(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0056239
(43) 공개일자 2024년04월30일

(51) 국제특허분류(Int. Cl.)
H04N 21/845 (2011.01) *G06F 40/279* (2020.01)
H04N 21/439 (2011.01) *H04N 21/44* (2024.01)
(52) CPC특허분류
H04N 21/845 (2013.01)
G06F 40/279 (2020.01)
(21) 출원번호 10-2022-0136628
(22) 출원일자 2022년10월21일
심사청구일자 없음

(71) 출원인
서강대학교산학협력단
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(72) 발명자
강석주
경기도 고양시 일산서구 후곡로 36 (일산3동, 후곡마을4단지아파트
조유빈
서울특별시 마포구 백범로 35(신수동)
(74) 대리인
유미특허법인

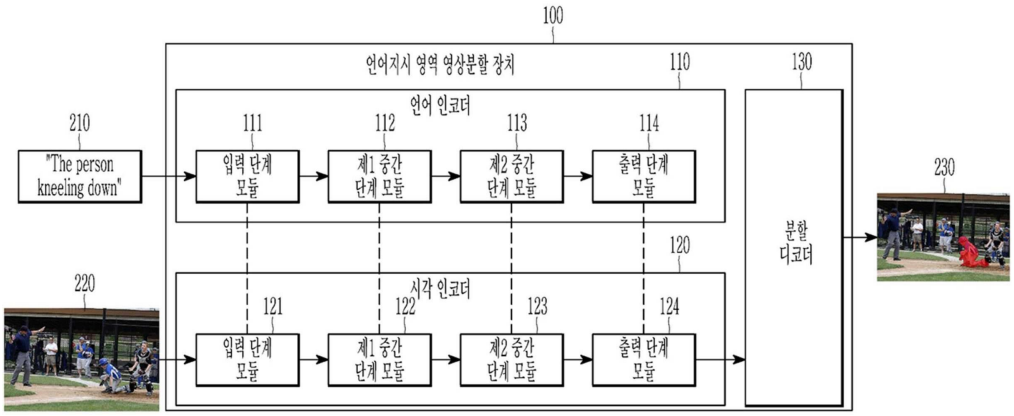
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **멀티모달 인코더 기반의 언어지시 영역 영상분할 방법 및 장치**

(57) 요 약

본 개시의 일 실시예에 따라, 컴퓨팅 장치에 의해 수행되는 언어지시 영역 영상분할 방법으로서, 언어표현 및 영상의 쌍을 입력받는 단계, 복수의 인코딩 단계들 각각에서, 선행하는 인코딩 단계로부터 획득되는 상기 영상에 관한 시각 특징을 기초로 상기 언어표현에 관한 언어 특징을 추출하는 단계, 상기 복수의 인코딩 단계들 각각에서, 선행하는 인코딩 단계로부터 획득되는 상기 언어 특징을 기초로 상기 시각 특징을 추출하는 단계, 및 상기 언어 특징 및 상기 시각 특징 중 적어도 일부를 기초로, 상기 영상에서 상기 언어표현과 대응되는 대상 영역을 분할하는 단계를 포함할 수 있다.

대 표 도



(52) CPC특허분류
H04N 21/4394 (2013.01)
H04N 21/44008 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711159124
과제번호 2020M3H4A1A02084899
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 나노미래소재원천기술개발사업
연구과제명 2축 신축성 AMLED용 TFT 및 LED 집적 기술
기 여 율 1/2
과제수행기관명 경희대학교 산학협력단
연구기간 2020.07.01~2024.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1375027429
과제번호 R2020040058
부처명 문화체육관광부
과제관리(전문)기관명 한국콘텐츠진흥원
연구사업명 문화기술연구개발
연구과제명 자율주행 차량 기반 다수 시나리오 실시간 인터랙티브 콘텐츠 및 플랫폼 기술 개발
기 여 율 1/2
과제수행기관명 한국전자기술연구원
연구기간 2020.07.01~2022.12.31

명 세 서

청구범위

청구항 1

컴퓨팅 장치에 의해 수행되는 언어지시 영역 영상분할 방법으로서,

언어표현 및 영상의 쌍을 입력받는 단계,

복수의 인코딩 단계들 각각에서, 선행하는 인코딩 단계로부터 획득되는 상기 영상에 관한 시각 특징을 기초로 상기 언어표현에 관한 언어 특징을 추출하는 단계,

상기 복수의 인코딩 단계들 각각에서, 선행하는 인코딩 단계로부터 획득되는 상기 언어 특징을 기초로 상기 시각 특징을 추출하는 단계, 및

상기 언어 특징 및 상기 시각 특징 중 적어도 일부를 기초로, 상기 영상에서 상기 언어표현과 대응되는 대상 영역을 분할하는 단계를 포함하는,

언어지시 영역 영상분할 방법.

청구항 2

제 1항에서,

상기 언어 특징을 추출하는 단계는,

상기 복수의 인코딩 단계들 각각을 수행하기 위한 입력 단계 모듈, 적어도 하나의 중간 단계 모듈, 및 출력 단계 모듈로 구성되는 언어 인코더에 의해 수행되는,

언어지시 영역 영상분할 방법.

청구항 3

제 2항에서,

상기 언어 특징을 추출하는 단계는,

상기 입력 단계 모듈에 의해, 상기 언어표현으로부터 초기 언어 특징을 추출하는 단계,

상기 초기 언어 특징을 기초로, 상기 초기 언어 특징에 상기 영상으로부터 추출되는 초기 시각 특징을 반영한 언어 특징이 추출되게끔 하는 단계, 및

후속하는 단계 모듈에 의해, 상기 초기 시각 특징을 반영한 언어 특징을 전달받는 단계를 포함하는,

언어지시 영역 영상분할 방법.

청구항 4

제 2항에서,

상기 언어특징을 추출하는 단계는,

상기 적어도 하나의 중간 단계 모듈에 의해, 상기 선행하는 인코딩 단계에서 추출된 선행 언어 특징을 전달받는 단계,

상기 선행 언어 특징에 대해 상기 선행하는 인코딩 단계에서 추출된 선행 시각 특징을 반영한 언어 특징을 추출하는 단계, 및

상기 선행 시각 특징을 반영한 언어 특징을 후속하는 인코딩 단계로 전달하는 단계를 포함하는,

언어지시 영역 영상분할 방법.

청구항 5

제 4항에서,

상기 선행 시각 특징은,

상기 선행 시각 특징에 관한 특징 맵에 대해 서로 다른 커널 사이즈를 가지는 풀링 레이어들을 적용하여, 서로 다른 시각적 관점들을 가지도록 생성되는 멀티 스케일(multi-scale) 시각적 토큰의 형태로 상기 선행 언어 특징에 반영되는,

언어지시 영역 영상분할 방법.

청구항 6

제 2항에서,

상기 언어특징을 추출하는 단계는,

상기 출력 단계 모듈에 의해, 상기 선행하는 인코딩 단계에서 추출된 선행 언어 특징을 전달받는 단계,

상기 선행 언어 특징에 대해 상기 선행하는 인코딩 단계에서 추출된 선행 시각 특징을 반영한 언어 특징을 추출하는 단계, 및

상기 선행 시각 특징을 반영한 언어 특징을 기초로, 상기 영상으로부터 상기 대상 영역을 식별하게끔 하는 단계를 포함하는,

언어지시 영역 영상분할 방법.

청구항 7

제 1항에서,

상기 시각 특징을 추출하는 단계는,

상기 복수의 인코딩 단계들 각각을 수행하기 위한 입력 단계 모듈, 적어도 하나의 중간 단계 모듈, 및 출력 단계 모듈로 구성되는 시각 인코더에 의해 수행되는,

언어지시 영역 영상분할 방법.

청구항 8

제 7항에서,

상기 시각 특징을 추출하는 단계는,

상기 입력 단계 모듈에 의해, 상기 영상으로부터 초기 시각 특징을 추출하는 단계,

상기 초기 시각 특징을 기초로, 상기 초기 시각 특징에 상기 언어표현으로부터 추출되는 초기 언어 특징을 반영하여, 상기 초기 언어 특징을 반영한 시각 특징 및 상기 초기 시각 특징을 반영한 언어 특징을 각각 추출하는 단계, 및

상기 초기 언어 특징을 반영한 시각 특징 및 상기 초기 시각 특징을 반영한 언어 특징 각각을 후속하는 인코딩 단계로 전달하는 단계를 포함하는,

언어지시 영역 영상분할 방법.

청구항 9

제 8항에서,

상기 각각 추출하는 단계는,

상기 초기 시각 특징을 쿼리(query)로 사용하고 상기 초기 언어 특징을 키(key)와 값(value)으로 사용한 멀티헤드 교차 어텐션(multi-head cross-attention)을 통해, 상기 초기 언어 특징을 반영한 시각 특징을 추출하는 단계, 및

상기 초기 언어 특징을 쿼리로 사용하고 상기 초기 시각 특징을 키와 값으로 사용한 멀티헤드 교차 어텐션을 통해, 상기 초기 시각 특징을 반영한 언어 특징을 추출하는 단계를 포함하는,

언어지시 영역 영상분할 방법.

청구항 10

제 7항에서,

상기 시각 특징을 추출하는 단계는,

상기 적어도 하나의 중간 단계 모듈에 의해, 상기 선행하는 인코딩 단계에서 추출된 선행 시각 특징을 전달받는 단계,

상기 선행 시각 특징에 대해 상기 선행하는 인코딩 단계에서 추출된 선행 언어 특징을 반영한 시각 특징을 추출하는 단계, 및

상기 선행 언어 특징을 반영한 시각 특징을 후속하는 인코딩 단계로 전달하는 단계를 포함하는,

언어지시 영역 영상분할 방법.

청구항 11

제 10항에서,

상기 후속하는 인코딩 단계로 전달하는 단계는,

상기 선행 언어 특징을 반영한 시각 특징을 다운 샘플링 하는 단계,

상기 다운 샘플링 된 시각 특징을 쿼리로 사용하고 상기 선행 언어 특징을 키와 값으로 사용한 멀티헤드 교차 어텐션을 통해, 상기 대상 영역에 대한 문맥(context) 정보가 보강된 시각 특징을 추출하는 단계, 및

상기 문맥 정보가 보강된 시각 특징을 상기 후속하는 인코딩 단계로 전달하는 단계를 포함하는,

언어지시 영역 영상분할 방법.

청구항 12

제 7항에서,

상기 시각특징을 추출하는 단계는,

상기 출력 단계 모듈에 의해, 상기 선행하는 인코딩 단계에서 추출된 선행 시각 특징을 전달받는 단계,

상기 선행 시각 특징에 대해 상기 선행하는 인코딩 단계에서 추출된 선행 언어 특징을 반영한 시각 특징을 추출하는 단계, 및

상기 선행 언어 특징을 반영한 시각 특징을 기초로, 상기 영상으로부터 상기 대상 영역을 식별하는 단계를 포함하는,

언어지시 영역 영상분할 방법.

청구항 13

언어표현으로부터 복수의 인코딩 단계들을 통해 언어 특징을 추출하는 언어 인코더 및 영상으로부터 상기 복수의 인코딩 단계들을 통해 시각 특징을 추출하는 시각 인코더를 포함하고, 상기 복수의 인코딩 단계들 각각에서 상기 언어 특징 및 상기 시각 특징이 인코더 간에 상호 교환되도록 구현되는 멀티모달 인코더, 및

상기 언어 특징 및 상기 시각 특징 중 적어도 일부를 기초로, 상기 영상에서 상기 언어표현과 대응되는 대상 영역을 분할하는 분할 디코더를 포함하는,

언어지시 영역 영상분할 장치.

