Bidirectional Convolutional LSTM for the Detection of Violence in Videos

UNIVERSITY OF MARYLAND

Alex Hanson*, Koutilya PNVR*, Sanjukta Krishnagopal, Larry Davis

*equal contribution

Goal

Propose Spatiotemporal Encoder architecture for violence detection on benchmark datasets: Bidirectional Convolutional LSTM (BiConvLSTM).

Contribution

- Run ablation studies to evaluate key modules of this architecture.
- Propose a simpler Spatial Encoder architecture that works on certain datasets.

Takeaways

- Matches state-of-the-art on two benchmark datasets.
- Demonstrates a need for larger, more complex datasets in the violence detection domain.

Model Architectures

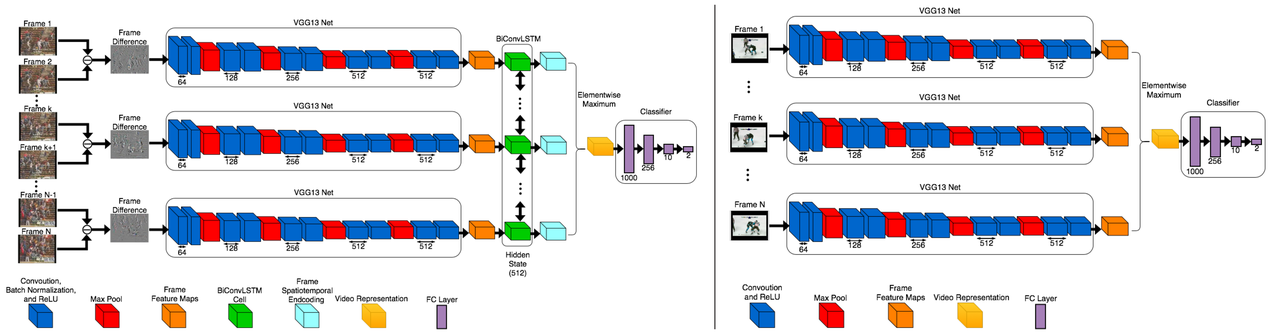

Figure 1: Spatiotemporal Encoder (left); Spatial Encoder (right)

Ablation

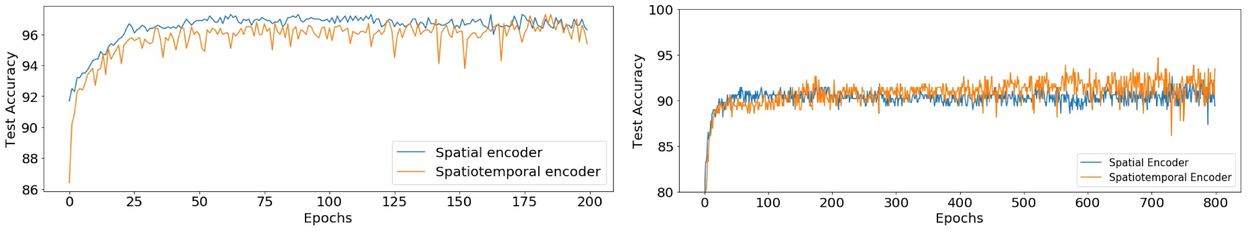

Figure 2: Comparing Spatiotemporal and Spatial Encoders on Hockey (left); Violent Flows (right) datasets

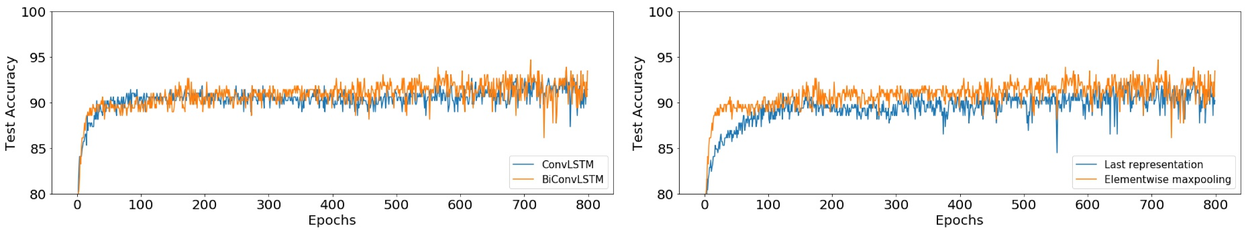

Figure 3: Ablation study of different modules on Violent Flows: BiConvLSTM vs ConvLSTM (left); Elementwise Maxpooling vs Last representation (right)

Method Comparison

Method	Hockey	Movies	Violent Flows
MoSIFT+HIK	90.9%	89.5%	-
ViF	82.9±0.14%	-	81.3±0.21%
MoSIFT+KDE+Sparse Coding	94.3±1.68%	-	89.05±3.26%
Deniz et al.	90.1±0%	98.0±0.22%	-
Gracia et al.	82.4±0.4%	97.8±0.4%	-
Substantial Derivative	-	96.89±0.21%	85.43±0.21%
Bilinski et al.[1]	93.4%	99%	**96.4%**
MoIWLD[2]	**96.8±1.04%**	-	93.19±0.12%
ViF+OViF	87.5±1.7%	-	88±2.45%
Three streams + LSTM	93.9	-	-
Proposed: Spatiotemporal Encoder	**96.54±1.01%**	**100±0%**	92.18±3.29%
Proposed: Spatial Encoder	**96.96±1.08%**	**100±0%**	90.63±2.82%
Swathikiran et al.[3]	97.1±0.55%*	**100±0%***	94.57±2.34%*
Proposed: Spatiotemporal Encoder	97.9±0.37%*	**100±0%***	**96.32±1.52%***
Proposed: Spatial Encoder	**98.1±0.58%***	**100±0%***	93.87±2.58%*

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. For Violent Flows, our method is comparable to existing methods.
**Following accuracy calculation from [3, 4]*

Results

- The accuracy of the Spatiotemporal Encoder architecture is comparable to existing recent methods on the Violent Flows dataset
- The simpler Spatial Encoder architecture is sufficient to match state-of-the-art accuracy on the Hockey and Movies datasets.

Code

Scan me

https://github.com/koutilya40192/BiConvLSTM_Violence_Detection

References

[1] Piotr Tadeusz Bilinski and François Brémond. Human violence recognition and detection in surveillance videos. *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 30–36, 2016.

[2] Tao Zhang, Wenjing Jia, Xiangjian He, and Jie Yang. Discriminative dictionary learning with motion weber local descriptor for violence detection. *IEEE Trans. Cir. and Sys. for Video Technol.*, 27(3):696–709, March 2017.

[3] Swathikiran Sudhakaran and Oswald Lanz. Learning to detect violent videos using convolutional long short-term memory. In *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*, pages 1–6. IEEE, 2017.

[4] Swathikiran Sudhakaran. Personal communication.