



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0105139
(43) 공개일자 2024년07월05일

- (51) 국제특허분류(Int. Cl.)
G06N 3/092 (2023.01) G06F 16/738 (2019.01)
G06F 16/783 (2019.01) G06F 18/213 (2023.01)
G06N 3/0455 (2023.01) G06N 3/047 (2023.01)
- (52) CPC특허분류
G06N 3/092 (2023.01)
G06F 16/738 (2019.01)
- (21) 출원번호 10-2022-0188043
(22) 출원일자 2022년12월28일
심사청구일자 2022년12월28일
- (71) 출원인
세종대학교산학협력단
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
- (72) 발명자
정철
서울특별시 광진구 능동로 209 세종대학교 대양A
I센터 609호
벨다 바니아
서울특별시 광진구 능동로 209 세종대학교 대양A
I센터 612호
- (74) 대리인
김현승

전체 청구항 수 : 총 15 항

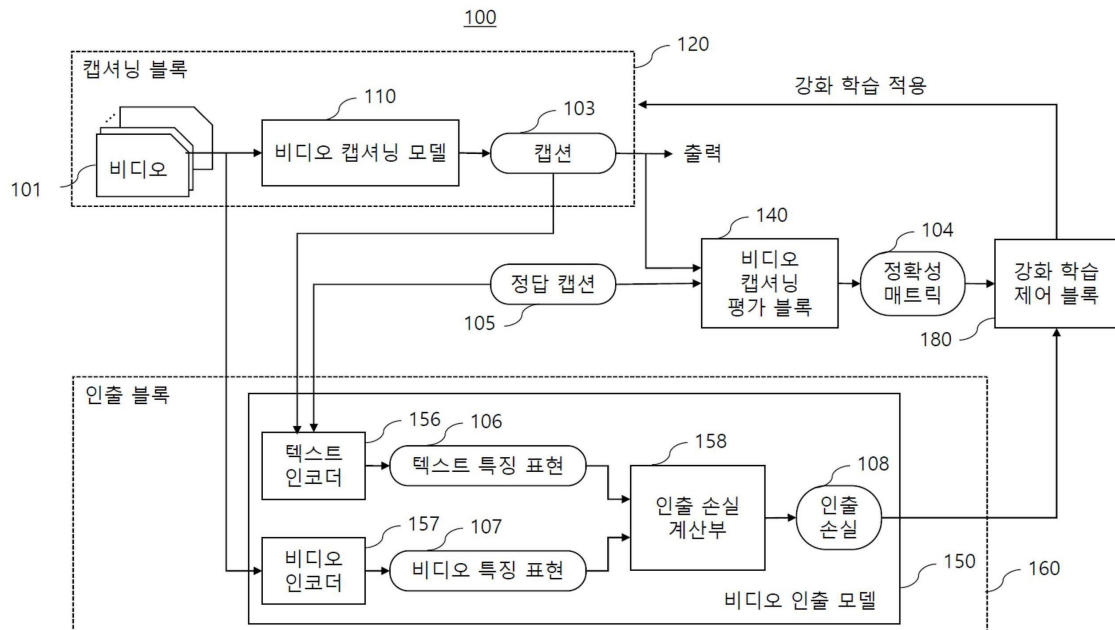
(54) 발명의 명칭 비디오 인출을 이용하는 비디오 캡서닝 프레임워크

(57) 요약

비디오 캡서닝을 위한 방법에 제공된다. 방법은, B개의 비디오 샘플 및 각자의 대응하는 정답 캡션을 포함하는 훈련 배치로써 비디오 캡서닝 모델을 훈련시키는 단계를 포함하되, 비디오 캡서닝 모델을 훈련시키는 단계는, 비디오 캡서닝 모델을 사용하여 B개의 비디오 샘플 중 임의의 특정한 것에 대해 캡션을 생성하는 단계와, 비디오

(뒷면에 계속)

대표도



인출 모델을 사용하여 제1 확률 분포 및 제2 확률 분포를 판정하는 단계(제1 확률 분포는 생성된 캡션을 입력으로서 수신하는 것에 응답하여 비디오 인출 모델이 B개의 비디오 샘플 중 어느 것을 출력할 것인지를 나타내고, 제2 확률 분포는 각자의 대응하는 정답 캡션 중에서 특정한 비디오 샘플에 대응하는 특정한 정답 캡션을 입력으로서 수신하는 것에 응답하여 비디오 인출 모델이 B개의 비디오 샘플 중 어느 것을 출력할 것인지를 나타냄)와, 제1 확률 분포 및 제2 확률 분포에 기반하여 보상을 계산하는 단계와, 보상을 사용하여 비디오 캡서닝 모델의 훈련에 강화 학습을 적용하는 단계를 포함한다.

(52) CPC특허분류

G06F 16/7844 (2019.01)

G06F 18/213 (2023.01)

G06N 3/0455 (2023.01)

G06N 3/047 (2023.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711159902
과제번호	2021-0-02067-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	차세대인공지능핵심원천기술개발
연구과제명	다목적 비디오 검색을 위한 차세대 인공신경망 기술 개발
기 여 율	70/100
과제수행기관명	세종대학교 산학협력단
연구기간	2022.04.01 ~ 2023.03.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711174042
과제번호	00156345
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성(R&D)
연구과제명	ICT혁신인재4.0(세종대학교)
기 여 율	30/100
과제수행기관명	세종대학교 산학협력단
연구기간	2022.07.01 ~ 2022.12.31

명세서

청구범위

청구항 1

비디오 캡셔닝(video captioning)을 위한 방법으로서,

B개의 비디오 샘플 및 각자의 대응하는 정답(ground-truth) 캡션을 포함하는 훈련 배치(training batch)로써 비디오 캡셔닝 모델을 훈련시키는 단계를 포함하되, 상기 비디오 캡셔닝 모델을 훈련시키는 단계는,

상기 비디오 캡셔닝 모델을 사용하여 상기 B개의 비디오 샘플 중 임의의 특정한 것에 대해 캡션을 생성하는 단계와,

비디오 인출 모델(video retrieval model)을 사용하여 제1 확률 분포 및 제2 확률 분포를 판정하는 단계 - 상기 제1 확률 분포는 상기 생성된 캡션을 입력으로서 수신하는 것에 응답하여 상기 비디오 인출 모델이 상기 B개의 비디오 샘플 중 어느 것을 출력할 것인지를 나타내고, 상기 제2 확률 분포는 상기 각자의 대응하는 정답 캡션 중에서 상기 특정한 비디오 샘플에 대응하는 특정한 정답 캡션을 입력으로서 수신하는 것에 응답하여 상기 비디오 인출 모델이 상기 B개의 비디오 샘플 중 어느 것을 출력할 것인지를 나타냄 - 와,

상기 제1 확률 분포 및 상기 제2 확률 분포에 기반하여 보상(reward)을 계산하는 단계와,

상기 보상을 사용하여 상기 비디오 캡셔닝 모델의 훈련에 강화 학습(Reinforcement Learning: RL)을 적용하는 단계를 포함하는,

방법.

청구항 2

제1항에 있어서,

상기 비디오 인출 모델은,

상기 생성된 캡션의 유사도 점수의 제1 리스트 및 상기 특정한 정답 캡션의 유사도 점수의 제2 리스트를 판정 - 상기 제1 리스트 내의 상기 유사도 점수 각각은 상기 B개의 비디오 샘플 중 각자의 것이 상기 생성된 캡션과 어느 정도로 매칭되는지를 나타내고, 상기 제2 리스트 내의 상기 유사도 점수 각각은 상기 B개의 비디오 샘플 중 각자의 것이 상기 특정한 정답 캡션과 어느 정도로 매칭되는지를 나타냄 - 하고,

상기 제1 리스트에 기반하여 상기 제1 확률 분포를 판정하고,

상기 제2 리스트에 기반하여 상기 제2 확률 분포를 판정하는 데에 사용되는,

방법.

청구항 3

제2항에 있어서,

상기 비디오 인출 모델은 텍스트 인코더(text encoder) 및 비디오 인코더(video encoder)를 포함하고, 상기 텍스트 인코더는 상기 생성된 캡션 및 상기 특정한 정답 캡션을 각각 공동 임베딩 공간(joint embedding space) 내의 제1 텍스트 특징 표현(text feature representation) 및 상기 공동 임베딩 공간 내의 제2 텍스트 특징 표현으로 맵핑하고, 상기 비디오 인코더는 상기 각자의 비디오 샘플을 상기 공동 임베딩 공간 내의 대응하는 비디오 특징 표현(video feature representation)으로 맵핑하고, 상기 제1 리스트 내의 상기 각각의 유사도 점수는 상기 대응하는 비디오 특징 표현에 대한 상기 제1 텍스트 특징 표현의 코사인 유사도(cosine similarity)이고, 상기 제2 리스트 내의 상기 각각의 유사도 점수는 상기 대응하는 비디오 특징 표현에 대한 상기 제2 텍스트 특징 표현의 코사인 유사도인,

방법.

청구항 4

제1항에 있어서,

상기 보상을 계산하는 단계는,

상기 제1 확률 분포 및 상기 제2 확률 분포에 기반하여 인출 손실을 계산하는 단계와,

상기 B개의 비디오 샘플 중 어느 것도 고려하지 않고서 상기 생성된 캡션 및 상기 특정한 정답 캡션에 기반하여 정확성 메트릭(accuracy metric)을 계산하는 단계와,

상기 인출 손실 및 상기 정확성 메트릭을 상기 보상으로 조합하는 단계를 포함하는,

방법.

청구항 5

제4항에 있어서,

상기 인출 손실을 계산하는 단계는 상기 생성된 캡션의 입력에 응답하여 상기 비디오 인출 모델에 의해 상기 특정한 비디오 샘플이 출력되는 확률 및 상기 특정한 정답 캡션의 입력에 응답하여 상기 비디오 인출 모델에 의해 상기 특정한 비디오 샘플이 출력되는 확률 간의 차이를 측정하는 단계를 포함하는,

방법.

청구항 6

제1항에 있어서,

상기 보상을 사용하여 상기 비디오 캡서닝 모델의 훈련에 RL을 적용하는 단계는,

상기 특정한 비디오 샘플에 대해 상기 생성된 캡션이 발생할 조건부 확률 및 상기 보상에 기반하여 RL 손실을 계산하는 단계와,

상기 RL 손실을 상기 비디오 캡서닝 모델의 훈련에 적용하는 단계를 포함하는,

방법.

