International Journal of Safety and Security Engineering
Vol. 12, No. 5, October, 2022, pp. 615-622
Journal homepage: http://iieta.org/journals/ijsse

Video Violence Detection Using LSTM and Transformer Networks Through Grid Search-Based Hyperparameters Optimization

Moch Arief Soeleman*, Catur Supriyanto, Dwi Puji Prabowo, Pulung Nurtantio Andono

Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Indonesia

Corresponding Author Email: arief22208@gmail.com

https://doi.org/10.18280/ijsse.120510

Received: 5 September 2022
Accepted: 19 October 2022

Keywords:
convolution neural networks, deep learning, long short-term memory (LSTM), transformer, video violence detection

ABSTRACT

The security system in public places can be improved by automatically detecting violence. Deep learning has recently gained popularity as a solution to classification problems, which improves the effectiveness of violent video detection. The authors extracted the features using a pretrained network, such as InceptionV3. To maximize the performance for violent video detection, the Grid Search approach was adopted to search for the optimal hyperparameter. The main goal is to evaluate how well LSTM and Transformer networks classify videos. The results show competitive performances in identifying violent videos, with the state-of-the-art methods. On the Hockey, Crowd, and AIRTLab datasets, LSTM outperformed Transformer with AUC scores of up to 0.976, 0.934, and 0.86, respectively.

1. INTRODUCTION

There may be crime or violence in public areas. Relying on human monitoring is difficult. Surveillance cameras can be used to keep an eye on what happens in public areas, especially if we want to detect violent activities automatically. The ability to automatically detect violence makes it simpler for the security forces to respond right away with help, and the recording can also be used as evidence in court.

Over the past ten years, research on the identification of violent video has increased. Their suggested techniques are based on learned and handcrafted features [1, 2]. The handcrafted features are carefully engineered by scientists. Examples of hand-crafted features include edge features and histogram features. Convolutional neural networks (CNNs) provide the foundation for learned features. The convolution layer of CNN is automatically used to extract the learned features.

By hand-crafting a feature, Lohithashva et al. [3] have created a method for detecting violence in videos. To identify violent incidents in a movie, the system combined Local Binary Pattern (LBP) and GLCM (Gray Level Co-occurrence Matrix) as feature extraction techniques. In their study, the system was evaluated using a variety of classifiers, including Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Discriminant Analysis (DA), and Logistic Regression (LR).

Violent Flows (ViF) [4], Oriented ViF (OViF) [5], Motion Weber Local Descriptor (MoWLD) [6], and Histogram of Optical flow Magnitude and Orientation (HOMO) [7] are only a few of the methods for video violence detection that are based on optical flow orientation and magnitude. Hassner et al. [4] proposed ViF by exploiting the magnitude of pixel flow and classified the video clip using a linear Support Vector Machine (SVM). To overcome the limitation of ViF, Zhou et al. [8] put forward OViF, and implemented it on non-crowded violence video clips. The technique relies on motion magnitude and motion orientation of a pixel in two consecutive frames. On violent videos with plenty of people, however, OViF was not successful [8].

Zhang et al. [6] expanded Weber Local Descriptor into MoWLD (WLD). To create the MoWLD descriptor, MoWLD combined the optical flow and WLD histograms. Kernel Density Estimation is then used to remove the extracted MoWLD features (KDE). SVM evaluations of the proposed MoWLD were conducted using the Hockey, Crowd, and BEHAVE datasets. Additionally, Mahmoodi and Salajeghe [7] presented the HOMO optical flow-based feature descriptor. The optical flow between each frame was calculated and each frame was transformed into a grayscale image. To extract the feature's histogram and send it to the SVM classifier, HOMO evaluated the change's magnitude and direction. The performance of optical flow-based algorithms in terms of violence detection is still inferior to deep learning-based systems.

CNNs, a deep learning-based technique, have emerged as an additional option for video violence identification. Ullah et al. [1] suggested utilizing 3D CNN to extract spatiotemporal features. For the Hockey dataset and the Crowd dataset, they obtained 3D CNN accuracy of 96% and 98%, respectively. Asad et al. [9] proposed to combine the spatial information of each frame and send them to long short-term memory (LSTM). The features are extracted using a pretrained network called VGG16. Their study had an accuracy rate of 98.8% for the Hockey dataset and 97.1% for the Crowd dataset. LSTM networks for classification were also proposed by Shoaib and Sayed [10]. ResNet 101 was adopted to extract the features of each frame. Region of Interest (ROI) was used to localized the human body and detects the key points. On the Weizman, KTH, and Custom datasets, respectively, the results on three datasets demonstrate that their proposed method obtained 77.4%, 95.7%, and 88.2% accuracies. In another video surveillance system, LSTM also works for anomaly detection in crowd situations [11]. Compared to optical flow-based

method, CNN evolved to have better accuracy for violence video detection. However, the network requires a complex structure to mine the learned deep features.

