(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0084725
(43) 공개일자 2022년06월21일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01) *G06N 3/04* (2006.01)
(52) CPC특허분류
G06N 3/08 (2013.01)
G06N 3/0454 (2013.01)
(21) 출원번호 10-2020-0174480
(22) 출원일자 2020년12월14일
심사청구일자 없음

(71) 출원인
주식회사 엔씨소프트
서울특별시 강남구 테헤란로 509 (삼성동)
이화여자대학교 산학협력단
서울특별시 서대문구 이화여대길 52 (대현동, 이화여자대학교)
(72) 발명자
강제원
서울특별시 서대문구 이화여대길 52 아산공학관 524호
김나영
서울특별시 서대문구 이화여대길 52 아산공학관 524호
하성종
경기도 성남시 분당구 대왕판교로 644번길 12
(74) 대리인
함영욱

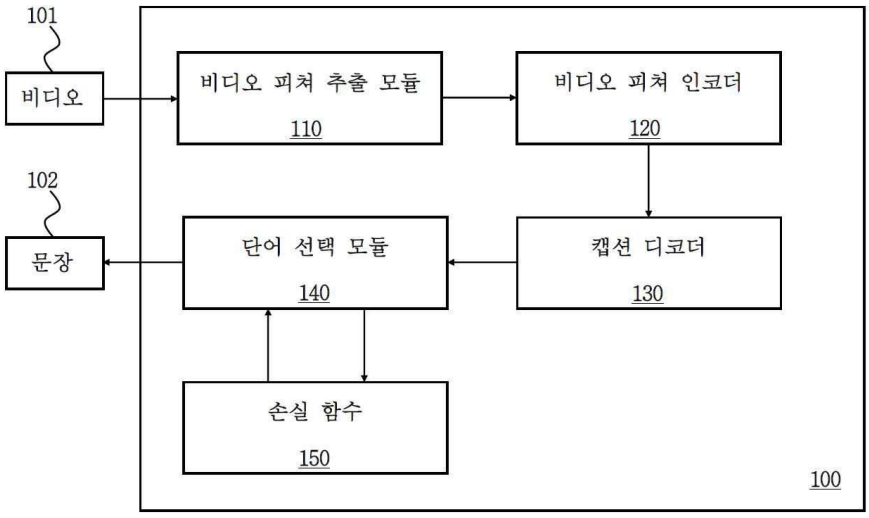
전체 청구항 수 : 총 18 항

(54) 발명의 명칭 **비디오 캡셔닝 학습 장치, 비디오 캡셔닝 학습 방법, 비디오 캡셔닝 장치 및 비디오 캡셔닝 방법**

(57) 요 약

본 발명의 일실시예에 따르면, 비디오 캡셔닝(captioning) 학습 장치에 있어서, 적어도 하나의 프로세서를 포함하고, 상기 적어도 하나의 프로세서는, 비디오 피쳐 및 템퍼럴 어텐션(temporal attention)을 기초로 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합을 생성하고, 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합 및 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피쳐를 기초로 단어를 생성한다.

대 표 도 - 도1



명 세 서

청구범위

청구항 1

비디오 캡셔닝(captioning) 학습 장치에 있어서,

적어도 하나의 프로세서를 포함하고,

상기 적어도 하나의 프로세서는,

비디오 피쳐 및 템퍼럴 어텐션(temporal attention)을 기초로 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합을 생성하고,

상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합 및 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피쳐를 기초로 단어를 생성하는 비디오 캡셔닝(captioning) 학습 장치.

청구항 2

제1항에 있어서,

상기 적어도 하나의 프로세서는,

제1 딥 뉴럴 네트워크(Deep Neural Networks)를 이용하여 상기 비디오 피쳐를 인코딩하여 제1 아웃풋(output)을 생성하고,

상기 제1 아웃풋(output)에 미리 설정된 함수를 적용하여 템퍼럴 어텐션(temporal attention)을 생성하고,

상기 비디오 피쳐와 상기 템퍼럴 어텐션(temporal attention)을 결합하여 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합을 생성하는 비디오 캡셔닝(captioning) 학습 장치.

청구항 3

제2항에 있어서,

상기 적어도 하나의 프로세서는,

상기 제1 아웃풋(output)을 입력으로 하는 학습이 완료된 제1 풀리 커넥티드 레이어(Fully Connected Layer)와 매칭되는 제1 파라미터의 적어도 하나의 파라미터 값을 결정하기 위해 제1 풀리 커넥티드 레이어(Fully Connected Layer)를 학습시키는 비디오 캡셔닝(captioning) 학습 장치.

청구항 4

제1항에 있어서,

상기 적어도 하나의 프로세서는,

상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피쳐에 미리 설정된 함수를 적용하여 어텐션(attention)을 생성하고,

상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합과 상기 어텐션(attention)을 결합하여 워드 어텐디드 피쳐(word attended features)를 생성하고,

상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)를 이용하여 상기 워드 어텐디드 피쳐(word attended features) 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피쳐를 디코딩하여 제2 아웃풋(output)을 생성하고,

상기 제2 아웃풋(output)을 기초로 미리 설정된 단어집에서 단어를 선택하는 비디오 캡셔닝(captioning) 학습

장치.

청구항 5

제4항에 있어서,

상기 적어도 하나의 프로세서는,

상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처를 결합(concatenate)하고,

상기 결합(concatenate)한 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처를 입력으로 하는 학습이 완료된 제2 풀리 커넥티드 레이어(Fully Connected Layer)와 매칭되는 제2 파라미터의 적어도 하나의 파라미터 값을 결정하기 위해 제2 풀리 커넥티드 레이어(Fully Connected Layer)를 학습시키는 비디오 캡셔닝(captioning) 학습 장치.

청구항 6

제4항에 있어서,

상기 적어도 하나의 프로세서는,

상기 단어집과 매칭되도록 상기 제2 아웃풋(output)의 차원을 변화시키기 위한 제3 파라미터를 결정하는 비디오 캡셔닝(captioning) 학습 장치.

청구항 7

비디오 피처 및 템퍼럴 어텐션(temporal attention)을 기초로 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합을 생성하는 동작; 및

상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합 및 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처를 기초로 단어를 생성하는 동작

을 포함하는 비디오 캡셔닝(captioning) 학습 방법.

청구항 8

제7항에 있어서,

상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합을 생성하는 동작은,

제1 딥 뉴럴 네트워크(Deep Neural Networks)를 이용하여 상기 비디오 피처를 인코딩하여 제1 아웃풋(output)을 생성하는 동작;

상기 제1 아웃풋(output)에 미리 설정된 함수를 적용하여 템퍼럴 어텐션(temporal attention)을 생성하는 동작; 및

상기 비디오 피처와 상기 템퍼럴 어텐션(temporal attention)을 결합하여 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합을 생성하는 동작

을 포함하는 비디오 캡셔닝(captioning) 학습 방법.

청구항 9

제8항에 있어서,

상기 제1 아웃풋(output)에 미리 설정된 함수를 적용하여 템퍼럴 어텐션(temporal attention)을 생성하는 동작은,

상기 제1 아웃풋(output)을 입력으로 하는 학습이 완료된 제1 풀리 커넥티드 레이어(Fully Connected Layer)와 매칭되는 제1 파라미터의 적어도 하나의 파라미터 값을 결정하기 위해 제1 풀리 커넥티드 레이어(Fully Connected Layer)를 학습시키는 동작

을 더 포함하는 비디오 캡셔닝(captioning) 학습 방법.

청구항 10

제7항에 있어서,

상기 단어를 생성하는 동작은,

상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처에 미리 설정된 함수를 적용하여 어텐션(attention)을 생성하는 동작;

상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합과 상기 어텐션(attention)을 결합하여 워드 어텐디드 피처(word attended features)를 생성하는 동작;

상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)를 이용하여 상기 워드 어텐디드 피처(word attended features) 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처를 디코딩하여 제2 아웃풋(output)을 생성하는 동작; 및

상기 제2 아웃풋(output)을 기초로 미리 설정된 단어집에서 단어를 선택하는 동작

을 포함하는 비디오 캡셔닝(captioning) 학습 방법.

청구항 11

제10항에 있어서,

상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처에 미리 설정된 함수를 적용하여 어텐션(attention)을 생성하는 동작은,

상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처를 결합(concatenate)하는 동작; 및

상기 결합(concatenate)한 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처를 입력으로 하는 학습이 완료된 제2 풀리 커넥티드 레이어(Fully Connected Layer)와 매칭되는 제2 파라미터의 적어도 하나의 파라미터 값을 결정하기 위해 제2 풀리 커넥티드 레이어(Fully Connected Layer)를 학습시키는 동작

을 더 포함하는 비디오 캡셔닝(captioning) 학습 방법.

청구항 12

제10항에 있어서,

상기 제2 아웃풋(output)을 기초로 미리 설정된 단어집에서 단어를 선택하는 동작은,

상기 단어집과 매칭되도록 상기 제2 아웃풋(output)의 차원을 변화시키기 위한 제3 파라미터를 결정하는 동작

을 더 포함하는 비디오 캡셔닝(captioning) 학습 방법.

청구항 13

비디오 캡셔닝(captioning) 장치에 있어서,

적어도 하나의 프로세서를 포함하고,

상기 적어도 하나의 프로세서는,

비디오 피처 및 템퍼럴 어텐션(temporal attention)을 기초로 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합을 생성하고,

상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합 및 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처를 기초로 단어를 생성하는 비디오 캡셔닝(captioning) 장치.

청구항 14

제13항에 있어서,

상기 적어도 하나의 프로세서는,

제1 딥 뉴럴 네트워크(Deep Neural Networks)를 이용하여 상기 비디오 피처를 인코딩하여 제1 아웃풋(output)을 생성하고,

상기 제1 아웃풋(output)에 미리 설정된 함수를 적용하여 템퍼럴 어텐션(temporal attention)을 생성하고,

상기 비디오 피처와 상기 템퍼럴 어텐션(temporal attention)을 결합하여 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합을 생성하는 비디오 캡셔닝(captioning) 장치.

청구항 15

제13항에 있어서,

상기 적어도 하나의 프로세서는,

상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처에 미리 설정된 함수를 적용하여 어텐션(attention)을 생성하고,

상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합과 상기 어텐션(attention)을 결합하여 워드 어텐디드 피처(word attended features)를 생성하고,

상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)를 이용하여 상기 워드 어텐디드 피처(word attended features) 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처를 디코딩하여 제2 아웃풋(output)을 생성하고,

상기 제2 아웃풋(output)을 기초로 미리 설정된 단어집에서 단어를 선택하는 비디오 캡셔닝(captioning) 장치.

청구항 16

비디오 피처 및 템퍼럴 어텐션(temporal attention)을 기초로 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합을 생성하는 동작; 및

상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합 및 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처를 기초로 단어를 생성하는 동작

을 포함하는 비디오 캡셔닝(captioning) 방법.

청구항 17

제16항에 있어서,

상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합을 생성하는 동작은,

제1 딥 뉴럴 네트워크(Deep Neural Networks)를 이용하여 상기 비디오 피처를 인코딩하여 제1 아웃풋(output)을 생성하는 동작;

상기 제1 아웃풋(output)에 미리 설정된 함수를 적용하여 템퍼럴 어텐션(temporal attention)을 생성하는 동작; 및

상기 비디오 피처와 상기 템퍼럴 어텐션(temporal attention)을 결합하여 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합을 생성하는 동작

을 포함하는 비디오 캡셔닝(captioning) 방법.

청구항 18

제16항에 있어서,

상기 단어를 생성하는 동작은,

상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피쳐에 미리 설정된 함수를 적용하여 어텐션(attention)을 생성하는 동작;

상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합과 상기 어텐션(attention)을 결합하여 워드 어텐디드 피쳐(word attended features)를 생성하는 동작;

상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)를 이용하여 상기 워드 어텐디드 피쳐(word attended features) 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피쳐를 디코딩하여 제2 아웃풋(output)을 생성하는 동작; 및

상기 제2 아웃풋(output)을 기초로 미리 설정된 단어집에서 단어를 선택하는 동작

을 포함하는 비디오 캡셔닝(captioning) 방법.