청구항 7

제6항에 있어서,

상기 캡션은 예측된 단어 확률의 세트의 시퀀스에 의해 나타내어지고, 상기 비디오 캡서닝 모델을 훈련시키는 단계는,

예측된 단어 확률의 세트의 상기 시퀀스 및 상기 특정한 정답 캡션에 기반하여 음의 로그 우도(Negative Log-Likelihood: NLL) 손실을 계산하는 단계와,

상기 RL 손실 및 상기 NLL 손실에 기반하여 상기 비디오 캡서닝 모델의 파라미터를 조정하는 단계를 더 포함하는,

방법.

청구항 8

비디오 캡서닝을 위한 장치로서,

프로세서와,

상기 프로세서에 의해 실행되는 경우에 상기 프로세서로 하여금 B개의 비디오 샘플 및 각자의 대응하는 정답 캡션을 포함하는 훈련 배치로써 비디오 캡서닝 모델을 훈련시키는 동작을 수행하게 하는 명령어가 저장된 컴퓨터 판독가능 저장 매체를 포함하되, 상기 동작은,

상기 비디오 캡서닝 모델을 사용하여 상기 B개의 비디오 샘플 중 임의의 특정한 것에 대해 캡션을 생성하는 것과,

비디오 인출 모델을 사용하여 제1 확률 분포 및 제2 확률 분포를 판정하는 것 - 상기 제1 확률 분포는 상기 생

성된 캡션을 입력으로서 수신하는 것에 응답하여 상기 비디오 인출 모델이 상기 B개의 비디오 샘플 중 어느 것을 출력할 것인지를 나타내고, 상기 제2 확률 분포는 상기 각자의 대응하는 정답 캡션 중에서 상기 특정한 비디오 샘플에 대응하는 특정한 정답 캡션을 입력으로서 수신하는 것에 응답하여 상기 비디오 인출 모델이 상기 B개의 비디오 샘플 중 어느 것을 출력할 것인지를 나타냄 - 과,

상기 제1 확률 분포 및 상기 제2 확률 분포에 기반하여 보상을 계산하는 것과,

상기 보상을 사용하여 상기 비디오 캡셔닝 모델의 훈련에 강화 학습(Reinforcement Learning: RL)을 적용하는 것을 포함하는,

장치.

청구항 9

제8항에 있어서,

상기 비디오 인출 모델은,

상기 생성된 캡션의 유사도 점수의 제1 리스트 및 상기 특정한 정답 캡션의 유사도 점수의 제2 리스트를 판정 - 상기 제1 리스트 내의 상기 유사도 점수 각각은 상기 B개의 비디오 샘플 중 각자의 것이 상기 생성된 캡션과 어느 정도로 매칭되는지를 나타내고, 상기 제2 리스트 내의 상기 유사도 점수 각각은 상기 B개의 비디오 샘플 중 각자의 것이 상기 특정한 정답 캡션과 어느 정도로 매칭되는지를 나타냄 - 하고,

상기 제1 리스트에 기반하여 상기 제1 확률 분포를 판정하고,

상기 제2 리스트에 기반하여 상기 제2 확률 분포를 판정하는 데에 사용되는,

장치.

청구항 10

제9항에 있어서,

상기 비디오 인출 모델은 텍스트 인코더 및 비디오 인코더를 포함하고, 상기 텍스트 인코더는 상기 생성된 캡션 및 상기 특정한 정답 캡션을 각각 공동 임베딩 공간 내의 제1 텍스트 특징 표현 및 상기 공동 임베딩 공간 내의 제2 텍스트 특징 표현으로 맵핑하고, 상기 비디오 인코더는 상기 각자의 비디오 샘플을 상기 공동 임베딩 공간 내의 대응하는 비디오 특징 표현으로 맵핑하고, 상기 제1 리스트 내의 상기 각각의 유사도 점수는 상기 대응하는 비디오 특징 표현에 대한 상기 제1 텍스트 특징 표현의 코사인 유사도이고, 상기 제2 리스트 내의 상기 각각의 유사도 점수는 상기 대응하는 비디오 특징 표현에 대한 상기 제2 텍스트 특징 표현의 코사인 유사도인,

장치.

청구항 11

제8항에 있어서,

상기 보상을 계산하는 것은,

상기 제1 확률 분포 및 상기 제2 확률 분포에 기반하여 인출 손실을 계산하는 것과,

상기 B개의 비디오 샘플 중 어느 것도 고려하지 않고서 상기 생성된 캡션 및 상기 특정한 정답 캡션에 기반하여 정확성 메트릭(accuracy metric)을 계산하는 것과,

상기 인출 손실 및 상기 정확성 메트릭을 상기 보상으로 조합하는 것을 포함하는,

장치.

청구항 12

제11항에 있어서,

상기 인출 손실을 계산하는 것은 상기 생성된 캡션의 입력에 응답하여 상기 비디오 인출 모델에 의해 상기 특정한 비디오 샘플이 출력되는 확률 및 상기 특정한 정답 캡션의 입력에 응답하여 상기 비디오 인출 모델에 의해

상기 특정한 비디오 샘플이 출력되는 확률 간의 차이를 측정하는 것을 포함하는, 장치.

청구항 13

제8항에 있어서,

상기 보상을 사용하여 상기 비디오 캡셔닝 모델의 훈련에 RL을 적용하는 것은,

상기 특정한 비디오 샘플에 대해 상기 생성된 캡션이 발생할 조건부 확률 및 상기 보상에 기반하여 RL 손실을 계산하는 것과,

상기 RL 손실을 상기 비디오 캡셔닝 모델의 훈련에 적용하는 것을 포함하는,

장치.

청구항 14

제13항에 있어서,

상기 캡션은 예측된 단어 확률의 세트의 시퀀스에 의해 나타내어지고, 상기 비디오 캡셔닝 모델을 훈련시키는 것은,

예측된 단어 확률의 세트의 상기 시퀀스 및 상기 특정한 정답 캡션에 기반하여 음의 로그 우도(Negative Log-Likelihood: NLL) 손실을 계산하는 것과,

상기 RL 손실 및 상기 NLL 손실에 기반하여 상기 비디오 캡셔닝 모델의 파라미터를 조정하는 것을 더 포함하는, 장치.

청구항 15

컴퓨터 프로세서에 의해 실행되는 경우에 상기 컴퓨터 프로세서로 하여금 제1항 내지 제7항 중 어느 한 항에 기재된 방법을 수행하게 하는 컴퓨터 실행가능 명령어가 저장된 컴퓨터 판독가능 저장 매체.

발명의 설명

기술 분야

[0001] 본 개시는 비디오 캡셔닝(video captioning)에 관련된 것으로서, 구체적으로 비디오 인출(video retrieval)을 이용하는 비디오 캡셔닝 프레임워크에 관련된다.

배경 기술

[0002] 비디오 캡셔닝은 비디오 내의 시각적 콘텐츠(visual content)를 이해하여 비디오를 텍스트로써 자동적으로 기술하는 작업이다. 비디오 캡셔닝은 실생활의 다양한 분야, 예컨대, 로봇틱스(robotics), 수화 번역(sign language translation), 시각 장애인에 대한 보조 등에 응용될 수 있다. 이와 관련하여, 현재 비디오 캡셔닝 분야의 많은 작업은 인간에게 자연스러운 정답(ground truth)에 가능한 한 가깝게 정확한 문장(sentence)을 출력하게 하는 데에 노력을 기울이고 있다. 그러한 정확성(accuracy)은 종래의 평가 메트릭, 예컨대, BLEU(Bilingual Evaluation Understudy), METEOR(Metric for Evaluation of Translation with Explicit Ordering), ROUGE(Recall Oriented Understudy of Gisting Evaluation), CIDEr(Consensus-based Image Description Evaluation) 등에 의해 측정될 수 있다. 이들 메트릭은 n-그램 중첩 매칭(n-gram overlap matching)에 기반하므로 문장 정확성을 평가하는 데에 적합하다. 다만, 그러한 메트릭에 따라 출력 문장을 정답 문장과 비교하여 고성능의 평가가 나오더라도 출력 문장의 변별성(distinctiveness)을 보장하는 것은 아니다. 다시 말해, 기존의 몇몇 비디오 캡셔닝 방안은 상이한 비디오가 주어질 때 동일한 문장을 제공할 수도 있다.

발명의 내용

해결하려는 과제

[0003] 비디오 캡서닝이 본 문서에 개시된다.

과제의 해결 수단

[0004] 예에서, 비디오 캡서닝(video captioning)을 위한 방법은, B개의 비디오 샘플 및 각자의 대응하는 정답(ground-truth) 캡션을 포함하는 훈련 배치(training batch)로써 비디오 캡서닝 모델을 훈련시키는 단계를 포함하되, 위 비디오 캡서닝 모델을 훈련시키는 단계는, 위 비디오 캡서닝 모델을 사용하여 위 B개의 비디오 샘플 중 임의의 특정한 것에 대해 캡션을 생성하는 단계와, 비디오 인출 모델(video retrieval model)을 사용하여 제1 확률 분포 및 제2 확률 분포를 판정하는 단계(위 제1 확률 분포는 위 생성된 캡션을 입력으로서 수신하는 것에 응답하여 위 비디오 인출 모델이 위 B개의 비디오 샘플 중 어느 것을 출력할 것인지를 나타내고, 위 제2 확률 분포는 위 각자의 대응하는 정답 캡션 중에서 위 특정한 비디오 샘플에 대응하는 특정한 정답 캡션을 입력으로서 수신하는 것에 응답하여 위 비디오 인출 모델이 위 B개의 비디오 샘플 중 어느 것을 출력할 것인지를 나타냄)와, 위 제1 확률 분포 및 위 제2 확률 분포에 기반하여 보상(reward)을 계산하는 단계와, 위 보상을 사용하여 위 비디오 캡서닝 모델의 훈련에 강화 학습(Reinforcement Learning: RL)을 적용하는 단계를 포함한다.

[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 따르면, 비디오 캡서닝 모델(video captioning model)의 훈련에서 출력 캡션(caption)의 정확성을 향상하지 않고서도 출력 캡션의 변별성을 개선할 수 있다.

[0007] 본 개시에 따르면, 비디오 인출 모델(video retrieval model)을 이용하여 훈련 비디오 샘플에 대한 출력 캡션의 유사도 및 훈련 비디오 샘플에 대한 정답 캡션 간의 유사도를 비교함으로써 보상을 계산하고, 이에 따라 변별성 목적 함수(objective function)를 최적화하도록 강화 학습(Reinforcement Learning: RL)을 비디오 캡서닝 모델(video captioning model)의 훈련에 적용할 수 있다.

