



(19) 대한민국특허청(KR)  
(12) 등록특허공보(B1)

(45) 공고일자 2023년11월03일  
(11) 등록번호 10-2598678  
(24) 등록일자 2023년11월01일

(51) 국제특허분류(Int. Cl.)  
G06F 40/174 (2020.01) G06F 16/33 (2019.01)  
G06F 16/583 (2019.01) G06F 16/783 (2019.01)  
G06F 40/284 (2020.01) G06N 3/04 (2023.01)  
G06N 3/08 (2023.01)  
(52) CPC특허분류  
G06F 40/174 (2020.01)  
G06F 16/3347 (2019.01)  
(21) 출원번호 10-2021-0147768  
(22) 출원일자 2021년11월01일  
심사청구일자 2021년11월01일  
(65) 공개번호 10-2023-0063003  
(43) 공개일자 2023년05월09일  
(56) 선행기술조사문헌  
KR1020200106115 A\*  
KR1020210114074 A\*  
KR1020210053864 A  
US20180329892 A1  
\*는 심사관에 의하여 인용된 문헌

(73) 특허권자  
연세대학교 산학협력단  
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)  
(72) 발명자  
박상현  
서울특별시 양천구 오목로 300, 204동 3701호(목동, 현대하이페리온2)  
이지은  
서울특별시 서대문구 동교로 291, 101동 1004호(연희동, 연희동임광아파트)  
박진욱  
서울특별시 노원구 덕릉로 613, 301동 307호(중계동, 우성3차아파트)  
(74) 대리인  
특허법인(유한)아이시스

전체 청구항 수 : 총 3 항

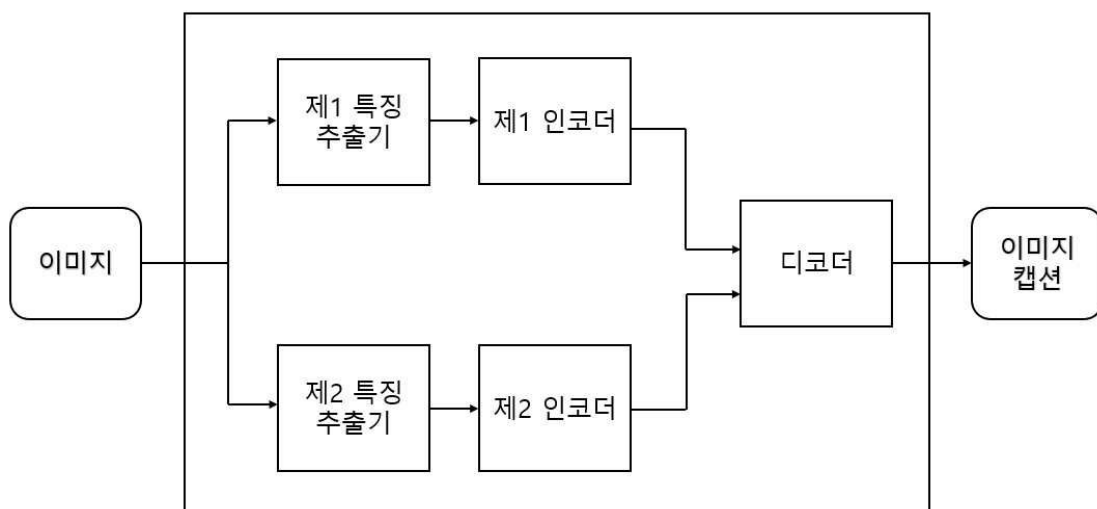
심사관 : 김경완

(54) 발명의 명칭 이미지 캡션 생성 방법 및 장치

(57) 요약

이미지 캡션 생성 방법 및 장치가 제공된다. 이미지 캡션 생성 방법은 복수의 특징 추출기에서 이미지로부터 복수의 특징 벡터를 추출하는 단계, 복수의 인코더에서 상기 복수의 특징 벡터 각각에 기초하여 상기 이미지에 대한 복수의 이미지 특징 값을 생성하는 단계, 및 디코더에서 상기 복수의 이미지 특징 값을 기초로 상기 이미지에 대한 이미지 캡션을 생성하는 단계를 포함한다.

대표도 - 도2



이미지 캡션 생성 장치

(52) CPC특허분류

*G06F 16/583* (2019.01)

*G06F 16/783* (2019.01)

*G06F 40/284* (2020.01)

*G06N 3/045* (2023.01)

*G06N 3/08* (2023.01)

이 발명을 지원한 국가연구개발사업

과제고유번호 1711103018

과제번호 2017-0-00477-004

부처명 과학기술정보통신부

과제관리(전문)기관명 정보통신기획평가원

연구사업명 정보통신방송연구개발사업

연구과제명 (SW 스타랩) IoT 환경을 위한 고성능 플래시 메모리 스토리지 기반 인메모리 분산 DBMS 연구개발

기 여 율 1/1

과제수행기관명 연세대학교 산학협력단

연구기간 2021.01.01 ~ 2021.12.31

공지예외적용 : 있음

---

## 명세서

### 청구범위

#### 청구항 1

이미지 캡션 생성 방법으로서,

복수의 특징 추출기에서 이미지로부터 복수의 특징 벡터를 추출하는 단계;

복수의 인코더에서 상기 복수의 특징 벡터를 각각 입력받아 상기 이미지에 대한 복수의 이미지 특징 값을 각각 생성하는 단계; 및

디코더에서 상기 복수의 이미지 특징 값을 기초로 상기 이미지에 대한 이미지 캡션을 생성하는 단계를 포함하되,

상기 복수의 특징 추출기는 서로 다른 유형의 특징 추출기이고,

상기 복수의 이미지 특징 값은 제1 특징 값 및 제2 특징값을 포함하고,

상기 디코더는 상기 제1 특징 값을 입력받아 처리하는 제1 멀티-헤드 주의 계층, 상기 제1 멀티-헤드 주의 계층의 출력값 및 상기 제2 특징값을 입력받아 처리하는 제2 멀티-헤드 주의 계층, 상기 제2 멀티-헤드 주의 계층을 출력값을 입력받아 처리하는 마스크 멀티-헤드 주의 계층(Masked Multi-Head Attention) 및 상기 마스크 멀티-헤드 주의 계층의 출력을 입력받는 포지션-와이즈 순방향 네트워크를 포함하는 것을 특징으로 하는 이미지 캡션 생성 방법.

#### 청구항 2

삭제

#### 청구항 3

삭제

#### 청구항 4

삭제

#### 청구항 5

삭제

#### 청구항 6

삭제

#### 청구항 7

삭제

#### 청구항 8

제1 항에 있어서,

상기 복수의 특징 추출기는 Faster R-CNN 모델인 제1 특징 추출기 및 Mask R-CNN 모델인 제2 특징 추출기를 포함하는 것을 특징으로 하는 이미지 캡션 생성 방법.

#### 청구항 9

삭제

#### 청구항 10

제8 항에 있어서,

상기 복수의 인코더는 제1 인코더 및 제2 인코더를 포함하고,

상기 제1 인코더는 상기 제1 특징 추출기로부터 추출된 제1 특징 벡터에 기초하여 상기 제1 특징 값을 생성하고,

상기 제2 인코더는 상기 제2 특징 추출기로부터 추출된 제2 특징 벡터에 기초하여 상기 제2 특징 값을 생성하고,

상기 제1 인코더 및 상기 제2 인코더 각각에서 상기 제1 특징 벡터 및 상기 제2 특징 벡터가 동일한 차원을 갖도록 선형 변환을 수행하는 단계를 더 포함하는 것을 특징으로 하는 이미지 캡션 생성 방법.

#### 청구항 11

삭제

#### 청구항 12

삭제

#### 청구항 13

삭제

#### 청구항 14

삭제

#### 청구항 15

삭제

#### 청구항 16

삭제

#### 청구항 17

삭제

#### 청구항 18

삭제

#### 청구항 19

삭제

### 발명의 설명

### 기술 분야

[0001] 이하 설명하는 기술은 이미지 캡션 생성 방법 및 장치에 관한 것이다.

### 배경 기술

[0002] 이미지 캡션 생성(Image Captioning)은 주어진 이미지의 구성요소를 파악하여, 장면을 묘사해주는 자연어를 자동으로 생성하는 작업이다.

[0003] 이미지 캡션 생성은 여러 어플리케이션에 적용될 수 있다. 예를 들어, 시각 장애인들을 위해 이미지를 실시간으로 설명해주거나, 웨어러블 카메라를 통한 라이프로그 사진의 캡션을 자동으로 생성할 수 있다. 또한, 드론을

통한 범죄 및 안전 사고를 감지하여, 스마트시티 분야에서도 활용될 수 있다.

## 선행기술문헌

### 특허문헌

[0004] (특허문헌 0001) 미국등록특허 US10,915,701호

## 발명의 내용

### 해결하려는 과제

- [0005] 이미지 캡션 생성 모델의 성능은 특징 추출기로 뽑아내는 이미지 정보의 우수성과 관련이 있다. 추출된 이미지 정보의 우수성이 낮을수록 언어 모델이 생성하는 캡션은 일반적인 문장이 된다는 한계가 있다.
- [0006] 기존의 이미지 캡션 생성 모델들은 인코더에서 하나의 특징 추출기를 통해 단일 관점으로 이미지의 정보를 추출하여, 이미지 정보의 우수성이 낮을 수 있다. 또한, 특징 추출기를 통해서 우수한 이미지 정보를 뽑아내더라도, 캡션을 생성하는 언어 모델의 성능이 떨어지는 경우, 캡션의 질이 낮을 수 있다.
- [0007] 따라서, 높은 우수성을 갖는 이미지 정보를 추출하고, 추출된 이미지 정보로부터 높은 질의 캡션을 생성하기 위한 이미지 캡션 생성 방법 및 장치에 대한 연구가 필요하다.

### 과제의 해결 수단

- [0008] 본 발명의 일 양태에 따른 이미지 캡션 생성 방법은 복수의 특징 추출기에서 이미지로부터 복수의 특징 벡터를 추출하는 단계, 복수의 인코더에서 상기 복수의 특징 벡터 각각에 기초하여 상기 이미지에 대한 복수의 이미지 특징 값을 생성하는 단계, 및 디코더에서 상기 복수의 이미지 특징 값을 기초로 상기 이미지에 대한 이미지 캡션을 생성하는 단계를 포함한다.
- [0009] 본 발명의 일 양태에 따른 이미지 캡션 생성 방법은 상기 디코더에서 자가 교열 트랜스포머(self-revising transformer)를 사용하는 것을 특징으로 한다.
- [0010] 본 발명의 일 양태에 따른 이미지 캡션 생성 방법은 상기 자가 교열 트랜스포머에서 제1 마스크 멀티-헤드 주의 기제 기법(Masked Multi-Head Attention), 스택 멀티-모달 주의 기제 기법 (Stacked Multimodal Attention), 및 제2 마스크 멀티-헤드 주의 기제 기법을 순차적으로 사용하는 것을 특징으로 한다.
- [0011] 본 발명의 일 양태에 따른 이미지 캡션 생성 방법은 상기 제1 마스크 멀티-헤드 주의 기제 기법 및 상기 제2 마스크 멀티-헤드 주의 기제 기법은 현재 단계 이전의 데이터만을 기초로 멀티-헤드 주의 기제 기법을 사용하는 것을 특징으로 한다.
- [0012] 본 발명의 일 양태에 따른 이미지 캡션 생성 방법은 상기 자가 교열 트랜스포머에서 상기 제2 마스크 멀티-헤드 주의 기제 기법을 사용한 이후에, 포지션-와이즈 순방향 네트워크(Position-wise Feed-Forward Network)를 사용하는 것을 특징으로 한다.
- [0013] 본 발명의 일 양태에 따른 이미지 캡션 생성 방법은 상기 스택 멀티-모달 주의 기제 기법에서 상기 복수의 이미지 특징 값들을 기초로 멀티-헤드 주의 기제 기법(Multi-Head Attention)을 연속적으로 적용하는 것을 특징으로 한다.
- [0014] 본 발명의 일 양태에 따른 이미지 캡션 생성 방법은 상기 복수의 특징 추출기는 제1 특징 추출기 및 제2 특징 추출기를 포함하되, 상기 제1 특징 추출기 및 상기 제2 특징 추출기는 서로 상이한 지역 기반 합성곱 신경망(R-CNN: Region based Convolution Neural Network) 모델을 사용하는 것을 특징으로 한다.
- [0015] 본 발명의 일 양태에 따른 이미지 캡션 생성 방법은 상기 제1 특징 추출기 및 상기 제2 특징 추출기 중 하나의 특징 추출기는 Faster R-CNN 모델을 사용하고, 상기 제1 특징 추출기 및 상기 제2 특징 추출기 중 다른 하나의 특징 추출기는 Mask R-CNN 모델을 사용하는 것을 특징으로 한다.
- [0016] 본 발명의 일 양태에 따른 이미지 캡션 생성 방법에서 상기 복수의 인코더는 제1 인코더 및 제2 인코더를 포함하되, 상기 제1 인코더에서 상기 제1 특징 추출기로부터 추출된 제1 특징 벡터에 기초하여 제1 이미지 특징 값

을 생성하는 단계, 및 상기 제2 인코더에서 상기 제2 특징 추출기로부터 추출된 제2 특징 벡터에 기초하여 제2 이미지 특징 값을 생성하는 단계를 더 포함하는 것을 특징으로 한다.

