



(19) 대한민국특허청(KR)

(12) 등록특허공보(B1)

(45) 공고일자 2024년01월08일

(11) 등록번호 10-2622435

(24) 등록일자 2024년01월03일

(51) 국제특허분류(Int. Cl.)

G06V 10/764 (2022.01) G06V 10/46 (2022.01)

G06V 10/774 (2022.01) G06V 10/80 (2022.01)

G06V 30/19 (2022.01)

(52) CPC특허분류

G06V 10/764 (2023.08)

G06V 10/467 (2023.08)

(21) 출원번호 10-2023-0047711

(22) 출원일자 2023년04월11일

심사청구일자 2023년04월11일

(56) 선행기술조사문헌

US20230104127 A1

(뒷면에 계속)

(73) 특허권자

고려대학교산학협력단

서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)

(72) 발명자

김진규

서울시 광진구 아차산로 599 신동아파밀리에 107동 801호

박노경

서울특별시 서초구 잠원로 202-11 3동 1404호

(74) 대리인

특허법인위솔

전체 청구항 수 : 총 10 항

심사관 : 강현일

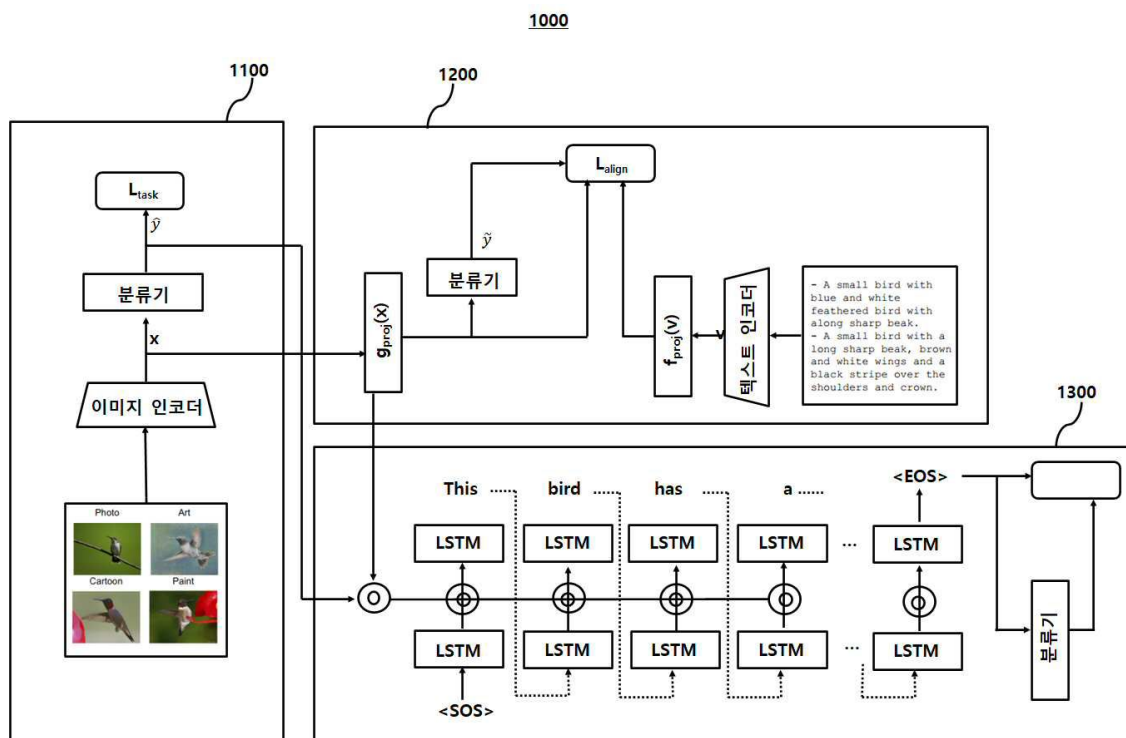
(54) 발명의 명칭 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치 및 방법

(57) 요약

본 발명의 일 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치는 학습용 이미지를 입력 받아 이로부터 시각적 특징자(Visual Feature)를 생성하고, 상기 생성한 시각적 특징자를 이용하여 상기 학습용 이미지의 분류 결과를 생성하여 출력하며, 제1 손실 함수(L_{task})를 이용하여 학습을 진행하는 이미지 분류부, 학습

(뒷면에 계속)

대표도 - 도6



용 텍스트를 입력 받아 이로부터 텍스트적 특징자(Textual Feature)를 생성하고, 상기 시각적 특징자와 텍스트적 특징자를 맵핑시키며, 제2 손실 함수(L_{align})를 이용하여 학습을 진행하는 이미지-텍스트 결합부 및 상기 시각적 특징자 및 상기 생성한 학습용 이미지의 분류 결과를 입력 받아 이에 근거하여 상기 학습용 이미지를 설명하는 텍스트(Text)를 생성하며, 제3 손실 함수(L_{exp1})를 이용하여 학습을 진행하는 텍스트 설명 생성부를 포함하며, 상기 학습용 이미지는, 하나 이상의 클래스(Class)에 대하여 하나 이상의 도메인으로 표현한 하나 이상의 이미지이며, 상기 학습용 텍스트는, 상기 하나 이상의 클래스에 대하여 하나 이상의 텍스트로 설명한 하나 이상의 텍스트이고, 상기 학습용 이미지 및 학습용 텍스트는 하나의 학습용 데이터셋에 포함된다.

(52) CPC특허분류

G06V 10/774 (2023.08)
G06V 10/806 (2023.08)
G06V 30/19147 (2023.08)

(56) 선행기술조사문헌

CN115410031 A
CN110443293 A
KR102279797 B1
KR1020230014034 A
CN115205592 A

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193663
과제번호	2020-0-01819-004
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성
연구과제명	ICT명품인재양성(고려대학교)
기 여 율	30/100
과제수행기관명	고려대학교산학협력단
연구기간	2023.01.01 ~ 2023.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193454
과제번호	2022-0-00043-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	사람중심인공지능핵심원천기술개발
연구과제명	개성 형성이 가능한 에이전트 플랫폼 기술 개발
기 여 율	50/100
과제수행기관명	한국전자기술연구원
연구기간	2023.01.01 ~ 2023.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1345347378
과제번호	2021R1A6A1A13044830
부처명	교육부
과제관리(전문)기관명	한국연구재단
연구사업명	이공학학술연구기반구축
연구과제명	융합소프트웨어연구소
기 여 율	20/100
과제수행기관명	고려대학교
연구기간	2022.03.01 ~ 2023.02.28

공지예외적용 : 있음

명세서

청구범위

청구항 1

학습용 이미지를 입력 받아 이로부터 시각적 특징자(Visual Feature)를 생성하고, 상기 생성한 시각적 특징자를 이용하여 상기 학습용 이미지의 분류 결과를 생성하여 출력하며, 제1 손실 함수(L_{task})를 이용하여 학습을 진행하는 이미지 분류부;

학습용 텍스트를 입력 받아 이로부터 텍스트적 특징자(Textual Feature)를 생성하고, 상기 시각적 특징자와 텍스트적 특징자를 맵핑시키며, 제2 손실 함수(L_{align})를 이용하여 학습을 진행하는 이미지-텍스트 결합부; 및

상기 시각적 특징자 및 상기 생성한 학습용 이미지의 분류 결과를 입력 받아 이에 근거하여 상기 학습용 이미지를 설명하는 텍스트(Text)를 생성하며, 제3 손실 함수(L_{expl})를 이용하여 학습을 진행하는 텍스트 설명 생성부;

를 포함하며,

상기 학습용 이미지는,

하나 이상의 클래스(Class)에 대하여 하나 이상의 도메인으로 표현한 하나 이상의 이미지이며,

상기 학습용 텍스트는,

상기 하나 이상의 클래스에 대하여 하나 이상의 텍스트로 설명한 하나 이상의 텍스트이고,

상기 학습용 이미지 및 학습용 텍스트는 하나의 학습용 데이터셋에 포함되는,

텍스트를 활용한 도메인 비특이적인 이미지 분류 장치에 있어서,

상기 텍스트 설명 생성부는,

제1 LSTM(Long Short Term Memory) 레이어 및 제2 LSTM 레이어를 포함하며,

상기 제2 LSTM 레이어는,

상기 제1 LSTM 레이어의 출력 토큰과 상기 시각적 특징자 및 학습용 이미지의 분류 결과를 입력 받아 상기 텍스트 설명 생성부가 보유하고 있는 각 단어의 softmax 확률 값을 생성하여 가장 높은 확률 값을 가진 단어 토큰을 출력하고,

상기 단어 토큰의 출력은,

상기 제2 LSTM 레이어가 출력한 단어 코튼이 EOS(End Of Sentence token) 또는 사전에 설정한 샘플링 단어 토큰의 개수를 만족할 때까지 반복되는,

텍스트를 활용한 도메인 비특이적인 이미지 분류 장치.

청구항 2

제1항에 있어서,

상기 이미지 분류부는,

상기 입력 받은 학습용 이미지로부터 시각적 특징자를 생성하는 이미지 인코더로 ResNet을 포함하는,

텍스트를 활용한 도메인 비특이적인 이미지 분류 장치.

청구항 3

제1항에 있어서,

상기 이미지 분류부는,

하기 수학적 식 1에 따른 제1 손실 함수(L_{task})를 이용하여 학습을 진행하는,
 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치.

$$\text{수학적 식 1}(L_{task}) = -\sum_i y_i \log(\hat{y}_i)$$

여기서 상기 y 는 정답 라벨만 1이고 나머지는 0인 one-hot 벡터, $\hat{y}$ 은 상기 시각적 특징자를 이용하여 생성한 상기 학습용 이미지의 분류 결과인 softmax 분포를 의미함

청구항 4

제1항에 있어서,

상기 이미지-텍스트 결합부는,

상기 입력 받은 학습용 텍스트로부터 텍스트적 특징자를 생성하는 텍스트 인코더로 CLIP(Contrastive Language-Image Pre-training)의 텍스트 인코더를 포함하는,

텍스트를 활용한 도메인 비특이적인 이미지 분류 장치.

청구항 5

제1항에 있어서,

상기 이미지-텍스트 결합부는,

상기 입력 받은 학습용 텍스트로부터 텍스트적 특징자를 생성하는 텍스트 인코더로 상기 텍스트 설명 생성부를 이용할 수 있는,

텍스트를 활용한 도메인 비특이적인 이미지 분류 장치.

