



(19) 대한민국특허청(KR)

(12) 등록특허공보(B1)

(45) 공고일자 2022년08월05일

(11) 등록번호 10-2430542

(24) 등록일자 2022년08월03일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01) G06N 3/04 (2006.01)
G06N 5/02 (2006.01)

(52) CPC특허분류
G06N 3/08 (2013.01)
G06K 9/6267 (2022.01)

(21) 출원번호 10-2021-0042691

(22) 출원일자 2021년04월01일

심사청구일자 2021년04월01일

(56) 선행기술조사문헌

Zhenwei He and Lei Zhang, "Multi-adversarial Faster-RCNN for Unrestricted Object Detection," arXiv:1907.10343v2 [cs.CV] 7 Sep 2019 (2019.09.07.)*

*는 심사관에 의하여 인용된 문헌

(73) 특허권자

한밭대학교 산학협력단

대전광역시 유성구 동서대로 125 (덕명동)

(72) 발명자

이현빈

대전광역시 서구 둔산로 155, 109동 703호(둔산동, 크로바아파트)

경민영

대전광역시 서구 청사서로 70, 202동 801호(월평동, 무궁화아파트)

(74) 대리인

특허법인오암

전체 청구항 수 : 총 5 항

심사관 : 송근배

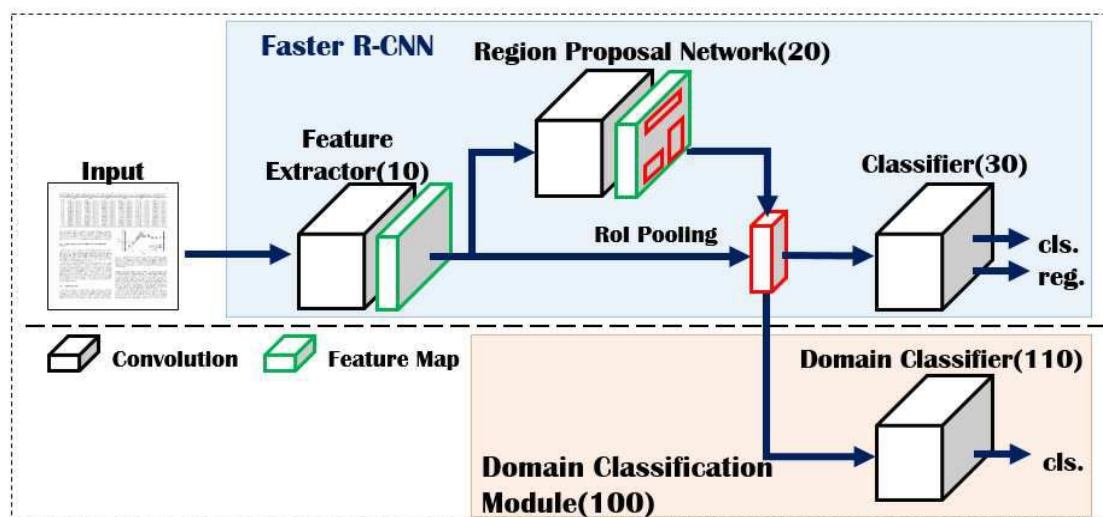
(54) 발명의 명칭 문서 이미지의 표 오검출을 줄이기 위한 도메인 분류 학습 방법 및 시스템

(57) 요약

검출 단계를 후보 영역 검출과 클래스 분류로 구분하는 검출기로서, 객체 검출을 위한 특징을 추출하는 특징 추출기(Feature Extractor), RPN(Region Proposal Network, RPN) 그리고 관심영역(Region-of-Interest, RoI) 기반의 분류기(Classifier)를 포함하는 도메인 분류 학습 시스템에 있어서, 상기 특징 추출기는 입력 이미지의 특

(뒷면에 계속)

대표도 - 도1



정을 추출하여 특징맵(feature map) 형태로 출력하고, RPN에서는 특징맵을 통해 객체 후보 영역인 바운딩 박스의 좌표를 추론하고, 상기 분류기는 특징맵 내 모든 픽셀마다 정의된 앵커(anchor)를 기준으로 생성되는 각 바운딩 박스를 대상으로 클래스를 분류하고 상기 좌표를 보정하되, 상기 도메인 분류 학습 시스템은 표 영역을 학습 목표인 소스 도메인으로 정의하고, 배경 영역을 타겟 도메인으로 정의하며, 정의된 두 도메인을 도메인 분류 모델에 기반하여 학습을 통해 분류하여 상기 분류기에 제공하고, 상기 도메인 분류 모델은 타겟 도메인인 배경 영역에 대해 잘못 예측한 표 후보 영역에 대한 손실값을 계산하여 분류하는 도메인 분류 모듈을 더 포함한다.

(52) CPC특허분류

G06N 3/04 (2013.01)

G06N 5/025 (2019.01)

G06V 10/46 (2022.01)

명세서

청구범위

청구항 1

검출 단계를 후보 영역 검출과 클래스 분류로 구분하는 검출기로서, 객체 검출을 위한 특징을 추출하는 특징 추출기(Feature Extractor), RPN(Region Proposal Network, RPN) 그리고 관심영역(Region-of-Interest, RoI) 기반의 분류기(Classifier)를 포함하는 도메인 분류 학습 시스템에 있어서,

상기 특징 추출기는 입력 이미지의 특징을 추출하여 특징맵(feature map) 형태로 출력하고, RPN에서는 특징맵을 통해 객체 후보 영역인 바운딩 박스의 좌표를 추론하고, 상기 분류기는 특징맵 내 모든 픽셀마다 정의된 앵커(anchor)를 기준으로 생성되는 각 바운딩 박스를 대상으로 클래스를 분류하고 상기 좌표를 보정하되,

상기 도메인 분류 학습 시스템은,

표 영역을 학습 목표인 소스 도메인으로 정의하고, 배경 영역을 타겟 도메인으로 정의하며,

상기 도메인 분류 학습 시스템은, 도메인 분류 모듈을 더 포함하고,

상기 도메인 분류 모듈은, 정의된 두 도메인을 도메인 분류 모델에 기반하여 학습을 통해 분류하여 상기 분류기에 제공하고, 상기 도메인 분류 모듈은 타겟 도메인인 배경 영역에 대해 잘못 예측한 표 후보 영역에 대한 손실 값을 계산하되, 상기 손실값은 상기 도메인 분류 모델에 기반하는 학습인 모델학습을 위해 정의되는 손실 함수 L_{det} 를 이용하여 바운딩 박스별 도메인 판별 및 레이블 정의하여 오차를 계산하고,

상기 도메인 분류 모듈은,

모듈 내 도메인 분류기를 더 포함하고,

상기 도메인 분류기는, RPN의 바운딩 박스를 입력받아 각 박스별 도메인을 판별하는 도메인 분류 학습 시스템.

