Journal of Advanced Research in Technology and Management Sciences
Volume: 04 Issue: 04 ISSN: 2582-3078 July 2022
Available online at: http://www.jartms.org

ANOMALY DETECTION USING VGG 16 ARCHITECTURE

[1]KALLEPALLI ROHIT KUMAR, [2]DR.NISARG GANDHEWAR

Department of Computer Science and Engineering
Dr A.P.J. Abdul kalam University, Indore-452010(INDIA)

Abstract

There has been a recent uptick in the installation of high-tech video surveillance equipment in public areas. One of the primary uses for gathered video features is safety monitoring, made possible by the implementation of deep learning and machine learning techniques. In this work, our primary focus is on anomaly detection in situations with a large number of people, both indoors and outdoors. In this study, we describe the VGG 16 architecture for the detection of abnormalities occurring in surveillance cameras using classifiers. The VGG 16 architecture was implemented with state-of-the-art classifiers like random forest, J48, decision tree, and SVM. Experiments revealed that the suggested model of using VGG 16 has a relatively low computational burden while still producing satisfactory results, with an accuracy of 80.86% for the random forest classifier.
Keywords: Classifier, VGG 16, SVM, Anomaly, Random Forest, Architecture

1.0 Introduction

Data science and machine learning models use a large amount of data to identify and give valuable indicators for predicting the outcome of the classifier. In general, the models are built to give accurate state-of-the-art predictions for what purpose they are built for. The use of video surveillance is the most important tool for maintaining the calm and safety of a public place. The practise of using closed-circuit television (CCTV) is common in extremely busy areas, such as shopping malls, train stations, and other outdoor events. Because people are getting more worried about safety and security, each of these public places now has its own set of CCTV cameras.[1][2] The tracking devices are able to produce a wealth of aesthetically meaningful data. A crucial component of contemporary video surveillance is the capacity to recognise abnormal occurrences in situations when no humans are involved in the process in a successful, speedy, and efficient manner.

Camera surveillance systems are rapidly becoming an indispensable component of the advanced industries and smart cities of the current day. They are able to monitor high-risk regions for potential security threats, such as airports, bus stops, borders, and industrial zones, among other places. For traditional surveillance systems to work, there must always be someone on duty to keep an eye on the monitors and look for any potential security lapses. Because of this, maintaining a high degree of focus at all times is required, which increases the risk of making mistakes. [3] Intelligent surveillance technologies are gradually replacing their more archaic equivalents in the surveillance infrastructure. They are driven by computer vision algorithms, which give them the ability to automatically identify any abnormalities that may be present in a scene. New methods for collecting data on individuals passing through the field of surveillance cameras are being developed. This is done in order to collect information that can be used for the analysis. The data set that was made by using these methods can be used for a lot of different things, like regulating traffic in the same area and keeping track of, surveying, and counting the population.

The security infrastructures of emerging nations are now undergoing upgrades in order to better protect and manage both public and private crowds. Finding anomalies is fraught with peril when done in a crowded environment. Considering that the anomaly is responsible for injuries and damage to property in the public realm. Abnormality detection is frequently required in order to protect the health of the population as well as the environment whenever an anomaly takes place in a busy location. [9][10] When anything out of the ordinary is discovered, an alerting device must be employed to tell the people in the area. The notification system is available in a number of different formats, such as tones, speech, and text.

The warning system is able to quickly send out a message or tone if it detects something out of the ordinary in the crowd. Protection at a reasonable cost is something the government needs, especially in private and public spaces where there are a lot of people congregated together. Protection is necessary for individuals in big groups as well as during public and private events. As a direct result of this, the computer vision technique that is based on deep learning is able to provide a wide array of effective solutions for both private and public safety.

