Bidirectional Convolutional LSTM for the Detection of Violence in Videos

Alex Hanson(✉) , Koutilya PNVR , Sanjukta Krishnagopal ,
and Larry Davis

University of Maryland, College Park, MD 20740, USA
hanson@cs.umd.edu, koutilya@terpmail.umd.edu, sanjukta@umd.edu,
lsd@umiacs.umd.edu

Abstract. The field of action recognition has gained tremendous traction in recent years. A subset of this, detection of violent activity in videos, is of great importance, particularly in unmanned surveillance or crowd footage videos. In this work, we explore this problem on three standard benchmarks widely used for violence detection: the Hockey Fights, Movies, and Violent Flows datasets. To this end, we introduce a Spatiotemporal Encoder, built on the Bidirectional Convolutional LSTM (BiConvLSTM) architecture. The addition of bidirectional temporal encodings and an elementwise max pooling of these encodings in the Spatiotemporal Encoder is novel in the field of violence detection. This addition is motivated by a desire to derive better video representations via leveraging long-range information in both temporal directions of the video. We find that the Spatiotemporal network is comparable in performance with existing methods for all of the above datasets. A simplified version of this network, the Spatial Encoder is sufficient to match state-of-the-art performance on the Hockey Fights and Movies datasets. However, on the Violent Flows dataset, the Spatiotemporal Encoder outperforms the Spatial Encoder.

Keywords: Violence detection · Convolutional LSTM ·
Bidirectional LSTM · Action recognition · Fight detection ·
Video surveillance

1 Introduction

In recent years, the problem of human action recognition from video has gained momentum in the field of computer vision [14,26,36]. Despite its usefulness, the specific task of violence detection has been comparatively less studied. However, violence detection has huge applicability in public security and surveillance markets. Surveillance cameras are deployed in large numbers, particularly in schools, prisons etc. Problems such as lack of personnel and slow response arise, leading to a strong demand for automated violence detection systems. Additionally, with the surge in easy-to-access data uploaded to social media sites and

---

A. Hanson and K. PNVR—Equal contribution.

across the web, it is imperative to develop automated methods to childproof the internet. Hence, in recent years, focus has been directed towards solving this problem [1,7,8,32,37].

1.1   Contributions and Proposed Approach

In this work, we propose a Bidirectional Convolutional LSTM (BiConvL-STM) [17,30,38] architecture, called the Spatiotemporal Encoder, to detect violence in videos. Our architecture builds on existing ConvLTSM architectures in which we include bidirectional temporal encodings and elementwise max pooling, novel in the field of violence detection. We encode each video frame as a collection of feature maps via a forward pass through a VGG13 network [29]. We then pass these feature maps to a BiConvLSTM to further encode them along the video's temporal direction, performing both a pass forward in time and in reverse. Next, we perform an elementwise maximization on each of these encodings to create a representation of the entire video. Finally, we pass this representation to a classifier to identify whether the video contains violence. This extends the architecture of [32], which uses a Convolutional LSTM (ConvLSTM) by encoding temporal information in both directions. We speculate that access to both future and past inputs from a current state allows the BiConvLSTM to understand the context of the current input, allowing for better classification on heterogeneous and complex datasets. We validate the effectiveness of our networks by running experiments on three standard benchmark datasets commonly used for violence detection, namely, the Hockey Fights dataset (HF), the Movies dataset (M), and the Violent Flows dataset (VF). We find that our architecture matches state-of-the-art on the Hockey Fights [23] and Movies [23] datasets and performs comparably with other methods on the Violent Flows [15] dataset. Surprisingly, a simplified version of our architecture, called the Spatial Encoder, also matches state-of-the-art on Hockey Fights and Movies, leading us to speculate that these datasets may be comparatively smaller and/or simpler for the task of violence detection.

This paper is outlined as follows. Section 2 provides more detail about the model architectures we propose. Section 3 describes the datasets used in this work. Section 4 summarizes the training methodology. And Sect. 5 presents our experimental results and ablation studies.

1.2   Related Work

Early work in the field includes [22], where violent scenes in videos were recognized by using flame and blood detection and capturing the degree of motion, as well as the characteristic sounds of violent events. Significant work has been done on harnessing both audio and video features of a video in order to detect and localize violence [9]. For instance, in [18], a weakly supervised method is used to combine auditory and visual classifiers in a co-training way. While incorporating audio in the analysis may often be more effective, audio is not often

available in public surveillance videos. We address this problem by developing an architecture for violence detection that does not require audio features.

Additionally, violence is a rather broad category, encompassing not only person-person violence, but also crowd violence, sports violence, fire, gunshots, physical violence etc. In [21], crowd violence is detected using Latent Dirichlet Allocation (LDA) and Support Vector Machines (SVMs). Violence detection through specific violence-related object detection such as guns is also a current topic of research [24].

Several existing techniques use inter-frame changes for violence detection, in order to capture fast motion changing patterns that are typical of violent activity. [5] proposed the use of acceleration estimates computed from the power spectrum of adjacent frames as an indicator of fast motion between successive frames. [32] proposed a deep neural network for violence detection by feeding in frame differences. [10] proposed using blob features, obtained by subtracting adjacent frames, as the feature descriptor.

Other methods follow techniques such as motion tracking and position of limbs etc. to identify spatiotemporal interest points and extract features from these points. These include Harris corner detector [2], Motion Scale-Invariant Feature Transform (MoSIFT) [35]. MoSIFT descriptors are obtained from salient points in two parts: the first is an aggregated Histogram of Gradients (HoG) which describe the spatial appearance. The second part is an aggregated Histogram of optical Flow (HoF) which indicates the movement of the feature point. [37] used a modified version of motion-Weber local descriptor (MoIWLD), followed by sparse representation as the feature descriptor.

