Module 4 | Introduction to CNN

08 November 2025 10:07

CNN stands for Convolutional Neural Networks

What is convolutional neural network? How will you explain this to a kid?

©afpc

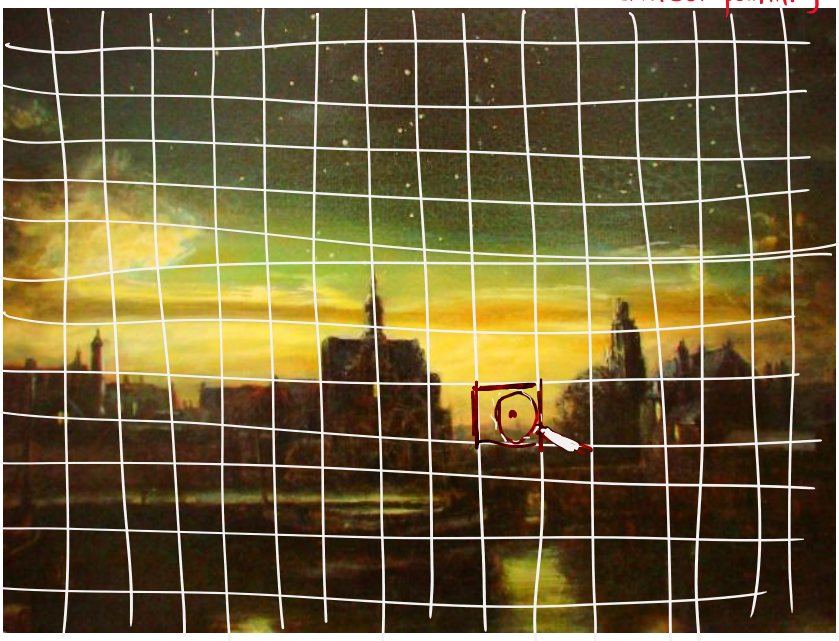
Detective: SHERLOCK HOLMES (SH)

→ detective

↓

to validate a painting whether it is fake or not

Vermeer painting



① Painting as a puzzle

SH considers the painting as a big puzzle made up of tiny squares called pixels. These pixels are basically

Colors → Numbers → R : [0-255]
G : [0-255]
B : [0-255]

② SH's Magnifying Glass ⟷ CONVOLUTION

↳ also known as filtering

SH has a magnifying glass with which he looks at small/local sections of the painting to understand the finer details.

→ Zooming-in

↓

"convolution step is like zooming-in into the raw image"

SH scans the whole painting from left to right, top to bottom → to find the important lines, corners or edges etc.

③ What patterns does SH see?

* In the very 1st layer, SH looks for the simple things such as fine lines, edges etc.
* In the second layer, he tries to find shapes or patterns and
  in subsequent layers, he tries to combine these simple patterns to find even more complex patterns.

fine lines, edges, corners etc.

