LMM(Large Multimodal Model)을 활용한 In-Context Learning 기반 이상 상황 탐지 및 분류

이하리 [1], 문진영 [2]
[1] 과학기술연합대학원대학교 인공지능전공 석사과정
[2] 과학기술연합대학원대학교 인공지능전공 교수

leehari@etri.re.kr, jymoon@etri.re.kr

Anomaly Detection and Classification Based on In-Context Learning Using LMM

Ha-Ri Lee[1], Jin-Young Moon[2]
[1]Dept. of Artificial Intelligence, University of Science and Technology
[2]Electronics and Telecommunications Research Institute

요 약

본 연구는 In-context learning 을 적용한 LMM 을 이용하여 감시 카메라 비디오 데이터를 기반으로 이상 상황을 탐지하고 이에 대한 범죄 클래스를 분류하는 방법을 제안한다. 특히 VTimeLLM[1] 모델을 사용하여 비디오 데이터를 분석하고, '정상' 및 '비정상' 이벤트를 분류한다. 추가적으로 '비정상' 이벤트는 13 개의 범죄 클래스 중 하나로 분류된다. 본 연구에서 zero-shot 과 few-shot 학습 기법을 적용하여 기존 방법들과 정량적으로 비교 실험을 수행했다. 실험 결과 LMM 과 In-context learning 을 결합한 방식이 기존 방법들과 비교해 이상 상황 탐지 성능이 개선되었다.

1. 서론

감시 카메라 비디오 데이터는 공공장소와 도심 환경에서 일어나는 사건과 상황을 실시간으로 기록하여 범죄 예방 및 사건 분석에 중요한 도구로 사용된다. 그러나 감시 카메라 비디오 데이터를 분석하고 이해하는 데에는 여러 도전 과제가 존재한다.

첫째, 비디오 데이터는 이미지와 시간 흐름에 따른 변화가 포함된 복합적인 정보를 담고 있어 단순한 텍스트나 정적 이미지와는 다른 방식으로 처리되어야 한다. 둘째, 다양한 환경 조건(예: 조명 변화, 날씨, 카메라 각도 등)이 비디오 품질에 영향을 미쳐, 정확한 분석을 어렵게 만든다. 마지막으로, 비정상적인 상황은 종종 예측하기 어려운 방식으로 발생하며, 범죄와 같은 특정 이벤트를 감지하기 위한 고도의 추론 능력이 요구된다.

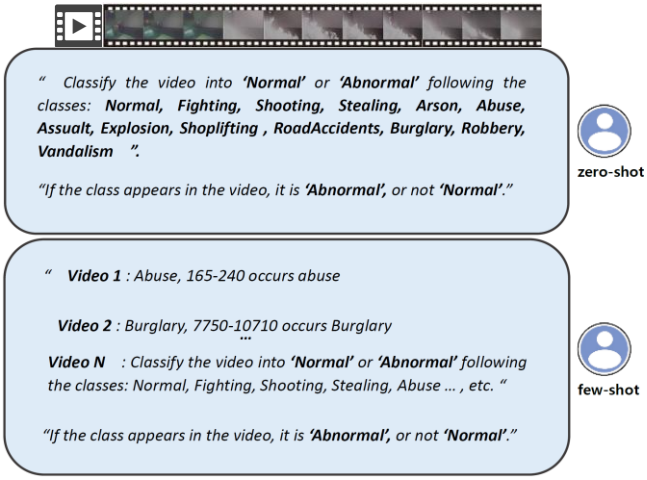
이러한 이유로 감시 카메라 비디오 데이터를 효과적으로 이해하고 분석하는 것은 기술적으로 복잡한 작업이다. 기존의 비디오 분석 방법론은 이러한 복잡성을 완전히 해결하지 못했기 때문에, 이를 보완할 수 있는 새로운 접근법이 필요하다.

최근 NLP (Natural Language Processing) 분야에서 방대한 양의 텍스트 데이터를 학습한 LLM (Large Language Model)은 언어 이해와 언어 이해 및 생성에 탁월한 성능을 보여주고 있다. LLM(Large Language Model) 의 성능이 지속적으로 향상됨에 따라 최근 텍스트 외에도 이미지, 오디오 등 여러 유형의 데이터를 통합하여 처리할 수 있는 LMM(Large Multimodal Model)이 주목받고 있다.

