International Journal of Research in Engineering, Science and Management
Volume-2, Issue-4, April-2019
www.ijresm.com | ISSN (Online): 2581-5792

Video Anomaly Detection through Live Surveillance

Vina Lomte[1], Satish Singh[2], Siddharth Patil[3], Siddheshwar Patil[4], Durgesh Pahurkar[5]
[1]*Professor & HoD, Department of Computer Engineering, RMD Sinhgad School of Engineering, Pune, India*
[2,3,4,5]*Student, Department of Computer Engineering, RMD Sinhgad School of Engineering, Pune, India*

***Abstract*: Anomaly detection is a method of identifying an abnormal activity through the live video sequence. Our proposed system uses Autoencoders for detecting anomaly activities. Nowadays there is a need for video anomaly detecting systems to prevent crimes and bad thigs from happening. In this paper we provide a system that detects the anomalous events by implementing the ConvLSTM auto encoder which gives the results depending on the type of the events. We achieve an accuracy of 74% on avenue dataset.**

***Keywords*: Long Short Term Memory, Convolutional Neural Networks, autoencoders, anomaly detection.**

1. Introduction

Nowadays, video anomaly detection is used to monitor public places in order to control crime and abnormal activities such as burglary, fight, abuse, etc. There is always a big risk to anomaly because of unmannered behavior of humans. The goal of anomaly detection is to minimize risk and instantly detect the anomaly happening in the surrounding. The successful detection will result in the detection of the type of anomaly. Anomaly detection is the unmannered behavior differing significantly from the major of the data. Abnormal behavior contains the issues like noise, overcrowding of people, happening of none structured event, different abnormal events. The hardest task is tracking in real time processing and dynamic environment that have lot of moving objects. The model we used in our proposed work is ConvLSTM. The CNN Long Short-Term Memory Network or CNN LSTM for short is an LSTM architecture specifically designed for sequence prediction problems with spatial inputs, like images or videos. The CNN Long Short-Term Memory Network or CNN LSTM for short is an LSTM architecture specifically designed for sequence prediction problems with spatial inputs, like images or videos. ConvLSTM is basically a sequential model which The model type that used is Sequential. Sequential is the easiest way to build a model for autoencoders. It allows us to build a model layer by layer. Each layer has weights that correspond to the layer the follows it.

A CNN LSTM can be defined by adding CNN layers on the front end followed by LSTM layers with a Dense layer on the output. It is helpful to think of this architecture as defining two sub-models: the CNN Model for feature extraction and the LSTM Model for interpreting the features across time steps. conceptually there is a single CNN model and a sequence of LSTM models, one for each time step. We want to apply the CNN model to each input image and pass on the output of each input image to the LSTM as a single time step.

2. Related work

Video anomaly detection is a field which is getting lot of attention these days due to increasing crime rates and availability of video surveillance for most parts of the city. Many researchers have come up with working models which despite of having some advantages are not suitable for real world application due to their high learning complexity and unacceptable error rates.

In [1] the authors proposed the model in which they can predict whether event is normal or abnormal based on chaos model. Model can adapt to the environment, such as lighting and other background changes, it does not need a lot of training data. Model can update parameters according to the slow changes of the scene, and model can achieve real-time update effect. But it can only predict for an anomaly being abnormal or normal and is suitable for a specific environment.

In [2] the authors describe predefined set of relevant normal events for detecting an abnormal event. Model is able to detect and locate multiple abnormal events in the scene at the same time, it reaches to 252 frames/sec. But it lacks in accuracy that reaches only upto 65%-68%, which by far cannot be accepted.

In [3] the authors implemented a model with low complexity, accurate and reliable anamoly detection and localization. The training process is every time consuming and every non dominant object is considered as an anomaly activity. But training is time consuming, every non dominant object is considered as an anomaly (person in a car is considered an anomaly if other all people are pedestrians).

In [4] the authors describe sparse dictionary to detect anomaly detection activities. It has outperformed every other model in terms of anomaly detection. Although false alarm rate is reduced but not completely eliminated and it also needs high computational resources and training time. Accuracy still needs improvement in terms of True negatives.

3. Our contribution

Our contribution to this field are as follows:

- We propose a sequential auto encoder ConvLSTM model which takes in the input video frames as a continuous sequence and detect the normal or abnormal events based on the computed Euclidean distance loss between input frame and reconstructed frame.
- We try to find the type of anomaly occurred based on reconstruction values above threshold, this will help in prioritizing and planning actions against various type of crime or abnormal situations.

4. Proposed system

A. *Problem statement*

To process real time video and generate alerts(notifications) when some predefined type of anomaly is generated.

