DenseBox: Unifying Landmark Localization with End to End Object Detection

0. Abstract

물체 탐지에 대해 어떻게 완전한 하나의 fully CNN(FCN)이 수행될 수 있는가? 우리는 이미지의 모든 위치와 스케일을 통해 bounding 박스와 객체 클래스 confidence를 직접 예측하는 통합 엔드투엔드 FCN 프레임워크인 DenseBox를 소개한다. 우리의 Contribution은 두 가지다. 첫째, 하나의 FCN이 신중하게 설계되고 최적화되면 여러 개의 서로 다른 개체를 매우 정확하고 효율적으로 감지할 수 있음을 보여 준다. 둘째, 우리는 멀티태스킹 학습 중에 랜드마크의 지역화와 통합할 때, DenseBox가 객체 감지 기능을 더욱 향상시킨다는 것을 보여준다. 우리는 MALF 얼굴검출과 KITTI 차량검출을 포함한 공공 벤치마크 데이터셋에 대한 실험 결과를 제시하는데, 이는 DenseBox가 얼굴과 자동차와 같은 도전적인 물체를 감지하는 최첨단 시스템임을 나타낸다.

1. Introduction

우리의 일상 생활은 사물 탐지 사례로 가득 차 있다. 운전 중 주변 차량 확인, 사람 찾기, 낯익은 얼굴 국소화 등이 모두 물체 탐지의 예다. 물체 탐지는 컴퓨터 시력의 핵심 문제들 중 하나이다. CNN(Convolutional nerve networks) [18]의 성공 이전에, 물체 검출은 대개 가능한 모든 위치와 영상 척도에서 추출한 수공예 형상[5, 25, 4]에 분류기를 적용하는 슬라이딩 윈도우 기반 방법[11, 39]에 의해 해결된다. 최근에는 완전 경사진 신경망(FCN) [24] 기반 방법 [34, 8, 29]이 개체 탐지 분야에 혁명을 가져온다. 이러한 FCN 프레임워크도 슬라이딩 윈도우 패션을 따르지만 모델 파라미터와 이미지 기능을 처음부터 학습하는 엔드 투 엔드 접근 방식은 감지 성능을 크게 향상시킨다.

R-CNN [15, 14]은 FCN 기반 방법을 넘어 객체 감지에 대한 발생률을 더욱 향상시킨다. 개념적으로, R-CNN은 두 단계를 포함한다. 첫 번째 단계는 영상에서 모든 잠재적 bounding 상자 후보를 생성하기 위해 지역 제안 방법을 사용한다. 그런 다음 두 번째 단계는 CNN 분류기를 적용하여 모든 제안에 대해 서로 다른 대상을 구분한다. R-CNN이 일반 물체 탐지를 위한 새로운 최첨단 시스템이 되지만[9, 33] 각 후보 박스의 해상도가 낮고 문맥이 부족하면 그것들에 대한 분류 정확도가 현저히 떨어지기 때문에 사람 얼굴이나 먼 차와 같은 작은 물체[27]를 탐지하는 것은 매우 어렵다. 더욱이, R-CNN 파이프라인의 두 가지 서로 다른 단계는 공동으로 최적화할 수 없으므로, R-CNN에 엔드투엔드 교육을 적용하는 데 어려움을 겪게 된다.

이 작품에서 우리는 다음과 같은 한 가지 질문에 초점을 맞춘다. 개체 감지 시 1단계 FCN이 수행할 수 있는 성능은? 이를 위해 제안서 생성이 필요 없고 훈련 중 엔드투엔드로 최적화할 수 있

는 새로운 FCN 기반 객체 검출기 DenseBox를 제시한다. 기존의 많은 슬라이딩 윈도우 패션 FCN 검출 프레임워크와 유사하지만 [34, 8, 29], DenseBox는 작은 스케일과 무거운 폐색 하에서 물체를 감지하도록 보다 세심하게 설계되었다. 우리는 DenseBox를 훈련시키고 감지 성능을 향상시키기 위해 조심스러운 음극 채굴 기술을 적용한다. 이를 더욱 향상시키기 위해, 우리는 공동 다과제 학습[1]을 통해 획기적인 현지화를 시스템에 더욱 통합한다. 랜드마크 현지화의 유용성을 검증하기 위해, KITTI 차량 감지 데이터셋[13]에 대한 일련의 키포인트에 수동으로 주석을 달며, 이후 주석을 공개한다.

