(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0085694
(43) 공개일자 2021년07월08일

(51) 국제특허분류(Int. Cl.)
G06F 40/169 (2020.01) *G06K 9/46* (2006.01)
G06N 3/08 (2006.01)
(52) CPC특허분류
G06F 40/169 (2020.01)
G06K 9/46 (2013.01)
(21) 출원번호 10-2019-0179027
(22) 출원일자 2019년12월31일
심사청구일자 없음

(71) 출원인
한국전력공사
전라남도 나주시 전력로 55(빛가람동)
성균관대학교산학협력단
경기도 수원시 장안구 서부로 2066 (천천동, 성균관대학교내)
(72) 발명자
류은석
서울특별시 강남구 선릉로69길 20, e편한세상 아파트 106동 303호
박은수
경기도 의정부시 호동로 72, 201동 1202호 (호원동, 호원가든 2차 아파트)
(74) 대리인
특허법인아주

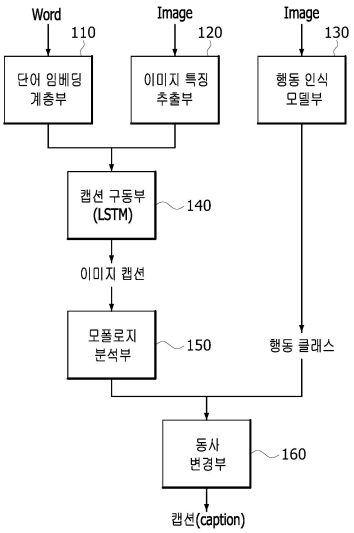
전체 청구항 수 : 총 7 항

(54) 발명의 명칭 **이미지 캡셔닝 장치 및 방법**

(57) 요 약

본 발명은 이미지 캡셔닝 장치에 관한 것으로, 이미지 캡셔닝을 위한 학습된 단어들을 저장하는 단어 임베딩 계층부; 이미지를 입력받아 기 학습된 특징 추출 방식에 따라 이미지의 특징을 추출하는 이미지 특징 추출부; 상기 이미지를 동일하게 입력받아 기 학습된 행동 인식 방식에 따라 이미지에서 행동 특징을 구분하는 행동 인식 모델부; 상기 이미지 특징 추출부에서 추출된 특징에 대응하는 연관 단어들을 상기 단어 임베딩 계층부에서 추출하여 캡셔닝을 수행하는 캡션 구동부; 상기 캡션 구동부에서 캡셔닝된 문장에서 각 단어에 대한 품사를 분석하는 모폴로지 분석부; 및 상기 캡셔닝된 문장에서 동사에 대하여, 상기 행동 인식 모델부에서 추출한 이미지의 행동에 일치하는 동사인지 체크하여, 지정된 체크 조건을 만족하지 않을 경우, 상기 캡셔닝된 문장에서 동사를 상기 행동에 대응하는 동사로 최종 캡셔닝 문장을 변경하는 동사 변경부;를 포함한다.

대 표 도 - 도2



(52) CPC특허분류
G06N 3/08 (2013.01)

명 세 서

청구범위

청구항 1

이미지 캡셔닝을 위한 학습된 단어들을 저장하는 단어 임베딩 계층부;

이미지를 입력받아 기 학습된 특징 추출 방식에 따라 이미지의 특징을 추출하는 이미지 특징 추출부;

상기 이미지를 동일하게 입력받아 기 학습된 행동 인식 방식에 따라 이미지에서 행동 특징을 구분하는 행동 인식 모델부;

상기 이미지 특징 추출부에서 추출된 특징에 대응하는 연관 단어들을 상기 단어 임베딩 계층부에서 추출하여 캡셔닝을 수행하는 캡션 구동부;

상기 캡션 구동부에서 캡셔닝된 문장에서 각 단어에 대한 품사를 분석하는 모폴로지 분석부; 및

상기 캡셔닝된 문장에서 동사에 대하여, 상기 행동 인식 모델부에서 추출한 이미지의 행동에 일치하는 동사인지 체크하여, 지정된 체크 조건을 만족하지 않을 경우, 상기 캡셔닝된 문장에서 동사를 상기 행동에 대응하는 동사로 최종 캡셔닝 문장을 변경하는 동사 변경부;를 포함하는 것을 특징으로 하는 이미지 캡셔닝 장치.

청구항 2

제 1항에 있어서, 상기 캡션 구동부는,

복수의 LSTM(Long Short Term Memory)을 포함하는 것을 특징으로 하는 이미지 캡셔닝 장치.

