



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0175437
(43) 공개일자 2024년12월20일

(51) 국제특허분류(Int. Cl.)
G06V 20/52 (2022.01) G06V 10/44 (2022.01)
G06V 20/40 (2022.01)
(52) CPC특허분류
G06V 20/52 (2023.08)
G06N 3/0455 (2023.01)
(21) 출원번호 10-2023-0075379
(22) 출원일자 2023년06월13일
심사청구일자 2023년06월13일

(71) 출원인
세종대학교산학협력단
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
(72) 발명자
백성욱
서울특별시 강남구 영동대로 22, 813동 1303호(일원동, 디에이치 자이 개포)
이미영
서울특별시 강남구 언주로85길 13, 102호 (역삼동, 경남아파트)
(74) 대리인
특허법인위더피플

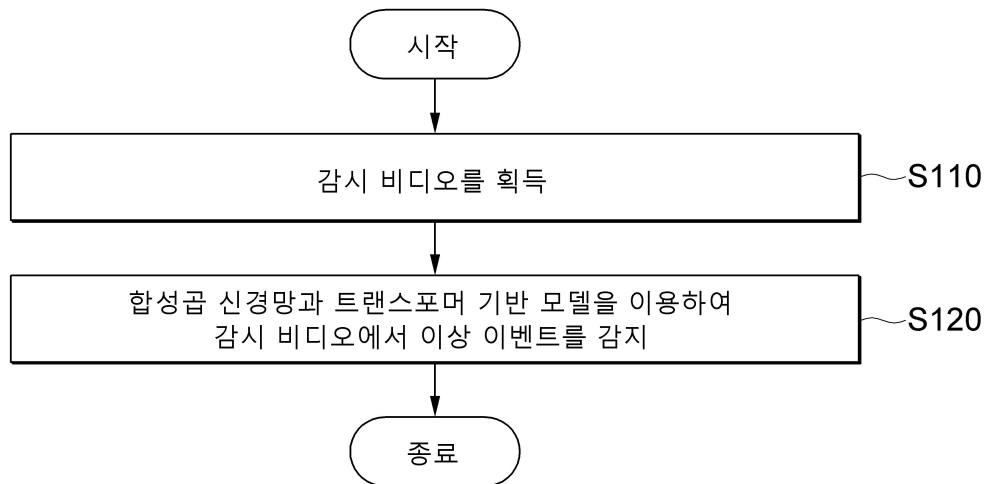
전체 청구항 수 : 총 15 항

(54) 발명의 명칭 하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 방법, 이를 수행하는 장치 및 컴퓨터 프로그램

(57) 요약

본 발명의 바람직한 실시예에 따른 하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 방법, 이를 수행하는 장치 및 컴퓨터 프로그램은, 하이브리드 합성곱 신경망(hybrid CNN)과 트랜스포머(transformer) 메커니즘을 이용하여 감시 비디오에서 이상한 사건을 감지함으로써, 공간적 특징과 시간적 특징을 효과적으로 학습하여 감시 비디오에서 이상한 사건을 감지하는 성능을 향상시킬 수 있다.

대표도 - 도2



(52) CPC특허분류

G06N 3/0464 (2023.01)

G06V 10/454 (2023.08)

G06V 20/44 (2022.01)

G06V 20/46 (2022.01)

(72) 발명자

와심올라

서울특별시 광진구 능동로 209 세종대학교 대양AI
센터 411호

사미 올라 칸

서울특별시 광진구 능동로 209 세종대학교 대양AI
센터 411호

김민제

서울특별시 광진구 천호대로110길 83-5, 101호(능
동, 에코빌)

이수민

경기도 안산시 단원구 선부로 166, 120동 509호(선
부동, 공작한양아파트)

이 발명을 지원한 국가연구개발사업

과제고유번호 1711188556

과제번호 2023R1A2C1005788

부처명 과학기술정보통신부

과제관리(전문)기관명 한국연구재단

연구사업명 개인기초연구(과기정통부)

연구과제명 커넥티드 비전 환경에서 범죄/사고 예방을 위한 이상행동 상황인식 기술

기 여 율 1/1

과제수행기관명 세종대학교

연구기간 2023.03.01 ~ 2024.02.29

명세서

청구범위

청구항 1

감시 비디오를 획득하는 단계; 및

합성곱 신경망(CNN)과 트랜스포머(transformer) 기반 모델을 이용하여 상기 감시 비디오에서 이상 이벤트를 감지하는 단계;

를 포함하는 하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 방법.

청구항 2

제1항에서,

상기 이상 이벤트 감지 단계는,

미리 학습되어 구축된 상기 합성곱 신경망을 이용하여 상기 감시 비디오에서 공간적 특징을 추출하고, 미리 학습되어 구축된 상기 트랜스포머 기반 모델을 이용하여 상기 공간적 특징을 기반으로 상기 감시 비디오에서 상기 이상 이벤트를 감지하는 것으로 이루어지는,

하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 방법.

청구항 3

제2항에서,

상기 트랜스포머 기반 모델은,

훈련 비디오에서 상기 합성곱 신경망을 통해 추출된 상기 공간적 특징을 기반으로 다중 헤드 셀프-어텐션(multi-head self-attention)을 이용하여 상기 훈련 비디오의 시공간적 특징을 학습하는 모델인,

하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 방법.

청구항 4

제3항에서,

상기 트랜스포머 기반 모델은,

상기 공간적 특징을 기반으로 공간적 표현을 생성하는 공간적 어텐션(spatial attention) 레이어, 및 상기 공간적 어텐션 레이어를 통해 생성된 공간적 표현을 기반으로 시간적 표현을 생성하는 시간적 어텐션(temporal attention) 레이어를 포함하는,

하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 방법.

청구항 5

제4항에서,

상기 트랜스포머 기반 모델은,

상기 공간적 특징을 토대로 획득된 패치를 선형 투영(linear projection)하여 임베딩 벡터를 획득하는 입력 임베딩(input embeddings) 레이어를 더 포함하는,

하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 방법.

청구항 6

제5항에서,

상기 공간적 어텐션 레이어는,
상기 입력 임베딩 레이어를 통해 생성된 상기 임베딩 벡터를 기반으로 공간적 표현을 생성하는,
하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 방법.

청구항 7

제1항에서,
상기 합성곱 신경망은, EfficientNetV2인,
하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 방법.

청구항 8

제1항 내지 제7항 중 어느 한 항에 기재된 하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 방법을 컴퓨터에서 실행시키기 위하여 컴퓨터 판독 가능한 저장 매체에 저장된 컴퓨터 프로그램.

청구항 9

하이브리드 합성곱 신경망(hybrid CNN)과 트랜스포머(transformer) 메커니즘을 이용하여 감시 비디오에서 이상한 사건을 감지하기 위한 하나 이상의 프로그램을 저장하는 메모리; 및
상기 메모리에 저장된 상기 하나 이상의 프로그램에 따라 상기 하이브리드 합성곱 신경망과 상기 트랜스포머 메커니즘을 이용하여 상기 감시 비디오에서 이상한 사건을 감지하기 위한 동작을 수행하는 하나 이상의 프로세서;
를 포함하며,
상기 프로세서는,
감시 비디오를 획득하고,
합성곱 신경망(CNN)과 트랜스포머 기반 모델을 이용하여 상기 감시 비디오에서 이상 이벤트를 감지하는,
하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 장치.

청구항 10

제9항에서,
상기 프로세서는,
미리 학습되어 구축된 상기 합성곱 신경망을 이용하여 상기 감시 비디오에서 공간적 특징을 추출하고,
미리 학습되어 구축된 상기 트랜스포머 기반 모델을 이용하여 상기 공간적 특징을 기반으로 상기 감시 비디오에서 상기 이상 이벤트를 감지하는,
하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 장치.

청구항 11

제10항에서,
상기 트랜스포머 기반 모델은,
훈련 비디오에서 상기 합성곱 신경망을 통해 추출된 상기 공간적 특징을 기반으로 다중 헤드 셀프-어텐션(multi-head self-attention)을 이용하여 상기 훈련 비디오의 시공간적 특징을 학습하는 모델인,
하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 장치.

청구항 12

제11항에서,
상기 트랜스포머 기반 모델은,

상기 공간적 특징을 기반으로 공간적 표현을 생성하는 공간적 어텐션(spatial attention) 레이어, 및 상기 공간적 어텐션 레이어를 통해 생성된 공간적 표현을 기반으로 시간적 표현을 생성하는 시간적 어텐션(temporal attention) 레이어를 포함하는,

하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 장치.

청구항 13

제12항에서,

상기 트랜스포머 기반 모델은,

상기 공간적 특징을 토대로 획득된 패치를 선형 투영(linear projection)하여 임베딩 벡터를 획득하는 입력 임베딩(input embeddings) 레이어를 더 포함하는,

하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 장치.

청구항 14

제13항에서,

상기 공간적 어텐션 레이어는,

상기 입력 임베딩 레이어를 통해 생성된 상기 임베딩 벡터를 기반으로 공간적 표현을 생성하는,

하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 장치.

청구항 15

제9항에서,

상기 합성곱 신경망은, EfficientNetV2인,

하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 장치.

발명의 설명

기술 분야

[0001] 본 발명은 하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 방법, 이를 수행하는 장치 및 컴퓨터 프로그램에 관한 것으로서, 더욱 상세하게는 감시 비디오에서 이상한 사건을 감지하는, 방법, 장치 및 컴퓨터 프로그램에 관한 것이다.

