(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0093221
(43) 공개일자 2024년06월24일

(51) 국제특허분류(Int. Cl.)
G06N 3/096 (2023.01) *G06F 16/432* (2019.01)
G06F 16/532 (2019.01) *G06F 16/9032* (2019.01)
G06F 40/284 (2020.01) *G06N 3/0455* (2023.01)
G06N 3/0895 (2023.01)
(52) CPC특허분류
G06N 3/096 (2023.01)
G06F 16/434 (2019.01)
(21) 출원번호 10-2022-0176260
(22) 출원일자 2022년12월15일
심사청구일자 2022년12월15일

(71) 출원인
한양대학교 산학협력단
서울특별시 성동구 왕십리로 222(행당동, 한양대학교내)
(72) 발명자
한경식
경기도 성남시 수정구 위례동로 61, 5607동 1401호
진승완
서울특별시 성동구 왕십리로 222 FTC 507호
(뒷면에 계속)
(74) 대리인
특허법인(유한)아이시스

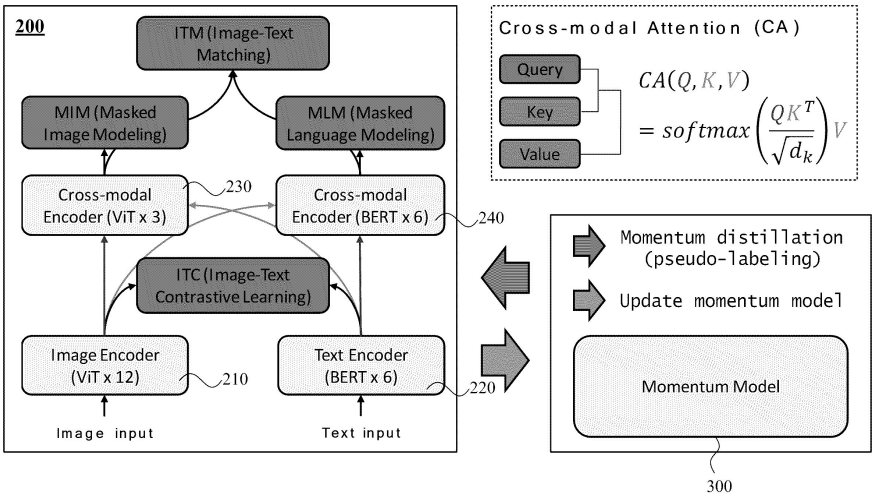
전체 청구항 수 : 총 8 항

(54) 발명의 명칭 **이미지-텍스트 검색 모델 구축 방법 및 서비스 장치**

(57) 요 약

이미지-텍스트 검색 모델 구축 방법.은 학습장치는 학습데이터 중 이미지 및 텍스트 쌍을 선택하고, 상기 이미지를 이미지 인코더에 입력하고, 상기 텍스트를 텍스트 인코더에 입력하는 단계, 상기 학습장치는 상기 이미지 인코더와 상기 텍스트 인코더에 대한 ITC(image-text contrastive learning)를 수행하는 단계, 상기 학습장치는 상기 이미지 인코더가 출력하는 이미지 임베딩을 입력받는 제1 크로스 모달 인코더에 대한 MIM(Masked Image Modeling)을 수행하는 단계, 상기 학습장치는 상기 텍스트 인코더가 출력하는 텍스트 임베딩을 입력받는 제2 크로스 모달 인코더에 대한 MLM(Masked Language Modeling)을 수행하는 단계 및 상기 학습장치는 상기 제1 크로스 모달 인코더가 출력하는 이미지 임베딩과 상기 제2 크로스 모달 인코더가 출력하는 텍스트 임베딩에 대한 ITM(Image-Text Matching) 학습을 수행하는 단계를 포함한다.

대 표 도



(52) CPC특허분류
G06F 16/532 (2019.01)
G06F 16/90332 (2019.01)
G06F 40/284 (2020.01)
G06N 3/0455 (2023.01)
G06N 3/0895 (2023.01)

(72) 발명자
최호영
경기도 수원시 영통구 중부대로 604, 계룡리슈빌 612호

노태형
서울특별시 성동구 살곶이길 348, 306호

이 발명을 지원한 국가연구개발사업
과제고유번호 1711152849
과제번호 2020-0-01373-003
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성
연구과제명 인공지능대학원지원(한양대학교)
기 여 율 40/100
과제수행기관명 한양대학교산학협력단
연구기간 2022.01.01 ~ 2022.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711160226
과제번호 2018-0-01431-005
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성
연구과제명 MR-IoT융합 기반의 재난대응 인공지능 응용기술
기 여 율 30/100
과제수행기관명 아주대학교산학협력단
연구기간 2018.06.01 ~ 2023.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711160466
과제번호 2022-0-00240-001
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 실감콘텐츠핵심기술개발
연구과제명 비언어적 요소 기반 XR콘텐츠 상호 적용 기술개발
기 여 율 30/100
과제수행기관명 연세대학교 산학협력단
연구기간 2022.04.01 ~ 2022.12.31

명 세 서

청구범위

청구항 1

학습장치는 학습데이터 중 이미지 및 텍스트 쌍을 선택하고, 상기 이미지를 이미지 인코더에 입력하고, 상기 텍스트를 텍스트 인코더에 입력하는 단계;