우리의 공헌은 두 배다. 첫째, 우리는 한 개의 완전히 복잡한 신경망이 신중하게 설계되고 최적화된다면, 매우 정확하고 효율적으로 무거운 폐색을 가진 다른 척도 아래의 물체를 탐지할 수 있다는 것을 증명한다. 둘째, 우리는 멀티태스킹 학습을 통해 랜드마크 현지화에 통합할 때 DenseBox가 개체 감지 정확도를 더욱 향상시킨다는 것을 보여준다. 우리는 MALF(Multi-Atribe Labeled Faces) 얼굴 검지[42] 및 KITTI 차량 검지[13]를 포함한 공공 벤치마크 데이터 세트에 대한 실험 결과를 제시하며, 이는 DenseBox가 얼굴 검지와 자동차 검출을 위한 최첨단 시스템임을 나타낸다.

2. Related Work

물체 탐지에 관한 문헌은 광범하다.

얼굴검출과 같은 검출 과제에 신경망을 응용하는 것 역시 오랜 역사를 가지고 있다.

최근에, 몇몇 논문들은 물체를 찾기 위해 깊은 경사진 신경망을 사용하는 알고리즘을 제안한다[34, 8, 29].

그러나 대부분의 첨단 물체 감지 접근법[26, 20, 8, 14, 41]은 탐지를 두드러진 개체 제안 생성과 지역 제안 분류의 두 단계로 나누는 R-CNN에 의존한다.

개체 탐지는 종종 랜드마크 현지화, 포즈 추정 및 의미 분할과 같은 다중 작업 학습과 관련되어 있다.

3. DenseBox for Detection

전체 감지 시스템은 그림 1에 설명되어 있다.

3.1. GT Generation

배경화면을 연결하는데 대부분의 계산 시간이 걸릴 것이기 때문에 훈련을 위해 전체 이미지를 네

트워크에 넣을 필요는 없다.

일반적으로 말해서, 우리의 제안된 네트워크는 세분화 같은 방식으로 훈련된다.

한 패치에 여러 개의 면이 발생하는 경우, 해당 얼굴이 패치의 중심에서 면에 비례하여 스케일 범위(예: 설정에서 0.8 ~ 1.25)에 있으면 양으로 유지된다.

3.2. Model Design

그림 3에 표시된 우리의 네트워크 아키텍처는 이미지 분류에 사용되는 VGG 19 모델[35]에서 파생되었다.

3.2.1. Multi-Level Feature Fusion.

최근의 연구는 다른 연결 계층의 형상을 사용하는 것이 가장자리 감지 및 분할과 같은 작업에서 성능을 향상시킬 수 있다는 것을 보여준다.

3.3. Multi-Task Training

우리는 DenseBox를 초기화하기 위해 ImageNet 사전 훈련된 VGG 19 네트워크를 사용한다.

Fast R-CNN처럼 우리 네트워크는 두 개의 형제 생산 지점을 가지고 있다.

여기서 우리는 L2손실을 얼굴 및 자동차 감지 작업에서 모두 사용한다.

출력 bounding 상자 회귀 손실의 두 번째 분기(Lloc로 표시됨).

3.3.1. Balance Sampling

부정적인 샘플을 선택하는 과정은 학습에 있어 중요한 부분 중 하나이다.

3.3.1.1. Ignoring Gray Zone

회색 영역은 양과 음의 영역에 정의된다.

3.3.1.2. Hard Negative Mining

SVM에서 하드 음의 채굴 절차와 유사하게, 우리는 무작위 표본보다 나쁘게 예측된 표본을 검색함으로써 학습을 더 효율적으로 만든다.

3.3.1.3. Loss with Mask

이제 각 샘플에 대한 마스크를 플래그의 함수로 정의할 수 있다.

3.3.1.4. Other Implementation Details

훈련에서 입력 패치는 특정 스케일로 중심에 있는 물체를 포함하는 경우 "긍정 패치"로 간주된다. 우리는 훈련에서 미니 배치 SGD를 사용하며 배치 크기는 10으로 설정되어 있다.

3.4. Refine With Landmark Localization

이 파트에서, 우리는 완전히 복잡한 구조 덕분에 몇 개의 층을 쌓는 것 만으로도 DenseBox에서 획기적인 지역화가 이루어질 수 있다는 것을 보여준다.

최종 산출물은 분기를 정제하며, 분류 점수 지도와 랜드마크 지역 지도를 입력으로 하여 검출 결과의 미세화를 목표로 한다.

