Two-Stage Spatio-Temporal Vision Transformer for the Detection of Violent Scenes

Mihai Gabriel Constantin
University Politehnica of Bucharest
Bucharest, Romania
mihai.constantin84@upb.ro

Bogdan Ionescu
University Politehnica of Bucharest
Bucharest, Romania

2022 14th International Conference on Communications (COMM) | 978-1-6654-9485-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/COMM54429.2022.9817200

***Abstract*—The rapid expansion and adoption of CCTV systems brings with itself a series of problems that, if remain unchecked, have the potential of hindering the advantages brought by such systems and reduce the effectiveness of this type of system in security surveillance scenarios. The possibly vast quantities of data associated with a CCTV system that covers a city or problematic areas of that city, venues, events, industrial sites or even smaller security perimeters can overwhelm the human operators and make it hard to distinguish important security events from the rest of the normal data. Therefore, the creation of automated systems that are able to provide operators with accurate alarms when certain events take place is of paramount importance, as this can heavily reduce their workload and improve the efficiency of the system. In this regard, we propose a Two-Stage Vision Transformer-based (2SViT) system for the detection of violent scenes. In this setup, the first stage handles frame-level processing, while the second stage processes temporal information by gathering frame-level features. We train and validate our proposed Transformer architecture on the popular XD-Violence dataset, while testing some size variations for the architecture, and show good results when compared with baseline scores.**

***Index Terms*—violence detection, violent behavior, surveillance, Vision Transformers, deep neural networks,**

I. Introduction

The topic of violence detection gained a considerate amount of attention in the literature, being analyzed in several scenarios, starting from security-related scenarios like security surveillance [1] and smart city applications [2], but also targeting content moderation applications like parental video filtering [3]. While violence detection may be seen as part of the larger action recognition domain, its large domain of applicability and immediate added value makes it stand out, and a large number of papers deal with violence on its own, without integrating it in the larger action recognition domain. This is also due to the multimodal nature of violence, considering that there are scenarios where violence may only be present in the audio modality, but also due to the wide range of actions that are considered violent. For example, while pointing a gun at someone is a visually different action compared with a fight breaking out between two or more people, the two actions both represent some type of violent behaviour.

Another interesting aspect of violence detection is related to the concept itself. In annotating video data for violence detection there is an inherent subjectivity that may cause human subjects to disagree with regards to the degree or to the presence of violence in certain video scenes. This is handled in several ways in the literature by dataset creators, ranging from creating several definitions for violence and using master annotators for solving the inconsistencies in the annotations [3], to splitting the dataset according to several concepts related to violence like “Abuse”, “Car Accident”, “Explosion”, “Fighting”, “Riot”, and “Shooting” [4], [5]. These different approaches also allow researchers interested in creating violence detectors to select the type of data that is most appropriate for their particular use cases.

The remainder of the paper is structured as follows. Section II analyzes related work and the current state-of-the-art in automated violence detection. The proposed method is presented in Section III, while the dataset used for training and validating the system, as well as the results are presented in Section IV. Finally, Section V presents the main conclusions and analyzes the proposed work.

II. Related Work

Some of the defining aspects of violence detection presented in Section I, such as multimodality and the possibility of subjective annotations, create an interesting landscape,

This work has been funded by the Ministry of Research, Innovation, and Digitization, UEFISCDI, projects IDENEO, Solutions Axis, grant 29SOL/2021 and DeepVisionRomania, Solutions Axis, grant 28SOL/2021.

978-1-6654-9485-4/22/$31.00 ©2022 IEEE

where deep neural networks do not necessarily represent the state of the art approach. Papers that analyze the state-of-the-art on violence detection [1], [3] show that methods based on feature extraction or more traditional classifiers [6], [7] are still published and can still achieve good results in the last 5 years, even after the great successes recorded by deep learning approaches in other domains and benchmarking activities like ILSVRC [8]. However, in recently published papers, there is a noticeable trend favoring the use of certain types of deep learning-based approaches.

Simple deep neural network or multilayer perceptron approaches, without any convolutional or temporal operations, are most often used for processing pre-computed features that are then fed into the input of the networks [9], [10].