청구항 3

제 1항에 있어서, 상기 동사 변경부는,

지정된 단어의 손실(loss) 값이 지정된 값 보다 작고, 상기 지정된 단어의 품사가 동사이며, 상기 행동 인식 모델부에서 인식한 행동 인식 정확도가 지정된 기준보다 큰 경우인지 체크하고, 상기 체크 조건을 모두 만족하는 경우 상기 행동 인식 모델부에서 인식한 행동에 대응하는 단어로 최종 캡션에서 동사를 변경하는 것을 특징으로 하는 이미지 캡셔닝 장치.

청구항 4

제 1항에 있어서, 상기 캡셔닝 문장의 복수의 품사에 해당하는 단어를 변경하거나 추가하기 위하여,

이미지에서 얼굴을 인식하는 얼굴 인식 모델; 및 이미지를 촬영한 카메라의 위치 데이터를 이용하여 위치를 인식하는 위치 데이터 모델;을 더 포함하고,

상기 캡셔닝 문장의 기 지정된 위치에 얼굴 인식된 사용자(user)의 이름을 추가하고, 상기 캡셔닝 문장의 기 지정된 위치에 위치 데이터(Location)의 명칭을 더 추가하도록 구현된 것을 특징으로 하는 이미지 캡셔닝 장치.

청구항 5

단어 임베딩 계층부에 이미지 캡셔닝을 위한 학습된 단어들을 저장하는 단계;

이미지 특징 추출부가 이미지를 입력받아 기 학습된 특징 추출 방식에 따라 이미지의 특징을 추출하는 단계;

행동 인식 모델부가 상기 이미지를 동일하게 입력받아 기 학습된 행동 인식 방식에 따라 이미지에서 행동 특징

을 구분하는 단계;

캡션 구동부가 상기 이미지 특징 추출부에서 추출된 특징에 대응하는 연관 단어들을 상기 단어 임베딩 계층부에서 추출하여 캡셔닝을 수행하는 단계;

모폴로지 분석부가 상기 캡션 구동부에서 캡셔닝된 문장에서 각 단어에 대한 품사를 분석하는 단계; 및

동사 변경부가 상기 캡셔닝된 문장에서 동사에 대하여, 상기 행동 인식 모델부에서 추출한 이미지의 행동에 일치하는 동사인지 체크하여, 지정된 체크 조건을 만족하지 않을 경우, 상기 캡셔닝된 문장에서 동사를 상기 행동에 대응하는 동사로 최종 캡셔닝 문장을 변경하는 단계;를 포함하는 것을 특징으로 하는 이미지 캡셔닝 방법.

청구항 6

제 5항에 있어서, 상기 캡셔닝된 문장에서 동사에 대하여, 상기 행동 인식 모델부에서 추출한 이미지의 행동에 일치하는 동사인지 체크하는 단계에서,

상기 동사 변경부는,

지정된 단어의 손실(loss) 값이 지정된 값 보다 작고, 상기 지정된 단어의 품사가 동사이며, 상기 행동 인식 모델부에서 인식한 행동 인식 정확도가 지정된 기준보다 큰 경우인지 체크하고, 상기 체크 조건을 모두 만족하는 경우 상기 행동 인식 모델부에서 인식한 행동에 대응하는 단어로 최종 캡션에서 동사를 변경하는 것을 특징으로 하는 이미지 캡셔닝 방법.

청구항 7

제 5항에 있어서, 상기 캡셔닝 문장의 복수의 품사에 해당하는 단어를 변경하거나 추가하기 위하여,

얼굴 인식 모델을 통해 이미지에서 얼굴을 인식하는 단계; 및 위치 데이터 모델을 통해 이미지를 촬영한 카메라의 위치 데이터를 이용하여 위치를 인식하는 단계;를 더 포함하고,

상기 캡셔닝 문장의 기 지정된 위치에 얼굴 인식된 사용자(user)의 이름을 추가하고, 상기 캡셔닝 문장의 기 지정된 위치에 위치 데이터(Location)의 명칭을 더 추가하도록 구현된 것을 특징으로 하는 이미지 캡셔닝 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 이미지 캡셔닝 장치 및 방법에 관한 것으로, 보다 상세하게는 자연어 처리 기반의 사용자의 행동 인식 묘사에 특화된 이미지 캡셔닝 장치 및 방법에 관한 것이다.

배 경 기 술

[0003] 최근 고성능 GPU(Graphic Processing Unit)의 사용으로 처리 가능한 연산량이 대폭 증가함에 따라, 패턴을 인식하는데 연산량이 많이 필요한 딥 러닝 기술이 계속 연구 되고 있다.

[0004] 예컨대 합성곱신경망(Convolution neural network, CNN)과 같은 신경망의 발달과 함께 객체 인식, 이미지 분류 등과 같은 이미지 프로세싱 연구가 상당히 빠른 속도로 진행되고 있다. 가령 헬스케어 분야의 딥 러닝 기술 적용으로 인하여, 사람이 포함된 영상 이해, 상황 인식과 같은 연구가 심도 있게 진행되어 오면서, 딥 러닝 기반의 영상 캡셔닝의 중요도가 부각되어 오고 있다.

