(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0121507
(43) 공개일자 2023년08월18일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2023.01) *G06N 3/04* (2023.01)
(52) CPC특허분류
G06N 3/08 (2023.01)
G06N 3/047 (2023.01)
(21) 출원번호 10-2022-0018494
(22) 출원일자 2022년02월11일
심사청구일자 2022년02월11일

(71) 출원인
세종대학교산학협력단
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
(72) 발명자
정철
서울특별시 광진구 능동로 209 세종대학교
헨드리아 윌리 피트라
서울특별시 광진구 능동로 209 세종대학교
(74) 대리인
김현승

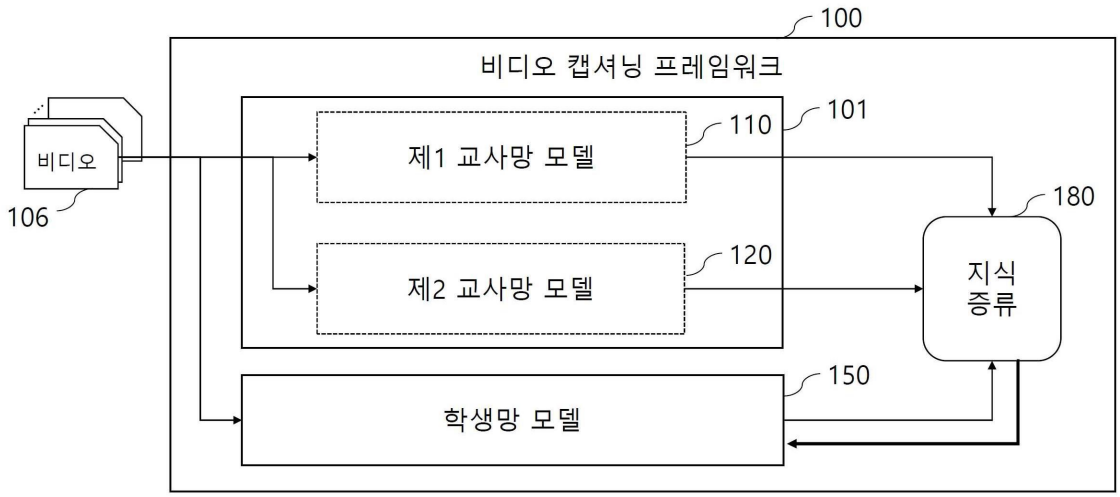
전체 청구항 수 : 총 17 항

(54) 발명의 명칭 **그래프 기반 비디오 캡셔닝을 위한 지식 증류**

(57) 요 약

그래프 기반 비디오 캡셔닝을 위한 지식 증류의 방법이 제공된다. 방법은, 적어도 하나의 교사망 모델 및 학생망 모델에 입력 비디오를 제공하는 단계와, 적어도 하나의 교사망 모델을 통해 입력 비디오에 대해 그래프 기반 비디오 캡셔닝을 수행하여 적어도 하나의 교사망 출력을 생성하는 단계(적어도 하나의 교사망 모델은 입력 비디오에 대해 액션 그래프 기반 비디오 캡셔닝을 수행하여 제1 교사망 출력을 생성하는 제1 교사망 모델 및 입력 비디오에 대해 술어 그래프 기반 비디오 캡셔닝을 수행하여 제2 교사망 출력을 생성하는 제2 교사망 모델 중 적어도 하나를 포함함)와, 학생망 모델을 통해 입력 비디오에 대해 비디오 캡셔닝을 수행하여 학생망 출력을 생성하는 단계와, 적어도 하나의 교사망 출력 및 학생망 출력에 기반하여 총 증류 손실을 계산하는 단계와, 계산된 총 증류 손실에 기반하여 학생망 모델을 훈련시키는 단계를 포함한다.

대 표 도



이 발명을 지원한 국가연구개발사업

과제고유번호	1711139507
과제번호	2021-0-02067-001
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	차세대인공지능핵심원천기술개발(R&D)
연구과제명	다목적 비디오 검색을 위한 차세대 인공신경망 기술 개발
기 여 율	1/1
과제수행기관명	세종대학교 산학협력단
연구기간	2021.07.01 ~ 2022.03.31

명 세 서

청구범위

청구항 1

그래프 기반 비디오 캡셔닝(video captioning)을 위한 지식 증류(knowledge distillation)의 방법으로서,

적어도 하나의 교사망 모델(teacher network model) 및 학생망 모델(student network model)에 비디오를 입력하는 단계와,

상기 적어도 하나의 교사망 모델을 통해, 상기 입력된 비디오에 대해 그래프 기반 비디오 캡셔닝을 수행하여 적어도 하나의 교사망 출력을 생성하는 단계 - 상기 적어도 하나의 교사망 모델은 상기 입력된 비디오에 대해 액션(action) 그래프 기반 비디오 캡셔닝을 수행하여 제1 교사망 출력을 생성하는 제1 교사망 모델 및 상기 입력된 비디오에 대해 술어(predicate) 그래프 기반 비디오 캡셔닝을 수행하여 제2 교사망 출력을 생성하는 제2 교사망 모델 중 적어도 하나를 포함함 - 와,

상기 학생망 모델을 통해, 상기 입력된 비디오에 대해 비디오 캡셔닝을 수행하여 학생망 출력을 생성하는 단계와,

상기 적어도 하나의 교사망 출력 및 상기 학생망 출력에 기반하여 총 증류 손실을 계산하는 단계와,

상기 계산된 총 증류 손실에 기반하여 상기 학생망 모델을 훈련시키는 단계를 포함하는

방법.

청구항 2

제1항에 있어서,

상기 적어도 하나의 교사망 출력 및 상기 학생망 출력 각각은 상기 입력된 비디오를 위한 각자의 후보 캡션을 예측된 단어 확률의 세트의 시퀀스(sequence)로서 나타내는 소프트 레이블(soft label) 예측을 포함하는,

방법.

청구항 3

제1항에 있어서,

상기 적어도 하나의 교사망 모델은 각자의 상이한 그래프 기반 비디오 캡셔닝을 수행하여 복수의 교사망 출력을 생성하는 복수의 교사망 모델이고, 상기 복수의 교사망 모델은 상기 제1 교사망 모델 및 상기 제2 교사망 모델 중 상기 적어도 하나를 포함하고, 상기 증류 손실을 계산하는 단계는,

상기 복수의 교사망 출력을 각각 상기 학생망 출력과 비교하여 복수의 증류 손실을 계산하는 단계와,

상기 복수의 교사망 출력으로부터 복수의 성능 메트릭을 계산하는 단계 - 상기 복수의 성능 메트릭 각각은 상기 복수의 교사망 모델 중 각자의 것의 성능 메트릭임 - 와,

상기 총 증류 손실을 상기 복수의 증류 손실의 가중 합(weighted sum)으로서 계산하기 위해, 상기 복수의 증류 손실을 각각 상기 복수의 성능 메트릭에 기반하여 가중화하는 단계를 포함하는,

방법.

청구항 4

제3항에 있어서,

상기 복수의 성능 메트릭 각각은 상기 복수의 교사망 모델 중 각자의 것으로부터 출력된 후보 캡션 및 상기 입력된 비디오에 대해 주어진 기준 캡션 간의 대응관계 또는 유사성을 측정하는,

방법.

청구항 5

제1항에 있어서,

상기 액션 그래프 기반 비디오 캡셔닝을 수행하여 상기 제1 교사망 출력을 생성하는 것은,

장면을 나타내는, 상기 입력 비디오의 프레임의 시퀀스로부터, 상기 장면 내에 나타나는 검출된 객체 간의 인식된 액션을 나타내는 액션 그래프를 구축하는 것과,

상기 액션 그래프에 기반하여 상기 제1 교사망 출력을 생성하는 것을 포함하는,

방법.

청구항 6

제5항에 있어서,

상기 액션 그래프를 구축하는 것은, 상기 인식된 액션을 다수의 인식가능한 액션에 걸쳐 정해진, 예측된 액션 확률의 세트로서 나타내는 다차원(multi-dimensional) 특징을 상기 액션 그래프의 엣지(edge)에 할당하기 위해, 상기 다수의 인식가능한 액션으로의 분류를 수행하도록 사전훈련된 액션 인식 모델을 사용하는 것을 포함하고,

상기 액션 그래프에 기반하여 상기 제1 교사망 출력을 생성하는 것은, 상기 입력된 비디오의 특징 표현의 시퀀스를 추출하기 위해 상기 액션 그래프에 그래프 신경망(graph neural network)을 적용하는 것과, 상기 특징 표현의 상기 시퀀스로부터 상기 제1 교사망 출력을 생성하기 위해 시퀀스 대 시퀀스 모델(sequence-to-sequence model)을 사용하는 것을 포함하는,

방법.

청구항 7

제1항에 있어서,

상기 술어 그래프 기반 비디오 캡셔닝을 수행하여 상기 제2 교사망 출력을 생성하는 것은,

장면을 나타내는, 상기 입력 비디오의 프레임으로부터, 상기 장면 내에 나타나는 객체의 쌍의 검출된 시각적 관계를 나타내는 술어 그래프를 구축하는 것과,

상기 술어 그래프에 기반하여 상기 제2 교사망 출력을 생성하는 것을 포함하는,

방법.

청구항 8

제7항에 있어서,

상기 술어 그래프를 구축하는 것은, 상기 검출된 시각적 관계를 다수의 검출가능한 시각적 관계에 걸쳐 정해진, 예측된 술어 확률의 세트로서 나타내는 다차원 특징을 상기 술어 그래프의 엣지에 할당하기 위해, 상기 다수의 검출가능한 시각적 관계로의 분류를 수행하도록 사전훈련된 시각적 관계 검출 모델을 사용하는 것을 포함하고,

상기 술어 그래프에 기반하여 상기 제2 교사망 출력을 생성하는 것은, 상기 입력된 비디오의 특징 표현의 시퀀스를 추출하기 위해 상기 술어 그래프에 그래프 신경망을 적용하는 것과, 상기 특징 표현의 상기 시퀀스로부터 상기 제2 교사망 출력을 생성하기 위해 시퀀스 대 시퀀스 모델을 사용하는 것을 포함하는,

방법.

청구항 9

그래프 기반 비디오 캡셔닝을 위한 지식 증류의 장치로서,

프로세서와,

상기 프로세서에 의해 실행되는 경우에 상기 장치로 하여금 동작을 수행하게 하는 명령어가 저장된 컴퓨터 판독가능 저장 매체를 포함하되, 상기 동작은,

적어도 하나의 교사망 모델 및 학생망 모델에 비디오를 입력하는 것과,

상기 적어도 하나의 교사망 모델을 통해, 상기 입력된 비디오에 대해 그래프 기반 비디오 캡셔닝을 수행하여 적어도 하나의 교사망 출력을 생성하는 것 - 상기 적어도 하나의 교사망 모델은 상기 입력된 비디오에 대해 액션 그래프 기반 비디오 캡셔닝을 수행하여 제1 교사망 출력을 생성하는 제1 교사망 모델 및 상기 입력된 비디오에 대해 술어 그래프 기반 비디오 캡셔닝을 수행하여 제2 교사망 출력을 생성하는 제2 교사망 모델 중 적어도 하나를 포함함 - 과,

상기 학생망 모델을 통해, 상기 입력된 비디오에 대해 비디오 캡셔닝을 수행하여 학생망 출력을 생성하는 것과,

상기 적어도 하나의 교사망 출력 및 상기 학생망 출력에 기반하여 총 증류 손실을 계산하는 것과,

상기 계산된 총 증류 손실에 기반하여 상기 학생망 모델을 훈련시키는 것을 포함하는,

장치.

청구항 10

제9항에 있어서,

상기 적어도 하나의 교사망 출력 및 상기 학생망 출력 각각은 상기 입력된 비디오를 위한 각자의 후보 캡션을 예측된 단어 확률의 세트의 시퀀스로서 나타내는 소프트 레이블 예측을 포함하는,

장치.

청구항 11

제9항에 있어서,

상기 적어도 하나의 교사망 모델은 각자의 상이한 그래프 기반 비디오 캡셔닝을 수행하여 복수의 교사망 출력을 생성하는 복수의 교사망 모델이고, 상기 복수의 교사망 모델은 상기 제1 교사망 모델 및 상기 제2 교사망 모델 중 상기 적어도 하나를 포함하고, 상기 증류 손실을 계산하는 것은,

상기 복수의 교사망 출력을 각각 상기 학생망 출력과 비교하여 복수의 증류 손실을 계산하는 것과,

상기 복수의 교사망 출력으로부터 복수의 성능 메트릭을 계산하는 것 - 상기 복수의 성능 메트릭 각각은 상기 복수의 교사망 모델 중 각자의 것의 성능 메트릭임 - 과,

상기 총 증류 손실을 상기 복수의 증류 손실의 가중 합으로서 계산하기 위해, 상기 복수의 증류 손실을 각각 상기 복수의 성능 메트릭에 기반하여 가중화하는 것을 포함하는,

장치.

청구항 12

제11항에 있어서,

상기 복수의 성능 메트릭 각각은 상기 복수의 교사망 모델 중 각자의 것으로부터 출력된 후보 캡션 및 상기 입력된 비디오에 대해 주어진 기준 캡션 간의 대응관계 또는 유사성을 측정하는,

장치.