발명의 설명

기 술 분 야

[0001] 아래의 실시예들은 비디오 캡셔닝 학습 장치, 비디오 캡셔닝 학습 방법, 비디오 캡셔닝 장치 및 비디오 캡셔닝 방법에 관한 것이다.

배 경 기 술

[0002] 머신 러닝(machine learning)은 인공 지능의 한 분야로, 패턴인식과 컴퓨터 학습 이론의 연구로부터 진화한 분야이며, 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야를 말한다. 머신 러닝의 핵심은 표현(representation)과 일반화(generalization)에 있다. 표현이란 데이터의 평가이며, 일반화란 아직 알 수 없는 데이터에 대한 처리이다. 이는 전산 학습 이론 분야이기도 하다.

[0003] 딥 러닝(deep learning)은 여러 비선형 변환기법의 조합을 통해 높은 수준의 추상화를 시도하는 기계학습(machine learning) 알고리즘의 집합으로 정의되며, 큰 틀에서 사람의 사고방식을 컴퓨터에게 가르치는 기계학습의 한 분야라고 이야기할 수 있다.

[0004] 비디오 캡셔닝(Video Captioning)이란 영상이 주어질 때 네트워크가 영상의 내용에 대해서 문장을 자동으로 생성하는 기술을 말한다.

발명의 내용

해결하려는 과제

[0005] 본 발명의 실시예에 따르면, 비디오 피쳐와 템퍼럴 어텐션(temporal attention)을 결합하여 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합을 생성할 수 있는 비디오 캡셔닝 학습 장치, 비디오 캡셔닝 학습 방법, 비디오 캡셔닝 장치 및 비디오 캡셔닝 방법을 제공할 수 있다.

[0006] 또한, 본 발명의 다른 실시예에 따르면, 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합 및 이전 히든 스테이트(hidden state)가 출력한 피쳐를 기초로 단어를 생성할 수 있는 비디오 캡셔닝 학습 장치, 비디오 캡셔닝 학습 방법, 비디오 캡셔닝 장치 및 비디오 캡셔닝 방법을 제공할 수 있다.

[0007] 또한, 본 발명의 또 다른 실시예에 따르면, 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합과 어텐션을 결합하여 워드 어텐디드 피쳐를 생성할 수 있는 비디오 캡셔닝 학습 장치, 비디오 캡셔닝 학습 방법, 비디오 캡셔닝 장치 및 비디오 캡셔닝 방법을 제공할 수 있다.

과제의 해결 수단

[0008] 본 발명의 일실시예에 따르면, 비디오 캡셔닝(captioning) 학습 장치에 있어서, 적어도 하나의 프로세서를 포함하고, 상기 적어도 하나의 프로세서는, 비디오 피쳐 및 템퍼럴 어텐션(temporal attention)을 기초로 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합을 생성하고, 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합 및 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피쳐를 기초로 단어를 생성한다.

[0009] 또한, 상기 적어도 하나의 프로세서는, 제1 딥 뉴럴 네트워크(Deep Neural Networks)를 이용하여 상기 비디오 피쳐를 인코딩하여 제1 아웃풋(output)을 생성하고, 상기 제1 아웃풋(output)에 미리 설정된 함수를 적용하여 템퍼럴 어텐션(temporal attention)을 생성하고, 상기 비디오 피쳐와 상기 템퍼럴 어텐션(temporal attention)을 결합하여 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합을 생성할 수 있다.

[0010] 또한, 상기 적어도 하나의 프로세서는, 상기 제1 아웃풋(output)을 입력으로 하는 학습이 완료된 제1 풀리 커넥티드 레이어(Fully Connected Layer)와 매칭되는 제1 파라미터의 적어도 하나의 파라미터 값을 결정하기 위해 제1 풀리 커넥티드 레이어(Fully Connected Layer)를 학습시킬 수 있다.

[0011] 또한, 상기 적어도 하나의 프로세서는, 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피쳐에 미리 설정된 함수를 적용하여 어텐션(attention)을 생성하고, 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합과 상기 어텐션(attention)을 결합하여 워드 어텐디드 피쳐(word attended features)를 생성하고, 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)를 이용하여 상기 워드 어텐디드 피쳐(word attended features) 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피쳐를 디코딩하여 제2 아웃풋(output)을 생성하고, 상기 제2 아웃풋(output)을 기초로 미리 설정된 단어집에서 단어를 선택할 수 있다.

[0012] 또한, 상기 적어도 하나의 프로세서는, 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피쳐를 결합(concatenate)하고, 상기 결합(concatenate)한 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피쳐를 입력으로 하는 학습이 완료된 제2 풀리 커넥티드 레이어(Fully Connected Layer)와 매칭되는 제2 파라미터의 적어도 하나의 파라미터 값을 결정하기 위해 제2 풀리 커넥티드 레이어(Fully Connected Layer)를 학습시킬 수 있다.

[0013] 또한, 상기 적어도 하나의 프로세서는, 상기 단어집과 매칭되도록 상기 제2 아웃풋(output)의 차원을 변화시키기 위한 제3 파라미터를 결정할 수 있다.

[0014] 본 발명의 다른 실시예에 따르면, 비디오 피쳐 및 템퍼럴 어텐션(temporal attention)을 기초로 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합을 생성하는 동작 및 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합 및 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피쳐를 기초로 단어를 생성하는 동작을 포함한다.

[0015] 또한, 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합을 생성하는 동작은, 제1 딥 뉴럴 네트워크(Deep Neural Networks)를 이용하여 상기 비디오 피쳐를 인코딩하여 제1 아웃풋(output)을 생성하는 동작, 상기 제1 아웃풋(output)에 미리 설정된 함수를 적용하여 템퍼럴 어텐션(temporal attention)을 생성하는 동작 및 상기 비디오 피쳐와 상기 템퍼럴 어텐션(temporal attention)을 결합하여 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합을 생성하는 동작을 포함할 수 있다.

[0016] 또한, 상기 제1 아웃풋(output)에 미리 설정된 함수를 적용하여 템퍼럴 어텐션(temporal attention)을 생성하는 동작은, 상기 제1 아웃풋(output)을 입력으로 하는 학습이 완료된 제1 풀리 커넥티드 레이어(Fully Connected Layer)와 매칭되는 제1 파라미터의 적어도 하나의 파라미터 값을 결정하기 위해 제1 풀리 커넥티드 레이어(Fully Connected Layer)를 학습시키는 동작을 더 포함할 수 있다.

[0017] 또한, 상기 단어를 생성하는 동작은, 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피쳐에 미리 설정된 함수를 적용하여 어텐션(attention)을 생성하는 동작, 상기 템퍼럴 어텐션(temporal

attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합과 상기 어텐션(attention)을 결합하여 워드 어텐디드 피쳐(word attended features)를 생성하는 동작, 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)를 이용하여 상기 워드 어텐디드 피쳐(word attended features) 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피쳐를 디코딩하여 제2 아웃풋(output)을 생성하는 동작 및 상기 제2 아웃풋(output)을 기초로 미리 설정된 단어집에서 단어를 선택하는 동작을 포함할 수 있다.

[0018] 또한, 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피쳐에 미리 설정된 함수를 적용하여 어텐션(attention)을 생성하는 동작은, 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피쳐를 결합(concatenate)하는 동작 및 상기 결합(concatenate)한 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피쳐를 입력으로 하는 학습이 완료된 제2 풀리 커넥티드 레이어(Fully Connected Layer)와 매칭되는 제2 파라미터의 적어도 하나의 파라미터 값을 결정하기 위해 제2 풀리 커넥티드 레이어(Fully Connected Layer)를 학습시키는 동작을 더 포함할 수 있다.

[0019] 또한, 상기 제2 아웃풋(output)을 기초로 미리 설정된 단어집에서 단어를 선택하는 동작은, 상기 단어집과 매칭되도록 상기 제2 아웃풋(output)의 차원을 변화시키기 위한 제3 파라미터를 결정하는 동작을 더 포함할 수 있다.

발명의 효과

[0020] 본 발명의 일실시예에 따르면, 비디오 피쳐와 템퍼럴 어텐션(temporal attention)을 결합하여 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합을 생성할 수 있는 효과가 있다.

[0021] 또한, 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합 및 이전 히든 스테이트(hidden state)가 출력한 피쳐를 기초로 단어를 생성할 수 있는 효과가 있다.

[0022] 또한, 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합과 어텐션을 결합하여 워드 어텐디드 피쳐를 생성할 수 있는 효과가 있다.

도면의 간단한 설명

[0023] 도 1은 일실시예에 따른 비디오 캡셔닝(captioning) 학습 장치의 구성을 나타내는 도면이다.

도 2는 일실시예에 따른 비디오 캡셔닝(captioning) 학습 방법을 나타내는 플로우 차트이다.

도 3은 일실시예에 따른 비디오 캡셔닝(captioning) 장치의 구성을 나타내는 도면이다.

도 4는 일실시예에 따른 비디오 캡셔닝(captioning) 방법을 나타내는 플로우 차트이다.

도 5는 일실시예에 따른 비디오 캡셔닝(captioning) 장치에 포함된 비디오 피쳐 인코더와 캡션 디코더의 동작을 나타내는 도면이다.

도 6은 일실시예에 따라 비디오 캡셔닝(captioning) 장치가 비디오 캡셔닝(video captioning)을 수행하는 모습을 나타내는 모습이다.

도 7은 본 발명의 일실시예를 구현하기 위한 예시적인 컴퓨터 시스템의 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0024] 본 명세서에 개시되어 있는 본 발명의 개념에 따른 실시 예들에 대해서 특정한 구조적 또는 기능적 설명들은 단지 본 발명의 개념에 따른 실시 예들을 설명하기 위한 목적으로 예시된 것으로서, 본 발명의 개념에 따른 실시 예들은 다양한 형태들로 실시될 수 있으며 본 명세서에 설명된 실시 예들에 한정되지 않는다.

[0025] 본 발명의 개념에 따른 실시 예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시 예들을 도면에 예시하고 본 명세서에 상세하게 설명하고자 한다. 그러나, 이는 본 발명의 개념에 따른 실시 예들을 특정한 개시 형태들에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물, 또는 대체물을 포함한다.

[0026] 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 이탈되지 않은 채, 제1구성요소는 제2구성요소로 명명될 수 있고, 유사하게 제2구성요소는 제1구성요소로도 명명될 수 있다.

[0027] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다. 구성요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0028] 본 명세서에서 사용한 용어는 단지 특정한 실시 예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다.

[0029] 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 설명된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0030] 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 의미가 있다.

[0031] 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.

[0032] 이하의 설명에서 동일한 식별 기호는 동일한 구성을 의미하며, 불필요한 중복적인 설명 및 공지 기술에 대한 설명은 생략하기로 한다.

[0033] 이하, 첨부한 도면을 참조하여 본 발명의 바람직한 실시 예를 설명함으로써, 본 발명을 상세히 설명한다.

[0034] 도 1은 일실시예에 따른 비디오 캡셔닝(captioning) 학습 장치의 구성을 나타내는 도면이다.

[0035] 도 1을 참조하면, 일실시예에 따른 비디오 캡셔닝(captioning) 학습 장치(100)는 비디오 피쳐 추출 모듈(110), 비디오 피쳐 인코더(120), 캡션 디코더(130), 단어 선택 모듈(140) 및 손실 함수(150)를 포함한다.

[0036] 일실시예에 따라 비디오 캡셔닝(captioning) 학습 장치(100)에 포함된 비디오 피쳐 추출 모듈(110), 비디오 피쳐 인코더(120), 캡션 디코더(130), 단어 선택 모듈(140) 및 손실 함수(150)는 상호 연결되어 있으며, 상호 데이터를 전송하는 것이 가능하다.

