Anomaly Detection

Dinesh Nariani AU1920128

Henil Shah AU1940205

Devanshu Magiawala AU1940190

Abstract — The detection of abnormal occurrences is critical in video content analysis and is a difficult task when monitoring surveillance fields. Surveillance cameras can capture a wide range of actual irregularities. The paper highlights a learning algorithm that trains on both normal and anomalous videos to learn to detect anomalies. To avoid spending time annotating anomalous segments or clips in training films, the proposed learning anomaly uses the deep multiple instance ranking framework and weakly labeled training movies, where the training labels (anomalous or normal) are at the video level rather than the clip level. In our approach, we use multiple instance learning (MIL) to automatically develop a deep anomaly ranking model that predicts high anomaly scores for anomalous video segments by treating normal and anomalous movies as bags and video segments as instances.

Keywords — Anomaly event detection, Multiple instance learning Framework, C3d Model, 3D convolutions, activity recognition, feature extraction, video surveillance, computer vision.

I. Introduction

The use of cameras for surveillance has been increasing with time at public places to ensure safety. One of the main tasks for video surveillance is the detection of anomalies such as crimes, illegal activities, fire etc. These types of anomalous events rarely occur as compared with normal activities. So, in order to slice off the waste of time and labor, developing and deploying an intelligent computer vision algorithm for anomaly detection through video surveillance is in great need. So, the primary target for these anomaly detection systems is to identify and detect anomalies and generate a signal which is different from the signal generated through normal patterns. All these real world anomalies happen for a short duration and are complicated, due to which it is not possible to list down all possible anomalous events. So, it is necessary and appropriate that anomaly detection algorithms do not depend on any prior information about the events. Also, in these types of algorithms, as the situation captured by the cameras can change over time, these could lead to high false positives for different normal behaviors.

II. Literature Survey

Abnormal detection is one of the most difficult and time-consuming tasks in computer vision. A number of attempts have been made in surveillance systems to detect abnormal activity.[1] Yihao Zhang et al.proposes anomaly detection in traffic video with the information provided in the HEVC compressed domain.In High Efficiency Video Coding (HEVC).[2] Tian Wanga et al. propose event detection based on moment feature descriptor and classification. The feature descriptor extracts the optical flow and computes the histogram of optical flow orientations(HOFO).The hidden Markov model(HMM) is proposed to classify the events due to the probabilistic property. HOFO feature descriptor is based on the movement and if the index of the HOFO is at a low level.[3]

Fan Jianget et al. propose a multi-sample-based similarity measure in which HMM training and distance measurement are performed on many samples. A unique dynamic hierarchical clustering (DHC) algorithm is used to collect these numerous training data sets.[4] Xun Tanget et al. offer a method for recognising anomalous occurrences in

crowded settings based on sparsely coded motion attention.[5]

III. Dataset

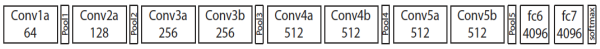
The dataset used for the proposed approach is the anomaly detection dataset from the University of Central Florida (UCF). The dataset contains a total of 1692 surveillance films, with 900 anomalous videos and 792 normal videos. Arrest, Arson, Assault, Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism are among the 13 real-world abnormalities and anomalous incidents covered in the videos. We are considering 5 videos for each anomaly, 65 videos for normal videos, and 10 videos from each abnormality for testing purposes for our project. We are using multiple instance learning to annotate our dataset..

IV. Multiple Instance Learning

MIL eliminates the requirement for correct temporal annotations. The specific temporal position of anomalous events in videos is uncertain in MIL. Only video-level labels identifying the presence of an anomaly in the entire video are required. A video with anomalies is labeled as positive, while a video with no anomalies is labeled as negative. Then, we represent a positive video as a positive bag B_a, with different temporal segments forming individual instances in the bag (p^1, p^2,..., p^m), where m is the number of instances in the bag. We presume that the anomaly exists in at least one of these circumstances. Similarly, the negative video is represented by a negative bag B_n, in which temporal segments generate negative instances (n^1, n^2, . . . , n^m).

