ISSN 2959-6157

From CNN to GAN: The evolution of deep learning techniques in image processing

Yanghao Wu

Chengdu Institute of Educational Science Affiliated School,Chengdu,611100,China

Abstract:

In recent years, the technique for image processing has developed to improve the visual effects of images. This paper introduces ways to express images and basic operations. For the rest part of this paper, we elaborate on the current techniques for image processing including CNNs, RNNs, and GANs. each of them has different benefits and drawbacks. This paper also have further introduction about the theory and structure of neural networks above, and the reason for creating GANs, which is two adversarial network to improve the visual effects.

Keywords: Image Processing, deep learning, Convolutional Neural Network , Generative Adversarial Networks

1 Introduction

Many of the techniques, including image enhancement, encoding, decoding, restoration, and compression, of image processing developed in the 1960s at the Bell laboratories,Jet Propulsion Laboratories, et al. At that moment, image processing was applied to satellite imagery, wire-photo standards conversion, medical imaging, and character recognition, aiming to improve the visual effects for human beings. Before the emergence of cheaper computers and hardware becoming accessible in the 1970s, the cost of image processing was desperately high. This revolutionary improvement led image processing to real-time. With the coming of fast computers and signal processing becoming accessible in the 2000s, digital image processing has become the most common method to process images and the cheapest as well. In this article, we introduce and analyze the basic concepts of a network. There are four sections in total. The second section, which is the first section of the body, talks about the basic techniques for image operations including transformation, translation, zooming, and rotation et al. In the third section, we will introduce the RNNs, a deep learning model trained to process and convert a sequential data input into a specific sequential data output, including the input layer, output layer, and hidden layer. in this Part, the utility and application of the Convolution and pooling layers will also be covered. However, most of the time there are noticeable discrepancies with the real images, to solve this issue, GANS was born at the right moment, an adversarial network comprised of a discriminator and a generator. These two systems will optimize with each other, til the generated image cannot be recognized as an AI-generating image by human beings.

2 The basic concept of image

In the field of image processing, understanding the basic concepts of different types of images is essential. This section will explore three fundamental types of images: grayscale images, binary images, and indexed images.

2.1 Categories of image

Grayscale images are a type of image that uses 8 binary numbers to represent Iuminance and light intensity by pixel value ranging from 0 to 255, where 0 corresponds to black, 255 corresponds to white, and the values in between represent varying shades of gray. This type of image is particularly effective for simple image processing tasks such as edge detection and morphological processing. For example, in edge detection, the gradient of intensity values can be utilized to identify boundaries within an image.

Binary images are another fundamental category of images, distinguished by their representation using only two colors for each pixel. Topically, these two colors are black and white(usually 0 for black and 1 for white), which is widely used to represent the simple morphological information of images. For instance, binary images are widely used in scenarios where simple morphological information is sufficient, such as in shape analysis, object recognition, and various forms of threshold tasks. The stark contrast in binary images makes them ideal for representing the structural or geometric features of objects within an image.

Indexed images consist of data matrixes, which are always represented by RGB 3 channels(Red, Green, Blue). In this format, each channel corresponds to the intensity of one of the primary colors, combining to produce a wide range of colors. It is worth mentioning that indexed images are usually used in photographs, video, and image editing, providing people with vigorous visual effects. However, The main disadvantage of using indexed color is the limited set of simultaneous colors per image. This limitation can reduce color fidelity and affect the quality of images, particularly in contexts where a broad spectrum of color is necessary, such as in high-definition photos or graphics design.

2.2 Ways to express image

Images can be expressed in various ways depending on the context and application, with each method offering distinct advantages. One of the most common methods is by using a pixel matrix, where each pixel's color intensity is represented by an integer value. For grayscale and binary images, a single matrix suffices, with grayscale images using a range of values(typically from 0 to 255) to represent different shades of gray, and binary images using just two values to represent black and white. For an indexed image, we need three matrixes to express the image by RGB channel. Express images by using a pixel matrix is one of the most common ways to express images which is the most visualised way for image input and output.