청구항 14

제 13항에서,

상기 언어 인코더 및 상기 시각 인코더는,

상기 복수의 인코딩 단계들 각각을 수행하기 위한 입력 단계 모듈, 적어도 하나의 중간 단계 모듈, 및 출력 단계 모듈을 각각 포함하는,

언어지시 영역 영상분할 장치.

청구항 15

제 14항에서,

상기 언어 인코더의 상기 입력 단계 모듈은,

상기 입력 단계 모듈에 의해, 상기 언어표현으로부터 초기 언어 특징을 추출하고, 그리고

상기 시각 인코더의 상기 입력 단계 모듈은,

상기 초기 언어 특징을 기초로, 상기 초기 언어 특징에 상기 영상으로부터 추출되는 초기 시각 특징을 반영한 언어 특징 및 상기 초기 언어 특징을 반영한 시각 특징을 추출하고, 상기 초기 언어 특징을 반영한 시각 특징 및 상기 초기 시각 특징을 반영한 언어 특징을 상기 언어 인코더 및 상기 시각 인코더의 후속하는 단계 모듈로 각각 전달하는,

언어지시 영역 영상분할 장치.

청구항 16

제 15항에서,

상기 초기 언어 특징을 반영한 시각 특징은,

상기 초기 시각 특징을 쿼리(query)로 사용하고 상기 초기 언어 특징을 키(key)와 값(value)으로 사용한 멀티헤드 교차 어텐션(multi-head cross-attention)을 통해 추출되고,

상기 초기 시각 특징을 반영한 언어 특징은,

상기 초기 언어 특징을 쿼리로 사용하고 상기 초기 시각 특징을 키와 값으로 사용한 멀티헤드 교차 어텐션을 통해 추출되는,

언어지시 영역 영상분할 장치.

청구항 17

제 14항에서,

상기 언어 인코더의 상기 적어도 하나의 중간 단계 모듈은,

선행하는 인코딩 단계에서 추출된 선행 언어 특징을 전달받아, 상기 선행 언어 특징에 대해 상기 선행하는 인코딩 단계에서 추출된 선행 시각 특징을 반영한 언어 특징을 추출하고, 상기 선행 시각 특징을 반영한 언어 특징을 상기 시각 인코더의 대응되는 단계 모듈 및 상기 언어 인코더의 후속하는 단계 모듈로 각각 전달하는,

언어지시 영역 영상분할 장치.

청구항 18

제 17항에서,

상기 선행 시각 특징은,

상기 선행 시각 특징에 관한 특징 맵에 대해 서로 다른 커널 사이즈를 가지는 풀링 레이어들을 적용하여, 서로 다른 시각적 관점들을 가지도록 생성되는 멀티 스케일(multi-scale) 시각적 토큰의 형태로 상기 선행 언어 특징에 반영되는,

언어지시 영역 영상분할 장치.

청구항 19

제 14항에서,

상기 시각 인코더의 상기 적어도 하나의 중간 단계 모듈은,

선행하는 인코딩 단계에서 추출된 선행 시각 특징을 전달받아, 상기 선행 시각 특징에 대해 상기 선행하는 인코딩 단계에서 추출된 선행 언어 특징을 반영한 시각 특징을 추출하고, 상기 선행 언어 특징을 반영한 시각 특징을 상기 언어 인코더 및 상기 시각 인코더의 후속하는 단계 모듈로 각각 전달하는,

언어지시 영역 영상분할 장치.

청구항 20

제 19항에서,

상기 선행 언어 특징을 반영한 시각 특징은,

다운 샘플링 된 후, 상기 다운 샘플링 된 시각 특징을 쿼리로 사용하고 상기 선행 언어 특징을 키와 값으로 사용한 멀티헤드 교차 어텐션을 통해 추출되는, 상기 대상 영역에 대한 문맥(context) 정보가 보강된 시각 특징의 형태로 상기 후속하는 단계 모듈로 전달되는,

언어지시 영역 영상분할 장치.

발명의 설명

기 술 분 야

[0001] 본 개시는 멀티모달 인코더 기반의 언어지시 영역 영상분할 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 최근 들어 컴퓨터 비전 또는 자연어 처리와 같은 단일 모달(modal) 분야에 대한 연구를 넘어, 시각-언어 멀티모달(multimodal) 연구가 주목을 받으며 활발히 연구되고 있다. 시각-언어 멀티모달 연구 분야 중 하나인 언어지시 영역 영상 분할은, 여러 객체가 존재하는 영상 내에서 언어 표현으로 묘사된 특정 목표 객체만을 분할 해내는 기술이다. 이 기술은 복잡하고 모호한 자연어 표현을 알맞은 의미로 이해하여 다루어야 하는 문제와, 서로 다른 공간에 임베딩 된 시각 특징과 언어 특징을 같은 공간으로 투영 및 융합해야 하는 문제가 존재한다.

[0003] 기존의 방법론의 네트워크 구조는 크게 두 분류로 나누어진다. 첫 번째 구조는 시각 인코더와 언어 인코더 각각에서 특징을 추출한 후 멀티모달 융합 모듈을 통해 두 특징을 융합하는 방식이다. 이러한 구조는 시각 특징과 언어 특징의 정렬을 멀티모달 융합 모듈에만 의존하게 되어, 풍부한 정보를 지닌 멀티모달 특징을 추출할 수 없는 한계가 있다.

[0004] 이를 개선한 두 번째 구조는 언어 인코더에서 언어 특징을 먼저 추출한 후, 시각 인코더의 중간 계층마다 언어 특징을 주입하여 언어를 통해 인지하는 시각 인코더를 구성하는 방식이다. 이 방식은 인코더 내에서의 시각 특징과 언어 특징의 초기 융합을 진행하는 것이 멀티모달 특징을 보다 효과적으로 정렬할 수 있음을 입증했다.

[0005] 하지만 이러한 방식은, 시각 인코딩 과정에서는 단계를 나누어 진행하는 것과 달리 언어 인코딩 과정은 단계를 나누지 않아 언어 인코더에서 시각 특징을 고려하지 않기 때문에, 복잡하고 모호한 언어 표현을 주어진 영상에 적합한 의미로 이해하지 못하여 목표 객체를 잘못 인지하게 되는 한계가 있다.

[0006] 이에, 언어 및 시각의 두 모달리티를 상호 고려하여 영상으로부터 언어지시 영역을 보다 정확하게 분할할 수 있는 기술이 요구된다.

발명의 내용

해결하려는 과제

[0007] 해결하고자 하는 과제는, 인코더 내 모든 단계에서 두 모달리티가 서로 정보를 주고 받는 교차 인지(cross-aware) 인코딩 기반의 멀티모달 인코더를 기반으로, 영상으로부터 언어지시 영역을 분할하는 방법 및 장치를 제공하는 것이다. 상기 과제 이외에도 구체적으로 언급되지 않은 다른 과제를 달성하는 데 사용될 수 있다.

과제의 해결 수단

[0008] 본 발명의 실시예에 따르면, 컴퓨팅 장치에 의해 수행되는 언어지시 영역 영상분할 방법으로서, 언어표현 및 영상의 쌍을 입력받는 단계, 복수의 인코딩 단계들 각각에서, 선행하는 인코딩 단계로부터 획득되는 상기 영상에 관한 시각 특징을 기초로 상기 언어표현에 관한 언어 특징을 추출하는 단계, 상기 복수의 인코딩 단계들 각각에서, 선행하는 인코딩 단계로부터 획득되는 상기 언어 특징을 기초로 상기 시각 특징을 추출하는 단계, 및 상기 언어 특징 및 상기 시각 특징 중 적어도 일부를 기초로, 상기 영상에서 상기 언어표현과 대응되는 대상 영역을 분할하는 단계를 포함할 수 있다.

[0009] 상기 언어 특징을 추출하는 단계는, 상기 복수의 인코딩 단계들 각각을 수행하기 위한 입력 단계 모듈, 적어도 하나의 중간 단계 모듈, 및 출력 단계 모듈로 구성되는 언어 인코더에 의해 수행될 수 있다.

[0010] 상기 언어 특징을 추출하는 단계는, 상기 입력 단계 모듈에 의해, 상기 언어표현으로부터 초기 언어 특징을 추출하는 단계, 상기 초기 언어 특징을 기초로, 상기 초기 언어 특징에 상기 영상으로부터 추출되는 초기 시각 특징을 반영한 언어 특징이 추출되게끔 하는 단계, 및 후속하는 단계 모듈에 의해, 상기 초기 시각 특징을 반영한 언어 특징을 전달받는 단계를 포함할 수 있다.

[0011] 상기 언어특징을 추출하는 단계는, 상기 적어도 하나의 중간 단계 모듈에 의해, 상기 선행하는 인코딩 단계에서 추출된 선행 언어 특징을 전달받는 단계, 상기 선행 언어 특징에 대해 상기 선행하는 인코딩 단계에서 추출된 선행 시각 특징을 반영한 언어 특징을 추출하는 단계, 및 상기 선행 시각 특징을 반영한 언어 특징을 후속하는 인코딩 단계로 전달하는 단계를 포함할 수 있다.

[0012] 상기 선행 시각 특징은, 상기 선행 시각 특징에 관한 특징 맵에 대해 서로 다른 커널 사이즈를 가지는 풀링 레이어들을 적용하여, 서로 다른 시각적 관점들을 가지도록 생성되는 멀티 스케일(multi-scale) 시각적 토큰의 형태로 상기 선행 언어 특징에 반영될 수 있다.

[0013] 상기 언어특징을 추출하는 단계는, 상기 출력 단계 모듈에 의해, 상기 선행하는 인코딩 단계에서 추출된 선행 언어 특징을 전달받는 단계, 상기 선행 언어 특징에 대해 상기 선행하는 인코딩 단계에서 추출된 선행 시각 특징을 반영한 언어 특징을 추출하는 단계, 및 상기 선행 시각 특징을 반영한 언어 특징을 기초로, 상기 영상으로부터 상기 대상 영역을 식별하게끔 하는 단계를 포함할 수 있다.

[0014] 상기 시각 특징을 추출하는 단계는, 상기 복수의 인코딩 단계들 각각을 수행하기 위한 입력 단계 모듈, 적어도 하나의 중간 단계 모듈, 및 출력 단계 모듈로 구성되는 시각 인코더에 의해 수행될 수 있다.

[0015] 상기 시각 특징을 추출하는 단계는, 상기 입력 단계 모듈에 의해, 상기 영상으로부터 초기 시각 특징을 추출하는 단계, 상기 초기 시각 특징을 기초로, 상기 초기 시각 특징에 상기 언어표현으로부터 추출되는 초기 언어 특징을 반영하여, 상기 초기 언어 특징을 반영한 시각 특징 및 상기 초기 시각 특징을 반영한 언어 특징을 각각 추출하는 단계, 및 상기 초기 언어 특징을 반영한 시각 특징 및 상기 초기 시각 특징을 반영한 언어 특징 각각을 후속하는 인코딩 단계로 전달하는 단계를 포함할 수 있다.

[0016] 상기 각각 추출하는 단계는, 상기 초기 시각 특징을 쿼리(query)로 사용하고 상기 초기 언어 특징을 키(key)와 값(value)으로 사용한 멀티헤드 교차 어텐션(multi-head cross-attention)을 통해, 상기 초기 언어 특징을 반영한 시각 특징을 추출하는 단계, 및 상기 초기 언어 특징을 쿼리로 사용하고 상기 초기 시각 특징을 키와 값으로 사용한 멀티헤드 교차 어텐션을 통해, 상기 초기 시각 특징을 반영한 언어 특징을 추출하는 단계를 포함할 수 있다.

