International Journal for Modern Trends in Science and Technology, 8(01): 147-152, 2022

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0801025
Available online at: http://www.ijmtst.com/vol8issue01.html

Convolution Neural Network (CNN) for Video Processing: A Survey

Palash Feddewar[1] | Dr. Bharti Deshmukh[2]

[1]Assistant Professor, Department of Computer Science, Prerna College of Commerce, Nagpur, MS, India.
[2]Assistant Professor, Department of Computer Application, Prerna College of Commerce, Nagpur, MS, India.

To Cite this Article
Palash Feddewar and Dr. Bharti Deshmukh. Convolution Neural Network (CNN) for Video Processing: A Survey. *International Journal for Modern Trends in Science and Technology* 2022, *8* pp. 147-152. https://doi.org/10.46501/IJMTST0801025

Article Info
Received: 02 December 2021; Accepted: 03 January 2022; Published: 08 January 2022.

ABSTRACT

The past decade focused on Image Processing. Recently it is being found that many have shown their keen interest in Video Processing. Convolution Neural Network (CNN) showed extraordinary results in the field of image processing; now we wish to bring the same computation power in video processing too. This paper deals with the detailed study of CNN which is used for video processing. This research paper is expected to provide a review on how CNN is applicable for video processing. We also have discussed the architecture of CNN. This paper also highlights the different models of CNN present for the task of video processing and their functionality.

KEYWORDS:CNN, Video Processing, ANN.

1.INTRODUCTION

CNN has become very popular in the world due to its computational power. CNN is capable of dealing with almost all computer vision problems. At its earlier phase CNN was used for the purpose of image processing. It was applied for the task of face recognition, emotion recognition, object tracking, action recognition. The reason CNN is used to solve almost all possible problems of computer vision is the high accuracy that it provides for all the problems.

LeNet was the pioneering work in Convolutional Neural Networks by LeCun et al. in 1990 [1]. CNN can perform this task with the help of the neurons and their interconnections present inside the layers. Neurons are basically the learning blocks used to extract the features of the input data. With the increase in the complexity of the network the depth of these interconnections is increased, increasing the computational power of CNN. For example, AlexNet - one of the most representative CNNs, has 650K neurons and 60M related parameters [2]. CNN uses various BP(Back-propagation) algorithms to achieve this complex task.

Video processing remained untouched for a handful of years. Traditional video processing involves the image processing techniques which were used for feature extraction from the videos and then produce the desired results. The major task in video processing involves the extraction of spatio-temporal features. The other challenging part in video processing is the size of a video and its dimensions.

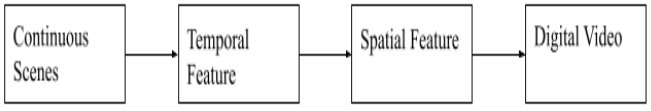

Figure 1- Video Data

Figure 1 gives the rough idea of video data. Most of the fundamental research of computer vision today focuses on images, focusing less on sequences of images, i.e., video frames. However, video data provides deeper situational understanding because a series of images gives various information about the subject. For example, we can track an object through an optical flow of the sequence of images and predict its next action [3]

2.RELATED WORK

In a paper proposed by [4] presented a Systematic Literature Review (SLR) of video processing using DL. They investigated the applications, functionalities, techniques, datasets, issues, and challenges by formulating the relevant research questions (RQs). They categorize the deep learning technique for video processing as CNN, DNN, and RNN based.

Paper [5] provided a review of advancement in applying DL techniques for STDM. They first categorize spatio-temporal data into 5 types, and then introduced the DL models widely used in STDM. Next, they classified literature on the basis of types of spatio-temporal data, data mining tasks, and also the DL models, followed by the diffrent applications of deep learning for STDM in different domains including transportation, on-demand service, climate & weather analysis, human mobility, location-based social network, crime analysis, and neuroscience. Finally, they conclude the limitations of current research and point out future research directions.

[6] discusses the basics of ANN as tool for recognition of complex pattern and image processing tasks. Also, some applications of the CNN tool we will present OCR based text translation and biometric based uni modal and multimodal person identification systems.

[7] presented CNN EXPLAINER, which is an interactive visualization tool, which is designed for amateurs to learn CNN and examine it. It is a foundational deep learning model architecture. There tool addressed key challenges that they faced while learning about CNN, which they had identified from interviews with instructors and a survey with past students.CNN EXPLAINER helps users understand the inner workings of CNNs, and is engaging and enjoyable to use.

[8] presented a survey of recent advances in CNN architecture design taking into account the three periods first by improving accuracy, next minimizing the number of parameters using squeeze architecture, then CNN model adapted for embedded and mobile systems.