B. *System architecture*

Architecture as shown in the Fig. 1, In our proposed work, we first pre-processed the input video frames from the dataset which are then stored in the image dump as a separate file. This image dump contains the weights of the different pre-processed input frames. This image dump is then used in training our ConvLSTM model which takes in the input as the image dump. We are using auto-encoder for reducing dimensions of the input video frame while processing to require less computational resources. While training we generate values for reconstruction threshold of Convolutional LSTM (Long Short term memory) for normal and abnormal events the proposed model takes in the input video sequence and passes the video sequence to the spatial encoder part of the model. After this model applies both encoder and decoder on the input video frame and further passes it to the spatial decoder. Spatial Encode includes the convolution of the input frames in which the previous video frames are also combined. In Spatial Decoder the same convoluted input video frames are deconvoluted which gives us the final reconstruction of input video sequence.

During testing and operation phase we provide an unknown video sequence to the model and it alarms the system user of any anomaly occurring.

C. *Mathematical model*

Convolutional LSTM: Fully connected Long short term memory(FC-LSTM) is variant of recurrent neural networks used for SS learning based on history. The major drawback of FC-LSTM in handling spatio-temporal data is its usage of full connection in input to state and state to state transitions in which no spatial information is encoded.

To overcome this problem, a distinguishing feature of our design is that all the inputs $x_1, \ldots, x_t$, cell outputs $c_1, \ldots, ct$, hidden states $h_1, \ldots, h_t$, and gates i_t, f_t, o_t of the ConvLSTM are 3D tensors whose last two dimensions are spatial dimensions. A variant of the LSTM architecture, namely Convolutional Long Short-term Memory (ConvLSTM) model was introduced by Shi et al. in [5] and has been recently utilized by Patraucean et al. in [6] for video frame prediction. Compared to the usual fully connected LSTM (FC-LSTM), ConvLSTM has its matrix operations replaced with convolutions. By using convolution for both input-to-hidden and hidden-to-hidden connections, ConvLSTM requires fewer weights and yield better spatial feature maps. The formulation of the ConvL STM unit can be summarized with (1) through (5).

$$i_t = (W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (1)$$

$$f_t = (W_{xf}\, x_t + W_{hf}\, h_{t-1} + W_{cf}\, c_{t-1} + b_f) \quad (2)$$

$$c_t = f_t\, c_{t-1} + it*\tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (3)$$

$$o_t = (W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (4)$$

$$h_t = o_t*\tanh(c_t) \quad (5)$$

The input is fed in as images, while the set of weights for every connection is replaced by convolutional filters (the symbol denotes a convolution operation). This allows ConvLSTM work better with images than the FC-LSTM due to its ability to propagate spatial characteristics temporally through each ConvLSTM state.

The optimizer controls the learning rate. We have used ‘adam’ as our optmizer. Adam is generally a good optimizer to use for many cases. The adam optimizer adjusts the learning rate throughout training. The learning rate determines how fast the optimal weights for the model are calculated. A smaller learning rate may lead to more accurate weights (up to a certain point), but the time it takes to compute the weights will be longer. For our loss function, we will use ‘mean_squared_error’. It is calculated by taking the average squared difference between the predicted and actual values. It is a popular loss function for regression problems. The closer to 0 this is, the better the model performed.

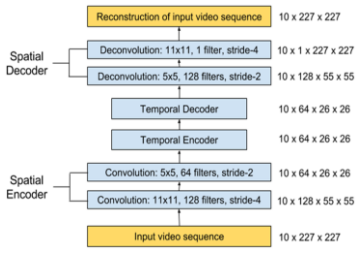

Fig. 1. Shows the architecture of this system

5. Algorithm

- Begin
- Preprocess the input frames from the dataset.
- Store the preprocessed input frames into image dump.
- Train the Autoencoder(ConvLSTM) by setting the epochs 30 with batch size = 1.

- Test and run the model threshold = 0.00065 if loss > threshold: Anomaly detected on bunch frames.
  else: Normal bunch detected, End

6. Result

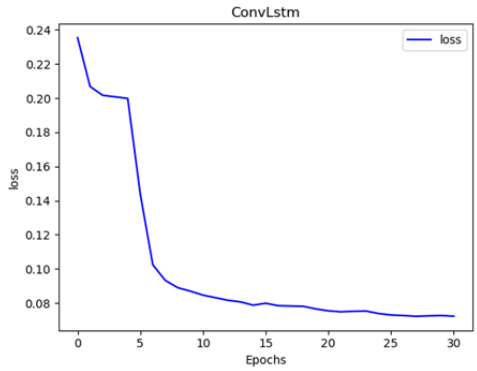

Fig. 2. Epochs vs. loss

As we can see from the above graph that as we keep increasing the number of epochs(iteration) for training our ConvLSTM model, the loss of some important features gets decreased.

