



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0120219
(43) 공개일자 2023년08월17일

(51) 국제특허분류(Int. Cl.)

G06N 3/08 (2023.01) G06F 16/783 (2019.01)
G06F 17/16 (2006.01) G06F 40/268 (2020.01)
G06F 40/284 (2020.01) G06V 10/762 (2022.01)
G06V 10/774 (2022.01) G06V 30/18 (2022.01)
G06V 30/24 (2022.01)

(52) CPC특허분류

G06N 3/08 (2023.01)
G06F 16/7837 (2019.01)

(21) 출원번호 10-2022-0016380

(22) 출원일자 2022년02월08일

심사청구일자 2022년02월08일

(71) 출원인

이화여자대학교 산학협력단

서울특별시 서대문구 이화여대길 52 (대현동, 이화여자대학교)

(72) 발명자

강제원

서울특별시 마포구 독막로 145, 104동 904호(창전동, 서강쌍용예가아파트)

이주희

인천광역시 부평구 길주남로10번길 21, 114동 1703호(부평동, 래미안부평)

(74) 대리인

특허법인(유한)아이시스

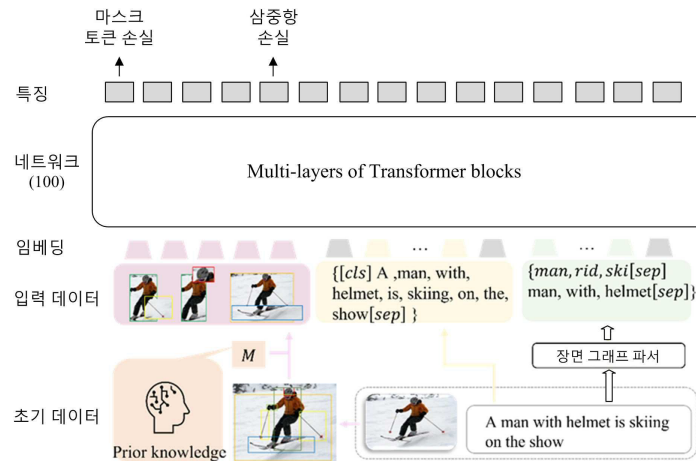
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 객체 간의 상관 관계에 기반한 비전-언어 사전학습 방법

(57) 요약

객체 간의 상관 관계에 기반한 비전-언어 사전학습 방법은 모델 생성 장치가 영상과 문장 쌍으로 구성된 데이터는 입력받는 단계, 상기 모델 생성 장치가 상기 영상에 대한 객체 탐지를 수행하여 서로 클래스가 다른 객체들의 영역을 구분하고, 상기 객체들의 위치 및 클래스 발생 빈도에 따른 통합적 객체에 대한 특징 벡터를 추출하는 단계, 상기 모델 생성 장치가 상기 문장에 대한 토큰을 추출하고, 상기 영상에 대한 장면 그래프 분석을 이용하여 생성된 장면 그래프로부터 관계 구문에 대한 토큰화된 텍스트 시퀀스를 추출하는 단계 및 상기 모델 생성 장치가 상기 영상에서 추출한 특징 벡터, 상기 문장에서 추출한 토큰 및 상기 토큰화된 텍스트 시퀀스를 임베딩하여 비전-언어 사전학습 모델에 입력하여 사전 학습을 수행하는 단계를 포함한다.

대표도 - 도3



(52) CPC특허분류

G06F 17/16 (2013.01)

G06F 40/268 (2020.01)

G06F 40/284 (2020.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711126409
과제번호	2020-0-01460-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성(R&D)
연구과제명	Simulated Reality 콘텐츠 구현을 위한 핵심기술 연구 및 인력양성
기 여 율	1/2
과제수행기관명	이화여자대학교산학협력단
연구기간	2021.01.01 ~ 2021.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711140190
과제번호	2019R1C1C1010249
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)(R&D)
연구과제명	고화질 비디오의 장기간 예측과 해석 가능한 딥러닝 모델 기반의 비디오 신호처리
알고리즘 개발	
기 여 율	1/2
과제수행기관명	이화여자대학교
연구기간	2021.03.01 ~ 2022.02.28

명세서

청구범위

청구항 1

모델 생성 장치가 영상과 문장 쌍으로 구성된 데이터는 입력받는 단계;

상기 모델 생성 장치가 상기 영상에 대한 객체 탐지를 수행하여 서로 클래스가 다른 객체들의 영역을 구분하고, 상기 객체들의 위치 및 클래스 발생 빈도에 따른 통합적 객체에 대한 특징 벡터를 추출하는 단계;

상기 모델 생성 장치가 상기 문장에 대한 토큰을 추출하고, 상기 영상에 대한 장면 그래프 분석을 이용하여 생성된 장면 그래프로부터 관계 구문에 대한 토큰화된 텍스트 시퀀스를 추출하는 단계; 및

상기 모델 생성 장치가 상기 영상에서 추출한 특징 벡터, 상기 문장에서 추출한 토큰 및 상기 토큰화된 텍스트 시퀀스를 임베딩하여 비전-언어 사전학습 모델에 입력하여 사전 학습을 수행하는 단계를 포함하되,

상기 비전-언어 사전학습 모델은 트랜스포머 블록들로 구성된 다수의 계층을 포함하는 객체 간의 상관 관계에 기반한 비전-언어 사전학습 방법.

청구항 2

제1항에 있어서,

상기 특징 벡터를 추출하는 단계는

상기 모델 생성 장치가 상기 객체들의 영역 사이의 거리를 기준으로 거리 행렬을 생성하는 단계;

상기 모델 생성 장치가 사전에 구축한 상관관계 행렬에서 상기 객체들의 카테고리를 기준으로 부분 행렬을 추출하는 단계;

상기 모델 생성 장치가 상기 거리 행렬과 상기 부분 행렬을 결합하여 선행 행렬을 생성하는 단계; 및

상기 모델 생성 장치가 상기 선행 행렬을 기준으로 상관 관계가 큰 상위 N개의 객체 쌍에 대한 상기 특징 벡터를 생성하는 단계를 포함하되,

상기 상관관계 행렬을 영상 데이터 세트에서 하나의 이미지에 동시에 등장하는 객체의 빈도를 기준으로 생성되는 객체 간의 상관 관계에 기반한 비전-언어 사전학습 방법.

청구항 3

제2항에 있어서,

상기 상관관계 행렬은 시각적 상관관계 행렬 및 언어적 상관관계 행렬로 구성되고,

상기 선행 행렬은 아래 수식으로 표현되는 객체 간의 상관 관계에 기반한 비전-언어 사전학습 방법.

$$M = \lambda V + \theta R + (1 - \lambda)S$$

(M은 상관관계 행렬, V는 시각적 상관관계 부분 행렬, S는 언어적 상관관계 부분 행렬, λ 및 θ 는 값 조절을 위한 상수값, R는 상기 거리 행렬임)

청구항 4

제1항에 있어서,

상기 모델 생성 장치가

상기 장면 그래프로부터 <주어-관계-목적어> 성분을 추출하여 상기 관계 구문을 추출하는 객체 간의 상관 관계에 기반한 비전-언어 사전학습 방법.

청구항 5

제1항에 있어서,

상기 모델 생성 장치는

상기 영상에서 검출한 객체, 상기 관계 구문 및 상기 문장에서 마스킹되지 않은 토큰 정보를 이용하여 상기 문장에서 마스킹된 토큰을 예측하도록 상기 비전-언어 사전학습 모델을 학습하는 객체 간의 상관 관계에 기반한 비전-언어 사전학습 방법.

청구항 6

제1항에 있어서,

상기 모델 생성 장치는

상기 영상에서 추출한 특징 벡터와 상기 문장에서 추출한 언어 벡터 쌍을 교환하여 오답 쌍을 생성하고, 정답 쌍과의 유사도를 비교하여 유사도가 높을수록 가깝게 표현하도록 상기 비전-언어 사전학습 모델을 학습하는 객체 간의 상관 관계에 기반한 비전-언어 사전학습 방법.