[0017] 본 발명의 일 양태에 따른 이미지 캡션 생성 방법에서 상기 제1 인코더 및 상기 제2 인코더 각각에서 상기 제1 특징 벡터 및 상기 제2 특징 벡터가 동일한 차원을 갖도록 선형 변환을 수행하는 단계를 더 포함하는 것을 특징으로 한다.

[0018] 본 발명의 일 양태에 따른 이미지 캡션 생성 방법은 상기 디코더에서 상기 제1 이미지 특징 값 및 상기 제2 이미지 특징 값을 입력으로 받는 단계, 상기 디코더에서 상기 제1 이미지 특징 값을 기초로 제1 멀티-헤드 주의 기제 기법을 적용하는 단계, 및 상기 디코더에서 상기 제1 멀티-헤드 주의 기제 기법의 결과 값과 상기 제2 이미지 특징 값을 기초로 제2 멀티-헤드 주의 기제 기법을 적용하는 단계를 더 포함하는 것을 특징으로 한다.

[0019] 본 발명의 일 양태에 따른 이미지 캡션 생성 방법은 상기 디코더에서 상기 제2 멀티-헤드 주의 기제 기법의 결과 값에 마스크 멀티-헤드 주의 기제 기법을 적용하는 단계를 더 포함하는 것을 특징으로 한다.

[0020] 본 발명의 일 양태에 따른 이미지 캡션 생성 장치는 복수의 특징 추출기, 복수의 인코더, 및 디코더를 포함한다. 상기 복수의 특징 추출기는 이미지로부터 복수의 특징 벡터를 추출하고, 상기 복수의 인코더는 상기 복수의 특징 벡터 각각에 기초하여 상기 이미지에 대한 복수의 이미지 특징 값을 생성하고, 상기 디코더는 상기 복수의 이미지 특징 값을 기초로 상기 이미지에 대한 이미지 캡션을 생성한다.

### 발명의 효과

[0021] 이하에서 설명하는 기술은 추출된 이미지 정보의 우수성 및 추출된 이미지 정보로부터 생성된 캡션의 질을 향상시킬 수 있다.

[0022] 구체적으로, 본 발명은 다중 관점을 가진 자가 교열 트랜스포머에 기반한 이미지 캡션 생성 장치를 사용하여 효율적으로 이미지 캡션을 생성할 수 있다.

[0023] 예를 들어, 본 발명의 몇몇 실시예에 따른 이미지 캡션 생성 방법 및 장치는 이미지 특징 추출기를 복합적으로 사용하여 다중 관점의 이미지 정보를 추출할 수 있다. 또한, 다중 관점 인코더를 사용하여 이미지 정보 사이의 중요도를 계산할 수 있다.

[0024] 예를 들어, 본 발명의 몇몇 실시예에 따른 이미지 캡션 생성 방법 및 장치는 두 가지의 특징 추출기(예를 들어, Faster R-CNN 및 Mask R-CNN)를 적용하여 같은 이미지에서 서로 다른 정보를 얻어, 부족한 정보를 상호 보완시킬 수 있다. 또한, 트랜스포머 기반의 다중 관점 인코더를 통해서 얻은 이미지 정보의 중요도를 산정하여 업데이트 함으로써, 높은 질의 캡션을 생성할 수 있다.

[0025] 예를 들어, 본 발명의 몇몇 실시예에 따른 이미지 캡션 생성 방법 및 장치는 자가 교열 트랜스포머(Self-Revising Transformer)를 사용하여 캡션의 중복성을 제거하고 정확성을 향상시킬 수 있다. 구체적으로, 본 발명에서 제안되는 자가 교열 트랜스포머는 언어 모델의 역할을 강화하여 단어의 중복된 표현을 억제하고, 생성된 문장을 재구축하여 연결성을 높은 문장을 생성할 수 있다.

### 도면의 간단한 설명

[0026] 도 1은 본 발명의 몇몇 실시예에 따른 이미지 캡션 생성 장치를 나타낸다.

도 2는 본 발명의 몇몇 실시예에 따른 이미지 캡션 생성 장치를 설명하기 위한 구조도이다.

도 3은 본 발명의 몇몇 실시예에 따른 이미지 캡션 생성 장치를 적용한 일 예를 나타낸다.

도 4은 본 발명의 몇몇 실시예에 따른 다중 관점 인코더를 설명하기 위한 블록도이다.

도 5는 본 발명의 몇몇 실시예에 따른, 인코더와 디코더의 결합 방법을 나타낸다.

도 6은 스택 멀티모달 주의 기제 기법의 일 예를 나타내는 블록도이다.

도 7은 본 발명의 몇몇 실시예에 따른 이미지 캡션 생성 장치의 디코더를 설명하기 위한 블록도이다.

도 8은 본 명세서에서 사용되는 모델 손실 함수(Model Loss Function)의 흐름도이다.

도 9는 특징 추출기에 따른 이미지 캡션 생성 모델이 생성한 문장 예시를 나타낸다.

도 10은 자가 교열 트랜스포머의 유무에 따른 이미지 캡션 생성 장치의 성능을 설명하기 위한 예시이다.

도 11은 본 발명의 몇몇 실시예에 따른 이미지 캡션 생성 방법의 흐름도이다.

### 발명을 실시하기 위한 구체적인 내용

- [0027] 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.
- [0028] 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.
- [0029] 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.
- [0030] 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.
- [0031] 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.
- [0032] 이하에서, 이미지 캡션 생성 방법에 대해 설명한다. 이하에서 설명되는 이미지 캡션 생성 방법은 본 발명의 몇몇 실시예에 따른 이미지 캡션 생성 방법 및 장치에 활용될 수 있다.
- [0033] 이미지 캡션 생성에서는 딥러닝(Deep Learning) 기술이 활용될 수 있다. 예를 들어, 이미지 캡션 생성에서는 합성곱 신경망(Convolution Neural Network) 모델 및 순환 신경망(Recurrent Neural Network) 모델이 사용될 수 있다.
- [0034] 합성곱 신경망 모델은 컴퓨터 비전 분야의 이미지 정보를 얻는 머신 러닝 기법이다. 순환 신경망 모델은 기계 번역(Machine Translation)에서 일반적으로 사용되는 머신 러닝 기법이다.
- [0035] 이미지 캡션 생성 모델은 인코더-디코더 구조를 가질 수 있다. 인코더는 사전 학습된 합성곱 신경망 모델을 전이 학습(Transfer Learning)하여 이미지의 오브젝트 분류 및 특징을 얻는 특징 추출기(Feature Extractor)로 사용될 수 있다.
- [0036] 예를 들어, 합성곱 신경망 모델은 사전 학습된 이미지 분류(Image Classification) 모델을 포함할 수 있다. 예를 들어, 객체 탐지(Object Detection)를 위한 Faster R-CNN을 또는 Faster R-CNN을 일부 수정한 모델을 합성곱 신경망 모델로 사용할 수 있다.
- [0037] 디코더는 순환 신경망 모델에 속하는 LSTM(Long Short-Term Memory)이나 GRU(Gated Recurrent Unit)를 통해 인코더에서 얻은 이미지 특징 정보와 결합하여 자연어로 이루어진 캡션을 생성할 수 있다.
- [0038] 또한, 이미지 캡션 생성 방법에서 주의 기제 기법(Attention)이 사용될 수 있다. 주의 기제 기법은 신경망의 맥락에서 인지 주의(cognitive attention)를 모방하는 기술이다. 주의 기제 기법은 입력 데이터의 중요한 부분을 향상시키고 나머지는 희미하게 만들 수 있다. 주의 기제 기법은 전체 데이터 중 작지만 중요한 부분에 더 많은 컴퓨팅 리소스를 할당할 수 있다. 주의 기제 기법은 트랜스포머에 사용될 수 있다. 또한, 주의 기제 기법은 합



성급 신경망을 기반으로 하는 컴퓨터 비전 시스템에 적용될 수 있다.

- [0039] 이미지 캡션 생성 모델에서 주의 기제 기법은 단어와 이미지 사이의 관계 정보를 사용하여 자연어를 생성할 때 이미지의 특정 부분을 집중해줌으로써, 이미지 캡션 생성 모델은 이미지 안의 객체를 더 잘 추출하도록 할 수 있다.
- [0040] 기계 번역 관련 기술이 발전함에 따라, 이미지 캡션 생성 모델의 성능이 향상될 수 있다. 예를 들어, 시퀀스 (Sequence)에서 시퀀스를 생성하는 기계 번역 모델이 이미지 캡션 생성 모델에 적용될 수 있다. 이 경우, 이미지 캡션 생성 모델은 인코더-디코더 구조를 적용한 모델을 포함할 수 있다. 인코더-디코더 구조에서 인코더는 입력 값이 이미지인 것에 따라 순환 신경망 모델 대신 합성곱 신경망 모델을 사용할 수 있다.
- [0041] 그러나 디코더에서 순환 신경망 모델을 사용하는 경우, 장기 의존성(Long-Term Dependency)로 인한 정보 손실이 발생할 수 있다. 이를 보완하기 위해 이미지 캡션 생성 모델에 하드 주의 기제 기법(Hard Attention)과 소프트 주의 기제 기법(Soft Attention)을 사용할 수 있다.
- [0042] 소프트 주의 기제 기법으로서 적응형 주의 기제 기법이 사용될 수 있다. 예를 들어, 적응형 주의 기제 기법 (Adaptive Attention)을 통해 캡션을 생성할 때, 이미지 정보와 언어 모델 중 어떤 정보를 편중할지 게이트를 통해 결정할 수 있다.
- [0043] 트랜스포머 모델은 자연어 처리 분야에서 사용되는 모델로서, 높은 성능의 기계 번역을 제공할 수 있다. 트랜스포머 모델은 기존의 자연어처리에서 일반적으로 사용된 순환 신경망 모델 대신 주의 기제 기법을 사용한다. 따라서, 입력 문장 사이의 관계 정보를 추가로 갖을 수 있게 되었고, 훈련 시간을 크게 단축할 수 있다.
- [0044] 이미지 캡션 생성 모델에서, 트랜스포머 모델은 적절히 수정되어 사용될 수 있다. 예를 들어, 기하학 주의 기제 기법(Geometric Attention)을 통해 이미지 안의 객체 간의 관계에 대한 정보를 결합할 수 있다. 예를 들어, 이미지 캡션 생성 모델은 트랜스포머 모델의 디코더 구조를 차용하고 멀티-레벨 지도학습을 적용할 수 있다.
- [0045] 한편, 이미지 캡션 생성 모델의 성능은 특징 추출기로 뽑아내는 이미지 정보의 우수성과 관련이 있다. 추출된 이미지 정보의 우수성이 낮을수록 언어 모델이 생성하는 캡션은 일반적인 문장이 된다는 한계가 있다.
- [0046] 기존의 이미지 캡션 생성 모델들은 인코더에서 하나의 특징 추출기를 통해 단일 관점으로 이미지의 정보를 추출하여, 이미지 정보의 우수성이 낮을 수 있다. 또한, 특징 추출기를 통해서 우수한 이미지 정보를 뽑아내더라도, 캡션을 생성하는 언어 모델의 성능이 떨어지는 경우, 캡션의 질이 낮을 수 있다.
- [0047] 따라서, 높은 우수성을 갖는 이미지 정보를 추출하고, 추출된 이미지 정보로부터 높은 질의 캡션을 생성하기 위한 이미지 캡션 생성 방법 및 장치에 대한 연구가 필요하다.
- [0048] 이하에서, 본 발명의 몇몇 실시예에 따른 이미지 캡션 생성 방법 및 장치에 대해 설명한다. 이하의 도면은 본 발명의 구체적인 실시예를 설명하기 위해 작성된 것이다. 도면에 나타난 특정 장치의 명칭은 예시적인 것으로, 본 발명의 기술적 사상이 하기 도면에서 사용되는 특정 명칭에 한정되는 것은 아니다.
- [0049] 또한, 본 명세서에서, 이미지 캡션 생성 장치 또는 방법은 이미지 캡션 생성 모델로 표현될 수 있다. 예를 들어, 이미지 캡션 생성 장치는 이미지 캡션 생성 모델을 사용하는 장치일 수 있다. 또한, 이미지 캡션 생성 방법은 이미지 캡션 생성 모델이 동작하는 방법일 수 있다.
- [0050] 도 1은 본 발명의 몇몇 실시예에 따른 이미지 캡션 생성 장치를 나타낸다. 도 1을 참조하면, 본 발명에 따른 이미지 캡션 생성 장치는 이미지를 입력으로 받아, 이미지를 잘 설명하는 이미지 캡션을 생성할 수 있다.
- [0051] 도 1의 이미지 캡션 생성 장치에서 사용되는 이미지 캡션 생성 모델은 입력 이미지의 특징 간의 관계 정보를 얻기 위하여 다중 관점 인코더와 순환 신경망의 장기 의존성 한계를 극복한 트랜스포머 모델 기반의 디코더를 사용할 수 있다.
- [0052] 구체적으로, 본 발명의 이미지 캡션 생성 모델은 일반적인 이미지 캡션 생성 모델과는 상이하게, 단일 특징 추출기가 아닌 다중 특징 추출기를 사용하여, 다양한 관점의 이미지의 정보를 추출할 수 있다. 또한, 본 발명의 이미지 캡션 생성 모델에 포함된 디코더는 기존의 트랜스포머와는 상이한, 언어 모델의 역할을 강화한 자가 교열 트랜스포머를 사용할 수 있다. 나아가, 본 발명의 몇몇 실시예에 따른 이미지 캡션 생성 모델은 SCST를 통해 최적화되어, 더욱 높은 성능을 가질 수 있다.
- [0053] 본 발명의 몇몇 실시예에 따른 이미지 캡션 생성 모델은 순환 신경망을 사용하는 전통적인 방법과는 달리 주의



기계 기법만으로 훈련하는 기계 번역 모델인 트랜스포머 모델의 구조를 사용할 수 있다.