3.CONVOLUTION NEURAL NETWORK (CNN) ARCHITECTURE

Prior to CNN, ANN (artificial Neural Network) was used. ANN was then improved in 3 major aspects to produce CNN. Those 3 aspects are:

A.Sparse Connectivity

In traditional ANN in order to get n output from m input matrix multiplication m x n was done. In CNN only K inputs are interconnected with n output, where k<n. This correction reduces the time taken by the algorithm remarkably.

B. Sharing of Parameters / Weights

In ANN the weight vector associated with the input is used only once. for the next iteration it is updated. Which means that the same weight cannot be reused. On the contrary in CNN, components of the kernel are moved with the activation function to be used everywhere through the network.

C.Equi-variant Representation

The property in which the overall output corresponding to the input remains the same even if the sequence of operation is interchanged is called Equi-variance.

CNN is basically a Neural Network; and a neural network is simply collection and combination of neurons organized in layered manner. Each neuron is associated with some weights and a bias. CNN consists of the following basic blocks.

1. A Tensor, a n-dimensional matrix.
2. A Neuron, it is a function which takes n input and returns single output.
3. A layer is a collection of artificial neurons that perform the same operation.
4. Kernel weights, each neuron is trained with some weights to identify and classify the data set during

the training phase. They are also associated with some bias.

CNN can also be viewed as a layered architecture. It consists of the following layers:

i. *Input Layer*- An input layer is the first layer of any neural network. The input to the network is given through this layer. The input layer has no computation power; it just takes the input from the user and passes the input to the next layer in the network.

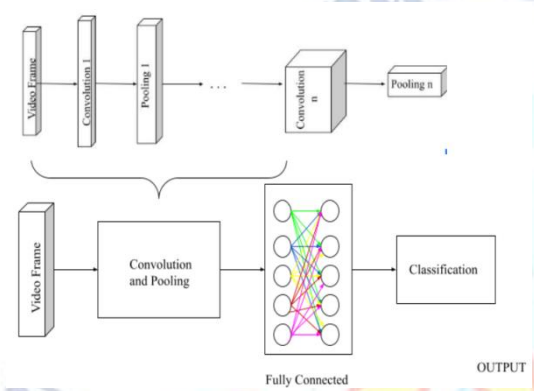

Figure2-Basic CNN Architecture

ii. *Convolution Layer*- Convolution layer can be viewed as the second layer of any CNN architecture. While in many CNN models it is viewed as the first layer for a reason that, Input layer does not compute any results. It just passes the input to the next layer. We can say that the convolution layer is the first computation layer in a CNN architecture. The convolution layer is foundation of CNN as most of the computation is done in this layer.

Here dot product is performed with the predetermined kernels. The dot product can be between the input and the filters or it can also be of the filters and the output of the previous layer. Convolution involves formation of activation map. The neurons that are arranged in the same feature map share their weights as a result the CNN is more efficient than the ANN. These intermediate results are then summed up with the bias to get an output of the convolution layer.

The size of the kernel (filter) entirely depends on the network architecture. There is no direct way to find the exact size of kernel for the specific input. Multiple sizes are tested and the optimum filter is selected. The size of kernels can be 2x2, 3x3, 5x5, 7x7, etc. The minimum the size of the filter, higher is the accuracy, higher computation time. Similarly filters of greater size need less time for computation, but compromising the accuracy of the network.

i. Padding- The convolution layer faces some problems while convolution with the kernels/filters. If we consider an (n x n) input and we apply the (k x k) size filter on it, then after convolution we will get the result in the form (n - k + 1, n - k + 1). For example, if we take input of size (8 x 8) and we apply a filter/kernel of size (3 x 3) the output of convolution will be (6 x 6). Which means every time we do convolution the input shrinks? The second problem is that the corner pixels and the pixels at the edge are used less compared to the middle process during the convolution. As a result, the information in the corner and edge pixel is not preserved as well as it gets preserved in the case of the middle pixel. To overcome this problem padding is done. Padding simply involves addition of 0 to the input. This is a widely used technique. It is known as zero padding. It helps in preserving the data at the border.

