REAL-WORLD VIDEO ANOMALY DETECTION BY EXTRACTING SALIENT FEATURES

Yudai Watanabe[1†], *Makoto Okabe*[1], *Yasunori Harada*[2] *and Naoji Kashima*[2]

[1]Shizuoka University [2]Chubu Electric Power Co., Inc.

ABSTRACT

We propose a lightweight and highly accurate method for detecting anomalies in videos. Existing methods use multiple-instance learning (MIL) to determine the normal/abnormal status of each segment of the video. Recent successful researches argue that it is important to learn the temporal relationships among segments to achieve high accuracy, instead of focusing on only a single segment. We analyzed the existing methods that have been successful in recent years, and found that while it is indeed important to learn all segments together, the temporal relationships among them are irrelevant to achieving high accuracy. Based on this finding, we do not use the MIL framework, but instead introduce a self-attention mechanism to automatically extract features that are important for determining normal/abnormal from all input segments. As a result, the neural network with 1.3% of the number of parameters of the existing method can achieve the comparable or better accuracy than the existing method.

Index Terms— Weakly supervised anomaly detection

1. INTRODUCTION

The number of surveillance cameras in the world is increasing every year, and they are used for crime prevention in cities and for safety confirmation in factories, power plants, and other large-scale facilities. However, since it is difficult for humans to see and confirm all of these videos, there is an urgent need to develop technology that enables artificial intelligence to analyze the videos and automatically detect abnormal events on behalf of humans. Since abnormal events are rarely observed, many methods have been proposed that use only the normal state as training data, and judge whether the input video is normal or abnormal based on the criterion of how much it deviates from the learned normal state when inferring [1, 2, 3, 4, 5]. However, these methods can only detect abnormalities based on low-level features such as differences in the appearance and velocity in the video. Therefore, recently, a method has been proposed to train an anomaly detector using a weakly supervised dataset that contains both normal and abnormal videos [6, 7].

In the weakly supervised dataset, each video is labeled as normal or abnormal. That is, the videos labeled as normal contain only the normal state throughout all frames. On the other hand, the videos labeled as abnormal contain a mixture of normal and abnormal frames. By using such a dataset, it is no longer necessary to label each frame in the video as normal or abnormal, thus reducing the labeling effort.

Many existing methods treated multiple consecutive frames as a single short-term segment and used multiple instance learning (MIL) to determine normal/abnormal for each segment[6]. However, recent successful methods have argued that it is important to learn the temporal relationships among segments by taking all segments from the video as input together, rather than focusing on only a single segment [7]. To confirm this, we analyzed these methods on a dataset of video segments randomly sorted in time. We then found that while it is indeed important to train all segments together, the temporal relationship among them is irrelevant for high accuracy.

Based on this finding, we propose a novel model that does not rely on MIL, but instead takes all segments as input and has a self-attention mechanism to automatically extract features important for determining normal/abnormal from them. Despite the fact that the proposed neural network has 1.3% of the number of parameters of the existing method [7], the proposed method can achieve the comparable or better accuracy than the existing method [7]. We report the frame-level detection accuracy using benchmark datasets (UCF-Crime, ShanghaiTech).

2. RELATED WORK

In the real world, most of what we can observe are normal states, and abnormal events are rarely observed. For this reason, many anomaly detection methods have been developed using unsupervised learning approaches that learn only the normal state. In inference, the input video is judged to be normal or abnormal based on how much it deviates from the learned normal state. The normal state is learned using Gaussian mixture models [1], sparse dictionary learning [8], autoencoders [2, 3, 4, 9], etc. A method that uses multi-task learning has also been proposed [10].

Recently, a number of methods have been proposed to learn anomaly detectors using weakly supervised datasets containing both normal and abnormal videos [6, 7, 11, 12, 13, 14, 15]. Usually, to train a frame-by-frame anomaly detector, we need to label and train every frame of the video, which

[†]Current affiliation: Panasonic System Networks R&D Lab. Co., Ltd.

is expensive to label. Sultani et al. developed an anomaly detector by introducing MIL, which considers a video as a bag and selects the segment with the highest anomaly score from the bag for training [6].