상기 학습장치는 상기 이미지 인코더와 상기 텍스트 인코더에 대한 ITC(image-text contrastive learning)를 수행하는 단계;

상기 학습장치는 상기 이미지 인코더가 출력하는 이미지 임베딩을 입력받는 제1 크로스 모달 인코더에 대한 MIM(Masked Image Modeling)을 수행하는 단계;

상기 학습장치는 상기 텍스트 인코더가 출력하는 텍스트 임베딩을 입력받는 제2 크로스 모달 인코더에 대한 MLM(Masked Language Modeling)을 수행하는 단계; 및

상기 학습장치는 상기 제1 크로스 모달 인코더가 출력하는 이미지 임베딩과 상기 제2 크로스 모달 인코더가 출력하는 텍스트 임베딩에 대한 ITM (Image-Text Matching) 학습을 수행하는 단계를 포함하는 이미지-텍스트 검색 모델 구축 방법.

청구항 2

제1항에 있어서,

상기 학습장치는

상기 제1 크로스 모달 인코더에서 상기 이미지의 패치 토큰을 쿼리로 사용하고, 상기 텍스트의 단어 토큰을 키 및 값으로 사용하는 셀프 어텐션을 상기 MIM 전에 수행하는 단계; 및

상기 제2 크로스 모달 인코더에서 상기 이미지의 패치 토큰을 쿼리로 사용하고, 상기 텍스트의 단어 토큰을 키 및 값으로 사용하는 셀프 어텐션을 상기 MIM 전에 수행하는 단계를 더 포함하는 이미지-텍스트 검색 모델 구축 방법.

청구항 3

제1항에 있어서,

상기 이미지 인코더는 ViT(Vision Transformer)의 인코더들을 포함하고,

상기 텍스트 인코더는 BERT (Bidirectional Encoder Representations from Transformers)의 인코더들로 포함하는 이미지-텍스트 검색 모델 구축 방법.

청구항 4

제1항에 있어서,

상기 학습장치는 상기 ITM 학습에서 상기 제1 크로스 모달 인코더가 출력하는 이미지 임베딩과 상기 제2 크로스 모달 인코더가 출력하는 텍스트 임베딩의 유사도를 예측하도록 학습하는 이미지-텍스트 검색 모델 구축 방법.

청구항 5

사용자가 입력하는 쿼리값을 입력받는 인터페이스 장치;

이미지 및 텍스트를 입력받아 유사도를 결정하는 검색 모델, 이미지 세트 미 텍스트 세트를 저장하는 저장장치; 및

상기 쿼리값이 이미지인 경우 상기 이미지를 패치 토큰으로 변환하여 상기 검색 모델에 입력하고 상기 텍스트 모델에 선택한 어느 하나의 텍스트를 단어 토큰으로 변환하여 상기 검색 모델에 입력하고, 상기 검색 모델의 출

력하는 값을 기준으로 상기 이미지와 상기 어느 하나의 텍스트가 유사한지 결정하는 연산장치를 포함하되,

상기 검색 모델은

이미지의 패치 토큰을 입력받아 이미지 임베딩하는 이미지 인코더;

텍스트의 단어 토큰을 입력받아 텍스트 임베딩하는 텍스트 인코더;

상기 이미지 인코더의 출력을 입력받아 대응하는 단어를 예측하는 제1 크로스 모달 인코더;

상기 텍스트 인코더의 출력을 입력받아 대응하는 이미지 패치를 예측하는 제2 크로스 모달 인코더; 및

상기 제1 크로스 모달 인코더가 출력하는 이미지 임베딩과 상기 제2 크로스 모달 인코더가 출력하는 텍스트 임베딩을 입력받아 서로의 유사도를 출력하는 매칭 계층을 포함하는 이미지-텍스트 검색을 수행하는 서비스 장치.

청구항 6

제5항에 있어서,

상기 이미지 인코더와 상기 텍스트 인코더는 학습 과정에서 ITC(image-text contrastive learning)를 통해 학습되는 이미지-텍스트 검색을 수행하는 서비스 장치.

청구항 7

제5항에 있어서,

상기 제1 크로스 모달 인코더는 학습과정에서 입력되는 이미지의 패치 토큰을 쿼리로 사용하고, 입력되는 텍스트의 단어 토큰을 키 및 값으로 사용하는 셀프 어텐션을 수행하고, MIM(Masked Image Modeling)을 수행하여 학습되는 이미지-텍스트 검색을 수행하는 서비스 장치.

청구항 8

제5항에 있어서,

상기 제2 크로스 모달 인코더는 학습과정에서 입력되는 텍스트의 단어 토큰을 쿼리로 사용하고, 입력되는 이미지의 패치 토큰을 키 및 값으로 사용하는 셀프 어텐션을 수행하고, MLM(Masked Language Modeling)을 수행하여 학습되는 이미지-텍스트 검색을 수행하는 서비스 장치.

발명의 설명

기 술 분 야

[0001] 이하 설명하는 기술은 이미지와 텍스트 사이의 크로스 모달 검색 기법에 관한 것이다.

