



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0030307
(43) 공개일자 2024년03월07일

(51) 국제특허분류(Int. Cl.)
G06F 40/284 (2020.01) G06F 40/16 (2020.01)
G06F 40/58 (2020.01) G06N 20/00 (2019.01)
G06V 10/22 (2022.01) G06V 30/10 (2022.01)
(52) CPC특허분류
G06F 40/284 (2020.01)
G06F 40/16 (2020.01)
(21) 출원번호 10-2022-0109147
(22) 출원일자 2022년08월30일
심사청구일자 2022년08월30일

(71) 출원인
세종대학교산학협력단
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
(72) 발명자
유성준
서울특별시 광진구 능동로 209, 대양AI센터 717호(군자동)
구영현
서울특별시 광진구 능동로 209, 대양AI센터 717호(군자동)
하유진
서울특별시 광진구 능동로 209, 대양AI센터 410호(군자동)
(74) 대리인
양성보

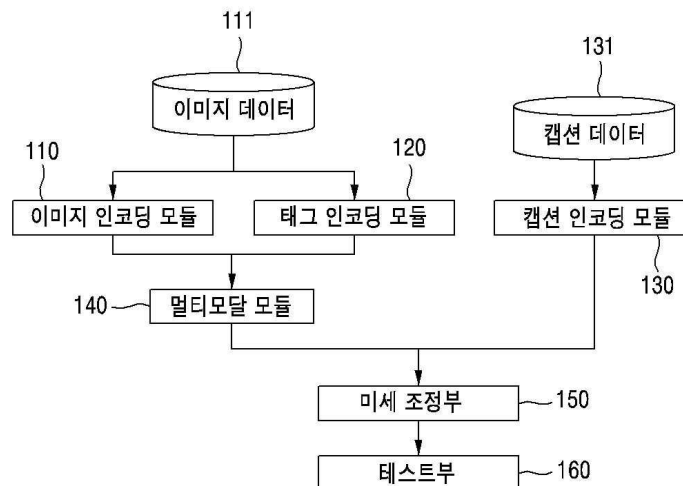
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 사전 학습 모델을 사용한 멀티모달 기반 이미지 캡셔닝 방법 및 시스템

(57) 요약

사전 학습 모델을 사용한 멀티모달 기반 이미지 캡셔닝 방법 및 시스템이 제시된다. 본 발명에서 제안하는 사전 학습 모델을 사용한 멀티모달 기반 이미지 캡셔닝 시스템은 입력 이미지로부터 시각적 특징을 추출하여 시각적 토큰 임베딩으로 변환하는 이미지 인코딩 모듈, 입력 이미지로부터 객체를 탐지하여 객체 태그를 획득한 후 태그 토큰 임베딩으로 변환하는 태그 인코딩 모듈, 캡션 데이터를 전처리 한 후 인코딩을 수행하여 텍스트 모달리티를 생성하는 캡션 인코딩 모듈, 상기 시각적 토큰 임베딩 및 상기 태그 토큰 임베딩을 결합하여 멀티모달 임베딩으로 결합하는 멀티모달 모듈, 상기 결합된 멀티모달 임베딩에 대하여 사전 학습된 언어 모델을 미세 조정하는 미세 조정부 및 상기 미세 조정된 사전 학습된 언어 모델을 새로운 이미지에 대해 테스트하는 테스트부를 포함한다.

대표도 - 도1



(52) CPC특허분류

G06F 40/58 (2020.01)

G06N 20/00 (2021.08)

G06V 10/22 (2023.08)

G06V 30/10 (2023.08)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711152786
과제번호	2021-0-00755-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	SW컴퓨팅산업원천기술개발
연구과제명	데이터 규모 확장과 정확성 향상을 위한 다크 데이터 분석 기술 개발
기 여 율	1/1
과제수행기관명	(주)와이즈넷
연구기간	2022.01.01 ~ 2022.12.31

명세서

청구범위

청구항 1

입력 이미지로부터 시각적 특징을 추출하여 시각적 토큰 임베딩으로 변환하는 이미지 인코딩 모듈;
 입력 이미지로부터 객체를 탐지하여 객체 태그를 획득한 후 태그 토큰 임베딩으로 변환하는 태그 인코딩 모듈;
 캡션 데이터를 전처리 한 후 인코딩을 수행하여 텍스트 모달리티를 생성하는 캡션 인코딩 모듈;
 상기 시각적 토큰 임베딩 및 상기 태그 토큰 임베딩을 결합하여 멀티모달 임베딩으로 결합하는 멀티모달 모듈;
 상기 결합된 멀티모달 임베딩에 대하여 사전 학습된 언어 모델을 미세 조정하는 미세 조정부; 및
 상기 미세 조정된 사전 학습된 언어 모델을 새로운 이미지에 대해 테스트하는 테스트부
 를 포함하는 이미지 캡서닝 시스템.

청구항 2

제1항에 있어서,

상기 이미지 인코딩 모듈은,

CLIP(Contrastive Language-Image Pre-training) 이미지 인코더를 통해 입력 이미지로부터 시각적 특징을 추출하고, 상기 추출된 시각적 특징을 프로젝션 네트워크(Projection Network)를 통해 시각적 토큰 임베딩으로 변환하는

이미지 캡서닝 시스템.

청구항 3

제1항에 있어서,

상기 태그 인코딩 모듈은,

객체 탐지 모델을 통해 입력 이미지로부터 객체 태그와 신뢰도 점수를 추출하고, 상기 추출된 객체 태그가 외국어인 경우 한국어로 번역하여, 상기 번역된 객체 태그를 숫자로 변환하는 인코딩 과정을 거친 후 임베딩 계층을 통해 컴퓨터 판독 가능한 벡터인 태그 토큰 임베딩으로 변환하는

이미지 캡서닝 시스템.

청구항 4

제1항에 있어서,

상기 캡션 인코딩 모듈은,

사전 학습된 언어 모델을 통해 자연어 형태의 텍스트인 캡션 데이터를 숫자로 변환하는 인코딩을 수행하고, 상기 캡션 데이터에서 특수문자를 제거한 후 상기 캡션 데이터를 최대 길이에 맞도록 조정한 후, 토큰라이저를 통해 상기 캡션 데이터를 인코딩된 캡션 토큰으로 변환하는

이미지 캡서닝 시스템.

청구항 5

제1항에 있어서,

상기 멀티모달 모듈은,

상기 이미지 인코딩 과정을 통해 생성된 시각적 토큰 임베딩과 태그 인코딩 과정을 통해 생성된 태그 토큰 임베

당의 서로 다른 모달리티의 임베딩을 결합하기 위해 연결(Concatenation)을 적용하여 상기 시각적 토큰 임베딩과 태그 토큰 임베딩을 결합하여 멀티모달 임베딩을 생성하는

이미지 캡서닝 시스템.

청구항 6

제1항에 있어서,

상기 미세 조정부는,

상기 멀티모달 모듈을 통해 생성된 멀티모달 임베딩과 상기 캡션 인코딩 모듈로부터 인코딩된 캡션 토큰을 입력받아 사전 학습 언어 모델을 미세 조정하고,

상기 사전 학습 언어 모델에 대해 현재 시점의 실제값(Ground Truth)을 다음 시점의 입력값으로 사용하여 모델을 훈련하는 교사 강제(Teacher Forcing) 방식으로 학습하는

이미지 캡서닝 시스템.

