감시 동영상에서의 시간 모듈을 통한 약한 지도 학습 이상 감지

*임희정, 설상훈
고려대학교 전기전자공학부
e-mail : *hjlim@mpeg.korea.ac.kr*, *sull@korea.ac.kr*

Weakly-supervised Anomaly Detection with Temporal module in Surveillance Videos

*Hee-Jeong Lim, Sang-Hoon Sull
School of Electrical Engineering
Korea University

Abstract

In the anomaly detection in surveillance videos, the temporal change is one of the important concepts. To this end, this paper proposes anomaly detection network that considers various relationships along the time axis. In particular, the classifier considering the temporal order calculates anomaly scores more accurately. We experiment our network with the widely used ShanghaiTech dataset and show improved performance result.

I. 서론

감시 동영상에서의 딥러닝을 이용한 이상 감지 연구는 이를 필요로 하는 다양한 산업에서 인력을 대체할 수 있어 최근 활발히 연구되고 있다[1,2,3]. 감시 동영상의 이상 감지의 목표는 동영상 내의 비정상적인 모양, 모션 등의 이상 징후가 나타나는 부분을 알아내는 것이다. 다양한 특성이 이상 징후가 될 수 있기 때문에, 이를 전부 고려하는 모델을 설계하는 것은 어려운 문제이다. 하지만, 동영상에서 어떤 이상이 발생하던지 시간에 따라 정상 프레임들과 큰 차이가 발생한다. 따라서, 최근 많은 연구들이 시간 축으로의 변화를 고려한 방법들을 연구하고 있다[4,5,6].

많은 연구 중 [4]은 과거와 미래에 대해 dilated convolution으로 지역적 정보를, self-attention으로 전역적 정보를 추출하여 clip-level classifier가 이상 프레임들을 판별한다. 하지만, clips 간의 시간 순서가 중요한 이상은 판별하기 어렵다. 예시로, 직진 차선에서 후진을 하는 경우를 들 수 있다.

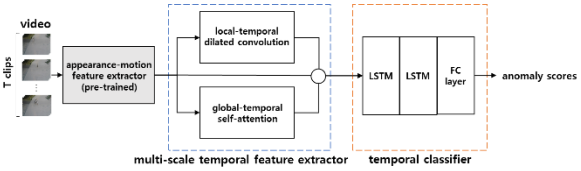
따라서 본 논문에서는 시간 순서를 고려한 temporal classifier을 제안함으로써 [4]의 방법을 보완한다. 제안 모듈은 과거 프레임들에 대한 feature를 입력으로 현재 프레임의 이상 정도를 알아내는 LSTM classifier로 구성되어 있다. 실험에서 시간 순서를 고려한 우리의 방법을 ShanghaiTech 데이터 셋을 통해 향상된 성능을 보임을 확인했다.

II. 본론

2.1 네트워크 구조

본 논문이 제안하는 네트워크 구조는 그림1과 같다. Action recognition에서 학습된 I3D[7]를 통해 영상을 일정 간격으로 겹치지 않게 나눈 T개의 clips에 대한 pre-trained feature들을 전처리로 구하고 이를 입력으로 사용한다. Dilated convolution과 self-attention을 통해 pre-trained feature들로부터 지역

적, 전역적 시간 변화가 고려된 $\mathbb{R}^{T\times D}$ temporal feature들을 추출한다[1]. 이를 입력으로 temporal classifier를 통해 $\mathbb{R}^{T\times 1}$개의 이상 scores를 얻는다.



2.2 Temporal classifier

본 논문에서는 시간 순서도 고려하여 이상을 잘 판별하기 위한 temporal classifier를 제안한다. 이는 2개의 LSTM들을 쌓은 stacked LSTM과 하나의 FC layer로 구성되어 있다. Classifier는 T개의 step으로 이루어져 있으며, 한번의 step에서 D차원 feature를 입력 받고 해당 시점의 score를 출력한다.

$$score_i = f_\phi(s_\theta(F)_{\le i}), \qquad 1 \le i \le T \qquad (1)$$

f_ϕ 는 temporal classifier, s_θ 는 multi-scale temporal feature extractor, F 는 pre-trained features, i 는 step이다.

