A Contrastive Learning Framework for Weakly Supervised Video Anomaly Detection

Hyeon Jeong Park, * Je Hyeong Hong
Department of Electronic Engineering, Hanyang University
hjung4337@hanyang.ac.kr, *jhh37@hanyang.ac.kr

Summary

Weakly-supervised learning is a widely adopted approach in video anomaly detection whereby only video labels are utilized instead of expensive frame-level annotations. Since the success of multi-instance learning (MIL), almost all recent approaches are based on maximizing the margin between the set of abnormal video snippets and those of normal video snippets. In this work, we present a simple contrastive approach for weakly supervised video anomaly detection (WS-VAD) with aims to enhance the performance of existing models. The method is generic in nature and introduces a loss function to encourage attraction of output features from the same video class and repel those from different video classes. Experimental results demonstrate our method can be applied to existing algorithms to improve detection accuracy in public video anomaly dataset.

1. Introduction

In recent years, CCTV cameras have been extensively deployed to enforce public security in various countries. However, analyzing and monitoring CCTV videos still mostly rely on expensive manual efforts requiring human interventions. Even with multiple monitoring personnel the task is difficult as i) anomalous actions occur only rarely and ii) there are usually several CCTV streams to monitor simultaneously. Consequently, the demand for intelligent surveillance systems is increasing.

Currently, video anomaly detection (VAD) is one of the most widely-used approach for solving aforementioned task. It aims to detect unusual patterns, appearances, or situations (such as violence, car accident, and explosion) from RGB and optical flow data obtained from CCTV videos. While the use of videos can benefit from the underlying temporal information, it also stands a major challenge that frame-level annotations are generally missing, which may not provide enough cues for training an accurate VAD model [2,3].

To tackle above problem, recent studies have focused on developing weakly-supervised approaches which do not require frame-level ground truths. One of the foundational algorithms is multi-instance learning (MIL), which considers VAD as a regression problem estimating an abnormal score for each snippet (set of frames) of video. It aims to maximize abnormal scores of abnormal snippets than those of normal snippets. To this end, only one snippet with the highest abnormal score is selected from each anomalous and normal video, then they construct a positive and a negative bag, respectively. Then, the margin between the instances of the positive bag and those of the negative bags is maximized. Since it has shown significantly enhanced performance in weakly-supervised settings, most existing methods are based on it.

In this paper, we present a simple contrastive approach for WS-VAD not only to attract features of the same class but to repel features of the different classes to learn discriminative representations for the clear decision boundary between anomalous and normal situations. It is worth noting that this contrastive loss can be applied to an off-the-shelf training scheme. We apply the self-supervised-based contrastive loss proposed by Supcon loss [5].

The rest of the paper is organized as follows: Section 2 presents the related works on WS-VAD. Section 3 provides a detailed description of the simple contrastive framework for WS-VAD proposed here, and Section 4 presents the experimental setup and results. Finally, Section 5 concludes our study.

2. Related work

2.1 Weakly-supervised Video Anomaly Detection

Most VAD methods earlier than 2017 were based on unsupervised settings that utilized only normal training videos. For instance, the methods that used sparse representation to learn the dictionary of normal behaviors were mainly proposed [1]. In this case, the patterns which have large reconstruction errors are considered anomalies during testing. After this, [2] proposed a deep MIL ranking loss to predict anomaly scores by leveraging the abnormal samples with video-level labels. Technically, the MIL ranking loss is designed to maximize the anomaly score of abnormal instances compared to normal instances. As they showed substantially improved performance over the unsupervised approaches, most existing WS-VAD methods are based on this loss. On the other hand, [2] proposed robust temporal feature magnitude learning to recognize anomalous situations efficaciously. They considered the feature magnitudes more powerful indicators than anomaly scores for effective training. Therefore, they designed the loss function that enables the feature magnitudes of the abnormal instances to be larger than those of the normal instances. However, the usefulness of the feature similarity between abnormal and normal instances has been relatively disregarded in the previous works. In this work, we present a simple contrastive approach based on the feature similarities with an aim to enhance the performance of existing models.