Transformer has recently emerged as a leading network for video violence identification. Vaswani et al. [12] proposed the Transformer network, which tries to eliminate the convolutional and recurrent parts of CNN's sequence model. Transformer network for Violence Detection has been applied by Abdali [13]. On the Real-life Violence dataset (RLVS), the proposed transformer received a score of 96.25% accuracy. It is crucial to understand how well these CNN architectures perform in detecting violence in videos because LSTM and Transformer are the two newest trends in CNN-based violence detection. However, no research has been done comparing the effectiveness of the Transformer and LSTM structures for video violence identification.

This study seeks to assess and compare LSTM and Transformer networks for violence video recognition through in-depth analysis. Each video clip's features are extracted using the InceptionV3 network and supplied into the proposed LSTM and Transformer networks as input. This study uses Grid Search to find the optimum hyperparameter for LSTM and Transformer networks in order to deliver the greatest performance.

The remainder of the paper is structured as follows: Some related works are included in Section 2. The experimental design is presented in depth in Section 3 while the results are discussed in Section 4. The last Section wraps up the project and offers ideas for additional development.

2. LITERATURE REVIEW

2.1 Convolution neural networks

CNNs have long been utilized to detect violent acts in videos. Multiple layers are present in CNNs for feature extraction and classification. Despite having a sophisticated architecture, CNNs are more accurate than conventional machine learning algorithms. CNNs use a variety of combination processes, including pretrained networks, data augmentation, and call-back on training phase, to achieve high accuracy. Pretrained networks are used in this study to retrieve each frame's features. Pretrained network is an architecture of CNNs which have been trained on ImageNet database, which contains a million images on 1,000 object classes.

Pretrained CNNs include ResNet50, InceptionV3, Xception, VGG16, and VGG19, among others. InceptionV3 outperformed ResNet50 and Xception in a comparison study by Xiao et al. [14] on the accuracy of breast cancer identification. When compared to VGG16, VGG19, Xception, and ResNet50, InceptionV3 produced the greatest results, according to our earlier research [15] on violence detection. Table 1 lists some related studies on the identification of violence in videos using deep learning. The earlier publications validated their suggested methodologies using several well-known publicly available datasets. Even though LSTM is the most often used deep learning architecture for violence detection, other architectures, such as the Transformer architecture, should be compared.

2.2 Transformer

Three modules make up a transformer network: patch embedding, encoder, and multi-layer perceptron (MLP) [16]. In patch embedding, there are reshape and 2D convolution. Normalization and fully connected layers constitute MLP. Transformer is capable of generalizing the model and has produced positive results on the ImageNet dataset.

2.3 Grid search

Grid Search is a methodical or systematic approach to obtain the best hyperparameter [17]. The initialized hyperparameter is searched for in every conceivable combination through grid search. Other hyperparameter optimization techniques include evolutionary algorithms and Bayesian optimization. The evaluation of these hyperparameter optimization methods on neural networks revealed that the genetic algorithm outperformed Grid Search, Bayesian optimization, and other techniques [17]. Grid Search is the most straightforward method to construct, despite the fact that it is ineffective for big parameters [18]. Grid Search has been successful in locating the appropriate hyperparameters for machine learning algorithms in addition to deep learning [19].

Table 1. Related works

Reference	Proposed method	Year of publication	Dataset name	Results
Zhou et al. [2]	FightNet	2017	Hockey, Movie, Violent Intercation Dataset (VID)	Fusing RGB, optical flow, and acceleration improve the performance of violence detection.
Sernani et al. [20]	C3D-SVM, LSTM	2021	Hockey, Crowd Violence, AIRTLab	3D CNNs perform better than 2D CNNs.
Jain et al. [21]	Inception-Resnet-V2	2020	Hockey, Movie, Real-life violence	Applying Dynamic Image on violence detection leads to better results.
Samuel R. et al. [22]	Bidirectional LSTM	2019	Violent Intercation Dataset (VID), football stadium	Each frame extracts the features from violence model, human part model, and negative model, with a violence detection rate of 94.5%.
Sudhakaran and Lanz [23]	LSTM, ConvLSTM	2017	Hockey, Movie, Crowd Violence	ConvLSTM is better than LSTM with fewer parameters, and does exceptionally well in preventing overfitting.

