**ABSTRACT**

*Precision farming uses modern technologies such as satellite imagery, UAV images or field mapping to improve crop quality and profitability. Moreover, it optimizes the use of traditional resources. Therefore, this agricultural management system contributes to the development of sustainable agriculture, allowing to solve both economic and ecological problems, which are becoming more acute.*

*The ability to automatically monitor agricultural fields is an important capability in precision farming, enabling steps towards more sustainable agriculture. Precise, high-resolution monitoring is a key prerequisite for targeted intervention and the selective application of argo-chemicals.*

*The main goal of this study is developing a model for crop/weed segmentation and mapping framework that processes multispectral images obtained from an unmanned aerial vehicle (UAV) using UNET - a fully convolutional neural network (CNN).*

*Unmanned aerial vehicles (UAVs) or drones have been developed significantly over the past two decades, for a wide variety of applications such as surveillance, geographic studies, fire monitoring, security, military applications, search and rescue, agriculture, etc. They are able to cover large areas in a matter of a few minutes.*

*In agriculture, UAV-based remote sensing technologies are increasingly used to collect valuable data that could be used to achieve many precision agriculture applications, including crop/plant classification.*

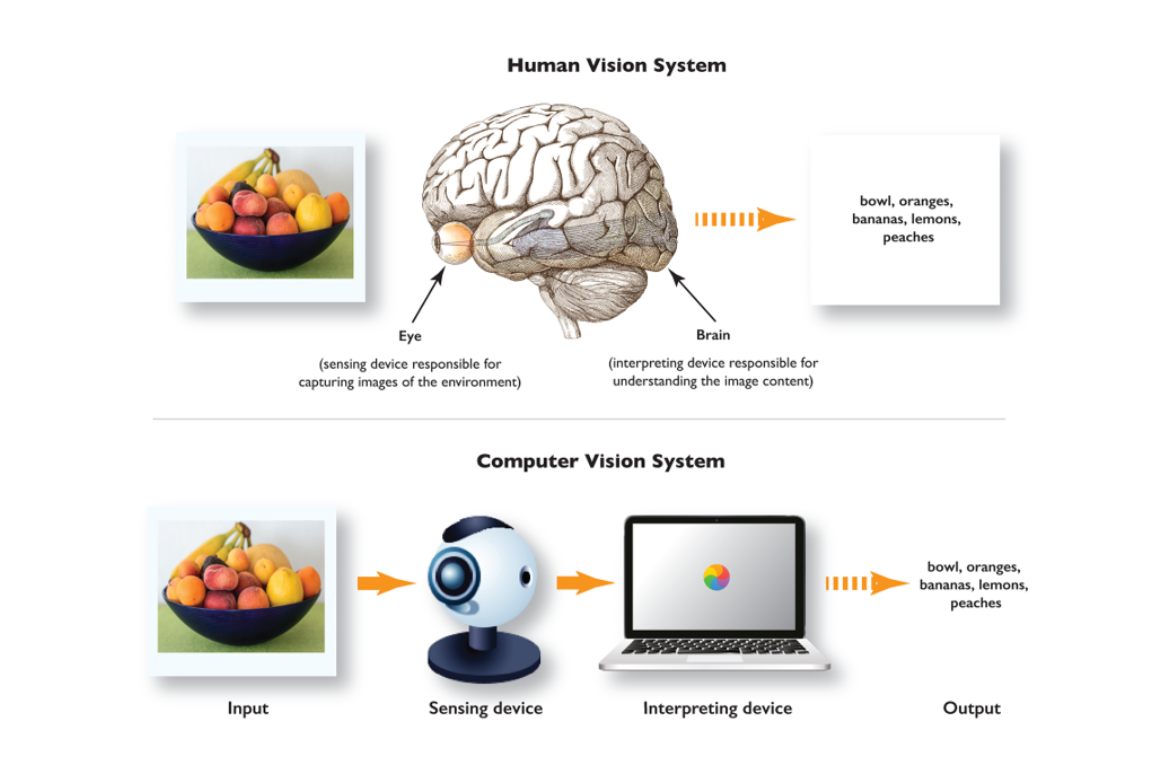
*In order to process these data accurately, powerful tools and algorithms such as Deep Learning approaches are necessary to handle the task. Recently, Convolutional Neural Network (CNN) has emerged as a powerful tool for image processing tasks achieving remarkable results making it the state-of-the-art technique for vision applications.*

