Violence Detection using Vision Transformers

Project Review

Sakthiprian S

Roll No.: EC21B1075

under the supervision of
Dr. Umarani Jayaraman

Department of Electronics & Communication Engineering
Indian Institute of Information Technology
Design and Manufacturing, Chennai, Tamil-Nadu-600127, India

May 5, 2025

Seminar Outline

Problem Statement

Background:

- Public safety concerns due to increasing violent incidents in public spaces.
- Need for automated surveillance systems to detect violence in real-time.
- Traditional CNN-based models struggle with long-range dependencies and complex action recognition.

Objectives:

- Develop a **Vision Transformer (ViT)-based model** for detecting violence in videos.
- Enhance **classification accuracy** by leveraging self-attention mechanisms in ViTs.
- Fine-tune the model on an open-source dataset to benchmark its performance against existing approaches.

Selected Literature Survey I

CrimeNet: A Deep Learning Approach to Violence Detection

CrimeNet is a deep neural network designed for violence detection in surveillance footage. It integrates CNN and LSTM layers, where the CNN extracts spatial features, and the LSTM models temporal dependencies.

Edge Deployment of Vision Transformers

A study explored the use of pre-trained Vision Transformers (ViTs) for video violence detection in edge computing environments. Hybrid ViTs improved accuracy by 2-3% over CNN/LSTM models.

Selected Literature Survey II

Video Vision Transformers for Violence Detection

A deep learning framework using Video Vision Transformers (ViViT) was introduced for detecting violence in videos.

Lightweight Transformers for Indoor Surveillance

A lightweight transformer model was tailored for detecting violence in indoor surveillance environments, tackling issues like occlusions and limited datasets.

JOSENet: Joint Stream Embedding Network

JOSENet introduces a dual spatiotemporal stream processing model leveraging RGB frames and optical flow.

Available Open-Source Datasets

Dataset	Classes	Hours	Files
UCF Crime	14	128	1,900
UBI Fights	2	80	1,000
RWF-2000	2	$\sim$3	2,000
XD Violence	7	217	4,754
NTU CCTV Fights	2	1,417.68	1,000

Dataset Preparation I

- **Dataset:** RWF-2000 – 2000 surveillance videos labeled as Fight / Non-Fight (1000 each).
- Each video has 150 frames; original resolution frames are extracted.
- Clips of **16 consecutive frames** are created from each video (non-overlapping).
- Remaining 6 frames at the end are discarded to maintain uniform clip size.
- This results in a total of **16,000 clips (256,000 frames)** for training and **1,792 clips (28,672 frames)** for validation.
- Dataset is organized as:
  - `clips16/train/Fight`, `clips16/train/NotFight`
  - `clips16/val/Fight`, `clips16/val/NotFight`

Training Process I

This study explores two approaches for violence detection in video:

- **Optical Flow + ViT**
- **Spatiotemporal Attention with TimeSformer**

Optical Flow + ViT

- **Optical Flow:** Estimates motion between consecutive frames over 16-frame clips.
- **Flow Accumulation:** Momentum-based update:

$$\text{acc\_flow} = \alpha \cdot \text{curr\_flow} + (1 - \alpha) \cdot \text{prev\_acc\_flow}, \quad \alpha = 0.7$$

- **Model Input:** 16th RGB frame + accumulated flow.
- **Output:** Binary classification – Fight / NotFight.

Training Process II

ViT + Optical Flow Training Setup:

- Model: ViT-Base (patch size: 16, image size: 224).
- Optimizer: AdamW (LR: 1e-4)
- Scheduler: ReduceLROnPlateau.
- Batch Size: 128, Epochs: 20, VRAM: 22.5 GB, Time/Epoch: 15 mins.

ViT + Optical Flow Training Details:

- Dataset: Clips with 16 frames per segment.
- GPU: Google Colab L4.

Training Process III

Spatiotemporal Attention with TimeSformer

- **TimeSformer:** Extends ViT with temporal attention to capture motion across frames using divided space-time attention.
- **Input:** 16 RGB frames per clip, split into patches.
- **Output:** Binary classification – Fight / NotFight.

TimeSformer Training Setup:

- Model: TimeSformer-Base (16-frame input, patch size: 16, image size: 224).
- Optimizer: AdamW (LR: 1e-4), Scheduler: ReduceLROnPlateau.

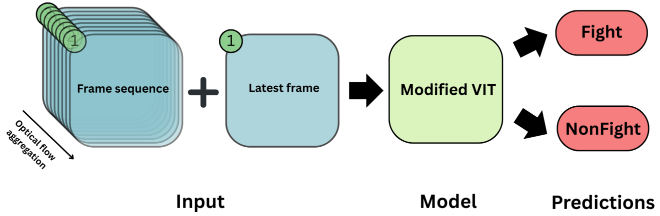
Training Process IV

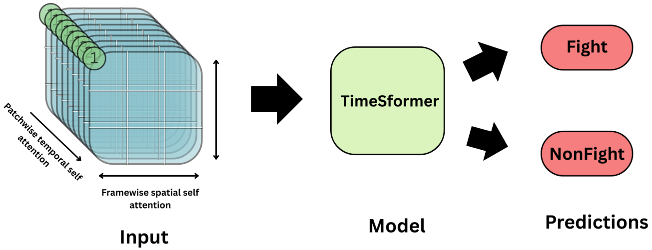
- Batch Size: 14, Epochs: 14, VRAM: 23.8 GB, Time/Epoch: 50 mins.

