

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)(11) 공개번호 10-2025-0023864
(43) 공개일자 2025년02월18일

(51) 국제특허분류(Int. Cl.)

G06V 40/20 (2022.01) G06V 10/62 (2022.01)
G06V 10/774 (2022.01) G06V 10/94 (2022.01)
G06V 20/40 (2022.01)

(52) CPC특허분류

G06V 40/20 (2022.01)
G06N 3/0455 (2023.01)

(21) 출원번호 10-2023-0105125

(22) 출원일자 2023년08월10일

심사청구일자 2023년08월10일

(71) 출원인

계명대학교 산학협력단

대구광역시 달서구 달구벌대로 1095, 계명대학교
산학협력관 201호(신당동)

(72) 발명자

고병철

대구광역시 수성구 범어천로 107

안다습

대구광역시 북구 침산로21길 39

(74) 대리인

김건우

전체 청구항 수 : 총 16 항

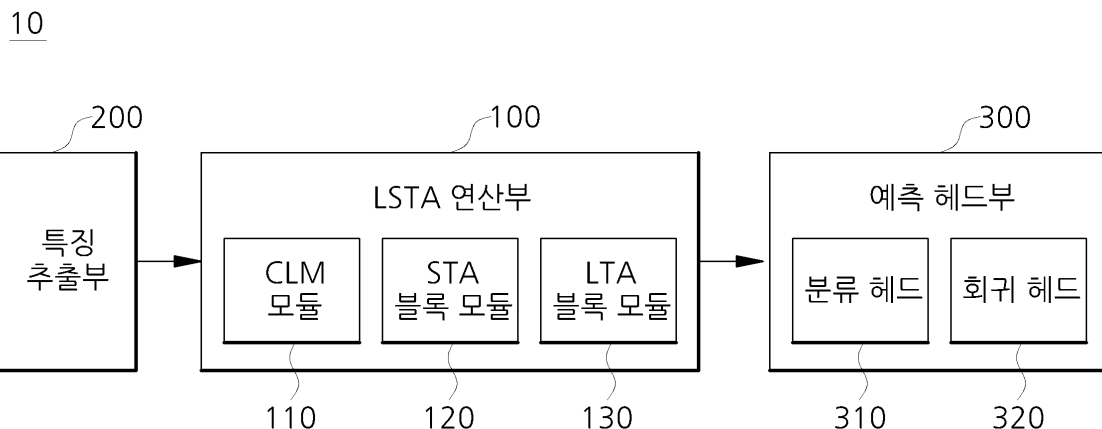
(54) 발명의 명칭 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치 및 방법

(57) 요약

본 발명은 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치에 관한 것으로서, 보다 구체적으로는 트랜스포머 기반으로 시간적 행동(temporal action)을 검출하는, 검출 장치로서, 입력 비디오로부터 추출된 특징을 사용해 특징 피라미드 구조를 생성하여 여러 규모의 시간적 특징을 통합하고, 장기와 단기의 시간적 특징 간의 관

(뒷면에 계속)

대표도 - 도1



계를 학습하여 장단기 시간적 주의(long- and short-term temporal attention; LSTA)를 연산하는 LSTA 연산부를 포함하되, 상기 LSTA 연산부는, 상기 입력 비디오에서 추출된 특징을 사용해 장기 메모리를 생성하는 CLM(compact long-term memory, 소형 장기 메모리) 모듈; 입력 특징을 나누어 단기 시간적 특징을 생성하고, 상기 CLM 모듈에서 구성된 장기 메모리와 상기 단기 시간적 특징을 사용해 장기와 단기의 시간적 특징 간의 관계를 학습하고 단기 시간적 특징을 강화하는 STA(short-term temporal attention) 블록 모듈; 및 장기 시간적 특징 전체에서 필요한 부분을 동적으로 학습하고 강화하는 LTA(long-term temporal attention) 블록 모듈을 포함하며, 상기 STA 블록 모듈이 복수의 레벨을 이루고 상위 레벨의 STA 블록 모듈은 하위 레벨의 STA 블록 모듈의 출력 특징을 입력 특징으로 사용하며, 각 레벨의 상기 STA 블록 모듈의 입력 및 출력 특징이 상기 특징 피라미드 구조를 이루는 것을 그 구성상의 특징으로 한다.

또한, 본 발명은 비디오의 장단기 의존성을 이용한 시간적 행동 검출 방법에 관한 것으로서, 보다 구체적으로는 트랜스포머 기반으로 시간적 행동을 검출하는, 컴퓨터에서 각 단계가 수행되는 검출 방법으로서, 입력 비디오로부터 추출된 특징을 사용해 특징 피라미드 구조를 생성하여 여러 규모의 시간적 특징을 통합하고, 장기와 단기의 시간적 특징 간의 관계를 학습하여 장단기 시간적 주의(long- and short-term temporal attention; LSTA)를 연산하는 LSTA 연산부를 포함하는 LSTA 모델을 학습하는 학습 단계; 및 상기 학습된 LSTA 모델을 이용해 입력 비디오로부터 시간적 행동을 예측하는 예측 단계를 포함하되, 상기 LSTA 연산부는, 상기 입력 비디오에서 추출된 특징을 사용해 장기 메모리를 생성하는 CLM(compact long-term memory, 소형 장기 메모리) 모듈; 입력 특징을 나누어 단기 시간적 특징을 생성하고, 상기 CLM 모듈에서 구성된 장기 메모리와 상기 단기 시간적 특징을 사용해 장기와 단기의 시간적 특징 간의 관계를 학습하고 단기 시간적 특징을 강화하는 STA 블록 모듈; 및 장기 시간적 특징 전체에서 필요한 부분을 동적으로 학습하고 강화하는 LTA(long-term temporal attention) 블록 모듈을 포함하며, 상기 STA 블록 모듈이 복수의 레벨을 이루고 상위 레벨의 STA 블록 모듈은 하위 레벨의 STA 블록 모듈의 출력 특징을 입력 특징으로 사용하며, 각 레벨의 상기 STA 블록 모듈의 입력 및 출력 특징이 상기 특징 피라미드 구조를 이루는 것을 그 구성상의 특징으로 한다.

본 발명에서 제안하고 있는 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치 및 방법에 따르면, 시퀀스 분석 작업에서 중요한 장기 및 단기 시간적 특징 간의 상호 작용을 활용하는 장단기 시간적 주의 기반의 새로운 TAD 모델을 도입하되, 장기 메모리를 고려하여 단기 시간적 특징을 학습하고 강화하는 STA 블록 모듈과, 장기 시간적 특징 전체에서 필요한 부분을 동적으로 학습하고 강화하는 LTA 블록 모듈을 포함함으로써, 비디오의 장단기 의존성을 이용해 시간적 행동 검출의 성능을 향상시킬 수 있다.

(52) CPC특허분류

G06N 3/08 (2023.01)
G06V 10/62 (2023.08)
G06V 10/774 (2023.08)
G06V 10/955 (2022.01)
G06V 20/46 (2022.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1345363225
과제번호	2022R1I1A3058128
부처명	교육부
과제관리(전문)기관명	한국연구재단
연구사업명	지역대학우수과학자지원사업
연구과제명	자기-지도학습 기반의 경량 시공간 그래프 트랜스포머연구
기 여 율	1/1
과제수행기관명	계명대학교
연구기간	2023.03.01 ~ 2024.02.29

명세서

청구범위

청구항 1

트랜스포머 기반으로 시간적 행동(temporal action)을 검출하는, 검출 장치로서,

입력 비디오로부터 추출된 특징을 사용해 특징 피라미드 구조를 생성하여 여러 규모의 시간적 특징을 통합하고, 장기와 단기의 시간적 특징 간의 관계를 학습하여 장단기 시간적 주의(long- and short-term temporal attention; LSTA)를 연산하는 LSTA 연산부(100)를 포함하되,

상기 LSTA 연산부(100)는,

상기 입력 비디오에서 추출된 특징을 사용해 장기 메모리를 생성하는 CLM(compact long-term memory, 소형 장기 메모리) 모듈(110);

입력 특징을 나누어 단기 시간적 특징을 생성하고, 상기 CLM 모듈(110)에서 구성된 장기 메모리와 상기 단기 시간적 특징을 사용해 장기와 단기의 시간적 특징 간의 관계를 학습하고 단기 시간적 특징을 강화하는 STA(short-term temporal attention) 블록 모듈(120); 및

장기 시간적 특징 전체에서 필요한 부분을 동적으로 학습하고 강화하는 LTA(long-term temporal attention) 블록 모듈(130)을 포함하며,

상기 STA 블록 모듈(120)이 복수의 레벨을 이루고 상위 레벨의 STA 블록 모듈(120)은 하위 레벨의 STA 블록 모듈(120)의 출력 특징을 입력 특징으로 사용하며, 각 레벨의 상기 STA 블록 모듈(120)의 입력 및 출력 특징이 상기 특징 피라미드 구조를 이루는 것을 특징으로 하는, 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치.

청구항 2

제1항에 있어서,

백본 네트워크를 통해 입력 비디오로부터 특징을 추출하고, 특징 임베딩을 통해 추출된 특징을 토근화하는 특징 추출부(200)를 더 포함하는 것을 특징으로 하는, 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치.

청구항 3

제1항에 있어서, 상기 CLM 모듈(110)은,

학습 가능한 두 개의 빈 쿼리(Query, Q)와 트랜스포머 모델의 디코더를 사용해 압축된 상기 장기 메모리를 생성하는 것을 특징으로 하는, 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치.