청구항 7

이미지 인코딩 모듈을 통해 입력 이미지로부터 시각적 특징을 추출하여 시각적 토큰 임베딩으로 변환하는 단계;

태그 인코딩 모듈을 통해 입력 이미지로부터 객체를 탐지하여 객체 태그를 획득한 후 태그 토큰 임베딩으로 변환하는 단계;

캡션 인코딩 모듈을 통해 캡션 데이터를 전처리 한 후 인코딩을 수행하여 텍스트 모달리티를 생성하는 단계;

멀티모달 모듈을 통해 상기 시각적 토큰 임베딩 및 상기 태그 토큰 임베딩을 결합하여 멀티모달 임베딩으로 결합하는 단계;

미세 조정부를 통해 상기 결합된 멀티모달 임베딩에 대하여 사전 학습된 언어 모델을 미세 조정하는 단계; 및

테스트부를 통해 상기 미세 조정된 사전 학습된 언어 모델을 새로운 이미지에 대해 테스트하는 단계

를 포함하는 이미지 캡서닝 방법.

청구항 8

제7항에 있어서,

상기 입력 이미지로부터 시각적 특징을 추출하여 시각적 토큰 임베딩으로 변환하는 이미지 인코딩 모듈은,

CLIP(Contrastive Language-Image Pre-training) 이미지 인코더를 통해 입력 이미지로부터 시각적 특징을 추출하고, 상기 추출된 시각적 특징을 프로젝션 네트워크(Projection Network)를 통해 시각적 토큰 임베딩으로 변환하는

이미지 캡서닝 방법.

청구항 9

제7항에 있어서,

상기 입력 이미지로부터 객체를 탐지하여 객체 태그를 획득한 후 태그 토큰 임베딩으로 변환하는 태그 인코딩 모듈은,

객체 탐지 모델을 통해 입력 이미지로부터 객체 태그와 신뢰도 점수를 추출하고, 상기 추출된 객체 태그가 외국어인 경우 한국어로 번역하여, 상기 번역된 객체 태그를 숫자로 변환하는 인코딩 과정을 거친 후 임베딩 계층을 통해 컴퓨터 판독 가능한 벡터인 태그 토큰 임베딩으로 변환하는

이미지 캡서닝 방법.

청구항 10

제7항에 있어서,

상기 캡션 데이터를 전처리 한 후 인코딩을 수행하여 텍스트 모달리티를 생성하는 캡션 인코딩 모듈은,
사전 학습된 언어 모델을 통해 자연어 형태의 텍스트인 캡션 데이터를 숫자로 변환하는 인코딩을 수행하고, 상기 캡션 데이터에서 특수문자를 제거한 후 상기 캡션 데이터를 최대 길이에 맞도록 조정한 후, 토큰라이저를 통해 상기 캡션 데이터를 인코딩된 캡션 토큰으로 변환하는
이미지 캡서닝 방법.

청구항 11

제7항에 있어서,

상기 멀티모달 모듈을 통해 상기 시각적 토큰 임베딩 및 상기 태그 토큰 임베딩을 결합하여 멀티모달 임베딩으로 결합하는 단계는,

상기 이미지 인코딩 과정을 통해 생성된 시각적 토큰 임베딩과 태그 인코딩 과정을 통해 생성된 태그 토큰 임베딩의 서로 다른 모달리티의 임베딩을 결합하기 위해 연결(Concatenation)을 적용하여 상기 시각적 토큰 임베딩과 태그 토큰 임베딩을 결합하여 멀티모달 임베딩을 생성하는

이미지 캡서닝 방법.

청구항 12

제7항에 있어서,

상기 미세 조정부를 통해 상기 결합된 멀티모달 임베딩에 대하여 사전 학습된 언어 모델을 미세 조정하는 단계는,

상기 멀티모달 모듈을 통해 생성된 멀티모달 임베딩과 상기 캡션 인코딩 모듈로부터 인코딩된 캡션 토큰을 입력 받아 사전 학습 언어 모델을 미세 조정하고,

상기 사전 학습 언어 모델에 대해 현재 시점의 실제값(Ground Truth)을 다음 시점의 입력값으로 사용하여 모델을 훈련하는 교사 강제(Teacher Forcing) 방식으로 학습하는

이미지 캡서닝 방법.

발명의 설명

기술 분야

[0001] 본 발명은 사전 학습 모델을 사용한 멀티모달 기반 이미지 캡서닝 방법 및 시스템에 관한 것이다.

배경 기술

[0002] 이미지 캡서닝(Image Captioning)은 이미지의 내용을 설명하는 시각-언어(Vision-and-Language) 작업(Task) 중 하나로, 이미지가 입력으로 주어질 때 캡션을 생성하는 모델이다. 이 기술은 자막 생성, 이미지 검색 및 시력 약자를 위한 지원과 같이 다양한 분야에서 활용하기 위해 연구되고 있다.

[0003] 일반적으로 이미지 캡서닝 모델은 이미지로부터 특징을 추출하는 인코더와 특징을 입력 받아 캡션을 생성하는 디코더로 구성된다. 최근에는 이미지뿐만 아니라 이미지로부터 탐지된 객체 태그를 함께 사용해 모델의 성능을 개선하는 연구들이 있다. 최근 이미지 캡서닝 연구에서 인코더로 CLIP(Contrastive Language-Image Pre-training)의 이미지 인코더를 사용하고 디코더로 사전 학습된 언어 모델을 사용한다. 여기서, CLIP은 수억 개의 이미지-텍스트 쌍의 데이터 셋에서 훈련되어 일반화 능력이 뛰어나고 이미지 인코더 및 텍스트 인코더로 구성된 모델이다. 또한, 이미지 캡서닝 모델의 성능을 개선하기 위해 이미지뿐만 아니라 텍스트 정보를 활용한다.

[0004] 기존에는 사전 학습된 Faster R-CNN을 사용해 이미지로부터 객체 태그를 탐지했으나, Faster R-CNN은 이미지로부터 탐지할 수 있는 객체 태그의 수가 대략 1,600개로 적은 문제점이 있다. 이러한 이미지 캡서닝 연구는 영어권에서 활발히 연구되는 반면 국내 한국어 이미지 캡서닝 연구는 충분히 이루어지지 않았고 최신 알고리즘이 적용되지 않아 성능이 저조하다. 한국어 이미지 캡서닝 모델은 여전히 인코더로 사전 학습된 이미지 분류 모델을 사용하고 디코더로 LSTM을 사용하고 최신 알고리즘이 적용되지 않아 성능이 저조한 문제점이 있다.

선행기술문헌

특허문헌

- [0005] (특허문헌 0001) "Adaptive Attention 기반 한국어 & 영어 이미지 캡션 생성 기술 개발", 한국통신학회 학술대회논문집, 1,021 - 1,022 (2 pages), 2020.02.

발명의 내용

해결하려는 과제

- [0006] 본 발명이 이루고자 하는 기술적 과제는 한국어 이미지 캡셔닝 모델의 성능을 개선하는 방법을 제안하고 이를 영어 이미지 캡셔닝 연구에 적용하여 방법의 타당성을 검증하고자 한다. 기존 한국어 이미지 캡셔닝 연구에서 디코더로 사전 학습된 언어 모델을 사용한 사례가 없다. 따라서 본 발명에서는 인코더로서 CLIP의 이미지 인코더 및 디코더로 사전 학습된 언어 모델인 KE-T5의 조합을 통해 사전 학습 모델을 사용한 멀티모달 기반 이미지 캡셔닝 방법 및 시스템을 제공하는데 있다. 또한, 모델의 성능을 개선하기 위해 이 미지뿐만 아니라 텍스트 정보를 함께 사용하는 방법을 적용한다.