청구항 7

영상과 문장 쌍으로 구성된 데이터 세트를 입력받는 입력장치;

트랜스포머 블록들로 구성된 다수의 계층을 포함하는 비전-언어 사전학습 모델을 저장하는 저장장치; 및

상기 데이터 세트 각각에 대하여 상기 영상에 대한 객체 탐지를 수행하여 서로 클래스가 다른 객체들의 영역을 구분하고, 상기 객체들의 위치 및 클래스 발생 빈도에 따른 통합적 객체에 대한 특징 벡터를 추출하고, 상기 문장에 대한 토큰을 추출하고, 상기 영상에 대한 장면 그래프 분석을 이용하여 생성된 장면 그래프로부터 관계 구문에 대한 토큰화된 텍스트 시퀀스를 추출하고, 추출한 상기 특징 벡터, 상기 문장에서 추출한 토큰 및 상기 토큰화된 텍스트 시퀀스를 임베딩하고, 상기 비전-언어 사전학습 모델에 입력하여 사전 학습을 수행하는 연산장치를 포함하되,

상기 비전-언어 사전학습 모델은 트랜스포머 블록들로 구성된 다수의 계층을 포함하는 객체 간의 상관 관계에 기반한 비전-언어 사전학습 모델을 구축하는 모델 생성 장치.

청구항 8

제7항에 있어서,

상기 연산장치는

상기 객체들의 영역 사이의 거리를 기준으로 생성된 거리 행렬 및 사전에 구축한 상관관계 행렬에서 상기 객체들의 카테고리를 기준으로 생성한 부분 행렬을 결합하여 선행 행렬을 생성하고, 상기 선행 행렬을 기준으로 상관 관계가 큰 상위 N개의 객체 쌍에 대한 상기 특징 벡터를 생성하되,

상기 상관관계 행렬을 영상 데이터 세트에서 하나의 이미지에 동시에 등장하는 객체의 빈도를 기준으로 생성되는 객체 간의 상관 관계에 기반한 비전-언어 사전학습 모델을 구축하는 모델 생성 장치.

청구항 9

제8항에 있어서,

상기 상관관계 행렬은 시각적 상관관계 행렬 및 언어적 상관관계 행렬로 구성되고, 상기 선행 행렬은 아래 수식으로 표현되는 객체 간의 상관 관계에 기반한 비전-언어 사전학습 모델을 구축하는 모델 생성 장치.

$$M = \lambda V + \theta R + (1 - \lambda)S$$

(M은 상관관계 행렬, V는 시각적 상관관계 부분 행렬, S는 언어적 상관관계 부분 행렬, λ 및 θ 는 값 조절을 위한 상수값, R는 상기 거리 행렬임)

청구항 10

제7항에 있어서,

상기 연산장치가 상기 장면 그래프로부터 <주어-관계-목적어> 성분을 추출하여 상기 관계 구문을 추출하는 객체 간의 상관 관계에 기반한 비전-언어 사전학습 모델을 구축하는 모델 생성 장치.

발명의 설명

기술 분야

[0001] 이하 설명하는 기술은 비전-언어 사전학습 기법에 관한 것이다.

배경 기술

[0002] 비전-언어 사전학습(vision-language pre-training, VLP)은 대규모 이미지-텍스트 데이터를 통해 인코딩과 디코딩을 위한 교차 모달(cross modal) 표현을 학습하는 것을 목적으로 한다. 교차 모달 표현을 학습한 모델은 이미지 캡션, 질의 응답 태스크 등과 같은 애플리케이션에 대한 파인 튜닝(fine tuning)을 하여 특이적 애플리케이션에 대한 학습 소요시간 및 학습 데이터의 양을 줄이고 우수한 성능을 제공할 수 있다.

선행기술문헌

특허문헌

[0003] (특허문헌 0001) 한국등록특허 제10-2185777호

발명의 내용

해결하려는 과제

[0004] 종래 비전-언어 사전 학습은 단일 객체 입력을 토큰으로 사용하여 교차 모달 표현을 학습하기 때문에 복잡한 문장과 영상의 이해 및 표현 능력이 떨어진다. 이는 종래 사전 학습 기법은 단일 객체 입력을 사용하여 객체 사이의 통계적 상관 관계 내지 구조적 상관관계를 반영하지 못하였기 때문이다.

[0005] 이하 설명하는 기술은 객체 사이의 관계를 중심으로 시각 및 언어의 멀티모달 데이터의 공동 표현을 사전 학습하고자 한다.

과제의 해결 수단

[0006] 객체 간의 상관 관계에 기반한 비전-언어 사전학습 방법은 모델 생성 장치가 영상과 문장 쌍으로 구성된 데이터는 입력받는 단계, 상기 모델 생성 장치가 상기 영상에 대한 객체 탐지를 수행하여 서로 클래스가 다른 객체들의 영역을 구분하고, 상기 객체들의 위치 및 클래스 발생 빈도에 따른 통합적 객체에 대한 특징 벡터를 추출하는 단계, 상기 모델 생성 장치가 상기 문장에 대한 토큰을 추출하고, 상기 영상에 대한 장면 그래프 분석을 이용하여 생성된 장면 그래프로부터 관계 구문에 대한 토큰화된 텍스트 시퀀스를 추출하는 단계 및 상기 모델 생성 장치가 상기 영상에서 추출한 특징 벡터, 상기 문장에서 추출한 토큰 및 상기 토큰화된 텍스트 시퀀스를 임베딩하여 비전-언어 사전학습 모델에 입력하여 사전 학습을 수행하는 단계를 포함한다.

[0007] 객체 간의 상관 관계에 기반한 비전-언어 사전학습 모델을 구축하는 모델 생성 장치는 영상과 문장 쌍으로 구성된 데이터 세트를 입력받는 입력장치, 트랜스포머 블록들로 구성된 다수의 계층을 포함하는 비전-언어 사전학습 모델을 저장하는 저장장치 및 상기 데이터 세트 각각에 대하여 상기 영상에 대한 객체 탐지를 수행하여 서로 클래스가 다른 객체들의 영역을 구분하고, 상기 객체들의 위치 및 클래스 발생 빈도에 따른 통합적 객체에 대한 특징 벡터를 추출하고, 상기 문장에 대한 토큰을 추출하고, 상기 영상에 대한 장면 그래프 분석을 이용하여 생성된 장면 그래프로부터 관계 구문에 대한 토큰화된 텍스트 시퀀스를 추출하고, 추출한 상기 특징 벡터, 상기 문장에서 추출한 토큰 및 상기 토큰화된 텍스트 시퀀스를 임베딩하고, 상기 비전-언어 사전학습 모델에 입력하여 사전 학습을 수행하는 연산장치를 포함한다.

발명의 효과

[0008] 이하 설명하는 기술은 단일 객체가 아닌 영상 및 문장에 존재하는 객체들 사이의 상관관계를 반영하여 비전-언어 사전 학습을 수행한다. 따라서, 이하 설명하는 기술은 복잡한 문장이나 영상에 대한 표현을 학습하기 용이하다.

도면의 간단한 설명

- [0009] 도 1은 종래 비전-언어 사전학습 모델에 대한 예이다.
 도 2는 객체들이 포함된 이미지에 대한 예이다.
 도 3은 비전-언어 사전학습 모델에 대한 예이다.
 도 4는 영상에서 등장하는 규칙성을 설명하는 예이다.
 도 5는 지식 행렬을 이용하여 영상에서 통합된 객체를 추출하는 과정의 예이다.
 도 6은 문장에서 관계 구문을 추출하는 과정의 예이다.
 도 7은 비전-언어 사전학습 모델의 동작 과정의 예이다.
 도 8은 비전-언어 사전학습 모델을 구축하는 모델 생성 장치에 대한 예이다.