청구항 13

제9항에 있어서,

상기 액션 그래프 기반 비디오 캡셔닝을 수행하여 상기 제1 교사망 출력을 생성하는 것은,

장면을 나타내는, 상기 입력 비디오의 프레임의 시퀀스로부터, 상기 장면 내에 나타나는 검출된 객체 간의 인식된 액션을 나타내는 액션 그래프를 구축하는 것과,

상기 액션 그래프에 기반하여 상기 제1 교사망 출력을 생성하는 것을 포함하는,

장치.

청구항 14

제13항에 있어서,

상기 액션 그래프를 구축하는 것은, 상기 인식된 액션을 다수의 인식가능한 액션에 걸쳐 정해진, 예측된 액션 확률의 세트로서 나타내는 다차원 특징을 상기 액션 그래프의 엣지에 할당하기 위해, 상기 다수의 인식가능한 액션으로의 분류를 수행하도록 사전훈련된 액션 인식 모델을 사용하는 것을 포함하고,

상기 액션 그래프에 기반하여 상기 제1 교사망 출력을 생성하는 것은, 상기 입력된 비디오의 특징 표현의 시퀀스를 추출하기 위해 상기 액션 그래프에 그래프 신경망을 적용하는 것과, 상기 특징 표현의 상기 시퀀스로부터 상기 제1 교사망 출력을 생성하기 위해 시퀀스 대 시퀀스 모델을 사용하는 것을 포함하는,

장치.

청구항 15

제9항에 있어서,

상기 술어 그래프 기반 비디오 캡셔닝을 수행하여 상기 제2 교사망 출력을 생성하는 것은,

장면을 나타내는, 상기 입력 비디오의 프레임으로부터, 상기 장면 내에 나타나는 객체의 쌍의 검출된 시각적 관계를 나타내는 술어 그래프를 구축하는 것과,

상기 술어 그래프에 기반하여 상기 제2 교사망 출력을 생성하는 것을 포함하는,

장치.

청구항 16

제15항에 있어서,

상기 술어 그래프를 구축하는 것은 상기 검출된 시각적 관계를 다수의 검출가능한 시각적 관계에 걸쳐 정해진, 예측된 술어 확률의 세트로서 나타내는 다차원 특징을 상기 술어 그래프의 엣지에 할당하기 위해, 상기 다수의 검출가능한 시각적 관계로의 분류를 수행하도록 사전훈련된 시각적 관계 검출 모델을 사용하는 것을 포함하고,

상기 술어 그래프에 기반하여 상기 제2 교사망 출력을 생성하는 것은, 상기 입력된 비디오의 특징 표현의 시퀀스를 추출하기 위해 상기 술어 그래프에 그래프 신경망을 적용하는 것과, 상기 특징 표현의 상기 시퀀스로부터 상기 제2 교사망 출력을 생성하기 위해 시퀀스 대 시퀀스 모델을 사용하는 것을 포함하는,

장치.

청구항 17

컴퓨터 프로세서에 의해 실행되는 경우 상기 컴퓨터 프로세서로 하여금 제1항 내지 제8항 중 어느 한 항에 기재된 방법을 수행하게 하는 컴퓨터 실행가능 명령어가 저장된 컴퓨터 판독가능 저장 매체.

발명의 설명

기 술 분 야

[0001] 본 개시는 그래프 기반 비디오 캡셔닝(video captioning)을 위한 지식 증류(knowledge distillation)에 관한 것이다.

배 경 기 술

[0002] 비디오 캡셔닝은 비디오에서 일어나는 이벤트를 이해하여 비디오를 텍스트로써 자동적으로 기술하는 작업이다. 비디오 캡셔닝의 응용의 예는 인간-로봇 상호작용(interaction), 자동 비디오 자막처리(subtitling), 비디오 감시 시스템, 그리고 시각 장애인을 위해 그의 주변을 기술하는 것을 포함한다. 더 정확하고 상세한 캡션(caption)을 생성하는 데에는 비디오로부터 추출된 더 상세한 정보가 필요할 수 있다. 비디오로부터 추출된 묘사는 생성된 캡션의 품질에 상당한 영향을 미친다. 그러므로, 객체들을 이들 간의 상세한 관계와 함께 포착하는

(capture) 것이 비디오 캡셔닝에 중요하다. 다른 한편으로, 생성된 캡션은 정확하고 상세한 것뿐만 아니라, 몇몇 응용에서는 시기적절할 것이 요구될 수 있다. 느린 캡션 생성 프로세스는 사용자에게 문제를 야기할 수 있다. 아울러, 묘사를 생성하는 알고리즘이 복잡하면 캡션 생성 프로세스가 너무 느려질 수 있다. 따라서, 그러한 추론 국면에서 너무 알고리즘의 복잡도를 더하지 않는 것도 중요하다.

발명의 내용

해결하려는 과제

[0003] 그래프 기반 비디오 캡셔닝을 위한 지식 증류가 본 문서에 개시된다.

과제의 해결 수단

[0004] 예에서, 그래프 기반 비디오 캡셔닝(video captioning)을 위한 지식 증류(knowledge distillation)의 방법이 제공되는데, 방법은 다음을 포함한다: 적어도 하나의 교사망 모델(teacher network model) 및 학생망 모델(student network model)에 입력 비디오를 제공하는 단계; 적어도 하나의 교사망 모델을 통해 입력 비디오에 대해 그래프 기반 비디오 캡셔닝을 수행하여 적어도 하나의 교사망 출력을 생성하는 단계(적어도 하나의 교사망 모델은 입력 비디오에 대해 액션(action) 그래프 기반 비디오 캡셔닝을 수행하여 제1 교사망 출력을 생성하는 제1 교사망 모델 및 입력 비디오에 대해 술어(predicate) 그래프 기반 비디오 캡셔닝을 수행하여 제2 교사망 출력을 생성하는 제2 교사망 모델 중 적어도 하나를 포함함); 학생망 모델을 통해 입력 비디오에 대해 비디오 캡셔닝을 수행하여 학생망 출력을 생성하는 단계; 적어도 하나의 교사망 출력 및 학생망 출력에 기반하여 총 증류 손실을 계산하는 단계; 및 계산된 총 증류 손실에 기반하여 학생망 모델을 훈련시키는(training) 단계.

[0005] 전술된 개요는 상세한 설명에서 추가로 후술되는 몇몇 양상을 단순화된 형태로 소개하기 위해 제공된다. 이 개요는 청구된 주제(subject matter)의 중요 특징 또는 필수적 특징을 식별하도록 의도되지 않고, 청구된 주제의 범위를 정하는 데 사용되도록 의도되지도 않는다. 나아가, 청구된 주제는 본 명세서에서 논의되는 임의의 또는 모든 이점을 제공하는 구현에 한정되지 않는다.

발명의 효과

[0006] 본 개시에 따른 비디오 캡셔닝은 액션 그래프(action graph) 기반 비디오 캡셔닝 및/또는 술어 그래프(predicate graph) 기반 비디오 캡셔닝을 포함하는바, 개선된 성능을 갖는다.

[0007] 본 개시에 따르면, 엣지(edge)에 다차원 특징(multi-dimensional feature)이 할당된 그래프로써 객체 간의 상호작용이 표현될 수 있고, 이로부터 후보 캡션을 나타내는 예측이 획득될 수 있다.

[0008] 본 개시에 따르면, 비디오 캡셔닝을 위한 다양한 타입의 모델, 예컨대, 액션 인식 모델, 시각적 관계 검출 모델, 그래프 신경망(Graph Neural Network: GNN) 및 시퀀스 대 시퀀스(Sequence-to-Sequence: S2S) 모델에 적합한 동적 프레임워크(framework)가 제공된다.

[0009] 본 개시에 따르면, 서로 다른 특징을 갖는 여러 교사망 모델을 수반하는 지식 증류로써 학생망 모델을 훈련시키는 데에 이용가능한 비디오 캡셔닝 프레임워크가 제공된다.

도면의 간단한 설명

[0010] 도 1은 예시적인 비디오 캡셔닝 프레임워크를 보여주는 블록도이다.

도 2는 도 1의 제1 교사망 모델의 예를 보여주는 블록도이다.

도 3은 도 1의 제2 교사망 모델의 예를 보여주는 블록도이다.

도 4는 도 1의 학생망 모델의 예를 보여주는 블록도이다.

도 5는 도 2의 그래프 신경망(Graph Neural Network: GNN)에 의해 사용되는 액션 그래프의 예를 보여준다.

도 6은 도 3의 GNN에 의해 사용되는 술어 그래프의 예를 보여준다.

도 7은 도 1의 비디오 캡셔닝 프레임워크에서 지식 증류를 수행하기 위해 총 가중 증류 손실을 계산하는 것을 설명하기 위한 도면이다.

도 8은 도 1의 비디오 캡셔닝 프레임워크를 이용한 지식 증류를 위한 운영 환경의 예를 보여준다.

도 9는 그래프 기반 비디오 캡셔닝을 위한 지식 증류를 수행하는 예시적인 프로세스를 보여주는 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0011] 본 개시에서 사용되는 다양한 용어는 본 문서에서의 기능을 고려하여 상용 용어의 용어법으로부터 선택되는데, 이는 당업자의 의도, 준례, 또는 새로운 기술의 출현에 따라서 달리 인식될 수 있다. 특정한 사례에서, 몇몇 용어에는 상세한 설명에서 개진된 바와 같이 의미가 주어질 수 있다. 따라서, 본 문서에서 사용되는 용어는, 단순히 그 명칭에 의해서가 아니라, 본 개시의 맥락에서 그 용어가 갖는 의미와 일관되게 정의되어야 한다.

[0012] 본 문서에서 용어 "포함하다", "가지다" 등은 이후에 열거된 요소, 예컨대, 어떤 특징, 숫자, 단계, 동작, 구성요소, 정보 또는 이들의 조합의 존재를 명시하는 경우에 사용된다. 달리 표시되지 않는 한, 이런 용어 및 이의 변형은 다른 요소의 존재 또는 추가를 배제하도록 의도되지 않는다.

[0013] 본 문서에서 사용되는 바와 같이, 용어 "제1", "제2" 등은 몇 개의 서로 닮은 요소를 식별하도록 의도된다. 달리 기재되지 않는 한, 그러한 용어는 이들 요소의 또는 이들의 사용의 특정한 순서와 같은 한정을 부과하도록 의도된 것이 아니라, 단지 여러 요소를 따로따로 지칭하기 위해 사용된다. 예를 들면, 어떤 요소가 일례에서 용어 "제1"로써 참조될 수 있는 한편 동일한 요소가 다른 예에서 "제2" 또는 "제3"과 같은 상이한 서수로써 참조될 수 있다. 그러한 예에서, 이들 용어는 본 개시의 범위를 한정하지 않는 것이다. 또한, 여러 요소의 리스트에서 용어 "및/또는"을 사용하는 것은 열거된 항목 중 임의의 하나 또는 복수 개를 비롯하여 이들 항목의 모든 가능한 조합을 포함한다. 나아가, 단수 형태의 표현은 명확하게 달리 사용되지 않는 한, 복수 형태의 의미를 포함한다.

[0014] 첨부된 도면을 참조하여 본 개시의 소정의 예가 이제 상세히 기술될 것이다. 다만, 본 개시는 많은 상이한 형태로 체현될 수 있으며, 본 문서에 개진된 예에 한정되는 것으로 해석되어서는 안 된다. 오히려, 이들 예는 본 개시의 범위의 더 나은 이해를 제공하기 위해서 주어지는 것이다.

[0015] 도 1은 예시적인 비디오 캡셔닝 프레임워크(100)를 보여주는 블록도이다.

[0016] 도 1의 예에서, 비디오 캡셔닝 프레임워크(100)는 적어도 하나의 교사망 모델(가령, 아래에서 상세히 기술되는 바와 같은 제1 교사망 모델(110) 및/또는 제2 교사망 모델(120))을 포함하는 교사망 모델 세트(101) 및 학생망 모델(150)을 포함한다. 비디오 캡셔닝 프레임워크(100)의 지식 증류 모듈(180)은 교사망 모델 세트(101) 내의 하나 이상의 교사망 모델로부터의 비디오 캡셔닝 작업의 지식을 학생망 모델(150)에 증류하고, 증류된 지식으로써 학생망 모델(150)을 교사망 모델 세트(101)에 포함된 교사망 모델의 비디오 캡셔닝 작업을 모방하도록 훈련시키는 데에 사용된다. 다양한 예시적인 구현에서, 후술되는 바와 같이, 비디오 캡셔닝 프레임워크(100)는 단일 컴퓨팅 장치 상에 또는 여러 컴퓨팅 장치에 걸쳐 구현되고 컴퓨팅 장치를 통해 액세스가능하게(accessible) 될 수 있다.

[0017] 도시된 예에서, 비디오 캡셔닝 프레임워크(100)의 교사망 모델 세트(101) 내의 교사망 모델(가령, 제1 교사망 모델(110) 및 제2 교사망 모델(120) 중 적어도 하나) 및 학생망 모델(150)은 입력 비디오(106)를 수신하고, 이에 따라 각자의 출력을 제공한다. 예를 들어, 교사망 모델 세트(101) 내의 교사망 모델은 입력 비디오(106)에 대해 소정의 그래프를 기반으로 비디오 캡셔닝을 수행하여 교사망 출력을 생성할 수 있다. 또한, 학생망 모델(150)은 입력 비디오(106)에 대해 비디오 캡셔닝을 수행하여 학생망 출력을 생성할 수 있다.

