Volume 10, Issue 5
October 2021

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

Deep Learning Based Multi Class Human Violent Action Recognition

[1]Devang Jani and [2]Dr. Anand Mankodia

[1]Research Scholar, Ganpat University, Ganpat Vidyanagar, Mehsana. 384001, State: Gujarat, Country: India.

[2]Associate Professor, EC, Ganpat University - U. V. Patel College of Engineering, Ganpat Vidyanagar, Mehsana. 384001.

State: Gujarat, Country: India.

E-mail: [1]devangjani645@gmail.com , [2]anand.mankodia@ganpatuniversity.ac.in

ABSTRACT

With the accelerated growth of surveillance systems, human violent action recognition as a sub domain of abnormal activity recognition is in great demand in the field of computer vision. Early recognition of violent activities can be really helpful in governance of situations, collective security and safety, mitigating as well as prevention of adversaries. Actions which pose threat to human security and safety can be termed as violent events. On the contrary to manual inspection of such human violent actions with limited capabilities, deep learning can provide better and autonomous violent activity recognition. The proposed work represents fine-tuned deep learning model VGG16 which recognizes human violence from surveillance videos into multiple classes. The results obtained from experiments are improved and holds relevant to state of the art. Finally, this paper tentatively indicates which deep learning model can be more relevant for real time human action recognition based on defined resource parameters.

Keywords: *Surveillance Systems, Adversaries, Human Violent Action Recognition, Deep Learning, VGG16*

1. INTRODUCTION

It is due to technological advancement, there is a boom in deployment of surveillance systems on various private as well as public locations. The modern surveillance infrastructure enables better scope in governance, security, safety, risk management, prediction of events and prevention of adversaries etc. Further, environmental dynamics can be better understood through continuous surveillance. The accelerated growth of surveillance systems and their implied potential applications have drown researchers attention to object detection and tracking in the field of computer vision. To be precise, under broad area of object detection and tracking, there is hype for human action recognition as sub-domain of behavior understanding. Understanding human behavior through the lenses of technology can reveal previously unknown, hidden, potentially useful patterns which can help govern situations and places efficiently. Besides, it can play big role in security and safety of people. Thus, with the increase in crimes, there is a growing interest for violent action recognition in behavior understanding. In the context of human violent action recognition, violent behavior could be defined as events that pose threat to security and safety of human life. In other words, human actions such as Pushing, Kicking, Punching, Fighting, Stabbing, Gun-firing etc can be termed as violent behavior.

In the past few decades, there has been a significant research and developments in the area of human violent action recognition. The most of the current work can be divided into two groups, before and after the introduction of machine learning in the field of computer vision. As compared to traditional approaches, machine learning based human action recognition is proven to be more efficient as with increased ability to extract complex features. Human actions are versatile in nature. Humans are multi-tasking and also have ability to do same action in multiple ways. Traditional approaches mostly get limited results due to inefficient feature extraction and it also requires partial human intervention. Deep learning based approaches on the other hand can learn and extract features on their own automatically. In this paper, most popularly used deep learning models namely CNN, LSTM and VGG16 are explored in detail. [1-2]

Section 2 of this paper encloses related work that has been done in this area. Section 3, explains implementation of considered deep learning models. Section 4, consists of experimental results and discussion. And final in Section 5 concludes the paper.

1.1 Human Violent Activity Recognition

In general human violent activity recognition follows below mentioned steps as shown in Fig. 1. Initially surveillance videos are segmented into image frames and pre-processing is done to make data clean and uniform. Then from segmented data feature vectors are extracted. These feature vectors are dependent on application domain. These extracted features are then fed to human violent action recognition models. From training data, model learns and gives probability distribution for each class label. The label with average max probability score is assigned for data input.

2. RELATED WORK

In recent years, adequate amount of research has been done in violent activity recognition with growing state of the art approaches.