[0005] 여기서 이미지 캡셔닝 (Image captioning)이란, 입력된 이미지를 합성곱신경망을 통하여 특징을 추출하고, 학습된 단어 특징 공간에 매핑하여 입력된 이미지의 설명문을 생산하는 기술로서, 영상 이해 및 상황 인식에 가장 근접한 연구 중 하나이다.

[0006] 그런데 기존의 이미지 캡셔닝 모델을 사용하는 경우, 한 장의 이미지로 행동을 예측하는 부분에서 큰 어려움이 있을 수 있다. 예컨대 기존의 이미지 캡셔닝 모델을 사용하는 경우, 같은 행동이지만 다르게 예측될 수 있다.

가령 도 1에 도시된 바와 같이, 테이블에서 팔씨름을 하고 있는 이미지가 이미지 캡셔닝 장치에 입력되었다고 가정할 때, 상기 이미지에 대하여 “두 사람이 팔씨름을 하고 있다(Two men are doing arm wrestling)”및 “두 사람이 악수를 하고 있다(Two men are shaking hands)”로 다르게 예측될 수 있는 문제점이 있다.

[0007] 본 발명의 배경기술은 대한민국 공개특허 10-2017-0007747호(2017.01.20. 공개, 자연어 이미지 검색 기법)에 개시되어 있다.

발명의 내용

해결하려는 과제

[0009] 본 발명의 일 측면에 따르면, 본 발명은 상기와 같은 문제점을 해결하기 위해 창작된 것으로서, 자연어 처리 기반의 사용자의 행동 인식 묘사에 특화된 이미지 캡셔닝 장치 및 방법을 제공하는 데 그 목적이 있다.

과제의 해결 수단

[0011] 본 발명의 일 측면에 따른 이미지 캡셔닝 장치는, 이미지 캡셔닝을 위한 학습된 단어들을 저장하는 단어 임베딩 계층부; 이미지를 입력받아 기 학습된 특징 추출 방식에 따라 이미지의 특징을 추출하는 이미지 특징 추출부; 상기 이미지를 동일하게 입력받아 기 학습된 행동 인식 방식에 따라 이미지에서 행동 특징을 구분하는 행동 인식 모델부; 상기 이미지 특징 추출부에서 추출된 특징에 대응하는 연관 단어들을 상기 단어 임베딩 계층부에서 추출하여 캡셔닝을 수행하는 캡션 구동부; 상기 캡션 구동부에서 캡셔닝된 문장에서 각 단어에 대한 품사를 분석하는 모폴로지 분석부; 및 상기 캡셔닝된 문장에서 동사에 대하여, 상기 행동 인식 모델부에서 추출한 이미지의 행동에 일치하는 동사인지 체크하여, 지정된 체크 조건을 만족하지 않을 경우, 상기 캡셔닝된 문장에서 동사를 상기 행동에 대응하는 동사로 최종 캡셔닝 문장을 변경하는 동사 변경부;를 포함하는 것을 특징으로 한다.

[0012] 본 발명에 있어서, 상기 캡션 구동부는, 복수의 LSTM(Long Short Term Memory)을 포함하는 것을 특징으로 한다.

[0013] 본 발명에 있어서, 상기 동사 변경부는, 지정된 단어의 손실(loss) 값이 지정된 값 보다 작고, 상기 지정된 단어의 품사가 동사이며, 상기 행동 인식 모델부에서 인식한 행동 인식 정확도가 지정된 기준보다 큰 경우인지 체크하고, 상기 체크 조건을 모두 만족하는 경우 상기 행동 인식 모델부에서 인식한 행동에 대응하는 단어로 최종 캡션에서 동사를 변경하는 것을 특징으로 한다.

[0014] 본 발명에 있어서, 상기 캡셔닝 문장의 복수의 품사에 해당하는 단어를 변경하거나 추가하기 위하여, 이미지에서 얼굴을 인식하는 얼굴 인식 모델; 및 이미지를 촬영한 카메라의 위치 데이터를 이용하여 위치를 인식하는 위치 데이터 모델;을 더 포함하고, 상기 캡셔닝 문장의 기 지정된 위치에 얼굴 인식된 사용자(user)의 이름을 추가하고, 상기 캡셔닝 문장의 기 지정된 위치에 위치 데이터(Location)의 명칭을 더 추가하도록 구현된 것을 특징으로 한다.

