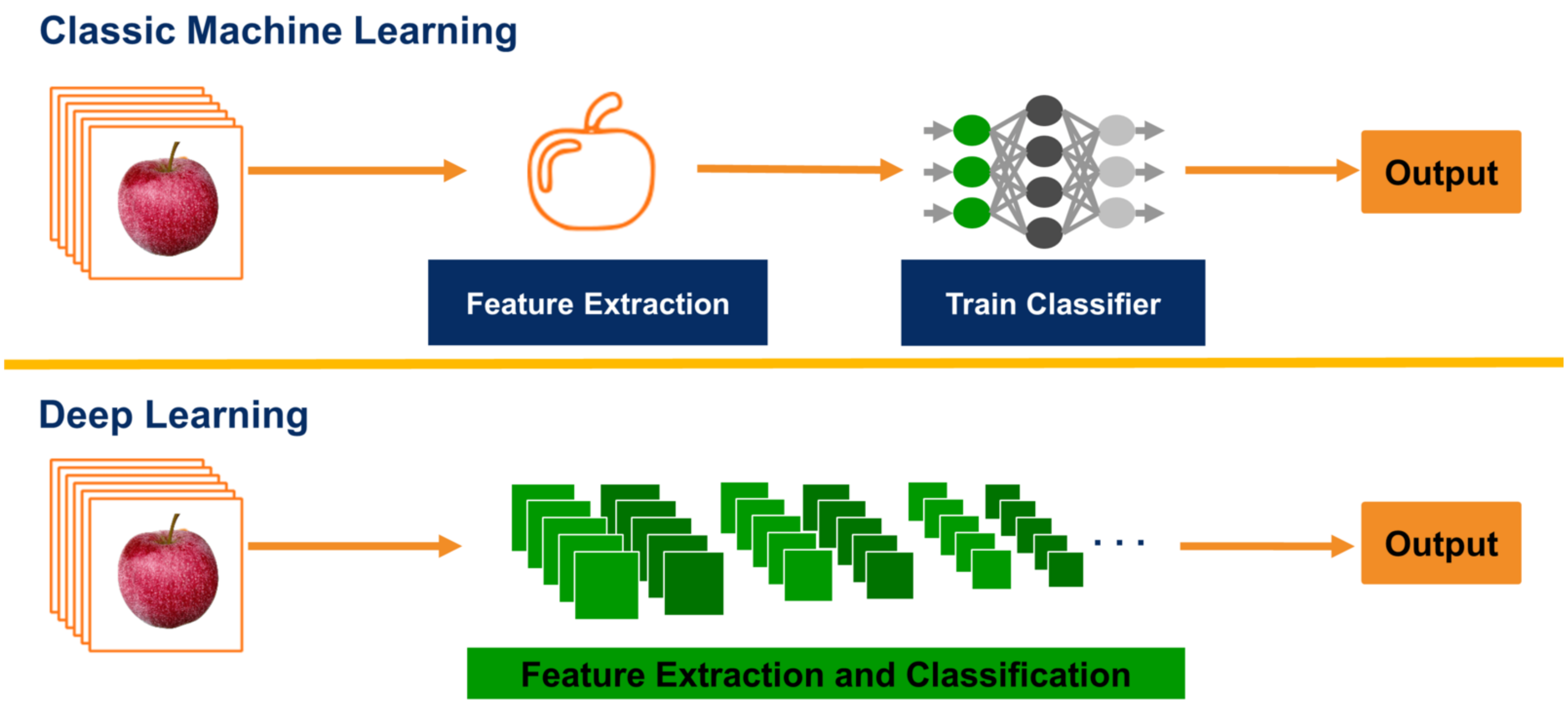
Self driving technology is giving rise to a whole new product: **Robotaxis**

Facial Recognition allows governments to monitor entire cities, finding a single person in a population of millions instantly

Creating images/artwork is a matter of seconds thanks to generative AI

https://blog.hireterra.com/machine-learning-in-computer-vision-484cbd84cabf

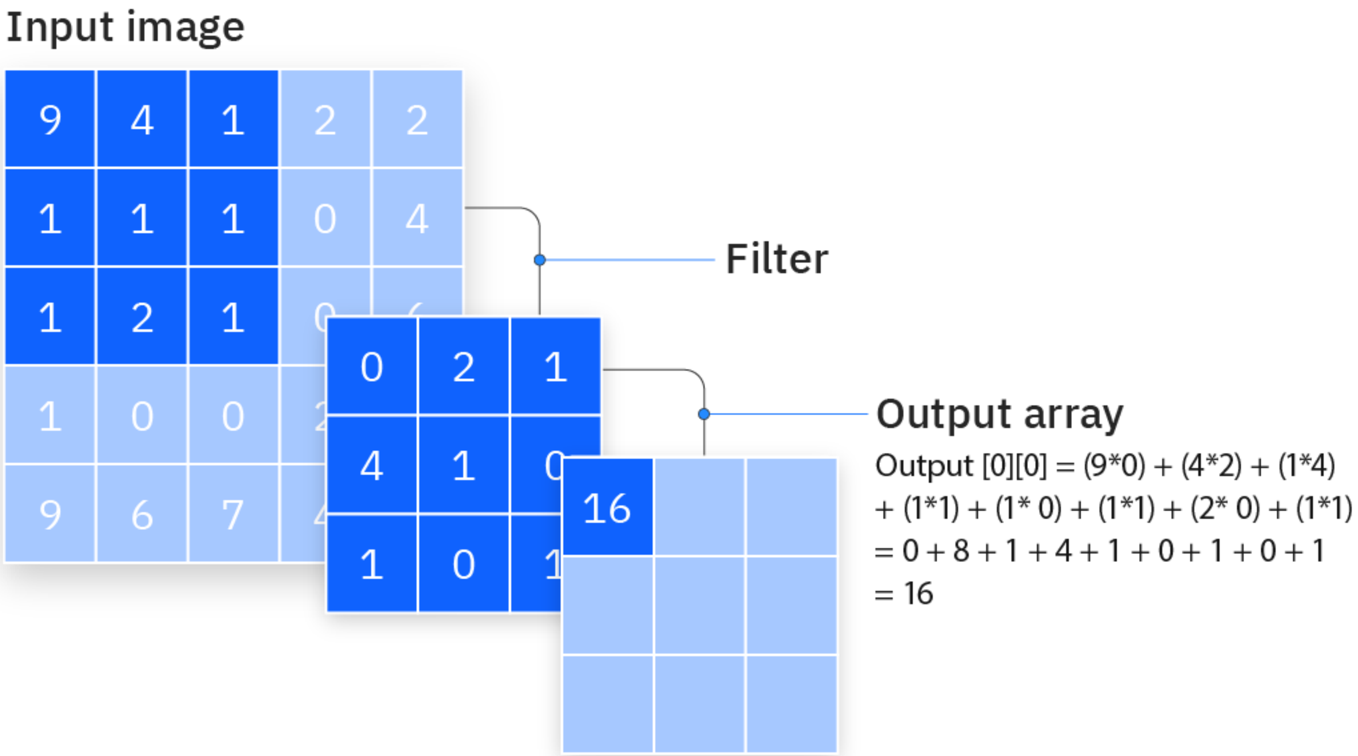
Traditional VS Deep Learning Based Approaches



Before Deep Learning, features needed to be manual extracted, then used to train a classifer

With Deep Learning, the model automatically learns to do feature extraction and applies it to perform the classification

Convolutional Layers: Kernels/Filters



Each Kernel is, in essence, a tiny feature detector.

During training, the filters learn to identify patterns within the image, such as edges, curves, textures, or even more complex shapes.