Nian Chi Tay, Tee Connie et.al demonstrated use of CNN for abnormal human behavior recognition. The experiment involved working with 5 public benchmark data sets covering range of human activities. Conventional feature information such as edge, shape, color was used. The CNN configuration consisted of 6 total 6 layers out of which 3 Conv, 2 FC and 1 Softmax layers were present.

Volume 10, Issue 5
October 2021

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

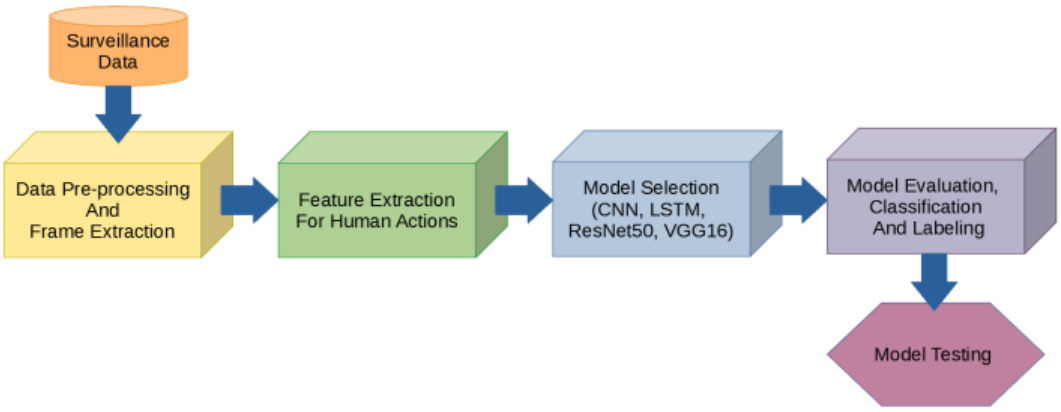

Figure 1. General Approach for Human Violent Action Recognition

For faster training ReLU was used. The proposed work demonstrated importance of choosing learning rate for training model and its impact of accuracy. [3]

S. Jothi Shri, S. Jothilakshmi came up with anomaly recognition/localization based on deep CNN in combination with GSM module which will produce alarm messages to send across. It was mainly focused on public crowd events such as protests, running, fighting, firing etc. The proposed method achieved close to 90% accuracy and was implemented on VGG16 baseline architecture. [4]

Ullah W., Ullah A. et.al presented new multi layer bi-directional LSTM framework (BD-LSTM) where two RNNs stacked together one for backward and one for forward pass was configured. The adaptability of method enables less reliance of available training data and can learn from growing data. On UCF Crime and UCFCrime2Local data sets, the proposed approach performed better as compared to TSN, MIL, SVM, weakly supervised, Optical flow, etc approaches. [5]

Mohamed Mostafa Soliman, Mohamed Hussein Kamal et.al proposed combination of VGG16 and LSTM deep network to extract spatial and temporal features related to human violent action recognition in crowded/non crowded surveillance places. The results of the experiment was comparable to the state of art with accuracy of 88.2% on hockey, movie, violent flow, RLVS data sets. [6]

Giorgio Molares, Itamar Salazar-Reque et.al also implemented similar combination of VGG16 and two deep conv LSTM networks for detecting violent robberies with accuracy of 96% on UCI-Crime dataset and newly proposed data set UCN-Crime dataset. [7]

Unnati Koppikar, C. Sujatha et.al carried out comparative analysis on two deep networks VGG16 and InceptionNet with UCF-Crime public data set for focusing on violent events such as theft, robbery, shoplifting, etc. The experiment Showed better performance of InceptionNet against VGG16 with 94.54% accuracy. [8]

To the best of our knowledge, usage of deep learning models in human violent action recognition is still finding its expression and not much of experiments have been carried out using the state of the art such as VGG16. Besides, there is less research available for uncommon events involving weapons such as shooting with gun, hit with object, stabbing etc. Public data-sets with armed violent events are scares. In proposed work, it is intended to cover events that include weapons along with other violent activities.

