Anomaly Detection in Surveillance Videos

Introduction

In order to improve public safety, surveillance cameras are increasingly used in public places, such as streets, intersections, shopping centers, etc. However, the monitoring and recognition of unsafe conditions has not been improved. As a result, there are obvious shortcomings in the use of surveillance cameras, and there are far more cameras than human monitors.

A key task of video surveillance is to detect anomalous events, such as illegal activities, traffic accidents, explosions and other abnormal accidents. Compared to normal events, anomalous events rarely occur. Therefore, there is an urgent need to develop intelligent computer vision algorithms for automatic video anomaly detection so as to reduce the waste of labor and time.

In practice, the purpose of the anomaly detection system is to signal in time when activities that deviate from the normal pattern are found, and to identify the time window in which the anomaly occurs. Therefore, anomaly detection can be regarded as a rough video understanding, and anomalies can be filtered out from the normal mode. When an anomalous event is detected, classification technology can be used to further classify it as one of the specific activities.

At present, some algorithms have been developed to detect specific anomalous events, such as traffic accident detectors. However, this kind of solution is difficult to generalize to other anomalous events detection. So in practice, this kind of use is limited.

In the real world, it is difficult to list all abnormal situations, because anomalous events are generally more complex and diverse. Therefore, anomaly detection should be performed with minimal supervision.

Motivation and contributions. There are many approaches that based on the assumption that any pattern that deviates from the learned normal patterns would be considered as an anomaly. However, this assumption may not hold true because it is very difficult to define a normal event which takes all possible normal patterns/behaviors into account. And the boundaries between normal and anomalous events are generally blurred.

We propose an anomaly detection algorithm using weakly labeled training videos. We only know the video-level labels, i.e. a video is normal or contains anomaly somewhere, but we do not know where. This is very efficient because we can easily annotate large numbers of videos just by assigning video-level labels. To formulate a weakly-supervised learning approach and to resolve the difficulty that the videos are long untrimmed surveillance videos with very large intra-class variance, we resort to the ResNet and LSTM solution. And there are several contributions:

• We propose a ResNet and LSTM solution to detect anomaly in surveillance videos. The CRNN model can utilize both the discriminative feature maps provided by the ResNet model and find the key frames for classification through the LSTM model.

• We use a large-scale video anomaly detection dataset consisting of 1900 real-world surveillance videos of 13 different anomalous events and normal activities captured by surveillance cameras.

• Experimental results on our dataset show that our proposed method achieves better performance compared to the state-of-the-art anomaly detection approaches.

• Our dataset serves a challenging benchmark for activity recognition on untrimmed videos, due to the complexity of activities and large intra-class variance. We provide results of baseline methods, C3D[4] and TCNN[5], on recognizing 13 different anomalous activities.

Related Work

Anomaly detection in video surveillance is one of the most challenging and long standing problems in computer vision. There are several attempts to detect anomalous events. Sultani *et al.*[1] used deep MIL framework with weakly labeled data to detect real-world anomalies in surveillance videos. This method is a MIL ranking loss with sparsity and smoothness constraints for a deep learning network to learn anomaly scores for video segments. Tran *et al.*[4] proposed an approach for spatiotemporal feature learning using deep 3-dimensional convolutional networks (3D ConvNets) trained on a large scale supervised video dataset. Hou *et al.*[5] proposed an end-to-end deep network called Tube Convolutional Neural Network (T-CNN) for action detection in videos. Rashmika *et al.*[7] proposed the incremental spatiotemporal learner (ISTL) to address challenges and limitations of anomaly detection and localization for real-time video surveillance. However, there exists a difficulty that the videos are long untrimmed surveillance videos with very large intra-class variance. We propose the ResNet-152 and LSTM solution, and the LSTM module in our model can help resolve this difficulty.

Approach

We use two approaches to detect anomaly in surveillance videos. One approach is 3D CNN, which is based on the existing paper[4]. We realize 3D CNN to use it as a baseline for our proposed approach. The second approach is the one we propose to solve the problem, that is, to use CRNN model to detect anomaly in surveillance videos.