발명을 실시하기 위한 구체적인 내용

- [0010] 이하 설명하는 기술은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세하게 설명하고자 한다. 그러나, 이는 이하 설명하는 기술을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 이하 설명하는 기술의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.
- [0011] 제1, 제2, A, B 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 해당 구성요소들은 상기 용어들에 의해 한정되지는 않으며, 단지 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 이하 설명하는 기술의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. 및/또는 이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.
- [0012] 본 명세서에서 사용되는 용어에서 단수의 표현은 문맥상 명백하게 다르게 해석되지 않는 한 복수의 표현을 포함하는 것으로 이해되어야 하고, "포함한다" 등의 용어는 설명된 특징, 개수, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함을 의미하는 것이지, 하나 또는 그 이상의 다른 특징들이나 개수, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 배제하지 않는 것으로 이해되어야 한다.
- [0013] 도면에 대한 상세한 설명을 하기에 앞서, 본 명세서에서의 구성부들에 대한 구분은 각 구성부가 담당하는 주기능 별로 구분한 것에 불과함을 명확히 하고자 한다. 즉, 이하에서 설명할 2개 이상의 구성부가 하나의 구성부로 합쳐지거나 또는 하나의 구성부가 보다 세분화된 기능별로 2개 이상으로 분화되어 구비될 수도 있다. 그리고 이하에서 설명할 구성부 각각은 자신이 담당하는 주기능 이외에도 다른 구성부가 담당하는 기능 중 일부 또는 전부의 기능을 추가적으로 수행할 수도 있으며, 구성부 각각이 담당하는 주기능 중 일부 기능이 다른 구성부에 의해 전담되어 수행될 수도 있음은 물론이다.
- [0014] 또, 방법 또는 동작 방법을 수행함에 있어서, 상기 방법을 이루는 각 과정들은 문맥상 명백하게 특정 순서를 기재하지 않은 이상 명기된 순서와 다르게 일어날 수 있다. 즉, 각 과정들은 명기된 순서와 동일하게 일어날 수도 있고 실질적으로 동시에 수행될 수도 있으며 반대의 순서대로 수행될 수도 있다.
- [0015] 이하 설명하는 기술은 비전-언어 사전학습 기법에 관한 것이다. 비전-언어 사전학습 모델은 종전 BERT(Bidirectional Encoder Representations from Transformers), GPT(Generative Pre-Training) 등과 같은 자연어 사전학습 모델로부터 영감을 받아 비전-언어에 대한 교차 모달 표현을 학습하도록 개발된 모델이다.
- [0016] 이하 모델 생성 장치 내지 모델 학습 장치가 비전-언어 사전학습을 수행한다고 설명한다. 모델 생성 장치는 이미지 및 텍스트 처리가 가능한 컴퓨터 장치를 의미한다. 예컨대, 모델 생성 장치는 PC, 네트워크상의 서버, 스마트폰 기기, 전용 프로그램이 임베딩된 칩셋 등으로 구현될 수 있다.

- [0018] 도 1은 종래 비전-언어 사전학습 모델에 대한 예이다. 도 1은 비전-언어 사전학습 모델 중 하나에 대한 예이다 (<https://arxiv.org/abs/1909.11059>). 도 1은 인코더와 디코더가 통합된 모델(unified encoder-decoder)이고, 다수의 트랜스포머(transformer) 블록들로 구성된 다중 계층 구조의 모델이다.
- [0019] 입력 데이터는 이미지 및 문장(sentence)이다. 각 입력 데이터는 토큰(token)으로 처리되어 입력된다. 이미지는 N 개의 관심영역으로 구분되고, 모델은 개별 영역의 특징을 입력받는다. 모델 생성 장치는 입력 이미지에서의 객체 탐지를 통해 N 개의 관심 영역을 구분하여 시각 토큰을 생성할 수 있다. 모델 생성 장치는 해당 영역의 특징을 임베딩하여 모델에 입력한다. 문장은 자연어 처리 모델 BERT와 유사하게 마스크된(masked) 토큰들로 처리되어 입력된다.
- [0020] 모델 생성 장치는 이미지와 텍스트 쌍이 입력되면, (i) 객체 탐지 모델을 이용하여 이미지에서 N 개의 객체를 추출하고 해당 객체가 위치하는 영역의 특징을 임베딩하여 VLP 모델에 입력하고, (ii) 텍스트 입력은 BERT와 같이 토큰화되고, 위치 임베딩을 거쳐 VLP모델에 입력된다.
- [0021] 한편, 입력 토큰의 시작은 특수 토큰인 [CLS]로 표시하고, 시각과 언어 입력의 구분은 [SEP]로, 종결은 [STOP]으로 표시한다.
- [0022] 모델 생성 장치는 양방향(bidirectional) 및 시퀀스 투 시퀀스(seq2seq) 마스크 언어 예측 작업으로 학습을 진행할 수 있다. 양방향 마스크 언어 예측은 BERT에서의 MLM (masked language model)과 같이 입력된 시각 토큰과 마스크 되지 않은 언어 토큰을 이용하여 [MASK]를 예측하는 것이다. 시퀀스 투 시퀀스 마스크 언어 예측은 이미지 캡셔닝에서의 단방향 시퀀스 예측을 따라 시각 토큰과 마스크 되지 않은 언어 토큰 중 [MASK] 이전의 토큰만을 이용하여 [MASK] 토큰으로 대체된 텍스트 토큰을 예측한다.
- [0023] 도 1과 같은 종래 VLP 모델은 입력 이미지의 객체 탐지의 결과인 시각 토큰을 입력받고, 입력 문장은 토큰화되어 입력받는다. 따라서, 종래 모델은 단일 객체에 대한 토큰을 사용하여 교차 모달 표현을 학습한다.
- [0025] 도 2는 객체들이 포함된 이미지에 대한 예이다. 도 2는 인간이 다수의 객체들로 구성된 영상 및 언어를 인지하는 과정을 설명하기 위한 예이다.
- [0026] 인간이 영상 및 문장을 이해할 때의 가장 하위 단위는 단일 객체이며, 실제 인간의 영상 이해는 선행 지식에 기반하여 통합된 객체를 통해 이루어진다. 예를 들어 도 2는 '사람', '헬멧', '고글', '집', '스키', '나무', '하늘' 등과 같은 단일 객체로 구성된다. 인간은 실제 "헬멧과 고글을 쓴 스키 탄 사람" / "눈 위의 세명의 사람"과 같이 통합된 객체를 기반으로 영상을 인지하고 해석한다. 이때 객체를 통합하는 기준은 선행지식으로, 객체 사이의 동시 발생 확률 및 위치를 고려하여 "나무 위의 구름과 헬멧"과 같은 관련 없는 조합을 기준으로 영상을 이해하지 않는다.
- [0027] 나아가, 영상을 설명하는 두 개의 문장인 (i) "헬멧과 고글을 쓴 사람이 눈 위에 있다."와 (ii) "고글을 쓴 사람이 집 앞에 있고 하늘에는 구름이 있다."를 살펴보면, 전자의 표현이 자연스러운 설명이라고 느껴진다.
- [0028] 즉, 인간이 영상을 이해하는데 중요한 선행 지식은 객체 사이의 동시 발생확률과 영상에서의 객체의 위치라고 할 수 있다.
- [0030] 멀티모달(Multimodal) 데이터는 서로 다른 형태의 정보로 이루어져 뚜렷한 특성이 구분되는 데이터다. 비전-언어 사전 학습은 대용량 이미지-텍스트 데이터와 트랜스포머의 셀프어텐션(self-attention)을 통해 비전-언어의 공동 표현을 비지도 학습하는 방법이다.
- [0031] 이미지와 텍스트로 구성된 멀티모델 데이터에서 이미지-텍스트 검색은 탐색하는 이미지 검색과 텍스트 검색 두 가지 하위 작업으로 구분된다. 모델 생성 장치는 주어진 이미지-텍스트 쿼리 내에서 타겟과의 유사도를 비교하여 가장 유사한 상위 K 개의 검색 결과를 확인한다. 이미지 텍스트 검색은 유사도를 비교를 위해 비전-언어의 공동 표현에 크게 의존한다.
- [0032] 영상 질의응답은 이미지를 기반으로 주어진 텍스트 질문에 알맞은 텍스트 답변을 출력하는 것을 목표로 한다. 영상을 보고 질문에 상응하는 적절한 대답을 출력하기 위해 질문과 영상에 대한 높은 수준의 이해 및 추론 능력이 필요하다.