[0018] 도시된 예에서, 교사망 모델 세트(101) 내의 적어도 하나의 교사망 모델로부터 제공되는 적어도 하나의 교사망 출력(가령, 제1 교사망 모델(110)로부터 제공되는 제1 교사망 출력 및/또는 제2 교사망 모델(120)로부터 제공되는 제2 교사망 출력) 및 학생망 모델(150)로부터 제공되는 학생망 출력 각각은 입력 비디오(106)를 위한 각자의 후보 캡션(즉, 임의의 후보 캡션에서 이용가능한 어휘(vocabulary)에 포함된 몇몇 단어의 특정 시퀀스(sequence), 예컨대, 문장(sentence))을 나타낸다. 몇몇 예시적인 구현에서, 이들 출력 각각은 해당 후보 캡션을 예측된 단어 확률의 세트(각각의 세트는 어휘 내의 모든 단어에 걸쳐 정해짐)의 시퀀스로서 나타낼 수 있다. 예를 들어, 만일 후보 캡션이 20개의 단어의 시퀀스인 경우, 후보 캡션은 어휘 내의 K개의 단어 각각에 대해 예측된 단어 확률의 적어도 20개의 세트의 시퀀스로서 표현될 수 있음(즉, 후보 캡션의 각각의 단어에 시퀀스 내의 상이한 세트가 대응할 수 있음)이 이해될 것이다. 이들 단어 확률 세트 각각은, 예컨대, 해당 망 모델(즉, 교사망 모델(110), 교사망 모델(120), 또는 다른 교사망 모델, 또는 학생망 모델(150))의 최종 층(이는 K개의 출력 유닛을 가짐)으로부터, 이른바 "온도" 하이퍼파라미터(hyperparameter) T가 있는 소프트맥스(softmax) 출

$$q_i = \frac{\exp(z_i/T)}{\sum_{j=1}^{K}\exp(z_j/T)}$$

력 에 따라 주어지는 확률 분포일 수 있다(여기서 i는 K개의 단어 중 어떤 것을 가리키는 인덱스이고, z_i 는 최종 층의 i번째 출력 유닛으로의 입력임). 특히, 그러한 망 모델 각각의 출력은,

$$q_i = \frac{\exp(z_i)}{\sum_{j=1}^{K}\exp(z_j)}$$

정규 소프트맥스 출력 으로서(즉, 하이퍼파라미터 T가 1로 설정됨) 산출된 단어 확률의 세트의 시퀀스로서 해당 후보 캡션을 나타내는 예측(이는 하드 예측(hard prediction)으로 지칭될 수 있음)을 포함하는 것 대신에, 또는 이에 더하여, 1보다 더 큰 하이퍼파라미터 T로써 부드러워진(softened) 소프트맥스 출력으로서 산출된 단어 확률의 세트의 시퀀스로서 해당 후보 캡션을 나타내는 소프트 레이블(soft label) 예측을 포함할 수 있다.

[0019] 도시된 예에서, 적어도 하나의 교사망 출력은 학생망 출력과 함께 지식 증류 모듈(180)에 의해 처리될 수 있다. 예를 들어, 소정의 손실 함수(loss function)(가령, 후술되는 제1 증류 손실, 제2 증류 손실, 총 가중 증류 손실(total weighted distillation loss), 또는 유사한 것)가 적어도 하나의 교사망 출력(가령, 제1 교사망 출력 및/또는 제2 교사망 출력)에 기반하여, 그리고 학생망 출력에 또한 기반하여 계산될 수 있고, 학생망 모델(150)을 훈련하는 데에 사용될 수 있다. 그러한 훈련은, 예컨대, 손실 함수를 최소화하는 방식으로, 학생망 모델(150)의 파라미터를 갱신하는 것을 포함할 수 있다.

[0020] 몇몇 예시적인 구현에서, 교사망 모델 세트(101)는 전술된 바와 같이 제1 교사망 모델(110) 및 제2 교사망 모델(120) 중 적어도 하나를 포함할 수 있다. 예를 들어, 제1 교사망 모델(110)은 입력 비디오(106)에 대해 액션 그래프 기반 비디오 캡셔닝을 수행하여 제1 교사망 출력을 생성할 수 있고, 제2 교사망 모델(110)은 입력 비디오(106)에 대해 술어 그래프 기반 비디오 캡셔닝을 수행하여 제2 교사망 출력을 생성할 수 있다. 제1 교사망 모델(110) 및 제2 교사망 모델(120)에 대한 더 상세한 설명이 아래에서 주어질 것이다. 선택적으로, 교사망 모델 세트(101)는 제1 교사망 모델(110) 및/또는 제2 교사망 모델(120)에 더하여 적어도 하나의 다른 교사망 모델(도시되지 않음)을 더 포함할 수 있는데, 그러한 추가적인 교사망 모델은 상이한 방식으로 그래프 기반 비디오 캡셔닝을 수행하여 자신의 교사망 출력을 생성할 수 있다.

[0021] 도 2는 제1 교사망 모델(110)의 예를 보여주는 블록도이다.

[0022] 도 2의 예에서, 제1 교사망 모델(110)은 객체 검출 모델(201), 액션 인식 모델(203), 그래프 신경망(Graph Neural Network: GNN)(205) 및 시퀀스 대 시퀀스(Sequence-to-Sequence: S2S) 모델(207)을 포함한다.

[0023] 도시된 예에서, 객체 검출 모델(201) 및 액션 인식 모델(203)은 장면(scene)을 나타내는, 입력 비디오(106)의 프레임의 시퀀스(sequence)로부터 액션 그래프를 구축하는 데에 사용된다. 예를 들어, 객체 검출 모델(201)은 입력 비디오(106)의 프레임에 대해 객체의 공간적 위치를 묘사하는 바운딩 박스(bounding box)를 설정하는 방식으로 객체를 검출할 수 있다. 이와 더불어, 액션 인식 모델(203)은 검출된 객체 간의 어떤 액션을 입력 비디오(106)의 몇몇 순차적 프레임으로부터 인식할 수 있다. 특정한 예에서, 액션 인식 모델(203)은 인간 객체(human object) 및 다른 인간 또는 비인간(non-human) 객체 간의 액션(가령, 악수를 하는 액션, 포옹을 하는 액션, 자전거를 타는 액션, 바이올린을 연주하는 액션, 접시를 닦는 액션, 또는 유사한 것)을 인식하도록 특정 데이터세트(가령, Kinetics-400)에 대해 사전훈련된 신경망(가령, 2차원 합성곱 신경망(Convolutional Neural Network: CNN), 3차원 CNN, 장시간 단기 메모리(Long Short-Term Memory: LSTM) 신경망 및/또는 다른 종류의 신경망)을 포함할 수 있고, 이에 따라 입력 비디오(106)에 의해 나타내어진 장면 내의 액션을 한정된 수의 인식가능한 액션(가령, Kinetics-400에 의해 커버되는(covered) 400개의 인간 액션 클래스) 중 하나로 분류할 수 있다.

[0024] 도시된 예에서, 객체 검출 모델(201)의 출력 및 액션 인식 모델(203)의 출력은 액션 그래프로 변환된다. 몇몇 예시적인 구현에서, 액션 그래프는 장면 내에 나타나는 검출된 객체(가령, 인간 객체 및 다른 객체) 간의 인식된 액션을 나타내도록 구성될 수 있다. 예를 들어, 검출된 객체를 나타내는 특징이 액션 그래프의 노드(node)에 할당될 수 있다. 또한, 인식된 액션을 나타내기 위해 액션 인식 모델(203)은 예측된 액션 확률의 세트를 출력할 수 있는데, 이 액션 확률 세트는 다수의 인식가능한 액션(가령, 액션 인식 모델(203)은 이들 액션으로의 분류를 수행하도록 사전훈련됨)에 걸쳐 정해진 확률 분포일 수 있다. 인식된 액션의 그러한 다차원(multi-dimensional) 표현은 액션 그래프의 엣지(edge) 특징으로서 설정될 수 있다. 도 5는 예시적인 액션 그래프(500)를 도시하는데, 이는 복수의 노드(이들 각각은 각자의 객체에 대응함) 및 이들 노드 중 두 개를 연결하는 엣지(이

들 각각은 각자의 액션에 대응함)를 포함한다. 구체적으로, 노드(501), 노드(502) 및 노드(503)는 각각 인간 객체 A1, 비인간 객체 A2 및 비인간 객체 A3에 대응하고, 엣지(512), 엣지(513) 및 엣지(523)는 각각 객체 A1 및 객체 A2 간의 액션, 객체 A1 및 객체 A3 간의 액션, 그리고 객체 A2 및 객체 A3 간의 액션에 대응한다. 도 5의 예에서, 만일 액션 인식 모델(203)이 데이터세트 Kinetics-400로써 사전훈련된 경우, 엣지(512)에는 400개의 타겟팅된(targeted) 인간 액션 각각의 예측된 확률을 나타내는 400차원의 영 아닌 벡터(non-zero vector)가 다차원 특징으로서 할당되고, 엣지(513)도 마찬가지인 반면에, 비인간 객체들 간의 액션은 Kinetics-400에 의해 추출되지 않으므로 엣지(523)에는 영 벡터(zero vector)가 다차원 특징으로서 할당된다.

[0025] 몇몇 예시적인 구현에서, 액션 그래프는 전술된 바와 같이 객체 특징(객체 검출 모델(201)로부터 출력됨) 및 액션 특징(액션 인식 모델(203)로부터 출력됨)에 기반하여 구성된 그래프(이는 어떤 시간 스텝에서 객체가 어떻게 연결되는지를 나타내는바, 공간적 그래프로 지칭될 수 있음)(가령, 그래프(500))에 그치지 않고, 시간적으로 확장될 수 있다. 예를 들어, 객체에 대응하는 노드 및 이들 노드를 연결하는 엣지를 갖는 시간적 그래프가 구성될 수 있는데, 이 엣지에는 어떤 시간 스텝에서 검출된 객체 및 다음 시간 스텝에서 검출된 동일한 또는 상이한 객체 간의 유사성을 나타내는 특징 값이 할당될 수 있다. 그러면, 액션 그래프는 공간적 그래프 및 시간적 그래프의 조합(즉, 시공간적 그래프(spatio-temporal graph))로서 구축될 수 있다.

[0026] 다시 도 2의 예로 돌아가면, 제1 교사망 출력은 다음과 같이 액션 그래프에 기반하여 생성된다. 우선, 액션 그래프에 GNN(205)이 적용되는데, 이는 입력 비디오(106)의 특징 표현(이는, 예컨대, 장면 내의 객체 및/또는 액션을 나타냄)의 시퀀스를 추출한다. 이후에, 추출된 시퀀스로부터 입력 비디오(106)를 위한 후보 캡션(즉, 단어의 시퀀스)을 나타내는 제1 교사망 출력(가령, 소프트 레이블)을 생성하는 데에 S2S 모델(207)이 사용된다.

[0027] 도 3은 제2 교사망 모델(120)의 예를 보여주는 블록도이다.

[0028] 도 3의 예에서, 제2 교사망 모델(120)은 시각적 관계 검출 모델(301), GNN(305) 및 S2S 모델(307)을 포함한다.

[0029] 도시된 예에서, 시각적 관계 검출 모델(301)은 장면을 나타내는, 입력 비디오(106)의 프레임으로부터 술어 그래프를 구축하는 데에 사용된다. 예를 들어, 시각적 관계 검출 모델(301)은 입력 비디오(106)의 단일 프레임으로부터 장면 내의 객체의 쌍을 이들 객체 간의 시각적 관계와 함께, 예컨대, "<객체 A - 술어 - 객체 B>"의 형태로, 검출할 수 있다. 술어의 예는 어떤 객체에 대한 다른 객체의 동작을 나타내는 것(가령, "착용하다")은 물론 객체 간의 다른 상호작용, 예컨대, 비교(가령, "더 크다"), 배치관계(가령, "옆에 있다") 등을 나타내는 것도 포함한다. 만일 주어진 비디오 프레임이 헬멧을 쓴 사람이 이륜 오토바이를 주행하는 장면을 나타내는 경우, <사람 - 타다 - 오토바이>, <사람 - 착용하다 - 헬멧> 및 <오토바이 - 갖다 - 바퀴>와 같은 시각적 관계가 검출될 수 있다. 만일 주어진 비디오 프레임이 헬멧을 쓴 사람이 이륜 오토바이 곁에 서 있는 장면을 나타내는 경우, <사람 - 옆에 있다 - 오토바이>, <사람 - 착용하다 - 헬멧> 및 <오토바이 - 갖다 - 바퀴>와 같은 시각적 관계가 검출될 수 있다. 이와 같이, 제2 교사망 모델(120)은 단일한 시각적 관계 검출 모델(301)을 사용하여, 그리고 가능하게는 하나의 비디오 프레임에 기반하여, 객체의 검출 및 객체 간의 상호작용(이는, 특히, 비인간 객체 간의 관계도 포함함)의 검출을 수행할 수 있다는 점에서 제1 교사망 모델(110)과 상이하다. 특정한 예에서, 시각적 관계 검출 모델(301)은 객체 및 술어를 위한 시각적 외양 모델을 학습함(가령, 객체를 분류하기 위한 CNN 및 술어를 분류하기 위한 다른 CNN을 훈련시킴)으로써, 그리고 사전훈련된 단어 벡터를 사용하여 의미상 유사한 관계를 단어 임베딩 공간(word embedding space) 내에 가까이 있도록 최적화함으로써, 객체 간의 시각적 관계를 검출하도록 훈련될 수 있다.