Additional work has used the Long Short-Term Memory (LSTM) [13] deep learning architecture to capture spatiotemporal features. [7] used LSTMs for feature aggregation for violence detection. The method consisted of extracting features from raw pixels using a CNN, optical flow images and acceleration flow maps followed by LSTM based encoding and a late fusion. Recently, [34] replaced the fully-connected gate layers of the LSTM with convolutional layers and used this improved model (named ConvLSTM) for predicting precipitation nowcasting from radar images with improved performance. This ConvLSTM architecture was also successfully used for anomaly prediction [19] and weakly-supervised semantic segmentation in videos [25].

Bidirectional RNNs are first introduced in [27]. Later, [12] proposed using the same for speech recognition task and was shown to perform better than an unidirectional RNN. Recently, bidirectional LSTMs were used in predicting network-wide traffic speed [3], framewise phoneme classification [11] etc. showing they are better in terms of prediction than unidirectional LSTMs. The same concept has been leveraged for tasks involving videos such as video-super resolution [16], object segmentation in a video [33] and learning spatiotemporal features for gesture recognition [37] and fine-grained action detection [30]. While several of these incorporate a convolutional module coupled with an RNN module, our architecture extends this by the inclusion of temporal encoding in both forward and backward temporal directions, through the use of a BiConvLSTM

and elementwise max pooling. We speculate that the access of future information from the current state is particularly beneficial in more heterogenous datasets.

2 Model Architecture

To appropriately classify violence in videos we sought to generate a robust video encoding to pass through a fully connected classifier network. We produce this video representation through a spatiotemporal encoder. This extracts features from a video that correspond to both spatial and temporal details via a Spatiotemporal Encoder (Sect. 2.1). The temporal encoding is done in both temporal directions, allowing access to future information from the current state. We also study a simplified version of the spatiotemporal encoder that encodes only spatial features via a simplified Spatial Encoder (Sect. 2.2). The architectures for both encoders are described below.

2.1 Spatiotemporal Encoder Architecture

The Spatiotemporal Encoder architecture is shown in Fig. 1. It consists of a spatial encoder that extracts spatial features for each frame in the video followed by a temporal encoder that allows these spatial feature maps to 'mix' temporally
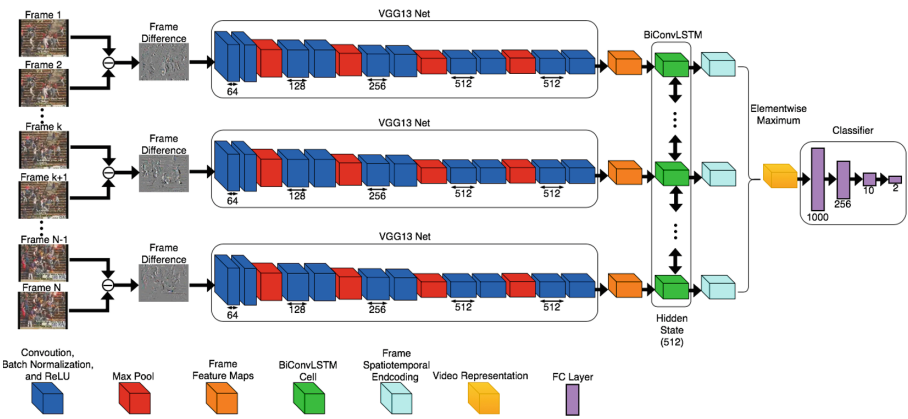

Fig. 1. The spatiotemporal encoder is comprised of three parts: a VGG13 network spatial encoder, a Bidirectional Convolution LSTM (BiConvLSTM), temporal encoder, and a classifier. Frames are resized to 224×224 and the difference between adjacent frames is used as input to the network. The VGG classifier and last max pooling layer is removed from VGG13 network (Blue and Red). The frame feature maps (Orange), are size $14 \times 14 \times 512$. The frame features are passed to the BiConvLSTM (Green) which outputs the frame spatiotemporal encodings (Cyan). An elementwise max pooling operation is performed on the spatiotemporal encoding to produce the final video representation (Gold). This video representation is then classified as violent or nonviolent via a fully connected classifier (Purple). (Color figure online)

to produce a spatiotemporal encoding at each time step. All of these encodings are then aggregated into a single video representation via an elementwise max pooling operation. This final video representation is vectorized and passed to a fully connected classifier.

Spatial Encoding: In this work, a VGG13 [29] convolutional neural network (CNN) model is used as the spatial encoder. The last max pool layer and all fully connected layers of the VGG13 net are removed, resulting in spatial feature maps for each frame of size $14 \times 14 \times 512$. Instead of passing video frames directly, adjacent frames were subtracted and used as input to the spatial encoder. This acts as pseudo-optical flow model and follows [28,32].

Temporal Encoding: A Bidirectional Convolutional LSTM (BiConvLSTM) is used as the temporal encoder, the input to which are the feature maps from the spatial encoder. We constructed the BiConvLSTM in such a way that the output from each cell is also $14 \times 14 \times 512$. The elementwise maximum operation is applied to these outputs as depicted in Fig. 1, thus resulting in a final video representation of size $14 \times 14 \times 512$.

A BiConvLSTM cell is essentially a ConvLSTM cell with two cell states. We present the functionality of ConvLSTM and BiConvLSTM in the following subsections.