However, one of the most popular approaches in the deep learning category is represented by the use of 3D Convolutional Neural Networks (3D-CNN). These types of networks are used to compute convolutions across several sequential frames, thus integrating motion information in the initial layers of the network. These types of approaches are either used by processing the video stream directly [11] or by integrating a pre-processing step that computes spatio-temporal features that are then sent to the 3D-CNN network [12]. It is also interesting to note that simple two-dimensional CNNs are also used in the literature, as they can accurately represent frame-level features, even though they lack any temporal information [9], [13], [14], while some datasets and benchmarking competitions [3] offer pre-computed convolutional features to the participants and interested parties, like fully connected layers from the popular AlexNet architecture [15].

Another important approach for video processing is represented by the popular Long short-term memory (LSTM) architectures. The authors in [16] propose an initial feature extraction process with the help of a CNN model, followed by a classification step that uses LSTM layers with convolutional gates (convLSTM). Bidirectional LSTM architectures, that have the added advantage of accessing information both in the forward and the reverse temporal directions, have been employed in [17]. Finally, the two approaches are combined by [18] in a CNN-BiLSTM approach for violence detection.

All these works represent important steps in the adoption of deep learning approaches for violence detection systems. However, little work has been done regarding the adoption of Transformer-type [19] networks in this domain. While the Transformer is a primarily language processing network, recent advances show that these networks can also be used for processing images [20] and videos [21], achieving state-of-the-art results [22]. Therefore, in this work we propose to develop a Two-Stage Transformer architecture, that uses a series of Spatial and Temporal Transformers for the detection of violent Scenes.

III. Proposed Method

A. Transformer Architecture

The proposed method is presented in Figure 1. The model, called 2SViT, can be interpreted as a Two-Stage Transformer approach, where the first stage is represented by a spatial processing transformer (ST), and the final stage is represented by a temporal processing transformer (TT). As is the norm in the current literature, the classification output is created via processing the output of the Transformer networks via a simple MLP classifier.

Similar to the ViT model [20], a set of initial pre-processing transformations must be applied on each frame in the targeted video. Therefore, given the initial image $I \in \mathbb{R}^{H\times W\times C}$, where H and W represent the height and width of the image and C represents the number of channels, the image will be flattened into sequences of total size $(N \times (P^2 \cdot C))$, where N represents the number of such sequences, and P represents the horizontal and vertical resolution of each corresponding sequence. A similar sequence is adopted for the Temporal Transformer stage of this architecture. This time, the outputs of the Spatial Transformers are sent to the input of the Temporal Transfomer, creating sequences of total size $M \times O_s$, where O_s represents the output size of the Spatial Transformer layers. Following these steps, the output of the Temporal Transformer layers O_l is processed by the MLP architecture, in order to create the final prediction.

B. Implementation Details

Several variations of this architecture are tested and validated. While in our initial experiments we tested the architecture size suggested in ViT [20] and BERT [23], namely ViT/BERT Base (12 layers), Large (24 layers) and Huge (32 layers), we soon found that the larger architectures have problems in converging. We believe this to indicate the need for larger datasets for training the system. On the other hand, we must take into account the fact that ViT is a one-stage architecture, that only processes images, and therefore under any of the three setups it will still have fewer trainable parameters than our proposed 2SViT.

Given these considerations, we adopt the following variations for the proposed architecture. For the ST stage, we will use 5, 10 and 15 layers, each with an attention structure with $N_H = 12$ heads, while for the TT stage we propose 5 and 10 layers, again each with an attention structure with $N_H = 12$ heads. For the final MLP architecture, we employ