3. IMPLEMENTATION

In the presented work, three well known deep learning approaches Convolutional Neural Network (CNN), Long Short Term Memory Network (LSTM), and VGG16 respectively were implemented. The experiments were conducted on PARAM Shavak super computer situated at Government Engineering College, Rajkot, Gujarat, India under the grant of GUJCOST. PARAM Shavak is developed by C-DAC and comprises of 2-5 Tera-Flops of peak computing power with 8TB of storage, 64 GB RAM, 2 Multi Core CPUs with 12 cores, 2 GPU accelerator cards NVIDIA K40 and NVIDIA P5000. [9] Approximately 12400 surveillance images consisting punching, kicking, hit with object, stabbing with knife, shooting with gun etc. human actions were considered for training from popular public data-set HMDB51 and NTU RBG+D dataset shown in figure 2. [10-11]

Initially, we made an attempt to train above models on Google Colab platform. [16], but even for as little data as 12400 images, model training time was exceeding default permitted session time that was 12 hours. Free open Google Colab platform comes with resource restrictions and it is limited for small scale ML training unless subscribed to Colab Pro which offers more resources for large scale ML training. Free version offers 1 GPU with 12 GB of RAM for 12 hours session. Further, after session expires model needs to be retrained from the beginning.

Volume 10, Issue 5
October 2021

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

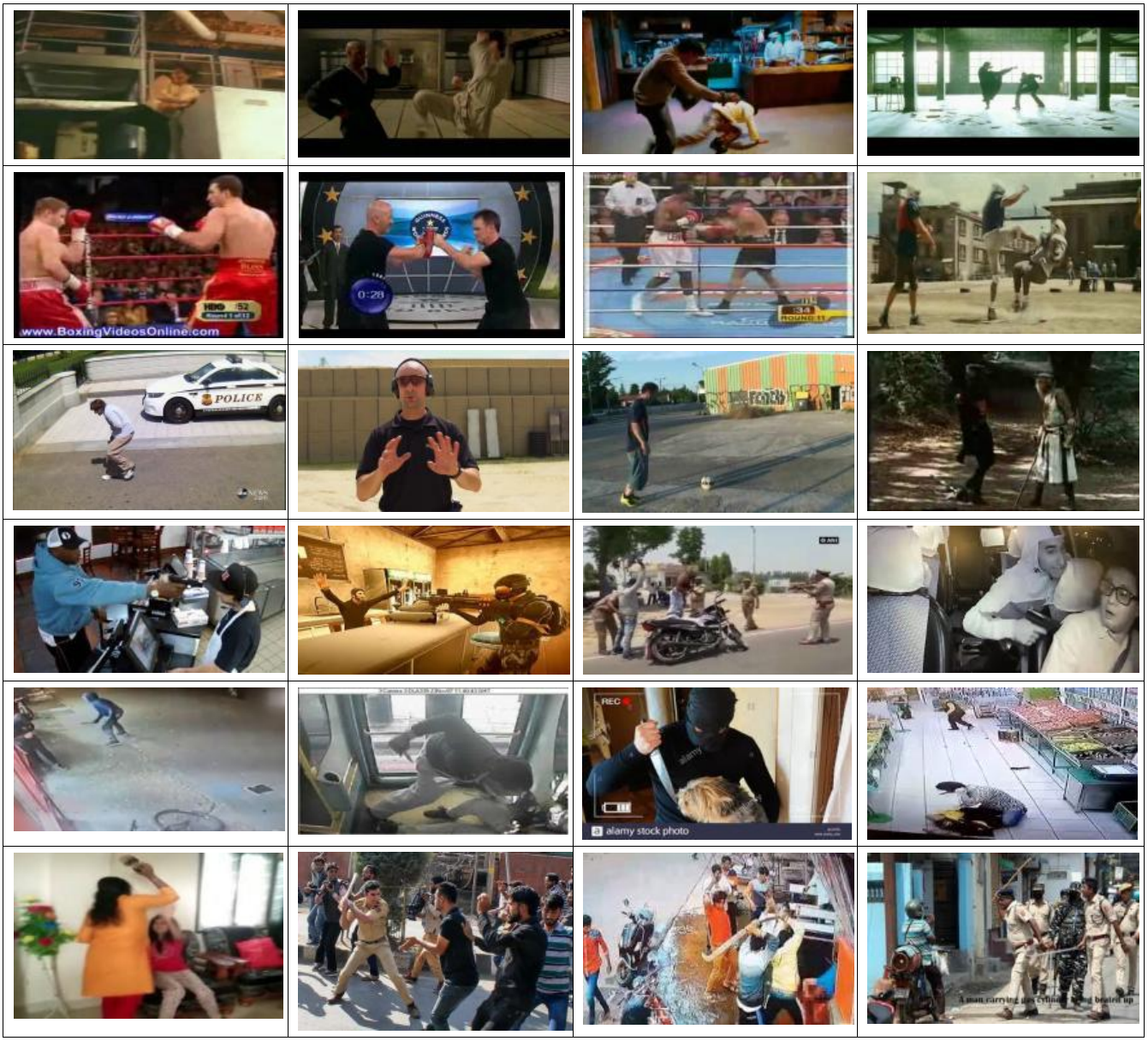

Figure 2. Sample Surveillance images consisting punching, kicking, hit with object, shooting with gun, stabbing with knife