과제의 해결 수단

- [0007] 일 측면에 있어서, 본 발명에서 제안하는 사전 학습 모델을 사용한 멀티모달 기반 이미지 캡셔닝 시스템은 입력 이미지로부터 시각적 특징을 추출하여 시각적 토큰 임베딩으로 변환하는 이미지 인코딩 모듈, 입력 이미지로부터 객체를 탐지하여 객체 태그를 획득한 후 태그 토큰 임베딩으로 변환하는 태그 인코딩 모듈, 캡션 데이터를 전처리 한 후 인코딩을 수행하여 텍스트 모달리티를 생성하는 캡션 인코딩 모듈, 상기 시각적 토큰 임베딩 및 상기 태그 토큰 임베딩을 결합하여 멀티모달 임베딩으로 결합하는 멀티모달 모듈, 상기 결합된 멀티모달 임베딩에 대하여 사전 학습된 언어 모델을 미세 조정하는 미세 조정부 및 상기 미세 조정된 사전 학습된 언어 모델을 새로운 이미지에 대해 테스트하는 테스트부를 포함한다.
- [0008] 본 발명의 실시예에 따른 이미지 인코딩 모듈은 CLIP(Contrastive Language-Image Pre-training) 이미지 인코더를 통해 입력 이미지로부터 시각적 특징을 추출하고, 상기 추출된 시각적 특징을 프로젝션 네트워크(Projection Network)를 통해 시각적 토큰 임베딩으로 변환한다.
- [0009] 본 발명의 실시예에 따른 태그 인코딩 모듈은 객체 탐지 모델을 통해 입력 이미지로부터 객체 태그와 신뢰도 점수를 추출하고, 상기 추출된 객체 태그가 외국어인 경우 한국어로 번역하여, 상기 번역된 객체 태그를 숫자로 변환하는 인코딩 과정을 거친 후 임베딩 계층을 통해 컴퓨터 판독 가능한 벡터인 태그 토큰 임베딩으로 변환한다.
- [0010] 본 발명의 실시예에 따른 캡션 인코딩 모듈은 사전 학습된 언어 모델을 통해 자연어 형태의 텍스트인 캡션 데이터를 숫자로 변환하는 인코딩을 수행하고, 상기 캡션 데이터에서 특수문자를 제거한 후 상기 캡션 데이터를 최대 길이에 맞도록 조정된 후, 토큰라이저를 통해 상기 캡션 데이터를 인코딩된 캡션 토큰으로 변환한다.
- [0011] 본 발명의 실시예에 따른 멀티모달 모듈은 상기 이미지 인코딩 과정을 통해 생성된 시각적 토큰 임베딩과 태그 인코딩 과정을 통해 생성된 태그 토큰 임베딩의 서로 다른 모달리티의 임베딩을 결합하기 위해 연결(Concatenation)을 적용하여 상기 시각적 토큰 임베딩과 태그 토큰 임베딩을 결합하여 멀티모달 임베딩을 생성한다.
- [0012] 본 발명의 실시예에 따른 미세 조정부는 상기 멀티모달 모듈을 통해 생성된 멀티모달 임베딩과 상기 캡션 인코딩 모듈로부터 인코딩된 캡션 토큰을 입력 받아 사전 학습 언어 모델을 미세 조정하고, 상기 사전 학습 언어 모델에 대해 현재 시점의 실체값(Ground Truth)을 다음 시점의 입력값으로 사용하여 모델을 훈련하는 교사 강제(Teacher Forcing) 방식으로 학습한다.
- [0013] 또 다른 일 측면에 있어서, 본 발명에서 제안하는 사전 학습 모델을 사용한 멀티모달 기반 이미지 캡셔닝 방법은 이미지 인코딩 모듈을 통해 입력 이미지로부터 시각적 특징을 추출하여 시각적 토큰 임베딩으로 변환하는 단계, 태그 인코딩 모듈을 통해 입력 이미지로부터 객체를 탐지하여 객체 태그를 획득한 후 태그 토큰 임베딩으로 변환하는 단계, 캡션 인코딩 모듈을 통해 캡션 데이터를 전처리 한 후 인코딩을 수행하여 텍스트 모달리티를 생

성하는 단계, 멀티모달 모듈을 통해 상기 시각적 토큰 임베딩 및 상기 태그 토큰 임베딩을 결합하여 멀티모달 임베딩으로 결합하는 단계, 미세 조정부를 통해 상기 결합된 멀티모달 임베딩에 대하여 사전 학습된 언어 모델을 미세 조정하는 단계 및 테스트부를 통해 상기 미세 조정된 사전 학습된 언어 모델을 새로운 이미지에 대해 테스트하는 단계를 포함한다.

발명의 효과

[0014] 본 발명의 실시예들에 따르면 이미지에 대한 설명을 자동으로 생성할 수 있고, 한국어 이미지 캡셔닝 모델의 성능을 개선할 수 있다. '한국어 C-T5' 모델의 경우 기존 한국어 이미지 캡셔닝의 벤치마크 모델과 비교해 CIDEr, SPICE, BLEU-4, METEOR 지표에서 각각 19.3, 3.3, 4.8, 3.2만큼 개선된 성능을 보였다. 그리고 '영어 C-T5' 모델의 경우 기존 영어 이미지 캡셔닝의 벤치마크 모델과 비교해 CIDEr 지표에서 119.9로 가장 높은 성능을 보였다. '영어 C-T5' 모델은 적은 노력에도 불구하고 사전 훈련(Pre-training)을 적용한 벤치마크 모델과 유사한 성능을 보였다.

도면의 간단한 설명

[0015] 도 1은 본 발명의 일 실시예에 따른 사전 학습 모델을 사용한 멀티모달 기반 이미지 캡셔닝 시스템의 구성을 나타내는 도면이다.

도 2는 본 발명의 일 실시예에 따른 이미지 인코딩 모듈을 설명하기 위한 도면이다.

도 3은 본 발명의 일 실시예에 따른 태그 인코딩 모듈을 설명하기 위한 도면이다.

도 4는 본 발명의 일 실시예에 따른 멀티모달 모듈을 설명하기 위한 도면이다.

도 5는 본 발명의 일 실시예에 따른 캡션 인코딩 모듈을 설명하기 위한 도면이다.

도 6은 본 발명의 일 실시예에 따른 미세 조정부를 설명하기 위한 도면이다.

도 7은 본 발명의 일 실시예에 따른 새로운 이미지에 대한 캡션 생성 과정을 설명하기 위한 예시도이다.

도 8은 본 발명의 일 실시예에 따른 사전 학습 모델을 사용한 멀티모달 기반 이미지 캡셔닝 방법을 설명하기 위한 흐름도이다.

도 9는 본 발명의 일 실시예에 따른 한국어 이미지 캡셔닝 모델과 벤치마크 모델 비교 결과이다.

도 10은 본 발명의 일 실시예에 따른 영어 이미지 캡셔닝 모델과 벤치마크 모델 비교 결과이다.