This process is called a convolution. Which is what gives CNNs its name

Filter looks for things like edges, shapes, and other features that we may not really understand

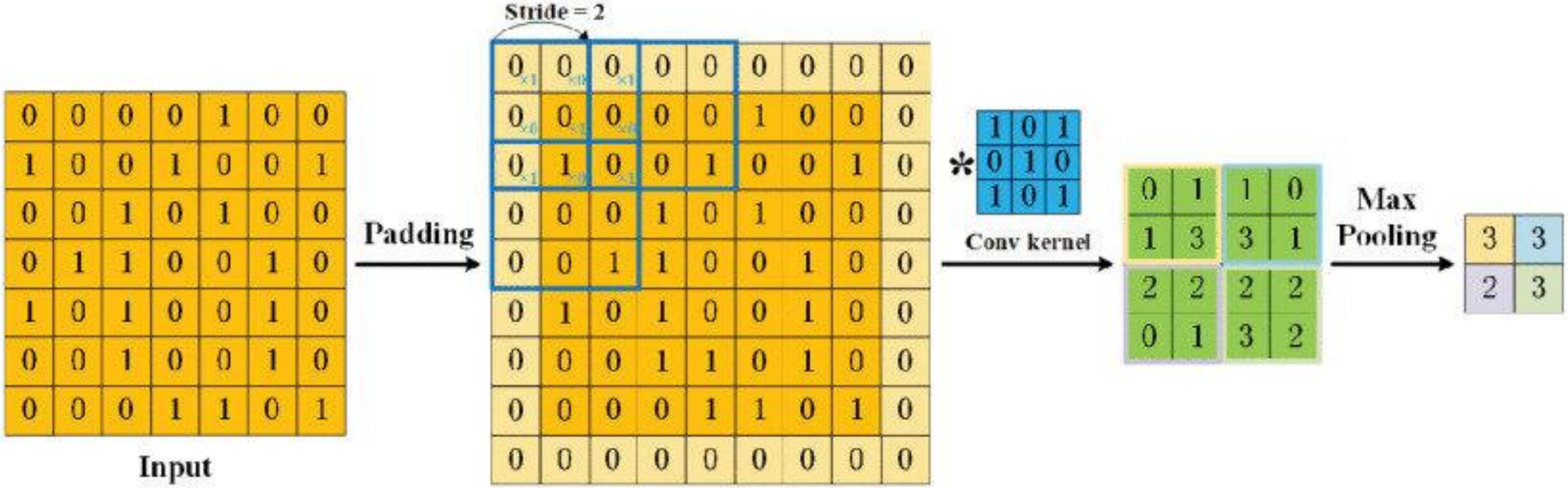
https://www.ibm.com/think/topics/convolutional-neural-networks#:~:text=The%20convolutional%20layer%20is%20the%20core%20building%20block%20of%20a,matrix%20of%20pixels%20in%203D.

Convolutional Layers: Kernels/Filters

$$\begin{bmatrix} 60 & 7 & 98 & 14 & 19 \\ 165 & 159 & 147 & 196 & 169 \\ 187 & 204 & 165 & 41 & 111 \\ 209 & 30 & 201 & 23 & 203 \\ 58 & 79 & 218 & 59 & 118 \end{bmatrix} * \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix} =$$

1. **Overlay:** At each position, the filter is placed on top of a section of the input image.
2. **Element-wise Multiplication:** Each element of the filter grid is multiplied with the corresponding pixel value from the overlapping part of the image.
3. **Summation:** The multiplied values are added together to produce a single number.
4. **Output (Feature Map):** This computed number becomes one pixel in the resulting output, commonly called a feature map or activation map.
5. **Sliding and Repetition:** The filter slides across the entire image (with a defined stride, i.e., the step size of movement), repeating the same process to create a complete feature map.

Convolutional Layers: Stride and Padding



Stride – distance the filter moves

Padding – adding numbers to the matrix

Especially for large neural networks with many layers, the deeper layers work in abstract ways that human cannot fathom

https://www.researchgate.net/figure/Summary-of-convolution-padding-stride-and-Max-Pooling_fig2_381448625

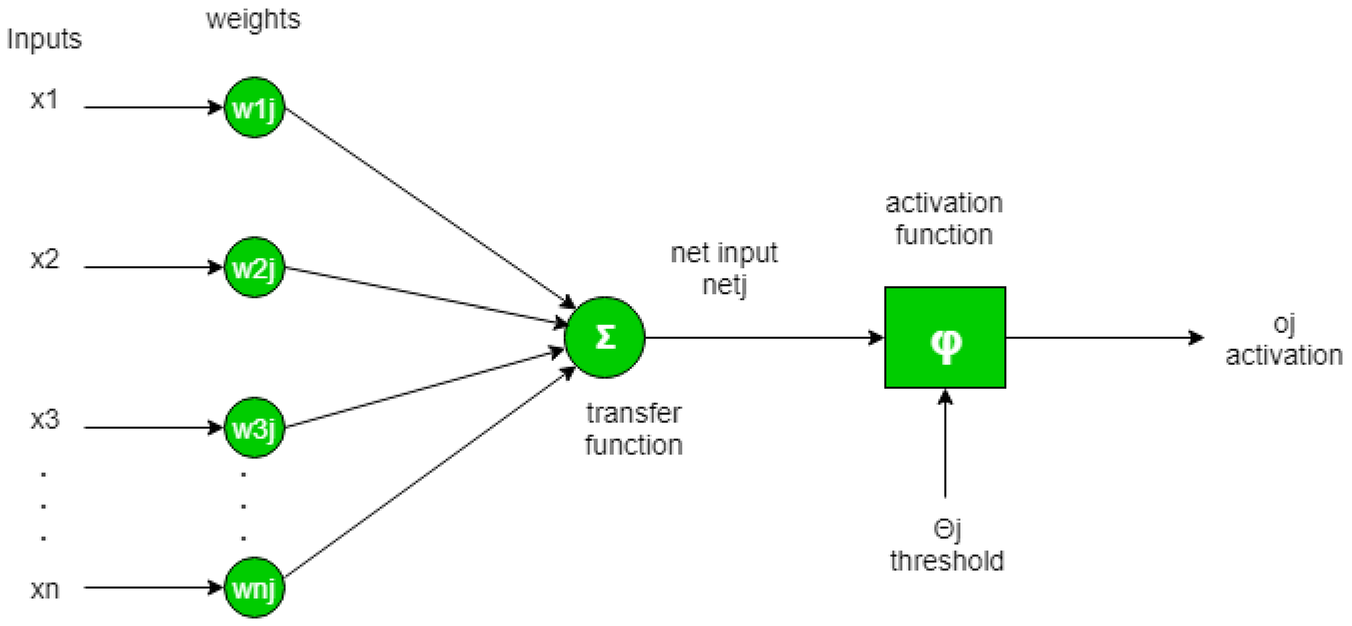
Convolutional Layers: Feature Maps

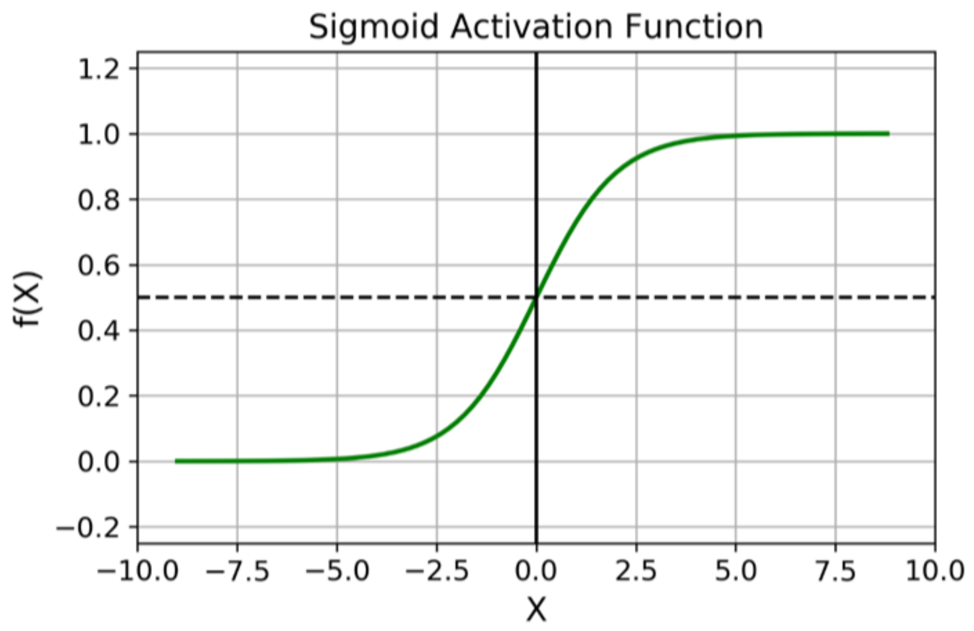
Feature map shows heatmap of what a CNN sees and uses to classify

Heatmap shows the different areas of the pictures highlighted, as the filters slide over the entirety of the input image

https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Activation Functions





Activation functions transform the input signals into the output according to the type used

They determine whether a neuron should activate or ‘fire’