V. C3d Model

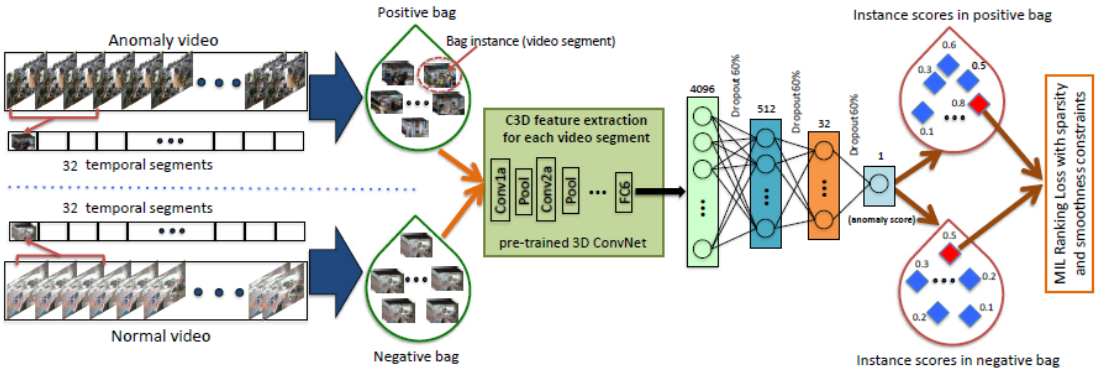
C3D convolutional neural networks are deep 3-dimensional convolutional neural networks with a homogeneous design consisting of 3 x 3 x 3 convolutional kernels followed by 2 x 2 x 2 pooling at each layer. They've been trained on large-scale supervised video datasets including UCF-101 and Sports 1M.



The first convolution layer of size 1x3x3 is depicted in the figure, followed by a pooling layer of size 1x2x2. This is done to maintain temporal information in the first layer and generate higher level temporal representations in later levels of the network. Every other convolution layer and pooling layer would be 3x3x3 and 2x2x2, respectively, with strides of 1 and 2. The fully connected layers have 4096 dimensions and softmax outputs that represent either 101 classes from the UCF-101 dataset or 487 classes from the Sports 1M dataset. For our project, we will look at the layer fc6 output, which is a 4096 feature vector with 16 frames.

VI. Implementation

We extract visual features from the C3D network's fully connected (FC) layer FC6. Before we compute features, we resize each video frame to 240 x 320 pixels and set the frame rate to 30 frames per second. After l2 normalization, we compute C3D characteristics for each 16-frame video clip. To get features for a video segment, we average all 16-frame clip features inside that segment. These characteristics (4096D) were fed into a 3-layer FC neural network. The first FC layer comprises 512 units, then 32 units, and finally one unit. Between FC layers, dropout regularization of 60% is utilized. We use ReLU activation and Sigmoid activation for the first and the last FC layers respectively, and employ an Adagrad optimizer with the initial learning rate of 0.001. The parameters of sparsity and smoothness constraint in the MIL ranking loss are set to $\lambda_1=\lambda_2 = 8\times10^{-5}$ and $\lambda_3 = 0.01$ for the best performance.



VII. Loss function

This is a regression problem, according to us. As a result, our ranking loss in the hinge-loss formulation is as follows:

$$l(\mathcal{B}_a, \mathcal{B}_n) = \max(0, 1 - \max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) + \max_{i \in \mathcal{B}_n} f(\mathcal{V}_n^i)).$$

One limitation of the above loss is that it ignores the abnormal video's underlying temporal structure. First, in real-world circumstances, abnormality frequently occurs for a short period of time. In this circumstance, the scores of the instances (segments) in the anomalous bag should be sparse, indicating that the anomaly may be contained in only a few segments. Second, because the video is a series of segments, the anomaly score should vary gradually between them. As a result, by minimizing the difference in scores for adjacent video segments, we impose temporal smoothness between anomaly scores of temporally adjacent video segments. The loss function is transformed by including the sparsity and smoothness restrictions on the instance scores.

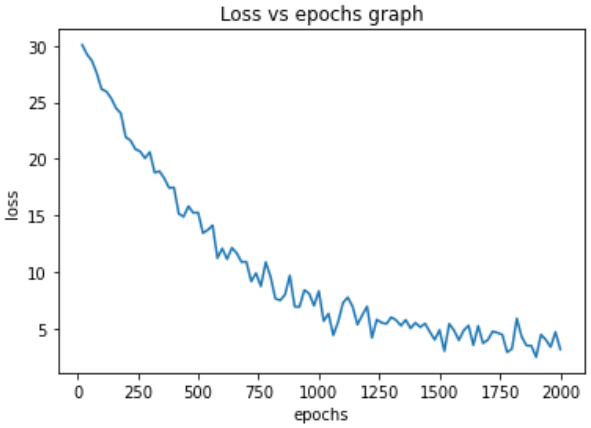
$$l(\mathcal{B}_a, \mathcal{B}_n) = \max(0, 1 - \max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) + \max_{i \in \mathcal{B}_n} f(\mathcal{V}_n^i))$$

$$+\lambda_1 \overbrace{\sum_{i}^{(n-1)} (f(\mathcal{V}_a^i) - f(\mathcal{V}_a^{i+1}))^2}^{①} + \lambda_2 \overbrace{\sum_{i}^{n} f(\mathcal{V}_a^i)}^{②},$$

where term 1 indicates temporal smoothness and term 2 indicates sparsity.