↓

to form shapes/structure

↓

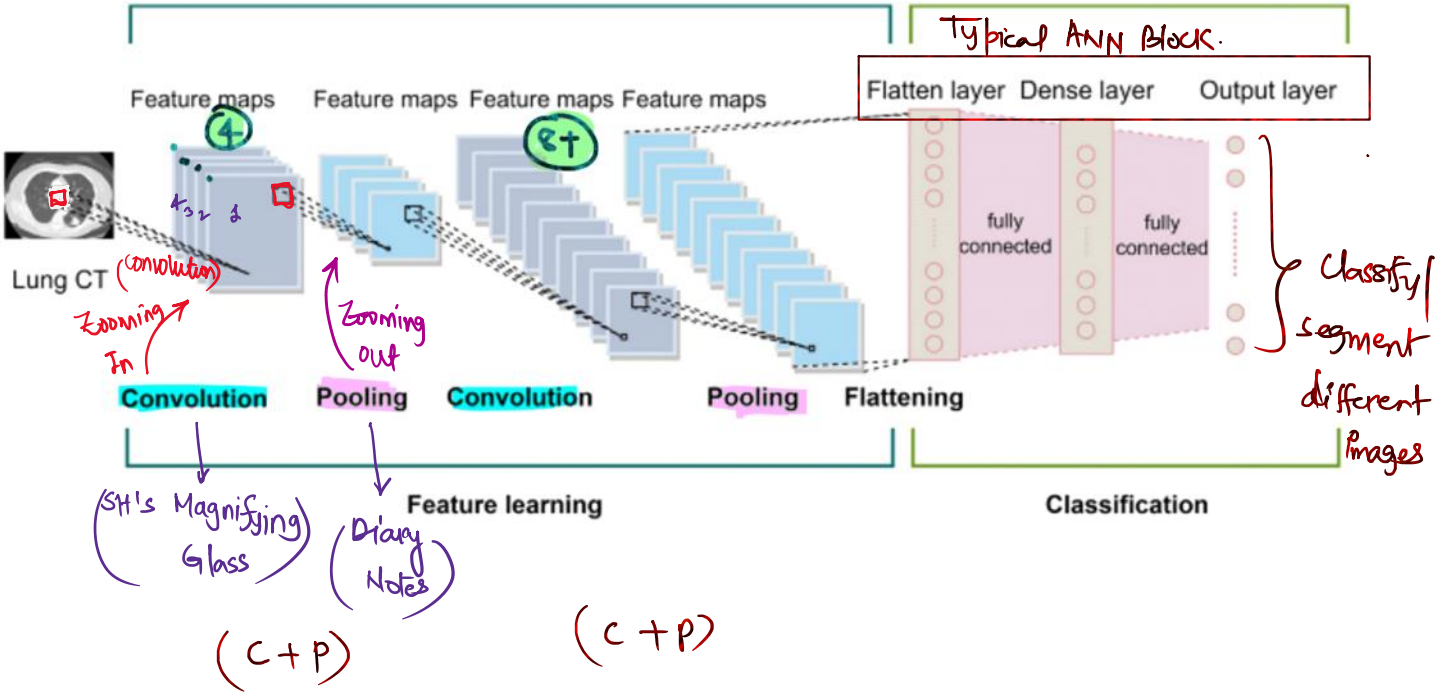
more complex patterns

④ SH simplifies the patterns → (POOLING)

to make these things easier and simpler, SH simplifies the painting case by keeping only the most imb ortant fea

simplifies the painting case by keeping only the most important features or shapes or patterns and this step is called POOLING in CNN

like Zooming out to see the big picture.



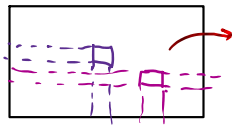
Step #1 CONVOLUTION / FILTERING

Step #2 POOLING

Textbook Definition

A convolutional neural network is a type of deep learning model specifically designed for tasks like image recognition / image classification, object detection and some other involving spatial data.

In the context of DL, spatial data refers to the data having structure or meaning tied to space or position



→ pixels are relatively positioned.

- Image: pixels are arranged in a 2D grid: W x H

  2800 x 1800

- video: 3D frames: time x width x height.
  (T x W x H)
- Satellite Maps: (latitude x longitude grids)
- Medical Scans: MRI, CT → (video)
  width x Height x Depth.

Real and Industrial Applications of CNN

Remote Sensing – It is the process of acquiring information about the Earth's surface without physical contact typically through satellites, drones or aircrafts etc

↓

to study things like weather and its patterns, oceans, farms, forest, cities etc.

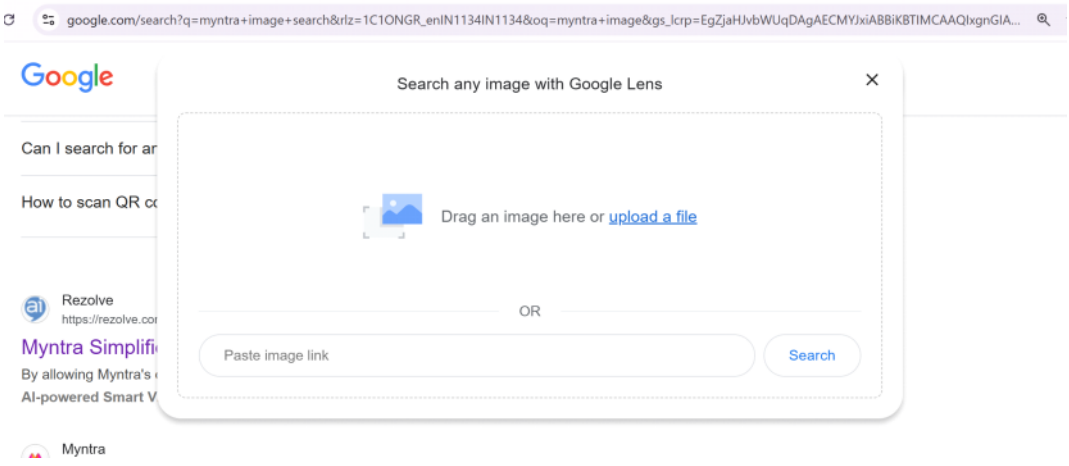
Google Lens:

