Neural Networks and Images

H.L.N.Himanshi

January 15, 2024

In this section, we'll talk about how neural networks process and recognize images. Neural networks have revolutionized the field of computer vision by enabling machines to recognize and analyze images. They have become increasingly popular due to their ability to learn complex patterns and features. Especially convolutional neural networks (CNN), are the most popular type of neural network used in image processing.
But also vision transformers (ViT) have become more and more popular in recent times due to the breakthrough achievements of generative pre-trained transformers (GPT) and other transformer-based architectures in natural language processing.

Neural Networks and Image Processing

Generally speaking, neural networks process and recognize images in different ways. It depends on the network architecture and the problem we must solve. Some of the most common problems that neural networks solve with images include:

- **Image classification** – includes assigning a label or category to an image. For example, whether an image contains a cat or a dog.
- **Object detection** – identifying and detecting objects within an image.
- **Image segmentation** – involves converting an image into a collection of regions of pixels represented by a mask or a labelled image.
- **Image generation** – generating new images based on certain criteria or characteristics.

Image Classification

Image classification is the most popular task in computer vision, where we train a neural network to assign a label or category to an input image. This can be accomplished using various techniques, but the most common are convolutional neural networks (CNN).

Convolutional Neural Networks

CNNs comprise multiple layers, including convolutional layers, pooling layers, and fully connected layers. The convolutional layers are the heart of the network and are responsible for learning features from the input image. Specifically, they apply a series of filters to the image, each capturing a particular pattern or feature, such as edges, textures, or shapes.
For example, in the image below, we see the matrix I to which we apply convolution with the filter K. It means that the filter K goes through the whole matrix I, and element-wise multiplication is applied between the corresponding elements of the matrix I and the filter K. After that, we sum the result of this element-wise multiplication into one number:

Convolutional Neural Networks

CNNs comprise multiple layers, including convolutional layers, pooling layers, and fully connected layers...

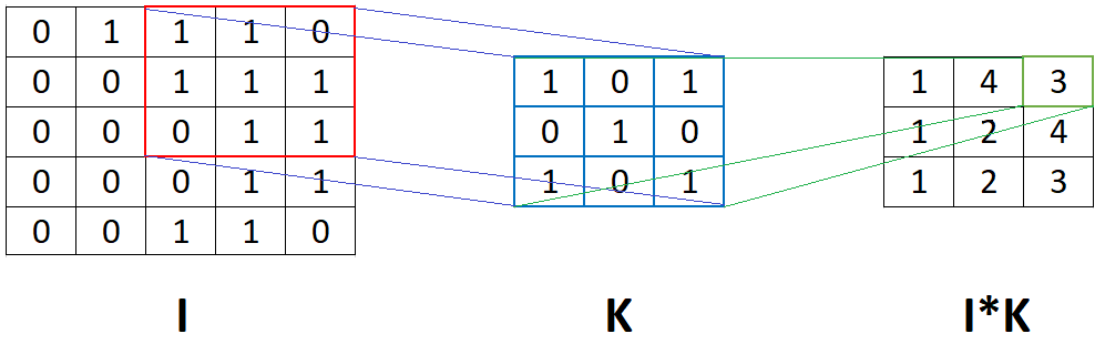

Figure: Convolution operation example.

ReLU Activation and Pooling

The ReLU activation function is commonly used after the convolutional layer, followed by a pooling layer. The pooling layer applies filters in the same way as the convolutional layer but only calculates the maximal or average item instead of convolution. In the image below, we can see the example of the convolutional layer, ReLU, and max pooling:

