AML Final Project

Title: Video Anomaly Detection in Deep Learning

Computer Vision

Maddi Vineeth Goud

Abstract

Typically, the problem of anomaly detection with weakly supervised video-level labels is formulated as a multiple instance learning (MIL) problem, where the goal is to find video snippets that contain anomalous events, with each video being a collection of such snippets. Even though current methods perform effectively in terms of detection, their recognition of positive instances, i.e., rare abnormal snippets in the abnormal videos, is heavily skewed by the predominately negative instances, particularly when the abnormal events are minute anomalies with barely discernible differences from normal events. Numerous approaches that disregard crucial video temporal dependencies make this problem worse. To solve this problem, from the state-of-the-art techniques (SOTA), I present the paper review of three techniques which are competing in the current research. A paper review is presented on the below techniques captured from the SOTA. Techniques namely Self-supervised Sparse Representation for Video Anomaly Detection (S3R), Robust Temporal Feature Magnitude Learning (RTFM), and Real-world Anomaly Detection in Surveillance Videos.

Acknowledgments

I would like to thank my Professor **CHAOJIANG WU**

Table of Contents

1. Video Anomaly Detection in Deep Learning:

Surveillance footage can contain a range of realistic anomalies, which can be detected in both normal and abnormal videos. Rather than annotating anomalous segments or clips in training videos, a deep learning framework called the multiple instances ranking framework can be used, which relies on weakly labeled training videos. These videos have labels for the entire video, rather than for individual clips. Our approach uses a self-attention module and the RTFM principle to develop a deep anomaly ranking model that assigns high anomaly scores to anomalous video segments. This eliminates the need for time-consuming manual annotation of anomalous segments or clips.

Scope:

Automated detection of unusual events, such as accidents, criminal activity, or unlawful behavior, is beneficial. Since anomalous events are uncommon and come in a variety of forms, manually recognizing them can be time-consuming and arduous, especially for lengthy video sequences. According to the computer vision field's annotations or presumptions on the training video sequences, recent approaches to the VAD task can be divided into unsupervised and weakly-supervised strategies.

State-Of-the-Art Techniques list in VAD (Video Anomaly Detection):

Rank	Model	ROC AUC	Decidability	EER	Paper	Code	Result	Year	Tags
1	MGFN	86.98			MGFN: Magnitude-Contrastive Glance-and-Focus Network for Weakly-Supervised Video Anomaly Detection			2022	
2	S3R	85.99			Self-supervised Sparse Representation for Video Anomaly Detection			2022	
3	WSAL	85.38			Localizing Anomalies from Weakly-Labeled Videos			2020	
4	Multi-stream Network with Late Fuzzy Fusion	84.48			A multi-stream deep neural network with late fuzzy fusion for real-world anomaly detection			2022	
5	RTFM	84.03	-	-	Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning			2021	
6	MIST	82.30			MIST: Multiple Instance Self-Training Framework for Video Anomaly Detection			2021	
7	DMRMs	81.91	-	-	Anomalous Event Recognition in Videos Based on Joint Learningof Motion and Appearance with Multiple Ranking Measures			2021	

Figure 1: SOTA Techniques.

2.Literature Review:

The detection of anomalies in computer vision is a challenging task, particularly in video surveillance applications for identifying hostility or violence. Various methods have been proposed to detect anomalies, including those that use multiple-instance learning (MIL), encoders and attention networks, magnitude learning, weakly supervised anomaly localization (WSAL), and sparse representation. Each method has its own strengths and weaknesses.

MIL is effective for weakly labeled training videos and uses a ranking loss with sparsity and smoothness constraints to learn anomaly scores for video segments. However, it requires video-level labels and does not know the exact temporal positions of abnormal events.

MIST uses multiple instance self-training with encoder and self-attention networks to analyze local temporal anomalies. It also includes a multiple instance pseudo label generator and a self-guided attention boosted feature encoder. However, it struggles to adapt to variation in duration of untrimmed videos and class imbalance.

RTFM focuses on magnitude learning of anomalies and trains a feature magnitude learning function to improve the robustness of MIL approach to the negative instances from abnormal videos. It also includes dilated convolutions and self-attention mechanisms to capture long- and short-range temporal dependencies. However, it does not perform well on subtle anomalies with minor differences from normal events.

WSAL combines a high-order context encoding model with dynamic fluctuation measurement to locate anomalous phenomena. However, it struggles with noise interference and lacks localization guidance in anomaly detection.

Finally, S3R introduces a self-supervised sparse representation framework that models the concept of anomaly at feature level by exploring the synergy between dictionary-based representation and self-supervised learning. It includes en-Normal and de-Normal modules to rebuild snippet-level features and filter out normal-event features. However, it also generates pseudo-normal/anomaly data that may not be reliable.

In conclusion, each method has its own advantages and disadvantages, and understanding the methodologies can help in building effective anomaly detection models.

Advantages:

• The proposed methods offer an MIL-based solution for anomaly detection using weakly labeled training videos.

