Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/57/20241336

Video anomaly detection based on hybrid attention mechanism

Ruifeng He[1,2,3], Mingtian Xie[1], Aixing He[1]

[1]College of Computer and Information Science, Southwest University, Chongqing, China

[2]heruifeng98@163.com
[3]corresponding author

Abstract. To improve the ability of video anomaly detection models to extract normal behavior features of samples and suppress abnormal behaviors, this paper proposes an unsupervised video anomaly detection model, which takes advantage of spatio-temporal feature fusion, storage module, attention mechanism, and 3D autoencoder model. The model utilizes autoencoder to capture scene feature maps to enhance anomaly feature extraction. These maps are merged with the original video frames, forming fundamental units constituting continuous sequences serving as the model's input. Moreover, the attention mechanism is integrated into the 3D convolutional neural network to strengthen the network's capability in extracting channel and spatial features from videos. Experimental validation is performed on a publicly accessible campus dataset, illustrating the model's superior accuracy in anomaly detection.

Keywords: anomaly detection, feature fusion, attention mechanism, autoencoder.

1. Introduction

Video anomaly behavior refers to a relatively rare type of behavior that often conflicts with common behavior and carries a certain degree of danger. Detecting these anomalies accurately and promptly is crucial for maintaining social security. The goal of video anomaly detection is to locate violations in time and space within video sequences, such as wrong-way driving, fighting, or crowd dispersal, by utilizing technologies such as computer vision and machine learning. Proper video anomaly detection methods can automatically learn normal behavior patterns in scenes and automatically detect deviations that significantly deviate from normal patterns. This not only significantly reduces labor costs but also has good timeliness and accuracy[1].

Unlike other supervised learning-based video behavior recognition tasks, video anomaly detection is not suitable for supervised learning methods due to the characteristics of vague definitions, rarity, and scene dependence of anomalous behavior. Therefore, most video anomaly detection adopts unsupervised learning methods. In related literature, some studies first use deep denoising autoencoder to reconstruct spatio-temporal cubes and use the output of fully connected layers as learned representations of video events[2]; and some studies use convolutional deep autoencoder to extract video features better[3]. These methods in the literature all use deep autoencoder and have achieved

DOI: 10.54254/2755-2721/57/20241336

good results. However, because videos have temporal continuity, the above methods only consider spatial feature extraction and ignore the correlation in time series.

Combining spatial and temporal features is crucial for anomaly detection tasks. This paper introduces a novel deep learning-based method, which aims to optimize anomaly detection by integrating spatial multiscale features from normal scenes with temporal information.

The structure of this method is illustrated in Figure 1. During the training process, the model assigns high weights to normal samples to focus on important information. However, during the testing phase, the extraction of features for abnormal behavior is not influenced by the same weights, which somewhat suppresses the generation of abnormal behavior. Additionally, the introduction of a memory module records the deep semantic features of different patterns in normal samples, thereby increasing the prediction error for abnormal samples.

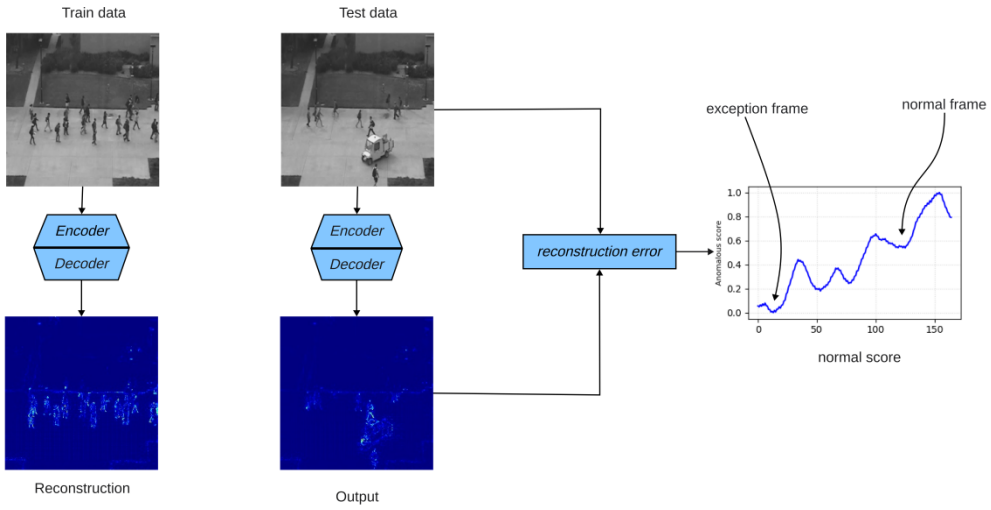

Figure 1. Anomaly detection reconstruction method diagram.