ConvLSTM: A ConvLSTM layer learns global, long-term spatiotemporal features of a video without shrinking the spatial size of the intermediate representations. This encoding takes place during the recurrent process of the LSTM. In a standard LSTM network the input is vectorized and encoded through fully connected layers, the output of which is a learned temporal representation. As a result of these fully connected layers, spatial information is lost. Hence, if one desires to retain that spatial information, the use of a convolutional operation instead of fully connected operation may be desired. The ConvLSTM does just that. It replaces the fully connected layers in the LSTM with convolutional layers. The ConvLSTM is utilized in our work such that the convolution and recurrence operations in the input-to-state and state-to-state transitions can make full use of the spatiotemporal correlation information. The formulation of the ConvLSTM cell is shown below:

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + b_i)$$
$$f_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + b_f)$$
$$o_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + b_o)$$
$$C_t = f_t \odot C_{t-1} + i_t \odot tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c)$$
$$H_t = o_t \odot tanh(C_t)$$

Where "*" denote the convolution operator, "$\odot$" denote the Hadamard product, "σ" is the sigmoid function and W_{x*}, W_{h*} are 2D Convolution kernels that corresponding to the input and hidden state respectively. The hidden

$(H_0, H_1, ..H_{t-1})$ and the cell states $(C_1, C_2, ..C_t)$ are updated based on the input $(X_1, X_2, ..X_t)$ that pass through i_t, f_t and o_t gate activations during each time sequence step. b_i, b_f, b_o and b_c are the corresponding bias terms.

BiConvLSTM: The BiConvLSTM is an enhancement to ConvLSTM in which two sets of hidden and cell states are maintained for each LSTM cell: one for a forward sequence and the other for a backward sequence in time. BiConvLSTM can thereby access long-range context in both directions of the time sequence of the input and thus potentially gain a better understanding of the entire video. Figure 2 illustrates the functionality of a BiConvLSTM Cell. It is comprised of a ConvLSTM cell with two sets of hidden and cell states. The first set (h_f, c_f) is for forward pass and the second set (h_b, c_b) is for backward pass. For each time sequence, the corresponding hidden states from the two sets are stacked and passed through a Convolution layer to get a final hidden representation for that time step. That hidden representation is then passed to the next layer in the BiConvLSTM module as input.
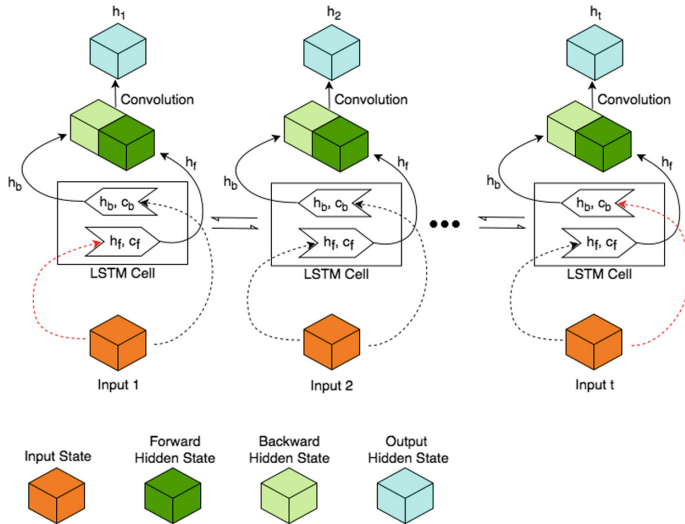

Fig. 2. Overview of a BiConvLSTM cell. The hidden and cell states are passed to the next LSTM cell in the direction of flow. Red dashed lines correspond to the first input in the time step for both the forward and backward hidden states. (Color figure online)

Classifier: The number of nodes in each layer in the fully connected classifier, ordered sequentially, are 1000, 256, 10, and 2. Each layer utilizes the hyperbolic tangent non-linearity. The output of the last layer is a binary predictor into classes violent and non-violent.

2.2   Spatial Encoder Architecture

Spatial Encoder is a simplified version of the Spatiotemporal Encoder architecture (Sect. 2.1) and is shown in Fig. 3. The temporal encoder is removed and elementwise max pooling is applied directly to the spatial features. Additionally, since we are interested in purely the spatial features in this architecture, adjacent frame differences are not used as input and frames are passed directly to the spatial encoder.
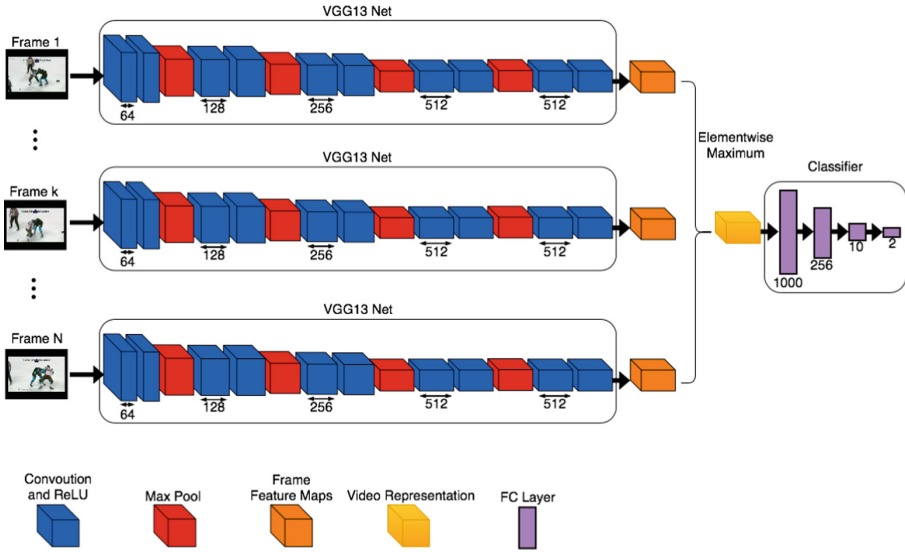

