Activity Recognition: Honours Project Brief

Thomas Scholtz, Stellenbosch University

February 2021

Contents

1 Introduction

The following document outlines methodology and architectures developed to recognise anomalous activity in CCTV video footage. The content will be presented in the form of short summarized versions of three research papers, namely: Learning Memory-guided normality for Anomaly Detection, Real-world Anomaly Detection in Surveillance Videos and Anomaly Detection in Video Sequence with Appearance-Motion Correspondence. The summaries will provide a broad overview of the objective of the paper and the components implemented in order to achieve the objective. This will be followed by a section mentioning some areas where there is potential for improvement of existing solutions or opportunities for further work to be done upon the solutions. Finally, the reimplementation or application of one of the three papers is discussed.

2 Paper Summaries

2.1 Anomaly Detection in Video Sequence with Appearance-Motion Correspondence by Trong-Nguyen Nguyen, Jean Meunier. DIRO, University of Montreal.[2]

This paper is concerned with anomaly detection in video footage via processing individual frames through a model of three connected neural networks. The proposed model should be able to flag anomalous activity in any relatively repetitive footage, irrespective of the context of the footage. The three networks, namely a common encoder, appearance decoder and motion decoder, are connected in such a way that input enters at the encoder and after being processed is then provided to the appearance and motion decoders – therefore both decoders share an encoder. The encoder combined with the appearance decoder creates a reconstructed image of the input frame with object recognition and thus operates as a Conv-AE. The encoder combined with the motion decoder makes a prediction of the movement of objects from the current frame to following frame and thus forms a U-Net. The model is optimized according to the difference between the outputted and original version of the frame as well as the optical flow together with an adversarial loss function (optical flow meaning the learned/predicted path of a recognized object). In this way, the model tracks objects and learns to predict their movement. Once the model has been trained on a certain dataset, it will be able to continue to monitor moving objects from that video stream and in the instance that the predicted movement of one of those objects differs from the ground truth significantly, it will be reflected by a large value in the loss function and flagged as anomalous activity. The model will also flag static sightings of strange, unfamiliar objects as anomalous.

Other concepts and processes employed by this paper include an inception module, batch-normalization, patch-based score estimation and skip connections. The model is depicted in Figure 1.

2.2 Real-world Anomaly Detection in Surveillance Videos by Waqas Sultani. Department of Computer Science, Information Technology University, Pakistan.[3]

This paper proposes anomaly detection in untrimmed video footage using weakly labeled training videos (labels only exist at video level and not at frame level meaning a video can be labeled anomalous or normal but frames are not labeled). The architecture implements a weakly-supervised learning approach in the form of a deep MIL (Multiple Instance Learning) framework. Whole videos are treated as bags and, as mentioned, a bag can be labeled as anomalous or normal. Videos are then broken down into short segments and these are referred to as instances within the video's bag. An anomalous bag is made up of a collection of normal segments and at least one anomalous segment and a normal bag is naturally made up of normal segments. Based on training, we learn an anomaly