2. Module Design

The structure of this method is illustrated in Figure 2. The framework is an encoding-decoding structure and mainly consists of a 3D autoencoder module, a 3DCBAM module and a memory module.

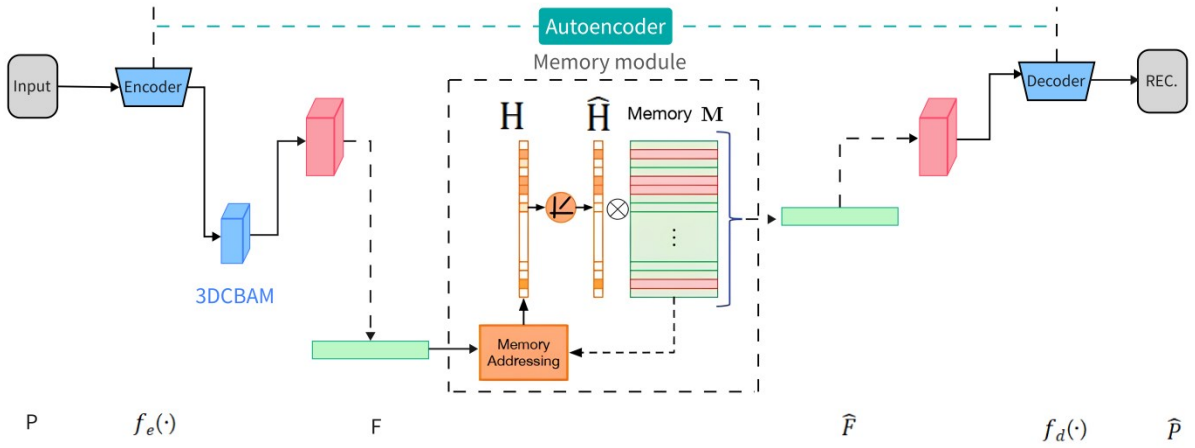

Figure 2. Architecture Diagram of Anomaly Behavior Detection Network.

DOI: 10.54254/2755-2721/57/20241336

2.1. Autoencoder

An autoencoder is a neural network model that comprises an encoder and a decoder. Its purpose is to map input data to itself. The encoder compresses the input data into a low-dimensional feature vector, aiming to minimize information loss, while the decoder reconstructs the low-dimensional feature vector back into the original data accurately. Throughout the training process, the model endeavors to minimize the disparity between the input and the reconstruction, thus learning effective features within the data during the mapping process[4].

2.2. 3DCBAM

The CBAM attention mechanism is a simple but highly effective attention module that can be used with feedforward convolutional neural networks[5]. It is made up of two parts: the channel attention module and the spatial attention module. CBAM can serialize feature maps created by convolutional neural networks and calculate attention maps in both channel and spatial dimensions. It then performs adaptive feature learning by multiplying the attention map and the feature map element-wise. This lightweight module can be embedded into any backbone network to improve performance. Researchers have attempted to apply CBAM to end-to-end training of 2D convolutional networks such as VGG, Inception, and ResNet[6]. To enhance the spatial feature utilization of 3D convolutional networks, this paper proposes the 3D-CBAM attention mechanism, with specific integration shown in Figure 3.

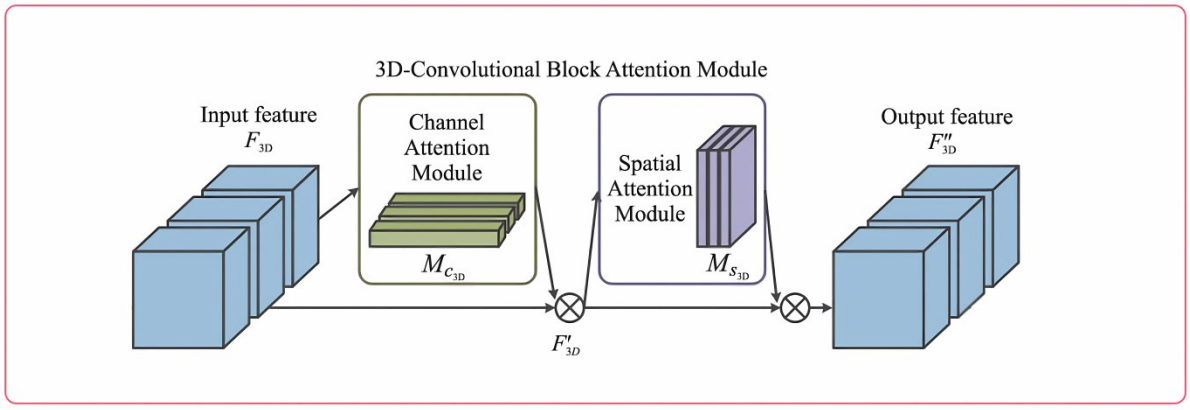

Figure 3. 3DCBAM Structure Diagram.