청구항 6

제1항에 있어서,

상기 이미지-텍스트 결합부는,

하기 수학적 식 2에 따른 제2 손실 함수(L_{align})를 이용하여 학습을 진행하는,

텍스트를 활용한 도메인 비특이적인 이미지 분류 장치.

$$\text{수학적 식 2}(L_{align}) = \|f_{proj}(v) - g_{proj}(x)\|_2 - \sum_i y_i \log(\hat{y}_i)$$

여기서 상기 f_{proj} 는 텍스트적 특징자를 위한 Projection Layer, g_{proj} 은 시각적 특징자를 위한 Projection Layer, y 는 정답 라벨만 1이고 나머지는 0인 one-hot 벡터, $\hat{y}$ 은 $g_{proj}(x)$ 로부터 얻은 softmax 분포를 의미함

청구항 7

삭제

청구항 8

제1항에 있어서,

상기 텍스트 설명 생성부는,

하기 수학적 식 3에 따른 제3 손실 함수(L_{expl})를 이용하여 학습을 진행하는,

텍스트를 활용한 도메인 비특이적인 이미지 분류 장치.

$$\text{수학식 3}(L_{\text{expl}}) = -\sum_t \log P(o_{t+1}|o_{0:t}, I, C) - E_{\tilde{o} \sim p(o|I, C)}[R(\tilde{o})]$$

여기서 상기 $\tilde{o} \sim p(o|I, C)$ 는 상기 학습용 이미지를 설명하는 텍스트를 생성하기 위해 샘플링한 문장, $p(o|I, C)$ 는 학습용 이미지 I와 클래스 C가 조건으로 주어졌을 때 description o에 대한 추정 분포, $R(\tilde{o})$ 는 $P(c|\tilde{o})$ 로 생성된 문장 $\tilde{o}$ 이 있을 때, 클래스 당 softmax 확률을 의미함

청구항 9

제1항에 있어서,

상기 장치는,

상기 제1 손실 함수(L_{task}), 제2 손실 함수(L_{align}) 및 제3 손실 함수(L_{expl})를 합친 전체 손실함수(L_{total})을 최소화하는 방향으로 학습을 진행하는,

텍스트를 활용한 도메인 비특이적인 이미지 분류 장치.

청구항 10

프로세서 및 메모리를 포함하는 장치가 텍스트를 활용하여 도메인 비특이적으로 이미지를 분류하는 방법에 있어서,

(a) 학습용 데이터셋을 입력 받아 전체 손실 함수(L_{total})를 최소화하는 방향으로 학습을 수행하는 제1 단계; 및

(b) 특정 도메인으로 표현된 실행용 이미지를 입력받아 상기 실행용 이미지의 분류 결과를 출력하는 제2 단계;

를 포함하며,

상기 학습용 데이터셋은,

하나 이상의 클래스(Class)에 대하여 하나 이상의 도메인으로 표현한 하나 이상의 학습용 이미지 및 상기 하나 이상의 클래스에 대하여 하나 이상의 텍스트로 설명한 하나 이상의 학습용 텍스트를 포함하고,

상기 제1 단계는,

상기 학습용 이미지를 입력 받아 이로부터 시각적 특징자(Visual Feature)를 생성하고, 상기 생성한 시각적 특징자를 이용하여 상기 학습용 이미지의 분류 결과를 생성하여 출력하며, 제1 손실 함수(L_{task})를 이용하여 학습을 진행하는 제1-1 단계;

상기 학습용 텍스트를 입력 받아 이로부터 텍스트적 특징자(Textual Feature)를 생성하고, 상기 시각적 특징자와 텍스트적 특징자를 맵핑시키며, 제2 손실 함수(L_{align})를 이용하여 학습을 진행하는 제1-2 단계; 및

상기 시각적 특징자 및 상기 생성한 학습용 이미지의 분류 결과를 입력 받아 이에 근거하여 상기 학습용 이미지를 설명하는 텍스트(Text)를 생성하며, 제3 손실 함수(L_{expl})를 이용하여 학습을 진행하는 제1-3 단계;

를 포함하는 텍스트를 활용하여 도메인 비특이적으로 이미지를 분류하는 방법에 있어서,

상기 제1-3 단계는,

제1 LSTM(Long Short Term Memory) 레이어 및 제2 LSTM 레이어를 통해 수행되며,

상기 제2 LSTM 레이어는,

상기 제1 LSTM 레이어의 출력 토큰과 상기 시각적 특징자 및 학습용 이미지의 분류 결과를 입력 받아 텍스트 설명 생성부가 보유하고 있는 각 단어의 softmax 확률 값을 생성하여 가장 높은 확률 값을 가진 단어 토큰을 출력하고,

상기 단어 토큰의 출력은,

상기 제2 LSTM 레이어가 출력한 단어 코튼이 EOS(End Of Sentence token) 또는 사전에 설정한 샘플링 단어 토큰

의 개수를 만족할 때까지 반복되는,

텍스트를 활용하여 도메인 비특이적으로 이미지를 분류하는 방법.

청구항 11

컴퓨팅 장치와 결합하여,

(AA) 학습용 데이터셋을 입력 받아 전체 손실 함수(L_{total})를 최소화하는 방향으로 학습을 수행하는 제1 단계; 및

(BB) 특정 도메인으로 표현된 실행용 이미지를 입력받아 상기 실행용 이미지의 분류 결과를 출력하는 제2 단계;를 실행하며,

상기 학습용 데이터셋은,

하나 이상의 클래스(Class)에 대하여 하나 이상의 도메인으로 표현한 하나 이상의 학습용 이미지 및 상기 하나 이상의 클래스에 대하여 하나 이상의 텍스트로 설명한 하나 이상의 학습용 텍스트를 포함하고,

상기 제1 단계는,

상기 학습용 이미지를 입력 받아 이로부터 시각적 특징자(Visual Feature)를 생성하고, 상기 생성한 시각적 특징자를 이용하여 상기 학습용 이미지의 분류 결과를 생성하여 출력하며, 제1 손실 함수(L_{task})를 이용하여 학습을 진행하는 제1-1 단계;

상기 학습용 텍스트를 입력 받아 이로부터 텍스트적 특징자(Textual Feature)를 생성하고, 상기 시각적 특징자와 텍스트적 특징자를 맵핑시키며, 제2 손실 함수(L_{align})를 이용하여 학습을 진행하는 제1-2 단계; 및

상기 시각적 특징자 및 상기 생성한 학습용 이미지의 분류 결과를 입력 받아 이에 근거하여 상기 학습용 이미지를 설명하는 텍스트(Text)를 생성하며, 제3 손실 함수(L_{expl})를 이용하여 학습을 진행하는 제1-3 단계;

를 포함하는 컴퓨터로 판독 가능한 매체에 저장된 컴퓨터 프로그램에 있어서,

상기 제1-3 단계는,

제1 LSTM(Long Short Term Memory) 레이어 및 제2 LSTM 레이어를 통해 수행되며,

상기 제2 LSTM 레이어는,

상기 제1 LSTM 레이어의 출력 토큰과 상기 시각적 특징자 및 학습용 이미지의 분류 결과를 입력 받아 텍스트 설명 생성부가 보유하고 있는 각 단어의 softmax 확률 값을 생성하여 가장 높은 확률 값을 가진 단어 토큰을 출력하고,

상기 단어 토큰의 출력은,

상기 제2 LSTM 레이어가 출력한 단어 코튼이 EOS(End Of Sentence token) 또는 사전에 설정한 샘플링 단어 토큰의 개수를 만족할 때까지 반복되는,

컴퓨터로 판독 가능한 매체에 저장된 컴퓨터 프로그램.

발명의 설명

기술 분야

[0001] 본 발명은 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치 및 이를 위한 방법에 관한 것이다. 보다 자세하게는 텍스트 정보를 활용하여 이미지의 도메인에 무관하게 이미지를 분류할 수 있는 장치 및 방법에 관한 것이다.

배경 기술

[0002] 최근 대량의 학습 데이터를 기반으로 한 이미지 분류 인공지능 모델들의 성능이 일취월장하고 있으며, 사람이 직접 분류하는 것보다 더 높은 분류 성능을 보여주는 경우까지 발생하고 있는바, 다양한 분야에서 널리 활용되고 있다.

[0003] 이와 같은 이미지 분류 인공지능 모델의 분류 성능은 해당 인공지능 모델이 어떠한 도메인(스타일)의 이미지를 가지고 학습을 수행했는지에 따라 달라질 수 있는바, 예를 들어, 스케치 도메인의 이미지를 가지고 학습을 수행한 이미지 분류 인공지능 모델에 수채화 도메인의 이미지가 입력된다면 분류 성능이 현저하게 하락하게 되며, 이러한 현상은 학습 이미지의 도메인과 실제 실행 이미지의 도메인의 차이가 큰 경우에 보다 심화된다.

[0004] 이와 같은 문제점을 해결하기 위해 최대한으로 다양한 도메인의 이미지를 수집하여 이미지 분류 인공지능 모델을 학습시키자는 업계 전문가들의 의견이 일부 대두되고 있으나, 현실에 존재하는 수없이 많은 종류의 도메인 그리고 이에 대한 이미지를 모두 수집하는 것은 현실적으로 불가능하며, 창작자들에 의해 기존에 없던 새로운 도메인까지 비일비재하게 생겨나고 있는바, 보다 근본적인 해결 방안이 요구된다. 본 발명은 이에 관한 것이다.

선행기술문헌

특허문헌

[0005] (특허문헌 0001) 대한민국 공개특허공보 제 10-2022-0138696호(2022.10.13)

발명의 내용

해결하려는 과제

[0006] 본 발명이 해결하고자 하는 기술적 과제는 일부 도메인의 이미지만을 학습한 경우라 할지라도 이와 상이한 도메인의 이미지까지 높은 정확도로 효과적으로 분류할 수 있는 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치 및 방법을 제공하는 것이다.

[0007] 본 발명의 기술적 과제들은 이상에서 언급한 기술적 과제들로 제한되지 않으며, 언급되지 않은 또 다른 기술적 과제들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0008] 상기 기술적 과제를 달성하기 위한 본 발명의 일 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치는 학습용 이미지를 입력 받아 이로부터 시각적 특징자(Visual Feature)를 생성하고, 상기 생성한 시각적 특징자를 이용하여 상기 학습용 이미지의 분류 결과를 생성하여 출력하며, 제1 손실 함수(L_{task})를 이용하여 학습을 진행하는 이미지 분류부, 학습용 텍스트를 입력 받아 이로부터 텍스트적 특징자(Textual Feature)를 생성하고, 상기 시각적 특징자와 텍스트적 특징자를 맵핑시키며, 제2 손실 함수(L_{align})를 이용하여 학습을 진행하는 이미지-텍스트 결합부 및 상기 시각적 특징자 및 상기 생성한 학습용 이미지의 분류 결과를 입력 받아 이에 근거하여 상기 학습용 이미지를 설명하는 텍스트(Text)를 생성하며, 제3 손실 함수(L_{expl})를 이용하여 학습을 진행하는 텍스트 설명 생성부를 포함하며,

[0009] 상기 학습용 이미지는, 하나 이상의 클래스(Class)에 대하여 하나 이상의 도메인으로 표현한 하나 이상의 이미지이며, 상기 학습용 텍스트는, 상기 하나 이상의 클래스에 대하여 하나 이상의 텍스트로 설명한 하나 이상의 텍스트이고, 상기 학습용 이미지 및 학습용 텍스트는 하나의 학습용 데이터셋에 포함된다.

[0010] 일 실시 예에 따르면, 상기 이미지 분류부는, 상기 입력 받은 학습용 이미지로부터 시각적 특징자를 생성하는 이미지 인코더로 ResNet을 포함할 수 있다.

[0011] 일 실시 예에 따르면, 상기 이미지 분류부는, 하기 수학적 식 1에 따른 제1 손실 함수(L_{task})를 이용하여 학습을 진행할 수 있다.