After data collection, it was found that collected data was not uniform and had so much of noise. In order to improve performance of the deep learning models, data pre-processing is crucial inevitable step. As a part of data pre-processing, manual sampling of surveillance images were done from the total surveillance data. After collecting sample images, moving average filter was used to remove excess noise from surveillance image data. Collected images were of different resolutions, so they needed to be re-scaled to a uniform resolution. Based on observation that many of collected sample data images had 320 x 240 resolutions, in presented work, all images were re-scaled to 320 x 240 resolution. There were few very small images and ignored for training. Sampled surveillance images then were equalized for each human violent action label in order to avoid skewness (over fitting, underfitting) in deep learning models.

Once data pre-processing was completed, it was fed to various deep learning models. The presented work was carried out on Tensor flow and Keras platform. First starting with CNN, 7 layers in which 2 Con2D, 2 MaxPooling, 1 Flatten, 2 Fully Connected Layers were configured. ReLu was used as activation function as shown in figure 3 (a). Mathematical description of ReLU non-linearity can be defined as function returns 0 for negative x values as shown in equation 1. Starting stride size was configured as 224*224. Learning rate for all deep learning approaches was fixed to 1E4 (000.1). Secondly for LSTM, feature engineering was done using small VGGNet. LSTM includes forget gate to overcome problem of vanishing gradient during feedback. VGG16 network was