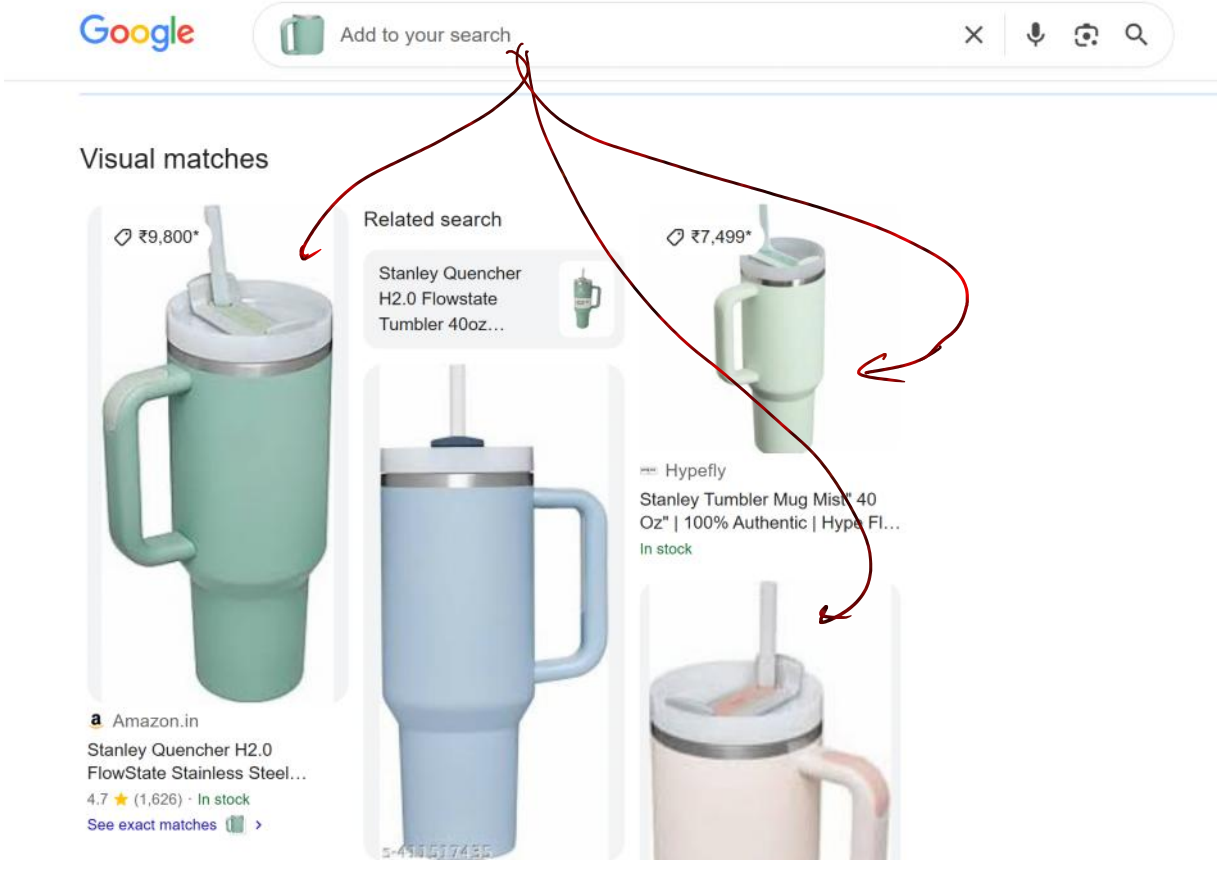
- to recognize objects in photos
- to detect and text from images (OCR)
- to identify landmarks
- to translate signs | languages in real time
- to do reverse image search → is a technique where you upload an image instead of text

[optical character Recognition
2015 → using 'Tesseract'
– (using R-programming)] API

image instead of text

"Myntra has this feature"





What happens when you "search by image"

① When an image is uploaded, the model uses deep neural networks to understand content

deep neural networks to understand content and semantics of the image.