[0012] 수학적 식 1(L_{task}) = $-\sum_i y_i \log(\hat{y}_i)$

[0013] 여기서 상기 y 는 정답 라벨만 1이고 나머지는 0인 one-hot 벡터, $\hat{y}$ 은 상기 시각적 특징자를 이용하여 생성한 상기 학습용 이미지의 분류 결과인 softmax 분포를 의미함

[0014] 일 실시 예에 따르면, 상기 이미지-텍스트 결합부는, 상기 입력 받은 학습용 텍스트로부터 텍스트적 특징자를

생성하는 텍스트 인코더로 CLIP(Contrastive Language-Image Pre-training)의 텍스트 인코더를 포함할 수 있다.

[0015] 일 실시 예에 따르면, 상기 이미지-텍스트 결합부는, 상기 입력 받은 학습용 텍스트로부터 텍스트적 특징자를 생성하는 텍스트 인코더로 상기 텍스트 설명 생성부를 이용할 수 있는 것일 수 있다.

[0016] 일 실시 예에 따르면, 상기 이미지-텍스트 결합부는, 하기 수학적 식 2에 따른 제2 손실 함수(L_{align})를 이용하여 학습을 진행할 수 있다.

[0017] 수학적 식 2(L_{align}) = $\|f_{proj}(v) - g_{proj}(x)\|_2 - \sum_i y_i \log(\tilde{y}_i)$

[0018] 여기서 상기 f_{proj} 는 텍스트적 특징자를 위한 Projection Layer, g_{proj} 은 시각적 특징자를 위한 Projection Layer, y 는 정답 라벨만 1이고 나머지는 0인 one-hot 벡터, $\tilde{y}$ 은 $g_{proj}(x)$ 로부터 얻은 softmax 분포를 의미함

[0019] 일 실시 예에 따르면, 상기 텍스트 설명 생성부는, 제1 LSTM(Long Short Term Memory) 레이어 및 제2 LSTM 레이어를 포함하며, 상기 제2 LSTM 레이어는, 상기 제1 LSTM 레이어의 출력 토큰과 상기 시각적 특징자 및 학습용 이미지의 분류 결과를 입력 받아 상기 텍스트 설명 생성부가 보유하고 있는 각 단어의 softmax 확률 값을 생성하여 가장 높은 확률 값을 가진 단어 토큰을 출력하고, 상기 단어 토큰의 출력은,

[0020] 상기 제2 LSTM 레이어가 출력한 단어 코드가 EOS(End Of Sentence token) 또는 사전에 설정한 샘플링 단어 토큰의 개수를 만족할 때까지 반복될 수 있다.

[0021] 일 실시 예에 따르면, 상기 텍스트 설명 생성부는, 하기 수학적 식 3에 따른 제3 손실 함수(L_{expl})을 이용하여 학습을 진행할 수 있다.

[0022] 수학적 식 3(L_{expl}) = $-\sum_t \log P(o_{t+1}|o_{0:t}, I, C) - E_{\tilde{o} \sim p(o|I, C)}[R(\tilde{o})]$

[0023] 여기서 상기 $\tilde{o} \sim p(o|I, C)$ 는 상기 학습용 이미지를 설명하는 텍스트를 생성하기 위해 샘플링한 문장, $p(o|I, C)$ 는 학습용 이미지 I와 클래스 C가 조건으로 주어졌을 때 description o에 대한 추정 분포, $R(\tilde{o})$ 는 $P(c|\tilde{o})$ 로 생성된 문장 $\tilde{o}$ 이 있을 때, 클래스 당 softmax 확률을 의미함

[0024] 일 실시 예에 따르면, 상기 장치는, 상기 제1 손실 함수(L_{task}), 제2 손실 함수(L_{align}) 및 제3 손실 함수(L_{expl})를 합친 전체 손실함수(L_{total})을 최소화하는 방향으로 학습을 진행할 수 있다.

[0025] 상기 기술적 과제를 달성하기 위한 본 발명의 또 다른 실시 예에 따른 프로세서 및 메모리를 포함하는 장치가 텍스트를 활용하여 도메인 비특이적으로 이미지를 분류하는 방법은 (a) 학습용 데이터셋을 입력 받아 전체 손실 함수(L_{total})를 최소화하는 방향으로 학습을 수행하는 제1 단계 및 (b) 특정 도메인으로 표현된 실행용 이미지를 입력받아 상기 실행용 이미지의 분류 결과를 출력하는 제2 단계를 포함하며, 상기 학습용 데이터셋은, 하나 이상의 클래스(Class)에 대하여 하나 이상의 도메인으로 표현한 하나 이상의 학습용 이미지 및 상기 하나 이상의 클래스에 대하여 하나 이상의 텍스트로 설명한 하나 이상의 학습용 텍스트를 포함하고, 상기 제1 단계는, 상기 학습용 이미지를 입력 받아 이로부터 시각적 특징자(Visual Feature)를 생성하고, 상기 생성한 시각적 특징자를 이용하여 상기 학습용 이미지의 분류 결과를 생성하여 출력하며, 제1 손실 함수(L_{task})를 이용하여 학습을 진행하는 제1-1 단계; 상기 학습용 텍스트를 입력 받아 이로부터 텍스트적 특징자(Textual Feature)를 생성하고, 상기 시각적 특징자와 텍스트적 특징자를 맵핑시키며, 제2 손실 함수(L_{align})를 이용하여 학습을 진행하는 제1-2 단계 및 상기 시각적 특징자 및 상기 생성한 학습용 이미지의 분류 결과를 입력 받아 이에 근거하여 상기 학습용 이미지를 설명하는 텍스트(Text)를 생성하며, 제3 손실 함수(L_{expl})를 이용하여 학습을 진행하는 제1-3 단계를 포함한다.

[0026] 상기 기술적 과제를 달성하기 위한 본 발명의 또 다른 실시 예에 따른 매체에 저장된 컴퓨터 프로그램은 컴퓨팅 장치와 결합하여, (AA) 학습용 데이터셋을 입력 받아 전체 손실 함수(L_{total})를 최소화하는 방향으로 학습을 수행하는 제1 단계 및 (BB) 특정 도메인으로 표현된 실행용 이미지를 입력받아 상기 실행용 이미지의 분류 결과를

출력하는 제2 단계를 실행하며, 상기 학습용 데이터셋은, 하나 이상의 클래스(Class)에 대하여 하나 이상의 도메인으로 표현한 하나 이상의 학습용 이미지 및 상기 하나 이상의 클래스에 대하여 하나 이상의 텍스트로 설명한 하나 이상의 학습용 텍스트를 포함하고, 상기 제1 단계는, 상기 학습용 이미지를 입력 받아 이로부터 시각적 특징자(Visual Feature)를 생성하고, 상기 생성한 시각적 특징자를 이용하여 상기 학습용 이미지의 분류 결과를 생성하여 출력하며, 제1 손실 함수(L_{task})를 이용하여 학습을 진행하는 제1-1 단계; , 상기 학습용 텍스트를 입력 받아 이로부터 텍스트적 특징자(Textual Feature)를 생성하고, 상기 시각적 특징자와 텍스트적 특징자를 맵핑시키며, 제2 손실 함수(L_{align})를 이용하여 학습을 진행하는 제1-2 단계 및 상기 시각적 특징자 및 상기 생성한 학습용 이미지의 분류 결과를 입력 받아 이에 근거하여 상기 학습용 이미지를 설명하는 텍스트(Text)를 생성하며, 제3 손실 함수(L_{expl})를 이용하여 학습을 진행하는 제1-3 단계를 포함한다.

발명의 효과

- [0027] 상기와 같은 본 발명에 따르면 학습 과정에서 이미지뿐만 아니라 텍스트 정보를 함께 사용하여 학습을 진행하는 바, 인간의 추론 과정을 적용하여 일부 도메인의 이미지만을 학습한 경우라 할지라도 이와 상이한 도메인의 이미지까지 높은 정확도로 효과적으로 분류할 수 있다는 효과가 있다.
- [0028] 본 발명의 효과들은 이상에서 언급한 효과들로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해 될 수 있을 것이다.

도면의 간단한 설명

- [0029] 도 1은 본 발명의 제1 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치가 포함하는 전체 구성을 예시적으로 도시한 도면이다.
- 도 2는 본 발명의 제2 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치가 이미지 분류를 위해 수행하는 동작 또는 연산을 기능적 구성으로 명명하여 도시한 내부 구조도이다.
- 도 3은 본 발명의 제2 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치에 있어서, 이미지 분류부의 내부 구조를 예시적으로 도시한 도면이다.
- 도 4는 본 발명의 제2 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치에 있어서, 이미지-텍스트 결합부의 내부 구조를 예시적으로 도시한 도면이다.
- 도 5는 본 발명의 제2 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치에 있어서, 텍스트 설명 생성부의 내부 구조를 예시적으로 도시한 도면이다.
- 도 6은 본 발명의 제2 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치에 있어서 전체 내부 구조를 예시적으로 도시한 도면이다.
- 도 7 내지 도 9는 본 발명의 제2 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치의 성능 평가 자료이다.
- 도 10은 본 발명의 제3 실시 예에 따른 텍스트를 활용하여 도메인 비특이적으로 이미지를 분류하는 방법의 대표적인 단계를 나타낸 순서도이다.

발명을 실시하기 위한 구체적인 내용

- [0030] 본 발명의 목적과 기술적 구성 및 그에 따른 작용 효과에 관한 자세한 사항은 본 발명의 명세서에 첨부된 도면에 의거한 이하의 상세한 설명에 의해 보다 명확하게 이해될 것이다. 첨부된 도면을 참조하여 본 발명에 따른 실시 예를 상세하게 설명한다.
- [0031] 본 명세서에서 개시되는 실시 예들은 본 발명의 범위를 한정하는 것으로 해석되거나 이용되지 않아야 할 것이다. 이 분야의 통상의 기술자에게 본 명세서의 실시 예를 포함한 설명은 다양한 응용을 갖는다는 것이 당연하다. 따라서, 본 발명의 상세한 설명에 기재된 임의의 실시 예들은 본 발명을 보다 잘 설명하기 위한 예시적인 것이며 본 발명의 범위가 실시 예들로 한정되는 것을 의도하지 않는다.
- [0032] 도면에 표시되고 아래에 설명되는 기능 블록들은 가능한 구현의 예들일 뿐이다. 다른 구현들에서는 상세한 설명의 사상 및 범위를 벗어나지 않는 범위에서 다른 기능 블록들이 사용될 수 있다. 또한, 본 발명의 하나 이상의

기능 블록이 개별 블록들로 표시되지만, 본 발명의 기능 블록들 중 하나 이상은 동일 기능을 실행하는 다양한 하드웨어 및 소프트웨어 구성들의 조합일 수 있다.

