(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0100916
(43) 공개일자 2025년07월04일

(51) 국제특허분류(Int. Cl.)
*G06N 3/0895* (2023.01) *G06F 16/532* (2019.01)
*G06F 16/538* (2019.01) *G06F 40/268* (2020.01)
*G06F 40/279* (2020.01) *G06N 3/0455* (2023.01)
*G06N 3/096* (2023.01)
(52) CPC특허분류
*G06N 3/0895* (2023.01)
*G06F 16/532* (2019.01)
(21) 출원번호 10-2023-0192027
(22) 출원일자 2023년12월27일
심사청구일자 2023년12월27일

(71) 출원인
**고려대학교 산학협력단**
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
**이정섭**
서울특별시 동대문구 무학로 184, 401호
**임희석**
경기도 수원시 영통구 영통로 460, 304동 1704호
(74) 대리인
**김등용**

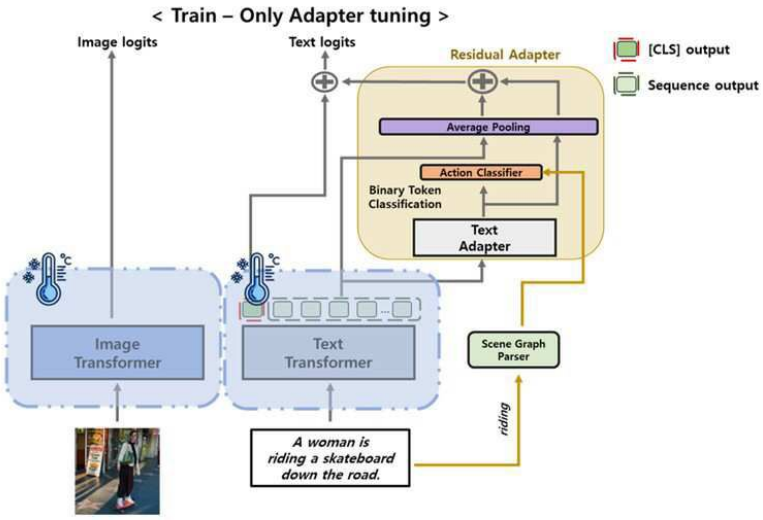
전체 청구항 수 : 총 5 항

(54) 발명의 명칭 **CLIP-어댑터 기반 이미지-텍스트 검색 장치 및 방법**

## (57) 요 약

CLIP(Contrastive Language-Image Pre-training)-어댑터 기반 이미지-텍스트 검색 장치 및 방법이 개시된다. 이미지-텍스트 검색 모델 생성 방법은 프로세서(Processor)를 포함하는 컴퓨팅 장치에 의해 수행되고, 이미지에 대한 이미지 로짓(logits)과 텍스트에 대한 텍스트 로짓이 생성되는 단계, 상기 텍스트에 포함된 동사가 예측되는 단계, 인공 신경망(Artificial Neural Network) 기반 모델인 텍스트 어댑터를 이용하여 상기 텍스트의 각 토큰에 대한 동사를 예측하는 단계, 상기 텍스트에 대한 최종 텍스트 로짓을 생성하는 단계, 및 상기 이미지 로짓과 상기 최종 텍스트 로짓을 통해 대조학습(Contrastive learning)을 수행하여, 상기 최종 텍스트 로짓과 상기 이미지 로짓이 동일한 사상에 매핑되도록 상기 텍스트 어댑터의 가중치를 업데이트하는 단계를 포함한다.

*대 표 도* - 도1



(52) CPC특허분류
*G06F 16/538* (2019.01)
*G06F 40/268* (2020.01)
*G06F 40/279* (2020.01)
*G06N 3/0455* (2023.01)
*G06N 3/096* (2023.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711193711
과제번호 2018-0-01405-006
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성
연구과제명 Human-inspired 복합지능 원천기술 연구 및 인력양성
기 여 율 1/2
과제수행기관명 고려대학교산학협력단
연구기간 2023.01.01 ~ 2023.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1345365402
과제번호 2021R1A6A1A03045425
부처명 교육부
과제관리(전문)기관명 한국연구재단
연구사업명 이공학학술연구기반구축
연구과제명 Human-inspired AI 연구소
기 여 율 1/2
과제수행기관명 고려대학교
연구기간 2023.03.01 ~ 2024.02.29