청구항 2

삭제

청구항 3

삭제

청구항 4

제1항에 있어서,

상기 도메인 분류 모듈에 있어서, 상기 손실 함수 L_{det} 는 RPN의 오차를 계산하는 손실 L_{rpn} 과 분류기의 오차를 계산하는 손실 L_{roi} 로 구성되며,

상기 복수의 손실은 예측한 바운딩 박스의 클래스가 얼마나 정확한지를 기준으로 오차를 측정하는 클래스피케이션 손실과 바운딩 박스의 좌표가 실제 객체의 위치와 얼마나 가까운지를 기준으로 오차를 측정하는 리그레션 손실을 포함하는 도메인 분류 학습 시스템.

청구항 5

제1항에 있어서,

상기 도메인 분류 모듈은

바운딩 박스별로 도메인을 얼마나 잘 판별했는지를 기준으로 오차를 계산하되,

상기 오차를 계산하기 위해 하기의 수학적 식 1과 같은 cross-entropy loss 기반의 손실 함수 L_{dom} 을 정의하고,

상기 도메인 분류기 h는 특징맵의 u,v 위치에 있는 바운딩 박스별로 도메인을 판별하며, 도메인 레이블 D_i 는 바운딩 박스가 소스 도메인일 경우에 0으로, 타겟 도메인일 경우에는 1로 정의되는 것을 특징으로 하는 도메인 분류 학습 시스템.

[수학적 식 1]

$$L_{dom} = -\frac{1}{N} \sum_{i=1}^N [D_i \log h_i^{(u,v)} + (1-D_i) \log (1-h_i^{(u,v)})]$$

청구항 6

제5항에 있어서,

상기 도메인 레이블의 선정 기준은

상기 바운딩 박스가 ground truth 레이블의 어느 표 영역과의 Intersection-over-Union(IoU) 비율이 설정된 상위 임계값 이상일 경우 소스 도메인으로 정의하고,

상기 바운딩 박스가 ground truth 레이블의 모든 표 영역과의 IoU 비율이 설정된 하위 임계값 이하일 경우 타겟 도메인으로 정의하는 것을 특징으로 하는 도메인 분류 학습 시스템.

청구항 7

삭제

청구항 8

삭제

청구항 9

검출 단계를 후보 영역 검출과 클래스 분류로 구분하는 검출기로서, 객체 검출을 위한 특징을 추출하는 특징 추출기(Feature Extractor), RPN(Region Proposal Network, RPN) 그리고 관심영역(Region-of-Interest, RoI) 기반의 분류기(Classifier)를 포함하는 도메인 분류 학습 시스템을 이용한 도메인 분류 학습 방법에 있어서,

상기 도메인 분류 학습 시스템은, 도메인 분류 모듈을 더 포함하고,

상기 도메인 분류 모듈은, 모듈 내 도메인 분류기를 더 포함하고,

입력 이미지의 특징을 추출하여 특징맵(feature map) 형태로 출력하는 단계;

상기 특징맵을 통해 객체 후보 영역인 바운딩 박스의 좌표를 추론하는 단계;

특징맵 내 모든 픽셀마다 정의된 앵커(anchor)를 기준으로 생성되는 각 바운딩 박스를 대상으로 클래스를 분류하고 상기 좌표를 보정하되, 상기 도메인 분류기는, RPN의 바운딩 박스를 입력받아 각 박스별 도메인을 판별하는 단계; 및

상기 도메인 분류 모듈은, 정의된 두 도메인을 도메인 분류 모델에 기반하여 학습을 통해 분류하여 상기 분류기에 제공하고, 상기 도메인 분류 모듈은 타겟 도메인인 배경 영역에 대해 잘못 예측한 표 후보 영역에 대한 손실값을 계산 및 분류하는 단계를 포함하고,

상기 손실값을 계산 및 분류하는 단계는, 손실함수를 이용하여 바운딩 박스별 도메인 판별 및 레이블 정의하여 오차 계산하는 단계를 더 포함하는 도메인 분류 학습 방법.

청구항 10

삭제

발명의 설명

기술분야

[0001] 본 발명은 도메인 분류 학습 시스템 및 방법에 관한 것으로, 더욱 상세하게는 문서 이미지 내 표 오검출을 줄이기 위한 도메인 분류 학습 시스템 및 방법에 관한 것이다.

배경기술

[0002] 문서 상에 존재하는 표(table)는 요약한 정보나 수치 데이터를 쉽게 알아볼 수 있도록 표현하기 위하여 사용된다. 특히, 조사, 분석, 실험, 평가 등의 내용을 포함한 문서는 핵심 내용을 체계적으로 정리하여 제시하거나, 비교, 분석, 평가 결과를 보기 쉽게 나타내기 위하여 표를 많이 사용한다. 따라서 다양한 문서 자동화 분야에서 문서를 분석할 때 표의 내용을 추출하여 검토하는 과정이 반드시 필요하다.

[0003] 하지만, 표의 형태가 다양하기 때문에 문서 분석 자동화 시스템 구현 시 모든 표를 정확히 검출해내기는 매우 어렵다.

[0004] 기존에 제안된 표 검출 기법의 대부분은 표의 형태적 특징을 추출하거나 정형화된 문서를 분석하는 등 연구자가 직접 설계한 특징(hand-crafted features)에 의존하기 때문에, 다양한 문서를 대상으로 일반화하여 적용하기엔 한계가 있다.

[0005] 이러한 문제를 해결하기 위해 최근에는 딥러닝을 활용한 기법들이 제안되어 검출 가능한 표의 형태와 문서의 종류를 늘려왔으며, TableBank를 학습한 Faster R-CNN 기반의 표 검출 기법의 경우, 문서 내 표의 위치를 정확하게 찾아내지만, 검출 성능을 떨어뜨리는 문제가 여전히 존재하며, 이러한 문제는 크게 오검출(mis-detection)과 미검출(un-detection) 그리고 부분검출(partial-detection)로 구분할 수 있다.