Most of the anomaly detection methods using weakly supervised datasets are based on MIL. Zaheer et al. proposed a method for detecting anomalies using global features of the entire video and local features of each segment using the attention mechanism and per-video clustering loss [14]. Tian et al. proposed a top-k strategy that calculates the difference in anomaly scores between segments and selects the top k segments with the highest scores for training [7]. These recent successful methods achieve high detection accuracy by efficiently using the features of multiple segments in the video instead of just a single segment.

3. METHOD

We propose a lightweight and highly accurate learning method for detecting anomalies in videos. The proposed method analyzes the entire video and automatically extracts and learns the features that are important for determining normal/abnormal.

Let $\mathcal{D} = \{(\mathbf{V}_i, y_i)\}$ be the dataset. where $\mathbf{V}_i$ is the ith video in the training dataset and y_i is the label attached to $\mathbf{V}_i$. $y_i = \{0, 1\}$, 0 indicates normal, and 1 indicates abnormal. In a video labeled as normal, only the normal state is recorded in all frames. On the other hand, the video labeled as abnormal contains a mixture of frames with normal and abnormal states. $\mathbf{V}_i$ is divided into T segments, and each segment is converted into a D-dimensional feature vector $\mathbf{F}_{i,j}$ by the feature extractor: $\mathbf{F}_{i,j}$ represents the jth feature vector of $\mathbf{V}_i$; the feature extractor used throughout all experiments was I3D [16], which has been trained on the Kineticts dataset.

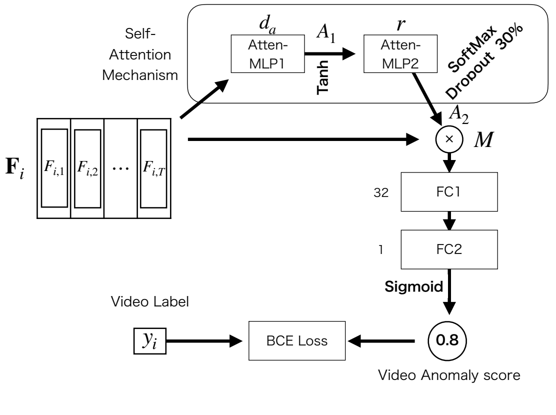

Fig. 1. An overview diagram of our method

Our method is a simple model that consists of a self-attention mechanism and two fully connected layers (Fig. 1). In the self-attention mechanism, the input $\mathbf{F}_i$ is transformed by Atten-MLP1 multilayer perceptron into a $d_a \times T$ matrix $\mathbf{A}_1$. At this time, the Tanh function is used for activation. $\mathbf{A}_1$ is converted to $\mathbf{A}_2$, a matrix of $r \times T$, by Atten-MLP2 multilayer perceptron. At this time, the softmax function is used for activation. In addition, a dropout regularization of 30% is also performed.

Using the weight matrix $\mathbf{A}_2$ obtained from the self-attention mechanism, we calculate $\mathbf{M} = \mathbf{F}_i \mathbf{A}_2^t$. After reshaping $\mathbf{M}$ into a vector of $D \times r$ dimensions, it is transformed into the anomaly score through two fully connected layers FC1 (32 units) and FC2 (1 unit). The activation function of FC1 is an identity function, and that of FC2 is a sigmoid function. The binary cross entropy (BCE) function was used as the loss function.

3.1. Motivation of our method

Several existing methods take into account the temporal relationships between segments in a video [6, 12, 7]. Sultani et al. introduced a term in the loss function to impose continuity of the anomaly score, as the anomaly score should vary continuously in the video [6]. Tian et al. introduced a multi-scale temporal network (MTN) to capture local and global temporal features [7]. Both of these methods achieve high accuracy in anomaly detection. Therefore, we investigated how the mechanism for capturing the temporal relationship between segments contributes to the high accuracy.