[0016] 본 발명의 다른 측면에 따른 이미지 캡셔닝 방법은, 단어 임베딩 계층부에 이미지 캡셔닝을 위한 학습된 단어들을 저장하는 단계; 이미지 특징 추출부가 이미지를 입력받아 기 학습된 특징 추출 방식에 따라 이미지의 특징을 추출하는 단계; 행동 인식 모델부가 상기 이미지를 동일하게 입력받아 기 학습된 행동 인식 방식에 따라 이미지에서 행동 특징을 구분하는 단계; 캡션 구동부가 상기 이미지 특징 추출부에서 추출된 특징에 대응하는 연관 단어들을 상기 단어 임베딩 계층부에서 추출하여 캡셔닝을 수행하는 단계; 모폴로지 분석부가 상기 캡션 구동부에서 캡셔닝된 문장에서 각 단어에 대한 품사를 분석하는 단계; 및 동사 변경부가 상기 캡셔닝된 문장에서 동사에 대하여, 상기 행동 인식 모델부에서 추출한 이미지의 행동에 일치하는 동사인지 체크하여, 지정된 체크 조건을 만족하지 않을 경우, 상기 캡셔닝된 문장에서 동사를 상기 행동에 대응하는 동사로 최종 캡셔닝 문장을 변경하는 단계;를 포함하는 것을 특징으로 한다.

[0017] 본 발명에 있어서, 상기 캡셔닝된 문장에서 동사에 대하여, 상기 행동 인식 모델부에서 추출한 이미지의 행동에 일치하는 동사인지 체크하는 단계에서, 상기 동사 변경부는, 지정된 단어의 손실(loss) 값이 지정된 값 보다 작고, 상기 지정된 단어의 품사가 동사이며, 상기 행동 인식 모델부에서 인식한 행동 인식 정확도가 지정된 기준보다 큰 경우인지 체크하고, 상기 체크 조건을 모두 만족하는 경우 상기 행동 인식 모델부에서 인식한 행동에

대응하는 단어로 최종 캡션에서 동사를 변경하는 것을 특징으로 한다.

[0018] 본 발명에 있어서, 상기 캡셔닝 문장의 복수의 품사에 해당하는 단어를 변경하거나 추가하기 위하여, 얼굴 인식 모델을 통해 이미지에서 얼굴을 인식하는 단계; 및 위치 데이터 모델을 통해 이미지를 촬영한 카메라의 위치 데이터를 이용하여 위치를 인식하는 단계;를 더 포함하고, 상기 캡셔닝 문장의 기 지정된 위치에 얼굴 인식된 사용자(user)의 이름을 추가하고, 상기 캡셔닝 문장의 기 지정된 위치에 위치 데이터(Location)의 명칭을 더 추가하도록 구현된 것을 특징으로 한다.

발명의 효과

[0020] 본 발명의 일 측면에 따르면, 본 발명은 자연어 처리 기반의 사용자의 행동 인식 묘사에 특화된 이미지 캡셔닝 장치 및 방법에 관한 것으로, 이미지 캡셔닝의 정확도를 향상시키는 효과가 있다.

도면의 간단한 설명

[0022] 도 1은 종래의 이미지 캡셔닝의 문제점을 설명하기 위하여 보인 예시도.

도 2는 본 발명의 일 실시예에 따른 이미지 캡셔닝 장치의 개략적인 구성을 보인 예시도.

도 3은 상기 도 2에 있어서, 동사 변경부의 동사 교체 과정을 설명하기 위하여 보인 예시도.

도 4는 상기 도 3에 있어서, 동사 변경을 위한 조건을 만족하는지 체크하는 알고리즘을 보인 예시도.

도 5는 상기 도 2에 있어서, 품사 분석을 위한 자연어 처리 모듈(NLTK)의 동사 태그의 종류를 보인 예시도.

도 6은 본 발명의 다른 실시예에 따라 복수의 품사에 해당하는 단어를 변경하거나 추가할 수 있는 캡셔닝 장치의 개략적인 구성을 보인 예시도.

발명을 실시하기 위한 구체적인 내용

[0023] 이하, 첨부된 도면을 참조하여 본 발명에 따른 이미지 캡셔닝 장치 및 방법의 일 실시예를 설명한다.

[0024] 이 과정에서 도면에 도시된 선들의 두께나 구성요소의 크기 등은 설명의 명료성과 편의상 과장되게 도시되어 있을 수 있다. 또한, 후술되는 용어들은 본 발명에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례에 따라 달라질 수 있다. 그러므로 이러한 용어들에 대한 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다.

[0025] 상술한 바와 같이 종래에는 한 장의 이미지를 입력하는 이미지 캡셔닝 과정에서 시간적 요소가 필요한 인간의 행동 관련 요소(즉, 시간이 경과됨에 따라 수행되는 행동)를 추측하는 데 어려움이 있었다. 따라서 본 실시예에서는 상기 종래의 문제점을 해결하기 위하여 행동인식 모델을 추가로 이용하여 캡셔닝을 수행함으로써, 이미지 캡셔닝의 정확도를 향상시키기 위한 것이다.

