(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2020년11월11일
(11) 등록번호 10-2177412
(24) 등록일자 2020년11월05일

(51) 국제특허분류(Int. Cl.)
G06K 9/62 (2006.01) *G06K 9/48* (2006.01)
G06N 3/08 (2006.01)
(52) CPC특허분류
G06K 9/6201 (2013.01)
G06K 9/481 (2013.01)
(21) 출원번호 10-2018-0159582
(22) 출원일자 2018년12월12일
심사청구일자 2018년12월12일
(65) 공개번호 10-2020-0075114
(43) 공개일자 2020년06월26일
(56) 선행기술조사문헌
JP2000148381 A*
KR1020100023787 A*
KR1020150078148 A*
KR1020160029330 A*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
주식회사 인공지능연구원
경기도 성남시 분당구 성남대로331번길 8, 13층 1301호(정자동, 킨스타워)
(72) 발명자
김성표
서울특별시 강남구 언주로30길 21, 에이동 4601호(도곡동, 아카데미스위트)
황형재
경기도 성남시 분당구 분당로343번길 11, 303호(분당동)
장태진
경기도 성남시 분당구 발이봉북로35번길 7-2, 202호(수내동)
(74) 대리인
특허법인 신지

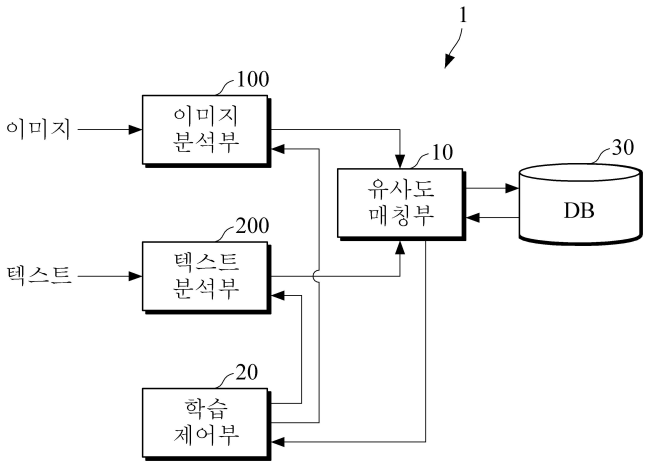
전체 청구항 수 : 총 6 항 심사관 : 강현일

(54) 발명의 명칭 **이미지와 텍스트간 유사도 매칭 시스템 및 방법**

(57) 요 약

본 발명은 이미지와 텍스트간 유사도 매칭 시스템으로, 입력 이미지의 상황 정보와 적어도 하나의 객체 정보가 반영된 이미지 특징 벡터를 생성하는 이미지 분석부와, 입력 텍스트의 상황 정보와 적어도 하나의 단어별 분석 정보가 반영된 텍스트 특징 벡터를 생성하는 텍스트 분석부와, 이미지 특징 벡터 및 텍스트 특징 벡터 간의 유사도를 산출하는 유사도 매칭부를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G06N 3/08 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	2017-0-0178
부처명	미래창조과학부
과제관리(전문)기관명	정보통신기술진흥센터
연구사업명	인공지능국가전략프로젝트 연구개발사업
연구과제명	(3세부) 비디오 이해를 위한 데이터 수집 및 보정자동화 시스템 개발
기 여 율	1/1
과제수행기관명	(주)코난테크놀로지
연구기간	2018.04.01 ~ 2018.12.31

명 세 서

청구범위

청구항 1

입력 이미지의 상황 정보와 적어도 하나의 객체 정보가 반영된 이미지 특징 벡터를 생성하는 이미지 분석부와,

입력 텍스트의 상황 정보와 적어도 하나의 단어별 분석 정보가 반영된 텍스트 특징 벡터를 생성하는 텍스트 분석부와,

이미지 특징 벡터 및 텍스트 특징 벡터 간의 유사도를 산출하는 유사도 매칭부와;

미리 산출된 유사도로 라벨링된 이미지-텍스트 쌍들인 훈련 데이터 셋을 각각 이미지 분석부 및 텍스트 분석부에 입력시킨 후, 유사도 매칭부에 의해 출력된 유사도와 유사도 레이블 간의 손실(Loss)를 줄여주는 방향으로 이미지 분석부 및 텍스트 분석부 각각을 구성하는 적어도 하나의 인공 신경망(Neural Network)들의 가중치(weight)를 조정하면서 학습시키는 학습 제어부를;

포함하는 이미지와 텍스트간 유사도 매칭 시스템.

청구항 2

제1 항에 있어서, 이미지 분석부는

입력 이미지로부터 적어도 하나의 객체를 추출하는 객체 인식 모듈과,

추출된 객체의 속성을 추출하는 객체 속성 인식 모듈 및 추출된 객체가 둘 이상일 경우, 객체 간 관계를 분석하는 객체 관계 인식 모듈 중 적어도 하나를 포함하는 이미지와 텍스트간 유사도 매칭 시스템.

청구항 3

제2 항에 있어서, 객체 인식 모듈은

적어도 하나의 객체의 입력 이미지 상의 위치 정보를 추출하되,

객체 관계 인식 모듈은

객체들 각각에 상응하는 위치 정보를 입력받아 위치 특징 벡터를 출력하는 완전 연결 레이어들과,

완전 연결 레이어들로부터 출력된 위치 특징 벡터들을 합산하는 합산부와,

합산된 특징 벡터를 입력받아 객체들 간의 관계 벡터를 출력하는 완전 연결 레이어를 포함하는 이미지와 텍스트간 유사도 매칭 시스템.

청구항 4

제2 항에 있어서, 이미지 분석부는

객체 속성 및 객체 간 관계 중 적어도 하나를 포함하는 특징 벡터를 분석하여 주목 위치가 반영된 벡터를 생성하는 주목 위치 분석 모듈을 더 포함하는 이미지와 텍스트간 유사도 매칭 시스템.

청구항 5

제1 항에 있어서, 텍스트 인식부는

상황 정보 및 적어도 하나의 단어별 분석 정보는 각각 회귀적 신경망에 의해 분석되는 이미지와 텍스트간 유사도 매칭 시스템.

청구항 6

제1 항에 있어서, 텍스트 인식부는

추출된 단어들 중 주목 위치가 표현된 벡터를 생성하는 주목 위치 분석 모듈을 더 포함하는 이미지와 텍스트간 유사도 매칭 시스템.

청구항 7

삭제

청구항 8

삭제

청구항 9

삭제

청구항 10

삭제

청구항 11

삭제

청구항 12

삭제

발명의 설명

기 술 분 야

[0001] 본 발명은 인공 지능 기술에 관한 것으로, 특히 이미지 및 텍스트를 매칭시키는 기술에 관한 것이다.

[0002] 본 연구는 미래창조과학부 및 정보통신기술진흥센터의 인공지능국가전략프로젝트 연구개발사업(과제고유번호: 2017-0-01781, 연구 과제명:(3 세부) 비디오 이해를 위한 데이터 수집 및 보정자동화 시스템 개발)의 일환으로 수행하였다.