배경 기술

[0002] 감시 시스템의 광범위한 적용으로 인해 영상에서 이상을 감지하는 것이 중요해지고 있다. 공공 안전 및 보안 문제를 해결해야 할 필요성이 증가함에 따라 기차역, 캠퍼스, 쇼핑몰과 같은 공공 장소에 감시 카메라를 설치하는 것이 주요 솔루션 중 하나이다. 최근 몇 년 동안 감시 시스템이 더 자주 배치되면서 이러한 시스템의 비용이 크게 감소했지만, 절도, 싸움, 교통사고, 학대 등과 같은 비정상적인 사건을 감지하는 것은 여전히 사람의 개입을 통해서만 달성할 수 있다. 모니터링 과정에서 사람의 실수를 방지하기 위해 영상에서 이상 이벤트를 감지하는 지능형 알고리즘을 개발하는 것은 인적 노동력을 줄이기 위한 중요한 과제이다. 이상 패턴은 일반적으로 모호하고 정의되지 않은 경우가 많으며, 주석이 달린 데이터 세트도 부족하여 이상 감지 작업을 더욱 어렵게 만든다. 따라서 감시 환경에서 이상 이벤트를 자동으로 감지할 수 있는 영상 이상 감지 시스템의 개발이 시급한 상황이다.

발명의 내용

해결하려는 과제

[0003] 본 발명이 이루고자 하는 목적은, 하이브리드 합성곱 신경망(hybrid CNN)과 트랜스포머(transformer) 메커니즘

을 이용하여 감시 비디오에서 이상한 사건을 감지하는, 하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 방법, 이를 수행하는 장치 및 컴퓨터 프로그램을 제공하는 데 있다.

[0004] 본 발명의 명시되지 않은 또 다른 목적들은 하기의 상세한 설명 및 그 효과로부터 용이하게 추론할 수 있는 범위 내에서 추가적으로 고려될 수 있다.

과제의 해결 수단

[0005] 상기의 기술적 과제를 달성하기 위한 본 발명의 바람직한 실시예에 따른 하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 방법은, 감시 비디오를 획득하는 단계; 및 합성곱 신경망(CNN)과 트랜스포머(transformer) 기반 모델을 이용하여 상기 감시 비디오에서 이상 이벤트를 감지하는 단계;를 포함한다.

[0006] 여기서, 상기 이상 이벤트 감지 단계는, 미리 학습되어 구축된 상기 합성곱 신경망을 이용하여 상기 감시 비디오에서 공간적 특징을 추출하고, 미리 학습되어 구축된 상기 트랜스포머 기반 모델을 이용하여 상기 공간적 특징을 기반으로 상기 감시 비디오에서 상기 이상 이벤트를 감지하는 것으로 이루어질 수 있다.

[0007] 여기서, 상기 트랜스포머 기반 모델은, 훈련 비디오에서 상기 합성곱 신경망을 통해 추출된 상기 공간적 특징을 기반으로 다중 헤드 셀프-어텐션(multi-head self-attention)을 이용하여 상기 훈련 비디오의 시공간적 특징을 학습하는 모델일 수 있다.

[0008] 여기서, 상기 트랜스포머 기반 모델은, 상기 공간적 특징을 기반으로 공간적 표현을 생성하는 공간적 어텐션(spatial attention) 레이어, 및 상기 공간적 어텐션 레이어를 통해 생성된 공간적 표현을 기반으로 시간적 표현을 생성하는 시간적 어텐션(temporal attention) 레이어를 포함할 수 있다.

[0009] 여기서, 상기 트랜스포머 기반 모델은, 상기 공간적 특징을 토대로 획득된 패치를 선형 투영(linear projection)하여 임베딩 벡터를 획득하는 입력 임베딩(input embeddings) 레이어를 더 포함할 수 있다.

[0010] 여기서, 상기 공간적 어텐션 레이어는, 상기 입력 임베딩 레이어를 통해 생성된 상기 임베딩 벡터를 기반으로 공간적 표현을 생성할 수 있다.

[0011] 여기서, 상기 합성곱 신경망은, EfficientNetV2일 수 있다.

[0013] 상기의 기술적 과제를 달성하기 위한 본 발명의 바람직한 실시예에 따른 컴퓨터 프로그램은 컴퓨터 판독 가능한 저장 매체에 저장되어 상기한 하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 방법 중 어느 하나를 컴퓨터에서 실행시킨다.

[0015] 상기의 기술적 과제를 달성하기 위한 본 발명의 바람직한 실시예에 따른 하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 장치는, 하이브리드 합성곱 신경망(hybrid CNN)과 트랜스포머(transformer) 메커니즘을 이용하여 감시 비디오에서 이상한 사건을 감지하기 위한 하나 이상의 프로그램을 저장하는 메모리; 및 상기 메모리에 저장된 상기 하나 이상의 프로그램에 따라 상기 하이브리드 합성곱 신경망과 상기 트랜스포머 메커니즘을 이용하여 상기 감시 비디오에서 이상한 사건을 감지하기 위한 동작을 수행하는 하나 이상의 프로세서;를 포함하며, 상기 프로세서는, 감시 비디오를 획득하고, 합성곱 신경망(CNN)과 트랜스포머 기반 모델을 이용하여 상기 감시 비디오에서 이상 이벤트를 감지한다.

[0016] 여기서, 상기 프로세서는, 미리 학습되어 구축된 상기 합성곱 신경망을 이용하여 상기 감시 비디오에서 공간적 특징을 추출하고, 미리 학습되어 구축된 상기 트랜스포머 기반 모델을 이용하여 상기 공간적 특징을 기반으로 상기 감시 비디오에서 상기 이상 이벤트를 감지할 수 있다.

[0017] 여기서, 상기 트랜스포머 기반 모델은, 훈련 비디오에서 상기 합성곱 신경망을 통해 추출된 상기 공간적 특징을 기반으로 다중 헤드 셀프-어텐션(multi-head self-attention)을 이용하여 상기 훈련 비디오의 시공간적 특징을 학습하는 모델일 수 있다.

[0018] 여기서, 상기 트랜스포머 기반 모델은, 상기 공간적 특징을 기반으로 공간적 표현을 생성하는 공간적 어텐션(spatial attention) 레이어, 및 상기 공간적 어텐션 레이어를 통해 생성된 공간적 표현을 기반으로 시간적 표현을 생성하는 시간적 어텐션(temporal attention) 레이어를 포함할 수 있다.

[0019] 여기서, 상기 트랜스포머 기반 모델은, 상기 공간적 특징을 토대로 획득된 패치를 선형 투영(linear projection)하여 임베딩 벡터를 획득하는 입력 임베딩(input embeddings) 레이어를 더 포함할 수 있다.

[0020] 여기서, 상기 공간적 어텐션 레이어는, 상기 입력 임베딩 레이어를 통해 생성된 상기 임베딩 벡터를 기반으로 공간적 표현을 생성할 수 있다.

[0021] 여기서, 상기 합성곱 신경망은, EfficientNetV2일 수 있다.

발명의 효과

[0022] 본 발명의 바람직한 실시예에 따른 하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 방법, 이를 수행하는 장치 및 컴퓨터 프로그램에 의하면, 하이브리드 합성곱 신경망(hybrid CNN)과 트랜스포머(transformer) 메커니즘을 이용하여 감시 비디오에서 이상한 사건을 감지함으로써, 공간적 특징과 시간적 특징을 효과적으로 학습하여 감시 비디오에서 이상한 사건을 감지하는 성능을 향상시킬 수 있다.

[0023] 본 발명의 효과들은 이상에서 언급한 효과들로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0024] 도 1은 본 발명의 바람직한 실시예에 따른 하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 장치를 설명하기 위한 블록도이다.

도 2는 본 발명의 바람직한 실시예에 따른 하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 방법을 설명하기 위한 흐름도이다.

도 3은 도 2에 도시한 이상 이벤트 감지 단계를 설명하기 위한 도면이다.

도 4는 본 발명의 바람직한 실시예에 따른 감시 이상 탐지 프레임워크를 설명하기 위한 도면이다.

도 5는 본 발명의 바람직한 실시예에 따른 트랜스포머 기반 모델의 상세 구성을 설명하기 위한 도면이다.

도 6은 본 발명의 바람직한 실시예에 따른 감시 이상 탐지 프레임워크의 AUC 성능을 설명하기 위한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0025] 이하, 첨부된 도면을 참조하여 본 발명의 실시예를 상세히 설명한다. 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나, 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하고, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다. 명세서 전체에 걸쳐 동일 참조 부호는 동일 구성 요소를 지칭한다.

[0026] 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어(기술 및 과학적 용어를 포함)는 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 공통적으로 이해될 수 있는 의미로 사용될 수 있을 것이다. 또한, 일반적으로 사용되는 사전에 정의되어 있는 용어들은 명백하게 특별히 정의되어 있지 않는 한 이상적으로 또는 과도하게 해석되지 않는다.

[0027] 본 명세서에서 "제1", "제2" 등의 용어는 하나의 구성 요소를 다른 구성 요소로부터 구별하기 위한 것으로, 이들 용어들에 의해 권리범위가 한정되어서는 아니 된다. 예컨대, 제1 구성 요소는 제2 구성 요소로 명명될 수 있고, 유사하게 제2 구성 요소도 제1 구성 요소로 명명될 수 있다.