- [0033] 영상 캡셔닝은 입력된 이미지 내용에 대한 자연어 설명을 생성하는 작업이다. 입력 이미지와 이전에 생성된 연속된 단어를 이용하여 출력 문장을 생성한다. 이 경우 질문을 통해 영상에서 집중해야 하는 영역을 알 수 있는 영상 질의응답 과 달리 전반적인 영상에 대한 이해와 요약 및 서술하는 능력이 필요하다.
- [0035] 도 3은 비전-언어 사전학습 모델(100)에 대한 예이다. 비전-언어 사전학습 모델(100)은 객체 사이의 관계를 중심으로 시각 및 언어의 멀티모달 데이터의 공동 표현을 학습하는 것을 목표로 한다.
- [0036] 네트워크(100)는 종전의 비전-언어 사전학습 모델과 유사하게 다수의 트랜스포머 블록으로 구성되는 계층들로 구성될 수 있다. 모델 생성 장치는 입력 이미지 및 텍스트를 토큰화하고 네트워크(100)에 입력한다.
- [0037] 초기 데이터는 이미지와 텍스트 쌍이다.
- [0038] 입력 데이터는 종래 토큰과 선행 지식을 이용하여 초기 데이터를 분석한 정보로 구성된다.
- [0039] 텍스트 입력 토큰은 장면그래프 구문 분석과 토큰화를 통해 얻는다.
- [0040] 장면 그래프 구분 분석은 장면 그래프(scene graph)를 이용하여 영상에 나타나는 객체들의 관계를 분석하는 기법이다(Xu, P. et al., 2020, A survey of scene graph: Generation and application, IEEE Trans. Neural Netw. Learn. Syst.). 장면 그래프는 영상 내 객체(objects)와 속성(attributes) 및 객체들 간의 관계(relationships)에 대한 정보를 갖는 그래프이다.
- [0041] 장면 그래프에서 추출 가능한 <주어-관계-목적어> 형태의 관계 구문은 인간의 시각 및 언어 이해 체계와 유사한 형태의 구조이다.
- [0042] 관계 구문은 이미지의 구조적인 설명이 가능하고 또한 관계 구문의 조합으로 문장의 생성이 가능하다. 따라서, 장면 그래프는 이미지와 문장을 서로 연결하는 수단으로 활용될 수 있다.
- [0043] 텍스트 입력 토큰은 입력된 문장을 토큰화하여 얻은 토큰과 장면 그래프 구분 분석을 통해 얻은 <주어-관계-목적어> 형태의 관계 구문으로 구성된다. 모델 생성 장치는 BERT와 같은 모델에서 사용되는 토큰화 과정을 거쳐 생성된 토큰과 별도의 장면 그래프 구분 분석을 통해 생성한 관계 구문을 네트워크(100)에 입력한다. 모델 생성 장치는 장면 그래프 파서를 사용하여 입력 영상에서 장면 그래프를 생성할 수 있다.
- [0044] 도 3과 같이 최종 텍스트 입력 토큰은 문장의 시작과 끝을 알리는 [CLS], [SEP]와 같은 특수 토큰이 토큰화된 텍스트 시퀀스에 추가되어 구성될 수 있다.
- [0045] 네트워크(100)에서 사용하는 손실 함수에 대해서는 후술한다.
- [0047] 이미지 입력 토큰은 입력 이미지에 객체 탐지를 진행하여 출력된 객체를 이용한다. 출력된 객체는 선행 지식으로 생성한 지식 행렬 M에 기반하여 통합된 객체로 임베딩하여 벡터로 표현될 수 있다.
- [0048] 모델 생성 장치는 선행 지식을 이용하여 이미지 입력 토큰을 생성한다. 선행 지식에 대하여 설명한다. 선행 지식은 객체 사이의 강한 규칙성이라고 할 수 있다. 강한 규칙성은 서로 클래스가 다른 객체들의 동시 발생 정도를 기반으로 결정될 수 있다.
- [0049] 도 4는 영상에서 등장하는 규칙성을 설명하는 예이다. 도 4(A)의 영상에서 '사람', '옷', '헬멧', '오토바이'는 동시 발견 확률이 크고, 영상에서 위치 관계(근접함)가 유사하다는 시각적 규칙성을 갖는다. 나아가, 언어적으로도 이러한 상관관계가 강하게 나타난다. 도 4(B)는 이미지를 설명하는 문장들을 보여준다. '파도'는 '서핑 보드' 또는 '서퍼'와 함께 등장하는 경향이 있다.
- [0050] 연구자는 학습 데이터와 학습에 사용되지 않지만 비슷한 양상을 보이는 영상 데이터 세트(visual genome, Krishna, R. et al., 2016, Visual genome: Connecting language and vision using crowd sourced dense image annotations. arXiv preprint arXiv:1602.07332. 참고)를 이용하여 시각적/언어적 행렬을 추출하였다. 즉, 영상 및 문장에서 상관 관계를 나타내는 행렬은 사전에 데이터 세트를 이용하여 구축될 수 있다.
- [0052] 시각적 상관관계는 다음과 같이 구할 수 있다. 단일 이미지 내에 등장하는 객체 o_i 에 대한 카테고리 c_i 가 등장할

때, 서로 다른 카테고리 c_i 와 c_j 가 함께 등장하는 빈도수 E_{ij} 를 구한다. 전체 데이터 세트에서 특정 카테고리의 등장 빈도의 불균형으로 인한 긴 꼬리 문제(long tail problem)를 해결하기 위해 카테고리 c_i 가 등장하는 총 횟수 N_i 으로 나누는 정규화를 수행할 수 있다. 최종적으로 시각적 상관 관계 행렬 V_{ij} 은 아래 수학적 식 1과 같이 얻을 수 있다.

수학적 식 1

$$V_{ij} = E_{ij}/N_i$$

추가적으로 객체 사이의 공간적 상관관계 표현을 위해 객체 사이의 유클리디안 거리를 구해 행렬 R 을 생성할 수 있다. 행렬 R 을 통해 가까운 곳에 위치하는 객체들에 가중을 줄 수 있다. 객체 탐지의 출력으로 얻은 영역 박스 b_i 로부터 객체의 센터 포인트(center point)를 얻을 수 있다. 즉, 이미지에서 탐지된 객체가 K 개라면, K 개의 센터 포인트를 얻고 점 사이의 음의 거리 행렬을 구하고 정규화하여 $0 \sim 1$ 의 크기를 갖는 거리 행렬 $R \in \mathbb{R}^{K \times K}$ 를 생성할 수 있다. 거리 행렬 R 은 자신과의 거리 성분을 제외하고, 거리가 가까운 객체에 해당하는 값이 1에 가깝게 설정될 수 있다. R 은 하이퍼파라미터 θ 를 곱해준 뒤 상관 관계 행렬에 더해주는 것으로 가까운 곳에 위치하는 객체들에 대한 가중을 부여할 수 있다(수학적 식 3 참고).

언어적 상관관계는 다음과 같이 구할 수 있다.

단일 문장 내에 등장하는 객체 o_i 에 대한 카테고리 c_i 가 등장할 때, 서로 다른 카테고리 c_i 와 c_j 가 함께 등장하는 빈도수 E'_{ij} 를 구한다. 특정 카테고리의 등장 빈도의 불균형으로 인한 긴 꼬리 문제를 해결하기 위해 카테고리 c_i 가 등장하는 총 횟수 N_i 으로 나누는 정규화를 수행할 수 있다. 최종적으로, 언어적 상관관계 행렬 S_{ij} 은 아래 수학적 식 2와 같이 얻을 수 있다.

수학적 식 2

$$S_{ij} = E'_{ij}/N_i$$

이제 지식 행렬 M 을 이용하여 초기 데이터에서 인식된 통합된 객체를 표현하는 벡터를 생성하는 과정을 설명한다.