*In this study, the UNet CNN architecture is reviewed to the UAV-based remote sensing image analysis for crop/plant classification to help researchers and farmers to decide what algorithms they should use according to their studied crops and the used hardware to classify different crop types accurately.*

## **Computer vision**

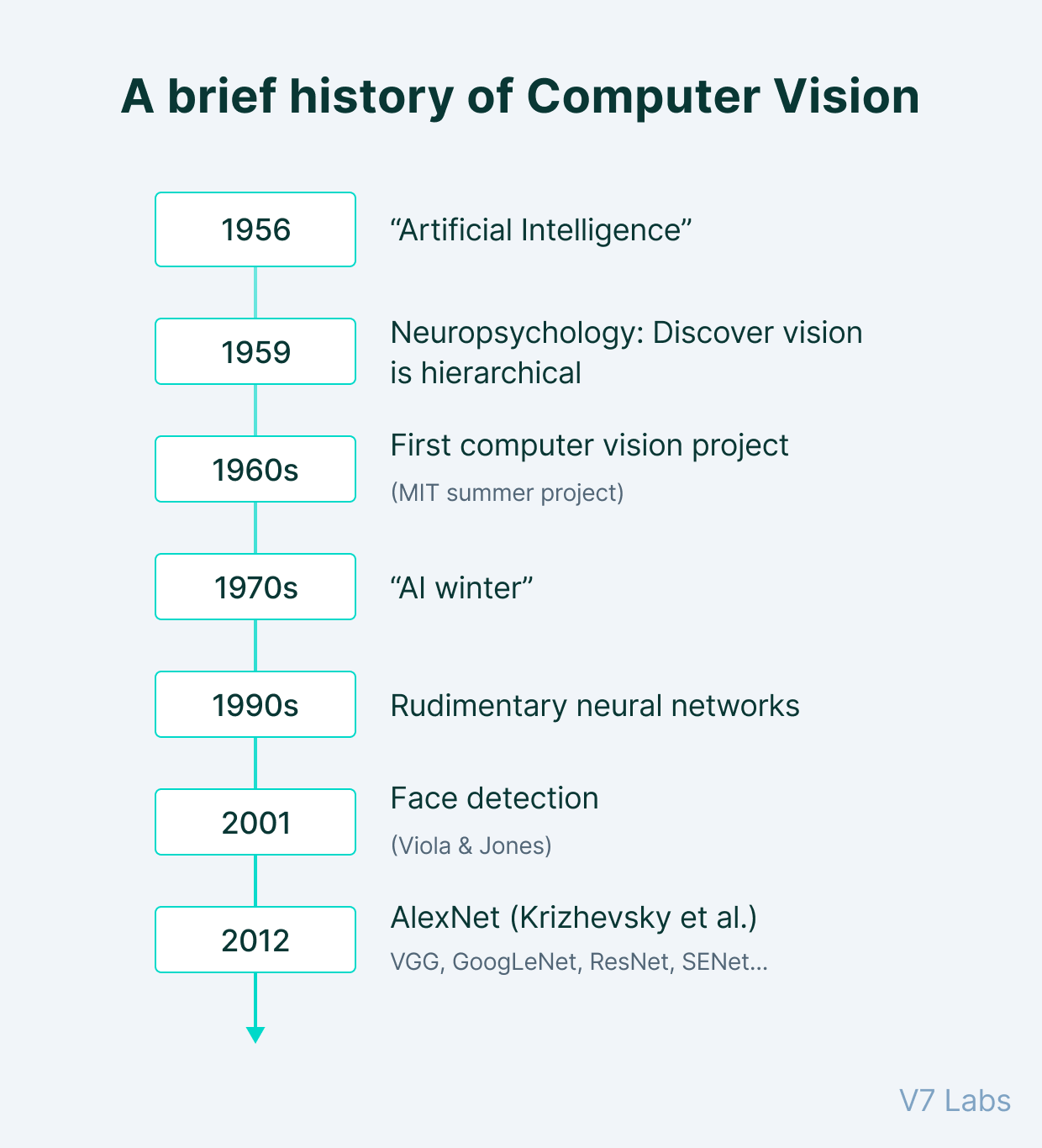
Computer vison is a field of artificial intelligence (AI) that deals with how computers can be made to gain high-level understanding from digital images, videos and other visual inputs.