[0028] 본 명세서에서 각 단계들에 있어 식별부호(예컨대, a, b, c 등)는 설명의 편의를 위하여 사용되는 것으로 식별 부호는 각 단계들의 순서를 설명하는 것이 아니며, 각 단계들은 문맥상 명백하게 특정 순서를 기재하지 않는 이상 명기된 순서와 다르게 일어날 수 있다. 즉, 각 단계들은 명기된 순서와 동일하게 일어날 수도 있고 실질적으로 동시에 수행될 수도 있으며 반대의 순서대로 수행될 수도 있다.

[0029] 본 명세서에서, "가진다", "가질 수 있다", "포함한다" 또는 "포함할 수 있다" 등의 표현은 해당 특징(예컨대, 수치, 기능, 동작, 또는 부품 등의 구성 요소)의 존재를 가리키며, 추가적인 특징의 존재를 배제하지 않는다.

- [0032] 이하에서 첨부한 도면을 참조하여 본 발명에 따른 하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 방법, 이를 수행하는 장치 및 컴퓨터 프로그램의 바람직한 실시예에 대해 상세하게 설명한다.
- [0034] 먼저, 도 1을 참조하여 본 발명의 바람직한 실시예에 따른 하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 장치에 대하여 설명한다.
- [0035] 도 1은 본 발명의 바람직한 실시예에 따른 하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 장치를 설명하기 위한 블록도이다.
- [0036] 도 1을 참조하면, 본 발명의 바람직한 실시예에 따른 하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 장치(이하 '감시 이상 탐지 장치'라 한다)(100)는 하이브리드 합성곱 신경망(hybrid CNN)과 트랜스포머(transformer) 메커니즘을 이용하여 감시 비디오에서 이상한 사건을 감지할 수 있다. 이에 따라, 본 발명은 공간적 특징과 시간적 특징을 효과적으로 학습하여 감시 비디오에서 이상한 사건을 감지하는 성능을 향상시킬 수 있다.
- [0037] 즉, 감시 비디오 이상 탐지(surveillance video anomaly detection, SVAD)는 객체 규모, 식별 및 예상치 못한 이벤트의 변화, 배경의 영향, 다양한 감시 상황에서 이상 이벤트의 광범위한 정의로 인해 까다로운 작업이다. 본 발명에서는 이상 탐지를 위한 엔드-투-엔드 하이브리드 합성곱 신경망과 비전 트랜스포머 기반 프레임워크를 제안한다. 제안된 프레임워크는 감시 비디오의 공간 및 시간 정보를 사용하여 이상 이벤트를 감지하며, 두 단계로 작동한다. 첫 번째 단계에서는 효율적인 백본 합성곱 신경망 모델을 사용하여 공간적 특징을 추출하고, 두 번째 단계에서는 이러한 특징을 트랜스포머 기반 모델에 전달하여 다양한 복잡한 감시 이벤트 간의 장기적인 시간적 관계를 학습한다. 백본 모델의 특징은 순차적 학습 모델에 공급되어 시간적 셀프-어텐션(self-attention)을 활용하여 어텐션 맵을 생성함으로써 제안된 프레임워크가 시공간적 특징을 효과적으로 학습하고 이상 이벤트를 탐지할 수 있도록 한다. 다양한 벤치마크 VAD 데이터 세트에 대한 실험 결과는 제안된 프레임워크의 유효성을 입증하며, 상하이테크(ShanghaiTech), UCSD Ped2, CUHK 애비뉴(avenue) 데이터 세트에서 각각 94.6%, 98.4%, 89.6%의 높은 AUC 값을 달성하여 다른 최신 접근법을 능가하는 성능을 보였다.
- [0039] 이를 위해, 감시 이상 탐지 장치(100)는 하나 이상의 프로세서(110), 컴퓨터 판독 가능한 저장 매체(130) 및 통신 버스(150)를 포함할 수 있다.
- [0040] 프로세서(110)는 감시 이상 탐지 장치(100)가 동작하도록 제어할 수 있다. 예컨대, 프로세서(110)는 컴퓨터 판독 가능한 저장 매체(130)에 저장된 하나 이상의 프로그램(131)을 실행할 수 있다. 하나 이상의 프로그램(131)은 하나 이상의 컴퓨터 실행 가능 명령어를 포함할 수 있으며, 컴퓨터 실행 가능 명령어는 프로세서(110)에 의해 실행되는 경우 감시 이상 탐지 장치(100)로 하여금 하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용하여 감시 비디오에서 이상한 사건을 감지하기 위한 동작을 수행하도록 구성될 수 있다.
- [0041] 컴퓨터 판독 가능한 저장 매체(130)는 하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용하여 감시 비디오에서 이상한 사건을 감지하기 위한 컴퓨터 실행 가능 명령어 내지 프로그램 코드, 프로그램 데이터 및/또는 다른 적합한 형태의 정보를 저장하도록 구성된다. 컴퓨터 판독 가능한 저장 매체(130)에 저장된 프로그램(131)은 프로세서(110)에 의해 실행 가능한 명령어의 집합을 포함한다. 일 실시예에서, 컴퓨터 판독 가능한 저장 매체(130)는 메모리(랜덤 액세스 메모리와 같은 휘발성 메모리, 비휘발성 메모리, 또는 이들의 적절한 조합), 하나 이상의 자기 디스크 저장 디바이스들, 광학 디스크 저장 디바이스들, 플래시 메모리 디바이스들, 그 밖에 감시 이상 탐지 장치(100)에 의해 액세스되고 원하는 정보를 저장할 수 있는 다른 형태의 저장 매체, 또는 이들의 적합한 조합일 수 있다.
- [0042] 통신 버스(150)는 프로세서(110), 컴퓨터 판독 가능한 저장 매체(130)를 포함하여 감시 이상 탐지 장치(100)의 다른 다양한 컴포넌트들을 상호 연결한다.
- [0043] 감시 이상 탐지 장치(100)는 또한 하나 이상의 입출력 장치를 위한 인터페이스를 제공하는 하나 이상의 입출력 인터페이스(170) 및 하나 이상의 통신 인터페이스(190)를 포함할 수 있다. 입출력 인터페이스(170) 및 통신 인터페이스(190)는

터페이스(190)는 통신 버스(150)에 연결된다. 입출력 장치(도시하지 않음)는 입출력 인터페이스(170)를 통해 감시 이상 탐지 장치(100)의 다른 컴포넌트들에 연결될 수 있다.

[0046] 그러면, 도 2 및 도 3을 참조하여 본 발명의 바람직한 실시예에 따른 하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 방법에 대하여 설명한다.

[0047] 도 2는 본 발명의 바람직한 실시예에 따른 하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 방법을 설명하기 위한 흐름도이고, 도 3은 도 2에 도시한 이상 이벤트 감지 단계를 설명하기 위한 도면이다.

[0048] 도 2를 참조하면, 감시 이상 탐지 장치(100)의 프로세서(110)는 감시 비디오를 획득할 수 있다(S110).

[0050] 그런 다음, 프로세서(110)는 합성곱 신경망과 트랜스포머 기반 모델을 이용하여 감시 비디오에서 이상 이벤트를 감지할 수 있다(S120).

[0051] 즉, 프로세서(110)는 미리 학습되어 구축된 합성곱 신경망을 이용하여 감시 비디오에서 공간적 특징을 추출할 수 있다.

[0052] 여기서, 합성곱 신경망은 EfficientNetV2일 수 있다.

[0053] 그리고, 프로세서(110)는 미리 학습되어 구축된 트랜스포머 기반 모델을 이용하여 합성곱 신경망을 통해 추출된 공간적 특징을 기반으로 감시 비디오에서 이상 이벤트를 감지할 수 있다.

[0054] 여기서, 트랜스포머 기반 모델은 훈련 비디오에서 합성곱 신경망을 통해 추출된 공간적 특징을 기반으로 다중 헤드 셀프-어텐션(multi-head self-attention)을 이용하여 훈련 비디오의 시공간적 특징을 학습하는 모델일 수 있다.

[0055] 즉, 트랜스포머 기반 모델은 도 3에 도시된 바와 같이, 입력 임베딩(input embeddings) 레이어, 공간적 어텐션(spatial attention) 레이어, 시간적 어텐션(temporal attention) 레이어, MLP 헤드(MLP head) 레이어, 및 클래스 예측(class prediction) 레이어를 포함할 수 있다.

[0056] 입력 임베딩 레이어는 공간적 특징을 토대로 획득된 패치를 선형 투영(linear projection)하여 임베딩 벡터를 획득할 수 있다.

[0057] 공간적 어텐션 레이어는 공간적 특징을 기반으로 공간적 표현을 생성할 수 있다. 즉, 공간적 어텐션 레이어는 입력 임베딩 레이어를 통해 생성된 임베딩 벡터를 기반으로 공간적 표현을 생성할 수 있다.

[0058] 시간적 어텐션 레이어는 공간적 어텐션 레이어를 통해 생성된 공간적 표현을 기반으로 시간적 표현을 생성할 수 있다.

[0061] 그러면, 도 4 내지 도 6을 참조하여 본 발명의 바람직한 실시예에 따른 하이브리드 합성곱 신경망과 트랜스포머 메커니즘을 이용한 감시 이상 탐지 동작에 대하여 보다 자세하게 설명한다.