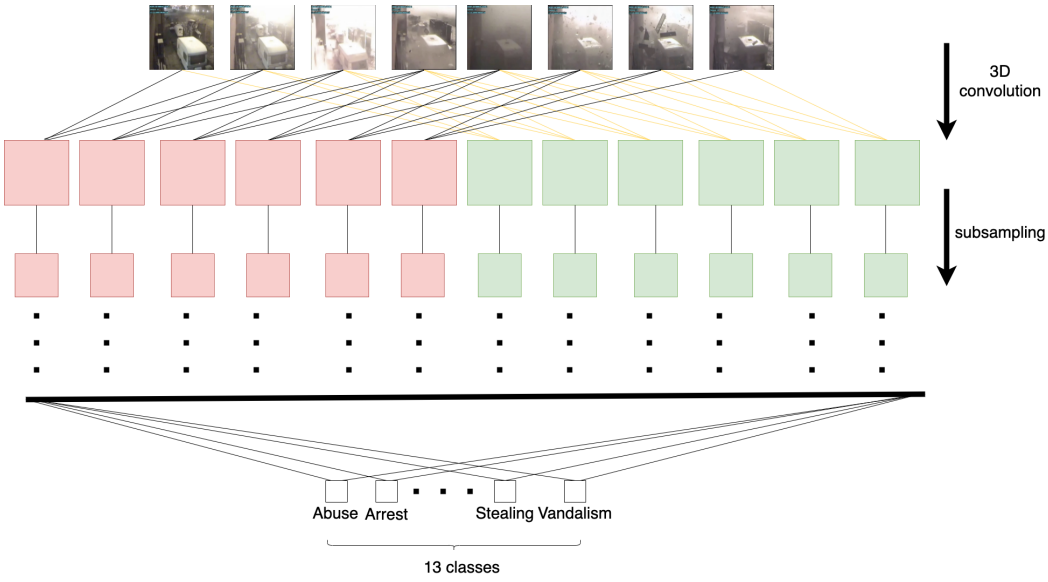
3D CNN

At present, many classifiers are using the features manually extracted from the original image, CNN is mainly used for dealing with 2D images. An easy way to think of is to treat each frame of video as a still image and use CNN to identify the anomaly event of a single frame. However, this method does not consider the encoded motion information of multiple consecutive frames.

In order to effectively combine the motion information in the video, we use 3D convolution which can be performed in CNN convolution layer to capture the distinguishing features along the spatial and temporal dimensions. 3D CNN model can generate multiple information channels from adjacent video frames, and perform convolution and down sampling in each channel respectively. 3D convolution consists of stacking consecutive frames into a cube and then using the convolution kernel inside the cube. With this structure, the feature maps in the convolution layer are connected to multiple adjacent frames in the previous layer to capture motion information. The value at a location of a feature map is locally perceived by convoluting the same location of a constant number of consecutive frames on the previous layer.

