Anomaly LVLM: LVLM을 활용한 사용자 맞춤형 비디오 이상 탐지 연구

안성현[O] 조영완 이기정 권세인 박상현[†]
연세대학교 컴퓨터과학과
{skd, jyy1551, rlwjd4177, seinkwon97, sanghyun}@yonsei.ac.kr

Anomaly LVLM: Customizable Video Anomaly Detection using Large Vision-Language Model

Sunghyun Ahn[O] Youngwan Jo Kijung Lee Sein Kwon Sanghyun Park[†]
Dept. of Computer Science, Yonsei University

요 약

스마트시티에서 지능형 비디오 감시 시스템은 이상 현상을 자동으로 감지하여 사고를 예방하고, 사고 발생 시 신속한 대응을 가능하게 한다. 하지만 기존 비디오 이상 탐지 모델들은 단항 분류 기법을 통해 학습되므로 학습 데이터의 정상 패턴에 의존하여 다양한 상황에 적용하기 어렵다는 한계점이 존재한다. 이를 해결하기 위해 LVLM을 활용하여 범용적으로 적용할 수 있는 사용자 맞춤형 비디오 이상 탐지 프레임워크인 Anomaly LVLM을 제안한다. 제안된 방법론의 검증을 위해 Customizable-STC 데이터셋을 제작하였으며, 이를 통해 Anomaly LVLM의 우수성을 입증하였다.

1. 서 론

스마트 시티에서 지능형 비디오 감시 시스템은 대중의 생명과 자산의 안전을 보호하기 위해 CCTV 비디오에서 이상 현상을 정확하게 감지하는 것을 목표로 한다[1].

일반적으로, 딥러닝 기반의 비디오 이상 탐지(Video Anomaly Detection, VAD) 방법들은 이상치가 드물고 이를 명확하게 정의하기 어렵기 때문에, 단항 분류(One Class Classification, OCC) 기법을 사용한다. OCC 방법은 정상 샘플만을 학습하고, 학습된 정상 패턴과 차이가 있는 샘플을 비정상으로 식별하는 방식이다[2].

그러나 해당 방법은 학습 데이터의 정상 패턴에 의존하므로 다양한 상황에 적용하기 어렵다는 한계가 존재한다. 예를 들어, 보행자 구역에 대해 학습된 모델은 차량이 통행하는 도로 상황에서 활용되기 힘들다. 따라서 사용자는 차도와 같은 다른 상황을 추가적으로 학습하거나, 별도의 AI 모델을 새롭게 만들어야 한다. 이러한 과정은 머신러닝에 대한 전문 지식과 고성능 장비, 학습에 필요한 데이터 수집을 요구하기 때문에 일반 사용자가 VAD를 쉽게 활용하기 어렵게 만든다.

따라서 우리는 사용자 맞춤형 비디오 이상 탐지 기술(Customizable VAD, C-VAD)을 개발하여 이러한 한계를 극복하였다. 제안하는 방식은 사용자가 정의한 텍스트를 비정상 현상으로 간주하고, 비디오에서 해당 현상을 지닌 프레임을 탐지한다. 이를 위해 방대한 이미지-텍스트 데이터로 학습된 대형 비전 언어 모델(Large Vision Language Model, LVLM)을 활용하여 다양한 현상에 적용할 수 있는 Anomaly LVLM 프레임워크를 제시한다. 이 프레임워크는 별도의 학습 과정 없이도 일반 사용자들이 쉽게 사용할 수 있는 범용적인 VAD 모델로 활용될 수 있다.

또한, 우리는 단순히 LVLM을 활용하지 않고 CCTV 비디오 데이터의 특성을 고려하였다. CCTV 영상은 전경보다 배경이 더 많이 포함되어 있어 작은 객체에 대한 올바른 분석이 어려울 수 있다. 이를 해결하기 위해 Jeong[3]이 제안한 WinCLIP 모델을 기반으로, 배경보다 객체가 강조된 영상을 제작하는 전처리 기법을 적용하였다.