배 경 기 술

[0004] 이미지-텍스트 매칭 기술은 이미지와 텍스트가 주어지면 서로간의 유사도을 계산하고, 계산된 유사도 기반으로 이미지와 텍스트가 연관되었는지의 여부를 판단하는 기술이다. 이러한 이미지-텍스트 매칭 기술은 다중 모드 검색(Multi-Modal Retrieval) 서비스에 주로 활용되는데, 동일한 도메인 내에서 검색하는 것과 달리 질문(query)으로부터 상이한 타입에서 타겟을 검색하는데 초점을 둔다. 즉, 이미지 또는 텍스트 검색 쿼리(Search Query)가 주어질 경우, 데이터베이스에서 가장 연관성이 있는 상응하는 텍스트 또는 이미지를 검색하는 것이 목적이다.

[0005] 이와 같은 이미지-텍스트 매칭을 위해 종래에는 이미지와 텍스트 전체적인 특징을 추출하고, 추출된 특징을 비교하여 유사도를 산출하였다. 예컨대, 이미지에 포함된 객체나 텍스트에 포함된 단어를 분석/비교하여 매칭하였다. 그런데, 동일한 종류의 객체가 포함되어 있더라도, 이미지는 객체의 컬러, 크기 및 갯수와 같은 속성에 따라 다시 구별될 수 있다. 또한, 이미지 상의 객체들 간의 관계에 의해서도 다시 구별될 수 있다. 따라서, 종래의 이미지-텍스트 매칭 기술만으로는 이러한 디테일한 요소들을 고려한 이미지-텍스트 매칭을 기대하기 어렵다.

발명의 내용

해결하려는 과제

[0007] 본 발명은 이미지와 텍스트 간의 디테일한 구별 요소들에 기반하여 정교하게 매칭할 수 있는 이미지와 텍스트간 유사도 매칭 시스템 및 방법을 제공한다.

과제의 해결 수단

[0009] 본 발명은 이미지와 텍스트간 유사도 매칭 시스템으로, 입력 이미지의 상황 정보와 적어도 하나의 객체 정보가 반영된 이미지 특징 벡터를 생성하는 이미지 분석부와, 입력 텍스트의 상황 정보와 적어도 하나의 단어별 분석 정보가 반영된 텍스트 특징 벡터를 생성하는 텍스트 분석부와, 이미지 특징 벡터 및 텍스트 특징 벡터 간의 유사도를 산출하는 유사도 매칭부를 포함한다.

[0010] 본 발명은 이미지와 텍스트간 유사도 매칭 방법으로, 입력 이미지의 상황 정보와 적어도 하나의 객체 정보가 반영된 이미지 특징 벡터를 생성하는 단계와, 입력 텍스트의 상황 정보와 적어도 하나의 단어별 분석 정보가 반영된 텍스트 특징 벡터를 생성하는 단계와, 이미지 특징 벡터 및 텍스트 특징 벡터 간의 유사도를 산출하는 단계를 포함한다.

발명의 효과

[0012] 본 발명은 이미지와 텍스트 간의 디테일한 구별 요소들에 기반하여 정교하게 매칭할 수 있다.

[0013]

도면의 간단한 설명

[0014] 도 1은 본 발명의 일 실시 예에 따른 이미지와 텍스트간 유사도 매칭 시스템의 블록 구성도이다.

도 2는 이미지와 텍스트 간의 유사도 매칭 예를 설명하기 위한 도면이다.

도 3은 본 발명의 일 실시 예에 따른 이미지 분석부의 개략적인 블록 구성이다.

도 4는 본 발명에 따른 분석 대상 이미지의 예시도이다.

도 5는 본 발명의 일 실시 예에 따른 객체 관계 인식 모듈의 블록 구성도이다.

도 6은 본 발명의 일 실시 예에 따른 텍스트 분석부의 개략적인 블록 구성이다.

도 7은 본 발명에 따른 단어 특징 추출부의 예시도이다.

도 8은 본 발명의 일 실시 예에 따른 이미지 분석 단계를 설명하기 위한 순서도이다.

도 9는 본 발명의 일 실시 예에 따른 텍스트 분석 단계를 설명하기 위한 순서도이다.

도 10은 본 발명의 일 실시 예에 따른 이미지와 텍스트간 유사도 매칭 방법을 활용하여 이미지-텍스트 간의 다중 모드 검색(Multi-Modal Retrieval) 서비스를제공하는 과정을 설명하기 위한 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0015] 이하 첨부된 도면을 참조하여, 바람직한 실시 예에 따른 이미지와 텍스트간 유사도 매칭 시스템 및 방법에 대해 상세히 설명하면 다음과 같다. 여기서, 동일한 구성에 대해서는 동일부호를 사용하며, 반복되는 설명, 발명의 요지를 불필요하게 흐릴 수 있는 공지 기능 및 구성에 대한 상세한 설명은 생략한다. 발명의 실시형태는 당업계에서 평균적인 지식을 가진 자에게 본 발명을 보다 완전하게 설명하기 위해서 제공되는 것이다. 따라서, 도면에서의 요소들의 형상 및 크기 등은 보다 명확한 설명을 위해 과장될 수 있다.

[0016] 첨부된 블록도의 각 블록과 흐름도의 각 단계의 조합들은 컴퓨터 프로그램인스트럭션들(실행 엔진)에 의해 수행될 수도 있으며, 이들 컴퓨터 프로그램 인스트럭션들은 범용 컴퓨터, 특수용 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장비의 프로세서에 탑재될 수 있으므로, 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장비의 프로세서를 통해 수행되는 그 인스트럭션들이 블록도의 각 블록 또는 흐름도의 각 단계에서 설명된 기능들을 수행하는 수단을 생성하게 된다.

[0017] 이들 컴퓨터 프로그램 인스트럭션들은 특정 방식으로 기능을 구현하기 위해 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장비를 지향할 수 있는 컴퓨터 이용가능 또는 컴퓨터 판독 가능 메모리에 저장되는 것도 가능하므로, 그 컴퓨터 이용가능 또는 컴퓨터 판독 가능 메모리에 저장된 인스트럭션들은 블록도의 각 블록 또는 흐름도의 각 단계에서 설명된 기능을 수행하는 인스트럭션 수단을 내포하는 제조 품목을 생산하는 것도 가능하다.

[0018] 그리고 컴퓨터 프로그램 인스트럭션들은 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장비 상에 탑재되는 것도 가능하므로, 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장비 상에서 일련의 동작 단계들이 수행되어 컴퓨터로 실행되는 프로세스를 생성해서 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장비를 수행하는 인스트럭션들은 블록도의 각 블록 및 흐름도의 각 단계에서 설명되는 기능들을 실행하기 위한 단계들을 제공하는 것도 가능하다.

[0019] 또한, 각 블록 또는 각 단계는 특정된 논리적 기능들을 실행하기 위한 하나 이상의 실행 가능한 인스트럭션들을 포함하는 모듈, 세그먼트 또는 코드의 일부를 나타낼 수 있으며, 몇 가지 대체 실시 예들에서는 블록들 또는 단계들에서 언급된 기능들이 순서를 벗어나서 발생하는 것도 가능함을 주목해야 한다. 예컨대, 잇달아 도시되어 있는 두 개의 블록들 또는 단계들은 사실 실질적으로 동시에 수행되는 것도 가능하며, 또한 그 블록들 또는 단계들이 필요에 따라 해당하는 기능의 역순으로 수행되는 것도 가능하다.

[0020] 이하, 첨부 도면을 참조하여 본 발명의 실시 예를 상세하게 설명한다. 그러나 다음에 예시하는 본 발명의 실시 예는 여러 가지 다른 형태로 변형될 수 있으며, 본 발명의 범위가 다음에 상술하는 실시 예에 한정되는 것은 아니다. 본 발명의 실시 예는 당업계에서 통상의 지식을 가진 자에게 본 발명을 보다 완전하게 설명하기 위하여 제공된다.

