CONVOLUTIONAL NEURAL NETWORKS

Abstract

Convolutional Neural Network (CNN) is one of the most important algorithms used for computer vision task. CNN integrates feature extraction process along with classification. It can automatically extract features from images. Feature extraction is done through convolution operation in CNN. CNN also employs pooling operation for dimensionality reduction. Over the years researchers have proposed different architectures of CNN. The architectures differ based on the numbers of convolution layers, size of filters, number of pooling layers and activation functions used in the CNN model. LeNet 5, AlexNet, VGG16, VGG19 are some of the CNN architectures discussed here.

Keywords: Convolutional Neural Network (CNN), LeNet 5, AlexNet, VGG16, VGG19.

Authors

Anuj Kumar Das
Department of Computer Science
Dudhnoi College
Assam, India

Research Scholar
Department of Computer Science & Engineering
Assam Don Bosco University
Assam, India

Dr. Syed Sazzad Ahmed
Department of Computer Science & Engineering
Assam Don Bosco University
Assam, India

I. INTRODUCTION

Artificial Neural Network (ANN) has revolutionized the world of AI to a great extent. ANN mimics the structure and working of the human brain. With the advent of low cost high performance computing, complex ANN structures were developed to solve problems through computers. Complex ANNs were developed by increasing the number of hidden layers. These ANNs have given rise to a new field in computer science known as deep learning. The increasing number of hidden layers in ANN can interpret data with a much more high accuracy. Different layers can have different activation functions to interpret different dimensions of data.

Convolutional Neural Network (CNN) is a deep neural network with multiple hidden layers. The most important operation of CNN is performed by the convolutional layers, from which its name is derived. The most common application of CNN is in computer vision. Computer Vision is an area of study in computer science where applications or devices were developed that can perceive objects or find patterns in images or video frames. This is done by processing the images or video frames for extracting features. Analysis of these features were done for information extraction or decision making. Techniques like detection, classification and segmentation are some of the common application of computer vision models. Computer Vision techniques are used in agriculture for monitoring of crop fields, prevention and control of crop diseases, classification and quality inspection of agricultural products. In transportation computer vision techniques are used for safety and navigation in autonomous vehicles. Computer vision techniques are also used in smart manufacturing, warehouse logistics, detecting anomalies with high accuracy in healthcare, etc.

II. CNN LAYERS

Convolutional Neural Networks (CNN) has a layered architecture. The most important layers that form the backbone of a CNN model are discussed below.

1. **Convolutional layer:** The convolutional layer in CNN is used for feature extraction. The pixels of an image whose features are to be extracted can be represented in the form of a matrix. The convolutional layer uses a kernel/ filter to perform the convolution operation. Kernel/filter is a matrix whose dimension is lesser than the dimension of the pixel matrix of an input image. Convolution is a mathematical operation where a kernel whose size is of the form mxm is placed over the input matrix. The values in the input matrix on which the kernel is placed is known as the receptive field. For each location of the kernel on the input image matrix, the corresponding values in the receptive field and that of the kernel are multiplied and added to get a value. The kernel then slides over to the next location in the input matrix to calculate the next value and so on. After performing the convolution operation a set of values are obtained in the form of a matrix, which is known as the feature map.

   Actually an image is represented in 3D matrix. An image has height, width and depth, where depth represents the color channels (RGB). So, convolution operation also needs to be done for the entire depth of the image. Convolution of image is done by using a 3D kernel. The output of this operation is a 2D feature map. In 3D convolution the 3D kernel is placed on the 3D matrix representing the input image and the corresponding values of the two matrices are multiplied and then all the values are added to get a single

value. After sliding the kernel over the entire image a 2D matrix is obtained, which is the feature map of the input image.

2. **Stride:** Stride in CNN refers to the amount of pixel shifts on the input image matrix by the kernel/filter. If the stride is 1 then the kernel shifts by 1 pixel value on the input image matrix. If the stride is 2 then the kernel will shift by 2 pixel values. Stride with higher number is used to reduce overlapping of the convoluted pixels with the next pixels used for convolution. Stride is also used for down sampling.

3. **Padding:** Padding refers to addition of extra pixels to the boundary of the input image. Padding is done to preserve the information of the boundary pixels of the image. This is done because we tend to lose the information of the boundary pixels during convolution operation. Padding is also done to up-sample the image, so that dimensionality of the image is not reduced after convolution. Generally the values of the padded pixels are set to 0.