청구항 4

제1항에 있어서, 상기 STA 블록 모듈(120)은,

미리 정의된 단기 윈도우(short-term window)를 사용해 입력 특징을 단기 메모리 특징으로 분할하고, 상기 분할된 단기 메모리 특징으로부터 쿼리를 계산하며, 상기 CLM 모듈(110)에서 생성된 장기 메모리로부터 키(Key, K) 및 값(Value, V)을 계산하여, 상기 쿼리, 키 및 값을 사용해 교차 주의(cross-attention) 연산을 수행하는 것을 특징으로 하는, 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치.

청구항 5

제4항에 있어서, 상기 STA 블록 모듈(120)은,

상기 교차 주의 연산을 통해, 상기 단기 메모리 특징을 강화하면서 상기 장기 메모리와 단기 메모리의 관계를 학습하여, 장기 시간 의존성을 인코딩하는 단기 시간적 특징을 생성하는 것을 특징으로 하는, 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치.

청구항 6

제5항에 있어서, 상기 STA 블록 모듈(120)은,

상기 분할 모듈에서 분할된 모든 상기 단기 메모리 특징에 대해 생성된 상기 단기 시간적 특징을 연결하여 출력 특징을 얻는 것을 특징으로 하는, 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치.

청구항 7

제1항에 있어서, 상기 LTA 블록 모듈(130)은,

상기 입력 비디오로부터 추출된 특징에 대한 기준점(reference point) 및 오프셋(offset)을 사용해 선택된 토큰으로부터 키와 값을 계산하고, 상기 STA 블록 모듈(120)의 출력 특징에서 생성된 쿼리 및 상기 계산된 키와 값을 사용해 교차 주의 연산을 적용해 장기 시간적 특징을 강화하는 것을 특징으로 하는, 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치.

청구항 8

제1항에 있어서,

상기 LTA 블록 모듈(130)의 출력으로부터 행동 레이블을 예측하는 분류 헤드(310)와 행동 경계(action boundary)를 예측하는 회귀 헤드(320)를 포함하여 구성되는 예측 헤드부(300)를 더 포함하는 것을 특징으로 하는, 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치.

청구항 9

트랜스포머 기반으로 시간적 행동을 검출하는, 컴퓨터에서 각 단계가 수행되는 검출 방법으로서,

입력 비디오로부터 추출된 특징을 사용해 특징 피라미드 구조를 생성하여 여러 규모의 시간적 특징을 통합하고, 장기와 단기의 시간적 특징 간의 관계를 학습하여 장단기 시간적 주의(long- and short-term temporal attention; LSTA)를 연산하는 LSTA 연산부(100)를 포함하는 LSTA 모델(10)을 학습하는 학습 단계(S100); 및

상기 학습된 LSTA 모델(10)을 이용해 입력 비디오로부터 시간적 행동을 예측하는 예측 단계(S200)를 포함하되,

상기 LSTA 연산부(100)는,

상기 입력 비디오에서 추출된 특징을 사용해 장기 메모리를 생성하는 CLM(compact long-term memory, 소형 장기 메모리) 모듈(110);

입력 특징을 나누어 단기 시간적 특징을 생성하고, 상기 CLM 모듈(110)에서 구성된 장기 메모리와 상기 단기 시간적 특징을 사용해 장기와 단기의 시간적 특징 간의 관계를 학습하고 단기 시간적 특징을 강화하는 STA(short-term temporal attention) 블록 모듈(120); 및

장기 시간적 특징 전체에서 필요한 부분을 동적으로 학습하고 강화하는 LTA(long-term temporal attention) 블록 모듈(130)을 포함하며,

상기 STA 블록 모듈(120)이 복수의 레벨을 이루고 상위 레벨의 STA 블록 모듈(120)은 하위 레벨의 STA 블록 모

들(120)의 출력 특징을 입력 특징으로 사용하며, 각 레벨의 상기 STA 블록 모듈(120)의 입력 및 출력 특징이 상기 특징 피라미드 구조를 이루는 것을 특징으로 하는, 비디오의 장단기 의존성을 이용한 시간적 행동 검출 방법.

청구항 10

제9항에 있어서, 상기 LSTA 모듈(10)은,

백본 네트워크를 통해 입력 비디오로부터 특징을 추출하고, 특징 임베딩을 통해 추출된 특징을 토근화하는 특징 추출부(200)를 더 포함하는 것을 특징으로 하는, 비디오의 장단기 의존성을 이용한 시간적 행동 검출 방법.

청구항 11

제9항에 있어서, 상기 CLM 모듈(110)은,

학습 가능한 두 개의 빈 쿼리(Query, Q)와 트랜스포머 모델의 디코더를 사용해 압축된 상기 장기 메모리를 생성하는 것을 특징으로 하는, 비디오의 장단기 의존성을 이용한 시간적 행동 검출 방법.

청구항 12

제9항에 있어서, 상기 STA 블록 모듈(120)은,

미리 정의된 단기 윈도우(short-term window)를 사용해 입력 특징을 단기 메모리 특징으로 분할하고, 상기 분할된 단기 메모리 특징으로부터 쿼리를 계산하며, 상기 CLM 모듈(110)에서 생성된 장기 메모리로부터 키(Key, K) 및 값(Value, V)을 계산하여, 상기 쿼리, 키 및 값을 사용해 교차 주의(cross-attention) 연산을 수행하는 것을 특징으로 하는, 비디오의 장단기 의존성을 이용한 시간적 행동 검출 방법.

청구항 13

제12항에 있어서, 상기 STA 블록 모듈(120)은,

상기 교차 주의 연산을 통해, 상기 단기 메모리 특징을 강화하면서 상기 장기 메모리와 단기 메모리의 관계를 학습하여, 장기 시간 의존성을 인코딩하는 단기 시간적 특징을 생성하는 것을 특징으로 하는, 비디오의 장단기 의존성을 이용한 시간적 행동 검출 방법.

청구항 14

제13항에 있어서, 상기 STA 블록 모듈(120)은,

상기 분할 모듈에서 분할된 모든 상기 단기 메모리 특징에 대해 생성된 상기 단기 시간적 특징을 연결하여 출력 특징을 얻는 것을 특징으로 하는, 비디오의 장단기 의존성을 이용한 시간적 행동 검출 방법.

청구항 15

제9항에 있어서, 상기 LTA 블록 모듈(130)은,

상기 입력 비디오로부터 추출된 특징에 대한 기준점(reference point) 및 오프셋(offset)을 사용해 선택된 토근으로부터 키와 값을 계산하고, 상기 STA 블록 모듈(120)의 출력 특징에서 생성된 쿼리 및 상기 계산된 키와 값을 사용해 교차 주의 연산을 적용해 장기 시간적 특징을 강화하는 것을 특징으로 하는, 비디오의 장단기 의존성을 이용한 시간적 행동 검출 방법.

청구항 16

제9항에 있어서, 상기 LSTA 모델(10)은,

상기 LTA 블록 모듈(130)의 출력으로부터 행동 레이블을 예측하는 분류 헤드(310)와 행동 경계(action boundary)를 예측하는 회귀 헤드(320)를 포함하여 구성되는 예측 헤드부(300)를 더 포함하는 것을 특징으로 하는, 비디오의 장단기 의존성을 이용한 시간적 행동 검출 방법.

발명의 설명**기술 분야**

[0001] 본 발명은 시간적 행동 검출에 관한 것으로서, 보다 구체적으로는 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치 및 방법에 관한 것이다.

배경 기술

[0002] 이 부분에 기술된 내용은 단순히 본 발명의 일실시예에 대한 배경 정보를 제공할 뿐 종래기술을 구성하는 것은 아니다.

[0004] TAD(시간적 행동 검출, Temporal Action Detection)는 인간의 행동을 예측하고 비디오 시퀀스의 시작 및 끝 프레임에 로컬화하는 것을 목표로 한다. TAD는 비디오 이해 분야에서 도전적이지만 필수적인 알고리즘으로 자율 주행 및 비디오 분석과 같은 분야에서 널리 사용된다. 비디오의 시간적 특성을 이해하는 것은 인간의 행동과 그 경계를 예측할 때 특히 인공 지능 모델에 중요하다. 이러한 이해를 달성하기 위해 시간적 특성의 학습 및 예측과 관련하여 앵커 및 제안 기반, 순환(recurrent) 및 컨볼루션(convolution) 모델과 같은 다양한 접근법이 연구되었다. 최근 긴-범위(long-range) 의존성 학습에서 다중 헤드 자기-주의 (Multi head self-attention; MHSA) (비특허문헌 0009)를 통합한 변압기(transformer) 기반 연구 (비특허문헌 0006, 0012)가 수행되었다.

[0006] 비디오 시퀀스에서 여러 동작을 인식하려면, 발생할 수 있는 다양한 시간 연속성에 대한 깊은 이해가 필요하다. 이러한 시간 연속성은 긴 기간 동안 유지되는 장기 시간 특징(long-term temporal features)과 상대적으로 짧은 기간 동안 유지되는 단기 시간 특징(short-term temporal features)을 기반으로 결정되거나 예측된다. 장기 시간 특징은 먼 과거의 사건이 현재의 사건에 영향을 미칠 때 중요하다. 예를 들어, 이전 동작으로서의 ‘요리’는 ‘음식을 먹는’ 동작을 예측하기 위한 중요한 단서이다. 단기 시간 특징은 TAD에서도 중요하다. 예를 들어, 각 프레임 또는 클립에서 발생하는 특정 동작은 비디오 시퀀스에서 동작 인식에 중요하다.