2.3 Score margin loss

Clip들 간의 feature들을 독립적으로 판별하여 score들을 얻는 MLP와 달리 LSTM은 clip들의 feature를 연속적으로 입력 받음으로써 순서 관계를 고려한다. 하지만, 이로 인해 이상 score가 smooth한 경향이 크다. 따라서 정상과 이상 score 차이를 벌리기 위해 [1]의 total loss에 score margin loss를 추가하였다.

$$diff = \frac{1}{k}\sum_{i\epsilon M_k} score_i - \frac{1}{k}\sum_{i\epsilon m_k} score_i \qquad (2)$$

$$L_{sm} = \begin{cases} \max(0, 1-diff) & ,if \\ 0 & , \end{cases} \qquad (3)$$

y 은 video-level 레이블이며 1일때 이상 레이블을 의미한다. M_k는 s_θ에서 얻은 temporal feature들 중 feature magnitude가 큰 순으로 k 개이며 m_k는 작은 순으로 k 개를 의미한다.

Ⅲ. 실험 결과

실험 환경은 TITAN RTX TU102에서 진행하였다. 데이터 셋은 weakly-supervised setting으로 변형한 ShanghaiTech 데이터 셋으로 실험을 진행하였다. 제안 방법의 stacked LSTM classifier은 baseline[4]의 classifier의 3개의 fc layers 중 앞 2개 layer들을 동일한 사이즈의 LSTM layer로 대체하였다. 또한, dropout layer와 ReLU activation 모두 그대로 사용하였다. Baseline의 오픈 코드의 구현과 성능이 해당 논문과 달라, 약간의 수정을 거친 후, 논문 및 코드에 공개된 모든 하이퍼파라미터를 그대로 사용했다. 제안 방법은 baseline과 같은 하이퍼파라미터를 사용했다.

표 1은 baseline과 제안 방법의 AUC 성능 비교이다. 성능의 편차가 커서 150 epoch을 5번 돌렸을 때 각각의 최고 수치를 평균한 값이다. Score margin loss는 기존의 gradient 크기와 비슷하도록 learning late를 0.00067로 주었다.

Baseline[4]	93.31
Baseline+Loss	93.57
Stacked LSTM	94.30
Stacked LSTM +Loss	94.36

표 1. baselines과의 AUC (%) 성능 비교

Ⅳ. 결론 및 향후 연구 방향

본 논문에서는 감시 동영상에서의 이상 감지를 위해 시간 순서를 고려하는 방법을 제안하였다. 제안하는 방법은 T개의 clips에 대한 temporal feature들을 시간 순서를 고려한 temporal classifier로 이상 여부를 분류하였다. 또한, 이를 잘 학습하기 위하여 이상과 이상이 아닌 clips에 대한 scores를 벌리도록 하는 loss를 제안하였다. 향후 연구에서는 감시 동영상을 이상 부분과 정상 부분을 모두 잘 표현하는 동영상 요약 기술을 볼 예정이다.

감사의 글

본 연구는 방위사업청과 국방과학연구소의 지원(UD190031RD)으로 한국과학기술원 미래 국방 인공지능 특화연구센터에서 수행되었습니다.

참고문헌

[1] Feng, Jia-Chang, Fa-Ting Hong, and Wei-Shi Zheng. "Mist: Multiple instance self-training framework for video anomaly detection." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

[2] Georgescu, Mariana-Iuliana, et al. "Anomaly detection in video via self-supervised and multi-task learning." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

[3] Lu, Yiwei, et al. "Few-shot scene-adaptive anomaly detection." *European Conference on Computer Vision*. Springer, Cham, 2020.

[4] Tian, Yu, et al. "Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning." *arXiv preprint arXiv:2101.10030* (2021).

[5] Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[6] Zhong, Jia-Xing, et al. "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

[7] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017.