마지막으로 C-VAD 작업을 정량적으로 평가하기 위해 VAD 벤치마크인 Shanghai Tech Campus(STC)[4]를 가공하여 Customizable-STC(C-STC) 데이터셋을 제작하였으며, 제안된 프레임워크의 성능을 검증하였다. 본 논문의 기여는 다음과 같다.

- 다양한 비정상 상황에 적용 가능한 Anomaly LVLM 프레임워크를 제안한다.
- CCTV 비디오에서 작은 객체를 효과적으로 감지할 수 있는 WinCLIP 기반 어텐션 방법을 제안한다.
- C-VAD 작업의 정량적 평가를 위해 C-STC 데이터셋을 구축하고, 이를 통해 Anomaly LVLM의 우수성을 입증한다.

2. 본 론

제안하는 프레임워크의 구조는 그림 1과 같다. 사용자로부터 비디오와 텍스트(text)를 입력받아 프레임 단위로 작업을 수행한다. 먼저, WinCLIP 기반 어텐션(WA)을 통해 텍스트와의 유사도가 높은 영역을 강조하는 전처리를 진행한다. 이후, LVLM은 원본 프레임, 전처리된 프레임, 프롬프트를 이용하여 0과 1 사이의 이상 점수를 산출한다.

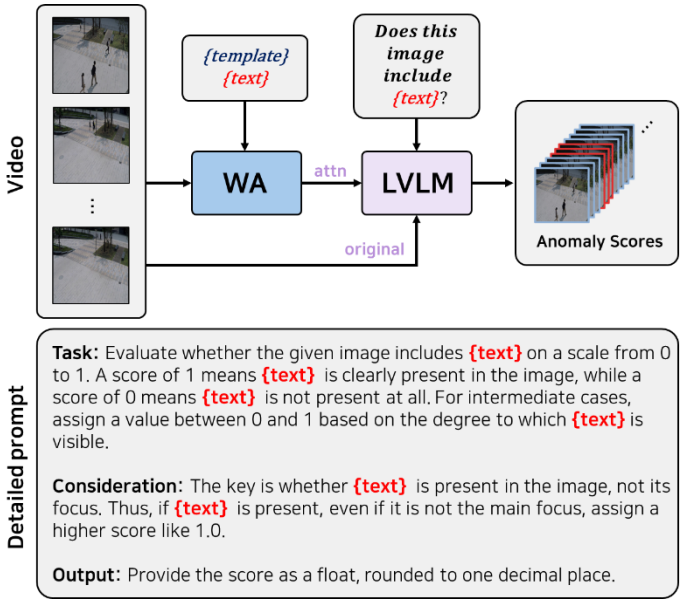

그림 1 Anomaly LVLM 프레임워크 구조

2.1 비디오 프레임 전처리

WA를 수행하는 방법은 그림 2와 같다. 입력 프레임을 여러 크기의 윈도우로 자르고, CLIP[5]의 이미지 인코더를 통과하여 윈도우 임베딩 맵들을 제작한다. 이러한 다중 스케일 맵들은 각각 프레임의 작은 영역과 큰 영역에 대한 특징들을 나타내고, 클래스 토큰은 프레임의 대표 특징을 나타낸다. 맵들은 각각 텍스트 임베딩과의 유사도가 계산되고, 집계되어 유사도 맵이 제작된다. 이 맵은 객체의 크기에 강건한 유사도를 가지며, 최종적으로 입력 프레임과 곱해져 텍스트와 관련된 영역을 강조할 수 있다.

$$f_i^{text} = \Phi_{\mathrm{WA}}(f_i, e^{text}) \quad (1)$$

수식 (1)의 f_i 는 i번째 프레임을 의미하고, e^{text} 는 텍스트 임베딩을 의미한다. e^{text} 를 생성하기 위해 Jeong이 제안한 템플릿[3]과 입력 텍스트를 결합한 후, CLIP의 텍스트 인코더를 통과시켰다. f_i^{text} 는 텍스트 영역이 강조된 i번째 프레임을 의미한다.