도면의 간단한 설명

[0008] 도 1은 비디오 캡서닝 모델(video captioning model)에 더하여 비디오 인출 모델(video retrieval model)을 포함하는 예시적인 비디오 캡서닝 프레임워크를 보여주는 블록도이다.

도 2는 도 1의 비디오 캡서닝 프레임워크에서 인출 손실을 계산하는 것을 설명하기 위한 도면이다.

도 3은 도 1의 비디오 캡서닝 프레임워크에서 최종 손실을 계산하는 것을 설명하기 위한 도면이다.

도 4는 도 1의 비디오 캡서닝 프레임워크를 이용하는 운영 환경의 예를 보여준다.

도 5는 비디오 캡서닝을 위한 프로세스의 예를 보여주는 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0009] 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 연관되게 정의되어야 한다.

[0010] 본 문서에서 용어 "포함하다", "가지다" 등은 이후에 열거된 요소, 예컨대, 어떤 특징, 숫자, 단계, 동작, 구성 요소, 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.

[0011] 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 참조

될 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.

- [0012] 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.
- [0013] 도면에 묘사된 개별 특징 중 몇몇은 서로 독립적으로 구현되거나 조합되어 구현될 수 있음이 고려된다. 다양한 예의 이들 특징 및 다른 특징은 첨부된 도면과 함께 이하의 상세한 설명으로부터 명백하게 될 것이다.
- [0014] 도 1은 예시적인 비디오 캡서닝 프레임워크(100)를 보여주는 아키텍처 도해이다.
- [0015] 도 1의 예에서, 비디오 캡서닝 프레임워크(100)는 비디오 캡서닝 모델(110)을 포함하는 캡서닝 블록(120), 비디오 인출 모델(150)을 포함하는 인출 블록(160), 비디오 캡서닝 평가 블록(140) 및 강화 학습(Reinforcement Learning: LR) 제어 블록(180)을 포함한다. 다양한 예시적인 구현에서, 후술되는 바와 같이, 비디오 캡서닝 프레임워크(100)는 단일 컴퓨팅 장치 상에 또는 여러 컴퓨팅 장치에 걸쳐 구현되고 다른 컴퓨팅 장치를 통해 액세스가능하게(accessible) 될 수 있다.
- [0016] 도시된 예에서, 캡서닝 블록(120)은 비디오 캡서닝 모델(110)을 훈련시키는 것을 포함하는 과제를 수행한다. 비디오 캡서닝 모델(110)의 훈련은 소정의 손실 함수(loss function)(가령, 후술되는 RL 손실, 최종 손실, 또는 유사한 것)를 최소화하는 방식으로, 비디오 캡서닝 모델(110)의 파라미터를 조정하는 것을 포함할 수 있다.
- [0017] 도 1에 묘사된 바와 같이, 비디오 캡서닝 모델(110)은 입력 비디오(101)를 처리하여 출력을 제공하는데, 이는 입력 비디오(101)를 위한 캡션(103)(즉, 총 D개의 단어로 이루어진 어휘(vocabulary)에 포함된 몇몇 단어의 특정 시퀀스(sequence), 예컨대, 문장(sentence))을 나타낸다. 예를 들어, 그러한 출력 캡션(103)은 예측된 단어 확률의 세트(각각의 세트는 어휘 내의 모든 단어에 걸쳐 정해짐)(가령, 예측된 단어 확률 벡터)의 시퀀스로서 나타내어질 수 있다. 만일 캡션이 20개의 단어의 시퀀스인 경우, 캡션은 어휘 내의 D개의 단어 각각에 대해 예측된 단어 확률의 적어도 20개의 세트의 시퀀스로서 표현될 수 있음(즉, 캡션의 각각의 단어에 시퀀스 내의 상이한 세트가 대응할 수 있음)이 이해될 것이다. 이들 단어 확률 세트 각각은, 예컨대, 비디오 캡서닝 모델(110)의 최종 층(final layer)(이는 D개의 출력 유닛을 가짐)으로부터 (가령, 소프트맥스(softmax) 출력에 따라) 주어지는 확률 분포일 수 있다.
- [0018] 몇몇 예에서, 비디오 캡서닝 모델(110)은 인코더-디코더 구조로 구현될 수 있다. 이 구조에서, 인코더는 비디오(101)의 프레임을 취하여 프레임 정보를 캡슐화하고(encapsulate), 디코더를 위한 초기 상태로서 전달할 수 있다. 그러면, 디코더는 초기 상태를 취하여 캡션(103)을 생성하기 시작할 수 있다.
- [0019] 도 1의 예에서, 비디오 캡서닝 평가 블록(140)은 정확성 메트릭(accuracy metric)(104)을 제공하는데, 이는 입력 비디오(101)로부터 생성된 캡션(103) 및 입력 비디오(101)에 대응하는 정답 캡션(105) 간의 대응관계 내지 유사성을 입력 비디오(101)에 대한 고려 없이 측정하는 평가 메트릭, 예컨대, CIDEr(Consensus-based Image Description Evaluation) 메트릭 또는 유사한 것이다.
- [0020] 도 1의 예에서, 인출 블록(160)은 캡션(103)의 변별성을 개선하도록 비디오 캡서닝 모델(110)을 훈련시키기 위해 도입된다.
- [0021] 도 1에 묘사된 바와 같이, 인출 블록(160)의 비디오 인출 모델(150)은 질의 텍스트(query text)의 입력에 응답하여 후보 비디오의 풀(pool)에서 질의 텍스트와 가장 잘 매칭되는 후보 비디오를 찾아 제공한다. 예를 들어, 비디오 인출 모델(150)은 그러한 풀 내의 후보 비디오를 질의 텍스트와 잘 매칭되는 순서로(예컨대, 후술되는 바와 같이 유사도 점수가 높은 순서로) 정렬하고, 정렬된 후보 비디오의 리스트를 제공할 수 있다.
- [0022] 몇몇 예에서, 도 1에 도시된 바와 같이, 비디오 인출 모델(150)은 텍스트 인코더(text encoder)(156) 및 비디오 인코더(video encoder)(157)를 포함할 수 있다. 텍스트 인코더(156)는 질의 텍스트(예컨대, 입력 비디오(101)에 대응하는 생성된 캡션(103)이든 또는 입력 비디오(101)에 대응하는 정답 캡션(105)이든)를 텍스트 및 비디오 간의 소정의 유사도 점수(similarity score)가 계산될 수 있는 공통의 공간(이는 "공동 임베딩 공간"(joint embedding space)으로도 지칭될 수 있음) 상으로 임베딩할 수 있고, 비디오 인코더(157)는 각각의 후보 비디오(가령, 입력 비디오(101))를 그러한 공간 상으로 임베딩할 수 있다. 구체적으로, 텍스트 인코더(156)는 질의 텍

스트를 공통 임베딩 공간 내의 텍스트 특징 표현(text feature representation)(106)(가령, 소정의 차원의 텍스트 특징 벡터)으로 맵핑할 수 있고, 비디오 인코더(157)는 각각의 후보 비디오를 공통 임베딩 공간 내의 비디오 특징 표현(video feature representation)(107)(가령, 동일한 차원의 비디오 특징 벡터)으로 맵핑할 수 있다. 그러면, 모든 텍스트-비디오 쌍에 대해 유사도 점수가 계산될 수 있고, 이에 따라 풀 내의 모든 후보 비디오는 질의 텍스트와 가장 관련 있는 것부터 가장 관련 없는 것까지 정렬될 수 있다. 그러한 유사도 점수 계산 및 후보 비디오 정렬은 비디오 인출 모델(150)(예컨대, 도 1에 도시된 인출 손실 계산부(158))에 의해 수행될 수 있다. 예를 들어, 인출 손실 계산부(158)는 비디오 v 에 대한 텍스트 x 의 유사도 점수를 다음과 같이 코사인 유사도(cosine similarity)로서 판정할 수 있다.

수학식 1

$$c(v, x) = \frac{(f_{TE}(x))^T f_{VE}(v)}{\|f_{TE}(x)\| \|f_{VE}(v)\|}$$

[0023]

[0024] 여기서, f_{TE} 는 텍스트 인코더(156)로부터 출력되는 텍스트 특징 표현(벡터)이고, f_{VE} 는 비디오 인코더(157)로부터 출력되는 비디오 특징 표현(벡터)이고, $\|a\|$ 는 벡터 a 의 L2 놈(norm)이고, a^T 은 벡터 a 의 전치(transpose)이다.

[0025]

비디오 인출 모델(150)은 비디오 캡셔닝 모델(110)의 훈련을 위해 이하에 기술되는 바와 같이 사용될 수 있다. 예시의 목적으로, 비디오 캡셔닝 모델(110)은 B 개의 비디오 샘플 및 각자의 대응하는 정답 캡션을 포함하는 훈련 배치(training batch)로써 훈련되는데, 입력 비디오(101)는 B 개의 비디오 샘플(이는 각각 101-1, 101-2, ..., 101- B 로 표기될 수 있음)로서 주어지고, 제 i 비디오 샘플(101- i)(단, $i=1, 2, \dots, B$)은 v_i 로 주어진다고 가정된다.

[0026]

우선, 입력 비디오(101)로서 v_i 가 제공됨에 응답하여 비디오 캡셔닝 모델(110)은 출력 캡션(103)으로서 $s_i=\{w_i^1, \dots, w_i^T\}$ 를 생성할 수 있는데, w_i^t 는 v_i 에 대응하는 생성된 캡션 s_i 가 시간 스텝(time step) t 에서 갖는 단어이다(단, $t=1, 2, \dots, T$). 도 1에 묘사된 바와 같이, 생성된 캡션 s_i 는 비디오 인출 모델(150)의 텍스트 인코더(156)에 입력될 수 있다. 그러면, 텍스트 인코더(156)는 생성된 캡션 s_i 를 공동 임베딩 공간 내의 제1 텍스트 특징 표현(이는 106-1로 표기될 수 있음)으로 맵핑할 수 있다.

[0027]

또한, 도 1에 묘사된 바와 같이, 텍스트 인코더(156)에는 v_i 에 대응하는 정답 캡션 $g_i=\{y_i^1, \dots, y_i^T\}$ 가 정답 캡션(105)으로서 입력될 수 있는데, y_i^t 는 정답 캡션 g_i 가 시간 스텝 t 에서 갖는 단어이다. 그러면, 텍스트 인코더(156)는 정답 캡션 g_i 를 공동 임베딩 공간 내의 제2 텍스트 특징 표현(이는 106-2로 표기될 수 있음)으로 맵핑할 수 있다.