↓

it turns the uploaded image into a vector representation

② Feature extraction via CNN model

↓

it produces an embedding vector – a high-dimensional numerical representation that captures

- object shapes
- colors
- textures
- composition
- context → mug with handle and straw

③ Semantic Understanding – to detect objects in the image (tumbler or cup or mug)
– to detect the text (using OCR)

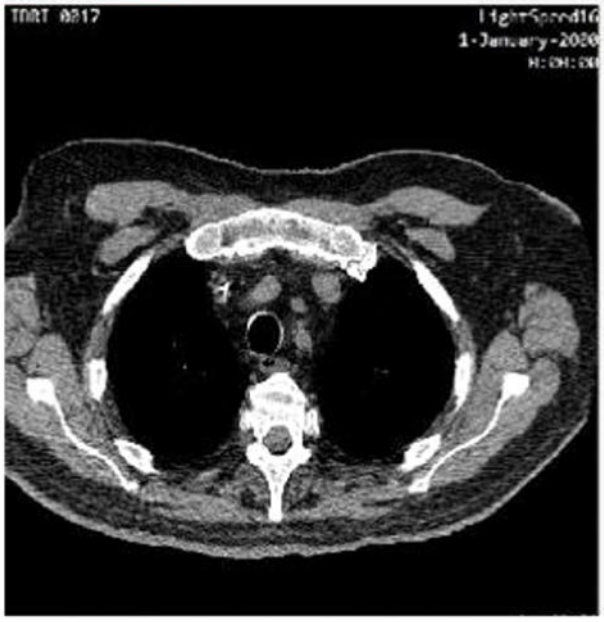
④ Vector Matching / Similarity Search

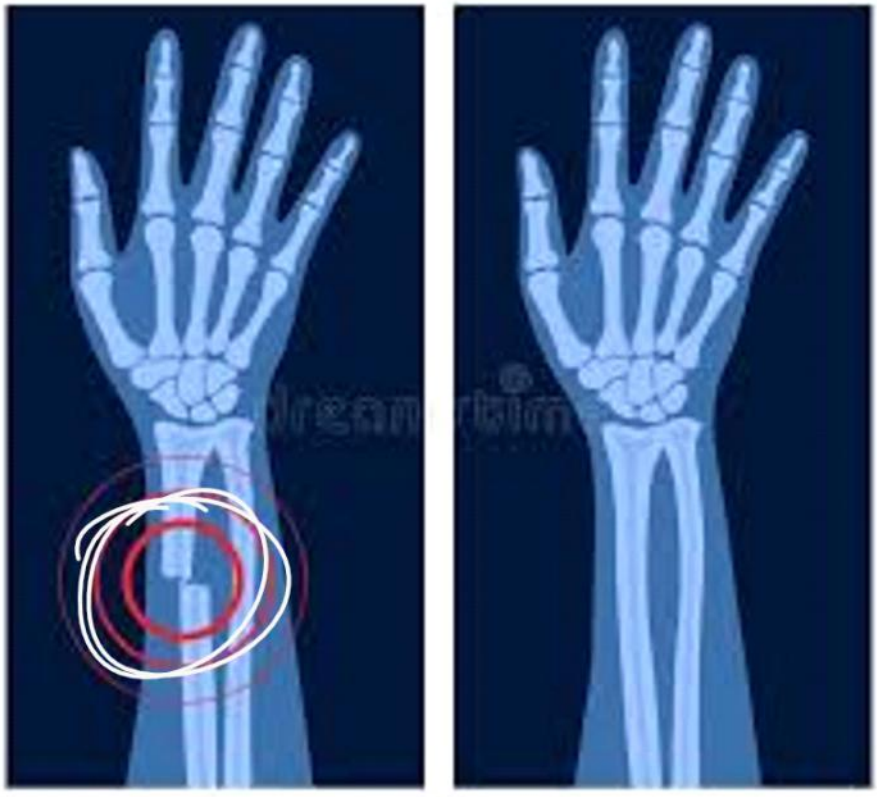
the uploaded image embedding is compared against billions of stored embeddings in Google's vector database.

Detecting a lung cancer using CNN

Advanced CNN model to detect / assist in finding lung cancer out of a CT scan.