[0062] 1. 개요

[0063] 보안 및 안전 구역을 제공하기 위해 스마트 시티에서 사용할 수 있는 잠재력으로 인해 감시 이상 탐지가 광범위하게 연구되고 있다. 이 작업은 주로 도난, 폭력, 사고 등과 같은 비정상적인 이벤트가 발생하는 시간대를 식별하는 것과 관련이 있다. 정상 비디오에 대해서만 학습되는 비지도 기반 이상 탐지 같은 원-클래스 분류기가 사용되었지만, 보다 효과적인 접근 방식은 정상 및 비정상 이벤트를 나타내는 레이블과 함께 시퀀스 또는 비디오 수준 주석이 있는 학습 샘플을 사용하는 약한 지도 학습을 사용한다. 이 약한 지도 학습 절차의 목표는 주석에 대한 사람의 노력을 줄이면서 이상 탐지의 정확도를 높이는 것이다. 그러나 약한 지도 학습 기반 이상 탐지의 한 가지 문제점은 전체 동영상에서 비정상이라고 레이블이 지정된 부분을 식별하는 것이다. 이 문제는 두

가지 이유로 발생한다. 첫째, 비정상 비디오의 대부분이 정상 이벤트로 구성되어 있어 학습 절차를 압도하여 특정 비정상 스니펫(snippet)을 식별하기 어렵고, 둘째, 비디오의 비정상 부분이 정상 이벤트와 다를 수 있어 정상과 비정상 비디오 세그먼트를 명확하게 구분하기 어렵다는 것이다. 감시 이상 탐지는 분류 문제의 하위 유형으로, 투-클래스 또는 원-클래스 탐지 문제로 축소할 수 있다.

[0064] 지난 20년 동안 여러 가지 비디오 이상 감지(video anomaly detection, VAD) 기법이 문헌에 제안되었다. 이상 움직임 탐지에 대한 초기 연구에서는 비디오 시퀀스에서 객체의 궤적을 추출하여 객체의 비정상적인 움직임 패턴을 탐지했다. 물체가 정상 궤적과 다른 방식으로 움직이면 이를 이상 움직임으로 간주했다. Cosar 등은 물체의 궤적(예: 속도, 방향, 신체 움직임)에 대한 평가와 픽셀 수준 분석(예: 모양)을 사용하는 비정상적인 행동 추정을 위한 비지도 모델을 개발했다. 이후 Piciarelli 등은 비디오 시퀀스에서 움직이는 물체에서 추출한 정상 궤적을 클러스터링하는 이상 탐지 프레임워크를 개발했다. 이러한 정상 물체 궤적을 학습하기 위해 단일 클래스 서포트 벡터 머신(support vector machine, SVM)을 사용했다. 그러나 이러한 방법은 물체가 가려진 경우에는 효과적이지 않다. 일부 연구에서는 물체 가려짐이라는 까다로운 문제를 해결하기 위해 이상 탐지에 전역 특징을 사용하기도 했다. 연구자들은 데이터의 정상 패턴을 학습하기 위해 비선형 원-클래스 서포트 벡터 머신(SVM)을 사용했다. 이상은 훈련된 모델에 의해 이상치 점수가 있는 것으로 예측된 이벤트 행동으로 설명되었다.

[0065] 이 분야에 대한 많은 노력에도 불구하고, 비디오 이상 감지(VAD)는 모호성과 비정상적인 비디오의 부족이라는 두 가지 주요 어려움으로 인해 여전히 어려운 과제로 남아 있다. 모호성은 의미적으로 고정되지 않고, 제한되지 않으며, 예측할 수 없고, 매우 가변적인 비정상 이벤트를 의미하므로 더욱 복잡하게 만든다. 더 어려운 두 번째 어려움은 비정상 이벤트는 일반적인 이벤트보다 발생 횟수가 훨씬 적기 때문에 가능한 모든 유형의 비정상 이벤트를 정확하게 정의하고 수집하는 것이 대부분의 보안 감시 환경에서 불가능하다는 것이다. 비정상 이벤트가 불규칙적으로 발생하기 때문에 적절한 수의 비정상 샘플을 수집하는 것 또한 어려운 과정이다. 따라서 보안 관제 환경에서의 비디오 이상 감지(VAD)는 일반적으로 이상값 탐지 문제로 간주된다. 비지도 설정은 정상 데이터만 사용할 수 있기 때문에 정상 패턴을 인식하도록 모델을 훈련하는 데 이상적이다. 비정상 프레임은 정상 패턴과 일치하지 않기 때문에 테스트 단계가 끝날 때 식별된다. 비디오 이상 감지(VAD) 시스템에는 훈련 단계에서 사용되는 지도 신호에 따라 약한 지도 방식과 비지도 방식으로 알려진 두 가지 유형이 있다. 약한 지도 비디오 이상 감지(VAD)가 모델을 훈련하려면 정상 데이터와 비정상 데이터를 모두 사용할 수 있어야 한다. 세분화된 이상(세그먼트 수준)을 감지하려면 모델이 중간 수준의 지도(비디오 수준에서 거친 주석)를 감지할 수 있어야 한다.

[0066] 연구자들은 최근 몇 년 동안 딥러닝 알고리즘을 사용하여 비디오 이상 감지(VAD)를 위한 다양한 심층 신경망을 개발했으며, 인상적인 결과를 얻었다. 예를 들어, 심층 신경망과 다중 인스턴스 학습을 사용하면 감시 환경에서 이상을 탐지할 수 있다. 그러나 약한 지도 기반 비디오 이상 감지(VAD) 기술은 괄목할 만한 발전을 이루었지만, (i)감시 영상에서 비정상적인 스니펫을 정확하게 인식하는 방법과 (ii)비디오 스니펫 간의 장기적인 시간적 의존성을 완전히 탐색하여 보다 강력한 비디오 이상 감지(VAD) 시스템을 생성하는 방법이라는 두 가지 문제에 여전히 직면하고 있다. 본 발명에서는 비디오 스니펫 간의 시간적 관계를 탐색하고 정상 샘플과 비정상 샘플 간의 분리 가능성을 향상시키는 시간적 관계 학습을 위한 새로운 약한 지도 방식의 하이브리드 합성곱 신경망(CNN) 및 트랜스포머 기반 모델을 제시한다. 또한 비디오 스니펫 간의 장기적인 시간적 종속성을 포착하기 위해 트랜스포머를 지원하는 시간적 관계 인코더를 설계하여 모델의 탐지 성능을 더욱 향상시킨다. 본 발명의 주요 기여는 다음과 같이 요약할 수 있다.

[0067] - 시공간적 시각 감시 데이터의 해석을 위해 합성곱 신경망과 트랜스포머 기반 모델을 기반으로 구축된 지능형 비디오 이상 감지 프레임워크를 제안한다. 이 프레임워크는 규모 변화와 이동에 변하지 않는 특징 표현을 학습하고 입력 데이터 간의 장기적인 의존 관계를 모델링하기 위해 합성곱 신경망과 트랜스포머의 장점을 결합한 하이브리드 접근 방식이다. 기존의 효율적인 방법과 달리 제안된 방법은 간단하고 모델을 쉽게 훈련할 수 있다. 이 방법은 기존 방법에서 처리하는 데이터의 3%만 사용하여 비디오 정보를 캡처한다.

[0068] - 움직임이 두드러지는 공간 영역과 움직임이 두드러지지 않는 공간 영역을 선택적으로 강조하는 시간적 셀프-어텐션 메커니즘을 적용한다. 제안된 프레임워크는 공간 어텐션 맵을 통해 시공간적 특징을 학습하고, 이를 통해 감시 이벤트를 인식하는 데 사용할 수 있다.

[0069] - 세 가지 벤치마크 비디오 이상 데이터 세트에 대한 실험 결과는 제안된 프레임워크가 기존 경쟁 제품과 비교하여 최첨단(state-of-the-art, SOTA) 성능을 달성한다는 것을 보여준다.

[0071] 2. 관련 문헌 검토

[0072] 복잡한 이상과 데이터 세트가 증가함에 따라 비디오 이상 감지(VAD) 시스템과 데이터 세트는 점점 더 복잡해지고 있다. 사용되는 공간적 및 시간적 구성 요소에 따라 일반적으로 비지도 또는 약한 지도라는 두 가지 범주로 비디오 이상 감지(VAD) 방법을 분류할 수 있다.

[0073] 2.1. 비지도 이상 감지

[0074] 최근까지 대부분의 이상 탐지 방법은 일반 훈련 데이터와 수작업으로 만든 특징을 사용하여 원-클래스 분류 문제를 처리했다. 딥 러닝이 널리 보급됨에 따라 사전 학습된 심층 신경망의 특징을 활용하기 위해 여러 가지 접근 방식이 적용되었다. 또한 일반 다양체(manifold)의 잠재 공간에 대한 제한을 사용하여 간결한 정규성 표현을 학습하는 연구도 일부 진행되었으며, 일부 접근 방식은 재구성 오류의 (적대적) 최소화를 통해 일반 샘플의 설명을 발견하기 위해 생성 모델에 의존하기도 했다. 이를 위해 일부 저자는 데이터 재구성을 위해 생성 모델을 사용하기도 했다. 이러한 접근 방식은 재구성 오류가 높은 샘플을 이상 영상으로 간주하며, 종종 보이지 않는 이상 영상/이미지는 잘 재구성하기 어렵다고 가정한다. 그 결과 학습 데이터가 과적합되어 비정상성에 대한 사전 지식이 부족하기 때문에 비정상적인 이벤트와 정상적인 이벤트를 구분하는 것이 거의 불가능해진다. 연구자들은 감시 이상 이벤트에 대해 시공간적 관계를 분리하고 외형과 움직임 정보를 별도로 획득하는 컨볼루션 자동 인코더를 제안했다. 최근에는 감시 이상 행위 탐지를 위한 스케leton 기반 특징 모델이 제안되었다. 이 모델에서는 자가 학습 회귀 기법을 적용하여 사람 골격의 특징을 기반으로 이상 점수를 높이고 감시 영상에서 사람 관련 이상 이벤트를 탐지한다.