VIII. Results

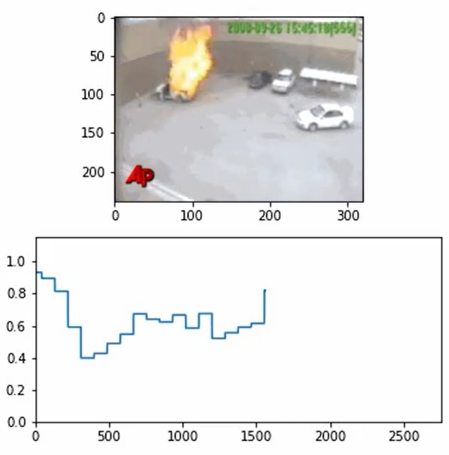
We have trained our model for 2000 iterations, batch size is 60, learning rate is 0.01 and we have got the sum of hinge-loss, sparsity loss and smoothness loss which is 5.38.



We generated some results on testing videos. Due to consideration of small sample size, we got some false negative results in anomalies while testing manually like abuse, arrest, assault, road accidents, robbery, shooting, shoplifting. While we are able to generate decent results with other anomalies.It is important to note that, despite the absence of segment-level annotations, the network can estimate the temporal position of an anomaly based on anomaly scores.As the iterations increases and the network sees more videos of the dataset, it will automatically learns to precisely localize anomaly.

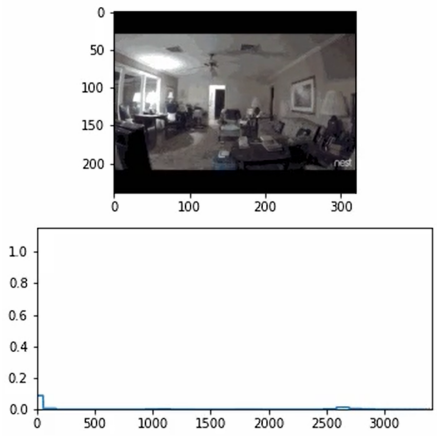
True Positive: -

Explosion Anomaly



False Negative: -

Abuse Anomaly



IX. Conclusion

Based on the model that is trained on the dataset that is taken, it is able to detect anomalies such as explosion, arson, vandalism, stealing, fighting etc. This model is still lagging in terms of detecting other anomalies such as abuse, arrest, shoplifting, assault, road accidents, robbery, shooting. By increasing the sample size and hyperparameter tuning, improvisation of results will be done.

X. References

1. S. Mohammadi, A. Perina, H. Kiani, and M. Vittorio.Angry Crowds: Detecting violent events in videos. In ECCV, 2016.
2. Yihao Zhang, Hongyang Chao.Abnormal Event Detection in Surveillance Video: ACompressed Domain Approach for HEVC, 2017.
3. Tian Wanga, MeinaQiaoa, Yingjun Dengb, Yi Zhouc, HuarenWang, Qi Yuan, and Hichem Snoussi.Abnormal Event Detection based on Analysis of Movement Information of Video Sequence, 2017.
4. Fan Jiang, Ying Wu, Aggelos K. Katsaggelos.Abnormal Event Detection From Surveillance Video By Dynamic Hierarchical Clustering, 2007.
5. Xun Tang, Shengping Zhang, Hongxun Yao.Sparse Coding Based Motion Attention For Abnormal Event Detection,2013.
6. Sultani, W. and Shah, M., 2022. Real-World Anomaly Detection in Surveillance Videos. [online] Openaccess.thecvf.com. Available at: <https://openaccess.thecvf.com/content_cvpr_2018/html/Sultani_Real-World_Anomaly_Detection_CVPR_2018_paper.html> [Accessed 20 March 2022].
7. Medium. 2022. *Deep Dive into Convolutional 3D features for action and activity recognition (C3D)*. [online] Available at: <https://medium.com/@nair.binum/quick-overview-of-convolutional-3d-features-for-action-and-activity-recognition-c3d-138f96d58d8f> [Accessed 20 March 2022].
8. Medium. 2022. *Deep Dive into Convolutional 3D features for action and activity recognition (C3D)*. [online] Available at: <https://medium.com/@nair.binum/quick-overview-of-convolutional-3d-features-for-action-and-activity-recognition-c3d-138f96d58d8f> [Accessed 20 March 2022].