Lung Cancer





https://www.schreibermd.com/humerus-fracture

③ In home / security cameras

Baby monitor camera

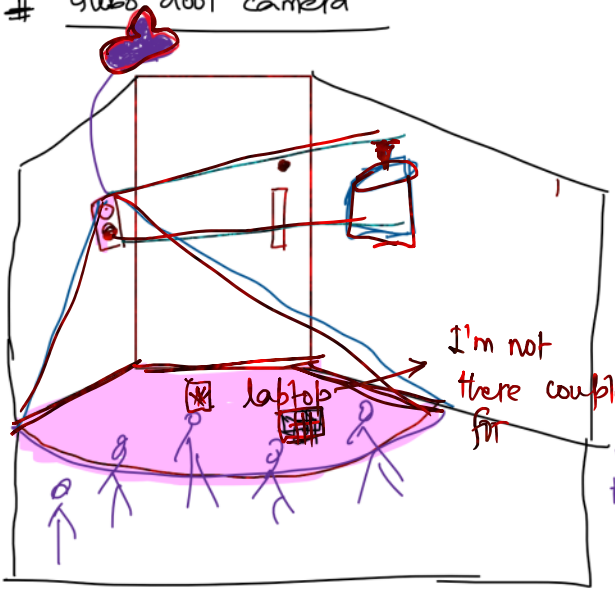
[Tapo] 2-5K

on ↳ if a baby is in distress - get triggers on the phone.

[Qubo] ↳ keep an eye on the babysitter

⇒ Qubo door camera

Qubo door camera

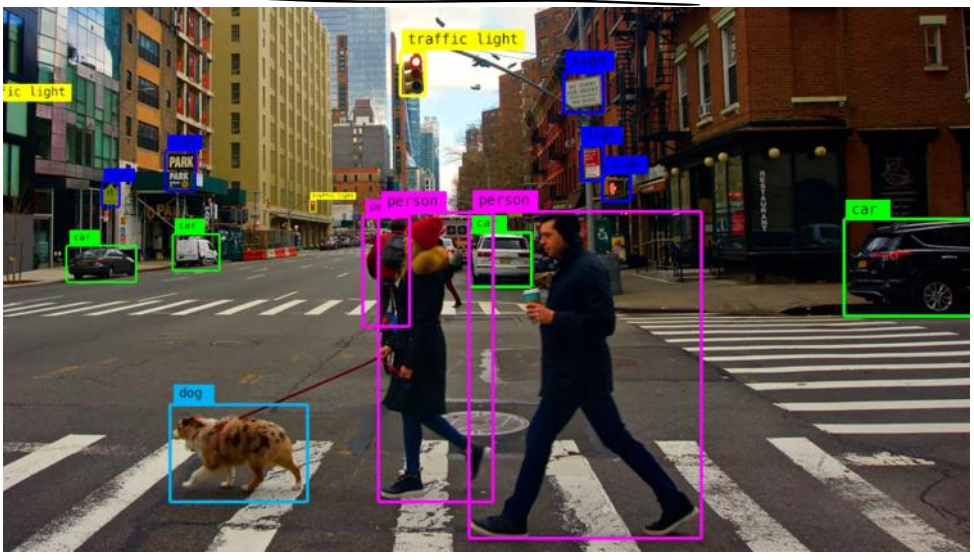


- challans for the traffic violations
  - without helmets
  - tripling
  - no seat belts.

- Digi Yatra - airport entry gate pass.
- Parking Management
- security system for face detection in surveillance.

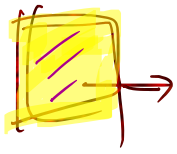
Object detection Algorithm - YOLO

→ You only look once. - (2016 Joseph Redmon)



Unlike image classification, object detection identifies what objects

Unlike image classification, object detection identifies what objects are present and where they are located with bounding boxes



cat → single label : Image classification
or
dog

↓

(one label per image)

Is computer vision (CV) a part of CNN or other way round ??

↓

This is not true ✗

CNNs are specialized type of deep learning algorithm (model) primarily designed for tasks involving image analysis and understanding.

CV is a field of AI that empowers computers to see, interpret and understand visual information from the world such as images, videos, any other visual inputs

library: opencv

- Image classification
- Object detection
- Image segmentation → (dividing an image into meaningful regions)
- facial recognition
- Object tracking
- motion analysis

* Pro-tip

CNN | YOLO | Computer vision.