도 5는 지식 행렬을 이용하여 영상에서 통합된 객체를 추출하는 과정(200)의 예이다.

모델 생성 장치는 초기 데이터를 입력받는다(210). 초기 데이터는 영상과 문장의 쌍으로 구성된다.

모델 생성 장치는 입력 영상에 대한 객체 탐지를 수행한다(220). 모델 생성 장치는 객체 탐지 결과로 서로 클래스가 다른 K 개의 객체 $o_1 \dots o_K$, 각 객체의 영역 박스의 좌표 $b_1 \dots b_K$ 및 각 객체의 카테고리 $c_1 \dots c_K$ 를 얻을 수 있다.

선행 지식은 일정한 크기를 갖는 시각적 상관관계 행렬 V , 언어적 상관관계 행렬 S 로 구성될 수 있다. 연구자는 1600개의 카테고리를 기준으로 동시 발생 빈도를 구한 1600×1600 크기의 행렬을 이용하였다.

모델 생성 장치는 입력 영상으로부터 얻은 객체 카테고리를 통해 $K \times K$ 부분 행렬인 V_{sub} 과 S_{sub} 를 추출한다(230).

- [0069] 모델 생성 장치는 객체의 영역 박스의 좌표로부터 얻은 거리 행렬 R 을 구한다(240).
- [0070] 모델 생성 장치는 V_{sub} , S_{sub} 및 R 을 이용하여 최종 선행 행렬을 생성한다(250). 최종 선행 행렬 M 은 아래 수학적 식 3과 같이 표현할 수 있다. 수학적 식 3에서 V 와 S 는 각각 전술한 부분 행렬을 의미한다.

수학적 식 3

[0072]
$$M = \lambda V + \partial R + (1 - \lambda)S$$

- [0074] 연구자는 실험적으로 각 파라미터를 $\lambda = 0.5$, $\partial = 0.2$ 로 설정하여 사용하였다.
- [0075] 모델 생성 장치는 선행행렬 M 을 통해 객체 간 상관 관계가 큰 객체 N 쌍을 선별하여 최종 선택된 객체들에 통합된 객체에 대한 비전 특징 $u_{i \dots N}$ 를 생성할 수 있다(260). 최종적인 통합 객체의 특징은 상관 관계가 큰 객체 N 개 쌍에 대한 벡터로 구성된다.
- [0077] 전술한 바와 같이 비전-언어 사전학습 모델(100)은 통합된 객체와 텍스트 사이의 정렬을 돕기 위해 <주어-관계-목적어> 형태의 구조화된 관계 구문을 사용한다.
- [0078] 통합된 객체와 관계 구문의 주어, 목적어는 서로 다른 모달리티를 갖지만 공통된 의미적 표현을 갖는다. 따라서 관계 구문은 두 개의 모달리티를 공동 최적화하는데 이용할 수 있다.
- [0079] 도 6은 문장에서 관계 구문을 추출하는 과정의 예이다. 관계 구문은 전술한 장면 그래프 파서를 이용하여 생성될 수 있다. 모델 생성 장치는 장면 그래프 파서를 이용하여 입력된 문장을 장면 그래프로 나타낼 수 있다. 또한, 모델 생성 장치는 관계 구문을 얻기 위해 장면 그래프를 구성하는 <주어-관계-목적어> 성분을 추출할 수 있다. 입력 문장은 하위 단어 토큰으로 토큰화된다. [CLS] 및 [SEP]와 같은 특수 토큰이 토큰화된 텍스트 시퀀스에 추가되어 도 6과 같이 최종 텍스트 시퀀스 $S = \{[CLS], w_1, \dots, w_k, [SEP]\}$ 를 얻는다. 모델 생성 장치는 입력 언어 벡터 t 를 만들기 위해서는 관계 구문과 문장 토큰을 연결하여 임베딩한다.
- [0081] 도 7은 비전-언어 사전학습 모델의 동작 과정(300)의 예이다. 모델 생성 장치는 전술한 바와 같이 입력 영상에서 선행 행렬 M 을 이용하여 통합된 객체에 대한 벡터들을 생성한다(310). 모델 생성 장치는 전술한 바와 같이 입력 문장에서 장면 그래프 분석을 이용하여 관계 구문을 생성한다(320). 모델 생성 장치는 생성한 벡터와 관계 구문을 각각 임베딩하여 비전-언어 사전학습 모델에 입력한다. 비전-언어 사전학습 모델은 영상의 객체와 문장의 교차 모달 관계를 학습하게 된다.

- [0083] 비전-언어 사전학습 모델의 학습 과정에 대하여 설명한다.
- [0084] 네트워크의 학습은 다음과 같이 구성한다. 네트워크는 비전-언어 사이 상호 의미 표현을 학습하기 위해 비지도 학습 방식을 사용한다.
- [0085] 학습을 위한 손실 함수로 마스크 토큰 손실과 삼중항 손실을 사용한다.
- [0086] 마스크 토큰 손실을 먼저 설명한다.
- [0087] 언어 도메인의 입력 토큰을 일정 확률로 [MASK] 토큰으로 교체한다.
- [0088] 학습은 입력된 통합된 비전 객체 u_v 와 관계 구문, 그리고 마스킹 되지 않은 문장 토큰 $t_{\setminus i}$ 의 정보 공유를 통해 [MASK] 토큰 t_i 를 예측하는 것을 목표로 한다.
- [0089] 학습에 사용되는 목적 함수는 음의 로그 우도를 최소화하는 것을 목표로 하는 손실함수 L_{ML} 를 사용하며 아래 수

학식 4와 같이 표현할 수 있다.

수학식 4

$$L_{MT} = -E_{v,n} \sim D \log p((t_i | t_{\setminus i}, u)$$

또한, 두 개의 모달리티 사이의 공동 표현 학습 방법으로 삼중항 손실(triplet loss)을 사용한다. 정답 이미지-텍스트 쌍 임베딩의 코사인 유사도 점수를 이용한다. 배치 B 내에서 랜덤으로 통합된 객체 비전 피쳐 u, 입력 언어 벡터 t 쌍을 교환하여 의미적 매칭이 되지 않는 오답 쌍 (u^i, t^j) 을 생성하여 정답 쌍 (u^i, t^i) 과 유사도 점수를 비교한다. 의미적 유사도가 높은 쌍을 가깝게, 낮은 쌍을 멀게 표현할 수 있도록 손실함수는 삼중항 손실을 사용하며 L_{tri} 이라 정의하였고, 아래 수학식 5와 같이 표현될 수 있다.

수학식 5

$$L_{tri} = \frac{1}{B} \sum_{I=1}^B \max_j [\delta + s(u^i, t^i) - s(u^i, t^j)]$$

이때 B는 배치 사이즈를 의미하고 $s(\dots)$ 는 코사인 유사도를 의미한다. δ 는 마진으로 연구자는 0.2로 설정했다.

최종적으로 사전학습을 위한 전체 손실함수는 마스크 토큰 손실과 트리플렛 손실을 최소화하는 것을 목적으로 한다. 전체 손실 함수 L_{Tot} 는 아래 수학식 6과 같이 표현할 수 있다.

수학식 6

$$L_{Tot} = L_{MT} + L_{tri}$$

전술한 비전-언어 사전학습 모델의 성능을 검증한 결과를 설명한다. 연구자는 학습된 비전-언어 사전학습 모델을 이용한 영상 캡셔닝에 대한 파인 튜닝을 수행하고 성능을 확인하였다. 모델 학습 및 검증 과정은 다음과 같다.

사전학습을 위한 데이터 셋으로 COCO, Flickr30k 및 VQA 2.0의 이미지-텍스트 쌍 3,691,045 개를 이용하였다. 객체 검출기로는 Detectron2를 이용하였다. 객체 추출 시 최대 검출 객체는 36개로 제한하였고, 통합된 객체는 연관도가 가장 높은 5개의 쌍을 선별하였다. 사전 학습을 위해 사용된 트랜스포머의 멀티 헤드는 총 6 개이고, 각 히든 레이어 사이즈는 768로 초기화되었다. 영상 피쳐가 BERT와 동일한 입력 임베딩 크기를 갖도록 선형 투영하였고, 최적화를 위해 Adam을 사용하였다.