[0008] 시간적 행동 검출과 관련하여, TAD는 객체 검출 모델과 유사하며 크게 1단계(one-stage) 방법(비특허문헌 0001, 0006, 0012)과 2단계(two-stage) 방법(비특허문헌 0008, 0011)으로 분류될 수 있다. 앵커 기반 방법과 같은 2단계 방법에서는 먼저 움직임 경계 후보를 생성한 후 움직임 예측을 수행한다. 2단계 방법은 속도 및 모델 크기 측면에서 추가적인 서버 모델의 요구로 인해 비효율적이다. 따라서 최근 연구에서는 2단계 접근법과 대조적으로 행동과 그 경계를 모두 직접 예측하는 1단계 방법을 사용하고 있다. 1단계 방법은 행동 경계를 추정할 때 모든 프레임에 대해 행동을 개별적으로 예측한다.

[0010] 한편, 트랜스포머(비특허문헌 0009)는 자연어 처리 작업에서 상당한 성공을 거두었고, 비전 트랜스포머(ViT) 아키텍처를 사용하여 비전 작업에 성공적으로 적용되었다. MHSA (multi-head self-attention) 메커니즘은 모든 입력 특징 간의 관계를 학습하여 긴-범위의 종속성(dependency)을 모델링하는 이점을 트랜스포머에 제공한다. 따라서 트랜스포머는 시간적 학습(temporal learning)에 매우 적합하다. ActionFormer (비특허문헌 0012)를 포함한 많은 트랜스포머 기반 TAD 모델 (비특허문헌 0006, 0012)은 이러한 이점을 활용하도록 제안되었다. ActionFormer는 특징 피라미드 구조와 경량 컨볼루션 디코더를 사용하여 성능을 향상시킨다. ActionFormer에 기반한 많은 모델이 제안되었다. 특히 ActionFormer 기반 모델인 TriDet (비특허문헌 0006)는 우수한 성능을

달성하기 위해 상대적 경계 모델링 기술을 추가한다.

- [0012] 전술한 배경 기술은 발명자가 본 발명의 도출을 위해 보유하고 있었거나, 본 발명의 도출 과정에서 습득한 기술 정보로서, 반드시 본 발명의 출원 전에 일반 공중에게 공개된 공지 기술이라 할 수는 없다.

선행기술문헌

비특허문헌

- [0013] (비특허문헌 0001) C. Lin, C. Xu, D. Luo, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Fu. Learning salient boundary feature for anchor-free temporal action localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3320-3329, 2021.
- (비특허문헌 0002) T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal ℓ loss for dense object detection. In Proceedings of the IEEE international conference on computer vision, pages 2980-2988, 2017.
- (비특허문헌 0003) X. Liu, Q. Wang, Y. Hu, X. Tang, S. Zhang, S. Bai, and X. Bai. End-to-end temporal action detection with transformer. IEEE Transactions on Image Processing, 31:5427-5441, 2022.
- (비특허문헌 0004) Z. Qing, H. Su, W. Gan, D. Wang, W. Wu, X. Wang, Y. Qiao, J. Yan, C. Gao, and N. Sang. Temporal context aggregation network for temporal action proposal refinement. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 485-494, 2021.
- (비특허문헌 0005) H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 658-666, 2019.
- (비특허문헌 0006) D. Shi, Y. Zhong, Q. Cao, L. Ma, J. Li, and D. Tao. Tridet: Temporal action detection with relative boundary modeling. arXiv preprint arXiv:2303.07347, 2023.
- (비특허문헌 0007) D. Shi, Y. Zhong, Q. Cao, J. Zhang, L. Ma, J. Li, and D. Tao. React: Temporal action detection with relational queries. In Computer Vision-ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part X, pages 105-121. Springer, 2022.
- (비특허문헌 0008) J. Tan, J. Tang, L. Wang, and G. Wu. Relaxed transformer decoders for direct action proposal generation. In Proceedings of the IEEE/CVF international conference on computer vision, pages 13526-13535, 2021.
- (비특허문헌 0009) A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.
- (비특허문헌 0010) M. Xu, Y. Xiong, H. Chen, X. Li, W. Xia, Z. Tu, and S. Soatto. Long short-term transformer for online action detection. Advances in Neural Information Processing Systems, 34:1086-1099, 2021.
- (비특허문헌 0011) M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. Ghanem. G-tad: Sub-graph localization for temporal action detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10156-10165, 2020.
- (비특허문헌 0012) C.-L. Zhang, J. Wu, and Y. Li. Actionformer: Localizing moments of actions with transformers. In Computer Vision-ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IV, pages 492-510. Springer, 2022.
- (비특허문헌 0013) C. Zhao, A. K. Thabet, and B. Ghanem. Video self-stitching graph network for temporal action localization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 13658-13667, 2021.

(비특허문헌 0014) Z. Zhu, W. Tang, L. Wang, N. Zheng, and G. Hua. Enriching local and global contexts for temporal action localization. In Proceedings of the IEEE/CVF international conference on computer vision, pages 13516-13525, 2021.

발명의 내용

해결하려는 과제

[0014] 본 발명은 기존에 제안된 방법들의 상기와 같은 문제점들을 해결하기 위해 제안된 것으로서, 시퀀스 분석 작업에서 중요한 장기 및 단기 시간적 특징 간의 상호 작용을 활용하는 장단기 시간적 주의(long- and short-term temporal attention; LSTA) 기반의 새로운 TAD 모델을 도입하되, 장기 메모리를 고려하여 단기 시간적 특징을 학습하고 강화하는 STA(short-term temporal attention) 블록 모듈과, 장기 시간적 특징 전체에서 필요한 부분을 동적으로 학습하고 강화하는 LTA(long-term temporal attention) 블록 모듈을 포함함으로써, 비디오의 장단기 의존성을 이용해 시간적 행동 검출의 성능을 향상시킬 수 있는, 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치 및 방법을 제공하는 것을 그 목적으로 한다.

[0016] 다만, 본 발명이 이루고자 하는 기술적 과제는 상기한 바와 같은 기술적 과제로 한정되지 않으며, 또 다른 기술적 과제들이 존재할 수 있고, 명시적으로 언급하지 않더라도 과제의 해결수단이나 실시 형태로부터 파악될 수 있는 목적이나 효과도 이에 포함됨은 물론이다.

과제의 해결 수단

[0017] 상기한 목적을 달성하기 위한 본 발명의 특징에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치는,

[0018] 트랜스포머 기반으로 시간적 행동(temporal action)을 검출하는, 검출 장치로서,

[0019] 입력 비디오로부터 추출된 특징을 사용해 특징 피라미드 구조를 생성하여 여러 규모의 시간적 특징을 통합하고, 장기와 단기의 시간적 특징 간의 관계를 학습하여 장단기 시간적 주의(long- and short-term temporal attention; LSTA)를 연산하는 LSTA 연산부를 포함하되,

[0020] 상기 LSTA 연산부는,

[0021] 상기 입력 비디오에서 추출된 특징을 사용해 장기 메모리를 생성하는 CLM(compact long-term memory, 소형 장기 메모리) 모듈;

[0022] 입력 특징을 나누어 단기 시간적 특징을 생성하고, 상기 CLM 모듈에서 구성된 장기 메모리와 상기 단기 시간적 특징을 사용해 장기와 단기의 시간적 특징 간의 관계를 학습하고 단기 시간적 특징을 강화하는 STA(short-term temporal attention) 블록 모듈; 및

[0023] 장기 시간적 특징 전체에서 필요한 부분을 동적으로 학습하고 강화하는 LTA(long-term temporal attention) 블록 모듈을 포함하며,

[0024] 상기 STA 블록 모듈이 복수의 레벨을 이루고 상위 레벨의 STA 블록 모듈은 하위 레벨의 STA 블록 모듈의 출력 특징을 입력 특징으로 사용하며, 각 레벨의 상기 STA 블록 모듈의 입력 및 출력 특징이 상기 특징 피라미드 구조를 이루는 것을 그 구성상의 특징으로 한다.

[0026] 바람직하게는,

[0027] 백본 네트워크를 통해 입력 비디오로부터 특징을 추출하고, 특징 임베딩을 통해 추출된 특징을 토큰화하는 특징 추출부를 더 포함할 수 있다.