# 명 세 서

## 청구범위

### 청구항 1

프로세서(Processor)를 포함하는 컴퓨팅 장치에 의해 수행되는, CLIP(Contrastive Language-Image Pre-training)-어댑터 기반 이미지-텍스트 검색 모델 생성 방법에 있어서,

이미지에 대한 이미지 로짓(logits)과 텍스트에 대한 텍스트 로짓이 생성되는 단계;

상기 텍스트에 포함된 동사가 예측되는 단계;

인공 신경망(Artificial Neural Network) 기반 모델인 텍스트 어댑터를 이용하여 상기 텍스트의 각 토큰에 대한 동사를 예측하는 단계;

상기 텍스트에 대한 최종 텍스트 로짓을 생성하는 단계; 및

상기 이미지 로짓과 상기 최종 텍스트 로짓을 통해 대조학습(Contrastive learning)을 수행하여, 상기 최종 텍스트 로짓과 상기 이미지 로짓이 동일한 사상에 매핑되도록 상기 텍스트 어댑터의 가중치를 업데이트하는 단계를 포함하는 이미지-텍스트 검색 모델 생성 방법.

### 청구항 2

제1항에 있어서,

상기 생성되는 단계는, CLIP 모델의 이미지 인코더를 이용하여 상기 이미지에 대한 상기 이미지 로짓을 생성하고, 상기 CLIP 모델의 텍스트 인코더를 이용하여 상기 텍스트에 대한 상기 텍스트 로짓을 생성하는,

이미지-텍스트 검색 모델 생성 방법.

### 청구항 3

제2항에 있어서,

상기 동사가 예측되는 단계는 장면 그래프 파서(Scene Graph Parser)를 이용하여 상기 텍스트 내의 동사를 예측하는,

이미지-텍스트 검색 모델 생성 방법.

### 청구항 4

제3항에 있어서,

상기 최종 텍스트 로짓을 생성하는 단계는,

상기 텍스트 인코더의 출력과 상기 텍스트 어댑터의 출력을 평균 풀링(Average pooling)한 후, 평균 풀인한 값과 상기 텍스트 인코더의 출력을 잔차 연결하여 상기 최종 텍스트 로짓을 생성하는,

이미지-텍스트 검색 모델 생성 방법.

### 청구항 5

제4항에 있어서,

상기 텍스트 로짓이 생성되는 단계 내지 상기 업데이트하는 단계는 반복 수행되는,

이미지-텍스트 검색 모델 생성 방법.

## 발명의 설명

### 기 술 분 야

[0001] 본 발명은 여러 텍스트 후보군에서 주어진 이미지에 대해 가장 유사한 k개의 텍스트를 검색하거나, 여러 이미지 후보군에서 주어진 텍스트에 대해 가장 유사한 k개의 이미지를 검색하는 이미지-텍스트 검색 장치 및 방법에 관한 것이다.

## 배 경 기 술

[0002] 이미지-텍스트 검색(Image-Text Retrieval)은 여러 텍스트 중에서 입력된 이미지를 가장 잘 설명하는 텍스트 k개를 검색하거나, 여러 이미지 중에서 입력된 텍스트를 가장 잘 설명하는 이미지 k개를 검색하는 시스템이다.

[0003] 이와 관련하여 OpenAI는 학습 없이 검색할 수 있는 제로샷 CLIP(Contrastive Language-Image Pre-training) 사전학습 모델을 제안했지만, CLIP은 동작이 포함된 이미지나 동작이 포함된 텍스트를 인식하는 능력이 떨어져 잘못된 이미지 혹은 텍스트를 검색하는 문제가 발생한다. 동작이 포함된 이미지 혹은 텍스트 검색을 효율적으로 하기 위해, 본 발명은 동작 인식 어댑터를 학습하는 방법을 추가하여 이미지-텍스트 검색 성능을 높이고자 한다.

[0004] 기존의 이미지-텍스트 검색은 기학습된 CLIP 모델을 사용한다. 특정 이미지-텍스트 검색을 더 잘 수행하기 위해서 CLIP 모델을 추가학습하는 방법이 주로 적용되는데, 이는 대량의 데이터로 CLIP 모델을 학습하므로 높은 학습 비용이 요구된다. 또한, 기존 CLIP의 사전학습 데이터가 동작을 설명하는 동사 보다 객체를 설명하는 명사에 취중되어 있기 때문에 행동에 초점을 맞춘 동작이 포함된 이미지 검색 성능이 비교적 떨어진다.