2.2 비디오 이상 탐지

우리는 LVLM으로 최첨단 모델인 Chat-UniVi[6]를 이용하여 VAD를 수행하였다. LVLM은 프레임과 프롬프트를 입력으로 받아 이상 점수를 반환하며, 이때 입력 프롬프트는 작업(Task), 고려 사항(Consideration), 출력(Output)의 형태로 구성된다. 먼저 Task는 LVLM이 수행해야 하는 작업으로, 프레임 내 텍스트 포함 여부를 평가하는 명령어이다. Consideration은 평가 시 고려 사항으로, 텍스트에 해당하는 요소가 프레임의 중심 내용이 아니더라도 높은 점수를 부여하라는 지침이다. Output은 평가 결과의 출력 형식을 지정하며, 이상 점수를 소수점 첫째 자리까지 표현하도록 명시한다. 입력 프레임으로는 원본 프레임과 전처리된 프레임이 각각 사용된다. LVLM이 응답한 결과들은 조합되어 최종 이상 점수가 산출된다.

$$ascore_i^{text} = \gamma_1 \cdot \Phi_{\mathrm{LVLM}}(f_i,\, p^{text}) + \gamma_2 \cdot \Phi_{\mathrm{LVLM}}(f_i^{text}, p^{text}) \quad (2)$$

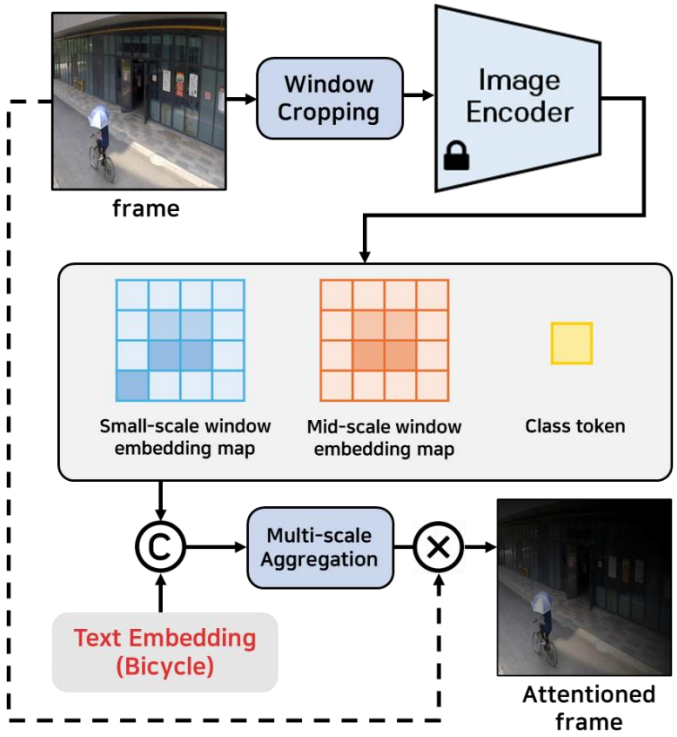

그림 2 WinCLIP 기반 어텐션 방법

수식 (2)의 $ascore_i^{text}$ 는 입력 텍스트에 대한 i번째 프레임의 이상 점수를 의미하고, p^{text} 는 프롬프트를 나타낸다. γ 는 하이퍼파라미터로, WA의 비율을 조정하는 값이다.

3. 실험 및 결과

3.1 실험 환경

제안 프레임워크가 C-VAD를 적절하게 수행하는지 평가하기 위해 C-STC 데이터셋을 구축하였다. C-STC 데이터셋은 STC의 비디오들을 11개의 비정상 유형으로 재분류하고, 각 유형에 대해 새로운 라벨을 할당한 것이다. 우리는 해당 데이터셋을 활용하여 행동(Action) 및 외형(Appearance) 이상 탐지 성능을 비교하였다. 실험은 NVIDIA GeForce RTX 3090이 장착된 Ubuntu 20.04 서버에서 Python 3.8로 진행하였다.