[0029] 바람직하게는, 상기 CLM 모듈은,

- [0030] 학습 가능한 두 개의 빈 쿼리(Query, Q)와 트랜스포머 모델의 디코더를 사용해 압축된 상기 장기 메모리를 생성할 수 있다.
- [0032] 바람직하게는, 상기 STA 블록 모듈은,
- [0033] 미리 정의된 단기 윈도우(short-term window)를 사용해 입력 특징을 단기 메모리 특징으로 분할하고, 상기 분할된 단기 메모리 특징으로부터 쿼리를 계산하며, 상기 CLM 모듈에서 생성된 장기 메모리로부터 키(Key, K) 및 값(Value, V)을 계산하여, 상기 쿼리, 키 및 값을 사용해 교차 주의(cross-attention) 연산을 수행할 수 있다.
- [0035] 더욱 바람직하게는, 상기 STA 블록 모듈은,
- [0036] 상기 교차 주의 연산을 통해, 상기 단기 메모리 특징을 강화하면서 상기 장기 메모리와 단기 메모리의 관계를 학습하여, 장기 시간 의존성을 인코딩하는 단기 시간적 특징을 생성할 수 있다.
- [0038] 더더욱 바람직하게는, 상기 STA 블록 모듈은,
- [0039] 상기 분할 모듈에서 분할된 모든 상기 단기 메모리 특징에 대해 생성된 상기 단기 시간적 특징을 연결하여 출력 특징을 얻을 수 있다.
- [0041] 바람직하게는, 상기 LTA 블록 모듈은,
- [0042] 상기 입력 비디오로부터 추출된 특징에 대한 기준점(reference point) 및 오프셋(offset)을 사용해 선택된 토큰으로부터 키와 값을 계산하고, 상기 STA 블록 모듈의 출력 특징에서 생성된 쿼리 및 상기 계산된 키와 값을 사용해 교차 주의 연산을 적용해 장기 시간적 특징을 강화할 수 있다.
- [0044] 바람직하게는,
- [0045] 상기 LTA 블록 모듈의 출력으로부터 행동 레이블을 예측하는 분류 헤드와 행동 경계(action boundary)를 예측하는 회귀 헤드를 포함하여 구성되는 예측 헤드부를 더 포함할 수 있다.
- [0047] 상기한 목적을 달성하기 위한 본 발명의 특징에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 방법은,
- [0048] 트랜스포머 기반으로 시간적 행동을 검출하는, 컴퓨터에서 각 단계가 수행되는 검출 방법으로서,
- [0049] 입력 비디오로부터 추출된 특징을 사용해 특징 피라미드 구조를 생성하여 여러 규모의 시간적 특징을 통합하고, 장기와 단기의 시간적 특징 간의 관계를 학습하여 장단기 시간적 주의(long- and short-term temporal attention; LSTA)를 연산하는 LSTA 연산부를 포함하는 LSTA 모델을 학습하는 학습 단계; 및
- [0050] 상기 학습된 LSTA 모델을 이용해 입력 비디오로부터 시간적 행동을 예측하는 예측 단계를 포함하되,
- [0051] 상기 LSTA 연산부는,
- [0052] 상기 입력 비디오에서 추출된 특징을 사용해 장기 메모리를 생성하는 CLM(compact long-term memory, 소형 장기 메모리) 모듈;
- [0053] 입력 특징을 나누어 단기 시간적 특징을 생성하고, 상기 CLM 모듈에서 구성된 장기 메모리와 상기 단기 시간적 특징을 사용해 장기와 단기의 시간적 특징 간의 관계를 학습하고 단기 시간적 특징을 강화하는 STA 블록 모듈; 및
- [0054] 장기 시간적 특징 전체에서 필요한 부분을 동적으로 학습하고 강화하는 LTA(long-term temporal attention) 블록 모듈을 포함하며,

- [0055] 상기 STA 블록 모듈이 복수의 레벨을 이루고 상위 레벨의 STA 블록 모듈은 하위 레벨의 STA 블록 모듈의 출력 특징을 입력 특징으로 사용하며, 각 레벨의 상기 STA 블록 모듈의 입력 및 출력 특징이 상기 특징 피라미드 구조를 이루는 것을 그 구성상의 특징으로 한다.
- [0057] 바람직하게는, 상기 LSTA 모델은,
- [0058] 백본 네트워크를 통해 입력 비디오로부터 특징을 추출하고, 특징 임베딩을 통해 추출된 특징을 토큰화하는 특징 추출부를 더 포함할 수 있다.
- [0060] 바람직하게는, 상기 CLM 모듈은,
- [0061] 학습 가능한 두 개의 빈 쿼리(Query, Q)와 트랜스포머 모델의 디코더를 사용해 압축된 상기 장기 메모리를 생성할 수 있다.
- [0063] 바람직하게는, 상기 STA 블록 모듈은,
- [0064] 미리 정의된 단기 윈도우(short-term window)를 사용해 입력 특징을 단기 메모리 특징으로 분할하고, 상기 분할된 단기 메모리 특징으로부터 쿼리를 계산하며, 상기 CLM 모듈에서 생성된 장기 메모리로부터 키(Key, K) 및 값(Value, V)을 계산하여, 상기 쿼리, 키 및 값을 사용해 교차 주의(cross-attention) 연산을 수행할 수 있다.
- [0066] 더욱 바람직하게는, 상기 STA 블록 모듈은,
- [0067] 상기 교차 주의 연산을 통해, 상기 단기 메모리 특징을 강화하면서 상기 장기 메모리와 단기 메모리의 관계를 학습하여, 장기 시간 의존성을 인코딩하는 단기 시간적 특징을 생성할 수 있다.
- [0069] 더더욱 바람직하게는, 상기 STA 블록 모듈은,
- [0070] 상기 분할 모듈에서 분할된 모든 상기 단기 메모리 특징에 대해 생성된 상기 단기 시간적 특징을 연결하여 출력 특징을 얻을 수 있다.
- [0072] 바람직하게는, 상기 LTA 블록 모듈은,
- [0073] 상기 입력 비디오로부터 추출된 특징에 대한 기준점(reference point) 및 오프셋(offset)을 사용해 선택된 토큰으로부터 키와 값을 계산하고, 상기 STA 블록 모듈의 출력 특징에서 생성된 쿼리 및 상기 계산된 키와 값을 사용해 교차 주의 연산을 적용해 장기 시간적 특징을 강화할 수 있다.
- [0075] 바람직하게는, 상기 LSTA 모델은,
- [0076] 상기 LTA 블록 모듈의 출력으로부터 행동 레이블을 예측하는 분류 헤드와 행동 경계(action boundary)를 예측하는 회귀 헤드를 포함하여 구성되는 예측 헤드부를 더 포함할 수 있다.

발명의 효과

- [0077] 본 발명에서 제안하고 있는 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치 및 방법에 따르면, 시퀀스 분석 작업에서 중요한 장기 및 단기 시간적 특징 간의 상호 작용을 활용하는 장단기 시간적 주의 기반의 새로운 TAD 모델을 도입하되, 장기 메모리를 고려하여 단기 시간적 특징을 학습하고 강화하는 STA 블록 모듈과, 장기 시간적 특징 전체에서 필요한 부분을 동적으로 학습하고 강화하는 LTA 블록 모듈을 포함함으로써, 비디오의 장단기 의존성을 이용해 시간적 행동 검출의 성능을 향상시킬 수 있다.

[0079] 더불어, 본 발명의 다양하면서도 유익한 장점과 효과는 상술한 내용에 한정되지 않으며, 본 발명의 구체적인 실시 형태를 설명하는 과정에서 보다 쉽게 이해될 수 있을 것이다.

도면의 간단한 설명

- [0080] 도 1은 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치의 LSTA 모델 구성을 도시한 도면.
- 도 2는 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치의 LSTA 모델 구조를 도시한 도면.
- 도 3은 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 방법의 흐름을 도시한 도면.
- 도 4는 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치 및 방법에서, CLM 모듈의 메커니즘을 도시한 도면.
- 도 5는 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치 및 방법에서, STA 블록 모듈의 메커니즘을 도시한 도면.
- 도 6은 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치 및 방법에서, LTA 블록 모듈의 메커니즘을 도시한 도면.
- 도 7은 THUMOS14 데이터셋에서 IoU 임계값의 변화에 따른 mAP(mean average precision) 값을 비교하여 표시한 도면.
- 도 8은 ActivityNet-1.3 데이터셋에서 IoU 임계값의 변화에 따른 mAP 값을 비교하여 표시한 도면.
- 도 9는 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치 및 방법에서, 각 모듈의 효과를 나타낸 도면.
- 도 10은 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치 및 방법에서, STA 블록 모듈에서 사용하는 단기 윈도우 크기에 따른 효과를 나타낸 도면.
- 도 11은 THUMOS14 데이터셋에 대한 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치 및 방법의 시각화 결과를 도시한 도면.

발명을 실시하기 위한 구체적인 내용

- [0081] 아래에서는 첨부한 도면을 참조하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본 발명의 실시예를 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며, 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.
- [0083] 명세서 전체에서, 어떤 부분이 다른 부분과 "연결"되어 있다고 할 때, 이는 "직접적으로 연결"되어 있는 경우뿐 아니라, 그 중간에 다른 소자를 사이에 두고 "간접적으로 연결"되어 있는 경우도 포함한다. 또한 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미하며, 하나 또는 그 이상의 다른 특징이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.
- [0085] 이하의 실시예는 본 발명의 이해를 돕기 위한 상세한 설명이며, 본 발명의 권리 범위를 제한하는 것이 아니다. 따라서 본 발명과 동일한 기능을 수행하는 동일 범위의 발명 역시 본 발명의 권리 범위에 속할 것이다.