발명을 실시하기 위한 구체적인 내용

[0016] 본 발명에서는 한국어 이미지 캡셔닝 모델의 성능을 개선하는 방법을 제안하고 이를 영어 이미지 캡셔닝 연구에 적용해 방법의 타당성을 검증한다. 한국어 이미지 캡셔닝 모델의 경우 인코더로 최근에 제안된 CLIP(Contrastive Language-Image Pre-training)의 이미지 인코더 그리고 디코더로 사전 학습 언어 모델인 KE-T5의 조합인 '한국어 C-T5' 모델을 제안하고 영어 이미지 캡셔닝 모델의 경우 인코더로 CLIP의 이미지 인코더 그리고 디코더로 사전 학습 언어 모델인 T5의 조합인 '영어 C-T5' 모델을 제안한다. 여기서 CLIP은 수억 개의 이미지-텍스트 쌍의 데이터 셋에서 훈련되어 일반화 능력이 뛰어난 모델이고, KE-T5는 트랜스포머(Transformer) 기반의 T5와 동일한 구조의 한국어 언어 모델이다. 또한, 모델의 성능을 개선하기 위해 이미지와 텍스트 정보를 함께 사용하는 방법을 적용하며 최대로 탐지할 수 있는 객체 태그의 수가 20,000개 이상인 개방형 어휘 객체 탐지 모델인 Detic(Detector with Image Classes)을 사용해 객체 태그를 탐지한다. 이하, 본 발명의 실시 예를 첨부된 도면을 참조하여 상세하게 설명한다.

[0018] 도 1은 본 발명의 일 실시예에 따른 사전 학습 모델을 사용한 멀티모달 기반 이미지 캡셔닝 시스템의 구성을 나타내는 도면이다.

[0019] 제안하는 사전 학습 모델을 사용한 멀티모달 기반 이미지 캡셔닝 시스템은 이미지 인코딩 모듈(110), 태그 인코딩 모듈(120), 캡션 인코딩 모듈(130), 멀티모달 모듈(140), 미세 조정부(150) 및 테스트부(160)를 포함한다.

[0020] 본 발명의 실시예에 따른 이미지 인코딩 모듈(110)은 입력 이미지(111)로부터 시각적 특징을 추출하여 시각적

토큰 임베딩으로 변환한다.

- [0021] 본 발명의 실시예에 따른 이미지 인코딩 모듈(110)은 CLIP(Contrastive Language-Image Pre-training) 이미지 인코더를 통해 입력 이미지(111)로부터 시각적 특징을 추출하고, 상기 추출된 시각적 특징을 프로젝션 네트워크(Projection Network)를 통해 시각적 토큰 임베딩으로 변환한다.
- [0022] 본 발명의 실시예에 따른 태그 인코딩 모듈(120)은 입력 이미지(111)로부터 객체를 탐지하여 객체 태그를 획득한 후 태그 토큰 임베딩으로 변환한다.
- [0023] 본 발명의 실시예에 따른 태그 인코딩 모듈(120)은 객체 탐지 모델을 통해 입력 이미지(111)로부터 객체 태그와 신뢰도 점수를 추출하고, 상기 추출된 객체 태그가 외국어인 경우 한국어로 번역하여, 상기 번역된 객체 태그를 숫자로 변환하는 인코딩 과정을 거친 후 임베딩 계층을 통해 컴퓨터 판독 가능한 벡터인 태그 토큰 임베딩으로 변환한다.
- [0024] 본 발명의 실시예에 따른 캡션 인코딩 모듈(130)은 캡션 데이터를 전처리 한 후 인코딩을 수행하여 텍스트 모달리티를 생성한다.
- [0025] 본 발명의 실시예에 따른 캡션 인코딩 모듈(130)은 사전 학습된 언어 모델을 통해 자연어 형태의 텍스트인 캡션 데이터를 숫자로 변환하는 인코딩을 수행하고, 상기 캡션 데이터에서 특수문자를 제거한 후 상기 캡션 데이터를 최대 길이에 맞추도록 조정 한 후, 토큰라이저를 통해 상기 캡션 데이터를 인코딩된 캡션 토큰으로 변환한다.
- [0026] 본 발명의 실시예에 따른 멀티모달 모듈(140)은 상기 시각적 토큰 임베딩 및 상기 태그 토큰 임베딩을 결합하여 멀티모달 임베딩으로 결합한다.
- [0027] 본 발명의 실시예에 따른 멀티모달 모듈(140)은 상기 이미지 인코딩 과정을 통해 생성된 시각적 토큰 임베딩과 태그 인코딩 과정을 통해 생성된 태그 토큰 임베딩의 서로 다른 모달리티의 임베딩을 결합하기 위해 연결(Concatenation)을 적용하여 상기 시각적 토큰 임베딩과 태그 토큰 임베딩을 결합하여 멀티모달 임베딩을 생성한다.
- [0028] 본 발명의 실시예에 따른 미세 조정부(150)는 상기 결합된 멀티모달 임베딩에 대하여 사전 학습된 언어 모델을 미세 조정한다.
- [0029] 본 발명의 실시예에 따른 미세 조정부(150)는 상기 멀티모달 모듈(140)을 통해 생성된 멀티모달 임베딩과 상기 캡션 인코딩 모듈로부터 인코딩된 캡션 토큰을 입력 받아 사전 학습 언어 모델을 미세 조정한다.
- [0030] 본 발명의 실시예에 따른 미세 조정부(150)는 상기 사전 학습 언어 모델에 대해 현재 시점의 실제값(Ground Truth)을 다음 시점의 입력값으로 사용하여 모델을 훈련하는 교사 강제(Teacher Forcing) 방식으로 학습한다.
- [0031] 본 발명의 실시예에 따른 테스트부(160)은 상기 미세 조정된 사전 학습된 언어 모델을 새로운 이미지에 대해 테스트한다.
- [0032] 이하, 도 2 내지 도 7을 참조하여 본 발명의 일 실시예에 따른 사전 학습 모델을 사용한 멀티모달 기반 이미지 캡셔닝 시스템의 각 구성에 관하여 더욱 상세히 설명한다.
- [0034] 도 2는 본 발명의 일 실시예에 따른 이미지 인코딩 모듈을 설명하기 위한 도면이다.
- [0035] 본 발명의 일 실시예에 따른 이미지 인코딩 모듈의 구조는 도 2와 같다. 입력 이미지(220)로부터 시각적 특징(230)을 추출하기 위해 CLIP(Contrastive Language-Image Pre-training)의 이미지 인코더(220)를 사용 한다. 여기서 본 발명의 실시예에 따른 CLIP의 이미지 인코더는 ResNet-50 구조를 가지는 모델일 수 있다. 본 발명의 실시예에 따른 CLIP(Contrastive Language-Image Pre-training)은 OpenAI에서 제안되었고 이미지 인코더와 텍스트 인코더를 4억 개의 이미지-텍스트 쌍의 데이터 셋에서 공동으로 사전 학습한 모델이다.
- [0036] 이와 같이, 추출된 시각적 특징(230)은 프로젝션 네트워크(Projection Network)(240)를 통해 시각적 토큰 임베딩(250)으로 변환된다. 여기서 본 발명의 실시예에 따른 프로젝션 네트워크는 2개의 선형(Linear) 계층과 Tanh(Hyperbolic Tangent) 활성화 함수로 이루어질 수 있다. 이는 CLIP 임베딩 공간의 시각적 특징을 언어 모델의 임베딩 공간으로 매핑(Mapping) 해주는 역할을 한다.