[0028]

나아가, 도 1에 묘사된 바와 같이, 비디오 인출 모델(150)의 비디오 인코더(157)에는 B 개의 비디오 샘플 각각이 입력될 수 있고, 비디오 인코더(157)는 제 j 비디오 샘플 v_j 를 공동 임베딩 공간 내의 비디오 특징 표현(107- j)으로 맵핑할 수 있다(여기서 $j=1, 2, \dots, B$).

[0029]

그러면, 도 2에 도시된 바와 같이, 비디오 인출 모델(150)의 인출 손실 계산부(158)는 훈련 배치 내의 B 개의 비디오 샘플 각각에 대한 생성된 캡션 s_i 의 유사도 점수 $c(v_j, s_i)$ 및 이들 비디오 샘플 각각에 대한 정답 캡션 g_i 의 유사도 점수 $c(v_j, g_i)$ 를 판정할 수 있다. 다시 말해, 생성된 캡션 s_i 의 유사도 점수의 제1 리스트(210)가, 그리고 정답 캡션 g_i 의 유사도 점수의 제2 리스트(220)가 획득될 수 있다. 각각의 리스트(210, 220) 내의 유사도 점수는 정규화될 수 있다. 예를 들어, 그러한 정규화는, 다음 두 식에서와 같이, 소프트맥스 함수를 사용함으로써 제1 리스트(210) 및 제2 리스트(220)를 각각 제1 확률 분포(230) 및 제2 확률 분포(240)로 맵핑하는 것일 수 있다.

수학식 2

$$P(v_j, s_i) = \frac{\exp(c(v_j, s_i))}{\sum_{j=1}^B \exp(c(v_j, s_i))}$$

[0030]

수학식 3

$$P(v_j, g_i) = \frac{\exp(c(v_j, g_i))}{\sum_{j=1}^B \exp(c(v_j, g_i))}$$

[0031]

[0032]

여기서 $\exp(\cdot)$ 는 지수 함수(exponential function)이다. 수학식 2에서 볼 수 있는 바와 같이, 각각의 비디오 샘플 v_j 에 대한 생성된 캡션 s_i 의 유사도 점수의 리스트(210)는 s_i 의 입력에 응답하여 비디오 인출 모델(150)이 v_j 를 출력할 확률 $P(v_j, s_i)$ 의 리스트인 확률 분포(230)로 맵핑될 수 있다. 마찬가지로, 수학식 3에서 볼 수 있는 바와 같이, 각각의 비디오 샘플 v_j 에 대한 정답 캡션 g_i 의 유사도 점수의 리스트(220)는 g_i 의 입력에 응답하여 비디오 인출 모델(150)이 v_j 를 출력할 확률 $P(v_j, g_i)$ 의 리스트인 확률 분포(240)로 맵핑될 수 있다.

[0033]

이들 두 확률 분포(230, 240)가 획득되므로, 확률 $P(v_i, s_i)$ (즉, 정규화된 유사도 점수 $c(v_i, s_i)$) 및 확률 $P(v_i, g_i)$ (즉, 정규화된 유사도 점수 $c(v_i, g_i)$) 간의 차이를 나타내는 손실 함수, 예컨대, 교차 엔트로피 손실(cross-entropy loss), 쿨백-라이블러 발산(Kullback-Leibler divergence), 또는 유사한 것이 훈련에 사용될 수 있다. 이와 관련하여, 예시로서, 인출 손실 계산부(158)는 다음과 같이 인출 손실(108) $L_{\text{retrieval}}$ 을 계산할 수 있다.

수학식 4

$$L_{\text{retrieval}}(s_i, g_i) = P(v_i, g_i) \log(P(v_i, s_i))$$

[0034]

[0035]

다음과 같은 점을 고려할 때, 인출 손실(108)은 캡션의 변별성이라는 측면에서 비디오 캡셔닝 모델(110)의 훈련에 대한 지침으로서 제공될 수 있다: 만일 생성된 캡션이 변별성이 있다면, 비디오 인출 모델(150)은 원래의 입력 비디오를 도로 성공적으로 인출할 것이며, 이는 입력 비디오가 그것의 대응하는 정답 캡션과 유사한 만큼 입력 비디오로부터 생성된 출력 캡션이 입력 비디오와 유사할수록 출력 캡션의 변별성이 높다는 것을 의미한다.

[0036]

이제, 도 1과 함께 도 3을 참조하여, 비디오 캡셔닝 모델(110)의 훈련이 상세히 논의된다.

[0037]

몇몇 예에서, 훈련 배치 내의 각각의 비디오 샘플 v_i 및 이에 대응하는 정답 캡션 g_i 에 대해, 비디오 캡셔닝 모델(110)은 각각의 시간 스텝 t 에서 어휘 내의 총 D 개의 단어에 걸쳐 확률 분포를 산출할 수 있다. 예를 들어, 비디오 캡셔닝 모델(110)은 비디오 샘플 v_i 및 정답 캡션 g_i 의 이전 단어 $\{y_i^1, \dots, y_i^{t-1}\}$ (이는 $y_i^{1:t-1}$ 로 표기될 수도 있음)에 기반하여 어휘 내의 단어 c_d 의 (정규화되지 않은) 예측 값 $m_{d,i}^t$ 를 계산할 수 있고, 이 예측 값은(예컨대, 다음 식에서와 같이 소프트맥스 함수를 사용하여) 예측된 단어 확률로 정규화될 수 있다.

수학식 5

$$p(c_d|y_i^{1:t-1}, \theta, v_i) = \frac{\exp(m_{d,i}^t)}{\sum_{r=1}^D \exp(m_{r,i}^t)}$$

[0038]

[0039] 여기서 θ 는 비디오 캡서닝 모델(110)의 파라미터를 나타낸다.

[0040] 몇몇 예에서, 비디오 캡서닝 모델(110)의 훈련에서의 사용을 위해 캡서닝 블록(120)은 다음과 같은 음의 로그 우도(Negative Log-Likelihood: NLL) 손실(330)을 계산할 수 있다.

수학식 6

$$L_{NLL}(\theta) = -\frac{1}{B} \sum_{i=1}^B \sum_{t=1}^T \log p(W_i^t = y_i^t | y_i^{1:t-1}, \theta, v_i)$$

[0041]

[0042] 여기서 W_i^t 는 비디오 샘플 v_i 에 대해 시간 스텝 t 에서 주어진 단어를 나타내는 랜덤 변수(random variable)이다.

수학식 6 내의 로그 항은 예측된 단어 확률 벡터 $[p(c_1|y_i^{1:t-1}, \theta, v_i), p(c_2|y_i^{1:t-1}, \theta, v_i), \dots, p(c_D|y_i^{1:t-1}, \theta, v_i)]$ 에 로그 함수(log function)를 적용함으로써 생성된 로그 확률 및 정답 캡션 g_i 가 시간 스텝 t 에서 갖는 정답 단어의 원핫 인코딩 벡터(one-hot encoding vector) 간의 내적(inner product) 연산으로 볼 수 있다.

[0043] 또한, 몇몇 예에서, RL 제어 블록(180)은 비디오 캡서닝 모델(110)의 훈련에 RL을 적용하는 소정의 동작을 수행할 수 있는데, 이는 비디오 캡서닝 모델(110)의 훈련에서 캡션(103)을 생성하는 것은 미분불가능한 연산을 수반한다는 점을 감안한 것이다. 이 RL에서, 비디오 캡서닝 모델(110)은 에이전트(agent)로서 작동하고, 비디오 샘플 v_i 은 상태(state)이며, 비디오 샘플 v_i 및 대응하는 정답 캡션 g_i 를 포함하는 환경에서 에이전트는 정책 π_θ 를 사용하여 확률 분포를 산출한다. 예를 들어, 각각의 시간 스텝 t 에서, 에이전트는 확률 분포 출력으로부터의 샘플링에 의해 단어를 선택하고, 캡션 생성이 완료되면 보상(reward)(310) r 을 받을 수 있다. 이 예에서, 각각의 비디오 샘플 v_i 에 대해 최적화될 목적 함수(objective function)은 다음과 같이 음의 기대되는 보상(negative expected reward)을 최소화하는 것일 수 있다.

수학식 7

$$L_{RL,i}(\theta) = -E_{W_i^1, \dots, W_i^T \sim \pi_{\theta|v_i}} [r(W_i^1, \dots, W_i^T)]$$

[0044]

[0045] 여기서 $\pi_{\theta|v_i}$ 는 v_i 에 대해 비디오 캡서닝 모델(110)에 의해 출력되는 모든 단어에 걸친 확률 분포이고,

$W_i^1, \dots, W_i^T \sim \pi_{\theta|v_i}$ 는 시간 스텝 t 에서 $\pi_{\theta|v_i}$ 로부터 샘플링된 W_i^t 를 나타내고, $r(W_i^1, \dots, W_i^T)$ 는 그러한 샘플링을 통해 형성된 캡션을 평가하는 보상 함수이다. 예를 들어, REINFORCE 알고리즘의 구현에서와 같이, 비디오 캡서닝 모델(110)의 최적의 파라미터를 구하기 위해 수학식 7에 정의된 것의 그라디언트(gradient)

를 사용할 수 있다. 이 그래디언트는, 다음과 같이, 각각의 비디오 샘플 v_i 에 대해 $\pi_{\theta|v_i}$ 로부터 오직 $s_i=\{w_i^1, \dots, w_i^T\}$ 를 샘플링함으로써 근사화될 수 있다.

수학식 8

$$\nabla_{\theta} L_{RL,i}(\theta) \approx -r(w_i^1, \dots, w_i^T) \nabla_{\theta} \log \pi_{\theta|v_i}(w_i^1, \dots, w_i^T)$$

각각의 단어 w_i^t 는, 다음과 같이, $\pi_{\theta|v_i}$ 에 의해 정의된 확률 분포 출력으로부터 가장 높은 확률을 갖는 단어로써 취함으로써 샘플링될 수 있다.

수학식 9

$$W_i^t = c_{d_{i,t}}$$

여기서 $d_{i,t}$ 는 비디오 샘플 v_i 에 대해 시간 스텝 t 에서 비디오 캡서닝 모델(110)에 의해 출력된 확률 분포에서 가장 높은 확률을 갖는, 어휘 내의 단어의 인덱스를 나타낸다.

수학식 10

$$d_{i,t} = \operatorname{argmax}_{d \in D} p(c_d | y_i^{1:t-1}, \theta, v_i)$$

수학식 8의 그래디언트를 계산하는 것은 다음과 같이 수학식 7에 대한 의사 손실(pseudo loss) $\hat{L}_{RL,i}$ 을 정의하고 이를 (예컨대, 수치 미분계산을 위해 이용가능한 autograd 라이브러리와 같은 자동 미분 툴을 사용하여) 계산함으로써 수행될 수 있고, 이는 수동 방식의 그래디언트 계산에 비해 더 효율적일 수 있다.