본 연구는 CVPR 2024 에서 소개된 LMM 인 VTimeLLM[1] 모델을 활용하여 감시 카메라 비디오 데이터를 분석하고, In-context learning 기법을 적용하여 이상 상황 탐지 및 범죄 분류 작업을 수행하는 것을 목표로 한다. 감시 카메라 비디오 데이터에서 이상 상황 탐지를 진행하여 '정상'과 '비정상' 으로 분류하고, 나아가 '비정상' 이벤트가 탐지된 비디오는 13 개의 범죄 클래스 중 어떤 범죄에 속하는지 분류한다. In-context learning 은 추가적인 미세 조정이나 학습 없이 추론 단계에서 입력 프롬프트의 맥락적인 의미(in-context)를 바탕으로 학습하여 주어진 태스크를 해결하는 방법론이다. VTimeLLM[1]에 In-context learning 의 zero-shot 과 few-shot 학습을 적용하여 기존 이상 상황 탐지 모델과 비교하여 6.95% 향상된 성능을 보였다.

2. 제안하는 방법

본 연구는 감시 카메라 비디오 데이터 세트인 UCF Crime[2]에 VTimeLLM[1] 모델을 활용하여 In-context learning 의 few-shot 과 zero-shot 학습의 두 가지 접근법을 적용했다. (그림 1) 은 이상 상황 탐지 및 이상 상황 범죄 클래스 분류에 대한 두 가지 접근법의 프롬프트 예시를 나타낸 것이다.



(그림 1) zero-shot & few-shot 프롬프트

2.1 zero-shot 학습

zero-shot 학습은 모델에게 정답 예시를 주지 않고 모델이 사전에 학습한 지식을 바탕으로 '비정상' 클래스에 대한 프롬프트를 입력하여 이상 상황 탐지와 이상 상황 범죄 클래스 분류에 대한 추론을 수행하도록 했다.

2.2 few-shot 학습

few-shot 학습은 모델에게 정답 예시를 소량 제공하여 학습시켰다. 13 개 범죄 클래스를 포함하는 UCF Crime[2] 데이터 세트에서 각 클래스의 10% 를 정답 예시로 사용했다. 정답 예시 프롬프트는 입력된 감시 카메라 비디오에서 범죄 이벤트가 발생하는 시간 구간과 그 시간 구간에 발생한 이벤트가 어떤 범죄 행위에 속하는지에 대한 정보를 제공하여 태스크를 수행한 후 zero-shot 과 few-shot 학습의 성능 차이를 비교했다.

2.3 데이터 세트

UCF Crime[2]은 약 128 시간 분량의 대규모 데이터 세트이며, 현실적인 이상 상황을 포함한 1900 개의 편집되지 않은(real-world) 감시 카메라 비디오로 구성되어 있다. <표 1> 은 각 클래스의 정의를 나타낸다. ('Abuse', 'Arrest', 'Arson', 'Assault', 'Burglary', 'Explosion', 'Fighting', 'Road Accident', 'Robbery', 'Shooting', 'Stealing', 'Shoplifting', 'Vandalism') 의 13 개 '비정상' 범죄 클래스와 일반적인 상황에 대한 '정상' 클래스인 총 14 개의 클래스로 구성된다. 또한 데이터 세트의 50% 이상이 3-4 분 길이의 동영상으로 구성되어 있으며 10 분 이내인 동영상은 200 여 개에 달하는 긴 비디오 세트다.