- [0038] 도 3은 본 발명의 일 실시예에 따른 태그 인코딩 모듈을 설명하기 위한 도면이다.
- [0039] 본 발명의 실시예에 따른 태그 인코딩 모듈의 구조는 도 3과 같다. 입력 이미지(310)로부터 객체 태그(342)와 신뢰도 점수(C Confidence Score)(341)를 얻기 위해 개방형-어휘 객체 탐지 모델인 Detic(320)을 사용한다. 여기서, Detic(Detector with image classes)은 MetaAI에서 제안되었고 탐지할 수 있는 객체의 어휘를 20,000개로 확장한 모델이다. 기존에 사용되었던 객체 탐지 모델인 사전 학습된 Faster R-CNN은 최대 탐지할 수 있는 객체의 수가 1,600개인 반면 Detic은 기존보다 12.5배인 20,000개 이상을 탐지할 수 있다.
- [0040] 본 발명의 실시예에 따르면, 객체 탐지 모델을 통해 탐지된 이미지(330)에 대하여 추출된 객체 태그(342) 중에서 신뢰도 점수(341)가 높은 순서대로 중복되지 않게 9개의 객체 태그를 선택할 수 있다. 객체 태그(342)는 영어로 이루어져 있으므로 한국어 이미지 캡셔닝 모델의 경우 한국어로 번역(350)하는 과정을 적용하고 영어 이미지 캡셔닝 모델의 경우 그대로 사용한다. 선택된 객체 태그(342)는 사전 학습 언어 모델의 토큰라이저(Tokenizer)를 사용해 태그를 숫자로 변환하는 인코딩(360) 과정을 거친다. 이후, 사전 학습 언어 모델의 임베딩 계층(370)을 통해 태그 토큰 임베딩(380)으로 변환된다. 하지만, 이미지로부터 추출되는 객체 태그가 적어 태그 토큰 임베딩의 개수가 9개보다 적은 경우가 존재할 수 있다. 이런 경우 그 차이만큼 0으로 이루어진 임베딩으로 채우는 패딩(Padding) 처리를 적용할 수 있다. 본 발명의 실시예에서는 태그 개수에 따른 모델의 성능을 비교하기 위해, 태그의 개수를 3/5/7/9와 같이 나누어 실험을 진행한다.
- [0042] 도 4는 본 발명의 일 실시예에 따른 멀티모달 모듈을 설명하기 위한 도면이다.
- [0043] 본 발명의 실시예에 따른 멀티모달 모듈의 구조는 도 4와 같다. 멀티모달은 이미지, 텍스트, 음성과 같은 다양한 종류의 데이터가 존재하는 것을 의미한다. 멀티모달 학습은 다양한 유형의 데이터를 알고리즘과 결합하여 더 높은 성능을 얻을 수 있는 학습 방식이며 일반적으로 단일 모달리티(Modality)를 사용하는 모델보다 높은 성능을 보여준다.
- [0044] 본 발명의 실시예에서는 이미지 모달리티와 텍스트 모달리티가 결합한 멀티모달을 사용한다. 여기서, 도 4와 같이 입력 이미지(411)에 대하여 이미지 인코더(412)를 통해 이미지 인코딩 과정을 수행하고, 추출된 시각적 특징 벡터(413)에 대하여 프로젝션 네트워크(Projection Network)(240)를 통해 시각적 토큰 임베딩(415)으로 변환하여 생성된 이미지 모달리티를 사용한다.
- [0045] 본 발명의 실시예에 따르면, 객체 태그(421)가 영어로 이루어져 있으므로 한국어 이미지 캡셔닝 모델의 경우 한국어로 번역(422)하는 과정을 적용하고 영어 이미지 캡셔닝 모델의 경우 그대로 사용한다. 선택된 객체 태그(421)는 사전 학습 언어 모델의 토큰라이저(Tokenizer)를 사용해 태그를 숫자로 변환하는 인코딩(423) 과정을 거친다. 이후, 사전 학습 언어 모델의 임베딩 계층(424)을 통해 태그 토큰 임베딩(425)으로 변환하여 생성된 텍스트 모달리티를 사용한다.
- [0046] 이와 같이, 서로 다른 모달리티의 임베딩을 병합하는 방식으로는 크게 요소별 합산(Element-Wise Summation)과 연결(Concatenation)이 있다. 요소별 합산을 적용하기 위해서는 각 임베딩의 형태(Shape)가 일치해야 하고, 그렇지 않으면 각 임베딩을 동일한 형태로 변환해야 한다.
- [0047] 반면에 연결은 각 임베딩을 단순히 연결하는 방식이다. 이미지 인코딩 과정과 태그 인코딩 과정을 통해 얻은 시각적 토큰 임베딩과 태그 토큰 임베딩은 형태 및 모달리티가 다르다. 또한, 각 모달리티의 특징을 유지하기 위해 연결 방식으로 두 임베딩을 병합해 멀티모달 임베딩(430)을 생성한다.
- [0049] 도 5는 본 발명의 일 실시예에 따른 캡션 인코딩 모듈을 설명하기 위한 도면이다.
- [0050] 도 5를 참조하여, 본 발명의 실시예에 따른 캡션 인코딩 모듈의 동작 과정을 설명한다. 일반적으로 사전 학습 언어 모델은 자연어로 이루어진 캡션을 그대로 입력 받지 않고 인코딩 과정을 거쳐서 각 단어를 고유한 정수로 매핑하여 입력 받는다. 먼저, 사전 학습 언어 모델의 토큰라이저를 사용해 캡션(510)을 인코딩(520)하여 인코딩된 캡션(530)으로 변환하며 구조는 도 5과 같다. 이러한 인코딩된 캡션(530)은 서로 다른 길이를 가지므로 모델 학습에서의 원활한 배치 처리를 위해 시퀀스(Sequence)의 길이를 모두 동일하게 만들어야 한다. 최대 길이보다 짧은 시퀀스의 경우 패딩 처리(padding)를 적용하고 최대 길이보다 긴 시퀀스의 경우 잘라내는 처리(truncate)를 적용한다(540). 여기서 최대 길이는 훈련 데이터 셋의 캡션을 인코딩된 캡션으로 변환하여 길이를 구한 후 길이 분포의 99.5%를 포함하는 길이를 최대 길이로 설정할 수 있다. 이러한 과정을 통해 인코딩된 캡션 토큰

(550)을 생성할 수 있다.