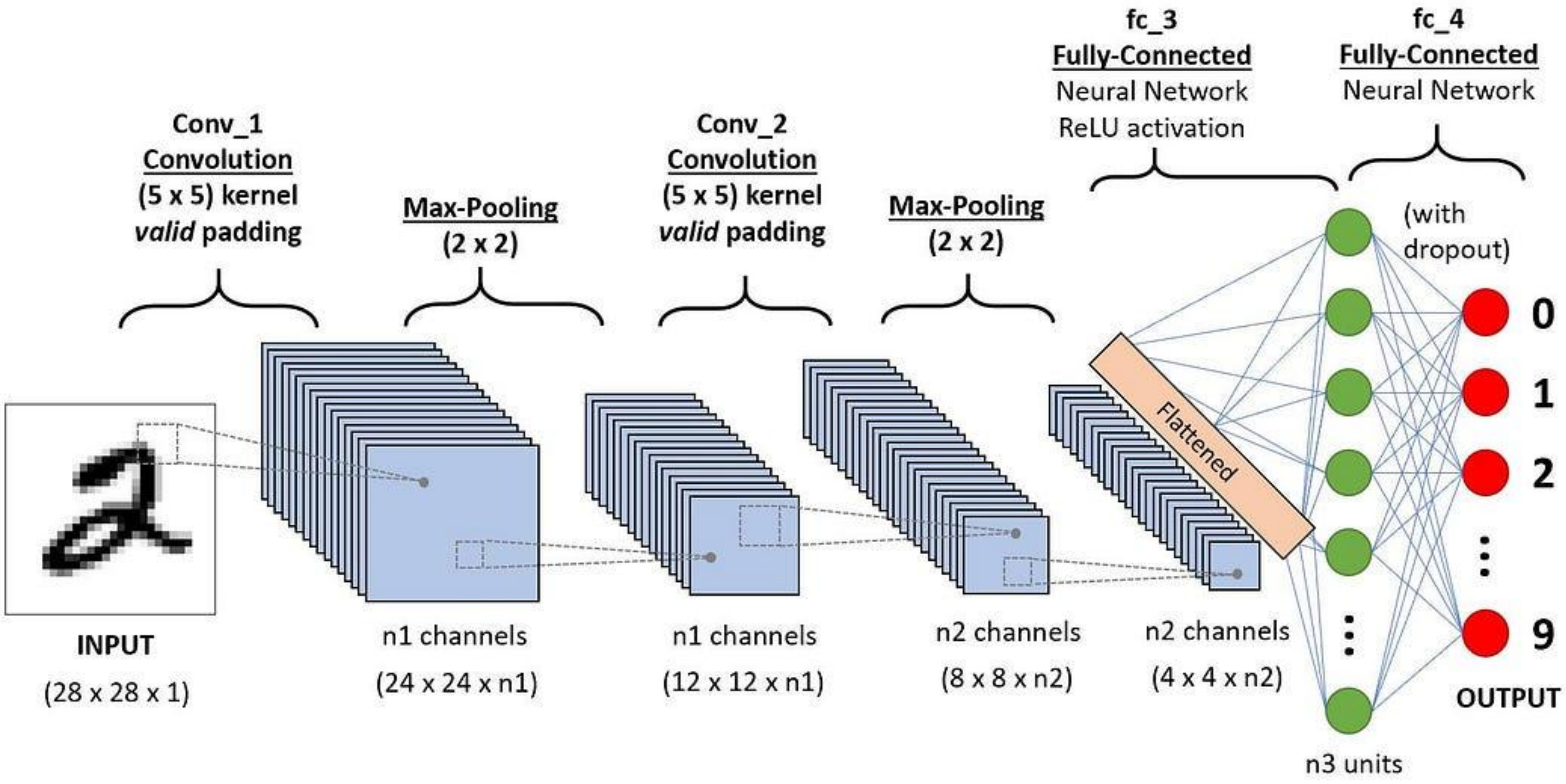
Without them, the neural network can only learn linear relationships

ImageNet: A Paradigm Shift

Big data may seem obvious now, but in the early 2010s it was a paradigm shift to focus on large amounts of high quality, labelled data

ImageNet researchers collected millions of images from the internet and performed data cleaning, labelling to prepare the images to train a CNN

Convolutional Neural Networks



https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

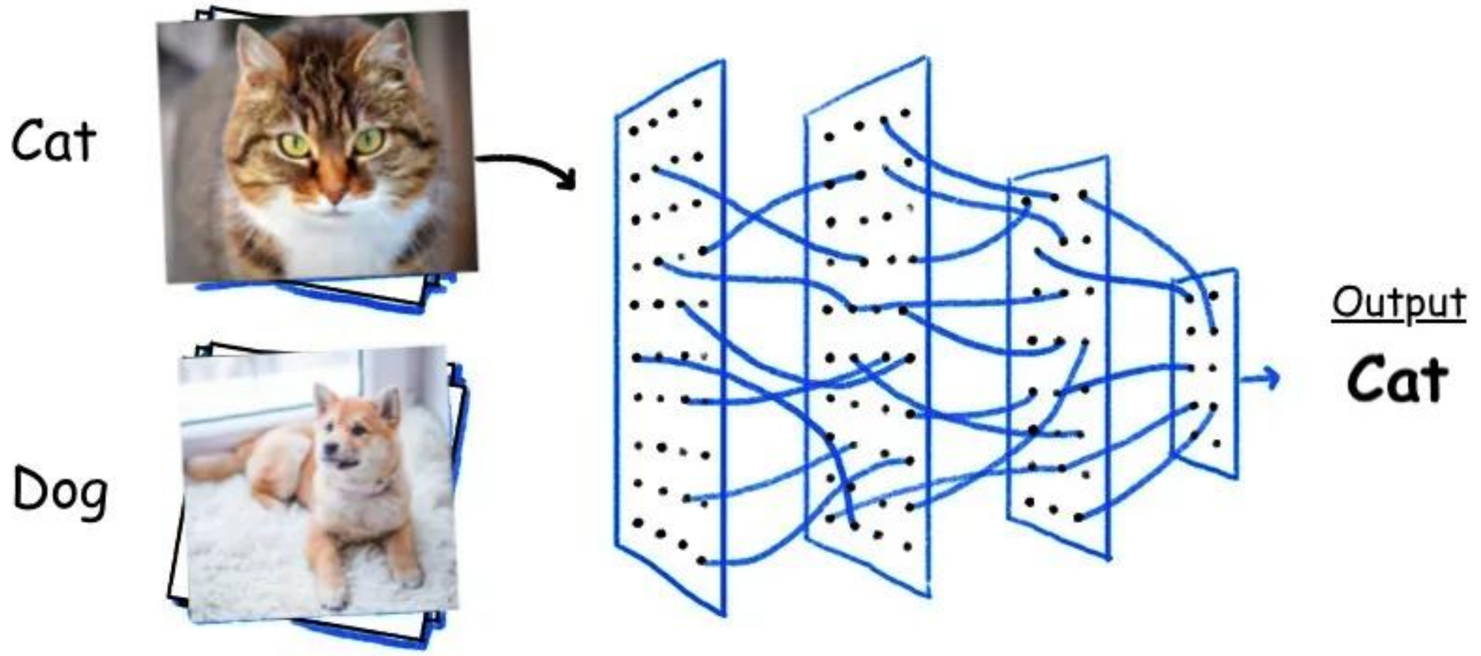
Convolutional Neural Networks

Convolution: The heart of a CNN is the convolution layer. This layer uses small, learnable filters (kernels) that slide across the image like spotlights. Each filter focuses on recognizing a specific feature, such as edges, curves, or textures. As it moves over the image, the filter calculates dot products between its weights and the corresponding image pixels, creating a feature map that highlights areas where the pattern is present.

Pooling: Pooling layers serve to downsample the feature maps. They reduce the amount of information while preserving the most important detected features. This makes training less computationally expensive and helps prevent overfitting, where a model becomes too focused on the training data.