- [0033] 또한, 어떤 구성요소들을 포함한다는 표현은 "개방형"의 표현으로서 해당 구성요소들이 존재하는 것을 단순히 지칭할 뿐이며, 추가적인 구성요소들을 배제하는 것으로 이해되어서는 안 된다.
- [0034] 나아가 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급될 때에는, 그 다른 구성요소에 직접적으로 연결 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 한다.
- [0035] 이하에서는 도면들을 참조하여 본 발명의 세부적인 실시 예들에 대해 살펴보도록 한다.
- [0036] 도 1은 본 발명의 제1 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치(100)가 포함하는 전체 구성을 예시적으로 도시한 도면이다.
- [0037] 그러나 이는 본 발명의 목적을 달성하기 위한 바람직한 실시 예일 뿐이며, 필요에 따라 일부 구성이 추가되거나 삭제될 수 있고, 어느 한 구성이 수행하는 역할을 다른 구성이 함께 수행할 수도 있음은 물론이다.
- [0038] 본 발명의 제1 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치(100)는 프로세서(10), 네트워크 인터페이스(20), 메모리(30), 스토리지(40) 및 이들을 연결하는 데이터 버스(50)를 포함할 수 있으며, 기타 본 발명의 목적을 달성함에 있어 요구되는 부가적인 구성들을 더 포함할 수 있음은 물론이라 할 것이다.
- [0039] 프로세서(10)는 각 구성의 전반적인 동작을 제어한다. 프로세서(10)는 CPU(Central Processing Unit), MPU(Micro Processor Unit), MCU(Micro Controller Unit) 또는 본 발명이 속하는 기술 분야에서 널리 알려져 있는 형태의 인공지능 프로세서 중 어느 하나일 수 있다. 아울러, 프로세서(10)는 후술할 본 발명의 제2 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치(1000)에서 설명할 각각의 동작을 수행하기 위한 적어도 하나의 애플리케이션 또는 프로그램에 대한 연산을 수행할 수 있다.
- [0040] 네트워크 인터페이스(20)는 본 발명의 제1 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치(100)의 유무선 인터넷 통신을 지원하며, 그 밖의 공지의 통신 방식을 지원할 수도 있다. 따라서 네트워크 인터페이스(20)는 그에 따른 통신 모듈을 포함하여 구성될 수 있다.
- [0041] 메모리(30)는 각종 정보, 명령 및/또는 정보를 저장하며, 본 발명의 제3 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 방법을 수행하기 위해 스토리지(40)로부터 하나 이상의 컴퓨터 프로그램(41)을 로드할 수 있다. 도 1에서는 메모리(30)의 하나로 RAM을 도시하였으나 이와 더불어 다양한 저장 매체를 메모리(30)로 이용할 수 있음은 물론이다.
- [0042] 스토리지(40)는 하나 이상의 컴퓨터 프로그램(41) 및 대용량 네트워크 정보(42)를 비임시적으로 저장할 수 있다. 이러한 스토리지(40)는 ROM(Read Only Memory), EPROM(Erasable Programmable ROM), EEPROM(Electrically Erasable Programmable ROM), 플래시 메모리 등과 같은 비휘발성 메모리, 하드 디스크(HDD), 보조 저장 매체(SSD), 착탈형 디스크, 또는 본 발명이 속하는 기술 분야에서 널리 알려져 있는 임의의 형태의 컴퓨터로 읽을 수 있는 기록 매체 중 어느 하나일 수 있다.
- [0043] 컴퓨터 프로그램(41)은 메모리(30)에 로드되어, 하나 이상의 프로세서(10)에 의해, (A) 학습용 데이터셋을 입력받아 전체 손실 함수(L_{total})를 최소화하는 방향으로 학습을 수행하는 제1 오퍼레이션 및 (B) 특정 도메인으로 표현된 실행용 이미지를 입력받아 상기 실행용 이미지의 분류 결과를 출력하는 제2 오퍼레이션을 실행할 수 있으며, 여기서 제1 오퍼레이션은 상기 학습용 이미지를 입력받아 이로부터 시각적 특징자(Visual Feature)를 생성하고, 상기 생성한 시각적 특징자를 이용하여 상기 학습용 이미지의 분류 결과를 생성하여 출력하며, 제1 손실 함수(L_{task})를 이용하여 학습을 진행하는 제1-1 오퍼레이션, 상기 학습용 텍스트를 입력받아 이로부터 텍스트적 특징자(Textual Feature)를 생성하고, 상기 시각적 특징자와 텍스트적 특징자를 맵핑시키며, 제2 손실 함수(L_{align})를 이용하여 학습을 진행하는 제1-2 오퍼레이션 및 상기 시각적 특징자 및 상기 생성한 학습용 이미지의 분류 결과를 입력받아 이에 근거하여 상기 학습용 이미지를 설명하는 텍스트(Text)를 생성하며, 제3 손실 함수(L_{expl})를 이용하여 학습을 진행하는 제1-3 오퍼레이션을 포함할 수 있다.
- [0044] 이상 간단하게 언급한 컴퓨터 프로그램(41)이 수행하는 오퍼레이션은 컴퓨터 프로그램(41)의 일 기능으로 볼 수 있으며, 보다 자세한 설명은 본 발명의 제2 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치(1000)에 대한 설명에서 후술하도록 한다.

- [0045] 데이터 버스(50)는 이상 설명한 프로세서(10), 네트워크 인터페이스(20), 메모리(30) 및 스토리지(40) 사이의 명령 및/또는 정보의 이동 경로가 된다.
- [0046] 이상 간단하게 설명한 본 발명의 제1 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치(100)는 독립된 디바이스의 형태, 예를 들어 전자 기기나 서버(클라우드 포함)의 형태일 수 있으며, 여기서 전자 기기는 한 장소에 고정 설치되어 사용하는 데스크톱 PC, 서버 디바이스 등과 같은 기기 뿐만 아니라, 스마트폰, 태블릿 PC, 노트북 PC, PDA, PMP 등과 같이 휴대가 용이한 포터블 기기 등이라도 무방한바, 프로세서(10)에 해당하는 CPU 등이 설치되고 네트워크 기능만 보유하고 있는 전자 기기라면 어떠한 것이라도 무방하다 할 것이다.
- [0047] 이하, 본 발명의 제1 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치(100)가 독립된 디바이스인 전자 기기 형태 중, "서버"의 형태임을 전제로 이미지 분류를 위해 프로세서(10)가 내부적으로 수행하는 각각의 동작 또는 연산을 기능적 구성의 명명을 통해 설명하도록 하는바, 이는 본 발명의 제2 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치(1000)에 관한 것이다.
- [0048] 도 2는 본 발명의 제2 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치(1000)가 이미지 분류를 위해 수행하는 동작 또는 연산을 기능적 구성으로 명명하여 도시한 내부 구조도이다.
- [0049] 도 2를 참조하면, 본 발명의 제2 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치(1000)는 이미지 분류부(1100), 이미지-텍스트 결합부(1200) 및 텍스트 설명 생성부(1300)를 포함할 수 있으며, 기타 본 발명의 목적을 달성하는데 요구되는 부가적인 구성을 더 포함할 수 있음은 물론이라 할 것이다.
- [0050] 이미지 분류부(1100), 이미지-텍스트 결합부(1200) 및 텍스트 설명 생성부(1300)는 앞서 언급한 바와 같이 프로세서(10)가 내부적으로 수행하는 각각의 동작 또는 연산을 기능적 구성으로 명명한 것이며, 이하 본 발명의 제2 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치(1000)를 "장치(1000)"라 명명하고, 여기서의 장치(1000)는 프로세서(10)와 동일한 의미가 될 수 있음을 밝혀두는 바이다.
- [0051] 장치(1000)는 크게 두 가지 경우로 동작할 수 있는바, 하나는 학습용 데이터셋을 이용하여 학습을 진행하는 경우이며, 또 다른 하나는 학습을 완료한 후, 특정 도메인으로 표현된 실행용 이미지를 입력 받아 실제 이미지 분류를 진행하는 경우인바, 전자의 경우부터 설명하도록 한다.
- [0052] 설명에 앞서, 장치(1000)의 학습 목표를 먼저 설명할 필요가 있는바, 장치(1000)의 학습은 본 발명의 목적에 부합하게 한정된 도메인의 학습용 이미지로 학습을 진행하여도 학습용 이미지에 포함되지 않은 도메인으로 표현된 이미지에 대해서까지 정확하게 분류할 수 있는 것을 목표로 하며, 이를 수식으로 나타내면 다음과 같다.
- [0053] 학습의 목표 =
$$\underset{\theta}{\text{Minimize}} \quad \sup_{T: D(S, T) \leq \rho} E[L_{\text{task}}(S; \theta)]$$
- [0054] 여기서 S는 학습용 데이터 셋에 포함되어 있는 도메인, T는 학습용 데이터셋에 포함되어 있지 않은 도메인, D(S, T)는 S와 T 사이의 차이를 의미하며 ρ 는 상한 값을 의미함
- [0055] 이하, 이상의 학습 목표를 달성하기 위한 각 구성의 동작에 대하여 설명하도록 한다.
- [0056] 도 3은 본 발명의 제2 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치(1000)에 있어서, 이미지 분류부(1100)의 내부 구조를 예시적으로 도시한 도면이다.
- [0057] 이미지 분류부(1100)는 학습용 이미지를 입력 받아 이로부터 시각적 특징자(Visual Feature)를 생성하고, 생성한 시각적 특징자를 이용하여 학습용 이미지의 분류 결과를 생성하여 출력한다.
- [0058] 여기서 학습용 이미지는 하나 이상의 클래스(Class)에 대하여 하나 이상의 도메인으로 표현한 하나 이상의 이미지일 수 있는바, 학습용 이미지는 학습용 데이터셋에 포함될 수 있다.
- [0059] 보다 구체적으로 학습용 이미지는 하나 이상의 클래스, 예를 들어 까치, 참새, 독수리, 매, 비둘기, 부엉이 등과 같이 조류에 속하는 새의 종류인 각각의 클래스에 대하여 다양한 도메인인 스케치, 수채화, 페인팅, 아트, 카툰 등으로 표현한 이미지일 수 있으며, 각각의 클래스에 대하여 하나의 도메인만으로 표현할 수도 있을 것이나, 다양한 도메인으로 표현한 학습용 이미지를 학습하는 것이 학습 성능을 향상시킬 수 있기에 바람직하다 할 것이다.
- [0060] 또한, 학습용 이미지는 이상과 같이 동일한 중, 예를 들어 조류에 속하는 각각의 클래스가 아니라 서로 다른 중

에 속하는 클래스에 대한 이미지라도 무방한바, 예를 들어, 독수리(조류), 호랑이(포유류), 상어(어류), 개구리(양서류) 등과 같이 서로 상이한 종에 속하는 클래스에 대한 이미지를 다양한 도메인으로 표현하여 학습용 이미지로 사용할 수 있을 것이나, 학습 성능은 동일한 종에 대한 학습용 이미지를 이용하여 학습하는 경우보다 일부 저하될 수도 있을 것이다.

[0061] 이상 설명한 학습용 이미지에 대한 설명(조류 등)은 학습 이미지의 종류에 대한 하나의 예시에 불과하며, 사용자에 따라 장치(1000)를 통해 분류하고자 하는 대상이 너무나 다양하게 상이할 것이기 때문에 학습용 이미지가 나타내는 클래스의 종류에는 제한이 없다 할 것이고, 단지 학습 과정에서 일정 수준 이상의 공통점이 있는 학습용 이미지를 함께 이용하여 학습을 진행하는 것이 학습 성능 향상에 이바지할 수 있다는 점만 명심하면 충분할 것이며, 이하 도 3에 예시적으로 도시된 학습용 이미지인 조류에 대한 클래스를 예로 설정하여 설명을 이어가도록 한다.