• They use a MIL ranking loss along with sparsity and smoothness constraints for a deep learning network to learn anomaly scores for video segments.

• Experimental results on a new dataset demonstrate that the proposed methods outperform the current state-of-the-art anomaly detection approaches.

• A self-guided attention boosted feature encoder is utilized to automatically focus on anomalous regions in frames while extracting task-specific representations.

• The RTFM models employ dilated convolutions and self-attention mechanisms to capture both long- and short-range temporal dependencies for more accurate feature extraction.

• A high-order context encoding model is introduced to effectively use temporal context and extract semantic representations while measuring dynamic fluctuations.

Disadvantages:

• While the requirement for providing precise temporal annotations is relaxed by MIL, the exact temporal positions of abnormal events in videos are not known.

• Only video-level labels showing the presence of an abnormality throughout the entire video are required.

• The main drawback of the MIST model is its inability to adapt to the variation in duration of untrimmed videos and handle class imbalance.

• The RTFM models can handle the variation in duration of untrimmed videos but are not designed to handle class imbalance.

• Since the training phase lacks information on anomaly positions, both methods may not accurately predict anomaly frames.

The proposed method is evaluated against state-of-the-art approaches using numerous benchmark datasets for weakly-supervised video anomaly identification. The outcomes show how useful and practical the suggested strategy is for practical applications.

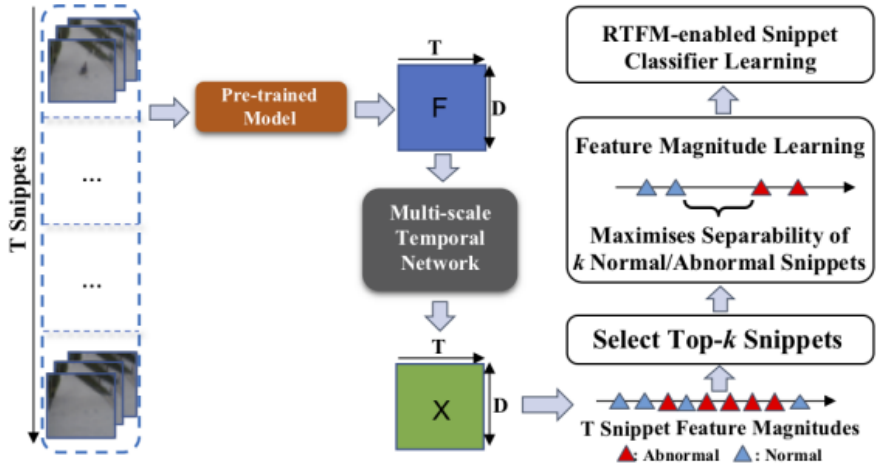

Figure 2: RTFM Architecture

Our suggested RTFM is given a T D feature matrix F that was taken from a video with T different video clips. Then, MTN generates X = s(F) by capturing the long- and short-range temporal dependencies between sample features. The top-k greatest magnitude feature snippets from abnormal and normal films are used to train a snippet classifier after we maximize the separability between abnormal and normal video features. Cutting-edge performance On a number of benchmark datasets for weakly-supervised video anomaly detection, the suggested method performs at the cutting edge, highlighting its efficacy and potential for use in practical situations.

For the MIST framework,

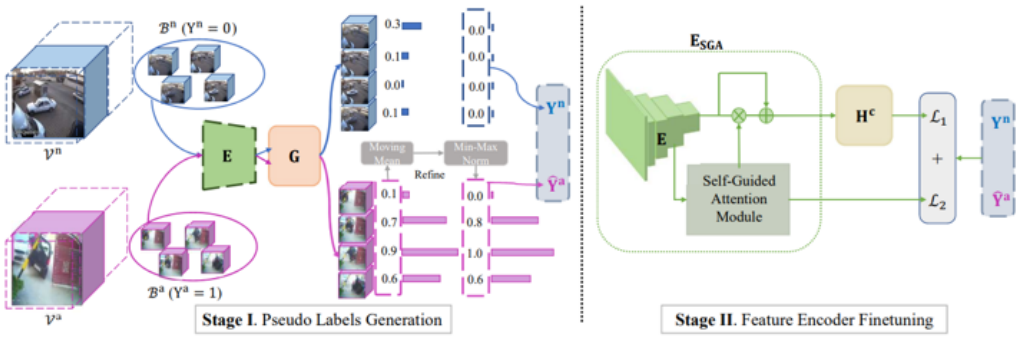

Figure 3: MIST Architecture

The objective is to find anomalous events in a video series given a binary anomaly label (anomalous or normal). The suggested MIST framework's specific goal is to increase a video anomaly detection model's detection accuracy by using labeled and unlabeled data.

To overcome the problem of poorly labeled anomaly data—where the anomaly label is only provided at the video-level—the suggested method makes use of multiple-instance learning. Better detection performance is achieved because to the MIST framework, which enables the model to learn from examples that are originally unannotated. On various benchmark datasets, the suggested solution performs better than the most advanced video anomaly detection techniques.