[0026] 도 2는 본 발명의 일 실시예에 따른 이미지 캡셔닝 장치의 개략적인 구성을 보인 예시도이다.

[0027] 도 2에 도시된 바와 같이, 본 실시예에 따른 이미지 캡셔닝 장치는, 단어 임베딩 계층부(110), 이미지 특징 추출부(120), 행동 인식 모델부(130), 캡션 구동부(140), 모폴로지 분석부(150), 및 동사 변경부(160)를 포함한다.

[0028] 상기 단어 임베딩 계층부(110)는 캡셔닝을 위한 단어들을 저장한다.

[0029] 예컨대 상기 단어 임베딩 계층부(110)는 학습된 단어들이 저장되는 일종의 데이터베이스에 해당한다.

[0030] 상기 이미지 특징 추출부(120)는 입력받은 이미지의 특징을 추출한다.

[0031] 상기 이미지 특징 추출부(120)가 이미지의 특징을 추출하는 방식은 이미 학습된 특징 추출 방식에 기초한다.

[0032] 상기 행동 인식 모델부(130)는 상기 이미지 특징 추출부(120)에 입력된 동일한 이미지를 입력받아 이미 학습된 행동 인식 방식에 기초하여 상기 이미지에서 행동 특징을 구분한다.

[0033] 상기 캡션 구동부(140)는 상기 이미지 특징 추출부(120)에서 추출된 특징에 대응하는 연관 단어들을 상기 단어

임베딩 계층부(110)에서 추출하여 이미지 캡셔닝을 수행한다. 즉, 상기 이미지의 특징을 문장 형태로 표시한다.

[0034] 상기 캡션 구동부(140)는 복수의 LSTM(Long Short Term Memory)을 포함하고, 상기 LSTM에서 실질적인 캡셔닝 동작을 수행한다.

[0035] 상기 모폴로지 분석부(150)는 상기 캡션 구동부(140)에서 캡셔닝된 문장에서 각 단어에 대한 품사를 분석한다. 예컨대 상기 캡셔닝된 문장에서 주어, 동사, 목적어, 및 보어 등을 구분한다.

[0036] 상기 동사 변경부(160)는 상기 캡셔닝된 문장에서 동사에 대하여, 상기 행동 인식 모델부(130)에서 추출한 이미지의 행동과 일치하는 동사인지 체크하여, 지정된 조건을 만족하지 않을 경우, 상기 캡셔닝된 문장에서 동사를 상기 행동에 대응하는 동사로 최종 캡셔닝 문장을 변경한다(도 3 참조).

[0037] 상기 캡션 구동부(140)의 동작에 대한 이해를 돕기 위하여 좀 더 구체적으로 설명한다.

[0038] 상기 캡션 구동부(140)는 합성곱신경망(CNN)을 사용하여 이미지의 특징을 인코딩하고, 상기 인코딩된 이미지의 특징은 단어의 특징이 임베딩 되어 있는 디코딩 단계에 입력된다. 그리고 상기 캡션 구동부(140)는 상기 디코딩 단계에서 입력된 특징 벡터 공간을 사용하여 상기 이미지의 특징에 매칭되는 단어를 추출한다. 이때 상기 캡션 구동부(140)는 입력받은 한 장의 이미지에서 행동, 및 상황을 포함한 복수의 이미지의 특징을 하나의 합성곱신경망(CNN)으로 추정해야 하므로, 정확도가 상당히 떨어질 수 있다.

[0039] 따라서 본 실시예에서는 행동 인식 데이터 셋으로 미리 학습된 행동 인식 모델부(130)를 통해 부족한 부분(예 : 행동에 관련된 특징)을 보완한다. 여기서 상기 부족한 부분(예 : 행동에 관련된 특징)은 동사에 해당하는 부분이므로, 동사에 해당하는 단어의 위치를 알기 위하여, 상기 모폴로지 분석부(150)가 상기 이미지 캡셔닝에서 출력된 캡션(즉, 캡셔닝된 문장)의 각 단어에 대한 품사(part-of-speech, POS)를 분석한다.

[0040] 그리고 상기 동사 변경부(160)가 상기 생성된 캡션(즉, 캡셔닝된 문장)의 손실 값으로 교체 여부를 판단한다.

[0041] 이때 동사의 손실 값이 아닌, 다른 품사의 손실 값이 낮아서 발생하는 오류를 방지하기 위하여, 도 3 및 도 4에 도시된 바와 같은 알고리즘을 이용하여 미리 지정된 조건을 만족하는 경우에 동사를 변경한다.