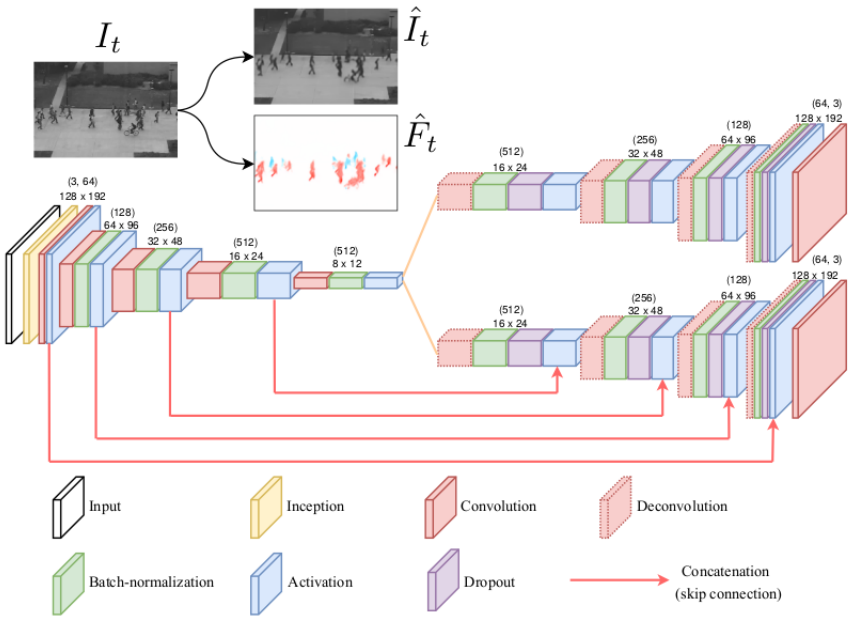

Figure 1: Conv-AE joined to U-Net through shared encoder. This combination of sub-networks forces the model to learn Appearance-Motion correspondence of objects.

ranking model which predicts high anomaly scores for anomalous segments.

In contrast to the existing methods, this paper formulates anomaly detection as a regression problem (we call it regression since we map feature vector to an anomaly score between 0 and 1). In typical supervised classification problems (where a frame must be labelled 1 or 0) we need labels of all examples to be available to learn a robust classifier. In the context of supervised anomaly detection, a classifier needs temporal annotations of each segment in videos. However, obtaining temporal annotations for videos is time consuming and laborious. MIL relaxes the assumption of having these accurate temporal annotations as it poses anomaly detection as regression problem where we subsequently rank video segments according to their anomaly scores and ideally have the highest scoring segment be the anomalous segment. In the absence of video segment level annotations, we propose a multiple instance ranking objective function where the max video segment anomaly score is taken over all video segments in each bag. Instead of enforcing ranking on every instance of the bag, we enforce ranking only on the two instances having the highest anomaly score respectively in the positive and negative bags. The segment corresponding to the highest anomaly score in the positive bag is most likely to be the true pos-

itive instance (anomalous segment). The segment corresponding to the highest anomaly score in the negative bag is the one that looks most similar to an anomalous segment but actually is a normal instance. These are typically difficult scenes to deal with in CCTV and including a strategy to account for them in the methodology is very useful in avoiding false alarms. Since the video is a sequence of segments, the anomaly score should vary smoothly between video segments. Therefore, temporal smoothness between anomaly scores of temporally adjacent video segments is enforced by minimizing the difference of scores for adjacent video segments. By training on a large number of positive and negative bags, we expect that the network will learn a generalized model to predict high scores for anomalous segments in positive bags. High false alarm rates seem to be one of the most challenging problems faced by anomaly detection architectures. This approach has a much lower false alarm rate than other methods, indicating a more robust anomaly detection system in practice. This validates that using both anomalous and normal videos for training helps our deep MIL ranking model to learn more general normal patterns.

Other significant contributions of this paper include 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 the new dataset which show that the proposed method achieves superior performance as compared to the state-of-the-art anomaly detection approaches.

2.3 Learning Memory-guided normality for Anomaly Detection by Hyunjong Park, Jongyoun Noh, Bumsub Ham. School of Electrical and Electronic Engineering, Yonsei University.[1]

This paper presents an unsupervised learning approach to anomaly detection in video sequences considering the diversity of normal patterns. The model works of the assumption that a single prototypical feature is not enough to represent various patterns of normal data. The researchers propose to use multiple prototypes to represent the diverse patterns of normal video frames for unsupervised anomaly detection. This is achieved by introducing a memory module recording prototypical patterns of normal data on the items in the memory. They also propose feature compactness and separateness losses to train the memory, ensuring the diversity (capturing normal objects from many different views) and discriminative (recognition and tracking) power of the memory items. Also presented is a new update scheme of the memory, when both normal and abnormal samples exist at test time.