표 1 Baseline과 Anomaly LVLM 간의 AUC 성능 비교

	Class	Baseline	Proposed	Improvement
Action	Skateboarding	0.6163	**0.6489**	+5.29%
	Throwing	0.9218	**0.9302**	+0.92%
	Running	0.5316	**0.5880**	+10.61%
	Loitering	0.6150	**0.7169**	+16.57%
	Jumping	0.8259	**0.8822**	+6.82%
	Falling	0.7727	**0.8325**	+7.74%
	Fighting	0.8449	**0.8789**	+4.03%
Appearance	Car	0.8846	**0.8846**	+0.00%
	Hand truck	0.9528	**0.9548**	+0.21%
	Bicycle	0.7219	**0.7219**	+0.00%
	Motorcycle	0.8759	**0.8759**	+0.00%
	Average	0.7784	**0.8104**	+4.74%

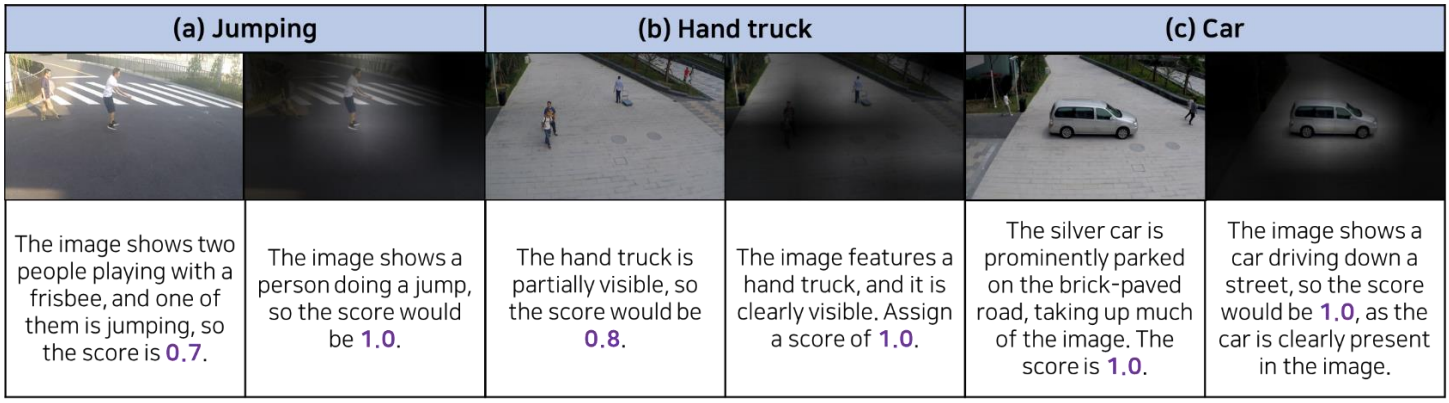

그림 3 Anomaly LVLM 생성 응답 예시

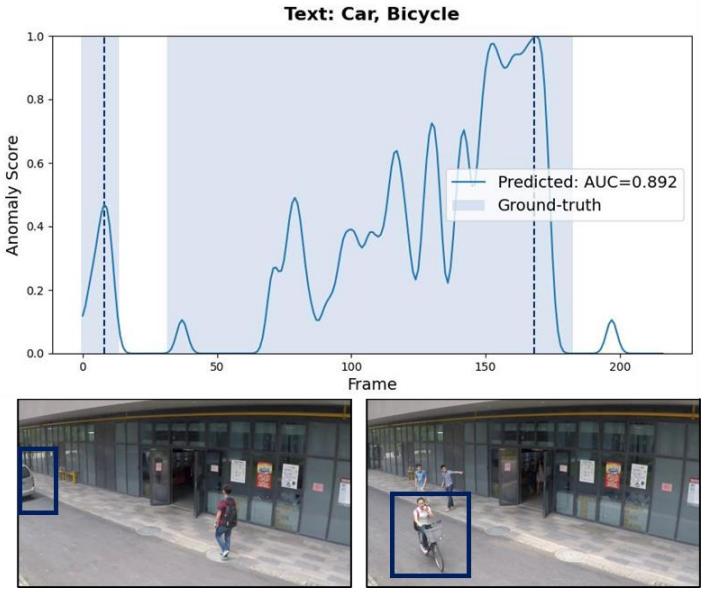

그림 4 다중 클래스 비디오의 이상 점수 시각화

3.2 실험 결과

표 1의 Area Under the ROC Curve(AUC)는 VAD 성능을 평가하는 정량적 지표로서, 이진 분류 모델이 정상 및 비정상 클래스를 얼마나 잘 구분하는지를 나타낸다. 제안된 프레임워크는 단일 LVLM(Baseline) 대비 평균 AUC가 4.74% 향상되었으며, 최종적으로 0.81의 준수한 성능을 달성하였다. 이는 WA가 작은 객체의 중요 정보를 강조하여, LVLM이 보다 정확한 이상 점수를 반환하는 데 기여했기 때문이다. 예를 들어, 그림 3의 (a)는 점프를 하는 한 명의 대상에 집중하여 더 높은 점수를 반환하였고, (b)는 작은 손수레(Hand truck) 영역을 강조하여 더 정확한 탐지가 가능했다. 하지만 Appearance 중 몇몇 클래스에 대해서는 성능 향상이 관찰되지 않았다. 이는 (c)와 같이 객체가 충분히 커서 WA를 수행하지 않고도 정확한 이상 탐지가 가능했기 때문이다.

마지막으로 그림 4는 다중 클래스에 대해 비디오의 이상 점수를 시각화한 것이다. 각 프레임에서는 두 개의 클래스 중 이상 점수가 더 높은 값이 최종적으로 할당된다. 이를 통해, 제안된 Anomaly LVLM이 다양한 비정상 유형에 대해서도 안정적인 이상 탐지 성능을 발휘함을 확인하였다.

4. 결 론