Volume 10, Issue 5
October 2021

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

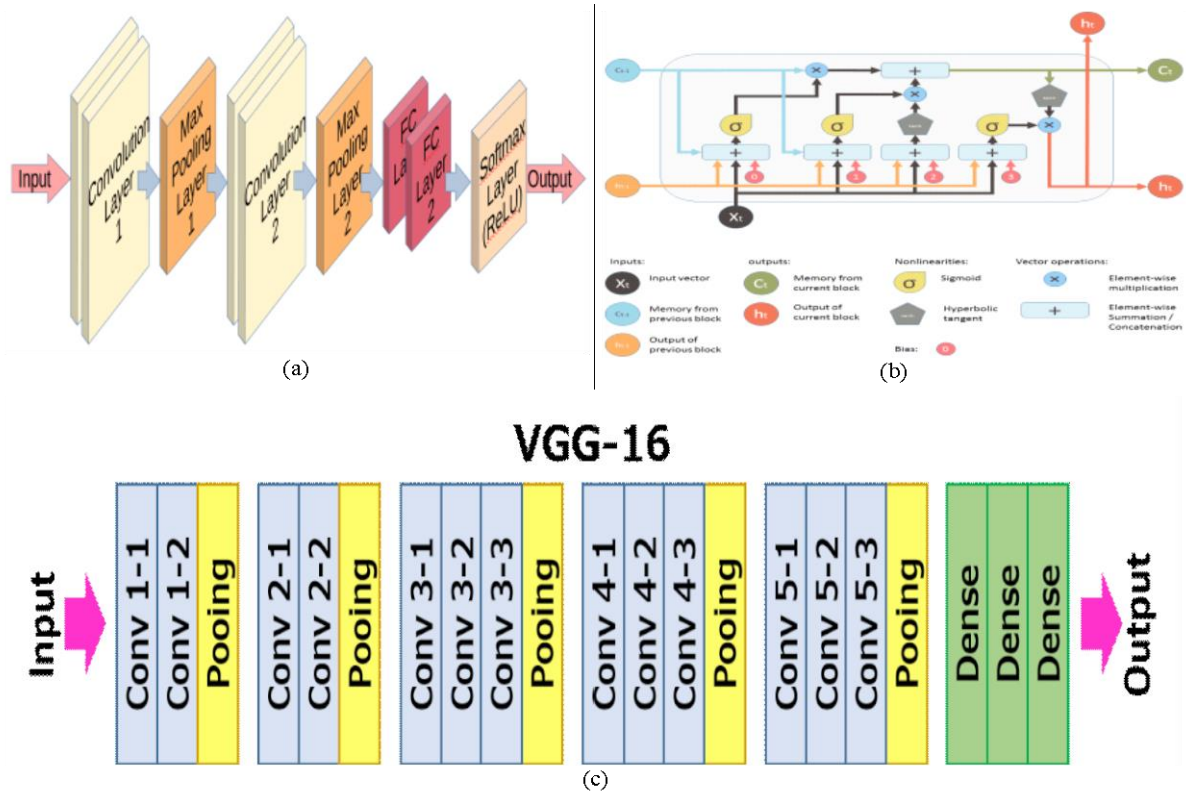

Figure 3. (a) CNN, (b) LSTM and (c) VGG16 architectures [12-13]

proposed as a part of ILSVRC2014 trained on ImageNet. [14 - 15] As shown in figure 3 (c), VGG16 is more complex deep CNN with 16 layers each layer having multiple sub layers and it also comes with multiple configured variants. Due to its nested sub layers, VGG16 enables more robust way to extract image features. VGG16 architecture comes with flexibility of data-augmentation. Due to availability of training data, data-augmentation flag was set to true in the proposed work. Three data augmentation variants zoom, shear, angle were experimented. Performance testing of all the models was carried out on unknown videos searched from Youtube consisting action key-words.

$$f(x) = max(0,x) \quad (1)$$

where, ReLU function $f(x)$ is defined on R –> R and x is output data.

4. RESULTS AND DISCUSSIONS

In order to evaluate the performance of deep learning models comparatively, precision, recall, f1-score, accuracy statistical metrics were considered. Precision refers to the ratio of correctly predicted instance to the total number of predicted instances. Precision gives an idea about how closely model classifies data to actual distribution. Recall on the other hand refers to the ratio of correctly predicted instances to the total actual correct instances. Recall indicates of actual data distribution, how close a classifier is with respected to predicted instances. Accuracy is a simple metric referring to correctly predicted instances to the total population instances. When it comes to evaluate performance of a classifier, statistically there is always a trade of when it comes to choosing precision and recall, and accuracy is good enough when datasets are balanced and equally distributed over class labels. However in reality it need not be the case as data set can be skewed. Hence, f1-score is used for overcome limitations of accuracy and trade of between precision and recall. F1-Score is harmonic mean of precision and recall which gives a weighted average among two smoothing out extremities on both metrics. Equations to compute above measures are given as below in equation 2-5.