배 경 기 술

[0002] 최근 인터넷 및 스마트 기기를 이용한 다양한 서비스가 급격하게 성장하고 있다. 다수의 서비스들은 멀티미디어 데이터를 이용한다. 멀티미디어 데이터는 대부분 이미지, 텍스트 등을 동시에 포함하는 멀티모달(multimodal) 데이터에 해당한다. 예컨대, 이커머스(e-commerce) 등이 대표적으로 멀티모달 데이터가 활용되는 도메인이다. 소비자는 쇼핑몰에서 이미지와 텍스트 검색을 통해 원하는 상품을 찾는다. 이에 멀티미디어 데이터를 이해하는 추천 알고리즘 기술이 개발되고 있다. 즉, 이미지와 텍스트 사이의 크로스 모달 검색(cross-modal retrieval) 기술 성능이 중요해지고 있다.

선행기술문헌

특허문헌

[0003] (특허문헌 0001) 한국공개특허 제10-2022-0107916호

발명의 내용

해결하려는 과제

[0004] 이미지와 텍스트 사이의 크로스 모달 검색 모델은 학습과정에서 서로 대응하는 이미지와 텍스트의 관계를 충분히 고려되어야만 한다. 그러나, 종래 기술은 특히 이미지 인코더의 학습 과정에서 텍스트를 반영하지 못했던 한계가 있었다.

[0005] 이하 설명하는 기술은 이미지 인코더 학습에 텍스트 특징을 반영하고, 동시에 텍스트 인코더 학습에 이미지 특징을 반영하는 크로스 모달 검색 모델을 제공하고자 한다.

과제의 해결 수단

[0006] 이미지-텍스트 검색 모델 구축 방법은 학습장치는 학습데이터 중 이미지 및 텍스트 쌍을 선택하고, 상기 이미지를 이미지 인코더에 입력하고, 상기 텍스트를 텍스트 인코더에 입력하는 단계, 상기 학습장치는 상기 이미지 인코더와 상기 텍스트 인코더에 대한 ITC(image-text contrastive learning)를 수행하는 단계, 상기 학습장치는 상기 이미지 인코더가 출력하는 이미지 임베딩을 입력받는 제1 크로스 모달 인코더에 대한 MIM(Masked Image Modeling)을 수행하는 단계, 상기 학습장치는 상기 텍스트 인코더가 출력하는 텍스트 임베딩을 입력받는 제2 크로스 모달 인코더에 대한 MLM(Masked Language Modeling)을 수행하는 단계 및 상기 학습장치는 상기 제1 크로스 모달 인코더가 출력하는 이미지 임베딩과 상기 제2 크로스 모달 인코더가 출력하는 텍스트 임베딩에 대한 ITM(Image-Text Matching) 학습을 수행하는 단계를 포함한다.

[0007] 이미지-텍스트 검색을 수행하는 서비스 장치는 사용자가 입력하는 쿼리값을 입력받는 인터페이스 장치, 이미지 및 텍스트를 입력받아 유사도를 결정하는 검색 모델, 이미지 세트 미 텍스트 세트를 저장하는 저장장치 및 상기 쿼리값이 이미지인 경우 상기 이미지를 패치 토큰으로 변환하여 상기 검색 모델에 입력하고 상기 텍스트 모델에 선택한 어느 하나의 텍스트를 단어 토큰으로 변환하여 상기 검색 모델에 입력하고, 상기 검색 모델의 출력하는 값을 기준으로 상기 이미지와 상기 어느 하나의 텍스트가 유사한지 결정하는 연산장치를 포함한다.

발명의 효과

[0008] 이하 설명하는 기술은 입력되는 텍스트 또는 이미지를 기준으로 대응되는 이미지 또는 텍스트를 정확하게 검색하여 서비스 플랫폼에서 소비자가 원하는 아이템을 추천할 수 있다.

도면의 간단한 설명

[0009] 도 1은 이미지와 텍스트의 크로스 모달 검색 서비스 시스템에 대한 예이다.

도 2는 검색 모델의 학습 과정에 대한 예이다.

도 3은 이미지와 텍스트의 크로스 모달 검색을 수행하는 서비스 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 주기

능 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.

[0016] 이하 설명하는 기술은 사용자가 입력하는 텍스트 또는 이미지를 기준으로 대응하는 이미지 또는 텍스트를 검색 결과로 추천하는 서비스에 적용할 수 있다. 이하 설명하는 기술은 대표적으로 이커머스와 같은 분야에 적용할 수 있다. 다만, 이하 설명하는 기술은 텍스트와 이미지 사이의 크로스 모달 검색이 활용되는 다양한 분야에 활용될 수 있다.

[0017] 이하 입력되는 이미지 또는 텍스트를 기준으로 대응하는 아이템을 검색 내지 추천하는 장치를 서비스 장치라고 명명한다. 서비스 장치는 텍스트 및 이미지 데이터 처리 및 학습 모델 추론 등이 가능한 장치이다. 컴퓨터 장치는 PC, 스마트기기, 네트워크의 서버, 데이터 처리 전용 칩셋 등의 형태를 가질 수 있다.

[0019] 도 1은 이미지와 텍스트의 크로스 모달 검색 서비스 시스템(100)에 대한 예이다. 도 1에서 서비스 서버(130)가 서비스 장치에 해당한다.

[0020] 사용자 단말(110)은 멀티모달 데이터를 이용한 서비스를 받는 사용자 장치이다. 예컨대, 사용자는 쇼핑몰에서 자신이 원하는 제품을 검색할 수 있다.