Fig. 3. The spatial encoder is comprised of two parts: a VGG13 network spatial encoder and a classifier. Frames are resized to 224×224 before provided as input to the network. The VGG classifier and last max pooling layer are removed from VGG13 network (Blue and Red). The frame feature maps (Orange), are size $14 \times 14 \times 512$. An elementwise max pooling operation is performed on the frame feature maps to produce the final video representation (Gold). This video representation is then classified as violent or nonviolent via a fully connected classifier (Purple). (Color figure online)

3   Data

Details about the three standard datasets widely used in this work are provided below. For all datasets, we downsampled each video to 20 evenly spaced frames as input to the network.

Hockey Fights dataset (HF) was created by collecting videos of ice hockey matches and contains 500 fighting and non-fighting videos. Almost all the videos in the dataset have a similar background and subjects (humans).

Movies dataset (M) consists of fight sequences collected from movies. The non-fight sequences are collected from publicly available action recognition datasets. The dataset is made up of 100 fight and 100 non-fight videos.

As opposed to the hockey fights dataset, the videos of the movies dataset are substantially different in their content.

Violent Flows dataset (VF) is a database of real-world, video footage of crowd violence, along with standard benchmark protocols designed to test both violent/non-violent classification and violence outbreak detection. The data set contains 246 videos. All the videos were downloaded from YouTube. The shortest clip duration is 1.04 s, the longest clip is 6.52 s, and the average length of a video clip is 3.60 s.

4 Training Methodology

For the spatial encoder, the weights were initialized as the pretrained ImageNet [4] weights for VGG13. For the Spatiotemporal Encoder, the weights of the BiConvLSTM cell and classifier were randomly initialized. Frame differences were taken for the Spatiotemporal Encoder architecture and frames were normalized to be in the range of 0 to 1. For both architectures, the learning rate was chosen to be 10^{-6}. A batch size of 8 video clips were used as input and the weight decay was set to 0.1. ADAM optimizer with default beta range $(0.5, 0.99)$ was used. Frames were selected at regular intervals and resized to 224×224. Additionally, random cropping (RC) and random horizontal flipping (RHF) data augmentations were used for the Hockey Fights and Movies clips, where as only RHF was applied to Violent Flows clips. Cross entropy loss was used during training. Furthermore, 5-fold cross validation was used to calculate performance.

5 Results

The following Subsects. 5.1 and 5.2 discuss the results and the corresponding model that obtained best performance for all three datasets.

5.1 Hockey Fights and Movies

The best performance for the Hockey Fights and Movies datasets was observed with the simpler Spatial Encoder Architecture depicted in Fig. 3 and described in Sect. 2.2. We obtained an accuracy of $96.96 \pm 1.08\%$ on the Hockey Fights dataset and an accuracy of $100 \pm 0\%$ on the Movies dataset, both of which match state-of-the-art. A comparison of our results with other recent work is given in Table 1. While our model performance was saturated at $100 \pm 0\%$ for the Movies dataset, it outperformed previous methods with comparable accuracy measures (Table 1 rows 1–11) by a statistically significant margin and hence, we believe, is a significant improvement.

These results, in contrast to most prior work, were attained without the use of a temporal encoding of the features. While the Spatiotemporal Encoder performed comparably to the Spatial Encoder, we observe that the additional level of complexity involved in utilizing the temporal features wasn't justified for datasets like Movies and Hockey Fight that are relatively more homogeneous than the Violent Flows dataset. We speculate that for certain domains, robust spatial information may be sufficient for violence classification.

Table 1. Performance comparison of different methods for Hockey Fights, Movies, and Violent Flows datasets. In the Hockey and Movies datasets our proposed methods match the state-of-the-art performance. In the case of the Violent Flows dataset, our method is comparable to existing methods. The best performance for each dataset and our proposed methods are highlighted in bold. Two methods for calculating accuracies are used here. Accuracy calculation of rows 1–11 are outlined in Sect. 5.3.

Method	Hockey	Movies	Violent Flows
MoSIFT + HIK [23]	90.9%	89.5%	-
ViF [15]	$82.9 \pm 0.14\%$	-	$81.3 \pm 0.21\%$
MoSIFT + KDE + Sparse coding [35]	$94.3 \pm 1.68\%$	-	$89.05 \pm 3.26\%$
Deniz et al. [6]	$90.1 \pm 0\%$	$98.0 \pm 0.22\%$	-
Gracia et al. [10]	$82.4 \pm 0.4\%$	$97.8 \pm 0.4\%$	-
Substantial derivative [20]	-	$96.89 \pm 0.21\%$	$85.43 \pm 0.21\%$
Bilinski et al. [1]	93.4%	99%	**96.4%**
MoIWLD [37]	$\mathbf{96.8 \pm 1.04\%}$	-	$93.19 \pm 0.12\%$
ViF + OViF [8]	$87.5 \pm 1.7\%$	-	$88 \pm 2.45\%$
Three streams + LSTM [7]	93.9	-	-
Proposed: spatiotemporal encoder	$\mathbf{96.54 \pm 1.01\%}$	$\mathbf{100 \pm 0\%}$	$92.18 \pm 3.29\%$
Proposed: spatial encoder	$\mathbf{96.96 \pm 1.08\%}$	$\mathbf{100 \pm 0\%}$	$90.63 \pm 2.82\%$
Swathikiran et al. [32]	$97.1 \pm 0.55\%*$	$\mathbf{100 \pm 0\%*}$	$94.57 \pm 2.34\%*$
Proposed: spatiotemporal encoder	$97.9 \pm 0.37\%*$	$\mathbf{100 \pm 0\%*}$	$\mathbf{96.32 \pm 1.52\%*}$
Proposed: spatial encoder	$\mathbf{98.1 \pm 0.58\%*}$	$\mathbf{100 \pm 0\%*}$	$93.87 \pm 2.58\%*$