<표 1> UCF Crime 데이터 세트의 클래스 정의

Abuse	어린이, 노인, 동물, 여성에 대한 나쁜, 잔인하거나 폭력적인 행동을 보여주는 비디오를 포함.
Robbery	사람들이 건물이나 집에 들어가 도둑질을 하려는 의도로 행동하는 비디오를 포함. (사람에 대한 폭력 사용은 포함되지 않음).
Stealing	도둑이 폭력 또는 폭력의 위협을 통해 불법적으로 돈을 빼앗는 장면을 보여주는 비디오를 포함. (총격 사건은 포함되지 않음).
Shooting	총으로 누군가를 쏘는 행위를 보여주는 비디오를 포함.
Shoplifting	사람들이 쇼핑객인 척하며 상점에서 물건을 훔치는 장면을 보여주는 비디오를 포함.
Assault	누군가를 갑자기 또는 폭력적으로 공격하는 장면을 보여주는 비디오를 포함. (공격당한 사람이 반격하지 않음)
Fighting	두 명 이상의 사람들이 서로 공격하는 장면을 보여주는 비디오를 포함.
Arson	사람들이 재산에 고의로 불을 지르는 장면을 보여주는 비디오를 포함.
Explosion	무언가가 폭발하는 파괴적인 사건을 보여주는 비디오를 포함. (사람에 의해 고의로 불을 지르거나 폭발을 일으키는 장면은 포함되지 않음).
Arrest	경찰이 사람들을 체포하는 장면을 보여주는 비디오를 포함.
Road Accident	차량, 보행자 또는 자전거 운전자가 관련된 교통사고를 보여주는 비디오를 포함.
Vandalism	공공 또는 사유 재산에 대한 고의적인 파괴나 손상을 포함하는 행동을 보여주는 비디오를 포함.
Normal	범죄가 발생하지 않은 비디오를 포함. (실내(예: 쇼핑몰)와 실외 장면, 낮과 밤 장면이 모두 포함)

2.4 베이스라인 모델

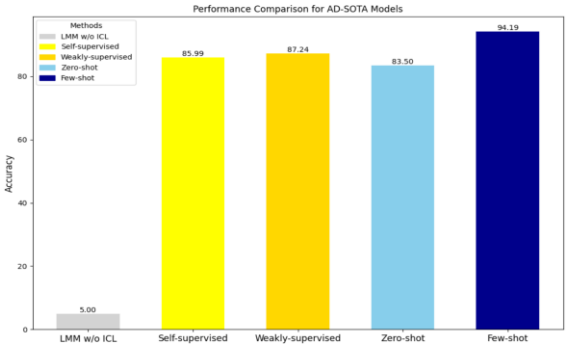
VTimeLLM[1] 은 정교한 비디오 순간 이해 및 시간 경계 추론을 목표로 설계된 LMM 이며 3 단계 학습 전략을 통해 시간 관련 작업과 비디오 대화에 뛰어난 성능을 보인다. 이 모델은 CLIP[3](Learning

<표 2> 이상 상황 탐지 실험 결과

	LMM w/o ICL	AD specific existing Model		VTimeLLM[1] w/ ICL (ours)	
		Self-supervised[3]	Weakly-supervised[4]	Zero-shot	Few-shot
Accuracy	5.00	85.99	87.24	83.50	94.19

Transferable Visual ModelsFrom Natural Language Supervision) 을 사용하여 비디오 프레임을 시각적으로 인코딩 한 후, 비주얼 어댑터를 통해 이러한 시각적 특징을 LLM 의 언어적 임베딩 공간에 정렬하는 방식을 채택하고 있다. 이후, 경계 인식 학습 단계를 통해 다중 이벤트가 포함된 비디오에서 각 이벤트의 시작과 종료 시간을 정확히 추론할 수 있는 능력을 갖추고, 마지막으로 고품질 대화 데이터 세트를 활용한 지시 조정 단계에서 인간의 의도를 이해하고 비디오 이벤트의 시간적 흐름을 정밀하게 파악할 수 있도록 훈련된다.

감시카메라 비디오 데이터는 여러 사건이나 활동을 포함하고 카메라에 포착된 사람 또는 객체의 행동을 이해할 수 있어야 한다. VTimeLLM[1]은 정교한 시간 경계 인식을 통해 감시카메라 비디오 데이터에서 특정 이벤트가 발생한 정확한 시작 시간과 종료 시간을 탐지할 수 있어 효과적임으로 VTimeLLM[1]을 연구에 사용하였다.