- [0054] 본 발명의 이미지 캡션 생성 모델에 포함된 트랜스포머 모델은 인코더 및 디코더를 이용하여 구현될 수 있다. 예를 들어, 트랜스포머 모델은 이하의 도 2에 개시된 복수의 인코더 (즉, 제1 인코더 및 제2 인코더) 및 디코더에 의해 구현될 수 있다.
- [0055] 트랜스포머 모델은 (1) 위치 인코딩, (2) 스케일 내적 주의 기제, (3) 멀티-헤드 주의 기제, 및 (4) 포지션-와이즈 순방향 네트워크와 같은 4가지의 핵심 네트워크로 구성될 수 있다.
- [0056] 위치 인코딩에 대해 설명한다. 자연어를 모델이 이해할 수 있도록 실수 벡터로 매핑하는 과정을 단어 임베딩 (Word Embedding)이라고 한다. 전통적인 순환 신경망 모델을 사용하는 자연어 처리 과정에서, 단어 임베딩을 통해 자연어를 벡터로 매핑을 할 수 있다.
- [0057] 트랜스포머 모델 또한 단어 임베딩이 필요하지만, 순차적이거나 합성곱 방법론이 아닌 주의 기제를 사용하는 병렬적인 방법론이기 때문에 입력 문장의 단어들의 순서들을 유지하기가 어려울 수 있다. 따라서, 단어의 위치 정보를 단어 임베딩에 추가로 제공할 필요가 있다. 트랜스포머 모델은 이 문제점을 보완하기 위해서 위치 인코딩 (Positional Encoding)을 통해 단어의 상대적, 절대적 위치에 대한 정보를 단어 임베딩에 포함시킬 수 있다.
- [0058] 본 발명의 몇몇 실시예에 따르면, 위치 인코딩은 생략될 수 있다. 예를 들어, 위치 정보가 중요하지 않은 이미지를 입력 값으로 사용하는 경우, 인코더에서 위치 인코딩을 적용하지 않을 수 있다. 이 경우, 문장을 입력 값으로 사용하는 디코더에는 위치 인코딩이 적용될 수 있다.
- [0059] 스케일 내적 주의 기제에 대해 설명한다.
- [0060] 주의 기제 기법이 핵심인 트랜스포머 모델은 내적 주의 기제 기법 (Dot-Product Attention)을 수정하여 사용할 수 있다.
- [0061] 내적 주의 기제 기법에서 사용되는 수식은 아래의 수학적식(1)과 같다.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

- [0062]
- [0063] 수학적식 (1)에서, 주어지는 입력 벡터는 Q(Query), K(Key), V(Value) 그리고  $d_k$  (Key의 크기)이다. Query와 Key의 내적을 구한 뒤 Key의 크기의 제곱근으로 나눠 줌으로써 벡터의 규모를 축소한다. 소프트맥스(Softmax) 함수를 사용하여 벡터 안의 인자들을 합이 1이 되는 분포로 변환하고, Value를 곱함으로써 Value에 대한 가중치를 얻는다.
- [0064] 트랜스포머 모델은 내적 주의 기법에서 추가적으로 Key의 크기로 벡터의 규모를 축소하는 규모 축소 내적 주의 기제 기법 (Scaled Dot-Product Attention)을 사용할 수 있다. Key의 크기가 커질수록 내적 결과 값의 규모가 커지기 때문에, 소프트맥스 함수를 취하게 되면 그라디언트 값이 급격히 줄어든다. 때문에 소프트맥스 함수를 취하기 전에 벡터의 규모를 축소해줄 필요가 있다.
- [0065] 멀티-헤드 주의 기제에 대해 설명한다.
- [0066] 일반적인 주의 기제 기법은 위의 수학적식 (1)을 한 번 연산하는 것으로 Value의 가중치를 얻는다.
- [0067] 그러나, 본 발명에 따른 트랜스포머 모델은 한 번의 연산으로 끝내지 않고 하나의 데이터로부터 복수의 다른 표현형을 얻은 후 합쳐서 다각도의 정보를 얻는 방법론을 사용할 수 있다.
- [0068] 구체적으로, 멀티-헤드 주의 기제 기법에서는, 아래의 수학적식 (2) 및 (3)과 같이, Query, Key, 및 Value를 h번 선형 투영을 하여 h 번의 스케일 내적 주의 기제를 계산할 수 있다.

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (2)$$

- [0069]
- $$MHAttention(Q, K, V) = [head_1, \dots, head_h]W^O \quad (3)$$
- [0070]

- [0071] 수학적식 (2)는 멀티-헤드 주의 기제 기법에서 각각의 head를 계산하기 위한 식이다. 수학적식 (2)에서는 Query,

Key, 및 value를 기초로 스케일 내적 주의 기제를 적용하여, 각각의 head (즉,  $head_i$ )를 계산할 수 있다.

[0072] 수학식 (3)은 모든 head를 취합하여 최종 값을 얻기 위한 식을 나타낸다. 수학식 (3)에서는, 수학식 (2)에서 계산된 h 개의 head를 취합하여, 멀티-헤드 주의 기제 기법의 최종 값을 계산할 수 있다.

[0073] 수학식 (2) 및 (3)에서, W는 모두 학습 가능한 매개변수 매트릭스(learnable parameter matrix)이다.

[0074] 이와 같은 수학식 (2) 및 (3)을 통해 단어들에 대한 다양한 주의 기제 정보를 얻을 수 있다.

[0075] 포지션-와이즈 순방향 네트워크에 대해 설명한다.

[0076] 본 발명에 따른 트랜스포머 모델의 레이어들은 서로 독립된 포지션-와이즈 순방향 네트워크(Position-wise Feed-Forward Network)를 포함할 수 있다. 포지션-와이즈 순방향 네트워크는 아래의 수학식 (4)와 같이 두 개의 선형 변환과 렐루(ReLU) 활성화 함수로 계산될 수 있다.

$$FFN(x) = \max(0, xW_1 + b_1) W_2 + b_2 \quad (4)$$

[0077]

[0078] 수식 (4)에서 W1과 W2는 아래와 같다.

$$W_1 \in \mathbb{R}^{d_{ff}}, W_2 \in \mathbb{R}^{d_{model}}$$

[0079]

[0080] 도 2는 본 발명의 몇몇 실시예에 따른 이미지 캡션 생성 장치를 설명하기 위한 구조도이다.

[0081] 도 2를 참조하면, 이미지 캡션 생성 장치는 제1 특징 추출기, 제2 특징 추출기, 제1 인코더, 제2 인코더, 및 디코더를 포함할 수 있다.

[0082] 도 1에서 이미지 캡션 생성 장치는 두 개의 특징 추출기 및 두 개의 인코더를 포함하는 것으로 개시되었으나, 본 발명은 이에 제한되지 않는다. 예를 들어, 본 발명에 따른 이미지 캡션 생성 장치는 n 개의 특징 추출기 및 n 개의 인코더를 포함할 수 있다. 이때, n은 3 이상의 자연수 일 수 있다.

[0083] 도 1 을 참조하면, 이미지 캡션 생성 장치는 이미지를 입력으로 받아 이미지 캡션을 생성할 수 있다.

[0084] 이미지 캡션 생성 장치에 이미지가 입력되는 경우, 제1 특징 추출기 및 제2 특징 추출기는 동일한 이미지를 입력으로 받을 수 있다. 제1 특징 추출기 및 제2 특징 추출기는 각각 이미지로부터 특징 벡터를 생성할 수 있다. 예를 들어, 제1 특징 추출기는 제1 특징 벡터를 생성하고, 제2 특징 추출기는 제2 특징 벡터를 생성할 수 있다.

[0085] 복수의 특징 추출기에서 생성된 특징 벡터는 각각 별개의 인코더로 입력될 수 있다. 예를 들어, 제1 특징 추출기에서 생성된 제1 특징 벡터는 제1 인코더로 입력될 수 있다. 제2 특징 추출기에서 생성된 제2 특징 벡터는 제2 인코더로 입력될 수 있다.

[0086] 복수의 인코더를 통해 각각의 특징 벡터를 인코딩한 결과 값은 디코더에 입력될 수 있다. 예를 들어, 제1 인코더는 제1 특징 벡터를 인코딩한 제1 결과 값을 출력할 수 있고, 출력된 제1 결과 값은 디코더에 입력될 수 있다. 제2 인코더는 제2 특징 벡터를 인코딩한 제2 결과 값을 출력할 수 있고, 출력된 제2 결과 값은 디코더에 입력될 수 있다.

[0087] 디코더는 입력된 제1 결과 값 및 제2 결과 값을 기초로, 이미지 캡션을 생성할 수 있다. 예를 들어, 디코더는 제1 결과 값 및 제2 결과 값을 동시에 사용하여, 이미지를 설명하는 이미지 캡션을 생성할 수 있다.

[0088] 도 3은 본 발명의 몇몇 실시예에 따른 이미지 캡션 생성 장치를 적용한 일 예를 나타낸다. 예를 들어, 도 3의 이미지 캡션 생성 장치는 도 1 및 도 2를 참조하여 설명된 이미지 캡션 생성 장치와 동일할 수 있다.

[0089] 도 3을 참조하면, 제1 특징 추출기 및 제2 특징 추출기는 각각 이미지로부터 제1 특징 벡터 및 제2 특징 벡터를 추출할 수 있다. 제1 인코더 및 제2 인코더는 각각 제1 특징 벡터 및 제2 특징 벡터를 입력으로 하여, 결과 값을 생성할 수 있다. 제1 인코더 및 제2 인코더로부터 출력된 결과 값은 디코더로 입력되고, 디코더는 이미지를 설명하기 위한 이미지 캡션을 생성할 수 있다.

[0090] 도 3에서 디코더에서 출력된 이미지 캡션은 "A plate of sliced oranges with a fork (포크와 함께 있는 얇게 썬 오렌지 한 접시)"이다. 즉, 본 발명에 따른 이미지 캡션 생성 장치에서 출력된 이미지 캡션은 입력된 이미지의 특징인 'a fork(포크)', 'sliced oranges(얇게 썬 오렌지)', 및 'a plate (접시)'를 포함하며, 이들을 잘

설명하는 문장으로 구성된 것을 확인할 수 있다.

- [0091] 이미지 특징 추출기에 대해 자세히 설명한다. 이하에서 설명되는 이미지 특징 추출기는 도 2의 제1 특징 추출기 또는 제2 특징 추출기의 구체적인 실시예일 수 있다.
- [0092] 이미지 특징 추출기는 이미지로부터 특징 벡터를 추출할 수 있다. 즉, 이미지 특징 추출기는 이미지를 벡터로 변환시킬 수 있다.
- [0093] 트랜스포머 모델은 기계 번역을 위한 것으로서 텍스트로부터 텍스트를 생성할 수 있다. 다만, 이미지 캡션 생성 모델은 이미지로부터 텍스트를 생성하는 것을 요구한다. 따라서, 이미지 캡션 생성 모델의 입력 값이 이미지이므로 트랜스포머 모델처럼 단어 임베딩을 구하는 것이 아니라, 이미지를 벡터로 변환하는 과정이 필요하다. 다시 말하면, 트랜스포머 모델을 사용하여 이미지로부터 텍스트를 생성하기 위해서는, 이미지를 벡터로 변환하는 과정이 필요하다.
- [0094] 본 발명에 따른 이미지 캡션 생성 장치는 다중 관점을 갖는 이미지 정보를 얻기 위해 복수의 이미지 특징 추출기를 사용할 수 있다. 예를 들어, 본 발명에 따른 이미지 캡션 생성 장치는 서로 다른 종류의 이미지 특징 추출기를 사용할 수 있다. 본 발명에 따른 이미지 캡션 생성 장치는 서로 다른 종류의 이미지 특징 추출기를 복합적으로 사용하여, 이미지로부터 서로 다른 종류의 이미지 특징 벡터를 추출할 수 있다.
- [0095] 다시 도2를 참조하면, 제1 특징 추출기 및 제2 특징 추출기는 서로 다른 종류의 이미지 특징 추출기일 수 있다.
- [0096] 예를 들어, 본 발명에 따른 이미지 캡션 생성 장치는, R-CNN 계열의 사전 학습된 합성곱 신경망 모델 객체인, Faster R-CNN과 Mask R-CNN을 사용할 수 있다. 다시 말하면, 제1 특징 추출기는 Faster R-CNN을 사용할 수 있고, 제2 특징 추출기는 Mask R-CNN을 사용할 수 있다.
- [0097] Faster R-CNN은 컴퓨터 비전 분야에서 사용되는 객체 탐지 모델이다. Faster R-CNN은 이미지에서 탐지되는 객체의 위치를 박스(bounding box)로 표시를 하고, 해당 객체의 클래스를 분류할 수 있다.
- [0098] Mask R-CNN은 인스턴스 분할(Instance Segmentation) 모델이다. Mask R-CNN은 Faster R-CNN과는 달리 객체를 박스로 구분하는 것뿐만 아니라, 박스 안의 해당 객체의 영역을 더 구체적으로 찾고, 객체의 클래스가 다른 클래스와 중복이 되 어도 서로 다른 클래스로 분류할 수 있다.
- [0099] Faster R-CNN으로 추출한 특징 벡터는 데이터베이스(예를 들어, Imagenet)으로 사전 학습된 후에 시각적 개념(Visual genome)으로 훈련하여 얻은 데이터를 사용할 수 있다.
- [0100] 다시 말하면, Faster R-CNN은 이미지 데이터베이스 (예를 들어, 대규모 계층적 이미지 데이터베이스인 Imagenet)를 이용하여 사전 학습될 수 있다. 사전 학습된 뒤에, Faster R-CNN는 시각적 개념(Visual genome) (즉, 클라우드소싱된 고밀도 이미지 주석을 사용하여 언어와 시각을 연결하는 방법)에 의해 훈련될 수 있다. 제1 특징 추출기는 사전에 훈련된 Faster R-CNN를 이용하여, 이미지로부터 특징 벡터를 추출할 수 있다.
- [0101] Mask R-CNN은 Faster R-CNN의 방법과 유사하게 벡터를 추출하여 적응형 평균 풀링(Adaptive Average Pooling)을 통해 차원을 축소할 수 있다.
- [0102] 다시 말하면, Mask R-CNN은 이미지 데이터베이스(예를 들어, Imagenet)를 이용하여 사전 학습되고, 시각적 개념(Visual genome)에 의해 훈련될 수 있다. 제2 특징 추출기는 사전에 훈련된 Mask R-CNN를 이용하여, 이미지로부터 특징 벡터를 추출하고, 적응형 평균 풀링을 통해, 추출된 특징 벡터의 차원을 축소시킬 수 있다.
- [0103] 각 합성곱 신경망 모델에서 추출된 특징 벡터는 선형 변환을 통해 입력 값 임베딩 차원  $d_{model}$  로 축소시킨다. 즉, 제1 특징 추출기 및 제2 특징 추출기는 이미지로부터 추출된 제1 특징 벡터 및 제2 특징 벡터의 차원을 선형 변환을 통해 입력 값 임베딩 차원  $d_{model}$  으로 축소시킬 수 있다.
- [0104] 제1 특징 추출기 및 제2 특징 추출기에서 추출된 제1 특징 벡터( $V_a$ ) 및 제2 특징 벡터( $V_b$ )는 각각 아래와 같이 표현될 수 있다.