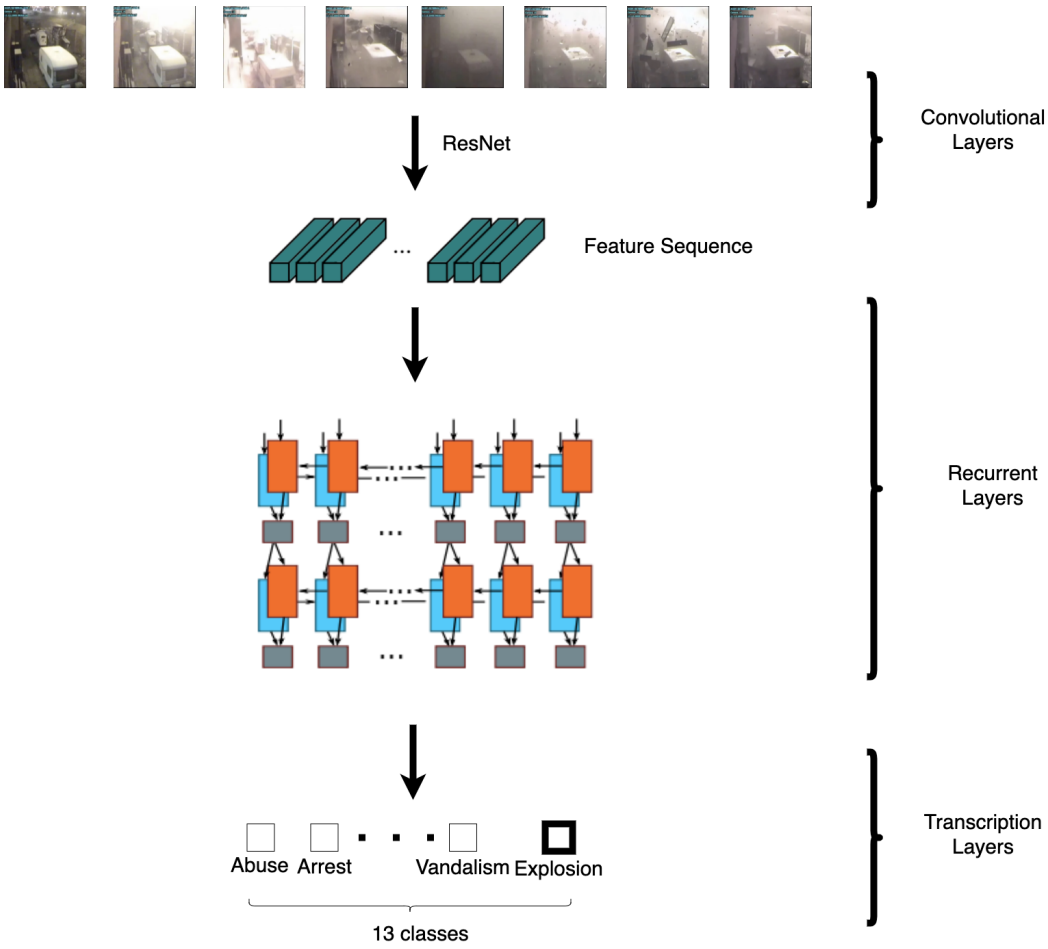
It is important to note that 3D convolution kernels can only extract one type of feature from cube, because the weights of the convolution cores are shared throughout the convolution process, so we present a variety of convolution kernels to extract a variety of features. Follow this intuition, the 3D CNN model consists of 2 convolution layers, 1 down-sampling layer and 3 fully connected layers. The cube for each 3D convolution kernel includes 28 consecutive frames, and the single-framed patch size is 240×320. The input of the 3D CNN model is limited to a small number of consecutive video frames (28 frames in our project), because as the size of the input window increases, the parameters that the model needs to train will also increase.

In short, 3D CNN can directly extract features from the original input, from the temporal and spatial dimensions in the video by performing 3D convolution.



CRNN

We propose to use CRNN model to detect anomaly in surveillance videos. This model is a pair of CNN encoder and RNN decoder. The convolution layer here is the famous ResNet-152 network for extracting convolutional feature maps of input images, converting an image of size 224×224 to a convolution feature vector of size 512. It encodes (meaning compressing dimension) every 2D image $x(t)$ into a 1D vector $z(t)$ by $f_{CNN}(x^{(t)} = z^{(t)})$. For the RNN decoder, we use a deep bi-directional LSTM[6] RNN to continue extracting sequence features based on convolution features, and softmax the output equal to the probability of each possible value. It receives a sequence input vectors $z(t)$ from the CNN encoder and outputs one dimensional sequence $h(t)$.



Implementation Details

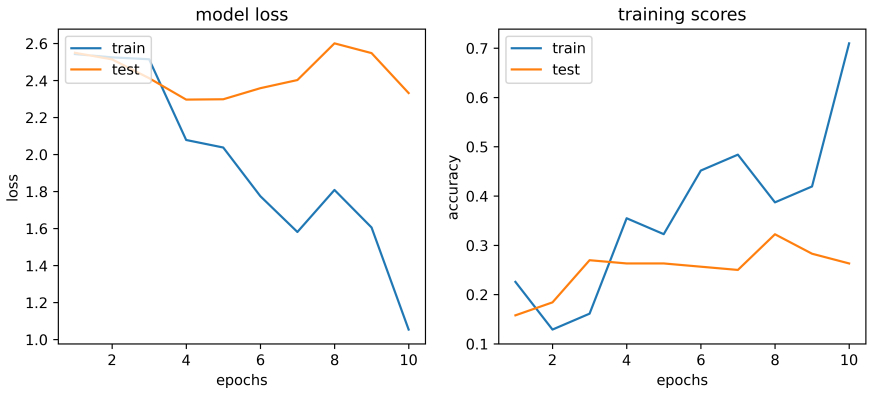
The experiments are conducted 10 runs. In each run, we randomly select 80% of the data from each category for training and the rest for testing. We report the average classification accuracy of the 10 runs. The classification accuracy on the test set is evaluated. Our model is based on the state-of-the-art CNN architecture ResNet-152, which are initialized with the weights pre-trained on ImageNet[2]. In addition, for the input video, we sample 60 successive frames with fps of 1. We resize each frame of the visual sample to the ResNet image size of 224×224 pixels. In our training, Adam[3] is adopted to automatically adjust the learning rate during optimization, with the initial learning rate set to 0.0002 and the model is trained with batch-size 32 for 150 epochs.