3. EXPERIMENTAL DESIGN

3.1 Datasets

This study uses the following three benchmark violence video datasets:

- The hockey [24] dataset includes 500 violent videos and 500 nonviolent ones. Each video clip consists of 50 frames with a 360 x 288 pixel resolution.
- 123 violent and 123 non-violent video clips are included in the violent crowd [4] dataset. Each clip comprises between 50 and 150 frames, each with a resolution of 320 x 240 pixels. The sample frame for the datasets is displayed in Figure 1.
- The two main directories in AIRTLab [25] are violent and non-violent. Each main directory contains the subdirectories cam1 and cam2. The video clips in cam2 were recorded with a different camera and point of view than those in cam1, which is the difference between the two cameras. 115 violent/cam1 video clips and 60 non-violent/cam1 video clips are used in this investigation.

3.2 Environment and setup

The proposed method for detecting video violence is depicted in Figure 2. First, we divide the video clip's frames and use InceptionV3 to extract the deep features. This study used Grid Search to find the optimal LSTM or Transformer network hyperparameters using all of the video clips. Batch size, epoch, optimizer, dropout, and learning rate are a few examples of hyperparameters. The setting for the hyperparameter is displayed in Table 2. Since the number of initialized hyperparameter values expands along with the computing of Grid Search, we tracked the value of a few hyperparameters based on prior research. In certain investigations, the dropout probability was chosen at between 0.2 and 0.5 [20, 26-29]. The learning rates are referred to Ullah's work [1], which is set at 0.0001 and 0.00001. The batch size is also determined in reference to Ullah's work [1]. We chose Adam as the optimization function in the fully linked layer, and we set the number of epochs to 100 [9]. We retrained the LSTM or Transformer using the retrieved hyperparameter. This study uses 5-fold cross-validation (CV) on Grid Search and Retrained steps.

The layers of the Transformer and LSTM are displayed in Tables 3 and 4. Table 3 begins with the LSTM layer, which was used as a feature extractor by InceptionV3. Then, using the rectified linear unit (ReLU) activation function, we constructed a fully linked layer with 1024 neurons. Implementation of dropout occurred at a rate of 0.20. The dropout layer and a further fully linked layer with sigmoid activation function were then added. As a final classification step, a fully connected sigmoid activation function was used.

The first layer in the transformer model depicted in Table 4 is frame position embedding. The input shape consists of frame, color channel, height, and width. The second layer is a transformer encoder that has a Gelu activation function and multi-head attention. We also included a layer for normalization.

These investigations were carried out with the Python programming language. The computer runs Ubuntu 18.04 and makes use of a GeForce RTX 3070 8GB GPU. TensorFlow 2.6.2 and Keras 2.6.0 were employed by our CNN. We set the seed for the Numpy and TensorFlow libraries in order to obtain consistent results across all executions.

(a) Hockey dataset

(b) Crowd dataset

(c) AIRTLab dataset

Figure 1. Samples video clips of each dataset. First row and second row show the violence and non-violence, respectively

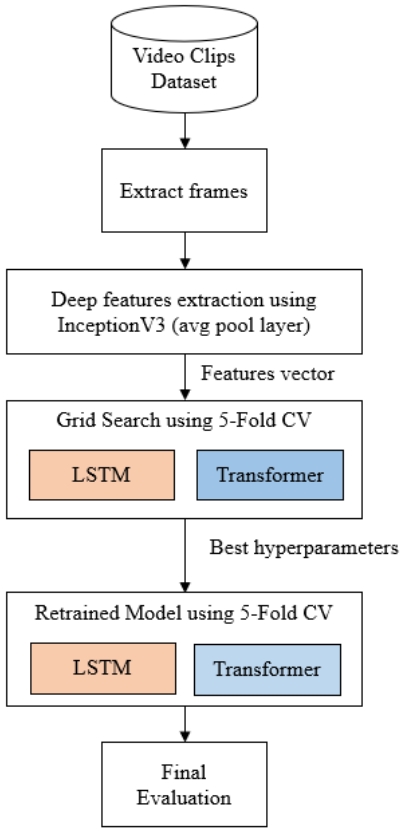


Figure 2. Flow of video violence classification

3.3 Experimental evaluation

In this study, measures including accuracy, standard deviation (SD), and area under the curve (AUC) were utilized to assess the suggested method against state-of-the-art references.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where, TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

Table 2. Hyperparameter settings

Hyperparameter	Value
Batch size	20, 100, 200
Dropout rate	0.2, 0.5
Learning rate	0.0001, 0.00001

Table 3. Layer structures of LSTM

Layer	Architecture	Output shape	Param #
LSTM	-	(None, 512)	
Dense	Relu	(None, 1024)	
Dropout	-	(None, 1024)	
Dense	Sigmoid	(None, 50)	
Dropout	-	(None, 50)	
Dense	Softmax	(None, 2)	