$$V_a = \{a_1, a_2, \dots, a_n\}, \quad V_b = \{b_1, b_2, \dots, b_m\}$$

$$a_i, b_j \in \mathbb{R}^{d_{model}}$$

[0105]

- [0106] 이미지 특징 벡터  $V$ 는 차원이  $d_{model}$  인  $a_i, b_j$  벡터로 구성될 수 있다. 이때,  $n$ 과  $m$ 은 이미지에서 선택된 객체 영역의 개수이다. 예를 들어,  $n$ 과  $m$ 은 10에서 100사이의 동적인 값일 수 있다.
- [0107] 이미지에서 추출한 특징 벡터  $V_a$ 와  $V_b$ 는 다른 합성곱 신경망 모델에서 얻어졌기 때문에 동일한 인코더를 사용하여 파라미터를 공유할 수 없다. 따라서, 도 2와 같이, 독립적인 두 개의 인코더를 사용하여 개별적인 복수의 결과값을 얻는 다중 관점 인코더(Multi-View Encoder)를 사용할 수 있다.
- [0108] 이하에서, 다중 관점 인코더에 대해 설명한다. 예를 들어, 이하에서 설명되는 다중 관점 인코더는 도 2의 제1 인코더 및 제2 인코더의 구체적인 실시예일 수 있다.
- [0109] 도 4은 본 발명의 몇몇 실시예에 따른 다중 관점 인코더를 설명하기 위한 블록도이다. 예를 들어, 도 4에 개시된 제1 특징 추출기, 제1 인코더, 제2 특징 추출기, 및 제2 인코더는 도 2에 개시된 것과 동일한 구성일 수 있다.
- [0110] 도 4에서, 제1 특징 추출기는 선형 변환기를 통해 제1 인코더와 연결될 수 있다. 또한, 제2 특징 추출기는 선형 변환기를 통해 제2 인코더와 연결될 수 있다.
- [0111] 도 4에서, 제1 특징 추출기 및 제2 특징 추출기는 각각 이미지를 입력으로 받아 제1 특징 벡터 및 제2 특징 벡터를 추출하여 제1 인코더 및 제2 인코더로 전달할 수 있다. 제1 특징 벡터( $V_a$ ) 및 제2 특징 벡터( $V_b$ )는 선형 변환기(Linear)를 통해 차원이  $d_{model}$  인 벡터로 변환되어, 제1 인코더 및 제2 인코더에 입력될 수 있다.
- [0112] 도 4의 다중 관점 인코더(즉, 제1 인코더 또는 제2 인코더)는 각각 멀티-헤드 주의 기제 기법(Multi-Head Attention)을 통해 이미지 정보 사이의 중요도를 기반으로 가중치를 업데이트할 수 있다. 그리고 포지션-와이즈 순방향 네트워크(Feed Forward)를 통과하여 결과 값(즉, 임베딩된 이미지 특징 값 (Embedded Image Feature s))을 구할 수 있다. 이때, 결과 값(즉, 임베딩된 이미지 특징 값)의 차원은  $d_{model}$ 로 이미지 특징 벡터  $V_a$  및  $V_b$ 의 차원과 동일할 수 있다.
- [0113] 이하에서, 멀티모달 트랜스포머 디코더에 대해 설명한다. 이하에서 설명되는 멀티모달 트랜스포머 디코더는 도 2의 디코더의 구체적인 실시예일 수 있다.
- [0114] 본 발명에 따른 이미지 캡션 장치는 다중 관점의 인코더에 의한 복수의 인코딩된 특징 값(즉, 임베딩된 이미지 특징 값)들을 사용할 수 있다. 따라서, 본 발명에 따른 이미지 캡션 장치의 디코더는 복수의 특징 값들을 모두 사용하기 위해, 수정된 주의 기제 기법을 사용할 수 있다.
- [0115] 트랜스포머 모델에서, 디코더 레이어는 두 종류의 멀티-헤드 주의 기제 기법(즉, Masked 멀티-헤드 주의 기제 기법 및 멀티모달(Multi Modal) 주의 기제 기법)을 포함할 수 있다.
- [0116] 첫 번째로, 디코더는 마스크를 사용하는 Masked 멀티-헤드 주의 기제 기법을 사용할 수 있다. 구체적으로, 멀티-헤드 주의 기제 기법에서, 디코더는 입력 값에 기초하여 자가-주의 기제(Self-Attention)를 통해 중간 값을 계산할 수 있다. 자가-주의 기제는 위의 수학적 식 (1)에서의 Query, Key 그리고 Value 가 모두 같은 값인 상황에서의 주의 기제 기법을 말한다. 디코더에서는 다음에 생성되는 단어를 예측하는 동작이 수행된다. 따라서, 현재 단계 이전의 데이터만을 계산하는 곳에 포함시켜, 데이터 사이의 인과 관계를 유지하기 위해 마스크를 사용하는 Masked 멀티-헤드 주의 기제 기법이 사용될 수 있다.
- [0117] 두 번째로, 디코더는 인코더의 결과 값과 이전 네트워크에서 나온 중간 값을 이용하여, 멀티-헤드 주의 기제 기법을 사용할 수 있다. 종래의 트랜스포머 모델은 단어와 단어 사이의 주의 기제 기법을 사용하는 반면에, 본 명세서에 따른 트랜스포머 모델은 이미지와 단어 사이의 멀티모달(Multi Modal) 주의 기제 기법을 사용할 수 있다.
- [0118] 이하에서, 스택 멀티모달 주의 기제 기법에 대해 설명한다. 이하에서 설명되는 멀티모달 주의 기제 기법은 도 2의 디코더에 의해 수행될 수 있다.
- [0119] 본 발명의 이미지 캡션 생성 장치에서, 인코더의 결과 값은 단일 값이 아닌 복수의 값이다. 따라서, 인코더의 결과 값과 디코더의 중간 값에 대한 멀티-헤드 주의 기제 기법의 네트워크가 복수 개 필요하다. 또한, 네트워크를 통해 얻은 값들을 적절히 결합하는 방법, 즉 복수 개의 멀티-헤드 주의 기제 네트워크를 구성하는 방법을 고안할 필요가 있다.
- [0120] 도 5는 본 발명의 몇몇 실시예에 따른, 인코더와 디코더의 결합 방법을 나타낸다. 구체적으로, 도 5는 이미지



특징 벡터  $V_a$ 와  $V_b$ 로부터 제1 및 제2 인코더를 통해 나온 결과 값  $M_a$ 와  $M_b$  (즉, 도 4의 임베딩된 이미지 특징 값(Embedded Image Features))을 결합하는 방법을 나타낸다. 보다 구체적으로, 도 5에 개시된 결합 방법은, 인코더로부터 출력된 이미지 특징 값을 결합하기 위한 방법으로서, 디코더에서 수행될 수 있다.

[0121] 도 5에서, "MHA with  $M_a$ ,  $M_b$ "는 각각의 임베딩된 이미지 특징 값( $M_a$ ,  $M_b$ )과 멀티모달 멀티-헤드 주의 기제를 수행하는 네트워크를 나타낸다. 예를 들어, "MHA with  $M_a$ "는 이미지 특징 벡터  $V_a$ 로부터의 결과 값인 임베딩된 이미지 특징 값  $M_a$ 를 기반으로 멀티모달 멀티-헤드 주의 기제를 수행하는 네트워크를 나타낸다. 예를 들어, "MHA with  $M_b$ "는 이미지 특징 벡터  $V_b$ 로부터의 결과 값인 임베딩된 이미지 특징 값  $M_b$ 를 기반으로 멀티모달 멀티-헤드 주의 기제를 수행하는 네트워크를 나타낸다. 이때, 이미지 특징 벡터  $V_a$  및  $V_b$ 는 각각 이미지로부터 도 2의 제1 특징 추출기 및 제2 특징 추출기에 의해 추출된 제1 특징 벡터  $V_a$  및 제2 특징 벡터  $V_b$ 를 나타낼 수 있다.

[0122] 도 5를 참조하면, 인코더와 디코더를 결합하는 방법으로 세 가지 방법이 제안된다.

[0123] 첫 번째로, 도 5(a)는 스택 주의 기제 기법 (Stacked Attention)을 나타낸다. 스택 주의 기제 기법은, 한 레이어 안에서 복수의 임베딩된 이미지 특징 값, 즉,  $M_a$ 와  $M_b$ ,에 대한 멀티모달 멀티-헤드 주의 기제를 연속적으로 쌓은 구조를 나타낸다. 이 경우, 한 레이어 안에서 서로 다른 이미지 특징 값( $M_a$ ,  $M_b$ )의 정보를 주의 기제 기법을 통해 순차적으로 결합하며 훈련을 진행할 수 있다.

[0124] 두 번째로, 도 5(b)는 스택 레이어 (Stack layer)를 나타낸다. 스택 레이어는 도 5(a)의 스택 주의 기제 기법과는 달리, 별개의 레이어에서 하나씩 멀티모달 멀티-헤드 주의 기제를 수행한다. 먼저 복수의 이미지 특징 값들 중 하나의 이미지 특징 값(즉,  $M_a$ )에 대해서 훈련을 진행하고, 그 이후에 다른 이미지 특징 값(즉,  $M_b$ )를 이용하여 모델을 학습시킬 수 있다. 이 경우, 각각의 레이어의 개수는 원래 수치의 절반( $N/2$ )에 해당될 수 있다.

[0125] 세 번째로, 도 5(c)는 병렬 주의 기제 기법 (Parallel Attention)을 나타낸다. 병렬 주의 기제 기법에서는, 복수의 이미지 특징 값들(즉,  $M_a$ 와  $M_b$ )에 대한 주의 기제를 동시에 수행한 뒤, 결과 값을 연결(Concatenate)하여 선형변환(Linear Transformation)을 통해 차원을 감소시켜서 레이어를 쌓는 구조이다.

[0126] 도 5(a)의 스택 주의 기제 기법은 이미지 특징 값 (즉,  $M_a$  및  $M_b$ )의 정보를 주의 기제 기법을 통해 결합하는 방법인 반면에, 도 5(c)의 병렬 주의 기제 기법은 이미지 특징 값 (즉,  $M_a$  및  $M_b$ )의 정보를 선형 변환을 통해 결합한다.

[0127] 본 발명의 이미지 캡션 생성 장치에 포함된 디코더는 위에서 설명된 세 가지의 결합 방법 즉, 스택 주의 기제 기법, 스택 레이어, 및 병렬 주의 기제 기법 중 적어도 하나를 사용할 수 있다. 다만, 이하에서는 설명의 편의상, 도 5(a)에 따른 스택 주의 기제 기법을 사용하는 것으로 가정한다.

[0128] 이하에서, 도 6 및 도 7을 참조하여, 스택 주의 기제 기법을 사용하는 디코더에 대해 설명한다. 이하에서 설명되는 디코더는 위의 도 2의 디코더와 동일한 구성일 수 있다.

[0129] 도 6은 스택 멀티모달 주의 기제 기법의 일 예를 나타내는 블록도이다.

[0130] 도 6를 참조하면, 인코더에 의해 출력된 임베딩된 이미지 특징 값들(즉,  $M_a$  및  $M_b$ )는 각각 멀티-헤드 주의 기제 기법으로 입력될 수 있다. 또한, 복수의 이미지 특징 값들 중 하나(예를 들어,  $M_a$ )에 기반한 멀티-헤드 주의 기제 기법의 출력 값은 복수의 이미지 특징 값들 중 다른 하나(예를 들어,  $M_b$ )을 위한 멀티-헤드 주의 기제 기법에 입력될 수 있다.