From the perspective of engineering, it seeks to automate tasks that the human visual system can do. Human sight has the advantage of lifetimes of context to train how to tell objects apart, how far away they are, whether they are moving and whether there is something wrong in an image.



Picture

According to Wikipedia, in the late 1960s, computer vision began at the universities which were pioneering artificial intelligence. It was meant to mimic the human visual system, as a stepping stone to endowing robots with intelligent behaviour. In 1966, it was believed that this could be achieved through a summer project, by attaching a camera to a computer and having it “describe what it saw”.



(PICTURE https://www.v7labs.com/blog/what-is-computer-vision)

The notion that machines vision must be derived from the animal vision was predominant as early as 1959 – when the neurophysiologists mentioned above tried to understand cat vision.

Since then, the history of computer vision is dotted with milestones formed by the rapid development of image capturing and scanning instruments complemented by state-of-the-art image processing algorithms’ design.

The 1960s saw the emergence of AI as an academic field of study, followed by the development of the first robust Optical Character Recognition system in 1974.

By the 2000s, the focus of Computer Vision has been shifted to much more complex topics, including:

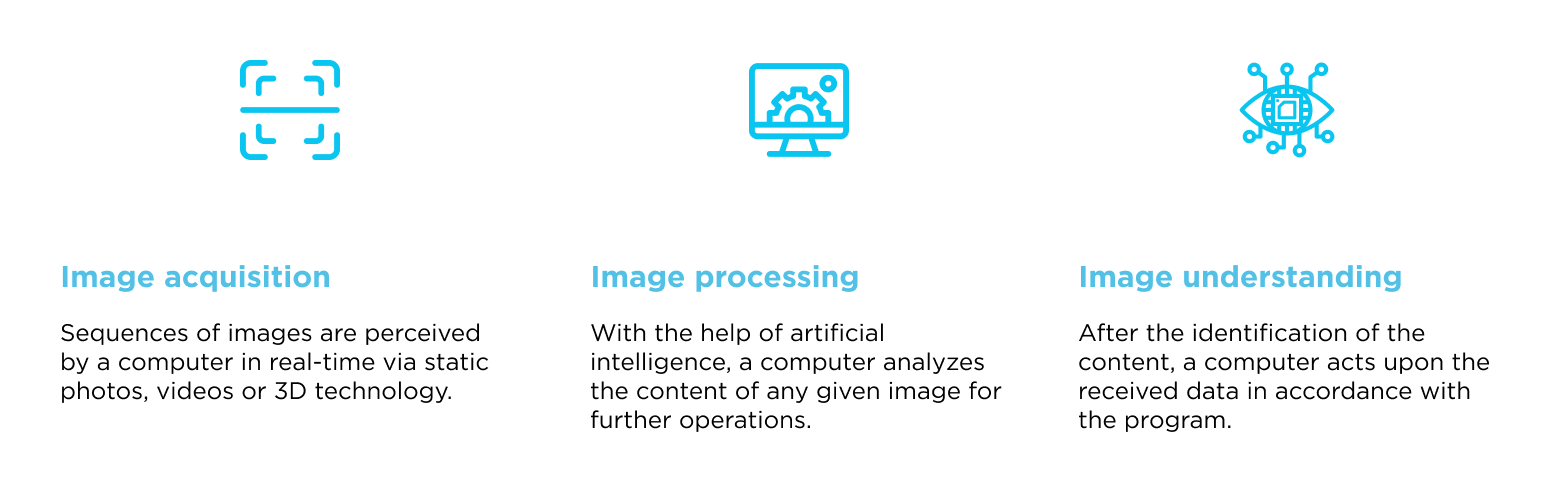
* Object identification
* Facial recognition
* Image segmentation
* Image classification

And more…

All of them have achieved commendable accuracies over the years.