Table 4. Layer structures of transformer

Layer	Architecture	Output shape	Param #
Frame position embedding	-	(None, None, 2048)	40960
Transformer encoder	-	(None, None, 2048)	16812036
Global max pooling1D	-	(None, 2048)	0
Dropout	-	(None, 2048)	0
Dense	Softmax	(None, 2)	4098

4. RESULTS AND ANALYSIS

The effectiveness of LSTM and Transformer on three video violence datasets is covered in this section. Accuracy and AUC are used to compare performance. Using Grid Search hyperparameter tuning, we assess the performance in the first comparison. Table 5 displays the best hyperparameter values obtained through Grid Search. On the Hocket dataset, the batch size value for both LSTM and Transformer is 200. The batch sizes for LSTM and Transformer on the Crowd and AIRTLab datasets are 20 and 100, respectively. Grid Search discovered that the optimal dropout rate is 0.5 and the best learning rate is 0.0001 mostly across the three datasets. The best accuracy was attained by LSTM and Transformer during hyperparameter tuning using Grid Search, as shown in Table 6. On the three datasets, LSTM surpasses Transformer in terms of accuracy, scoring 95%, 90.26%, and 80.57% for Hockey, Crowd, and AIRTLab, respectively. The two model's LSTM and Transformer are then retrained with the best hyperparameters for each dataset. With the aid of cross validation, the models are assessed. For AUC and accuracy, the findings are shown in Tables 7 and 8, respectively, along with the results for each fold. The mean AUC and accuracy for 5-fold cross validation are also displayed in the tables. The model based on LSTM outperformed Transformer in terms of AUC, with mean AUC values on the three datasets of 0.976, 0.934, and 0.86, respectively. On the Hockey, Crowd, and AIRTLab datasets, the mean accuracies of LSTM achieve classification results (94.6%, 89.86%, and 80.57%) better than Transformer.

The accuracy curves during the training procedure are shown in Figure 3. As seen in Table 7, the curves are derived from the optimal fold. The graphs show that the LSTM testing results outperform Transformer network, particularly for the Hockey and Crowd datasets. Transformer's testing results appear erratic over a period of 100 epochs.

We plot the receiver operating characteristic (ROC) curve, as seen in Figure 4, to assess the proposed model's performance graphically. The ROC curves for LSTM and Transformer on the three datasets are derived from the best AUC in Table 6. ROC curves for Hockey (top row), Crowd (second row), and AIRTLab (third row) are extracted from Fold 5, Fold 3, and Fold 2, respectively.

Table 5. Best hyperparameters from grid search

		Batch size	Dropout rate	Learning rate
Hockey	LSTM	200	0.2	0.0001
	Transformer	200	0.5	0.0001
Crowd	LSTM	20	0.5	0.0001
	Transformer	100	0.5	0.0001
AIRTLab	LSTM	20	0.2	0.0001
	Transformer	100	0.5	0.00001

Table 6. Best score (accuracy in %) of Grid Search. The best accuracies are highlighted in bold on each dataset

	LSTM	Transformer
Hockey	**95.00**	83.80
Crowd	**90.26**	82.11
AIRTLab	**80.57**	61.71

Table 7. AUC of each model on each dataset, for each fold of cross validation. The best accuracies are highlighted in bold on each model (row)

		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean
Hockey	LSTM	0.98	0.95	0.98	0.98	**0.99**	0.976
	Transformer	0.92	0.94	0.96	0.97	**0.98**	0.954
Crowd	LSTM	0.95	0.92	0.9	0.93	**0.97**	0.934
	Transformer	0.94	0.89	0.88	0.89	**0.98**	0.916
AIRTLab	LSTM	0.85	0.87	**0.88**	0.85	0.85	0.86
	Transformer	0.85	0.77	**0.95**	0.88	0.5	0.79

Table 8. Accuracy (in %) of each model on each dataset, for each fold of cross validation. The best accuracies are highlighted in bold on each model (row)

		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean
Hockey	LSTM	94.5	91	95.5	**97**	95	94.6
	Transformer	88.5	87.5	89	91	**94**	90
Crowd	LSTM	86	87.76	89.8	87.76	**97.96**	89.86
	Transformer	**88**	61.22	73.47	85.71	83.67	78.41
AIRTLab	LSTM	**82.86**	80	77.14	**82.86**	80	80.57
	Transformer	28.57	57.14	65.71	**68.57**	25.71	49.14