If for a classifier with binary class label value, TP is number of correctly classified instances as positive, TN is number of correctly classified instances as negative, FP is number of incorrectly classified instances as positive and FN is number of incorrectly classified instances as negative then,

$$\text{Precision, } P = TP / (TP + FP) \quad (2)$$

$$\text{Recall, } R = TP / (TP + FN) \quad (3)$$

$$\text{Accuracy, } A = TP / (TP + FP + TN + FN) \quad (4)$$

$$\text{F1- Score, } f = 2^* (\text{precision})(\text{recall})/(\text{precision} + \text{recall}) \quad (5)$$

Volume 10, Issue 5
October 2021

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

Table 1. Classification Matrix for CNN, LSTM, VGG16

Metric	CNN			LSTM			VGG16		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Normal	0.92	0.82	0.86	0.92	0.88	0.89	0.95	0.95	0.95
Kick	1.00	0.78	0.88	0.98	0.88	0.93	0.97	0.92	0.93
Punch	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99
Hitting With Object	0.94	0.88	0.90	0.94	0.90	0.91	0.96	0.96	0.96
Shooting With Gun	0.96	0.90	0.92	0.93	0.86	0.89	0.93	0.93	0.93
Stabbing With Knife	0.93	0.86	0.89	0.98	0.92	0.95	0.91	0.91	0.91
Accuracy	**--**	**--**	**0.91**	**--**	**--**	**0.93**	**--**	**--**	**0.94**

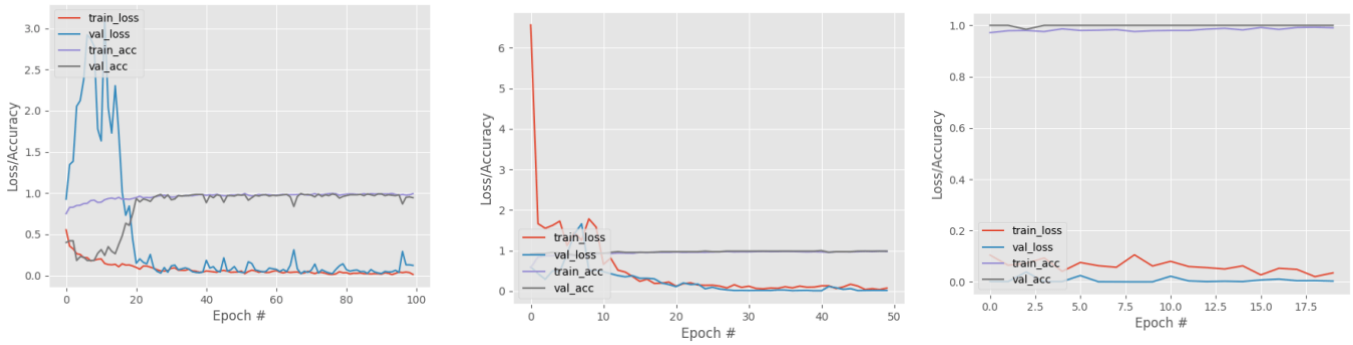

Figure (a) Figure (b) Figure (c)

Figure 4. Training Model AUC (Area Under the Curve) plots for (a) CNN (b) LSTM (c) VGG16 algorithms

Classification matrix consisting precision, recall and f1-score for CNN, LSTM and VGG16 are combined and enclosed in table 1. For each model, graphs were plotted against loss/accuracy and number of epochs.

The experimental results in Table 1 shows that VGG16 performs slightly better as compared to LSTM with respect to precision and recall values for each human action class even though having same accuracy of 94%. CNN is at the lowest among all three deep learning techniques with 91% accuracy.