- [0052] 도 6은 본 발명의 일 실시예에 따른 미세 조정부를 설명하기 위한 도면이다.
- [0053] 도 6을 참조하여, 본 발명의 실시예에 따른 미세 조정부를 통해 사전 학습 언어 모델을 미세 조정(Fine-tuning)하는 과정을 설명한다.
- [0054] 멀티모달 모듈을 통해 생성된 멀티모달 임베딩(610)과 캡션 인코딩 모듈을 통해 인코딩된 캡션 토큰(620)을 사용하여 사전 학습된 언어 모델(Pre-Trained Language Model; PTLM)(630)을 미세 조정한다.
- [0055] 본 발명의 실시예에 따르면, 사전 학습 언어 모델은 교사 강제(Teacher Forcing) 방식으로 학습할 수 있다. 여기서 교사 강제란 현재 시점의 실제값(Ground Truth)을 다음 시점의 입력값으로 사용하여 모델을 훈련하는 방식이다.
- [0056] 본 발명의 실시예에 따라 사용하는 사전 학습 언어 모델은 크게 GPT-2 및 KoGPT2인 GPT-2 계열의 모델과 T5 및 KE-T5인 T5 계열의 모델을 사용할 수 있다. 한국어 이미지 캡셔닝 모델의 경우 KoGPT2 및 KE-T5 모델을 사용하고 영어 이미지 캡셔닝 모델의 경우 GPT-2 및 T5 모델을 사용할 수 있다.
- [0057] 본 발명의 실시예에 따른 GPT-2 계열 모델을 미세 조정하는 과정을 설명한다. GPT-2 모델의 토큰라이저를 사용해 캡션을 인코딩 된 캡션 토큰으로 변환한다. 시퀀스의 길이를 동일하게 맞추기 위해 최대 길이를 정하여 패딩 처리 및 잘라내는 처리를 적용한다. KoGPT2 모델의 경우 최대 길이를 24로 정하고 GPT-2 모델의 경우 최대 길이를 22로 설정한다. GPT-2 모델을 미세 조정하기 위해 모델의 입력으로 멀티모달 임베딩과 인코딩 된 캡션 토큰을 사용한다.
- [0058] GPT-2 모델에 멀티모달 임베딩과 인코딩된 캡션 토큰을 함께 입력하기 위해서는 인코딩된 캡션 또한 임베딩으로 변환하는 과정을 거쳐야 한다. 따라서 GPT-2 모델의 임베딩 계층을 통해 인코딩 된 캡션 토큰을 캡션 임베딩으로 변환한다. 동일한 차원(Dimension)을 가지는 멀티모달 임베딩과 캡션 임베딩을 연결하여 GPT-2 모델을 Fine-tuning 하기 위한 입력으로 사용한다. GPT-2 모델을 미세 조정하는 과정에서 교사 강제 방식으로 학습시키며 크로스-엔트로피 손실 함수를 사용한다.
- [0059] 본 발명의 실시예에 따른 미세 조정이 적용된 GPT-2 모델을 사용해 새로운 이미지로부터 캡션을 추론(Inference)하는 과정을 설명한다. 미세 조정이 적용된 GPT-2 모델은 멀티모달 임베딩을 입력 받아 캡션을 생성하는 모델이다. 따라서 이미지로부터 멀티모달 임베딩을 얻기 위해 이미지 인코딩 과정, 태그 인코딩 과정 그리고 멀티모달 과정을 적용한다. 생성된 멀티모달 임베딩을 미세 조정이 적용된 GPT-2 모델에 입력하여 다음 토큰을 하나씩 생성한다. 여기서 캡션 생성을 위해 탐욕적 디코딩(Greedy Decoding) 방식을 적용해 각 타임-스텝(Time-Step)에서 가장 높은 확률을 가지는 토큰을 다음 토큰으로 선택한다. 이러한 방식은 짧은 문장을 생성할 때 잘 작동된다.
- [0060] 다음으로, 본 발명의 실시예에 따른 T5 계열 모델을 미세 조정하는 과정을 설명한다. T5 모델 또한 해당 토큰라이저를 사용해 캡션을 인코딩 된 캡션 토큰으로 변환한다. 시퀀스의 길이를 동일하게 맞추기 위해 최대 길이를 정하여 패딩 및 잘라내는 처리를 적용한다. KE-T5 모델의 경우 최대 길이를 22로 정하고 T5 모델의 경우 최대 길이를 29로 설정한다. T5 모델을 미세 조정하기 위해 모델의 입력으로 멀티모달 임베딩과 인코딩 된 캡션 토큰을 사용한다. T5 모델은 트랜스포머(Transformer)와 유사한 인코더-디코더 구조를 가지므로 GPT-2 모델의 미세 조정 과정과 다른 점이 있다. T5 모델의 인코더는 멀티모달 임베딩을 입력 받고 T5의 디코더는 인코딩 된 캡션 토큰을 입력 받는다. 그러므로 T5 모델의 임베딩 계층을 사용해 인코딩 된 캡션 토큰을 캡션 임베딩으로 변환할 필요가 없다. 대신, T5 모델은 미세 조정 과정에서 특정 작업에 대한 정보가 담긴 접두사(Prefix)와 함께 미세 조정되므로 본 발명에서도 'convert image to caption'이라는 작업 접두사를 사용한다. 작업 접두사는 멀티모달 임베딩과 함께 연결되어 T5 모델의 인코더에 입력되어야 한다. 따라서 자연어로 이루어진 작업 접두사를 T5 모델의 토큰라이저를 사용해 인코딩 된 접두사 토큰으로 변환한 후 T5 모델의 임베딩 계층을 통해 접두사 임베딩으로 변환한다. 동일한 차원을 가지는 접두사 임베딩과 멀티모달 임베딩은 서로 연결되어 T5 모델의 인코더에 입력되고 인코딩 된 캡션 토큰은 T5 모델의 디코더에 입력된다. T5 모델 또한 미세 조정하는 과정에서 교사 강제 방식으로 학습시키며 크로스-엔트로피 손실 함수를 사용한다.
- [0061] 본 발명의 실시예에 따른 미세 조정이 적용된 T5 모델을 사용해 새로운 이미지로부터 캡션을 추론하는 과정을 설명한다. T5 모델을 미세 조정하는 과정에서 작업 접두사를 사용했으므로 추론 과정에서도 작업 접두사를 사용

한다. 미세 조정이 적용된 T5 모델은 인코더에서 접두사 임베딩과 멀티모달 임베딩을 입력 받고 디코더로부터 캡션을 생성하는 모델이다. 따라서 접두사 임베딩을 얻은 후 이미지로부터 멀티모달 임베딩을 얻기 위해 이미지 인코딩 과정, 태그 인코딩 과정 그리고 멀티모달 과정을 적용한다. 생성된 접두사 임베딩과 멀티모달 임베딩을 연결한 후 미세 조정이 적용된 T5 모델의 인코더에 입력하고 디코더는 PAD 토큰을 시작 시퀀스 토큰으로 사용한다. 캡션 생성을 위해 GPT-2와 같이 탐욕적 디코딩 방식을 적용한다.

[0063] 도 7은 본 발명의 일 실시예에 따른 새로운 이미지에 대한 캡션 생성 과정을 설명하기 위한 예시도이다.

[0064] 도 7과 같이, 미세 조정된 T5 모델을 사용해 새로운 이미지(511)로부터 캡션(550)을 생성할 수 있다.

[0065] 새로운 이미지(511)로부터 이미지 인코딩 모듈(510)을 통해 생성된 이미지 모달리티와 태그 인코딩 모듈(520)을 통해 생성된 텍스트 모달리티를 멀티모달 모듈(530)을 통해 결합한다. 본 발명의 실시예에 따른 미세 조정된 T5 모델(540)은 인코더에서 멀티모달 임베딩(530)을 입력 받아 미세 조정된 T5 모델(540)을 통해 디코더로부터 캡션(550)을 생성할 수 있다.

[0067] 도 8은 본 발명의 일 실시예에 따른 사전 학습 모델을 사용한 멀티모달 기반 이미지 캡셔닝 방법을 설명하기 위한 흐름도이다.

[0068] 제안하는 사전 학습 모델을 사용한 멀티모달 기반 이미지 캡셔닝 방법은 이미지 인코딩 모듈을 통해 입력 이미지로부터 시각적 특징을 추출하여 시각적 토큰 임베딩으로 변환하는 단계(810), 태그 인코딩 모듈을 통해 입력 이미지로부터 객체를 탐지하여 객체 태그를 획득한 후 태그 토큰 임베딩으로 변환하는 단계(820), 캡션 인코딩 모듈을 통해 캡션 데이터를 전처리 한 후 인코딩을 수행하여 텍스트 모달리티를 생성하는 단계(830), 멀티모달 모듈을 통해 상기 시각적 토큰 임베딩 및 상기 태그 토큰 임베딩을 결합하여 멀티모달 임베딩으로 결합하는 단계(840), 미세 조정을 통해 상기 결합된 멀티모달 임베딩에 대하여 사전 학습된 언어 모델을 미세 조정하는 단계(850) 및 테스트부를 통해 상기 미세 조정된 사전 학습된 언어 모델을 새로운 이미지에 대해 테스트하는 단계(860)를 포함한다.

[0069] 단계(810)에서, 이미지 인코딩 모듈을 통해 입력 이미지로부터 시각적 특징을 추출하여 시각적 토큰 임베딩으로 변환한다.