- [0087] 또한, 본 발명의 각 실시예에 포함된 각 구성, 과정, 공정 또는 방법 등은 기술적으로 상호간 모순되지 않는 범위 내에서 공유될 수 있다.
- [0089] 또한, 본 발명에 있어서 단말, 장치 또는 디바이스가 수행하는 것으로 기술된 동작이나 기능 중 일부는 해당 단말, 장치 또는 디바이스와 연결된 서버에서 대신 수행될 수 있다. 마찬가지로, 서버가 수행하는 것으로 기술된 동작이나 기능 중 일부도 해당 서버와 연결된 단말, 장치 또는 디바이스에서 수행될 수도 있다.
- [0091] 특히, 본 발명의 각 실시예에 따른 시스템을 실행시키기 위한 수단으로는 애플리케이션(Application), 또는 웹 서버일 수 있으며, 이 애플리케이션, 또는 웹 서버를 기록한 기록매체를 읽을 수 있는 수단인 단말로는, 일반적인 데스크톱이나 노트북 등의 일반 PC뿐만 아니라, 스마트 폰, 태블릿 PC, 등의 모바일 단말기를 포함할 수 있다.
- [0093] 이하, 첨부된 도면을 참고하여 본 발명의 실시예들을 상세히 설명하도록 한다.
- [0095] 도 1은 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치의 LSTA 모델(10) 구성을 도시한 도면이고, 도 2는 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치의 LSTA 모델(10) 구조를 도시한 도면이다. 도 1 및 도 2에 도시된 바와 같이, 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치는, 입력 비디오로부터 추출된 특징을 사용해 특징 피라미드 구조를 생성하여 여러 규모의 시간적 특징을 통합하고, 장기와 단기의 시간적 특징 간의 관계를 학습하여 장단기 시간적 주의(LSTA)를 연산하는 LSTA 연산부(100)를 포함하여 구성될 수 있으며, 특징 추출부(200) 및 예측 헤드부(300)를 더 포함하여 구성될 수 있다. 보다 구체적으로, 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치는, LSTA 연산부(100), 특징 추출부(200) 및 예측 헤드부(300)를 포함하는 LSTA 모델(10)을 통해 입력 비디오로부터 시간적 행동을 예측할 수 있다.
- [0097] 여기서, LSTA 연산부(100)는, 입력 비디오에서 추출된 특징을 사용해 장기 메모리를 생성하는 CLM 모듈(110); 입력 특징을 나누어 단기 시간적 특징을 생성하고, CLM 모듈(110)에서 구성된 장기 메모리와 단기 시간적 특징을 사용해 장기와 단기의 시간적 특징 간의 관계를 학습하고 단기 시간적 특징을 강화하는 STA 블록 모듈(120); 및 장기 시간적 특징을 학습하고 강화하는 LTA 블록 모듈(130)을 포함하며, STA 블록 모듈(120)이 복수의 레벨을 이루고 상위 레벨의 STA 블록 모듈(120)은 하위 레벨의 STA 블록 모듈(120)의 출력 특징을 입력 특징으로 사용하며, 각 레벨의 STA 블록 모듈(120)의 입력 및 출력 특징이 특징 피라미드 구조를 이룰 수 있다.
- [0099] 즉, 도 2에 도시된 바와 같이, LSTA 모델(10)은 행동을 예측하고 행동 경계를 추정하기 위해 장기 및 단기 시간적 특징을 학습할 수 있다. 먼저, 특징 추출부(200)가 비디오 클립에서 특징 세트를 추출할 수 있다. 추출된 특징은 단기 특징을 향상시키기 위해 STA 블록 모듈(120)에 의해 생성된 L 레벨의 특징 피라미드를 사용하여 다양한 시간 척도를 고려할 수 있다. 그런 다음 특징 피라미드는 LTA 블록 모듈(130)을 통해 장기 특징이 강화되고 최종 출력이 생성될 수 있다. 마지막으로 공유된 예측 헤드부(300)는 분류 헤드(310) 및 회귀 헤드(320)를 포함하여, 모든 단계에서 작업 후보를 생성하는 데 사용될 수 있다.
- [0101] 도 3은 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 방법의 흐름을 도시한 도면이다. 도 3에 도시된 바와 같이, 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 방법은, 트랜스포머 기반으로 시간적 행동을 검출하는 컴퓨터에서 각 단계가 수행되는 검출 방법으로서, 입력 비디오로부터 추출된 특징을 사용해 특징 피라미드 구조를 생성하여 여러 규모의 시간적 특징을 통합하고, 장기와 단기의 시간적 특징 간의 관계를 학습하여 장단기 시간적 주의(LSTA)를 연산하는 LSTA 연산부(100)를 포함하는 LSTA 모델(10)을 학습하는 학습 단계(S100); 및 학습된 LSTA 모델(10)을 이용해 입력 비

디오로부터 시간적 행동을 예측하는 예측 단계(S200)를 포함하여 구현될 수 있다.

[0103] 본 발명에서는, 장기 및 단기 시간적 특징을 활용하기 위해 장기 및 단기 시간적 주의(long- and short-term temporal attention; LSTA)라고 불리는 새로운 메커니즘을 제안한다. 시퀀스 분석 작업에서 시간적 장기 및 단기 특징 간의 상호 작용의 중요성을 고려하여 LSTA 모델(10)을 설계했다. 이러한 메커니즘은 각 시간 단계에서 장기 및 단기 표현을 모두 통합하고 상호 작용을 수행하여 시간적 특징을 향상시켜 TAD 모델이 더 나은 성능을 달성할 수 있도록 한다. 본 발명의 LSTA 모델(10)은 기존 모델의 접근 방식을 따라 3D CNN 백본을 활용하여 다듬어지지 않은 비디오로부터 비디오 클립 특징을 구성할 수 있다. 그 후 비디오 클립 특징을 다운샘플링 하여 특징 피라미드 구조를 생성하여, 다중-시간 규모(multi-temporal scale) 특징의 계산을 용이하게 할 수 있다.

[0105] 장기 및 단기 시간적 특징의 학습을 더욱 향상시키기 위해 STA 블록 모듈(120)과 LTA 블록 모듈(130)을 이 프로세스에 통합할 수 있다. STA 블록 모듈(120)은 먼저 백본 네트워크에 의해 추출된 특징을 임의의 크기의 블록으로 나누어 단기 특징을 생성할 수 있다. STA 블록 모듈(120)에서는, 생성된 단기 특징과 CLM 모듈(110)에서 생성된 추가적인 소형 장기 메모리(CLM)를 활용하여 장기 및 단기 특징 간의 관계를 효과적으로 학습할 수 있다. 이 프로세스를 통해 STA 블록 모듈(120)은 장기 및 단기 시간적 특징 간의 관계를 학습하고 단기 특징을 강화할 수 있다. 이어서 LTA 블록 모듈(130)은 장기 관계에 중요한 특징을 효율적으로 학습하고 강화할 수 있다. 이를 통해 LTA 블록 모듈(130)은 출력의 중요한 특징을 다시 학습할 수 있다.

[0107] 본 발명은, 각 특징을 학습하고 향상시키기 위해 시퀀스 분석 작업에서 중요한 장기 및 단기 시간적 특징 간의 상호 작용을 활용하는 LSTA 기반의 새로운 TAD 모델을 도입하되, 장기 메모리를 고려하여 단기 시간적 특징을 학습하고 강화하는 STA 블록 모듈(120)과, 장기 시간적 특징 전체에서 필요한 부분을 동적으로 학습하고 강화하는 LTA 블록 모듈(130)을 포함함으로써, 비디오의 장단기 의존성을 이용해 시간적 행동 검출의 성능을 향상시킬 수 있다.

[0109] 이하에서는, LSTA 모델(10)의 세부적인 구성에 대해 상세히 설명하도록 한다.

[0111] 특징 추출부(200)는, 백본 네트워크를 통해 입력 비디오로부터 특징을 추출하고, 특징 임베딩을 통해 추출된 특징을 토근화할 수 있다. 도 2에 도시된 바와 같이, LSTA 모델(10)은 미리 학습된 백본 네트워크를 사용하여 입력 비디오에 대한 특징 $F \in \mathbb{R}^{T \times C}$ 을 추출할 수 있다. 여기서, T는 추출된 특징의 시간적 차원을 나타내고, C는 채널 차원을 나타낸다. 특징 임베딩을 통해, F는 $Z \in \mathbb{R}^{T \times D}$ 로 토근화되어 특징 피라미드 구조를 만드는 데 사용될 수 있다.

[0113] LSTA 연산부(100)는, 입력 비디오로부터 추출된 특징을 사용해 특징 피라미드 구조를 생성하여 여러 규모의 시간적 특징을 통합하고, 장기와 단기의 시간적 특징 간의 관계를 학습하여 장단기 시간적 주의를 연산할 수 있다. 보다 구체적으로, LSTA 연산부(100)는, 특징 피라미드 구조를 사용하여 여러 규모의 시간적 특징을 통합하고 단기 시간적 특징을 학습하고 강화하기 위해 새로운 STA 메커니즘을 사용할 수 있다.

[0115] 도 1 및 도 2에 도시된 바와 같이, LSTA 연산부(100)는 CLM 모듈(110), STA 블록 모듈(120), LTA 블록 모듈(130)을 포함하여 구성될 수 있다. 여기서, STA 블록 모듈(120)은 소형 장기 메모리(CLM)를 사용하여 구성할 수 있는 장기 메모리 E를 사용하여 단기 시간적 특징을 학습하고 강화할 수 있다. 그런 다음 이러한 특징은 LTA 블록 모듈(130)에 의해 처리되어 장기 시간적 특징을 동적으로 학습하고 강화하여 출력 $\hat{Z}$ 을 생성할 수 있다. 이하에서는, LSTA 연산부(100)의 세부적인 구성에 대해 상세히 설명하도록 한다.

[0117] CLM 모듈(110)은, 입력 비디오에서 추출된 특징을 사용해 장기 메모리를 생성할 수 있다. 보다 구체적으로, CLM 모듈(110)은, 학습 가능한 두 개의 빈 쿼리(Query, Q)와 트랜스포머 모델의 디코더를 사용해 압축된 장기 메모리를 생성할 수 있다.

[0119] 도 4는 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치 및 방법에서, CLM 모듈(110)의 메커니즘을 도시한 도면이다. 도 4에 도시된 바와 같이, 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치 및 방법의 CLM 모듈(110)은, 단기 메모리와 장기 메모리 간의 관계를 학습하기 위해 특징 임베딩 F를 사용하여 장기 메모리 E를 생성할 수 있다. F를 직접 사용하면 T^2 의 시간 복잡성이 발생하여 계산 시간이 크게 증가한다. 따라서 LSTR(Long Short-term TRansformer) (비특허문헌 0010)에서 영감을 받아 전체 시간 특성을 나타내는 학습 가능한 빈 쿼리 Q_1 및 $Q_2(Q_2 < Q_1 < T)$ 를 사용하여 압축된 장기 메모리 E를 구성할 수 있다.