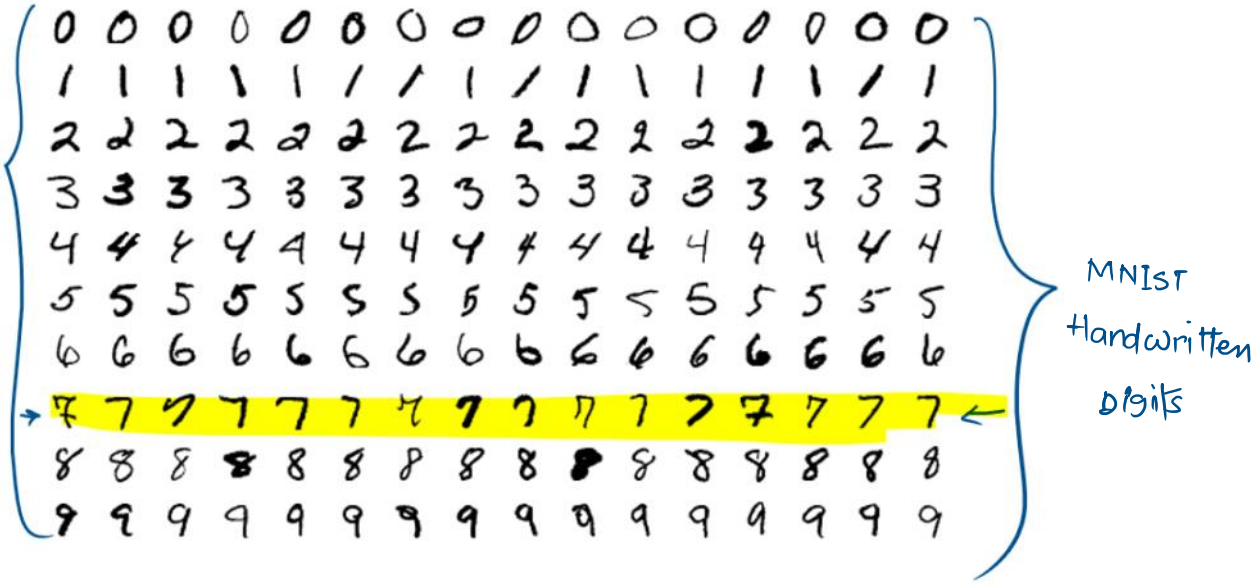
Task	Description
Image Classification	Identify what object is in an image (e.g., cat, dog)
Object Detection	Identify and locate multiple objects (bounding boxes)
Image Segmentation	Pixel-level classification (e.g., detect roads, cars)
Face Recognition	Identify or verify a person from an image
Pose Estimation	Detect human body parts and their positions
OCR (Optical Character Recognition)	Extract text from images
Visual Tracking	Follow moving objects in video streams
3D Reconstruction	Infer 3D structure from 2D images

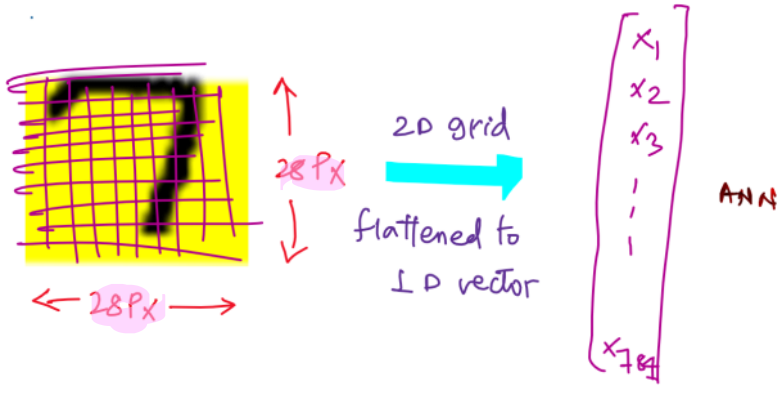
'CT angiography'

CIFAR-10 dataset

Why do we need CNN when we already have ANN ?
Or
How is CNN different from ANN ??

Reason # 1





Working of ANN

In ANN, each neuron in one layer is connected to

In ANN, each neuron in one layer is connected to every neuron in the next layer. To process an image in ANN, the image must be flattened into 1-D vector

©ope

Assume ANN is a Zigsaw scattered cards/ pieces

↓

and CNN is like a complete Zig saw picture

The issue with ANN is the process of flattening → as it ignores the 2-D spatial structure of the image

↓

Meaning the neural n/w doesn't explicitly captures the relationship b/w the neighbouring pixels.

↓

Idea is to maintain the structure of the image

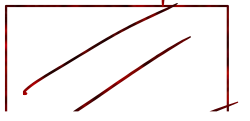
↓

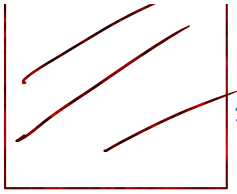
spatial relationships need to be preserved.

In MNIST, ANN would treat every pixel out of 786 pixels as a separate input without understanding how pixels near each other form edges or shapes etc.

(2) CNNs use reduced number of parameters through shared weights.