An overview of the proposed framework follows: we reconstruct input frames or predict future ones for unsupervised anomaly detection. We input four successive video frames to predict the fifth one for the prediction task. As the

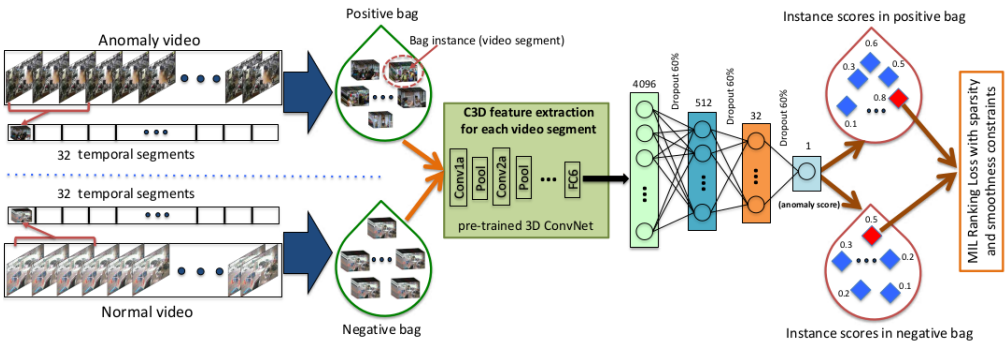

Figure 2: Deep MIL Framework applied to anomalous and normal bags of video segment instances. Given the positive (containing anomaly somewhere) and negative (containing no anomaly) videos, we divide each of them into multiple temporal video segments. Then, each video is represented as a bag and each temporal segment represents an instance in the bag. After extracting C3D features [37] for video segments, we train a fully connected neural network by utilizing a novel ranking loss function which computes the ranking loss between the highest scored instances (shown in red) in the positive bag and the negative bag.

prediction can be considered as a reconstruction of the future frame using previous ones, we use almost the same network architecture with the same losses for both tasks. The model mainly consists of three components: an encoder, a memory module, and a decoder. The encoder inputs a normal video frame and extracts query features. The features are then used to retrieve prototypical normal patterns in the memory items and to update the memory. We feed the query features and memory items aggregated (i.e., read) to the decoder for reconstructing the input video frame. We train our model using reconstruction, feature compactness and feature separateness losses end-to-end. For testing we use a weighted regular score in order to prevent the memory from being updated by abnormal video frames. We then compute the discrepancies between the input frame and its reconstruction and the distances between the query feature and the nearest item in the memory to quantify the extent of abnormalities in a video frame.

The remainder of the paper discusses details relating to the implementation of the network architecture as well as explanations of the feature construction, separation and compactness loss functions. The results of experimentation with this model on 3 benchmark datasets indicate that the model out-performs state-of-the-art methods in all 3 of the datasets on the prediction task.