[0017] 상기 시각 특징을 추출하는 단계는, 상기 적어도 하나의 중간 단계 모듈에 의해, 상기 선행하는 인코딩 단계에서 추출된 선행 시각 특징을 전달받는 단계, 상기 선행 시각 특징에 대해 상기 선행하는 인코딩 단계에서 추출된 선행 언어 특징을 반영한 시각 특징을 추출하는 단계, 및 상기 선행 언어 특징을 반영한 시각 특징을 후속하는 인코딩 단계로 전달하는 단계를 포함할 수 있다.

[0018] 상기 후속하는 인코딩 단계로 전달하는 단계는, 상기 선행 언어 특징을 반영한 시각 특징을 다운 샘플링 하는 단계, 상기 다운 샘플링 된 시각 특징을 쿼리로 사용하고 상기 선행 언어 특징을 키와 값으로 사용한 멀티헤드 교차 어텐션을 통해, 상기 대상 영역에 대한 문맥(context) 정보가 보강된 시각 특징을 추출하는 단계, 및 상기 문맥 정보가 보강된 시각 특징을 상기 후속하는 인코딩 단계로 전달하는 단계를 포함할 수 있다.

[0019] 상기 시각특징을 추출하는 단계는, 상기 출력 단계 모듈에 의해, 상기 선행하는 인코딩 단계에서 추출된 선행 시각 특징을 전달받는 단계, 상기 선행 시각 특징에 대해 상기 선행하는 인코딩 단계에서 추출된 선행 언어 특징을 반영한 시각 특징을 추출하는 단계, 및 상기 선행 언어 특징을 반영한 시각 특징을 기초로, 상기 영상으로부터 상기 대상 영역을 식별하는 단계를 포함할 수 있다.

[0020] 본 발명의 실시예에 따르면, 언어지시 영역 영상분할 장치 로서, 언어표현으로부터 복수의 인코딩 단계들을 통해 언어 특징을 추출하는 언어 인코더 및 영상으로부터 상기 복수의 인코딩 단계들을 통해 시각 특징을 추출하는 시각 인코더를 포함하고, 상기 복수의 인코딩 단계들 각각에서 상기 언어 특징 및 상기 시각 특징이 인코더 간에 상호 교환되도록 구현되는 멀티모달 인코더, 및 상기 언어 특징 및 상기 시각 특징 중 적어도 일부를 기초로, 상기 영상에서 상기 언어표현과 대응되는 대상 영역을 분할하는 분할 디코더를 포함할 수 있다.

[0021] 상기 언어 인코더 및 상기 시각 인코더는, 상기 복수의 인코딩 단계들 각각을 수행하기 위한 입력 단계 모듈, 적어도 하나의 중간 단계 모듈, 및 출력 단계 모듈을 각각 포함할 수 있다.

[0022] 상기 언어 인코더의 상기 입력 단계 모듈은, 상기 입력 단계 모듈에 의해, 상기 언어표현으로부터 초기 언어 특징을 추출하고, 그리고 상기 시각 인코더의 상기 입력 단계 모듈은, 상기 초기 언어 특징을 기초로, 상기 초기 언어 특징에 상기 영상으로부터 추출되는 초기 시각 특징을 반영한 언어 특징 및 상기 초기 언어 특징을 반영한 시각 특징을 추출하고, 상기 초기 언어 특징을 반영한 시각 특징 및 상기 초기 시각 특징을 반영한 언어 특징을 상기 언어 인코더 및 상기 시각 인코더의 후속하는 단계 모듈로 각각 전달할 수 있다.

[0023] 상기 초기 언어 특징을 반영한 시각 특징은, 상기 초기 시각 특징을 쿼리(query)로 사용하고 상기 초기 언어 특징을 키(key)와 값(value)으로 사용한 멀티헤드 교차 어텐션(multi-head cross-attention)을 통해 추출되고, 상기 초기 시각 특징을 반영한 언어 특징은, 상기 초기 언어 특징을 쿼리로 사용하고 상기 초기 시각 특징을 키와 값으로 사용한 멀티헤드 교차 어텐션을 통해 추출될 수 있다.

[0024] 상기 언어 인코더의 상기 적어도 하나의 중간 단계 모듈은, 선행하는 인코딩 단계에서 추출된 선행 언어 특징을 전달받아, 상기 선행 언어 특징에 대해 상기 선행하는 인코딩 단계에서 추출된 선행 시각 특징을 반영한 언어

특징을 추출하고, 상기 선행 시각 특징을 반영한 언어 특징을 상기 시각 인코더의 대응되는 단계 모듈 및 상기 언어 인코더의 후속하는 단계 모듈로 각각 전달할 수 있다.

[0025] 상기 선행 시각 특징은, 상기 선행 시각 특징에 관한 특징 맵에 대해 서로 다른 커널 사이즈를 가지는 풀링 레이어들을 적용하여, 서로 다른 시각적 관점들을 가지도록 생성되는 멀티 스케일(multi-scale) 시각적 토큰의 형태로 상기 선행 언어 특징에 반영될 수 있다.

[0026] 상기 시각 인코더의 상기 적어도 하나의 중간 단계 모듈은, 선행하는 인코딩 단계에서 추출된 선행 시각 특징을 전달받아, 상기 선행 시각 특징에 대해 상기 선행하는 인코딩 단계에서 추출된 선행 언어 특징을 반영한 시각 특징을 추출하고, 상기 선행 언어 특징을 반영한 시각 특징을 상기 언어 인코더 및 상기 시각 인코더의 후속하는 단계 모듈로 각각 전달할 수 있다.

[0027] 상기 선행 언어 특징을 반영한 시각 특징은, 다운 샘플링 된 후, 상기 다운 샘플링 된 시각 특징을 쿼리로 사용하고 상기 선행 언어 특징을 키와 값으로 사용한 멀티헤드 교차 어텐션을 통해 추출되는, 상기 대상 영역에 대한 문맥(context) 정보가 보강된 시각 특징의 형태로 상기 후속하는 단계 모듈로 전달될 수 있다.

발명의 효과

[0028] 본 발명의 실시예에 따르면, 멀티모달 인코더 내의 로우레벨(low-level)부터 하이레벨(high-level) 단계까지 두 모달리티의 융합을 진행하여, 두 모달리티의 특징을 같은 공간상에 효과적으로 정렬할 수 있다.

[0029] 본 발명의 실시예에 따르면, 각각의 모달리티는 각 단계에서 넘겨받은 상대 모달리티의 관점을 기초로 자신의 특징의 강인함(robustness)을 강화하여, 두 모달리티가 상호 이익을 얻을 수 있다.

[0030] 본 발명의 실시예에 따르면, 언어 인코딩 레이어가 시각 인코딩 레이어에서 넘겨받은 시각 특징을 기초로 시각적 정보를 지닌 토큰들을 생성함으로써, 시각적 정보를 고려하여 더욱 견고한 언어 특징을 추출할 수 있고, 궁극적으로 복잡하고 모호한 언어 표현에 대응하여 언어지시 영역을 정확하게 분할할 수 있다.

도면의 간단한 설명

[0031] 도 1은 본 개시의 몇몇 실시예에 따른 언어지시 영역 영상분할 장치를 나타낸 블록도이다.

도 2는 본 개시의 몇몇 실시예에 따른 멀티모달 인코더의 각 구성을 구체적으로 도시한 도면이다.

도 3은 본 개시의 몇몇 실시예에 따른 시각적 토큰의 생성 프로세스를 개략적으로 도시한 도면이다.

도 4는 본 개시의 몇몇 실시예에 따른 언어지시 영역 영상분할 방법의 성능을 종래의 방법들과 비교 도시한 도면이다.

도 5는 본 개시의 몇몇 실시예에 따른 언어지시 영역 영상분할 방법의 성능을 종래의 방법들과 비교 도시한 도면이다.

도 6은 본 개시의 몇몇 실시예에 따른 언어지시 영역 영상분할 방법의 순서도이다.

도 7은 본 개시의 몇몇 실시예에 따른 영상 분할 방법을 제공하는 컴퓨팅 장치를 나타낸 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0032] 아래에서는 첨부한 도면을 참고로 하여 본 개시의 실시예에 대하여 본 개시가 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그러나 본 개시는 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 개시를 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0033] 본 개시에서, 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 의미한다. 네트워크를 구성하는 장치들은 하드웨어나 소프트웨어 또는 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.

[0034] 도 1은 본 개시의 몇몇 실시예에 따른 언어지시 영역 영상분할 장치를 나타낸 블록도이다.

[0035] 도 1을 참조하면, 본 개시에 따른 언어지시 영역 영상분할 장치(100)는 언어 인코더(110)와 시각 인코더(120)를

포함하는 멀티모달(multimodal) 인코더 및 분할 디코더(130)를 포함할 수 있다. 다만 상술한 구성은 본 개시에 따른 언어지시 영역 영상분할 장치(100)를 구현하는 데 필수적인 것은 아니어서, 언어지시 영역 영상분할 장치(100)는 열거된 구성보다 많거나 적은 구성들로 구현될 수 있다.

[0036] 본 개시에 따른 언어지시 영역 영상분할 장치(100)의 언어 인코더(110) 및 시각 인코더(120) 각각은 분할된 인코딩 단계들을 수행하기 위한 복수의 단계 모듈들을 포함할 수 있다. 언어 인코더(110)는 분할된 언어 인코딩 단계들을 수행하기 위한 복수의 단계 모듈들(111, 112, 113, 114)을 포함할 수 있고, 시각 인코더(120) 또한 분할된 시각 인코딩 단계들을 수행하기 위한 복수의 단계 모듈들(121, 122, 123 124)을 포함할 수 있다.

[0037] 이는 단계를 분할하지 않고 특징을 인코딩 하는 종래의 언어 인코딩 프로세스와 구별된다. 본 개시에 따른 언어 인코더(110) 및 시각 인코더(120)는, 인코딩 프로세스의 분할된 각 단계에서 언어 인코딩 단계 모듈들(111, 112, 113, 114)과 시각 인코딩 단계 모듈들(121, 122, 123, 124) 간에 정보를 교환할 수 있다.

[0038] 본 개시에서는 각각의 인코더(110, 120)가 두 개의 중간 단계 모듈들(112, 113, 122, 123)을 포함하여 총 네 개의 단계 모듈들로 구성되는 일례를 도시하였으나 이에 한정되는 것은 아니며, 각각의 인코더(110, 120)는 이보다 적거나 많은 개수의 인코딩 단계 모듈들로 구성될 수 있다.

[0039] 이처럼 멀티모달 인코더의 하위 단계에서 상위 단계에 이르기까지 단계 별로 교차 인지 이중 인코딩을 수행함으로써, 종래에 언어 인코더(110) 및 시각 인코더(120) 각각에서 특징을 추출한 후 추가적인 멀티모달 융합 모듈을 사용하던 것에 비하여 보다 효과적으로 언어 특징 및 시각 특징을 상호 정렬할 수 있다.

[0040] 먼저 본 개시에 따른 언어지시 영역 영상분할 장치(100)의 언어 인코더(110)는, 시각 인코더(120)에 의해 영상으로부터 획득되는 시각 특징을 고려하여, 영상과 쌍을 이루는 언어표현의 의미를 해석하는 시각인지(visual-aware) 언어 인코딩을 수행할 수 있다.