[0030] 도시된 예에서, 시각적 관계 검출 모델(301)의 출력은 술어 그래프로 변환된다. 몇몇 예시적인 구현에서, 술어 그래프는 장면 내에 나타나는 객체의 쌍의 시각적 관계를 나타내도록 구성될 수 있다. 예를 들어, 검출된 객체를 나타내는 특징이 술어 그래프의 노드에 할당될 수 있다. 또한, 검출된 시각적 관계를 나타내기 위해 시각적 관계검출 모델(301)은 예측된 술어 확률의 세트를 출력할 수 있는데, 이 술어 확률 세트는 다수의 검출가능한 시각적 관계(가령, 시각적 관계 검출 모델(301)은 이들 시각적 관계로의 분류를 수행하도록 사전훈련됨)에 걸쳐 정해진 확률 분포일 수 있다. 검출된 시각적 관계의 그러한 다차원 표현은 술어 그래프의 엣지 특성으로서 설정될 수 있다. 도 6은 예시적인 술어 그래프(600)를 도시하는데, 이는 복수의 노드(이들 각각은 각자의 객체에 대응함) 및 이들 노드 중 두 개를 연결하는 엣지(이들 각각은 각자의 액션에 대응함)를 포함한다. 구체적으로, 노드(601), 노드(602) 및 노드(603)는 각각 객체 B1, 객체 B2 및 객체 B3에 대응하고, 엣지(612), 엣지(613) 및 엣지(614)는 객체 B1 및 객체 B2 간의 시각적 관계, 객체 B1 및 객체 B3 간의 시각적 관계, 그리고 객체 B2 및 객체 B3 간의 시각적 관계에 대응한다. 도 6에 도시된 바와 같이, 엣지(612), 엣지(613) 및 엣지(623)에는 대응하는 시각적 관계를 예측된 술어 확률의 세트로서 나타내는 다차원 특징이 할당된다.

[0031] 몇몇 예시적인 구현에서, 술어 그래프는 전술된 바와 같이 객체 특징 및 술어 특징(양자 모두 시각적 관계 검출 모델(301)로부터 출력됨)에 기반하여 구성된 공간적 그래프(가령, 그래프(600))에 그치지 않고, 공간적 그래프 및 시간적 그래프(여기에서 노드 간의 엣지에는 어떤 시간 스텝에서 검출된 객체 및 다음 시간 스텝에서 검출된 동일한 또는 상이한 객체 간의 유사성을 나타내는 특징 값이 할당될 수 있음)의 조합으로서 구축될 수 있다.

[0032] 다시 도 3의 예로 돌아가면, 제2 교사망 출력은 다음과 같이 술어 그래프에 기반하여 생성된다. 우선, 술어 그래프에 GNN(305)이 적용되는데, 이는 입력 비디오(106)의 특징 표현(이는, 예컨대, 장면 내의 객체 및/또는 술어를 나타냄)의 시퀀스를 추출한다. 이후에, 추출된 시퀀스로부터 입력 비디오(106)를 위한 후보 캡션(즉, 단어의 시퀀스)을 나타내는 제2 교사망 출력(가령, 소프트 레이블)을 생성하는 데에 S2S 모델(307)이 사용된다.

[0033] 도 4는 학생망 모델(150)의 예를 보여주는 블록도이다.

[0034] 도 4의 예에서, 학생망 모델(150)은 2차원 CNN(402), 3차원 CNN(403) 및 S2S 모델(407)을 포함한다. 학생망 모델(150)은 제1 교사망 모델(110) 및 제2 교사망 모델(120)에 비해 소형인 경량 모델일 수 있다. 다시 말해, 교사망 모델(110, 120)은 지식 증류를 통해 학생망 모델(150)로 압축된다고 볼 수 있다.

[0035] 도시된 예에서, 2차원 CNN(402) 및/또는 3차원 CNN(403)은 입력 비디오(106)의 특징 표현(이는, 예컨대, 입력 비디오(106)의 하나 이상의 프레임에 의해 나타내어진 장면 내의 객체, 액션 및/또는 술어를 나타냄)의 시퀀스를 추출하는 데에 사용된다. 그러면, 추출된 시퀀스로부터 입력 비디오(106)를 위한 후보 캡션(즉, 단어의 시퀀스)을 나타내는 학생망 출력을 생성하는 데에 S2S 모델(407)이 사용된다.

[0036] 이제, 비디오 캡셔닝 프레임워크(100)에서 교사망 모델 세트(101) 내의 적어도 하나의 교사망 모델로부터 학생망 모델(150)에 지식을 증류하기 위해 총 증류 손실을 계산하는 것을 설명한다.

[0037] 편의상, 제1 교사망 모델(110) 및 제2 교사망 모델(120) 양자로부터 학생망 모델(150)에 지식을 증류하는 것을 먼저 다룬다. 이 경우에, 총 증류 손실은 도 7에 도시된 바와 같이 총 가중 증류 손실(780)로서 계산될 수 있다. 도 7의 예에서, 제1 교사망 출력 및 학생망 출력의 비교에 기반하여 제1 증류 손실(711)이 계산되고, 제2 교사망 출력 및 학생망 출력의 비교에 기반하여 제2 증류 손실(712)이 계산된다. 예를 들어, 제1 증류 손실(711)은 제1 교사망 모델(110)로부터 출력된 소프트 레이블 및 학생망 모델(150)로부터 출력된 소프트 레이블 예측 간의 차이를 나타내는 교차 엔트로피 손실(cross-entropy loss)일 수 있고, 제2 증류 손실(712)은 제2 교사망 모델(120)로부터 출력된 소프트 레이블 및 학생망 모델(150)로부터 출력된 소프트 레이블 예측 간의 차이를 나타내는 교차 엔트로피 손실일 수 있다.

[0038] 또한, 제1 교사망 모델(110)의 비디오 캡셔닝 성능을 나타내는 제1 성능 메트릭(performance metric)(721)이 제1 교사망 출력에 기반하여 계산되고, 제2 교사망 모델(120)의 비디오 캡셔닝 성능을 나타내는 제2 성능 메트릭(722)이 제2 교사망 출력에 기반하여 계산된다. 예를 들어, 이들 성능 메트릭(721, 722) 각각은 각자의 교사망 모델(110, 120)로부터 출력된 후보 캡션 및 입력 비디오(106)에 대해 주어진 정답(ground truth) 내지 기준(reference) 캡션 간의 대응관계 내지 유사성을 측정하는 평가 메트릭, 예컨대, CIDEr(Consensus-based Image Description Evaluation), METEOR(Metric for Evaluation of Translation with Explicit Ordering), BLEU(Bilingual Evaluation Understudy)-4, ROUGE(Recall Oriented Understudy of Gisting Evaluation)-L, 또는 유사한 것일 수 있다. 또한, 제1 증류 손실(711) 및 제2 증류 손실(712)과 달리, 각각의 성능 메트릭(721, 722)의 계산은 학생망 출력에 의존하지 않을 수 있다.

[0039] 이후에, 제1 증류 손실(711) 및 제2 증류 손실(712)은 각각 제1 성능 메트릭(721) 및 제2 성능 메트릭(722)에 기반하여 가중화되고, 총 가중 증류 손실(780)은 제1 증류 손실(711) 및 제2 증류 손실(712)의 가중 합으로서 계산된다. 예를 들어, 총 가중 증류 손실(780)은 다음과 같이 표현될 수 있다.

수학식 1

[0040] $$L_{distill} = w_1 L_{distill\_action} + w_2 L_{distill\_predicate}$$

[0041] 여기서 $L_{distill}$은 총 가중 증류 손실(780), $L_{distill_action}$은 제1 증류 손실(711), $L_{distill_predicate}$은 제2 증류 손실(712)이고, w_1 및 w_2는 각각 제1 성능 메트릭(721) 및 제2 성능 메트릭(722)에 기반하여 정해진 가중치이다. 가

중치 w_1 및 w_2 각각은 대응하는 성능 메트릭(721, 722)이 더 큰 값을 가질수록 더 클 수 있다.

[0042] 이에 따라, 계산된 총 가중 증류 손실에 기반하여 학생망 모델(150)이 훈련된다.

[0043] 몇몇 예시적인 구현에서, 비록 도 7에 도시되지 않으나, 학생망 모델(150)의 후보 캡션 및 입력 비디오(106)를 위한 기준 캡션에 기반하여 학생 손실(student loss)이 또한 계산될 수 있다. 예를 들어, 학생 손실은 학생망 모델(150)로부터 출력된 하드 예측 및 기준 캡션을 나타내는 원핫 벡터(one-hot vector)인 하드 레이블 간의 차이를 나타내는 교차 엔트로피 손실일 수 있다.

[0044] 만일 제1 교사망 모델(110)이 액션 그래프 기반 비디오 캡셔닝을 수행하도록 사전훈련되고, 제2 교사망 모델(120)이 술어 그래프 기반 비디오 캡셔닝을 수행하도록 사전훈련된 경우, 학생망 모델(150)로의 지식 증류에서 발생하는 총 손실은 다음과 같이 정의될 수 있다.

수학식 2

[0045] $$L = \alpha L_{student} + \beta L_{distill}$$

[0046] 여기서 L은 총 손실, $L_{student}$은 학생 손실, $L_{distill}$은 총 가중 증류 손실(780)이고, α 및 β는 각각 이 식에서의 대응하는 손실을 가중화하는 하이퍼파라미터이다.

[0047] 만일 제1 교사망 모델(110) 및 제2 교사망 모델(120)이 학생망 모델(150)과 동시에 훈련되는 경우에, 총 손실은 다음과 같이 정의될 수 있다.

수학식 3

[0048] $$L = \alpha L_{student} + \beta L_{distill} + \gamma L_{teacher\_action} + \delta L_{teacher\_predicate}$$

[0049] 여기서 L은 총 손실, $L_{student}$은 학생 손실, $L_{distill}$은 총 가중 증류 손실(780)이고, $L_{teacher_action}$은 제1 교사망 모델(110)의 후보 캡션 및 입력 비디오(106)을 위한 기준 캡션에 기반하여 계산된 손실(예컨대, 제1 교사망 모델(110)로부터 출력된 하드 예측 및 주어진 하드 레이블 간의 차이를 나타내는 교차 엔트로피 손실)이고, $L_{teacher_predicate}$은 제2 교사망 모델(120)의 후보 캡션 및 입력 비디오(106)을 위한 기준 캡션에 기반하여 계산된 손실(예컨대, 제2 교사망 모델(120)로부터 출력된 하드 예측 및 주어진 하드 레이블 간의 차이를 나타내는 교차 엔트로피 손실)이고, α, β, γ 및 δ는 각각 이 식에서의 대응하는 손실을 가중화하는 하이퍼파라미터이다.

[0050] 나아가, 학생망 모델(150)로의 지식 증류를 위해 제1 교사망 모델(110) 및 제2 교사망 모델(120) 중 하나 및 적어도 하나의 다른 교사망 모델이 사용되는 경우에도, 총 증류 손실은 수학식 1과 유사하게 총 가중 증류 손실로서 계산될 수 있음(즉, 여러 교사망 출력을 각각 학생망 출력과 비교함으로써 계산된 여러 증류 손실을 각각 여러 교사망 모델의 성능 메트릭에 기반하여 가중화하는 방식으로 계산될 수 있음)이 이해될 것이다. 또한, 그러한 지식 증류의 총 손실은 수학식 2 또는 수학식 3과 마찬가지 방식으로 계산될 수 있다.

[0051] 한편, 제1 교사망 모델(110) 및 제2 교사망 모델(120) 중 단 하나로부터 학생망 모델(150)로 지식을 증류하는 경우에, 총 증류 손실은 성능 메트릭 기반의 가중화 없이 하나의 대응하는 교사망 출력 및 학생망 출력의 비교에 기반하여 (예컨대, 총 가중 증류 손실(780)이 아니라, 전술된 제1 증류 손실(711)로서, 아니면 전술된 제2 증류 손실(712)로서) 계산될 수 있다. 또한, 지식 증류의 총 손실은 수학식 2에서 위와 같은 증류 손실을 $L_{distill}$(즉, 총 가중 증류 손실(780)) 대신에 적용함으로써 계산될 수 있거나, 수학식 3에서 그러한 증류 손실을 $L_{distill}$ 대신에 적용하고 $L_{teacher\_action}$ 및 $L_{teacher\_predicate}$ 중에서 이 지식 증류와 무관한 것을 빠뜨림으로써 계산될 수 있다.

[0052] 도 8은 비디오 캡셔닝 프레임워크(100)을 이용한 지식 증류를 위한 운영 환경(800)의 예를 보여준다.

[0053] 예시적인 운영 환경(800)에서, 비디오 캡셔닝 프레임워크(100)는 컴퓨팅 장치(810) 상에 구현된다. 컴퓨팅 장치

(810)는 버스를 통해 서로 통신가능하게 커플링된 프로세서(processor)(811), 메모리(memory)(812) 및 통신 인터페이스(813)를 포함한다.

[0054] 도시된 예에서, 프로세서(811)는 컴퓨팅 장치(810)의 전반적인 동작을 제어한다. 예를 들어, 프로세서(811)는 (가령, 학생망 모델(150)을 훈련시키는 것과 관련하여) 본 문서에 기술된 동작 중 전부 또는 일부를 수행할 수 있다.

[0055] 도시된 예에서, 메모리(812)는 다양한 정보를 저장한다. 예를 들어, 메모리(812)는 비디오 캡셔닝 프레임워크(100)를 구현하고/거나 가동하는 데에서 사용되고/거나 산출되는 정보, 예컨대, 컴퓨팅 장치(810)의 동작 중에 사용되는 데이터 및/또는 컴퓨팅 장치(810)(가령, 프로세서(811))에 의해 실행되는 명령어의 세트를 비일시적(non-transitory) 형태로 저장할 수 있다.