[0006] 일반적으로 딥러닝 기반의 검출 기법은 검출 대상이 가지는 고유한 특징을 추출하는 것이 매우 중요하며, 영상 내 다양한 객체를 정확하게 분별하기 위해 각 객체별 학습 데이터가 균등하게 분포되는 것이 적절하다. 그러나 TableBank는 문서 내 표의 위치 정보만으로 구성되어 있기 때문에 학습 데이터의 분포가 불균형하며, 그 결과, 그래프나 그림과 같은 다른 객체를 분별하지 못하고 표로 잘못 인식하는 문제가 발생한다. 이러한 문제를 해결하기 위해 학습 데이터에 그래프와 그림에 대한 정보를 추가하는 작업이 요구되지만, 대량의 데이터를 일일이 레이블링하는 작업은 매우 많은 시간과 비용을 필요로 하는 문제점이 있었다.

[0007] 따라서, 진술한 문제를 해결하기 위하여 별도의 학습 데이터 추가없이 표 오검출 문제를 해결하기 위해 배경 영역에 대한 적절한 손실 값을 추가하고, 두 영역 간의 공통 특징 벡터를 제외한 표 영역만의 고유 특징 벡터를 학습할 수 있는 알고리즘을 이용한 도메인 분류 학습 시스템 및 방법에 대한 연구가 필요하게 되었다.

선행기술문헌

비특허문헌

[0008] (비특허문헌 0001) M. Li et al., "TableBank: Table Benchmark for Image-based Table Detection and Recognition," Proc. of the 12th Conference on Language Resources and Evaluation, pp. 1918-1925, 2020.

(비특허문헌 0002) S. Ren et al., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, No. 6, pp. 1137-1149, Jun, 2017.

발명의 내용

해결하려는 과제

[0009] 본 발명의 목적은 두 영역 간의 공통적인 특징 벡터를 분석하기 위해 도메인 간의 관계를 분석하는 도메인 적응 기술을 활용하고, 표 영역을 소스 도메인으로, 배경 영역을 타겟 도메인으로 정의하고 두 도메인을 분류하는 도메인 분류 모듈을 이용하는 도메인 분류 학습 방법 및 시스템을 제공하는 것이다.

과제의 해결 수단

[0010] 본 발명의 일 실시예에 따른 도메인 분류 학습 시스템은, 검출 단계를 후보 영역 검출과 클래스 분류로 구분하

는 검출기로서, 객체 검출을 위한 특징을 추출하는 특징 추출기(Feature Extractor), RPN(Region Proposal Network, RPN) 그리고 관심영역(Region-of-Interest, RoI) 기반의 분류기(Classifier)를 포함하는 도메인 분류 학습 시스템에 있어서, 상기 특징 추출기는 입력 이미지의 특징을 추출하여 특징맵(feature map) 형태로 출력하고, RPN에서는 특징맵을 통해 객체 후보 영역인 바운딩 박스의 좌표를 추론하고, 상기 분류기는 특징맵 내 모든 픽셀마다 정의된 앵커(anchor)를 기준으로 생성되는 각 바운딩 박스를 대상으로 클래스를 분류하고 상기 좌표를 보정하되, 상기 도메인 분류 학습 시스템은 표 영역을 학습 목표인 소스 도메인으로 정의하고, 배경 영역을 타겟 도메인으로 정의하며, 정의된 두 도메인을 도메인 분류 모델에 기반하여 학습을 통해 분류하여 상기 분류기에 제공하고, 상기 도메인 분류 모델은 타겟 도메인인 배경 영역에 대해 잘못 예측한 표 후보 영역에 대한 손실값을 계산하여 분류하는 도메인 분류 모듈을 더 포함한다.

[0011] 상기 도메인 분류 모듈은 RPN의 바운딩 박스를 입력받아 각 박스별 도메인을 판별하는 도메인 분류기를 더 포함한다.

[0012] 상기 도메인 분류 모듈은 상기 소스 도메인의 에러와 두 도메인 분포 간의 거리의 합으로 타겟 도메인의 에러 범위를 예측하되, 도메인 분포 간의 거리를 측정하기 위해 H-divergence를 사용하며, 상기 H-divergence는 영역 거리(domain distance)가 상기 도메인 분류기(Domain Classifier)의 오류율에 반비례함에 기반하여 도메인 분포 간의 거리를 측정할 수 있는 것을 특징으로 한다.

[0013] 상기 도메인 분류 모듈은 상기 손실값을 계산하기 위해 손실 함수 L_{det} 를 정의하고, 상기 손실 함수는 RPN의 오차를 계산하는 손실 L_{rpn} 과 분류기의 오차를 계산하는 손실 L_{roi} 로 구성되며, 상기 복수의 손실은 예측한 바운딩 박스의 클래스가 얼마나 정확한지를 기준으로 오차를 측정하는 클래스피케이션 손실과 바운딩 박스의 좌표가 실제 객체의 위치와 얼마나 가까운지를 기준으로 오차를 측정하는 리그레션 손실을 포함한다.

[0014] 상기 도메인 분류 모듈은 바운딩 박스별로 도메인을 얼마나 잘 판별했는지를 기준으로 오차를 계산하되, 상기 오차를 계산하기 위해 하기의 수학적 식 1과 같은 cross-entropy loss 기반의 손실 함수 L_{dom} 을 정의하고, 상기 도메인 분류기 h는 특징맵의 u,v 위치에 있는 바운딩 박스별로 도메인을 판별하며, 도메인 레이블 D_i 는 바운딩 박스가 소스 도메인일 경우에 0으로, 타겟 도메인일 경우에는 1로 정의되는 것을 특징으로 한다.

[0015] [수학적 식 1]

$$L_{dom} = -\frac{1}{N} \sum_{i=1}^N [D_i \log h_i^{(u,v)} + (1-D_i) \log (1-h_i^{(u,v)})]$$

[0017] 상기 도메인 레이블의 선정 기준은 상기 바운딩 박스가 ground truth 레이블의 어느 표 영역과의 Intersection-over-Union(IoU) 비율이 설정된 상위 임계값 이상일 경우 소스 도메인으로 정의하고, 상기 바운딩 박스가 ground truth 레이블의 모든 표 영역과의 IoU 비율이 설정된 하위 임계값 이하일 경우 타겟 도메인으로 정의하는 것을 특징으로 한다.