[0131] 스택 멀티모달 주의 기제 기법에서는 이하의 수학식 (5) 내지 (8)이 적용될 수 있다.

$$h_1^l = MHAttention(x^l, x^l, x^l) \quad (5)$$

$$h_2^l = MHAttention(h_1^l, M_a, M_a) \quad (6)$$

$$h_3^l = MHAttention(h_2^l, M_b, M_b) \quad (7)$$

$$x^l, h_{1,2,3}^l, M_{a,b} \in \mathbb{R}^{d_{model}} \quad (8)$$

[0132]

- [0133]  $x^1$ 은 디코더의 입력 값이다. 각 인자의 차원은 수학적 식 (8)과 같다.
- [0134] 디코더의 입력 값에 대해서 수학적 식 (5)를 계산하여  $h_1^1$ 을 구한다.
- [0135] 수학적 식 (6)에서 수학적 식 (5)의 결과 값( $h_1^1$ )과 이미지 특징 값(Ma)을 기반으로 멀티-헤드 주의 기제 기법을 수행하여  $h_2^1$ 을 구한다.
- [0136] 또한 수학적 식 (7)에서, 수학적 식 (6)의 결과 값( $h_2^1$ )과 이미지 특징 값(Mb)을 기반으로 멀티-헤드 주의 기제 기법을 수행하여  $h_3^1$ 을 구한다.
- [0137] 위의 수학적 식 (5) 내지 (8)과 같이, 본 발명은 다중 특징 추출기를 통해 얻은 서로 다른 이미지 특징 값들(즉, Ma 및 Mb)을 순차적으로 쌓아서 정보를 합치는 스택 멀티모달 주의 기제 기법(Stacked Multimodal Attention)을 사용할 수 있다. 예를 들어, 본 발명에 따른 이미지 캡션 생성기의 디코더는 위와 같은 스택 멀티모달 주의 기제 기법을 사용하여 훈련될 수 있다.
- [0138] 도 7은 본 발명의 몇몇 실시예에 따른 이미지 캡션 생성 장치의 디코더를 설명하기 위한 블록도이다. 도 7의 디코더는 도 2에 개시된 것과 동일한 것일 수 있다.
- [0139] 도 7에서 점선으로 표시된 스택 멀티모달 주의 기제 기법(Stacked Multimodal Attention)은 도 6의 스택 멀티모달 주의 기제 기법을 나타낸다.
- [0140] 도 7을 참조하면, 본 발명의 이미지 캡션 생성 장치의 디코더는 (i) 제1 Masked 멀티-헤드 주의 기제 기법 계층, (ii) 스택 멀티 모달 주의 기제 기법을 사용하는 계층, (iii) 제2 Masked 멀티-헤드 주의 기제 기법 계층, 및 (iv) 포지션-와이즈 순방향 네트워크 계층을 포함할 수 있다.
- [0141] 본 발명에서는 문장의 완성도를 높이기 위해 언어 모델이 강화된 자가 교열 트랜스포머(Self-Revising Transformer)이 제안된다. 즉, 본 발명에서 제안되는 자가 교열 트랜스포머는 도 4의 인코더 및 도 7의 디코더에 의해 구현될 수 있다. 구체적으로, 도 4의 인코더 및 도 7의 디코더는 도 6의 스택 멀티모달 주의 기제 기법으로 결합되어, 본 발명에 따른 자가 교열 트랜스포머를 형성할 수 있다.
- [0142] 기존의 트랜스포머 모델은 다음 단어를 예측할 때 입력 문장의 인코더 결과 값과 현재 단계의 단어들과 멀티-헤드 주의 기제 기법을 통해 구한다. 반면에, 이미지 캡션 생성 장치에서는, 인코더의 입력 값이 이미지 특징 벡터이기 때문에, 이미지의 정보를 통해서 다음 단어를 예측할 수 있다.
- [0143] 결과적으로 생성된 문장의 완성도가 떨어지거나 일반적인 문장이 되는 점을 보완해주기 위해서, 도 7과 같이, 디코더 레이어의 마지막 포지션-와이즈 순방향 네트워크를 계산하기 전에 추가적인 Masked 멀티-헤드 주의 기제 기법을 추가할 수 있다.
- [0144] 구체적으로, 본 발명의 이미지 캡션 생성 장치에 포함된 디코더에 Masked 주의 기제 기법을 추가함으로써, 이전 단계에서 얻은  $h_3^1$ 을 자가-주의 기제하고, 단어들 사이의 관계에 대한 정보를 한번 더 가중치에 업데이트할 수 있다.
- [0145] 따라서, 단어의 중복된 표현을 제재하고, 앞선 네트워크에서 생성된 문장을 재구성하여 문장의 연결성을 높일 수 있다.
- [0146] 이하에서, 본 발명의 몇몇 실시예에 따른 이미지 캡션 생성 장치의 훈련 방법에 대해 설명한다.
- [0147] 도 8은 본 명세서에서 사용되는 모델 손실 함수(Model Loss Function)의 흐름도이다.
- [0148] 도 8을 참조하면, 본 발명의 몇몇 실시예에 따른 이미지 캡션 생성 모델은 두 종류의 손실 함수(Loss Function)를 사용하여 학습될 수 있다. 구체적으로, 이미지 캡션 생성 모델은 라벨 스무딩(Label Smoothing)을 사용한 KLDiv(Kullback-Leibler divergence) 손실 함수에 의해 훈련된 뒤, 강화학습으로 모델을 최적화 시키는 SCST의 방법론을 적용하여 성능을 더욱 향상될 수 있다. 첫 번째로 라벨 스무딩이 적용된 KLDiv 손실함수는 아래의 수학적 식 (9) 및 (10)과 같이 계산될 수 있다.

$$L_{KLDiv} = -\sum_x p'(x) \log \frac{p'(x)}{q(x)} \quad (9)$$

$$p'(x) = (1 - \epsilon)p(x) + \frac{\epsilon}{X} \quad (10)$$

[0149]

[0150] 수학식 (10)는 라벨 스무딩을 적용하기 위한 것으로서, 정답을 정확하게 예측하지 않게 억압하는 방식으로 이미지 캡션 생성 모델을 정규화(Regularization)시킬 수 있다.

[0151] 수학식 (10)에서  $p(x)$ 는 정답 분포(Distribution)이고,  $q(x)$ 는 모델이 예측한 분포, 그리고  $\epsilon = 0.2$ 는 라벨 스무딩의 인자이다.

[0152] 두 번째로, 위의 손실 함수로 훈련된 모델은 강화학습을 이용한 SCST(Self-Critical Sequence Training)를 이용하여 최적화될 수 있다. SCST는 샘플링된 캡션과 추론 알고리즘으로 생성된 캡션의 사이의 CIDEr-D 스코어의 오차를 리워드로 주어 강화 학습하여 본 발명에 따른 이미지 캡션 생성 모델을 최적화시킬 수 있다.

$$L_{RL}(\theta) = -\mathbb{E}_{y^s \sim p(\theta)}[r(y^s)] \quad (11)$$

$$\nabla_{\theta} L_{RL}(\theta) \approx -(r(y^s) - r(\hat{y})) \nabla_{\theta} \log p_{\theta}(y^s) \quad (12)$$

[0153]

[0154] 수식 (11)은 SCST의 계산 식으로, 도 8과 같이 추론 알고리즘을 이용한 예측 캡션(Estimated Captions)과 샘플링된 캡션(Sampled Captions)의 CIDEr-D 스코어를 각각 구하여, 두 스코어의 오차를 구하는 리워드  $r$ 을 통해 본 발명에 따른 이미지 캡션 생성 모델을 훈련시킬 수 있다.

[0155] 그라디언트 값을 계산 하는 수식(12)에서  $y^s$ 는 샘플링된 예측 분포이다.

[0156] 또한,  $\hat{y}$  는 추론 알고리즘 탐욕 디코딩(Greedy Decoding)을 통해 얻은 예측 분포이다.

[0157] 이하에서, 본 발명의 몇몇 실시예에 따른 이미지 캡션 생성 방법 및 장치의 구체적인 실시예 및 유리한 효과에 대해 설명한다.

[0158] 먼저, 본 발명의 이미지 캡션 생성 모델에 사용되는 데이터셋에 대해 설명한다.

[0159] 본 발명의 몇몇 실시예에 따른 이미지 캡션 생성 장치를 위해, MSCOCO 데이터셋이 사용될 수 있다. MSCOCO 데이터셋은 평균적으로 한 이미지당 5개의 이미지를 설명하는 캡션을 포함한다. MSCOCO 데이터셋은 훈련 데이터셋으로 82,783장의 이미지, 평가 데이터셋으로 40,504장의 이미지를 포함한다.

[0160] 본 발명의 이미지 캡션 생성 장치를 위해, 총 123,287장의 이미지를 훈련, 평가, 그리고 성능 평가를 위한 테스트 데이터셋으로 세 종류로 분할하여 사용될 수 있다.

[0161] 또한, 본 발명의 이미지 캡션 생성 장치를 위해, "이미지 설명 생성을 위한 심층적인 시각적 의미 정렬 (Deep Visual-Semantic Alignments for Generating Image Descriptions, A. Karpathy and L. Fei-Fei)"에서 배포하는 위해 배포하는 분할된 데이터셋 정보가 사용될 수 있다. 상기 분할된 데이터셋 정보는 본 발명의 이미지 캡션 생성 장치의 성능 평가 비교를 위해 사용될 수 있다. 상기 분할된 데이터셋 정보는 113287장의 훈련 데이터셋, 각 5000장의 평가, 테스트 데이터셋의 분할 정보를 포함할 수 있다.

[0162] 단어 사전(Vocabulary)은 단어가 등장하는 최소 횟수가 5이상인 단어들로 구성하여 총 9487개 단어를 포함할 수 있다.

[0163] 본 발명의 이미지 캡션 생성 모델을 훈련시키기 위한 세부사항에 대해 설명한다.

[0164] 특징 추출기로는 Faster R-CNN과 Mask R-CNN 두 종류가 사용될 수 있다. 즉, 도 2의 제1 특징 추출기 및 제2 특징 추출기는 각각 Faster R-CNN과 Mask R-CNN일 수 있다.



- [0165] Faster R-CNN으로 얻은 특징 벡터는 2048차원이고 Mask R-CNN으로 얻은 특징 벡터는 4096차원일 수 있다. 이 두 가지의 특징 벡터들은 인코더에서  $d_{model} = 1024$ 의 차원으로 축소될 수 있다. Mask R-CNN 특징 벡터의 경우 4096에서  $d_{model}$ 로 직접적으로 축소하지 않고 한 단계를 더 거칠 수 있다.
- [0166] 본 발명의 이미지 캡션 생성 모델은 2개의 레이어를 사용할 수 있다. 본 발명의 이미지 캡션 생성 모델에서, 훈련 중 사용된 최적화 알고리즘으로 아담(Adam Optimizer)이 사용될 수 있다. 본 발명의 이미지 캡션 생성 모델에서 손실함수로는 라벨 스무딩을 정규화 방법으로 적용한 KLDiv 손실함수가 사용될 수 있으며, 이를 이용하여 25 예폭이 훈련될 수 있다.
- [0167] 본 발명의 이미지 캡션 생성 모델의 훈련이 끝난 후, CIDEr-D 스코어를 리워드로 사용하는 SCST 방법론을 이용하여 15 예폭을 추가로 훈련하여 모델이 최적화 될 수 있다.
- [0168] 본 발명의 이미지 캡션 생성 모델의 성능 평가를 위한 추론 알고리즘으로 빔 탐색 디코딩(Beam Search Decoding)이 사용될 수 있다. 이때, 빔의 크기는 2일 수 있다.
- [0169] 이하에서, 표 1 내지 표 4를 참조하여, 본 발명의 이미지 캡션 생성 모델의 유리한 효과에 대한 정량적 평가를 설명한다.
- [0170] 본 발명의 이미지 캡션 생성 장치는 BLEU, METEOR, ROUGE-L, CIDEr-D, 및 SPICE 스코어를 기초로 평가될 수 있다. 성능 비교에 사용된 모델들은 SCST로 최적화가 완료된 뒤, 스코어를 계산할 수 있다.