Work of Computer Vision

Basically, a digital image is a 2-dimensions matrix with a lot of elements within, which often known as pixels. Each element in the image is a set of pixels which have the different colour with other sets. Therefore, when do the image processing, these things are converted to the equations based on matrices and vectors.



(PICTURE https://digitalskynet.com/blog/Comprehensive-Introduction-to-Computer-Vision)

In terms of video, the essence of a video is many pictures run continuously together, such as 10 fps, 25fps, 30fps, 60fps, 120fps or more, thus applying computer vision to video is essentially running algorithms on a series of images continuously at high speed.

Some operations commonly used in computer vision based on a Deep Learning perspective include:

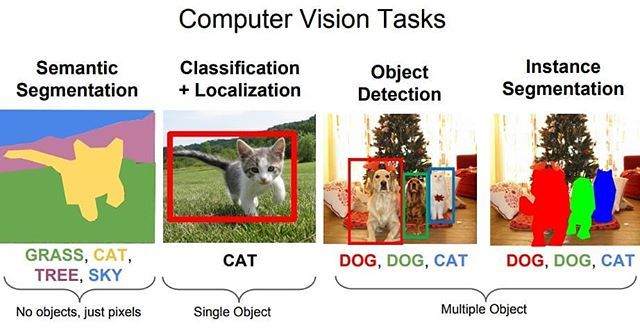
* Convolution: Convolution in computer vision is an operation in which a learnable kernel is “convolved” with the image. In other words—the kernel is slided across the image pixel by pixel, and an element-wise multiplication is performed between the kernel and the image at every pixel group.
* Pooling: Pooling is an operation used to reduce the dimensions of an image by performing operations at a pixel level. A pooling kernel slides across the image, and only one pixel from the corresponding pixel group is selected for further processing, thus reducing the image size., eg., Max Pooling, Average Pooling.
* Non-Linear Activations: Non-Linear activations introduce non-linearity to the neural network, thereby allowing the stacking of multiple convolutions and pooling blocks to increase model depth.

In essence, computer vision is used for tasks involving teaching computers to comprehend both digital images and visual information from the outside environment. This may entail taking information from such sources, processing it, and analysing it to make judgments.

Large-scale formalization of challenging problems into well-known, defensible problem statements was a hallmark of the development of machine vision.

Researchers from all over the world were able to recognize issues and effectively address them thanks to the topical division into well-defined areas with appropriate nomenclature.

The most common computer vision tasks that we frequently encounter in AI terminology include:



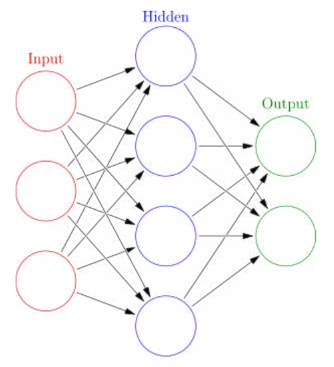
(PICTURE https://www.pinterest.co.kr/pin/861524603703486383/)