4. **Activation function:** Activation functions are used in convolutional layers to introduce non-linearity. This is done because images by itself are nonlinear in nature, but the convolution operation may make the image linear. So, to break the linearity, activation functions are used. ReLU is the most common activation function used in the convolutional layers of a CNN. ReLU stands for Rectified Linear Unit. ReLU returns 0 for a negative input and for a non-negative input it returns the value itself. (It is given by: $f(x) = \max(0,x)$.) Other than ReLU, sigmoid functions were also used as the activation function in convolutional layers of a CNN.

5. **Fully connected layers**: Multiple convolution and pooling layers were stacked alternatively in a CNN based on what architecture we are using. The final output of these layers is a matrix which is then flattened to one dimensional vector. These vectors are then fed to the fully connected layers. The fully connected layers perform the final classification operation. In most of the CNN architectures the last layer is a softmax layer that gives the probabilities of the different classes that the input may belong to.

III. CNN ARCHITECTURES

A number of CNN architectures exist, proposed by different researchers. The architectures vary based on the number of convolutional layers used, types of filters used, pooling layers and activation functions used. Some of these CNN architectures are discussed here.

1. **LeNet 5:** The LeNet5 model was proposed by LeCun et al. [1] in 1998 in his paper ‘Gradient Based Learning Applied to document Recognition’. The model was used for detecting hand written digits.

   The size of the input image to the model is of 32x32 pixels. Then a convolution layer is used to generate 6 feature maps of the size 28x28. These feature maps were subsampled using a pooling layer to generate 6 feature maps of size 14x14. A sigmoid function is used to introduce nonlinearity. Then a convolution layer is used to generate 16 feature maps of size 10x10. The generated feature maps are again subsampled with a pooling layer to obtain 16 feature maps of the size 5x5. Another convolutional layer is

used to generate 120 feature maps of size 1x1. Then a fully connected layer is used having 84 neurons.

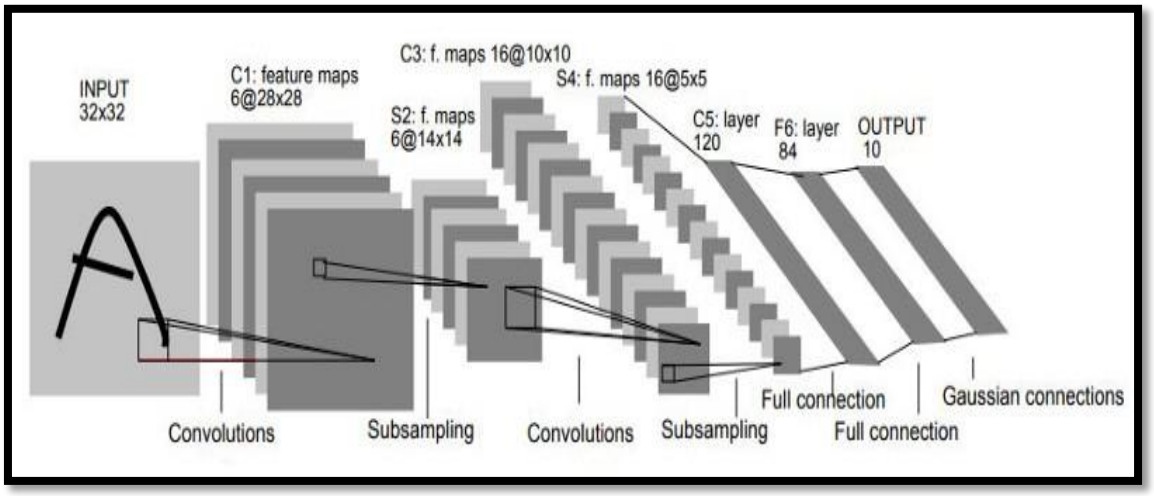

Figure 1: LeNet5 architecture (LeCun et al. [1]).

The activation function used till layer F6 in figure above is a sigmoid function. The final layer uses Eucledian Radial Basis function having 10 neurons representing 10 classes of output.