[0018] 상기 도메인 분류 모듈은 모델 학습시 하기의 수학적 식 2를 이용하여, 손실값을 줄이기 위한 Faster R-CNN의 파라미터 W와 b를 찾는 방식으로 최적화를 수행하는 것을 특징으로 한다.

[0019] [수학적 식 2]

$$\min_{W, b} \left[\frac{1}{n} \sum_{i=1}^n L_{det}^i(W, b) + \lambda R(W, b) \right]$$

[0021] 상기 도메인 분류 모듈은 hyper-parameter λ 가 중첩되어 동작하며, 하기의 수학적 식 3과 같이 도메인 분류기의 파라미터 u와 z를 최적화하여 도메인 판별 손실을 최소화하고, 도메인 간의 거리를 최대화할 수 있도록 두 도메인을 구분할 수 있는 특징을 추출하는 방향으로 Faster R-CNN의 파라미터를 최적화하는 것을 특징으로 한다.

[0022] [수학식 3]

$$R(W, b) = \min_{u, z} \left[\frac{1}{n} \sum_{i=1}^n L_{dom}^i(W, b, u, z) + \frac{1}{n'} \sum_{i=n+1}^{n'} L_{dom}^i(W, b, u, z) \right]$$

[0024] 본 발명의 일 실시예에 따른 도메인 분류 학습 방법은, 검출 단계를 후보 영역 검출과 클래스 분류로 구분하는 검출기로서, 객체 검출을 위한 특징을 추출하는 특징 추출기(Feature Extractor), RPN(Region Proposal Network, RPN) 그리고 관심영역(Region-of-Interest, RoI) 기반의 분류기(Classifier)를 포함하는 도메인 분류 학습 시스템을 이용한 도메인 분류 학습 방법에 있어서, 입력 이미지의 특징을 추출하여 특징맵(feature map) 형태로 출력하는 단계; 상기 특징맵을 통해 객체 후보 영역인 바운딩 박스의 좌표를 추론하는 단계; 특징맵 내 모든 픽셀마다 정의된 앵커(anchor)를 기준으로 생성되는 각 바운딩 박스를 대상으로 클래스를 분류하고 상기 좌표를 보정하는 단계; 및 배경 영역에 대해 잘못 예측한 표 후보 영역에 대한 손실값을 계산 및 분류하는 단계를 포함한다.

[0025] 상기 손실값을 계산 및 분류하는 단계는 손실함수를 이용하여 바운딩 박스별 도메인 판별 및 레이블 정의하여 오차 계산하는 단계를 더 포함한다.

발명의 효과

[0026] 본 발명은 별도의 학습 데이터 추가 없이 표 오검출율을 크게 낮출수 있고, 그 결과 기존 Faster R-CNN과 비교하여 향상된 표 검출 성능을 얻을 수 있는 장점이 있다.

[0027] 또한, 그래프나 그림이 표와 매우 유사하거나 글자가 일정 형태로 구성되어 있는 경우, 특징 추출기가 표 영역의 특징을 다양하게 추출할 수 있도록 반복된 학습에 의해 표 검출 성능을 개선할 수 있는 장점이 있다.

도면의 간단한 설명

[0028] 도 1은 본 발명의 일 실시예에 따른 도메인 분류 학습 시스템이 적용되는 Domain Classification Faster R-CNN 모델 구조를 보인 도면이다.

도 2는 본 발명의 일 실시예에 따른 도메인 분류 학습 방법의 순서도이다.

도 3은 학습 방식에 따라 달라지는 추출 가능한 특징 벡터 비교를 보여주는 도면이다.

도 4는 본 발명의 도메인 분류 학습 방법을 통한 실험 결과 비교 예시를 보여주는 도면이다.

도 5는 도메인 적응 개념을 설명하기 위한 도면이다.

도 6 및 도 7은 H-divergence 개념을 설명하기 위한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0029] 이하에서는 도면을 참조하여 본 발명의 구체적인 실시예를 상세하게 설명한다. 다만, 본 발명의 사상은 제시되는 실시예에 제한되지 아니하고, 본 발명의 사상을 이해하는 당업자는 동일한 사상의 범위 내에서 다른 구성요소를 추가, 변경, 삭제 등을 통하여, 퇴보적인 다른 발명이나 본 발명 사상의 범위 내에 포함되는 다른 실시예를 용이하게 제안할 수 있을 것이나, 이 또한 본원 발명 사상 범위 내에 포함된다고 할 것이다. 또한, 각 실시예의 도면에 나타나는 동일한 사상의 범위 내의 기능이 동일한 구성요소는 동일한 참조부호를 사용하여 설명한다.

[0031] 본 발명의 도메인 분류 학습은 별도의 학습 데이터 추가없이 표 오검출 문제를 해결하는 것이며, 이를 구현하기 위해 다음과 같은 두 가지 사항에 초점을 둔다.

[0032] 첫째, 학습 데이터가 표만으로 구성되어 있기 때문에, 기존에 정의된 Faster R-CNN의 손실 함수를 그대로 사용할 경우 배경 영역 즉, 그래프나 그림에 대한 충분한 오차를 계산하지 못한다. 따라서 배경 영역에 대한 적절한 손실값을 추가하는 알고리즘이 필요하다.

[0033] 두번째로는 표 오검출 문제가 발생하는 이유는 표 영역만을 학습한 모델이 추출한 특징 벡터(feature vector)에 그래프와 그림 영역의 특징 벡터가 일부 포함되기 때문으로 판단되며, 이와 같은 두 영역 간의 공통적인 특징

벡터를 제외하고 표 영역이 가지는 고유한 특징 벡터만을 학습할 수 있는 기법이 필요하다.