For the purpose of fair comparison with [32], we also present performance measured through the accuracy calculation of [32]. For more details refer to Sect. 5.3

5.2 Violent Flows

The best performance on the Violent Flows dataset was observed using the Spatiotemporal Encoder architecture shown in Fig. 1 and described in Sect. 2.1. Our accuracy on the Violent Flows dataset was $92.18 \pm 3.29\%$. While not state-of-the-art, this accuracy is comparable to existing recent methods as shown in Table 1. We noticed batch normalization caused a decrease in performance on the Violent Flows dataset. Hence, all reported accuracies for the Violent Flows dataset were obtained without applying batch normalization in the networks.

5.3 Accuracy Evaluation

Due to the small size of the datasets, we chose to employ 5-fold cross validation to evaluate model accuracies. We split each dataset into 5 equal sized and randomly partitioned folds. One fold is reserved for testing and the other four are used for training. The model is trained from scratch once for each test fold and hence five test accuracies are obtained per epoch of training. We calculate the mean per epoch of these accuracies and locate the epoch with maximal accuracy value. We then calculate the mean and standard deviation of all 100 test accuracies

that lie within a 10 epoch radius of this maximal accuracy. We report this as our overall model accuracy and standard deviation.

This contrasts the accuracy evaluation used in [32], where for each fold the maximum value over all epochs is obtained, and the mean of these values is reported [31]. For completeness, we report our accuracies using this evaluation method in Table 1 using a '*'.

As shown in Fig. 4, the mean test accuracy for the Hockey Fights dataset peaks at 97.3% for epoch 63. We take the mean and standard deviation of test accuracies from epoch 53 to epoch 73 and obtain an overall accuracy of $96.96 \pm 1.08\%$.
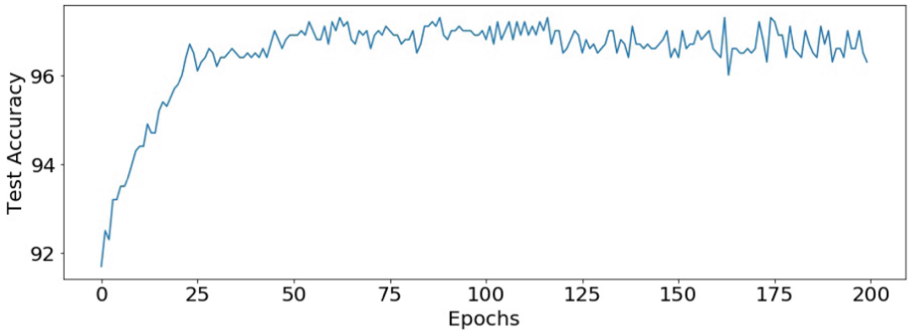
Fig. 4. Mean fold accuracy on Hockey evaluated using the spatial encoder architecture.

Figure 5 shows the mean test accuracy of the Violent Flows dataset to be 94.69 at epoch 710. The mean and standard deviation of test accuracies between epoch 700 and 720 produces an overall accuracy of $92.18 \pm 3.29\%$.

The Movies dataset converged to 100.0% after 3 epochs. Hence we report an overall accuracy of 100.0% for this dataset.
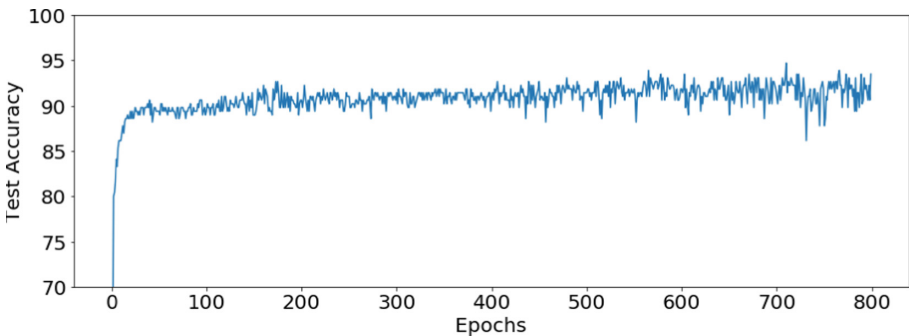

Fig. 5. Mean fold accuracy on Violent Flows evaluated using the spatiotemporal encoder architecture.

5.4   Ablation Studies

We conducted several ablation studies to determine how the boost in performance can be attributed to the key components in our Spatiotemporal Encoder Architecture. In particular, we examine the effects of using a VGG13 network pretrained on ImageNet to encode spatial features, the use of a BiConvLSTM network to refine these encodings temporally, and the use of elementwise max pooling to create an aggregate video representation. To baseline performance gains, we compare against architectural decisions made by the study that most closely resembles our work, [32].