캡션 토큰의 15%를 임의로 [MASK] 토큰으로 변경하고 해당 출력을 예측하였다. 트랜스포머 구조를 이용한 문장 생성을 위해 단방향의 시퀀셜한 구조를 따랐다. 단 방향 프로세스를 위해 어텐션 마스크를 통해 예측 토큰 이후의 내용은 사용하지 않았다. 추론작업은 [MASK] 토큰 추가와 예측을 반복 진행을 통해 이루어진다. [STOP] 토큰이 나올 때까지 문장을 생성하였다.

실험에서 빔 검색을 5로 설정하였고, COCO 이미지 캡션 데이터셋에 대해 결과를 보인다. 실험에서 빔 검색

(beam search)를 5로 설정하였고, COCO 이미지 캡션 데이터 세트로 검증하였다. 학습을 위해 20 에폭 동안 크로스 엔트로피 손실함수를 이용하였고, 이후 15 에폭 동안 SCST에서와 같이 CIDER 점수를 통해 최적화했다.

[0105] 모델의 정량적 평가를 위해 사전 학습된 모델을 영상 캡셔닝에 대해 파인 튜닝하여 얻은 성능을 종전 연구인 BUTD(Anderson, P. et al., 2018, Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition) 및 VLP(Zhou, L. et al., 2020, Unified vision-language pre-training for image captioning and vqa. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 07, pp. 13041-13049)와 비교했다. 아래 표 1은 검증 결과는 나타낸다. 평가 지표로 종래 연구에서 활용되었던 BLEU, METEOR, CIDEr 및 SPICE를 사용하였다 RE-VLP가 도 7에서 설명한 비전-언어 사전학습 모델을 의미한다. BUTD에 비해 RE-VLP가 METEOR, CIDEr, SPICE 평가지표에 대해 성능이 각각 1.29, 5.7, 1.4만큼 향상되었다. 또한, 객체 기반의 사전 학습을 진행한 VLP에 비해 RE-VLP가 METEOR, SPICE에 대해 각각 1.5, 0.4만큼의 성능 향상을 확인하였다.

표 1

	BLEU	METEOR	CIDEr	SPICE
BUTD	36.2	27.0	113.5	20.3
VLP	36.5	28.4	117.7	21.3
RE-VLP	35.13	28.29	119.2	2.17

[0109] 도 8은 비전-언어 사전학습 모델을 구축하는 모델 생성 장치(400)에 대한 예이다. 모델 생성 장치(400)는 물리적으로 다양한 형태로 구현될 수 있다. 예컨대, 모델 생성 장치(400)는 PC와 같은 컴퓨터 장치, 네트워크의 서버, 데이터 처리 전용 칩셋 등의 형태를 가질 수 있다.

[0110] 모델 생성 장치(400)는 저장장치(410), 메모리(420), 연산장치(430), 인터페이스 장치(440) 및 통신장치(450)를 포함할 수 있다.

[0111] 저장장치(410)는 학습 데이터를 저장할 수 있다.

[0112] 저장장치(410)는 사전에 마련한 상관관계 행렬을 저장할 수 있다. 상관관계 행렬은 전술한 바와 같이 학습 데이터 내지 다른 데이터 세트를 이용하여 사전에 클래스를 분류한 개수에 따라 일정한 크기의 행렬로 구축될 수 있다.

[0113] 저장장치(410)는 비전-언어 사전학습 모델을 저장할 수 있다. 비전-언어 사전학습 모델은 다층의 트랜스포머 블록들로 구성될 수 있다.

[0114] 메모리(420)는 모델 생성 장치(400)가 비전-언어 사전학습 모델을 구축하는 과정에서 생성되는 데이터 및 정보 등을 저장할 수 있다.

[0115] 인터페이스 장치(440)는 외부로부터 일정한 명령 및 데이터를 입력받는 장치이다. 인터페이스 장치(440)는 물리적으로 연결된 입력 장치, 외부 저장장치로부터 학습 데이터, 상관관계 행렬 등을 입력받을 수 있다. 인터페이스 장치(440)는 학습된 비전-언어 사전학습 모델을 외부 객체에 전달할 수도 있다.

[0116] 통신장치(450)는 유선 또는 무선 네트워크를 통해 일정한 정보를 수신하고 전송하는 구성을 의미한다. 통신장치(450)는 외부 객체로부터 학습 데이터, 상관관계 행렬 등을 수신할 수 있다. 또는 통신장치(450)는 학습된 비전-언어 사전학습 모델을 외부 객체에 송신할 수도 있다.

[0117] 인터페이스 장치(440) 및 통신장치(450)는 사용자 또는 다른 물리적 객체로부터 일정한 데이터를 주고받는 구성이므로, 포괄적으로 입출력장치라고도 명명할 수 있다. 필요한 데이터를 입력받는 기능에 한정하면 인터페이스 장치(440) 및 통신장치(450)는 입력장치라고 할 수도 있다.

[0118] 연산장치(430)는 저장장치(410)에 저장된 비전-언어 사전학습 모델을 학습 데이터를 이용하여 모델을 구축할 수 있다.

[0119] 연산장치(430)는 초기 데이터의 영상에서 객체 검출 프로그램을 이용하여 객체를 검출할 수 있다. 연산장치

(430)는 영상에서 특정 객체, 객체 영역, 객체 클래스를 추출할 수 있다. 연산장치(430)는 객체 영역의 중심을 기준으로 각 객체 사이의 거리를 기준으로 거리 행렬 R 을 산출한다. 연산장치(430)는 입력 영상으로부터 얻은 객체 카테고리 및 상관 관계 행렬을 이용하여 부분 행렬인 V_{sub} 과 S_{sub} 를 추출한다. 연산장치(430)는 V_{sub} , S_{sub} 및 R 을 이용하여 최종 선행 행렬을 생성한다. 연산장치(430)는 선행행렬 M 을 통해 객체 간 상관 관계가 큰 객체 N 쌍을 선별하여 최종 선택된 객체들에 통합된 객체에 대한 비전 특징 $u_i \dots N$ 를 생성한다.

[0120] 연산장치(430)는 초기 데이터의 문장에 대한 장면 그래프 분석을 통하여 전술한 바와 같이 <주어-관계-목적어> 구조의 관계 구문을 산출한다. 이후, 연산장치(430)는 [CLS] 및 [SEP]와 같은 특수 토큰이 토큰화된 텍스트 시퀀스에 추가하여 최종 텍스트 시퀀스를 생성할 수 있다.

[0121] 연산장치(430)는 비전 특징과 문장에 대한 최종 텍스트 시퀀스를 임베딩하여 비전-언어 사전학습 모델에 입력한다. 이때 비전-언어 사전학습 모델은 수식식 6에서 표현한 마스크 토큰 손실과 트리플렛 손실을 최소화하는 방향으로 학습된다.

[0122] 모델 생성 장치(400)는 다수의 학습 데이터에 대하여 비전-언어 사전학습 모델을 학습하는 과정을 반복하여 비전-언어 사전학습 모델에 대한 사전 학습을 수행한다.

[0123] 이후 모델 생성 장치(400)는 이미지 캡션과 같은 특정 애플리케이션을 위한 파인 튜닝을 수행하여 특정 애플리케이션에 대한 모델을 구축할 수 있다.

[0124] 연산장치(430)는 데이터를 처리하고, 일정한 연산을 처리하는 프로세서, AP, 프로그램이 임베디드된 칩과 같은 장치일 수 있다.

[0126] 또한, 상술한 바와 같은 비전-언어 사전학습 방법은 컴퓨터에서 실행될 수 있는 실행가능한 알고리즘을 포함하는 프로그램(또는 어플리케이션)으로 구현될 수 있다. 상기 프로그램은 일시적 또는 비일시적 판독 가능 매체(non-transitory computer readable medium)에 저장되어 제공될 수 있다.