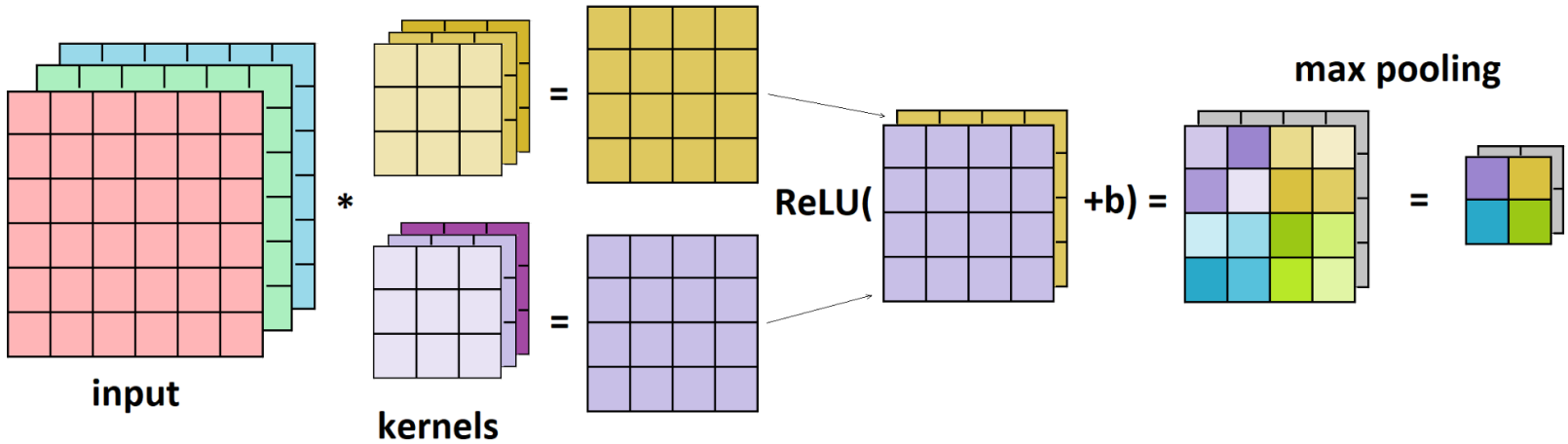

Figure: Illustration of a convolutional layer followed by ReLU activation and max pooling.

Object Detection

Object detection is detecting objects within an image or video by assigning a class label and a bounding box. For example, it takes an image as input and generates one or more bounding boxes, each with the class label attached.
Object detection is a combination of two tasks:

- Object localization
- Image classification

Algorithms for object localization identify an object's location in an image and indicate its position by drawing a box around it. These algorithms take an image containing one or more objects as input and provide the location of the objects by specifying the position, height, and width of the bounding boxes:

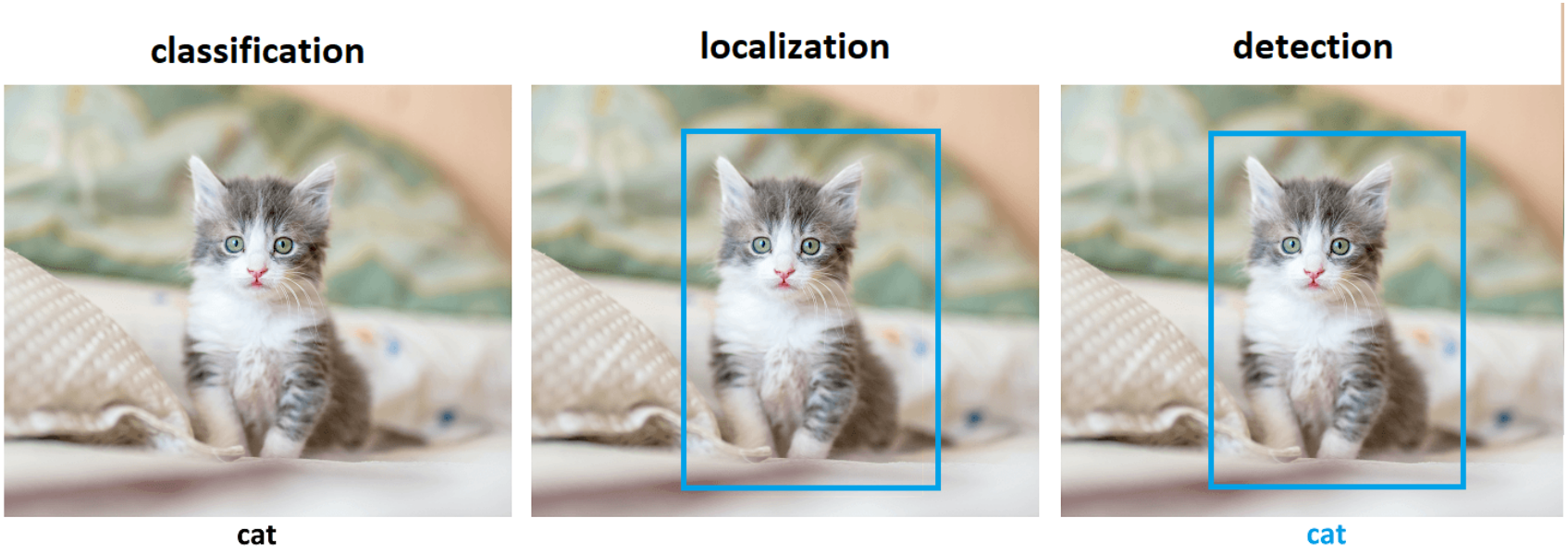

Figure: An example of object localization and detection.

CNNs in Object Detection

Similar to image classification, CNNs are commonly used for object detection. We can train the CNN on a dataset of labelled images, each with bounding boxes and class labels identifying the objects in the image. During training, the network learns to identify and classify objects in the image and locate them using bounding boxes. The most popular neural network architectures for object detection are:

- You Only Look Once (YOLO)
- Region-Based Convolutional Neural Networks (R-CNN, Fast R-CNN, etc.)
- Single Shot Detector (SSD)
- Retina-Net

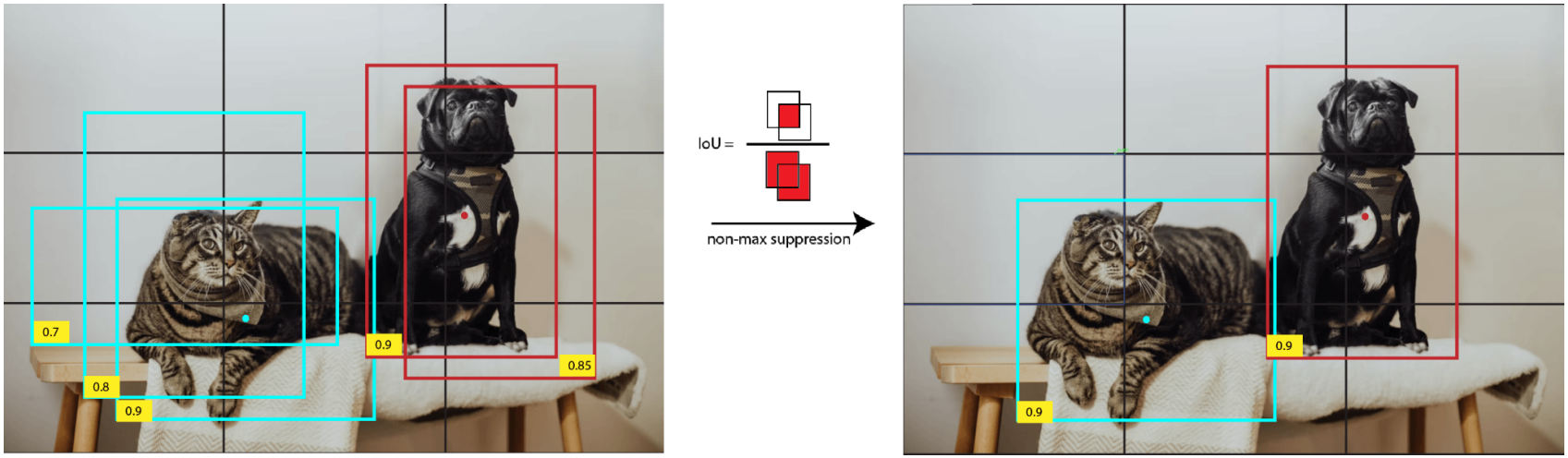

Figure: The process of YOLO object detection and non-max suppression.