[0037] 일실시예에 따라 비디오 캡셔닝(captioning) 학습 장치(100)는 비디오(101)를 획득할 수 있다. 이때, 비디오(101)는 일정 간격으로 선택된 프레임들의 집합일 수 있으나, 비디오(101)가 이에 한정되는 것은 아니다. 또한, 비디오(101)는 구간별로 구분되는 영상일 수 있으나, 비디오(101)가 이에 한정되는 것은 아니다.

[0038] 일실시예에 따라 비디오 피쳐 추출 모듈(110)은 비디오(101)를 획득할 수 있다.

[0039] 일실시예에 따라 비디오 피쳐 추출 모듈(110)은 비디오(101)에서 비디오 피쳐를 추출하기 위한 딥 뉴럴 네트워크(Deep Neural Networks)를 포함할 수 있다. 이때, 상기 딥 뉴럴 네트워크(Deep Neural Network)는 3D 합성곱 신경망(3D Convolutional Neural Networks, 3D CNN)일 수 있으나, 상기 딥 뉴럴 네트워크(Deep Neural Network)가 이에 한정되는 것은 아니다.

[0040] 일실시예에 따라 비디오 피쳐 추출 모듈(110)은 상기 딥 뉴럴 네트워크(Deep Neural Networks)를 이용하여 비디오(101)에서 비디오 피쳐를 추출할 수 있다. 이때, 상기 비디오 피쳐는 구간별로 구분된 비디오 피쳐 일 수 있으나, 상기 비디오 피쳐가 이에 한정되는 것은 아니다. 또한, 상기 비디오 피쳐는 비디오(101)의 입력 순서에 따라 순차적으로 추출될 수 있으나, 상기 비디오 피쳐의 순서가 이에 한정되는 것은 아니다.

[0041] 일실시예에 따라 비디오 피쳐 인코더(120)는 비디오 피쳐 추출 모듈(110)이 추출한 비디오 피쳐를 획득할 수 있다.

[0042] 일실시예에 따라 비디오 피쳐 인코더(120)는 제1 딥 뉴럴 네트워크(Deep Neural Networks)를 이용하여 반복적으로(iterative) 비디오(101)의 구간을 나타내는 상기 비디오 피쳐를 인코딩 하여 비디오(101)의 구간과 매칭되는 제1 아웃풋(output)을 생성할 수 있다. 이때, 상기 제1 딥 뉴럴 네트워크(Deep Neural Networks)는 장단기 메모리(Long Short-Term Memory, LSTM)일 수 있으나, 상기 제1 딥 뉴럴 네트워크(Deep Neural Networks)가 이에 한정되는 것은 아니다. 또한, 상기 제1 아웃풋(output)은 비디오(101)의 모든 구간에서 생성한 아웃풋일 수 있으나, 상기 제1 아웃풋(output)이 이에 한정되는 것은 아니다.

[0043] 일실시예에 따라 비디오 피쳐 인코더(120)는 하기 [수학식 1]을 이용하여 템퍼럴 어텐션(temporal attention, a_T)을 생성할 수 있다.

수학식 1

$$a_T = \sigma(W_T \cdot o_T + b_T), \quad a_T \in \mathbb{R}^{N_v}$$

[0044]
$$o_T = [o_1, \ldots, o_{N_v}]^{\mathrm{T}}$$

[0045] 여기서, σ는 시그모이드 함수 (Sigmoid)이고, W_T 및 b_T는 제1 풀리 커넥티드 레이어(Fully Connected Layer)의 학습에 의해 결정되는 제1 파라미터들이고, o_T는 제1 아웃풋(output)이다.

[0046] 일실시예에 따라 비디오 피쳐 인코더(120)는 제1 풀리 커넥티드 레이어(Fully Connected Layer)와 매칭되도록 하는 제1 파라미터들(W_T 및 b_T)을 제1 아웃풋(output)과 결합한 피쳐에 시그모이드 함수 (Sigmoid)(σ)를 적용하여 템퍼럴 어텐션(temporal attention, a_T)을 생성할 수 있다.

[0047] 일실시예에 따라 비디오 피쳐 인코더(120)는 하기 [수학식 2]를 이용하여 비디오 피쳐와 템퍼럴 어텐션(temporal attention, a_T)을 결합하여 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합 (x)을 생성할 수 있다.

수학식 2

[0048]
$$x = \sum_{t=1}^{N_v} a_{T_t} v_t$$

[0049] 여기서, a_{T_t}는 템퍼럴 어텐션(temporal attention)이고, v_t는 비디오 피쳐이다.

[0050] 일실시예에 따라 캡션 디코더(130)는 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피쳐의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피쳐에 미리 설정된 함수를 적용하여 어텐션을 생성할 수 있다. 이때, 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)는 장단기 메모리(Long Short-Term Memory, LSTM)일 수 있으나, 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)가 이에 한정되는 것은 아니다.

[0051] 일실시예에 따라 캡션 디코더(130)는 하기 [수학식 3]을 이용하여 어텐션(attention, a_{U_j})을 생성할 수 있다.

수학식 3

$$a_{U_j} = \sigma(W_U \cdot [h_j^u, x] + b_U)$$

[0052] $$W_U \in \mathbb{R}^{D_v \times 2D_v} \quad b_U \in \mathbb{R}^{D_v}$$

[0053] 여기서, σ는 시그모이드 함수 (Sigmoid)이고, W_U 및 b_U는 제2 풀리 커넥티드 레이어(Fully Connected Layer)의 학습에 의해 결정되는 제2 파라미터들이고, h_j^u는 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처이고, x는 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합이다.

[0054] 일실시예에 따라 캡션 디코더(130)는 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처(h_j^u)와 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합(x)을 결합(concatenate)할 수 있다.

[0055] 일실시예에 따라 캡션 디코더(130)는 제2 풀리 커넥티드 레이어(Fully Connected Layer)와 매칭되도록 하는 제2 파라미터들(W_U 및 b_U)을 상기 결합(concatenate)한 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))의 이전 히든 스테이트(hidden state)가 출력한 피처와 결합한 피처에 시그모이드 함수(Sigmoid)(σ)를 적용하여 어텐션(attention, a_{U_j})을 생성할 수 있다.

[0056] 일실시예에 따라 캡션 디코더(130)는 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))의 이전 히든 스테이트(hidden state)가 출력한 피처(h_j^u)의 수만큼 어텐션(attention, a_{U_j})을 반복적으로 생성할 수 있다.

[0057] 일실시예에 따라 캡션 디코더(130)는 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합(x)과 상기 어텐션(attention, a_{U_j})을 결합하여 워드 어텐디드 피처(word attended features)를 생성할 수 있다.

[0058] 일실시예에 따라 캡션 디코더(130)는 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합(x)과 상기 어텐션(attention, a_{U_j})을 결합하기 위하여 스칼라 곱(dot product)을 이용할 수 있으나, 캡션 디코더(130)가 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합(x)과 상기 어텐션(attention, a_{U_j})을 결합하기 위하여 사용하는 방법이 이에 한정되는 것은 아니다.

[0059] 일실시예에 따라 캡션 디코더(130)는 생성한 어텐션(attention, a_{U_j})의 수만큼 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합(x)과 상기 어텐션(attention, a_{U_j})을

결합하여 워드 어텐디드 피처(word attended features)를 반복적으로 생성할 수 있다.

[0060] 일실시예에 따라 캡션 디코더(130)는 하기 [수학식 4]를 이용하여 제2 아웃풋(output, o_{j+1}^u, h_{j+1}^u)을 생성할 수 있다.

수학식 4

[0061] $$o_{j+1}^u, h_{j+1}^u = \text{LSTM}_U([a_{U_j} \odot x, u_j], h_j^u)$$

[0062] 여기서, a_{U_j}는 어텐션이고, x는, 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합이고, u_j는 j번째 워드 피처 벡터이고, h_j^u는 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처이다.

[0063] 일실시예에 따라 캡션 디코더(130)는 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))를 이용하여 상기 워드 어텐디드 피처(word attended features) 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처(h_j^u)를 디코딩하여 제2 아웃풋(output, o_{j+1}^u, h_{j+1}^u)을 생성할 수 있다.

[0064] 일실시예에 따라 캡션 디코더(130)는 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))를 이용하여 상기 워드 어텐디드 피처(word attended features) 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처(h_j^u)를 디코딩하여 제2 아웃풋(output, o_{j+1}^u, h_{j+1}^u)을 반복적으로 생성할 수 있다.

[0065] 일실시예에 따라 캡션 디코더(130)는 상기 워드 어텐디드 피처(word attended features)의 수만큼 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))를 이용하여 상기 워드 어텐디드 피처(word attended features) 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처(h_j^u)를 디코딩하여 제2 아웃풋(output, o_{j+1}^u, h_{j+1}^u)을 반복적으로 생성할 수 있다.

[0066] 일실시예에 따라 캡션 디코더(130)는 하기 [수학식 5]를 이용하여 단어집에 대한 확률 값($\hat{u}_{j+1}$)을 생성할 수 있다.

수학식 5

[0067] $$\hat{u}_{j+1} = \text{Softmax}(W_o \cdot o_{j+1}^u)$$

[0068] 여기서, W_o는 학습에 의해 결정되는 제3 파라미터이고, o_{j+1}^u는 제2 아웃풋(output)이다.

[0069] 일실시예에 따라 단어 선택 모듈(140)은 제2 아웃풋(output, o_{j+1}^{u})의 수만큼 단어집에 대한 확률 값($\hat{u}_{j+1}$)을 반복적으로 생성할 수 있다.

[0070] 일실시예에 따라 캡션 디코더(130)는 학습을 통해 상기 단어집과 매칭되도록 상기 제2 아웃풋(output)의 차원을 변화시키기 위한 제3 파라미터를 결정할 수 있다.

[0071] 일실시예에 따라 단어 선택 모듈(140)은 캡션 디코더(130)가 생성한 단어집에 대한 확률 값($\hat{u}_{j+1}$)을 기초로 미리 설정된 단어집에서 단어를 선택할 수 있다.

[0072] 일실시예에 따라 단어 선택 모듈(140)은 단어집에서 가장 높은 확률 값을 가지는 단어를 선택할 수 있다.

[0073] 일실시예에 따라 단어 선택 모듈(140)은 선택한 단어를 생성할 수 있다.

[0074] 일실시예에 따라 손실 함수(150)는 정답(ground truth)을 획득할 수 있다.

[0075] 일실시예에 따라 손실 함수(150)는 상기 획득한 정답(ground truth)을 기초로 손실(loss)을 계산할 수 있다.

[0076] 일실시예에 따라 손실 함수(150)는 하기 [수학식 6]을 이용하여 손실(loss)을 계산할 수 있다.

수학식 6

[0077] $$L = \lambda_1 L_{var} + \lambda_2 L_{prob} + \lambda_3 R(W)$$

[0078] 여기서, λ_1, λ_2 및 λ_3는 스케일을 위한 계수이고, L_{var}는 분산에 관한 손실이고, L_{prob}는 정답 확률에 관한 손실이고, $R(W)$는 레귤러리제이션(regularization)이다.

[0079] 또한, 하기 [수학식 7]을 이용하여 분산에 관한 손실 L_{var}를 계산할 수 있다.

수학식 7

[0080] $$L_{var} = \frac{1}{N_u} \sum_{j=1}^{N_u} (\hat{u}_j - \bar{u}_j)^2$$

[0081] 여기서, $\hat{u}_j$는 생성된 단어 벡터이고, $\bar{u}_j$는 생성된 단어 벡터의 평균이다.

[0082] 일실시예에 따라 손실 함수(150)는 상기 계산한 손실을 기초로 비디오 캡셔닝(captioning) 학습 장치(100)에 포함된 적어도 하나의 네트워크를 업데이트 할 수 있다.