[0005] 이러한 문제를 해결하기 위해, 본 발명은 텍스트 인코더에 어댑터를 사용하며, 기존의 CLIP 모델은 파라미터의 그래디언트를 추적하지 않고 어댑터만 학습하는 방법을 차용하고, 동작에 관한 정보를 잘 인코딩 하기 위해 장면 그래프 파서(Scene Graph Parser)를 사용하여 동사에 더 집중할 수 있도록 추가적인 태스크를 수행하여 낮은 학습비용으로 높은 이미지-텍스트 검색을 수행하는 것을 목적으로 한다.

## 선행기술문헌

### 특허문헌

[0006] (특허문헌 0001) 대한민국 공개특허 제2020-0002332호 (2020.01.08. 공개)

(특허문헌 0002) 대한민국 공개특허 제2022-0147550호 (2022.11.03. 공개)

(특허문헌 0003) 대한민국 등록특허 제2151408호 (2020.10.26. 공고)

### 비특허문헌

[0007] (비특허문헌 0001) Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao, "CLIP-Adapter: Better Vision-Language Models wiht Feature Adapters," zrXiv:2110.04544v1. October 9, 2021.

## 발명의 내용

### 해결하려는 과제

[0008] 본 발명이 이루고자 하는 기술적인 과제는 CLIP-어댑터 기반 이미지-텍스트 검색 장치 및 방법을 제공하는 것이다.

### 과제의 해결 수단

[0009] 본 발명의 일 실시예에 따른 CLIP(Contrastive Language-Image Pre-training)-어댑터 기반 이미지-텍스트 검색 모델 생성 방법은 프로세서(Processor)를 포함하는 컴퓨팅 장치에 의해 수행되고, 이미지에 대한 이미지 로짓(logits)과 텍스트에 대한 텍스트 로짓이 생성되는 단계, 상기 텍스트에 포함된 동사가 예측되는 단계, 인공 신경망(Artificial Neural Network) 기반 모델인 텍스트 어댑터를 이용하여 상기 텍스트의 각 토큰에 대한 동사를 예측하는 단계, 상기 텍스트에 대한 최종 텍스트 로짓을 생성하는 단계, 및 상기 이미지 로짓과 상기 최종 텍스

트 로짓을 통해 대조학습(Contrastive learning)을 수행하여, 상기 최종 텍스트 로짓과 상기 이미지 로짓이 동일한 사상에 매핑되도록 상기 텍스트 어댑터의 가중치를 업데이트하는 단계를 포함한다.

## 발명의 효과

[0010] 본 발명의 실시예들에 의할 경우, 기존의 CLIP 모델을 사용하고, 소량의 이미지-텍스트 데이터를 통해 어댑터만을 학습하므로 80% 이상 학습 비용(GPU 사용량 및 학습 시간)을 절약할 수 있다.

[0011] 또한, 본 발명은 학습 시에만 동사를 예측하는 장면 그래프 파서(Scene Graph Parser)를 사용하고, 추론 시에 사용되지 않으므로 효율적인 이미지-텍스트 검색기를 구성할 수 있다.

[0012] 또한, 본 발명을 적용할 경우, 이미지-텍스트 검색 데이터셋인 Flickr 30k의 Recall@10에서 기존의 CLIP 모델 대비 45% 이상의 성능 개선을 보임을 확인하였다.

## 도면의 간단한 설명

[0013] 본 발명의 상세한 설명에서 인용되는 도면을 보다 충분히 이해하기 위하여 각 도면의 상세한 설명이 제공된다.

도 1은 본 발명의 일 실시예에 따른 이미지-텍스트 검색 방법을 설명하기 위한 개념도이다.

도 2는 본 발명의 일 실시예에 따른 이미지-텍스트 검색 방법을 설명하기 위한 흐름도이다.

## 발명을 실시하기 위한 구체적인 내용

[0014] 본 명세서에 개시되어 있는 본 발명의 개념에 따른 실시예들에 대해서 특정한 구조적 또는 기능적 설명들은 단지 본 발명의 개념에 따른 실시예들을 설명하기 위한 목적으로 예시된 것으로서, 본 발명의 개념에 따른 실시예들은 다양한 형태들로 실시될 수 있으며 본 명세서에 설명된 실시예들에 한정되지 않는다.

