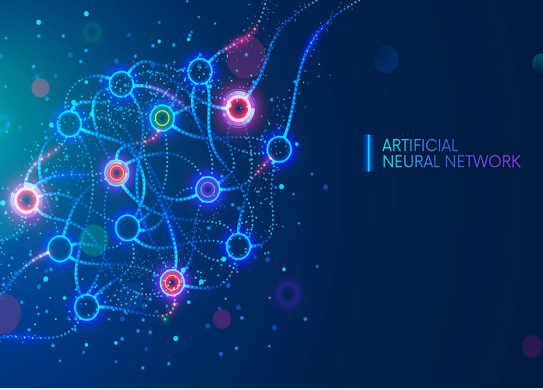




Computer-Vision

Dr. Seena Joseph

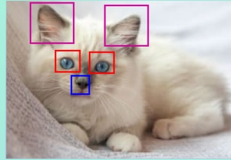
Computer Vision

- Replicate or surpass human visual perception capabilities
- Enable computers to interpret and understand the visual information.
- Allows machines to analyse and extract meaningful insights from images or videos
- Enabling machines to perform tasks such as
 - Image recognition
 - Object detection
 - Scene understanding
 - Image generation
- Applications
 - Autonomous vehicles
 - Medical imaging
 - Surveillance
 - Robotics
 - Augmented reality
 - Virtual reality
 - ...

Computer Vision

- Computer vision imparts visual capability to machines through cameras, data, models, and algorithms rather than retinas and the visual cortex.
- Modern computer vision algorithms rely on artificial neural networks
- Deep learning (DL) allows automatic feature extraction and can be utilized in many image processing tasks such as image classification, object detection and recognition, semantic classification
- Convolutional Neural Network (CNN) based DL architecture are well known for computer vision tasks.

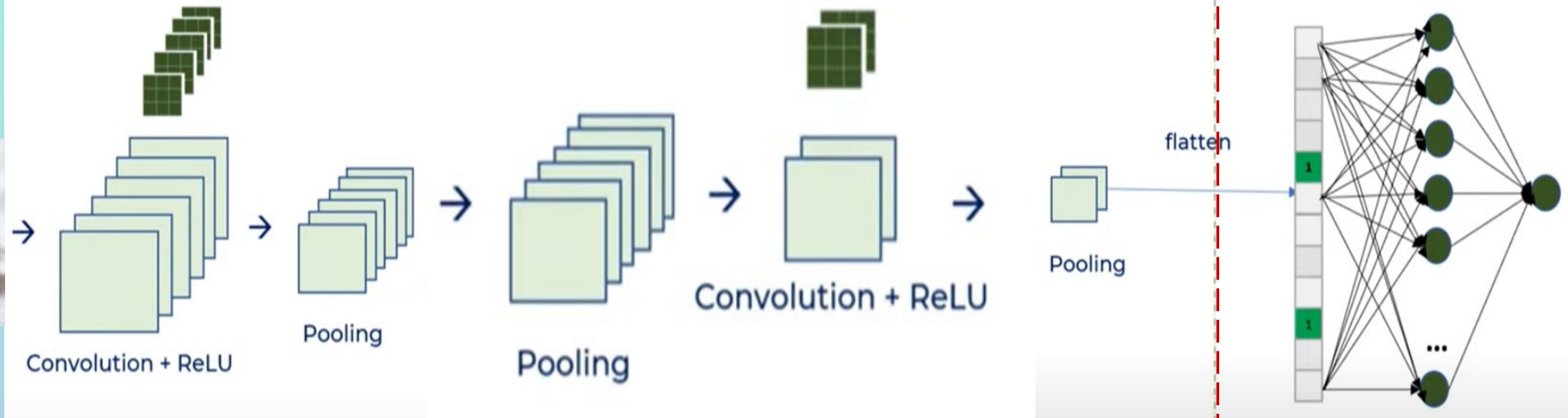
Computer Vision - CNN



Ear filter

Eye filter

Nose filter



Feature Extraction

Classification

Is this Cat ?