수학식 11

$$\hat{L}_{RL,i}(\theta) = -(r(w_i^1, \dots, w_i^T) \log \pi_{\theta|v_i}(w_i^1, \dots, w_i^T))$$

여기서 보상(310) $r(w_i^1, \dots, w_i^T) = r(s_i)$ 는 캡션 s_i 의 샘플링에 대해 주어지고, 인출 손실(108) $L_{\text{retrieval}}(s_i, g_i)$ 에 기반하여, 그리고 가능하게는 정확성 메트릭(104)에 또한 기반하여, (가령, 아래의 수학식 15에서와 같이) 정의되며, $\pi_{\theta|v_i}(w_i^1, \dots, w_i^T)$ 는 다음과 같이 주어진다.

수학식 12

$$\pi_{\theta|v_i}(w_i^1, \dots, w_i^T) = p(w_i^1, \dots, w_i^T | y_i^{1:T-1}, \theta, v_i)$$

예를 들어, 비디오 샘플 v_i 에 대해 캡션 $s_i = \{w_i^1, \dots, w_i^T\}$ 의 샘플링이 발생할 조건부 확률 $p(w_i^1, \dots, w_i^T | y_i^{1:t-1}, \theta, v_i)$ 는, 각각의 단어의 샘플링이 발생할 확률 $p(w_i^t)$ 가 역시 묵시적으로 이전의 정답 단어 $y_i^{1:t-1}$ 을 조건부로 한 것임을 감안하여, 다음과 같이 표현될 수 있다.

수학식 13

$$p(w_i^1, \dots, w_i^T | y_i^{1:T-1}, \theta, v_i) = \prod_{t=1}^T p(W_i^t = c_{d_{i,t}} | y_i^{1:t-1}, \theta, v_i)$$

그러면, 훈련 배치에 걸쳐 RL 손실(320) $\tilde{L}_{RL}$ 이 다음과 같이 정의될 수 있다.

수학식 14

$$\tilde{L}_{RL}(\theta) = -\frac{1}{B} \sum_{i=1}^B \hat{L}_{RL,i}(\theta)$$

RL 제어 블록(180)은 보상(310)이 반영된 RL 손실(320)을 계산하여 이를 비디오 캡셔닝 모델(110)의 훈련에서의 사용을 위해 제공할 수 있다. 이를 위해, RL 제어 블록(180)은 다음과 같이 보상(310)을 계산할 수 있다.

수학식 15

$$r(s_i) = CIDEr(s_i, g_i) - \lambda L_{retrieval}(s_i, g_i)$$

여기서 $L_{retrieval}(s_i, g_i)$ 은 수학식 4에 정의된 것이고, $CIDEr(s_i, g_i)$ 는 s_i 및 g_i 간의 CIDEr 메트릭이며, λ 는 $L_{retrieval}(s_i, g_i)$ 및 $CIDEr(s_i, g_i)$ 간의 밸런싱(balancing)을 위한 트레이드오프 하이퍼파라미터(tradeoff hyperparameter)이다. 인출 손실(108) $L_{retrieval}(s_i, g_i)$ 은 인출 손실 계산부(158)로부터 RL 제어 블록(180)에 제공될 수 있고, $CIDEr(s_i, g_i)$ 는 비디오 캡셔닝 평가 블록(140)으로부터 정확성 메트릭(104)으로서 RL 제어 블록(180)에 제공될 수 있다. 이와 같이, 인출 손실(108) 및 정확성 메트릭(104)을 보상(310)으로 조합하는 것은 생성된 캡션의 변별성을 개선하기 위해 비디오 캡셔닝 모델(110)의 훈련에서 인출 손실(108)을 사용할 뿐만 아니라, 그러한 캡션의 정확성을 유지하기 위해, 다시 말해, 그러한 캡션이 인간의 자연스러운 언어처럼 유창하게 하기 위해, CIDEr 메트릭과 같은 정확성 메트릭(104)을 사용하게 하며, 이를 RL 방식으로 비디오 캡셔닝 모델(110)의 훈련에 적용하는 것이다.

나아가, 비디오 캡셔닝 모델(110)의 훈련을 더욱 안정적이도록 보장하기 위해, 총 비디오 캡셔닝 손실(350)(이

는 최종 손실로도 지칭됨) L_{final} 이 다음과 같이 동일한 훈련 배치에 대해 계산된 RL 손실(320) $\tilde{L}_{RL}$ 및 NLL 손실(330) L_{NLL} 의 가중화된 조합(weighted combination)으로서 계산될 수 있다.

수학식 16

$$L_{final}(\theta) = (1 - \alpha)L_{NLL}(\theta) + \alpha\tilde{L}_{RL}(\theta)$$

여기서 α 는 L_{NLL} 및 $\tilde{L}_{RL}$ 간의 밸런싱을 위한 트레이드오프 하이퍼파라미터이다.

도 4는 비디오 캡서닝 프레임워크(100)을 이용하는 운영 환경(400)의 예를 보여준다.

예시적인 운영 환경(400)에서, 비디오 캡서닝 프레임워크(100)는 정보 처리 시스템, 예컨대, 컴퓨팅 장치(410) 상에 구현된다. 컴퓨팅 장치(410)는 버스를 통해 서로 통신가능하게 커플링된 프로세서(processor)(411), 메모리(memory)(412) 및 통신 인터페이스(413)를 포함한다.

도시된 예에서, 프로세서(411)는 컴퓨팅 장치(410)의 전반적인 동작을 제어한다. 예를 들어, 프로세서(411)는 (가령, 비디오 인출 모델(150)을 사용하여 보상을 판정하고 이를 기반으로 RL을 적용함으로써 비디오 캡서닝 모델(110)을 훈련시키는 것과 관련하여) 본 문서에 기술된 동작 중 전부 또는 일부를 수행할 수 있다.

도시된 예에서, 메모리(412)는 다양한 정보를 저장한다. 예를 들어, 메모리(412)는 비디오 캡서닝 프레임워크(100)를 구현하고/거나 가동하는 데에서 사용되고/거나 산출되는 정보, 예컨대, 컴퓨팅 장치(410)의 동작 중에 사용되는 데이터 및/또는 컴퓨팅 장치(410)(가령, 프로세서(411))에 의해 실행되는 명령어의 세트를 비일시적(non-transitory) 형태로 저장할 수 있다.

도시된 예에서, 통신 인터페이스(413)는 컴퓨팅 장치(410)로 하여금 다른 장치(가령, 다른 컴퓨팅 장치(420))에 데이터를 송신하고 이로부터 데이터를 수신하기 위해 네트워크(430)(가령, 인터넷)를 액세스할 수 있게 한다. 예를 들어, 컴퓨팅 장치(410)는, 통신 인터페이스(413)를 통해, 컴퓨팅 장치(420)(가령, 모바일 컴퓨팅 장치, 개인용 컴퓨터 또는 유사한 것과 같은 클라이언트 장치)로부터 비디오 캡서닝 모델(110)에 대한 요청을 수신하고, 전송된 바와 같이 훈련된 비디오 캡서닝 모델(110)을 컴퓨팅 장치(420)에 제공할 수 있다.

이제, 도 5를 참조하여, 비디오 캡서닝을 위한 예시적인 프로세스(500)가 설명된다. 예를 들어, 프로세스(500)는 비디오 캡서닝 프레임워크(100)가 구현된 컴퓨팅 장치(가령, 컴퓨팅 장치(410))에 의해 수행될 수 있다. 프로세스(500)의 다른 예시적인 흐름이 또한 고려된다. 예를 들어, 프로세스(500)는 도시되지 않은 추가적인 동작을 또한 포함할 수 있고/거나, 도 5와 관련하여 예시된 동작 중 일부를 포함하나 전부를 포함하지는 않을 수 있다.

동작(510)에서, 비디오 캡서닝 모델(가령, 비디오 캡서닝 모델(110))의 훈련을 위해, 비디오 캡서닝 모델을 사용하여 훈련 배치 내의 B개의 비디오 샘플 중 임의의 특정한 것(가령, v_i)에 대해 캡션(가령, s_i)이 생성된다. 훈련 배치 내에는 B개의 비디오 샘플에 각각 대응하는 B개의 정답 캡션이 포함될 수 있다.

동작(520)에서, 비디오 인출 모델(가령, 비디오 인출 모델(150))을 사용하여 제1 확률 분포(가령, 제1 확률 분포(230)) 및 제2 확률 분포(가령, 제2 확률 분포(240))가 판정되는데, 제1 확률 분포는 생성된 캡션을 입력으로서 수신하는 것에 응답하여 비디오 인출 모델이 B개의 비디오 샘플 중 어느 것을 출력할 것인지를 나타내고, 제2 확률 분포는 특정한 비디오 샘플에 대응하는 특정한 정답 캡션(가령, g_i)을 입력으로서 수신하는 것에 응답하여 비디오 인출 모델이 B개의 비디오 샘플 중 어느 것을 출력할 것인지를 나타낸다.

몇몇 예에서, 비디오 인출 모델은 생성된 캡션의 유사도 점수의 제1 리스트(가령, 제1 리스트(210)) 및 특정한 정답 캡션의 유사도 점수의 제2 리스트(가령, 제2 리스트(220))를 판정하고, 제1 및 제2 리스트에 각각 기반하여 제1 및 제2 확률 분포를 판정할 수 있다. 예를 들어, 제1 리스트 내의 유사도 점수 각각은 B개의 비디오 샘플

플 중 각자의 것이 생성된 캡션과 어느 정도로 매칭되는지를 나타낼 수 있고(가령, $c(v_j, s_i)$), 이는 생성된 캡션의 입력에 응답하여 비디오 인출 모델에 의해 각자의 비디오 샘플이 출력될 확률(가령, $P(v_j, s_i)$)로 정규화될 수 있다. 제2 리스트 내의 유사도 점수 각각은 B개의 비디오 샘플 중 각자의 것이 특정한 정답 캡션과 어느 정도로 매칭되는지를 나타낼 수 있고(가령, $c(v_j, g_i)$), 이는 특정한 정답 캡션의 입력에 응답하여 비디오 인출 모델에 의해 각자의 비디오 샘플이 출력될 확률(가령, $P(v_j, g_i)$)로 정규화될 수 있다.

[0074] 동작(530)에서, 제1 확률 분포 및 제2 확률 분포에 기반하여 보상이 계산된다.