[0075] 2.2. 약한 지도 이상 탐지

[0076] 라벨이 지정된 일부 비정상 샘플을 사용하는 것이 많은 비지도 접근법을 사용하는 것보다 성능 개선 측면에서 더 효과적이라는 것이 밝혀졌다. 안타깝게도 대규모 감시 비디오 데이터 세트의 프레임 수준 라벨 주석에 드는 비용은 엄청nam며, 이는 실현 가능하지 않다. 따라서 현재 SOTA 비디오 이상 감지(VAD)에 대한 접근 방식은 비용이 저렴한 비디오 수준 주석을 기반으로 하는 약한 지도 훈련 기법에 크게 의존하고 있다. Sultani 등은 비디오에서 이상을 감지하기 위해 비디오 레벨 라벨을 도입하고, 비디오 감시에 대한 이상 정보를 수집하기 위한 데이터 세트의 대규모 하위 집합인 UCF-Crime을 도입했다. 그 결과, 연구자들은 최근 몇 년 동안 이 분야에 상당한 관심을 보였다. 약한 지도 절차로 영상 이상을 탐지할 수 있는 몇 가지 기법이 있으며, 이러한 기법 대부분은 다중 인스턴스 학습(multi-instance learning, MIL) 프레임워크를 기반으로 구축된다. 안타깝게도 대부분의 다중 인스턴스 학습(MIL) 기반 접근 방식은 포지티브 백의 비정상 이벤트 수가 라벨에 있는 노이즈의 영향을 받을 수 있기 때문에 비정상 비디오 라벨을 활용할 수 없다. 이러한 노이즈는 정상 스니펫이 비정상 이벤트가 포함된 비디오에서 가장 높은 비정상 이벤트로 잘못 처리되는 데서 발생한다.

[0077] 이 문제를 해결하기 위해 Zhong 등은 잡음이 있는 레이블을 사용하여 이진 분류 문제로 재구성한 다음, 잡음을 제거하기 위해 그래프 합성곱 신경망(GCNN)을 적용했다. 이 논문에서 제시된 결과는 [W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6479-6488.]의 결과보다 더 정확했지만, 그래프 합성곱 신경망(GCNN)과 다중 인스턴스 학습(MIL)의 훈련은 계산 비용이 많이 들었다. 또한 그래프 합성곱 신경망(GCNN)과 다중 인스턴스 학습(MIL)을 훈련하면 제약되지 않은 잠재 공간(특징 공간의 어느 곳에서도 정상 또는 비정상 특징이 나타날 수 있음을 의미)이 발생하여 성능 불안정성을 유발할 수 있다. Wang 등은 다중 타임스케일 기법을 사용하여 감시 이상 예측을 위한 중간 융합 네트워크를 제안했다. 비디오 프레임 시퀀스는 서로 다른 샘플링 주파수를 사용하여 생성되었고, 콘텐츠와 모션은 두 개의 타임스케일 입력으로 표현되었다. 생성 네트워크는 정상 데이터로 학습되었기 때문에 비정상적인 이벤트를 잘 예측하지 못했다. 테스트 단계에서 생성된 프레임과 기준 프레임 사이에서 이상 이벤트 발생 여부를 판단하기 위해 PSNR 계산을 수행했다. 최근에는 감시 이상 이벤트 감지를 위해 이중 스트림 네트워크가 제안되었다. 저자들은 2D 합성곱 신경망을 자동 인코더로 사용하여 정제된 특징을 추출한 후 첫 번째 스트림의 순차 학습 모델에서 전달하고, 두 번째 스트림의 감시 이벤트에서 3D 특징을 추출했다. 이후 정확한 이상 탐지를 위해 융합 기법을 적용했다. [L. Sun, Y. Chen, W. Luo, H. Wu, and C. Zhang, "Discriminative clip mining for video anomaly detection," in 2020 IEEE International Conference on Image Processing (ICIP), 2020: IEEE, pp. 2121-2125.]의 저자들은 정상 이벤트와 비정상 이벤트를 구분하는 "변별적 비정상 클립 마이너"라는 프레임워크를 도입했다. 또한 차별적 세그먼트와 함께 다중 인스턴스 순위 모델을 통합했다. 어텐션 블록은 다중 인스턴스 학

습(MIL)과 시간적 맥락을 결합하는 모션 특징 학습을 위한 또 다른 메커니즘으로 제안되었다. 트랜스포머는 순수 트랜스포머 네트워크를 생성하여 비전 및 자연어 처리를 위한 통합 아키텍처를 제공한다. 이러한 아키텍처는 비전 작업에 필수적이라고 주장하는 지역성 표현을 학습하는 데 어려움을 겪기 때문에 비전 작업에 최적이지 아닐 수 있다. 본 발명에서는 강력한 귀납적 편향과 글로벌 모델링 기능을 갖춘 하이브리드 합성곱 신경망 및 트랜스포머 기반 프레임워크를 제안한다. 채널 기반 또는 공간적 어텐션을 사용하여 특징을 재계산하는 기존 프레임워크와 달리, 본 발명은 셀프-어텐션 메커니즘을 사용하여 적응적 특징을 개선하고 글로벌 관계를 학습하는 프레임워크를 제시한다.

[0079] 3. 본 발명에 따른 프레임워크

[0080] 도 4는 본 발명의 바람직한 실시예에 따른 감시 이상 탐지 프레임워크를 설명하기 위한 도면이다.

[0081] 도 4에서 볼 수 있듯이, 본 발명은 이상 탐지를 위한 새로운 합성곱 신경망 및 트랜스포머 기반 프레임워크를 제안한다. 이 프레임워크의 주요 구성 요소는 다음 하위 섹션에서 설명한다.

[0082] 3.1. 전처리 및 특징 추출

[0083] 중요 특징을 추출하기 전에, 감시 비디오 V_i 의 모든 프레임은 각 세그먼트 $\mathcal{F}(i, j)$ 가 K개의 겹치지 않는 프레임을 포함하도록 세그먼트 그룹으로 균등하게 나뉜다. 여기서 $i \in [1, n]$ 는 n개의 파일로 구성된 데이터 세트의 영상 인덱스이고, $j \in [1, m_i]$ 는 V_i 에서 m_i 세그먼트의 인덱스를 나타낸다. 각 감시 비디오에는 단일 이진 레이블[정상 = 0, 비정상 = 1]이 할당된다.

[0084] 딥러닝 분야의 연구자들은 최근 모델 아키텍처 및 훈련 기법의 개선과 하이퍼파라미터 최적화에 대한 개선 측면에서 상당한 진전을 이루었다. 최근 Bello 등은 훈련 및 확장에 대한 대체 접근 방식이 ResNet의 성능을 크게 향상시킬 수 있다는 사실을 입증했다. 합성곱 신경망을 사용하는 주된 목적은 가능한 한 효율적인 방식으로 리소스 사용을 극대화하는 것이다. 스케일링을 통해 모델의 정확도를 향상시킬 수 있지만, 이를 위해서는 더 많은 리소스가 필요하다. 이러한 관점에서 Tan과 Le는 네트워크의 깊이, 너비, 해상도 등 네트워크 차원의 스케일링을 연구하여 더 나은 성능을 가진 모델을 개발하는 것을 목표로 했다. 그들은 EfficientNets(B0-B7)이라는 패러다임 제품군을 도입했다. EfficientNet 제품군의 모델은 MobileNet 패러다임의 영향을 받았으며 역 잔여 블록(inverted residual block)(MBConv라고도 함)을 전략적 원칙으로 활용한다.