[0042] 도 3은 상기 도 2에 있어서, 동사 변경부의 동사 교체 과정을 설명하기 위하여 보인 예시도이고, 도 4는 상기 도 3에 있어서, 동사 변경을 위한 조건을 만족하는지 체크하는 알고리즘을 보인 예시도이다.

[0043] 참고로 도 3에서 S_n(n=0, 1, ...)은 단어를 의미하며, W_e는 단어 임베딩을 의미하는 것으로서, 상기 단어들은 이미 학습되어 있는 것이며, 상기 단어들이 학습된 공간을 사용하는 것을 단어(word) 임베딩이라고 한다. 그리고 도 3에서 W_eS_n은 다음 단어를 LSTM을 이용하여 출력하기 위해 워드 임베딩에 입력 단어를 입력하는 과정을 의미하고, 소프트맥스(Softmax)함수를 통해 출력된 값이 다음 LSTM 단계에 입력되는 과정을 반복한다. 즉, S_0를 입력하여 S_1을 출력하였을 경우, 다음 LSTM 단계에 상기 출력되었던 S_1을 입력하는 과정을 반복 수행한다.

[0044] 그리고 상기 도 3에서 동사 변경을 위한 조건은, 지정된 단어(예 : S_2)의 손실(loss) 값이 지정된 값(예 : -3)보다 작고, 상기 지정된 단어(예 : S_2)의 품사가 동사(VB : Base Form)이며, 상기 행동 인식 모델부(130)에서 인식한 행동 인식 정확도가 지정된 기준(예 : 80%)보다 큰 경우인지 체크한다.

[0045] 그리고 상기 지정된 조건을 모두 만족하는 경우, 상기 동사 변경부(160)는 상기 행동 인식 모델부(130)에서 인식한 행동에 대응하는 단어로 최종 캡션(예 : 캡셔닝된 문장)에서 동사를 변경한다.

[0046] 참고로 상기 모폴로지 분석부(150)에서 상기 이미지 캡션(예 : 이미지의 특징이 캡셔닝된 문장)의 품사 분석은 자연어 처리 모듈(NLTK : Natural Language ToolKit)을 사용하고, 상기 NLTK의 pos_tag 메소드를 사용하여, 사전에 NLTK에 입력된 단어와 매핑되어 있는 태그로 캡션내의 각 단어들에 태그를 넣을 수 있다. 예컨대 상기 NLTK 내에 있는 동사관련 태그는 총 6개로서, 도 5에 도시된 바와 같다. 상기 태그는 행동 인식 모델의 클래스에 인칭 동사가 없으므로, VBP, VBZ 태그를 제외한 나머지 동사 태그(시제에 대응하는 태그)들을 사용할 수 있다.

[0047] 이상으로 본 실시예는 이미지 캡셔닝 시 동사에 해당하는 단어를, 선행 학습된 행동 인식 모델을 사용하여 판단된 행동에 대응하는 단어로 변경함으로써, 이미지 캡셔닝의 정확도가 높아지게 하는 방법에 대해서 설명하였다.

[0048] 그러나 상기와 같은 방법은 단지 동사에만 한정하지 않고, 문장을 구성하는 다른 품사에 대해서도 적용할 수 있다.

[0049] 도 6은 본 발명의 다른 실시예에 따라 복수의 품사에 해당하는 단어를 변경하거나 추가할 수 있는 캡셔닝 장치의 개략적인 구성을 보인 예시도로서, 얼굴 인식 모델(210), 행동 인식 모델(220), 및 위치 데이터 모델(230)을 더 포함하고, 기 캡셔닝된 문장에서, 학습을 통해 기 지정된 위치에 얼굴 인식된 사용자(user)의 이름을 추가하고, 학습을 통해 기 지정된 동사를 행동(Action)에 더 적합하게 매칭되는 단어로 변경하며, 학습을 통해 기 지정된 위치에 위치(또는 공간) 데이터(Location)의 명칭을 더 추가하는 방식으로 캡셔닝의 정확도를 향상시키는 것이다.

[0050] 도 6에 도시된 바와 같이, 각 LSTM(Long Short Term Memory)에서 출력된 단어들의 품사를 분석하여 각각에 맞는 신경망 모델(예 : 얼굴 인식 모델, 행동 인식 모델, 위치 데이터 모델)(210, 220, 230)들의 출력값으로 교체하여 보다 정확한 캡셔닝을 수행할 수 있다. 가령, LSTM에서 출력된 단어가 명사라면, 얼굴 인식 모델을 통해 출력된 값인 유저의 이름을 명사와 교체할 수 있고, 카메라 디바이스의 위치 데이터를 이용하여 임의의 현장에서 근무하는 사용자의 행동을 캡셔닝 할 수 있다.