Fully Connected Layers: After multiple convolution and pooling operations, the final feature maps are flattened and fed into fully connected layers (similar to traditional neural networks). These layers perform the final classification task, using a softmax layer

Computer Vision Use Cases: Classification



Problem Definition: Assign a label or category to millions of images in dataset

Training data: millions of training images, high quality and labelled

Output: Probability distribution over possible classes, Cat or Dog

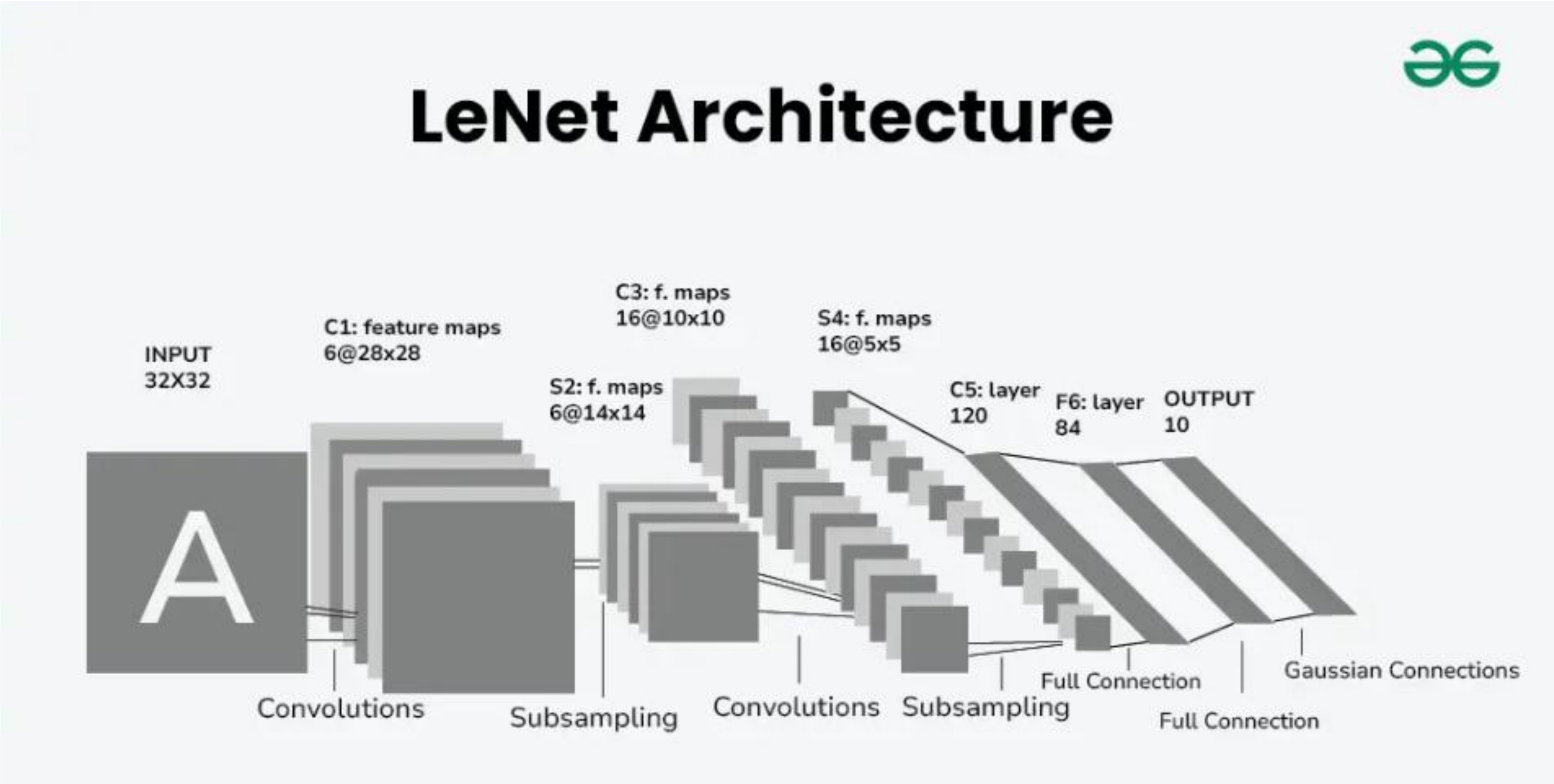
Input: images labelled cat or dog (supervised learning)

Notable Datasets:

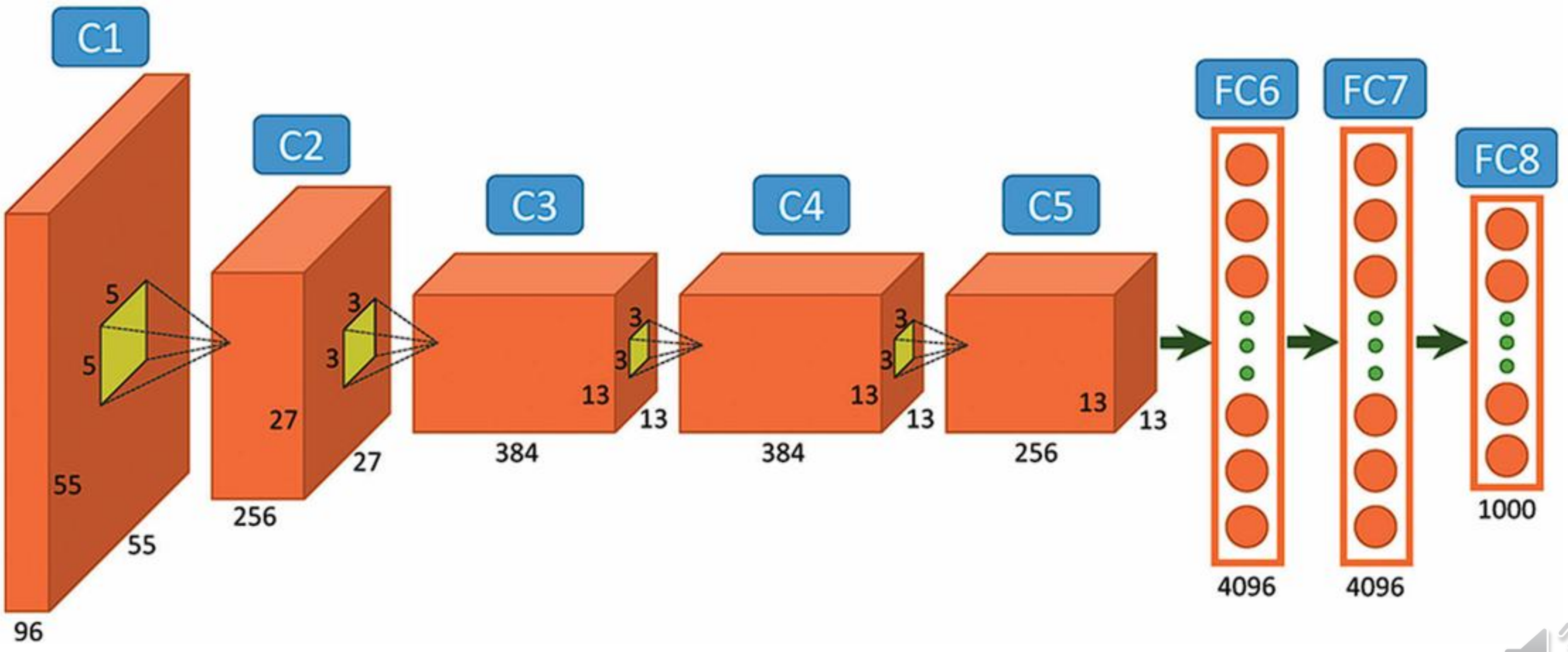
Imagenet

CIFAR

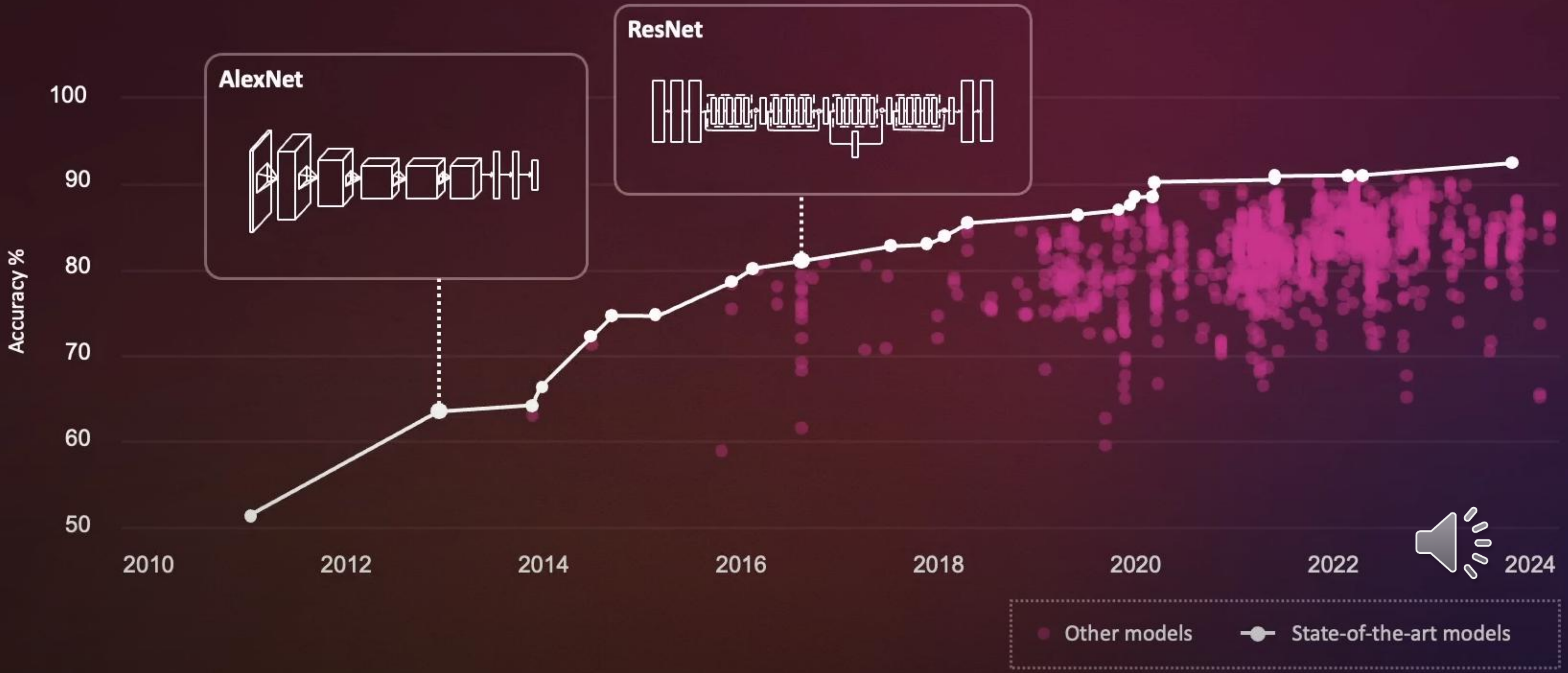
Classic Architectures: Lenet



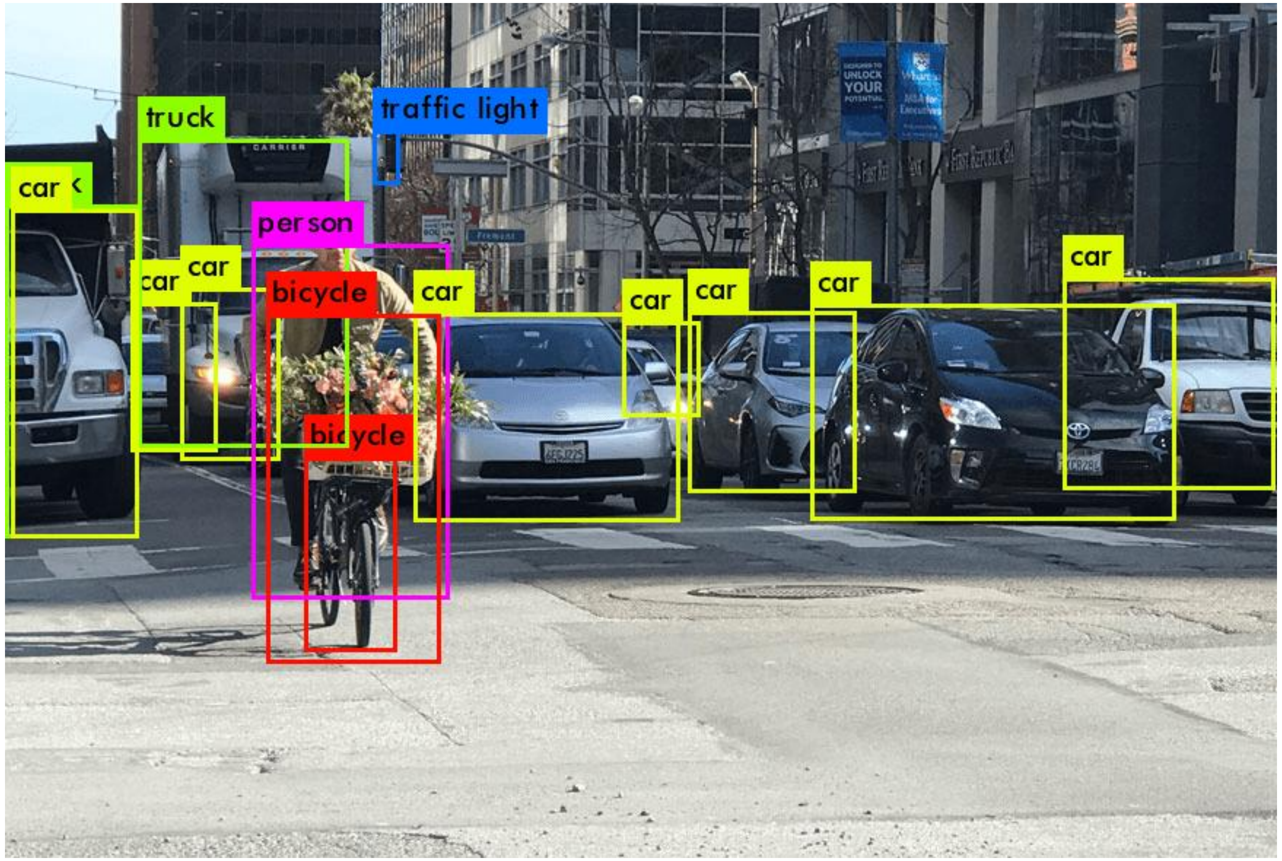
Classic Architectures: Alexnet



Top 1 Accuracy Model in Image Classification on IMAGENET Each Year



Computer Vision Techniques: Object Detection



Object Detection adds an extra dimension to classification: finding where the object is (bounding boxes)
On top of what the object is (classification)

Notable Architectures:
YOLO (you only look once)
R-CNN (region-based CNN)

Applications:
Autonomous Driving
Surveillance and Monitoring
Defect Detection in Manufacturing

https://www.wellable.co/blog/ai-chatbots-bridge-mental-healthcare-accessibility-gap/

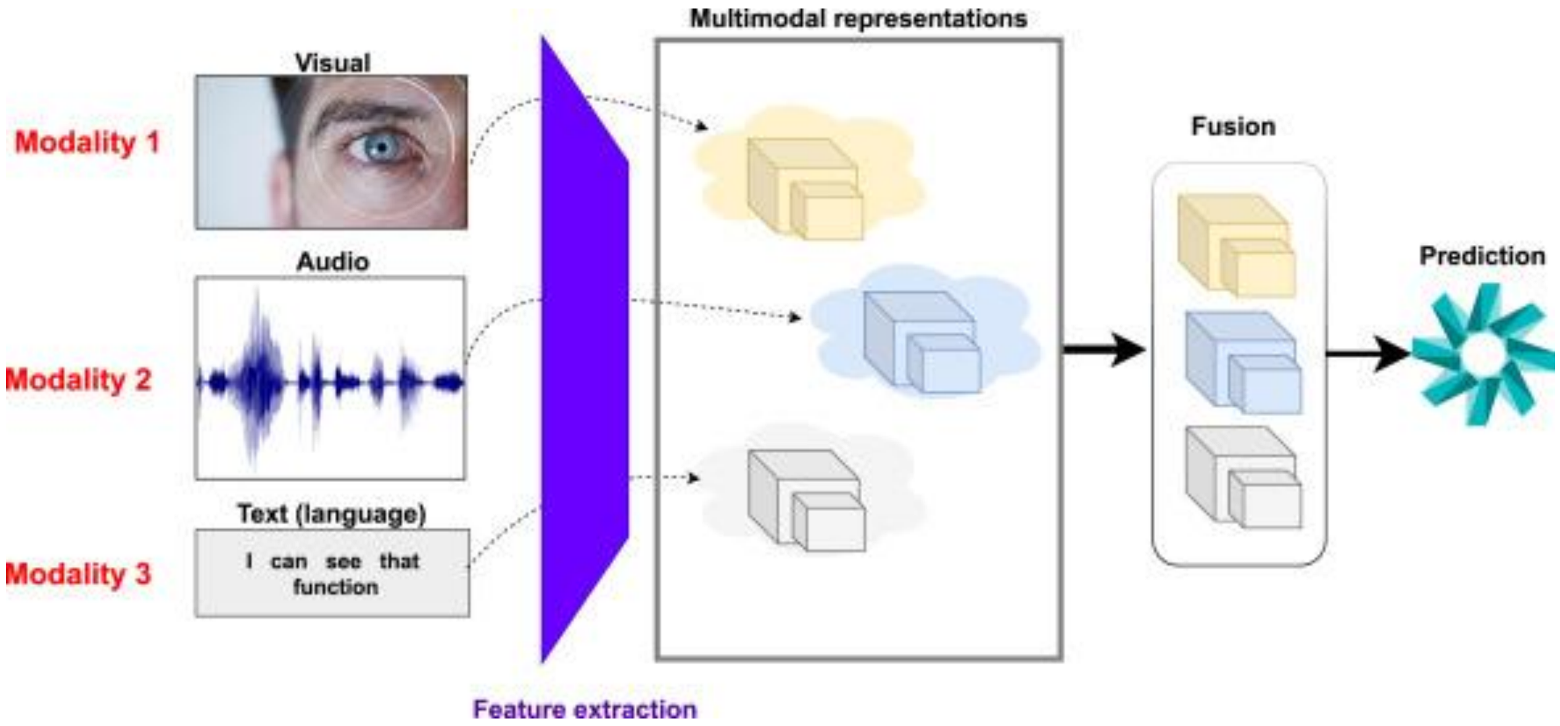
Computer Vision Use Cases: Content Generation