→ large input image ←





300 x 300 = 90,000 parameters

HL = 1

300*300=90,000

128

⇒ 90000 x 128 = parameters

Note: ANNs would require an enormous no. of parameters (weights) → making it very slow, memory-intensive and prone to overfitting

↓

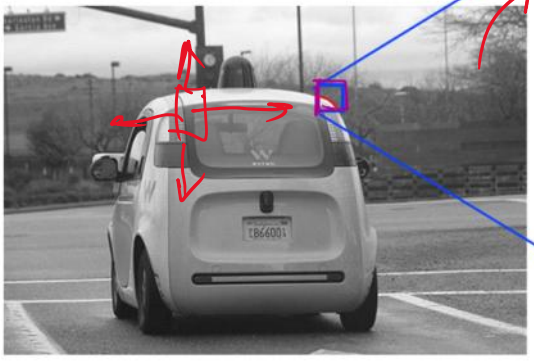
CNNs solve this problem by sharing weights (sliding the same filter across the image) and preserving the spatial information

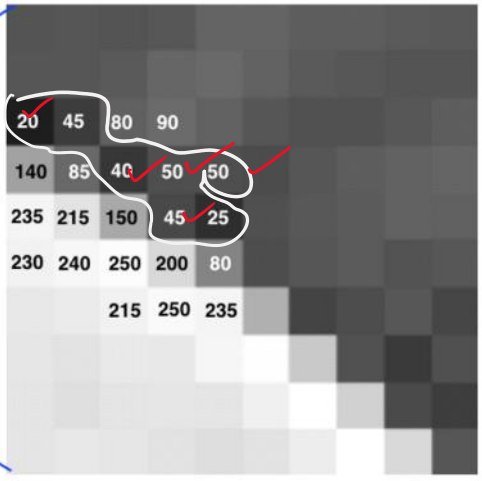
90000*128=11, 520,000

↓

[11.52 millions of parameters that too with just one hidden layer]

Given that in any image, nearby pixels are related, making it much faster and much more accurate





can a kid see the pattern?

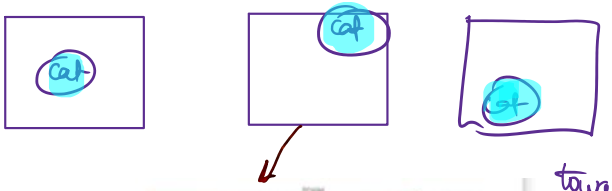
** Pro-tip

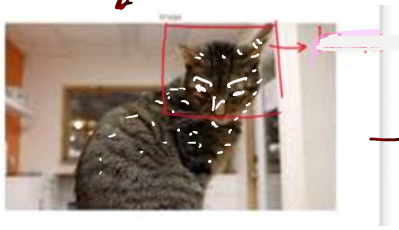
Reason#3 Translational and Rotational Invariance

Task: pls read about it !!!

Translational Invariance: The model correctly identifies an object if it moves (shifts) within the image.

Ex: a cat is still a cat whether it's in the top-left or bottom-right





towards top-right its not center-positioned.

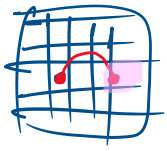
→ ANN → cat ??

will it correctly classify it as a cat or not.

In ANN, there's no spatial awareness and each input feature is treated independently

↓

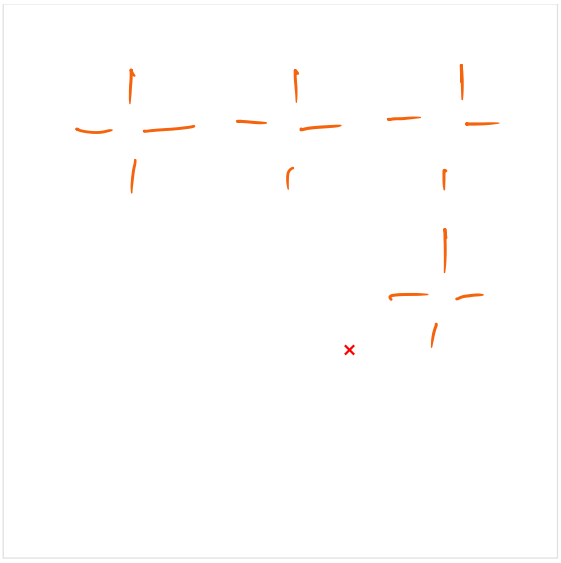
if a shift happens within an image let's say move the image 5 pixels to the right then pixels values go to completely different input neurons.



↓

ANN treats it as a totally new pattern.

So, a fully connected ANN has no built-in translational invariance.