[0075] 몇몇 예에서, 비디오 인출 모델(가령, 비디오 인출 모델(150)의 인출 손실 계산부(158))에 의해 다음 두 확률 간의 차이를 측정하는 방식으로 인출 손실(가령, 수학적 식 4에 정의된 바와 같은 인출 손실(108))이 계산될 수 있다: 생성된 캡션의 입력에 응답하여 비디오 인출 모델에 의해 특정한 비디오 샘플이 출력될 확률(가령, $P(v_i, s_i)$) 및 특정한 정답 캡션의 입력에 응답하여 비디오 인출 모델에 의해 특정한 비디오 샘플이 출력될 확률(가령, $P(v_i, g_i)$). 그러면, 인출 손실에 기반하여 보상(가령, 수학적 식 15에 정의된 바와 같은 보상(310))이 (가령, RL 제어 블록(180)에 의해) 계산될 수 있다. 예로서, 그러한 보상은 인출 손실뿐만 아니라, CIDEr 메트릭과 같은 정확성 메트릭(가령, B개의 비디오 샘플 중 어느 것도 고려하지 않고서 비디오 캡셔닝 평가 블록(140)에 의해 계산된 정확성 메트릭(104))에도 기반할 수 있다.

[0076] 동작(540)에서, 보상을 사용하여 비디오 캡셔닝 모델의 훈련에 강화 학습(Reinforcement Learning: RL)이 적용된다.

[0077] 몇몇 예에서, 동작(530)에서 계산된 보상에 기반하여, 그리고 특정한 비디오 샘플에 대한 생성된 캡션의 발생의 조건부 확률(가령, 수학적 식 12에 정의된 확률)에 또한 기반하여, RL 손실(가령, 수학적 식 14에 정의된 바와 같은 RL 손실(320))이 (가령, RL 제어 블록(180)에 의해) 계산될 수 있다. 이 RL 손실은 비디오 캡셔닝 모델의 훈련에서 비디오 캡셔닝 모델에 파라미터를 조정하는 데에 사용될 수 있다. 예를 들어, 그러한 조정은 RL 손실뿐만 아니라, B개의 비디오 샘플에 걸쳐서 정답 캡션 및 생성된 캡션 간의 차이를 나타내는 NLL 손실(가령, 수학적 식 6에 정의된 바와 같은 NLL 손실(330))에 기반하여, 예컨대, RL 손실 및 NLL 손실의 가중 합(weighted sum)(가령, 수학적 식 16에 정의된 최종 손실)이 최소화되도록 행해질 수 있다.

[0078] 다음은 비디오 캡셔닝에 관련된 다양한 예이다.

[0079] 예 1에서, 비디오 캡셔닝(video captioning)을 위한 방법은, B개의 비디오 샘플 및 각자의 대응하는 정답(ground-truth) 캡션을 포함하는 훈련 배치(training batch)로써 비디오 캡셔닝 모델을 훈련시키는 단계를 포함하되, 위 비디오 캡셔닝 모델을 훈련시키는 단계는, 위 비디오 캡셔닝 모델을 사용하여 위 B개의 비디오 샘플 중 임의의 특정한 것에 대해 캡션을 생성하는 단계와, 비디오 인출 모델(video retrieval model)을 사용하여 제1 확률 분포 및 제2 확률 분포를 판정하는 단계(위 제1 확률 분포는 위 생성된 캡션을 입력으로서 수신하는 것에 응답하여 위 비디오 인출 모델이 위 B개의 비디오 샘플 중 어느 것을 출력할 것인지를 나타내고, 위 제2 확률 분포는 위 각자의 대응하는 정답 캡션 중에서 위 특정한 비디오 샘플에 대응하는 특정한 정답 캡션을 입력으로서 수신하는 것에 응답하여 위 비디오 인출 모델이 위 B개의 비디오 샘플 중 어느 것을 출력할 것인지를 나타냄)와, 위 제1 확률 분포 및 위 제2 확률 분포에 기반하여 보상(reward)을 계산하는 단계와, 위 보상을 사용하여 위 비디오 캡셔닝 모델의 훈련에 강화 학습(Reinforcement Learning: RL)을 적용하는 단계를 포함한다.

[0080] 예 2는 예 1의 주제를 포함하는데, 위 비디오 인출 모델은, 위 생성된 캡션의 유사도 점수의 제1 리스트 및 위 특정한 정답 캡션의 유사도 점수의 제2 리스트를 판정하고(위 제1 리스트 내의 위 유사도 점수 각각은 위 B개의 비디오 샘플 중 각자의 것이 위 생성된 캡션과 어느 정도로 매칭되는지를 나타내고, 위 제2 리스트 내의 위 유사도 점수 각각은 위 B개의 비디오 샘플 중 각자의 것이 위 특정한 정답 캡션과 어느 정도로 매칭되는지를 나타냄), 위 제1 리스트에 기반하여 위 제1 확률 분포를 판정하고, 위 제2 리스트에 기반하여 위 제2 확률 분포를 판정하는 데에 사용된다.

[0081] 예 3은 예 2의 주제를 포함하는데, 위 제1 확률 분포를 판정하는 것은 위 제1 리스트 내의 위 유사도 점수를 위 제1 확률 분포로 정규화하는 것을 포함하고, 위 제2 확률 분포를 판정하는 것은 위 제2 리스트 내의 위 유사도 점수를 위 제2 확률 분포로 정규화하는 것을 포함한다.

[0082] 예 4는 예 2 또는 예 3의 주제를 포함하는데, 위 비디오 인출 모델은 텍스트 인코더(text encoder) 및 비디오 인코더(video encoder)를 포함하고, 위 텍스트 인코더는 위 생성된 캡션 및 위 특정한 정답 캡션을 각각 공동

임베딩 공간(joint embedding space) 내의 제1 텍스트 특징 표현(text feature representation) 및 위 공동 임베딩 공간 내의 제2 텍스트 특징 표현으로 맵핑하고, 위 비디오 인코더는 위 각자의 비디오 샘플을 위 공동 임베딩 공간 내의 대응하는 비디오 특징 표현(video feature representation)으로 맵핑하고, 위 제1 리스트 내의 위 각각의 유사도 점수는 위 대응하는 비디오 특징 표현에 대한 위 제1 텍스트 특징 표현의 코사인 유사도(cosine similarity)이고, 위 제2 리스트 내의 위 각각의 유사도 점수는 위 대응하는 비디오 특징 표현에 대한 위 제2 텍스트 특징 표현의 코사인 유사도이다.

- [0083] 예 5는 예 1 내지 예 4 중 임의의 것의 주제를 포함하는데, 위 보상을 계산하는 단계는, 위 제1 확률 분포 및 위 제2 확률 분포에 기반하여 인출 손실을 계산하는 단계와, 위 B개의 비디오 샘플 중 어느 것도 고려하지 않고서 위 생성된 캡션 및 위 특정한 정답 캡션에 기반하여 정확성 메트릭(accuracy metric)을 계산하는 단계와, 위 인출 손실 및 위 정확성 메트릭을 위 보상으로 조합하는 단계를 포함한다.
- [0084] 예 6은 예 5의 주제를 포함하는데, 위 인출 손실을 계산하는 단계는 위 생성된 캡션의 입력에 응답하여 위 비디오 인출 모델에 의해 위 특정한 비디오 샘플이 출력되는 확률 및 위 특정한 정답 캡션의 입력에 응답하여 위 비디오 인출 모델에 의해 위 특정한 비디오 샘플이 출력되는 확률 간의 차이를 측정하는 단계를 포함한다.
- [0085] 예 7은 예 1 내지 예 6 중 임의의 것의 주제를 포함하는데, 위 보상을 사용하여 위 비디오 캡서닝 모델의 훈련에 RL을 적용하는 단계는, 위 특정한 비디오 샘플에 대해 위 생성된 캡션이 발생할 조건부 확률 및 위 보상에 기반하여 RL 손실을 계산하는 단계와, 위 RL 손실을 위 비디오 캡서닝 모델의 훈련에 적용하는 단계를 포함한다.
- [0086] 예 8은 예 7의 주제를 포함하는데, 위 캡션은 예측된 단어 확률의 세트의 시퀀스에 의해 나타내어지고, 위 비디오 캡서닝 모델을 훈련시키는 단계는, 예측된 단어 확률의 세트의 위 시퀀스 및 위 특정한 정답 캡션에 기반하여 음의 로그 우도(Negative Log-Likelihood: NLL) 손실을 계산하는 단계와, 위 RL 손실 및 위 NLL 손실에 기반하여 위 비디오 캡서닝 모델의 파라미터를 조정하는 단계를 더 포함한다.
- [0087] 예 9는 예 8의 주제를 포함하는데, 위 조정은 위 RL 손실 및 위 NLL 손실의 가중 합(weighted sum)인 최종 손실이 최소화되도록 행해진다.
- [0088] 예 10은 예 8 또는 예 9의 주제를 포함하는데, 위 NLL 손실을 계산하는 것은 위 시퀀스 내의 각각의 세트 및 위 정답 캡션의 대응하는 단어의 원핫 벡터(one-hot vector) 간의 차이를 측정하는 것을 포함한다.
- [0089] 예 11에서, 비디오 캡서닝을 위한 장치는, 프로세서와, 위 프로세서에 의해 실행되는 경우에 위 프로세서로 하여금 B개의 비디오 샘플 및 각자의 대응하는 정답 캡션을 포함하는 훈련 배치로써 비디오 캡서닝 모델을 훈련시키는 동작을 수행하게 하는 명령어가 저장된 컴퓨터 판독가능 저장 매체를 포함하되, 위 동작은, 위 비디오 캡서닝 모델을 사용하여 위 B개의 비디오 샘플 중 임의의 특정한 것에 대해 캡션을 생성하는 것과, 비디오 인출 모델을 사용하여 제1 확률 분포 및 제2 확률 분포를 판정하는 것(위 제1 확률 분포는 위 생성된 캡션을 입력으로서 수신하는 것에 응답하여 위 비디오 인출 모델이 위 B개의 비디오 샘플 중 어느 것을 출력할 것인지를 나타내고, 위 제2 확률 분포는 위 각자의 대응하는 정답 캡션 중에서 위 특정한 비디오 샘플에 대응하는 특정한 정답 캡션을 입력으로서 수신하는 것에 응답하여 위 비디오 인출 모델이 위 B개의 비디오 샘플 중 어느 것을 출력할 것인지를 나타냄)과, 위 제1 확률 분포 및 위 제2 확률 분포에 기반하여 보상을 계산하는 것과, 위 보상을 사용하여 위 비디오 캡서닝 모델의 훈련에 강화 학습(Reinforcement Learning: RL)을 적용하는 것을 포함한다.
- [0090] 예 12는 예 11의 주제를 포함하는데, 위 비디오 인출 모델은, 위 생성된 캡션의 유사도 점수의 제1 리스트 및 위 특정한 정답 캡션의 유사도 점수의 제2 리스트를 판정하고(위 제1 리스트 내의 위 유사도 점수 각각은 위 B개의 비디오 샘플 중 각자의 것이 위 생성된 캡션과 어느 정도로 매칭되는지를 나타내고, 위 제2 리스트 내의 위 유사도 점수 각각은 위 B개의 비디오 샘플 중 각자의 것이 위 특정한 정답 캡션과 어느 정도로 매칭되는지를 나타냄), 위 제1 리스트에 기반하여 위 제1 확률 분포를 판정하고, 위 제2 리스트에 기반하여 위 제2 확률 분포를 판정하는 데에 사용된다.
- [0091] 예 13은 예 12의 주제를 포함하는데, 위 제1 확률 분포를 판정하는 것은 위 제1 리스트 내의 위 유사도 점수를 위 제1 확률 분포로 정규화하는 것을 포함하고, 위 제2 확률 분포를 판정하는 것은 위 제2 리스트 내의 위 유사도 점수를 위 제2 확률 분포로 정규화하는 것을 포함한다.
- [0092] 예 14는 예 12 또는 예 13의 주제를 포함하는데, 위 비디오 인출 모델은 텍스트 인코더 및 비디오 인코더를 포함하고, 위 텍스트 인코더는 위 생성된 캡션 및 위 특정한 정답 캡션을 각각 공동 임베딩 공간 내의 제1 텍스트 특징 표현 및 위 공동 임베딩 공간 내의 제2 텍스트 특징 표현으로 맵핑하고, 위 비디오 인코더는 위 각자의 비