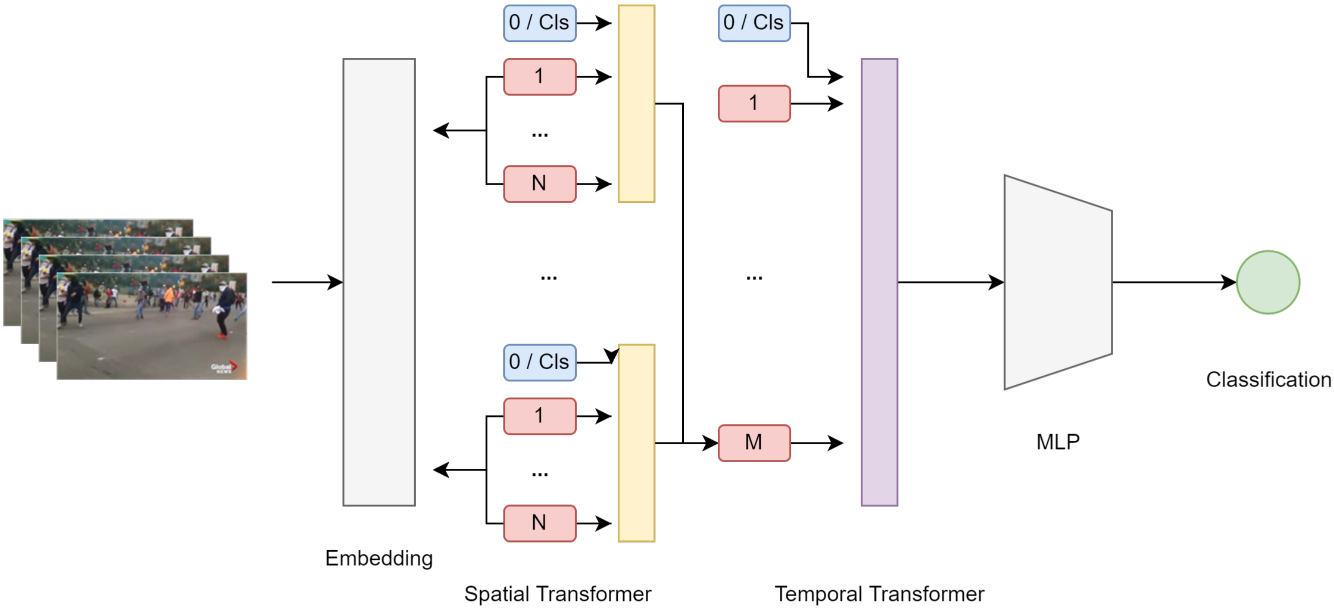

Fig. 1. The proposed Two-Stage Transformer architecture. The first stage handles the spatial processing (Spatial Transformer or ST) of the video information, while the second one encodes temporal data (Temporal Transformer or TT). The network classifies samples by processing Temporal Transformer outputs through an MLP architecture. The figure shows the architecture variations with regards to the N and M sizes of the Transformer architectures.

TABLE I
RESULTS OF THE PROPOSED ARCHITECTURE IN A 2SViT-ST5/TT5 SETUP, WITH VARYING SIZES FOR THE MLP LAYERS. LAYER I IS THE ONE THAT PROCESSES THE OUTPUT FROM THE TT LAYERS, WHILE LAYER II REPRESENTS AN INTERMEDIAT LAYER AND LAYER III REPRESENTS THE LAST LAYER BEFORE CLASSIFICATION.

Layer I	512	512	512	512	256	256	256	128
Layer II	512	256	256	128	256	256	128	128
Layer III	512	256	128	128	256	128	128	128
AP	63.11	63.18	62.84	61.58	63.29	62.77	62.42	62.28

a set of 3 fully connected layers, of varying sizes: 128, 256, 512.

The training protocol we adopt is inspired by ViT [20] and BEiT [24]. An Adam optimizer [25] is employed, with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Furthermore, a batch size of 64 segments is used in training the data, with a cosine learning schedule and minimal learning rate of $1e-6$.

IV. Experimental Results

We performed training and testing on the popular XD-Violence dataset [5]. XD-Violence contains 217 hours of visual data, corresponding to 4,754 untrimmed videos, with an almost even split of 2,405 videos containing violence and 2,349 non-violent videos. Furthermore, the training set consists of 3,954 videos, while the testing set is composed of 800 videos. Each violent scene in the videos is marked by a start and end time, and the dataset contains violence divided into six separate events: “Abuse”, “Car Accident”, “Explosion”, “Fighting”, “Riot”, and “Shooting”. Performance is measured using the official Average Precision metric (AP).

