Detection of abnormality in CCTV Footage: Computer Vision

[1]D.Krishna Sai, [2]K.Sahithi, [3]D.Sameera (Ph.D.)
[1]UG Student, IIITDM KURNOOL,AP, India. [2]BVRIT, Narsapur, TS, India.
[3]Assistant Professor, Dept. of IT, BVRIT, Narsapur,T S, India

Abstract

Our project aims at detecting any abnormal behaviour in CCTV footage to solve many accidents, crimes, security issues, etc. CCTV footage gives the best information to detect any real-time abnormal behaviour and hence the CCTV footage is incorporated in the proposed project. Dataset is prepared by adding abnormalities by the use of Gaussian blur/noise, to replace the training set of normal images with abnormal images and to make sure that there is no bias in the dataset. To get better performance from our dataset, feature normalization is harnessed. Moreover, we used a Convolutional Neural Network (CNN) as our model to convolute and predict the images in our dataset, which achieves better performance than any other neural network existing today. The proposed neural network performs optimization based on Nadam and the experimental outcomes show that the proposed neural network offers superior results compared to that of its peers. Therefore, it has been observed that the proposed neural network obtains better accuracy than conventional neural networks.

1. Introduction

Now almost every video camera is prominently used for surveillance in areas that may need monitoring for security purposes. For this purpose we use closed-circuit Television (CCTV) but as we observe in day to day life the footage from the CCTV cameras goes unscanned as there is lots of data to be parsed. We propose that by the use of detection of abnormality in CCTV footage in real-time we can solve many security issues, accidents, crimes, etc which are not observed by man and which require attention in any case of emergency.

In this project, we develop an optimized neural network that takes grey-scale images from the footage of CCTV as inputs and outputs a prediction of abnormality in a binary form, inspired from (20,21). Furthermore, this trained model can be used in various real-world applications for the detection of abnormality. Therefore, the trained model only provides the classification of abnormality in the footage but does not tell about the part of the image that contains abnormality.

2. Related Work

We take artistry from Ler’s and Decker’s paper Exploration of Anomaly detection through CCTV Cameras: Computer Vision (1) by utilizing the image processing techniques for creating random noises and blurs on the images of the CCTV footage to create the class of abnormality. These image processing techniques are used to create the class of abnormality which is not present in the dataset employed.
To create the abnormalities in the image, Ler et al use the technique of adding abnormalities randomly by choosing half of the frames to have abnormalities added to them. Specifically, they have added the Gaussian noise/blur with 1/2 probability to each of the images they have used for training. One novelty in the Ler et al is that the adding of Gaussian blur/noise to make the class of abnormality. Although the saliency (see equation (8) of (7)) detection method is not used in this project, it could be the most important incorporation for better projects.

To create abnormalities in the training set, we use a similar method as adding Gaussian blur/noise with different parameters with the same probability ratio as described in Wilson Ler, Sean Decker’s paper “Exploration of Anomaly Detection through CCTV cameras: Computer Vision” (1). Some statistics were analyzed from (5) to detect abnormalities in densely crowded scenes. More detailed survey of detection of abnormality can be found in (13,14).

To explain the results obtained from our model we can use different evaluation metrics of a neural network. The suggested model is validated on the standard dataset, and the experimental results indicate that the model has better results than its competent models.

3. Dataset and Features

Our model of the neural network is constructed on UC San Deigo's Statistical Visual Computing Lab's UCSD Anomaly Detection Dataset Peds1 (2). This dataset of CCTV footage is acquired with a stationary camera mounted at an elevation above pedestrian walkways. From the study of crowd behaviour (4), the crowd density in this footage is irregular which varies from densely to sparsely crowded. The sparsity of the crowd can be measured by Kullback-Liebler (KL) divergence (6) for a clearer picture. And in the dataset, the abnormalities are due to the non-pedestrians on the pathway, which include skaters, bikers, people walking in the grass that surrounds the pathway, and golf carts.