2. AlexNet

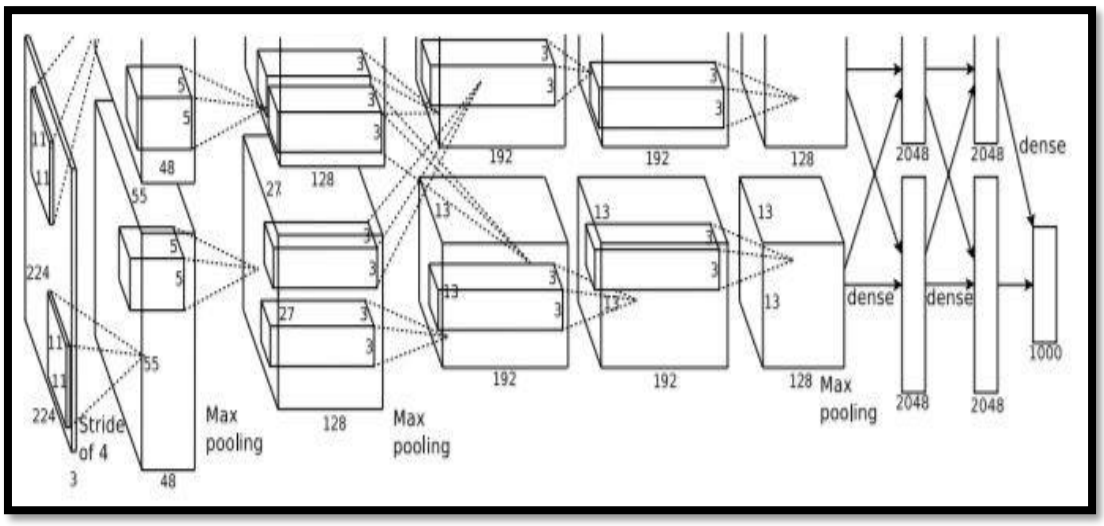

Figure 2: AlexNet architecture (Krizhevsky et al. [2]).

The AlexNet architecture was proposed by Alex Krizhevsky et al. [2]. The model was used in the ILSVRC-2012 competition. ILSVRC stands for ImageNet Large-Scale Visual Recognition Challenge. It is a competition held annually that evaluates algorithms for object detection and image classification in large scale.

The AlexNet architecture contains 5 convolutional layers and 3 fully connected layers. The input image is of the size 224x224x3. The first convolutional layer uses 96 kernels of size 11x11x3 with a stride of 4 pixels. Then a max-pooling layer is used with stride 2. Then another convolutional layer uses 1156 kernels of size 5x5x48. After that a max-pooling layer of stride 2 is used. Then 3 convolutional layers are used without any max-pooling layers in between. The third convolutional layer uses 384 kernels of size 3x3x256. The fourth convolution layer uses 384 kernels of 3x3x172 size and fifth layer uses 256 kernels of size 3x3x192. Another max-pooling layer is applied. The final output of these connected convolutional and max-pooling layers are then input to three fully connected layers. The first two fully connected layers have 4096 neurons. The third layer uses 1000 units for classification into 1000 classes. The final layer is a softmax layer.

The model was able to achieve remarkable results for image classification. It was also observed that if a single convolution layer was removed, then the performance of the model also deteriorates. So, the depth of the model is essential for achieving good results.

3. **VGG16:** VGG16 was proposed by Karen Simonyan and Andrew Zisserman [3] and published in the paper 'Very Deep Convolutional Networks for Large-Scale Image Recognition'. The researchers were of the notion that increasing the depth of a CNN model makes its performance better. So, they have made the architecture of VGG16 deeper by adding more hidden layers. The model was trained and tested on the ImageNet dataset.

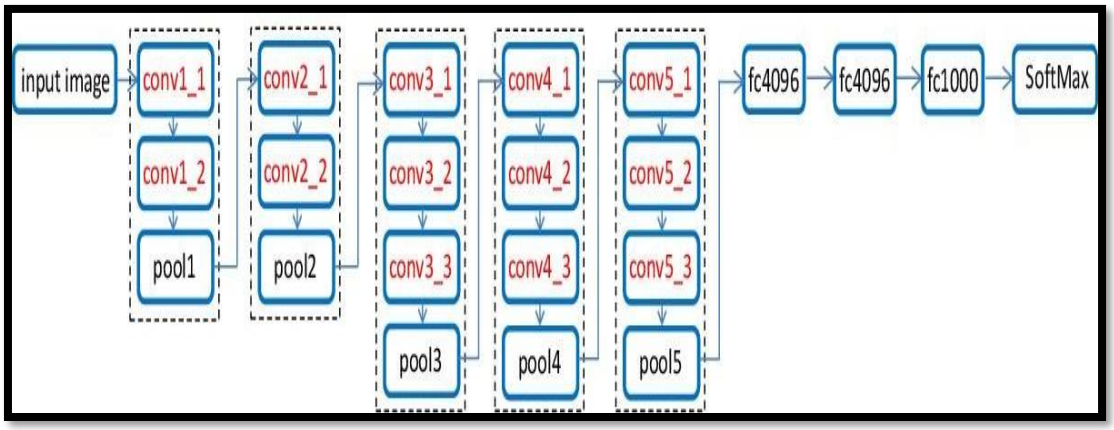

Figure 3: VGG16 architecture (Yu et al. [4]).