Sometimes, especially within the field of machine learning and computer vision, images are expressed as vectors. This involves flattening the pixel matrix into a one-dimensional vector, where each element corresponds to a pixel's intensity value in the original image. This vectorized form is particularly useful as input for algorithms, which treat images as feature vectors. This approach simplifies the application of mathematical and statistical methods.

To facilitate efficient image transfer and storage, especially over the internet, images are often expressed in a compressed format. Common formats such as JPEG and PNG are used to encode images in a way that reduces their file size. JPEG achieves this through lossy compression, which reduces file size by discarding some image data, while PNG uses lossless compression, preserving all image data while still achieving some level of size reduction.

2.3 Basic operating

2.3.1 Geometric transformation

Basic operation is applied for observation from the other perspectives, focusing on a specific area, and for the consistent format of input and output. When people need to observe images from other perspectives, rotation of images is useful. We can choose to rotate the image for a certain degree clockwise or counterclockwise, and whether change the size of the image. In addition for some systems, the input of image needs a uniform format, including the same degree. For instance, the input of many image recognition systems and the input of many edge detectors need to rotate the image to a certain degree. Sometimes, the size of the image is too big or too small, especially in the preprocessing of images for input and output of the network. zooming is useful to be applied. Zooming is Changing the size of images by increasing the resolution ratio and reducing the revolution ratio. Zooming is achieved by using methods including Nearest neighbor interpolation, bilinear interpolation, etc. To digitize this process We can illustrate the process of zooming mathematically by using the zooming factor

$$x^{'} = sx * x, y^{'} = sy * y \quad (1)$$

where sx and sy are the zooming factors. Moreover, sometimes we also need to change the position of the image, we need to translate the image to the other position. The essence of translation is moving the pixels of images for a certain distance in a certain direction. After translation, the image will appear in a new place, even in the blank area, which is used in data strengthening and image correction tasks. To make the process more comprehensive, we can digitalize the process. The result of translation can also be represented mathematically by the following equation.

$$x^{'} = x + tx, y^{'} + y + ty \quad (2)$$

where tx and ty are the translation factors.

2.3.2 strengthening and recovery of image

To meet the ideal quality of the image, strengthening, and recovery is necessary. Common techniques include filter and denoise. To increase the quality of images, a filter is commonly used. The filter is achieved by image convolution, which is widely used to denoise. filter including median and nearest filter, both of them aiming to achieve different purposes. The median filtering denoising method has been widely used for impulse noise. However, as the noise density increases beyond 50There also may be some noise like black dots. To improve the quality of the image, we need to remove the noise from our image, which is called denoise. To be more specific, denoise is removing the part of the image that it should not exist, including some black dots to improve the quality of the images. Which is widely used in the fields including medical image analysis and satellite image analysis. We strengthen the visual effects by adjusting the contrast of images, which can be achieved by using histogram stretch and histogram equalization. The techniques of denoise can be used in the post-processing of photography and processing of medical images. For instance, the visual effect is significantly increased after denoising. There is another way to strengthen the visual effects of images called contrast control, which makes the darkness more dark and the brightness more bright. Controlling the image’s contrast can make the image more vivid and vigorous. Contrast control can be achieved by applying contrast stretching.

2.3.3 Iamge Segregation

Sometimes we need to segregate the image into different parts, especially in the field of image detection. Image segregation can be applied, which is a process in image processing and computer vision that involves dividing or separating an image into different regions or segments based on certain criteria or features. It is quite difficult to achieve this technique. We will introduce two different ways to achieve image segregation. Region growth is an effective way for image segregation, defined by following rules.