On a frame-by-frame basis, convolution neural networks are able to be utilised for the purpose of detecting abnormalities in video events (CNN). A CNN model that has been started and implemented with high-resolution video event frames is a component of the framework that has been suggested. The CNN model was built with the help of a large amount of training data. [11][12] Automated anomaly detection is highly helpful for reducing the quantity of data that has to be manually evaluated. This is accomplished by focusing attention on a small section of the data while disregarding enormous quantities of data that is irrelevant to the investigation at hand. In this paper, we will look at how the problem formulations and approach methods used in anomaly detection research relate to the automation of surveillance. [13][14][15]

1.1 VGG 16

Figure 1 displays the VGG16's internal architecture. There are a total of 13 convolutional layers and 3 fully connected layers in this architecture. Rectification (ReLU) non-linearity is built into all buried layers. Each convolutional kernel in this architecture has a size of 33. There are five distinct groups that may be made up of the thirteen convolutional layers. There are two distinct levels in Block 1. This cluster has 64 available channels. The second block also features two convolutional layers, but this time there are 128 channels. The third block has 512 channels and consists of three convolutional layers. The final two blocks each consist of 512 channels and three convolutional layers. A max-pooling layer of size 22 is placed after each block. When the fifth block is in place, the next three layers can be established and they are all interconnected. Each of the first two can hold 4096 channels. The 3,000-channel tier is the most impressive. It is possible to adjust the number of channels in fully linked layers to better suit a variety of datasets and activities. Figure 2 shows the VGG 16 model used for the study.

Journal of Advanced Research in Technology and Management Sciences
Volume: 04 Issue: 04 ISSN: 2582-3078 July 2022
Available online at: http://www.jartms.org

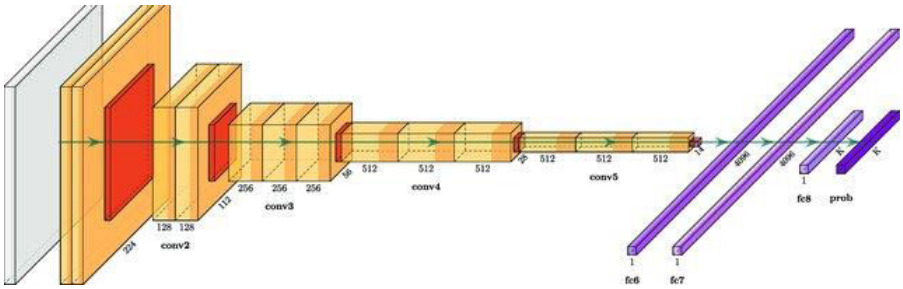

Figure 1: VGG 16 Architecture

The primary ways in which VGG16 improves upon the prior CNN architecture are by means of two new additions. First, it shrinks the convolutional kernel to a manageable 3x3. When compared to large kernels, smaller kernels can significantly reduce the network's computational and parameter requirements. While a larger convolutional kernel can collect more spatial data, this limitation can be worked around by using a larger number of smaller kernels. Moreover, VGG16's architecture demonstrates that a CNN's performance is influenced by its depth.

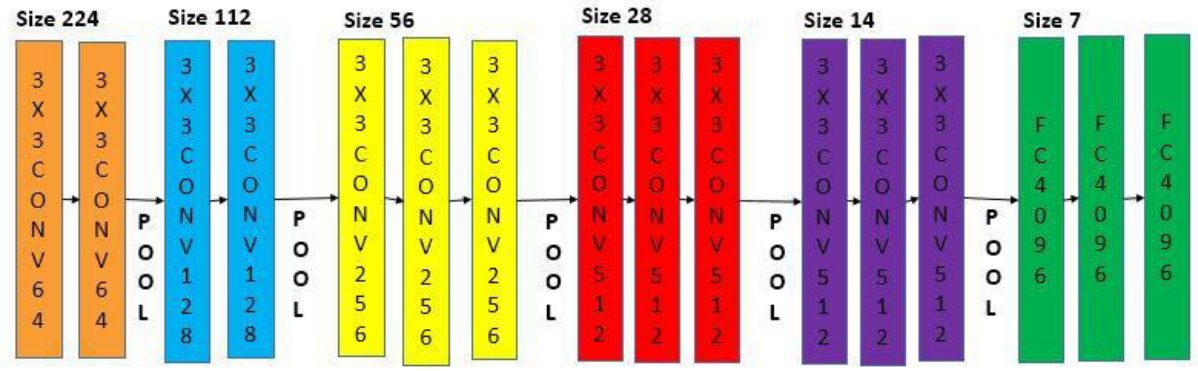

Figure 2 VGG 16 model

2.0 Related Work

One of the most well-known examples of deep learning architecture is the Convolutional Neural Network (CNN). When it comes to extracting spatial characteristics from photos and movies, it excels. Fifty years ago, the structure of the animal visual brain provided the inspiration for what would later become CNN. The researchers discovered that the arrangement of cells in the visual cortex of animals allows them to see light and make spatial perceptions. "Neocognitron" was the first neural network model to employ a similar architecture. This research presents an unsupervised approach to visual pattern identification using a multi-layer neural network. This model does not necessitate a common weight, unlike state-of-the-art CNN models. For the purpose of voice identification, the Time-delay Neural Network (TDNN) was created in 1987. One way to look at this model is as a pioneering attempt at a one-dimensional convolutional neural network structure. TDNNs are more computationally efficient since their weights are shared over a time dimension. Initially