[0015] 본 발명의 개념에 따른 실시예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시예들을 도면에 예시하고 본 명세서에서 상세하게 설명하고자 한다. 그러나, 이는 본 발명의 개념에 따른 실시예들을 특정한 개시 형태들에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물, 또는 대체물을 포함한다.

[0016] 제1 또는 제2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로만, 예컨대 본 발명의 개념에 따른 권리 범위로부터 벗어나지 않은 채, 제1 구성 요소는 제2 구성 요소로 명명될 수 있고 유사하게 제2 구성 요소는 제1 구성 요소로도 명명될 수 있다.

[0017] 어떤 구성 요소가 다른 구성 요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성 요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성 요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성 요소가 다른 구성 요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는 중간에 다른 구성 요소가 존재하지 않는 것으로 이해되어야 할 것이다. 구성 요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0018] 본 명세서에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로서, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 본 명세서에 기재된 특징, 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0019] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0020] 이하, 본 명세서에 첨부된 도면들을 참조하여 본 발명의 실시예들을 상세히 설명한다. 그러나, 특허출원의 범위

가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일한 참조 부호는 동일한 부재를 나타낸다.

[0021] 도 1은 본 발명의 일 실시예에 따른 이미지-텍스트 검색 방법을 설명하기 위한 개념도이고, 도 2는 본 발명의 일 실시예에 따른 이미지-텍스트 검색 방법을 설명하기 위한 흐름도이다.

[0022] 도 1과 도 2를 참조하면, 이미지-텍스트 검색 모델 생성 방법 등으로 명명될 수도 있는 이미지-텍스트 검색 방법은 적어도 프로세서(Processor) 및/또는 메모리(Memory)를 포함하는 컴퓨팅 장치에 의해 수행될 수 있다. 다시 말하면, 이미지-텍스트 검색 방법을 이루는 단계들 중 적어도 일부는 컴퓨팅 장치에 포함되는 프로세서의 동작으로 이해될 수도 있다. 컴퓨팅 장치는 PC(Personal Computer), 서버(server), 랩탑 컴퓨터, 태블릿 PC 등을 포함할 수 있다. 실시예에 따라, 컴퓨팅 장치는 물리적으로 분리된 복수의 장치로 구현되거나, 클라우드 환경에서 구현되는 것도 가능하다.

[0023] 이미지 로짓(Image logits)과 텍스트 로짓(Text logits)이 생성된다(S110). 본 발명은 기존의 사전학습 모델인 CLIP(Contrastive Language-Image Pre-training) 모델을 이용한다. CLIP 모델은 텍스트 인코더와 이미지 인코더로 구성된다. CLIP의 텍스트 인코더는 학습 데이터에 포함되는 텍스트를 입력받고 이에 대응하는 텍스트 로짓을 생성할 수 있다. CLIP의 이미지 인코더는 학습 데이터에 포함되는 이미지를 입력받고 이에 대응하는 이미지 로짓을 생성할 수 있다.

[0024] 텍스트에 포함된 동사가 예측된다(S120). 동사의 예측은 장면 그래프 파서(Scene Graph Parser)를 이용하여 수행될 수 있다. 예시적으로, 주어진 텍스트가 "A woman is riding a skateboard down the road."일 때, 장면 그래프 파서는 텍스트 내의 동사로 "riding"을 예측할 수 있다.

[0025] 텍스트 어댑터를 이용하여 주어진 텍스트의 각 토큰에 대한 동사 예측을 이진 분류로 수행한다(S130). 텍스트 어댑터는 임의의 인공 신경망(Artificial Neural Network, ANN) 모델일 수 있다. 실시예에 따라, 텍스트 어댑터는 CLIP의 텍스트 인코더에 의해 생성된 텍스트 로짓을 수신하고, 수신된 텍스트 로짓의 각 토큰에 대한 동사 예측을 이진 분류로 수행한다. 예컨대, 토큰이 [A, Woman, is, riding, a, skateboard, down, the, road, .]일 때, 텍스트 어댑터는 각 토큰이 동사에 해당하는지 여부를 예측할 수 있다. 이때, 정답 레이블은 장면 그래프 파서의 예측 결과를 이용하여, [0, 0, 0, 1, 0, 0, 0, 0, 0, 0]일 수 있다.