[0070] 본 발명의 실시예에 따른 이미지 인코딩 모듈은 CLIP(Contrastive Language-Image Pre-training) 이미지 인코더를 통해 입력 이미지로부터 시각적 특징을 추출하고, 상기 추출된 시각적 특징을 프로젝션 네트워크(Projection Network)를 통해 시각적 토큰 임베딩으로 변환한다.

[0071] 단계(820)에서, 태그 인코딩 모듈을 통해 입력 이미지로부터 객체를 탐지하여 객체 태그를 획득한 후 태그 토큰 임베딩으로 변환한다.

[0072] 본 발명의 실시예에 따른 태그 인코딩 모듈은 객체 탐지 모델을 통해 입력 이미지로부터 객체 태그와 신뢰도 점수를 추출하고, 상기 추출된 객체 태그가 외국어인 경우 한국어로 번역하여, 상기 번역된 객체 태그를 숫자로 변환하는 인코딩 과정을 거친 후 임베딩 계층을 통해 컴퓨터 판독 가능한 벡터인 태그 토큰 임베딩으로 변환한다.

[0073] 단계(830)에서, 캡션 인코딩 모듈을 통해 캡션 데이터를 전처리 한 후 인코딩을 수행하여 텍스트 모달리티를 생성한다.

[0074] 본 발명의 실시예에 따른 캡션 인코딩 모듈은 사전 학습된 언어 모델을 통해 자연어 형태의 텍스트인 캡션 데이터를 숫자로 변환하는 인코딩을 수행하고, 상기 캡션 데이터에서 특수문자를 제거한 후 상기 캡션 데이터를 최대 길이에 맞도록 조정된 후, 토큰라이저를 통해 상기 캡션 데이터를 인코딩된 캡션 토큰으로 변환한다.

[0075] 단계(840)에서, 멀티모달 모듈을 통해 상기 시각적 토큰 임베딩 및 상기 태그 토큰 임베딩을 결합하여 멀티모달 임베딩으로 결합한다.

[0076] 본 발명의 실시예에 따른 멀티모달 모듈은 상기 이미지 인코딩 과정을 통해 생성된 시각적 토큰 임베딩과 태그 인코딩 과정을 통해 생성된 태그 토큰 임베딩의 서로 다른 모달리티의 임베딩을 결합하기 위해 연결(Concatenation)을 적용하여 상기 시각적 토큰 임베딩과 태그 토큰 임베딩을 결합하여 멀티모달 임베딩을 생성

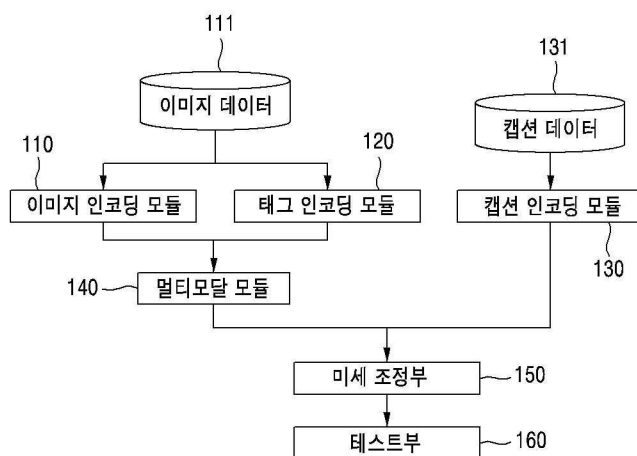
한다.

- [0077] 단계(850)에서, 미세 조정부를 통해 상기 결합된 멀티모달 임베딩에 대하여 사전 학습된 언어 모델을 미세 조정한다.
- [0078] 본 발명의 실시예에 따른 미세 조정부는 상기 멀티모달 모듈을 통해 생성된 멀티모달 임베딩과 상기 캡션 인코딩 모듈로부터 인코딩된 캡션 토큰을 입력 받아 사전 학습 언어 모델을 미세 조정한다.
- [0079] 본 발명의 실시예에 따른 미세 조정부는 상기 사전 학습 언어 모델에 대해 현재 시점의 실제값(Ground Truth)을 다음 시점의 입력값으로 사용하여 모델을 훈련하는 교사 강제(Teacher Forcing) 방식으로 학습한다.
- [0080] 단계(860)에서, 테스트부를 통해 상기 미세 조정된 사전 학습된 언어 모델을 새로운 이미지에 대해 테스트한다.
- [0082] 도 9는 본 발명의 일 실시예에 따른 한국어 이미지 캡셔닝 모델과 벤치마크 모델 비교 결과이다.
- [0083] 본 발명의 실시예에 따른 인코더로 사전 학습된 이미지 분류 모델인 ResNet-152 그리고 디코더로 적응형 어텐션을 적용한 LSTM의 조합을 사용하여 테스트를 수행하였다.
- [0084] 제안하는 한국어 이미지 캡셔닝 모델과 벤치마크 모델을 훈련하기 위해 사용한 데이터 셋은 기존 MS-COCO 데이터 셋의 한국어 버전을 사용하였다. AI 허브에서 한국어 MS-COCO 데이터 셋을 제공하지만 번역에 오류가 존재하여 수작업으로 번역 오류를 수정하였다. 영어 이미지 캡셔닝 모델은 영어 MS-COCO 데이터 셋을 사용하였다.
- [0085] 제안하는 모델과 벤치마크 모델을 평가하기 위해 CIDEr, SPICE, BLEU-4 그리고 METEOR 지표를 사용하였으며 각 지표는 높을수록 모델의 성능이 우수한 것을 의미한다. 도 9와 같이, 본 발명에서 제안하는 '한국어 C-T5' 모델의 경우 기존 한국어 이미지 캡셔닝의 벤치마크 모델과 비교해 CIDEr, SPICE, BLEU-4, METEOR 지표에서 각각 19.3, 3.3, 4.8, 3.2만큼 개선된 성능을 보였다.
- [0087] 도 10은 본 발명의 일 실시예에 따른 영어 이미지 캡셔닝 모델과 벤치마크 모델 비교 결과이다.
- [0088] 도 10과 같이, 본 발명에서 제안하는 '영어 C-T5' 모델의 경우 기존 영어 이미지 캡셔닝의 벤치마크 모델과 비교해 CIDEr 지표에서 119.9로 가장 높은 성능을 보였다. '영어 C-T5' 모델은 적은 노력에도 불구하고 사전 훈련(Pre-training)을 적용한 벤치마크 모델과 유사한 성능을 보였다.
- [0090] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.
- [0091] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치에 구체화(embodiment)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

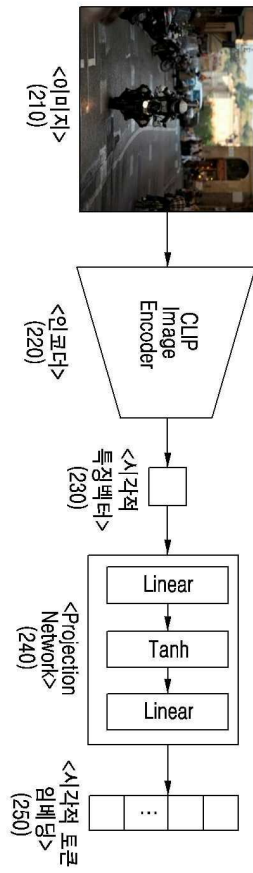
- [0092] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic 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 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.
- [0093] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.
- [0094] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