Spatial vs Spatiotemporal Encoders. This study examines the role of a temporal encoder during classification. The performance of the Spatial (Sect. 2.2) and Spatiotemporal (Sect. 2.1) Encoders are compared and illustrated in Figs. 6 and 7 for the Hockey and Violent Flows respectively. We see the temporal encoding is adding a slight boost in performance in the case of Violent Flows. However, the simpler Spatial Encoder architecture performs slightly better for the hockey dataset.
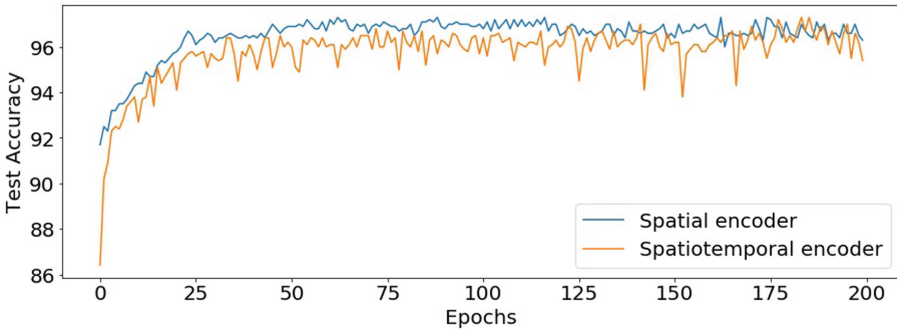

Fig. 6. Performance comparison between spatial and spatiotemporal encoders on the hockey dataset.

Elementwise Max Pooling vs. Last Encoding. In this study, we sought to determine the usefulness of aggregating the spatiotemporal encodings via the elementwise max pool operation. We did so by removing the elementwise max pooling operation and running classification on the last spatiotemporal frame representation. Figure 8 depicts that using elementwise max pool aggregation lead to significant improvement in performance.

ConvLSTM vs. BiConvLSTM. For this study, we evaluated the impact of bidirectionality of the BiConvLSTM on violence classification. We compared its performance to a ConvLSTM module and depict the accuracies of both in Fig. 9. BiConvLSTM yields a slightly higher classification accuracy.
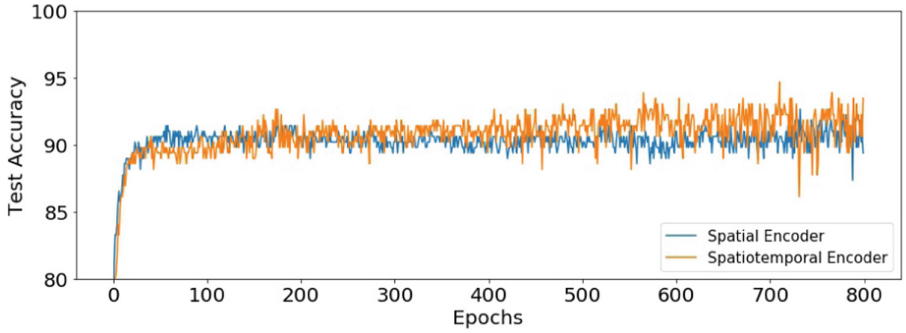
Fig. 7. Performance comparison between spatial and spatiotemporal encoders on the Violent Flows dataset.
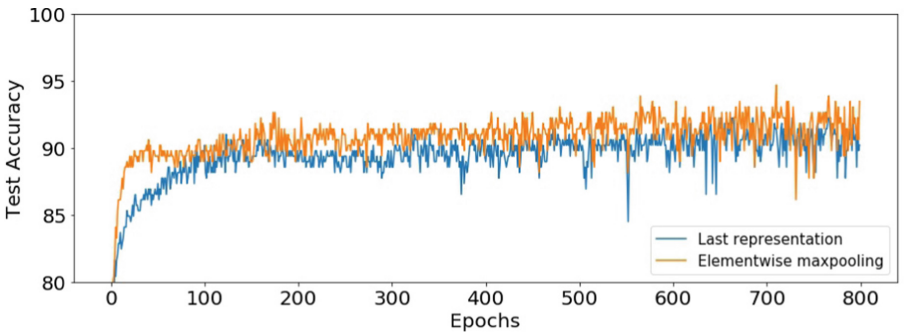

Fig. 8. Performance comparison between the feature aggregation techniques max pooling and last time sequence representation from the BiConvLSTM module on the Violent Flows dataset.
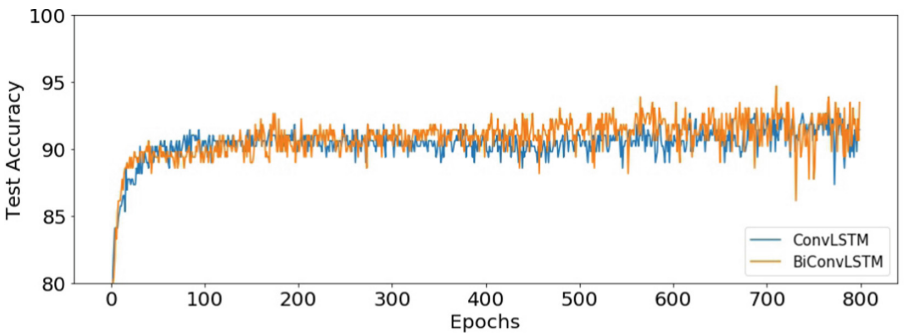

Fig. 9. Performance comparison between ConvLSTM and BiConvLSTM as temporal encoders on the Violent Flows dataset.