As shown in figure 4 (a), initially for CNN, training loss and validation loss metric values were really high and there were sharp fluctuations observed till 20 epochs. Hence, training accuracy and validation accuracy was so low initially. As number of epochs increases, CNN model stabilizes in learning about humans actions. From figure 3, it is clear that CNN is slow and requires almost up-to 100 epochs to classify actions better. The reason for CNN's poor performance is because its one directional, hence training loss and validation loss over epochs is higher. While in case for LSTM network, training loss drastically drops in initial epochs due to its bidirectional nature and capability to retain information for longer period as visible in figure 4 (b). That's why LSTMs are ideal for sequential data such as human action recognition through surveillance videos. As compared to CNN, LSTM AUC plot has lesser fluctuations during training and accuracy steadily increases in nearly 50 epochs. In case of fine tuned VGG16 AUC curves for training loss, validation loss and respective accuracy is nearly smooth and becomes stable nearly in 20 epochs as shown in figure 4 (c).

When it comes to assigning labels to unknown test data, VGG16 performs slower due to sub-layer architecture. Hence for real time human violent action recognition, faster deep learning framework is needed. Current set up was limited by infrastructure as well as limited training data. Screenshots of recognized human actions are shown in figure below.

Volume 10, Issue 5
October 2021

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

Figure 5(a). Human Violent Action Recognition through CNN

Figure 5(b). Human Violent Action Recognition through LSTM

Volume 10, Issue 5
October 2021

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

Figure 5(c). Human Violent Action Recognition through VGG16

As shown in figure 5(a),5(b) and 5(c), for given sample test videos, CNN can wrongly classify actions. While LSTM, VGG16 correctly classify the actions. This is due to limited training data. CNN, LSTM models require more data to perform well. Temporal differences of action recognition are not visible in figure 4. But Despite correct class labels, VGG16 recognizes slowly as action appears in surveillance footage.

5. CONCLUSION

In current accomplished work with regards to human violent activity recognition, it was implemented Three available state of the art deep learning approaches convolutional neural network CNN, Long Short Term Memory network LSTM and VGG16 (OxfordNet) on publicly available data-sets. Objectively compared the results of each method against each other to determine which approach gives better results. VGG16 also give relatively better performance than LSTM and CNN but is slow in recognition due to its multi sub-layer architecture. During the work it was learned that limited low quality training data can be bottleneck and significantly affect the performance of the deep learning models. Significant amount of time was invested in collecting and pre-processing of data before it can be used for training. To find high quality training data for violent events is still a challenging and time consuming task. Current work was conducted with limited data and infrastructure which is good enough for non real time violent activity detection but not good sufficient for real time applications. As due to limited moderate quality training data CNN and LSTM approaches couldn't perform well while VGG16 have data augmented capabilities which allow them to work efficiently with less data. It was also learned that how learning rate affects the accuracy of model. Also CNN, LSTM require more epochs in order to train as compared to VGG16. It also tried to implementing all three approaches on Google Colab, but even for less data as much as 12400 images model training time went beyond permitted session time that was 12 hours. So the trial with Google Colab platform was far from success. The future scope of work aims at using ResNet50 deep neural net architecture tuned for performance which is similar to VGG16 but it overcomes limitation of sub-layer processing and can be helpful for real time recognition of violent activities.

Acknowledgments

We would like to acknowledge our gratitude towards Gujcost for enabling us to use PARAM Shavak supercomputer deployed at Government Engineering College, Rajkot, Gujarat, India in order to carry out our research work without which it would have been more challenging and more time consuming.

REFERENCES

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in International Conference on Learning Representations (ICLR), 2015.

[2] Sepp Hochreiter and Jürgen Schmidhuber, Long Short-Term memory. Neural computation, 9(8):1735–1780, 1997.

[3] Nian Chi Tay , Tee Connie et.al, A Robust Abnormal Behavior Detection Method Using Convolutional Neural Network, Springer Nature Singapore Pte Ltd., 1-11, 2019.

