## **What is CNN?**

A Convolutional Neural Network (CNN) is a type of deep learning algorithm specialized for processing structured grid data, particularly images. It automatically learns features through filter (kernel) optimization using convolution operations—mathematical techniques rooted in linear algebra. It is commonly used in image classification and object recognition, with applications spanning facial recognition, self-driving car, medical imaging, etc.

CNNs architecture mimics the visual cortex of a human brain, as they have an excellent performance and effectiveness than the traditional neural networks in recognizing visual patterns. CNNs have enabled a scalable, end-to-end approach, while before they have to manually feature extract for image analysis. CNNs remain a foundational architecture in deep learning.

## **What is its do in Machine Learning?**

**2.1 Convolutional Layers**

Convolution layers are the first element in CNN, designed to extract features from input image. The layers apply numbers of learnable filter across the input in a sliding window. Each filter detects particular sequences in the image, such as edges, curves, or shapes. During the process, they compute the dot product between the weights and image patches to produce the feature maps that highlight the position of the patterns.

**2.2 Pooling Layers**

The pooling layer comes after the convolutional layer and sweep process across the input, but the purpose of it is to reduce the dimensionality through down sampling, which lowers computations and parameter count. Thus, it can improve the network’s overall efficiency and prevent overfitting by decrease the complexity of the model.

**2.3 Fully Connected Layers**

The result of that is then flattened into one-dimensional matrix generated by the last pooling layer so it can be passed into a completely linked layer for regression task or categorization task with corresponding activation function.

* 1. **Activation Functions**

ReLU activation functions are applied after convolution and pooling layers to introduce non-linearity, helping the network to learn complex patterns between the features in the image. After all the network process, using SoftMax activation function to convert the final outputs into probability result for each class, the model will then predict the class based on the highest probability score.

## **Examples of real-world examples**

3.1 Self-driving car

With the CNNs, humanity achieving the self-driving technology from concept to reality. They extract visual features by installing camera to capture information and using sensor suite to combine environmental data, so the car can recognize the situation in front of the car with modular pipeline that includes perception, localization, prediction, decision-making, and motion control.

3.2 Medical Imaging

CNNs are commonly used in medical imaging to improve diagnostic accuracy and helping clinical decision-making. In histopathology, they assist in identifying and classifying colorectal polyps, gastric epithelial tumors, and other malignancies. In radiology, they help detection of conditions automatically like pneumonia, pulmonary embolism, and rectal cancer. In photography, they used for analyzing retinal diseases, skin conditions, and detecting gastric and colorectal polyps during endoscopic procedures. Also, in surgical laparoscopy, they offer real-time assistance by interpreting anatomical structures and guiding safe dissection zones. These applications show how CNNs are transforming healthcare by enhancing image interpretation, improving workflow, and supporting better patient outcomes.

## **Alternatives for CNN**

4.1 Vision Transformers

Vision Transformers (ViTs) apply the Transformer model architecture, which originally developed for natural language processing (NLP), to handle image classification. They divide images into small patches and treat each patch as a token, similar to words in a sentence. Using self-attention mechanisms, ViTs learn the relationships between these patches to understand the overall structure and content of the image.

ViTs require explicit positional encoding and large datasets due to their minimal inductive biases, while CNNs naturally preserve spatial structure and perform better on smaller datasets. CNNs are generally more computationally efficient, whereas ViTs are more flexible but resource-intensive due to the self-attention mechanism.

4.3 Capsule Networks (CapsNets)

Capsule Networks (CapsNets) were created to solve the main limitations of CNNs, especially the low performance in handle geometric transformations. CapsNets focus on positional invariance and often lose important spatial relationships, it aims for equivariance, it can recognize objects even when the orientation or configuration of the input images changes. Groups of neurons that output matrices representing both the presence and pose of features. These matrices are refined using a squashing function to maintain direction and magnitude. Instead of max pooling, CapsNets use dynamic routing, which forwards information based on agreement between capsules, preserving spatial relationships and improving context awareness. Although more computationally intensive, CapsNets offer a more interpretable and transformation-resilient approach, making them ideal for complex tasks like medical imaging and autonomous systems.