표 1

Metric	B@1	B@4	M	R	C	S
LSTM[12]	-	31.9	25.5	54.3	106.3	-
SCST[20]	-	34.2	26.7	55.7	114	-
LSTM-A[22]	78.6	35.5	27.3	56.8	118.3	20.8
Up-Down[9]	79.8	36.3	27.7	56.9	120.1	21.4
RFNet[23]	79.1	36.5	27.7	57.3	121.9	21.2
ICSA[19]	80.2	38	28.6	58.4	<b>128.6</b>	22.1
GCN-LSTM[24]	80.5	38.2	28.5	58.3	127.6	22
SGAE[25]	<b>80.8</b>	38.4	28.4	58.6	127.8	22.1
ORT[17]	80.5	38.6	28.7	<b>58.7</b>	128.3	22.6
FRC-MRC+SRT	79.8	38.3	<b>29</b>	58.5	128.2	<b>22.8</b>
MRC-FRC+SRT	80.4	<b>38.7</b>	<b>29</b>	<b>58.7</b>	128.4	<b>22.8</b>

- [0171]
- [0172] 표 1은 본원의 이미지 캡션 생성 모델의 성능을 종래의 모델의 성능과 비교한 결과를 나타낸다.
- [0173] 본 발명의 이미지 캡션 생성 모델 (즉 자가 교열 트랜스포머 모델)은 SRT로 표시된다. FRC (Faster R-CNN) 및 MRC (Mask R-CNN)는 각각 본 발명의 이미지 캡션 생성 모델에서 사용된 특징 추출기를 나타낸다.
- [0174] 모델명은 스택 멀티모달 주의 기제 기법안의 레이어 하단에 위치한 특징 추출기의 이름을 앞에 표기한다.
- [0175] 예를 들어, 도 6의 스택 멀티모달 주의 기제 기법을 참조하면, Mask R-CNN를 통해 추출된 특징 벡터에 대해 멀티-헤드 주의 기제 기법이 먼저 적용되는 경우, MRC-FRC로 표기된다. 이 경우, Mask R-CNN를 통해 추출된 특징 벡터에 대한 멀티-헤드 주의 기제 기법이 적용된 결과 값은, Faster R-CNN를 통해 추출된 특징 벡터와 함께 멀티-헤드 주의 기제 기법의 입력으로 입력될 수 있다. 즉, 도 6에서, Mask R-CNN이 하단에 위치한 경우 MRC-FRC로 표기된다.
- [0176] 예를 들어, 도 6의 스택 멀티모달 주의 기제 기법을 참조하면, Faster R-CNN를 통해 추출된 특징 벡터에 대해 멀티-헤드 주의 기제 기법이 먼저 적용되는 경우, FRC-MRC로 표기된다. 이 경우, Faster R-CNN를 통해 추출된 특징 벡터에 대한 멀티-헤드 주의 기제 기법이 적용된 결과 값은, Mask R-CNN를 통해 추출된 특징 벡터와 함께 멀티-헤드 주의 기제 기법의 입력으로 입력될 수 있다. 즉, 도 6에서, Faster R-CNN이 하단에 위치한 경우 FRC-MRC로 표기된다.
- [0177] 또한, 자가 교열 트랜스포머는 SRT로 표기된다. 즉, 표 1에서 "MRC-FRC+SRT" 또는 "FRC-MRC+SRT"은 본 발명의 이미지 캡션 생성 장치에 적용되는 기제 학습 모델의 순서를 계략적으로 나타내기 위한 것이다.
- [0178] 다시 표 1을 참조하면, 종래의 모델과 비교했을 때, MRC-FRC+SRT 모델의 경우 BLEU 1 스코어와 CIDEr-D를 제외

한 모든 스코어에서 높은 성능을 보이는 것을 볼 수 있다. 또한, FRC-MRC+SRT 모델의 경우에도 대체적으로 우수한 성능을 보이는 것을 볼 수 있다.

[0179] 결과적으로 위 표 1을 통해, 본 발명의 이미지 캡션 생성 모델의 성능이 종래의 모델 성능 보다 우수하다는 것을 정량적으로 확인할 수 있다.

표 2

Model	KL Divergence Loss						CIDEr-D Score Optimization					
Feat. Extractor	B@1	B@4	M	R	C	S	B@1	B@4	M	R	C	S
FRC	71.6	29.4	24.9	48.2	99.2	19.6	80.1	38.3	28.7	58.5	127.2	22.6
MRC	58.1	18.3	18.4	39.2	58.8	12.4	66.6	25.3	21.9	47.3	82.8	14.8
FRC-FRC	68.7	27.9	24.2	48.5	99.4	18.4	75.1	34.5	27.7	54.1	123.3	21.6
FRC-MRC	<b>76.7</b>	<b>35.1</b>	<b>27</b>	<b>55.8</b>	<b>110.7</b>	<b>20.5</b>	79.8	38.3	<b>29</b>	58.5	<b>128.2</b>	<b>22.8</b>
MRC-FRC	70.8	29.8	24.9	49.5	101	19.2	<b>80.4</b>	<b>38.7</b>	<b>29</b>	<b>58.7</b>	<b>128.4</b>	<b>22.8</b>

[0180]

표 2는 본 발명의 이미지 캡션 생성 모델에 포함된 이미지 특징 추출기에 따른 성능을 나타낸다.

[0181]

[0182] 표 2에서 사용된 이미지 캡션 생성 모델은 KLDiv 손실 함수를 사용하여 훈련된 뒤, CIDEr-D 스코어를 이용하여 최적화 되었다. 비교 모델들의 디코더는 동일하게 자가 교열 트랜스포머를 사용하였다. 표 2는 각각의 모델을 SCST를 이용하여 각각의 모델을 최적화 시킨 뒤의 실험 결과를 나타낸다.

[0183] 표 2는 하나의 특징 추출기(즉, FRC 또는 MRC)만을 사용하는 이미지 캡션 생성 모델의 성능을 나타낸다. 또한, 표 2는 두개의 특징 추출기(즉, FRC-FRC, FRC-MRC, 또는 MRC-FRC)사용하는 이미지 캡션 생성 모델의 성능을 나타낸다. 표 2에서 FRC-MRC 또는 MRC-FRC는, 표 1을 참조하여 설명된 것과 같이, 스택 멀티모달 주의 기제 기법에서 멀티-헤드 주의 기제 기법이 적용되는 순서를 나타낸다.

[0184] 표 2를 참조하면, 특징 추출기로 하나의 MRC만을 사용한 이미지 캡션 생성 모델의 성능이 좋지 않은 것을 볼 수 있다. 이는 Mask R-CNN 모델은 인스턴스 분할 연구 분야의 특화되어 있기 때문이다. 따라서, Mask R-CNN 모델은 이미지 캡션 생성 연구의 특징 추출기로 사용하기 적합하지 않다는 것을 알 수 있다.

[0185] 그러나 Faster R-CNN과 Mask R-CNN의 다중 특징 추출기를 사용한 이미지 캡션 생성 모델들(즉, FRC-MRC 또는 MRC-FRC)의 성능이 향상된 것을 볼 수 있다. 이는 Faster R-CNN과 Mask R-CNN의 다중 특징 추출기를 사용한 이미지 캡션 생성 모델에서, 더 많은 이미지 정보가 디코더에 전달되었기 때문이다.

[0186] FRC-MRC 모델은 FRC 모델과 비교했을 때 대체적으로 성능이 상승했고, MRC-FRC 모델은 모든 스코어에서 성능이 향상되었다는 것을 볼 수 있다. 즉, 다중 관점을 갖는 모델이 단일 관점을 갖는 모델에 비해 성능이 향상된 성능을 갖는 다는 것을 볼 수 있다.

[0187] 나아가, 서로 다른 특징 추출기를 사용한 이미지 캡션 생성 모델(즉, FRC-MRC 또는 MRC-FRC)은 같은 특징 추출기를 사용한 이미지 캡션 생성 모델(즉, FRC-FRC 모델) 보다 높은 성능을 갖는 다는 것을 볼 수 있다.

[0188] 즉, 본 발명의 이미지 캡션 생성 모델은 서로 다른 종류의 특징 추출기를 사용함으로써, 보다 높은 성능을 갖는 다는 것을 볼 수 있다.

표 3

Model		CIDEr-D Score Optimization					
Feat. Extractor	Dec	B@1	B@4	M	R	C	S
FRC-MRC	Base	80.1	38.2	28.6	58.4	126.6	22.4
	+ SRT	79.8	38.3	29	58.5	128.2	22.8
MRC-FRC	Base	80.1	38.2	28.8	58.5	127.8	22.7
	+ SRT	80.4	38.7	29	58.7	128.4	22.8

[0189]

[0190] 표 3은 디코더 모델의 성능을 비교한 결과를 나타낸다. 구체적으로 표 3은 특징 추출기로 각각 FRC-MRC와 MRC-FRC를 사용하는 이미지 캡션 생성 모델에 대해서, 자가 교열 트랜스포머(SRT)를 사용한 모델과 사용하지 않는 모델의 성능을 비교한 결과를 나타낸다.

- [0191] 표 3에서 "Base"는 디코더로 SRT가 아닌 트랜스포머를 사용한 이미지 캡션 생성 모델을 나타낸다. 또한, "+SRT"는 디코더로 SRT를 사용한 이미지 캡션 생성 모델을 나타낸다.
- [0192] 표 3에서, 굵은 글씨로 표현된 결과 값은, 각각의 모델 안에서 성능이 높은 점수를 표시한 것이다. 두 모델 모두에서 자가 교열 트랜스포머를 사용한 모델이 우수한 결과를 갖는다는 것을 볼 수 있다.
- [0193] FRC-MRC 모델의 경우 BLEU 1 스코어를 제외한 모든 스코어가 높은 결과 값을 갖고, MRC-FRC 모델은 모든 스코어에서 성능이 향상되었다는 것을 볼 수 있다. 특히, CIDEr-D 스코어가 두 모델에서 크게 증가하였다는 것을 볼 수 있다.
- [0194] 따라서, 본 발명의 이미지 캡션 생성 모델은 자가 교열 트랜스포머를 사용함으로써 더욱 개선된 성능을 갖는다는 것을 볼 수 있다.

#### 표 4

Model		CIDEr-D Score Optimization					
Feat. Extractor	Methods	B@1	B@4	M	R	C	S
FRC-MRC	Stacked attn	79.8	38.3	29	58.5	128.2	22.8
MRC-FRC		<b>80.4</b>	<b>38.7</b>	<b>29</b>	<b>58.7</b>	<b>128.4</b>	<b>22.8</b>
FRC-MRC	Stacked layer	79.8	38.1	28.6	58.2	126.2	22.4
MRC-FRC		74.7	34.1	27.7	54	122.7	21.5
FRC+MRC	Parallel attn	78.7	36.9	28.2	56.6	128	21.9

- [0195]
- [0196] 표 4는 인코더와 디코더의 결합 방법에 따른 이미지 캡션 생성 장치의 성능 결과를 나타낸다. 구체적으로, 표 4는 도 5를 참조하여 설명된 (1) 스택 주의 기제 기법 (Stacked Attention), (2) 스택 레이어, 및 (3) 병렬 주의 기제 기법 (Parallel Attention)을 적용한 이미지 캡션 생성 모델의 성능 결과를 나타낸다.
- [0197] 표 4를 참조하면, 도 5(a)에 따른 스택 주의 기제 기법을 적용한 이미지 캡션 생성 모델은 모든 스코어에서 다른 방법(즉, 도 5(b)에 따른 스택 레이어 및 도 5(c)에 따른 병렬 주의 기제 기법)을 사용한 이미지 캡션 생성 모델 보다 높은 성능을 갖는다는 것을 볼 수 있다.
- [0198] 스택 레이어를 사용한 이미지 캡션 생성 모델의 경우, FRC-MRC 모델을 사용한 모델이 MRC-FRC를 사용한 모델보다 높은 성능을 갖는다는 것을 볼 수 있다.
- [0199] 또한, 병렬 주의 기제 기법을 사용한 이미지 캡션 생성 모델의 종합적인 스코어들은, 스택 레이어를 사용한 이미지 캡션 생성 모델과 비슷한 성능을 갖는다는 것을 볼 수 있다.
- [0200] 예를 들어, 병렬 주의 기제 기법을 사용한 이미지 캡션 생성 모델은 높은 METEOR (M) 스코어를 갖는 반면에, 다른 스코어는 FRC-MRC 모델과 비슷하지만 낮은 점수를 갖는다는 것을 볼 수 있다.
- [0201] 따라서, 이미지 캡션 생성 모델에서, 스택 주의 기제 기법을 사용하여 인코더와 디코더를 결합하는 것이 다른 방법(즉, 스택 레이어 또는 병렬 주의 기제 기법)을 사용한 것 보다 더 효과적이라는 것을 볼 수 있다.
- [0202] 이하에서, 도 9 및 도 10를 참조하여, 본 발명의 이미지 캡션 생성 모델의 유리한 효과에 대한 정성적 평가를 설명한다.
- [0203] 도 9는 특징 추출기에 따른 이미지 캡션 생성 모델이 생성한 문장 예시를 나타낸다. 구체적으로, 도 9는 본 발명의 이미지 캡션 생성 모델 및 종래의 이미지 캡션 생성 모델에 의해 생성된 문장의 예시를 포함한다. 도 9(a)는 이미지 캡션 생성 모델이 올바른 캡션을 생성한 예시를 나타낸다. 도 9(b)는 이미지 캡션 생성 모델이 잘못된 캡션을 생성한 예시를 나타낸다. 도 9에서 사용된 이미지 캡션 생성 모델의 디코더는 모두 자가 교열 트랜스포머를 사용할 수 있다.
- [0204] 도 9를 참조하면, 다중 특징 추출기를 사용한 모델이 주어진 이미지의 특징들을 더욱 세부적으로 포착하여 서술한 이미지 캡션을 생성한다는 것을 볼 수 있다.
- [0205] 예를 들어, 도 9(a)의 첫 번째 이미지를 참조하면, 세 가지 모델(즉, FRC+SRT, FRC-MRC+SRT, 및 MRC-FRC+SRT)은 모두 이미지를 잘 설명한 캡션을 생성하였다는 것을 볼 수 있다. 구체적으로, FRC+SRT 모델은 새가 작다는 것과 벤치가 하얗다는 특징을 포함한 이미지 캡션을 생성한 것을 볼 수 있다. FRC-MRC+SRT 모델은 새가 작다는 것과 벤치가 나무로 만들어 졌다는 특징을 포함한 이미지 캡션을 생성한 것을 볼 수 있다. 또한, MRC-

FRC+SRT 모델은 노란 새가 나무로 만들어진 벤치에 있다는 특징을 포함한 것을 볼 수 있다.

