Trimble / Bilberry : AI Engineer technical exercise

Part 2 : Paper Review : Ano-Graph Learning Normal Scene Contextual Graphs to Detect Video Anomalies

---

candidate : Rabah MOULAI

20 mars 2023

1 Introduction

Detecting anomalies in video streams is one of the most vital and indispensable tasks. However, video anomaly detection has proven to be a difficult task due to its unsupervised learning procedure and the great spatio-temporal complexity existing in real-world scenarios.

In the absence of anomalous training samples, state-of-the-art methods attempt to extract features that fully capture normal behaviors in spatial and temporal domains using different approaches such as autoencoders or generative adversarial networks. But, these approaches completely ignore or, poorly model the spatiotemporal interactions that exist between objects.

To address this problem, the authors of this article propose a novel and efficient method called AnoGraph for learning and modeling the interaction between objects. In order to encapsulate the interactions in a semantic space and perform anomaly detection from this semantic space.

1.1 Proposed Method

The goal of their approach is to find anomalies in the video while automatically explaining the reason and environmental context of the detector's response.To this end, a spatio-temporal graph (STG) is constructed by considering each node as an entity feature extracted from a real-time object detection while edges are made based on their interplays. Once the graph is built, self-supervised learning is performed on this graph in order to encapsulate the interactions into a semantic representation space for different normal and abnormal settings. The interest of this approach is that they take into account the interactions of objects across time and space in modeling contexts for video captioning tasks.

they propose to create a specific type of STG that captures spatio-temporal interactions while being trained using a self-supervised learning method that aims to train models without providing labeled data as input. Instead, they seek to label the data by finding and exploiting the relationships (or correlations) between different input signals. Thanks to this, they can obtain a normal semantic representation in the spatial and temporal domains without the need to fine-tune any parameter such as the number of frames but also in a completely unsupervised manner. The authors claim that their method can understand normal spatio-temporal relations of objects and discern abnormal behaviors based on them. Features of the objects are considered as nodes of the graph. Spatial edges are undirected, while directed edges correspond to temporal interactions. At test time, a discriminator is trained to distinguish between normal and abnormal graphs using an anomaly score.

In detail, in a first step, a real-time object detector such as Convolutional Neural Network detectors (R-CNN, Faster-RCNN, YOLO,...) is exploited to detect objects that exist in every frame. Thus, a high-level representation of the video for each frame and each object belonging to it is obtained in the form of a feature vector. Every object is viewed as a node in the graph and some spatial edges are built based on the Intersection Over Union (IOU) of the object's bounding boxes. This leads to a better modeling of objects' interactions thorough space. Where, for modeling temporal relations, consecutive frames' dynamics are considered by calculating the cosine similarity. The result of both modeling is a spatio-

temporal graph (STG) which resumes the adjacency matrix that includes the relational information between nodes (objects).Once the space-time structure is established, the entire graph is passed to a Graph Convolutional network (GCN) encoder to obtain the node embedding. This embedding attempts to capture the local and global information existing in the graph for each node or precisely centered around each node rather than the node itself. Based on the self-supervised learning method on graph embedding, they learn node representations containing both local and global information of all spatio-temporal interactions of objects. Then, a discriminator is trained to assign a probability score based on the presence of the patch-level summary and the global graph-level summary for the normal and abnormal graphs. At the test time, the videos are concatenated with each other, and an entire graph is extracted. Then, the graph is passed to the encoder and the embedding at the time of testing is extracted. Finally, the discriminator discerns the abnormal and normal embedding of the nodes using their deviation from the summary vector obtained during the training process.

1.2 Results and conclusion

This approach has been tested on different open source datasets like **Avenue, ShanghaiTech, UCSDPed2, ADOC and Street Scene** Comparing with an exhaustive set of SOTA approaches, including generative, SSL and AE-based methods. the results of this method is comparable or even better compared with the other considered methods

In this work, they have proposed a novel anomaly detection approach based on a SSL method and a STG, presenting comprehensive results on five datasets Avenue, ShanghaiTech, UCSD-Ped2, ADOC, and Street Scene. Experiments and ablation studies show that they not only pass SOTA by a large margin but also they are more flexible, data efficient, and robust compared to others.

1.3 My interest for this article

Computer vision is a broad discipline that has grown rapidly in recent years due to the use of deep learning in areas such as image classification and object detection. Convolutional neural networks are the most commonly used application but are generally used as black boxes without necessarily understanding what is going on inside. Recently, GNNs have also been used in this area. Although GNN applications in the computer vision field are still in their infancy, they have a huge potential for the coming years especially since they manage to capture the semantic aspect. And this same semantic aspect that interested me in this article.

Article : Masoud Pourreza, Mohammadreza Salehi, and Mohammad Sabokrou. Anograph : Learning normal scene contextual graphs to detect video anomalies. arXiv preprint arXiv :2103.10502, 2021.