3.5. Comparison

DenseBox의 하이라이트는 회귀 문제로서 객체 감지를 프레임화하고 엔드투엔드 검출 프레임워크를 제공한다는 것이다.

4. Experiments

이 절에서는 MALF[42] 및 KITTI[13] 차량 감지 과제에 대한 DenseBox의 성능을 시연한다.

4.1. MALF Detection Task

MALF 탐지 테스트 데이터 세트에는 인터넷에서 수집된 5,000개의 이미지가 포함되어 있다.

4.1.1. Training and Testing

우리는 그림 5에 표시된 72개의 랜드마크로 주석을 단 81,024개의 얼굴을 가진 31,337개의 인터넷 수집 이미지 섹션 3에 설명된 두 가지 모델을 훈련한다.

4.2. KITTI Car Detection Task

KITTI 객체 감지 벤치마크는 7481개의 교육 영상과 7518개의 테스트 영상으로 구성된다.

4.2.1. Training and Testing

얼굴 감지 작업뿐만 아니라, 우리는 KITTI 물체 감지 훈련 세트에서 두 가지 모델(하나는 랜드마크가 없고 다른 하나는 랜드마크가 있는 모델)을 훈련한다.

4.2.2. Results.

표 1은 DenseBox와 다른 방법의 결과를 보여준다.

5. Conclusion

우리는 탐지를 위한 통합된 엔드 투 엔드 감지 파이프라인인 DenseBox를 제공했다. 이 공연은 획기적인 정보를 통합함으로써 쉽게 향상될 수 있다. 우리는 또한 DenseBox의 차이와 기여를 강조하면서 우리의 방법과 다른 관련 물체 감지 시스템을 분석한다. DenseBox는 얼굴 감지 및 차량 감지 작업 모두에서 인상적인 성능을 달성하여 제안서 생성에 실패할 수 있는 상황에 적합한 높은 성능을 입증한다. DenseBox의 핵심 문제는 속도다. 이 논문에서 제시된 원래의 DenseBox는 하나의 이미지를 처리하는데 몇 초가 필요하다. 그러나 이것은 우리의 후기 버전에서 다루어졌다.

우리는 KITTI와 얼굴 검지에 대한 실시간 탐지 시스템을 설명하는 또 다른 논문인 DenseBox2를 발표할 것이다.

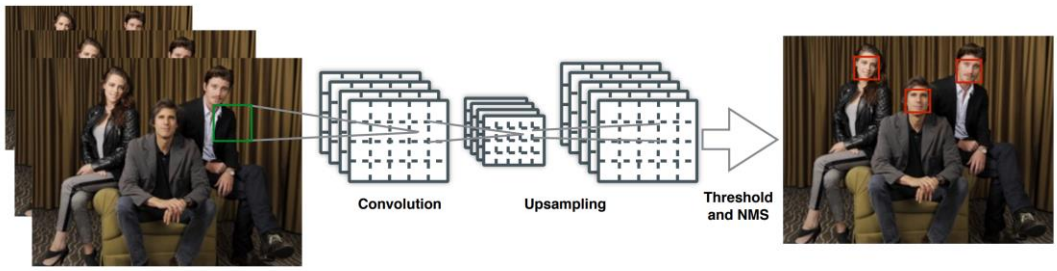


Figure 1: **The DenseBox Detection Pipeline.** 1) Image pyramid is fed to the network. 2) After several layers of convolution and pooling, upsampling feature map back and apply convolution layers to get final output. 3) Convert output feature map to bounding boxes , and apply non-maximum suppression to all bounding boxes over the threshold.

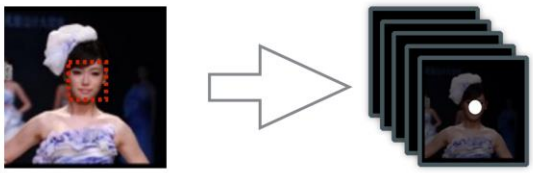


Figure 2: **The Ground Truth Map in Training .** The left image is the input patch, and the right one is its ground truth map.

act as fully connected layers in a sliding-window fashion.

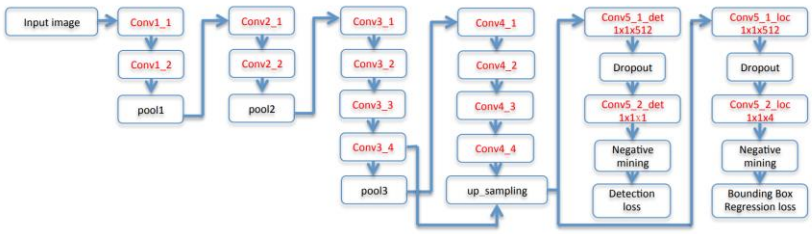


Figure 3: **Network architecture of DenseBox.** The rectangles with red names contain learnable parameters.