- [0206] 그러나, 다른 예를 참조하면, FRC+SRT 모델은 이미지에 대한 캡션을 대체적으로 잘 생성하지만, 다중 특징 추출기를 쓴 모델이 생성한 캡션의 퀄리티가 더 높은 경우가 많다는 것을 볼 수 있다.
- [0207] 예를 들어, 도 9(a)에 포함된 두 번째 이미지 및 세 번째 이미지를 참조하면, FRC-MRC+SRT 모델 및 MRC-FRC+SRT 모델 세트기의 정확한 색상과 개수 및 양 이외의 동물인 개를 포함하는 이미지 캡션을 생성한 반면에, FRC+SRT 모델이 생성한 캡션은 이미지에 대해 비교적 일반적인 특성만을 포함하는 이미지 캡션을 생성한 것을 볼 수 있다. 즉, 이미지 캡션 생성 장치에서 다중 특징 추출기를 사용하는 경우, 이미지의 특징을 포착하는 성능이 향상된다는 것을 볼 수 있다.
- [0208] 도 9(b)는 이미지 캡션 생성 모델이 잘못된 캡션을 생성한 예시들을 나 타낸다.
- [0209] 도 9(b)를 참조하면, 첫 번째 이미지에서 FRC+SRT 모델과 MRC-FRC+SRT 모델은 이미지 안의 객체를 세부적으로 찾아낸 것을 볼 수 있다. 다만, 캡션의 주체가 자전거를 타는 사람인 이미지에서, 중요하지 않는 특징인 서핑 보드를 포착하여 캡션 생성에 영향을 끼쳐, 틀린 묘사를 한 것을 볼 수 있다. 또한, 이미지 캡션 생성 모델이 이미지의 객체를 다른 객체와 혼동하여 이미지 캡션을 잘못 생성한 것을 볼 수 있다.
- [0210] 두 번째 이미지에서는 이미지에 없는 바나나와 물고기라는 객체를 인식하여 잘못된 캡션을 생성한 것을 볼 수 있다. 또한, 세 번째 이미지는 인물이 여자임에도 불구하고 남성으로 인식하고, 우산을 야구 방망이로 인식한 것을 볼 수 있다.
- [0211] 도 10은 자가 교열 트랜스포머의 유무에 따른 이미지 캡션 생성 장치의 성능을 설명하기 위한 예시이다.
- [0212] 도 10에서 Base는 자가 교열 트랜스포머(SRT) 없이 MRC-FRC 특징 추출기를 포함한 이미지 캡션 생성 장치에 의해 생성된 이미지 캡션을 나타낸다. 또한, 도 8에서 +SRT는 자가 교열 트랜스포머 및 MRC-FRC 특징 추출기를 포함한 이미지 캡션 생성 장치에 의해 생성된 이미지 캡션을 나타낸다.
- [0213] 도 10을 참조하면, 자가 교열 트랜스포머를 포함한 이미지 캡션 생성 장치에서 생성된 이미지 캡션이 이미지를 더욱 자세히 설명하고 있다는 것을 볼 수 있다.
- [0214] 구체적으로, 도 10의 왼쪽에서 첫 번째 이미지는 남자가 파란색 곰인형을 안고 누워있는 특징을 포함한다.
- [0215] 자가 교열 트랜스포머(SRT)를 사용하지 않은 이미지 캡션 생성 모델이 생성한 이미지 캡션은 "곰인형을 안고 의자에 누워있는 남자 (a man laying in a chair holding a teddy bear)"인 것을 볼 수 있다.
- [0216] 반면에, 자가 교열 트랜스포머를 사용한 이미지 캡션 생성 모델이 생성한 이미지 캡션은 "곰인형을 안고 침대에 누워있는 남자("a man laying on a bed with a teddy bear)"인 것을 볼 수 있다.
- [0217] 즉, 자가 교열 트랜스포머(SRT)를 사용하지 않은 이미지 캡션 생성 모델은 "의자에 누워있다"와 같은 어색한 표현을 포함한 이미지 캡션을 생성하였으나, 본 발명에 따른 자가 교열 트랜스포머를 사용한 이미지 캡션 생성 모델은 "침대에 누워있다"와 같은 적절한 이미지 캡션을 생성하는 것을 볼 수 있다. 즉, 본 발명에 따른 자가 교열 트랜스포머를 사용한 이미지 캡션 생성 모델은 단어 사이의 연결성을 개선한 이미지 캡션을 생성할 수 있다는 것을 볼 수 있다.
- [0218] 도 10의 왼쪽에서 두 번째 이미지를 참조하면, 두 모델에서 각각 생성된 이미지 캡션은 어순이 다르지만 대체로 비슷한 단어들을 활용하고 있다는 것을 볼 수 있다.
- [0219] 도 10의 왼쪽에서 세 번째 이미지에서, 자가 교열 트랜스포머를 사용하지 않은 이미지 캡션 생성 모델(Base)은 이미지의 객체(즉, 우산(umbrella))를 잘못 포착한 것을 볼 수 있다.
- [0220] 반면에, 본 발명에 따른 자가 교열 트랜스포머를 사용한 이미지 캡션 생성 모델은 이미지에 포함된 객체와 유사한 객체(즉, 판잣집(shack))를 포착한 것을 볼 수 있다.
- [0221] 구체적으로, 데이터베이스에 포함된 세 번째 이미지의 이미지 캡션은 "정자" 또는 "오두막"이라는 표현을 포함한다. 한편, 자가 교열 트랜스포머를 사용하지 않은 이미지 캡션 생성 모델은 우산을 포함한 이미지 캡션을 생성한 반면에, 본 발명에 따른 제안하는 자가 교열 트랜스포머를 사용하지 않은 이미지 캡션 생성 모델은 "정자" 또는 "오두막"과 비슷한 단어인 판잣집을 포함한 이미지 캡션을 생성하였다.
- [0222] 도 10의 마지막 이미지를 참조하면, 자가 교열 트랜스포머를 사용하지 않은 이미지 캡션 생성 모델(Base)이 생성한 이미지 캡션은 "탁자 위에 앉아 있는 곰 인형들(a group of stuffed teddy bears sitting on a table)"인



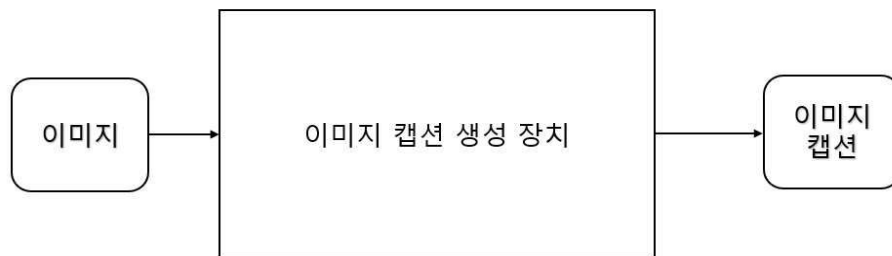
것을 볼 수 있다.

- [0223] 또한, 본 발명에 따른 자가 교열 트랜스포머를 사용한 이미지 캡션 생성 모델이 생성한 이미지 캡션은 "책과 함께 탁자 위에 앉아 있는 봉제 완구들(a group of stuffed animals sitting on a table with books)"인 것을 볼 수 있다.
- [0224] 즉, 본 발명에 따른 자가 교열 트랜스포머를 사용한 이미지 캡션 생성 모델이 이미지를 더 잘 설명하는 이미지 캡션을 생성한다는 것을 볼 수 있다.
- [0225] 다시 말하면, 본 발명에 따른 자가 교열 트랜스포머를 사용한 이미지 캡션 생성 모델이 문장의 연결성 및 이미지 설명의 정확도가 높은 이미지 캡션을 생성하는 것을 볼 수 있다.
- [0226] 이하, 도 11을 참조하여, 본 발명의 몇몇 실시예에 따른 이미지 캡션 생성 방법에 대해 설명한다.
- [0227] 도 11은 본 발명의 몇몇 실시예에 따른 이미지 캡션 생성 방법의 흐름도이다. 예를 들어, 이미지 캡션 생성 방법은 위의 도 1 및 도 2의 이미지 캡션 생성 장치에 의해 수행될 수 있다.
- [0228] 단계 S1101에서, 복수의 특징 추출기는 이미지로부터 복수의 특징 벡터를 추출할 수 있다.
- [0229] 예를 들어, 복수의 특징 추출기는 제1 특징 추출기 및 제2 특징 추출기를 포함할 수 있다. 제1 특징 추출기는 이미지로부터 제1 특징 벡터를 추출할 수 있다. 또한, 제2 특징 추출기는 이미지로부터 제2 특징 벡터를 추출할 수 있다.
- [0230] 제1 특징 추출기 및 제2 특징 추출기는 서로 상이한 지역 기반 합성곱 신경망 (R-CNN: Region based Convolution Neural Network) 모델을 사용할 수 있다.
- [0231] 예를 들어, 상기 제1 특징 추출기 및 상기 제2 특징 추출기 중 하나의 특징 추출기는 Faster R-CNN 모델을 사용할 수 있다. 또한, 상기 제1 특징 추출기 및 상기 제2 특징 추출기 중 다른 하나의 특징 추출기는 Mask R-CNN 모델을 사용할 수 있다.
- [0232] 단계 S1102에서, 복수의 인코더는 상기 복수의 특징 벡터 각각에 기초하여 상기 이미지에 대한 복수의 이미지 특징 값을 생성할 수 있다.
- [0233] 예를 들어, 복수의 인코더는 제1 인코더 및 제2 인코더를 포함할 수 있다.
- [0234] 상기 제1 인코더는 상기 제1 특징 추출기로부터 추출된 제1 특징 벡터에 기초하여 제1 이미지 특징 값을 생성할 수 있다. 또한, 상기 제2 인코더는 상기 제2 특징 추출기로부터 추출된 제2 특징 벡터에 기초하여 제2 이미지 특징 값을 생성할 수 있다.
- [0235] 상기 제1 인코더 및 상기 제2 인코더는 각각 상기 제1 특징 벡터 및 상기 제2 특징 벡터가 동일한 차원을 갖도록 선형 변환을 수행할 수 있다.
- [0236] 단계 S1103에서, 디코더는 상기 복수의 이미지 특징 값을 기초로 상기 이미지에 대한 이미지 캡션을 생성할 수 있다.
- [0237] 예를 들어, 상기 디코더는 상기 제1 이미지 특징 값 및 상기 제2 이미지 특징 값을 입력으로 받을 수 있다. 상기 디코더는 상기 제1 이미지 특징 값을 기초로 제1 멀티-헤드 주의 기제 기법을 적용할 수 있다. 상기 디코더는 (i) 상기 제1 멀티-헤드 주의 기제 기법의 결과 값과 (ii) 상기 제2 이미지 특징 값을 기초로 제2 멀티-헤드 주의 기제 기법을 적용할 수 있다.
- [0238] 또한, 상기 디코더는 상기 제2 멀티-헤드 주의 기제 기법의 결과 값에 마스크 멀티-헤드 주의 기제 기법을 적용할 수 있다.
- [0239] 본 발명의 몇몇 실시예에 따르면, 상기 디코더는 자가 교열 트랜스포머(self-revising transformer)를 사용할 수 있다.
- [0240] 예를 들어, 상기 자가 교열 트랜스포머는 제1 마스크 멀티-헤드 주의 기제 기법(Masked Multi-Head Attention), 스택 멀티-모달 주의 기제 기법 (Stacked Multimodal Attention), 및 제2 마스크 멀티-헤드 주의 기제 기법을 순차적으로 사용할 수 있다.
- [0241] 예를 들어, 상기 제1 마스크 멀티-헤드 주의 기제 기법 및 상기 제2 마스크 멀티-헤드 주의 기제 기법은 현재 단계 이전의 데이터만을 기초로 멀티-헤드 주의 기제 기법을 사용하는 것일 수 있다.

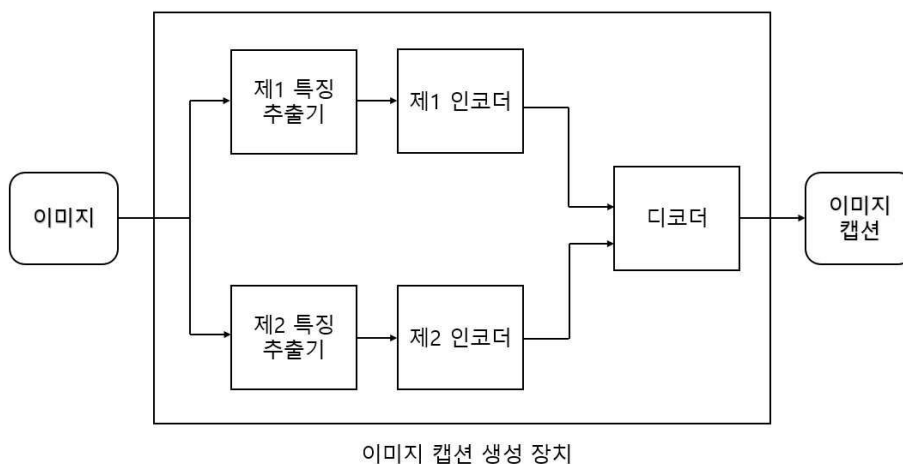
- [0242] 예를 들어, 상기 자가 교열 트랜스포머는 상기 제2 마스크 멀티-헤드 주의 기체 기법을 사용한 이후에, 포지션-와이즈 순방향 네트워크(Position-wise Feed-Forward Network)를 사용할 수 있다.
- [0243] 예를 들어, 상기 자가 교열 트랜스포머에서 사용되는 상기 스택 멀티모달 주의 기체 기법(Multi-Head Attention)을 연속적으로 적용하는 것을 의미할 수 있다.
- [0244] 또한, 상술한 바와 같은 이미지 캡션 생성 방법은 컴퓨터에서 실행될 수 있는 실행가능한 알고리즘을 포함하는 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.
- [0245] 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.
- [0246] 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.
- [0247] 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