AlexNet vs. VGG13. The aim of this study was to understand the affect of different spatial encoder architectures on the classification performance. For this we chose AlexNet and VGG13 Net pretrained on ImageNet as spatial encoders. Figure 10 shows the performance comparison for the two encoders. It is apparent that VGG13 is performing appreciably better than AlexNet.
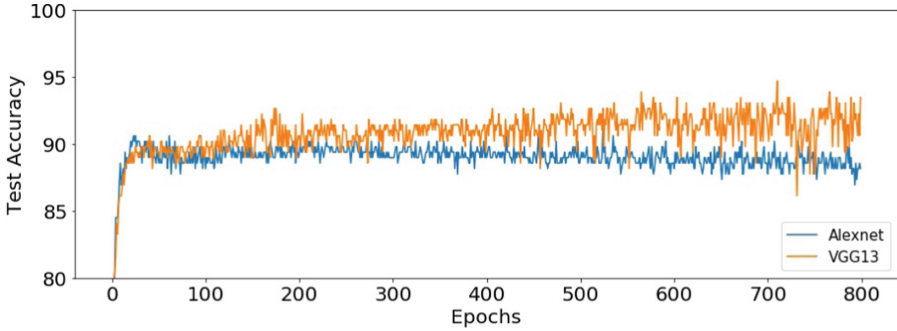

Fig. 10. Performance comparison between AlexNet and VGG13 pretrained models as spatial encoders on Violent Flows dataset.

6   Conclusions

We have proposed a Spatiotemporal Encoder architecture and a simplified Spatial Encoder for supervised violence detection. The former performs reasonably well on all the three benchmark datasets whereas the later matches state-of-the-art performance on the Hockey Fights and Movies datasets. We presented various ablation studies that demonstrate the significance of each module in the spatiotemporal encoder model and provide grounding for our architectures.

While several studies have used ConvLSTMs for video related problems, our contribution of introducing bidirectional temporal encodings and the element-wise max pooling of those encodings facilitates better context-based representations. Hence, our Bidirectional ConvLSTM performs better for more heterogeneous and complex datasets such as the Violent Flows dataset compared to the ConvLSTM architecture [32]. Based on the comparisons in the results section, it is not clear if there is a method that is consistently best. Current commonly used benchmark violence datasets are relatively small (a few hundred videos) compared to traditional deep learning dataset sizes. We anticipate that larger datasets may lead to better comparisons between methods. This may constitute an interesting future course of study.

Additionally, we were surprised by the performance of the Spatial Encoder Architecture. Violence detection is a difficult problem, but we speculate that some datasets may be easier than others. Pause a movie or hockey match at just the right frame and it is likely that a human user will be able to tell if a fight

scene or brawl is taking place. We hypothesize that the same is true for a neural network. A specific frame may fully encode violence in a video for a particular domain. We speculate that this is why our Spatial Encoder Architecture was able to match state-of-the-art on the Hockey Fights and Movies datasets. For more complex datasets and scenes with rapidly changing violence features, it is important to understand the context of the frame in the whole video, i.e., both the past video trajectory and future video trajectory leading outwards from that frame. This is particularly true for longer or more dynamic videos with greater heterogeneity; the same sequence of frames could go one of several directions in the future. It is for this reason that we believe our novel contributions to the architecture, the 'Bi' in the BiConvLSTM and elementwise max pooling, are beneficial to develop better video representations, and we speculate that our architecture may perform well on more dynamic and heterogeneous datasets. We anticipate further investigation into this may lead to fruitful results.

References

1. Bilinski, P.T., Brémond, F.: Human violence recognition and detection in surveillance videos. In: 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 30–36 (2016)
2. Chen, D., Wactlar, H., Chen, M.Y., Gao, C., Bharucha, A., Hauptmann, A.: Recognition of aggressive human behavior using binary local motion descriptors. In: 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2008, pp. 5238–5241. IEEE (2008)
3. Cui, Z., Ke, R., Wang, Y.: Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction. CoRR abs/1801.02143 (2018)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255, June 2009. https://doi.org/10.1109/CVPR.2009.5206848
5. Deniz, O., Serrano, I., Bueno, G., Kim, T.K.: Fast violence detection in video. In: 2014 International Conference on Computer Vision Theory and Applications (VISAPP), vol. 2, pp. 478–485. IEEE (2014)
6. Déniz-Suárez, O., Serrano, I., García, G.B., Kim, T.K.: Fast violence detection in video. In: 2014 International Conference on Computer Vision Theory and Applications (VISAPP), vol. 2, pp. 478–485 (2014)
7. Dong, Z., Qin, J., Wang, Y.: Multi-stream deep networks for person to person violence detection in videos. In: Tan, T., Li, X., Chen, X., Zhou, J., Yang, J., Cheng, H. (eds.) CCPR 2016. CCIS, vol. 662, pp. 517–531. Springer, Singapore (2016). https://doi.org/10.1007/978-981-10-3002-4_43
8. Gao, Y., Liu, H., Sun, X., Wang, C., Liu, Y.: Violence detection using oriented violent flows. Image Vis. Comput. **48**(C), 37–41 (2016). https://doi.org/10.1016/j.imavis.2016.01.006
9. Giannakopoulos, T., Kosmopoulos, D., Aristidou, A., Theodoridis, S.: Violence content classification using audio features. In: Antoniou, G., Potamias, G., Spyropoulos, C., Plexousakis, D. (eds.) SETN 2006. LNCS (LNAI), vol. 3955, pp. 502–507. Springer, Heidelberg (2006). https://doi.org/10.1007/11752912_55