About the hardware, we use four P100 and train the models on Google Cloud Platform. The main framework we use is PyTorch, and the related software libraries and packages we use include OpenCV, Numpy, Matplotlib, Scikit-learn, etc. We use a large-scale data set --- UCF crime data set. It consists of 1900 long, untrimmed real-world surveillance videos with 13 real anomalies, including abuse, arrest, arson, attack, traffic accident, burglary, explosion, fight, robbery, shooting, theft, shop theft and vandalism. The total duration is 128 hours. About functionalities, we have trained a model that can detect anomalies in surveillance video, which can assist manual screening. And when talking about the limitations, there are two limitations, one is that the screening accuracy may not be high when considering the accuracy of the model, which affects the efficiency, and the other is that there may be false positives due to the recall rate.

Experiments and Observations

In this section, we evaluate the proposed CRNN model. We provide two state-of-the-art baselines for video classification on our dataset based on 5-fold cross validation. For the first baseline, we construct a 4096-D feature vector by averaging C3D[4] features from each 16-frames clip followed by an L2-normalization. The feature vector is used as an input to a nearest neighbor classifier. The second baseline is the Tube Convolutional Neural Network (TCNN)[5] which introduces the tube of interest (ToI) pooling layer to replace the 5-th and 3d-max-pooling layer in C3D pipeline. The ToI pooling layer aggregates features from all clips and outputs one feature vector for a whole video. The accuracy is given in the following Table. From the results, we have the following observation: these state-of-the-art video classification methods perform poor on this dataset. This is because the videos are long untrimmed surveillance videos with very large intra-class variance, and the LSTM module in our model can help resolve that difficulty.

METHOD	ACCURACY
C3D	23.0
TCNN	28.4
CRNN (ours)	**35.5**



Conclusion

In this project, we have proposed an effective video anomaly detection method in surveillance videos based on ResNet and LSTM. The CRNN model can utilize both the discriminative feature maps provided by the ResNet model and find the key frames for classification through the LSTM model. The extensive experiments demonstrate that the CRNN model achieves 7.1% performance improvements as compared to the best state-of-the-art video classification approach. In future studies, we plan to extend the CRNN model to investigate attention mechanism that can better concentrate on the key frames in video segments.

Bibliography

[1] Sultani, Waqas, C. Chen and M. Shah. "Real-World Anomaly Detection in Surveillance Videos", 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018): 6479-6488.

[2] Deng, Jia, W. Dong, R. Socher, L. Li, K. Li and Li Fei-Fei. "ImageNet: A large-scale hierarchical image database", CVPR (2009).

[3] Kingma, Diederik P. and Jimmy Ba. "Adam: A Method for Stochastic Optimization", CoRR abs/1412.6980 (2015): n. pag.

[4] Tran, Du, Lubomir D. Bourdev, R. Fergus, L. Torresani and Manohar Paluri. "Learning Spatiotemporal Features with 3D Convolutional Networks", 2015 IEEE International Conference on Computer Vision (ICCV) (2015): 4489-4497.

[5] Hou, Rui, C. Chen and M. Shah. "Tube Convolutional Neural Network (T-CNN) for Action Detection in Videos", 2017 IEEE International Conference on Computer Vision (ICCV) (2017): 5823-5832.

[6] Hochreiter, S. and J. Schmidhuber. "Long Short-Term Memory", Neural Computation 9 (1997): 1735-1780.

[7] Rashmika Nawaratne, Damminda Alahakoon, Daswin De Silva, and Xinghuo Yu. "Spatiotemporal Anomaly Detection Using Deep Learning for Real-Time Video Surveillance", IEEE Trans. Ind. Informat., vol. 16, no. 1, pp. 393-402, Jan. 2020.