[0062] 이미지 분류부(1100)은 입력 받은 학습용 이미지로부터 시각적 특징자를 생성하는 이미지 인코더(Encoder, Featurizer, 1101)를 포함할 수 있으며, 도 3을 참조하면, 학습용 이미지가 이미지 인코더를 통과하여 시각적 특징자인 x 가 출력되고 있음을 확인할 수 있다.

[0063] 여기서 이미지 인코더(1101)는 사전 학습된 공지된 이미지 인코더를 사용할 수 있는바, 예를 들어 CNN(Convolution Neural Network), 오토인코더(AutoEncoder), ResNet(Residual Neural Network), 삼 네트워크(Siamese Neural Network), 트랜스포머 네트워크(Transformer Network) 등 어떠한 것이라도 무방하다 할 것이다.

[0064] 한편, 이미지 분류부(1100)는 생성한 시각적 특징자를 이용하여 이미지를 분류하는 분류기(1102)를 포함할 수 있는바, 여기서 분류기(1102)는 공지된 Classifier Layer(Linear Layer)일 수 있다.

[0065] 이상 설명한 바와 같은 이미지 분류부(1100)의 동작을 쉽게 설명하면 학습용 이미지를 입력 받아 이로부터 시각적 특징자를 생성하고, 생성한 시각적 특징자를 이용하여 학습용 이미지의 분류 결과를 생성하여 출력하는 것인바, 장치(1000)의 학습이 완료된 이후, 실행용 이미지를 입력 받아 해당 이미지를 분류하는 경우에는 다른 구성들을 제외하고 오직 이미지 분류부(1100)만 동작한다 할 것이다.

[0066] 한편, 이미지 분류부(1100)는 학습용 이미지를 입력 받아 올바르게 정확한 클래스로 분류하는 것이 목표이기 때문에 다음과 같은 제1 손실 함수(L_{task})를 이용하여 학습을 진행할 수 있다.

[0067] 제1 손실 함수(L_{task}) = $-\sum_i y_i \log(\hat{y}_i) = L_{task}(y, \hat{y}) = L_{task}(S; \theta)$

[0068] 여기서 y 는 정답 라벨만 1이고 나머지는 0인 one-hot 벡터, $\hat{y}$ 은 시각적 특징자를 이용하여 생성한 학습용 이미지의 분류 결과인 softmax 분포, S 는 학습용 데이터 셋에 포함되어 있는 도메인, θ 는 장치(1000)의 파라미터를 의미함

[0069] 이상과 같은 제1 손실 함수(L_{task})는 이미지 분류 로스와 동일하게 Cross-Entropy Loss(교차 엔트로피 손실)가 된다 할 것이다.

[0070] 도 4는 본 발명의 제2 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치(1000)에 있어서, 이미지-텍스트 결합부(1200)의 내부 구조를 예시적으로 도시한 도면이다.

[0071] 이미지-텍스트 결합부(1200)는 학습용 텍스트를 입력 받아 이로부터 텍스트적 특징자(Textual Feature)를 생성하고, 시각적 특징자와 텍스트적 특징자를 맵핑시킨다.

[0072] 여기서 학습용 텍스트는 하나 이상의 클래스에 대하여 하나 이상의 텍스트로 설명한 하나 이상의 텍스트일 수 있는바, 학습용 텍스트 역시 학습용 이미지와 마찬가지로 학습용 데이터셋에 포함될 수 있다.

[0073] 앞서 학습용 이미지에 대하여 조류에 대한 클래스를 예로 설정하여 설명을 이어간다고 했던바, 그에 따른 학습용 텍스트는 각각의 클래스인 까치, 참새, 독수리, 매, 비둘기, 부엉이 등에 대하여 해당 클래스를 설명할 수 있는 텍스트이며, 까치를 예를 든다면 학습용 텍스트는 "검정색 부리에 전체적으로 검정색 깃털을 가진 새", "검정색 부리에 몸통은 흰색 깃털을 포함하고, 머리와 꼬리는 검정색 깃털로 이루어진 새" 등과 같은 텍스트일 수 있다.

- [0074] 이와 같은 학습용 텍스트는 학습용 이미지가 나타내고 있는 클래스에 대하여 상세하게 묘사된 텍스트일수록 학습 성능을 향상시킬 수 있을 것인바, 예를 들어 학습용 텍스트가 "검정색 깃털을 가진 새"와 같이 지나치게 간단하게 묘사한 텍스트라 한다면, 검정색 깃털을 가진 새가 너무나 많기에 특정 클래스로 분류할 수 없을 것이기 때문이다.
- [0075] 한편, 장치(1000), 보다 구체적으로 이미지-텍스트 결합부(1200)가 학습에 학습용 텍스트인 텍스트 정보를 활용하는 것은 일반적인 이미지 분류 인공지능 모델이 학습용 이미지의 도메인과 상이한 도메인의 이미지가 입력되었을 때 분류 성능이 저하되는 현상을 방지하기 위하여 인간의 추론 과정을 응용하여 적용한 본 발명만의 독자적인 특징인바, 예를 들어 인간이 책에 수채화 도메인으로 그려진 코끼리라는 동물을 보며 코끼리라는 동물의 시각적 특징을 "코끼리는 몸통이 크고 긴 상아로 된 하얀 어금니와 큰 귀를 가진 식물을 주식으로 하는 포유류"라는 텍스트(인간이기 때문에 사고일 수 있음)적 특징으로 매칭시켜 기억했다면, 그 이후에는 코끼리를 어떠한 도메인으로 표현했든지와 무관하게 이미지로부터 코끼리를 나타내는 시각적 특징을 파악하여 이에 매칭되어 있는 기억시킨 텍스트적 특징을 통해 정확하게 코끼리라는 동물을 분류할 수 있기 때문이며, 이를 위한 구성이 이미지-텍스트 결합부(1102)이다.
- [0076] 그에 따라 이미지-텍스트 결합부(1200)는 학습용 테스트로부터 텍스트적 특징자를 생성하는 텍스트 인코더(1201)를 포함할 수 있으며, 도 5를 참조하면, 학습용 텍스트가 텍스트 인코더를 통과하여 텍스트적 특징자인 v (일종의 textual pivot)가 출력되고 있음을 확인할 수 있다.
- [0077] 여기서 텍스트 인코더(1201)는 사전 학습된 공지된 텍스트 인코더를 사용할 수 있는바, 예를 들어 BoW(Bag-of-Words) 인코더, TF-IDF(Term Frequency-Inverse Document Frequency) 인코더, 워드 임베딩(Wprd Embedding) 인코더, 문자 수준(CharacterLevel) 인코더, 트랜스포머 네트워크(Transformer Network) 등 어떠한 것이라도 무방하다 할 것이며, 트랜스포머 네트워크 중 63만개의 파라미터로 이루어진 CLIP(Contrastive Language Image Pre-training)의 텍스트 인코더를 사용할 수 있고, 더 나아가, 텍스트 인코더(1201)로서 후술할 텍스트 설명 생성부(1300)를 이용할 수도 있는바, 이에 대해서는 후술하도록 한다.
- [0078] 텍스트 인코더(1201)의 동작을 보다 구체적으로 설명하면, 학습용 텍스트인 가변 길이의 문장을 입력 받아 고정 길이의 Latent vector인 v 를 출력하며, 여기서 Latent vector를 pivot이라 하여 장치(1000)는 이미지 인코더(1101)가 해당 pivot과 잘 매칭(align)되는 시각적 특징자 x 를 생성할 수 있도록 학습을 진행한다. 본 발명에서 인간의 추론 과정을 적용하여 텍스트가 시각적 의미를 반영하여 클래스를 구분할 수 있는 단서들을 포함하고 있을 것이라 가정하기 때문에 텍스트적 특징자와 시각적 특징자를 맵핑(align)하는 과정을 통해 이미지 인코더(1101)는 지식을 내재화할 수 있는 것이다. 이를 위해 이미지-텍스트 결합부(1200)은 다음과 같은 제2 손실 함수(L_{align})를 이용하여 학습을 진행할 수 있다.
- [0079] 제2 손실 함수(L_{align}) = $\|f_{proj}(v) - g_{proj}(x)\|_2 - \sum_i y_i \log(\hat{y}_i)$
- [0080] 여기서 f_{proj} 는 텍스트적 특징자를 위한 Projection Layer, g_{proj} 은 시각적 특징자를 위한 Projection Layer, y 는 정답 라벨만 1이고 나머지는 0인 one-hot 벡터, $\hat{y}$ 은 $g_{proj}(x)$ 가 분류기(1202)를 통과하여 얻은 softmax 분포를 의미하며, 여기서 분류기(1202)에 대한 설명은 앞선 이미지 분류부(1100)에서의 분류기(1102)에 대한 설명으로 대체하도록 한다.
- [0081] 이상과 같은 제2 손실 함수(L_{align})에서 첫 번째 항은 텍스트적 특징자와 시각적 특징자의 차이, 쉽게 이야기하면 텍스트로 나타내어 지는 특징과 이미지로 나타내어지는 특징의 차이이며, 해당 차이가 작아질수록 학습용 텍스트와 학습용 이미지가 잘 매칭된 것으로 볼 수 있으므로 첫 번째 항만을 이용하여 학습을 진행할 수도 있을 것이나, 본 발명에서는 모든 점이 동일한 점에 프로젝션(projection)되는 경우와 같이 솔루션이 무너지는 경우를 피하기 위한 대응책으로 두 번째 항까지 별도로 둬으로써 학습 성능 향상에 이바지한다는 특징까지 가진다 할 것이다.
- [0082] 도 5는 본 발명의 제2 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치(1000)에 있어서, 텍스트 설명 생성부(1300)의 내부 구조를 예시적으로 도시한 도면이다.
- [0083] 텍스트 설명 생성부(1300)는 시각적 특징자 및 이미지 분류부(1100)로부터 학습용 이미지의 분류 결과를 입력 받아 이에 근거하여 학습용 이미지를 설명하는 텍스트(Text)를 생성한다.