Multi-Level Feature Fusion. Recent works[2, 22] indicate that using features from different convolution layers can enhance performance in task such as edge detection and segmentation. Part-level feature focus on local details of object to find discriminative appearance parts, while object-level or

3.4 Refine with Landmark Localization.

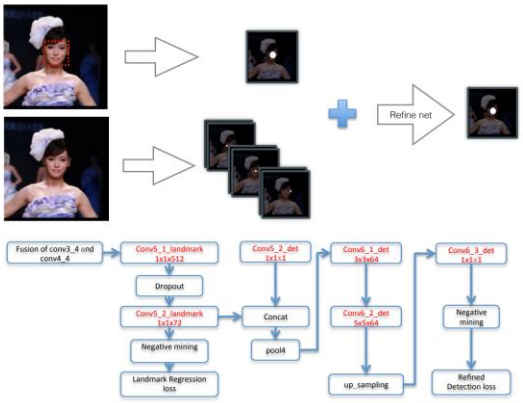


Figure 4: **Top:** The pipeline of DenseBox with landmark localization. **Bottom:** The network structure for landmark localization.

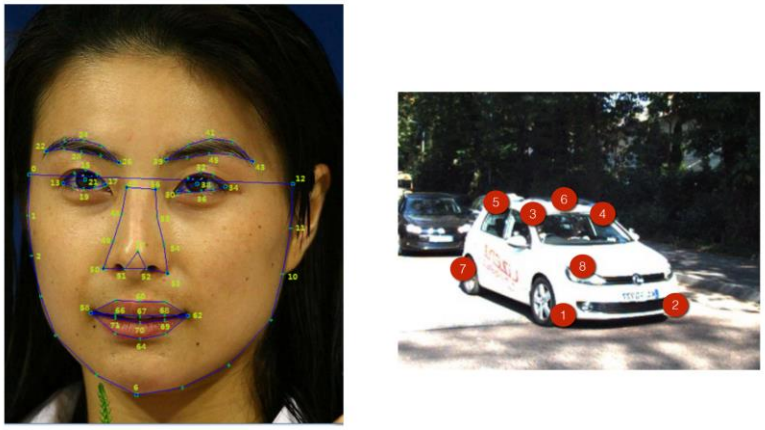
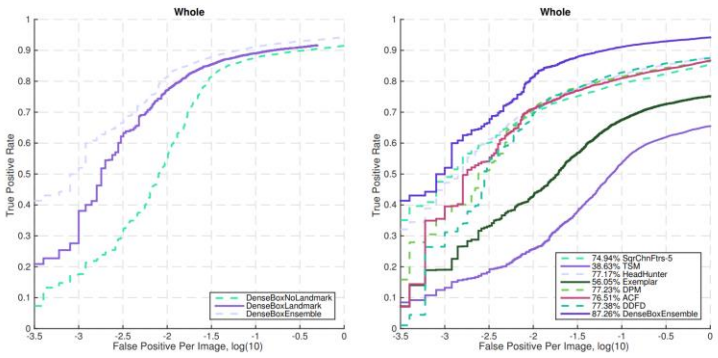


Figure 5: **Left:** 72 landmarks for face. **Right:** 8 landmarks for car.



(a) (b)

Figure 6: **Result on MALF dataset.** (a) Comparison of different versions of DenseBox; (b)The curves and mean recall rate of DenseBox and other methods;

Method	*Moderate*	*Easy*	*Hard*
Regionlets [23]	76.45%	84.75%	59.70%
AOG [19]	74.26%	84.24%	60.51%
3DVP [40]	75.77%	87.46%	65.38%
spCov_LBP	77.40%	87.19%	60.60%
DeepInsight	84.40%	84.59%	76.09%
NIPS ID 331	87.14%	88.33%	76.11%
DJML	88.79%	91.31%	77.73%
DenseBox (without landmark)	85.07%	82.33%	76.27%
DenseBox (withlandmark)	85.74%	83.63%	76.71%

Table 1: The Averate Presision on KITTI Car Detection Task

Figure 7: **Examples on both the MALF detection set KITTI car detection set.** The numbers above the bounding boxes are the confidence score. Our system works very well in complex scene where objects are small and highly occluded. However, it still could miss some objects and generate false alarm.