[0051] 상기와 같이 본 실시예는 기존의 이미지 캡셔닝의 문제점인 행동 묘사를 보다 정확하게 묘사할 수 있는 캡셔닝을 수행할 수 있도록 함으로써, 가령 현장 근무자의 행동에 대한 정보를 보다 정확하게 수집할 수 있고, 행동 인식 모델을 위험 행동 등에 초점에 맞추어 학습 시킨다면 근무자의 위험 행동 묘사가 가능하여 위험 상황 관리를 용이하게 하는 효과가 있다.

[0052] 참고로 본 실시예에서 행동 인식을 위한 데이터 셋(예 : Flickr 8k 데이터 셋)의 이미지의 설명 데이터에서 자연어 처리 모듈을 사용하여, 동사 부분만을 출력 후 데이터 셋을 만들었다. 상기와 같이 동사 부분만을 모아 만든 데이터 셋의 클래스의 개수는 총 1523가지이며, 각 동사 클래스별 이미지의 개수는 전반적으로 10장 내외다. 이때 인칭 동사(예 : is, are, etc.)는 제외하였으며, 실험을 통해 이미지 캡셔닝 데이터 셋으로 학습할 경우, 행동 인식 관련 부분은 상당히 많은 동사 클래스와 각 동사 클래스의 이미지 개수는 적은 상태로 학습을 진행하였으며, 본 실시예에서 기존의 이미지 캡셔닝 모델로 정확하게 판단하기 힘든 행동 인식 부분은 선행 학습된 행동 인식 모델을 사용하여 이미지 캡셔닝의 정확도가 높아지는 것을 확인할 수 있다.

[0053] 이상으로 본 발명은 도면에 도시된 실시예를 참고로 하여 설명되었으나, 이는 예시적인 것에 불과하며, 당해 기술이 속하는 분야에서 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 따라서 본 발명의 기술적 보호범위는 아래의 특허청구범위에 의해서 정하여져야 할 것이다. 또한 본 명세서에서 설명된 구현은, 예컨대, 방법 또는 프로세스, 장치, 소프트웨어 프로그램, 데이터 스트림 또는 신호로 구현될 수 있다. 단일 형태의 구현의 맥락에서만 논의(예컨대, 방법으로서만 논의)되었더라도, 논의된 특징의 구현은 또한 다른 형태(예컨대, 장치 또는 프로그램)로도 구현될 수 있다. 장치는 적절한 하드웨어, 소프트웨어 및 펌웨어 등으로 구현될 수 있다. 방법은, 예컨대, 컴퓨터, 마이크로프로세서, 집적 회로 또는 프로그래밍 가능한 로직 디바이스 등을 포함하는 프로세싱 디바이스를 일반적으로 지칭하는 프로세서 등과 같은 장치에서 구현될 수 있다. 프로세서는 또한 최종-사용자 사이에 정보의 통신을 용이하게 하는 컴퓨터, 셀 폰, 휴대용/개인용 정보 단말기(personal digital assistant: "PDA") 및 다른 디바이스 등과 같은 통신 디바이스를 포함한다.