The model takes an input image of size 224x224x3. The architecture contains 16 weight layers of which 13 are convolutional layers and 3 are fully connected layers. The convolutional layers use kernel of size 3x3 and stride of 1. Thee convolutional layers are stacked into a group of 5. The first group of convolutional layers uses 64 filters. The second group uses 128 filters. The third group uses 256 filters. The fourth and fifth group uses 512 filters each. Pooling layers are used after each group of convolutional layers. Pooling applied is max-pooling of size 2x2 with stride 2. Three fully connected layers were used. The first two layers have 4096 neurons each and the third layer has 1000 neurons. The final layer is a Softmax layer. ReLU is used as the activation function in all the hidden layers.

4. **VGG19:** The VGG19 architecture is an extension of the VGG16 architecture with three additional convolutional layers, making the number of convolutional layers 16 in VGG19. The VGG19 also has three fully connected layers. So, the VGG19 architecture has 19 weight layers.

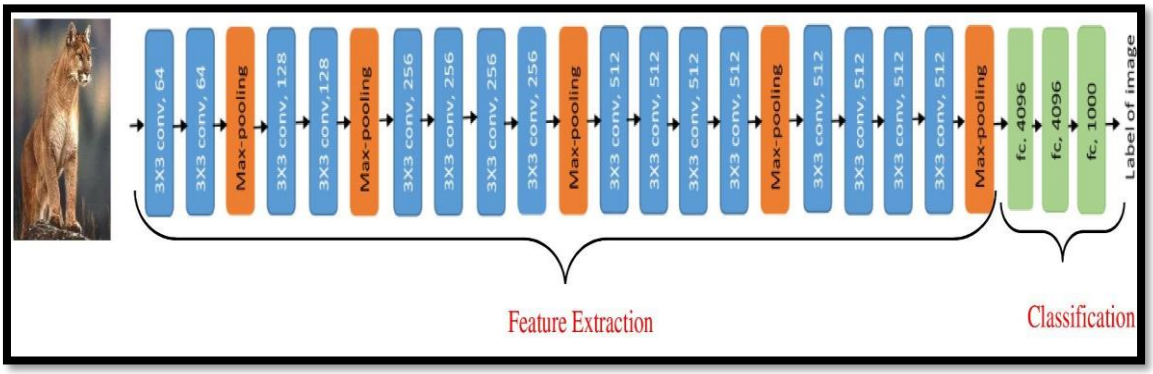

Figure 4: VGG19 architecture (Bansal et al. [5]).

CNNs have revolutionized the field of computer vision. Different architectures were developed by adding more convolutional and pooling layers, making the model much deeper. CNN removes the need to manually extract features from images. Convolutional layers in CNN uses varying filter size to detect different features of images. Here we have studied the architectures of only a very few CNN models. Over the years a number of CNN architectures have been proposed by researchers which have shown to be highly effective in computer vision tasks. CNN is a hot topic in the field of research. With the passage of time much more efficient CNN architectures will be developed that will be able to perform complex computer vision tasks with higher level of performance in the future.

REFERENCES

[1] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998 Nov;86(11):2278-324.
[2] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012;25.
[3] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014 Sep 4.
[4] Yu W, Yang K, Bai Y, Xiao T, Yao H, Rui Y. Visualizing and comparing AlexNet and VGG using deconvolutional layers. InProceedings of the 33 rd International Conference on Machine Learning 2016 Jun 21.
[5] Bansal M, Kumar M, Sachdeva M, Mittal A. Transfer learning for image classification using VGG19: Caltech-101 image data set. Journal of Ambient Intelligence and Humanized Computing. 2021 Sep 17:1-2.
[6] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. InEuropean conference on computer vision 2014 Sep 6 (pp. 818-833). Springer, Cham.