Image Segmentation with Neural Networks

Neural networks are a popular tool for image segmentation, which is the process of partitioning an image into multiple segments, or sets of pixels. There are several types of image segmentation that can be performed using neural networks:

- **Semantic segmentation** - Assigning a class label to every pixel in the image.
- **Instance segmentation** - Distinguishing individual objects within the same class.
- **Boundary detection** - Identifying the edges of objects within an image.
- **Panoptic segmentation** - Combining semantic and instance segmentation to provide a comprehensive understanding of the scene.

Semantic Segmentation

Semantic segmentation involves assigning a class label to every pixel in the image. This means that if there are two or more objects of the same class in the image, semantic segmentation will return a single mask that includes all objects of that class.

Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) are a class of artificial intelligence algorithms used in unsupervised machine learning, implemented by a system of two neural networks contesting with each other in a zero-sum game framework. This technique can generate photographs that look at least superficially authentic to human observers, having many realistic characteristics.

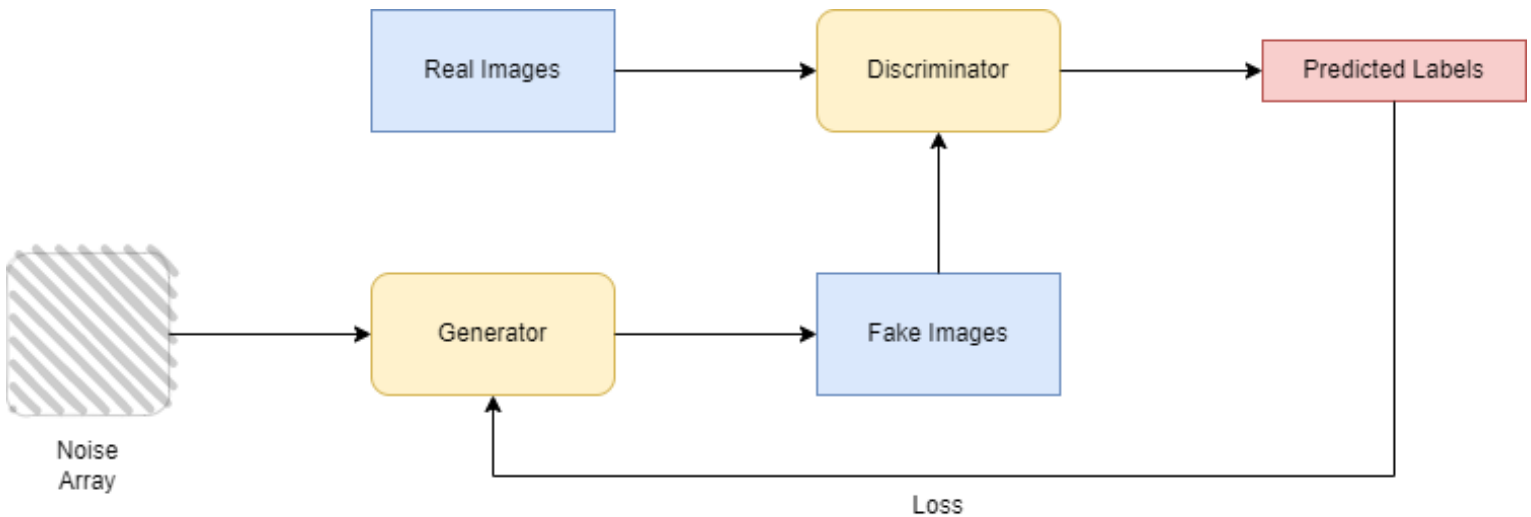

Figure: Diagram of a Generative Adversarial Network architecture.

The generator network generates new data instances, while the discriminator network evaluates them. The process is a competition

Variational Autoencoders (VAEs)

Variational Autoencoders (VAEs) are a type of neural network that consists of two main parts: an encoder and a decoder. The encoder network takes the input data and compresses it into a lower-dimensional representation in the latent space. This encoded data is then passed to the decoder network, which attempts to reconstruct the input data from the latent space representation. By sampling different points from the latent space, VAEs can generate new images that share characteristics with the input images they were trained on. This makes VAEs particularly useful for tasks like image generation, where we want to create new images that are variations of a learned dataset.