* Image segmentation:
  + Since the publication of the ImageNet dataset in 2010, one of the most researched subjects has been image classification.
  + Image categorization is the most common computer vision work undertaken by both novices and specialists, and the problem formulation is relatively straightforward.
  + The aim is to categorize a set of given photographs into a set of predefined classes using only a set of already identified sample images.
  + Image classification deals with processing the complete image and tagging it with a particular label, in contrast to more complicated issues like object identification and image segmentation, which need the localization (or assignment of coordinates for) of the features they discover.
* Object detection:
  + As the name suggests, object detection pertains to the identification and location of objects using bounding boxes.
  + Whenever they occur, class-specific characteristics are searched for in an image or video and identified via object detection. These classes can be whatever the detection model has been trained on, including persons, animals, or vehicles.
  + In the past, methods for object detection used features from images like as Haar, SIFT, and HOG to identify features and categorize them using traditional machine learning techniques.
* Image segmentation:
  + In addition to taking a long time and being mostly inaccurate, this approach has severe restrictions on the number of items that may be detected.
  + As a result, Deep Learning models like YOLO, RCNN, and SSD are frequently used for this task. These models use millions of parameters to overcome these restrictions.
  + Object recognition, often referred to as object classification, is frequently used in conjunction with object detection.
* Face and person recognition:
  + To show that a machine can distinguish an object from its surroundings and/or from another object in the same image, an image is segmented into smaller sections called sub-objects.
  + A "segment" of an image is a specific class of objects that the neural network has recognized in the image and may be extracted from using a pixel mask.
  + This well-known area of computer vision has been extensively researched using both conventional image processing techniques like clustering-based segmentation and watershed algorithms as well as well-liked contemporary deep learning architectures like PSPNet, FPN, U-Net, SegNet, etc.
  + The primary object being recognized in facial recognition, a subset of object detection, is the human face.
  + While facial recognition conducts not only detection but also recognition of the discovered face, it is comparable to object detection in that task, where features are detected and localized.
  + The placement of these landmarks and common characteristics, such as the eyes, mouth, and nose, are used by facial recognition systems to categorise a face.
  + Facial recognition techniques based on traditional image processing techniques include Haar Cascades, which is readily available through the OpenCV library. In works like FaceNet, more reliable Deep Learning-based algorithmic techniques are described.
* Edge detection:
  + The task of identifying boundaries in objects is known as edge detection.
  + Mathematical techniques that help detect abrupt changes or discontinuities in the brightness of the image are used in an algorithm to carry out this task. Edge detection is primarily carried out by conventional image processing-based methods like Canny Edge detection and by convolutions with specially built edge detection filters. It is frequently used as a data pre-processing step for numerous jobs.
  + Additionally, edges in an image provide us with essential information about the contents of the image. For this reason, all deep learning algorithms inherently perform edge detection to capture global low-level features with the aid of learnable kernels.
* Image restoration:
  + Picture restoration is the process of restoring or reconstructing old, fading image hard copies that were improperly captured and preserved, resulting in the loss of image quality.
  + Typical picture restoration techniques use mathematical methods to reduce additive noise, but occasionally, reconstruction calls for significant adjustments, necessitating additional analysis and the use of image inpainting.
  + Image inpainting involves filling in damaged areas of an image with the aid of generative models that estimate the message that the image is attempting to convey. After the restoration procedure, the subject of the image (if it was originally black and white) is frequently coloured in a way that is as realistic as feasible.
* Feature matching:
  + In computer vision, features are areas of an image that provide the most information about a certain object in the image.
  + While corners and even more localized, sharp details, such as edges, can function as features, edges are powerful markers of object detail and, therefore, crucial features. We can relate the features of similar regions in one image to those in another image by using feature matching.
  + In computer vision tasks like object identification and camera calibration, feature matching is used. The following steps are typically taken in this order when performing feature matching:
    - Detection of features
    - Formation of local descriptors
    - Feature matching
* Scene reconstruction:
  + Scene reconstruction, one of the most challenging issues in computer vision, is the digital 3D reconstruction of an item from an image.
  + Most scene reconstruction techniques function approximately by creating a point cloud at the object's surface and reconstructing a mesh from this point cloud.
* Video motion analysis:
  + The study of moving objects or animals and the trajectory of their bodies is known as video motion analysis, and it is a task in machine vision.
  + The task of object identification, tracking, segmentation, and pose estimation are particularly important components of motion analysis as a whole.
  + Motion analysis is used to count and track microorganisms like bacteria and viruses, as well as to analyse human motion in fields like manufacturing, sports, medicine, intelligent video analytics, and physical therapy.

c

## **Artificial Neural Network Definition**

According to Wikipedia, Artificial Neural Networks (ANNs) or connectionist systems are computing systems vaguely inspired by the biological neural networks that constitute animal brains. Such system “learn” (i.e. progressively improve performance on) tasks by considering example, generally without task-specific programming. For instance, in image recognition, they might learn to identify images that contain cats by analysing example images that have been manually labelled as “cat” or “no cat” and using the results to identify cats in other images. They do this without any a priori knowledge about cats, e.g., that they have fur, tails, whiskers and cat-like faces. Instead, they evolve their own set of relevant characteristics from the learning material that they process.