[0041] 구체적으로 언어 인코더(110)의 단계 모듈들(111, 112, 113, 114) 각각은, 시각 인코더(120)의 단계 모듈들(121, 122, 123, 124) 각각에서 획득되는 멀티스케일 시각적 토큰들을 사용하여 멀티스케일 시각인지 언어 어텐션(Multi-scale Visual-aware Language Attention; MVLA)을 수행할 수 있고, 이를 통해 다양한 시각적 관점에서 언어 표현의 의미를 해석하여 언어 특징의 강인함을 향상시키고 언어 표현의 복잡성과 모호성을 극복할 수 있다.

[0042] 언어 인코더(110)와 시각 인코더(120) 간의 이러한 교차인지 이중 인코딩은, 상호 모달리티 특징들을 효과적으로 정렬하기 위해, 각 인코더의 하위 단계에서부터 상위 단계까지 각 단계들에서 수행될 수 있다.

[0043] 언어 인코더(110)를 구성하는 언어 인코딩 단계 모듈들(111, 112, 113, 114) 각각은 언어 특징을 추출하기 위한 트랜스포머 기반 언어 모델인 BERT 모델을 기반으로 설계될 수 있다. 언어 인코더(110)의 입력 단계 모듈(111)은 언어 전용(language-only) 인코딩 모듈로서, 멀티헤드 셀프 어텐션(multi-head self-attention) 모듈과 피드 포워드(feed forward) 모듈로 구성될 수 있고, 한편 언어 인코더(110)의 다른 단계 모듈들(112, 113, 114)은 시각인지 언어 인코딩 모듈들로서, 멀티헤드 셀프 어텐션 모듈, MVLA 모듈, 및 피드 포워드 모듈로 구성될 수 있다.

[0044] 언어 인코딩 단계 모듈들(112, 113, 114) 내의 MVLA 모듈은 강인한(robust) 언어 특징을 추출할 수 있다. 이전 단계 모듈에서 획득한 시각인지 언어 특징은 먼저 멀티헤드 셀프 어텐션 모듈로 전송되어 장기적 종속성(long-term dependencies)을 학습할 수 있고, 그 후 MVLA 모듈은 다양한 시각 관점들을 참고하여 언어 특징의 강인함(robustness)을 강화할 수 있다.

[0045] MVLA 모듈은 언어 특징을 쿼리로 사용하고, 큰 단위부터 작은 단위까지의(coarse-to-fine grained) 시각적 토큰들을 키와 값으로 사용하여 멀티헤드 교차 어텐션을 수행할 수 있다. 큰 단위부터 작은 단위까지의 시각적 토큰들은 멀티스케일 시각적 토큰 임베딩을 수행함으로써, 이전 시각 인코딩 단계에서 가져온 언어인지 시각 특징으로부터 생성될 수 있다.

[0046] 이와 같은 MVLA 모듈을 통해, 더 풍부한 시각적 정보를 참고하여 언어 특징을 향상시킬 수 있다. 시각인지 언어 특징은 피드 포워드 모듈을 수행한 후 시각 인코딩 단계 모듈 및 다음 언어 인코딩 단계 모듈로 전달될 수 있다. 이러한 방식으로 각 단계에서 교차 인지을 수행함으로서, 본 개시에 따른 언어지시 영역 영상분할 방법은 시각 특징 및 언어 특징들을 효과적으로 정렬하고, 각 모달리티의 특징을 상호 향상시킬 수 있다.

[0047] 본 개시에 따른 언어지시 영역 영상분할 장치(100)의 시각 인코더(120)는, 언어 인코더(110)에 의해 언어표현으로부터 획득되는 언어 특징을 고려하여, 언어표현과 쌍을 이루는 영상 내에서 언어표현에 대응되는 대상 영역을

해석하는 언어인지(language-aware) 시각 인코딩을 수행할 수 있다.

[0048] 구체적으로 시각 인코더(120)의 각 인코딩 단계 모듈들(121, 122, 123, 124)은, 언어 인코더(110)의 각 단계 모듈들(111, 112, 113, 114)에서 획득되는 강인한 언어 특징을 사용하여 언어인지 멀티헤드 어텐션(Language-aware Multi-Head Attention; LMHA)을 수행함으로서, 시각적인 대상(object) 영역의 컨텍스트 정보를 풍부하게 할 수 있다.

[0049] 시각 인코더(120)와 언어 인코더(110) 간의 이러한 교차인지 이중 인코딩은 상호 모달리티 특징들을 효과적으로 정렬하기 위해, 각 인코더의 하위 단계에서부터 상위 단계까지 각 단계 모듈들에서 수행될 수 있다.

[0050] 시각 인코더(120)를 구성하는 시각 인코딩 단계 모듈들(121, 122, 123, 124) 중 출력 모듈(124)을 제외한 나머지 모듈들 각각은, 트랜스포머 인코딩 모듈, 사전(pre-) LMHA 모듈 및 사후(post-) LMHA 모듈을 포함할 수 있다.

[0051] 한편 시각 인코더(120)의 입력 모듈(121)은 사전 LMHA 모듈 대신 SCA(Sharing Cross-Attention; 공유 교차 어텐션) 모듈을 포함할 수 있으며, SCA 모듈은 언어 인코더(110)의 입력 모듈(111)에서 추출된 언어 특징을 키와 값으로 사용하고, 시각 인코더(120)의 입력 모듈(121) 내 트랜스포머 인코딩 모듈에서 인코딩 된 시각 특징을 키와 값으로 사용한 멀티헤드 교차 어텐션(multi-head cross-attention)을 수행하여, 언어인지 시각 특징 뿐만 아니라 언어 인코더(110)의 입력 모듈(111)에서는 추출될 수 없는 시각인지 언어 특징도 추출할 수 있다.

[0052] SCA 모듈을 통해, 동일한 시각 및 언어 특징들에 대한 어텐션 스코어를 반복적으로 계산하지 않고도 각 교차인지 특징(cross-aware feature)을 효율적으로 추출할 수 있다.

[0053] 획득된 시각인지 언어 특징은 다음 단계의 언어 인코딩 단계 모듈(112)로 넘어갈 수 있고, 언어인지 시각 특징은 다운 샘플링 되어 언어 인코딩 단계 모듈(112) 및 다음 단계의 시각 인코딩 모듈(122) 모두에 제공될 수 있다.

[0054] 각각의 시각 인코딩 단계 모듈(121, 122, 123, 124) 내의 사후 LMHA 모듈은, 다운 샘플링 된 언어인지 시각 특징에 객체 영역에 대한 풍부한 컨텍스트 정보를 전파(propagate)하기 위해, 다운 샘플링 된 언어인지 시각 특징을 쿼리로 사용하고 언어 특징을 키와 값으로 사용하여 멀티헤드 교차 어텐션을 수행할 수 있다.

[0055] 사후 LMHA 모듈을 통해 시각 피처 맵의 각 픽셀과 관련된 언어 정보를 참조하여, 시각 특징의 컨텍스트 정보를 보강(enriched)할 수 있다. 멀티헤드 프로세스는 또한 다양한 관점(perspectives)에서 정보를 수집하여 시각 특징의 강인함(robustness)을 향상시킬 수 있다. 향상된(evolved) 시각 특징은 언어 인코딩 단계모듈 및 다음 단계의 시각 인코딩 모듈 모두에 제공될 수 있다.

[0056] 한편 시각 인코더(120)를 구성하는 시각 인코딩 단계 모듈들(121, 122, 123, 124) 중 입력 단계 모듈(121)을 제외한 나머지 단계 모듈들은, 시각인지 언어 특징이 언어 인코딩 단계 모듈들로부터 획득되기 때문에, 각각의 시각 인코딩 단계 모듈들에서 SCA 대신 사전 LMHA 모듈이 사용될 수 있다.

[0057] 본 개시에 따른 언어지시 영역 영상분할 장치(100)의 분할 디코더(130)는, 멀티모달 인코더로부터 수신한 언어인지 시각 특징을 기초로 대상 객체에 대한 최종 분할 마스크를 생성할 수 있다.

[0058] 본 개시의 분할 디코더(130)는, 멀티모달 인코더의 효율성을 검증하기 위해 단순한 구조로 설계될 수 있다. 가령 분할 디코더(130)는 3개의 디코더 블록을 포함하며, 3X3 컨볼루션, 배치 정규화(batch normalization) 및 ReLU 함수가 스태킹 된(staked) 구조를 가질 수 있다.

[0059] 디코더 특징은 쌍선형 보간(bilinear interpolation)에 의해 업 샘플링 된 다음, 멀티모달 인코더의 언어인지 시각 특징과 연결(concatenating)된 후 다음 디코더 블록에 공급될 수 있다. 최종 예측 마스크는 1X1 컨볼루션을 수행하여 이진 클래스 마스크로 투영(projected)될 수 있다. 네트워크 훈련에는 이진 교차 엔트로피 손실함수가 사용될 수 있다.

[0060] 도 2는 본 개시의 몇몇 실시예에 따른 멀티모달 인코더의 각 구성을 구체적으로 도시한 도면이다.

[0061] 도 2를 참고하여, 먼저 언어 인코더(110)의 각 단계 모듈들의 구성을 설명한다. 언어 인코더(110)의 입력 단계 모듈(111)은 언어 전용(language-only) 인코딩 모듈로서, 멀티헤드 셀프 어텐션 모듈(111a)과 피드 포워드 모듈(111b)로 구성될 수 있다.

[0062] 언어 특징을 시각 인코더(120)의 입력 단계 모듈(121)로 전달하기 위해, 언어 인코더(110)의 입력 단계 모듈

(111)은 먼저 수식 1로 표현되는 고차원 언어 특징을 추출하여 단어 간의 상관 관계를 학습할 수 있다. 여기서 L과 D는 각각 언어 표현의 길이와 특징 차원을 의미한다.

[0063] (수식 1)

$$F_{l1} \in \mathbb{R}^{L \times D}$$

[0064]

[0065] 한편 언어 인코더(110)의 다른 단계 모듈들(112, 113, 114)은 시각인지 언어 인코딩 모듈로서, 멀티헤드 셀프 어텐션 모듈(112a, 113a, 114a), MVLA 모듈(112b, 113b, 114b), 및 피드 포워드 모듈(112c, 113c, 114c)로 구성될 수 있다.

[0066] 언어 인코딩 단계 모듈들(112, 113, 114) 내의 MVLA 모듈(112b, 113b, 114b)은 강인한 언어 특징을 추출할 수 있다. 이전 단계에서 획득한 시각인지 언어 특징은 먼저 멀티헤드 셀프 어텐션 모듈(112a, 113a, 114a)로 전송되어 장기적 종속성을 학습할 수 있고, 그 후 MVLA 모듈(112b, 113b, 114b)은 다양한 시각 관점들을 참고하여 언어 특징의 강인함을 강화할 수 있다.

[0067] MVLA 모듈(112b, 113b, 114b)은 언어 특징을 쿼리로 사용하고, 큰 단위부터 작은 단위까지의(coarse-to-fine grained) 시각적 토큰들을 키와 값으로 사용하여 멀티헤드 교차 어텐션을 수행할 수 있다. 큰 단위부터 작은 단위까지의 시각적 토큰들은 멀티스케일 시각적 토큰 임베딩을 수행함으로써, 이전 시각 인코딩 단계에서 가져온 언어인지 시각 특징 F_{mi}^{2} 로부터 생성될 수 있다.

[0068] 멀티스케일 시각적 토큰 임베딩에서 F_{mi}^{2} 는 먼저 2D 피처 맵으로 재구성(reshaped)될 수 있다. 그 후 다양한 커널 사이즈를 가진 평균 풀링 레이어(average pooling layer)를 적용하여, 작은 단위의(fine-grained) 시각 특징 F_{mi}^{2} 로부터 멀티스케일의 큰 단위의(coarse-grained) 시각 특징을 획득할 수 있다.