Specifically, in $\mathbf{F_i}$, the set of feature vectors obtained from the training video $\mathbf{V_i}$, the feature vectors are ordered by default as $\{\mathbf{F}_{i,1}, \mathbf{F}_{i,2}, \cdots, \mathbf{F}_{i,T}\}$, but we randomly rearranged this order to create a new dataset and used it for training. The results of the experiments using the UCF-Crime dataset [6] are shown in Table1. In both methods [6, 7], there was no degradation in accuracy due to random reordering of feature vectors.

Method	Reorder	AUC(%)
Sultani et al. [6]		81.39
	✓	81.54
RTFM [7]		84.30
	✓	84.26

Table 1. Comparison of AUC performance using the UCF-Crime dataset [6], where the feature vectors of each video are randomly reordered.

This result indicates that capturing the temporal relationship between segments does not contribute to the accuracy of the anomaly detector. On the other hand, both Sultani et al.'s term in the loss function [6], which imposes temporal continuity, and MTN or top-k strategy [7] in Tian et al.'s method, have the effect of encouraging more feature vectors to be involved in the training process. In MIL, only a small number of selected feature vectors can be involved in the training process, so it is likely that these systems are trying to efficiently extract the features that are important for determining normal/abnormal. Based on the above observations, we hypothesized that the high anomaly detection accuracy achieved by these methods is due to the fact that they have a mechanism that allows all segments from the video to be trained together and extracts salient features from them efficiently.

Our method (Fig. 1) is designed based on the above insights. Our method is not an MIL framework but it is a model that takes all segments in a video as input and determines whether the video is normal or abnormal. Since we do not use MIL, the extraction of salient features can be achieved with a simple self-attention mechanism. For the self-attention mechanism, we introduce a model inspired by Lin et al.'s method [17]: their method targets sentence classification and can deal with variable length input. We adopted this mechanism because we divide the video $\mathbf{V}_i$ into T segments during training, but the number of segments during inference should be different dependent on the length of the input video. Our method is accurate and also lightweight because it does not have any mechanism to capture temporal relationships.

3.2. Inference

Let $\mathbf{V}^e$ be the video to which we want to apply our method for anomaly detection. First, we divide $\mathbf{V}^e$ into N segments $\{\mathbf{V}^e_1, \mathbf{V}^e_2, \cdots, \mathbf{V}^e_N\}$ using 16 consecutive frames as one segment. Each segment $\mathbf{V}^e_i$ is converted by I3D [16] into a D-dimensional feature vector $\mathbf{F}^e_i$. Let the set of feature vectors be $\mathbf{F}^e = \{\mathbf{F}^e_1, \mathbf{F}^e_2, \cdots, \mathbf{F}^e_N\}$. Let l be the split size and we divide $\mathbf{F}^e$ into $m = N/l$ bags for every l segments:

$$\begin{aligned}&\mathbf{F}^e = \{\mathbf{B}_1, \mathbf{B}_2, \cdots, \mathbf{B}_m\}\\ &= \{\{\mathbf{F}^e_1, \cdots, \mathbf{F}^e_l\}, \{\mathbf{F}^e_{l+1}, \cdots, \mathbf{F}^e_{2l}\} \cdots \{\mathbf{F}^e_{N-l+1}, \cdots, \mathbf{F}^e_N\}\}.\end{aligned}$$

Next, we input $\mathbf{B}_1, \mathbf{B}_2, \cdots, \mathbf{B}_m$ one by one into our anomaly detector, and we obtain the inference result $\mathbf{S}^v = \{s^v_1, s^v_2, \cdots, s^v_m\}$. s^v_i is used as the anomaly score for each segment of $\{\mathbf{V}^e_{(i-1)l+1}, \cdots, \mathbf{V}^e_{il}\}$. When it is necessary to produce the results of frame-level anomaly detection, the obtained scores are assigned to all frames in each segment.

4. EXPERIMENTS

To evaluate our method, we conducted experiments using two weakly supervised datasets for anomaly detection (UCF-Crime dataset [6] and ShanghaiTech dataset [5]).

4.1. Dataset and evaluation measure

UCF-Crime dataset contains 13 types of anomalies. The total number of videos is 1900, and the total duration of the videos is 128 hours. Of the 1900 videos, 1610 are training videos and 290 are test videos. Each training video is labeled as normal or abnormal for each video. Each test video is labeled as normal or abnormal for each frame.