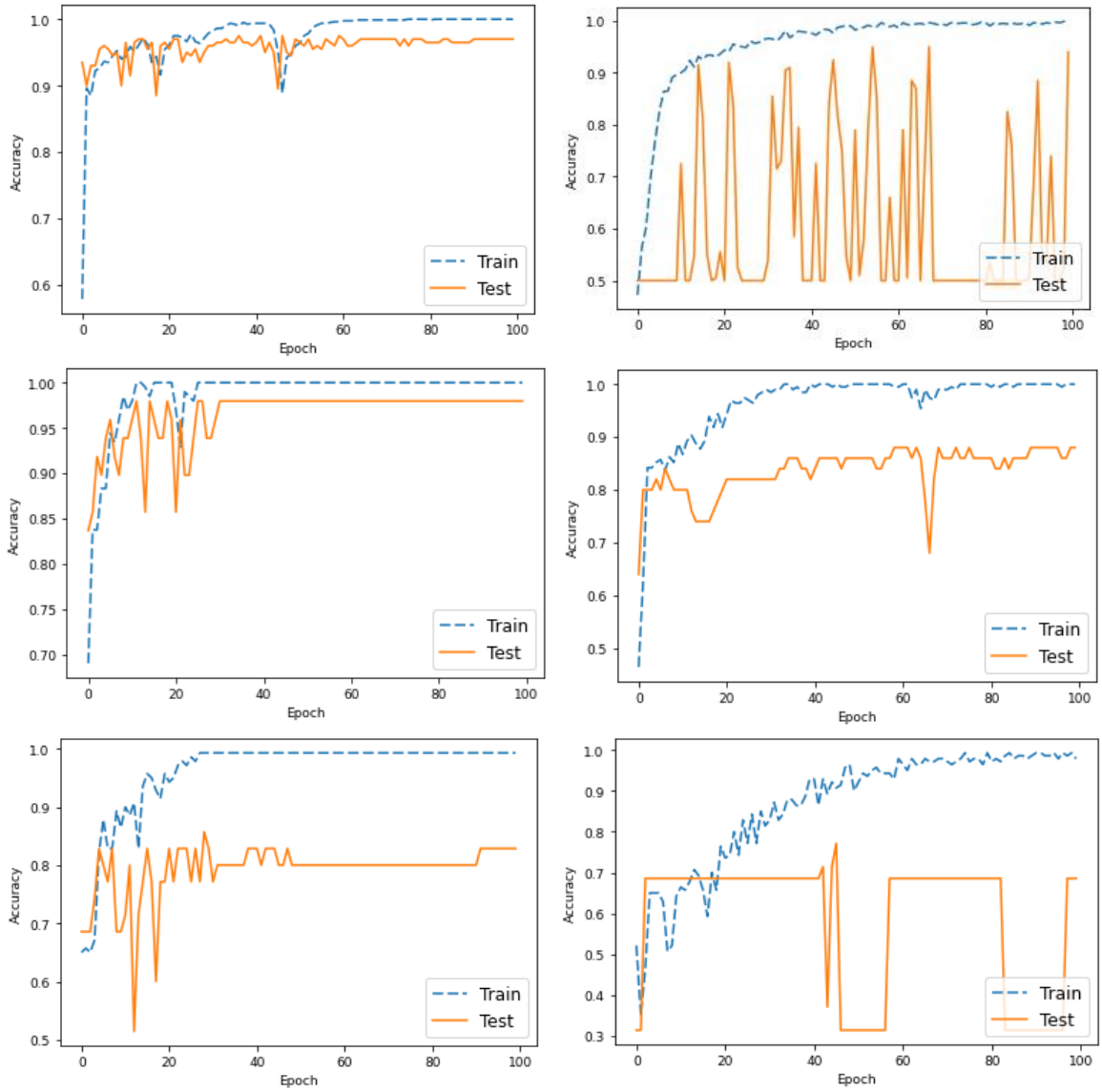
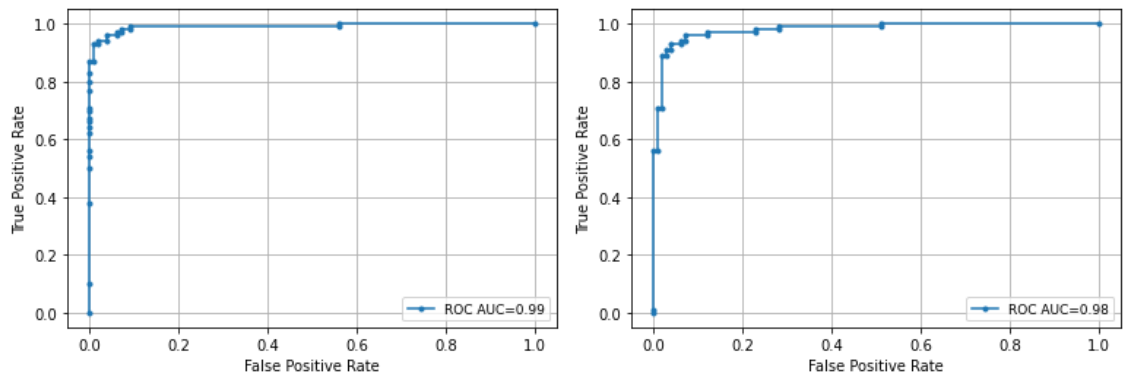