[0069] 각각의 멀티스케일 시각적 토큰들은, 서로 다른 크기의 수용 필드(receptive fields)에서 특징을 추출하여, 서로 다른 시각적 관점을 가질 수 있다. 가령 스트라이드(stride)가 1인 3X3 커널 사이즈와, 스트라이드가 2인 5X5 커널 사이즈가 채택될 수 있으며, 다만 이에 한정되는 것은 아니다.

[0070] 큰 단위의 시각 특징은 상대적 위치 인코딩(relative positional encoding)을 위한 깊이별 컨볼루션(depthwise convolution)에 의해 임베딩 되고, 작은 단위의 시각 특징과 연결(concatenated)될 수 있다. 그 후 큰 단위부터 작은 단위까지의 시각 특징을 평면화(flattening)하여 큰 단위부터 작은 단위까지의 시각적 토큰들을 생성할 수 있고, 이는 시각인지 언어 특징 F_{ti} 의 추출을 위해 사용될 수 있다.

[0071] 이와 같은 MVLA 모듈(112b, 113b, 114b)을 통해, 더 풍부한 시각적 정보를 참고하여 언어 특징을 향상시킬 수 있다. 피드 포워드 모듈(112c, 113c, 114c)를 수행한 후, 시각인지 언어 특징은 시각 단계 모듈 및 다음 언어 단계 모듈로 전달될 수 있다.

[0072] 이러한 방식으로 각 단계에서 교차 인지을 수행함으로서, 본 개시에 따른 언어지시 영역 영상분할 방법은 시각 특징 및 언어 특징들을 효과적으로 정렬하고, 각 모달리티의 특징을 상호 향상시킬 수 있다.

[0073] 다음으로 도 2를 참고하여, 시각 인코더(120)의 각 단계 모듈들의 구성을 설명한다. 시각 인코더(120)를 구성하는 시각 인코딩 단계 모듈들(121, 122, 123, 124) 중 출력 단계 모듈(124)을 제외한 나머지 단계 모듈들 각각은, 트랜스포머 인코딩 모듈(121a, 122a, 123a, 124a), 사전 LMHA 모듈(122b, 123b, 124b) 및 사후 LMHA 모듈(121c, 122c, 123c)을 포함할 수 있다.

[0074] 여기서 트랜스포머 인코딩 모듈들(121a, 122a, 123a, 124a)은 Swin 트랜스포머(Liu et al. 2021)의 각 단계에 해당할 수 있다. 한편 i-번째 시각 인코딩 단계 모듈의 사전 LMHA 모듈(122b, 123b, 124b)에서 추출한 언어인지 시각 특징과 사후 LMHA 모듈(121c, 122c, 123c)에서 추출한 언어인지 시각 특징은 각각 하기 수식 2와 같이 정의될 수 있다. 여기서 H, W, C는 각각 특징 맵(feature map)의 높이와 너비, 그리고 트랜스포머의 i-번째 단계

의 채널 수를 의미한다:

[0075] (수식 2)

$$pre\text{-}LMHA\ vision\ features = F_{mi}^{1} \in \mathbb{R}^{H_i W_i \times C_i}$$

[0076]

$$post\text{-}LMHA\ vision\ features = F_{mi}^{2} \in \mathbb{R}^{H_{i+1} W_{i+1} \times C_{i+1}}$$

[0077]

[0078] 시각 인코더(120)의 입력 단계 모듈(121)은 사전 LMHA 레이어 대신 SCA(Sharing Cross-Attention; 공유 교차 어텐션) 모듈(121b)을 포함할 수 있으며, SCA 모듈(121b)은 언어 인코더(110)의 입력 단계 모듈(111)에서 추출된 언어 특징 F_{l1} 과 시각 인코더(120)의 입력 단계 모듈(121)의 트랜스포머 인코딩 모듈(121a)에서 인코딩 된 시각 특징 $F_{v1} \in \mathbb{R}^{H_1 W_1 \times C_1}$ 을 사용할 수 있다.

[0079] SCA 모듈(121b)은 언어인지 시각 특징인 F_{m1}^{1} 뿐만 아니라, 언어 인코더(110)의 입력 단계 모듈(111)에서는 추출될 수 없는 시각인지 언어 특징인 $F_{t1} \in \mathbb{R}^{L \times D}$ 또한 추출하기 위해 공유 어텐션 스코어 행렬(shared attention score matrix) $Att \in \mathbb{R}^{H_1 W_1 \times L}$ 을 사용하여 멀티헤드 교차 어텐션(multi-head cross-attention)을 수행한다.

[0080] SCA 모듈(121b) 통해, 동일한 시각 및 언어 특징들에 대한 어텐션 스코어를 반복적으로 계산하지 않고도 각 교차인지 특징(cross-aware feature)을 효율적으로 추출할 수 있다. SCA 모듈(121b)은 하기 수식 3의 프로세스에 따라 언어인지 시각 특징을 획득할 수 있다:

[0081] (수식 3)

$$Q_v^j = F_{v1} W_q^j, K_l^j = F_{l1} W_k^j, Att^j = \frac{Q_v^j K_l^{jT}}{\sqrt{d}},$$

[0082]

$$V_l^j = F_{l1} W_{vl}^j, A_v^j = softmax\left(Att^j\right) V_l^j,$$

[0083]

$$M_v = MultiHead(Q, K, V_l) = [A_v^0, \ldots, A_v^n] W_v,$$

[0084]

$$F_{c1}^{1} = M_v + F_{v1}, F_{m1}^{1} = FFN(F_{c1}^{1}) + F_{v1}$$

[0085]

[0086] 수식 3에서 Q_v, K_l, V_l 및 d는 각각 시각 특징들의 쿼리(queries), 언어 특징들의 키(keys) 및 값(values), 그리고 키의 차원(dimension of keys)을 나타내며 d의 값은 가령 64일 수 있으나 이에 한정되는 것은 아니다. 또한 F_{c1}^{1} 은 중간 특징(intermediate features)이고, W_q , W_k , W_{vl} 및 W_v는 투영(projection)에 대해 학습 가능한 매개변수이며, FFN(*)은 피드 포워드(feed-forward) 레이어이다. j(=1, 2, ..., n)는 헤드의 수를 나타낸다.

[0087] 즉, 수식 3은 시각 특징 F_{v1} 을 쿼리로 사용하고, 언어 특징 F_{l1} 을 키와 값으로 사용하여 멀티헤드 교차 어텐션을 구현할 수 있다. SCA 모듈(121b)은 하기 수식 4의 프로세스에 따라 시각인지 언어 특징을 획득할 수 있다:

[0088] (수식 4)

$$V_v^j = F_{v1} W_{vv}^j, A_l^j = softmax\left(Att^{jT}\right) V_v^j$$

[0089]

$$M_l = MultiHead(K_l, Q_v, V_v) = [A_l^0, \ldots, A_l^n] W_l,$$

[0090]

$$F_{t1} = FFN(M_l + F_{l1}) + F_{l1}$$

[0091]

[0092] 수식 4에서 여기서 V_v는 시각 특징의 값을 나타내고 W_{vv} 및 W_l은 투영에 대해 학습 가능한 매개변수이다. 수식 3에서 키로 사용되는 K_l은 수식 4에서 쿼리로 사용될 수 있다. 수식 3에서 구한 어텐션 스코어 행렬 Att는 수식 4에서 전치되어, F_{l1}을 쿼리로, F_{v1}을 키와 값으로 사용하여 멀티헤드 교차 어텐션을 구현할 수 있다. 수식 3 및 수식 4에서 언어인지 시각 특징 F_{m1}^1과 시각인지 언어 특징 F_{t1}은 수집된 멀티모달 정보를 정제(refine)하는 피드 포워드 모듈을 수행한 후 획득될 수 있다.

[0093] 획득된 시각인지 언어 특징 F_{t1}은 다음 단계의 언어 인코딩 단계 모듈(112)로 넘어갈 수 있고, 언어인지 시각 특징 F_{m1}^1은 다운 샘플링될 수 있다. 다운 샘플링 된 시각 특징 $\tilde{F}_{m1}^1 \in \mathbb{R}^{H_2 W_2 \times C_2}$은 언어 인코딩 단계 모듈 및 다음 단계의 시각 인코딩 모듈 모두에 제공되며, 객체 영역의 풍부한 컨텍스트 정보를 캡처할 필요가 있다.

[0094] 각각의 시각 인코딩 단계 모듈(121, 122, 123, 124) 내의 사후 LMHA 모듈(121c, 122c, 123c)은, 다운 샘플링 된 시각 특징에 컨텍스트 정보를 전파(propagate)하기 위해, $\tilde{F}_{m1}^1$을 쿼리로 사용하고 F_{l1}을 키와 값으로 사용하여 멀티헤드 교차 어텐션을 수행할 수 있다. 사후 LMHA 모듈(121c, 122c, 123c)은 하기 수식 5의 프로세스를 따를 수 있다:

[0095] (수식 5)

$$F_{ci}^2 = MHCA\left(Q: \tilde{F}_{mi}^1, K: F_{li}, V: F_{li}\right) + \tilde{F}_{mi}^1,$$

[0096]

$$F_{mi}^2 = FFN(F_{ci}^2) + \tilde{F}_{mi}^1$$

[0097]

[0098] 수식 5에서 F_{ci}^2는 중간 특징(intermediate features)이고 MHCA(*)는 멀티헤드 교차 어텐션 모듈이다. 이 모듈을 통해 시각 피처 맵의 각 픽셀과 관련된 언어 정보를 참조하여, 시각 특징의 컨텍스트 정보를 보강(enriched)할 수 있다. 멀티헤드 프로세스는 또한 다양한 관점(perspectives)에서 정보를 수집하여 시각 특징의 강인함(robustness)을 향상시킬 수 있다. 향상된(evolved) 시각 특징 F_{mi}^2는 언어 인코딩 단계 모듈 및 다음 단계의 시각 인코딩 모듈 모두에 제공될 수 있다.

[0099] 한편 시각 인코더(120)를 구성하는 시각 인코딩 단계 모듈들(121, 122, 123, 124) 중 입력 단계 모듈(121)을 제외한 나머지 모듈들은, 시각인지 언어 특징이 언어 인코딩 단계 모듈들로부터 획득되기 때문에, 각각의 시각 인코딩 단계 모듈들에서 SCA 모듈 대신 사전 LMHA 모듈(122b, 123b, 124b)이 사용될 수 있다. 사전 LMHA 모듈(122b, 123b, 124b)은 하기 수식 6의 프로세스를 따를 수 있다. 중간 특징인 F_{ci}^1은 후술할 분할 디코더(130)에

서 사용될 수 있으며, 한편 시각 인코더(120)의 출력 단계 모듈(124)에서 추출되는 F_{m4}^{1} 는 다운 샘플링 없이 분할 디코더(130)에 입력될 수 있다:

[0100] (수식 6)

$$F_{ci}^{1} = MHCA(Q{:}F_{vi}, K{:}F_{li}, V{:}F_{li}) + F_{vi},$$

[0101]

$$F_{mi}^{1} = FFN(F_{ci}^{1}) + F_{vi}$$

[0102]

[0103] 도 3은 본 개시의 몇몇 실시예에 따른 시각적 토큰의 생성 프로세스를 개략적으로 도시한 도면이다.