2.2. Contrastive Learning

Recently, contrastive learning has greatly facilitated in various computer vision tasks such as object classification and action recognition. It allows a model to learn the discriminative features by not only pulling the representations of the same instance close but pushing those of different instances far away. [4] suggested the simple framework for contrastive learning, that generates different views of an image - meaning of positive pairs -through augmentation and achieved outstanding performance in object classification in a self-supervised manner. However, since it does not use ground truth labels at all, it has the disadvantage of learning to repel each other even if there are images of the same class among the batch data. To mitigate this problem, [5] proposed supervised contrastive loss that utilized the ground truth labels when constructing positive pairs. Since it is a way to perform more accurate pairing rather than learning to classify labels, we adopt this function.

3. Proposed Method

The main purpose of the proposed method is to demonstrate that contrastive loss can be effectively applied to the existing WS-VAD algorithms such as MIL framework as shown in Fig. 1.

3.1 Formulation

The purpose of anomaly detection is to estimate the anomaly status of a video and localize the anomalies in the video sequence if exist. In the weakly supervised scenario, a video V and its corresponding video-level annotation $y \in \{0,1\}$ are given, where the case $y = 1$ means there exists an anomaly (V_a) otherwise $y = 0$ indicates that there is no anomaly (V_n). Usually, each video is divided into 32 snippets according to [2]. Then, the features of those snippets are extracted by using the pretrained models with action recognition databases, $\mathbf{X}=\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{32}\}$.

3.2 MIL with contrastive learning

MIL assumed that abnormal snippets should have higher abnormal scores than those normal snippets. As frame-level labels are not available in WS-VAD, the top-1 scores are selected from each normal and abnormal video for comparison as follows:

$$\max_i f(\mathbf{X}_a^i) > \max_i f(\mathbf{X}_n^i) \quad (1)$$

where i means the index of mini-batch and $f(\cdot)$ is the MIL model. Specifically, to distinguish the abnormal and normal instances, Eq. (1) is modified to the hinge-loss formulation as below:

$$L_{mil} = \max\left(0, 1 - \max_i f(\mathbf{X}_a^i) + \max_i f(\mathbf{X}_n^i)\right) \quad (2)$$

In practice, since a model usually has the sigmoid operation after the last layer, the range of the outputs is restricted from zero to one. Hence, it is possible to calculate the derivatives of L_{mil}.

However, maximizing the margin between the set of abnormal video snippets and those of normal video snippets cannot assure that the model can learn discriminative features, since normal and abnormal features can lie at any

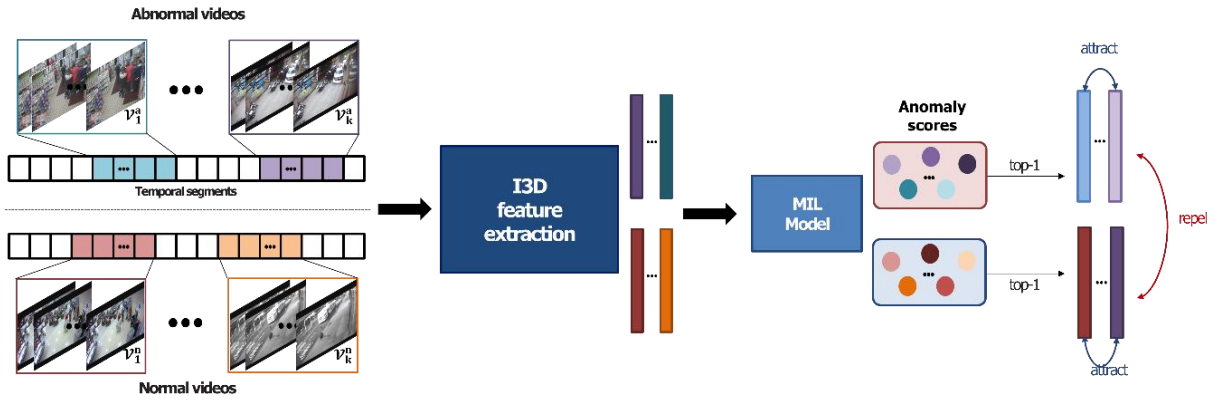

Figure 1. The overall framework of the proposed method. After abnormal and normal videos are divided into 32 segments, the features are extracted by I3D model pretrained with action recognition datasets. Then, MIL model which is the MLP with three layers predicts the anomaly scores of each snippet and scores and the features that have the highest anomaly scores are selected. The MIL model is trained by MIL ranking loss and contrastive loss.