Figure 3. Accuracy curves between training and testing during 100-epochs on Hockey (first row), Crowd (second row), and AIRTLab (third row) datasets. The left and right sides are for the LSTM and Transformer models



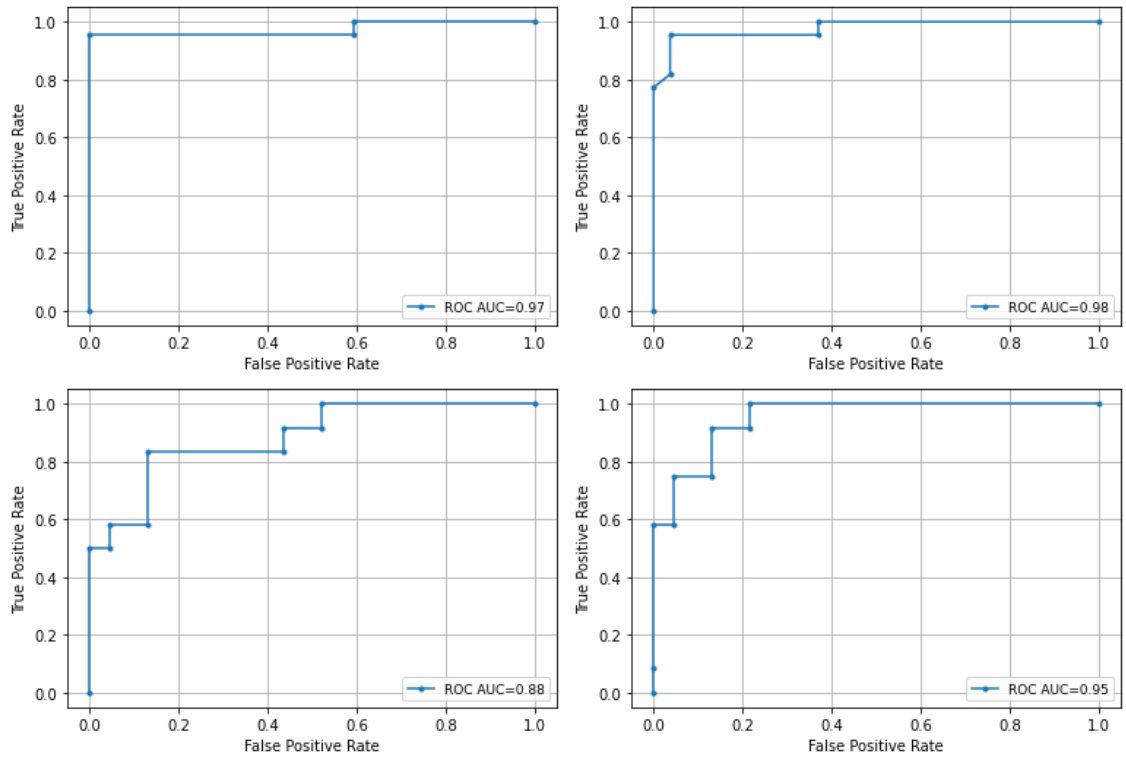


Figure 4. AUC and ROC Curve on Hockey (first row), Crowd (second row), and AIRTLab (third row) datasets. The left and right sides are for the LSTM and Transformer models

Table 9. State of the art comparison on three datasets

Dataset	Related Works	AUC
Hockey	ViF [4]	0.8801
	OViF [5]	0.9193
	DiMOLIF [29]	0.9323
	LBP+GLCM [3]	0.9360
	HOMO [7]	0.9518
	3D CNN [1]	0.970
	MoWLD [6]	0.9758
	LHOG+LHOF [8]	0.9798
	C3D+SVM [20]	0.9962
	C3D+FC [20]	0.9927
	ConvLSTM [20]	0.9931
	Ours (Transformer)	**0.954**
	Ours (LSTM)	**0.976**
Crowd	HOMO [7]	0.8284
	ViF [4]	0.8804
	DiMOLIF [29]	0.8925
	OViF [5]	0.9182
	LBP+GLCM [3]	0.93
	MoWLD [6]	0.9408
	ConvLSTM [20]	0.9443
	LHOG+LHOF [8]	0.9703
	3D CNN [1]	0.98
	C3D+FC [20]	0.9994
	C3D+SVM [20]	1
	Ours (Transformer)	**0.916**
	Ours (LSTM)	**0.934**
AIRTLab	C3D+SVM [20]	0.993
	C3D+FC [20]	0.9894
	ConvLSTM [20]	0.9967
	Ours (Transformer)	**0.79**
	Ours (LSTM)	**0.86**