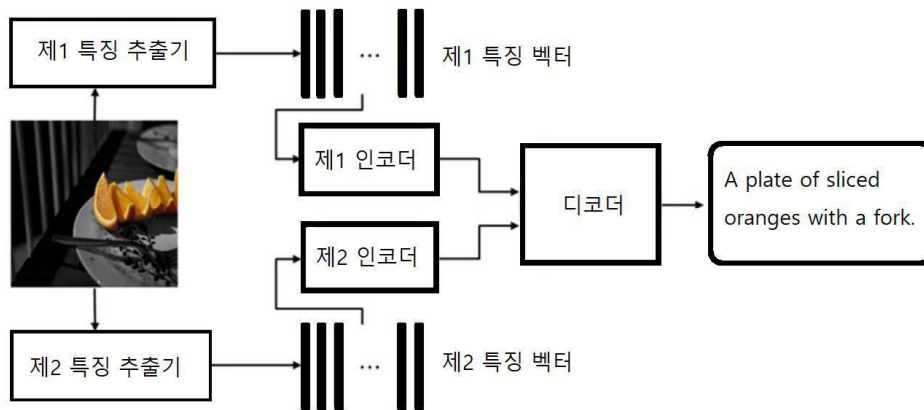
### 도면1



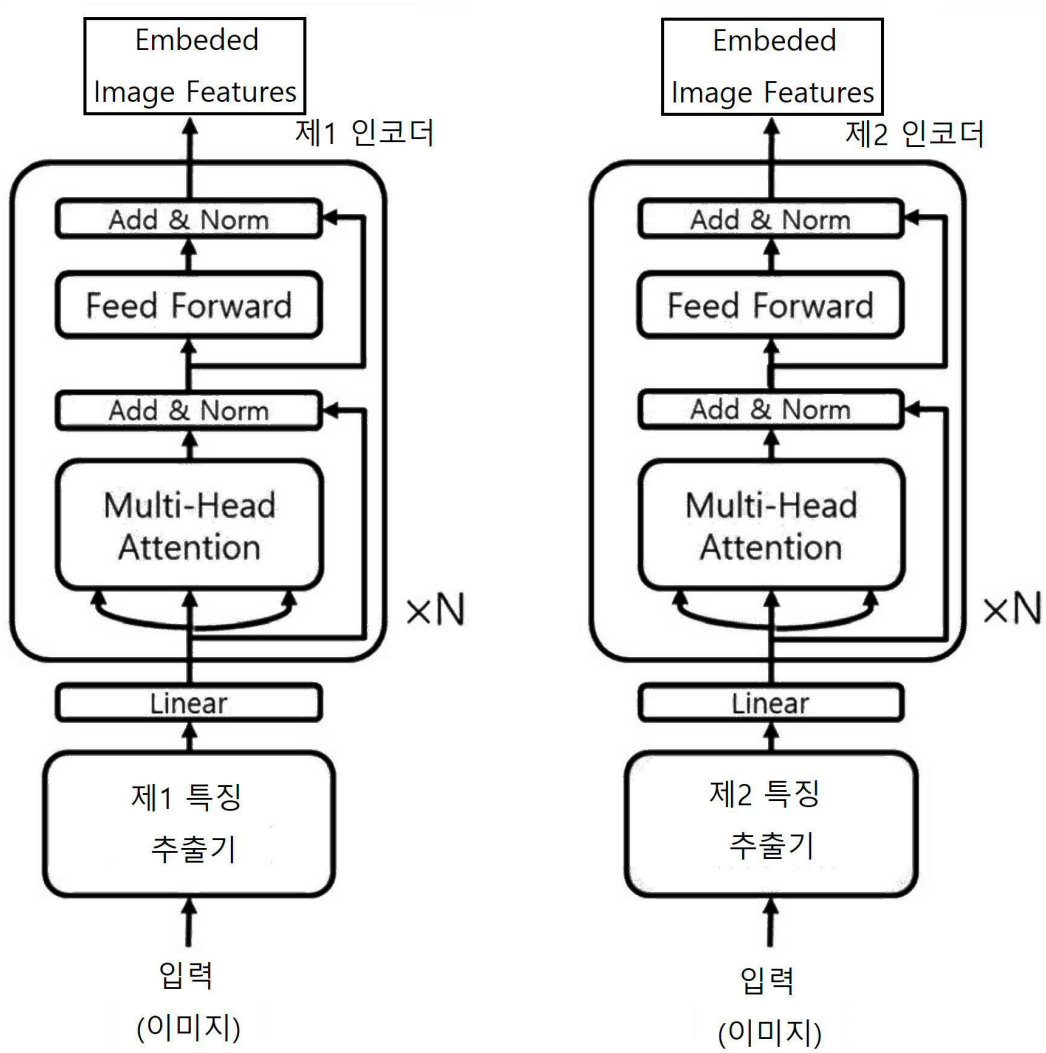
### 도면2



도면3

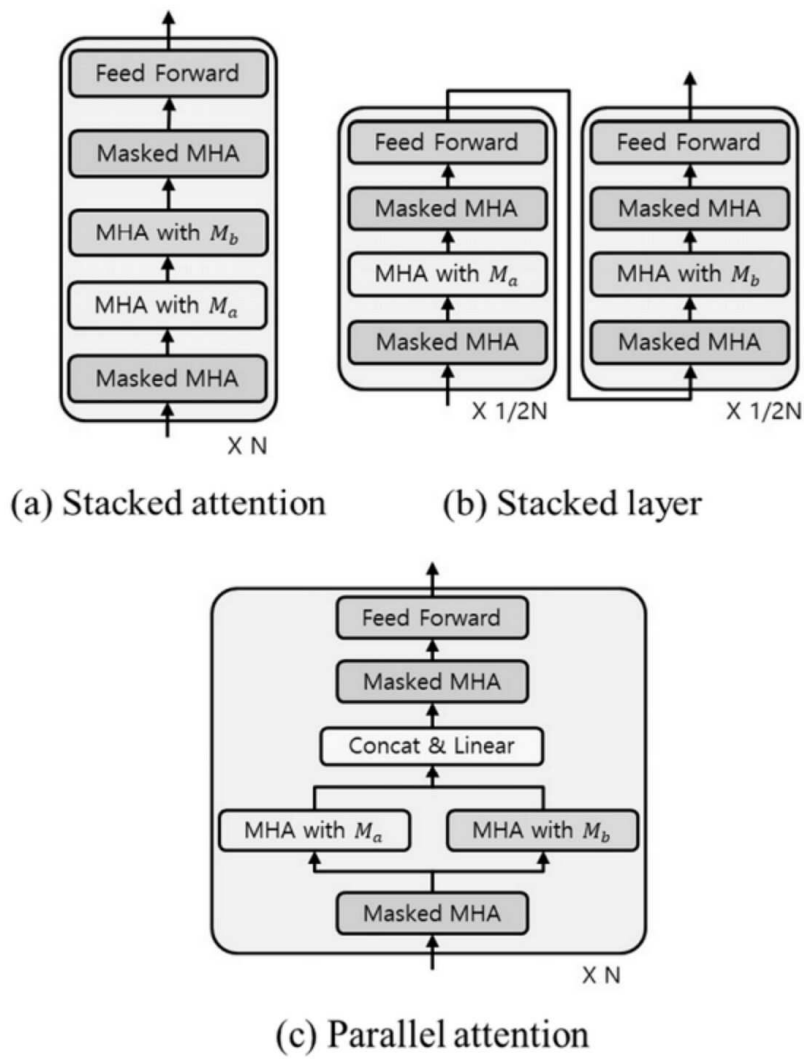


도면4

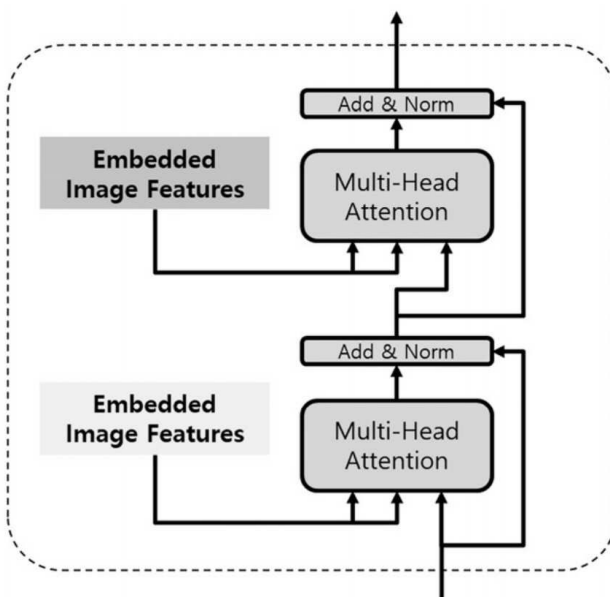




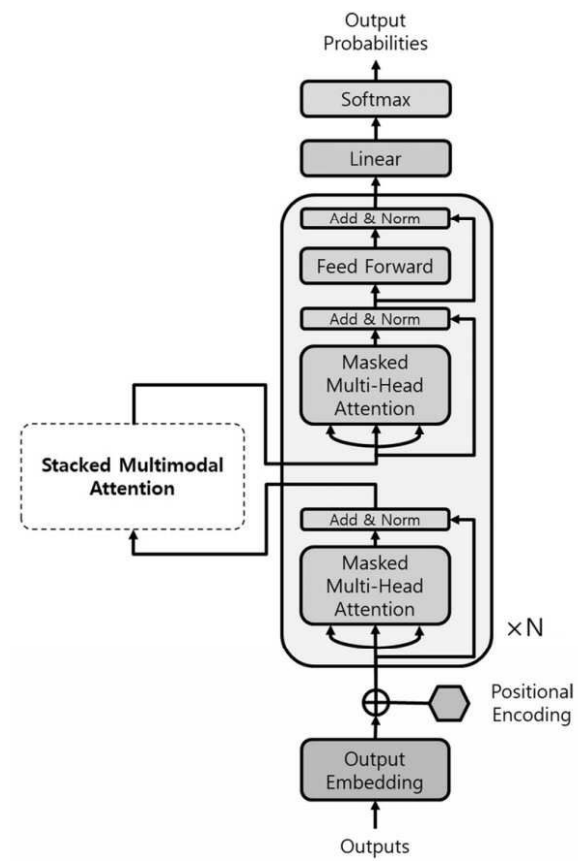
도면5



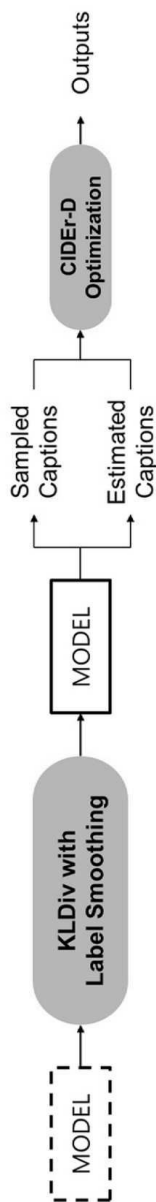
도면6









도면7



도면8







도면9

	<b>Ground-Truth</b>	A yellow bird sitting on a white piece of wood.		<b>Ground-Truth</b>	A couple of people are riding bicycles on the beach.
	<b>FRC+SRT</b>	a small bird sitting on top of a white bench		<b>FRC+SRT</b>	a young boy riding a bike with a surfboard on the beach
	<b>FRC-MRC+SRT</b>	a small bird sitting on a wooden bench		<b>FRC-MRC+SRT</b>	a group of people riding bikes on a beach
	<b>MRC-FRC+SRT</b>	a yellow bird perched on top of a wooden bench		<b>MRC-FRC+SRT</b>	two people riding bikes on the beach with a surfboard
	<b>Ground-Truth</b>	Two blue planes flying next to each other in a blue sky.		<b>Ground-Truth</b>	Two brown bears playing in a field together
	<b>FRC+SRT</b>	a group of fighter jets flying in the sky		<b>FRC+SRT</b>	two brown bears playing with a banana in the grass
	<b>FRC-MRC+SRT</b>	two blue and yellow fighter jets flying in the sky		<b>FRC-MRC+SRT</b>	a couple of bears sitting on top of a grass covered field
	<b>MRC-FRC+SRT</b>	two blue fighter jets flying in formation in the sky		<b>MRC-FRC+SRT</b>	two brown bears are playing with a fish
	<b>Ground-Truth</b>	A man that is standing in front of a group of sheep		<b>Ground-Truth</b>	Woman wearing red sash and hat holding umbrella
	<b>FRC+SRT</b>	a man standing next to a herd of sheep		<b>FRC+SRT</b>	a young boy wearing a baseball uniform holding a bat
	<b>FRC-MRC+SRT</b>	a man is holding a dog and sheep		<b>FRC-MRC+SRT</b>	a man in a baseball uniform holding a bat
	<b>MRC-FRC+SRT</b>	a man standing in a field with sheep and a dog		<b>MRC-FRC+SRT</b>	a young boy wearing a baseball uniform holding a bat

(a) Correct Examples

(b) Incorrect Examples

도면10

	<b>Base</b>	a man laying in a chair holding a teddy bear
	<b>+SRT</b>	a man laying on a bed with a teddy bear
	<b>Base</b>	a man standing on a tennis court with a racket
	<b>+SRT</b>	a man holding a tennis racket on a tennis court
	<b>Base</b>	a woman standing in front of an umbrella
	<b>+SRT</b>	a woman sitting on top of a small shack
	<b>Base</b>	a group of stuffed teddy bears sitting on a table
	<b>+SRT</b>	a group of stuffed animals sitting on a table with books

도면11