[0021] 사용자 단말(110)은 사용자가 입력하는 텍스트를 서비스 서버(130)에 전달한다. 서비스 서버(130)는 사전에 학습된 검색 모델을 저장한다. 이미지와 텍스트 사이의 크로스 모달 검색은 사전에 학습된 검색 모델에서 수행된다. 서비스 서버(130)는 사용자가 입력한 텍스트를 검색 모델에 입력하고, 검색 모델이 출력하는 값(이미지)을 기준으로 대응하는 아이템을 결정한다. 서비스 서버(130)가 검색 모델의 출력값으로 결정된 아이템을 사용자 단말(110)에 전달한다.

[0022] 또는 사용자 단말(110)은 사용자가 입력하는 이미지를 서비스 서버(130)에 전달한다. 서비스 서버(130)는 사전에 학습된 검색 모델을 저장한다. 이미지와 텍스트 사이의 크로스 모달 검색은 사전에 학습된 검색 모델에서 수행된다. 서비스 서버(130)는 사용자가 입력한 이미지를 검색 모델에 입력하고, 검색 모델이 출력하는 값(텍스트)을 기준으로 대응하는 아이템을 결정한다. 서비스 서버(130)가 검색 모델의 출력값으로 결정된 아이템을 사용자 단말(110)에 전달한다.

[0024] 이하 전술한 검색 모델의 구조 및 학습 과정에 대하여 상세하게 설명한다. 도 2는 검색 모델의 학습 과정에 대한 예이다. 검색 모델(200)은 이미지 인코더(210), 텍스트 인코더(220), 제1 크로스 모달 인코더(230) 및 제2 크로스 모달 인코더(240)를 포함한다.

[0025] 검색 모델의 학습 과정은 학습장치가 수행한다고 설명한다. 학습 장치는 모델 개발자가 검색 모델 구축에 이용하는 컴퓨터 장치에 해당한다. 학습데이터는 대응하는 이미지와 텍스트 쌍들로 구성될 수 있다. 대응하는 이미지와 텍스트는 서로 정답 데이터로 사용된다.

[0026] 학습과정은 다음과 같은 단계로 구성된다. (i) 학습장치는 단일 모달리티 인코더(210 및 220)에 대한 이미지와 텍스트에 대한 대조학습(contrastive learning)을 수행한다. 학습장치는 ITC(image-text contrastive learning)를 수행하여 각 인코더가 서로 다른 모달리티로부터 임베딩된 벡터들에 대하여 대응하는 쌍(positive pair)인 경우 가깝게 임베딩하고, 대응하지 않는 쌍(negative pair)인 경우 멀게 임베딩하도록 만든다. 따라서, 이미지 인코더(210) 및 텍스트 인코더(220)는 각각 크로스 모달 학습에 적합한 형태로 입력 데이터를 임베딩하게 된다. (ii) 학습장치는 각 모달리티에 해당하는 크로스 모달 인코더(230 및 240)를 학습시킨다. 학습장치는 각 크로스 모달 인코더에 대하여 대응하는 모달리티를 예측하도록 한다. 학습장치는 크로스 모달 인코더 230 및

240 각각에 대하여 MIM(Masked Image Modeling) 및 MLM(Masked Language Modeling)을 수행한다. (iii) 마지막으로 학습장치는 ITM(Image-Text Matching) 학습을 통해 크로스 모달 학습이 적용된 이미지와 텍스트 임베딩 벡터의 매칭 여부를 학습하게 한다. 나아가, 학습장치는 학습 과정에서 모멘텀 모델(momentum mobel, 300)을 이용하여 수도 라벨(pseudo label)을 생성하고, 크로스 모달 인코더의 학습 과정에서 과적합 문제를 완화할 수도 있다.

[0028] 이하 학습 과정을 구체적으로 설명한다.

[0029] 이미지 인코더(210) 및 텍스트 인코더(220)는 각각 서로 다른 임베딩 공간에서 입력된 이미지와 텍스트를 임베딩한다.

[0030] 이미지 인코더(210)는 ViT(Vision Transformer)를 사용할 수 있다. 연구자는 이미지 인코더(210)를 ViT의 인코더 12개로 구성하였다. 학습장치는 입력 이미지를 일련의 패치(patch) 토큰들로 변환하여 이미지 인코더(210)에 입력한다. 따라서 이미지 인코더(210)는 패치 토큰 단위로 임베딩 작업을 수행한다.

[0031] 텍스트 인코더(220)는 BERT (Bidirectional Encoder Representations from Transformers)를 사용할 수 있다. 연구자는 텍스트 인코더(220)를 6개의 BERT 인코더로 구성하였다. 학습장치는 입력 텍스트를 일련의 단어 토큰으로 변환하여 텍스트 인코더(220)에 입력한다.

[0032] 학습장치는 이미지 인코더(210)와 텍스트 인코더(220)에 대하여 단일 모딜리티에 기반한 ITC를 수행한다. 연구자는 65,536개의 이미지와 텍스트가 저장된 큐(queue)에서 추출하면서 이미지 인코더(210)와 텍스트 인코더(220)에 대한 ITC를 수행하였다. 학습장치는 정답

[0033] 학습장치는 입력되는 이미지-텍스트 쌍이 대응하는 쌍(positive pair)인 경우 가깝게 임베딩하고, 대응하지 않는 쌍(negative pair)인 경우 멀게 임베딩하도록 유도한다. 이 과정에서 학습장치는 하드 네거티브(hard negative)인 쌍을 선정하여 ITM 학습 과정에 활용한다.

[0035] 제1 크로스 모달 인코더(230)과 제2 크로스 모달 인코더(240)의 학습 과정에 대하여 설명한다.