Table I presents a grid-search approach to finding the best MLP setup for classification. We vary the size of the three fully connected layers, while keeping the same setup for both the ST and TT. While the best setup we found with this experiment is a 256, 256, 256 setup, achieving an AP of 63.29, the difference between them is marginal. At most a difference of 2.7% is recorder here between the 256, 256, 256 setup and the 512, 128, 128 setup.

Following this set of experiments, we perform another grid search for the best ST and TT variations. The results of this process are described in Table II. This time we report a more significant change in results, with the lowest

TABLE II
RESULTS OF THE PROPOSED ARCHITECTURE, IN A MLP SETUP WITH 256, 256, 256 SIZED LAYERS, WHILE VARYING THE ST AND TT DIMENSIONS.

	2SViT-ST5	2SViT-ST10	2SViT-ST15
2SViT-TT5	63.29	68.15	70.49
2SViT-TT10	71.18	**77.33**	77.15

TABLE III
RESULTS ON THE XD-VIOLENCE DATASET, COMPARED WITH THE FOUR BASELINE SYSTEMS SELECTED BY THE AUTHORS OF THE DATASET, AS PRESENTED IN [5].

Method	AP
XD-Violence-Baseline [5]	50.78
OCSVM [26]	27.25
Hasan et al. [27]	30.77
Sultani et al. [28]	73.20
2SViT-ST5/TT5	63.29
2SViT-ST10/TT10	**77.33**

AP value being 63.29, achieved by the 2SViT-ST5/TT5 architecture, while the best performance is achieved by the 2SViT-ST10/TT10 architecture, with an AP of 77.33. This represents a significant increase of over 20%, indicating that the most important factor in enhancing the overall system performance is changing and adapting the Transformer architectures.

Finally, we present an analysis of the results in a larger context. Table III compares the results of our proposed system with the results of baseline methods chosen by the authors of the XD-Violence dataset. The proposed baselines are as follows: two SVM-based approaches (XD-Violence-Baseline [5] and OCSVM [26]), an autoencoder based model (Hasan et al. [27]) and a deep anomaly ranking model (Sultani et al. [28]). As the table presents, our best performing model variant, the 2SViT-ST10/TT10 surpasses these chosen baseline systems, and this would also be true for one other variant (2SViT-ST15/TT10), while two variants would score second in this comparison, after Sultani et al., namely 2SViT-ST5/TT10 and 2SViT-ST15/TT10.

It is also important to note that, while several of the baseline models like XD-Violence-Baseline take full advantage of the multimodality of the data, the fact that our visual-only model surpassed those baseline systems is in no way an indication that the audio modality does not have any useful information. More likely, this is an indication of the high performance of Vision Transfomers in general, and most likely these results may be augmented in the future, when we could take into account the audio signal as well.

V. CONCLUSIONS

In this paper we presented a violence detection Vision Transformer-based network, that uses two stages for processing videos. The first stage is dedicated to processing spatial information, representing a Spatial Transformer, while the second stage processes sequences of consecutive frames, thus representing a Temporal Transformer. The final layers of the architecture is composed of a multilayer perceptron classifier that collects data from the last layers of the Temporal Transformer and performs the classification. Training and testing were performed on the popular XD-Violence dataset, showing a performance increase of 5.64% over the baseline systems chosen by the authors of the dataset. Our experiments show that, while the composition of the final MLP classification layers may have little influence on the final score, the setup of the two Transformer stages greatly contributes to variations in the final performance of the proposed system.

Given all these, we conclude that Transformer architectures can be successfully used in the detection of violent scenes. Future avenues of research in this domain may include the addition of an audio processing branch in the system, testing a set of popular Transformer architectures, and changing the way the Spatial and Temporal components of the network interact.