implemented in the computer vision field in 1988, CNN has been used extensively since then. For the purpose of analysing medical images, it is suggested that a CNN with only two dimensions be used. But at the time, there just wasn't enough data or processing power to make widespread use of CNNs.

In the last 15 years, CNNs have exploded in popularity thanks to the advancement of deep learning and computation. In order to train CNNs more quickly and effectively, we have taken advantage of GPU-accelerated computing approaches. Regular NNs use matrix multiplication, but in CNNs we use convolution instead. By decreasing the total number of weights in the network, complexity is reduced. Additionally, the photos can be uploaded directly to the network as raw inputs, skipping the feature extraction process in the traditional learning algorithms. Due to the effective training of the hierarchical layers, CNNs are the first truly successful deep learning architectures. Standard backpropagation algorithms are used to boost performance, and the CNN architecture takes advantage of geographical correlations to reduce the number of network parameters. The CNN model benefits from needing little in the way of pre-processing, which is another plus. The subsequent chapters introduce some convolutional neural networks that will be used in later ones. Ryan et al. [4] proposed a different kind of visual representation, namely the textures of optical flow. The suggested representation uses uniformity measurements of a flow field to identify out-of-place objects like bicycles, automobiles, and skateboarders, and can be paired with spatial information to identify various forms of anomaly. [5] Nogueira et al. proposed an inexpensive deep learning strategy for real-time people counting inside stores and mapping out busy areas. We utilise a Convolutional Neural Network (CNN) regression model for supervised learning. In addition, there was a depiction of the image in four different channels. They implemented the approach of foreground and background detection that takes human behavioural quirks into account. [6] Vijeikis et al. proposed a system for detecting violent incidents in CCTV footage. The model is a U-Net-like network that use MobileNet V2 as an encoder, and then LSTM for temporal feature extraction and classification. [7] Majji et al. suggested a model for surveillance video behaviour modelling to find departures from conventional online behaviour. Data marking in training allows for automatic behaviour profiling and online anomaly sampling and detection without human interaction. [8] Amina and Binu proposed an algorithm to study the occurrence of anomalies in crowd surveillance. [11] A hire et al. proposed a VGG 16 based CNN model to detect anomalies.

3.0 Proposed Work

In this body of work, there has been widespread acceptance of the definition of a basic deep neural network model for the identification of anomalies. The Deep Neural Network (DNN) is a mathematical model that performs an analysis of a large number of categorised video input frames using broad strokes. The fundamental nervous system that may be found in animal brains serves as the inspiration for the layout of neural networks. The CNN Network is a multi-perspective network that is completely connected to itself. Within the crucial network, both weights and biases are referred to as network neurons. Two of the most important aspects of this network are the dot product and the neuron weight.

The pooling layer and the totally connected layer are both crucial components of a convolutional neural network. The CNN network layers are responsible for carrying out the primary purpose of the convolution network. During this period, each neuron that is part of

the network layer is linked to a tiny piece of frame data that is located nearby. The term "receptive field" refers to the limited space occupied by an individual. In the proposed system, the CNN model is responsible for identifying abnormalities in the video events. In order to extract the features of an anomalous instance, frames from photos are employed. A standard CNN classifier is used to analyse the characteristics that have been averaged from the video frames and then input into the system. The CNN model is trained using anomalous characteristics taken from the various layers of the model, each of which is constructed with a unique kernel for the purpose of anomaly detection. In a layer that is completely linked, the feature representations are created one after the other with the help of a feed-forward process. In the proposed system, the CNN model is responsible for identifying abnormalities in the video events. In order to extract the features of an anomalous instance, frames from photos are employed. A standard CNN classifier is used to analyse the characteristics that have been averaged from the video frames and then input into the system. The CNN model is trained using anomalous characteristics taken from the various layers of the model, each of which is constructed with a unique kernel for the purpose of anomaly detection. In a layer that is completely linked, the feature representations are created one after the other with the help of a feed-forward process.