ShanghaiTech dataset contains 437 videos captured by 13 fixed cameras. Of the 437 videos, 307 are normal videos and 130 are videos with anomalies. The original dataset was intended for the development of an anomaly detector based on unsupervised learning. However, Zhong et al. labeled each video so that it could be used as a dataset for weakly supervised learning [13]. We constructed a weakly supervised dataset using the same procedure as Zhong et al.[13] and conducted experiments.

Evaluation measure: we used the Area Under the Curve (AUC) with the Receiver Operating Characteristic (ROC) curve, which is calculated based on the frame-level anomaly detection accuracy, as in the existing studies [6, 11, 7, 13, 14, 15]. A larger AUC value indicates a more accurate anomaly detector.

4.2. Implementation detail

Our method was developed and evaluated using PyTorch. Radam [18] was used as the optimization algorithm, and the learning rate was set to 0.001. The batch size was set to 64. As in the existing method [6], mini-batches were created so that each mini-batch contained an equal number of normal and abnormal videos. The hyperparameter T was set to 32. As in the existing method [7, 15], 10-crop augmentation was performed for each video.

4.3. Results on UCF-Crime

Table 2 shows the frame-level AUC performance on the UCF-Crime dataset. The split size l is set to 28. Although our method is a simple model, it achieves 0.61% higher detection accuracy than RTFM [7], which has the highest accuracy among the existing methods.

4.4. Results on ShanghaiTech

Table 3 shows the frame-level AUC performance on the ShanghaiTech dataset. The split size l is set to 21. Our is 1.49% inferior to RTFM [7], which has the highest accuracy among the existing methods, but our method outperforms the other existing methods. In addition, the accuracy of more

Method	Feature Type	AUC(%)
Sultani et al. [6]	C3D RGB	75.41
GCN-Anomaly [13]	TSN RGB	82.12
CLAWS Net [14]	C3D RGB	83.03
Wu et al. [11]	I3D RGB	82.44
MIST [15]	I3D RGB(Fine)	82.30
RTFM [7]	I3D RGB	84.30
Ours ($d_a = 64, r = 3$)	I3D RGB	84.74
Ours ($d_a = 128, r = 7$)		84.91

Table 2. Comparison of frame-level AUC performance on UCF-Crime dataset.

Method	Feature Type	AUC(%)
GCN-Anomaly [13]	TSN RGB	84.44
CLAWS Net [14]	C3D RGB	89.67
RTFM [7]	I3D RGB	97.21
Ours ($d_a = 64, r = 3$)	I3D RGB	95.72

Table 3. Comparison of frame-level AUC performance on ShanghaiTech dataset.

than 95% is achieved, indicating that a sufficiently practical anomaly detector can be trained.

4.5. Comparison of the number of trainable parameters

Table 4 shows the number of parameters that can be trained in the model for each method. Our method is an extremely lightweight model with a much smaller number of parameters than the existing methods. Even though the number of parameters is only 1.3% of RTFM [7] when $d_a = 64$ and $r = 3$, our method achieves higher accuracy than RTFM [7] in Table 2, and achieves a comparable high accuracy in Table 3. It is also the lightest model among the existing methods even when $d_a = 128$ and $r = 7$, which achieves even higher accuracy.

Method	Number of Parameters
Sultani et al. [6]	2,114,113
Wu et al. [11]	769,155
RTFM [7]	24,718,849
Ours ($d_a = 64, r = 3$)	328,004
Ours ($d_a = 128, r = 7$)	721,992

Table 4. Comparison of the number of trainable parameters.