디오 샘플을 위 공동 임베딩 공간 내의 대응하는 비디오 특징 표현으로 맵핑하고, 위 제1 리스트 내의 위 각각의 유사도 점수는 위 대응하는 비디오 특징 표현에 대한 위 제1 텍스트 특징 표현의 코사인 유사도이고, 위 제2 리스트 내의 위 각각의 유사도 점수는 위 대응하는 비디오 특징 표현에 대한 위 제2 텍스트 특징 표현의 코사인 유사도이다.

- [0093] 예 15는 예 11 내지 예 14 중 임의의 것의 주제를 포함하는데, 위 보상을 계산하는 것은, 위 제1 확률 분포 및 위 제2 확률 분포에 기반하여 인출 손실을 계산하는 것과, 위 B개의 비디오 샘플 중 어느 것도 고려하지 않고서 위 생성된 캡션 및 위 특정한 정답 캡션에 기반하여 정확성 메트릭(accuracy metric)을 계산하는 것과, 위 인출 손실 및 위 정확성 메트릭을 위 보상으로 조합하는 것을 포함한다.
- [0094] 예 16은 예 15의 주제를 포함하는데, 위 인출 손실을 계산하는 것은 위 생성된 캡션의 입력에 응답하여 위 비디오 인출 모델에 의해 위 특정한 비디오 샘플이 출력되는 확률 및 위 특정한 정답 캡션의 입력에 응답하여 위 비디오 인출 모델에 의해 위 특정한 비디오 샘플이 출력되는 확률 간의 차이를 측정하는 것을 포함한다.
- [0095] 예 17은 예 11 내지 예 16 중 임의의 것의 주제를 포함하는데, 위 보상을 사용하여 위 비디오 캡서닝 모델의 훈련에 RL을 적용하는 것은, 위 특정한 비디오 샘플에 대해 위 생성된 캡션이 발생할 조건부 확률 및 위 보상에 기반하여 RL 손실을 계산하는 것과, 위 RL 손실을 위 비디오 캡서닝 모델의 훈련에 적용하는 것을 포함한다.
- [0096] 예 18은 예 17의 주제를 포함하는데, 위 캡션은 예측된 단어 확률의 세트의 시퀀스에 의해 나타내어지고, 위 비디오 캡서닝 모델을 훈련시키는 것은, 예측된 단어 확률의 세트의 위 시퀀스 및 위 특정한 정답 캡션에 기반하여 음의 로그 우도(Negative Log-Likelihood: NLL) 손실을 계산하는 것과, 위 RL 손실 및 위 NLL 손실에 기반하여 위 비디오 캡서닝 모델의 파라미터를 조정하는 것을 더 포함한다.
- [0097] 예 19는 예 18의 주제를 포함하는데, 위 조정은 위 RL 손실 및 위 NLL 손실의 가중 합인 최종 손실이 최소화되도록 행해진다.
- [0098] 예 20은 예 18 또는 예 19의 주제를 포함하는데, 위 NLL 손실을 계산하는 것은 위 시퀀스 내의 각각의 세트 및 위 정답 캡션의 대응하는 단어의 원핫 벡터 간의 차이를 측정하는 것을 포함한다.
- [0099] 예 21에서, 컴퓨터 프로세서에 의해 실행되는 경우에 위 컴퓨터 프로세서로 하여금 예 1 내지 예 10 중 임의의 것의 방법을 수행하게 하는 컴퓨터 실행가능 명령어가 저장된 컴퓨터 판독가능 저장 매체가 제공된다.
- [0100] 특정한 예에서, 본 문서에서 언급된 장치, 디바이스, 시스템, 머신 등은 임의의 적합한 유형은 컴퓨팅 장치이거나, 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.
- [0101] 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.

[0102] 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 대체 또는 치환되더라도 적절한 결과가 달성될 수 있다.

[0103] 그러므로, 본 개시의 범주는 개시된 그 형태에 한정되어서는 안 되며, 후술하는 청구항 및 이의 균등물에 의해 정해져야 한다.