Diffusion Models: Gradually add noise to images, then learn to reverse the process
- Stable Diffusion
- Midjourney
- DALL-E

Generative Adversarial Networks (GANs):
•Two competing networks:
  - Generator: Creates images
  - Discriminator: Tries to detect fake images

•Training is like a minimax game:
  - Generator tries to fool discriminator
  - Discriminator tries to spot fakes

Computer Vision Use Cases: Multimodal Models



Multimodal deep learning deals with the fusion of multiple data types, such as text, image, video, audio etc

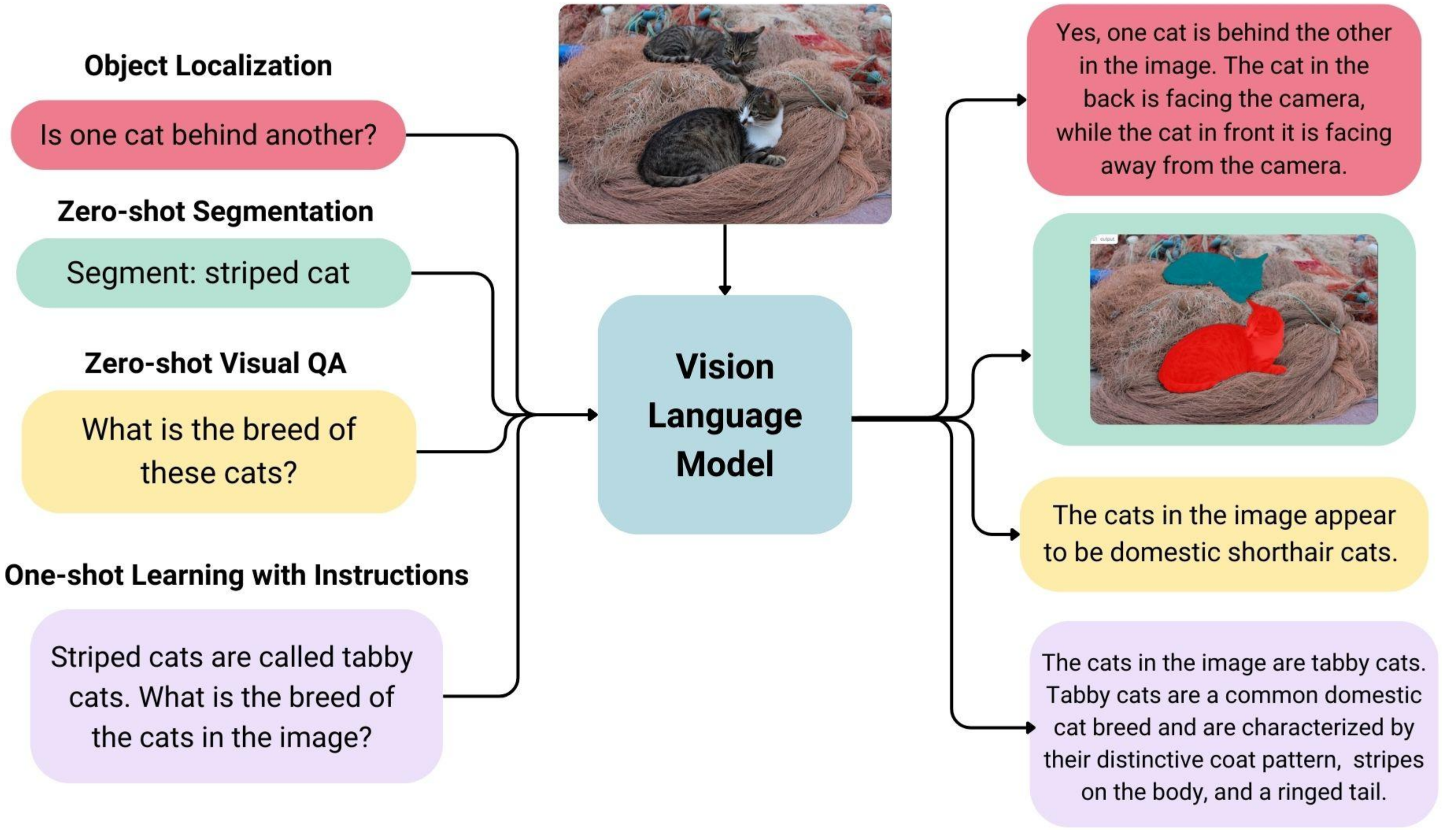
MMDP models can consist of many neural networks, each specializing in analyzing a particular modality. The output of these networks is then combined to create a joint representation of the multimodal data

Example:

mixture of experts architecture

Vision LLM

Vision Language Models

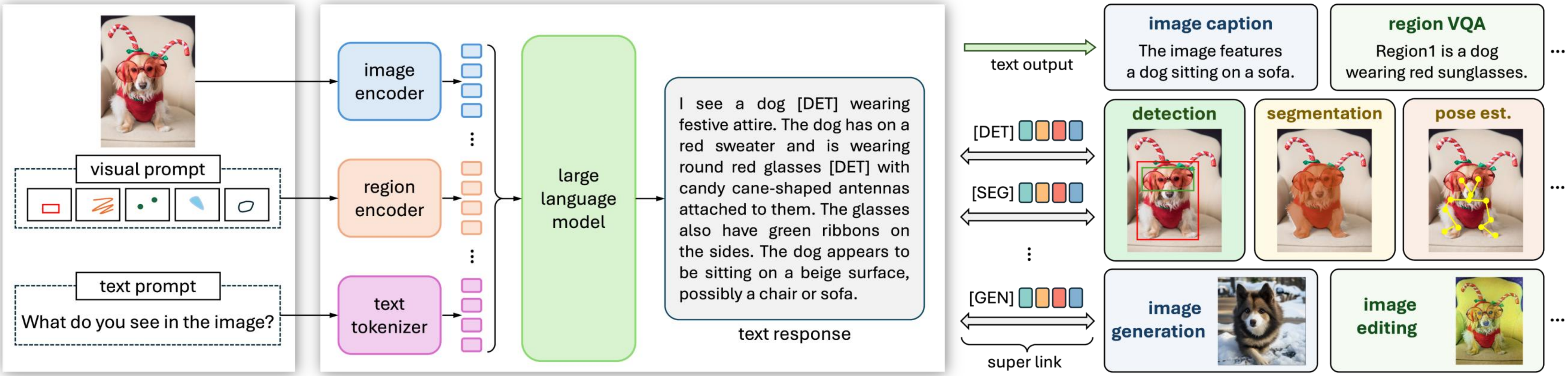


Vision Language Models (VLMs) are multi-modal models that can learn from images and text, they take images and text inputs and generate text outputs.

The reverse process can be done also, by generating images/videos from text inputs

