Understanding AI Video Surveillance: What Makes Anomaly Detection So Difficult?

Elena Ramlow
Analytics, Georgetown University

Abstract

Anomaly detection can be classified as both an object detection and action recognition problem, as violence is recognized both by interaction and objects present. Video analysis and action recognition are at the forefront of the field of artificial intelligence. However, they are the most difficult to find information on and to comprehend. This project explores the data processing and model building that goes into video object detection and recognition and sheds light on the mechanisms that enterprise "AI Video Surveillance" is built on. The goal is to better understand AI applications in video surveillance and evaluate their abilities.

1 Introduction

With artificial intelligence continuing to be at the forefront of data science, select applications have dominated the trend. Object detection and image related neural networks are extremely popular and easy to reproduce. However, the counterpart in video analysis has not been as forthcoming.

One specific potential benefit to using AI on video data would be anomaly detection in video surveillance, which could automate the process of threat detection in video and expedite the response, including intervention and contacting emergency services. However, models able to successfully detect anomaly in video surveillance are not publicly available for analysis and replication.

Currently marketed AI video surveillance is only available in the enterprise realm, where the details of the models are not disclosed. However, the potential of these models cannot be thoroughly assessed without an understanding of what goes into object detection and action recognition neural network models and how they can be used in video surveillance.

To date there are few open-source examples neural networks trained on video data and even fewer dealing with anomaly detection. Whereas for most data science problems there exist a multitude of tutorials and example code, for anomaly detection in video surveillance the general audience is mainly left in the dark, while those who employ models in the professional sector keep their state-of-the-art methodology as proprietary technology.

This project seeks to explore the realistic applicability of machine learning models to anomaly detection in video surveillance and assess two state of the art object detection models in their application to annotated frames of video data.

2 Background

Researchers at the University of Central Florida Center for Research in Computer Vision have gathered two large video datasets for applications in machine learning. The first is UCF101, a dataset of human actions which contains over 27 hours of video data depicting 101 human actions (Soomro, Zamir, and Shah, 2012). This dataset can be used in action recognition models for differentiation between the action classes. While resulting models using UCF101 are interesting in their ability for baseline action recognition, the real-world applications for simply distinguishing between a variety of human behavior are limited. However, it is important to note as it represents one of the first large-scale video datasets available.

The second dataset compiled by the UCF CRCV has much greater implications for applying machine learning to real-word issues. This dataset, UCF-Crime, contains 1900

untrimmed surveillance videos depicting 13 anomalies. The original use of the dataset was to make video level classification of anomalous behavior, including both whether an anomaly occurred at any point in the video and what type of anomaly was depicted (Sultani, Chen, and Shah, 2018). In annotating the dataset for the project at hand one key issue arose—many of the videos depicting anomalous behavior were uploaded to YouTube from a site known as "Live Leak" with the site's logo in the corner of the video. Live Leak is well known for its horrific content depicting all types of disturbing behavior and incidents. The issue in implementing a model at the video level is that the presence of this logo correlates with anomalous behavior, introducing a confounding variable.

As for object detection and action recognition in videos, currently employed methods can perform only low-level functions, relying on pre-defined events and behaviors, without any machine learning (Ko, 2011). Researchers have begun experimenting with applying machine learning techniques to video surveillance, but the existing publicly available research, code and models is lacking.

A key issue in translating object detection in images to video is the transfer from two dimensions to three dimensions—incorporating the spatiotemporal features of video, which are paramount to action recognition and defining behavior. Two papers, Hou, Chen, Sukthankar, and Shah (2019) and Ji, Xu, Yang, and Yu (2012) introduce a 3D CNN that performs video action segmentation, which perform very well but require complex and computationally taxing models. Similarly, Veeriah, Zhuang, and QI (2015) introduce an RNN that utilizes LSTM applied to three-dimensional data. What all available research suffers from is a lack of computational efficiency and the need for highly complex models that are difficult to understand. Furthermore, given their complexity there is little room for reproduction.

3 Method

The UCF Crime dataset, of 1900 untrimmed real world surveillance videos, is downloaded and analyzed at the frame level, versus at the video level as in Sultani et al. (2019). The videos are cleaned, annotated, and fed into two object detection models. The object detection models are retrained using transfer learning and performance is measured by the precision of predicting Pascal VOC boxes onto frames.

3.1 Data

The video dataset is downloaded from the UCF Crime Dataset Dropbox repository. All 1900 videos are downloaded and cleaned. Cleaning is done in Python to resize videos to 240 x 360 pixels and set the frames per second at 30. Since videos must be annotated by hand, only 48 videos classified as "Abuse" are used.

The 48 videos are cleaned, processed, and frames are extracted. To simulate the real-world application issue of variability in anomaly only 25% of the data is used in training, while the other 75% is held out for testing. While this will likely decrease the accuracy and ability of the models, it is a better representation of the generalizability and application of anomaly detection models in video, when anomalies contain high variability and low consistency across occurrences.

To annotate the videos, before splitting them into individual frames, the open source tool CVAT is used. Bounding boxes are drawn on key frames within each video using interpolation to ensure that the path of the anomalous behavior is correctly followed. An example of how the CVAT tool works is depicted below in Figure 1.