부호의 설명

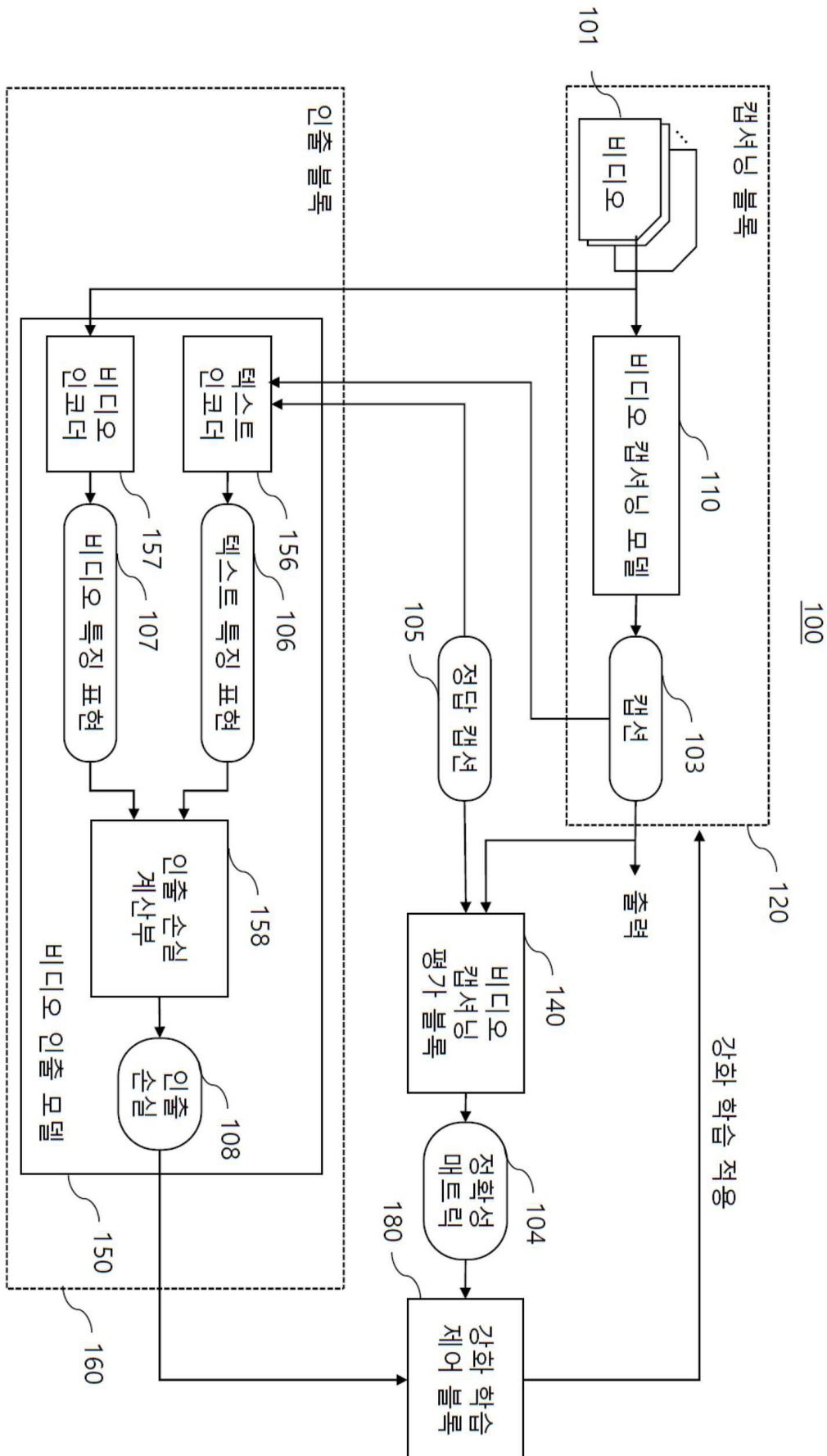
[0104] 100: 비디오 캡서닝 프레임워크

110: 비디오 캡서닝 모델

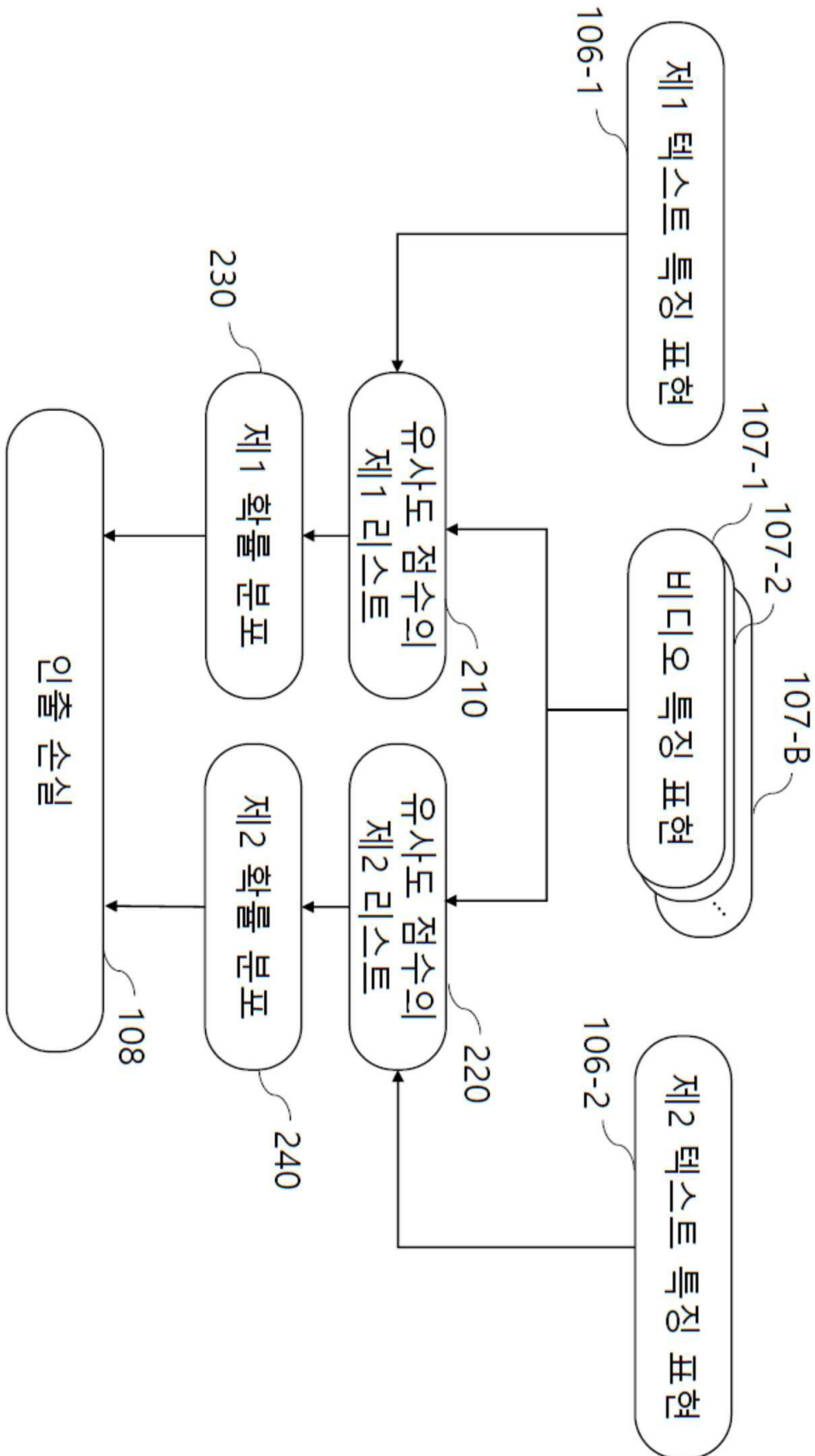
150: 비디오 인출 모델

도면

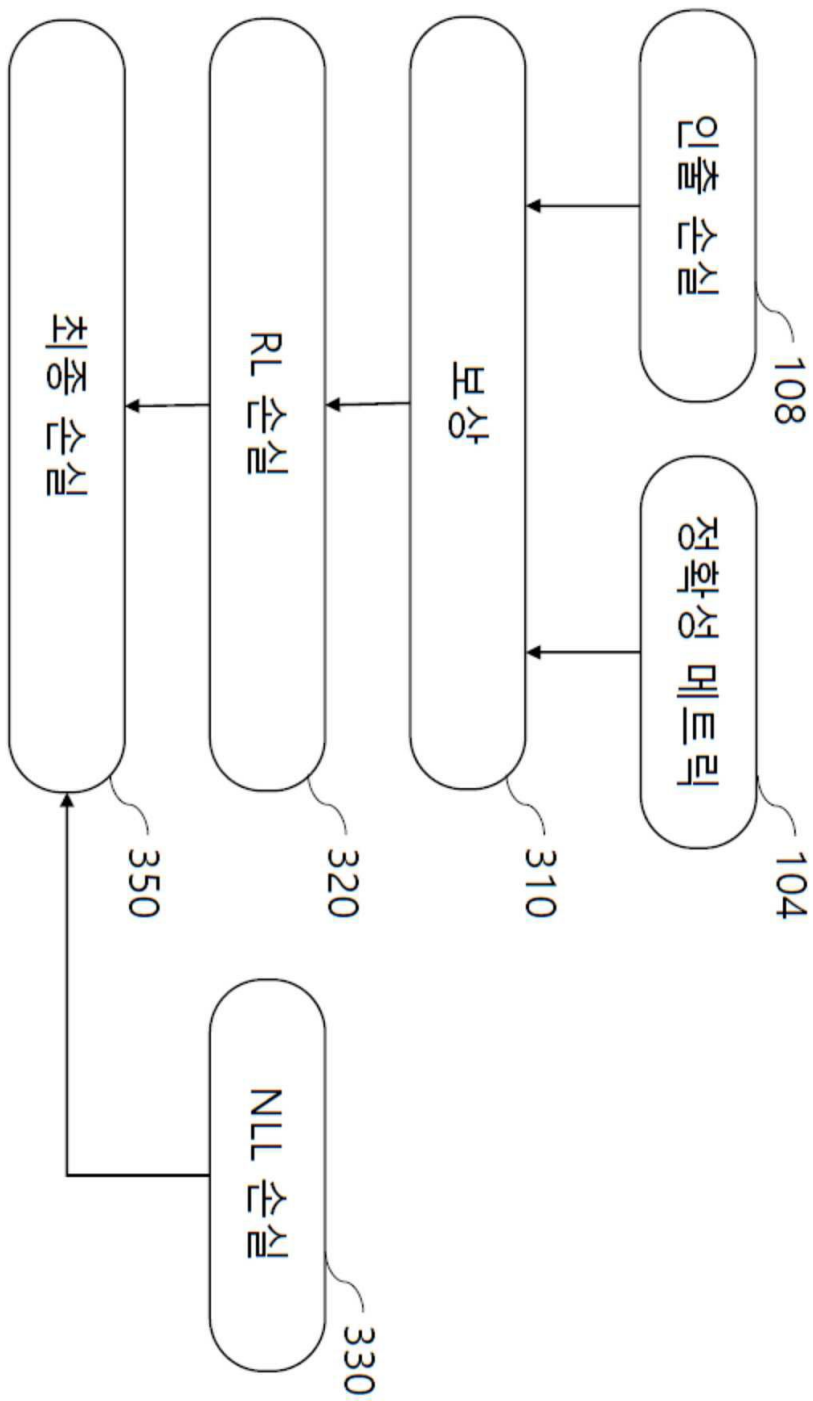
도면1



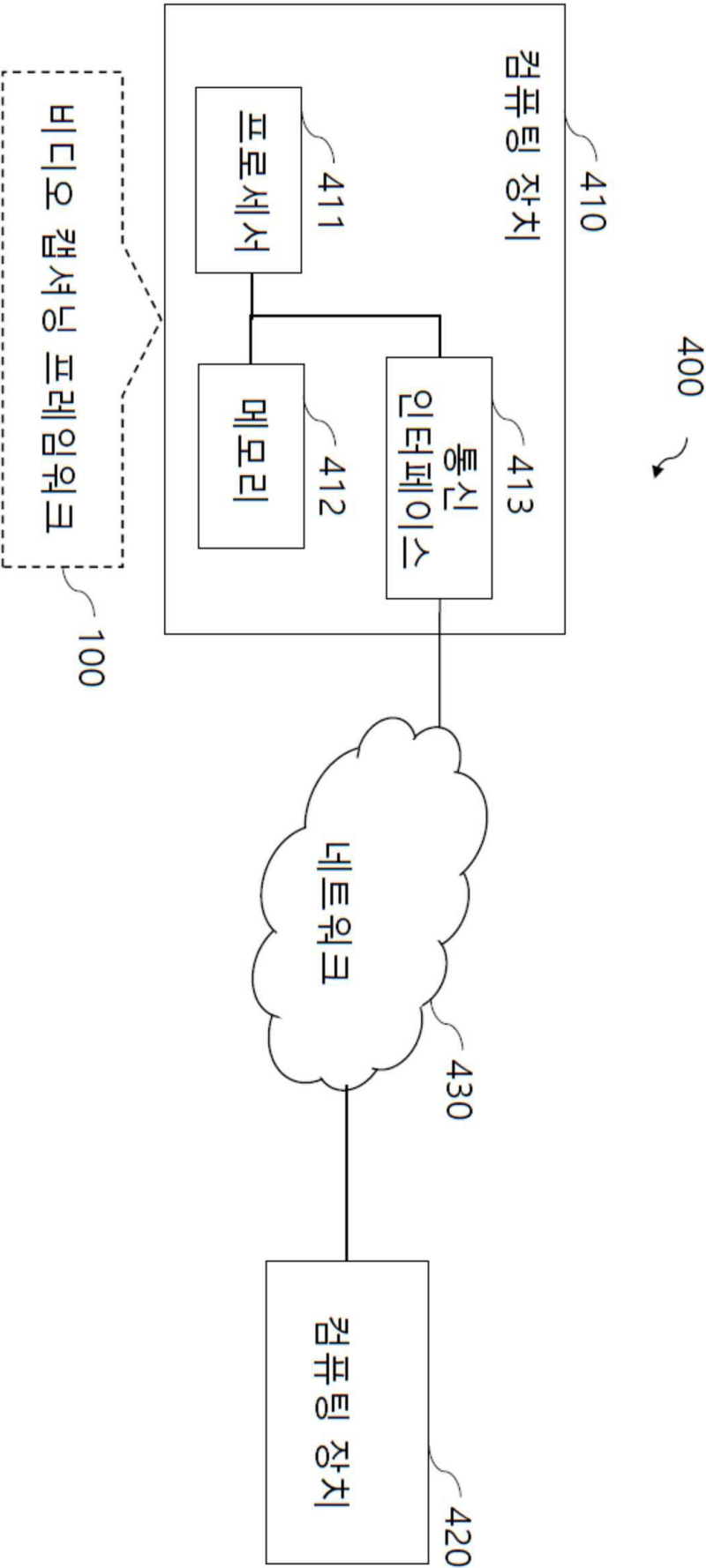
도면2



도면3



도면4



도면5