4.0 Evaluation and Results

When we use neural networks, one of the things that we commonly aim to do is reduce the amount of inaccuracy. As a consequence of this, the objective function is frequently referred to as a cost function or a loss function, and the total that is generated by the loss function is simply referred to as "Model loss." The objective function, also known as the criteria, is the function whose value we are interested in minimising or optimising. When we try to minimise it, we may refer to it as the cost function, the loss function, or the error function. In this case, the loss function that is applied in order to do a comparison between the actual output and the anticipated output and whose ideal value is close to zero is the category cross entropy. If using an optimizer results in an incorrect comparison, the weight of the model will be adjusted, and the output will be made more accurate. Accuracy is a measurement of how well the model performs in all of the different classes. When the weight of each category is considered equally, it might be beneficial. The formula for determining it is the total number of forecasts divided by the total number of accurate predictions. A variable is used to hold the answer that is obtained by dividing the total number of values in the matrix by the total number of true positives and true negatives. The test dataset consisted of 500 image frames. There were 300 known anomalies and 200 normal. The experimental analysis is given in table 1. The VGG 16. Table 1 gives the number of records in the training and test datasets. The training dataset consists of 350 records of type anomaly and 188 records of type normal. The test dataset contains 143 records of type anomaly and 87 records of normal type.

Training Dataset		**Test Dataset**	
Dataset Type	**Number of Records**	**Dataset Type**	**Number of Records**
Anomaly	350	Anomaly	143
Normal	188	Normal	87
Total	538	Total	230

Table 1: Training and Test dataset records

	Classified Anomaly	Classified Normal
Anomaly	118	26
Normal	13	73

Table 2: Confusion Matrix

The efficiency of the classifier is computed with the given performance measures [9]. Table 3 contains the experimental performance of standard classifiers on VGG 16 architecture. The performance measures computed are as follows: $\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$ 1

$$\text{Precision} = \frac{TP}{(TP + FP)} \quad 2$$

$$\text{FMeasure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad 3$$

$$\text{Recall} = \frac{TP}{(TP + FN)} \quad 4$$

Classifier	Accuracy	Precision	Recall	f Measure
Random Forest	0.8086	0.8636	0.7972	0.8290
SVM	0.651	0.424	0.651	0.513
J48	0.739	0.678	0.739	0.742
Decision Tree	0.7382	0.735	0.738	0.736
Proposed Method	0.8304	0.8194	0.733	0.773

Table 3: Performance measure

Table 4 gives the comparison of the existing literature used in the study with the proposed methodology. The proposed method has a better accuracy.

Literature	Accuracy
Majji et a. [7] 2022	78.7
Shahbaz et al [13] 2020	76.4
Ryan et al. [4] 2011	72.2
Khan et al. [14] 2022	81.5
Proposed	83.04
Nogueira et al. [5]	69.3
Amina et al. [8] 2022	82.3
Saleem et al. [15] 2022	67.2
Vijeikis et al. [6] 2022	82
Mahdi et al. [ ] 2021	72.6

Table 4 :Comparison with exiting literature

Figure 3: Classifier Accuracy

Conclusion

The VGG 16 architecture was used and standard machine learning classifiers like random forest, SVM, J48 and decision tree were analysed on the provided dataset. The accuracy of the random forest was 80.86 with a precision value of 86. The recall and Fmeasure were 79 and 82. The accuracy of the proposed method is better than random forest. Because it makes use of deep learning techniques, the methodology that has been developed may recognise unusual patterns. The intricacy of video data makes it extremely difficult to comprehend unusual occurrences in video sequences, making this an extremely challenging endeavour. Convolution is used in this study to reduce the amount of work that must be done on the computer without compromising the quality

of the detection. When it comes to finding unusual things, the proposed CNN model is about 83.04% accurate as a whole.