(PICTURE <https://www.studocu.com/en-au/document/monash-university/computation-intelligence-and-ai/artificial-neural-network-wikipedia/1689550>) (https://www.investopedia.com/terms/n/neuralnetwork.asp#:~:text=A%20neural%20network%20is%20a,organic%20or%20artificial%20in%20nature.)

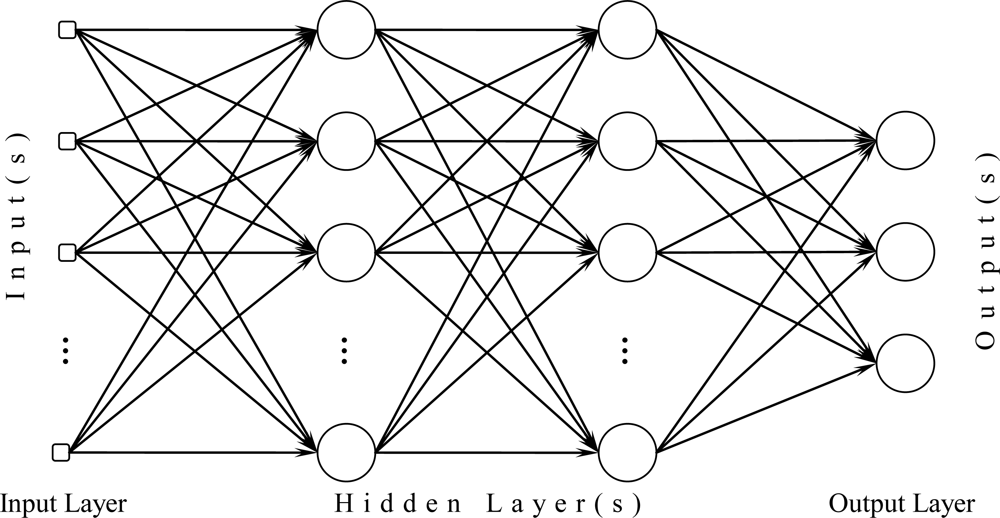
Artificial neurons are a group of interconnected units or nodes that form the foundation of an ANN (a simplified version of biological neurons in an animal brain). Each link (a condensed version of a synapse) between artificial neurons is capable of transmitting a signal. The artificial neuron that receives the signal can process it before signalling any associated artificial neurons.

In typical ANN implementations, the output of each artificial neuron is determined by a non-linear function of the sum of its inputs. The signal at a link between artificial neurons is a real number. Typically, the weight of artificial neurons and connections changes as learning progresses. The weight alters a connection's signal intensity by increasing or decreasing it. Artificial neurons might have a threshold that must be crossed in order for the signal to be transmitted. Artificial neurons are often arranged in layers. On their inputs, distinct layers may change those inputs in various ways. Signals move through the layers, perhaps more than once, from the first (input) to the final (output).

The original intent of the ANN technique was to approach issues in a manner similar to that of the human brain. However, over time, attention shifted to coordinating with particular tasks, which caused biological abnormalities. ANNs have been utilized for a range of activities, including medical diagnosis, playing board and video games, speech recognition, social network filtering, and computer vision.

## **Convolutional neural network**

A form of Artificial Neural Network (ANN) known as a Convolutional Network (CNN, or ConvNet) is used in deep learning and is most frequently used to evaluate visual imagery. Based on the shared-weight architecture of the convolution kernels or filters that slide along input features and produce translation-equivariant responses known as feature maps, CNNs are also known as Shift Invariant or Space Invariant Artificial Neural Networks (SIANN). Contrary to popular belief, most convolutional neural networks are not translation invariant because of the input. They are used in image and video recognition, recommender systems, image classification, image segmentation, and medical image analysis in addition to financial time series and natural language processing.