[0085] 그 결과 저자들은 깊이, 너비, 해상도 비율에 개별적으로 초점을 맞춰 EfficientNet을 위한 복합 스케일링 메커니즘을 개발했다. ImageNet 데이터 세트에서 EfficientNet 모델 제품군(여기서 B0은 기준 모바일 크기를 나타내고 B7은 가장 높은 해상도를 가짐)은 경쟁적으로 성능을 발휘하며, 모델의 매개변수 수와 규모가 증가함에 따라 성능이 향상된다. Tan과 Le는 이러한 문제점을 해결하기 위해 더 빠른 학습 속도를 가진 더 작은 합성곱 신경망 모델을 제안했으며, 이들의 합성곱 신경망 모델은 EfficientNetV2로 알려져 있다. 머신 러닝 분야의 연구자들은 최적의 네트워크 토폴로지, 즉 특정 작업에서 최고의 성능을 달성하는 데 가장 효과적인 네트워크 토폴로지를 결정하기 위해 신경 아키텍처 검색(neural architecture search, NAS)을 비롯한 여러 가지 방법을 사용해 왔다. Tan과 Le는 파라미터 효율성과 훈련 속도를 최적화하기 위해 훈련 인식 NAS와 스케일링 기술을 통합했다. 검색 공간에서 모델을 찾기 위해 이미지 크기에 따라 정규화를 적응적으로 조정하는 프로그레시브 러닝이라는 새로운 방법뿐만 아니라 Fused-MBConv와 같은 새로운 연산이 적용되었다. 이러한 기능은 시각적 콘텐츠를 효율적으로 표현할 수 있어 복잡한 비디오 시퀀스를 더 쉽게 기억하고 학습할 수 있다고 생각한다. 합성곱 신경망 모델은 한 번에 한 프레임씩 처리하도록 훈련되지만, 비디오 데이터에는 일련의 프레임이 포함될 수 있다. 본 발명에서는 비디오 스트림에서 10개의 프레임을 가져와 특징 추출을 위한 백본 모델로 EfficientNetV2를 사용한다. 이 모델은 감시 이벤트의 원시 표현으로서 고차원 특징 벡터를 제공한다. 이 방법은 감시에서 이상 탐지에 중점을 두지만, 비디오의 이벤트 이해와 관련된 다양한 작업에 쉽게 적용할 수 있다.

[0086] 3.2. 시간적 시퀀스 학습

[0087] 도 5는 본 발명의 바람직한 실시예에 따른 트랜스포머 기반 모델의 상세 구성을 설명하기 위한 도면이다.

[0088] 시각적 트랜스포머(visual transformer, ViT)에 대한 최근 연구[A. Dosovitskiy et al., "An image is worth

16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.]에서는 셀프-어텐션 패러다임이 이미지의 효과적인 해석을 필요로 한다는 것을 보여주었다. 이 맥락에서 이미지는 일련의 단어로 간주되며, 포함된 모든 단어는 이미지 패치의 선형 투영을 나타낸다. 다양한 어텐션 영역에 사용할 수 있는 다양한 모델이 있지만, 셀프-어텐션만으로 작동하는 모델은 전적으로 셀프-어텐션 블록에 기반한다. 따라서 시간적 의존성을 여러 가지 방식으로 표현할 수 있는 트랜스포머 기반 모델을 제안하는데, 이는 어텐션 메커니즘을 시간적 의존성으로 확장하기 쉽고 구현이 간편하기 때문이다. 제안한 프레임워크는 컨볼루션 연산자보다 어텐션 메커니즘을 동시에 더 잘 활용할 수 있기 때문에 시간적 정보를 처리할 때 특히 유리하다.

[0089] 본 발명의 목적은 드물게 서브샘플링된 시간적 데이터를 사용하여 정확한 예측을 하고 단기 및 장기 의존성을 모두 얻을 수 있는 모델을 개발하는 것이다. 이 경우 2D 컨볼루션 필터는 이미지의 국소적 연관성을 구현하고 물체 감지 및 인식에 사용하기에 바람직한 특성을 제공하기 때문에 이미지 형성에 적합하다. 장기 종속성을 모델링할 때 트랜스포머는 컨볼루션 모델에 비해 상당한 이점이 있다. 또한, 적절한 수의 헤드를 가진 다중 헤드 셀프-어텐션(multi-head self-attention) 레이어를 사용하여 장거리 상호 작용을 직접 모델링할 수 있지만, 많은 컨볼루션 레이어가 필요하기 때문에 비용이 많이 든다. 트랜스포머의 빌딩 블록을 구성하는 셀프-어텐션 레이어는 순서에 독립적이다. 비디오 시퀀스는 순차적인 프레임으로 구성되므로 트랜스포머 모델은 이 정보를 고려해야 한다. 이는 비디오 내부에 프레임 위치를 임베딩하는 데 사용되는 위치 인코딩 및 임베딩 레이어를 통해 수행된다.

[0090] **입력 임베딩(Input Embeddings):** 공간-시간 트랜스포머에 대한 입력은 $\mathcal{H} \times \mathcal{W}$ 크기의 원본 비디오에서 가져온 RGB 프레임 $\mathcal{F}$ 의 세트인 $\mathcal{X} \in \mathcal{R}^{\mathcal{H} \times \mathcal{W} \times 3 \times \mathcal{F}}$ 이다. 이 입력 블록의 초기 단계에서는 각 프레임이 겹치지 않는 세그먼트로 분할된다. $\mathcal{H} \times \mathcal{W}$ 크기의 프레임에 해당하는 $\mathcal{P} \times \mathcal{P}$ 사이의 패치 수는 $\mathcal{N} = \mathcal{H}\mathcal{W}/\mathcal{P}^2$ 이다. 이러한 패치를 벡터로 평탄화하고 이를 임베딩 벡터로 선형 투영하면 다음의 [수학식 1]과 같은 결과가 나온다.

[0091] [수학식 1]

$$[0092] \mathcal{Z}_{(\mathcal{P}, \mathcal{F}_i)}^{(0)} = \mathbb{E} \mathcal{X}_{(\mathcal{P}, \mathcal{F}_i)} + e_{(\mathcal{P}, \mathcal{F}_i)}^{pos}$$

[0093] 위의 [수학식 1]에서, 입력 $\mathcal{X}_{(\mathcal{P}, \mathcal{F}_i)} \in \mathcal{R}^{3\mathcal{P}^2}$ 와 임베딩 벡터 $\mathcal{Z}_{(\mathcal{P}, \mathcal{F}_i)} \in \mathcal{R}^{\mathcal{D}}$ 는 위치 임베딩 벡터 $e_{(\mathcal{P}, \mathcal{F}_i)}^{pos}$ 와 행렬 $\mathbb{E}$ 를 통해 연결된다. 패치 및 프레임 인덱스는 $\mathcal{P}$ 와 $\mathcal{F}_i$ 이며, 각각 $\mathcal{P} = 1, 2, \dots, \mathcal{N}$ 와 $\mathcal{F}_i = 1, 2, \dots, \mathcal{F}$ 의 값을 갖는다. 트랜스포머 모델은 각 프레임의 정보를

$\mathcal{Z}_{(0, 0)}^{(0)} \in \mathcal{R}^{\mathcal{D}}$ 학습 가능한 토큰으로 인코딩한 다음 비디오에 포함된 프레임 시퀀스에 걸쳐 한 프레임에서 다른 프레임으로 시간적으로 전파함으로써 분류 도구로 사용할 수 있다.

[0094] **다중 헤드 셀프 어텐션 모듈(Multi-head Self-Attention Module, MSA):** 제안된 트랜스포머 기반 프레임워크는 $\mathcal{L}$ 개의 MSA 블록으로 구성된다. 모든 블록헤드 $\mathcal{H} = [1, \dots, \mathcal{A}]$ 와 $\ell \in [1, \dots, \mathcal{L}]$ 에서, 각 패치는 값(value), 키(key), 쿼리 벡터(query vector)로 변환된다. 앞 블록의 출력은 입력으로 사용되는 $\mathcal{Z}_{(\mathcal{P}, \mathcal{F}_i)}^{\ell-1}$ 이다.

[0095] [수학식 2]

$$Q_{(\mathcal{P}, \mathcal{F}_i)}^{(\ell, \mathcal{H})} = W_Q^{(\ell, \mathcal{H})} \mathcal{LN}(Z_{(\mathcal{P}, \mathcal{F}_i)}^{\ell-1}) \in \mathcal{R}^{\mathcal{D}_{\mathcal{H}}}$$

[0096]

[0097] [수학식 3]

$$K_{(\mathcal{P}, \mathcal{F}_i)}^{(\ell, \mathcal{H})} = W_K^{(\ell, \mathcal{H})} \mathcal{LN}(Z_{(\mathcal{P}, \mathcal{F}_i)}^{\ell-1}) \in \mathcal{R}^{\mathcal{D}_{\mathcal{H}}}$$

[0098]

[0099] [수학식 4]

$$V_{(\mathcal{P}, \mathcal{F}_i)}^{(\ell, \mathcal{H})} = W_V^{(\ell, \mathcal{H})} \mathcal{LN}(Z_{(\mathcal{P}, \mathcal{F}_i)}^{\ell-1}) \in \mathcal{R}^{\mathcal{D}_{\mathcal{H}}}$$

[0100]

[0101] 여기서 $\mathcal{LN}$ 은 정규화(normalisation) 레이어이고, 각 어텐션 헤드의 차원은 $\mathcal{D}_{\mathcal{H}} = \mathcal{D} / \mathcal{A}$ 로 표시된다. 어텐션 가중치는 도트 프로덕트(dot product) 방법을 사용하여 쿼리와 키를 비교하여 계산된다. 패치

$(\mathcal{P}, \mathcal{F}_i)$ 에 대한 셀프-어텐션 가중치 $\mathcal{A}_{(\mathcal{P}, \mathcal{F}_i)}^{(\ell, \mathcal{H})} \in \mathcal{R}^{\mathcal{N}^{\mathcal{F}} + \mathcal{F}}$ 는 다음의 [수학식 5]와 같이 표현된다.