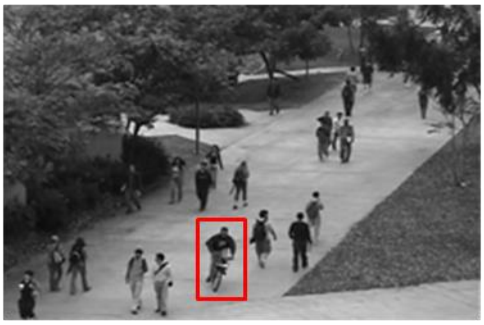


Figure.1. Example images of abnormalities from the clips of UCSD Anomaly Detection Dataset. The Image consists of a golf cart on the pathway

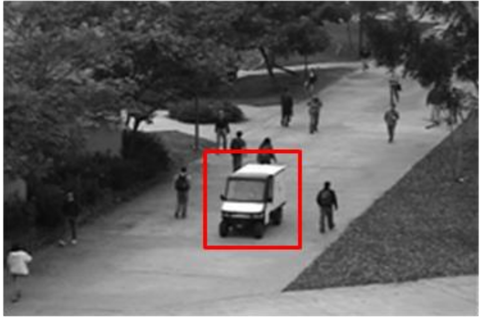


Figure.2. Image consists of a biker in between the pedestrians on the pathway.

Figure.3. Example of a normal image from the clips of UCSD Anomaly Detection Dataset.

Coming to the description of the dataset. It is already well organized into training and testing sets, where the training set consists of 34 video samples, and the testing set consists of 36 video samples, with each of

200 grayscale and no alpha channel tiff images in series. The video samples present in the training set does not contain any abnormality. Moreover, most of the video samples present in the testing set contain abnormality at some subset of a sample. The testing set contains labels with appropriate frames as either normal or abnormal.

3.1 Data Preparation

As the training set comprises only the class of normal images, there is a need to create an abnormality for training our model. To create abnormalities in the training set we randomly add Gaussian noise/blur with a probability of 1/2 to each of the images in half of the training set. We choose the size of the window of Gaussian blur/noise to be 50x50 pixels as it replicates the size of a golf cart in the image which is the maximum possible size of an abnormality. An Example of adding gaussian blur/noise is given in Fig3.

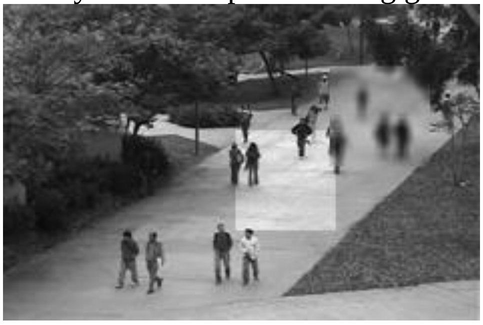


Figure.4. Example of a training image that had two squares (50x50) of error added to it, one of blur and one of noise

3.2 Feature Normalization

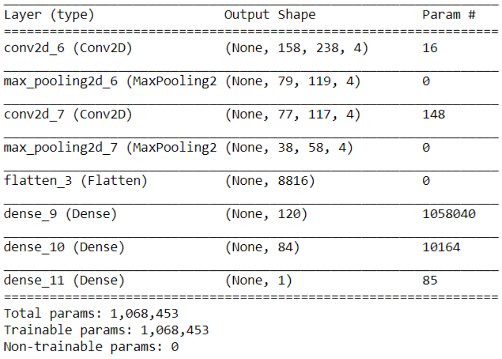
Almost everywhere to get a better classification performance, there is a need for normalization of features known as Min-Max scaling. The feature X is normalized to Xn using the formula:

$$X_n = (X - X_{min}) /(X_{max}-X_{min})$$

Where, Xmax and Xmin are the maximum and minimum values of the features respectively.