- [0034] 본 발명의 도메인 분류 학습 시스템에서는, 두 영역 간의 공통적인 특징 벡터를 분석하기 위하여, 도메인 적응(Domain Adaptation) 기술을 활용하며, 이를 구현하기 위해 표 영역과 배경 영역을 도메인으로 정의하고 두 도메인을 분류하는 도메인 분류 모듈을 구비한다.
- [0035] 즉 도메인 분류 모듈은 표 영역을 학습 목표인 소스 도메인으로 정의하고, 배경 영역을 타겟 도메인으로 정의하며, 정의된 두 도메인을 도메인 분류 모델에 기반하여 학습을 통해 분류하여 상기 분류기에 제공하고, 상기 도메인 분류 모듈은 타겟 도메인인 배경 영역에 대해 잘못 예측한 표 후보 영역에 대한 손실값을 계산하여 분류한다.
- [0036] 이를 위한 도메인 분류 학습 시스템은 검출 단계를 후보 영역 검출과 클래스 분류로 구분하는 검출기로서, 객체 검출을 위한 특징을 추출하는 특징 추출기(Feature Extractor, 10), RPN(Region Proposal Network, 20) 그리고 관심영역(Region-of-Interest, RoI) 기반의 분류기(Classifier, 30)를 포함할 수 있다.
- [0037] 또한 상기 특징 추출기(10)는 입력 이미지의 특징을 추출하여 특징맵(feature map) 형태로 출력하고, RPN(20)에서는 특징맵을 통해 객체 후보 영역인 바운딩 박스의 좌표를 추론하고, 상기 분류기(30)는 특징맵 내 모든 픽셀마다 정의된 앵커(anchor)를 기준으로 생성되는 각 바운딩 박스를 대상으로 클래스를 분류하고 상기 좌표를 보정한다.
- [0039] 도 1은 Faster R-CNN에 도메인 분류 모듈을 적용한 새로운 Faster R-CNN 모델(Domain Classification Faster R-CNN)의 구조를 나타내며, 도메인 분류 모듈을 통해 현재 모델이 잘못 예측한 표 후보 영역에 대한 적절한 손실 값을 계산할 수 있다.
- [0040] 또한 도메인 적응(Domain Adaptation)은 타겟(target) 도메인 즉, 학습 데이터의 레이블이 없거나 불충분한 상태에서 정확한 추론이 가능하도록 모델을 학습하는 기술이며, 주로 레이블이 충분한 소스(source) 도메인과의 관계를 분석하여 두 도메인 간의 공통적인 특징을 찾는 것을 목적으로 수행된다.
- [0041] 도메인 적응은 공통적 특징을 찾기 위해 도 5에서 예시를 든 바와 같이, 소스 도메인인 영화 리뷰와 타겟 도메인인 도서 리뷰의 특징 공간(feature space)에서 공통적인 영역을 찾기 위해, 두 도메인 분포 간의 거리를 측정하는 것이 중요하다.
- [0043] 본 발명의 시스템을 적용하기 위한 딥러닝 알고리즘의 일례로 사용되는 Faster R-CNN은 더욱 정확한 객체 검출을 수행하기 위해 검출 단계를 후보 영역 검출과 클래스 분류로 구분한 2단(two-stage) 검출기로서, 도 1을 참조하면, 크게 특징 추출기(10), RPN(20) 그리고 후보영역 기반의 분류기(30)로 구성된다.
- [0044] 모델의 전체 구조는 도 1의 위쪽 영역에 나타난 바와 같으며, 도 1의 왼쪽을 기준으로 이미지를 입력받아 각 네트워크가 순차적으로 동작한다.
- [0045] 우선 특징 추출기(10)에서는 입력 이미지의 특징을 추출하여 특징맵(feature map) 형태로 출력하고, RPN(20)에서는 특징맵을 통해 객체 후보 영역인 바운딩 박스의 좌표를 추론한다.
- [0046] 해당 바운딩 박스는 특징맵 내 모든 픽셀마다 정의된 앵커(anchor)를 기준으로 생성된다.
- [0047] 마지막으로 분류기(30)에서는 각 바운딩 박스를 대상으로 클래스를 분류하고 좌표를 보정한다.
- [0048] 모델 학습을 위해 정의된 손실 함수 L_{det} 는 다음 수학식 1과 같이, RPN(20)의 오차를 계산하는 L_{rpn} 과 분류기(30)의 오차를 계산하는 L_{roi} 로 구성되어 있다.

수학식 1

$$L_{det} = L_{rpn} + L_{roi}$$

여기서, L_{rpn} 과 L_{roi} 모두 클래스피케이션 손실(classification loss) L_{cls} 과 리그레션 손실(regression loss) L_{reg} 을 포함하며, L_{cls} 는 예측한 바운딩 박스의 클래스가 얼마나 정확한지를 기준으로 오차를 측정

하며, L_{reg} 는 바운딩 박스의 좌표가 실제 객체의 위치와 얼마나 가까운지를 기준으로 오차를 측정한다.

[0051] 보다 더 자세한 설명을 위해 하기의 수학적 식 2는 L_{reg} 의 손실 함수를 나타낸 것으로 i 는 앵커의 인덱스를 나타내며, p_i 는 앵커 i 의 객체유무를 예측한 값을 나타낸다.

[0052] ground truth 레이블인 p_i^* 는 앵커 i 에 객체가 있을 경우 즉, positive 앵커로서 1로 설정되며, 그렇지 않을 경우 즉, negative 앵커로서 0으로 설정된다. t_i^* 는 앵커 i 를 대상으로 예측한 바운딩 박스의 좌표를 나타내며, ground truth 레이블인 t_i^* 는 positive 앵커의 바운딩 박스의 좌표로 설정된다. 하기 수학적 식 2를 통해 L_{reg} 는 p_i^* 가 1일 경우 즉, 앵커 i 에 객체가 있을 경우에만 활성화되는 것을 확인할 수 있으며, p_i^* 가 0일 경우 즉, 앵커 i 에 객체가 없는 경우에는 오직 L_{cls} 만을 고려한 학습이 수행된다.

수학적 식 2

[0053]
$$L_{reg}(p_i, t_i) = L_{cls}(p_i, p_i^*) + p_i^* L_{reg}(t_i, t_i^*)$$

[0054] 도메인 적응 문제를 해결하는 기존 방법의 대다수는 소스 도메인과 타겟 도메인이 서로 관련되어 있다는 가정 하에 소스 도메인의 예러와 두 도메인 분포 간의 거리의 합으로 타겟 도메인의 예러 범위를 예측한다.

[0055] 따라서 도메인 분포 간의 거리를 정확하게 측정하는 것이 중요하며, 이를 위해 본 발명에서는 H-divergence를 사용한다.