https://huggingface.co/blog/vlms

Vision LLM

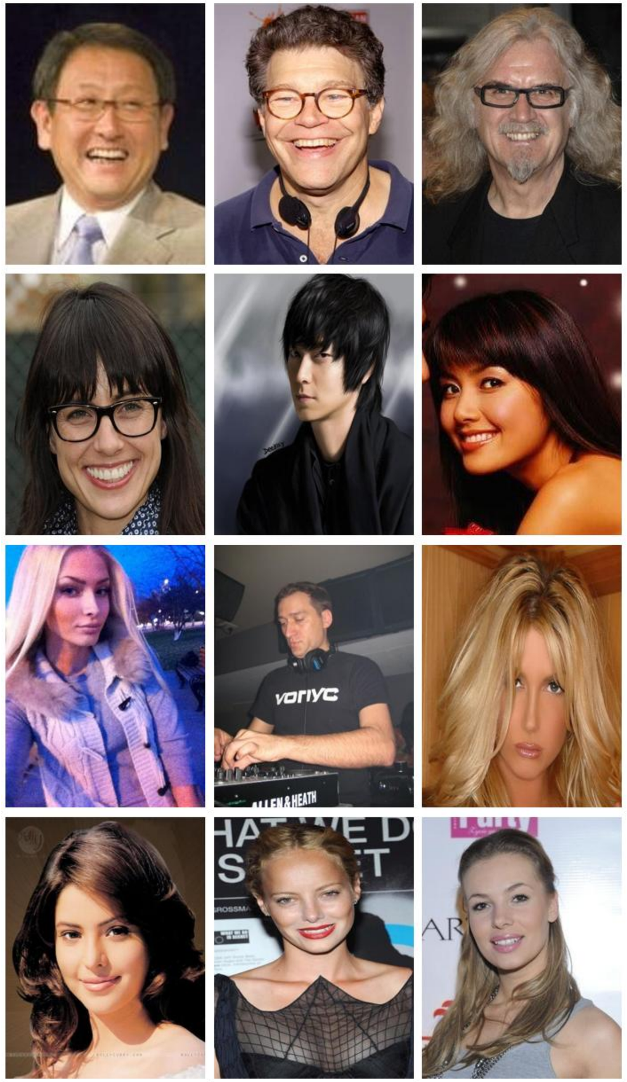


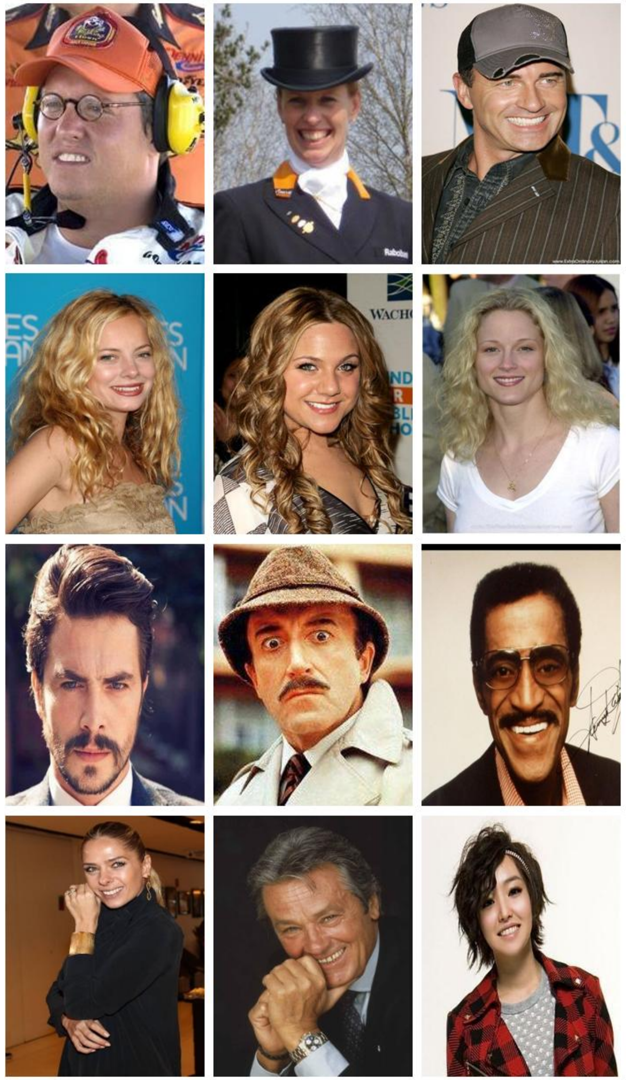
Vision LLM is a generalist multimodal LLM supporting vision language tasks, such as visual understanding, perception and generation

Each encoder-decoder pair is responsible for a specific task, such as image captioning, object detection, segmentation and pose detector.

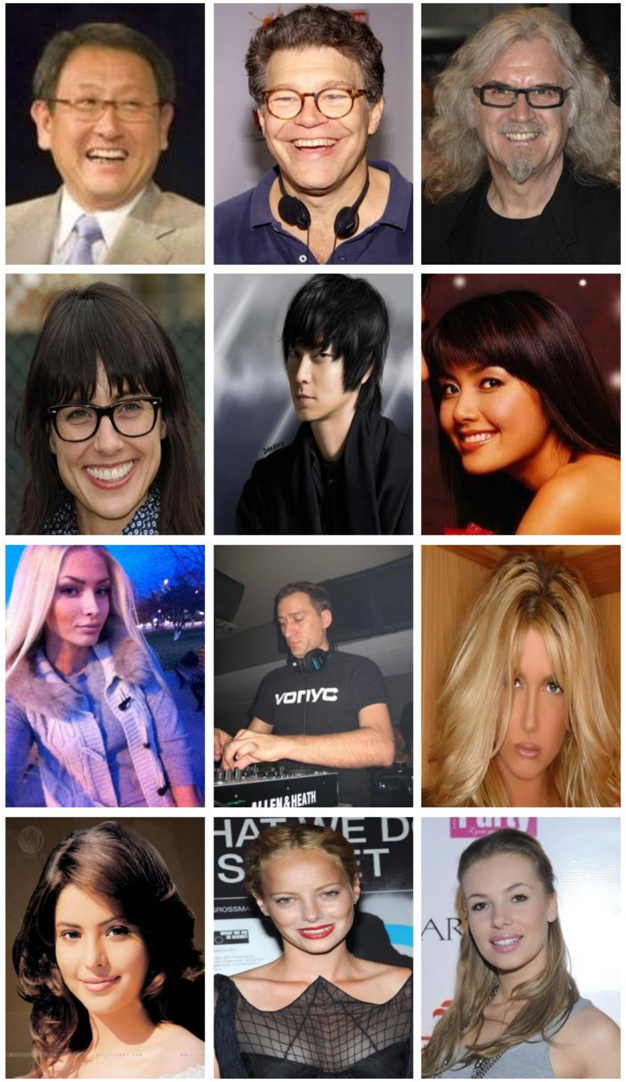
Multimodal models are getting closer to how human brain processes many different types of data

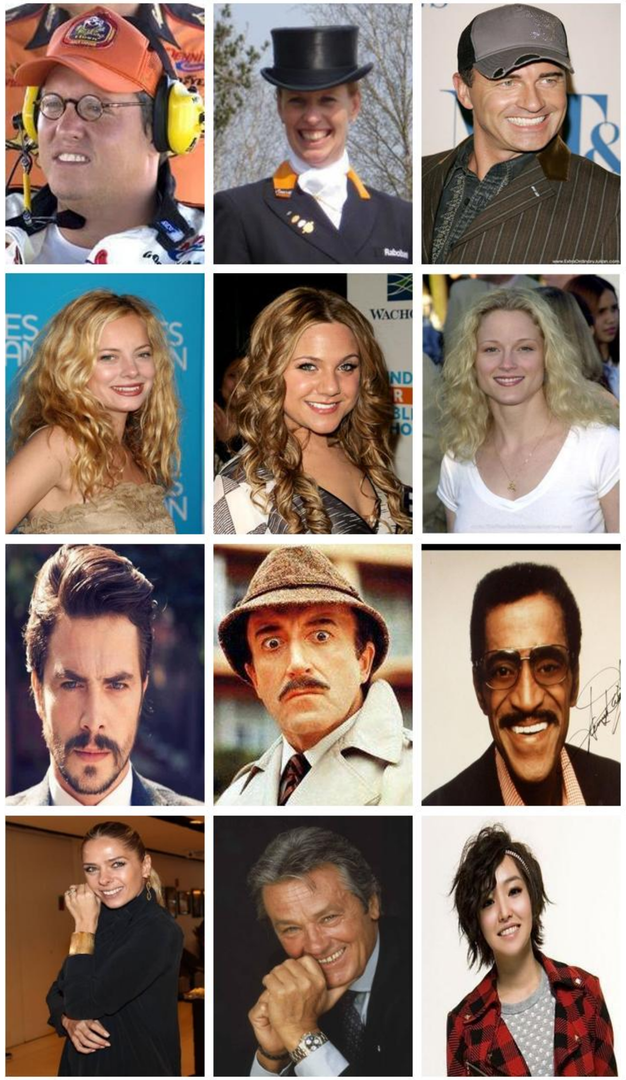
Do you recognize any of these celebrities?