ANN is really bad in handling of translations

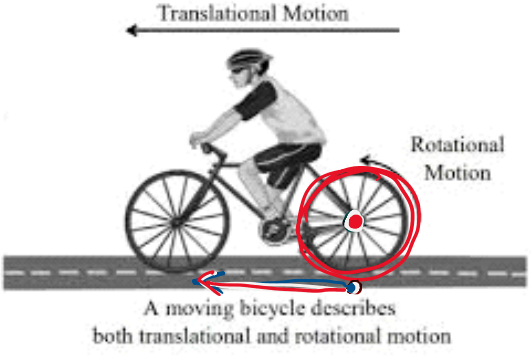
CNNs were specifically designed to handle spatial patterns – edges, shapes and textures – using shared convolutional filters

In CNNs, translational invariance is the core strength. convolutional filters (kernels) slide over the image and they learn to detect local patterns such as edges corners, shapes regardless of their position.

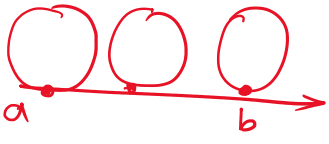
→ So the same filter is applied across the entire image and hence if a "cat ear" pattern is detected at the top-right or bottom-left, it activates the same set of neurons.

↓

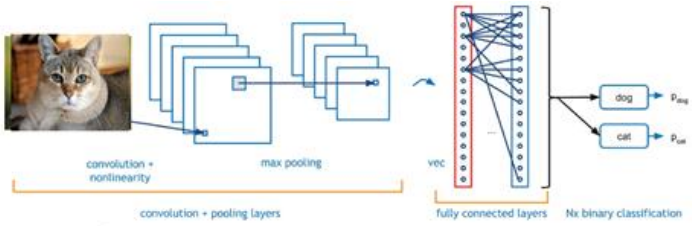
This provides translational invariance.

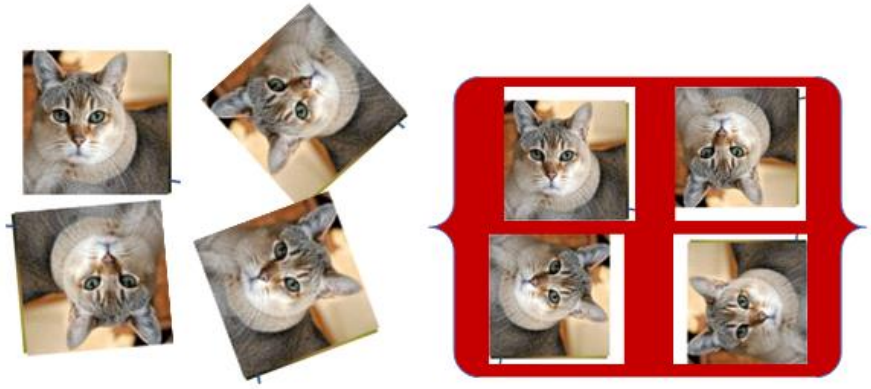


A moving bicycle describes both translational and rotational motion



Rotational Invariance:





It's a CAT no matter how does it look like after rotating

If an image is rotated say by 15°, 45°, 90°, 180°, 270° etc; the CNN should still recognize it as of the same class

However, CNNs are not rotationally invariant. In principle, standard CNN models don't perform well on rotated images as input.

↓

To some extent, CNN models can still handle small/slight rotations (upto 45°) because local features like edges or corners often activate the same filters.

To achieve the rotational invariance in the CNN model,

Data Augmentation :- add rotational versions of images during training

Take one original image → rotate it by some standard angles say 15°, 30°, 45°, 90°, 180°, 270° etc and add back to training dataset.

↑ AI summary for notes

... back to training dataset.

↑ AI summary for notes

Aspect	Example Visualization	Behavior in ANN	Behavior in CNN
Translational Invariance →	Object (e.g., a square or digit) moves across positions in the image	ANN fails — each pixel position is treated as unique; moving object changes input pattern completely	CNN recognizes the same object anywhere — shared filters slide across image (weight sharing)
Rotational Invariance →	Object is rotated at different angles (0°, 45°, 90°, etc.)	ANN fails — rotated input changes feature order entirely	CNN partially invariant — small rotations are tolerated due to local features; large rotations need augmentation or special architectures
Mechanism	–	Fully connected neurons (no spatial context)	Convolution + Pooling layers preserve spatial features
Result	–	Model sensitive to position and orientation	Model robust to object location; partly robust to rotation