[0036] 이미지 인코더(210) 및 텍스트 인코더(220)는 각각 이미지 임베딩과 텍스트 임베딩을 산출한다. 제1 크로스 모달 인코더(230)는 이미지 인코더(210)가 산출하는 이미지 임베딩을 입력받는다. 학습장치는 제1 크로스 모달 인코더(230)가 텍스트를 반영한 이미지 임베딩을 산출하도록 학습한다. 제2 크로스 모달 인코더(240)는 텍스트 인코더(210)가 산출하는 텍스트 임베딩을 입력받는다. 학습장치는 제2 크로스 모달 인코더(240)가 이미지를 반영한 텍스트 임베딩을 산출하도록 학습한다.

[0037] 제1 크로스 모달 인코더(230)과 제2 크로스 모달 인코더(240) 각각은 셀프 어텐션(self-attention) 과정에서 Q(Query), K(Key) 및 V(Value)를 서로 다른 모달리티로 구성한다. 셀프 어텐션은 크로스 모달 임베딩을 가능하게 하는 특징을 부여한다.

[0038] 제1 크로스 모달 인코더(230)는 아래 수학식 1과 같이 셀프 어텐션 과정(CA_I)에서 이미지 패치 토큰을 쿼리(Q)로 입력받고, 단어 토큰을 키(K) 및 값(V)으로 입력받아 연산을 수행한다. 제1 크로스 모달 인코더(230)는 이미지 패치 토큰(Q)과 단어 토큰(K)의 유사도를 값(V)에 반영한다.

수학식 1

[0040] $$\text{Cross} - \text{modal Attention}\ (CA_I) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

[0041] d_k는 Q와 K 벡터의 차원에 해당한다. 연산식 자체는 자연어 처리 모델에서의 셀프 어텐션과 유사하다. 다만, 제1 크로스 모달 인코더(230)에서 셀프 어텐션은 서로 다른 모달리티를 갖는 Q(이미지)와 K(단어)를 입력받아 V(유사도)를 결정한다.

[0043] 제2 크로스 모달 인코더(230)는 아래 수학식 2와 같이 셀프 어텐션 과정(CA_T)에서 단어 토큰을 쿼리(Q)로 입력받고, 이미지 패치 토큰을 키(K) 및 값(V)으로 입력받아 연산을 수행한다. 제1 크로스 모달 인코더(230)는 단어 토큰(Q)과 이미지 패치 토큰(K)의 유사도를 값(V)에 반영한다.

수학식 2

[0045] $$\text{Cross} - \text{modal Attention } (CA_T) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

[0046] d_k는 Q와 K 벡터의 차원에 해당한다. 연산식 자체는 자연어 처리 모델에서의 셀프 어텐션과 유사하다. 다만, 제2 크로스 모달 인코더(240)에서 셀프 어텐션은 서로 다른 모달리티를 갖는 Q(단어)와 K(이미지)를 입력받아 V(유사도)를 결정한다.

[0048] 학습장치는 제1 크로스 모달 인코더(230)에 대하여 MIM (Masked Image Modeling)을 수행한다.

[0049] 학습장치는 입력 이미지 패치 토큰들 중 일부를 마스크(mask) 토큰으로 변경하고, 제1 크로스 모달 인코더(230)가 마스크 토큰의 본래 클래스(정답)를 예측하도록 학습한다.

[0050] 연구자는 마스크 토큰의 본래 클래스를 VQ-VAE(Vector Quantised-Variational AutoEncoder) 연구(Aaron van den Oord et al., Neural Discrete Representation LearningNeural Information Processing Systems 30, NIPS 2017)에서 이용한 이미지 패치별 코드로 이용하였다. 연구자는 사전 학습된 코드를 사용하였다.

[0051] 정리하면, 제1 크로스 모달 인코더(230)는 텍스트를 반영한 이미지 임베딩을 하고, 임베딩의 마스크 위치에 해당하는 특징을 통해 입력 이미지의 마스크 위치에 해당하는 패치 데이터를 예측한다. MIM을 통해 제1 크로스 모달 인코더(230)는 스스로 입력 이미지 데이터에 대한 정답 데이터를 생성하며 이미지 데이터의 분포를 학습하게 된다.

[0052] 학습장치는 제2 크로스 모달 인코더(230)에 대하여 MLM (Masked Language Modeling)을 수행한다.

[0053] 학습장치는 입력 단어 토큰들 중 일부를 마스크(mask) 토큰으로 변경하고, 제2 크로스 모달 인코더(240)가 마스크 토큰의 본래 클래스(정답)를 예측하도록 학습한다. 마스크 토큰의 본래 클래스는 단어 토큰의 인덱스에 해당한다.

[0054] 정리하면, 제2 크로스 모달 인코더(240)는 이미지를 반영한 텍스트 임베딩을 하고, 임베딩의 마스크 위치에 해당하는 특징을 통해 입력 텍스트의 마스크 위치에 해당하는 단어 데이터를 예측한다. MLM을 통해 제2 크로스 모달 인코더(240)는 스스로 입력 텍스트 데이터에 대한 정답 데이터를 생성하며 텍스트 데이터의 분포를 학습하게 된다.