[0022] 도 1은 본 발명의 일 실시 예에 따른 이미지와 텍스트간 유사도 매칭 시스템의 블록 구성도이고, 도 2는 이미지와 텍스트 간의 유사도 매칭 예를 설명하기 위한 도면이다.

[0023] 도 1을 참조하면, 이미지와 텍스트간 유사도 매칭 시스템(이하 '시스템'으로 기재함)(1)은 이미지와 텍스트 각각에 대한 전반적 내용과 세부적인 내용이 모두 반영된 특징 벡터를 생성하고, 생성된 이미지 특징 벡터와 텍스트 특징 벡터를 이용하여 유사도를 매칭한다.

[0024] 이를 위해, 시스템(1)은 크게 이미지 분석부(100), 텍스트 분석부(200) 및 유사도 매칭부(10)를 포함한다.

[0025] 이미지 분석부(100)는 입력 이미지의 상황 정보와 적어도 하나의 객체 정보가 반영된 이미지 특징 벡터를 생성한다. 여기서, 입력 이미지는 단일 이미지를 지칭하는 것일 수도 있으며, 또는 시계열적으로 연속된 일련의 이미지, 즉, 동영상을 지칭하는 것일 수도 있다. 이미지 분석부(100)에 대한 상세한 설명은 이하 도 3 내지 도 5를 참조하여 후술하기로 한다.

[0026] 텍스트 분석부(200)는 입력 텍스트의 상황 정보와 적어도 하나의 단어별 분석 정보가 반영된 텍스트 특징 벡터를 생성한다. 텍스트 분석부(100)에 대한 상세한 설명은 이하 도 6 및 도 7을 참조하여 후술하기로 한다.

[0027] 유사도 매칭부(10)는 이미지 분석부(100)에 의해 생성된 이미지 특징 벡터 및 텍스트 분석부(200)에 텍스트 특징 벡터 간의 유사도를 산출한다. 일 실시 예에 따라, 유사도 매칭부(10)는 이미지-텍스트 상의 유사도 $\boldsymbol{S(x,y)}$ 를 판단하는데 일반적으로 코사인 유사도가 많이 사용된다.

[0028] 학습 제어부(20)는 미리 라벨링된 이미지-텍스트 쌍들인 훈련 데이터 셋을 각각 이미지 분석부(100) 및 텍스트 분석부(200)에 입력시킨 후, 유사도 매칭부(10)에 의해 출력된 유사도와 미리 정의된 데이터 내의 서로 연관된 이미지-텍스트 쌍의 비교를 통해 손실(Loss) $\boldsymbol{l}$ 을 줄여주는 방향으로 이미지 분석부(100) 및 텍스트 분석부(200) 각각을 구성하는 적어도 하나의 인공 신경망(Neural Network)들의 가중치(weight)를 조정하면서 학습시킨다. 학습 제어부(20)는 다음의 <수학식 1>과 같이 손실 $\boldsymbol{l}$ 을 계산한다.

수학식 1

[0029] $$\boldsymbol{l(I,T) = [\alpha - S(I,T) + S(I,\hat{T})]_+ + [\alpha - S(I,T) + S(\hat{I},T)]_+}$$

[0030] <수학식 1>에서 I 는 이미지이고, T 는 텍스트이고, I 와 T 는 미리 정의된 서로 연관된 이미지-텍스트 쌍이다. $\hat{I}$ 는 T 와 연관이 없는 오답 이미지이고, $\hat{T}$ 는 I 와 연관이 없는 오답 텍스트이다. 그리고, $S(x,y)$ 는 x 및 y간의 유사도 산출 함수로 전술한 바와 같이 코사인 유사도가 많이 사용된다. 또한, α 는 마진(margin)을 나타내는 하이퍼 파라미터(hyper parameter)로, 서로 연관된 쌍의 유사도와 연관이 없는 쌍의 유사도간의 차이(gap)를 보장하기 위한 것이다. 예컨대, α 가 없으면, $S(I,T)$ 및 $S(I,\hat{T})$ 이 모두 '0.5' 이어도 손실 l 은 '0'이 되므로, 연관된 쌍과 연관이 없는 쌍의 유사도가 동일해지도록 학습이 된다. 반면, α 가 0.5라면 $S(I,T)$ 가 0.7일 때, $S(I,\hat{T})$ 는 0.2가 되어야 손실 l 이 0이 되므로 연관된 쌍과 연관이 없는 쌍의 유사도의 차이를 만들도록 학습이 된다. 한편, <수학식 1>에서 $[F]_+$ 는 다음의 <수학식 2>와 같이 정의된다.

수학식 2

[0031]
$$[F]_+ = \begin{cases} F & if\ F > 0 \\ 0 & if\ F \leq 0 \end{cases}$$

[0032] 전술한 바와 같이 학습 제어부(20)는 손실 l 을 줄이는 방향으로 학습이 되며, 서로 연관된 이미지 I 와 텍스트 T 간의 유사도를 높이고, 서로 연관이 없는 이미지 $\hat{I}$ 와 텍스트 T 그리고 서로 연관이 없는 텍스트 $\hat{T}$ 와 이미지 I 간의 유사도를 낮추도록 학습이 된다.

[0033] 학습 제어부(20)는 미리 산출된 유사도로 라벨링된 이미지-텍스트 쌍들인 훈련 데이터 셋을 각각 이미지 분석부(100) 및 텍스트 분석부(200)에 입력시킨 후, 유사도 매칭부(10)에 의해 출력된 유사도와 유사도 레이블 간의 손실(Loss)를 줄여주는 방향으로 이미지 분석부(100) 및 텍스트 분석부(200) 각각을 구성하는 적어도 하나의 인공 신경망(Neural Network)들의 가중치(weight)를 조정하면서 학습시킨다.

[0034] 한편, 전술한 바와 같이 학습되어 설계된 시스템(1)은 이미지-텍스트 간의 다중 모드 검색(Multi-Modal Retrieval) 서비스를 제공할 수 있다. 이를 위해, 시스템(1)은 데이터베이스(DB)(30)를 더 포함할 수 있다. DB(30)에는 이미지 분석부(100) 또는 텍스트 분석부(200) 각각에 의해 미리 생성된 특징 벡터들이 매핑된 복수의 이미지들또는 텍스트들을 저장할 수 있다.

[0035] 그러면, 시스템(1)에 검색 쿼리(Search Query)로 이미지 (또는 텍스트)가 입력됨에 따라, 이미지 분석부(100)(또는 텍스트 분석부(200))에 의해 이미지 특징 벡터(또는 텍스트 특징 벡터)가 생성되고, 유사도 매칭부(10)는 생성된 이미지 특징 벡터(또는 텍스트 특징 벡터)와 DB(30)에 저장된 텍스트 특징 벡터들(또는 이미지 특징 벡터들)를 비교하여, 저장된 텍스트 특징 벡터들(또는 이미지 특징 벡터들)에 매핑된 텍스트들(이미지들)을 유사도 내림차순으로 정렬한 검색 결과를 출력한다. 도 2를 참조하면, (a) 사각형은 이미지 특징 벡터를 나타내고, 삼각형은 텍스트 특징 벡터를 나타내는데, 시스템(1)에 의해 (b)에 도시된 바와 같이 서로 연관된 동일하게 음영 표시된 이미지-텍스트 쌍(정답)이 매칭될 수 있다.