[0083] 일실시예에 따라 손실 함수(150)는 역전파(backpropagation)를 이용하여 비디오 캡셔닝(captioning) 학습 장치(100)에 포함된 적어도 하나의 네트워크를 업데이트 할 수 있다.

[0084] 다른 실시예에 따라 비디오 캡셔닝(captioning) 학습 장치(100)는 상기 계산한 손실을 기초로 비디오 캡셔닝(captioning) 학습 장치(100)에 포함된 적어도 하나의 네트워크를 업데이트 할 수 있다.

[0085] 다른 실시예에 따라 비디오 캡셔닝(captioning) 학습 장치는 역전파(backpropagation)를 이용하여 비디오 캡셔

닝(captioning) 학습 장치(100)에 포함된 적어도 하나의 네트워크를 업데이트 할 수 있다.

[0086] 일실시예에 따라 단어 선택 모듈(140)은 선택한 단어를 기초로 문장(102)을 생성하여 출력할 수 있다.

[0087] 다른 실시예에 따라 비디오 캡셔닝(captioning) 학습 장치는 단어 선택 모듈(140)이 선택한 단어를 기초로 문장(102)을 생성하여 출력할 수 있다.

[0088] 여기서 사용된 '모듈'이라는 용어는 논리적인 구성 단위를 나타내는 것으로서, 반드시 물리적으로 구분되는 구성 요소가 아니라는 점은 본 발명이 속하는 기술분야의 당업자에게 자명한 사항이다.

[0089] 도 2는 일실시예에 따른 비디오 캡셔닝(captioning) 학습 방법을 나타내는 플로우 차트이다.

[0090] 도 2를 참조하면, 비디오 캡셔닝(captioning) 학습 장치가 비디오 피처 및 템퍼럴 어텐션(temporal attention)을 기초로 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합을 생성한다(200).

[0091] 이때, 비디오 캡셔닝(captioning) 학습 장치는 비디오에서 비디오 피처를 추출하기 위한 딥 뉴럴 네트워크(Deep Neural Networks)를 포함할 수 있다. 이때, 상기 딥 뉴럴 네트워크(Deep Neural Network)는 3D 합성곱 신경망(3D Convolutional Neural Networks, 3D CNN)일 수 있으나, 상기 딥 뉴럴 네트워크(Deep Neural Network)가 이에 한정되는 것은 아니다.

[0092] 또한, 비디오 캡셔닝(captioning) 학습 장치는 상기 딥 뉴럴 네트워크(Deep Neural Networks)를 이용하여 비디오에서 비디오 피처를 추출할 수 있다. 이때, 상기 비디오 피처는 구간별로 구분된 비디오 피처 일 수 있으나, 상기 비디오 피처가 이에 한정되는 것은 아니다. 또한, 상기 비디오 피처는 비디오(101)의 입력 순서에 따라 순차적으로 추출될 수 있으나, 상기 비디오 피처의 순서가 이에 한정되는 것은 아니다.

[0093] 또한, 비디오 캡셔닝(captioning) 학습 장치는 제1 딥 뉴럴 네트워크(Deep Neural Networks)를 이용하여 반복적으로(iterative) 비디오의 구간을 나타내는 상기 비디오 피처를 인코딩 하여 비디오의 구간과 매칭되는 제1 아웃풋(output)을 생성할 수 있다. 이때, 상기 제1 딥 뉴럴 네트워크(Deep Neural Networks)는 장단기 메모리(Long Short-Term Memory, LSTM)일 수 있으나, 상기 제1 딥 뉴럴 네트워크(Deep Neural Networks)가 이에 한정되는 것은 아니다. 또한, 상기 제1 아웃풋(output)은 비디오의 모든 구간에서 생성한 아웃풋일 수 있으나, 상기 제1 아웃풋(output)이 이에 한정되는 것은 아니다.

[0094] 또한, 비디오 캡셔닝(captioning) 학습 장치는 제1 풀리 커넥티드 레이어(Fully Connected Layer)와 매칭되도록 하는 제1 파라미터들을 제1 아웃풋(output)과 결합한 피처에 시그모이드 함수 (Sigmoid)를 적용하여 템퍼럴 어텐션(temporal attention)을 생성할 수 있다.

[0095] 또한, 비디오 캡셔닝(captioning) 학습 장치는 비디오 피처와 템퍼럴 어텐션(temporal attention)을 결합하여 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합을 생성할 수 있다.

[0096] 비디오 캡셔닝(captioning) 학습 장치가 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처에 미리 설정된 함수를 적용하여 어텐션(attention)을 생성한다(210).

[0097] 이때, 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)는 장단기 메모리(Long Short-Term Memory, LSTM)일 수 있으나, 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)가 이에 한정되는 것은 아니다.

[0098] 또한, 상기 미리 설정된 함수는 시그모이드 함수 (Sigmoid)일 수 있으나, 상기 미리 설정된 함수가 이에 한정되는 것은 아니다.

[0099] 또한, 비디오 캡셔닝(captioning) 학습 장치는 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처와 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합을 결합(concatenate)할 수 있다.

[0100] 또한, 비디오 캡셔닝(captioning) 학습 장치는 제2 풀리 커넥티드 레이어(Fully Connected Layer)와 매칭되도록 하는 제2 파라미터들을 상기 결합(concatenate)한 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))의 이전 히든 스테이트(hidden state)가 출력한 피처와 결합한 피처에 시그모이드 함수 (Sigmoid)를 적용하여 어텐션(attention)을 생성할 수 있다.

[0101] 또한, 비디오 캡셔닝(captioning) 학습 장치는 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))의 이전 히든 스테이트(hidden state)가 출력한 피처의 수만큼 어텐션(attention)을 반복적으로 생성할 수 있다.

[0102] 비디오 캡셔닝(captioning) 학습 장치가 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합과 상기 어텐션(attention)을 결합하여 워드 어텐디드 피처(word attended features)를 생성한다(220).

[0103] 이때, 비디오 캡셔닝(captioning) 학습 장치는 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합과 상기 어텐션(attention)을 결합하기 위하여 스칼라 곱(dot product)을 이용할 수 있으나, 비디오 캡셔닝(captioning) 학습 장치가 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합과 상기 어텐션(attention)을 결합하기 위하여 사용하는 방법이 이에 한정되는 것은 아니다.

[0104] 또한, 비디오 캡셔닝(captioning) 학습 장치는 생성한 어텐션(attention)의 수만큼 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합과 상기 어텐션(attention)을 결합하여 워드 어텐디드 피처(word attended features)를 생성할 수 있다.

[0105] 비디오 캡셔닝(captioning) 학습 장치가 상기 워드 어텐디드 피처(word attended features) 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처를 기초로 단어를 생성한다(230).

[0106] 이때, 비디오 캡셔닝(captioning) 학습 장치는 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))를 이용하여 상기 워드 어텐디드 피처(word attended features) 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처를 디코딩하여 제2 아웃풋(output)을 생성할 수 있다.

[0107] 또한, 비디오 캡셔닝(captioning) 학습 장치는 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))를 이용하여 상기 워드 어텐디드 피처(word attended features) 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처를 디코딩하여 제2 아웃풋(output)을 반복적으로 생성할 수 있다.

[0108] 또한, 비디오 캡셔닝(captioning) 학습 장치는 상기 워드 어텐디드 피처(word attended features)의 수만큼 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))를 이용하여 상기 워드 어텐디드 피처(word attended features) 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처를 디코딩하여 제2 아웃풋(output)을 반복적으로 생성할 수 있다.

[0109] 또한, 비디오 캡셔닝(captioning) 학습 장치는 상기 제2 아웃풋(output)을 기초로 단어집에 대한 확률 값을 생성할 수 있다.

[0110] 또한, 비디오 캡셔닝(captioning) 학습 장치는 상기 제2 아웃풋(output)의 수만큼 단어집에 대한 확률 값을 생성할 수 있다.

[0111] 또한, 비디오 캡셔닝(captioning) 학습 장치는 학습을 통해 상기 단어집과 매칭되도록 상기 제2 아웃풋(output)의 차원을 변화시키기 위한 제3 파라미터를 결정할 수 있다.

[0112] 또한, 비디오 캡셔닝(captioning) 학습 장치는 상기 생성한 단어집에 대한 확률 값을 기초로 미리 설정된 단어집에서 단어를 선택할 수 있다.

[0113] 또한, 비디오 캡셔닝(captioning) 학습 장치는 상기 단어집에서 가장 높은 확률 값을 가지는 단어를 선택할 수 있다.

[0114] 비디오 캡셔닝(captioning) 학습 장치가 네트워크를 업데이트 한다(240).

[0115] 이때, 비디오 캡셔닝(captioning) 학습 장치는 정답(ground truth)을 획득할 수 있다.

[0116] 또한, 비디오 캡셔닝(captioning) 학습 장치는 상기 획득한 정답(ground truth)을 기초로 손실(loss)을 계산할 수 있다.

[0117] 또한, 비디오 캡셔닝(captioning) 학습 장치는 분산에 관한 손실을 계산할 수 있다.

[0118] 또한, 비디오 캡셔닝(captioning) 학습 장치는 상기 계산한 손실을 기초로 상기 비디오 캡셔닝(captioning) 학습 장치에 포함된 적어도 하나의 네트워크를 업데이트 할 수 있다.

[0119] 또한, 비디오 캡셔닝(captioning) 학습 장치는 역전파(backpropagation)를 이용하여 비디오 캡셔닝(captioning) 학습 장치에 포함된 적어도 하나의 네트워크를 업데이트 할 수 있다.

[0120] 또한, 비디오 캡셔닝(captioning) 학습 장치는 상기 선택한 단어를 기초로 문장을 생성하여 출력할 수 있다.

[0121] 도 3은 일실시예에 따른 비디오 캡셔닝(captioning) 장치의 구성을 나타내는 도면이다.

[0122] 도 3을 참조하면, 일실시예에 따른 비디오 캡셔닝(captioning) 장치(300)는 비디오 피처 추출 모듈(310), 비디오 피처 인코더(320), 캡션 디코더(330) 및 단어 선택 모듈(340)을 포함한다.

[0123] 일실시예에 따라 비디오 캡셔닝(captioning) 장치(300)에 포함된 비디오 피처 추출 모듈(310), 비디오 피처 인코더(320), 캡션 디코더(330) 및 단어 선택 모듈(340)은 상호 연결되어 있으며, 상호 데이터를 전송하는 것이 가능하다.

[0124] 일실시예에 따라 비디오 캡셔닝(captioning) 장치(300)는 비디오(301)를 획득할 수 있다. 이때, 비디오(301)는 일정 간격으로 선택된 프레임들의 집합일 수 있으나, 비디오(301)가 이에 한정되는 것은 아니다. 또한, 비디오(301)는 구간별로 구분되는 영상일 수 있으나, 비디오(301)가 이에 한정되는 것은 아니다.

[0125] 일실시예에 따라 비디오 피처 추출 모듈(310)은 비디오(301)를 획득할 수 있다.

[0126] 일실시예에 따라 비디오 피처 추출 모듈(310)은 비디오(301)에서 비디오 피처를 추출하기 위한 딥 뉴럴 네트워크(Deep Neural Networks)를 포함할 수 있다. 이때, 상기 딥 뉴럴 네트워크(Deep Neural Network)는 3D 합성곱 신경망(3D Convolutional Neural Networks, 3D CNN)일 수 있으나, 상기 딥 뉴럴 네트워크(Deep Neural Network)가 이에 한정되는 것은 아니다.