[0055] 학습장치는 ITM (Image-Text Matching) 학습을 수행한다. ITM은 별도의 계층들로 구성된 모델을 이용한다. 해당 계층은 순서대로 덴스 레이어(dense layer), GELU(Gaussian error linear unit) 및 덴스 레이어로 구성될 수 있다. 설명의 편의를 위하여 ITM을 수행하는 계층을 매칭 계층이라고 명명한다. 매칭 계층은 제1 크로스 모달 인코더(230) 및 제2 크로스 모달 인코더(240)가 출력하는 이미지 임베딩과 텍스트 임베딩을 입력받는다. 학습장치는 ITM에서 이미지와 텍스트가 서로 정답(positive pair)인 경우 1로 예측하도록(출력하도록) 학습하고, 오답인 경우 0으로 예측하도록 학습한다. 즉, 학습장치는 매칭 계층이 제1 크로스 모달 인코더(230) 및 제2 크로스 모달 인코더(240)가 최종적으로 출력하는 이미지 임베딩과 텍스트 임베딩의 유사도를 예측하게 한다.

[0056] 학습장치는 하드 라벨에 기반하여 MIM 및 MLM 학습을 하였다. 따라서, 검색 모델은 과적합 문제가 발생할 여지가 있다. 이 경우 학습장치는 모멘텀 모델을 이용하여 검색 모델의 가중치를 조정할 수 있다.

[0057] 모멘텀 모델(300)은 검색 모델(200)과 동일한 구조를 갖는다. 다만, 모멘텀 모델(300)은 각각의 파라미터에 대해 검색 모델(200)과 다른 가중치를 갖는다. 모멘텀 모델(300)은 가중이 갱신은 그레디언트 기법을 이용하지 않

고, EMA (Exponential Moving Average) 기법으로 업데이트된다. 연구자는 검색 모델(200)의 학습 과정에서 각 가중치에 대하여 이전 단계의 검색 모델(200)의 가중치 0.5%와 이전 단계의 모멘텀 모델(300)의 가중치 99.5%를 더하여 모멘텀 모델(300)의 가중치를 업데이트하였다. 이때 가중치 비율은 실험적으로 결정될 수 있다.

[0058] 검색 모델(200)의 학습 과정에서 모멘텀 모델(300)은 검색 모델(200)의 가중치를 0.5%씩을 사용하여 조금씩 업데이트하면서 ITC, MLM 및 MIM 학습 과정에서 입력 데이터에 대한 수도 라벨을 생성한다.

[0059] 검색 모델(200)의 ITC, MLM 및 MIM 학습 과정에서 실제 정답값과 모멘텀 모델(300)의 수도 라벨을 동시에 사용한다. 예컨대, 검색 모델(200)의 학습 단계에서 최종 정답은 "실제 정답 ×(1-α)+ 모멘텀 모델(300)의 수도 정답 ×α"로 결정할 수 있다.

[0060] 학습이 완료된 검색 모델(200)은 입력되는 텍스트와 이미지의 유사도를 예측하게 된다. 따라서, 서비스 장치는 검색 모델을 이용하여 사용자가 입력하는 텍스트(또는 이미지)와 데이터베이스에 저장된 이미지들(이미지인 경우 텍스트들)의 유사도를 결정한다. 서비스 장치는 입력된 텍스트와 유사도가 1로 예측되는 이미지(들)를 추천 아이템으로 결정할 수 있다.

[0062] 연구자는 제안한 모델(CRLM로 표기)과 종래 모델의 성능을 비교하였다. 연구자는 Fashion-Gen 데이터 세트를 이용하여 각각 ITR (Image-to-Text Retrieval) 및 TIR (Text-to-Image Retrieval) 성능을 평가하였다. 아래 표 1은 제안 모델과 종래 모델의 성능 평가 결과이다. 표 1은 정확도 %를 나타낸다. 표 1을 살펴보면, 제안 모델이 종래 모델에 비하여 높은 성능을 보인다.

표 1

[0064]

Model	ITR (Image-to-Text Retrieval)			TIR (Text-to-Image Retrieval)		
	R@1	R@5	R@10	R@1	R@5	R@10
VSE	4.01	11.03	22.14	4.35	12.76	20.91
VSE++	4.59	14.99	24.10	4.60	16.89	28.99
SCAN	4.59	16.50	26.60	4.30	13.00	22.30
PFAN	4.29	14.90	24.20	6.20	20.79	31.52
ViLBERT	20.97	40.49	48.21	21.12	37.23	50.11
FashionBERT	23.96	46.31	52.12	26.75	46.48	55.74
ImageBERT	22.76	41.89	50.77	24.78	45.20	55.90
OSCAR	23.39	44.67	52.55	25.10	49.14	56.68
Kaleido-BERT	27.99	60.09	68.37	33.88	60.60	68.59
EI-CLIP	38.70	72.20	84.25	40.06	71.99	82.90
ALBEF	45.40	71.10	82.80	48.00	76.50	85.70
CRLM (제안모델)	56.90	77.60	87.30	43.80	71.00	83.20

[0066] 도 3은 이미지와 텍스트의 크로스 모달 검색을 수행하는 서비스 장치(400)에 대한 예이다. 서비스 장치(400)는 저장장치(410), 메모리(420), 연산장치(430), 인터페이스 장치(440) 및 통신장치(450)를 포함할 수 있다.

[0067] 저장장치(410)는 전술한 검색 모델(200)을 저장할 수 있다. 검색 모델은 사전에 학습된 모델이다.