4.6. Analysis on the hyperparameters d_a and r

We defined d_a and r as hyperparameters in Sec. 3. By changing the hyperparameters, the number of trainable parameters changes, and the detection accuracy also changes. Therefore, using the UCF-Crime dataset, we investigated how changes in hyperparameters affect the detection accuracy. The split size l was set to 32. The results are shown in Fig. 2. The points surrounded in red represent the highest detection accuracy for each d_a. Focusing only on r, if r is larger than 3, detection accuracy is almost stable and its variation is small. Focusing on the relationship between d_a and r, if d_a is increased, r should also be increased to some extent to obtain better detection accuracy. However, if r is made too large, the detection accuracy will decrease. Up to 256, better detection accuracy could be obtained by increasing d_a, but when d_a was increased to 512, the overall detection accuracy decreased regardless of the value of r. This may be due to overfitting caused by making the model too large.

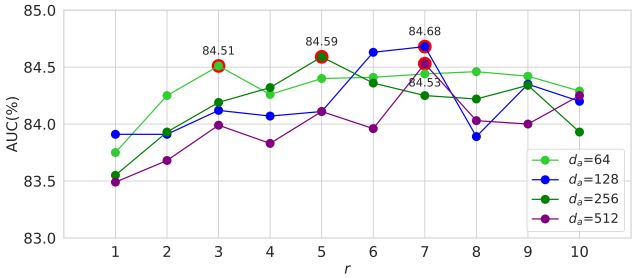

Fig. 2. Relationship between hyperparameters (d_a and r) and AUC performance on the UCF-Crime dataset.

4.7. Analysis on the split size l

We defined the split size l in Sec. 3.2. Since our method divides N feature vectors into $m = N/l$ bags during inference, each bag contains l feature vectors. Since the detection accuracy varies depending on the split size l, we used the UCF-Crime dataset to investigate how the change in l affects the detection accuracy. The result for $d_a = 64$ and $r = 3$ and the result for $d_a = 128$ and $r = 7$ are shown in Fig. 3. The points surrounded in red represent the highest detection accuracy for each hyperparameter set. The detection accuracy is increased until l is around 16. Since our method analyzes the entire video, it requires a certain length of the video, which means that a certain split size is necessary. On the other hand, when the split size l is larger than 16, the detection accuracy is stable, indicating that our method can analyze the whole video efficiently if the video is long enough.

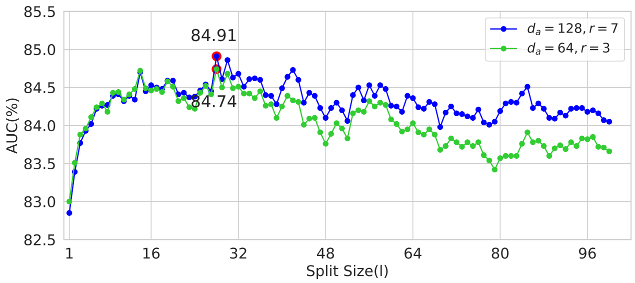

Fig. 3. Relationship between the split size l and AUC performance during inference on the UCF-Crime dataset

5. CONCLUSION

We have proposed a lightweight and accurate weakly supervised learning method for anomaly detection from video. Since MIL is not used, the extraction of salient features can be achieved with a simple self-attention mechanism. We show that the proposed model is simple and lightweight, yet achieves the comparable or better accuracy than the existing method.

6. REFERENCES

[1] Arslan Basharat, Alexei Gritai, and Mubarak Shah, "Learning object motion patterns for anomaly detection and improved object detection," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[2] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1705–1714.

[3] Trong-Nguyen Nguyen and Jean Meunier, "Anomaly detection in video sequence with appearance-motion correspondence," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1273–1283.

[4] 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.

[5] 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.

[6] 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.

[7] Yu Tian, 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*, 2021.

[8] 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.

[9] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao, "Object-centric auto-encoders and dummy anomalies for abnormal event detection in video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7842–7851.

[10] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah, "Anomaly detection in video via self-supervised and multi-task learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12742–12752.

[11] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang, "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *European Conference on Computer Vision*. Springer, 2020, pp. 322–339.

[12] 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.

[13] 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.

[14] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee, "Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 358–376.

[15] 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.

[16] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[17] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio, "A structured self-attentive sentence embedding," *arXiv preprint arXiv:1703.03130*, 2017.

[18] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han, "On the variance of the adaptive learning rate and beyond," *arXiv preprint arXiv:1908.03265*, 2019.