수학식 1

[0121]
$$E = \text{Decoder}(Q_2, \text{Decoder}(Q_1, F))$$

[0123] 수학식 1에서 Decoder는 LSTR을 따르는 트랜스포머 모델 (비특허문헌 0009)의 디코더를 의미한다.

[0125] STA 블록 모듈(120)은, 입력 특징을 나누어 단기 시간적 특징을 생성하고, CLM 모듈(110)에서 구성된 장기 메모리와 생성된 단기 시간적 특징을 사용해 장기와 단기의 시간적 특징 간의 관계를 학습하고 단기 시간적 특징을 강화할 수 있다. 보다 구체적으로, STA 블록 모듈(120)은, 미리 정의된 단기 윈도우를 사용해 입력 특징을 단기 메모리 특징으로 분할하고, 분할된 단기 메모리 특징으로부터 쿼리를 계산하며, CLM 모듈(110)에서 생성된 장기 메모리로부터 키(Key, K) 및 값(Value, V)을 계산하여, 쿼리, 키 및 값을 사용해 교차 주의 연산을 수행할 수 있다. STA 블록 모듈(120)은, 교차 주의 연산을 통해, 단기 메모리 특징을 강화하면서 장기 메모리와 단기 메모리의 관계를 학습하여, 장기 시간 의존성을 인코딩하는 단기 시간적 특징을 생성할 수 있다. STA 블록 모듈(120)은, 분할 모듈에서 분할된 모든 단기 메모리 특징에 대해 생성된 단기 시간적 특징을 연결하여 출력 특징을 얻을 수 있다.

[0127] 도 5는 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치 및 방법에서, STA 블록 모듈(120)의 메커니즘을 도시한 도면이다. 도 5에 도시된 바와 같이, 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치 및 방법의 STA 블록 모듈(120)은, 단기 시간적 특징을 학습하고 강화하기 위해 단기 시간적 주의(Short-term attention; STA)를 학습할 수 있다. 임의의 크기의 시간 윈도우를 사용하는 것과 같은 단순한 접근법을 학습에 사용할 수도 있지만, 본 발명에서는 단기 및 장기 시간 정보 간의 상호 작용을 학습하기 위해 장기 메모리 쿼리를 적용할 수 있다.

[0129] 먼저, 입력 Z는 크기 S의 미리 정의된 단기 윈도우를 사용하여 단기 메모리 특징 $Z_{\text{short}} = \{Z_i\}_{i=1}^I$ 로 분할되며, 여기서 $I = T/S$ 는 STA 연산에서 쿼리 Q로 사용될 수 있다. STA에서는 교차 주의 연산을 통해 단기 메모리 특징이 강화되는 동안, 단기 메모리와 장기 메모리의 관계를 학습할 수 있다. 이때, 장기 메모리 E는 키(K)와 값(V) 쌍으로 사용되고, 분할된 단기 메모리 특징 Z_{short} 는 쿼리 Q로 사용될 수 있다. 이 연산은 다음 수학식 2 및 3과 같이 공식화될 수 있다.

수학식 2

$$Q = Z_{short} \cdot W_q, K = E \cdot W_k, V = E \cdot W_v$$

수학식 3

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_h}}\right) \cdot V$$

여기서, K, V는 학습 가능한 매개 변수 W_k 및 W_v 를 갖는 선형 투영 층(linear projection layer)을 사용하여 E로부터 계산될 수 있다. 또한, Z_{short} 는 학습 가능한 매개 변수 W_q 를 갖는 선형 투영 층을 통해 적용되는 Q로 사용될 수 있다. 또한 Q, K 및 V는 교차 주의 연산을 수행하여 장기 시간 의존성을 인코딩하는 단기 시간적 특징이 생성될 수 있다. 이 프로세스를 I번 반복한 후 모든 출력된 단기 시간적 특징을 연결하여 최종 출력 특징 Z를 얻을 수 있다.

LTA 블록 모듈(130)은, 장기 시간적 특징 전체에서 필요한 부분을 동적으로 학습하고 강화할 수 있다. 보다 구체적으로, LTA 블록 모듈(130)은, 입력 비디오로부터 추출된 특징에 대한 기준점(reference point) 및 오프셋(offset)을 사용해 선택된 토큰으로부터 키와 값을 계산하고, STA 블록 모듈(120)의 출력 특징에서 생성된 쿼리 및 계산된 키와 값을 사용해 교차 주의 연산을 적용해 장기 시간적 특징을 강화할 수 있다.

도 6은 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치 및 방법에서, LTA 블록 모듈(130)의 메커니즘을 도시한 도면이다. 도 6에 도시된 바와 같이, 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치 및 방법의 LTA 블록 모듈(130)은, 특징 임베딩 F에 대한 장기 시간적 특징을 강화하기 위해 STA 블록 모듈(120)에서 강화된 단기 시간적 특징을 활용할 수 있다. 앞서 언급한 바와 같이 STA 블록 모듈(120)에서 F에 대해 전체 시간 T에 걸쳐 주의를 적용하는 것은 계산 비용이 많이 든다. 따라서 본 발명에서는 F에 영향을 미치는 필수 토큰을 유연하게 선택할 수 있는 변형 주의(deformable attention)를 가진 LTA를 제안한다.

먼저, 전체 특징 F에 대한 기준점 p를 생성하고 컨볼루션 레이어에서 학습된 오프셋 θ 를 사용하여 필요한 토큰 좌표값을 생성하여 K와 V를 계산할 수 있다. 그런 다음 STA 블록 모듈(120)의 출력인 Z에서 생성된 Q와 앞서 계산된 K, V로 교차 주의를 적용하여 장기 시간적 특징을 강화할 수 있다. LTA는 다음 수학식 4 내지 7을 통해 정의할 수 있다:

수학식 4

$$\tilde{Z} = \phi(F; p + \Delta p), \Delta p = \theta(F)$$

수학식 5

[0145]
$$Q = Z \cdot \hat{W}_q, K = \tilde{Z} \cdot \hat{W}_k, V = \tilde{Z} \cdot \hat{W}_v$$

수학식 6

[0147]
$$\text{LTA}(Z, F) = \text{Attention}(Q, K, V)$$

수학식 7

[0149]
$$\hat{Z} = \text{LTA}(Z, F)$$

[0151] 여기서, ϕ 는 기준점 p 와 오프셋 θ 를 이용하여 토큰 좌표값을 학습하고 생성하는 함수이며, $\hat{W}_q, \hat{W}_k, \hat{W}_v$ 는 Q, K, V 를 생성하기 위해 사용되는 학습 가능한 파라미터이다. LTA 블록 모듈(130)의 출력, 즉 $\hat{Z}$ 로부터 이하에서 상세히 설명할 예측 헤드부(300)가 행동 클래스 레이블과 행동 경계를 예측하기 위해 두 개의 헤드를 사용할 수 있다.

[0153] 예측 헤드부(300)는, LTA 블록 모듈(130)의 출력으로부터 행동 레이블을 예측하는 분류 헤드(310)와 행동 경계(action boundary)를 예측하는 회귀 헤드(320)를 포함할 수 있다. 여기서, 분류 헤드(310)는 $\hat{Z}$ 를 사용하여 행동 레이블을 예측하고, 회귀 헤드(320)는 $\hat{Z}$ 를 사용하여 행동 경계를 예측할 수 있다. 최종 결과는 $Y=(y_t, y_t^s, y_t^e)$ 로 표현되며, 여기서 y_t 는 시간 t 에서의 행동 레이블의 예측 확률이고 y_t^s 와 y_t^e 는 시간 t 에서 행동의 시작과 종료까지의 예측 거리를 각각 나타낸다.

[0155] 분류 헤드(310)는, 세 개의 컨볼루션, 두 개의 계층 노름(norm), ReLU 활성화 함수로 구성될 수 있으며, 각 프레임에 대한 C 행동 클래스를 예측할 수 있다. 이진 분류를 위해 초점 손실(focal loss) L_{class} (비특허문헌 0002)를 이용해 이 헤드를 학습시킬 수 있는데, 이는 데이터 불균형에 강건하다.

[0157] 회귀 헤드(320)는, 분류 헤드(310)와 유사한 구조를 가지며 각 프레임에서 행동 경계까지의 거리를 예측할 수 있다. 회귀 헤드(320)는 DIOU 손실 (비특허문헌 0005)을 사용하는 L_{reg} 로 정규화될 수 있다. 총 학습 손실 L_{total} 에 대한 공식은 다음과 같이 정의될 수 있다:

수학식 8

$$\mathcal{L}_{total} = (\mathcal{L}_{class} + \mathbb{C}_t \mathcal{L}_{reg}) / T_+$$

여기서, T_+ 는 양의 샘플 수를 나타내고, $\mathbb{C}_t$ 는 시간 스텝 t 에서 전경과 배경을 구분하는 지시 함수이다.

실험

여기서는 실험에 사용된 데이터세트와 제안된 모델의 하이퍼 파라미터 등 구현 세부 사항을 소개한다. 제안된 LSTA 모델(10)의 경쟁력 있는 성능을 입증하기 위해 기존 SoTA 모델을 사용하여 비교 실험을 수행했다. 또한 LSTA 모델(10)의 각 모듈을 분석하기 위해 절제 연구를 수행했다.