[0036] 도 3은 본 발명의 일 실시 예에 따른 이미지 분석부의 개략적인 블록 구성이고, 도 4는 본 발명에 따른 분석 대상 이미지의 예시도이고, 도 5는 본 발명의 일 실시 예에 따른 객체 관계 인식 모듈의 블록 구성도이다.

[0037] 도 3을 참조하면, 이미지 분석부(100)는 상황 특징 추출부(110), 객체 특징 추출부(120) 및 결합부(130)를 포함한다.

[0038] 상황 특징 추출부(110)는 입력된 이미지로부터 유추될 수 있는 전반적인 상황 정보를 표현하는 벡터를 생성한다. 여기서, 상황 정보는 시간, 장소 및 사건 정보 등을 포함하는 배경을 통해 유추할 수 있는 정보를 포

함한다. 예컨대, 상황 특징 추출부(110)는 도 3에 도시된 입력 이미지로부터 "축구" 와 같은 사건 정보 또는 "축구장"과 같은 장소 정보를 표현하는 벡터를 생성하여 출력한다. 이러한 상황 정보를 표현하는 벡터를 둘 이상이 생성되어 출력될 수도 있다.

[0039] 객체 특징 추출부(120)는 입력된 이미지로부터 추출되는 적어도 하나의 객체에 대한 정보를 추출한다. 본 발명의 일 실시 예에 따라, 객체 정보는 단순히 입력 이미지 상에 존재하는 객체의 종류를 구별하는 명칭 뿐만 아니라, 더 나아가 동일한 종류의 객체를 더 상세하게 구별해낼 수 있는 구별 요소로써 상세 정보를 추출해낸다. 이를 위해, 일 실시 예에 따라, 객체 특징 추출부(120)은 객체 인식 모듈(121), 객체 관계 인식 모듈(122) 및 객체 속성 인식 모듈(123)을 포함한다. 추가적으로, 주목 위치 분석 모듈(124)을 더 포함한다.

[0040] 객체 인식 모듈(121)은 이미지에 포함된 적어도 하나의 객체를 추출하여, 추출된 객체들 각각을 표현하는 벡터를 생성한다. 도 4를 참조하면, 이미지에서 경계 박스(bounding box)들(41, 42, 43, 44)에 포함된 4개의 객체들 각각을 표현하는 4개의 벡터들이 생성된다. 즉, 2개의 벡터들은 사람(혹은 축구선수)(41, 43)을 표현하고, 1개의 벡터는 (축구)공(42)을 표현하고, 나머지 1개의 벡터는 잔디장(44)을 표현한다. 여기서, 객체 인식 모듈(121)은 Faster R-CNN, SSD 등 다양한 인공 신경망 알고리즘에 의해 설계될 수 있다.

[0041] 객체 관계 인식 모듈(122)은 추출된 객체가 둘 이상일 경우, 객체 간 관계를 분석한다. 도 5를 참조하면, 객체 관계 인식 모듈(122)은 객체 1의 위치 정보 벡터를 입력받아 출력하는 제1 완전 연결(Fully-Connected, FC) 레이어(122a)와, 객체 2의 위치 정보 벡터를 입력받아 출력하는 제2 완전 연결(Fully-Connected, FC) 레이어(122b)와, 제1 완전 연결 레이어(122a) 및 제2 완전 연결 레이어(122b)를 결합하는 결합부(122c)와, 결합된 하나의 위치 정보 벡터를 입력받아 객체 1 및 객체 2 간의 관계를 표현하는 벡터를 생성하여 출력하는 제3 완전 연결(Fully-Connected, FC) 레이어(122d)를 포함한다. 여기서, 위치 정보는 추출된 객체들 각각의 입력 이미지 상의 위치 정보로 중심 좌표(x, y) 및 크기(width, height) 정보일 수 있고, 이러한 위치 정보는 객체 인식 모듈(121)에 의해 추출될 수 있다.

[0042] 예컨대, 객체 관계 인식 모듈(122)은 도 4에 도시된 이미지에서 "우측 사람"(41)을 표현하는 벡터 및 "축구공"(42)을 표현하는 벡터를 입력받아, "사람이 공을 찬다"라는 관계를 표현하는 하나의 벡터를 생성하여 출력한다. 이때, 객체 관계 인식 모듈(122)은 인식된 객체들 중 두 객체들에 상응하는 벡터들을 조합 선택하여, 선택된 객체들 간의 관계 벡터를 산출한다. 따라서, 도 4에 도시된 바와 같이 4개의 객체들이 인식된 경우, 두 개의 객체들의 선택 조합쌍들이 6개가 생성되므로, 객체 관계 인식 모듈(122)은 6개의 객체들 간의 관계 벡터들을 생성하여 출력할 수 있다.

[0043] 객체 속성 인식 모듈(123)은 객체 인식 벡터에 객체의 속성 정보를 반영하는 벡터를 생성하여 출력한다. 즉, 추출된 객체를 표현하는 벡터를 속성 정보까지 반영된 벡터를 생성하여 출력한다. 이와 같이 객체의 속성까지 반영됨에 따라, 도 4에 도시된 사람(41, 42)에 대한 벡터들 2개 중 하나는 빨간색(표시안됨) 옷을 입은 사람(41), 다른 하나는 흰색 옷을 입은 사람(43)을 표현하는 벡터가 생성된다. 따라서, 객체들 간의 구별 요소를 하나 더 추가하게 되므로, 정밀한 이미지 매칭을 가능하게 한다.

[0044] 한편, 2018년에 발표된 논문, "Stacked Cross Attention for Image-Text Matching", Kuang-Huei Lee *et al.*에 개시된 바와 같이, 최근에는 이미지에서 객체를 인식한 뒤 객체의 주목 위치를 분석하여 이를 각 단어와 비교하는 추세이다. 따라서, 일 실시 예에 따라, 객체 특징 추출부(120)는 이미지에서 추출된 주목 위치를 분석하는 주목 위치 분석 모듈(124)을 더 포함하되, 전술한 논문과 같이 단순히 추출된 객체 특징 벡터로부터 주목 위치를 분석하는 것이 아니라, 객체 속성 및 객체 간 관계 중 적어도 하나를 포함하는 특징 벡터를 분석하여 주목 위치가 반영된 벡터를 생성한다. 즉, 최종적으로 관계 벡터들과 속성 벡터들을 이어 붙인 뒤, 어느 부분에 초점을 맞춰야 하는지(더 중요하게 봐야하는지)를 분석한다. 예컨대, 도 4에 도시된 바와 같이 빨간색 유니폼을 입은 선수('빨간색 옷을 입은 사람' 속성 벡터)(41)가 축구공을 차는 모습('사람이 축구공을 차는' 관계 벡터)(43)이 주요하므로 그 부분에 주목한다. 이러한 주목 위치 분석 모듈(124)은 완전 연결망(Fully-Connected Network, FCN)으로 설계될 수 있으나, 본 발명은 이에 한정되지 않는다.

[0045] 결합부(130)는 상황 특징 추출부(110)로부터 출력된 상황 특징 벡터 및 객체 특징 추출부(120)로부터 추출된 객체 특징 벡터를 결합하여 하나의 이미지 특징 벡터를 생성한다.

[0046] 도 6은 본 발명의 일 실시 예에 따른 텍스트 분석부의 개략적인 블록 구성이고, 도 7은 본 발명에 따른 단어 특징 추출부의 예시도이다.