[0056] 도시된 예에서, 통신 인터페이스(813)는 컴퓨팅 장치(810)로 하여금 다른 장치(가령, 컴퓨팅 장치(820))에 데이터를 송신하고 이로부터 데이터를 수신하기 위해 네트워크(830)(가령, 인터넷)를 액세스할 수 있게 한다. 예를 들어, 컴퓨팅 장치(810)는, 통신 인터페이스(813)를 통해, 컴퓨팅 장치(820)(가령, 모바일 컴퓨팅 장치, 개인용 컴퓨터 또는 유사한 것과 같은 클라이언트 장치)로부터 학생망 모델(150)에 대한 요청을 수신하고, 훈련된 학생망 모델(150)을 컴퓨팅 장치(820)에 제공할 수 있다.

[0057] 도 9는 그래프 기반 비디오 캡셔닝을 위한 지식 증류를 수행하는 예시적인 프로세스(900)를 보여주는 흐름도이다. 예를 들어, 프로세스(900)는 비디오 캡셔닝 프레임워크(100)가 구현된 컴퓨팅 장치(가령, 컴퓨팅 장치(810))에 의해 수행될 수 있다. 프로세스(900)의 다른 예시적인 흐름이 또한 고려된다. 예를 들어, 프로세스(900)는 도시되지 않은 추가적인 동작을 또한 포함할 수 있고/거나, 도 9에 도시된 동작 중 일부를 포함하나 전부를 포함하지는 않을 수 있다.

[0058] 동작(910)에서, 적어도 하나의 교사망 모델 및 학생망 모델에 비디오가 입력된다. 예를 들어, 상호작용하는 여러 객체를 포함한 장면을 나타내는 비디오 프레임의 시퀀스가 입력 비디오로서 제1 교사망 모델(110) 및 제2 교사망 모델(120) 중 적어도 하나 및 학생망 모델(150)에 제공될 수 있다.

[0059] 동작(920)에서, 입력된 비디오에 대해 그래프 기반 비디오 캡셔닝이 적어도 하나의 교사망 모델을 통해 수행되고, 이에 따라 적어도 하나의 교사망 출력이 생성된다. 몇몇 예시적인 구현에서, 입력된 비디오에 대해 액션 그래프 기반 비디오 캡셔닝이 제1 교사망 모델(110)을 통해 수행됨에 따라 제1 교사망 모델(110)에 의해 제1 교사망 출력이 생성되고 제공될 수 있다. 예를 들어, 제1 교사망 출력은 입력된 비디오에 대해 제1 교사망 모델(110)에 의해 예측된 후보 캡션을 단어 확률의 세트의 시퀀스의 형태로 나타내는 소프트 레이블 예측을 포함할 수 있다. 추가적으로 또는 대안적으로, 입력된 비디오에 대해 술어 그래프 기반 비디오 캡셔닝이 제2 교사망 모델(120)을 통해 수행됨에 따라 제2 교사망 모델(120)에 의해 제2 교사망 출력이 생성되고 제공될 수 있다. 예를 들어, 제2 교사망 출력은 입력된 비디오에 대해 제2 교사망 모델(120)에 의해 예측된 후보 캡션을 단어 확률의 세트의 시퀀스의 형태로 나타내는 소프트 레이블 예측을 포함할 수 있다.

[0060] 동작(930)에서, 입력된 비디오에 대해 학생망 모델)을 통해 비디오 캡셔닝이 수행되고, 이에 따라 학생망 출력이 생성된다. 몇몇 예시적인 구현에서, 입력된 비디오에 대해 학생망 모델(150)을 통해 비디오 캡셔닝이 수행됨에 따라 학생망 모델(150)에 의해 학생망 출력이 생성되고 제공된다. 예를 들어, 학생망 출력은 입력된 비디오에 대해 학생망 모델(150)에 의해 예측된 후보 캡션을 단어 확률의 세트의 시퀀스의 형태로 나타내는 소프트 레이블 예측을 포함할 수 있다.

[0061] 동작(940)에서, 적어도 하나의 교사망 출력 및 학생망 출력에 기반하여 총 증류 손실이 계산된다. 예를 들어, 그래프 기반 비디오 캡셔닝을 위해 제1 교사망 모델 및 제2 교사망 모델 양자 모두가 사용되는 경우, 총 증류 손실로서 총 가중 증류 손실이 계산될 수 있는데, 이는 수학식 1에서와 같이 제1 교사망 출력 및 학생망 출력 간의 제1 증류 손실(711) 및 제2 교사망 출력 및 학생망 출력 간의 제2 증류 손실(712)의 가중 합일 수 있다. 제1 증류 손실(711)의 가중치는 제1 교사망 모델(110)의 제1 성능 메트릭(721)에 기반하여 정해질 수 있고, 제2 증류 손실(712)의 가중치는 제2 교사망 모델(120)의 제2 성능 메트릭(722)에 기반하여 정해질 수 있다.

[0062] 동작(950)에서, 계산된 총 증류 손실에 기반하여 학생망 모델(150)이 훈련된다. 예를 들어, 총 증류 손실(가령, 수학식 2 또는 수학식 3에서와 같이 총 가중 증류 손실) 및 학생 손실로써 정의된 총 손실에 기반하여, 학생망 모델(150)의 파라미터의 세트가 갱신될 수 있다. 학생망 모델(150)의 훈련이 마무리되면, 학생망 모델(150) 단독으로 (가령, 컴퓨팅 장치(820)에 다운로드되어) 비디오 캡셔닝 작업(가령, 비디오 분류, 비디오 검색 등과 같은 다양한 타입의 다운스트림(downstream) 작업)을 위해 사용될 수 있다.

[0063] 다음은 그래프 기반 비디오 캡셔닝을 위한 지식 증류에 관련된 다양한 예이다.

[0064] 예 1에서, 그래프 기반 비디오 캡셔닝(video captioning)을 위한 지식 증류(knowledge distillation)의 방법은 다음을 포함한다: 적어도 하나의 교사망 모델(teacher network model) 및 학생망 모델(student network model)에 비디오를 입력하는 단계; 상기 적어도 하나의 교사망 모델을 통해, 상기 입력된 비디오에 대해 그래프 기반 비디오 캡셔닝을 수행하여 적어도 하나의 교사망 출력을 생성하는 단계(상기 적어도 하나의 교사망 모델은 상기 입력된 비디오에 대해 액션(action) 그래프 기반 비디오 캡셔닝을 수행하여 제1 교사망 출력을 생성하는 제1 교사망 모델 및 상기 입력된 비디오에 대해 술어(predicate) 그래프 기반 비디오 캡셔닝을 수행하여 제2 교사망 출력을 생성하는 제2 교사망 모델 중 적어도 하나를 포함함); 상기 학생망 모델을 통해, 상기 입력된 비디오에 대해 비디오 캡셔닝을 수행하여 학생망 출력을 생성하는 단계; 상기 적어도 하나의 교사망 출력 및 상기 학생망 출력에 기반하여 총 증류 손실을 계산하는 단계; 및 상기 계산된 총 증류 손실에 기반하여 상기 학생망 모델을 훈련시키는 단계.

[0065] 예 2는 예 1의 주제를 포함하는데, 상기 적어도 하나의 교사망 출력 및 상기 학생망 출력 각각은 상기 입력된 비디오를 위한 각자의 후보 캡션을 예측된 단어 확률의 세트의 시퀀스(sequence)로서 나타내는 소프트 레이블(soft label) 예측을 포함한다.

[0066] 예 3은 예 2의 주제를 포함하는데, 예측된 단어 확률의 상기 세트 각각은 상기 입력된 비디오를 위한 임의의 후보 캡션에서 이용가능한 어휘(vocabulary) 내의 다수의 단어에 걸쳐 정해지고, 상기 각자의 후보 캡션 내의 각각의 단어에 상기 시퀀스 내의 상이한 세트가 대응한다.

[0067] 예 4는 예 2 또는 예 3의 주제를 포함하는데, 상기 방법은 상기 학생망 모델의 상기 후보 캡션 및 상기 입력된 비디오를 위한 기준 캡션(reference caption)에 기반하여 학생 손실(student loss)을 계산하는 단계를 더 포함하고, 상기 학생망 모델은 상기 계산된 학생 손실에 또한 기반하여 훈련된다.

[0068] 예 5는 예 1 내지 예 4 중 임의의 것의 주제를 포함하는데, 상기 적어도 하나의 교사망 모델은 각자의 상이한 그래프 기반 비디오 캡셔닝을 수행하여 복수의 교사망 출력을 생성하는 복수의 교사망 모델이고, 상기 복수의 교사망 모델은 상기 제1 교사망 모델 및 상기 제2 교사망 모델 중 상기 적어도 하나를 포함한다.

[0069] 예 6은 예 5의 주제를 포함하는데, 상기 총 증류 손실을 계산하는 단계는 다음을 포함한다: 상기 복수의 교사망 출력을 각각 상기 학생망 출력과 비교하여 복수의 증류 손실을 계산하는 단계; 상기 복수의 교사망 출력으로부터 복수의 성능 메트릭을 계산하는 단계(상기 복수의 성능 메트릭 각각은 상기 복수의 교사망 모델 중 각자의 것의 성능 메트릭임); 및 상기 총 증류 손실을 상기 복수의 증류 손실의 가중 합(weighted sum)으로서 계산하기 위해, 상기 복수의 증류 손실을 각각 상기 복수의 성능 메트릭에 기반하여 가중화하는 단계.

[0070] 예 7은 예 6의 주제를 포함하는데, 상기 복수의 성능 메트릭 각각은 상기 복수의 교사망 모델 중 각자의 것으로부터 출력된 후보 캡션 및 상기 입력된 비디오에 대해 주어진 기준 캡션 간의 대응관계 또는 유사성을 측정한다.

[0071] 예 8은 예 1 내지 예 4 중 임의의 것의 주제를 포함하는데, 상기 적어도 하나의 교사망 모델은 상기 제1 교사망 모델 및 상기 제2 교사망 모델 양자 모두를 포함한다.

[0072] 예 9는 예 8의 주제를 포함하는데, 상기 총 증류 손실을 계산하는 단계는 다음을 포함한다: 상기 제1 교사망 출력을 상기 학생망 출력과 비교하여 제1 증류 손실을 계산하는 단계; 상기 제2 교사망 출력을 상기 학생망 출력과 비교하여 제2 증류 손실을 계산하는 단계; 상기 제1 교사망 출력으로부터 상기 제1 교사망 모델의 제1 성능 메트릭(performance metric)을 계산하는 단계; 상기 제2 교사망 출력으로부터 상기 제2 교사망 모델의 제2 성능 메트릭을 계산하는 단계; 및 상기 총 증류 손실을 상기 제1 증류 손실 및 상기 제2 증류 손실의 가중 합(weighted sum)으로서 계산하기 위해, 상기 제1 증류 손실 및 상기 제2 증류 손실을 각각 상기 제1 성능 메트릭 및 상기 제2 성능 메트릭에 기반하여 가중화하는 단계.

[0073] 예 10은 예 9의 주제를 포함하는데, 상기 제1 성능 메트릭 및 상기 제2 성능 메트릭 각각은 상기 제1 교사망 모델 및 상기 제2 교사망 모델 중 각자의 것으로부터 출력된 후보 캡션 및 상기 입력된 비디오에 대해 주어진 기준 캡션 간의 대응관계 또는 유사성을 측정한다.

[0074] 예 11은 예 1 내지 예 10 중 임의의 것의 주제를 포함하는데, 상기 액션 그래프 기반 비디오 캡셔닝을 수행하여 상기 제1 교사망 출력을 생성하는 것은 다음을 포함한다: 장면을 나타내는, 상기 입력 비디오의 프레임의 시퀀스로부터, 상기 장면 내에 나타나는 검출된 객체 간의 인식된 액션을 나타내는 액션 그래프를 구축하는 것; 및

상기 액션 그래프에 기반하여 상기 제1 교사망 출력을 생성하는 것.

[0075] 예 12는 예 11의 주제를 포함하는데, 상기 액션 그래프는 공간적 그래프(spatial graph) 및 시간적 그래프(temporal graph)로부터 생성된 시공간적 그래프(spatio-temporal graph)이고, 상기 공간적 그래프는 객체 검출 모델로부터 출력된 객체 특징 및 액션 인식 모델로부터 출력된 액션 특징으로부터 생성되고, 상기 시간적 그래프는 상기 객체 특징으로부터 생성된다.

[0076] 예 13은 예 11 또는 예 12의 주제를 포함하는데, 상기 검출된 객체는 인간 객체(human object) 및 다른 인간 또는 비인간(non-human) 객체이다.

[0077] 예 14는 예 11 내지 예 13 중 임의의 것의 주제를 포함하는데, 상기 액션 그래프를 구축하는 것은 다음을 포함한다: 상기 인식된 액션을 예측된 액션 확률의 세트로서 나타내는 다차원(multi-dimensional) 특징을 상기 액션 그래프의 엣지(edge)에 할당하기 위해, 액션 인식 모델을 사용하는 것.

[0078] 예 15는 예 14의 주제를 포함하는데, 예측된 확률의 상기 세트는 다수의 인식가능한 액션에 걸쳐 정해지고, 상기 액션 인식 모델은 상기 다수의 인식가능한 액션으로의 분류를 수행하도록 사전훈련된다.