[0127] 일실시예에 따라 비디오 피처 추출 모듈(310)은 상기 딥 뉴럴 네트워크(Deep Neural Networks)를 이용하여 비디오(301)에서 비디오 피처를 추출할 수 있다. 이때, 상기 비디오 피처는 구간별로 구분된 비디오 피처 일 수 있으나, 상기 비디오 피처가 이에 한정되는 것은 아니다. 또한, 상기 비디오 피처는 비디오(301)의 입력 순서에 따라 순차적으로 추출될 수 있으나, 상기 비디오 피처의 순서가 이에 한정되는 것은 아니다.

[0128] 일실시예에 따라 비디오 피처 인코더(320)는 비디오 피처 추출 모듈(310)이 추출한 비디오 피처를 획득할 수 있다.

[0129] 일실시예에 따라 비디오 피처 인코더(320)는 제1 딥 뉴럴 네트워크(Deep Neural Networks)를 이용하여 반복적으로(iterative) 비디오(301)의 구간을 나타내는 상기 비디오 피처를 인코딩 하여 비디오(301)의 구간과 매칭되는 제1 아웃풋(output)을 생성할 수 있다. 이때, 상기 제1 딥 뉴럴 네트워크(Deep Neural Networks)는 장단기 메모리(Long Short-Term Memory, LSTM)일 수 있으나, 상기 제1 딥 뉴럴 네트워크(Deep Neural Networks)가 이에 한정되는 것은 아니다. 또한, 상기 제1 아웃풋(output)은 비디오(301)의 모든 구간에서 생성한 아웃풋일 수 있으나, 상기 제1 아웃풋(output)이 이에 한정되는 것은 아니다.

[0130] 일실시예에 따라 비디오 피처 인코더(320)는 하기 [수학식 8]을 이용하여 템퍼럴 어텐션(temporal attention, a_T)을 생성할 수 있다.

수학식 8

$$a_T = \sigma(W_T \cdot o_T + b_T), \; a_T \in \mathbb{R}^{N_v}$$

$$o_T = [o_1, \ldots, o_{N_v}]^{\mathrm{T}}$$

[0131]

[0132] 여기서, σ는 시그모이드 함수 (Sigmoid)이고, W_T 및 b_T는 제1 풀리 커넥티드 레이어(Fully Connected Layer)의 학습에 의해 결정된 파라미터들이고, o_T는 제1 아웃풋(output)이다.

[0133] 일실시예에 따라 비디오 피처 인코더(320)는 학습에 의해 결정된 파라미터들(W_T 및 b_T)을 제1 아웃풋(output)과 결합한 피처에 시그모이드 함수(Sigmoid)(σ)를 적용하여 템퍼럴 어텐션(temporal attention, a_T)을 생성할 수 있다.

[0134] 일실시예에 따라 비디오 피처 인코더(320)는 하기 [수학식 9]를 이용하여 비디오 피처와 템퍼럴 어텐션(temporal attention, a_T)을 결합하여 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합 (x)을 생성할 수 있다.

수학식 9

[0135] $$x = \sum_{t=1}^{N_v} a_{T_t} v_t$$

[0136] 여기서, a_{T_t} 는 템퍼럴 어텐션(temporal attention)이고, v_t 는 비디오 피처이다.

[0137] 일실시예에 따라 캡션 디코더(330)는 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처에 미리 설정된 함수를 적용하여 어텐션을 생성할 수 있다. 이때, 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)는 장단기 메모리(Long Short-Term Memory, LSTM)일 수 있으나, 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)가 이에 한정되는 것은 아니다.

[0138] 일실시예에 따라 캡션 디코더(330)는 하기 [수학식 10]을 이용하여 어텐션(attention, a_{U_j})을 생성할 수 있다.

수학식 10

$$a_{U_j} = \sigma(W_U \cdot [h_j^u, x] + b_U)$$

[0139] $$W_U \in \mathbb{R}^{D_v \times 2D_v} \quad b_U \in \mathbb{R}^{D_v}$$

[0140] 여기서, σ 는 시그모이드 함수 (Sigmoid)이고, W_U 및 b_U 는 제2 풀리 커넥티드 레이어(Fully Connected Layer)의 학습에 의해 결정된 파라미터들이고, h_j^u 는 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처이고, x 는 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합이다.

[0141] 일실시예에 따라 캡션 디코더(330)는 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처(h_j^u)와 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의

가중치 합(x)을 결합(concatenate)할 수 있다.

[0142] 일실시예에 따라 캡션 디코더(330)는 학습에 의해 결정된 파라미터들(W_U 및 b_U)을 상기 결합(concatenate)한 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))의 이전 히든 스테이트(hidden state)가 출력한 피처와 결합한 피처에 시그모이드 함수(Sigmoid)(σ)를 적용하여 어텐션(attention, a_{U_j})을 생성할 수 있다.

[0143] 일실시예에 따라 캡션 디코더(330)는 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))의 이전 히든 스테이트(hidden state)가 출력한 피처(h_j^u)의 수만큼 어텐션(attention, a_{U_j})을 반복적으로 생성할 수 있다.

[0144] 일실시예에 따라 캡션 디코더(330)는 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합(x)과 상기 어텐션(attention, a_{U_j})을 결합하여 워드 어텐디드 피처(word attended features)를 생성할 수 있다.

[0145] 일실시예에 따라 캡션 디코더(330)는 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합(x)과 상기 어텐션(attention, a_{U_j})을 결합하기 위하여 스칼라 곱(dot product)을 이용할 수 있으나, 캡션 디코더(330)가 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합(x)과 상기 어텐션(attention, a_{U_j})을 결합하기 위하여 사용하는 방법이 이에 한정되는 것은 아니다.

[0146] 일실시예에 따라 캡션 디코더(330)는 생성한 어텐션(attention, a_{U_j})의 수만큼 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합(x)과 상기 어텐션(attention, a_{U_j})을 결합하여 워드 어텐디드 피처(word attended features)를 반복적으로 생성할 수 있다.

[0147] 일실시예에 따라 캡션 디코더(330)는 하기 [수학식 11]을 이용하여 제2 아웃풋(output, o_{j+1}^u, h_{j+1}^u)을 생성할 수 있다.

수학식 11

[0148] $$o_{j+1}^u, h_{j+1}^u = \mathrm{LSTM}_U([a_{U_j} \odot x, u_j], h_j^u)$$

[0149] 여기서, a_{U_j}는 어텐션이고, x는, 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합이고, u_j는 j번째 단어 피처 벡터이고, h_j^u는 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처이다.

[0150] 일실시예에 따라 캡션 디코더(330)는 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))를 이용하여 상기 워드 어텐디드 피처(word attended features) 및 상기 제2

딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피쳐(h_j^u)를 디코딩하여 제2 아웃풋(output, o_{j+1}^u, h_{j+1}^u)을 생성할 수 있다.

[0151] 일실시예에 따라 캡션 디코더(330)는 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))를 이용하여 상기 워드 어텐디드 피쳐(word attended features) 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피쳐(h_j^u)를 디코딩하여 제2 아웃풋(output, o_{j+1}^u, h_{j+1}^u)을 반복적으로 생성할 수 있다.

[0152] 일실시예에 따라 캡션 디코더(330)는 상기 워드 어텐디드 피쳐(word attended features)의 수만큼 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))를 이용하여 상기 워드 어텐디드 피쳐(word attended features) 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피쳐(h_j^u)를 디코딩하여 제2 아웃풋(output, o_{j+1}^u, h_{j+1}^u)을 반복적으로 생성할 수 있다.

[0153] 일실시예에 따라 캡션 디코더(330)는 하기 [수학식 12]를 이용하여 단어집에 대한 확률 값($\hat{u}_{j+1}$)을 생성할 수 있다.

수학식 12

[0154] $$\hat{u}_{j+1} = \text{Softmax}(W_o \cdot o_{j+1}^u)$$

[0155] 여기서, W_o 는 학습에 의해 결정된 제3 파라미터이고, o_{j+1}^u 는 제2 아웃풋(output)이다.

[0156] 일실시예에 따라 단어 선택 모듈(340)은 제2 아웃풋(output, o_{j+1}^u)의 수만큼 단어집에 대한 확률 값($\hat{u}_{j+1}$)을 반복적으로 생성할 수 있다.

[0157] 일실시예에 따라 캡션 디코더(330)는 상기 단어집의 차원과 2 아웃풋(output)의 차원이 매칭되도록 하기 위하여 학습을 통해 결정된 제3 파라미터를 상기 제2 아웃풋(output)과 결합할 수 있다.

[0158] 일실시예에 따라 단어 선택 모듈(340)은 캡션 디코더(330)가 생성한 단어집에 대한 확률 값($\hat{u}_{j+1}$)을 기초로 미리 설정된 단어집에서 단어를 선택할 수 있다.

[0159] 일실시예에 따라 단어 선택 모듈(340)은 단어집에서 가장 높은 확률 값을 가지는 단어를 선택할 수 있다.

[0160] 일실시예에 따라 단어 선택 모듈(340)은 선택한 단어를 생성할 수 있다.

[0161] 일실시예에 따라 단어 선택 모듈(340)은 선택한 단어를 기초로 문장(302)을 생성하여 출력할 수 있다.

[0162] 다른 실시예에 따라 비디오 캡셔닝(captioning) 장치는 단어 선택 모듈(340)이 선택한 단어를 기초로 문장(302)을 생성하여 출력할 수 있다.

[0163] 여기서 사용된 '모듈'이라는 용어는 논리적인 구성 단위를 나타내는 것으로서, 반드시 물리적으로 구분되는 구

성 요소가 아니라는 점은 본 발명이 속하는 기술분야의 당업자에게 자명한 사항이다.

[0164] 도 4는 일실시예에 따른 비디오 캡셔닝(captioning) 방법을 나타내는 플로우 차트이다.

[0165] 도 4를 참조하면, 비디오 캡셔닝(captioning) 장치가 비디오 피처 및 템퍼럴 어텐션(temporal attention)을 기초로 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합을 생성한다(400).

[0166] 이때, 비디오 캡셔닝(captioning) 장치는 비디오에서 비디오 피처를 추출하기 위한 딥 뉴럴 네트워크(Deep Neural Networks)를 포함할 수 있다. 이때, 상기 딥 뉴럴 네트워크(Deep Neural Network)는 3D 합성곱 신경망(3D Convolutional Neural Networks, 3D CNN)일 수 있으나, 상기 딥 뉴럴 네트워크(Deep Neural Network)가 이에 한정되는 것은 아니다.

[0167] 또한, 비디오 캡셔닝(captioning) 장치는 상기 딥 뉴럴 네트워크(Deep Neural Networks)를 이용하여 비디오에서 비디오 피처를 추출할 수 있다. 이때, 상기 비디오 피처는 구간별로 구분된 비디오 피처 일 수 있으나, 상기 비디오 피처가 이에 한정되는 것은 아니다. 또한, 상기 비디오 피처는 비디오(101)의 입력 순서에 따라 순차적으로 추출될 수 있으나, 상기 비디오 피처의 순서가 이에 한정되는 것은 아니다.

[0168] 또한, 비디오 캡셔닝(captioning) 장치는 제1 딥 뉴럴 네트워크(Deep Neural Networks)를 이용하여 반복적으로(iterative) 비디오의 구간을 나타내는 상기 비디오 피처를 인코딩 하여 비디오의 구간과 매칭되는 제1 아웃풋(output)을 생성할 수 있다. 이때, 상기 제1 딥 뉴럴 네트워크(Deep Neural Networks)는 장단기 메모리(Long Short-Term Memory, LSTM)일 수 있으나, 상기 제1 딥 뉴럴 네트워크(Deep Neural Networks)가 이에 한정되는 것은 아니다. 또한, 상기 제1 아웃풋(output)은 비디오의 모든 구간에서 생성한 아웃풋일 수 있으나, 상기 제1 아웃풋(output)이 이에 한정되는 것은 아니다.

[0169] 또한, 비디오 캡셔닝(captioning) 장치는 제1 풀리 커넥티드 레이어(Fully Connected Layer)의 학습에 의해 결정된 파라미터들을 제1 아웃풋(output)과 결합한 피처에 시그모이드 함수 (Sigmoid)를 적용하여 템퍼럴 어텐션(temporal attention)을 생성할 수 있다.