[0047] 도 6을 참조하면, 텍스트 분석부(200)는 임베딩부(210), 문장 특징 추출부(220), 단어 특징 추출부(230) 및 결

합부(240)를 포함한다.

[0048] 임베딩부(210)는 입력되는 텍스트를 단어별로 임베딩한다. 문장 특징 추출부(220)는 임베딩부(210)로부터 입력된 텍스트가 의미하는 전반적인 상황을 유추한 특징 벡터를 생성하여 출력한다. 예컨대, "가방으로 비를 피하며 거리를 걷고 있는 사람"이라는 텍스트가 입력되면, 비가 내리는 상황 등을 표현하는 벡터를 생성한다.

[0049] 단어 특징 추출부(230)는 회귀 분석 모듈(231) 및 주목 위치 분석 모듈(232)을 포함하여, 회귀적 신경망에 의해 분석되는 각 단어에 대하여 어느 단어에 집중해서 봐야하는지를 판단한다. 예컨대, "가방으로 비를 피하며 거리를 걷고 있는 사람"이라는 텍스트에서 비를 피하는 모습이지만 "가방"이라는 단어에 좀 더 집중된다.

[0050] 본 발명의 일 실시 예에 따라, 이러한 문장 특징 추출부(220) 및 회귀 분석 모듈(231)에서 사용되는 학습 알고리즘은 시퀀스 데이터 처리에 적합한 재귀적 신경망(Recurrent Neural Network : RNN) 모델 또는 LSTM(Long-Short Term Memory)을 사용하여 훈련될 수 있다. RNN(Recurrent Neural Network) 또는 LSTM(Long-Short Term Memory)은 시간의 흐름에 따라 변하는 시계열 데이터를 학습하고 인공지능을 예측하는 학습 알고리즘이다. RNN은 매순간의 데이터를 인공신경망 구조에 쌓아올린 것으로 딥 러닝 중 가장 깊은 네트워크 구조이다. 시계열 데이터의 예로는 본 발명에서와 같은 송수신 신호를 포함하여 주가, 사람의 움직임, 기후, 인터넷 접속자수, 검색어 등을 생각해 볼 수 있다. LSTM은 Long-Short term Memory란 게이트 유닛을 노드마다 배치하여 인공신경망이 너무 깊어서 오랜 시간 전의 데이터들을 까먹는 현상(vanishing gradient problem)을 해결한 알고리즘이다. 이러한 RNN 또는 LSTM을 사용함으로써, 학습 모델은 시간적인 샘플들의 연관성을 학습하게 된다.

[0051] 도 7을 참조하면, 연속적인 시간 샘플들이 입력됨에 따른 처리 과정의 이해를 돕기 위해, LSTM 유닛들(231a, 231b)은 시간 샘플링된 입력값들을 입력받는 형태로 펼쳐져 도시되어 있음을 유의하여야 한다. 즉, 각각 1, 2, 3, ..., 9에 상응하는 단어들을 LSTM 유닛들이 별도로 도시되어 있으나, 이는 하나의 LSTM 유닛들(231a, 231b)이 시간 샘플들의 순차적인 입력에 따른 순차적인 처리 과정을 도시하 것일 뿐이다.

[0052] 또한, 회귀 분석 모듈(231)은 LSTM 유닛들(231a, 231b)은 양방향(bidirection) 구조를 갖는다. 즉, 제1 LSTM 유닛(231b)은 순방향으로 입력되는 단어들을 학습하고, 제2 LSTM 유닛(231a)은 역방향으로 입력되는 단어들을 처리하여 출력한다. 그러면, 합산부(231c)는 제1 LSTM 유닛(231a) 및 제2 LSTM 유닛(231b)에 의해 출력된 출력값들을 합산한 후, 평균값을 산출하여 출력하게 된다.

[0053] 다음으로, 본 발명의 일 실시 예에 따른 이미지와 텍스트간 유사도 매칭 방법을 설명하기로 한다.

[0054] 본 발명에 따른 이미지와 텍스트간 유사도 매칭 방법은 이미지와 텍스트 각각에 대한 전반적 내용과 세부적인 내용이 모두 반영된 특징 벡터를 생성하고, 생성된 이미지 특징 벡터와 텍스트 특징 벡터를 이용하여 유사도를 매칭한다. 이를 위해, 입력 이미지의 상황 정보와 적어도 하나의 객체 정보가 반영된 이미지 특징 벡터를 생성하는 단계(도 8에 도시됨)와, 입력 텍스트의 상황 정보와 적어도 하나의 단어별 분석 정보가 반영된 텍스트 특징 벡터를 생성하는 단계(도 9에 도시됨)와, 생성된 이미지 특징 벡터 및 텍스트 특징 벡터 간의 유사도를 산출하는 단계를 포함한다.

[0055] 도 8은 본 발명의 일 실시 예에 따른 이미지 분석 단계를 설명하기 위한 순서도이다.

[0056] 도 8을 참조하면, 이미지 분석부(100)는 이미지가 입력(S310)됨에 따라, 입력된 이미지로부터 추출되는 적어도 하나의 객체에 대한 정보를 추출한다(S320~S340). 본 발명의 일 실시 예에 따라, 객체 정보는 단순히 입력 이미지 상에 존재하는 객체의 종류를 구별하는 명칭 뿐만 아니라, 더 나아가 동일한 종류의 객체를 더 상세하게 구별해낼 수 있는 구별 요소로써 상세 정보를 추출해낸다.

[0057] 이미지 분석부(100)는 이미지에 포함된 적어도 하나의 객체를 추출하여, 추출된 객체들 각각을 표현하는 벡터를 생성한다(S320).

[0058] 이미지 분석부(100)는 추출된 객체가 둘 이상일 경우, 객체 간 관계를 분석한다(S330). 즉, 이미지 분석부(100)는 객체 1의 위치 정보 벡터 및 객체 2의 위치 정보 벡터를 입력받아 객체 1 및 객체 2 간의 관계를 표현하는 벡터를 생성하여 출력한다. 여기서, 위치 정보는 추출된 객체들 각각의 입력 이미지 상의 위치 정보로 중심 좌표(x, y) 및 크기(width, height) 정보일 수 있고, 이러한 위치 정보는 객체 인식에 의해 추출될 수 있다. 이때, 인식된 객체들 중 두 객체들에 상응하는 벡터들을 조합 선택되어, 선택된 객체들 간의 관계 벡터들이 산출될 수 있다.

[0059] 이미지 분석부(100)는 객체 인식 벡터에 객체의 속성 정보를 반영하는 벡터를 생성하여 출력한다(S330). 즉, 추출된 객체를 표현하는 벡터를 속성 정보까지 반영된 벡터를 생성하여 출력한다. 이와 같이 객체의 속성까지 반

영됨에 따라, 객체들 간의 구별 요소를 하나 더 추가하게 되므로, 정밀한 이미지 매칭을 가능하게 한다.

[0060] 또한, 이미지 분석부(100)는 이미지에서 추출된 주목 위치를 분석한다(S340). 이때, 단순히 추출된 객체 특징 벡터로부터 주목 위치를 분석하는 것이 아니라, 객체 속성 및 객체 간 관계 중 적어도 하나를 포함하는 특징 벡터를 분석하여 주목 위치가 반영된 벡터를 생성한다. 즉, 최종적으로 관계 벡터들과 속성 벡터들을 이어 붙인 뒤, 어느 부분에 초점을 맞춰야 하는지(더 중요하게 봐야하는지)를 분석한다.