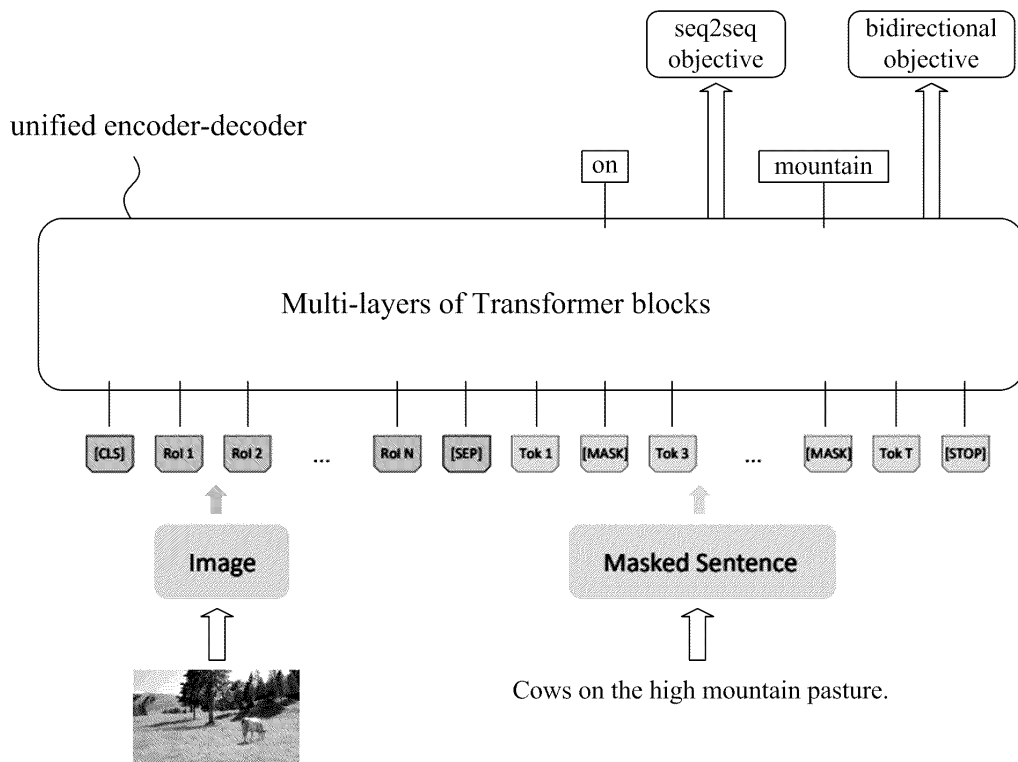
[0127] 비일시적 판독 가능 매체란 레지스터, 캐쉬, 메모리 등과 같이 짧은 순간 동안 데이터를 저장하는 매체가 아니라 반영구적으로 데이터를 저장하며, 기기에 의해 판독(reading)이 가능한 매체를 의미한다. 구체적으로는, 상술한 다양한 어플리케이션 또는 프로그램들은 CD, DVD, 하드 디스크, 블루레이 디스크, USB, 메모리카드, ROM(read-only memory), PROM(programmable read only memory), EPROM(Erasable PROM, EPROM) 또는 EEPROM(Electrically EPROM) 또는 플래시 메모리 등과 같은 비일시적 판독 가능 매체에 저장되어 제공될 수 있다.

[0128] 일시적 판독 가능 매체는 스태틱 램(Static RAM, SRAM), 다이내믹 램(Dynamic RAM, DRAM), 싱크로너스 디램(Synchronous DRAM, SDRAM), 2배속 SDRAM(Double Data Rate SDRAM, DDR SDRAM), 증강형 SDRAM(Enhanced SDRAM, ESDRAM), 동기화 DRAM(Synclink DRAM, SDRAM) 및 직접 램버스 램(Direct Rambus RAM, DRRAM)과 같은 다양한 RAM을 의미한다.

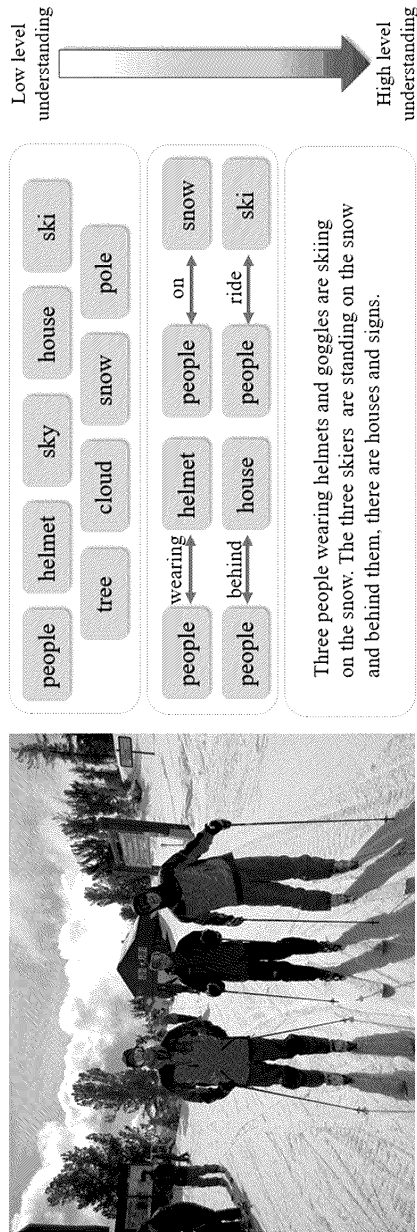
[0129] 본 실시예 및 본 명세서에 첨부된 도면은 전술한 기술에 포함되는 기술적 사상의 일부를 명확하게 나타내고 있는 것에 불과하며, 전술한 기술의 명세서 및 도면에 포함된 기술적 사상의 범위 내에서 당업자가 용이하게 유추할 수 있는 변형 예와 구체적인 실시예는 모두 전술한 기술의 권리범위에 포함되는 것이 자명하다고 할 것이다.