[0170] 또한, 비디오 캡셔닝(captioning) 장치는 비디오 피처와 템퍼럴 어텐션(temporal attention)을 결합하여 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합을 생성할 수 있다.

[0171] 비디오 캡셔닝(captioning) 장치가 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처에 미리 설정된 함수를 적용하여 어텐션(attention)을 생성한다(410).

[0172] 이때, 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)는 장단기 메모리(Long Short-Term Memory, LSTM)일 수 있으나, 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)가 이에 한정되는 것은 아니다.

[0173] 또한, 상기 미리 설정된 함수는 시그모이드 함수 (Sigmoid)일 수 있으나, 상기 미리 설정된 함수가 이에 한정되는 것은 아니다.

[0174] 또한, 비디오 캡셔닝(captioning) 장치는 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처와 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합을 결합(concatenate)할 수 있다.

[0175] 또한, 비디오 캡셔닝(captioning) 장치는 제2 풀리 커넥티드 레이어(Fully Connected Layer)의 학습에 의해 결정된 파라미터들을 상기 결합(concatenate)한 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))의 이전 히든 스테이트(hidden state)가 출력한 피처와 결합한 피처에 시그모이드 함수 (Sigmoid)를 적용하여 어텐션(attention)을 생성할 수 있다.

[0176] 또한, 비디오 캡셔닝(captioning) 장치는 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))의 이전 히든 스테이트(hidden state)가 출력한 피처의 수만큼 어텐션(attention)을 반복적으로 생성할 수 있다.

[0177] 비디오 캡셔닝(captioning) 장치가 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합과 상기 어텐션(attention)을 결합하여 워드 어텐디드 피처(word attended features)를 생성한다(420).

[0178] 이때, 비디오 캡셔닝(captioning) 장치는 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합과 상기 어텐션(attention)을 결합하기 위하여 스칼라 곱(dot product)을 이용할 수 있으나, 비디오 캡셔닝(captioning) 장치가 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합과 상기 어텐션(attention)을 결합하기 위하여 사용하는 방법이 이에 한정되는 것은 아니다.

[0179] 또한, 비디오 캡셔닝(captioning) 장치는 생성한 어텐션(attention)의 수만큼 상기 템퍼럴 어텐션(temporal attention)이 웨이트로 적용된 비디오 피처의 가중치 합과 상기 어텐션(attention)을 결합하여 워드 어텐디드 피처(word attended features)를 생성할 수 있다.

[0180] 비디오 캡셔닝(captioning) 장치가 상기 워드 어텐디드 피처(word attended features) 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처를 기초로 단어를 생성한다(430).

[0181] 이때, 비디오 캡셔닝(captioning) 장치는 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))를 이용하여 상기 워드 어텐디드 피처(word attended features) 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처를 디코딩하여 제2 아웃풋(output)을 생성할 수 있다.

[0182] 또한, 비디오 캡셔닝(captioning) 장치는 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))를 이용하여 상기 워드 어텐디드 피처(word attended features) 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처를 디코딩하여 제2 아웃풋(output)을 반복적으로 생성할 수 있다.

[0183] 또한, 비디오 캡셔닝(captioning) 장치는 상기 워드 어텐디드 피처(word attended features)의 수만큼 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))를 이용하여 상기 워드 어텐디드 피처(word attended features) 및 상기 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처를 디코딩하여 제2 아웃풋(output)을 반복적으로 생성할 수 있다.

[0184] 또한, 비디오 캡셔닝(captioning) 장치는 상기 제2 아웃풋(output)을 기초로 단어집에 대한 확률 값을 생성할 수 있다.

[0185] 또한, 비디오 캡셔닝(captioning) 장치는 상기 제2 아웃풋(output)의 수만큼 단어집에 대한 확률 값을 생성할 수 있다.

[0186] 또한, 비디오 캡셔닝(captioning) 장치는 상기 단어집의 차원과 2 아웃풋(output)의 차원이 매칭되도록 하기 위하여 학습을 통해 결정된 제3 파라미터를 상기 제2 아웃풋(output)과 결합할 수 있다.

[0187] 또한, 비디오 캡셔닝(captioning) 장치는 상기 생성한 단어집에 대한 확률 값을 기초로 미리 설정된 단어집에서 단어를 선택할 수 있다.

[0188] 또한, 비디오 캡셔닝(captioning) 장치는 상기 단어집에서 가장 높은 확률 값을 가지는 단어를 선택할 수 있다.

[0189] 또한, 비디오 캡셔닝(captioning) 장치는 상기 선택한 단어를 기초로 문장을 생성하여 출력할 수 있다.

[0190] 도 5는 일실시예에 따른 비디오 캡셔닝(captioning) 장치에 포함된 비디오 피처 인코더와 캡션 디코더의 동작을 나타내는 도면이다.

[0191] 도 5를 참조하면, 일실시예에 따른 비디오 캡셔닝(captioning) 장치는 비디오 피처 인코더(500) 및 캡션 디코더(520)를 포함할 수 있다.

[0192] 일실시예에 따라 비디오 피처 인코더(500)는 비디오에서 일정 간격으로 선택된 프레임(501, 502, 503, 504)을 획득할 수 있다.

[0193] 일실시예에 따라 비디오 피처 인코더(500)는 딥 뉴럴 네크워크(Deep Neural Networks)(예컨대, 3D 합성곱 신경망(3D Convolutional Neural Networks, 3D CNN) 등)(505)를 이용하여 획득한 비디오 프레임(501, 502, 503, 504)에서 구간을 나타내는 비디오 피처를 추출할 수 있다.

[0194] 일실시예에 따라 비디오 피처 인코더(500)는 제1 딥 뉴럴 네트워크(Deep Neural Networks)를 이용하여 반복적으로(iterative) 비디오의 구간을 나타내는 비디오 피처를 인코딩 하여 비디오의 구간과 매칭되는 제1 아웃풋

(output)(506)을 생성할 수 있다. 이때, 상기 제1 딥 뉴럴 네트워크(Deep Neural Networks)는 장단기 메모리(Long Short-Term Memory, LSTM)일 수 있으나, 상기 제1 딥 뉴럴 네트워크(Deep Neural Networks)가 이에 한정되는 것은 아니다. 또한, 상기 제1 아웃풋(output)은 비디오의 모든 구간에서 생성한 아웃풋일 수 있으나, 상기 제1 아웃풋(output)이 이에 한정되는 것은 아니다.

[0195] 일실시예에 따라 비디오 피처 인코더(500)는 학습에 의해 결정된 파라미터들을 제1 아웃풋(output)과 결합한 피처에 시그모이드 함수(Sigmoid)를 적용하여 템퍼럴 어텐션(temporal attention)(506)을 생성할 수 있다.

[0196] 일실시예에 따라 비디오 피처 인코더(500)는 비디오 피처와 템퍼럴 어텐션(temporal attention)(506)을 결합하여 템퍼럴 어텐션(temporal attention)(506)이 웨이트로 적용된 비디오 피처의 가중치 합(507)을 생성할 수 있다.

[0197] 일실시예에 따라 캡션 디코더(520)는 템퍼럴 어텐션(temporal attention)(506)이 웨이트로 적용된 비디오 피처의 가중치 합(507) 및 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처(예컨대, A, man, is, talking 등에 대한 피처)에 미리 설정된 함수를 적용하여 어텐션(521)을 생성할 수 있다. 이때, 제2 딥 뉴럴 네트워크(Deep Neural Networks)는 장단기 메모리(Long Short-Term Memory, LSTM)일 수 있으나, 제2 딥 뉴럴 네트워크(Deep Neural Networks)가 이에 한정되는 것은 아니다.

[0198] 일실시예에 따라 캡션 디코더(520)는 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처(예컨대, A, man, is, talking 등에 대한 피처)와 템퍼럴 어텐션(temporal attention)(506)이 웨이트로 적용된 비디오 피처의 가중치 합(507)을 결합(concatenate)할 수 있다.

[0199] 일실시예에 따라 캡션 디코더(520)는 학습에 의해 결정된 파라미터들을 상기 결합(concatenate)한 상기 템퍼럴 어텐션(temporal attention)(506)이 웨이트로 적용된 비디오 피처의 가중치 합(507) 및 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))의 이전 히든 스테이트(hidden state)가 출력한 피처(예컨대, A, man, is, talking 등에 대한 피처)와 결합한 피처에 시그모이드 함수(Sigmoid)를 적용하여 어텐션(attention)(521)을 생성할 수 있다.

[0200] 일실시예에 따라 캡션 디코더(520)는 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))의 이전 히든 스테이트(hidden state)가 출력한 피처(예컨대, A, man, is, talking 등에 대한 피처)의 수만큼 어텐션(attention)(521)을 반복적으로 생성할 수 있다.

[0201] 일실시예에 따라 캡션 디코더(520)는 템퍼럴 어텐션(temporal attention)(506)이 웨이트로 적용된 비디오 피처의 가중치 합(507)과 상기 어텐션(attention)(521)을 결합하여 워드 어텐디드 피처(word attended features)를 생성할 수 있다.

[0202] 일실시예에 따라 캡션 디코더(520)는 템퍼럴 어텐션(temporal attention)(506)이 웨이트로 적용된 비디오 피처의 가중치 합(507)과 어텐션(attention)(521)을 결합하기 위하여 스칼라 곱(dot product)을 이용할 수 있으나, 캡션 디코더(520)가 템퍼럴 어텐션(temporal attention)(506)이 웨이트로 적용된 비디오 피처의 가중치 합(507)과 어텐션(attention)(521)을 결합하기 위하여 사용하는 방법이 이에 한정되는 것은 아니다.

[0203] 일실시예에 따라 캡션 디코더(520)는 생성한 어텐션(attention)(521)의 수만큼 템퍼럴 어텐션(temporal attention)(506)이 웨이트로 적용된 비디오 피처의 가중치 합(507)과 어텐션(attention)(521)을 결합하여 워드 어텐디드 피처(word attended features)를 반복적으로 생성할 수 있다.

[0204] 일실시예에 따라 캡션 디코더(520)는 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))를 이용하여 워드 어텐디드 피처(word attended features) 및 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처(예컨대, A, man, is, talking 등에 대한 피처)를 디코딩하여 제2 아웃풋(output)을 생성(522)할 수 있다.

[0205] 일실시예에 따라 캡션 디코더(520)는 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))를 이용하여 상기 워드 어텐디드 피처(word attended features) 및 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테이트(hidden state)가 출력한 피처(예컨대, A, man, is, talking 등에 대한 피처)를 디코딩하여 제2 아웃풋(output)을 반복적으로 생성(522)할 수 있다.

[0206] 일실시예에 따라 캡션 디코더(520)는 상기 워드 어텐디드 피처(word attended features)의 수만큼 제2 딥 뉴럴 네트워크(Deep Neural Networks)(예컨대, 장단기 메모리(Long Short-Term Memory, LSTM))를 이용하여 상기 워드 어텐디드 피처(word attended features) 및 제2 딥 뉴럴 네트워크(Deep Neural Networks)의 이전 히든 스테

이트(hidden state)가 출력한 피쳐(예컨대, A, man, is, talking 등에 대한 피쳐)를 디코딩하여 제2 아웃풋(output)을 반복적으로 생성(522)할 수 있다.

[0207] 도 6은 일실시예에 따라 비디오 캡셔닝(captioning) 장치가 비디오 캡셔닝(video captioning)을 수행하는 모습을 나타내는 모습이다.

[0208] 도 6을 참조하면, 일실시예에 따라 비디오 캡셔닝(captioning) 장치는 구간별로 구분되는 비디오(610)에 해당하는 비디오 피쳐를 생성할 수 있다. 이때, 상기 비디오 피쳐는 구간별로 구분된 비디오 피쳐 일 수 있으나, 상기 비디오 피쳐가 이에 한정되는 것은 아니다.