Unlike 2D convolutional networks, 3D convolutional networks have an additional depth dimension. Therefore, when extracting spatial features, it is necessary to consider variations in the depth parameter. For a feature map of an intermediate 3D convolutional layer $F_{3D} \in R^{C\times H\times W\times D}$, where C is the number of channels, 3DCBAM sequentially derives the channel attention feature map $M_{c3D} \in R^{C\times 1\times 1\times 1}$ and the spatial attention feature map $M_{s3D} \in R^{H\times 1\times D\times W}$. The entire process can be represented by the following formulas:

$$F'_{3D} = F_{3D} \otimes M_{c3D}(F_{3D}) \tag{1}$$

$$F''_{3D} = F'_{3D} \otimes M_{s3D}(F'_{3D}) \tag{2}$$

The 3DCBAM model has a channel attention module that focuses on the channels that are important in determining the final classification results of the fused 3D network. It selects features that have a decisive impact on predictions. Figure 4 shows the specific steps involved. Firstly, the input feature map F_{3D}undergoes average pooling and max pooling operations along the depth, height, and width dimensions separately. Then, the features processed separately by the MLP are summed element-wise, followed by a sigmoid activation operation. The resulting channel attention feature map

DOI: 10.54254/2755-2721/57/20241336

M_{c3D} is multiplied element-wise with the input feature map F_{3D} to generate the final channel feature map F'_{3D}. This is expressed by the following formula:

$$M_{c3D}(F_{3D}) = \sigma(MLP(MaxPool3D(F_{3D}) + MLP(AvgPool3D(F_{3D})))$$
$$= \sigma((W_1(W_0(F_{max}^{C_1 0} {}_{avg}^{C} \tag{3}$$

In the equation, σ represents the sigmoid operation, and W needs to undergo ReLU activation. In this paper, the reduction ratio r is set to 4, which means that the number of channels C is transformed to C/4 during max-pooling and average-pooling operations, reducing the parameter count. Finally, the number of channels is restored to the original C through a fully connected layer.

2.3. Memory module

Sometimes, the generalization capability of the Autoencoder can be too strong. In such cases, if the input embedding to the decoder is composed of embedding from normal samples, it can be anticipated that the reconstructed images produced by the decoder will primarily consist of features from normal samples. In this way, by suppressing the generalization capability, the reconstructed images can be forced to be closer to normal samples[7].

The specific approach is as follows: all embedding obtained by encoding normal samples is stored in memory. When an image is input, its embedding is extracted using the encoder. Then, the similarity between the image's embedding and each embedding in the memory is calculated individually (e.g., using cosine similarity). Subsequently, the embedding in the memory is weighted and averaged using the similarities as weights, resulting in a new embedding. This new embedding will possess two characteristics: it will be relatively close to the original image's embedding and composed of features from normal samples. This new embedding is then input into the decoder, enabling the generation of images close to the original image and resembling normal samples.

The memory module essentially stores a matrix of size N×C, where C is consistent with the dimensionality of the encoding result Z. Each row in the memory is denoted by m, representing a storage item. Given the encoding result F, the memory network obtains $\hat{H}$ based on a soft addressing vector H (1xN), and H is also computed based on F (each item being non-negative). Here, N is a hyperparameter, and empirical evidence shows that 3DCAE-MEM is not sensitive to N, with larger values generally yielding better performance.

$$\hat{F} = hM = \sum_{i=1}^{N} h_i m_i \tag{4}$$

3. Experiment analysis

3.1. Experiment dataset

The Avenue[8], UCSDPed[9], and ShanghaiTech[10] datasets were used for video anomaly detection. These datasets have predefined training and testing sets, and anomaly events only exist during testing.

3.2. Evaluation standard

In this section, the most commonly used evaluation metric in video anomaly detection, the Area Under the Curve (AUC), is employed for assessment. AUC focuses solely on the overall performance without considering the specific scores of positive and negative samples. Therefore, it effectively avoids the subjectivity introduced by empirical threshold setting during evaluation, making it particularly suitable for assessing the performance of tasks with imbalanced positive and negative samples.

3.3. Result

Table 1 presents a comparison of the AUC performance metrics between the 3DCAE-MEM model and other mainstream methods across different datasets. It's evident from the table that in terms of

DOI: 10.54254/2755-2721/57/20241336

reconstruction algorithms, the proposed method in this paper performs similarly to memAE[11] and outperforms other traditional algorithms. This is attributed to the fact that memAE also adopts a model structure based on memory modules, which offers certain advantages in feature extraction.