[0026] 텍스트에 대한 최종 텍스트 로짓이 생성된다(S140). 최종 텍스트 로짓은 텍스트 어댑터의 출력과 텍스트 인코더의 [CLS] 값에 대한 평균 풀링(Average pooling)을 수행하고, 텍스트 인코더의 텍스트 로짓과 평균 풀링의 결과를 잔차연결함으로써, 최종 텍스트 로짓이 생성될 수 있다. 즉, 텍스트 어댑터의 구조는 잔차-스타일 특징 혼합(Residual-style feature blending)을 사용할 수 있다.

[0027] 텍스트 어댑터의 가중치가 업데이트된다(S150). 생성된 두 로짓, 즉 이미지 로짓과 최종 텍스트 로짓을 통해 대조학습(Contrastive learning)을 수행하여, 최종 텍스트 로짓과 이미지 로짓이 동일한 사상에 비슷한 벡터 값으로 매핑되도록 학습을 수행할 수 있다. 이때, CLIP의 구성들(즉, 이미지 인코더와 텍스트 인코더)에 대한 가중치 업데이트는 수행하지 않고(freeze), 텍스트 어댑터만을 대상으로 업데이트가 수행될 수 있다. 텍스트 어댑터를 학습할 때, 텍스트 어댑터의 손실은 이미지 분류 손실 및 동사 예측 이진분류 손실을 a:b(a와 b는 임의의 실수로써, 예시적인 값은 각각 8과 2임)로 할 수 있다. 즉, 텍스트 어댑터의 최종 손실은 0.8×Cross Entropy Loss + 0.2×Binary Token Classification Loss일 수 있다.

[0028] 상술한 S110 단계 내지 S150 단계를 반복수행함으로써, 이미지-텍스트 검색 모델이 생성될 수 있다.

[0029] 생성된 이미지-텍스트 검색 모델을 이용하여, 이미지-텍스트 검색을 수행하기 위해서는, 적어도 하나의 이미지 대한 이미지 로짓과 적어도 하나의 텍스트에 대한 최종 텍스트 로짓이 생성될 수 있다.

[0030] 이때, 주어진 이미지에 대해 가장 유사한 k(k는 임의의 자연수) 개의 텍스트를 검색하는 경우, 하나의 이미지와 k 보다 많은 복수의 텍스트가 주어질 수 있다. 결국, 이미지에 대한 로짓과 가장 유사도가 높은 k 개의 최종 텍스트 로짓에 대응하는 텍스트를 유사한 텍스트로 선택할 수 있다. 유사도는 코사인 유사도와 같은 임의의 유사도 산출 기법이 이용될 수 있다.

[0031] 반대로, 주어진 텍스트에 대해 가장 유사한 k 개의 이미지를 검색하는 경우, 하나의 텍스트와 k 보다 많은 복수의 이미지가 주어질 수 있다. 결국, 텍스트에 대한 최종 텍스트 로짓과 가장 유사도가 높은 k 개의 이미지 로짓에 대응하는 이미지를 유사한 이미지로 선택할 수 있다.

[0032] 또한, 학습과정에서 이용된, 장면 그래프 파서 등은 추론 과정에서는 활용되지 않는다.

[0033] 이상에서 설명된 장치는 하드웨어 구성 요소, 소프트웨어 구성 요소, 및/또는 하드웨어 구성 요소 및 소프트웨어 구성 요소의 집합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성 요소는, 예를 들어, 프로세서, 콘트롤러, ALU(Arithmetic Logic Unit), 디지털 신호 프로세서(Digital Signal Processor), 마이크로컴퓨터, FPA(Field Programmable array), PLU(Programmable Logic Unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(Operation System, OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술 분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(Processing Element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(Parallel Processor)와 같은, 다른 처리 구성(Processing Configuration)도 가능하다.