[0056] H-divergence는 영역 거리(domain distance) $d_H(S, T)$ 가 도메인 분류기(Domain Classifier, 110) h의 오류율에 반비례함에 기반하며, 이는 최적의 h의 오류가 높아서 두 도메인을 구별하기 어려울 경우 도메인 분포 간의 거리가 충분히 가깝다는 것을 의미한다.

[0057] 예컨대 도 6 및 도 7을 참조하면, 소스 도메인인 영화 리뷰와 타겟 도메인인 도서 리뷰에 대해서 도메인 분류기(110)가 잘 분류해내려면, 도메인 분포 간 거리가 멀어야 하며, 도메인 분포 간 거리가 가까울 수록 공통 영역이 많기 때문에 분류가 어렵게 되는 것이다.

[0059] 하기의 수학적 식 3은 H-divergence를 수식으로 나타낸 것으로, $\mathbf{x}$ 는 특징 벡터이고, err_S 와 err_T 는 각각 소스 도메인과 타겟 도메인에서의 h의 예측 오류를 의미한다.

수학적 식 3

[0060]
$$d_H(S, T) = 2(1 - \min_{h \in H} (err_S(h(x)) + err_T(h(x))))$$

[0061] H-divergence를 활용하여 두 도메인 간의 거리를 최소화할 수 있는 특징 벡터를 찾기 위해 도메인 분류기(110) h의 오류율을 최대화하는 특징 추출기(10) f를 학습하는 방식이 존재하며, 본 발명에서는 하기의 수학적 식 4에 나타낸 바와 같이, 두 개의 네트워크가 서로 적대적으로 학습함으로써 두 도메인의 공통적인 특징을 추출할 수 있는 도메인 적응 모델을 구현하였다.

수학식 4

$$\min_f d_H(S, T) \Leftrightarrow \max_f \min_{h \in H} \{err_S(h(x)) + err_T(h(x))\}$$

[0062]

[0063]

본 발명에서는 표 검출을 학습하기 위해 일 실시예로 TableBank 데이터셋을 사용한다. 하지만 TableBank는 문서 이미지 내 표의 위치 정보만으로 구성되어 있기 때문에, Faster R-CNN의 학습 방식을 그대로 적용할 경우 표를 제외한 모든 영역이 negative 앵커로 설정되어 이미지 내 대부분의 영역이 리그레션 손실(regression loss)이 비활성화된 채 학습을 수행하게 된다.

[0064]

이러한 학습 과정은 결과적으로 표 오검출 문제를 발생시키며, 이와 같은 문제를 해결하기 위하여 배경 영역에 대한 적절한 손실값을 추가하는 알고리즘이 필요하다.

[0065]

본 발명에서는 도 1와 같이 표 영역과 배경 영역을 구분하는 도메인 분류 모듈을 기존 Faster R-CNN에 추가한 Domain Classification Faster R-CNN 모델을 제안하며, 표 영역이 가지는 고유한 특징을 추출할 수 있도록 학습 방식을 설계하였다.

[0066]

도메인 분류 모듈(Domain Classification Module)은 도 1의 아래쪽 영역과 같이 구성되며, 모듈 내 도메인 분류기(110)는 RPN(20)의 바운딩 박스를 입력받아 각 박스별 도메인을 판별한다.

[0067]

본 발명에서는 ground truth 레이블을 기준으로 표 영역을 소스 도메인 S로, 배경 영역을 타겟 도메인 T로 정의한다. 두 도메인 간의 H-divergence $d_H(S, T)$ 는 수학식 5와 같이 계산할 수 있으며, 도메인 분류기(110) h는 인덱스 i의 바운딩 박스 X_i 를 입력받고, 이를 소스 도메인으로 예측할 경우 0을, 타겟 도메인으로 예측할 경우 1을 출력한다.

[0068]

$I[a]$ 는 a가 참이면 1을 출력하고, 그렇지 않으면 0을 출력하는 지시 함수(indicator function)로써 도메인 분류기(110) h가 각 도메인에 대해 잘못된 예측을 수행할 경우 1을 출력한다.

수학식 5

$$d_H(S, T) = 2 \left(1 - \min_{h \in H} \left[\frac{1}{n} \sum_{i=1}^n I[h(X_i) = 1] + \frac{1}{n'} \sum_{i=n+1}^N I[h(X_i) = 0] \right] \right)$$

[0069]

[0070]

도메인 분류 모듈의 오차는 바운딩 박스별로 도메인을 얼마나 잘 판별했는지를 기준으로 정하며, 하기의 수학식 6과 같이 cross-entropy loss 기반의 손실 함수 L_{dom} 을 정의할 수 있다.

[0071]

도메인 분류기(110) h는 특징맵의 u, v 위치에 있는 바운딩 박스별로 도메인을 판별하며, 도메인 레이블 D_i 는 바운딩 박스가 소스 도메인일 경우에 0으로, 타겟 도메인일 경우에는 1로 정의된다.

[0072]

도메인 레이블의 선정 기준은 다음과 같다. i) 바운딩 박스가 ground truth 레이블의 어느 표 영역과의 Intersection-over-Union(IoU) 비율이 상위 임계값인 0.7 이상일 경우 소스 도메인으로 정의한다. ii) 바운딩 박스가 ground truth 레이블의 모든 표 영역과의 IoU 비율이 하위 임계값인 0.3 이하일 경우 타겟 도메인으로 정의한다.

수학식 6

$$L_{dom} = -\frac{1}{N} \sum_{i=1}^N [D_i \log h_i^{(u,v)} + (1-D_i) \log(1-h_i^{(u,v)})]$$

[0073]

[0074] Domain Classification Faster R-CNN 모델의 학습은 최적화 문제와 같이 손실 값을 줄이기 위한 Faster R-CNN의 파라미터 W와 b를 찾는 방식으로 수행되며, 하기의 수학식 7과 같이 표현할 수 있다.

[0075] 도메인 분류 모듈(100)은 hyper-parameter λ 로 가중되어 동작하며, 하기의 수학식 8과 같이 도메인 분류기(110)의 파라미터 u와 z를 최적화하여 도메인 판별 손실을 최소화하는 동시에 도메인 간의 거리를 최대화할 수 있도록 즉, 두 도메인을 구분할 수 있는 특징을 추출하는 방향으로 Faster R-CNN의 파라미터를 최적화한다.