10. Gracia, I.S., Suarez, O.D., Garcia, G.B., Kim, T.K.: Fast fight detection. PLoS One **10**(4), e0120448 (2015)

11. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM networks. In: Proceedings of the 2005 IEEE International Joint Conference on Neural Networks, vol. 4, pp. 2047–2052, July 2005. https://doi.org/10.1109/IJCNN.2005.1556215

12. Graves, A., Jaitly, N., Mohamed, A.R.: Hybrid speech recognition with deep bidirectional LSTM. In: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) (2013)

13. Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., Schmidhuber, J.: LSTM: a search space odyssey. IEEE Trans. Neural Netw. Learn. Syst. **28**(10), 2222–2232 (2017)

14. Guo, G., Lai, A.: A survey on still image based human action recognition. Pattern Recogn. **47**(10), 3343–3361 (2014)

15. Hassner, T., Itcher, Y., Kliper-Gross, O.: Violent flows: real-time detection of violent crowd behavior. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–6, June 2012. https://doi.org/10.1109/CVPRW.2012.6239348

16. Huang, Y., Wang, W., Wang, L.: Video super-resolution via bidirectional recurrent convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 1015–1028 (2018). https://doi.org/10.1109/TPAMI.2017.2701380

17. Huang, Y., Wang, W., Wang, L.: Bidirectional recurrent convolutional networks for multi-frame super-resolution. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 28, pp. 235–243. Curran Associates, Inc. (2015). http://papers.nips.cc/paper/5778-bidirectional-recurrent-convolutional-networks-for-multi-frame-super-resolution.pdf

18. Lin, J., Wang, W.: Weakly-supervised violence detection in movies with audio and video based co-training. In: Muneesawang, P., Wu, F., Kumazawa, I., Roeksabutr, A., Liao, M., Tang, X. (eds.) PCM 2009. LNCS, vol. 5879, pp. 930–935. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-10467-1_84

19. Medel, J.R., Savakis, A.E.: Anomaly detection in video using predictive convolutional long short-term memory networks. CoRR abs/1612.00390 (2016)

20. Mohammadi, S., Kiani, H., Perina, A., Murino, V.: Violence detection in crowded scenes using substantial derivative. In: 2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, August 2015. https://doi.org/10.1109/avss.2015.7301787

21. Mousavi, H., Mohammadi, S., Perina, A., Chellali, R., Murino, V.: Analyzing tracklets for the detection of abnormal crowd behavior. In: 2015 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 148–155. IEEE (2015)

22. Nam, J., Alghoniemy, M., Tewfik, A.H.: Audio-visual content-based violent scene characterization. In: Proceedings of the 1998 International Conference on Image Processing, ICIP 1998 (Cat. No. 98CB36269), vol. 1, pp. 353–357, October 1998. https://doi.org/10.1109/ICIP.1998.723496

23. Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., Sukthankar, R.: Violence detection in video using computer vision techniques. In: Real, P., Diaz-Pernil, D., Molina-Abril, H., Berciano, A., Kropatsch, W. (eds.) CAIP 2011. LNCS, vol. 6855, pp. 332–339. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23678-5_39

24. Olmos, R., Tabik, S., Herrera, F.: Automatic handgun detection alarm in videos using deep learning. Neurocomputing **275**, 66–72 (2018)

25. Patraucean, V., Handa, A., Cipolla, R.: Spatio-temporal video autoencoder with differentiable memory. arXiv preprint arXiv:1511.06309 (2015)
26. Peng, X., Schmid, C.: Multi-region two-stream R-CNN for action detection. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 744–759. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_45
27. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Sig. Process. **45**(11), 2673–2681 (1997). https://pdfs.semanticscholar.org/4b80/89bc9b49f84de43acc2eb8900035f7d492b2.pdf
28. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 27, pp. 568–576. Curran Associates, Inc. (2014). http://papers.nips.cc/paper/5353-two-stream-convolutional-networks-for-action-recognition-in-videos.pdf
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015). http://arxiv.org/abs/1409.1556
30. Singh, B., Marks, T.K., Jones, M., Tuzel, O., Shao, M.: A multi-stream bidirectional recurrent neural network for fine-grained action detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1961–1970, June 2016. https://doi.org/10.1109/CVPR.2016.216
31. Sudhakaran, S.: Personal communication
32. Sudhakaran, S., Lanz, O.: Learning to detect violent videos using convolutional long short-term memory. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6. IEEE (2017)
33. Tokmakov, P., Alahari, K., Schmid, C.: Learning video object segmentation with visual memory. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 4491–4500 (2017)
34. Xingjian, S., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.C.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Advances in Neural Information Processing Systems, pp. 802–810 (2015)
35. Xu, L., Gong, C., Yang, J., Wu, Q., Yao, L.: Violent video detection based on MoSIFT feature and sparse coding. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3538–3542. IEEE (2014)
36. Yeung, S., Russakovsky, O., Mori, G., Fei-Fei, L.: End-to-end learning of action detection from frame glimpses in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2678–2687 (2016)
37. Zhang, T., Jia, W., He, X., Yang, J.: Discriminative dictionary learning with motion weber local descriptor for violence detection. IEEE Trans. Cir. Sys. Video Technol. **27**(3), 696–709 (2017). https://doi.org/10.1109/TCSVT.2016.2589858
38. Zhang, Y., Chan, W., Jaitly, N.: Very deep convolutional networks for end-to-end speech recognition. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4845–4849 (2017)