LEARNING TASK-SPECIFIC REPRESENTATION FOR VIDEO ANOMALY DETECTION WITH SPATIAL-TEMPORAL ATTENTION

Yang Liu [1] Jing Liu [1] Xiaoguang Zhu [2] Donglai Wei [1] Xiaohong Huang [3] Liang Song [1]

[1]Fudan University [2]Shanghai Jiao Tong University [2]Shanghai University of Finance and Economics

Abstract

The automatic detection of abnormal events in surveillance videos with weak supervision has been formulated as a multiple instance learning task, which aims to localize the clips containing abnormal events temporally with the video-level labels. However, most existing methods rely on the features extracted by the pre-trained action recognition models, which are not discriminative enough for video anomaly detection. In this work, we propose a spatial-temporal attention mechanism to learn inter- and intra-correlations of video clips, and the boosted features are encouraged to be task-specific via the mutual cosine embedding loss. Experimental results on standard benchmarks demonstrate the effectiveness of the spatial-temporal attention, and our method achieves superior performance to the state-of-the-art methods.

Introduction

Video anomaly detection (VAD) aims to detect abnormal events in surveillance videos. Since abnormal events are rare and diverse, it is almost impossible to collect and label all kinds of anomalies for modeling. Therefore, most of the existing methods formulate VAD as an unsupervised or a weakly supervised task. Unsupervised methods use normal samples to learn a model of 'normality', and the anomaly is detected by measuring its deviation to the learned model. Due to the lack of observation of abnormal events, the unsupervised models may not learn the essential difference between normal and anomaly. In contrast, the weakly supervised VAD (ws-VAD) detects anomalies by comparing the normal and abnormal clips with the video-level labels.

The ws-VAD has been formulated as a multiple instance learning task, which uses the easy-to-obtain video-level labels to localize the abnormal clips. Sultani *et al.* [1] firstly proposed a MIL ranking model. The objective is that the maximum score of clips from an abnormal video should be greater than that of a normal video. Most of the existing methods directly use the spatial-temporal features extracted by the pre-trained models. However, the features extracted by the pre-trained models are not discriminative enough to distinguish normal and abnormal events. In addition, previous works always treat the video clips cut from the same video as independent instances, ignoring the inter-connections between adjacent video clips.

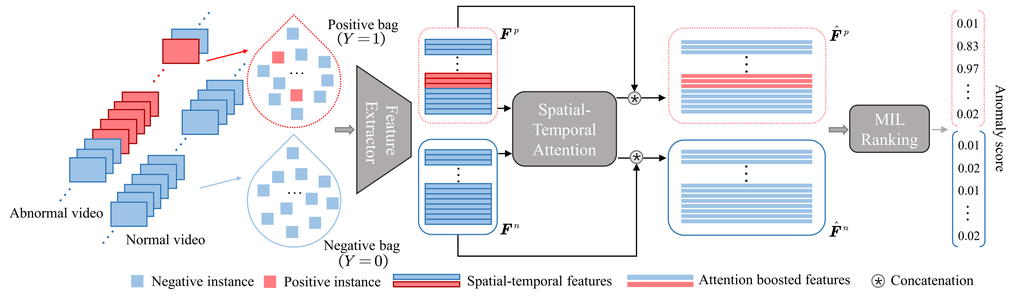
To obtain the task-specific spatial-temporal features, we propose **spatial-temporal attention (STA)** to explore the inter- and intra-correlations between video clips. The STA module can capture the global contextual spatial-temporal correlations through a recurrent crisscross attention operation.

Contributions

- We propose spatial-temporal attention to obtain task-specific features for ws-VAD. The global spatial-temporal correlations of all video clips can be captured via the easy-to-plugin recurrent attention operations.
- We propose an STA augmented multiple instance ranking model and introduce a mutual cosine loss to encourage the model to learn the prototypical patterns of normal events.
- Experimental results on three standard benchmarks demonstrate the effectiveness of the STA, and our model outperforms the state-of-the-art methods on the UCF-crime dataset.