$$(a)\bigcup_{i=1}^{n} R_i = R. \tag{3}$$

$$(b)\ R_i \text{ is a connected region, } \mathrm{i} = 1, 2, \ldots, \mathrm{n} \tag{4}$$

$$(c)\ R_i \bigcap R_j = \varnothing, i \neq j \tag{5}$$

$$(d)\ P\left(R_i\right) = \text{TRUE for } i = 1, 2, \ldots, n. \tag{6}$$

$$(e)\ P\left(R_i \bigcup R_j\right) = \text{FALSE for any adjacent region } R_i \text{ and } R_j. \tag{7}$$

$P(Ri)$ is apredicate that satisfies the purpose of segregation, for instance, $P(Ri)$ is set as true when every point in this area share the same gray-scale, is set as false when there are two points have different grayscale in this area. The other way to achieve image segregation is the watershed algorithm, Watershed by topographic distance. Intuitively, a drop of water falling on a topographic relief flows towards the ”nearest” minimum. The ”nearest” minimum is that minimum which lies at the end of the path of steepest descent. In terms of topography, this occurs if the point lies in the catchment basin of that minimum. The previous definition does not verify this condition.

2.3.4 Image Edge Detection

Image edge detection is a pivotal technique in computer vision that identifies boundaries within digital images by detecting significant changes in pixel intensity or color. This process simplifies image analysis by highlighting the contours of objects, facilitating applications such as object recognition, medical imaging, autonomous vehicle navigation, and image enhancement. Algorithms like the Sobel operator, Canny edge detector, Prewitt operator, Laplace of Gaussian (LoG), and Roberts Cross operator are commonly used to achieve edge detection.

2.3.5 Object Detection of Image

Object detection is widely used in a lot of fields including video monitor and auto-driving. The mainstream algorithm to achieve object detection is SVM. Support vector machines are supervised max-margin models with associated learning algorithms that analyze data for classification and regression analysis. The most significant development of this technique is attributed to an essay ”Attention is all you need” which introduces the YOLO algorithm to this system, which dramatically increases the efficiency and accuracy of object detection. To be more specific, YOLO is a kind of object detection algorithm that transforms object detection tasks into regression. With a single forward pass, YOLO can simultaneously predict the categories and bounding box positions of multiple objects in an image.

3 Convolutional Neural Network

Convolutional Neural Networks (CNN) have become a cornerstone in the field of image processing, enabling the efficient training and optimization of models for various applications. CNN [1]plays an important role in optimizing and training models, especially in modern applications. CNN facilitates the transfer and sharing of image data, enabling distributed image processing tasks and real-time analysis.

3.1 The Architecture of Computer Network

The architecture of the CNN consists of three main types of layers: an input layer[2], a hidden layer, and an output layer. This is why we need to design an algorithm to connect each layer to achieve our purpose.

3.1.1 Tnput layer

The input layer of a CNN is responsible for receiving raw data in a format that the network can process. This usually involves converting images, text, or audio into matrices of vectors. For images, each pixel’s intensity is typically represented by a matrix, with different matrices for each color channel(e.g., RGB).

3.1.2 Hidden layer

The hidden layers of a CNN are where the network learns to detect various features in the input data. The significant and essential difference between the RNN network and the CNN network is the addition of the convolution layer .The convolution layer aims to detect all the features for input in image, text, and audio, which can be expressed as

$$(I * K = \mathrm{sumsumI}(i, j)\,\mathrm{times}\,K(x-i, y-j) \tag{8}$$

where I represent the input image, K is the convolutional kernel (filter) (x,y) denotes the coordinates in the output feature map. The operation involves sliding the filter over the input image and computing the dot product at each position. To reduce the spatial dimensions of the feature maps, which helps to minimize computational load and control overfitting, the addition of the pooling layer plays an important part. The most common pooling operation is shown in formula 9 after pooling, pad zero into the image so that the input and output share the same scale.