수학식 7

$$\min_{W, b} \left[\frac{1}{n} \sum_{i=1}^n L_{det}^i(W, b) + \lambda R(W, b) \right]$$

[0076]

수학식 8

$$R(W, b) = \min_{u, z} \left[\frac{1}{n} \sum_{i=1}^n L_{dom}^i(W, b, u, z) + \frac{1}{n'} \sum_{i=n+1}^N L_{dom}^i(W, b, u, z) \right]$$

[0077]

[0079] 이와 같은 학습 방식은 기존의 두 도메인 간의 거리를 최소화할 수 있도록 도메인 분류기(110)의 오류율을 최대화하는 학습 방식과 상반되는 형태이며, 도 3에 나타난 것처럼 기존에 제안된 도메인 적응 기반의 학습 방식(도 3의 (a) 학습 방식)은 두 도메인 간의 공통된 특징을 찾기 위해 모델을 최적화하였지만, 본 발명에서 제안하는 도메인 분류 기반의 학습 방식(도 3의 (b) 학습 방식)은 두 도메인 간의 공통된 특징을 배제하도록 모델을 최적화한다.

[0080] 이러한 학습 방식을 통해 얻을 수 있는 이점은 학습 데이터가 불균형한 상태에서도 클래스가 가지는 고유한 특징을 추출할 수 있다는 것이며, 기존에 정의된 손실 함수가 각 클래스에 해당하는 오차만을 계산하는 방식으로, 클래스의 개수가 적을 경우, 특히, TableBank와 같이 단일 클래스(표 영역)로 구성되어 있을 경우 다른 클래스(배경 영역)에 대한 충분한 오차를 계산하는 것이 어렵기 때문에, 본 발명에서 제안하는 학습 방식을 통해 배경 영역에 해당하는 적절한 손실 값을 추가하여 표 영역이 가지는 고유한 특징을 추출하는 것이 가능하다.

[0082] 본 발명은 Domain Classification Faster R-CNN 모델을 기반으로 하여 도메인 적응 모듈을 학습하기 위해 필요한 바운딩 박스별 도메인 레이블은 학습 중 자동으로 선정되기 때문에 기존 Faster R-CNN 모델과 마찬가지로 End-to-End 학습이 가능하며, 도메인 적응 모듈은 모델 학습 시에만 활성화되기 때문에 추론 시에는 해당 모듈을 제거하는 것이 가능하다.

[0083] 따라서 실제 추론에 필요한 연산과 시간은 기존 Faster R-CNN 모델과 동일하다.

[0084] 도 4에는 본 발명의 도메인 분류 학습 시스템을 통하여 실험한 결과를 기존 학습 방법과 비교하여 보여주고 있다.

[0085] 본 발명의 도메인 분류 학습 시스템은 도 4 하단의 proposed 부분에 해당하며, 기존 학습 방법은 도 4 상단의 faster R-CNN으로 표시된 부분이며, 결과적으로 기존 학습방법은 도표나 그림에 대해서도 표로 오인식하나, 본 발명의 경우가 표영역에 대한 분류된 학습을 통해 표 만을 더욱 잘 인식하는 것을 볼 수 있다.

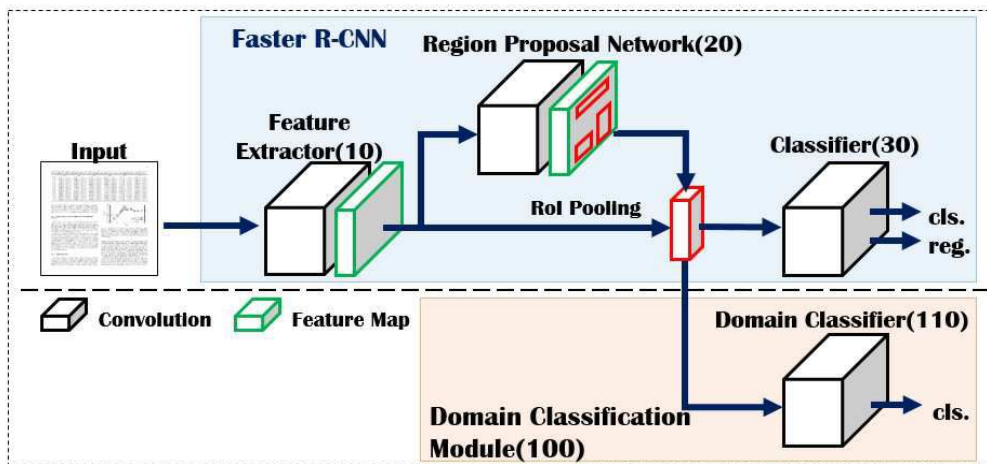
부호의 설명

[0086]

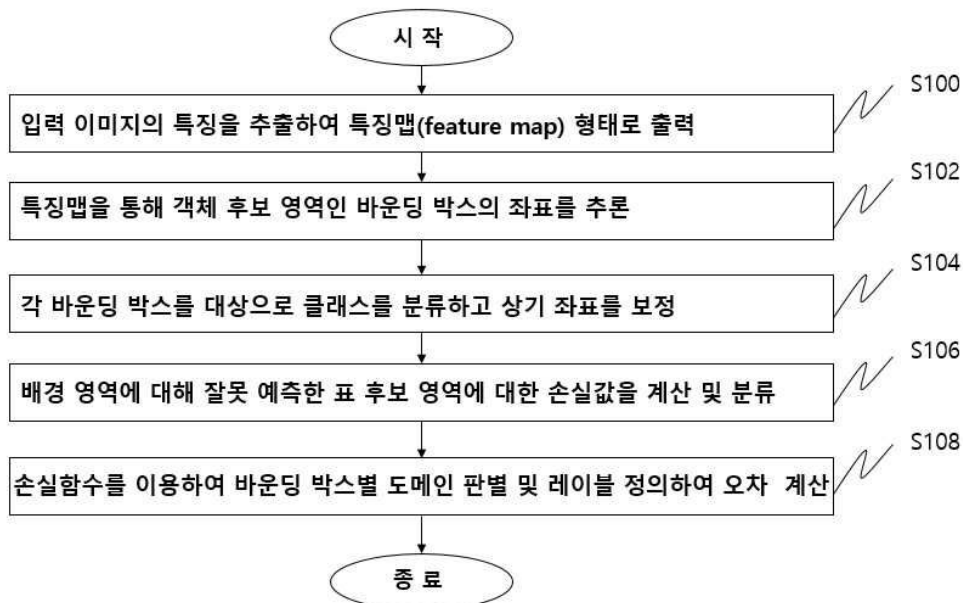
- 10 ; 특징 추출기
- 20 ; RPN
- 30 ; 분류기
- 100 ; 도메인 분류 모듈
- 110 ; 도메인 분류기