[0102] [수학식 5]

$$\mathcal{A}_{(\mathcal{P}, \mathcal{F}_i)}^{(\ell, \mathcal{H})} = \mathcal{SM} \left(Q_{(\mathcal{P}, \mathcal{F}_i)}^{(\ell, \mathcal{H})}{}^T / \sqrt{\mathcal{D}_{\mathcal{H}}} \cdot [\mathcal{K}_{(0, \mathcal{F}_i)}^{(\ell, \mathcal{H})} \{ \mathcal{K}_{(\mathcal{P}', \mathcal{F}_i')}^{(\ell, \mathcal{H})} \}] \right)$$

[0103]

[0104] 여기서 $\mathcal{SM}$ 은 SoftMax 활성화 함수이고, $\mathcal{K}_{(0, \mathcal{F}_i)}^{(\ell, \mathcal{H})}$ 는 프레임 $\mathcal{F}_i$ 의 클래스 토큰과 관련된 키 값을 나타낸다.

[0105] **시간적 어텐션(Temporal Attention):** 공간적 어텐션은 동일한 이미지 간의 패치를 비교하여 이미지의 패치 정보를 강력하게 표현하는 데 사용된다. 도 5에서 볼 수 있듯이, 프레임 시퀀스 내의 시간적 정보는 시간적 어텐션 메커니즘 없이는 캡처할 수 없다. 공간적 어텐션 메커니즘의 중추가 영상을 잘 표현하지만, 시간적 어텐션 메커니즘을 추가하면 이러한 표현이 크게 향상된다. 본 발명에 따른 모델에서는 공간적 어텐션 레이어가 생성한 표현에 시간적 어텐션 레이어를 적용한다. 이 시간적 블록에서는 시퀀스의 입력으로 공간적 어텐션과 동일한

프레임 임베딩 벡터를 사용하여 행렬 $\mathcal{X}_{time} \in \mathcal{R}^{\mathcal{F} \times \mathcal{D}}$ 에 쌓아 올린다. 시퀀스 분류에 사용되는 토

큰 이전과 마찬가지로 인덱스 $\mathcal{F}_i = 0$ 에 첨부된다. 이 경우 입력 시퀀스 데이터는 쿼리 $Q_{\mathcal{F}_i}^{(\ell, \mathcal{H})}$, 키 $K_{\mathcal{F}_i}^{(\ell, \mathcal{H})}$ 및 값 $V_{\mathcal{F}_i}^{(\ell, \mathcal{H})}$ 벡터에 투영된다. 이 단계의 시간적 어텐션 레이어는 프레임 인덱스만 처리한다.

[0106] [수학식 6]

$$\mathcal{A}_{(\mathcal{P}, \mathcal{F}_i)}^{(\ell, \mathcal{H})} = \mathcal{SM} \left(Q_{\mathcal{F}_i}^{(\ell, \mathcal{H})}{}^T / \sqrt{\mathcal{D}_{\mathcal{H}}} \cdot [\mathcal{K}_0^{(\ell, \mathcal{H})} \{ \mathcal{K}_{\mathcal{F}_i'}^{(\ell, \mathcal{H})} \}] \right)$$

[0107]

[0108] 비디오 시퀀스의 최상위 레이어를 정규화한 후, 비디오 시퀀스의 최상위 레이어에 대한 분류 임베딩을 사용하여

비디오 시퀀스에 대한 임베디드 벡터를 얻을 수 있다.

[수학식 7]

$$y = \mathcal{LN} \left(\mathcal{Z}_{(0)}^{(\mathbb{L}_{time})} \right) \in \mathcal{R}^D$$

여기서 $\mathbb{L}_{time}$ 은 시간적 어텐션 레이어의 수를 나타낸다. 추가 레이어로 MLP를 적용하여 클래스 수와 동일한 차원의 벡터를 생성한다. 본 발명에서 제안한 프레임워크는 합성곱 신경망과 트랜스포머 모델의 장점을 효율적으로 계승하여 스케일/시프트 불변 특징 해석을 발견하고 입력 데이터의 복잡한 감시 이벤트 간의 장기적인 의존성을 모델링할 수 있다.

4. 실험 결과 및 평가

도 6은 본 발명의 바람직한 실시예에 따른 감시 이상 탐지 프레임워크의 AUC 성능을 설명하기 위한 도면이다.

이 섹션에서는 본 발명이 제안한 프레임워크의 성능을 평가하기 위해 상하이테크(ShanghaiTech), UCSD Ped2, CUHK 애비뉴(avenue) 등 여러 벤치마크 이상 탐지 데이터 세트에 대한 다양한 실험 결과를 보고한다. 또한 본 발명이 제안한 프레임워크의 우수성을 테스트하기 위해 다른 SOTA SVAD 방법과 비교한다.

- 상하이테크(ShanghaiTech) 데이터 세트: 총 437개의 감시 비디오 클립이 포함된 가장 큰 규모의 비디오 이상 감지(VAD) 벤치마크 데이터 세트 중 하나이다. 별도의 훈련 및 테스트 세트가 제공되며, 각각 330개와 107개의 비디오로 다양한 카메라 각도와 다양한 조명 조건에서 캡처된 13개의 비정상 장면이 포함되어 있다. 이 데이터 세트에는 추격, 점프, 싸움, 오토바이 등과 같은 130개의 비정상적인 이벤트가 포함되어 있다.

- 애비뉴(avenue) 데이터 세트: 이 데이터 세트는 640×360픽셀 해상도의 37개 동영상으로 구성되어 있으며, 각각 16개와 21개의 동영상이 포함된 훈련 세트와 테스트 세트로 나뉜다. 비정상적인 이벤트는 던지기, 배회, 사람 뛰기 등으로 레이블이 지정된다. 각 비디오 클립은 2분 길이이며 픽셀 수준 주석을 포함한다. 총 비디오 프레임 수는 약 31,000개이며, 모든 비디오는 고정식 카메라로 녹화되었다.

- UCSD Ped2 데이터 세트: 여기에는 총 38개의 비디오와 12개의 다양한 이상 이벤트가 포함되어 있으며, 고정된 시각 센서로 캡처되었다. 이 비디오 클립의 군중 밀도는 다양하다. 데이터 세트는 16개의 클립으로 구성된 훈련 세트와 12개의 클립으로 구성된 테스트 세트로 나뉘며, 총 4,950개의 프레임으로 구성되어 있다.

4.1. 상하이테크(ShanghaiTech) 데이터 세트에서 제안된 프레임워크의 실험 평가

상하이테크(ShanghaiTech) 데이터 세트에서 본 발명이 제안한 프레임워크는 모든 SOTA 기법[5, 6, 12, 27, 49, 50, 53, 56, 64-68]을 능가하는 성능을 보였으며, 가장 최근 기법보다 1.63% 높은 평균 AUC 94.6%를 달성했다. 자세한 실험 결과는 도 6에 도시된 표에 나와 있다.

4.2. UCSD Ped2 데이터 세트에서 제안된 프레임워크의 실험 평가

도 6에 도시된 표는 기존 SOTA 기법[20, 27, 54-63, 66-69]과 비교하여 제안한 프레임워크의 우수성을 보여주는 UCSD-Ped2 데이터 세트에 대한 실험 결과를 보여준다. 본 발명의 방법은 97.8%의 AUC 점수를 달성했으며, 이는 가장 최근의 방법보다 0.8% 포인트 높은 수치이다.

4.3. 애비뉴(avenue) 데이터 세트에서 제안된 프레임워크의 실험 평가

애비뉴(Avenue) 데이터 세트에서 제안한 프레임워크의 AUC 성능은 도 6에 도시된 표에 나와 있다. 본 발명이 제안한 프레임워크는 다른 SOTA 기법[20, 27, 54-57, 60-63, 66-68, 71-73]보다 더 나은 성능을 보였으며, 가장 최근의 방법보다 0.1% 높은 89.6%의 AUC를 기록했다. 이는 본 발명에 따른 모델이 감시 이상 탐지에 대한 장기적인 시간적 종속성을 다른 SOTA 기법보다 더 효과적으로 포착했음을 나타낸다.

본 실시예들에 따른 동작은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨

터 판독 가능한 저장 매체에 기록될 수 있다. 컴퓨터 판독 가능한 저장 매체는 실행을 위해 프로세서에 명령어를 제공하는데 참여한 임의의 매체를 나타낸다. 컴퓨터 판독 가능한 저장 매체는 프로그램 명령, 데이터 파일, 데이터 구조 또는 이들의 조합을 포함할 수 있다. 예컨대, 자기 매체, 광기록 매체, 메모리 등이 있을 수 있다. 컴퓨터 프로그램은 네트워크로 연결된 컴퓨터 시스템 상에 분산되어 분산 방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수도 있다. 본 실시예를 구현하기 위한 기능적인(Functional) 프로그램, 코드, 및 코드 세그먼트들은 본 실시예가 속하는 기술 분야의 프로그래머들에 의해 용이하게 추론될 수 있을 것이다.