- [0084] 이를 위해 텍스트 설명 생성부(1300)는 두 개의 LSTM(Long Short Term Memory) 레이어를 포함하며, 이를 각각 제1 LSTM 레이어(1301) 및 제2 LSTM 레이어(1302)라고 한다.
- [0085] 제1 LSTM 레이어(1301)에서는 이전에 생성한 토큰 o_{t-1} 을 입력으로 받고 hidden state를 업데이트하여 출력 z_t 를 생성하며, 해당 출력은 프로젝션된 시각적 특징자와 이미지 분류부(1100)가 포함하는 분류기(1102)의 학습용 이미지의 분류 결과인 예측값을 붙인 벡터와 함께 제2 LSTM 레이어(1302)로 입력될 수 있다.
- [0086] 이를 제2 LSTM 레이어(1302) 입장에서 설명하면 제1 LSTM 레이어(1301)의 출력 토큰 z_t 과 프로젝션된 시각적 특징자 및 학습용 이미지의 분류 결과(예측값)를 입력 받는 것이며, 제2 LSTM 레이어(1302)는 입력 받은 입력값에 기초하여 텍스트 설명 생성부(1300), 보다 구체적으로 제2 LSTM 레이어(1302)가 보유하고 있는 각 단어의 softmax 확률 값 $p(o_t)$ 을 생성하여 가장 높은 확률 값을 가진 단어 토큰을 출력하고, 단어 토큰의 출력을 제2 LSTM 레이어(1302)가 출력한 단어 코튼이 EOS(End Of Sentence token) 또는 사전에 설정한 샘플링 단어 토큰의 개수를 만족할 때까지 반복한다.
- [0087] 이 때, 텍스트 설명 생성부(1300)는 학습용 이미지가 나타내는 클래스를 조금 더 잘 분류할 수 있는, 조금 더 잘 설명할 수 있는 텍스트를 생성하기 위해 강화 학습에 기반한 제3 손실 함수(L_{expl})를 이용하여 학습을 진행한다.
- [0088] 제3 손실 함수(L_{expl}) = $-\sum_t \log P(o_{t+1}|o_{0:t}, I, C) - E_{\tilde{o} \sim P(o|I, C)}[R(\tilde{o})]$
- [0089] 여기서 $\tilde{o} \sim P(o|I, C)$ 는 학습용 이미지를 설명하는 텍스트를 생성하기 위해 샘플링한 문장, $P(o|I, C)$ 는 학습용 이미지 I와 클래스 C가 조건으로 주어졌을 때 description o에대한 추정 분포, $R(\tilde{o})$ 는 $P(c|\tilde{o})$ 로 생성된 문장 $\tilde{o}$ 이 있을 때, 클래스 당 softmax 확률을 의미함
- [0090] 한편, 앞서 이미지-텍스트 결합부(1200)에 대한 설명에서 텍스트 인코더(1201)로 텍스트 설명 생성부(1300)를 이용할 수 있다고 했던바, 텍스트 설명 생성부(1300)가 자연어(텍스트)로 장치(1000)의 선택에 대한 근거를 정당화시킬 수 있는, 쉽게 이야기하면 텍스트 설명 생성부(1300)가 장치(1000)의 학습용 이미지 분류 결과를 뒷받침해줄 수 있는 텍스트를 생성하기에 두 개의 LSTM 레이어를 통해 softmax 확률 값이 가장 큰 단어 토큰이 출력되어 가면서 마지막으로 업데이트되는 hidden state를 텍스트적 특징자 v로 볼 수 있기 때문이다.
- [0091] 지금까지 본 발명의 제2 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치(1000)가 이미지 분류를 위해 수행하는 동작 또는 연산을 기능적 구성으로 명명하여 학습용 데이터셋을 이용해 학습을 진행하는 경우에 대하여 설명하였다. 이미지 분류부(1100), 이미지-텍스트 결합부(1200), 텍스트 설명 생성부(1300) 각각의 학습 진행에 적용되는 손실 함수가 제1 손실 함수(L_{task}), 제2 손실 함수(L_{align}), 제3 손실 함수(L_{expl})임은 앞서 설명하였으며, 장치(1000)의 전체 측면에서 본다면 전체 손실 함수(L_{total})는 다음과 같이 세 개의 손실 함수 전체를 합친 것일 수 있고 학습은 전체 손실 함수(L_{total})을 최소화시키는 최적화 과정으로 진행될 수 있을 것이며, 도 6에 이와 관련된 장치(1000)의 전체 내부 구조를 예시적으로 도시해 놓았다.
- [0092] 전체 손실 함수(L_{total}) = 제1 손실 함수(L_{task}) + λ_{align} 손실 함수(L_{align}) + λ_{expl} 제3 손실 함수(L_{expl})
- [0093] 여기서 λ_{align} 와 λ_{expl} 는 각각 alignment Loss와 explanation Loss의 강도를 조절하는 하이퍼파라미터를 의미함
- [0094] 이상 장치(1000)의 학습에 관련된 구성들의 설명을 마무리하겠으며, 이번에는 장치(1000)가 학습을 완료한 후, 특정 도메인으로 표현된 실행용 이미지를 입력 받아 실제 이미지 분류를 진행하는 경우에 대하여 간단히 설명하도록 한다.
- [0095] 앞서 언급한 바와 같이 학습이 완료된 장치(1000)가 실제 이미지 분류를 진행하는 경우에는 이미지 분류부(1100)만 동작하며, 장치(1000)의 학습 성능 향상에 이바지하는 이미지-텍스트 결합부(1100)과 텍스트 설명 생성부(1200)는 동작하지 않는다. 학습 과정에서 이들의 동작으로 인해 이미지 분류부(1100)가 영향을 받아 학습이 진행되었기 때문이다.
- [0096] 그렇기 때문에 실제 이미지 분류에 대한 설명은 매우 간단한바, 특정 도메인(여기서 특정 도메인은 학습용 이미지의 도메인일 수도 있고, 아닐 수도 있음. 아닌 경우가 본 발명의 목적에 보다 부합함)으로 표현된 실제 이미

지를 입력 받아 이미지 인코더(1101)가 시각적 특징자 x 를 생성하고, 생성한 시각적 특징자 x 가 분류기(1102)를 통과하여 이미지 분류 결과를 생성해 출력하게 되는데 여기서 이미지 인코더(1101) 및 분류기(1102)가 인간의 추론 과정을 적용하여 입력 받은 이미지로부터 텍스트적 특징과 잘 매칭(aligned)되는 시각적 특징자 x 를 생성하고 이로부터 인간의 사고 개념과 유사하게 이미지 분류 결과를 출력하도록 학습되어 있기 때문이다.

[0097] 지금까지 본 발명의 제2 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치(1000)에 대하여 설명하였다. 본 발명에 따르면, 학습 과정에서 이미지뿐만 아니라 텍스트 정보를 함께 사용하여 학습을 진행하는바, 인간의 추론 과정을 적용하여 일부 도메인의 이미지만을 학습한 경우라 할지라도 이와 상이한 도메인의 이미지까지 높은 정확도로 효과적으로 분류할 수 있다.

[0098] 본 발명의 제2 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치(1000)의 성능 평가 자료인 도 7 내지 도 9중, 우선 도 7을 참조하면, 종래 기술인 ERM 모델에 따른 이미지 분류 결과는 서로 다른 클래스를 의미하는 서로 다른 색상이 동일한 도메인을 의미하는 같은 무늬끼리 뭉쳐져 있어 클래스 분류가 잘 되지 않은 것을 확인할 수 있으나, 본 발명(텍스트 인코더로 CLIP 텍스트 인코더를 사용한 경우)인 도 8을 참조하면, 동일한 클래스를 의미하는 같은 색상이 뭉쳐져 있고 도메인을 의미하는 무늬가 흩어져 있는 것을 확인할 수 있는바, 도메인과 무관하게 클래스별로 이미지 분류가 잘 되었음을 확인할 수 있다.

[0099] 또한, 도 9를 참조하면, 본 발명의 예측 정확도 평균이 종래 기술들의 예측 정확도 평균보다 높은 것을 확인할 수 있는바, 하나만 예를 들어 설명하면, 텍스트 인코더(1201)로 CLIP 텍스트 인코더를 사용한 경우의 본 발명에서 카툰, 아트, 페인트 도메인 이미지로 학습하였을 때, 포토 도메인의 이미지를 입력 받은 경우의 예측 정확도가 74.6, 텍스트 인코더(1201)로 텍스트 설명 생성부(1300)를 사용한 경우의 본 발명에서 동일하게 테스트했을 때 예측 정확도가 74.3으로 종래 기술들을 동일하게 테스트 했을 때의 예측 정확도보다 현저하게 높음을 확인할 수 있다.

[0100] 마지막으로, 본 발명의 제1 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치(100)와 본 발명의 제2 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치(1000)는 도 10에 예시적으로 도시한 바와 같이 본 발명의 제3 실시 예에 따른 텍스트를 활용하여 도메인 비특이적인 이미지를 분류하는 방법으로 구현할 수 있으며, 이 경우 본 발명의 제3 실시 예에 따른 텍스트를 활용하여 도메인 비특이적인 이미지를 분류하는 방법은 (a) 학습용 데이터셋을 입력 받아 전체 손실 함수(L_{total})를 최소화하는 방향으로 학습을 수행하는 제1 단계 및 (b) 특정 도메인으로 표현된 실행용 이미지를 입력받아 상기 실행용 이미지의 분류 결과를 출력하는 제2 단계를 포함하며, 상기 학습용 데이터셋은, 하나 이상의 클래스(Class)에 대하여 하나 이상의 도메인으로 표현한 하나 이상의 학습용 이미지 및 상기 하나 이상의 클래스에 대하여 하나 이상의 텍스트로 설명한 하나 이상의 학습용 텍스트를 포함하고, 상기 제1 단계는, 상기 학습용 이미지를 입력 받아 이로부터 시각적 특징자(Visual Feature)를 생성하고, 상기 생성한 시각적 특징자를 이용하여 상기 학습용 이미지의 분류 결과를 생성하여 출력하며, 제1 손실 함수(L_{task})를 이용하여 학습을 진행하는 제1-1 단계, 상기 학습용 텍스트를 입력 받아 이로부터 텍스트적 특징자(Textual Feature)를 생성하고, 상기 시각적 특징자와 텍스트적 특징자를 맵핑시키며, 제2 손실 함수(L_{align})를 이용하여 학습을 진행하는 제1-2 단계 및 상기 시각적 특징자 및 상기 생성한 학습용 이미지의 분류 결과를 입력 받아 이에 근거하여 상기 학습용 이미지를 설명하는 텍스트(Text)를 생성하며, 제3 손실 함수(L_{expl})를 이용하여 학습을 진행하는 제1-3 단계를 포함할 수 있을 것이다.

[0101] 더 나아가, 본 발명의 제1 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치(100)와 본 발명의 제2 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치(1000)는 본 발명의 제4 실시 예에 따른 컴퓨터로 판독 가능한 매체에 저장된 컴퓨터 프로그램으로 구현할 수도 있는바, 이 경우 컴퓨팅 장치와 결합하여 (AA) 학습용 데이터셋을 입력 받아 전체 손실 함수(L_{total})를 최소화하는 방향으로 학습을 수행하는 제1 단계 및 (BB) 특정 도메인으로 표현된 실행용 이미지를 입력받아 상기 실행용 이미지의 분류 결과를 출력하는 제2 단계를 실행하며, 상기 학습용 데이터셋은, 하나 이상의 클래스(Class)에 대하여 하나 이상의 도메인으로 표현한 하나 이상의 학습용 이미지 및 상기 하나 이상의 클래스에 대하여 하나 이상의 텍스트로 설명한 하나 이상의 학습용 텍스트를 포함하고, 상기 제1 단계는, 상기 학습용 이미지를 입력 받아 이로부터 시각적 특징자(Visual Feature)를 생성하고, 상기 생성한 시각적 특징자를 이용하여 상기 학습용 이미지의 분류 결과를 생성하여 출력하며, 제1 손실 함수(L_{task})를 이용하여 학습을 진행하는 제1-1 단계, 상기 학습용 텍스트를 입력 받아 이로부터 텍스트적 특징자(Textual Feature)를 생성하고, 상기 시각적 특징자와 텍스트적 특징자를 맵핑시키며, 제2 손실 함수(L_{align})를 이용하여 학습을 진행하는 제1-2 단계 및 상기 시각적 특징자 및 상기 생성한 학습용 이

미지의 분류 결과를 입력 받아 이에 근거하여 상기 학습용 이미지를 설명하는 텍스트(Text)를 생성하며, 제3 손실 함수(L_{expl})를 이용하여 학습을 진행하는 제1-3 단계를 포함할 수 있을 것이다.