[0079] 예 16은 예 11 내지 예 15 중 임의의 것의 주제를 포함하는데, 상기 액션 그래프에 기반하여 상기 제1 교사망 출력을 생성하는 것은 다음을 포함한다: 상기 입력된 비디오의 특징 표현의 시퀀스를 추출하기 위해 상기 액션 그래프에 그래프 신경망(graph neural network)을 적용하는 것; 및 특징 표현의 상기 시퀀스로부터 상기 제1 교사망 출력을 생성하기 위해 시퀀스 대 시퀀스 모델(sequence-to-sequence model)을 사용하는 것.

[0080] 예 17은 예 1 내지 예 16 중 임의의 것의 주제를 포함하는데, 상기 술어 그래프 기반 비디오 캡셔닝을 수행하는 것은 다음을 포함한다: 장면을 나타내는, 상기 입력 비디오의 프레임으로부터, 상기 장면 내에 나타나는 객체의 쌍의 검출된 시각적 관계를 나타내는 술어 그래프를 구축하는 것; 및 상기 술어 그래프에 기반하여 상기 제2 교사망 출력을 생성하는 것.

[0081] 예 18은 예 17의 주제를 포함하는데, 상기 술어 그래프는 공간적 그래프 및 시간적 그래프로부터 생성된 시공간적 그래프이고, 상기 공간적 그래프는 시각적 관계 검출 모델로부터 출력된 객체 특징 및 술어 특징으로부터 생성되고, 상기 시간적 그래프는 상기 객체 특징으로부터 생성된다.

[0082] 예 19는 예 17 또는 예 18의 주제를 포함하는데, 상기 술어 그래프를 구축하는 것은 다음을 포함한다: 상기 검출된 시각적 관계를 예측된 술어 확률의 세트로서 나타내는 다차원 특징을 상기 술어 그래프의 엣지에 할당하기 위해, 시각적 관계 검출 모델을 사용하는 것.

[0083] 예 20은 예 19의 주제를 포함하는데, 예측된 술어 확률의 상기 세트는 다수의 검출가능한 시각적 관계에 걸쳐 정해지고, 상기 시각적 관계 검출 모델은 상기 다수의 검출가능한 시각적 관계로의 분류를 수행하도록 사전훈련된다.

[0084] 예 21은 예 17 내지 예 20 중 임의의 것의 주제를 포함하는데, 상기 술어 그래프에 기반하여 상기 제2 교사망 출력을 생성하는 것은 다음을 포함한다: 상기 입력된 비디오의 특징 표현의 시퀀스를 추출하기 위해 상기 술어 그래프에 그래프 신경망을 적용하는 것; 및 상기 특징 표현의 상기 시퀀스로부터 상기 제2 교사망 출력을 생성하기 위해 시퀀스 대 시퀀스 모델을 사용하는 것.

[0085] 예 22는 예 1 내지 예 21 중 임의의 것의 주제를 포함하는데, 상기 제1 교사망 모델은 상기 액션 그래프 기반 비디오 캡셔닝을 수행하도록 사전훈련되고, 상기 제2 교사망 모델은 상기 술어 그래프 기반 비디오 캡셔닝을 수행하도록 사전훈련된다.

[0086] 예 23은 예 1 내지 예 21 중 임의의 것의 주제를 포함하는데, 상기 방법은 상기 입력된 비디오를 사용하여 상기 제1 교사망 모델 및 상기 제2 교사망 모델을 훈련시키는 단계를 더 포함한다.

[0087] 예 24에서, 그래프 기반 비디오 캡셔닝을 위한 지식 증류의 장치는 다음을 포함한다: 프로세서; 및 상기 프로세서에 의해 실행되는 경우에 상기 장치로 하여금 예 1 내지 예 23 중 임의의 것에 따른 방법을 수행하게 하는 명령어가 저장된 컴퓨터 판독가능 저장 매체.

[0088] 예 25에서, 컴퓨터 프로세서에 의해 실행되는 경우 상기 컴퓨터 프로세서로 하여금 예 1 내지 예 23 중 임의의 것에 따른 방법을 수행하게 하는 컴퓨터 실행가능 명령어가 저장된 컴퓨터 판독가능 저장 매체가 제공된다.

[0089] 특정한 예에서, 본 문서에서 언급된 장치, 디바이스, 시스템, 머신 등은 임의의 적합한 유형은 컴퓨팅 장치이거

나, 이를 포함하거나, 이에 구현될 수 있다. 컴퓨팅 장치는 프로세서 및 프로세서에 의해 판독가능한 컴퓨터 판독가능 저장 매체를 포함할 수 있다. 프로세서는 컴퓨터 판독가능 저장 매체 내에 저장된 하나 이상의 명령어를 실행할 수 있다. 프로세서는 또한 컴퓨터 판독가능 저장 매체 내에 저장된 다른 정보를 판독할 수 있다. 추가로, 프로세서는 컴퓨터 판독가능 저장 매체 내에 새로운 정보를 저장할 수 있고 컴퓨터 판독가능 저장 매체 내에 저장된 어떤 정보를 갱신할 수 있다. 프로세서는, 예컨대, 중앙 처리 유닛(Central Processing Unit: CPU), 디지털 신호 프로세서(Digital Signal Processor: DSP), 그래픽 처리 유닛(Graphics Processing Unit: GPU), 프로세서 코어(processor core), 마이크로프로세서(microprocessor), 마이크로제어기(microcontroller), 필드 프로그램가능 게이트 어레이(Field-Programmable Gate Array: FPGA), 애플리케이션 특정 집적 회로(Application Specific Integrated Circuit: ASIC), 다른 하드웨어 및 로직 회로, 또는 이의 임의의 적합한 조합을 포함할 수 있다. 컴퓨터 판독가능 저장 매체는 다양한 정보, 예컨대, 프로세서에 의해 수행될 수 있는 프로세서 실행가능(processor executable) 명령어의 세트 및/또는 다른 정보로써 인코딩된다. 예를 들어, 컴퓨터 판독가능 저장 매체는 프로세서에 의해 실행되는 경우 컴퓨팅 장치(가령, 프로세서)로 하여금 본 문서에 개시된 몇몇 동작을 수행하게 하는 컴퓨터 프로그램 명령어 및/또는 그러한 동작에서 사용되는 정보, 데이터, 변수, 상수, 데이터 구조, 기타 등등이 내부에 저장될 수 있다. 컴퓨터 판독가능 저장 매체는, 예컨대, 판독 전용 메모리(Read-Only Memory: ROM), 랜덤 액세스 메모리(Random-Access Memory: RAM), 휘발성(volatile) 메모리, 비휘발성(non-volatile) 메모리, 착탈가능(removable) 메모리, 비착탈가능(non-removable) 메모리, 플래시(flash) 메모리, 솔리드 스테이트(solid-state) 메모리, 다른 타입의 메모리 디바이스, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD-ROM, DVD와 같은 광 기록 매체, 플롭티컬 디스크와 같은 자기-광 매체, 다른 타입의 저장 디바이스 및 저장 매체, 또는 이의 임의의 적합한 조합을 포함할 수 있다.

[0090] 특정한 예에서, 본 문서에 기술된 동작, 기법, 프로세스, 또는 이의 어떤 양상이나 부분은 컴퓨터 프로그램 제품 내에 구체화될 수 있다. 그러한 컴퓨터 프로그램은 컴퓨터에 의해 실행될 수 있는 어떤 유형의 (가령, 컴파일형(compiled) 또는 해석형(interpreted)) 프로그래밍 언어, 예컨대, 어셈블리(assembly), 기계어(machine language), 프로시저형(procedural) 언어, 객체지향(object-oriented) 언어 등등으로 구현될 수 있고, 하드웨어 구현과 조합될 수 있다. 컴퓨터 프로그램 제품은 컴퓨터 판독가능 저장 매체의 형태로 배포될 수 있거나 온라인으로 배포될 수 있다. 온라인 배포를 위해, 컴퓨터 프로그램 제품의 일부 또는 전부가 서버(가령, 서버의 컴퓨터 판독가능 저장 매체) 내에 일시적으로 저장되거나 일시적으로 생성될 수 있다.

[0091] 이상의 설명은 상세하게 몇몇 예를 예시하고 기술하기 위해 제시되었다. 본 개시의 범주에서 벗어나지 않고서 위의 교시에 비추어 많은 수정 및 변형이 가능함을 당업자는 응당 이해할 것이다. 다양한 예에서, 전술된 기법이 상이한 순서로 수행되고/거나, 전술된 시스템, 아키텍처, 디바이스, 회로 및 유사한 것의 컴포넌트 중 일부가 상이한 방식으로 결합 또는 조합되거나, 다른 컴포넌트 또는 이의 균등물에 의해 대치 또는 치환되더라도 적절한 결과가 달성될 수 있다.

[0092] 그러므로, 본 개시의 범주는 개시된 그 형태에 한정되어서는 안 되며, 후술하는 청구항 및 이의 균등물에 의해 정해져야 한다.