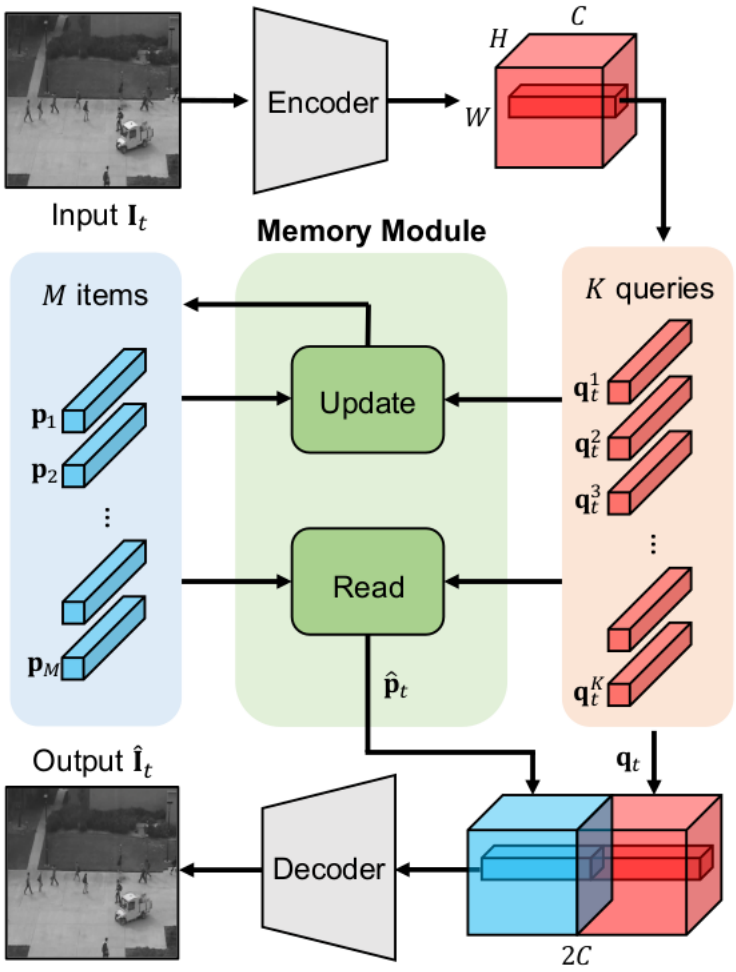

Figure 3: Overview of framework for reconstructing a video frame. The model mainly consists of three parts: an encoder, a memory module, and a decoder.

3 Potential for Improvement and Extended Work

A persistent problem faced by anomaly detection architectures seems to be the high rate of false alarms triggered which in some cases can significantly reduce the quality of a system when networks are not able to find solutions to reducing the false alarm rate. This problem does also depend on the dataset in question i.e., a dataset of pedestrians on a walkway with an occasional bicycle passing by may consider a bicycle to be an anomaly and whether this is a correct diagnosis would depend on the intention of those overseeing the CCTV supervision. Another issue that arises is the missed detection of anomalies due to

poor/alternative lighting or due to the anomaly being of a subtle nature which would not be picked up by a computer vision network lacking general intelligence. Many anomaly detecting frameworks operate 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 or impossible to define a normal event which takes all possible normal patterns/behaviors into account. The above comments lead us to the overall issue of how normality differs vastly from normality, as do anomalies differ vastly from other anomalies. This is the root of problems in anomalous activity monitoring as the domain which the network operates on is extremely unpredictable.

The problem of anomaly detection may be simplified by classifying the surrounding environment as either crowded (a lot of objects in motion and a relatively erratic background) or uncrowded (static background, occasional activity). Based on this classification, the framework applies differing processes on top of the anomaly score prediction. In the case of a crowded scene, anomaly scores before an 'anomaly spike' are compared to anomaly scores afterwards (in a sliding window fashion). The length/time of comparison is inversely proportional to the height of the peak anomaly score in the spike. The thought behind this is that an anomaly has an effect on an overall environment (and a false positive has none) therefore, in the case of a true anomaly and a relatively crowded environment, the behaviour of other components in the scene should change compared to their behaviour before the anomaly. This solves the problem of normal, yet unusual, occurences being flagged (as the environment will not respond to them) as well as subtle, yet anomalous, occurences being missed (as the environment should respond to an anomaly). The window is inverse proportional such that a large enough anomaly peak will need very little extra monitoring however a slight anomaly peak will be further examined. The difference in anomaly scores from the window ahead and window behind the peak will contribute towards the classification of normal/anomalous action.

The above mentioned technique will not be applied to static scenes as the environment will show no change in the case of an anomaly. In this case I propose a motion prediction scheme. The anomalies seen in these instances are of the type of muggings, petty theft, burglary, loitering or perhaps violence in situations where there are no other people around. The environment will often be static in these scenes and remain unchanged after the anomaly, especially in the case of subtle activity. I expect the best way to identify such an anomaly is to learn the typical motion of moving objects in these situations and apply an adversarial loss function to flag unpredictable movement e.g. a burglary in an empty house will be a non-crowded environment. The model should have learnt the motion of the residents as casual, consistent walking and in the case of a human 'sneaking', loitering or climbing through a window the model should flag the activity as anomalous.

I would like to propose a combination of the above work as well as an exten-

sion to the methodology in order to better deal with the variance of anomalous and normal activity. The proposed method is intended to be an extension to the architecture discussed in 2.2 where videos are treated as bags and anomaly scores are predicted for segments. Figure 4 depicts graphs of the anomaly scores for segments over time and depicts a case where an anomaly is missed and one where a false positive is given.