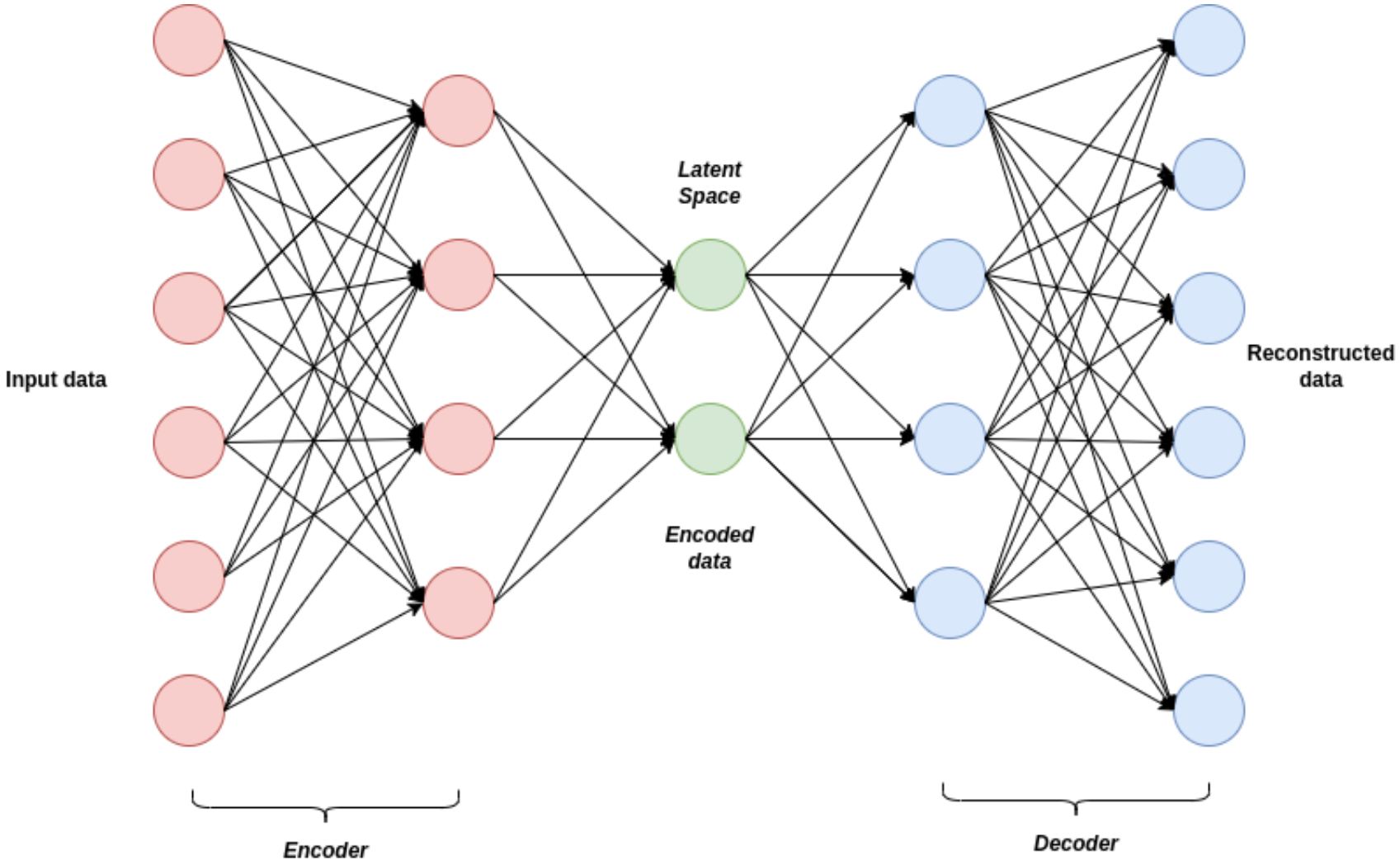

Figure: The architecture of a Variational Autoencoder, showing the encoder and decoder components.