In order to detect video anomalies, the study suggests a multiple instance self-training framework (MIST) that makes use of both labeled and unlabeled data.

Through self-supervised learning, the MIST framework is intended to increase the detection precision of a specific video anomaly detection model.

3. Applications of Anomaly Detection in Current Research:

Anomaly detection is the act of locating data points or patterns that differ from a system's or process's typical behavior. It has been demonstrated that deep learning approaches are highly good at spotting abnormalities across a variety of businesses.

Deep learning-based anomaly detection in many industries has the following prospective and existing applications:

Finance: Deep learning has the potential to identify fraud in areas such as stock trading, credit card transactions, and financial transactions. By training deep learning algorithms, patterns of fraudulent activity or abnormal transactions can be detected in real-time.

Manufacturing: Deep learning has the potential to detect abnormalities in manufacturing procedures, including equipment malfunctions and production line errors. By processing machine sensor data, deep learning models can be trained to identify patterns that indicate potential issues.

Healthcare: Deep learning algorithms can be employed to detect abnormalities in medical information, including imaging test outcomes and patient vital signs. As an instance, deep learning algorithms can be employed to examine patient data and detect patterns that might suggest potential health hazards, or to recognize cancers in medical imaging.

Cybersecurity: Deep learning is useful in detecting abnormalities in network traffic for cybersecurity purposes, such as identifying viruses or cyberattacks. One example is the utilization of deep learning models to scrutinize network data, identifying patterns that signify potential risks.

Energy: Deep learning can be used to spot irregularities in usage trends, such as rapid increases or decreases. Deep learning models, for instance, can be used to track patterns in energy data from smart meters that could point to energy wastage or equipment malfunction.

By enabling real-time detection of unexpected behavior, which can result in more effective operations, increased safety, and lower costs, deep learning-based anomaly detection has the potential to transform numerous industries.

4. Potential Future Developments

- **Better Accuracy:** Deep learning models are constantly developing, thus new breakthroughs in the future are probably going to make anomaly detection in videos more accurate. This could be accomplished by advances in model interpretability, availability of training data, and neural network topologies.
- **Real-time Processing**: The potential for real-time processing of video data is growing as powerful computing resources become more widely available. Real-time anomaly

identification and reaction would be possible as a result, enhancing security and safety across numerous industries.

- **Cross-domain Application**: Deep learning models that have been trained on a particular type of video data can be used across domains with similar properties. A model that was trained on traffic camera footage, for instance, may possibly be used with security camera footage.

5.limitations:

- **Limited Training Data:** In order to produce correct results, deep learning models require a large amount of training data. However, gathering labeled training data might take a lot of time and money for video anomaly detection.
- **Lack of Interpretability:** Because deep learning models can be challenging to understand, it might be difficult to understand how and why a model is making a particular prediction.
- **Temporal dependencies** are tough to address since video data is inherently temporal and anomalies may not be obvious in a single frame, which may be a problem in instances where reasons for decisions are required. Deep learning systems may struggle to understand temporal dependencies, which could make anomaly identification less precise.

6.Potential Solutions:

- • **Transfer Learning:** Transfer learning can be used to discover anomalies in video data by utilizing built-in, pre-trained models. In order to achieve high accuracy, less labeled training data may be required.
- **Explainable AI:** Explainable AI techniques can be applied to deep learning models to boost transparency, making it easier to understand how and why a model is making a particular prediction.
- **Attention Mechanisms**: By utilizing attention mechanisms, deep learning models may more accurately depict the temporal dependencies in video data. By focusing on specific regions of the video, attention mechanisms can improve the accuracy of anomaly identification.

Citations:

- Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6479-6488. 2018.

- Feng, Jia-Chang, Fa-Ting Hong, and Wei-Shi Zheng. "Mist: Multiple instance self-training framework for video anomaly detection." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 14009-14018. 2021.
- Tian, Yu, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W. Verjans, and Gustavo Carneiro. "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 4975-4986. 2021.
- Lv, Hui, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. "Localizing anomalies from weakly-labeled videos." IEEE transactions on image processing 30 (2021): 4505-4515.
- Wu, Jhih-Ciang, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. "Self-supervised Sparse Representation for Video Anomaly Detection." In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII, pp. 729-745. Cham: Springer Nature Switzerland, 2022.
- Chen, Yingxian, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. "MGFN: Magnitude-Contrastive Glance-and-Focus Network for Weakly-Supervised Video Anomaly Detection." arXiv preprint arXiv:2211.15098 (2022).

References:

1. Mist: Multiple instance self-training framework for video anomaly detection -- https://arxiv.org/pdf/2104.01633v1.pdf
2. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning -- https://arxiv.org/pdf/1801.04264v3.pdf
3. Localizing anomalies from weakly labeled videos -- https://arxiv.org/pdf/2008.08944v3.pdf
4. Self-supervised Sparse Representation for Video Anomaly Detection -- https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136730727.pdf
5. MGFN: Magnitude-Contrastive Glance-and-Focus Network for Weakly-Supervised Video Anomaly Detection -- https://arxiv.org/abs/2211.15098