[0061] 이미지 분석부(100)는 입력된 이미지로부터 유추될 수 있는 전반적인 상황 정보를 표현하는 벡터를 생성한다(S350). 여기서, 상황 정보는 시간, 장소 및 사건 정보 등을 포함하는 배경을 통해 유추할 수 있는 정보를 포함한다.

[0062] 마지막으로, 이미지 분석부(100)는 상황 특징 벡터 및 객체 특징 벡터를 결합하여 하나의 이미지 특징 벡터를 생성한다(S360).

[0063] 도 9는 본 발명의 일 실시 예에 따른 텍스트 분석 단계를 설명하기 위한 순서도이다.

[0064] 도 9를 참조하면, 텍스트 분석부(200)는 텍스트가 입력됨(S410)에 따라, 단어별로 임베딩한다(S420).

[0065] 텍스트 분석부(200)는 회귀적 신경망에 의해 분석되는 각 단어에 대하여 어느 단어에 집중해서 봐야하는지를 판단한다(S430~S440). 예컨대, "가방으로 비를 피하며 거리를 걷고 있는 사람"이라는 텍스트에서 비를 피하는 모습이지만 "가방"이라는 단어에 좀 더 집중된다.

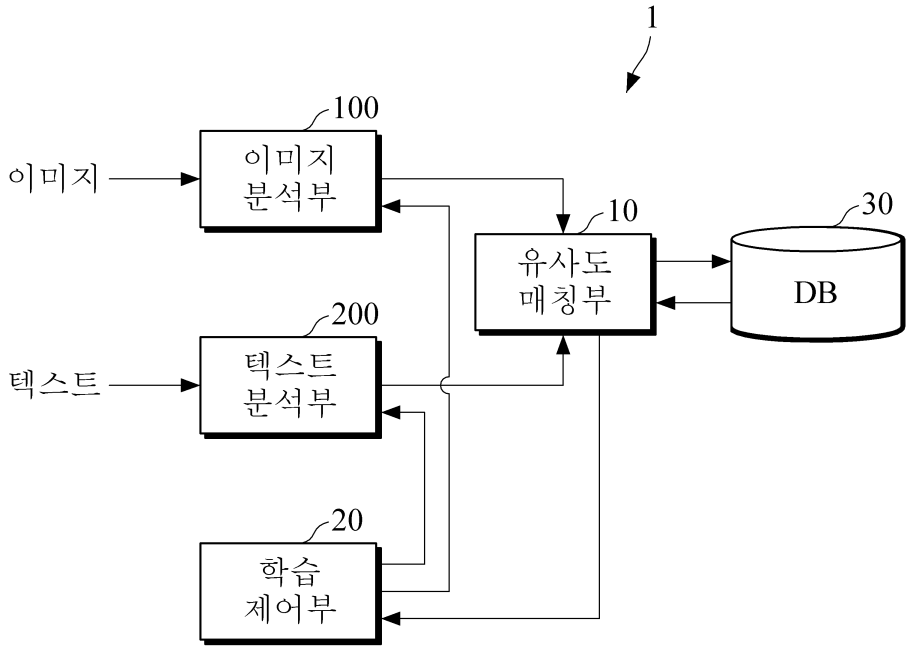
[0066] 텍스트 분석부(200)는 입력된 텍스트가 의미하는 전반적인 상황을 유추한 특징 벡터를 생성하여 출력한다(S440).

[0067] 도 10은 본 발명의 일 실시 예에 따른 이미지와 텍스트간 유사도 매칭 방법을 활용하여 이미지-텍스트 간의 다중 모드 검색(Multi-Modal Retrieval) 서비스를제공하는 과정을 설명하기 위한 순서도이다.

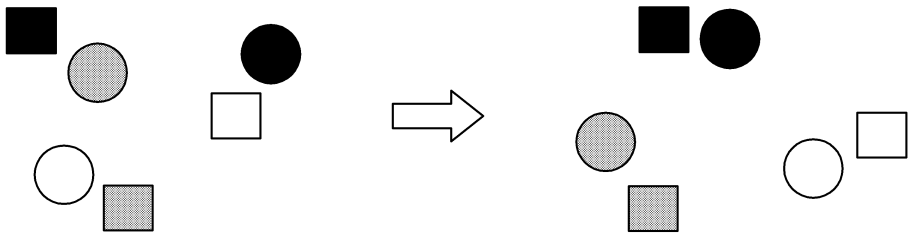
[0068] 도 10을 참조하면, 검색 쿼리(Search Query)로 이미지 (또는 텍스트)가 입력(S510)됨에 따라, 시스템(1)은 이미지 특징 벡터(또는 텍스트 특징 벡터)를 생성한다(S520). 시스템(1)은 생성된 이미지 특징 벡터(또는 텍스트 특징 벡터)와 DB(30)에 저장된 텍스트 특징 벡터들(또는 이미지 특징 벡터들)을 이용하여 유사도를 산출한다(S530). 그런 후, 시스템(1)은 저장된 텍스트 특징 벡터들(또는 이미지 특징 벡터들)에 매핑된 텍스트들(이미지들)을 유사도 내림차순으로 정렬한 검색 결과를 출력한다(S540).