References:

[1] Bendali-Braham, M., Weber, J., Forestier, G., Idoumghar, L. and Muller, P.A., 2021. Recent trends in crowd analysis: A review. Machine Learning with Applications, 4, p.100023.
[2] S´anchez, F.L., Hupont, I., Tabik, S. and Herrera, F., 2020. Revisiting crowd behaviour analysis through deep learning: Taxonomy, anomaly detection, crowd emotions, datasets, opportunities and prospects. Information Fusion, 64, pp.318-335
[3] Li, Xuelong, Mulin Chen, and Qi Wang. "Quantifying and detecting collective motion in crowd scenes." IEEE Transactions on Image Processing 29 (2020): 5571-5583.
[4] D. Ryan, S. Denman, C. Fookes, and S. Sridharan, "Textures of optical flow for real-time anomaly detection in crowds," in *2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Klagenfurt, Austria, Aug. 2011, pp. 230–235. doi: 10.1109/AVSS.2011.6027327.

[5] V. Nogueira, H. Oliveira, J. Augusto Silva, T. Vieira, and K. Oliveira, "RetailNet: A Deep Learning Approach for People Counting and Hot Spots Detection in Retail Stores," in *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, Rio de Janeiro, Brazil, Oct. 2019, pp. 155–162. doi: 10.1109/SIBGRAPI.2019.00029.

[6] R. Vijeikis, V. Raudonis, and G. Dervinis, "Efficient Violence Detection in Surveillance," *Sensors*, vol. 22, no. 6, p. 2216, Mar. 2022, doi: 10.3390/s22062216.

[7] V. Majji, D. V. Kakollu, M. S. Kumar, K. N. Soujanya, D. B. Kanthamma, and D. G. B. Rao, "Videobehavior Possible Identification And Recognition Of Abnormalities And Normal Behavior Profiling For Anomaly Detection Using Cnn Model," . *Vol.*, no. 14, p. 7, 2022.

[8] Amina P1 and Binu L. S, "Real Time Crowd Analysis and Anomaly Detection," In Review, preprint, Sep. 2022. doi: 10.21203/rs.3.rs-1977255/v1.

[9] Nair, V., Kosal Ram, P. G., & Sundararaman, S. (2019). Shadow detection and removal from images using machine learning and morphological operations. *The Journal of Engineering*, *2019*(1), 11–18. https://doi.org/10.1049/joe.2018.5241

[10] Rajesh Banala, D.Upender,: "Remote Home Security System Based on Wireless Sensor Network Using NS2", International Journal of Computer Science and Electronics Engineering, India, Vol. 2 Issue 2 (2012).

[11] Monali Ahire, Devarshi Borse, Amey Chavan, Shubham Deshmukh, Favin Fernandes
M. Ahire, D. Borse, A. Chavan, S. Deshmukh, and F. Fernandes, "Suspicious and Anomaly Detection," p. 7.

[12] C. Dhiman and D. K. Vishwakarma, "A review of state-of-the-art techniques for abnormal human activity recognition," *Engineering Applications of Artificial Intelligence*, vol. 77, pp. 21–45, Jan. 2019, doi: 10.1016/j.engappai.2018.08.014.

[13] A. Shahbaz, V.-T. Hoang, and K.-H. Jo, "Convolutional Neural Network based Foreground Segmentation for Video Surveillance Systems," in *IECON 2019 - 45th Annual Conference of the IEEE Industrial Electronics Society*, Lisbon, Portugal, Oct. 2019, pp. 86–89. doi: 10.1109/IECON.2019.8927776.

[14] A. A. Khan *et al.*, "Crowd Anomaly Detection in Video Frames Using Fine-Tuned AlexNet Model," *Electronics*, vol. 11, no. 19, p. 3105, Sep. 2022, doi: 10.3390/electronics11193105.

[15] G. Saleem, U. I. Bajwa, R. Hammad Raza, F. H. Alqahtani, A. Tolba, and F. Xia, "Efficient anomaly recognition using surveillance videos," *PeerJ Computer Science*, vol. 8, p. e1117, Oct. 2022, doi: 10.7717/peerj-cs.1117.

[16] M. Mahdi, A. J Mohammed, M. M. Jafer,"Unusual Activity Detection in Surveillance Video Scene: Review "Journal of Al-Qadisiyah for Computer Science and Mathematics Vol. 13(3) 2021 , pp Comp. 92–98, DOI : https://doi.org/10.29304/jqcm.2021.13.3.848