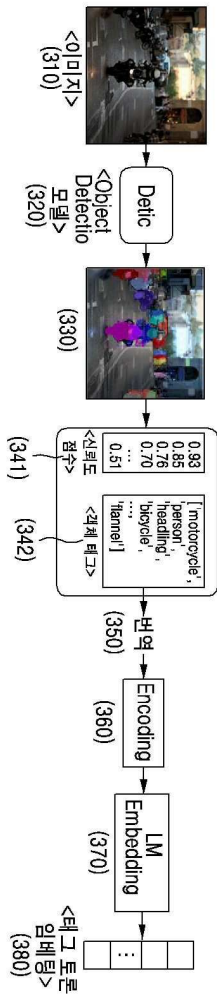
도면1



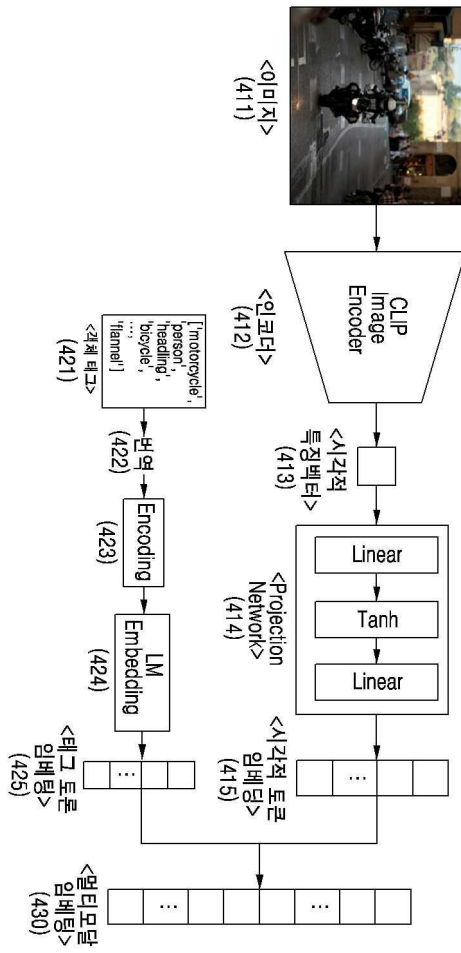
도면2



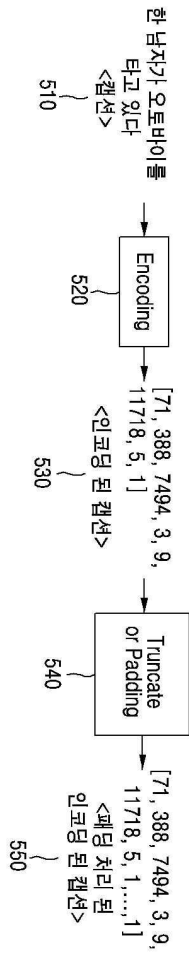
도면3



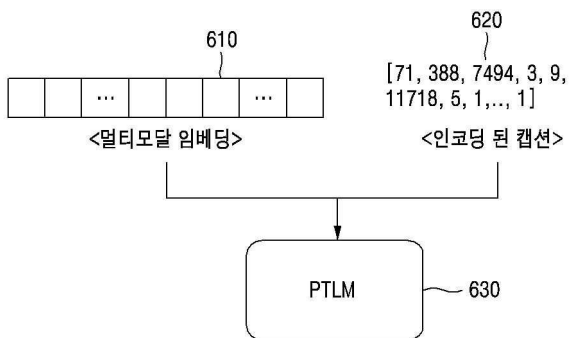
도면4



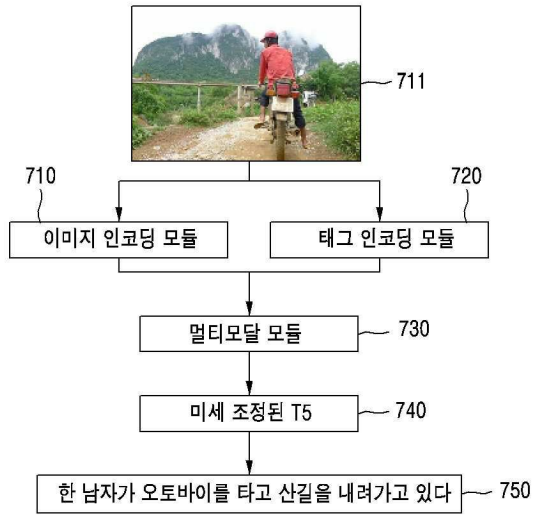
도면5



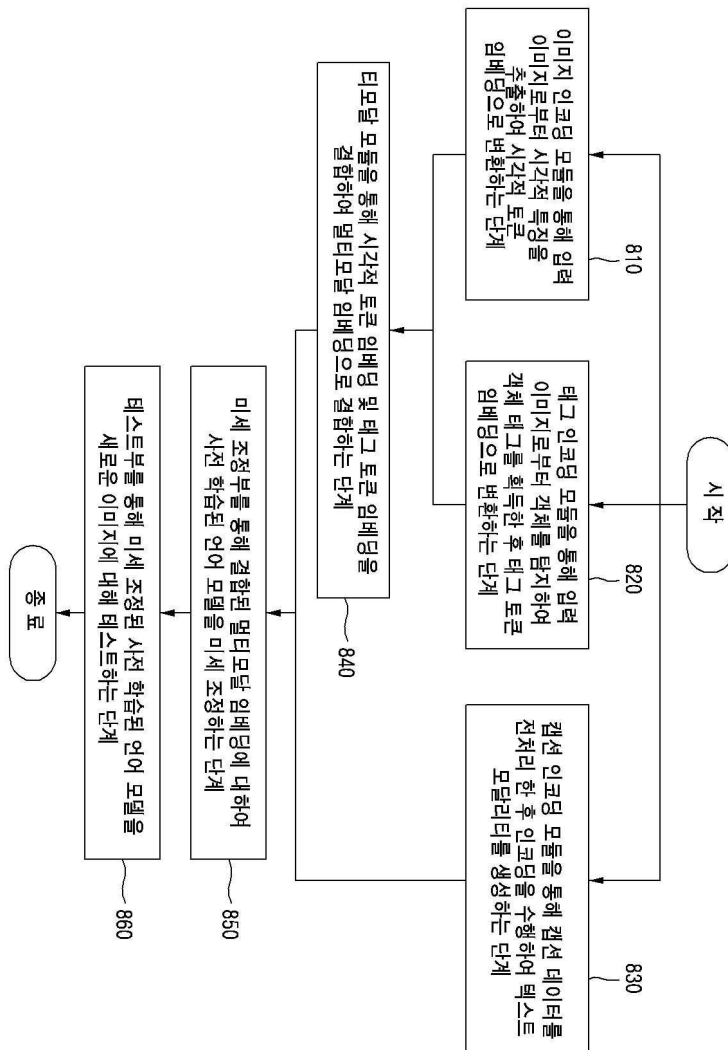
도면6



도면7



도면8



도면9

구분	인코더	디코더	CIDEr	SPICE	BLEU-4	METEOR
벤치마크 모델	ResNet-152	LSTM	79.1	18.6	20.6	26.2
제안하는 모델	CLIP의 이미지 인코더	KE-T5	98.4	21.9	25.4	29.4

도면10

구분	인코더	디코더	CIDEr	SPICE	BLEU-4	METEOR
벤치마크 모델	ResNet-152	LSTM	108.5	-	33.2	26.6
제안하는 모델	CLIP의 이미지 인코더	T5	119.9	22.1	35.0	28.7