(1) 실험 설정

데이터세트

본 발명에서는 THUMOS14 데이터세트 및 ActivityNet-1.3 데이터세트를 사용하여 실험을 수행했다. THUMOS14 데이터세트는 20개의 행동 클래스와 413개의 비디오 시퀀스로 구성된다. 데이터세트는 유효성 검사 세트를 사용하여 200개의 비디오 시퀀스와 테스트 세트를 사용하여 평가되었다. ActivityNet-1.3 데이터세트는 20,000개의 비디오 시퀀스와 200개의 행동 클래스를 포함하는 대규모 데이터세트이다. 데이터세트는 학습용 10,024개, 유효성 검사용 4,926개, 테스트용 5,044개의 비디오로 구성된다.

구현 세부 정보

이 실험에서는 PyTorch 소프트웨어를 사용하여 제안된 LSTA 모델(10)을 구현했다. ActionFormer (비특허문헌 0012)와 유사하게, 본 실험에서는 사전 학습된 two-stream I3D 및 R(2+1)D에서 사전 추출된 특징을 사용했다. THUMOS14 데이터세트 실험의 경우 AdamW를 옵티마이저로 선택하고 배치 크기 8, 학습률 0.0001, weight decay 0.05로 30개의 에포크에 대해 모델을 학습했다. ActivityNet-1.3 데이터세트 실험의 경우 배치 크기 32, 학습률 0.001, weight decay 0.05, 10개의 에포크에 대해 학습했다. Soft-NMS를 후처리에 사용했으며, 실험은 단일 GeForce RTX 3090 GPU가 있는 시스템을 사용하여 수행되었다.

평가 메트릭

본 실험에서는 IoU(intersection over union) 임계값과 mAP(mean average precision)를 사용하여 성능을 평가했다. THUMOS14 데이터세트의 경우 IoU 임계값을 [0.3:0.1:0.7]로 보고했다. ActivityNet-1.3 데이터세트의 경우 IoU 임계값 [0.5,0.75,0.95]에서의 결과를 보고하고, 평균 mAP를 [0.5:0.05:0.95]에서 계산했다.

(2) 최첨단 방법과의 비교

ThUMOS14 데이터세트

도 7은 THUMOS14 데이터셋에서 IoU 임계값의 변화에 따른 mAP 값을 비교하여 표시한 도면이다. 표에서 “[]”안의 숫자는 선행기술문헌 번호를 나타낸다. 여기서, THUMOS14 데이터세트에 대한 실험을 위해 백본 네트워크로 I3D를 사용했다. 도 7에 나타난 바와 같이 제안된 LSTA 모델(10)(“LS-TAD”)은 기존 SoTA 모델과 비교하여 경쟁력 있는 성능을 보여줄 수 있다. 특히 LSTA 모델(10)은 평균 mAP가 68.2%로 차이가 나서 두 번째로 높은 성능을 보였고, 트랜스포머 기반 TAD의 기준 모델인 ActionFormer(비특허문헌 0012)와 비교할 때 1.4% 성능이 더 좋았다. 제안된 모델은 최고 성능의 TriDet 모델보다 1.1% 낮은 성능을 보였지만 큰 차이는 없었

며, IoU 임계값을 0.6 또는 0.7로 설정했을 때도 비슷한 성능이 관찰되었다.

[0180] ActivityNet-1.3 데이터셋

[0181] 도 8은 ActivityNet-1.3 데이터셋에서 IoU 임계값의 변화에 따른 mAP 값을 비교하여 표시한 도면이다. ActivityNet-1.3 데이터셋에서 수행된 실험의 경우 R(2+1)D를 백본 네트워크로 사용했다. 전체적으로 본 발명의 LSTA 모델(10)은 기존 SoTA 모델과 비교하여 경쟁력 있는 성능을 달성했다. LSTA 모델(10)은 ActivityNet-1.3에서 TriDet보다 평균 mAP가 0.3% 더 높았다. 이 결과는 THUMOS14와 ActivityNet-1.3 사이의 데이터셋 특성의 차이 때문으로 생각된다. 장기 및 단기 시간적 특징을 학습할 때는 더 복잡한 시간 패턴을 학습해야 하므로, 데이터가 많을수록 모델은 더 정확한 시간적 특징을 학습할 수 있고, 작은 데이터셋의 경우 모델이 일반화하기가 어려워 성능이 저하된다. 따라서 실험 결과를 통해 볼 수 있듯이 THUMOS14 데이터셋보다 ActivityNet-1.3 데이터셋에서 수행된 실험은 더 나은 성능을 보일 것이다. 학습을 위해 SlowFast를 백본 네트워크로 사용하는 2단계 모델인 TCANet(비특허문헌 0004)이 가장 좋은 성능을 보였다. 더 나은 백본 네트워크를 사용할 때 성능 향상을 기대할 수 있으므로, 본 발명의 LSTA 모델(10)에 SlowFast를 백본으로 적용하였을 때 더 나은 성능을 달성할 것으로 기대된다.

[0183] (3) 절제 실험(Ablation study)

[0184] 여기서는, 여러 절제 실험의 결과를 바탕으로 LSTA 모델(10)의 성능을 검증한다.

[0186] 각 모듈의 효과

[0187] 도 9는 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치 및 방법에서, 각 모듈의 효과를 나타낸 도면으로, CLM 모듈(110), STA 블록 모듈(120), LSTA 블록 모듈(120) 각각이 전체 성능에 어떠한 영향을 미치는지를 나타낸 것이다. 도 9에서 확인할 수 있는 바와 같이, 제안한 모든 모듈을 함께 사용하였을 때 ActivityNet-1.3에서 가장 좋은 평균 mAP 성능인 37.1%를 달성하였다. 구체적으로는 CLM 모듈(110)을 적용하였을 때 성능이 다소 향상되었으며, STA 블록 모듈(120)과 LTA 블록 모듈(130)을 별도로 사용하였을 때보다 두 모듈을 함께 사용하였을 때 성능이 더 좋아졌다. 따라서 STA와 LTA를 유연하게 결합하여 장기적인 시간적 특징을 효과적으로 학습할 수 있다고 결론지을 수 있다.

[0189] 단기 창 크기 수의 효과

[0190] 도 10은 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치 및 방법에서, STA 블록 모듈(120)에서 사용하는 단기 윈도우 크기에 따른 효과를 나타낸 도면이다. 도 10에서 확인할 수 있는 바와 같이, 실험 결과는 단기 윈도우의 크기가 12일 때 LSTA 모델(10)의 성능이 가장 우수하며, 윈도우 크기가 12에서 벗어날수록 성능이 점차 감소하는 것으로 나타났다. 따라서 본 발명에서는 고정된 윈도우 크기인 12를 이용하여 실험을 진행하였다. 비디오 시퀀스에서 동작의 길이는 데이터셋에 따라 다를 수 있기 때문에 각 데이터셋의 특성에 따라 다른 윈도우 크기가 성능에 영향을 줄 수 있다.

[0192] 시각화

[0193] 도 11은 THUMOS14 데이터셋에 대한 본 발명의 일실시예에 따른 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치 및 방법의 시각화 결과를 도시한 도면이다. 도 11에서 보라색 막대와 노란색 막대는 각각 행동 경계에 대한 실제 값과 LSTA 모델(10)의 예측 결과를 나타낸다. 실험 결과로부터 모델이 행동 경계에 대한 세밀한 예측을 생성하여 행동과 행동 경계 모두에 대한 정확한 예측을 보여줌을 알 수 있다.

[0195] 전술한 바와 같이, 본 발명에서 제안하고 있는 비디오의 장단기 의존성을 이용한 시간적 행동 검출 장치 및 방법에 따르면, 시퀀스 분석 작업에서 중요한 장기 및 단기 시간적 특징 간의 상호 작용을 활용하는 장단기 시간적 주의 기반의 새로운 TAD 모델을 도입하되, 장기 메모리를 고려하여 단기 시간적 특징을 학습하고 강화하는

STA 블록 모듈(120)과, 장기 시간적 특징 전체에서 필요한 부분을 동적으로 학습하고 강화하는 LTA 블록 모듈(130)을 포함함으로써, 비디오의 장단기 의존성을 이용해 시간적 행동 검출의 성능을 향상시킬 수 있다.

[0197] 한편, 본 발명은 다양한 통신 단말기로 구현되는 동작을 수행하기 위한 프로그램 명령을 포함하는 컴퓨터에서 판독 가능한 매체를 포함할 수 있다. 예를 들어, 컴퓨터에서 판독 가능한 매체는, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media) 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치를 포함할 수 있다.

[0199] 이와 같은 컴퓨터에서 판독 가능한 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 이때, 컴퓨터에서 판독 가능한 매체에 기록되는 프로그램 명령은 본 발명을 구현하기 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 예를 들어, 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해 실행될 수 있는 고급 언어 코드를 포함할 수 있다.

[0201] 전술한 본 발명의 설명은 예시를 위한 것이며, 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명된 각 구성요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명된 구성 요소들도 결합된 형태로 실시될 수 있다.