[0102] 중복 서술을 위해 자세히 기재하지는 않았지만 본 발명의 제1 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치(100) 및 본 발명의 제2 실시 예에 따른 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치(1000)에 적용된 모든 기술적 특징은 본 발명의 제3 실시 예에 따른 텍스트를 활용하여 도메인 비특이적인 이미지를 분류하는 방법 및 제4 실시 예에 따른 컴퓨터로 판독 가능한 매체에 저장된 컴퓨터 프로그램에 모두 동일하게 적용될 수 있음은 물론이라 할 것이다.

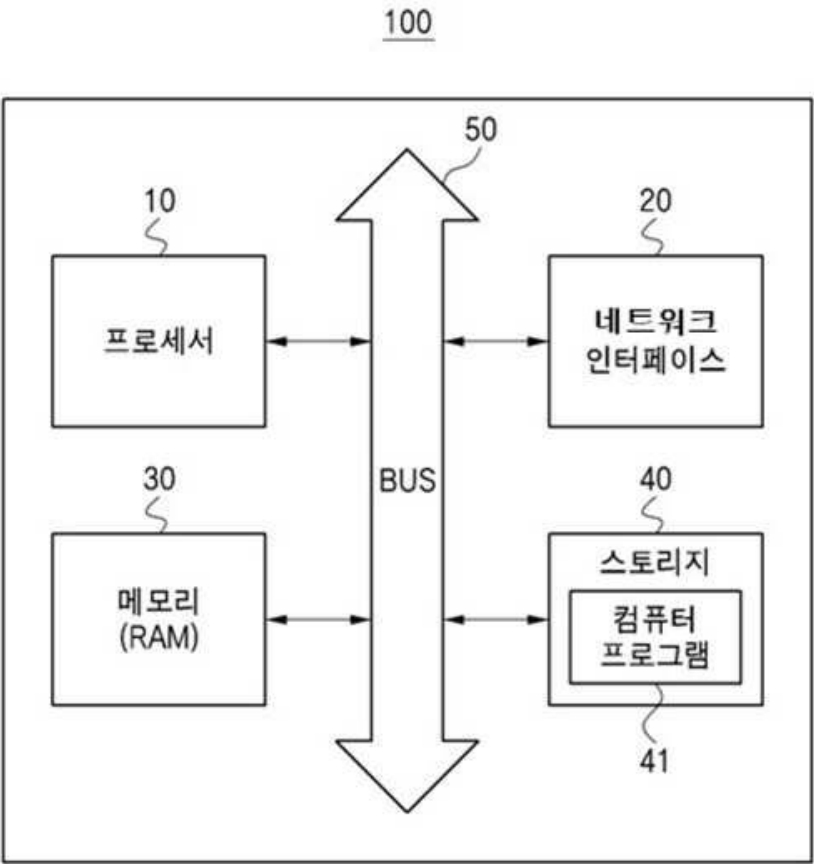
[0103] 이상 첨부된 도면을 참조하여 본 발명의 실시 예들을 설명하였지만, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 그 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 실시될 수 있다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시 예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다.

부호의 설명

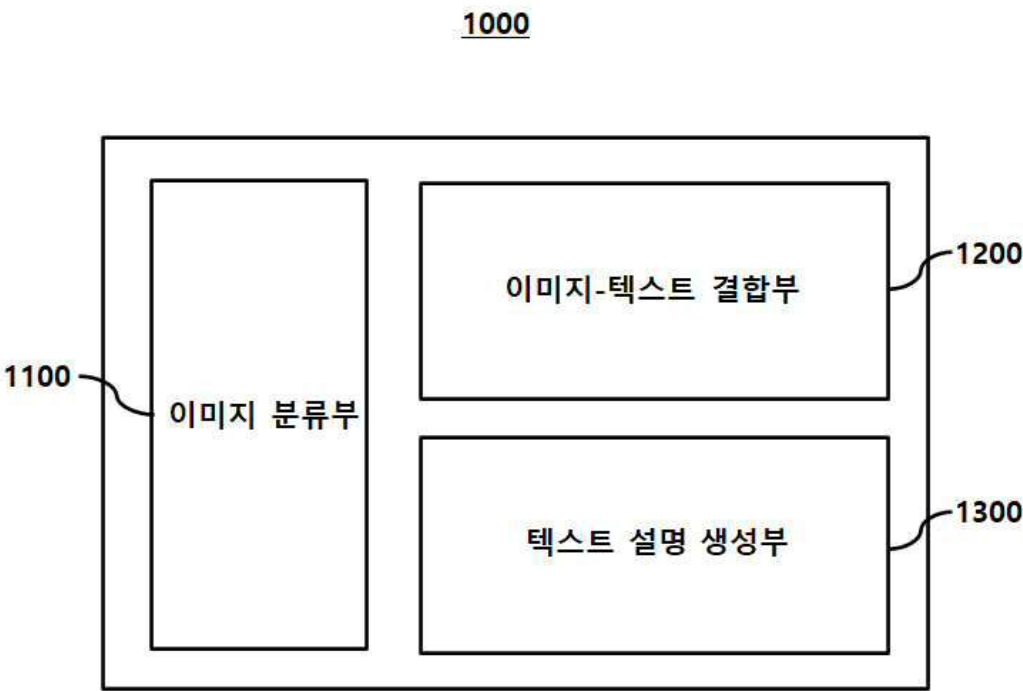
[0104] 10: 프로세서
20: 네트워크 인터페이스
30: 메모리
40: 스토리지
41: 컴퓨터 프로그램
50: 정보 버스
100, 1000: 텍스트를 활용한 도메인 비특이적인 이미지 분류 장치
1100: 이미지 분류부
1101: 이미지 인코더
1102: 분류기
1200: 이미지-텍스트 결합부
1201: 텍스트 인코더
1202: 분류기
1300: 텍스트 설명 생성부
1301: 제1 LSTMT 네트워크
1302: 제2 LSTM 네트워크