[0104] 언어 인코더(110)의 MVLA 모듈들(112b, 113b, 114b)은 언어 특징을 쿼리로 사용하고, 큰 단위부터 작은 단위까지의 시각적 토큰들을 키와 값으로 사용하여 멀티헤드 교차 어텐션을 수행할 수 있다. 큰 단위부터 작은 단위까지의 시각적 토큰들을 사용함으로써, 더 풍부한 시각적 정보를 참고하여 언어 특징을 향상시킬 수 있다.

[0105] 여기서 큰 단위부터 작은 단위까지의 시각적 토큰들은 멀티스케일 시각적 토큰 임베딩을 수행함으로써, 이전 시각 인코딩 단계에서 가져온 언어인지 시각 특징으로부터 생성될 수 있다.

[0106] 도 3을 참고하면, 멀티스케일 시각적 토큰 임베딩에서 언어인지 시각 특징은 먼저 2D 피처 맵으로 재구성(reshaped)될 수 있다. 그 후 다양한 커널 사이즈를 가진 평균 풀링 레이어(average pooling layer)를 적용하여, 작은 단위의(fine-grained) 언어인지 시각 특징으로부터 큰 단위의(coarse-grained) 시각 특징을 획득할 수 있다.

[0107] 각각의 멀티스케일 시각적 토큰들은, 서로 다른 크기의 수용 필드(receptive fields)에서 특징을 추출하여, 서로 다른 시각적 관점을 가질 수 있다. 가령 스트라이드(stride)가 1인 3X3 커널 사이즈와, 스트라이드가 2인 5X5 커널 사이즈가 채택될 수 있으며, 다만 이에 한정되는 것은 아니다.

[0108] 큰 단위의 시각 특징은 상대적 위치 인코딩(relative positional encoding)을 위한 깊이별 컨볼루션(depthwise convolution)에 의해 임베딩 되고, 작은 단위의 시각 특징과 연결(concatenated)될 수 있다. 그 후 큰 단위부터 작은 단위까지의 시각 특징을 평면화(flattening)하여 큰 단위부터 작은 단위까지의 시각적 토큰들을 생성할 수 있고, 이는 시각인지 언어 특징의 추출을 위해 사용될 수 있다.

[0109] 도 4 및 도 5는 본 개시의 몇몇 실시예에 따른 언어지시 영역 영상분할 방법의 성능을 종래의 방법들과 비교 도시한 도면이다.

[0110] 본 개시에서 제안된 언어지시 영역 영상분할 방법에 대해, 잘 알려진 세 가지 벤치마크 데이터셋인 RefCOCO, RefCOCO+및 G-Ref을 사용하여, 언어지시 영역 영상분할 방법의 각 구성요소에 대한 제거 연구(ablation study) 및 종래의 최고 수준 모델(state-of-the-art; SOTA)들에 대한 성능 비교를 각각 수행하였다.

[0111] 데이터셋 RefCOCO(Kazemzadeh et al. 2014), RefCOCO+(Kazemzadeh et al. 2014) 및 G-Ref(Mao et al. 2016)는 언어지시 영역 영상분할(referring image segmentation)에 널리 사용되는 데이터셋으로, 객체에 대한 언어표현 및 이와 쌍을 이루는 영상들로 구성된다.

[0112] RefCOCO+의 언어표현은 절대 위치에 대한 단어를 포함하지 않아, RefCOCO 대비 난이도가 높으며, RefCOCO 및 RefCOCO+의 각 언어표현은 평균 3.5개의 단어를 포함하는 것에 비해 G-Ref의 언어표현은 평균 8.4단어를 포함하여 복잡도가 높으므로, RefCOCO 및 RefCOCO+ 대비 난이도가 더 높은 데이터셋에 해당한다.

[0113] 성능 평가는 상술한 세 가지 데이터셋에서, 언어표현에 해당하는 대상 객체를 영상 내에서 얼마나 정확하게 하이라이트하는지 여부를 기준으로 수행되었다. 하이라이트의 정확도에 관한 지표로는, 전체 교차 결합(overall intersection-over-union; oIoU) 측정항목(metric), 평균 교차 결합(mean intersection-over-union; mIoU) 측정항목 및 정밀도(precision)가 사용되었으며, 임계값으로는 0.5, 0.7 및 0.9이 사용되었다.

[0114] 여기서 oIoU는 전체 교차 영역(total intersection regions)과 모든 테스트 세트의 총 합집합 영역(total union regions) 간의 비율이고, mIoU는 예측 마스크와 모든 테스트 세트의 정답(ground truth) 사이의 평균 IoU 값이며, 정밀도는 IoU 점수가 임계값보다 높은 테스트 세트의 백분율로 계산된다.

[0115] 먼저 언어지시 영역 영상분할 방법의 주요 구성요소인 SCA(Sharing cross-attention; 공유 교차 어텐션) 모듈,

MVLA(Multi-scale Visual-aware Language Attention Module; 멀티스케일 시각인지 언어 어텐션) 모듈, 및 LMHA(Language-aware Multi-Head Attention module; 언어인지 멀티헤드 어텐션) 모듈 각각에 대한 제거 연구(ablation study)를 수행하였다.

표 1

[0116]

SCA	MVLA		P@0.5	P@0.7	P@0.9	mIoU	oIoU
	MS	SS					
?	*		**85.23**	**77.18**	**35.73**	**75.01**	**72.90**
*		*	84.64	76.31	34.95	74.49	72.58
*			83.97	74.75	33.02	73.67	72.12
	*		84.18	75.20	33.83	74.24	72.36
		*	83.93	74.36	32.85	73.87	72.06
			83.31	73.89	32.77	73.36	71.59

표 2

[0117]

LMHA		P@0.5	P@0.7	P@0.9	mIoU	oIoU
head	post					
multi	*	**85.23**	**77.18**	**35.73**	**75.01**	**72.90**
multi		82.21	70.81	28.98	71.74	70.28
single	*	84.56	76.45	34.90	74.50	72.26

[0118] 먼저 시각 인코더(120)의 SCA 모듈은, 언어 인코더(110)의 입력 단계 모듈(111)은 다른 언어 인코딩 단계 모듈들과 달리 시각인지 언어 특징을 추출할 수 없으므로, 이를 대신 추출하여 언어 인코더(110)의 다음 단계 모듈로 전달하는 역할을 한다. SCA 모듈에 관한 제거 연구에서는 시각 인코더(120)의 입력 모듈(121)에서 SCA를 시각인지 언어 특징을 추출하지 않는 LMHA로 교체하였다.

[0119] 표 1을 참고하면, 이 경우에 mIoU와 oIoU 점수는 각각 0.77%와 0.54% 감소하였다. 이러한 결과는, 언어 인코더(110)의 두 번째 단계 모듈(112)이 시각 인코더(120)의 입력 모듈(121) 내 SCA로부터 시각인지 언어 특징을 획득하는 것이 성능상 보다 유리함을 시사한다.

[0120] 다음으로 언어 인코더(110)의 MVLA 모듈과 관련하여, 표 1을 참고하면 MVLA가 없는 경우 mIoU 및 oIoU 점수는 각각 1.34% 및 0.78% 감소하였다. 또한 MVLA가 멀티스케일(MS) 시각적 토큰 대신 단일스케일(single-scale; SS) 시각적 토큰을 사용하는 경우에, mIoU 및 oIoU 점수는 각각 0.52% 및 0.32% 감소하였다. 이러한 결과는, 다양한 시각적 정보와 함께 멀티스케일 시각적 토큰을 사용하는 것이 언어 특징의 강인함(robustness)을 향상시키는 데에 보다 유리함을 시사한다.

[0121] 한편 BERT와 동일하게 단일 단계 언어 인코딩 프로세스로 구성되어 교차 인지 프로세스를 수행하지 않는 모델의 평가를 위해 SCA와 MVLA를 모두 제거하였으며, 표 1을 참고하면 이 경우 mIoU 및 oIoU 점수가 각각 1.65% 및 1.31% 감소하여 성능이 상당히 저하됨을 알 수 있다. 이러한 결과는, 각 모달리티의 상호 이익을 위한 교차인지 이중 인코딩 구조가 멀티 모달리티의 특징들을 정렬하는 데에 훨씬 효과적임을 시사한다.

[0122] 다음으로 시각 인코더(120)의 LMHA 모듈과 관련하여 표 2를 참고하면, 먼저 사후 LMHA를 제거한 경우에 mIoU 및 oIoU 점수는 각각 1.39% 및 2.62% 감소하여 가장 큰 성능 감소를 나타낸다. 이러한 결과는 다음 시각 인코딩 단계 및 언어 단계 모두에 전달되는 다운 샘플링 된 시각 특징에 풍부한 컨텍스트 정보를 전파(propagate)하는 것이 효과적임을 시사한다.

[0123] 한편 멀티헤드 어텐션 대신 싱글헤드 어텐션을 적용한 경우에 mIoU 및 oIoU 점수는 각각 0.51% 및 0.64% 감소하였다. 이러한 결과는 다양한 관점에서 수집된 풍부한 컨텍스트 정보가 시각 특징에 유리함을 시사한다.

표 3

[0124]

Method	RefCOCO			RefCOCO+			G-Ref		#Params
	val	test A	test B	val	test A	test B	val	test	
RRN	55.33	57.26	53.93	39.75	42.15	36.11	-	-	-

MAttNet	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61	-
BRINet	60.98	62.99	59.21	48.17	52.32	42.11	-	-	241.2
CMPC	61.36	64.53	59.64	49.56	53.44	43.23	-	-	118.7
LSCM	61.47	64.99	59.55	49.34	53.12	43.50	-	-	127.9
MCN	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40	-
EFN	62.76	65.69	59.67	51.50	55.24	43.01	-	-	232.8
LTS	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25	-
VLT	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65	-
ReSTR	67.22	69.30	64.45	55.78	60.44	48.27	-	-	122.9
CRIS	70.47	73.18	66.10	62.27	68.08	53.68	59.87	60.36	161.0
LAVT	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09	227.7
MECS-S	**72.02**	**74.92**	**67.59**	**61.99**	**67.49**	**53.22**	**62.25**	**63.58**	**151.5**
MECS-B	**72.90**	**75.92**	**69.54**	**62.94**	**69.29**	**55.14**	**63.42**	**64.23**	**207.6**

[0125] 다음으로 본 개시의 언어지시 영역 영상분할 방법의 성능을 종래의 최고 수준 모델(state-of-the-art; SOTA)과 비교하였으며, 성능 비교의 정량적 결과는 표 3에서, 그리고 성능 비교의 정성적 결과는 도 4 및 도 5에서 각각 도시한다.

[0126] 성능 비교에서는 서로 다른 크기의 시각 모델(Swin-S 및 Swin-B(Liu et al. 2021))을 각각 포함하는 두 가지 버전의 언어지시 영역 영상분할 방법(이하 각각 “MECS-S” 및 “MECS-B”)이 사용되었다. 성능 비교의 결과, 본 개시에 따른 언어지시 영역 영상분할 방법은 모든 데이터셋에 대해 종래의 모든 최고수준 모델들을 능가하는 것으로 나타났다.

[0127] 먼저 정량적 성능 비교를 살펴보면, RefCOCO 및 RefCOCO+ 데이터셋에서 SOTA인 LAVT와 비교하여, 본 개시에 따른 언어지시 영역 영상분할 방법(MECS-B)은 매개변수가 적음에도 불구하고 더 높은 성능을 보였다. 나아가, RefCOCO에서 본 개시에 따른 언어지시 영역 영상분할 방법(MECS-S)은 각 항목에서 각각 1.55%, 1.74% 및 1.49%로 다른 SOTA인 CRIS를 능가하였다.