(PICTURE <https://www.google.com/search?q=2.3.+Convolutional+neural+network&source=lnms&tbm=isch&sa=X&ved=2ahUKEwiuksigiPX4AhUHfnAKHWXPALUQ_AUoAXoECAIQAw&biw=736&bih=780&dpr=1.25#imgrc=LI59tI77iMdWBM&imgdii=zwIZfLyynjv19M>)

Multilayer perceptrons are regularized variants of CNNs. Typically, multilayer perceptrons refer to completely linked networks, meaning that every neuron in one layer is coupled to every neuron in the following layer. These networks are vulnerable to overfitting data because of their "full connectedness." Regularizing parameters during training (such as weight decay) or cutting connectivity are common methods of regularization, or preventing overfitting (skipped connections, dropout, etc.) CNNs adopt a different strategy for regularization: they take advantage of the hierarchical pattern in the data and combine smaller and simpler patterns imprinted in their filters to create patterns of increasing complexity. CNNs are at the lowest end of the connectivity and complexity spectrum as a result.

Convolutional networks were influenced by biological processes [9][10][11][12] because of the way that neurons are connected to one another and how the visual cortex of animals is set up. The receptive field is a constrained area of the visual field where individual cortical neurons are exclusively responsive to stimuli there. Each neuron's receptive area partially encloses the whole visual field because to this arrangement.

Compared to other image classification methods, CNNs employ a comparatively low amount of pre-processing. Consequently, whereas in conventional algorithms these filters are manually constructed, the network learns to optimize the filters (or kernels) through automatic learning. A significant benefit is that feature extraction is free from the influence of human interaction and prior knowledge.

*The so-called "fully convolutional network," which was first put forth by Long, Shelhamer, and Darrell, is the ancestor of the U-Net architecture. [2]*

*The fundamental idea is to add additional layers to a typical contractual network, replacing pooling operations with upsampling operators. Thus, the resolution of the output is increased by these layers. Based on this knowledge, a succeeding convolutional layer can subsequently learn to put together a precise output. [1]*

*There are many feature channels in the upsampling portion of U-Net, which is a significant improvement that enables the network to pass context information to higher resolution layers. As a result, the expansive path produces a u-shaped design because it is roughly symmetric to the contracting component. There are no fully connected layers in the network; it just utilizes the valid portion of each convolution. [2] The missing context is extrapolated by mirroring the input image in order to forecast the pixels in the border region of the image. To apply the network to large images, this tiling technique is crucial since otherwise, the GPU RAM would impose a resolution restriction.*

*One significant change in U-Net is the abundance of feature channels in the upsampling section, which enables the network to relay context information to higher resolution layers. As a result, the expansive path produces a u-shaped structure because it is roughly symmetric to the contracting section. The network just use the legitimate portion of each convolution, without*

*UNet is an architecture for a fully convolutional neural network that specializes in image segmentation, also call semantic segmentation. Image segmentation or semantic segmentation is a procedure where it is not only predicts whether something specific is on a picture like a dog and a cat, but it is also able to create a mask that shows where the image that specific object is located and its dimensions. For easier imagination, it is able to think of semantic segmentation as a variation of classification where basically every pixel in an image gets assigned a class that it belongs to which. Then kind of forms the mask unit is fully convolutional since it contains only convolutional layers and does not contain any dense layer or also call fully connected layers.*

*It was initially developed in Germany by researchers at the university of Freiburg for medical image segmentation*

*Unet is initially developed as a single class segmentation model, but it can also be used for multi-class segmentation*