[4] Ullah, W., Ullah, A., Haq, I.U. et al. CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks. Multimed Tools Appl (2020). https://doi.org/10.1007/s11042-020-09406-3.

[5] S. Jothi Shri, S. Jothilakshmi, Anomaly Detection in Video Events using Deep Learning, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-9 July, 1-4, 2019.

[6] Mohamed Mostafa Soliman, Mohamed Hussein Kamal et.al, Violence Recognition from Videos using Deep Learning Techniques, International Conference on Intelligent Computing and Information Systems (ICICIS) , IEEE, 1-6, 2019.

[7] Giorgio Molares, Itamar Salazar-Reque et.al, Detecting Violent Robberies in CCTV Videos Using Deep Learning, International Federation for Information Processing (IFIP), Springer Nature, 1-10, 2019.

[8] Unnati Koppikar, C. Sujatha et.al, Real-World Anomaly Detection Using Deep Learning, Intelligent Computing and Communication, Advances in Intelligent Systems and Computing 1034, Springer Nature Singapore Pvt Ltd., 1-10, 2020.

[9] http://www.gecrj.cteguj.in/labs

[10] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, Gang Wang, "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[11] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A Large Video Database for Human Motion Recognition. ICCV, 2011.

[12] VGG16 – Convolutional Network for Classification and Detection.

[13] Understanding LSTM and its diagrams, https://medium.com/mlreview/understanding-lstm-and-its-diagrams-37e2f46f1714

[14] Large Scale Visual Recognition Challenge 2014, http://www.image-net.org/challenges/LSVRC/2014/results

[15] ImageNet Dataset, http://www.image-net.org/

[16]GoogleCollab, https://colab.research.google.com/notebooks/intro.ipynb

[17] Amira Ben Mabrouk, Ezzeddine Zagrouba, Abnormal behavior recognition for intelligent video surveillance systems: a review, Expert Systems With Applications (2017), doi: 10.1016/j.eswa.2017.09.029.

[18] Kelathodi Kumaran Santhosh, Debi Prosad Dogra, Partha Pratim Roy, Anomaly Detection in Road Traffic Using Visual Surveillance: A Survey, Preprint, 2019.

AUTHOR PROFILES

Devang Jani has completed B.E. in EC from Shantilal Shah Engineering College, Bhavnagar, Bhavnagar Univesity in 2008. He has Completed Master of Engineering in EC from Shantilal Shah Engineering College, Bhavnagar, Gujarat Technological University, Gujarat in 2016. He is pursuing Ph.D. in EC from U. V. Patel College of Engineering, Ganpat University, Gujarat. Currently, he is working as an Assistant Professor in EC Department at Shantilal Shah Engineering College,Bhavnagar Gujarat. He has more than 12 years of teaching experience. He has published more than 5 papers in various journals and conferences. His research interests include Image & Video Processing, Machine Learning & Deep Learning.

Dr. Anand Mankodia received B.E. (Electronics) degree in 2000 from Dharmsinh Desai Institute of Technology (DDIT), Nadiad, Master of Engineering (M.E.) in 2009 from Gujarat University and Ph.D. degree in 2016 from Ganpat University, Gujarat. He is currently working as an Associate Professor at EC Department, U. V. Patel College of Engineering, Ganpat University, Mehsana, Gujarat. His research interests include Image & Video Processing, Embedded Systems, Machine Learning and Deep Learning. He has published more than 20 research articles in peer-reviewed International journals and renowned International conferences. He has participated in more than 30 different FDP, Seminar & Workshops. He has also organized more than 10 seminars, workshops and symposiums at the institute sponsored by various funding agencies like CARS, GUJCOST, GEDA, DST, etc. He has supervised more than 100 UG & 10 PG students during their project work. Currently 6 Ph.D. Scholars registered under him at Ganpat University. He has more than 20 years of teaching experience.