ACKNOWLEDGMENT

This work has been funded by the Ministry of Research, Innovation, and Digitization UEFISCDI, projects IDENEO, "Identifying the Unusual Behavior of People in Video Streams", Solutions Axis, grant 29SOL/2021, PN-III-P2-2.1-SOL-2021-0024 and DeepVisionRomania, "Identifying People in Video Streams using Silhouette Biometrics", Solutions Axis, grant 28SOL/2021, PN-III-P2-2.1-SOL-2021-0114.

REFERENCES

[1] Muhammad Ramzan, Adnan Abid, Hikmat Ullah Khan, Shahid Mahmood Awan, Amina Ismail, Muzamil Ahmed, Mahwish Ilyas, and Ahsan Mahmood. A review on state-of-the-art violence detection techniques. *IEEE Access*, 7:107560–107575, 2019.

[2] Ling Tian, Hongyu Wang, Yimin Zhou, and Chengzong Peng. Video big data in smart city: Background construction and optimization for surveillance video processing. *Future Generation Computer Systems*, 86:1371–1382, 2018.

[3] Mihai Gabriel Constantin, Liviu Daniel Stefan, Bogdan Ionescu, Claire-Hélene Demarty, Mats Sjoberg, Markus Schedl, and Guillaume Gravier. Affect in multimedia: Benchmarking violent scenes detection. *IEEE Transactions on Affective Computing*, 2020.

[4] Claire-Hélène Demarty, Cédric Penet, Markus Schedl, Ionescu Bogdan, Vu Lam Quang, and Yu-Gang Jiang. The mediaeval 2013 affect task: violent scenes detection. In *MediaEval 2013 Working Notes*, page 2, 2013.

[5] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European Conference on Computer Vision*, pages 322–339. Springer, 2020.

[6] Tobias Senst, Volker Eiselein, Alexander Kuhn, and Thomas Sikora. Crowd violence detection using global motion-compensated lagrangian features and scale-sensitive video-level representation. *IEEE transactions on information forensics and security*, 12(12):2945–2956, 2017.

[7] Sarita Chaudhary, Mohd Aamir Khan, and Charul Bhatnagar. Multiple anomalous activity detection in videos. *Procedia Computer Science*, 125:336–345, 2018.

[8] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[9] Zuxuan Wu, Yu-Gang Jiang, Jun Wang, Jian Pu, and Xiangyang Xue. Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 167–176, 2014.

[10] Zihan Meng, Jiabin Yuan, and Zhen Li. Trajectory-pooled deep convolutional networks for violence detection in videos. In *International Conference on Computer Vision Systems*, pages 437–447. Springer, 2017.

[11] Chunhui Ding, Shouke Fan, Ming Zhu, Weiguo Feng, and Baozhi Jia. Violence detection in video by using 3d convolutional neural networks. In *International Symposium on Visual Computing*, pages 551–558. Springer, 2014.

[12] Fath U Min Ullah, Amin Ullah, Khan Muhammad, Ijaz Ul Haq, and Sung Wook Baik. Violence detection using spatiotemporal features with 3d convolutional neural network. *Sensors*, 19(11):2472, 2019.

[13] Guankun Mu, Haibing Cao, and Qin Jin. Violent scene detection using convolutional neural networks and deep audio features. In *Chinese Conference on Pattern Recognition*, pages 451–463. Springer, 2016.

[14] Hao Ye, Zuxuan Wu, Rui-Wei Zhao, Xi Wang, Yu-Gang Jiang, and Xiangyang Xue. Evaluating two-stream cnn for video classification. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 435–442, 2015.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[16] Swathikiran Sudhakaran and Oswald Lanz. 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)*, pages 1–6. IEEE, 2017.

[17] E Fenil, Gunasekaran Manogaran, GN Vivekananda, T Thanjaivadivel, S Jeeva, A Ahilan, et al. Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional lstm. *Computer Networks*, 151:191–200, 2019.

[18] Rohit Halder and Rajdeep Chatterjee. Cnn-bilstm model for violence detection in smart surveillance. *SN Computer science*, 1(4):1–9, 2020.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[21] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8741–8750, 2021.

[22] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, 2021.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[24] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[26] Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. *Advances in neural information processing systems*, 12, 1999.

[27] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016.

[28] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018.