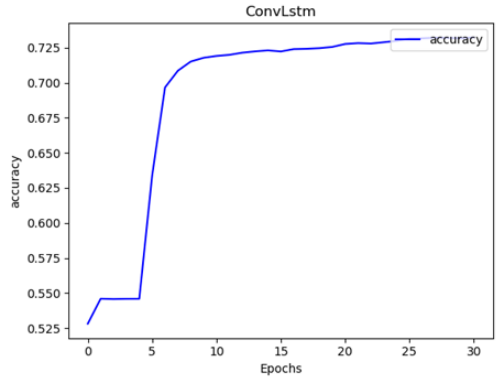

Fig. 3. Epochs vs. Accuracy

As we can see from the above graph that as we keep increasing the number of epochs(iteration) for training our ConvLSTM model, we get the better accuracy. The more the number of epochs, the better the accuracy we get. And the accuracy is found to be near around 73%. The confusion matrix for the conv LSTM is as shown.

	Abnormal	Normal
Abnormal	15	20
Normal	7	53

The accuracy of our proposed conv LSTM model was found to be 74%.

7. Conclusion and future applications

Our method which is the Conv LSTM model being sequential is able to predict classify the normal and abnormal events without determining normal and abnormal events beforehand and will be more suitable for real-time application such as live surveillance and crime detection. This system can be further integrated with audio sensor to increase accuracy or confirm results of video processing sys-tem. We can further collaborate with government and integrate this system with public surveillance to detect public crimes and get faster crime response than what it is currently. This system can also help private organizations and stores to monitor large amount of employees/people.

Acknowledgement

We would like to acknowledge and thank The Chinese University of Hong Kong for making the avenue dataset publicly available. Avenue Dataset contains 16 training and 21 testing video clips. The videos are captured in CUHK campus avenue with 30652 (15328 training, 15324 testing) frames in total.

References

[1] Zhaohui Luo, Weisheng He, Minghui Liwang, Lianfen Huang and Yifeng Zhao Real-Time Detection Algorithm of Abnormal Behavior in Crowds Based on Gaussian Mixture Model, August 22-25, 2017.
[2] S. Maryam Masoudirad and Jawad Hadadnia Anomaly Detection in Video Using Two-part Sparse Dictionary in 170 FPS, April 19-20, 2017.
[3] Mohammad Sabokrou, Mahmood Fathy, Mojtaba Hoseini, Reinhard Klette Real-Time Anomaly Detection and Localization in Crowded Scenes, 26 October 2015.
[4] Waqas Sultani, Chen Chen, Mubarak Shah Real-world Anomaly Detection in Surveillance Videos, 31 Mar 2018.
[5] Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.k., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. pp. 802-810. NIPS'15, MIT Press, Cambridge, MA, USA (2015).
[6] Patraucean, V., Handa, A., Cipolla, R.: Spatio-temporal video autoencoder with dierentiable memory. International Conference On Learning Repre-sentations (2015), 1-10 (2016).
[7] Cong, Y., Yuan, J., Liu, J.: Sparse reconstruction cost for abnormal event detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 3449-3456 (2011).
[8] Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S.: Learning temporal regularity in video sequences. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 733-742 (June 2016)
[9] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu. Violence detection using oriented violent flows. Image and Vision Computing, 2016.
[10] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. Learning temporal regularity in video sequences. In CVPR, June 2016.
[11] A. Sodemann, M. P. Ross, and B. J. Borghetti, A review of anomaly detection in automated surveillance, Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 42, no. 6, pp. 1257-1272, 2012.
[12] L. Kratz, and K. Nishino," Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models." pp. 1446-1453.
[13] §. Yi, H. Li, and X. Wang, Understanding pedestrian behaviors from stationary crowd groups, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 34883496, June 2015.
[14] R. A. A. Rupasinghe, S. G. M. P. Senanayake, D. A. Padmasiri, M. P. B. Ekanayake, G. M. R. I. Godaliyadda, and J. V. Wijayakulasooriya, Modes of clustering for motion pattern analysis in video surveillance, in 2016 IEEE International Conference on Information and Automation for Sustainability (ICIAfS), Dec 2016, pp. 16.

[15] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In NIPS, pages 577584, Cambridge, MA, USA, 2002. MIT Press.
[16] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In CVPR, 2016.
[17] A. Gordo, J. Almazan, J. Revaud, and D. Larlus. Deep image retrieval: Learning global representations for image search. In ECCV, 2016.
[18] M. J. Roshtkhari, and M. D. Levine, an on-line, real-time learning method for detecting anomalies in videos using spatiotemporal com-positions, Computer Vision and Image Understanding, vol. 117, no. 10, pp. 1436-1452, 2013.
[19] W. Hu, T. Tan, L. Wang, and S. Maybank, A survey on visual surveillance of object motion and behaviors, Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 34, no. 3, pp. 334-352, 2004.
[20] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, Toward a practical face recognition system: Robust alignment and illumination by sparse representation, Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 34, no. 2, pp. 372-386, 2012.
[21] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, Anomaly detection in crowded scenes, 2010.