Computer Vision - CNN

- CNN Architectures

- LeNet
- AlexNet
- ZF Net
- GoogLeNet
- VGGNet
- ResNet
- YOLO
-

Computer Vision – Case study

Select any one of the CNN architecture for computer vision and present the following details:

- Introduction to the model
- Architecture overview (Explain the sequence of layers, the size and shape of feature maps at each layer, and any specialized modules or operations used)
- Implementation details , including the programming framework (pyTorch, TensorFlow)
- Training and optimization
- Performance Evaluation (Dataset, Metrics)
- Application

Computer Vision

Convolutional kernel has three dimensions : $L \times W \times D$ (Length x Width x Depth)

$L \times W$ know as the size of the kernel and normally it is 3×3 or 5×5

Number of channels or depth of the feature map is the number of feature maps output from each convolutional layer in the CNN; it is same as the number of sheets of the feature map.

number of channel is directly proportional to the feature extraction ability of CNN, however it increases the computational complexity .

Convolutional operation : process of sliding the kernel over the image ; multiplying the corresponding pixel value with the filter value and summing up to obtain a feature map.

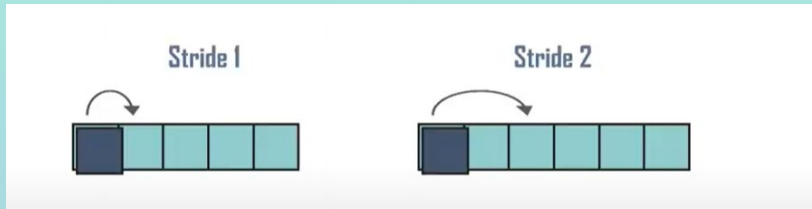
Computer Vision

Convolutional operation

Stride : movement of the filter over the input

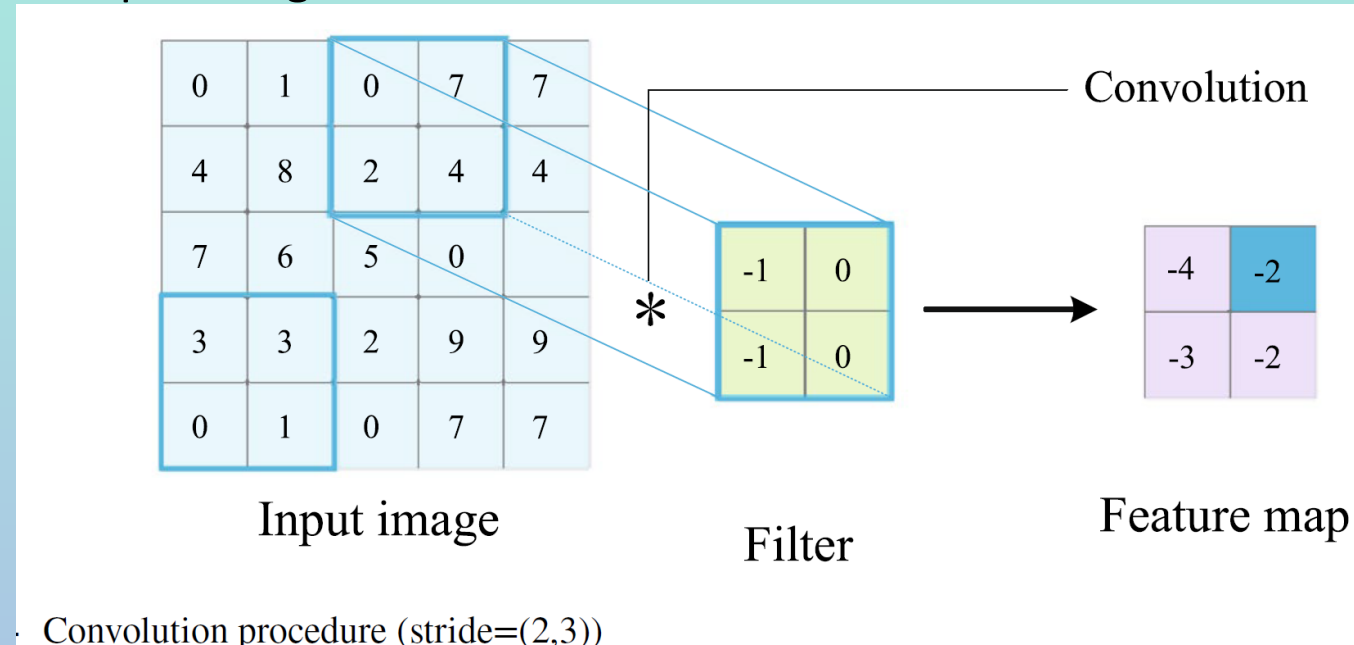
Image along the height and width dimension