[0209] 일실시예에 따라 비디오 캡셔닝(captioning) 장치는 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 상기 구간별로 구분된 비디오 피쳐를 인코딩 하여 아웃풋을 생성할 수 있다.

[0210] 일실시예에 따라 비디오 캡셔닝(captioning) 장치는 상기 아웃풋을 기초로 템퍼럴 어텐션(temporal attention)을 생성할 수 있다.

[0211] 일실시예에 따라 비디오 캡셔닝(captioning) 장치는 상기 템퍼럴 어텐션(temporal attention)을 기초로 상기 구간별로 구분된 비디오 피쳐의 시간에 따른 중요도의 변화를 결정할 수 있다.

[0212] 일실시예에 따라 비디오 캡셔닝(captioning) 장치는 상기 템퍼럴 어텐션(temporal attention)을 기초로 상기 구간별로 구분된 비디오 피쳐의 구간별 중요도를 결정할 수 있다.

[0213] 일실시예에 따라 비디오 캡셔닝(captioning) 장치는 상기 템퍼럴 어텐션(temporal attention)을 기초로 동일한 구간에 속한 비디오 피쳐의 상기 동일한 구간 내에서의 시간에 따른 중요도를 결정할 수 있다.

[0214] 일실시예에 따라 비디오 캡셔닝(captioning) 장치는 비디오 피쳐와 매칭되는 상기 템퍼럴 어텐션(temporal attention)을 디스플레이(600) 할 수 있다.

[0215] 일실시예에 따라 비디오 캡셔닝(captioning) 장치는 상기 템퍼럴 어텐션(temporal attention)을 기초로 결정한 상기 비디오 피쳐의 중요도에 따라 상기 템퍼럴 어텐션(temporal attention)이 디스플레이 되는 색을 변화시킬 수 있다.

[0216] 일실시예에 따라 비디오 캡셔닝(captioning) 장치는 상기 템퍼럴 어텐션(temporal attention)을 기초로 결정한 상기 비디오 피쳐의 중요도가 높은 경우, 상기 비디오 피쳐와 매칭되는 상기 템퍼럴 어텐션(temporal attention)을 미리 설정된 제1 색(예컨대, 붉은색) 계열의 색으로 디스플레이(601, 602, 603) 할 수 있다.

[0217] 일실시예에 따라 비디오 캡셔닝(captioning) 장치는 상기 템퍼럴 어텐션(temporal attention)을 기초로 결정한 상기 비디오 피쳐의 중요도가 중간인 경우, 상기 비디오 피쳐와 매칭되는 상기 템퍼럴 어텐션(temporal attention)을 미리 설정된 제2 색(예컨대, 노랑색) 계열의 색으로 디스플레이(604, 605) 할 수 있다.

[0218] 일실시예에 따라 비디오 캡셔닝(captioning) 장치는 상기 템퍼럴 어텐션(temporal attention)을 기초로 결정한 상기 비디오 피쳐의 중요도가 낮은 경우, 상기 비디오 피쳐와 매칭되는 상기 템퍼럴 어텐션(temporal attention)을 미리 설정된 제3 색(예컨대, 푸른색) 계열의 색으로 디스플레이(606, 607) 할 수 있다.

[0219] 일실시예에 따라 비디오 캡셔닝(captioning) 장치는 상기 비디오 피쳐 및 상기 템퍼럴 어텐션(temporal attention)을 기초로 생성한 문장(620)을 디스플레이 할 수 있다.

[0220] 도 7은 본 발명의 일실시예를 구현하기 위한 예시적인 컴퓨터 시스템의 블록도이다.

[0221] 도 7을 참조하면, 본 발명의 일실시예를 구현하기 위한 예시적인 컴퓨터 시스템은 정보를 교환하기 위한 버스 또는 다른 커뮤니케이션 채널(701)을 포함하고, 프로세서(702)는 정보를 처리하기 위하여 버스(701)와 연결된다.

[0222] 컴퓨터 시스템(700)은 정보 및 프로세서(702)에 의해 처리되는 명령들을 저장하기 위하여 버스(701)와 연결된 RAM(Random Access Memory) 또는 다른 동적 저장 장치인 메인 메모리(703)를 포함한다.

[0223] 또한, 메인 메모리(703)는 프로세서(702)에 의한 명령들의 실행동안 임시변수들 또는 다른 중간 정보를 저장하기 위해 사용될 수 있다.

[0224] 컴퓨터 시스템(700)은 프로세서(702)에 대한 정적인 정보 또는 명령들을 저장하기 위하여 버스(701)에 결합된 ROM(Read Only Memory) 및 다른 정적 저장장치(704)를 포함할 수 있다.

[0225] 마그네틱 디스크, 집(zip) 또는 광 디스크 같은 대량 저장장치(705) 및 그것과 대응하는 드라이브 또한 정보 및 명령들을 저장하기 위하여 컴퓨터 시스템(700)에 연결될 수 있다.

[0226] 컴퓨터 시스템(700)은 엔드 유저(end user)에게 정보를 디스플레이 하기 위하여 버스(701)를 통해 음극선관 또는 엘씨디 같은 디스플레이 장치(710)와 연결될 수 있다.

[0227] 키보드(720)와 같은 문자 입력 장치는 프로세서(702)에 정보 및 명령을 전달하기 위하여 버스(701)에 연결될 수 있다.

[0228] 다른 유형의 사용자 입력 장치는 방향 정보 및 명령 선택을 프로세서(702)에 전달하고, 디스플레이(710) 상의 커서의 움직임을 제어하기 위한 마우스, 트랙볼 또는 커서 방향 키들과 같은 커서 컨트롤 장치(730)이다.

[0229] 통신 장치(740) 역시 버스(701)와 연결된다.

[0230] 통신 장치(740)는 지역 네트워크 또는 광역망에 접속되는 것을 서포트 하기 위하여 모뎀, 네트워크 인터페이스 카드, 이더넷, 토큰 링 또는 다른 유형의 물리적 결합물과 연결하기 위해 사용되는 인터페이스 장치를 포함할 수 있다. 이러한 방식으로 컴퓨터 시스템(700)은 인터넷 같은 종래의 네트워크 인프라 스트럭쳐를 통하여 다수의 클라이언트 및 서버와 연결될 수 있다.

[0231] 여기서 사용된 '장치'라는 용어는 논리적인 구성 단위를 나타내는 것으로서, 반드시 물리적으로 구분되는 구성요소가 아니라는 점은 본 발명이 속하는 기술분야의 당업자에게 자명한 사항이다.

[0232] 이상에서, 본 발명의 실시예를 구성하는 모든 구성 요소들이 하나로 결합되거나 결합되어 동작하는 것으로 설명되었다고 해서, 본 발명이 반드시 이러한 실시예에 한정되는 것은 아니다. 즉, 본 발명의 목적 범위 안에서라면, 그 모든 구성 요소들이 적어도 하나로 선택적으로 결합하여 동작할 수도 있다.

[0233] 또한, 그 모든 구성 요소들이 각각 하나의 독립적인 하드웨어로 구현될 수 있지만, 각 구성 요소들의 그 일부 또는 전부가 선택적으로 조합되어 하나 또는 복수 개의 하드웨어에서 조합된 일부 또는 전부의 기능을 수행하는 프로그램 모듈을 갖는 컴퓨터 프로그램으로서 구현될 수도 있다. 그 컴퓨터 프로그램을 구성하는 코드들 및 코드 세그먼트들은 본 발명의 기술 분야의 당업자에 의해 용이하게 추론될 수 있을 것이다.

[0234] 이러한 컴퓨터 프로그램은 컴퓨터가 읽을 수 있는 저장매체(Computer Readable Media)에 저장되어 컴퓨터에 의하여 읽혀지고 실행됨으로써, 본 발명의 실시예를 구현할 수 있다. 컴퓨터 프로그램의 저장매체로서는 자기 기록매체, 광 기록매체, 등이 포함될 수 있다.

[0235] 또한, 이상에서 기재된 "포함하다", "구성하다" 또는 "가지다" 등의 용어는, 특별히 반대되는 기재가 없는 한, 해당 구성 요소가 내재될 수 있음을 의미하는 것이므로, 다른 구성 요소를 제외하는 것이 아니라 다른 구성 요소를 더 포함할 수 있는 것으로 해석되어야 한다.

[0236] 기술적이거나 과학적인 용어를 포함한 모든 용어들은, 다르게 정의되지 않는 한, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 사전에 정의된 용어와 같이 일반적으로 사용되는 용어들은 관련 기술의 문맥 상의 의미와 일치하는 것으로 해석되어야 하며, 본 발명에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0237] 본 발명에서 개시된 방법들은 상술된 방법을 달성하기 위한 하나 이상의 동작들 또는 단계들을 포함한다. 방법 동작들 및/또는 단계들은 청구항들의 범위를 벗어나지 않으면서 서로 상호 교환될 수도 있다. 다시 말해, 동작들 또는 단계들에 대한 특정 순서가 명시되지 않는 한, 특정 동작들 및/또는 단계들의 순서 및/또는 이용은 청구항들의 범위로부터 벗어남이 없이 수정될 수도 있다.

[0238] 본 발명에서 이용되는 바와 같이, 아이템들의 리스트 중 "그 중 적어도 하나" 를 지칭하는 구절은 단일 멤버들을 포함하여, 이들 아이템들의 임의의 조합을 지칭한다. 일 예로서, "a, b, 또는 c: 중의 적어도 하나" 는 a, b, c, a-b, a-c, b-c, 및 a-b-c 뿐만 아니라 동일한 엘리먼트의 다수의 것들과의 임의의 조합 (예를 들어, a-a, a-a-a, a-a-b, a-a-c, a-b-b, a-c-c, b-b, b-b-b, b-b-c, c-c, 및 c-c-c 또는 a, b, 및 c 의 다른 임의의 순서화한 것) 을 포함하도록 의도된다.

[0239] 본 발명에서 이용되는 바와 같이, 용어 "결정하는"는 매우 다양한 동작들을 망라한다. 예를 들어, "결정하는"는 계산하는, 컴퓨팅, 프로세싱, 도출하는, 조사하는, 룩업하는 (예를 들어, 테이블, 데이터베이스, 또는 다른 데이터 구조에서 룩업하는), 확인하는 등을 포함할 수도 있다. 또한, "결정하는"은 수신하는 (예를 들면, 정보를 수신하는), 액세스하는 (메모리의 데이터에 액세스하는) 등을 포함할 수 있다. 또한, "결정하는"은 해결하는,

선택하는, 고르는, 확립하는 등을 포함할 수 있다.

[0240] 이상의 설명은 본 발명의 기술 사상을 예시적으로 설명한 것에 불과한 것으로서, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자라면 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 다양한 수정 및 변형이 가능할 것이다.