[0068] 저장장치(410)는 입력되는 텍스트와 매칭되는지 확인하기 위한 이미지 세트를 저장할 수 있다. 저장장치(410)는 입력되는 이미지와 매칭되는지 확인하기 위한 텍스트 세트를 저장할 수 있다.

[0069] 저장장치(410)는 텍스트 기반 이미지 검색 또는 이미지 기반 텍스트 검색을 제어하는 프로그램을 저장할 수 있다.

[0070] 메모리(420)는 텍스트 기반 이미지 검색 또는 이미지 기반 텍스트 검색을 제어하는 과정에서 생성되는 데이터 및 정보 등을 저장할 수 있다.

[0071] 인터페이스 장치(440)는 외부로부터 일정한 명령 및 데이터를 입력받는 장치이다. 인터페이스 장치(440)는 물리적으로 연결된 입력 장치 또는 외부 저장장치로부터 사용자가 입력한 텍스트(또는 이미지)를 입력받을 수 있다. 사용자가 입력하는 값은 쿼리값이라고 명명할 수도 있다. 인터페이스 장치(440)는 텍스트 기반한 이미지 검색 결과(또는 이미지 기반한 텍스트 검색 결과)를 외부 객체에 전달할 수도 있다.

[0072] 통신장치(450)는 유선 또는 무선 네트워크를 통해 일정한 정보를 수신하고 전송하는 구성을 의미한다. 통신장치(450)는 외부 객체로부터 사용자가 입력한 텍스트(또는 이미지)를 수신할 수 있다. 또는 통신장치(450)는 텍스트 기반한 이미지 검색 결과(또는 이미지 기반한 텍스트 검색 결과)를 사용자 단말과 같은 외부 객체에 송신할 수도 있다.

[0073] 인터페이스 장치(440)는 통신장치(450)를 통해 수신된 데이터를 내부로 전달하는 구성일 수 있다.

[0074] 입력 데이터가 이미지인 경우, 연산 장치(430)는 입력 이미지를 다수의 패치들로 구분하고, 일련의 패치 토큰들로 변환하여 검색 모델의 이미지 인코더에 입력한다. 또한, 연산 장치(430)는 사전에 보유한 텍스트 세트 중 특정 텍스트를 단어 단위 토큰으로 변환하여 검색 모델의 텍스트 인코더에 입력한다. 검색 모델은 현재 입력된 이미지(패치)와 텍스트의 유사도를 예측하는 값(예컨대, 1 또는 0)을 출력한다. 예컨대, 연산장치(430)는 검색 모델이 출력하는 값이 1이면 현재 이미지와 텍스트가 대응 내지 유사하다고 판단한다. 연산장치(430)는 사용자기 입력한 이미지에 대응되는 텍스트(추천 아이템)를 결정할 수 있다. 연산장치(430)는 텍스트 세트 중 복수의 텍스트를 추천 아이템으로 결정할 수도 있다.

[0075] 입력 데이터가 텍스트인 경우, 연산 장치(430)는 입력 텍스트를 단어 토큰들로 변환하여 검색 모델의 텍스트 인코더에 입력한다. 또한, 연산 장치(430)는 사전에 보유한 이미지 세트 중 특정 이미지를 패치 단위 토큰으로 변환하여 검색 모델의 이미지 인코더에 입력한다. 검색 모델은 현재 입력된 텍스트와 이미지(패치)의 유사도를 예측하는 값(예컨대, 1 또는 0)을 출력한다. 예컨대, 연산장치(430)는 검색 모델이 출력하는 값이 1이면 현재 텍스트와 이미지가 대응 내지 유사하다고 판단한다. 연산장치(430)는 사용자기 입력한 텍스트에 대응되는 이미지(추천 아이템)를 결정할 수 있다. 연산장치(430)는 이미지 세트 중 복수의 이미지를 추천 아이템으로 결정할 수도 있다.

[0076] 연산 장치(430)는 데이터를 처리하고, 일정한 연산을 처리하는 프로세서, AP, 프로그램이 임베디드된 칩과 같은 장치일 수 있다.

[0078] 또한, 상술한 바와 같은 크로스 모달 검색 방법, 이미지 기반 텍스트 검색 방법 및 텍스트 기반 이미지 검색 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.

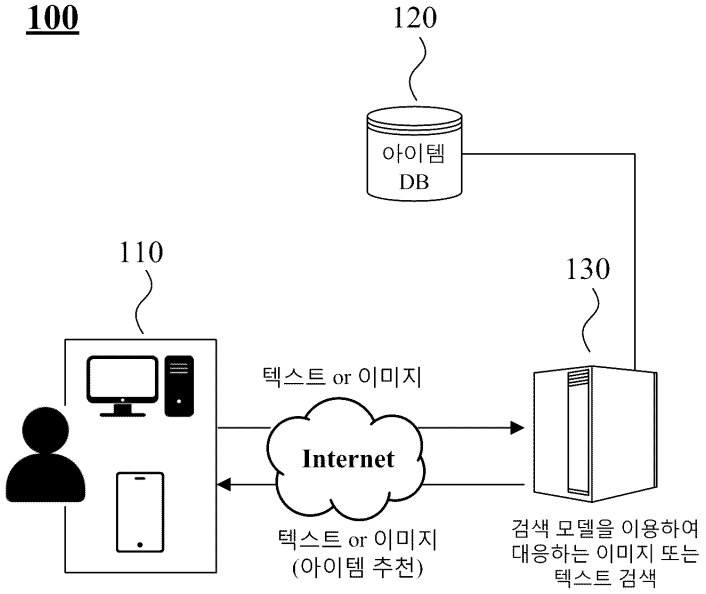
[0079] 비일시적 판독 가능 매체란 레지스터, 캐쉬, 메모리 등과 같이 짧은 순간 동안 데이터를 저장하는 매체가 아니라 반영구적으로 데이터를 저장하며, 기기에 의해 판독(reading)이 가능한 매체를 의미한다. 구체적으로는, 상술한 다양한 어플리케이션 또는 프로그램들은 CD, DVD, 하드 디스크, 블루레이 디스크, USB, 메모리카드, ROM(read-only memory), PROM (programmable read only memory), EPROM(Erasable PROM, EPROM) 또는 EEPROM(Electrically EPROM) 또는 플래시 메모리 등과 같은 비일시적 판독 가능 매체에 저장되어 제공될 수 있다.