Number of steps the filter takes to slide over the input image



Impact of stride

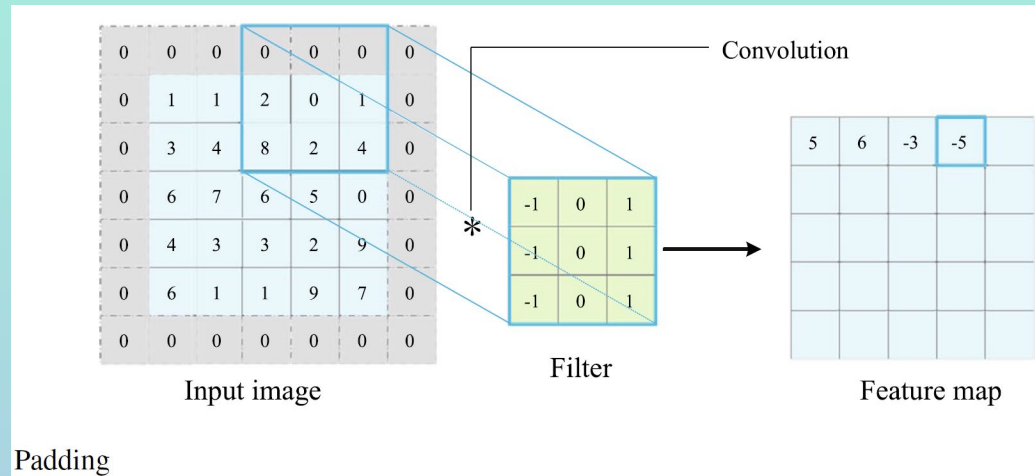
- Output data size (Dimensionality reduction)
- Computational Speed ?
- Model capacity ?



Source: (Zhao et al., 2024)

Computer Vision

- Padding is the process of adding a certain number of pixels to the edges of the input data so that the size of the output data can match the input data.
- Padding is used to avoid shrinking of feature map size at each layer.



Source: (Zhao et al., 2024)

How to calculate the feature map dimension based on stride and padding ?