0	0	0	0
0			0
0			0
0	0	0	0

Figure 3- Image after Zero Padding

ii. Kernel Size- Selecting appropriate size kernel/filter is an important part of this process. Selection of proper hyperparameter to slide on an input in order to extract appropriate features so as to classify the given input. The size of filter may vary according to the network architecture and dataset but it is greatly observed that the smaller the size of filter the better is the performance.

iii. Stride- stride refers to the number of pixels the kernel/filter to be shifted. If for example we have taken shride as 1 for a 2 x 2 filter it simply means that after the dot product of 2 x 2 the kernel/filter will be shifted one pixel to the right. It works on the similar principle as kernel size; i.e., smaller the parameter of stride more feature is extracted from an input image, higher is the dimension of the output, and larger the parameter of stride less

feature is extracted from the input and lesser is the dimension of the output

4.ACTIVATION FUNCTION

The convolution layer is followed by an activation function. The CNN has harness various activation functions to express complex features. An activation function is the unit which is responsible for determining what information to be transmitted to the next layer. This activation function is also referred to as a nonlinear function. A simple activation function looks like:

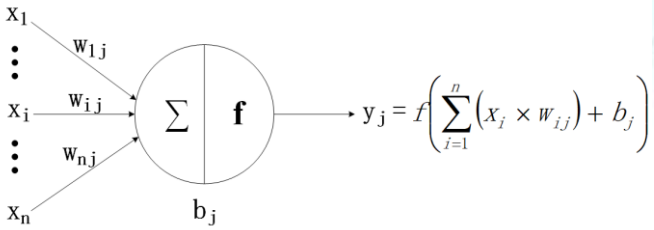

Figure 4- Activation Function [9]

CNN is rich in the form of activation functions. Some of the activations are named here, sigmoidal, tanh, ReLU, Sigmoidal, Leaky ReLU, Softmax.

The equation of Softmax function used to generate the final classification probability can be given as:

$$P_i = \frac{e^{ai}}{\Sigma_{N=1}^{10} e^{an}} \quad (1)$$

Where a_i is the i^{th}neuron.[10]

ReLU activation units are used in the output of the convolution layer and stores the output ReLU operation in the feature mapped array (F_T) in Eq. 2.[11]

$$F_T = \text{ReLU}(x_i) = \max(0, x_i) \quad (2)$$

Slight modification to ReLU resulted in Leaky ReLU given by Equation (3). [12]

$$F(x) = \begin{cases} x, x \leq 0 \\ 0.01x, otherwise \end{cases} \quad (3)$$

The values of each layer is calculated by function of its predecessor layers. The first layer is given by [13]

$$h^{(1)} = g^{(1)}(W^{(1)}x + b^{(1)}) \quad (4)$$

The following layers are given by

$$h^{(i+1)} = g^{(i+1)}(W^{(i+1)}h^{(i)} + b^{(i+1)}) \quad (5)$$

i. *Pooling Layer-* In a CNN architecture a Convolution Layer is followed by a Pooling Layer. This Pooling Layer reduces the spatial dimensionality of the features extracted by convolution Layer. The major Significance of the Pooling Layer is that it reduces the dimensionality without loss of information. Pooling also helps in generalization and reducing the problem of overfitting of data. This results in reducing the complexity of the network.There are different types of techniques available for pooling such as, max pooling, L2 Pooling, average pooling, multiscale, stochastic pooling, spectral pooling, overlapping, orderless pooling, spatial pyramid pooling, etc.

ii. *Fully Connected Layer-* Fully-connected layer is usually at the end of the Network. Unlike the convolution and pooling layer, it is not used for local extraction. It is called fully connected because here all the neurons are interconnected to other neurons in the network. The arrangement of neurons is not spatial in this layer. This layer is also called a Flatten Layer. Here outputs corresponding to the features extracted are computed. Mathematically it can be seen as the scalar product of the kernel/filter and the extracted features. It converts the n dimensional output to one dimension as an output.

5.CNN FOR VIDEO PROCESSING

The classical CNN model is used for image processing. In the early stages traditional image processing algorithms were used for video processing. After seeing the performance results of CNN on images, CNN was used for video processing. But still the complete video was not passed to the CNN model. As the video is a sequence of frames. Those frames are passed one by one to the CNN model and image processing is done on those video frames. In video processing techniques, temporal and spatial information from videos is exploited. [14]

Recently, CNN is being developed for actual video processing. Few CNN are mentioned below:

A. Spatial- Temporal CNN

This model was built to overcome the above-mentioned problem. Spatial- Temporal CNN is able to overcome this problem due to broad network connectivity.

i. *Single Frame-* This is used to extract feature frame by frame i.e from single frame.

ii. *Early Fusion-* In this strategy spatial stream and the temporal stream are combined at the beginning of

the network. It combines the frames and their optical flow to produce channel input.

iii. *Mid Fusion*-Here still frames and their stacted optical flow are given to spatial stream and temporal stream. And these two streams extract their own motion and appearance features. And finally, these extracted motion feature vectors and appearance feature vectors are combined to form spatio temporal feature.

iv. *Late Fusion*- In the late fusion model it places two different single-frame networks with shared parameters at an interval of 15 frames. Then these two frames are then combined at fully connected layer.