To exemplify the effectiveness of the suggested models, Table 9 provides examples of numerous cutting-edge video violence detection techniques. With regard to the hockey and crowd datasets, our LSTM model outperforms earlier methods like ViF [4], OViF [5], DiMOLIF [29], and HOMO [7] by a wide margin in terms of AUC. Meanwhile, our LSTM model only outperforms on the Hockey dataset when compared to MoWLD [6] and 3D CNN [1]. MoWLD outperforms our LSTM, but only marginally. The top models in Hockey, Crowd, and AIRTLab are still C3D [20] and ConvLSTM [20] according to the table.

5. CONCLUSIONS

This paper compares the LSTM and Transformer models for video violence detection. Each video clip's features are extracted using the InceptionV3 pretrained network. The LSTM and Transformer are then provided the features to handle the spatiotemporal features. Next, Grid Search was utilized to discover the ideal LSTM and Transformer hyperparameters, including Batch size, dropout rate, and learning rate. Based on the optimal hyperparameter values, the LSTM and Transformer were retrained and evaluated in this work. The results of the experiments showed that LSTM remains superior to Transformer. Our suggested models perform comparably on three publicly available video violence datasets, Hockey, Crowd, and AIRTLab. One of the potential future possibilities for violence detection in videos can be seen as a result of this paper's evaluation of two deep learning architectures. The best design enables developers to create applications that monitor human activity in public spaces and guarantee the comfort and safety of the general public. The poor performance of violence detection in the AIRTLab dataset is one of the study's limitations. We intend to assess channel-separated networks (CSNs) on violence detection for further research. Different convolutional blocks in the network

can be used to extract spatial and spatiotemporal characteristics. We want to test the suggested approach on other datasets of violent videos.

ACKNOWLEDGMENT

This work was supported by DRPM-DIKTI Number: 312/E4.1/AK.04.PT/2021. The scheme of applied research for two years 2021-2022.

REFERENCES

[1] Ullah, F.U., Ullah, A., Muhammad, K., Haq, I.U., Baik, S.W. (2019). Violence detection using spatiotemporal features with 3D convolutional neural network. Sensors, 19(11): 1-15. https://doi.org/10.3390/s19112472

[2] Zhou, P., Ding, Q., Luo, H., Hou, X. (2017). Violent interaction detection in video based on deep learning. Journal of Physics: Conf. Series, 844: 012044. https://doi.org/10.1088/1742-6596/844/1/012044

[3] Lohithashva, B., Aradhya, V.M., Guru, D. (2020). Violent video event detection based on integrated LBP and GLCM texture features. Revue d'Intelligence Artificielle, 34(2): 179-187. https://doi.org/10.18280/ria.340208

[4] Hassner, T., Itcher, Y., Kliper-Gross, O. (2012). Violent flows: Real-time detection of violent crowd behavior. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1-6. https://doi.org/10.1109/CVPRW.2012.6239348

[5] Gao, Y., Liu, H., Sun, X., Wang, C., Liu, Y. (2016). Violence detection using Oriented VIolent Flows. Image and Vision Computing, 48-49: 37-41. https://doi.org/10.1016/j.imavis.2016.01.006

[6] Zhang, T., Jia, W., Yang, B., Yang, J., He, X., Zheng, Z. (2017). MoWLD: A robust motion image descriptor for violence detection. Multimed Tools Appl, 76: 1419-1438. https://doi.org/10.1007/s11042-015-3133-0

[7] Mahmoodi, J., Salajeghe, A. (2019). A classification method based on optical flow for violence detection. Expert Systems with Applications, 127: 121-127. https://doi.org/10.1016/j.eswa.2019.02.032

[8] Zhou, P., Ding, Q., Luo, H., Hou, X. (2018). Violence detection in surveillance video using low-level features. PLoS One, 13(10). https://doi.org/10.1371/journal.pone.0203668

[9] Asad, M., Yang, J., He, J., Shamsolmoali, P., He, X. (2021). Multi-frame feature-fusion-based model for violence detection. The Visual Computer, 37: 1415-1431. https://doi.org/10.1007/s00371-020-01878-6

[10] Shoaib, M., Sayed, N. (2021). A deep learning based system for the detection of human violence in video data. Traitement du Signal, 38(6): 1623-1635. https://doi.org/10.18280/ts.380606

[11] Horii, H. (2020). Crowd behaviour recognition system for evacuation support by using machine learning. International Journal of Safety and Security Engineering, 10(2): 243-246. https://doi.org/10.18280/ijsse.100211