부호의 설명

[0055] 110 : 단어 임베딩 계층부

120 : 이미지 특징 추출부

130 : 행동 인식 모델부

140 : 캡션 구동부

150 : 모폴로지 분석부

160 : 동사 변경부

도면

도면1

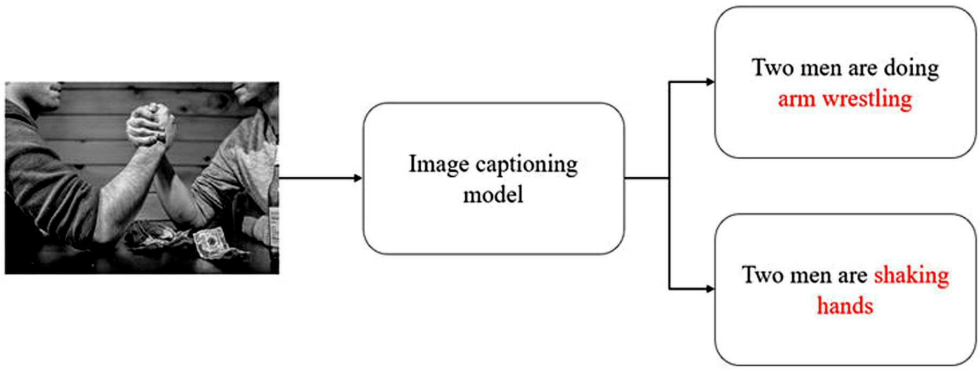
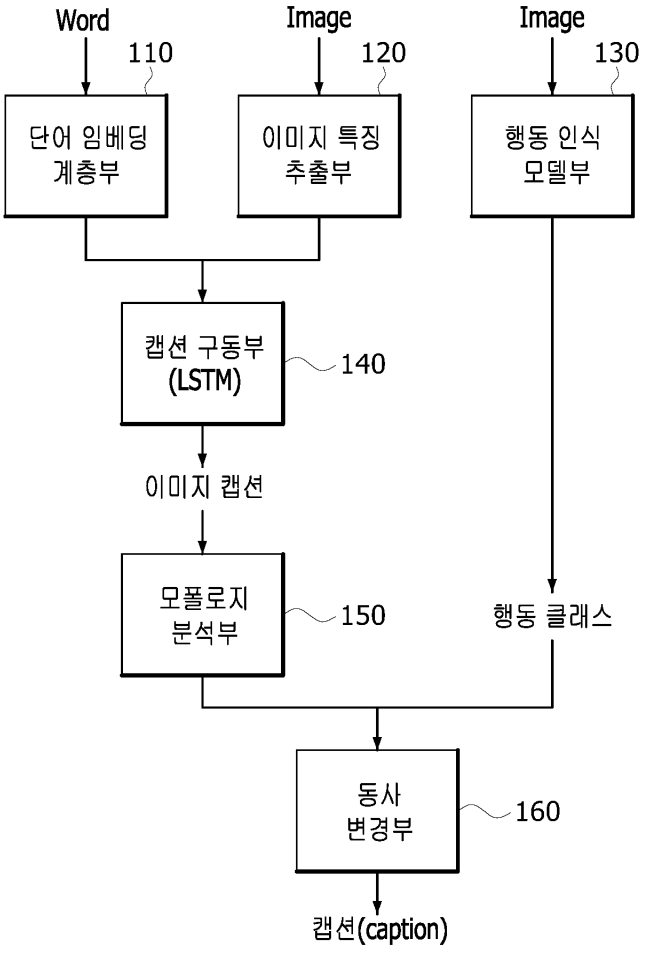
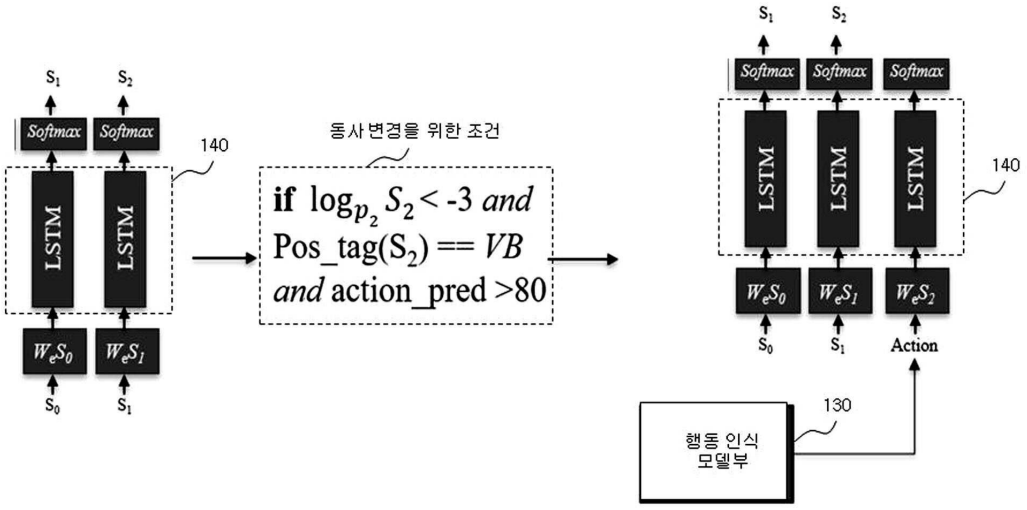


Image captioning model
Two men are doing arm wrestling
Two men are shaking hands

도면2


Word
110
Image
120
Image
130
단어 임베딩 계층부
이미지 특징 추출부
행동 인식 모델부
캡션 구동부 (LSTM)
140
이미지 캡션
모폴로지 분석부
150
행동 클래스
동사 변경부
160
캡션(caption)

도면3



도면4

M_D: image captioning model prediction
$M_C(pos\_tag)$: part of speech of predicted caption
M_L: loss value of image captioning model
V: max length of caption
Vcondition: verb condition (e.g. VB, VBD, etc.)
A_C: predicted action class
A_D: accuracy of action classification model
C: caption

```
for i < V do
M_C, M_L ← M_D (image, cap)
    if M_L <(score/i) then
      if M_C(pos_tag) == V_condition then
      if A_D > 0.8 then
            C ← Ac
  end if
      end if
    else
  C ← MC
end if
  i ← i + 1
end for
```

도면5

Tag	Description	Example
VB	base form	'take'
VBD	past tense	'took'
VBG	gerund/present participle	'taking'
VBN	past participle	'taken'
VBP	non−3rd person singular present	'take'
VBZ	3rd person singular present	'takes'

도면6