부호의 설명

[0093] 100: 비디오 캡셔닝 프레임워크

101: 교사망 모델 세트

110: 제1 교사망 모델

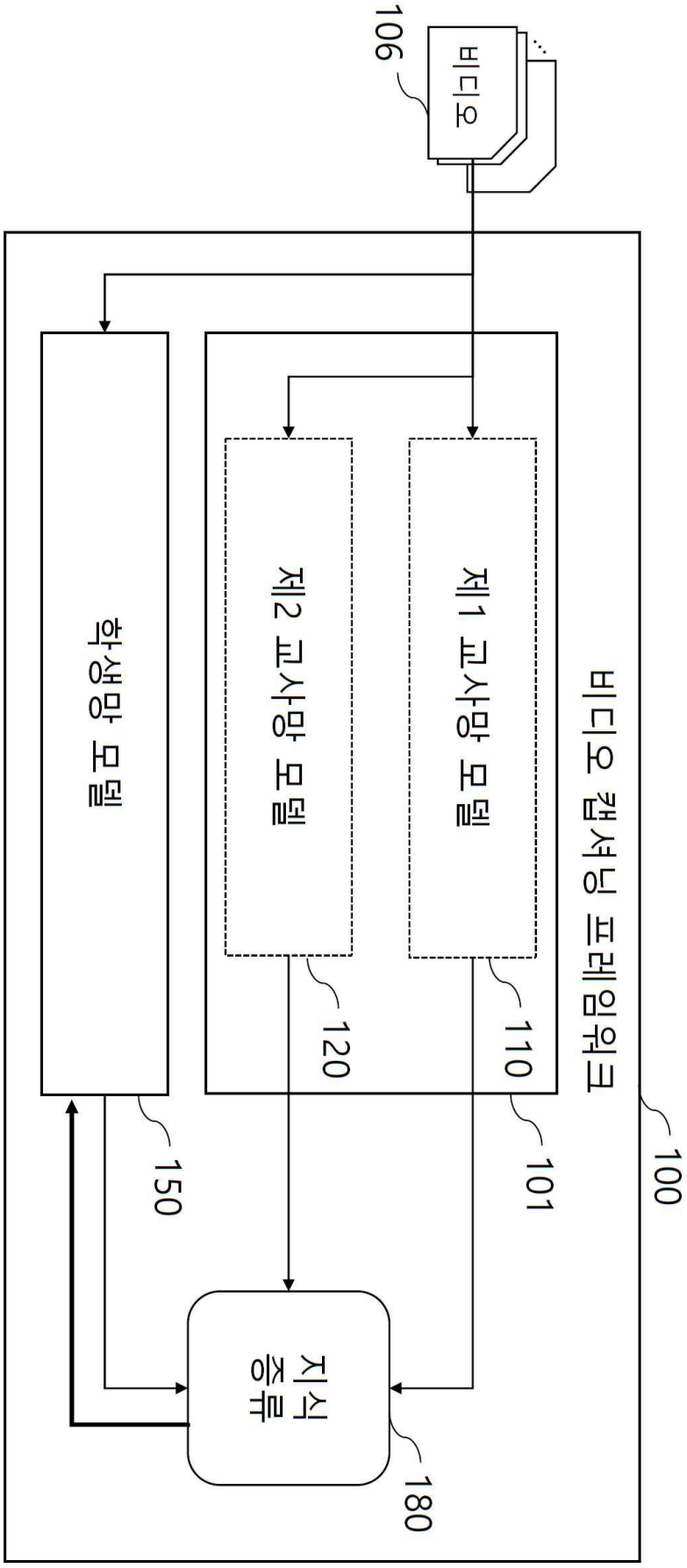
120: 제2 교사망 모델

150: 학생망 모델

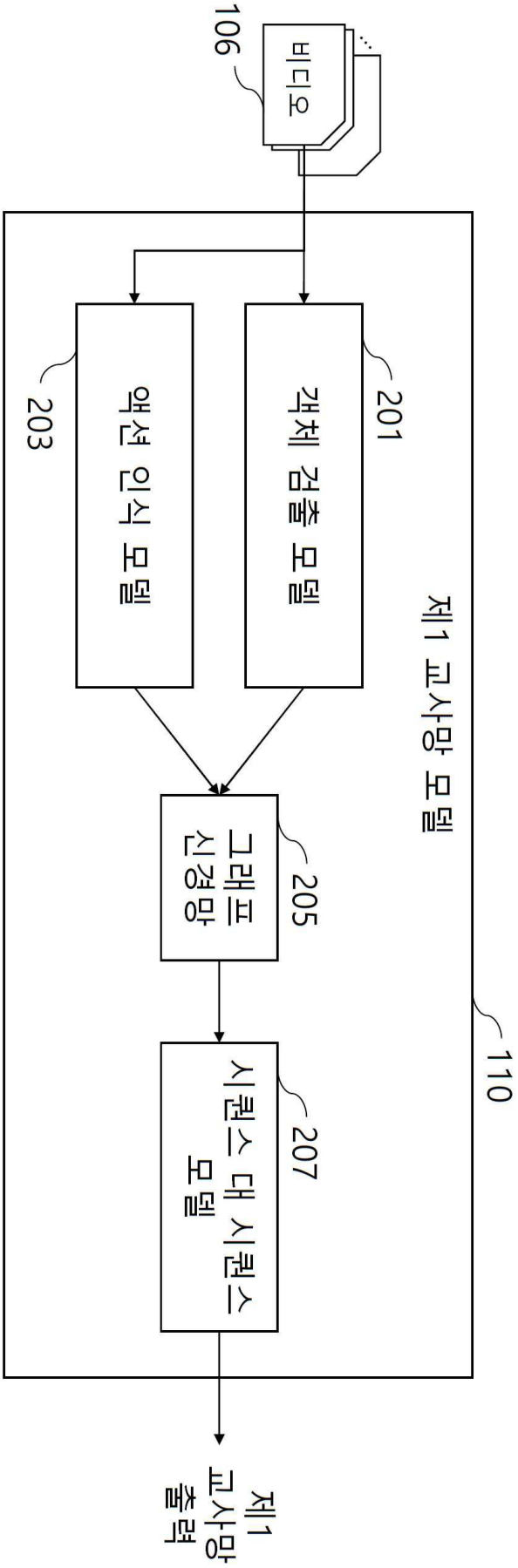
180: 지식 증류 모듈

도면

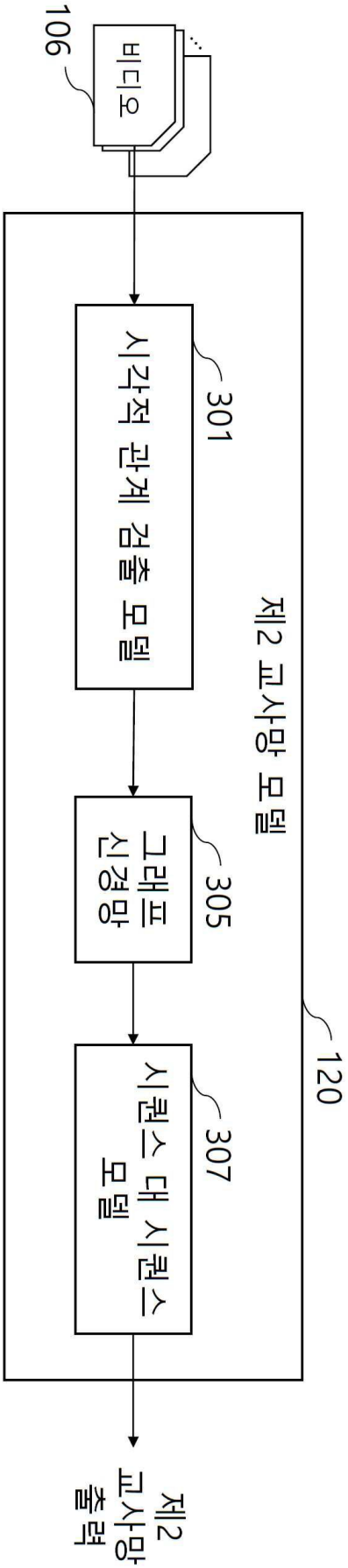
도면1


106
비디오
비디오 캡셔닝 프레임워크
제1 교사망 모델
제2 교사망 모델
학생망 모델
110
120
150
101
100
지식 증류
180

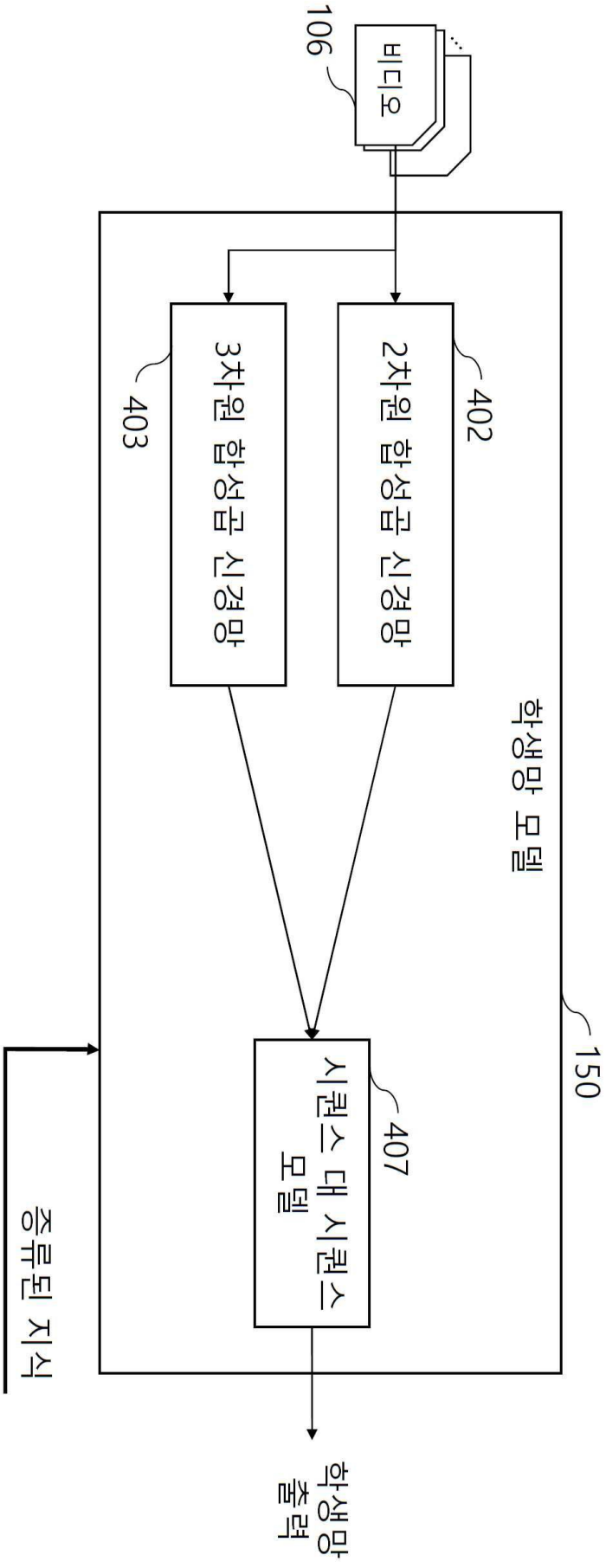
도면2


106
비디오
203
액션 인식 모델
201
객체 검출 모델
제1 교사망 모델
205
그래프 신경망
207
시퀀스 대 시퀀스 모델
110
제1 교사망 출력

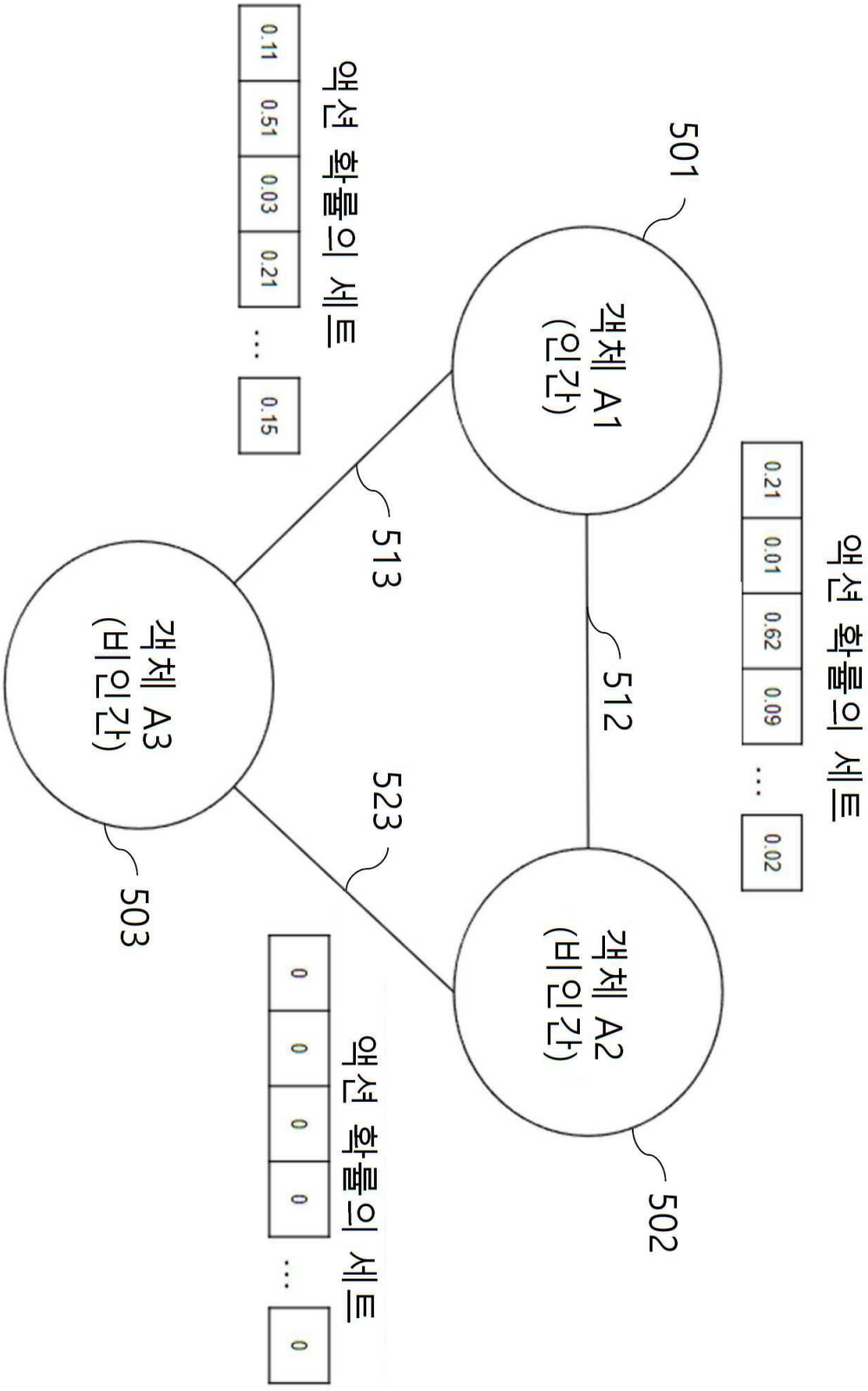
도면3


106
비디오
301
시각적 관계 검출 모델
305
그래프 신경망
307
시퀀스 대 시퀀스 모델
120
제2 교사망 모델
제2 교사망 출력

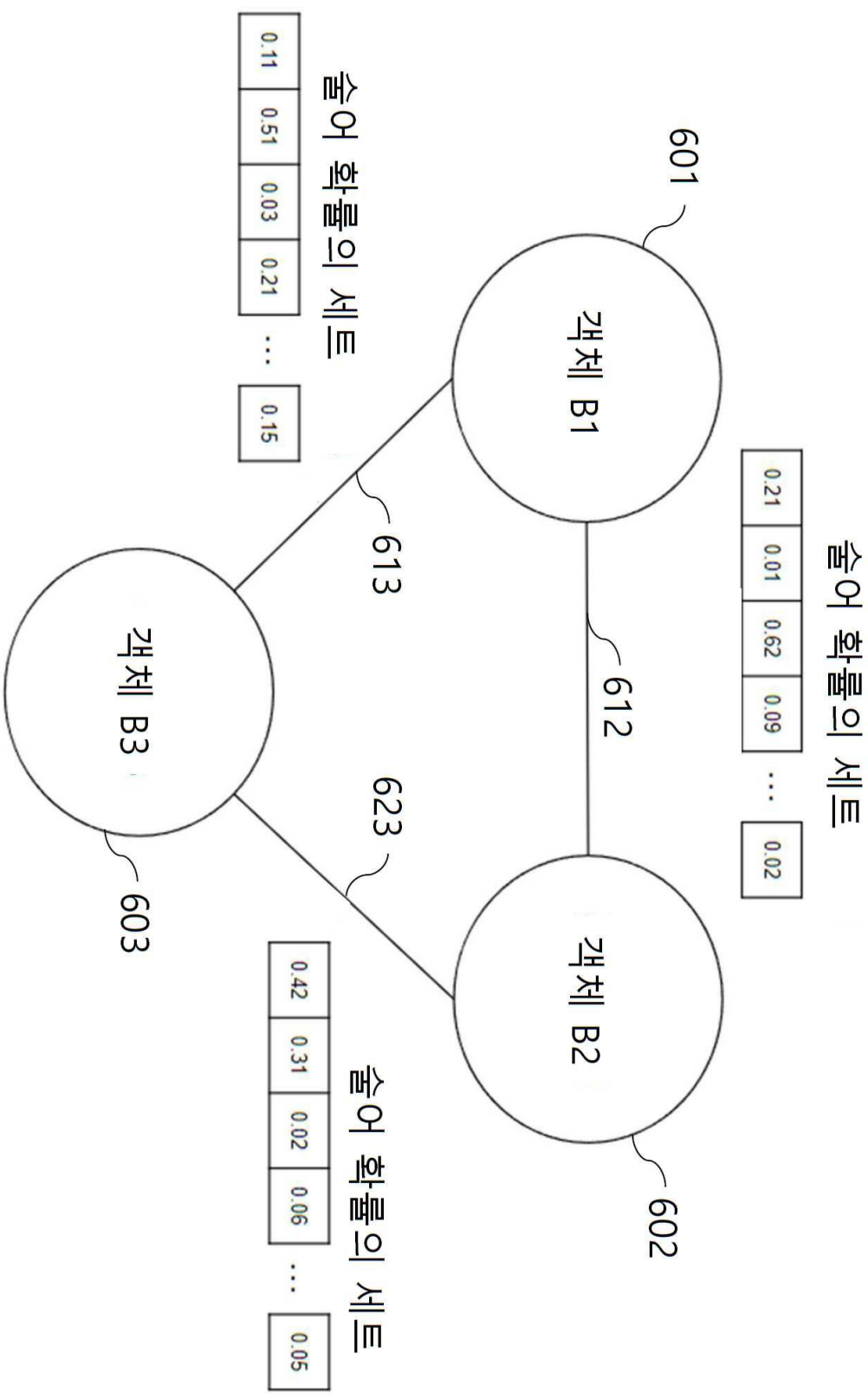
도면4


106
비디오
2차원 합성곱 신경망
402
3차원 합성곱 신경망
403
학생망 모델
150
시퀀스 대 시퀀스 모델
407
증류된 지식
학생망 출력