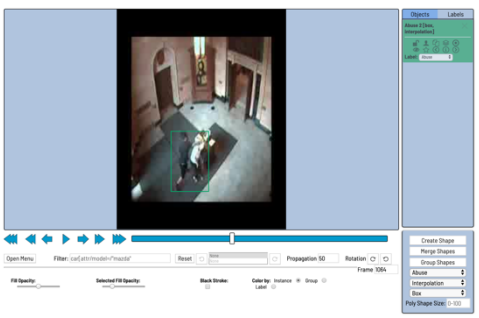

Figure 1. Annotation using CVAT

3.2 Models

Two pretrained models, a Faster RCNN and an SSD, from the Tensorflow detection zoo are used for transfer learning. This means that the final layers of the pretrained models are retrained on the new video data and given a new class to detect—the Abuse class as defined in the dataset.

Both models rely on Inception v2 feature extraction and are trained on the COCO dataset. This should provide the most consistency across models and allow for better comparison between the structures.

The Faster RCNN (recurrent convolutional neural network) model starts by feeding the image into a convolutional neural network to produce a feature map. The activation map is then run through a regional proposal network that outputs important regions within the image. Features from these regions are extracted for the proposals using RoI pooling, which are then fed into fully connected layers to produce an output path and predicted bounding box coordinates.

In comparison, the SSD (single shot detector) model combines the region proposals and regional classifications to simultaneously predict the bounding box and class as the image is processed. Essentially, the model considers every bounding box in every location to speed up the processing time. In comparison to the Faster RCNN, the SSD model performs much faster but with less accuracy.

3.3 Metrics

The models are evaluated on training, validation, and testing data based on the average precision of Pascal VOC box predictions on each frame. For validation and testing 500 images are used and for the training precision 1,000 are used. Frames with the predicted bounding boxes are also plotted to provide visual analysis, although the actual precision is not apparent in the visual.

4 Results

Neither of the models perform well on the task. On the 1000 training examples, the SSD average precision was 0.085 and the Faster RCNN was 0.056. The average validation precision, on 500 frames, was 0.095 for the Faster RCNN and 0.068 for the SSD. Average testing precision, again on 500 frames, was 0.012 for the SSD and 0.007 for the Faster RCNN. Overall, the SSD was more accurate than the Faster RCNN, but both models performed poorly.

Figure 3. Faster RCNN

Figure 4. SSD

As an example, the same frame that is depicted above for annotation with CVAT is fed into each model for prediction. Visually, the Faster RCNN prediction box appears near where the anomaly is occurring in the frame. Interestingly, the box is in the direction that the offender both enters and exits the frame from. The SSD fails to predict the presence of Abuse in the frame.

Interestingly, the SSD took longer to train, with loss holding constant at approximately 5 after 10,000 steps, while the Faster RCNN trained quickly, in about 800 steps. The SSD model was designed to increase speed at the cost of accuracy, yet the results of the two models contradict that. Likely, this is due to the dataset being poorly formatted for the SSD model and overall difficult to train.

5 Discussion

The object detection models did not perform well on identifying and locating "Abuse" within frames of a video. This is likely due to the inability of object detection models to

associate frames and track actions from one frame to the next. The movement and interaction of people are important features in recognizing violence and anomalies. Action recognition models making use of segmentation would be better suited.

The other main issue of the project was the extremely limited amount of training data used. While this certainly reflects the difficulty in applying a trained model to the real-world scenario where anomalous situations are not consistent and often appear extremely different from one another, it does result in a very poorly performing model. In order to improve the model to the point of usability many more videos would have to be annotated by hand and normalized by averaging across multiple annotators. Annotating the videos by hand is extremely time consuming and given the difficult nature of the videos can cause distress. The ethics of having to watch violent and alarming content for annotating videos to be used in a machine learning model should be considered and emphasized.

This project would be improved by applying the same transfer learning techniques to action recognition models, which would require different annotation and formatting of the video data. If the quantity of training data was increased, the annotating procedure improved, and the models employed expanded, results could show how machine learning is able to accurately classify anomalous behavior. However, in its current state the project fails to meet that go, instead highlighting the challenge in attempting to do so.

6 Conclusion

Modern techniques for applying object detection and action recognition to video surveillance rely on complex models that are difficult to comprehend and reproduce. While the lack of open source projects relating to video surveillance in comparison to available enterprise models that claim to successfully applying AI to video surveillance may suggest an overstating of enterprise abilities, the difficulty and complexity of the models is likely at fault.

In order for accurate video anomaly detection models to be produced and publicized by researchers, large datasets of annotated videos must be constructed with protocols in place for ensuring objective annotation and some method of peer-review or annotation averaging to reduce variability.

As the field of artificial intelligence develops, anomaly detection in video surveillance may become the next popular topic for online tutorials and reproduction of models. However, in its current state models are too complex and difficult to train and the available data too limited for any widescale increase in available open-source modelling. The seemingly "black box" nature of using artificial intelligence in video applications stems from the complexity of three-dimensional data and the increased computational needs that come from introducing spatiotemporal elements.

References

Hou, R., Chen, C., Sukthankar, R., & Shah, M. (2019). An efficient 3D CNN for action/object segmentation. *ArXiv*.

Ji, S., Xu, W., Yang, M., & Yu, K. (2012). 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 221-231. doi: 10.1109/TPAMI.2012.59

Ko, T. (2011). "A Survey on Behavior Analysis in Video Surveillance Applications" Ch. 16 of Video Surveillance.

Sekachev, B., Manovich, N., & Zhavoronkov, A. (2019). Computer Vision Annotation Tool. *Zenodo*. doi: 10.5281/zenodo.3497106

Soomro, K., Roshan Zamir, A., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. *ArXiv.*

Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. *The IEEE Conference on Computer Vision and Pattern Recognition.* 6479-6488.

Veeriah, V., Zhuang, N., & Qi, G.J. (2015). Differential recurrent neural networks for action recognition. *THE IEEE International Conference on Computer Vision*, 4041-4049.