4. Classification

The model we have concentrated on is a convolutional neural network (CNN) inspired from (8,24), which is majorly used to extract features from the images and predict. This neural network architecture is motivated by (3,9,10). Each layer in a neural network is the addition and multiplication of weights and biases learned to give an output when provided input. Our convolutional neural network consists of two sets of convolution layers with an and a max pool layer, with the first convolution layer comprising 1 to 1 layer to maintain most of the features to the next layer and followed by a max pool window of 2x2. The second convolution layer convolves input with a 3x3 polynomial. The last three of the network are all dense layers with activation functions and biases.This is also a memory efficient implementation inspired from the architecture in (25). Our CNN model can be expressed as in Fig 4.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_6 (Conv2D)            (None, 158, 238, 4)       16
_________________________________________________________________
max_pooling2d_6 (MaxPooling2 (None, 79, 119, 4)        0
_________________________________________________________________
conv2d_7 (Conv2D)            (None, 77, 117, 4)        148
_________________________________________________________________
max_pooling2d_7 (MaxPooling2 (None, 38, 58, 4)         0
_________________________________________________________________
flatten_3 (Flatten)          (None, 8816)              0
_________________________________________________________________
dense_9 (Dense)              (None, 120)               1058040
_________________________________________________________________
dense_10 (Dense)             (None, 84)                10164
_________________________________________________________________
dense_11 (Dense)             (None, 1)                 85
=================================================================
Total params: 1,068,453
Trainable params: 1,068,453
Non-trainable params: 0
```

Fig4: The complete description of our convolution neural network with the number of parameters at every layer.

The convolution layer makes the process of feature extraction from the images provided but as to say that the convolution process is mathematically complex and the formula is given as:

$$(f * g)(t) \overset{\text{def}}{=} \int_{-\infty}^{\infty} f(\tau)\, g(t-\tau)\, d\tau$$

The most prominent thing in our model is the use of ReLu as an activation function in the layers of convolution and dense. The formula is:

$$R(z) = \max(0,z)$$

As the weights are not initialized initially, the weights are being learned in the neural network by the process of backpropagation. This approach has also helped to generate better results in different applications(15).

5. Experimental Results and Comparisons

The experiments are carried out on a PC with 16 GB of RAM, 2.8GHz Core-i7 processor and NVIDIA GTX 1060 GPU 3GB running on windows 10 operating system. The proposed neural network is emulated in Tensor flow.

The important attribute of the abnormality detection is that the normal and abnormal images are equal in number as the normal images were converted to abnormal images using Gaussian blur/noise and as there is an equal amount of normal and abnormal images, we will focus on the accuracy of the model of how good it predicts a given image. Although the frames present in the video sample are predicted and not the entire video sample, which leads to a frame-wise prediction process. We utilized the accuracy metric in our neural network because it is the ratio of correct prediction to the total number of predictions. The simple formula of accuracy is given as :

$$\text{Accuracy} = (TP+TN)/TP+TN+FP+FN$$

The settings for the inputs and outputs for our neural network is shown in the Table 1.

Total number of images(14000)		Training images or frames(6800)		Testing images or frames(7200)	
Normal	Abnormal	Normal	Abnormal	Normal	Abnormal
7000	7000	3400	3400	3600	3600

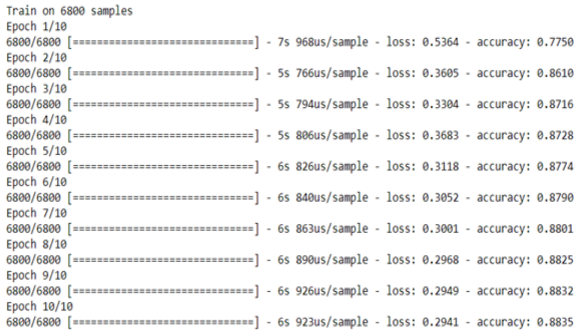
Table .1 Settings of Inputs and Outputs for our 'Dataset' (2)

So in this work, we consider Normal and abnormal classes as positive and negative classes. After the normalization process, our neural network is compiled with 'Nadam', it is a momentum-based optimizer and it performs better than a conventional optimizer. 'binary cross-entropy' is used as the loss function and is the most useful parameter in the case of binary classification. Moreover, the learning rate is set to 0.01, which is typical for a neural network and momentum-based optimizers.

The improvement of classification accuracy was achieved from (11, 12).

6. Results and Comparisons

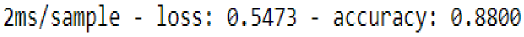
Our convolutional neural network was able to converge in around 10epochs and the training accuracy achieved is around 88%. The detailed description of the Training is shown in Fig5.