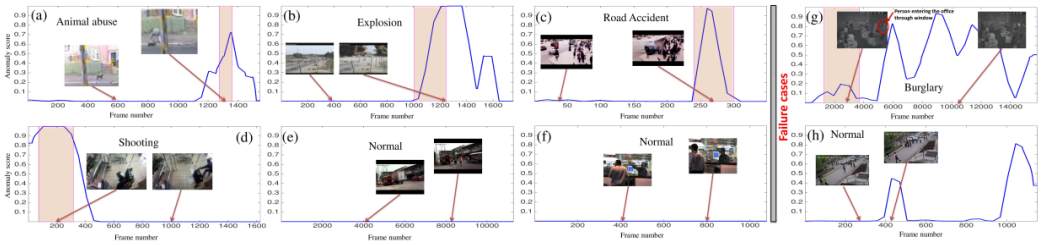

Figure 4: Qualitative results of 2.2 method on testing videos. Colored window shows ground truth anomalous region. (a), (b), (c) and (d) show videos containing animal abuse (beating a dog), explosion, road accident and shooting, respectively. (e) and (f) show normal videos with no anomaly. (g) and (h) present two failure cases of our anomaly detection method.

Adding to the issue of vastly differing anomalous and normal footage is the fact that anomalous activity is not confined to one region of a CCTV frame and may occur in opposing corners of the frame each time they occur. I imagine this makes it difficult to use conventional neural networks as the process of finding consistent features emerging from certain neurons associated with an anomaly is complicated by the anomaly appearing in different regions of the frame each time and therefore the network would struggle to adjust to find strong weights for a certain group of neurons associated with anomalous activity. The idea of moving/focusing the camera to put suspected anomalies at the center of the frame in real-time comes to mind when considering this issue (just as a human's attention is drawn to suspicious activity which is then given attention and further examined.). I expect that this idea is unrealistic given the datasets available and further complications that will arise from camera-jitter but I would like to research the problem of anomalies appearing in different regions and how it should be combated.

4 Proposed Brief

The paper summarized in section 2.2 seems like an intuitive approach with what seems to be the most conventional methodology in terms of the deep learning processes used. This paper takes a higher level approach where it is not as concerned with the fine-grained details within frames (which I find tends to

become very abstract and difficult to build on the strategy of the system) but rather lets the computer learn the intricate nuances behind anomalies and treats the problem as somewhat of a logistic regression problem where a positive result is a rare anomalous activity and a negative result is common activity. Te paper also provides a solution to dealing with false positives that are seen as very convincing by the network via learning from the mots anomalous seeming normal segments. The architecture of the paper seems to perform very well in the experiments and the anomalies detected by it are real-world and more relevant to society than those detected in the other papers. The paper does have the drawback that it requires weakly labeled data (at video level), however there is such a dataset with an extensive amount of quality, real-world footage provided in the paper. I expect to be able to develop a good understanding of the workings of this model to the degree that I may be able to optimize results, suggest and implement new techniques to complement the existing solution and apply the overall architecture to a certain dataset/application.

My proposed work for this project is to reimplement the paper mentioned in section 2.2 such that I can reproduce the quality of results found by the researchers. Thereafter I would like to investigate the extended functionality proposed in the above section and use it in experimentation on benchmark tests in order to see if the methods hold any merit. Thereafter, if time allows, I would propose that the final architecture is applied to a specific situation where the model can be well trained and become sufficiently capable in that domain e.g hijacking recognition.

References

[1] Bumsub Ham. School of Electrical Hyunjong Park, Jongyoun Noh and Yonsei University. Electronic Engineering. Learning memory-guided normality for anomaly detection.

[2] University of Montreal. Trong-Nguyen Nguyen, Jean Meunier. DIRO. Anomaly detection in video sequence with appearance-motion correspondence.

[3] Pakistan. Waqas Sultani. Department of Computer Science, Information Technology University. Real-world anomaly detection in surveillance videos.