[0203] 본 발명의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

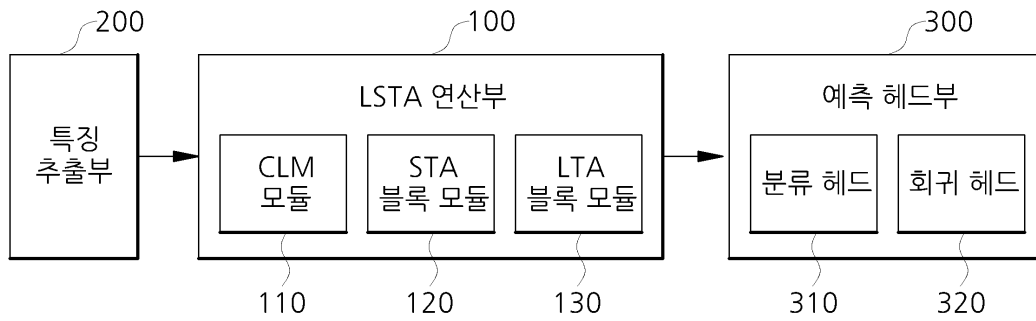
부호의 설명

[0204] 10: LSTA 모델
100: LSTA 연산부
110: CLM 모듈
120: STA 블록 모듈
130: LTA 블록 모듈
200: 특징 추출부
300: 예측 헤드부
310: 분류 헤드
320: 회귀 헤드
S100: 학습 단계
S200: 예측 단계

도면

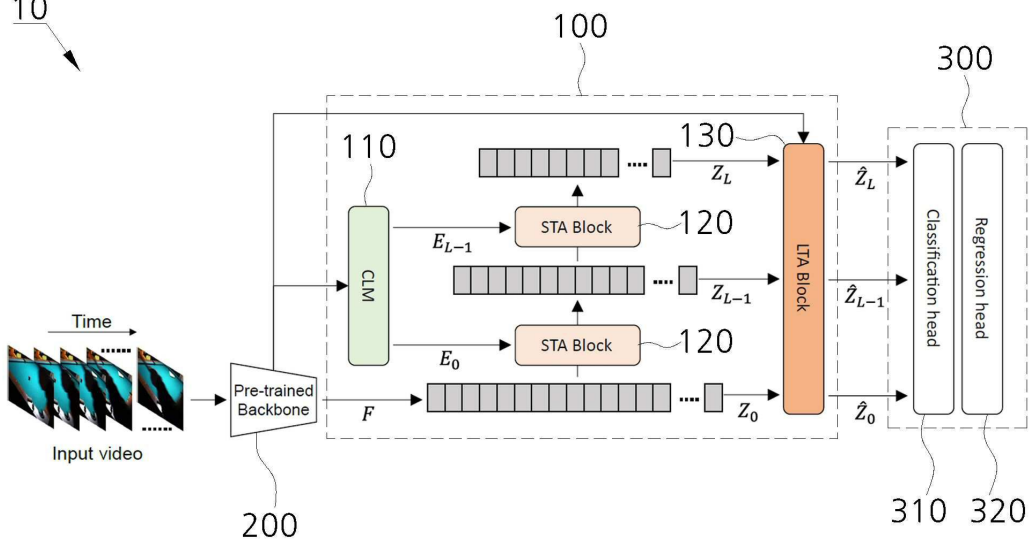
도면1

10

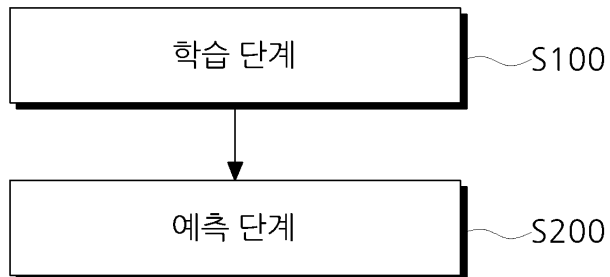


도면2

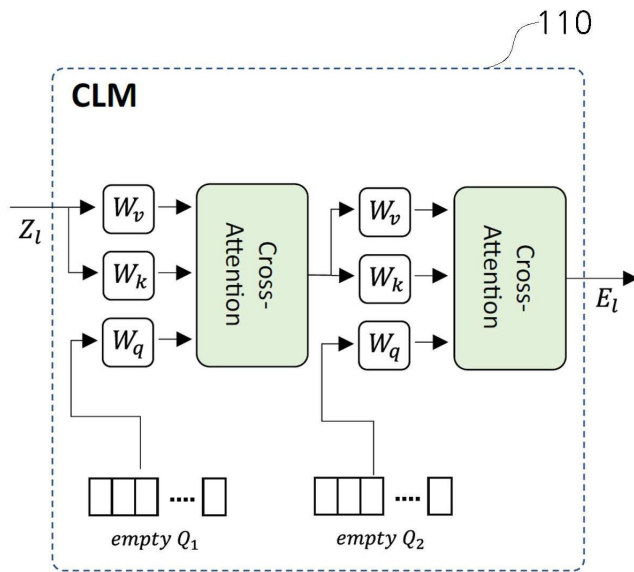
10



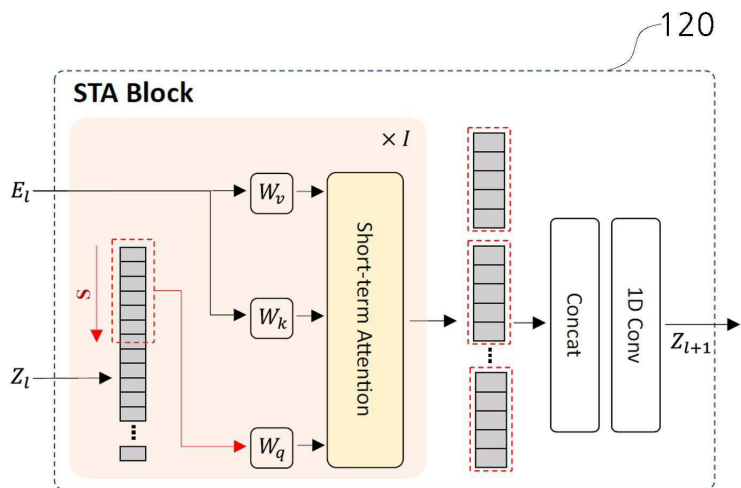
도면3



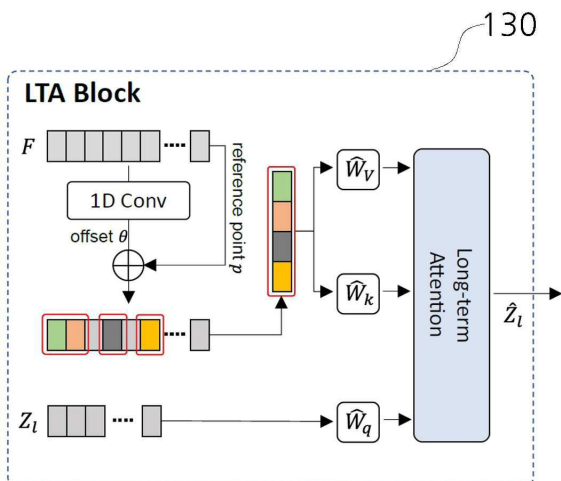
도면4



도면5



도면6



도면7

Method	0.3	0.4	0.5	0.6	0.7	Avg.
TCANet [04]	60.6	53.2	44.6	36.8	26.7	44.3
RTD-Net [08]	68.3	62.3	51.9	38.8	23.7	49.0
VSGN [13]	66.7	60.4	52.4	41.0	30.4	50.2
ContextLoc [14]	68.3	63.8	54.3	41.8	26.2	50.9
AFSD [01]	67.3	62.4	55.5	43.7	31.1	52.0
ReAct [07]	69.2	65.0	57.1	47.8	35.6	55.0
TadTR [03]	74.8	69.1	60.1	46.6	32.8	56.7
ActionFormer [12]	82.1	77.8	71.0	59.4	43.9	66.8
TriDet [06]	83.6	80.1	72.9	62.4	47.4	69.3
LS-TAD	81.1	77.9	72.1	62.3	47.4	68.2

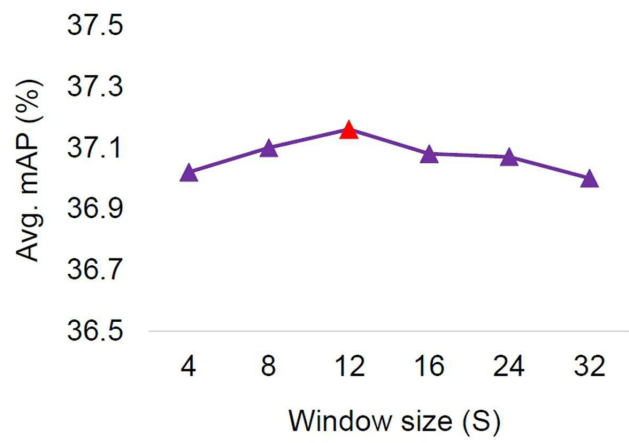
도면8

Method	Backbone	0.50	0.75	0.95	Avg.
ReAct [07]	TSN	49.6	33.0	8.6	32.6
G-TAD [11]	TSN	50.4	34.6	9.0	34.1
AFSD [01]	I3D	52.4	35.2	6.5	34.3
TadTR [03]	TSN	51.3	35.0	9.5	34.6
TadTR [03]	R(2+1)D	53.6	37.5	10.5	36.8
VSGN [13]	I3D	52.3	35.2	8.3	34.7
TCANet+BMN [04]	TSN	52.3	36.7	6.9	35.5
TCANet+BMN [04]	SlowFast	54.3	39.1	8.4	37.6
ActionFormer [12]	R(2+1)D	54.7	37.8	8.4	36.6
TriDet [06]	R(2+1)D	54.7	38.0	8.4	36.8
LS-TAD	R(2+1)D	55.2	38.1	9.3	37.1

도면9

Method	CLM	STA	LTA	Avg.
Ours			✓	36.7
Ours		✓	✓	36.9
Ours	✓	✓		36.9
Ours	✓	✓	✓	37.1

도면10



도면11