Generative Adversarial Networks





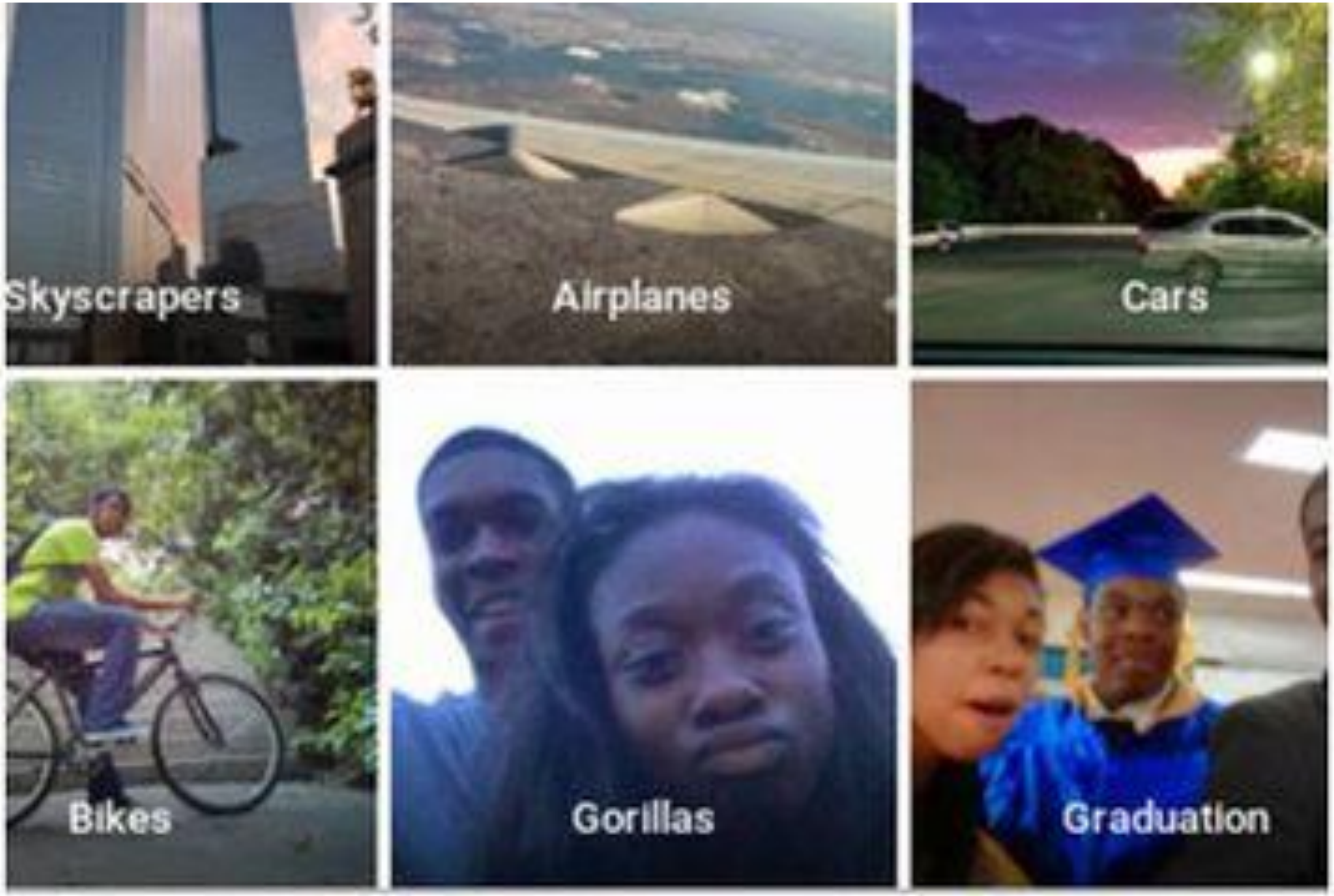
These celebrity pictures are generated by a GAN (generative adversarial network), trained on images of real celebrities

The fake celebrities have many features that look similar or are completely inspired by the real training images, you can see the uncanny resemblance

The features are learnt from the model, translated into the feature map and used to generate these fake pictures

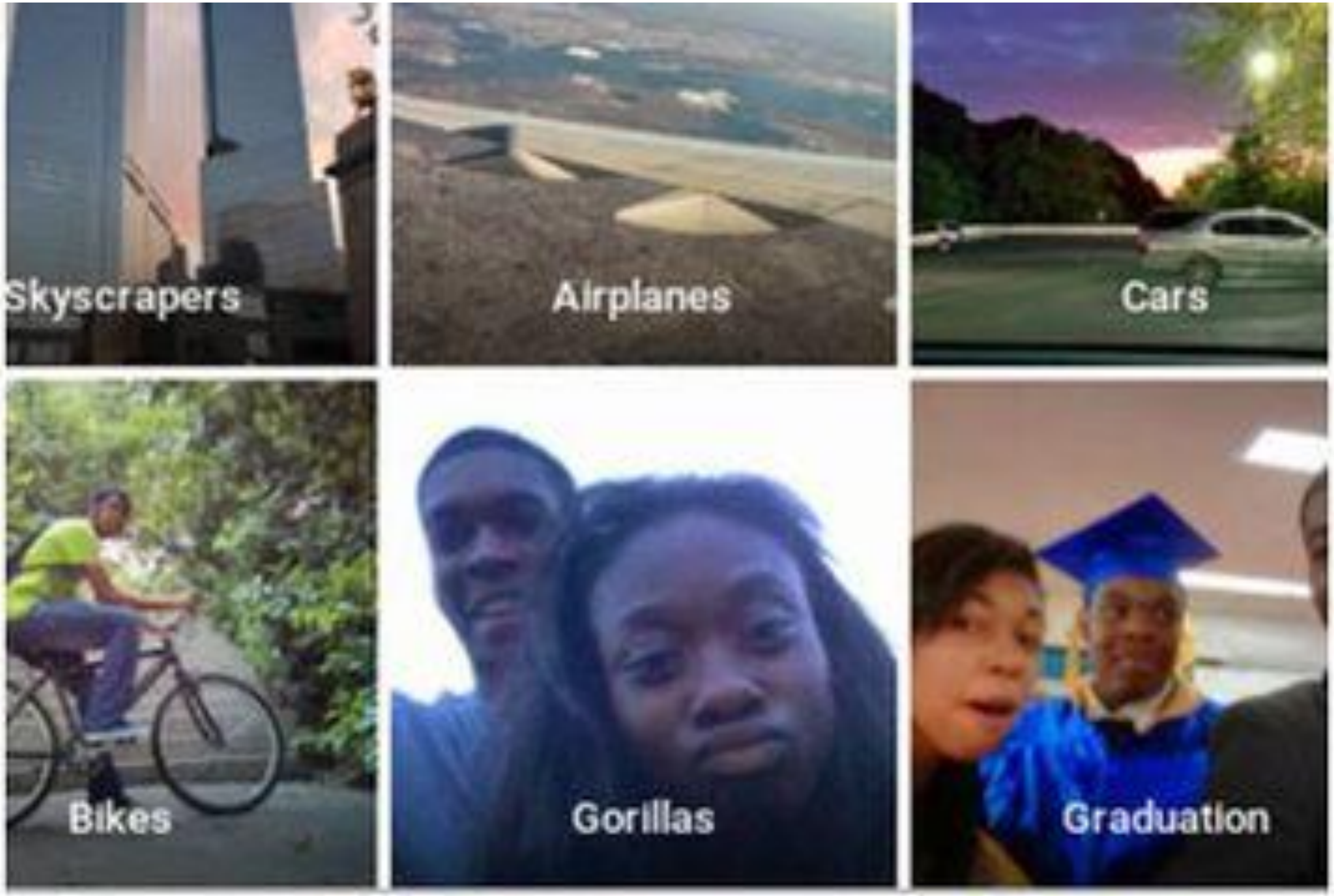
Try to generate your own fake content with GANs!

Notice anything out of place?



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Notice anything out of place?



Google Photos wrongly labelled African Americans as Gorilla

Result was a PR nightmare that took much effort to appease

AI is not perfect, and we need you to help regulate them, as Responsible AI Engineers

References for Chapter 11 – Computer Vision

1. https://www.datacamp.com/blog/top-machine-learning-use-cases-and-algorithms
2. https://www.databricks.com/resources/ebook/big-book-of-machine-learning-use-cases/thank-you?scid=7018Y000001Fi19QAC&utm_source=google&utm_adgroup=141597893652&utm_offer=big-book-of-machine-learning-use-cases&utm_term=machine+learning+use+cases&gad_source=1&gclid=CjwKCAiAxqC6BhBcEiwAlXp45zG
3. https://www.researchgate.net/publication/351021675_Artificial_intelligence_in_cancer_diagnostics_and_therapy_Current_perspectives-G9y0tvxwNF2eskPqGlVAsxxtPXDibjGQBobW-_5A4ZhFFsDKTRoCWT8QAvD_BwE
4. https://www.heavy.ai/technical-glossary/fraud-detection-and-prevention
5. Andrew Ng’s Machine Learning course https://www.coursera.org/learn/machine-learning/lecture/Q8Vvp/supervised-learning-part-2
6. https://cloud.google.com/discover/what-is-unsupervised-learning
7. https://blog.aspiresys.com/data-and-analytics/customer-segmentation-empowered-by-machine-learning-reap-the-benefits-of-ai-to-serve-your-customers-better/
8. https://www.v7labs.com/blog/supervised-vs-unsupervised-learning

Note: All online articles were accessed between Oct to Nov 2024

The End
Questions?