```
Train on 6800 samples
Epoch 1/10
6800/6800 [==============================] - 7s 968us/sample - loss: 0.5364 - accuracy: 0.7750
Epoch 2/10
6800/6800 [==============================] - 5s 766us/sample - loss: 0.3605 - accuracy: 0.8610
Epoch 3/10
6800/6800 [==============================] - 5s 794us/sample - loss: 0.3304 - accuracy: 0.8716
Epoch 4/10
6800/6800 [==============================] - 5s 806us/sample - loss: 0.3683 - accuracy: 0.8728
Epoch 5/10
6800/6800 [==============================] - 6s 826us/sample - loss: 0.3118 - accuracy: 0.8774
Epoch 6/10
6800/6800 [==============================] - 6s 840us/sample - loss: 0.3052 - accuracy: 0.8790
Epoch 7/10
6800/6800 [==============================] - 6s 863us/sample - loss: 0.3001 - accuracy: 0.8801
Epoch 8/10
6800/6800 [==============================] - 6s 890us/sample - loss: 0.2968 - accuracy: 0.8825
Epoch 9/10
6800/6800 [==============================] - 6s 926us/sample - loss: 0.2949 - accuracy: 0.8832
Epoch 10/10
6800/6800 [==============================] - 6s 923us/sample - loss: 0.2941 - accuracy: 0.8835
```

Fig5: The description of training of our convolution neural network with loss and accuracy values at every epoch.

With our neural network trained, we have tested it on a set of test video samples of around 7200 frames, then we have achieved testing loss and accuracy as:



```
2ms/sample - loss: 0.5473 - accuracy: 0.8800
```

And on the empirical basis, we say that the accuracy reduces as we increase the size of our testing dataset. Although on a holistic view, our model or neural network is trained without any bias towards a class examples used in training or testing, it performs poorly on the biased dataset. From the table1 we see that there is no bias in the training and testing datasets. Also, there could be few changes to be done for generalizing the model or neural network.

If we compare our neural network with that of Wilson Ler, Sean Decker's paper "Exploration of Anomaly Detection through CCTV cameras: Computer Vision"(1) we can see that they failed to predict false negatives and so the model is biased. And also their testing accuracy of the network on 10000 images is 48% which is less compared to our model where the testing accuracy on 7200 images is 88%. This proves that our model performs efficiently than existing models for the purpose of abnormality detection. To conjecture that the model is able to perfectly identify unbiased dataset. Also, we could identify that the model fails to predict the anomalies which are very much less than the size of a golf cart, and by decreasing the size of the window of Gaussian blur/noise our model does not give satisfactory results.

7. Conclusion

This project proposes an efficient model to predict the abnormalities in a frame of CCTV footage. The convolutional neural network (model) does a feature extraction process and a prediction process. But the normalization is done prior to the prediction manually by using the given formula. The classification accuracy of Dataset (2) is 88%.

In future works, we could make our model better by using RGB images in our training and testing sets. Also by having a huge amount of dataset with an equal amount of class examples and by preventing the use of the addition of abnormalities by the use of Gaussian blur/noise. As another possibility, auto encoders (17),(16) and variants of auto encoders(18),(19)can also be used in place of CNNs. From our observation, we can also use our model in the fields of medical imaging to detect abnormalities in the scanned images. The CT(Computed Tomography) and MRI(Magnetic Resonance Imaging) scans are also low-resolution images of gray-scale. Therefore, there is a possibility of utilizing our model for CT and MRI scans dataset for the betterment of the prediction of medical images or scans.

Mobile-based Active Authentication (22,23) is another field of binary classification which is gaining interest recently,

8. References

[1] Wilson Ler, and Sean Decker. “Exploration of Anomaly Detection through CCTV cameras: Computer Vision” Stanford University, Stanford, CA 94305.

[2] Anomaly Detection and Localization in Crowded Scenes.” Anomaly Detection and Localization in Crowded Scenes, Statistical Visual Computing Lab: UC San Diego, http://www.svcl.ucsd.edu/projects/anomaly/.

[3] Perera, Pramuditha, and Vishal M. Patel. “Learning Deep features for One-Class Classification.” IEEE Transactions on Image Processing 28.11 (2019): 5450-5463. Crossref. Web.

[4] D. Helbing and P. Moln´ar. Social force model for pedestrian dynamics. *Physical Review E*, 51(5):4282–4286, May 1995.

[5] Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *CVPR09*, pages 1446–1453, 2009.

[6] S. Kullback. *Information Theory and Statistics*. Dover Publications, New York, 1968.

[7] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *CVPR*, pages 935–942, 2009.

[8] P. Samangouei and R. Chellappa. Convolutional neural networks for attribute-based active

[9] authentication on mobile devices. In IEEE International Conference on Biometrics: Theory, Applications and Systems, Sept 2016.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems 25, pages 1097–1105, 2012.

[11] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image
a. recognition. CoRR, 2014.

[12] D. A. Clifton, S. Hugueny, and L. Tarassenko. Novelty detection with multivariate extreme value statistics. Journal of signal processing systems, 65(3):371–389, 2011.

[13] S. J. Roberts. Novelty detection using extreme value statistics. IEE Proceedings-Vision, Image and Signal Processing, 146(3):124–129, 1999.

[14] M. Markou and S. Singh. Novelty detection: a review – part 1: statistical approaches. Signal Processing, 83(12):2481 – 2497, 2003.

[15] M. Markou and S. Singh. Novelty detection: a review – part 2: neural network based approaches. Signal Processing, 83(12):2499 – 2521, 2003.

[16] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In E. P. Xing and T. Jebara, editors,Proceedings of the 31st International Conference on Machine Learning, volume 32, pages 647–655, Bejing, China, 22–24 Jun 2014.

[17] R. Hadsell, S. Chopra, and Y. Lecun. Dimensionality reduction by learning an invariant mapping. In In Proc. Computer Vision and Pattern Recognition Conference (CVPR06. IEEE Press, 2006.

[18] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th International Conference on
a. Machine Learning, pages 1096– 1103, 2008.

[19] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research, 11:3371–3408, 2010.

[20] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In International Conference on Learning Representations.

[21] D. Bankman, L. Yang, B. Moons, M. Verhelst and B. Murmann, "An Always-On 3.8 μ J/86% CIFAR-10 Mixed-Signal Binary CNN Processor With All Memory on Chip in nm CMOS," in *IEEE Journal of Solid-State Circuits*, vol. 54, no. 1, pp. 158-172, Jan. 2019, doi: 10.1109/JSSC.2018.2869150.

[22] Xiaofan Lin, Cong Zhao, Wei Pan, “Towards Accurate Binary Convolutional Neural Network. CoRR abs/1711.11294(2017).

[23] V. M. Patel, R. Chellappa, D. Chandra, and B. Barbello. Continuous user authentication on

[24] mobile devices: Recent progress and remaining challenges. IEEE Signal Processing Magazine, 33(4):49–61, July 2016.

[25] P. Perera and V. M. Patel. Face-based multiple user active authentication on mobile devices. Transactions on Information Forensics and Security, 14(5):1240–1250, 2019

[26] P. Perera and V. M. Patel. Deep transfer learning for multiple class novelty detection. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019

[27] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. arrel. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22Nd ACM International Conference on Multimedia, pages 675–678, 2014.