도면

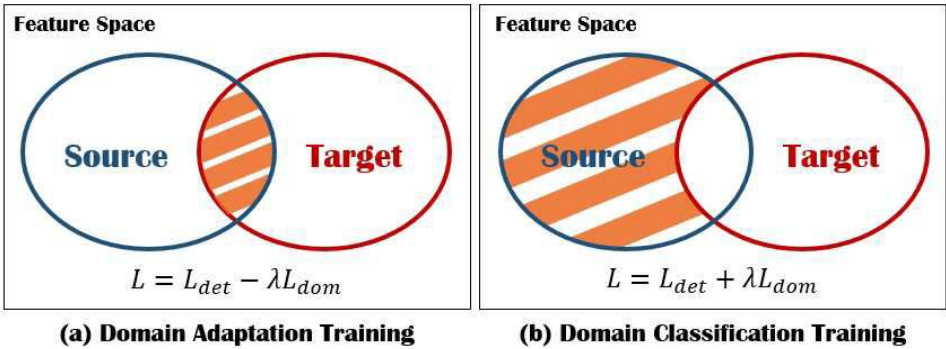
도면1



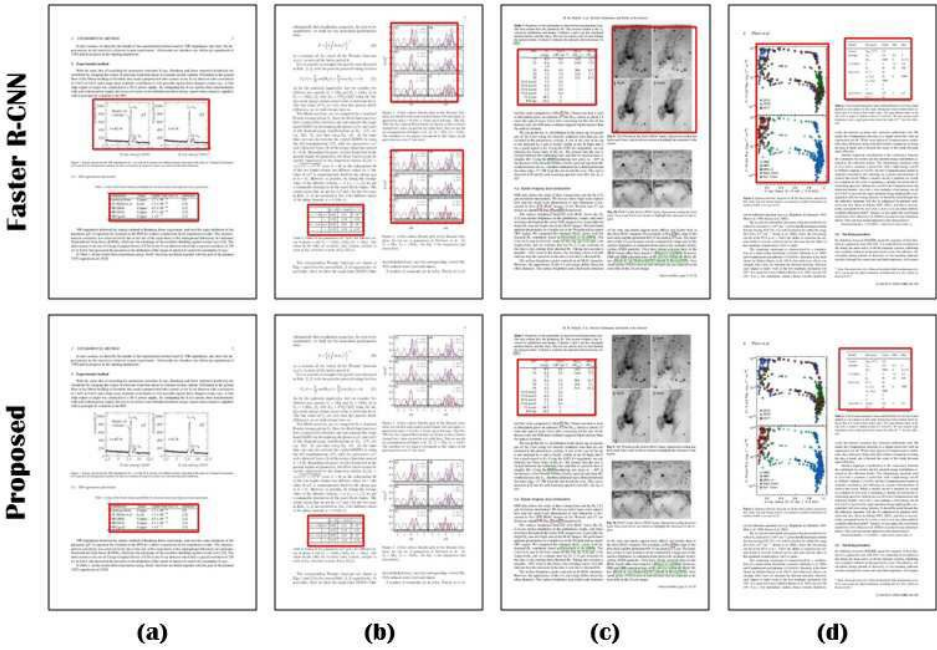
도면2



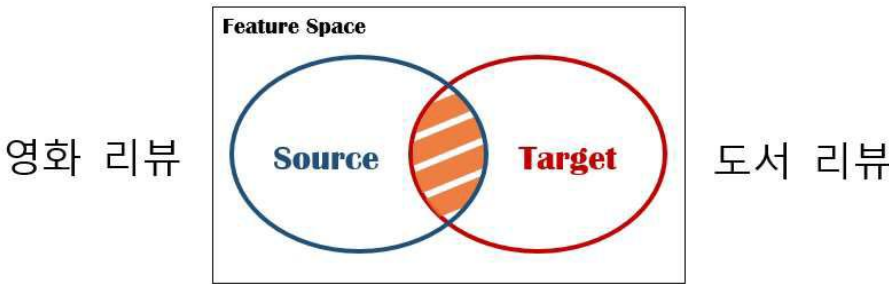
도면3



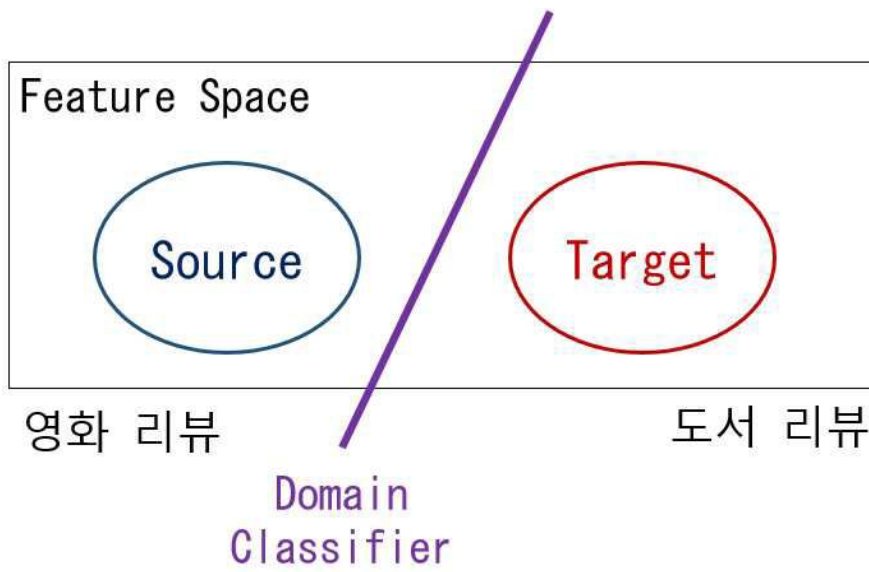
도면4



도면5



도면6



도면7