[0080] 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, SLDRAM) 및 직접 램버스 램(Direct Rambus RAM, DRRAM) 과 같은 다양한 RAM을 의미한다.

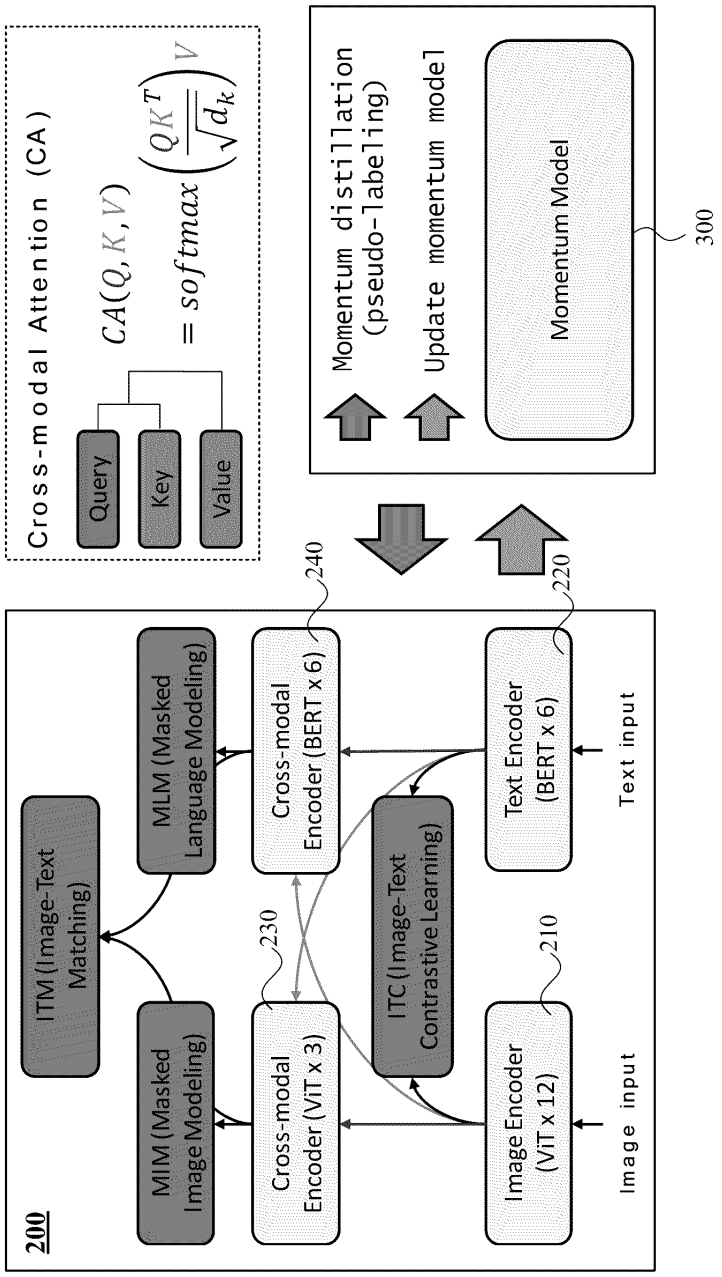
[0081] 본 실시예 및 본 명세서에 첨부된 도면은 전술한 기술에 포함되는 기술적 사상의 일부를 명확하게 나타내고 있는 것에 불과하며, 전술한 기술의 명세서 및 도면에 포함된 기술적 사상의 범위 내에서 당업자가 용이하게 유추할 수 있는 변형 예와 구체적인 실시례는 모두 전술한 기술의 권리범위에 포함되는 것이 자명하다고 할 것이다.

도면

도면1


100
120
아이템
DB
110
130
텍스트 or 이미지
Internet
텍스트 or 이미지
(아이템 추천)
검색 모델을 이용하여
대응하는 이미지 또는
텍스트 검색

도면2


Cross-modal Attention (CA)
$CA(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$
Query
Key
Value
Momentum distillation (pseudo-labeling)
Update momentum model
Momentum Model
300
200
ITM (Image-Text Matching)
MLM (Masked Language Modeling)
MIM (Masked Image Modeling)
Cross-modal Encoder (BERT x 6)
240
Cross-modal Encoder (ViT x 3)
230
ITC (Image-Text Contrastive Learning)
Text Encoder (BERT x 6)
220
Image Encoder (ViT x 12)
210
Text input
Image input

도면3