Computer Vision

- Pooling Layer

Pooling layer will be added to down sample the feature map

Common pooling functions: Max pooling and average pooling

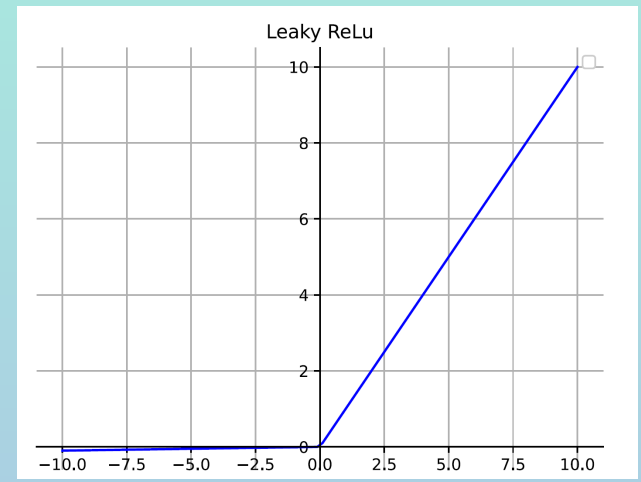
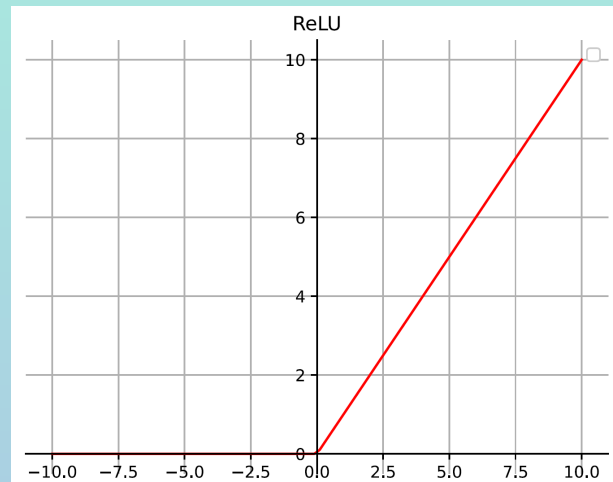
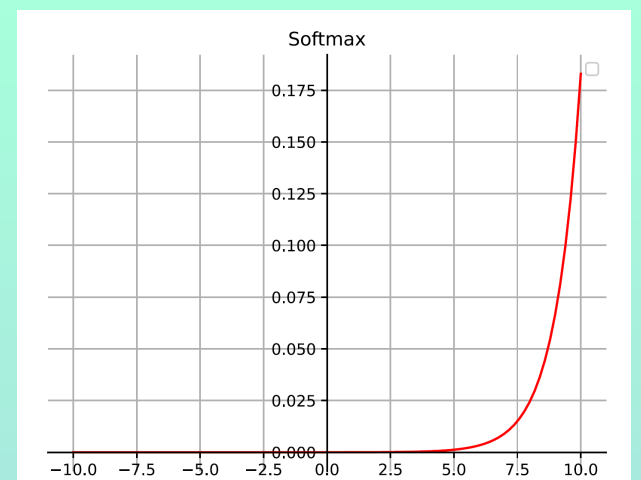
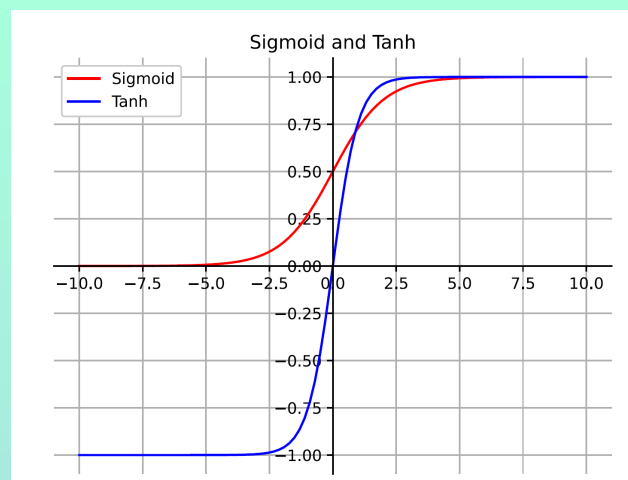
Impact of Pooling:

- Compress the data and parameters
- Reduction of feature map dimension
- Minimize overfitting
- Makes the network is invariant to smaller morphological changes
 - size of the pooling window can impact the network's performance and sensitivity to morphological changes

Computer Vision

- Activation Function

- Sigmoid
- Tanh
- Softmax
- ReLu
- Leaky ReLu



Source: (Zhao et al., 2024)

Computer Vision

Optimization: Gradient Descent

An optimization algorithm to minimise the objective function by iteratively adjusting the parameters in the opposite direction of the gradient of the objective function.