$$\text{MaxPool}(x, y) = \text{Max}\, I(x+i, y+j) \tag{9}$$

These two operations are linear. This is why it is essential to add an activate function in this process. The most common activate functions are the sigmoid function and softmax function. To convert a complex function to a simple expression, we can transform it by a sigmoid function, with values between 0 and 1. It is defined as:

$$1/(1+e(-z)) \tag{10}$$

As for the softmax function, converts raw scores (logits) into probabilities for each class. It is defined as:

$$\text{Softmax}(z) = e^z / \left(\text{sum}\left(e^z i\right)\right) \tag{11}$$

This ensures that the output values sum to 1, providing a probability distribution over classes. It is worth noticing, that there is an activated layer is necessary after every hidden layer, otherwise, the superposition of each layer will lose its meaning because it is a linear process. The process of the Rnns network can be visualized as such

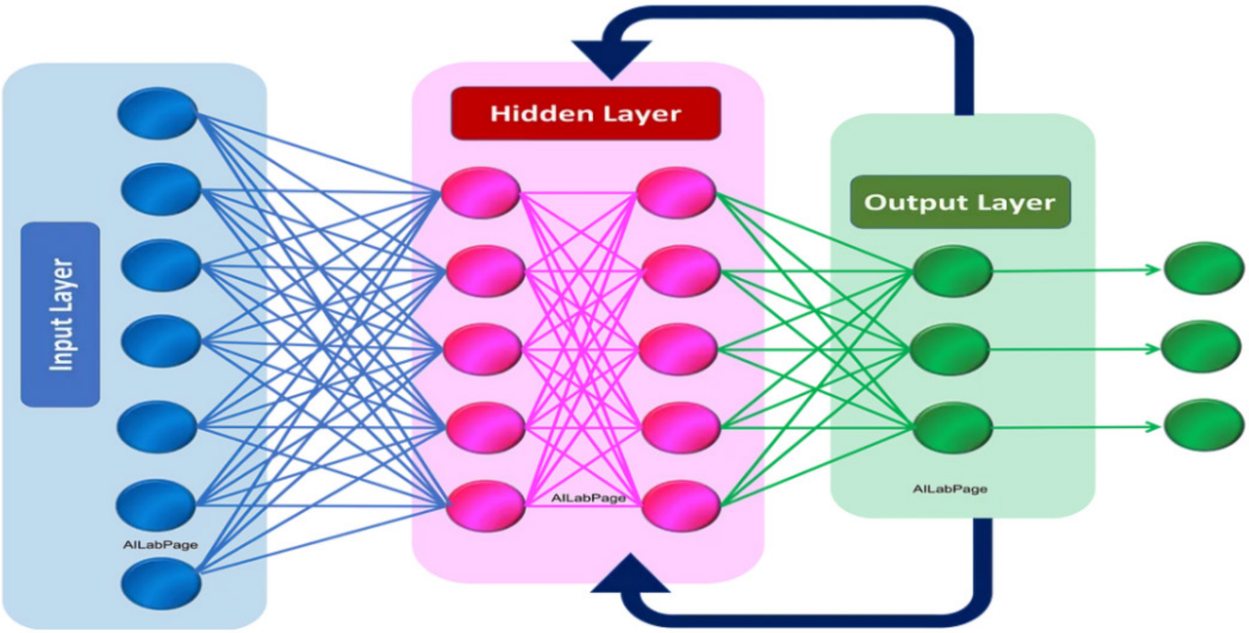

Figure 1: Enter Caption

3.1.3 The normalization of output

The output is in the form of a vector or a matrix, which is hard for the computer to further analysis, especially for the classification. It is necessary to find an intuitive way to the expression output. In this way, we can not only classify by the output but also know the possibility of each category. The formula for calculation is as follows

$$\frac{e^z}{\sum_j e^{z_j}} \tag{12}$$

This layer for normalization is named softmax, which turns the network as follows