도면

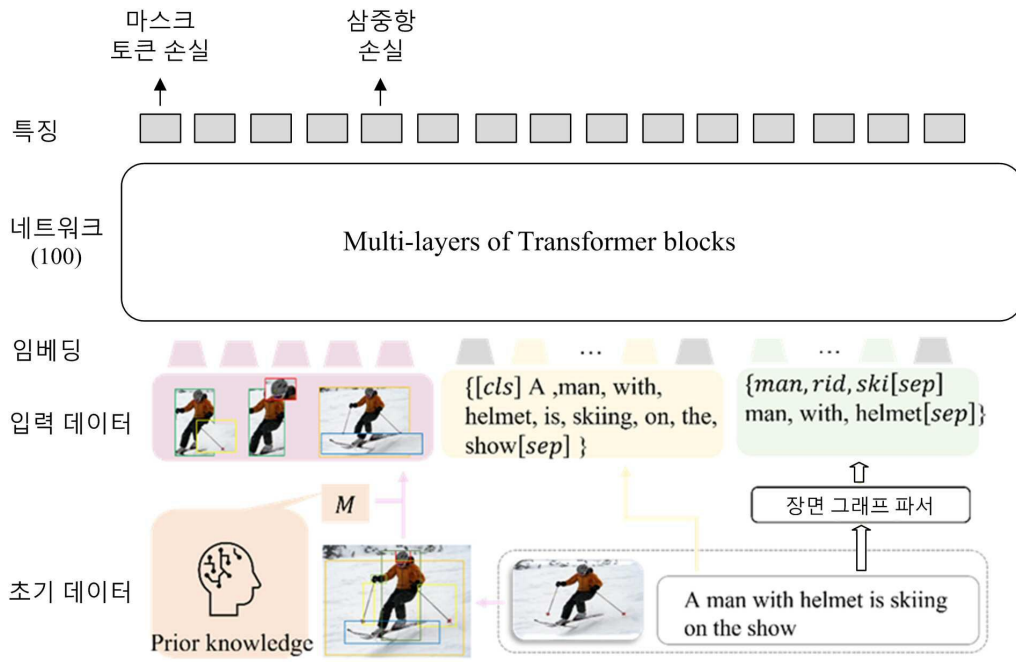
도면1



도면2



도면3



도면4



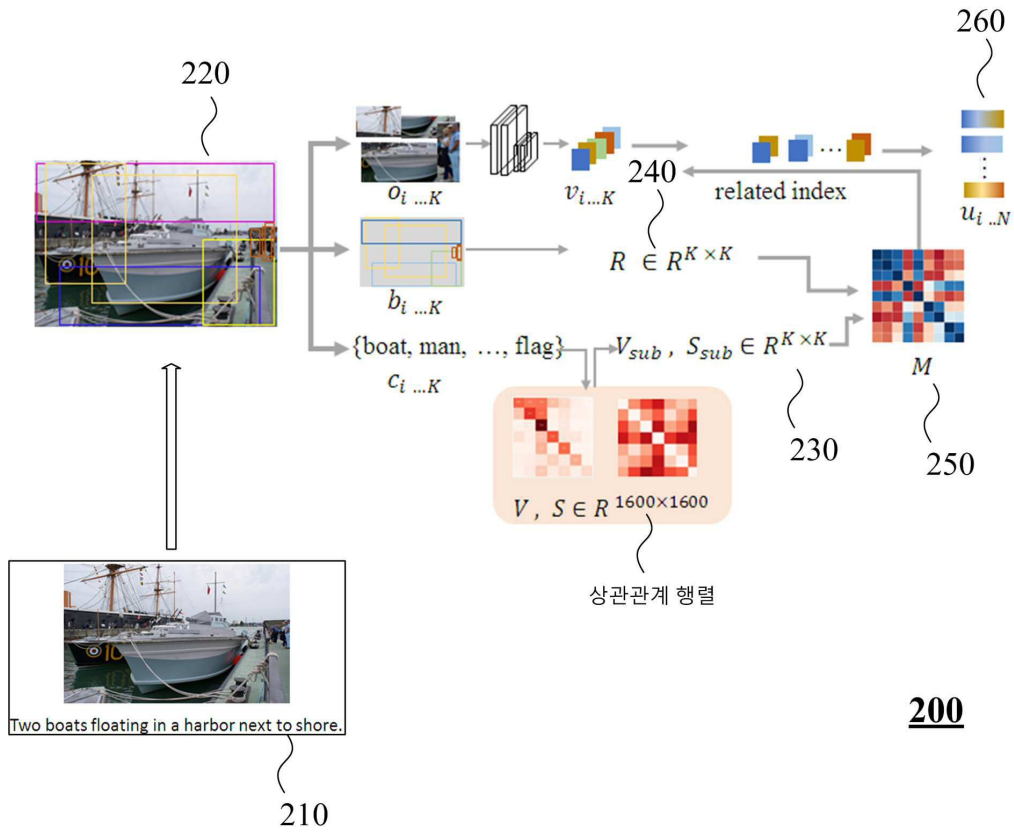
(A)



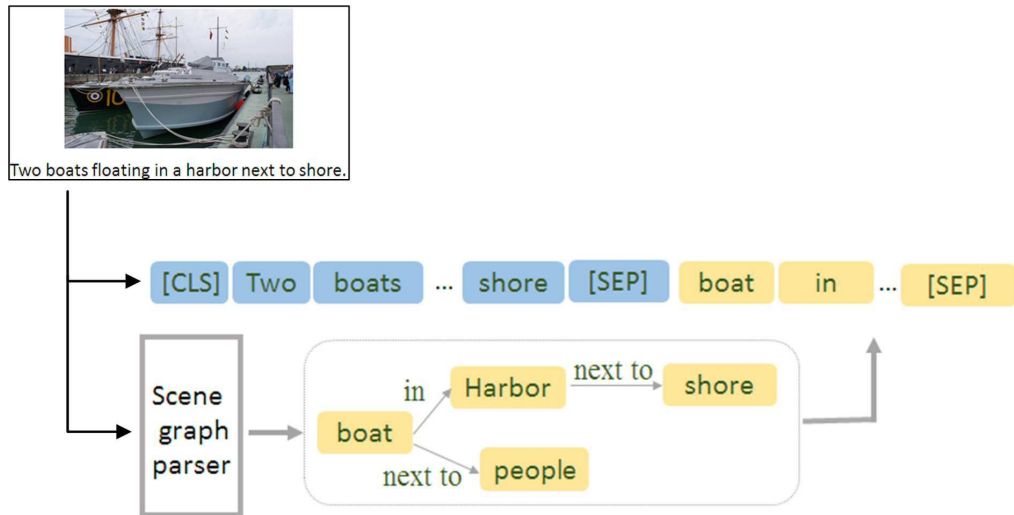
A young boy **surfing** in low **waves**.
A young boy is standing on a **surfboard** and riding a **wave**.
A surfer rides his **surfboard** on some very small **waves**.
A young boy is standing on a **surfboard** in the **water**.
A young boy is standing on a **surfboard** in the **ocean**.

(B)

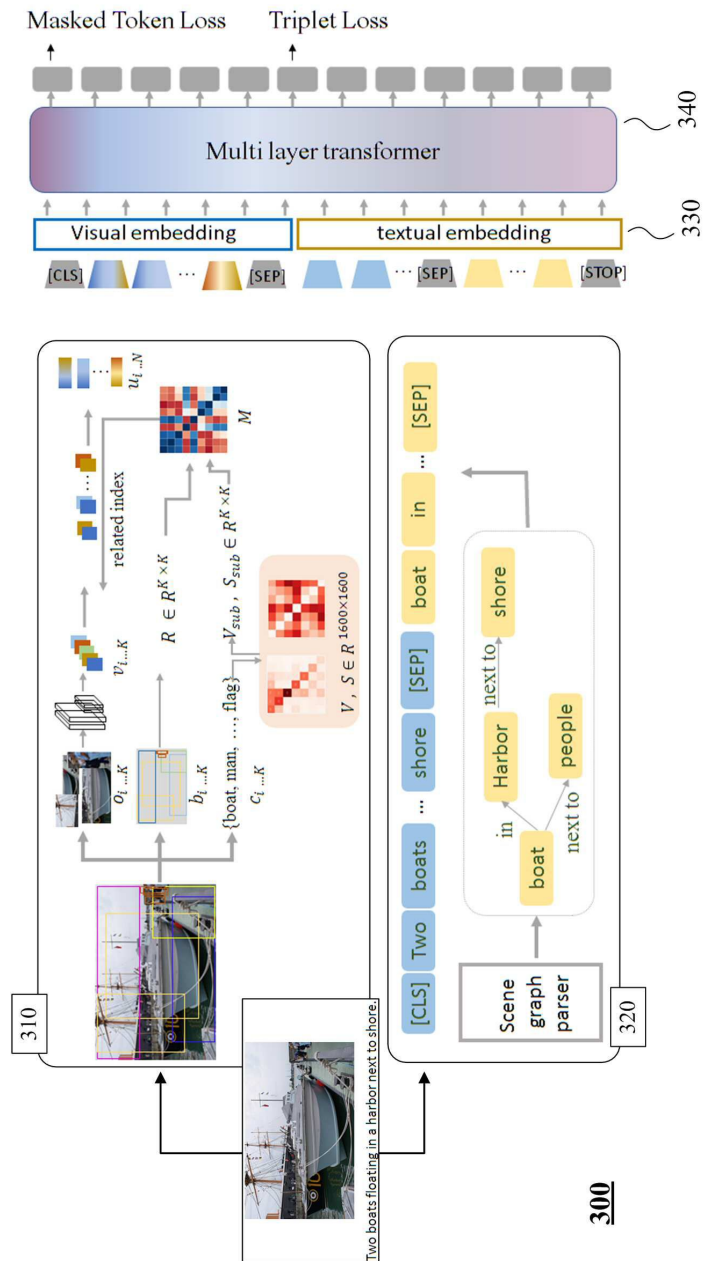
도면5



도면6



도면7



도면8