도면

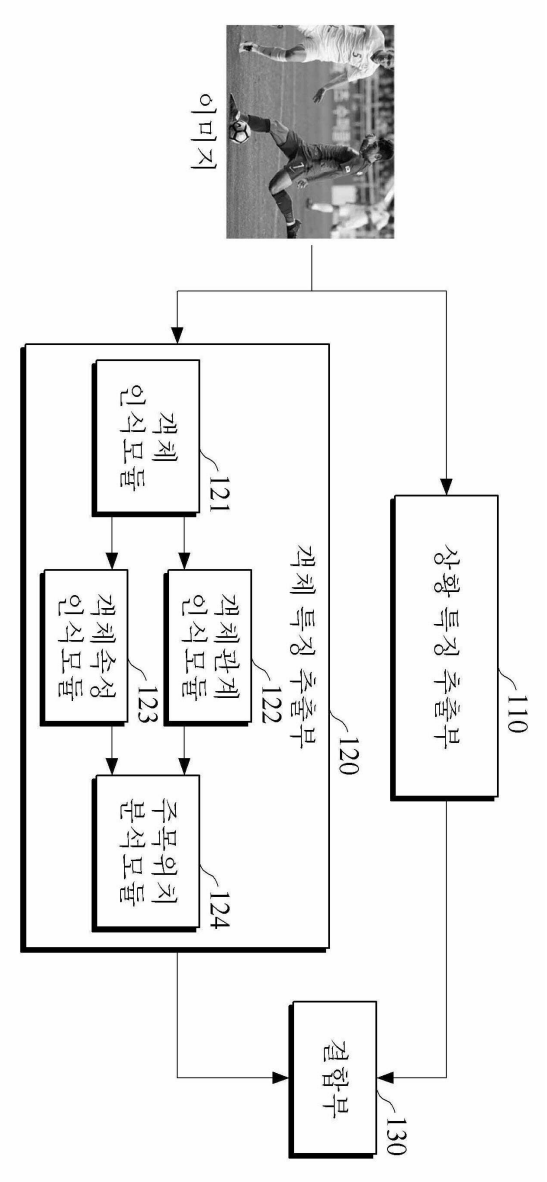
도면1



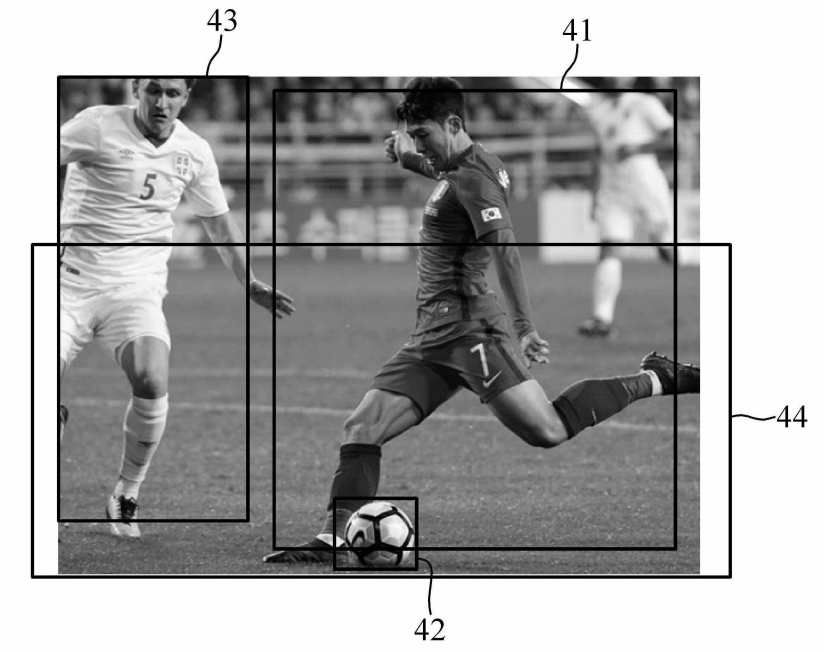
도면2



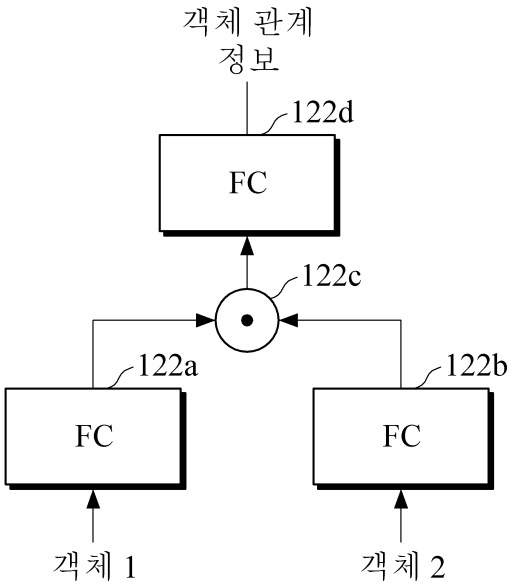
도면3


이미지
객체 인식모듈
121
객체속성 인식모듈
123
객체관계 인식모듈
122
주목위치 분석모듈
124
객체 특징 추출부
120
상황 특징 추출부
110
결합부
130