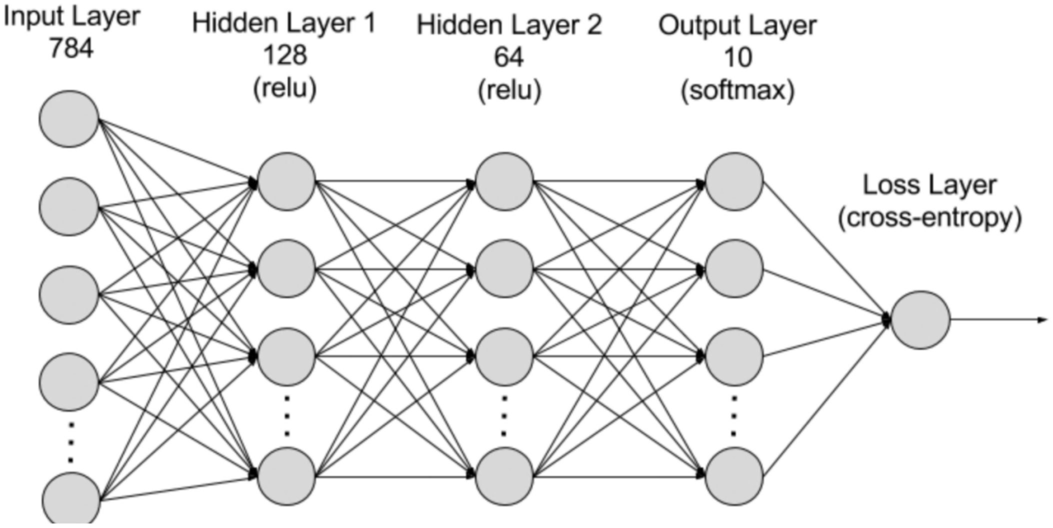

Figure 2: Enter Caption

3.2 The way to measure the output

Even though the softmax function helps optimize the output, we still need a program to measure the output for further optimization. Therefore, We need to quantificate the result for analysis. One of the ways to measure output is called cross-entropy loss which is a loss function determined as follows, which measures the difference between the predicted probability distribution and the actual distribution (one-hot encoded true labels).

$$textCross - EntropyLoss = -\sum_{i=1}^{N} y_i \log\left(p_i\right) \qquad (13)$$

where *yi* is the true label (1 for the correct class, 0 otherwise), *pi* is the predicted probability for class *i*, and *N* is the number of classes. The other way to measure the output is accuracy, it also can be expressed mathematically as follows, which measures the proportion of correct predictions out of all predictions made. Mathematically

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \qquad (14)$$

4 Generative Adversarial Network

For a long time, people have struggled to find a way to generate high-quality images, which can confuse humans with real images. However, attributed to the black box of AI generation, it is desperate for humans to modify the system for generations. In this case, the GAN system comes true. GAN is a system a Generative Adversarial Network (GAN) is a type of deep learning model designed for generating new data samples that resemble a given dataset. GANs consist of two neural networks: a generator and a discriminator, which are trained simultaneously through a process of adversarial competition. A generator is a training system to generate images. This network creates fake data samples from random noise. Its goal is to produce data that is indistinguishable from real data. The loss function is the most important part of this network which aims to produce data that looks like real data, such that the discriminator is tricked into thinking it's real.

Mathematically, the generator's loss function is expressed as :

$$\mathcal{L}_G = -\mathbb{E}_{z\sim p_z(z)}[\log D(G(z))] \qquad (15)$$

where: *z* represents the noise vector sampled from a prior distribution *pz*(*z*). *G*(*z*) denotes the generated sample from the generator given the noise *z*. *D*(*G*(*z*)) is the discriminator's probability that the generated sample *G*(*z*) is real. The generator's objective is to maximize *D*(*G*(*z*)), making the discriminaor believe that the generated data is real. Thus, the loss function encourages the generator to produce samples that the discriminator cannot easily distinguish from real ones. The system that plays an adversarial part in work is discriminator which is a network that evaluates data samples and determines whether they are real (from the dataset) or fake (produced by the generator). Its goal is to classify data as real or fake accurately. It outputs a probability score indicating whether a given input is from the real data distribution or generated by the generator. For training this system, the loss function is essential The loss function for the discriminator in a Generative Adversarial Network (GAN) is given by:

$$\text{Loss}_D = -\left[\mathbb{E}_{x\sim\text{real}}[\log D(x)] + \mathbb{E}_{z\sim\text{noise}}[\log(1 - D(G(z)))]\right]$$

where: E*x*~real[log *D*(*x*)] represents the expected value of the logarithm of the discriminator's output for real data samples x. $\mathrm{E}z\sim\mathrm{noise}[\log(1 - D(G(z)))]$ represents the expected value of the logarithm of one minus the discriminator's output for fake data samples $G(z)$, where G is the generator and z is a latent variable.

5 Conclusion

This essay introduces the nowadays' methods to process and generate images. Initially, this essay introduces the main categories of images. There are three main categories of images, including grayscale image, binary image and indexed image. The different ways to express images have intrinsic advantages for image processing. Pixel matrix expression helps digitize the intensity of colors, expressing images by vectors is easier for input of image processing, and expressing images by compressed format is easier for storage. Next, this essay introduces the basic image operations. to modify the sizes, directions, and position, we apply geometric operations. To strengthen the quality of images, it is necessary to remove the noise in the image for example black dot, or the image is not recognizable, the image will be clearer for humans by applying strengthening. Sometimes, a single image may contain several targets, image segregation can be applied to segregate the image into different part, which is an important pre-processing for image detection. To achieve image segregation, it is necessary to find the edge of an object. In this case, people have optimized methods to detect edges, which are Sobel operator, Canny edge detector, Prewitt operator, Laplace of Gaussian (LoG), and Roberts Cross operator. In the field of object detection, to increase efficiency and accuracy, the algorithm YOLO is applied. By applying RNNs network, image feature extraction is achieved, so that it will be easier for computer to analyze the information that these images contain. Next, this essay introduces CNNs and RNNS, including the theories and purpose of these two networks. In addition, the difference between these two network. An important component of network is the definition of loss function, and the acitivate function, so that this essay introduces the common ways to define loss function, and the three major activate functions. Apart from this, this essay further introduces how to organize these different layers. Next, to improve the quality of generated images to achieve the purpose that human is confused with generated images and real image, GANs comes. This essay introduces the structure and the theory of GANs. This essay introduces the discriminator and the generator as well, and hw they work.

References

[1] Muhammad Waqas and Usa Wannasingha Humphries. A critical review of RNN and LSTM variants in hydrological time series predictions[J]. ScienceDirect, 2024.

[2] Karuna R. Dongur Pushpa Tandekar, Shrawan Kumar Purve. Digital Image Processing: Its History and Application[J]. IJARCCE, 2022.

[3] Cheng Yu, Wenmin Wang, Roberto Bugiolacchi. Improving generative adversarial network inversion via fine-tuning GAN encoders[J]. ScienceDirect, 2024.

[4] Xiangwei Zheng, Lifeng Zhang, Chunyan Xu, Xuanchi Chen, Zhen Cui. An attribution graph-based interpretable method for CNNs[J]. ScienceDirect, 2024.

[5] Ismail Akg¨ul. A Pooling Method Developed for Use in Convolutional Neural Networks[J]. ScienceDirect, 2024.

[6] Mukul, Singh S, Nishi. The Origins of Digital Image Processing Application areas in Digital Image Processing Medical Images[J]. Ijert, 2018.

[7] B R Q A, C M U, B K K, et. al. Hyperspectral document image processing: Applications, challenges and future prospects[J]. ScienceDirect, 2019

[8] MD S P, MSc D D, MD S G, et. al. Kiosk 5R-FA-03 - Automated Cardiac MRI Plane Prescription Through Cnn-based Landmark Detection[C]: Journal of Cardiovascular Magnetic Resonance, ELSEVIER, 2024.

[9] A H L, A C Q, A C L, et. al. Chapter 18 - Generative adversarial network (GAN) assisted IoT search engine for disaster damage assessment[M]. Elsevier Shop, 2024.