place in the feature space. Therefore, we apply the contrastive loss with an aim to attract the features of the same video classes and take apart the features of the different video classes for a clearer decision boundary. For contrastive learning, we construct a positive feature set including all abnormal categories to learn generalized features for anomalies, since there are diverse abnormal events in the real world. After the features that have the highest anomaly scores are selected to construct the positive and negative pairs, contrastive learning is performed by Supcon loss:

$$L_{supcon} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \frac{exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{a \in A(i)} exp\,(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)} \qquad (3)$$

Where $P(i)$ and $A(i)$ means positive pair sets and negative pair sets, respectively. $\mathbf{z}$ is an embedding feature of x extracted from MIL model and τ is a temperature parameter.

Since Supcon loss mainly perform metric learning, it cannot assure the abnormal snippet has a higher anomaly score than that of normal snippet. Therefore, we simply combine the MIL ranking loss and Supcon loss with hyper-parameter for balancing them as follow:

$$L_{total} = L_{mil} + \lambda\, L_{supcon} \qquad (4)$$

4. Experimental Results

4.1 Experimental Settings

We test our proposed method on the UCF-Crime dataset which is a large-scale dataset with a total of 128 hours. It contains 1,900 untrimmed actual surveillance videos with 13 classes of abnormal events [2]. The training set includes 1,610 videos with video-level labels, and the test set consists of 290 videos with frame-level labels. We used the features extracted from pretrained I3D [6] with a Kinetic database. We employed the same model with [2] that has three layers (512-32-1). Then, we trained the model for 75 epochs with Adam optimizer. As a hyper-parameter setting, 32 videos were selected from each abnormal and normal video for constructing a mini-batch, and the initial learning rate is 0.001. Lastly, the lambda (λ) for balancing between MIL loss and Supcon loss and the temperature of SupCon loss is set to 0.4 and 0.07, respectively. Similar to the previous methods, we use the frame-level area under the ROC curve (AUC) as the evaluation metric. A larger AUC indicates better performance.

4.2 Experimental Results

The AUC results on UCF-Crime are illustrated in Table 1. The original paper of MIL [2] used the 4096-d features extracted by C3D [7] which utilizes only RGB frames during training action classes. Whereas we used the features extracted by I3D that showed significantly improved action recognition performance by utilizing RGB and optical flow images since C3D features that were utilized in [2] are not available. Therefore, we also trained the original MIL ranking loss with the I3D features for a fair comparison.

Consequently, we obtained an AUC of 85.16%,

which was improved by 0.85% over the existing method. On the other hand, 84.3% of [3] is a result of using the relatively complex model to consider temporal information from Resnet-based I3D features by using only RGB frames. However, we showed slightly better performance with a simple architecture.

Table 1. Performance comparison on UCF-Crime

Methods	AUC (%)
Original MIL [2]	75.41
Retrained MIL [2]	84.31
RTFM [3]	84.30
Ours	**85.16**

5. Conclusion

We present a simple contrastive approach for weakly supervised video anomaly detection (WS-VAD) to improve performance of existing models. We added a contrastive loss function to encourage attraction of positive pairs meaning that have the same video classes and repel negative pairs meaning that have different video classes. The experimental results on the public video anomaly dataset demonstrated our method can be applied to existing algorithms to improve detection accuracy.

Acknowledgement

This work was supported in part by Institute of information and Communications Technology Planning and Evaluation (No.2020-0-01373, Hanyang University, Department of Artificial Intelligence) funded by the government (Ministry of Science and ICT) in 2022, in part by Samsung Advanced Institute of Technology, Samsung Electronics Co., Ltd.

Reference

[1] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2013.

[2] 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 (CVPR), June 2018.

[3] 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 (ICCV), pages 4975-4986, October 2021.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daum´e III and Aarti Singh, editors, Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 1597-1607. PMLR, 13-18 July 2022.

[5] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems (NeuRIPS), volume 33, pages 18661-18673. Curran Associates, Inc., 2020.

[6] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4724-4733, 2017

[7] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In 2015 International Conference on Computer Vision, ICCV 2015, Proceedings of the IEEE International Conference on Computer Vision, pages 4489-4497. Institute of Electrical and Electronics Engineers Inc., Feb. 2015. 15th IEEE International Conference on Computer Vision, ICCV, 2015.