B. Multi-Resolutions CNN

This CNN model works at a much faster rate compared to other models, but on the contrary, it requires more time for training and also demands a high configuration GPU computer. This layer has two different streams for computation of spatial features called context stream and Fovea stream.

i. *Context stream*- This stream learns on low-resolution frames, 89 x 89 downsized clips at half original resolution.

ii. *Fovea stream*- It learns on high-resolution stream and is operated only in the middle portion of a frame. 89 x 89 center crop of the original resolution.

C. 3-D Convolution Neural Network

The 3D- CNN is similar to the same 2D - CNN which we have discussed in the earlier section. They differ in the following point.

i. *3D- Convolution Layer*- In traditional CNN dot product of input and kernel/filter is done which is basically dot product of 2 x 2 matrix. 3D CNN also use the same logic but here the multiplication is done in pair of 2D matrix.

ii. *3D- Max Pooling Layer*- In traditional 2D max-pooling layer (2 x 2) filter is used which can be viewed as a square. In 3D max pooling layer (2 x 2 x 2) filter is used which can be viewed as a cube.

6.FUTURE SCOPE AND CONCLUSION

This paper is an introduction to the CNN (Convolution Neural Network). In this survey paper we tried to explain the working of CNN in a manner which can be understood even to an amateur. CNN has bought a revolutionary change in Computer Vision. We have also taken in consideration a few CNN architectures which are used for video processing and discussed their working. In its early stages CNN was used for image processing. Due to its outstanding results CNN is being used even for Video Processing

REFERENCES

[1] B. B. Le Cun, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a backpropagation network", in NIPS. Citeseer, 1990.

[2] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", In: Proceedings of the Advances in Neural Information Processing Systems, 2012, 1097-1150.

[3] A. Taha, H. H. Zayed, M. E. Khalifa, and E.-S. M. El-Horbaty, ``Exploring behavior analysis in video surveillance applications," Int. J. Comput. Appl., vol. 93, no. 14, 2014, pp. 2232, doi: 10.5120/16283-6045.

[4] V. Sharma, M. Gupta, A. Kumar and D. Mishra, "Video Processing Using Deep Learning Techniques: A Systematic Literature Review," in IEEE Access, vol. 9, pp. 139489-139507, 2021, doi: 10.1109/ACCESS.2021.3118541.

[5] S. Wang, J. Cao and P. Yu, "Deep Learning for Spatio-Temporal Data Mining: A Survey," in IEEE Transactions on Knowledge and Data Engineering, doi: 10.1109/TKDE.2020.3025580.

[6] G. Sarker, "A Survey on Convolution Neural Networks," 2020 IEEE REGION 10 CONFERENCE (TENCON), 2020, pp. 923-928, doi: 10.1109/TENCON50793.2020.9293902.

[7] Z. J. Wang et al., "CNN Explainer: Learning Convolutional Neural Networks with Interactive Visualization," in IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 2, 2021, pp. 1396-1406, doi: 10.1109/TVCG.2020.3030418.

[8] A.Elhassouny and F.Smarandache, "Trends in deep convolutional neural Networks architectures: a review", IEEE/ICCSRE2019,Agadir, Morocco, 978-1-7281-0827-8/19/$31.00©2019 IEEE, 2019.

[9] Z. Li, F. Liu, W. Yang, S. Peng and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," in IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2021.3084827.

[10] Z. Qin, F.YuChenchen L. X. Chen, "How Convolutional Neural Networks See TheWorld- A Survey of Convolutional NeuralNetwork Visualization Methods", Mathematical Foundations of Computing American Institute of Mathematical Sciences, Volume 1, Number 2, 2018, pp. 149–180 doi:10.3934/mfc.2018008.

[11] M. Sahu andR. Dash. A Survey on Deep Learning: Convolution Neural Network (CNN). 10.1007/978-981-15-6202-0_32, 2021.

[12] N. Aloysius and M. Geetha, "A review on deep convolutional neural networks,"International Conference on Communication and Signal Processing (ICCSP), 2017, pp. 0588-0592, doi: 10.1109/ICCSP.2017.8286426.

[13] A. Elhassouny and F. Smarandache, "Trends in deep convolutional neural Networks architectures: a review," 2019

International Conference of Computer Science and Renewable Energies (ICCSRE), 2019, pp. 1-8, doi: 10.1109/ICCSRE.2019.8807741.

[14] A. Khan1, A. Sohail1, U.Zahoora, and A. S. Qureshi, "A Survey of the Recent Architectures of Deep Convolutional Neural Networks",ArtifIntell Rev 53, 5455–5516, 2020. https://doi.org/10.1007/s10462-020-09825-6