TimeSformer Training Details:

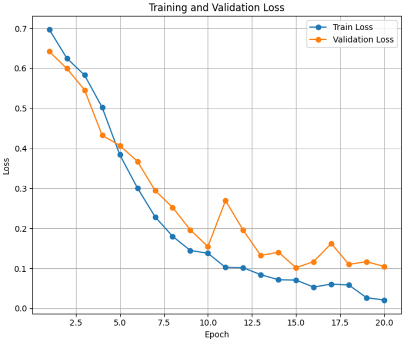
- Dataset: 16-frame clips from RWF-2000.
- GPU: Google Colab L4.

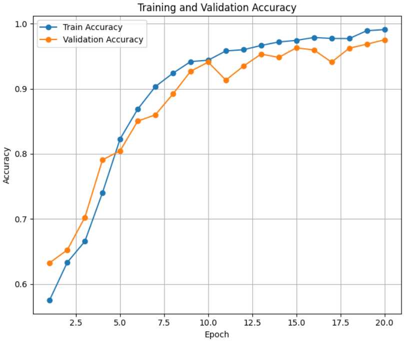
Working of both models



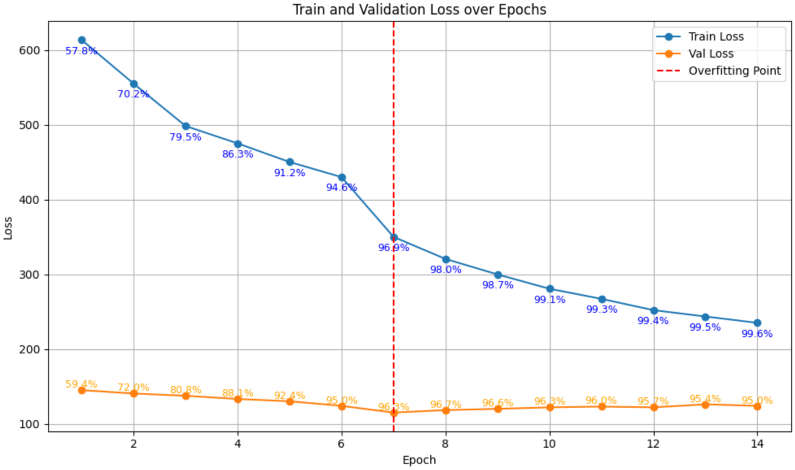


ViT + Optical Flow Accumulation Training Results





TimeSformer model Training Results



Summary & Conclusion

Summary

- Developed a **ViT-based model** for violence detection using the **RWF-2000 dataset** with two approaches: **Optical Flow + ViT** and **TimeSformer**.
- Achieved **97.2** percent validation accuracy on optical flow model and **95.2** percent validation accuracy in timesformer model.

Conclusion and future scope

- Both **Optical Flow + ViT** and **TimeSformer** models show great potential for **violence detection in videos**.
- High accuracy on validation sets indicates effective **motion and spatiotemporal feature extraction**.
- Future work: Train on bigger datasets and longer clip lengths for more accuracy.

Selected References I

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N.: 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale', arXiv preprint arXiv:2010.11929.

[2] Li, Y., Wang, X., Zhang, J., & Li, H.: 'MViTv2: Improved Multiscale Vision Transformers for Image and Video Recognition', Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[3] Park, J., Kim, D., & Lee, H.: 'Spatiotemporal Feature Learning for Video-Based Violence Detection Using Deep Learning', IEEE Transactions on Image Processing, 33, 2024, pp. 2154–2168.

[4] Zhou, L., Yang, X., & Sun, J.: 'Histogram of Oriented Tracklets (HoT) for Abnormal Activity Recognition in Surveillance Videos', CVPR Workshops, 2022.

[5] Rendón-Segador, F. J., Álvarez-García, J. A., Salazar-González, J. L., & Tommasi, T.: 'CrimeNet: Neural Structured Learning using Vision Transformer for Violence Detection', Sensors, 24(16), 2023, p. 5429. https://www.mdpi.com/1424-8220/24/16/5429

[6] Wang, T., Liu, H., & Zhang, C.: 'Reinforcement Learning-Based Mixture of Vision Transformers for Video Analysis', NeurIPS, 2023.

[7] Chen, F., Xu, R., & Zhao, M.: 'Multi-Scale Bottleneck Transformer for Multimodal Violence Detection', Journal of Artificial Intelligence Research, 72, 2023, pp. 1285–1302.

[8] Rahman, M., Hasan, R., & Ahmed, S.: 'Lightweight Vision Transformers for Real-Time Violence Detection in Indoor Surveillance', Pattern Recognition Letters, 168, 2023, pp. 56–65.

Selected References II

[9] Patel, A., & Singh, P.: 'VioNet: Vision Transformer and 3D Neural Network Fusion for Violence Detection in Videos', ICML Proceedings, 2023.

[10] Kim, J., & Choi, S.: 'JOSENet: Joint Stream Embedding Network for Self-Supervised Violence Detection', Pattern Analysis and Applications, 26(3), 2023, pp. 234–248.

[11] Singh, S., Dewangan, S., Krishna, G. S., Tyagi, V., Reddy, S., & Medi, P. R.: 'Video Vision Transformers for Violence Detection', arXiv preprint arXiv:2209.03561, 2022. https://arxiv.org/abs/2209.03561

[12] Bertasius, G., Wang, H., & Torresani, L.: 'Is Space-Time Attention All You Need for Video Understanding?', Proceedings of the International Conference on Machine Learning (ICML), 2021, pp. 813–824. https://arxiv.org/abs/2102.05095

[13] Senadeera, D. C., Yang, X., Kollias, D., & Slabaugh, G.: 'CUE-Net: Violence Detection Video Analytics with Spatial Cropping, Enhanced UniformerV2 and Modified Efficient Additive Attention', arXiv preprint arXiv:2404.18952, 2024. https://arxiv.org/abs/2404.18952

Thank You