G J v b c m a k y p t h s r

A b c g h j k m p r s t v y

1. Gray, H., & Nuri, K. R. (2020). Differing visions of agriculture: Industrial‐chemical vs. small farm and urban organic production. American Journal of Economics and Sociology, 79(3), 813-832.
2. Gu, D., Andreev, K., & Dupre, M. E. (2021). Major trends in population growth around the world. China CDC weekly, 3(28), 604.
3. Jámbor, A., Czine, P., & Balogh, P. (2020). The impact of the coronavirus on agriculture: first evidence based on global newspapers. Sustainability, 12(11), 4535.
4. van Loon, M. P., Hijbeek, R., Ten Berge, H. F., De Sy, V., Ten Broeke, G. A., Solomon, D., & van Ittersum, M. K. (2019). Impacts of intensifying or expanding cereal cropping in sub‐Saharan Africa on greenhouse gas emissions and food security. Global change biology, 25(11), 3720-3730.
5. Beacham, A. M., Vickers, L. H., & Monaghan, J. M. (2019). Vertical farming: a summary of approaches to growing skywards. The Journal of Horticultural Science and Biotechnology, 94(3), 277-283.
6. Goodman, W., & Minner, J. (2019). Will the urban agricultural revolution be vertical and soilless? A case study of controlled environment agriculture in New York City. Land use policy, 83, 160-173.
7. Chamorro Martinez, J. A., Cue La Rosa, L. E., Feitosa, R. Q., Sanches, I. D. A., & Happ, P. N. (2021). Fully convolutional recurrent networks for multidate crop recognition from multitemporal image sequences. ISPRS Journal of Photogrammetry and Remote Sensing, 171, 188-201.
8. Mazzia, V., Comba, L., Khaliq, A., Chiaberge, M., & Gay, P. (2020). UAV and machine learning based refinement of a satellite-driven vegetation index for precision agriculture. Sensors, 20(9), 2530.
9. Adrian, J., Sagan, V., & Maimaitijiang, M. (2021). Sentinel SAR-optical fusion for crop type mapping using deep learning and Google Earth Engine. ISPRS Journal of Photogrammetry and Remote Sensing, 175, 215-235.
10. Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. (2017). Deep learning classification of land cover and crop types using remote sensing data. IEEE Geoscience and Remote Sensing Letters, 14(5), 778-782.
11. Mazzia, V., Khaliq, A., & Chiaberge, M. (2019). Improvement in land cover and crop classification based on temporal features learning from Sentinel-2 data using recurrent-convolutional neural network (R-CNN). Applied Sciences, 10(1), 238.
12. Yan, Y., & Ryu, Y. (2021). Exploring Google Street View with deep learning for crop type mapping. ISPRS Journal of Photogrammetry and Remote Sensing, 171, 278-296.
13. Yang, S., Gu, L., Li, X., Jiang, T., & Ren, R. (2020). Crop classification method based on optimal feature selection and hybrid CNN-RF networks for multi-temporal remote sensing imagery. Remote Sensing, 12(19), 3119.
14. Piiroinen, R., Heiskanen, J., Mõttus, M., & Pellikka, P. (2015). Classification of crops across heterogeneous agricultural landscape in Kenya using AisaEAGLE imaging spectroscopy data. International Journal of Applied Earth Observation and Geoinformation, 39, 1-8.
15. Ashapure, A., Jung, J., Chang, A., Oh, S., Yeom, J., Maeda, M., ... & Smith, W. (2020). Developing a machine learning based cotton yield estimation framework using multi-temporal UAS data. ISPRS Journal of Photogrammetry and Remote Sensing, 169, 180-194.
16. Yang, Q., Shi, L., Han, J., Zha, Y., & Zhu, P. (2019). Deep convolutional neural networks for rice grain yield estimation at the ripening stage using UAV-based remotely sensed images. Field Crops Research, 235, 142-153.
17. Théau, J., Gavelle, E., & Ménard, P. (2020). Crop scouting using UAV imagery: a case study for potatoes. Journal of Unmanned Vehicle Systems, 8(2), 99-118.
18. Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. ISPRS Journal of Photogrammetry and Remote Sensing, 66(3), 247-259.
19. Gao, Z., Luo, Z., Zhang, W., Lv, Z., & Xu, Y. (2020). Deep learning application in plant stress imaging: a review. AgriEngineering, 2(3), 29.
20. Hasan, M., Tanawala, B., & Patel, K. J. (2019, March). Deep learning precision farming: Tomato leaf disease detection by transfer learning. In Proceedings of 2nd international conference on advanced computing and software engineering (ICACSE).
21. Sa, I., Popović, M., Khanna, R., Chen, Z., Lottes, P., Liebisch, F., ... & Siegwart, R. (2018). WeedMap: A large-scale semantic weed mapping framework using aerial multispectral imaging and deep neural network for precision farming. Remote Sensing, 10(9), 1423.
22. Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.
23. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
24. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).
25. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).