Regularization : Batch normalization

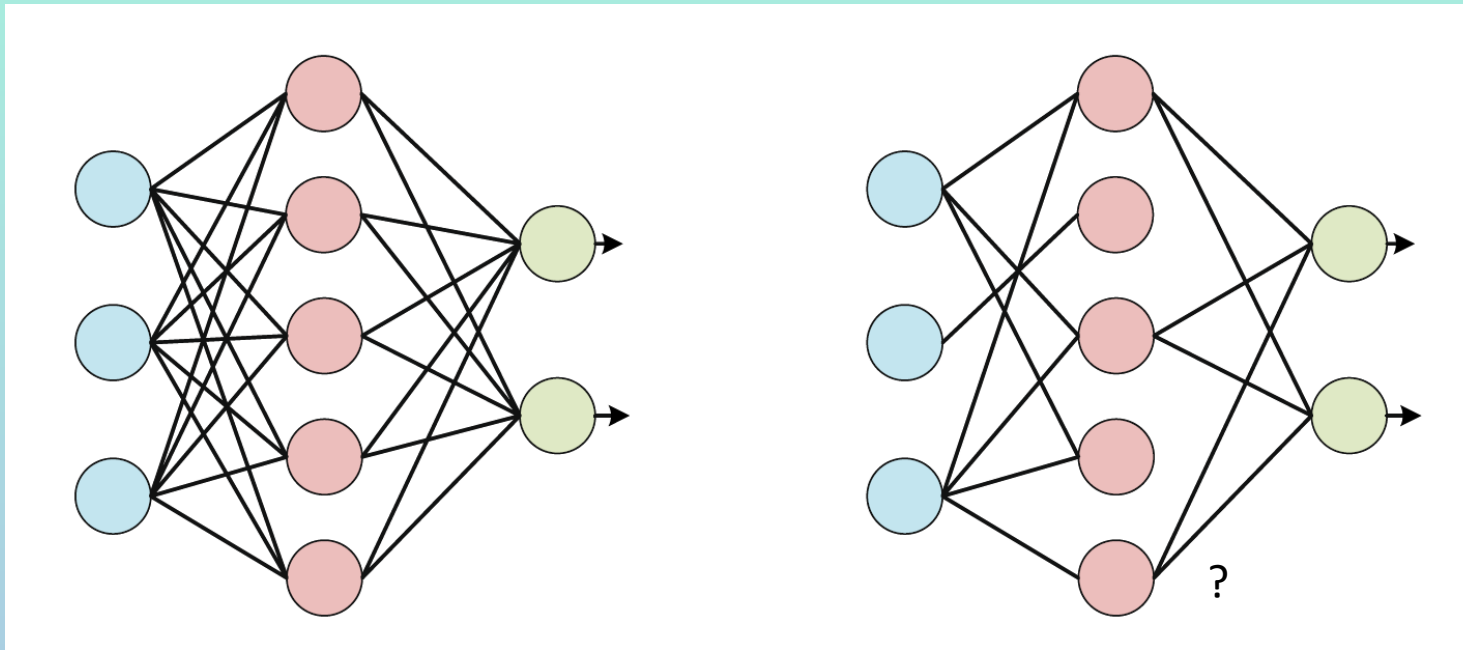
A technique used in deep learning to normalize the inputs to each layer of a neural network, ensuring stable and efficient training.

Computer Vision

Comparison	Gradient Descent	Batch Normalization
Purpose	Optimize the model parameters	Stabilize and speed up the training process
Objective	Minimize the error or loss function	Ensure the consistent inputs to each layer
Mechanism	Adjust the model parameter based on gradient of the loss function	Normalizes and scales inputs within each layer
Optimization	Aims to reduce the error of the model	Mitigate issues like vanishing gradient and accelerates convergence
Usage	General purpose optimization algorithm used in various machine learning models	Specific to deep learning architectures, especially neural networks

Computer Vision

Dropout and fully connected layer



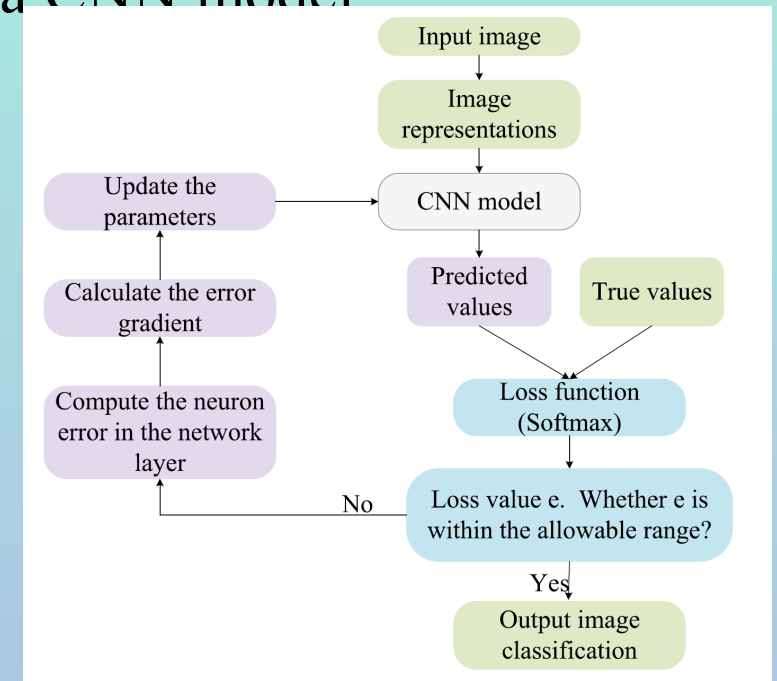
Source: (Zhao et al., 2024)

Computer Vision

Subtask: Image classification

Purpose is to differentiate between distinct classes of objects such as flowers, birds, animal, figures, vehicles based on the image properties.