우리는 다양한 비정상 상황에 적용할 수 있는 Anomaly LVLM을 제안하고, C-STC 데이터셋을 구축하여 제안된 프레임워크의 우수성을 입증하였다. 또한, WA 기법을 통해 CCTV 영상 내 작은 객체에 집중함으로써 LVLM을 효과적으로 활용하였다. 향후 연구에서는 행동 이상 유형을 보다 정교하게 판단하기 위해 시간 정보를 활용한 추가적인 연구를 진행할 계획이다.

Acknowledgements

이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원 (IITP-2017-0-00477, (SW스타랩) IoT 환경을 위한 고성능 플래시 메모리 스토리지 기반 인메모리 분산 DBMS 연구개발, RS-2020-II201361, 인공지능대학원지원(연세대학교))과 국토교통부의 스마트시티 혁신인재육성사업으로 지원을 받아 수행된 연구임.

† 교신 저자: sanghyun@yonsei.ac.kr

참 고 문 헌

[1] R. Nayaket et al., “A comprehensive review on deep learningbased methods for video anomaly detection”, Image Vis. Comput., vol.106 Art. no.104078., 2021.

[2] Hong, S et al., “Making anomalies more anomalous: Video anomaly detection using a novel generator and destroyer”, IEEE Access, vol.12, pp.36712-36726, 2024.

[3] Jeong, J et al., “WinCLIP: Zero-/Few-Shot Anomaly Classification and Segmentation”, Proceedings of the IEEE conference on computer vision and pattern recognition, pp.19606-19616, 2023.

[4] Liu, W et al., “ Future frame prediction for anomaly detection–a new baseline”, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6536-6545, 2018.

[5] Radford A et al., “Learning transferable visual models from natural language supervision”, International conference on machine learning. PMLR, 2021.

[6] Jin, P et al., “Chat-univi: Unified visual representation empowers large language models with image and video understanding”, pp.13700-13710, 2024.