(그림 2) 이상 상황 탐지 실험 결과

3. 실험 결과

3.1 이상 상황 탐지

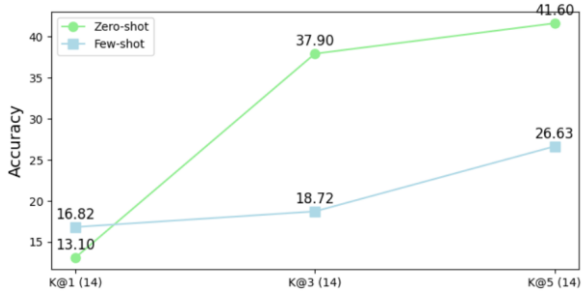
<표 2> 과 (그림 2) 는 모델에 입력된 UCF Crime[2] 데이터 세트를 '정상'과 '비정상'으로 분류하는 이상 상황 탐지 실험 결과를 나타낸다. VTimeLLM[1]을 In-context learning 없이 단독으로 사용한 경우, 이상 탐지 성능은 5% 로 매우 낮게 나타났다. In-context learning 을 적용한 후 zero-shot 학습은 83.50% , few-shot shot 학습은 94.19%으로 성능이 크게 향상되었다.

few-shot 학습 결과의 경우, 기존 이상 상황 탐지 모델 중 좋은 성능을 냈던 self-supervised 방식인 S3R (Self-Supervised Sparse Representation)[4] 모델과 weakly--supervised 방식인 BN-WVAD (BatchNorm-based Weakly Supervised Video Anomaly Detection)[5] 모델에 비해 정확도가 8.20% 와 6.95% 향상되었다.

3.2 이상 상황 범죄 클래스 분류

<표 3> 와 (그림 3) 은 '비정상' 이벤트를 탐지한 비디오 데이터에 대해 13 개의 범죄 클래스로 분류한 실험 결과이다. UCF-Crime[2] 데이터 세트의 경우, 단일 클래스 내에서 여러 클래스가 발생하는 경향이 있었다. 예를 들어, 'Robbery' 클래스는 'Shooting' 혹은 'Assault' 이벤트 가 발생한 후 'Stealing'이 나타나는 다중 클래스 이벤트가 자주 관찰되었다. 이러한 데이터 세트의 특징을 고려하여 모델의 성능을 평가하기 위해 Top-K 평가 방식을 사용했다. 이는 모델이 예측한 답이 상위 K 개의 후보군 안에 포함되는지 여부를 평가하는 방법이다.

실험 결과, K 값이 증가할수록 성능이 향상되는 것을 확인할 수 있었다. 또한 이상 상황 탐지의 실험 결과와 달리 zero-shot 학습이 few-shot 학습에 비해 더 나은 성은을 보여 K@1, K@3, K@5 에서 각각 13.10%, 37.90%, 41.60% 의 정확도를 기록했다.



(그림 3) 이상 상황 범죄 클래스 실험 결과

<표 3> 이상 상황 범죄 클래스 실험 결과

	Top-K Accuracy		
	K@1	K@3	K@5
Zero-shot	13.10	37.90	41.60
Few-shot	16.82	18.72	26.63

4. 결론

본 연구에서는 이상 상황 탐지 및 이상 상황 범죄 클래스 분류에 In-context learning 의 zero-shot 과 few-shot 학습을 적용하여 기존 방법들과 비교해 성능을 개선하였다. LMM 은 일반적으로 다양한 응용

분야에서 좋은 성능을 발휘하지만, 특정 도메인이나 복잡한 문제를 다룰 때는 In-context learning 을 함께 적용하는 것이 성능 향상에 중요한 역할을 한다는 것을 알 수 있었다.

향후 연구에서 이상 탐지 범죄 클래스 분류의 성능을 강화하기 위해 프롬프트 최적화 및 다중 클래스 문제 해결을 위한 데이터 세트 확장에 중점을 둘 것이다.

사 사

본 연구는 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2020-0-00004, 장기 시각 메모리 네트워크 기반의 예지형 시각지능 핵심기술 개발)

참고문헌

[1] B. Huang, X. Wang, H. Chen, Z. Song, W. Zhu, "VTimeLLM: Empower LLM to Grasp Video Moments," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.14271-14280, 2024.

[2] W. Sultani, C. Chen, and M. Shah, “Real-World Anomaly Detection in Surveillance Videos,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[4] J.-C. Wu, H.-Y. Hsieh, D.-J. Chen, C.-S. Fuh, T.-L. Liu, "Self-Supervised Sparse Representation for Video Anomaly Detection," Proceedings of the European Conference on Computer Vision (ECCV), 2022.

[5] Y. Zhou, Y. Qu, X. Xu, F. Shen, J. Song, H. Shen, "BatchNorm-based Weakly Supervised Video Anomaly Detection," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.