[0241] 따라서, 본 발명에 개시된 실시예들은 본 발명의 기술 사상을 한정하기 위한 것이 아니라 설명하기 위한 것이고, 이러한 실시예에 의하여 본 발명의 기술 사상의 범위가 한정되는 것은 아니다. 본 발명의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 발명의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

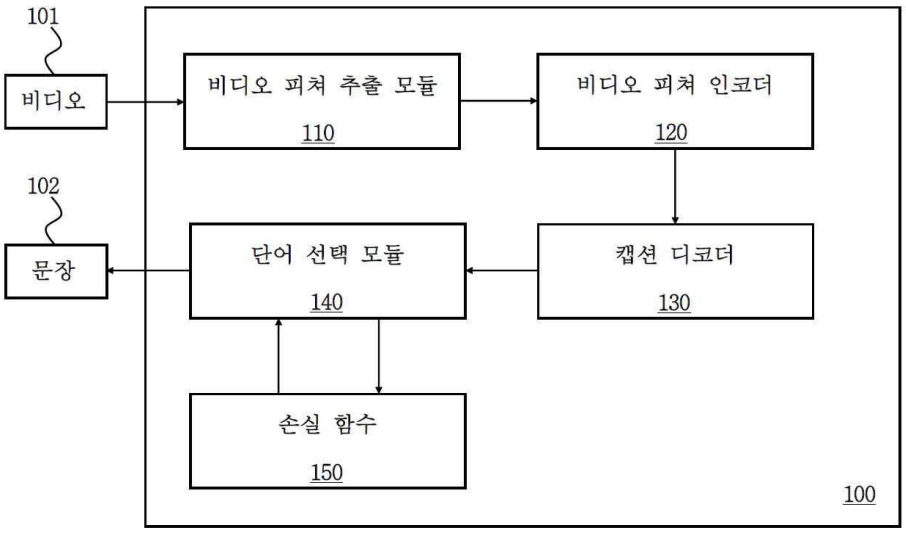
부호의 설명

[0242] 100... 비디오 캡셔닝(captioning) 학습 장치

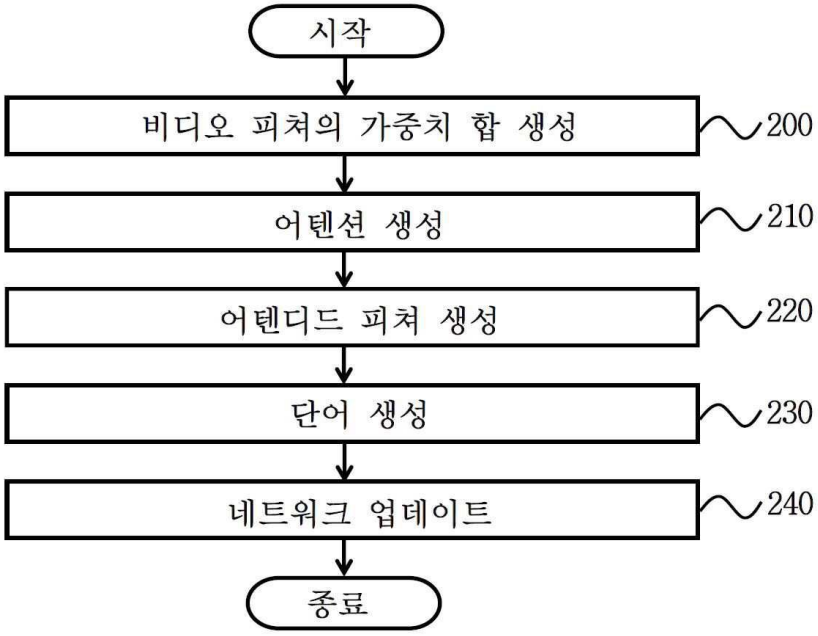
200... 비디오 캡셔닝(captioning) 장치

도면

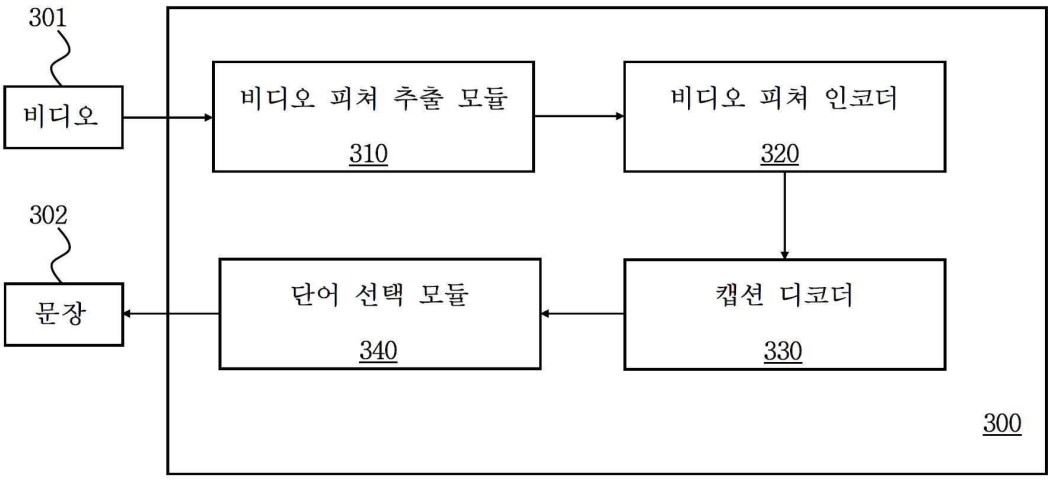
도면1



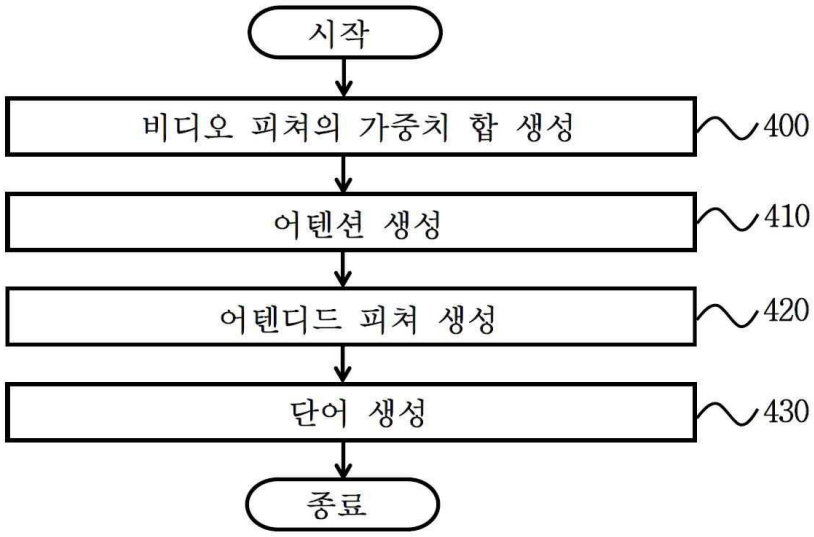
도면2


시작
비디오 피처의 가중치 합 생성
200
어텐션 생성
210
어텐디드 피처 생성
220
단어 생성
230
네트워크 업데이트
240
종료

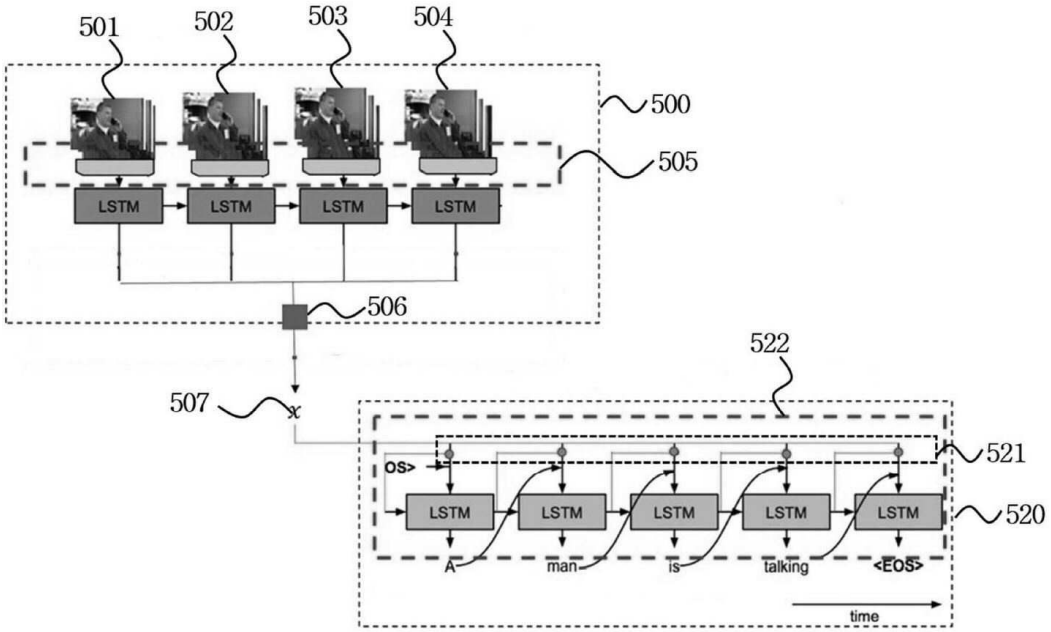
도면3


301
비디오
비디오 피처 추출 모듈
310
비디오 피처 인코더
320
302
문장
단어 선택 모듈
340
캡션 디코더
330
300

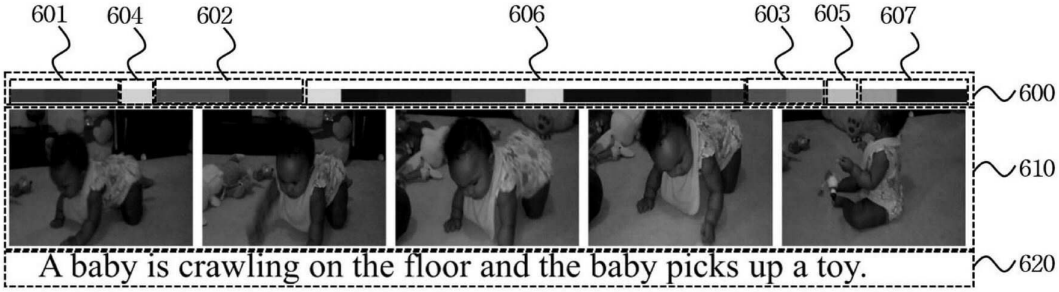
도면4


시작
비디오 피쳐의 가중치 합 생성
400
어텐션 생성
410
어텐디드 피쳐 생성
420
단어 생성
430
종료

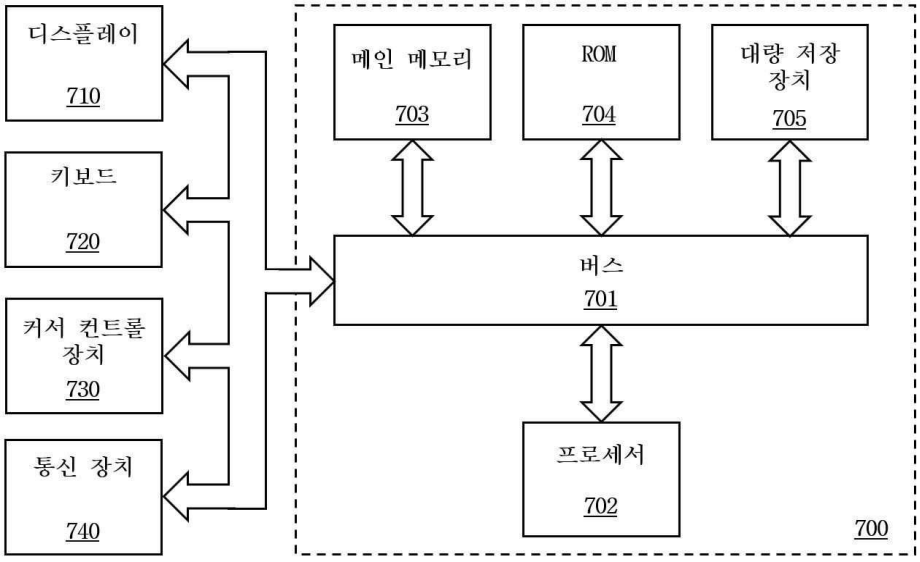
도면5


501
502
503
504
500
505
LSTM
506
507
522
521
520
<OS>
A
man
is
talking
<EOS>
time

도면6


601
604
602
606
603
605
607
600
610
620
A baby is crawling on the floor and the baby picks up a toy.

도면7


디스플레이
710
키보드
720
커서 컨트롤 장치
730
통신 장치
740
메인 메모리
703
ROM
704
대량 저장 장치
705
버스
701
프로세서
702
700