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2017). Attention is all you need. Neural Information Processing Systems (NIPS). https://doi.org/10.48550/arXiv.1706.03762

[13] Abdali, A.R. (2021). Data efficient video transformer for violence. IEEE International Conference on Communication, Networks and Satellite (Comnetsat). https://doi.org/10.1109/COMNETSAT53002.2021.9530829

[14] Xiao, T., Liu, L., Li, K., Qin, W., Yu, S., Li, Z. (2018). Comparison of transferred deep neural networks in ultrasonic breast masses discrimination. BioMed Research International, 2018: 1-9. https://doi.org/10.1155/2018/4605191

[15] Soeleman, M.A., Supriyanto, C., Prabowo, D.P. (2021). An empirical study of CNN-LSTM on class imbalance datasets for violence video detection. The 2021 International Conference on Computer, Control, Informatics and Its Applications, pp. 81-85. https://doi.org/10.1145/3489088.3489126

[16] Li, S., Wu, C., Xiong, N. (2022). Hybrid architecture based on CNN and transformer for strip steel surface defect classification. Electronics, 11(8): 1200. https://doi.org/10.3390/electronics11081200

[17] Alibrahim, H., Ludwig, S.A. (2021). Hyperparameter optimization: Comparing genetic algorithm against grid search and Bayesian optimization. In IEEE Congress on Evolutionary Computation (CEC), pp. 1551-1559. https://doi.org/10.1109/CEC45853.2021.9504761

[18] Dufour, J.M., Neves, J. (2019). Chapter 1 - Finite-sample inference and nonstandard asymptotics with Monte Carlo tests and R. Handbook of Statistics, 41: 3-31. https://doi.org/10.1016/bs.host.2019.05.001

[19] Marco, R., Ahmad, S.S., Ahmad, S. (2021). Empirical analysis of software effort preprocessing techniques. The International Journal of Intelligent Engineering and Systems, 14(6): 554-567.

[20] Sernani, P., Falcionelli, N., Tomassini, S., Contardo, P., Dragoni, A.F. (2021). Deep learning for automatic violence detection: Tests on the AIRTLab dataset. IEEE Access, 9: 160580-160595. https://doi.org/10.1109/ACCESS.2021.3131315

[21] Jain, A., Vishwakarma, D.K.. (2020). Deep NeuralNet for violence detection using motion features from dynamic images. The Third International Conference on Smart Systems and Inventive Technology. https://doi.org/10.1109/ICSSIT48917.2020.9214153

[22] Samuel R., D.J., Fenil, E., Manogaran G., Vivekananda, G., Thanjaivadivel, M., Jeeva, S. Ahilan, A. (2019). Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM. Computer Networks, 151: 191-200. https://doi.org/10.1016/j.comnet.2019.01.028

[23] Sudhakaran, S., Lanz, O. (2017). Learning to detect violent videos using convolutional long short-term memory. 14th IEEE International Conference on Advanced Video and Signal Based Surveillance. https://doi.org/10.1109/AVSS.2017.8078468

[24] Nievas, E.B., Suarez, O.D., García, G.B., Sukthankar, R. (2011). Violence detection in video using computer vision techniques. In: Real, P., Diaz-Pernil, D., Molina-Abril, H., Berciano, A., Kropatsch, W. (eds) Computer Analysis of Images and Patterns. CAIP 2011. Lecture Notes in Computer Science, vol 6855. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-23678-5_39

[25] Bianculli, M., Falcionelli, N., Sernani, P., Tomassini, S.,

Contardo, P., Lombardi, M., Dragoni, A.F. (2020). A dataset for automatic violence detection in videos. Data in Brief, 33: 106587. https://doi.org/10.1016/j.dib.2020.106587

[26] Mensa, E., Colla, D., Dalmasso, M., Giustini, M., Mamo, C., Pitidis A., Radicioni, D.P. (2020). Violence detection explanation via semantic roles embeddings. BMC Medical Informatics and Decision Making, 20: 263. https://doi.org/10.1186/s12911-020-01237-4

[27] Baba, M., Gui, V., Cernazanu, C., Pescaru, D. (2019). A sensor network approach for violence detection in smart cities using deep learning. Sensors, 19(7): 1676. https://doi.org/10.3390/s19071676

[28] Sumon, S.A, Goni, R., Hashem, N.B, Shahria, T., Rahman, R.M. (2020). Violence detection by pretrained modules with different deep learning approaches. Vietnam Journal of Computer Science: 7(1): 19-40. https://doi.org/10.1142/S2196888820500013

[29] Mabrouk, A.B., Zagrouba, E. (2017). Spatio-temporal feature using optical flow based distribution for violence detection. Pattern Recognition Letters, 92: 62-67. https://doi.org/10.1016/j.patrec.2017.04.015