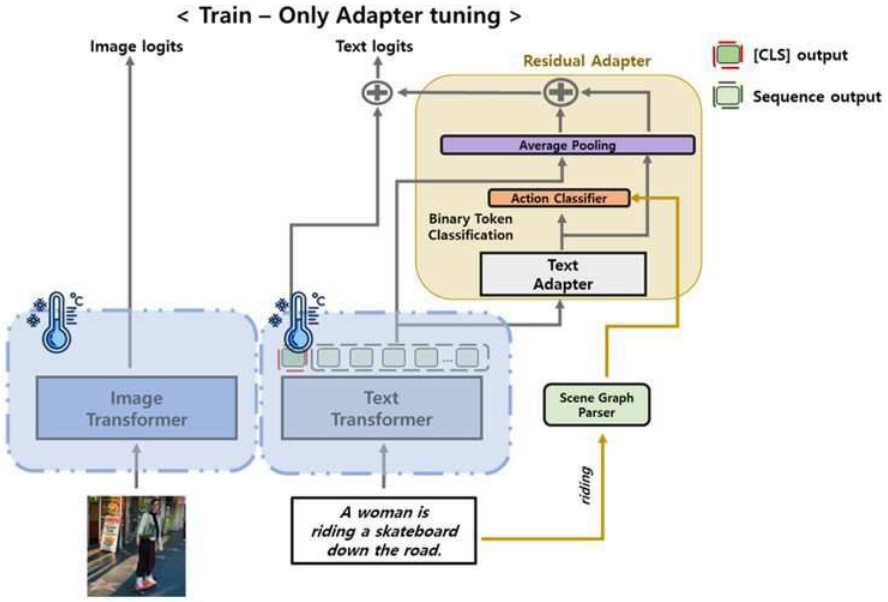
[0034] 소프트웨어는 컴퓨터 프로그램(Computer Program), 코드(Code), 명령(Instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(Collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성 요소(Component), 물리적 장치, 가상 장치(Virtual Equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(Signal Wave)에 영구적으로, 또는 일시적으로 구체화(Embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0035] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(Magnetic Media), CD-ROM, DVD와 같은 광기록 매체(Optical Media), 플롭티컬 디스크(Floptical Disk)와 같은 자기-광 매체(Magneto-optical Media), 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

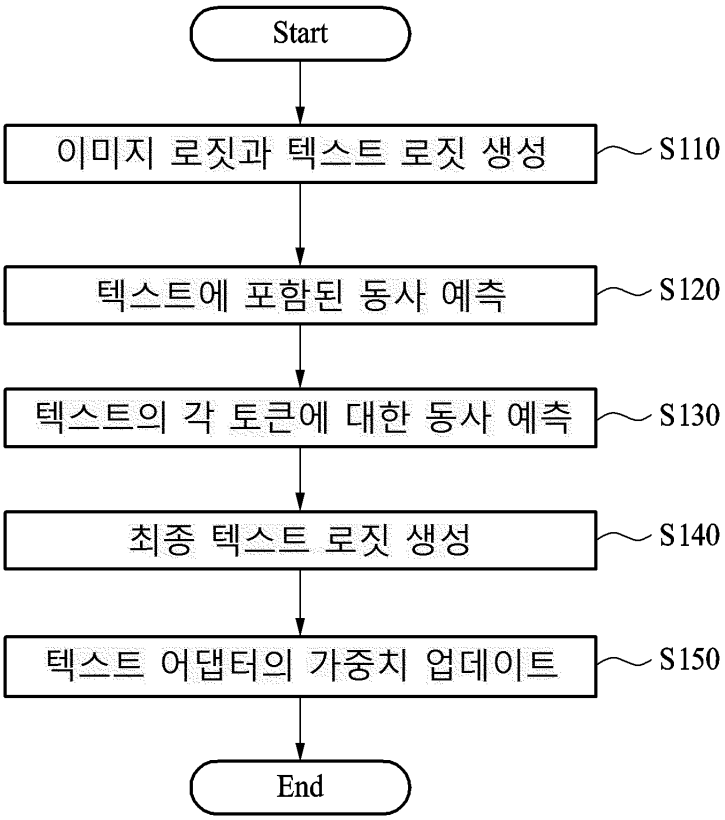
[0036] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성 요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성 요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

## 도면

### 도면1


< Train – Only Adapter tuning >
Image logits
Text logits
Residual Adapter
[CLS] output
Sequence output
Average Pooling
Action Classifier
Binary Token Classification
Text Adapter
Image Transformer
Text Transformer
Scene Graph Parser
riding
A woman is riding a skateboard down the road.

### 도면2


Start
이미지 로짓과 텍스트 로짓 생성
S110
텍스트에 포함된 동사 예측
S120
텍스트의 각 토큰에 대한 동사 예측
S130
최종 텍스트 로짓 생성
S140
텍스트 어댑터의 가중치 업데이트
S150
End