[0128] 또한 매개변수 복잡도가 상대적으로 작은 본 개시에 따른 언어지시 영역 영상분할 방법(MECS-S)은 G-Ref에서 두 번째로 좋은 성능결과를 나타냈다. 이러한 결과는, 본 개시에 따른 언어지시 영역 영상분할 방법이 멀티 모달리티 특징들을 효과적으로 융합하며 이전의 일반적인 구조보다 길고 복잡한 언어 표현의 의미를 보다 정확히 이해함을 나타냅니다.

[0129] 다음으로 정성적 성능 비교를 살펴보면, 먼저 도 4는 다른 SOTA 모델들인 VLT와 LAVT의 분할 결과를 본 개시에 따른 언어지시 영역 영상분할 방법의 분할 결과와 비교하여 도시한다. VLT는 싱글모달리티 인코더에서 특징 추출 후 추가적인 융합 모듈을 사용하는 대표적인 모델이고, LAVT는 언어인지 시각 트랜스포머와 조기 융합을 사용한 또 다른 대표적인 모델이다.

[0130] 도 4를 참고하면, 복잡한 표현과 밀집된 객체들을 포함하는 복잡한 영상에 대한 예제 (a)와 (b)에서, 본 개시에 따른 언어지시 영역 영상분할 방법은 종래의 두 SOTA 모델들에 비해 대상 영역을 더 정확하게 하이라이트 하였다. 이러한 결과는, 본 개시에 따른 언어지시 영역 영상분할 방법이 시각 및 언어 특징의 컨텍스트 정보를 풍부하게 하여 멀티 모달리티 특징 정렬을 향상시킬 수 있음을 시사한다.

[0131] 모호한 표현을 포함하는 다른 네 가지 예제의 경우, 본 개시에 따른 언어지시 영역 영상분할 방법은 언어표현에 해당하는 대상을 올바르게 분할한 반면, 다른 SOTA 모델들은 언어표현에 해당하는 대상을 잘못 예측하고 영역을 불확실하게 분할하였다.

[0132] 마찬가지로 도 5를 참고하면, 동일한 대상을 설명하는 다양한 언어표현이 주어졌을 때 본 개시에 따른 언어지시 영역 영상분할 방법은 대상을 일관되게 예측한 반면, 종래의 SOTA 모델은 LAVT는 일관되지 않게 대상을 예측하였다. 이러한 결과는, 본 개시에 따른 언어지시 영역 영상분할 방법이 언어 특징의 강인함을 강화하여 다양한 언어 표현을 처리할 수 있음을 시사한다.

[0133] 도 6은 본 개시의 몇몇 실시예에 따른 언어지시 영역 영상분할 방법에 대한 순서도이다.

[0134] 도 6을 참고하면, 먼저 본 개시에 따른 언어지시 영역 영상분할 장치(100)는 언어표현 및 영상의 쌍을 입력받을 수 있다(S110).

[0135] 다음으로, 본 개시에 따른 언어지시 영역 영상분할 장치(100)는 복수의 인코딩 단계들 각각에서, 선행하는 인코딩 단계로부터 획득되는 영상에 관한 시각 특징을 기초로 언어표현에 관한 언어 특징을 추출할 수 있다(S120).

[0136] 다음으로, 본 개시에 따른 언어지시 영역 영상분할 장치(100)는 복수의 인코딩 단계들 각각에서, 선행하는 인코딩 단계로부터 획득되는 언어 특징을 기초로 시각 특징을 추출할 수 있다(S130).

[0137] 다음으로, 먼저 본 개시에 따른 언어지시 영역 영상분할 장치(100)는 언어 특징 및 시각 특징 중 적어도 일부를 기초로, 영상에서 언어표현과 대응되는 대상 영역을 분할할 수 있다(S140).

[0138] 도 7은 본 개시의 몇몇 실시예에 따른 언어지시 영역 영상분할 방법을 제공하는 컴퓨팅 장치를 나타낸 블록도이다.

[0139] 여기서 언어지시 영역 영상분할 방법을 제공하는 컴퓨팅 장치(10)는, 전술한 언어지시 영역 영상분할 장치(100)이거나, 또는 언어지시 영역 영상분할 방법을 제공하기 위해 언어지시 영역 영상분할 장치(100)와 통신적으로 연결되는 임의의 단말(미도시)일 수 있다. 다만 이에 한정되는 것은 아니다.

[0140] 도 7을 참조하면, 본 개시에 따른 컴퓨팅 장치(10)는 하나 이상의 프로세서(11), 프로세서(11)에 의하여 수행되는 프로그램을 로드하는 메모리(12), 프로그램 및 각종 데이터를 저장하는 스토리지(13), 및 통신 인터페이스(14)를 포함할 수 있다. 다만, 상술한 구성 요소들은 본 개시에 따른 컴퓨팅 장치(10)를 구현하는데 있어서 필수적인 것은 아니어서, 컴퓨팅 장치(10)는 위에서 열거된 구성요소들 보다 많거나, 또는 적은 구성요소들을 가질 수 있다. 예컨대 컴퓨팅 장치(10)는 출력부 및/또는 입력부(미도시)를 더 포함하거나, 또는 스토리지(13)가 생략될 수도 있다.

[0141] 프로그램은 메모리(12)에 로드될 때 프로세서(11)로 하여금 본 개시의 다양한 실시예에 따른 방법/동작을 수행하게끔 하는 명령어들(instructions)을 포함할 수 있다. 즉, 프로세서(11)는 명령어들을 실행함으로써, 본 개시의 다양한 실시예에 따른 방법/동작들을 수행할 수 있다. 프로그램은 기능을 기준으로 묶인 일련의 컴퓨터 판독 가능 명령어들로 구성되고, 프로세서에 의해 실행되는 것을 가리킨다.

[0142] 프로세서(11)는 컴퓨팅 장치(10)의 각 구성의 전반적인 동작을 제어한다. 프로세서(11)는 CPU(Central Processing Unit), MPU(Micro Processor Unit), MCU(Micro Controller Unit), GPU(Graphic Processing Unit) 또는 본 개시의 기술 분야에 잘 알려진 임의의 형태의 프로세서 중 적어도 하나를 포함하여 구성될 수 있다. 또한, 프로세서(11)는 본 개시의 다양한 실시예들에 따른 방법/동작을 실행하기 위한 적어도 하나의 애플리케이션 또는 프로그램에 대한 연산을 수행할 수 있다.

[0143] 메모리(12)는 각종 데이터, 명령 및/또는 정보를 저장한다. 메모리(12)는 본 개시의 다양한 실시예들에 따른 방법/동작을 실행하기 위하여 스토리지(13)로부터 하나 이상의 프로그램을 로드할 수 있다. 메모리(12)는 RAM과 같은 휘발성 메모리로 구현될 수 있을 것이나, 본 개시의 기술적 범위는 이에 한정되지 않는다.

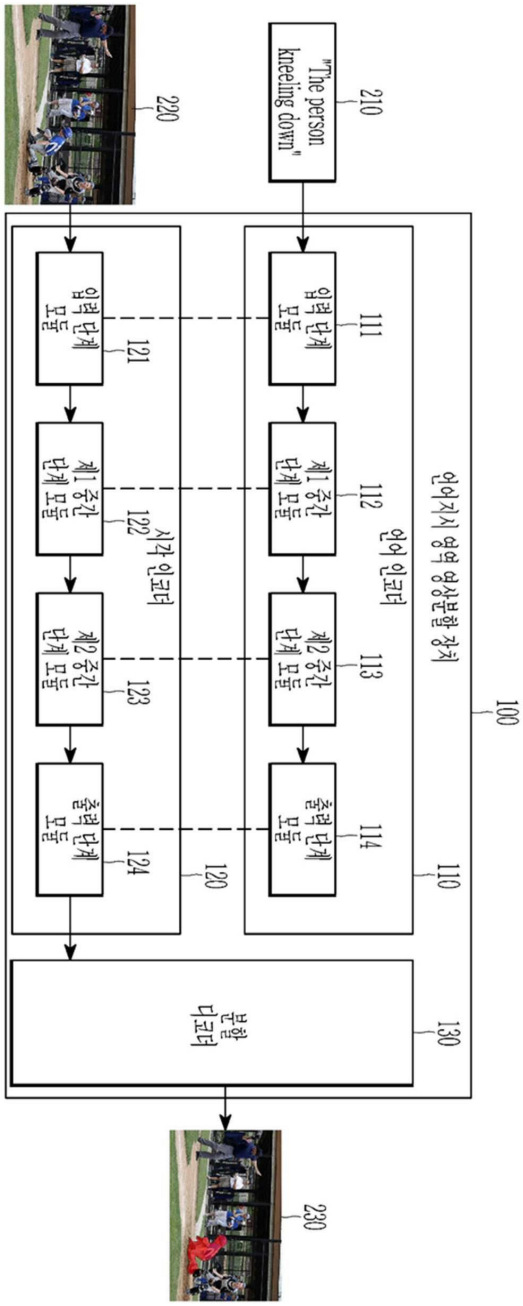
[0144] 스토리지(13)는 프로그램을 비임시적으로 저장할 수 있다. 스토리지(13)는 ROM(Read Only Memory), EPROM(Erasable Programmable ROM), EEPROM(Electrically Erasable Programmable ROM), 플래시 메모리 등과 같은 비휘발성 메모리, 하드 디스크, 착탈형 디스크, 또는 본 개시가 속하는 기술 분야에서 잘 알려진 임의의 형태의 컴퓨터로 읽을 수 있는 기록 매체를 포함하여 구성될 수 있다. 통신 인터페이스(14)는 유/무선 통신 모듈일 수 있다.

[0145] 이상에서 설명한 본 개시의 실시예는 장치 및 방법을 통해서만 구현이 되는 것은 아니며, 본 개시의 실시예의 구성에 대응하는 기능을 실현하는 프로그램 또는 그 프로그램이 기록된 기록 매체를 통해 구현될 수도 있다.

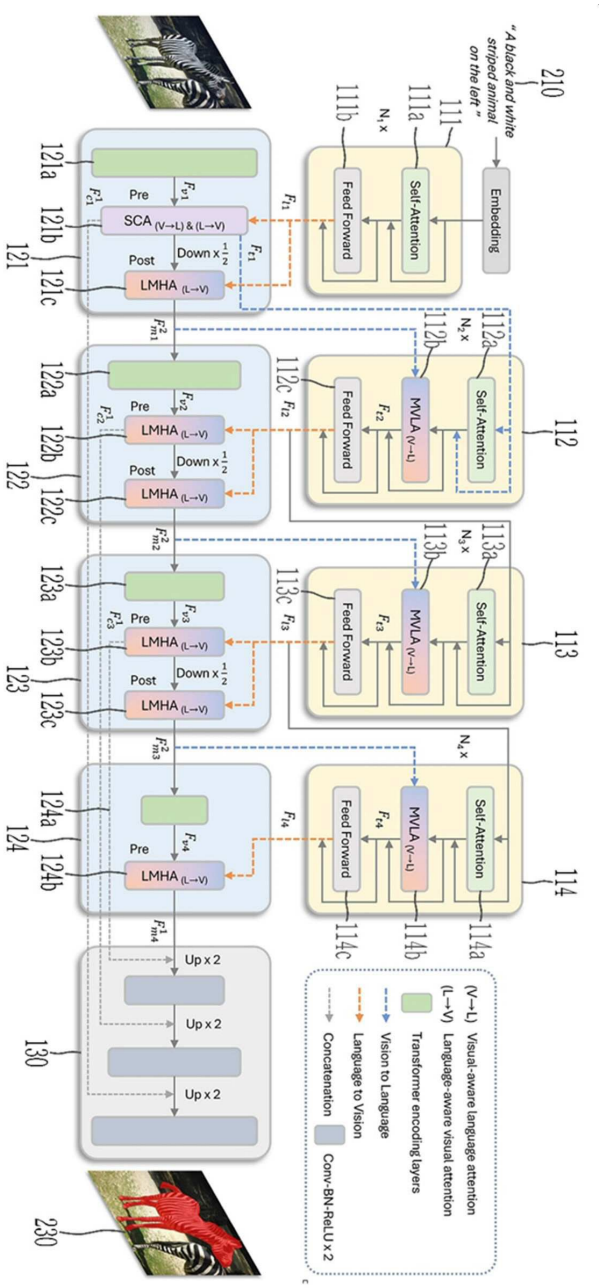
[0146] 이상에서 본 개시의 실시예에 대하여 상세하게 설명하였지만 본 개시의 권리범위는 이에 한정되는 것은 아니고 다음의 청구범위에서 정의하고 있는 본 개시의 기본 개념을 이용한 당업자의 여러 변형 및 개량 형태 또한 본 개시의 권리범위에 속하는 것이다.