The dataflow diagram of image classification based on a CNN model



Source: (Zhao et al., 2024)

Computer Vision

Image classification : CNN Models

AlexNet : First deep CNN model in modern times

Dual GPU, parallel convolutional operation

Visual Geometry Group (VGG)

small convolutional layer

increased network layer

VGG-16

GoogLeNet

Nine Inception V1 modules

Five down sampling layers

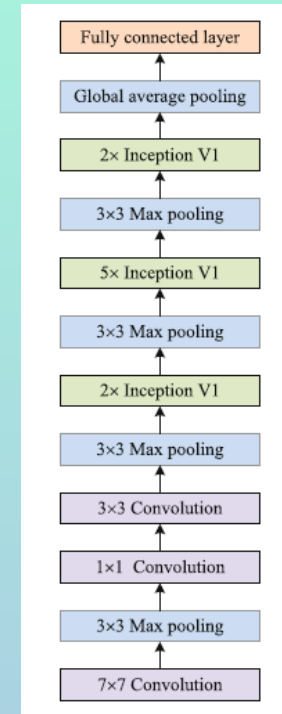
Deeper network layers but lesser parameters compared to VGG

Use sparse connection

Global average Pooling

.....

.....



Source: (Zhao et al., 2024)

Computer Vision

Performance comparison of different models

Models	Year	Depth	Parameters/M	Error rate/%
AlexNet Alom et al. (2018)	2012	8	60	16.4
VGG Sengupta et al. (2019)	2014	19	138	7.3
GoogLeNet Khan et al. (2019)	2015	22	4	6.7
ResNet Wightman et al. (2021)	2016	152	25.6	3.6
SENet Jin et al. (2022)	2017	152	27.5	2.3
MobileNetV1 Howard et al. (2017)	2017	–	4.2	5–10
MobileNetV2 Sandler et al. (2018)	2018	–	3.47	3–6

Top 5 error rate is reported for all architectures

Source: (Zhao et al., 2024)

Computer Vision

Subtask :Object detection

Object tracking; behaviour recognition

Traditional methods are dependent on the sliding window approach with feature descriptors such as Haar, SIFT (Scale Invariant Feature Transform), SURF (Speeded up Robust Features)

Modern methods

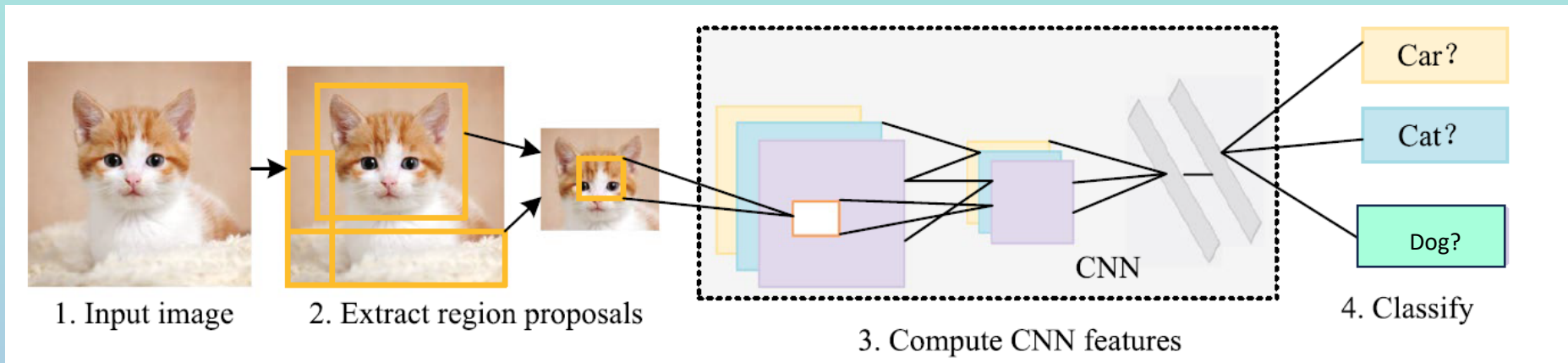
Regions based (Two stage model)

Regression based (One stage model)