도면

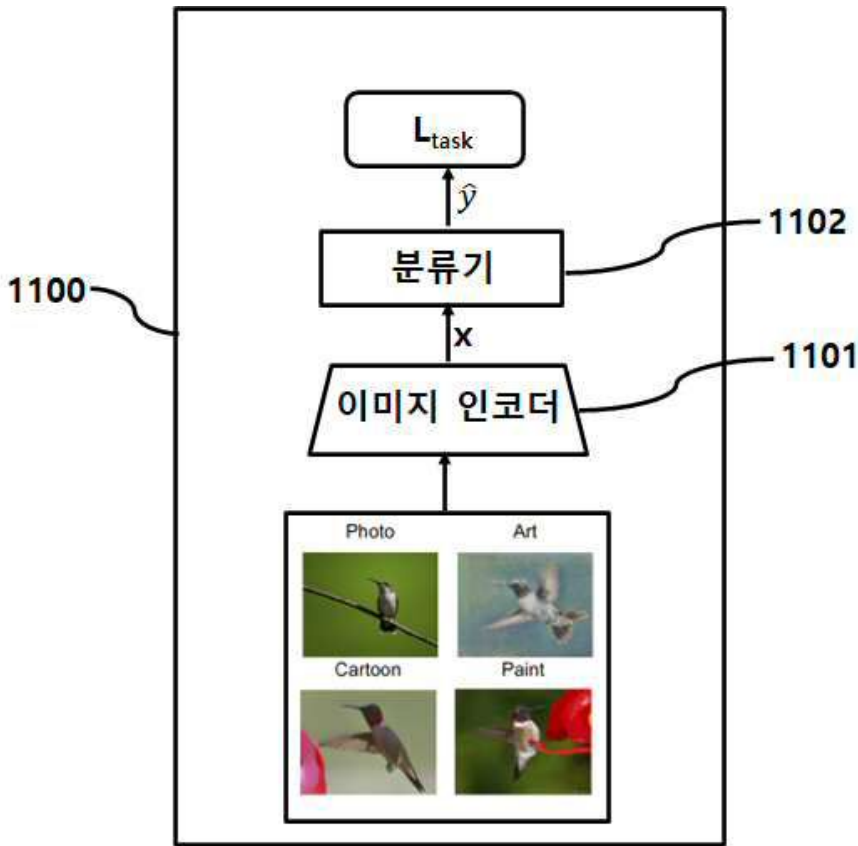
도면1



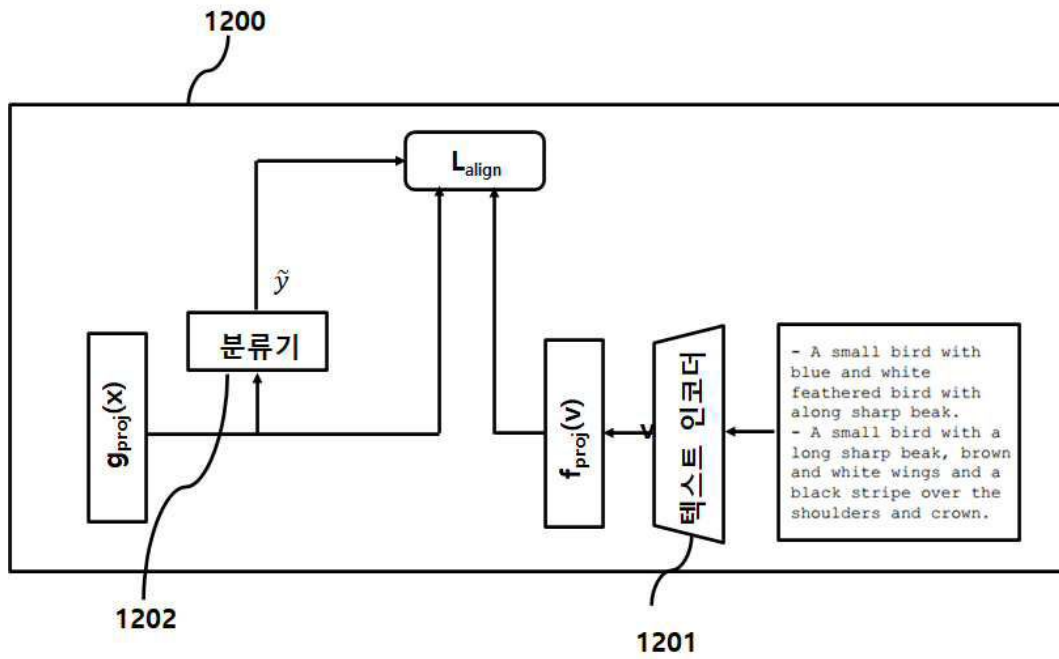
도면2



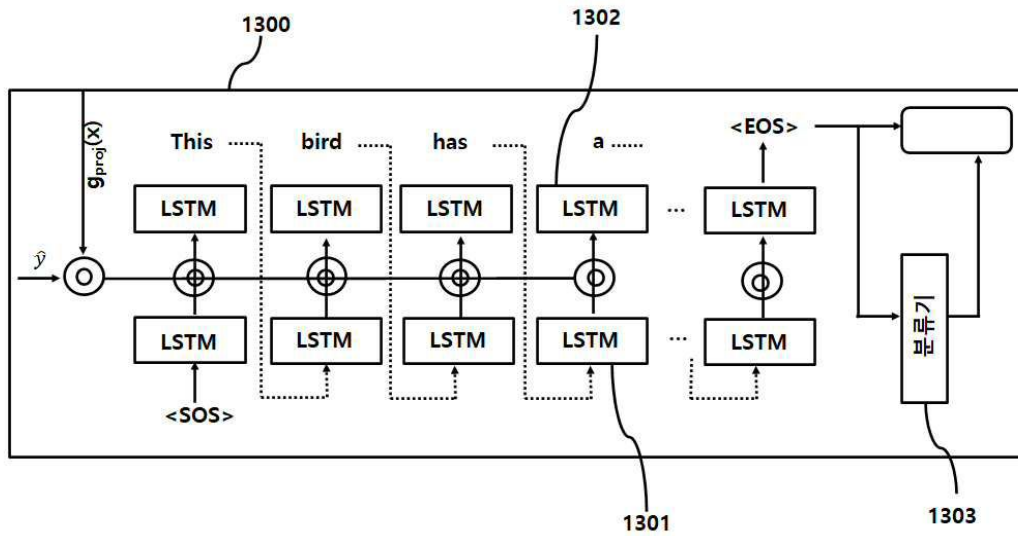
도면3



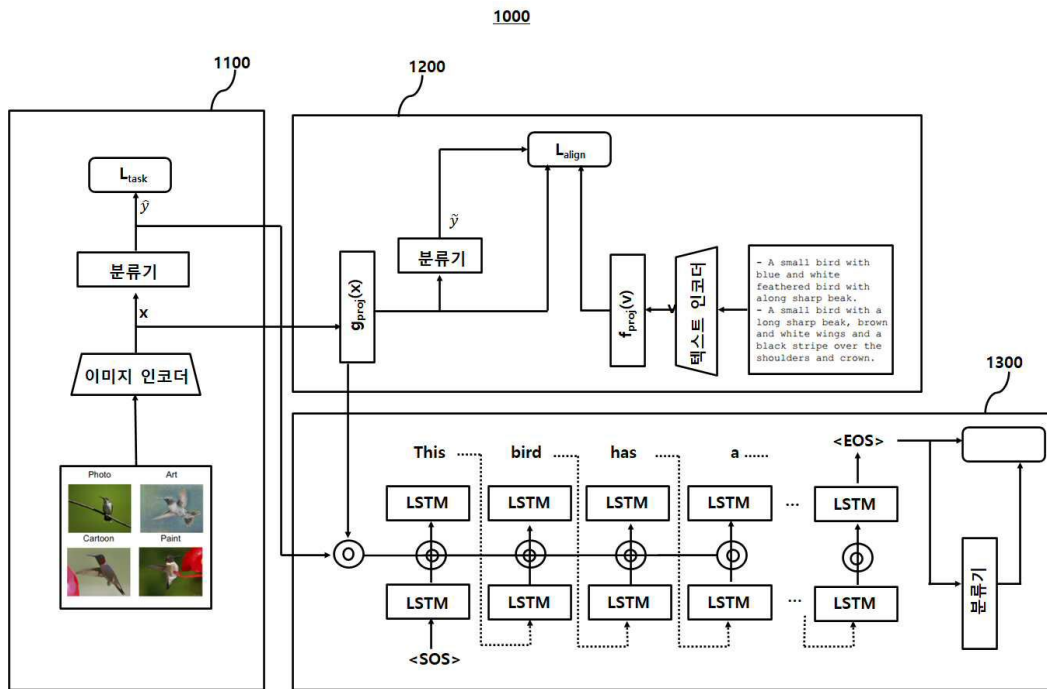
도면4



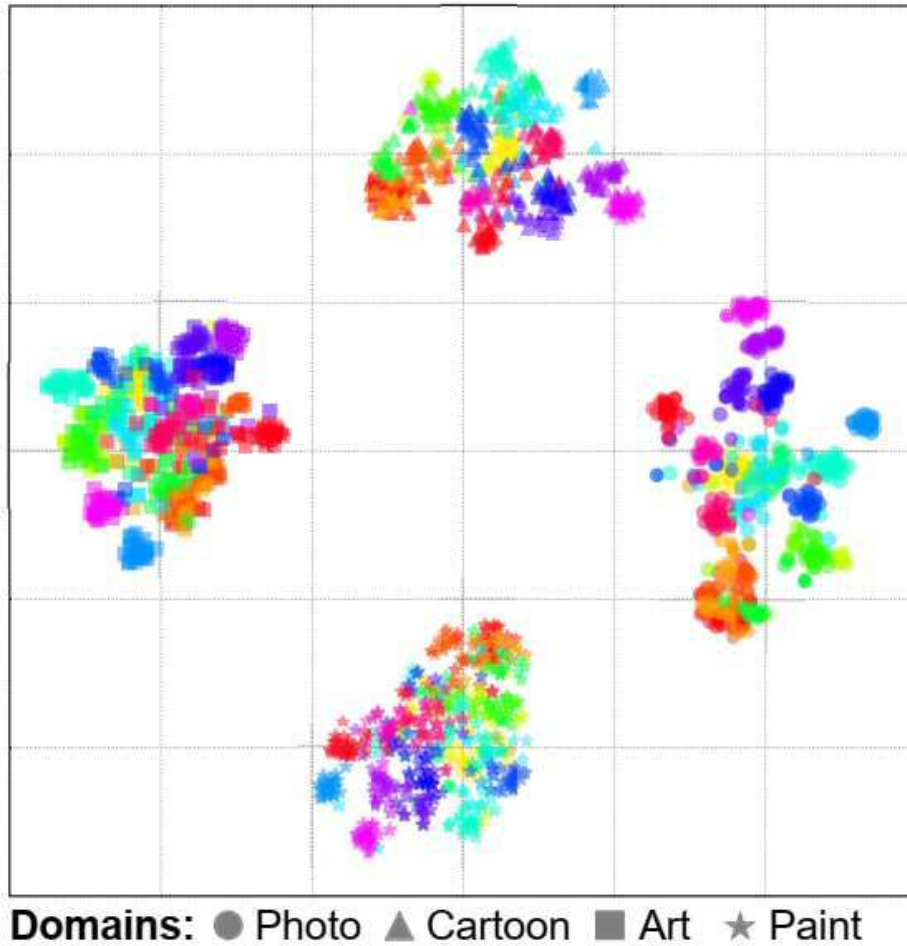
도면5



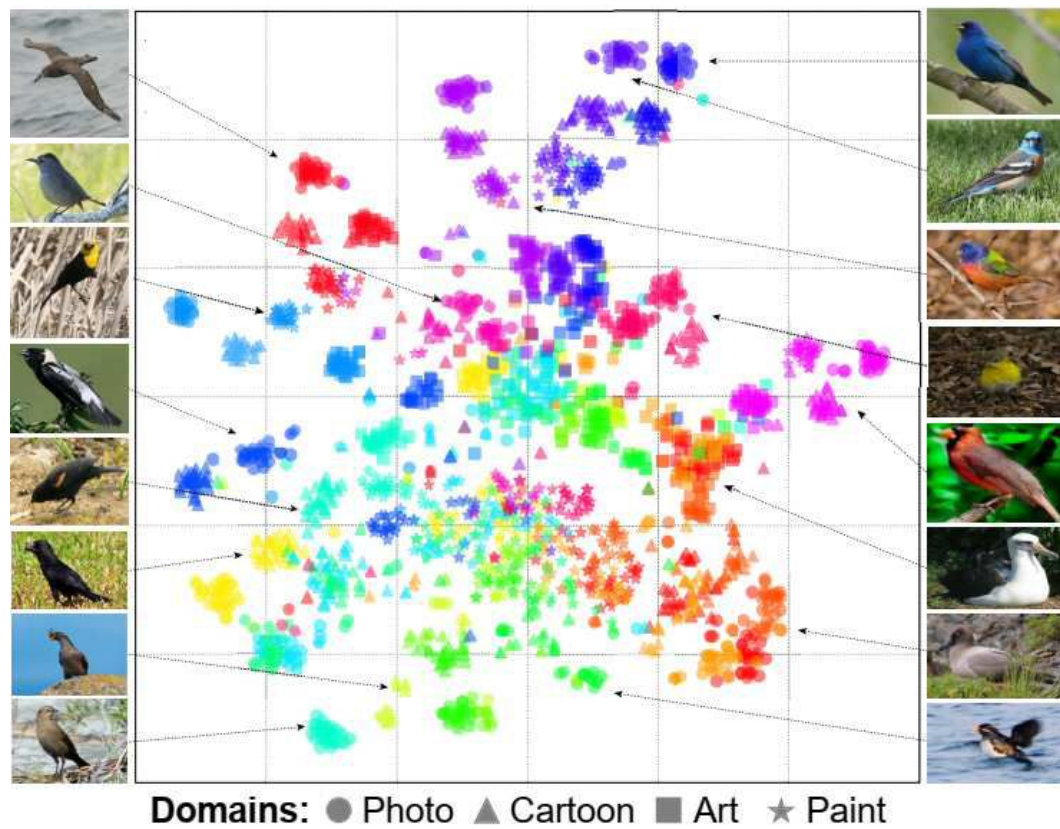
도면6



도면7



도면8



도면9

Model	<i>D</i>	<i>A</i>	<i>M</i>	<i>T</i>	Target Domain				Avg
					Photo	Cartoon	Art	Paint	
본 발명 / PTE	✓			✓	74.6	64.2	52.2	37.0	57.0
본 발명 / STE	✓			✓	74.3	63.9	50.0	38.1	56.6
CORAL	✓				72.2	63.5	50.3	35.8	55.4
SD					71.3	62.2	50.8	34.8	54.7
SagNet	✓	✓			67.4	60.7	44.0	34.2	51.6
MixStyle			✓		59.0	56.7	50.3	35.8	50.4
Mixup			✓		67.1	55.9	51.1	27.2	50.3
DANN	✓	✓			67.5	57.0	42.8	30.6	49.5
CDANN	✓	✓			65.3	55.2	43.2	30.5	48.6
VREx	✓				63.9	54.9	38.6	30.1	46.9
ERM					62.5	53.2	37.4	29.0	45.5
ARM					62.3	51.2	38.2	28.4	45.0
GroupDRO	✓				60.9	54.8	36.5	27.0	44.8
IRM					60.6	51.6	36.5	30.3	44.8

도면10