도면

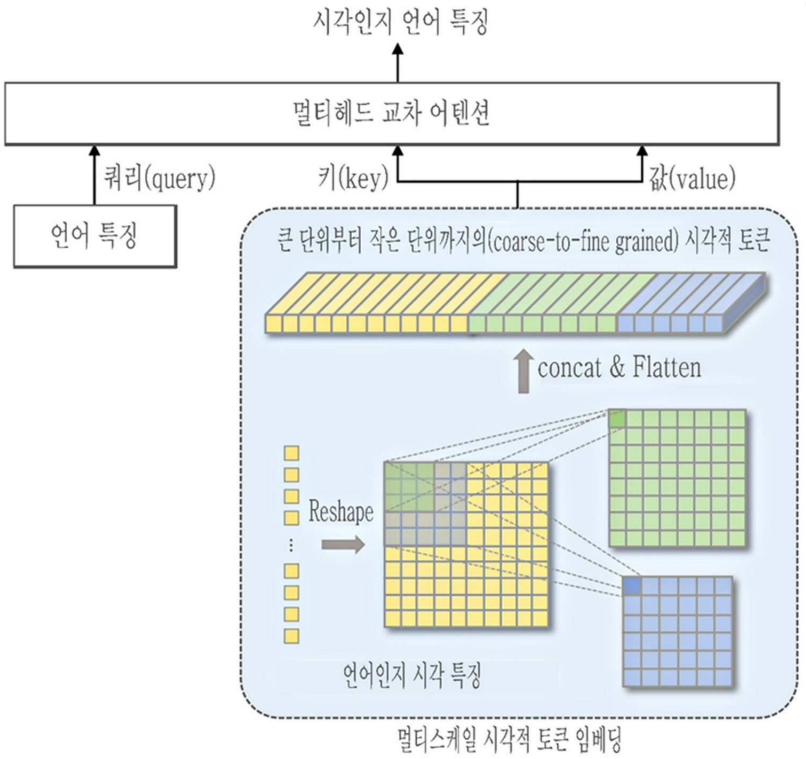
도면1


"The person kneeling down"
210
220
입력 단계 모듈
111
제1 중간 단계 모듈
112
제2 중간 단계 모듈
113
출력 단계 모듈
114
언어 인코더
110
입력 단계 모듈
121
제1 중간 단계 모듈
122
제2 중간 단계 모듈
123
출력 단계 모듈
124
시각 인코더
120
분할 디코더
130
언어지시 영역 영상분할 장치
100
230

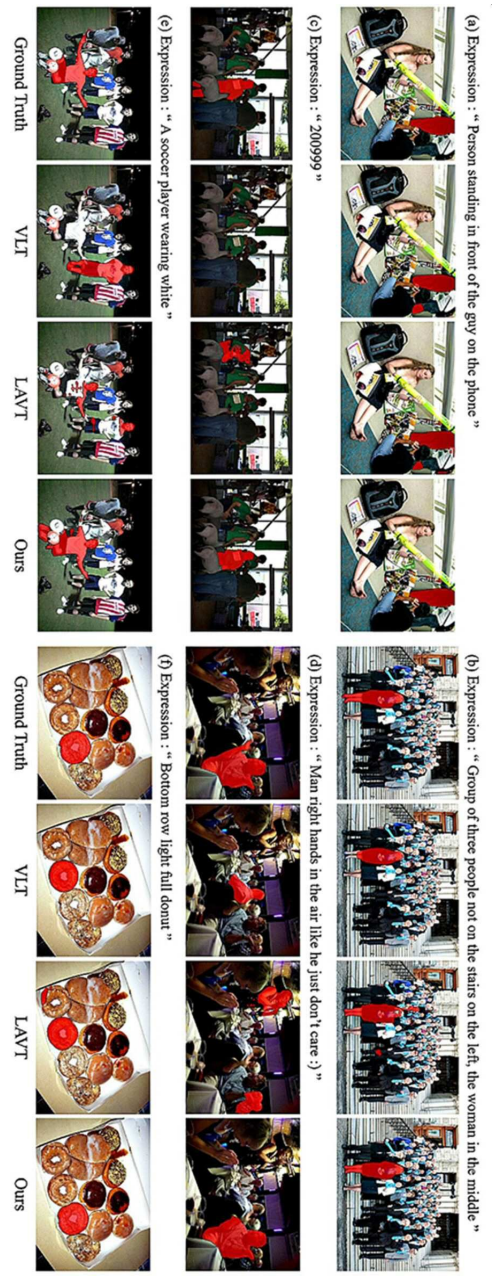
도면2


"A black and white striped animal on the left."
210
Embedding
111
111a
111b
Self-Attention
Feed Forward
112
112a
112b
112c
MVLA (V→L)
113
113a
113b
113c
114
114a
114b
114c
121
121a
121b
121c
Pre
SCA (V→L) & (L→V)
Post
Down
LMHA (L→V)
122
122a
122b
122c
123
123a
123b
123c
124
124a
124b
Up x 2
130
230
(V→L) Visual-aware language attention
(L→V) Language-aware visual attention
Transformer encoding layers
Vision to Language
Language to Vision
Concatenation
Conv-BN-ReLU x 2

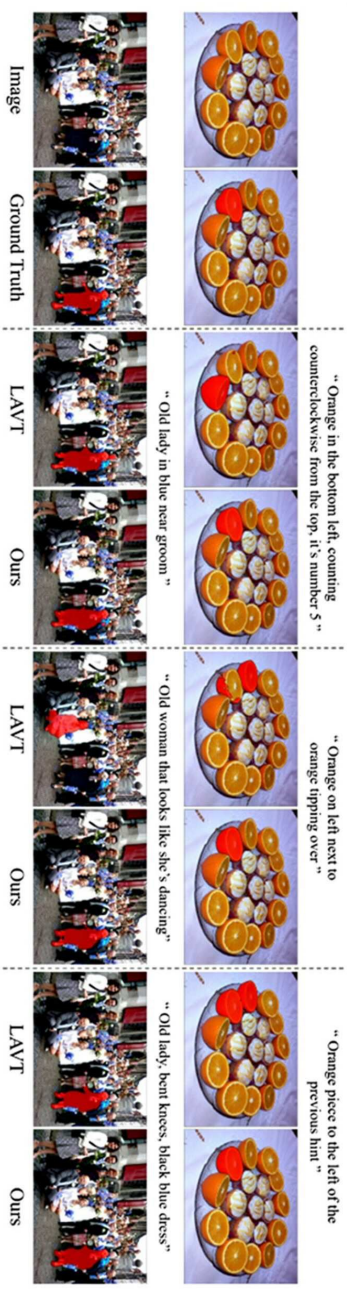
도면3


시각인지 언어 특징
멀티헤드 교차 어텐션
쿼리(query)
키(key)
값(value)
언어 특징
큰 단위부터 작은 단위까지의(coarse-to-fine grained) 시각적 토큰
concat & Flatten
Reshape
언어인지 시각 특징
멀티스케일 시각적 토큰 임베딩

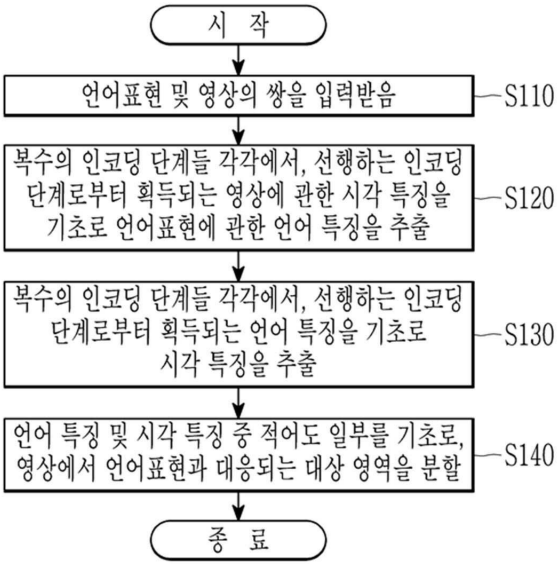
도면4


(a) Expression : " Person standing in front of the guy on the phone "
(b) Expression : " Group of three people not on the stairs on the left, the woman in the middle "
(c) Expression : " 200999 "
(d) Expression : " Man right hands in the air like he just don't care :) "
(e) Expression : " A soccer player wearing white "
(f) Expression : " Bottom row light full donut "
Ground Truth
VLT
LAVT
Ours

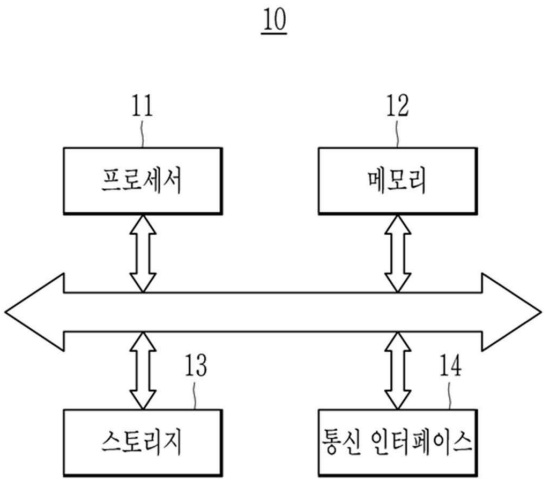
도면5


Image
Ground Truth
LAVT
Ours
LAVT
Ours
LAVT
Ours
" Orange in the bottom left, counting counterclockwise from the top, it's number 5 "
" Old lady in blue near groom "
" Orange on left next to orange tipping over "
" Old woman that looks like she's dancing"
" Orange piece to the left of the previous hint "
" Old lady, bent knees, black blue dress"

도면6


시 작
언어표현 및 영상의 쌍을 입력받음
S110
복수의 인코딩 단계들 각각에서, 선행하는 인코딩 단계로부터 획득되는 영상에 관한 시각 특징을 기초로 언어표현에 관한 언어 특징을 추출
S120
복수의 인코딩 단계들 각각에서, 선행하는 인코딩 단계로부터 획득되는 언어 특징을 기초로 시각 특징을 추출
S130
언어 특징 및 시각 특징 중 적어도 일부를 기초로, 영상에서 언어표현과 대응되는 대상 영역을 분할
S140
종 료

도면7


10
11
프로세서
12
메모리
13
스토리지
14
통신 인터페이스