In this study, a Channel-Spatial Mixed Attention 3D-CBAM module is added before the memory module to better extract both global and local information from the feature maps. Since only normal data are involved in the training process, the 3D-CBAM mixed attention mechanism often cannot effectively extract features related to abnormal behavior in the test set. This prevents encoder features from directly concatenating into the decoder, thus resulting in abnormal behavior generating normal results. The memory mechanism at the bottleneck records prototype patterns of normal data, allowing for the constraint of abnormal behavior features during testing. This helps reduce the model's generalization ability while also improving accuracy.

Table 1. Performance comparison of different methods.

	Method	Ped1	Ped2	Avenue	SH.Tech
Non-Recon.	HOFME	0.727	0.875	–	–
	MPPCA+SF	0.742	0.613	–	–
	Conv-AE	0.81	0.9	0.702	–
	ST-AE	**0.899**	0.874	0.803	–
	Frame-Pred	–	**0.954**	**0.849**	**0.728**
Recon.	AE	–	0.917	0.810	0.697
	MemAE-nonSpar	–	0.929	0.821	0.688
	MemAE	–	**0.941**	**0.833**	0.712
	3DCAE-MEM	**0.901**	0.936	0.828	**0.736**

4. Conclusions

This paper proposes a fusion model aimed at addressing some limitations of existing video anomaly detection methods. The model introduces a memory module on top of the AE network and integrates the 3D-CBAM attention mechanism to enhance feature recognition accuracy. Through this fusion, the model can more effectively capture key features in videos, thereby improving the performance of anomaly detection. Experiments conducted on campus datasets demonstrate the model's excellent performance in detecting abnormal behaviors.

However, the model also has some limitations. The main issue is that the model requires fixed parameters for the same scene, and changing scenes necessitate retraining the model. This limits the flexibility and generality of the model in practical applications. The authors plan to address these issues further by developing a universal anomaly detection model applicable to most scenarios. This may involve optimizing the model structure, adapting parameters automatically, and employing other techniques to enhance the model's universality and applicability, thereby better addressing the needs of anomaly detection in different scenarios.

References

[1] Sultani W, Chen C, Shah M. (2018) Real-World Anomaly Detection in Surveillance Videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City. pp. 6479-6488. https://doi.org/10.48550/arXiv.1801.04264.

[2] Xu D, Yan Y, Ricci E, Sebe N. (2017) Detecting anomalous events in videos by learning deep representations of appearance and motion. Computer Vision and Image Understanding. 156: 117-127. https://doi.org/10.48550/arXiv.1510.01553.

[3] Hasan M, Choi J, Neumann J, et al. (2016) Learning Temporal Regularity in Video Sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas. pp. 733-742. https://doi.org/10.48550/arXiv.1604.04574.

[4] Deepak K, Chandrakala S, et al. (2021) Residual spatiotemporal autoencoder for unsupervised video anomaly detection. Signal, Image and Video Processing. 15: 215–222. https://doi.org/10.1007/s11760-020-01740-1.

[5] Woo S, Park J, Lee JY, et al. (2018) CBAM: Convolutional Block Attention Module. Proceedings of the European Conference on Computer Vision. Munich. pp. 3-19. https://doi.org/10.48550/arXiv.1807.06521.

[6] Zhou J T, Zhang L, et al. (2020) Attention-Driven Loss for Anomaly Detection in Video Surveillance. IEEE Transactions on Circuits and Systems for Video Technology. 30: 4639 - 4647. https://doi.org/10.1109/TCSVT.2019.2962229.

[7] Luo W X, Liu W, Lian D Z, et al. (2021) Video Anomaly Detection with Sparse Coding Inspired Deep Neural Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 43: 1070 - 1084. 10.1109/TPAMI.2019.2944377.

[8] Lu C, Shi J P, Jia J Y. (2013) Abnormal Event Detection at 150 FPS in MATLAB. In: Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney. pp. 2720-2727. 10.1109/ICCV.2013.338.

[9] Mahadevan V, Li W X, Bhalodia V, Vasconcelos N. (2010) Anomaly detection in crowded scenes. In: Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco. pp. 1975-1981. 10.1109/CVPR.2010.5539872.

[10] Luo W X, Liu W, Gao S. A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework. (2017) In: Proceedings of the IEEE International Conference on Computer Vision. Venice. pp. 341-349. 10.1109/ICCV.2017.45.

[11] Gong D, Liu Y, Le V, et al. (2019) Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection. In: Proceedings of the IEEE International Conference on Computer Vision. Seoul. pp. 1705-1714. 10.1109/ICCV.2019.00179.