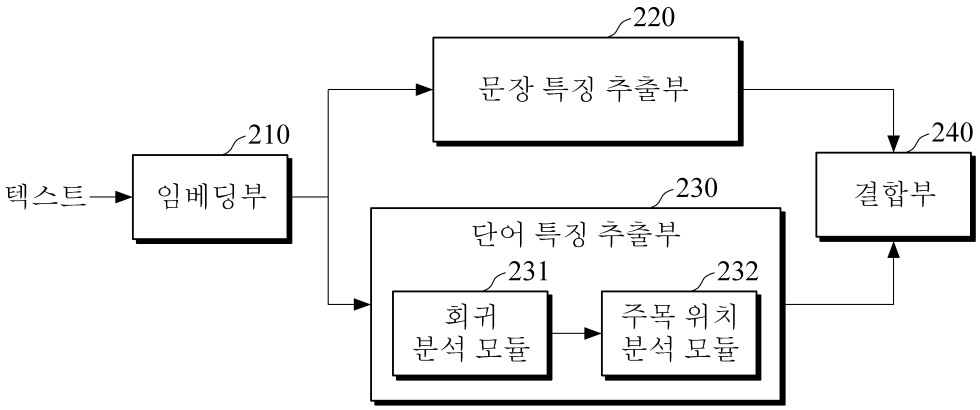
도면4



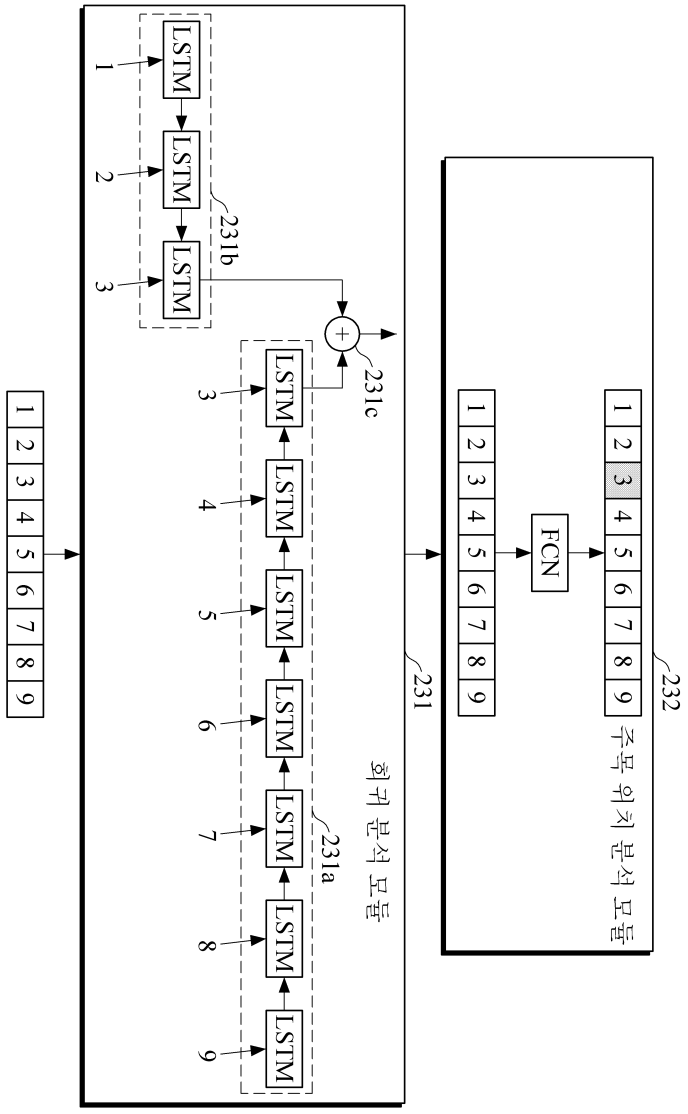
도면5



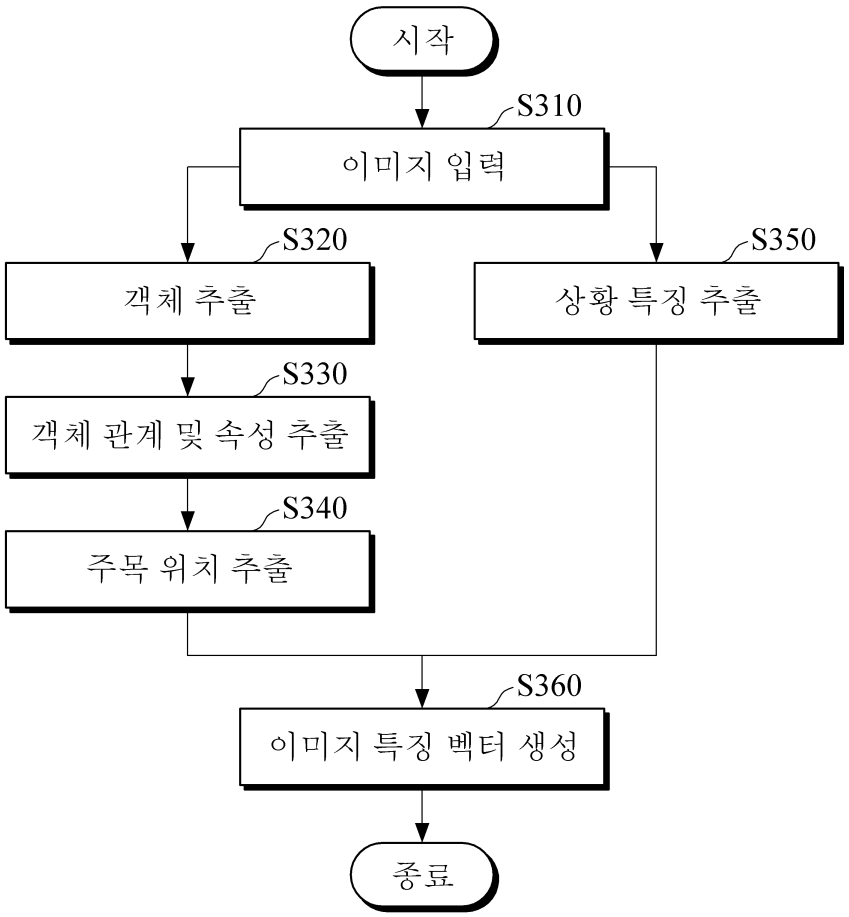
도면6


220
문장 특징 추출부
210
텍스트
임베딩부
240
결합부
230
단어 특징 추출부
231
회귀
분석 모듈
232
주목 위치
분석 모듈

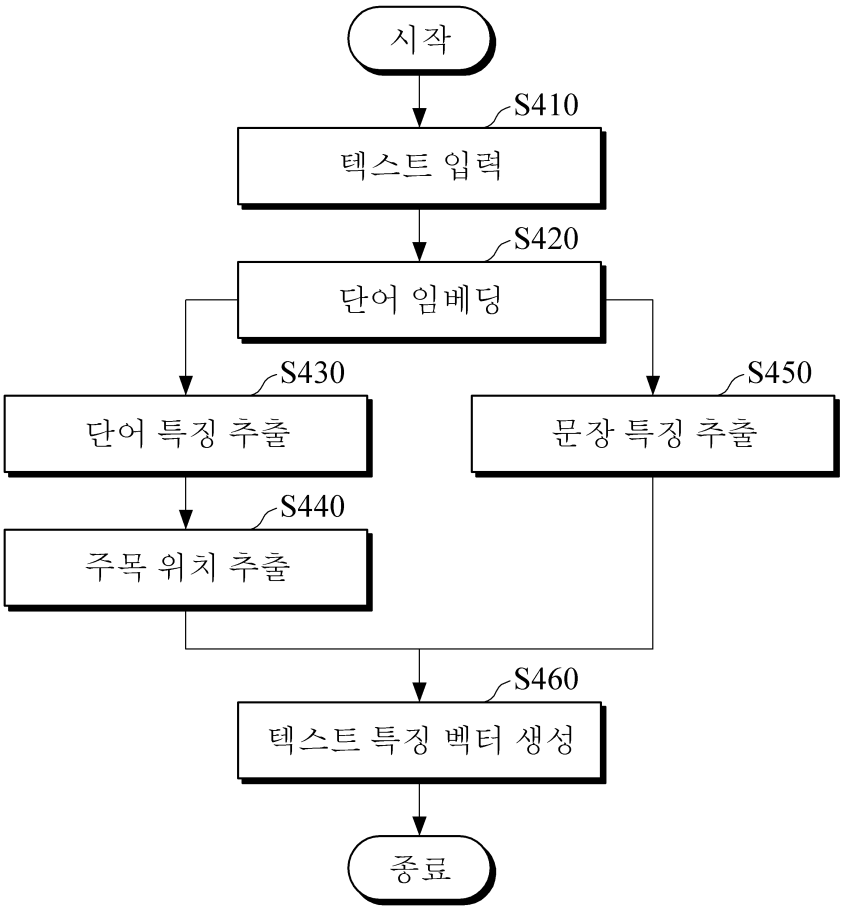
도면7


LSTM
231b
231c
231a
231
회귀 분석 모듈
FCN
232
주목 위치 분석 모듈

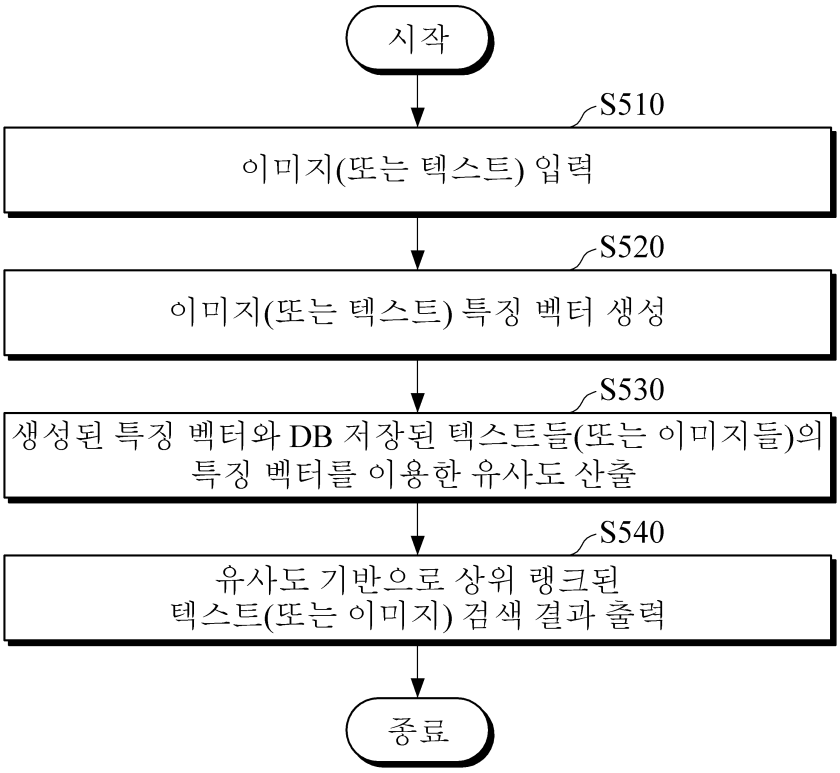
도면8


시작
S310
이미지 입력
S320
객체 추출
S350
상황 특징 추출
S330
객체 관계 및 속성 추출
S340
주목 위치 추출
S360
이미지 특징 벡터 생성
종료

도면9


시작
S410
텍스트 입력
S420
단어 임베딩
S430
단어 특징 추출
S450
문장 특징 추출
S440
주목 위치 추출
S460
텍스트 특징 벡터 생성
종료

도면10


시작
S510
이미지(또는 텍스트) 입력
S520
이미지(또는 텍스트) 특징 벡터 생성
S530
생성된 특징 벡터와 DB 저장된 텍스트들(또는 이미지들)의 특징 벡터를 이용한 유사도 산출
S540
유사도 기반으로 상위 랭크된 텍스트(또는 이미지) 검색 결과 출력
종료