Computer Vision

Object detection – Two stage detector

Example : Region based CNN (R-CNN)

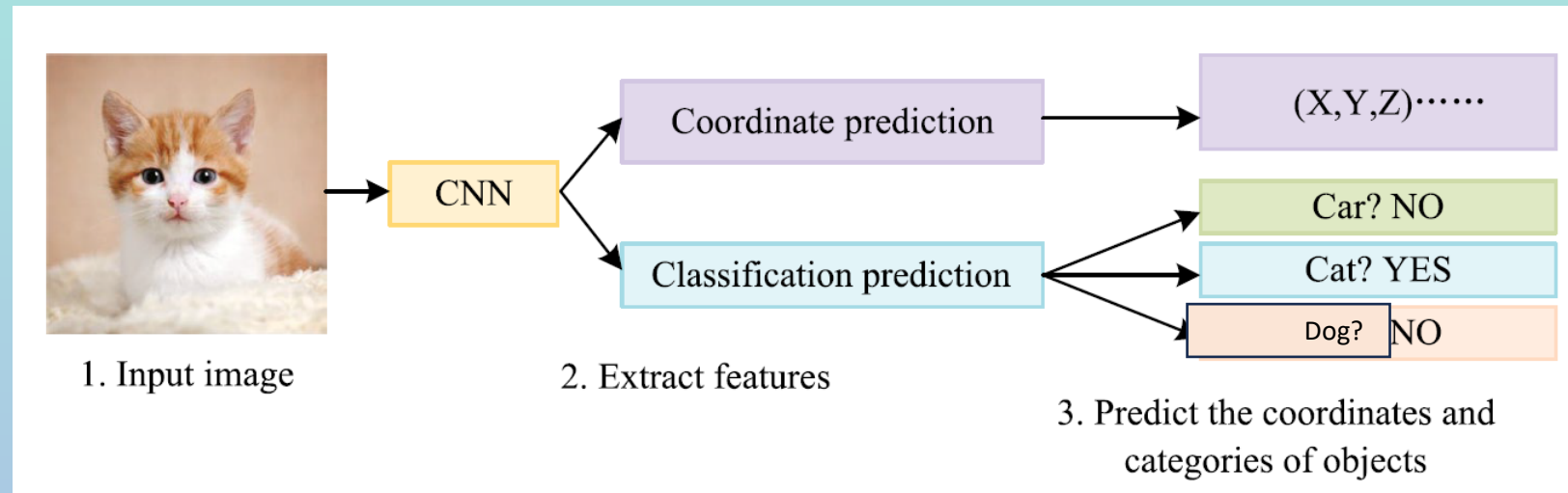


Source: (Zhao et al., 2024)

Computer Vision

Object detection – One stage detector

Example : YOLO



Source: (Zhao et al., 2024)

Computer Vision

- CNN – a deep feed forward multilayered hierarchical network
- Multiple layers for feature transformation
- Underlying feature representation of original data is gradually transformed to higher level feature representation
- Processed data is fed into a prediction function for final classification
- AlexNet – Image classification model
- Visual Geometry Group (VGG)
- GoogLeNet
- Residual Net (ResNet)
- Squeeze and Excitation Network (SENet)
- MobileNet
- Disadvantages of these models – overfitting and increased error rate especially with deeper and complex model (Zhao et al., 2024)
- Techniques such as network pruning, knowledge distilling and tensor decomposition are recommended to overcome the aforementioned shortcomings.

Computer Vision

- CNN – a deep feed forward multilayered hierarchical network
- Multiple layers for feature transformation
- Underlying feature representation of original data is gradually transformed to higher level feature representation
- Processed data is fed into a prediction function for final classification
- AlexNet – Image classification model
- Visual Geometry Group (VGG) - Image classification model
- GoogLeNet - Image classification model
- Residual Net (ResNet)
- Squeeze and Excitation Network (SENet)
- MobileNet
- Disadvantages of these models – overfitting and increased error rate especially with deeper and complex model (Zhao et al., 2024)
- Techniques such as network pruning, knowledge distilling and tensor decomposition are recommended to overcome the shortcomings.

Reference

- Zhao, X., Wang., Zhang Y., Han X., Deveci M., and Parmar M (2024). A review of convolutional neural network in computer vision, Artificial Intelligence Review, 57 (99).