[0128] 본 실시예들은 본 실시예의 기술 사상을 설명하기 위한 것이고, 이러한 실시예에 의하여 본 실시예의 기술 사상의 범위가 한정되는 것은 아니다. 본 실시예의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 실시예의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

[0129]

100 : 감시 이상 탐지 장치,

110 : 프로세서,

130 : 컴퓨터 판독 가능한 저장 매체,

131 : 프로그램,

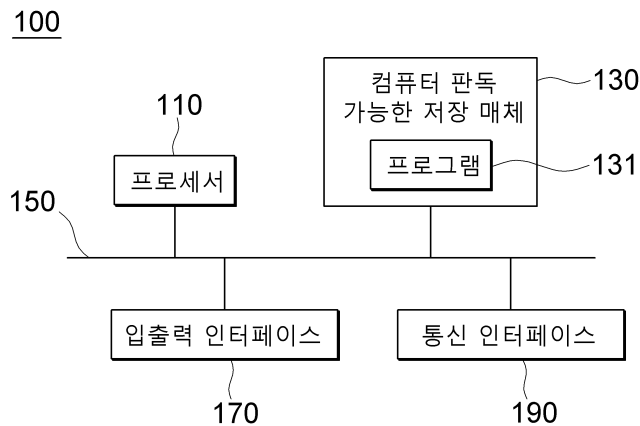
150 : 통신 버스,

170 : 입출력 인터페이스,

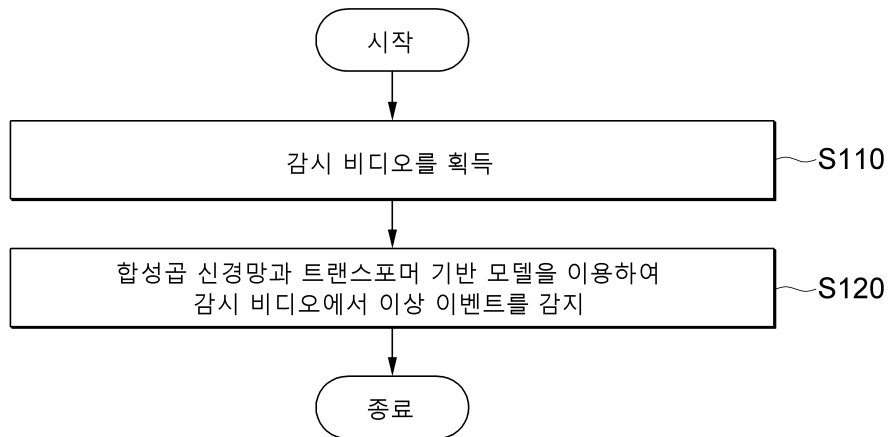
190 : 통신 인터페이스

도면

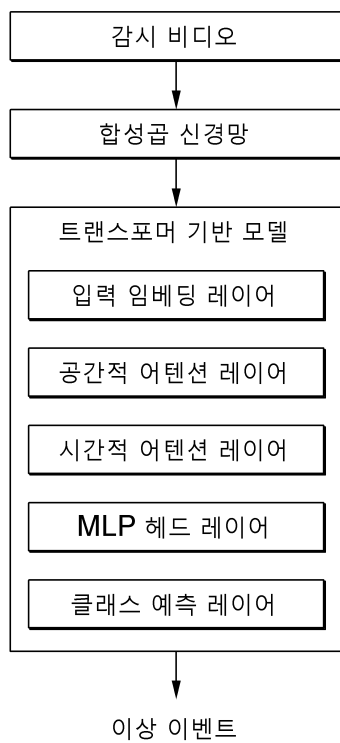
도면1



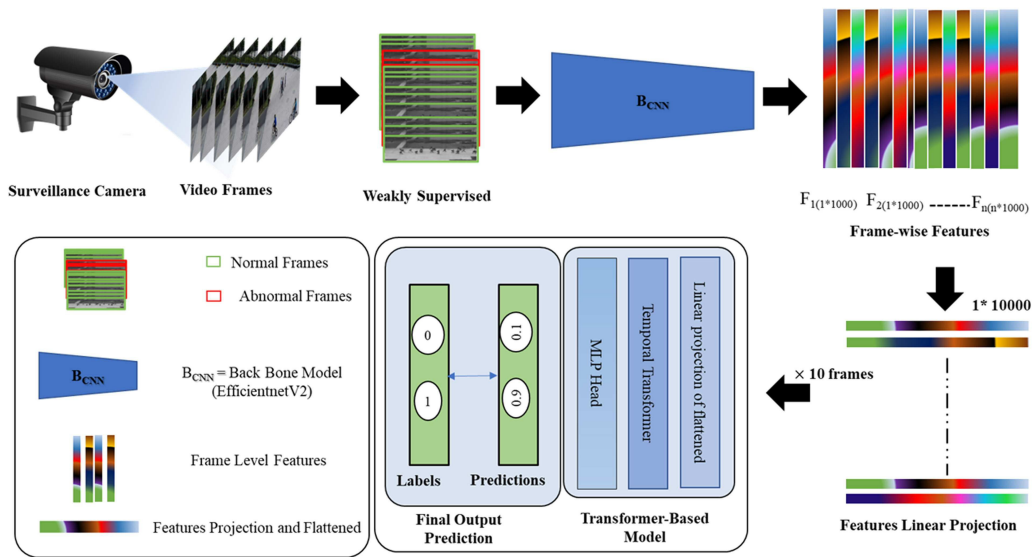
도면2



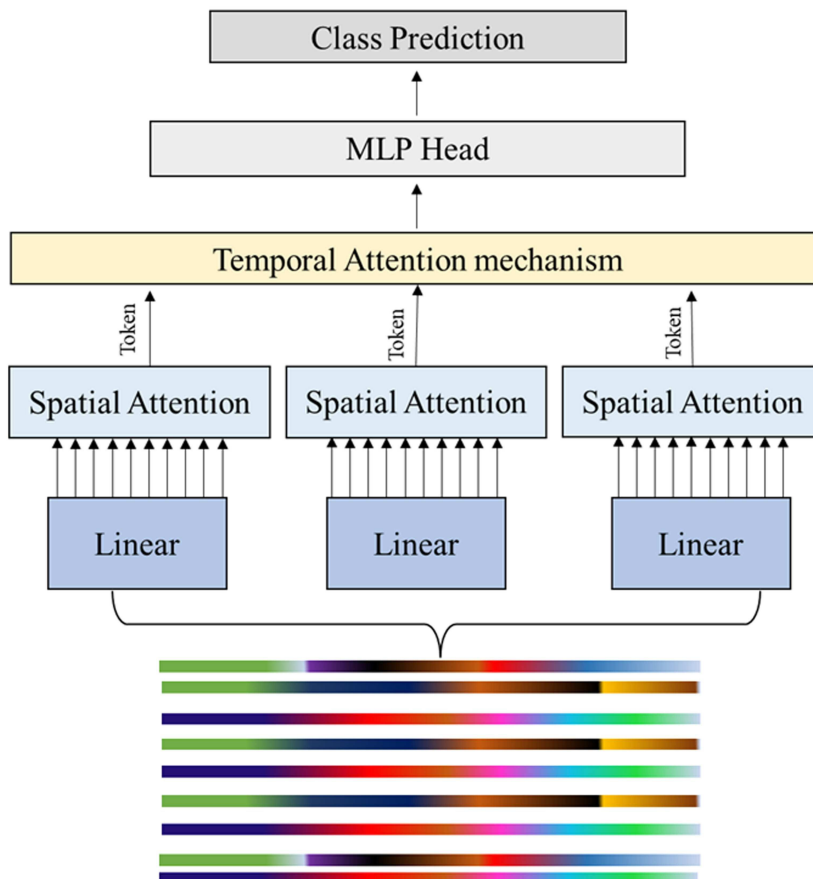
도면3



도면4



도면5



도면6

Technique	Backbone	AUC (%)		
		Ped2	Avenue	ShanghaiTech
Conv-AE [49]	Autoencoders	-	-	60.9
CLSTM-AE [50]	ConvLSTM with autoencoder	-	-	55.0
TSC [12]	Temporally coherent sparse coding with RNN	-	-	67.9
OGNet [51]	Adversarial network	-	-	69.26
RW [6]	C3D	-	-	71.5
FFP [5]	Generator (U-Net)	-	-	72.8
MLEP [52]	Margin learning module	-	-	73.40
MNAD [27]	Memory module	97.0	88.5	51.13
sAdaNet [53]	Autoencoder with attention mechanism	-	-	70.0
AMAE [54]	CNN-based reconstruction model	96.2	86.9	-
AnomalyNet [55]	Sparse long short-term memory	94.9	86.1	-
AdaNet [53]	Autoencoder with attention mechanism	-	-	70.0
AR-Net [33]	I3D	-	-	91.24
MemAE [56]	Memory-augmented autoencoder	94.1	83.3	71.2
ISTL [57]	Fuzzy aggregation	91.1	76.8	-
ADLNET [58]	Attention-driven loss	96.1	-	-
BMAN [59]	Multi-scale and attention schemes	96.6	-	-
MESDnet [60]	Multiple encoders	95.6	86.3	-
STCaE [61]	Two-stream framework	92.9	83.5	73.2
AMMCN [62]	Memory mechanism for both motion and appearance	96.6	86.6	73.7
RUVAD [63]	ConvGRU	96.3	88.3	-
CL [64]	Continual learning	-	86.4	71.6
ClusterAE [65]	Autoencoder	-	-	73.3
sRNN-AE [66]	Temporal coherence-based sparse coding	92.2	83.4	69.6
ITAEGM [67]	Two encoders with single decoder	96.8	86.0	73.0
SSAGAN [68]	GAN	97.6	88.8	74.3
DPU [69]	Meta-learning and few-shot	96.9	89.5	73.8
MTDNN [70]	I3D	-	-	92.97
SIGnet [20]	Siamese generative network	96.2	86.8	-
IFN [35]	Intermediate fused network	96.0	86.3	74.5
STD [28]	Spatio-temporal dissociation	96.7	87.1	73.7
Proposed method		98.4	89.6	94.6