Methods

Architecture



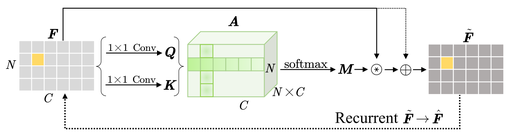
The architecture of the STA augmented MIL ranking framework is shown in the Figure above. We apply the well-trained 3D convolution model as feature extractor and feed features into the STA module to capture the global spatial-temporal correlations through a recurrent criss-cross attention. The bd-RNN based regression model output the anomaly score directly.

Spatial-temporal attention

Firstly, we obtain the query map $\boldsymbol{Q}$ and key map $\boldsymbol{K}$ via the 1×1 convolution. The vector of the i-th row and j-th column of query map $\boldsymbol{Q}$ is denoted by $\boldsymbol{q}_{(i,j)}$, obviously $1 \leq i \leq N$, $1 \leq j \leq C$ and $\boldsymbol{q}_{(i,j)} \in \mathbb{R}^D$, then we obtain criss-cross attention map $\boldsymbol{A}^{i,j}$ by computing the cosine similarity between $\boldsymbol{q}_{(i,j)}$ and vector $\boldsymbol{k}_{(m,n)}$ in the $\boldsymbol{K}$ that are in the same row or column as $\boldsymbol{q}_{(i,j)}$:

$$\boldsymbol{A}^{i,j}_{(m,n)} = \begin{cases} \dfrac{\boldsymbol{q}_{(i,j)}\boldsymbol{k}^T_{(m,n)}}{\|\boldsymbol{q}_{(i,j)}\|\|\boldsymbol{k}_{(m,n)}\|} & \text{if } m = i \text{ or } n = j, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

After traversing all vectors in the spatial dimension of $\boldsymbol{Q}$, we obtain $N \times C$ criss-cross attention maps, denoted by $\boldsymbol{A} = \{\boldsymbol{A}^{1,1}, \cdots, \boldsymbol{A}^{N,C}\} \in \mathbb{R}^{N \times C \times (N \times C)}$. Then, we perform softmax operation to obtain the spatial-temporal attention map $\boldsymbol{M}$:



$$\boldsymbol{M}^{i,j} = \exp\left(\boldsymbol{A}^{i,j}\right) \circledcirc \sum_{m=1}^{N}\sum_{n=1}^{C} \exp\left(\boldsymbol{A}^{m,n}\right), \tag{2}$$

where $\circledcirc$ indicates the pixel-wise division. We add up $\boldsymbol{M}^{i,j} \circledast \boldsymbol{F}$ and $\boldsymbol{F}_{(i,j)}$ to obtain the aggregated features $\tilde{\boldsymbol{F}}_{(i,j)}$, where $\circledast$ indicates the pixel-wise multiplication. We repeat the above process once to establish the connection between any two pixels, denoted by $\tilde{\boldsymbol{F}} \to \hat{\boldsymbol{F}}$. The STA is easy-to-plugin with a complexity of space and time of $\mathbf{O}((N \times C) \times (N + C))$.

Training loss

To enable $\hat{\boldsymbol{F}}^n$ to record the prototypical patterns while ignoring the diversity, we introduce a mutual cosine embedding loss $\mathcal{L}_{MCE}$, to obtain a more compact feature representation of $\hat{\boldsymbol{F}}^n$ while keeping the features of abnormal clips in $\hat{\boldsymbol{F}}^p$ away:

$$\mathcal{L}_{MCE} = 1 - \underset{1 \leq i < j \leq N}{\text{Avg}}\left(\frac{\hat{\boldsymbol{f}}^n_i \hat{\boldsymbol{f}}^n_j}{\|\hat{\boldsymbol{f}}^n_i\|\|\hat{\boldsymbol{f}}^n_j\|}\right) + \underset{1 \leq i < j \leq N}{\text{Avg}}\left(\min\left(\frac{\hat{\boldsymbol{f}}^p_i \hat{\boldsymbol{f}}^p_j}{\|\hat{\boldsymbol{f}}^p_i\|\|\hat{\boldsymbol{f}}^p_j\|}, \xi\right)\right), \tag{3}$$

where $\text{Avg}(\cdot)$ denotes the mean value. The ξ denotes a margin. We apply the MIL ranking loss $\mathcal{L}_{MIL}$ to optimize the regression model:

$$\mathcal{L}_{MIL} = \max\left(0, 1 - \max_{1 \leq i \leq N} r\left(\hat{\boldsymbol{f}}^p_i\right) + \max_{1 \leq j \leq N} r\left(\hat{\boldsymbol{f}}^n_j\right)\right) + \lambda_1 \sum_{i=1}^{N-1}\left(r\left(\hat{\boldsymbol{f}}^p_{i+1}\right) - r\left(\hat{\boldsymbol{f}}^p_i\right)\right)^2 + \lambda_2 \sum_{i=1}^{N} r\left(\hat{\boldsymbol{f}}^p_i\right), \tag{4}$$

where $r(\cdot)$ denote the anomaly score. The first part is ranking loss, used to make the maximum score of instance in the positive bag higher than that in the negative bag. The last two parts are smoothness loss and sparse loss, which are used to encourage the smoothness and sparsity of scores, respectively. Balanced by the λ_{MCE}, the total loss $\mathcal{L}_{total}$ is as follows:

$$\mathcal{L}_{total} = \lambda_{MCE}\mathcal{L}_{MCE} + \mathcal{L}_{MIL}. \tag{5}$$

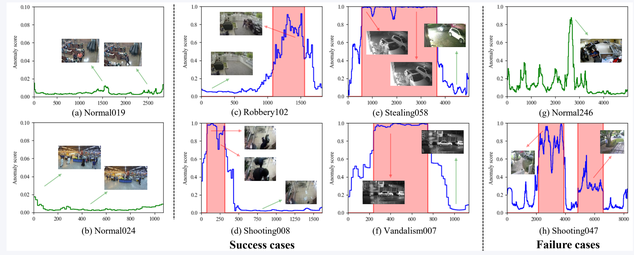
Experiments

Table 1. Results of Quantitative Frame-level AUC Comparison.

Supervision	Method	Feature type	Frame-level AUC(%)		
			UCF-crime	ShanghaiTech weakly†	UCSD Ped2†
None	Hassan *et al.* [2]	-	50.6	60.9	-
	Lu *et al.* [3]	-	65.5	-	-
	StackRNN [4]	-	-	68.0	92.2
	Frame-Pred [5]	-	-	72.8	95.4
	Mem-Guided [6]	-	-	70.5	**97.0**
Video-level labels	Sultani *et al.* [1]	$C3D^{(RGB)}$	75.4	86.3	-
		$I3D^{(RGB)}$	77.9	87.7*	91.8*
	Zhang *et al.*[7]	$I3D^{(RGB)}$	78.7	82.5	-
	Zhong *et al.* [8]	$C3D^{(RGB)}$	81.1	76.4	-
		$TSN^{(RGB)}$	82.1	84.4	-
	MIST [9]	$C3D^{(RGB)}$	81.4	<u>93.1</u>	-
		$I3D^{(RGB)}$	<u>82.3</u>	**94.8**	-
	Ours	$C3D^{(RGB)}$	81.6	88.7	92.3
		$I3D^{(RGB)}$	**83.0**	90.2	<u>96.7</u>

Table 2. Results of ablation studies.

Model	Negative ↓	Positive ↑	Score Gap ↑	Frame-level AUC (%)
Ours	0.21	0.84	0.63	83.0
Ours w/o mutual cosine embedding loss	0.33	0.75	0.42	79.8
Ours w/ FC	0.24	0.80	0.56	82.2
Ours w/ FC w/o mutual cosine embedding loss	0.35	0.76	0.41	79.4



References

[1] Waqas Sultani, Chen Chen, and Mubarak Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6479–6488.

[2] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis, "Learning temporal regularity in video sequences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 733–742.

[3] Cewu Lu, Jianping Shi, and Jiaya Jia, "Abnormal event detection at 150 fps in matlab," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2720–2727.

[4] Weixin Luo, Wen Liu, and Shenghua Gao, "A revisit of sparse coding based anomaly detection in stacked rnn framework," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 341–349.

[5] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao, "Future frame prediction for anomaly detection–a new baseline," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6536–6545.

[6] Hyunjong Park, Jongyoun Noh, and Bumsub Ham, "Learning memory-guided normality for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14372–14381.

[7] Jiangong Zhang, Laiyun Qing, and Jun Miao, "Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 4030–4034.

[8] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1237–1246.

[9] Jia-Chang Feng, 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*, 2021, pp. 14009–14018.

[10] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.