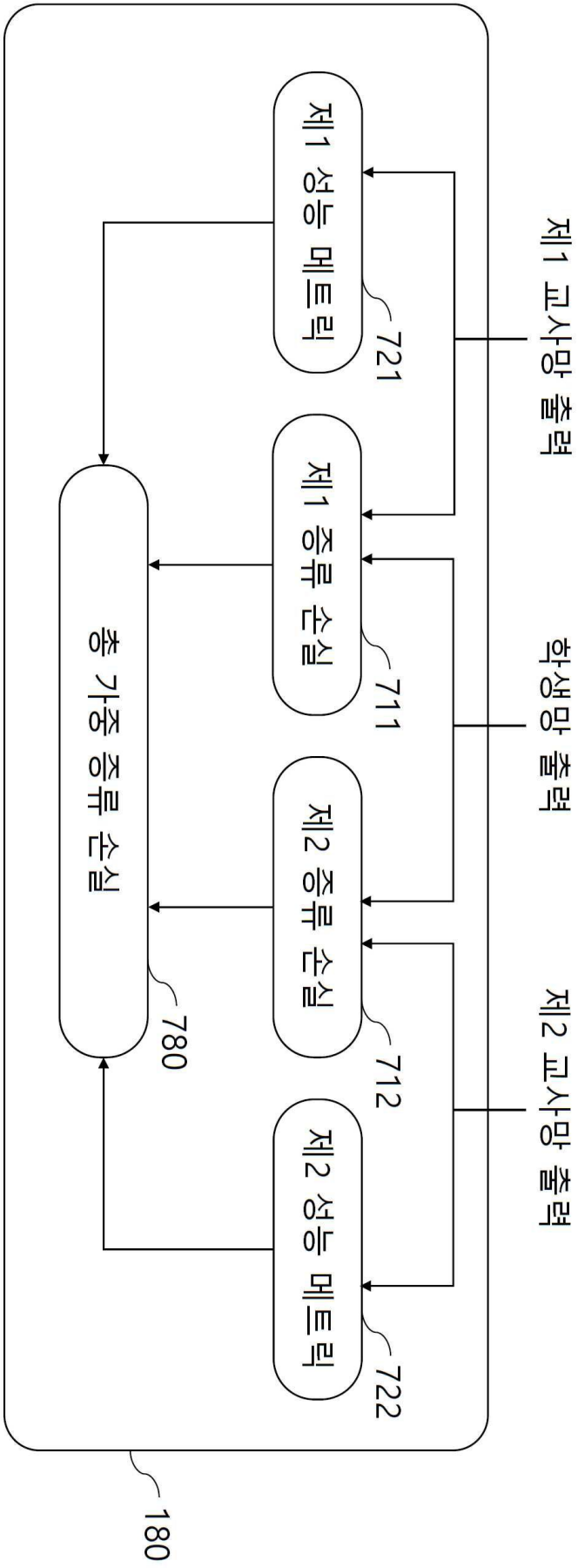
도면5


501
객체 A1
(인간)
액션 확률의 세트
0.11
0.51
0.03
0.21
...
0.15
513
512
0.21
0.01
0.62
0.09
...
0.02
객체 A3
(비인간)
503
523
객체 A2
(비인간)
502
액션 확률의 세트
0
0
0
0
...
0

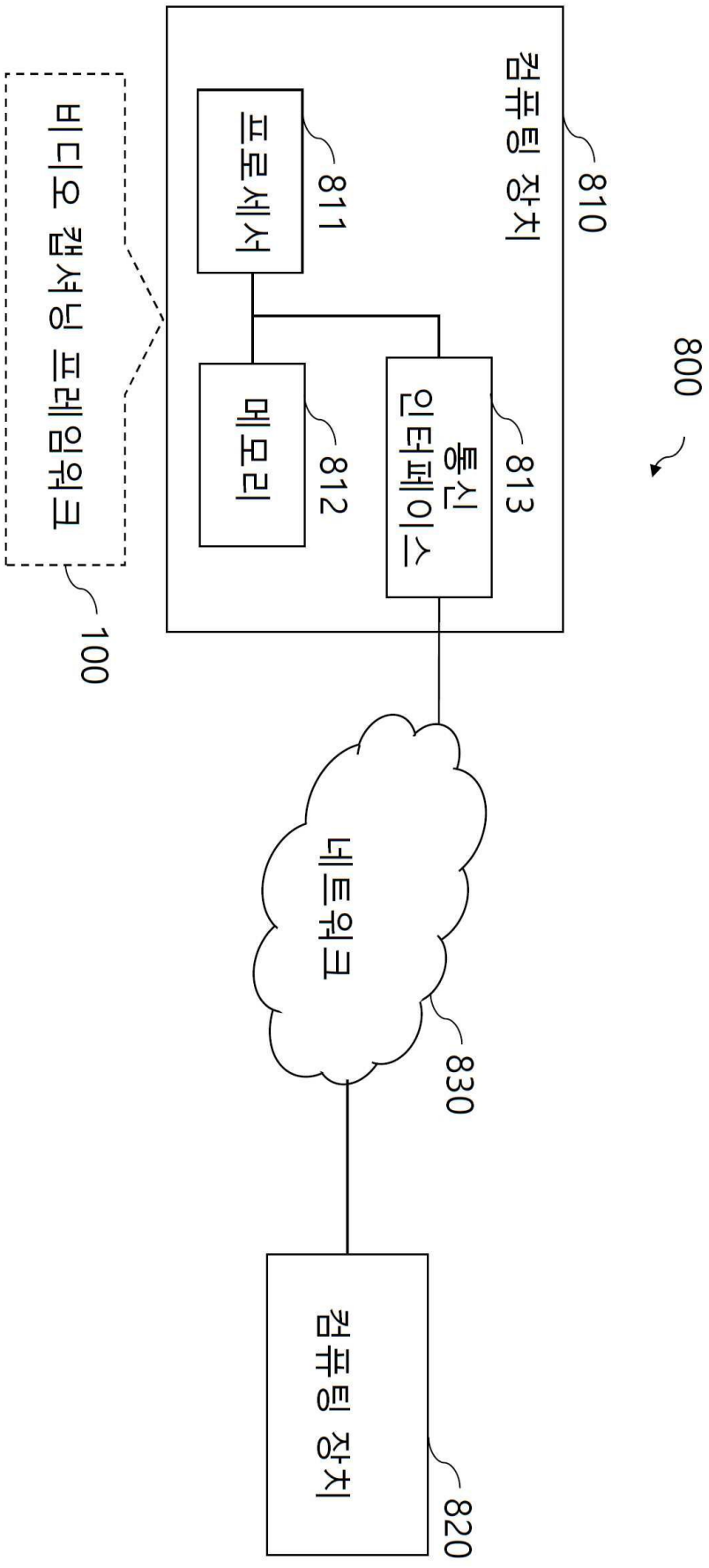
도면6


601
객체 B1
613
612
623
객체 B3
603
객체 B2
602
숱어 확률의 세트
0.11
0.51
0.03
0.21
0.15
숱어 확률의 세트
0.21
0.01
0.62
0.09
0.02
숱어 확률의 세트
0.42
0.31
0.02
0.06
0.05

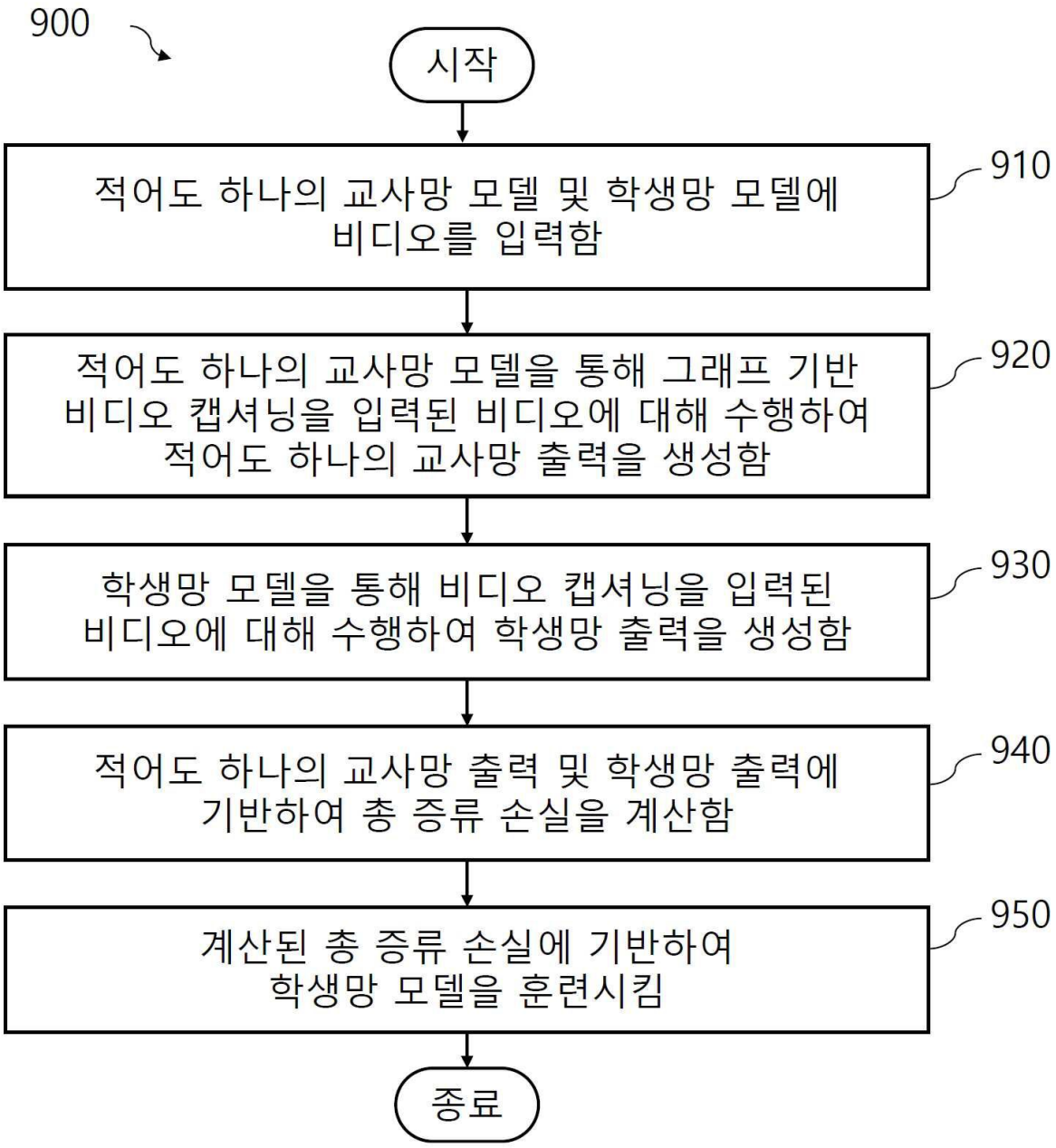
도면7


제1 성능 메트릭
721
제1 중류 손실
711
제2 중류 손실
712
제2 성능 메트릭
722
총 가중 중류 손실
780
180
제1 교사망 출력
학생망 출력
제2 교사망 출력

도면8


800
컴퓨팅 장치
810
프로세서
811
메모리
812
통신 인터페이스
813
비디오 캡쳐링 프레임워크
100
네트워크
830
컴퓨팅 장치
820

도면9


900
시작
적어도 하나의 교사망 모델 및 학생망 모델에 비디오를 입력함
910
적어도 하나의 교사망 모델을 통해 그래프 기반 비디오 캡셔닝을 입력된 비디오에 대해 수행하여 적어도 하나의 교사망 출력을 생성함
920
학생망 모델을 통해 비디오 캡셔닝을 입력된 비디오에 대해 수행하여 학생망 출력을 생성함
930
적어도 하나의 교사망 출력 및 학생망 출력에 기반하여 총 증류 손실을 계산함
940
계산된 총 증류 손실에 기반하여 학생망 모델을 훈련시킴
950
종료