

Microsoft COCO: Common Objects in Context

Abstract

- 328k장의 image에 250만 개 labeled instance
- 해당 논문은 PASCAL, ImageNet, SUN과 비교해 dataset에 대한 자세한 통계 분석 제시
- Deformable Parts Model 이용한 bounding box와 segmentation detection 결과에 대한 baseline의 성능 분석 제공

Introduction

- computer vision의 주요 목표 중 하나는 visual scenes 이해하는 것
- 수 많은 image 포함하는 ImageNet dataset은 최근 object detection 연구에서 큰 발전 가능하게 했으며 object attributes, scene attributes, keypoints, 3D scene information 포함하는 dataset 만들
- 어떤 dataset이 scene understanding이라는 목표 향해 진보 하는가?
scene understanding에서의 아래 3가지 핵심 연구 문제 해결 위한 새로운 dataset 소개

challenge

1. object가 image 중앙 배경에 방해 받지 않아야 함
2. 여러 object 포함하는 natural image 찾는 것
3. dataset의 모든 object 범주의 모든 instance에 label 지정되고, 완전히 segmentation 되어야 함



Fig. 2. Example of (a) iconic object images, (b) iconic scene images, and (c) non-iconic images. In this work we focus on challenging non-iconic images.

- 3가지 목표 만족하는 대규모 dataset 만들기 위해 Amazon Mechanical Turk 광범위하게 사용해 data 수집
- contextual relationships와 non-iconic object views 포함하는 많은 image set 수집, hierarchical labeling approach 사용해 특정 object의 category 포함하는 것으로 labeling 진행
- COCO dataset에는 91개 공통 object category 포함됨
 - 그 중 82개는 5,000개 이상의 label 지정된 instance 있음

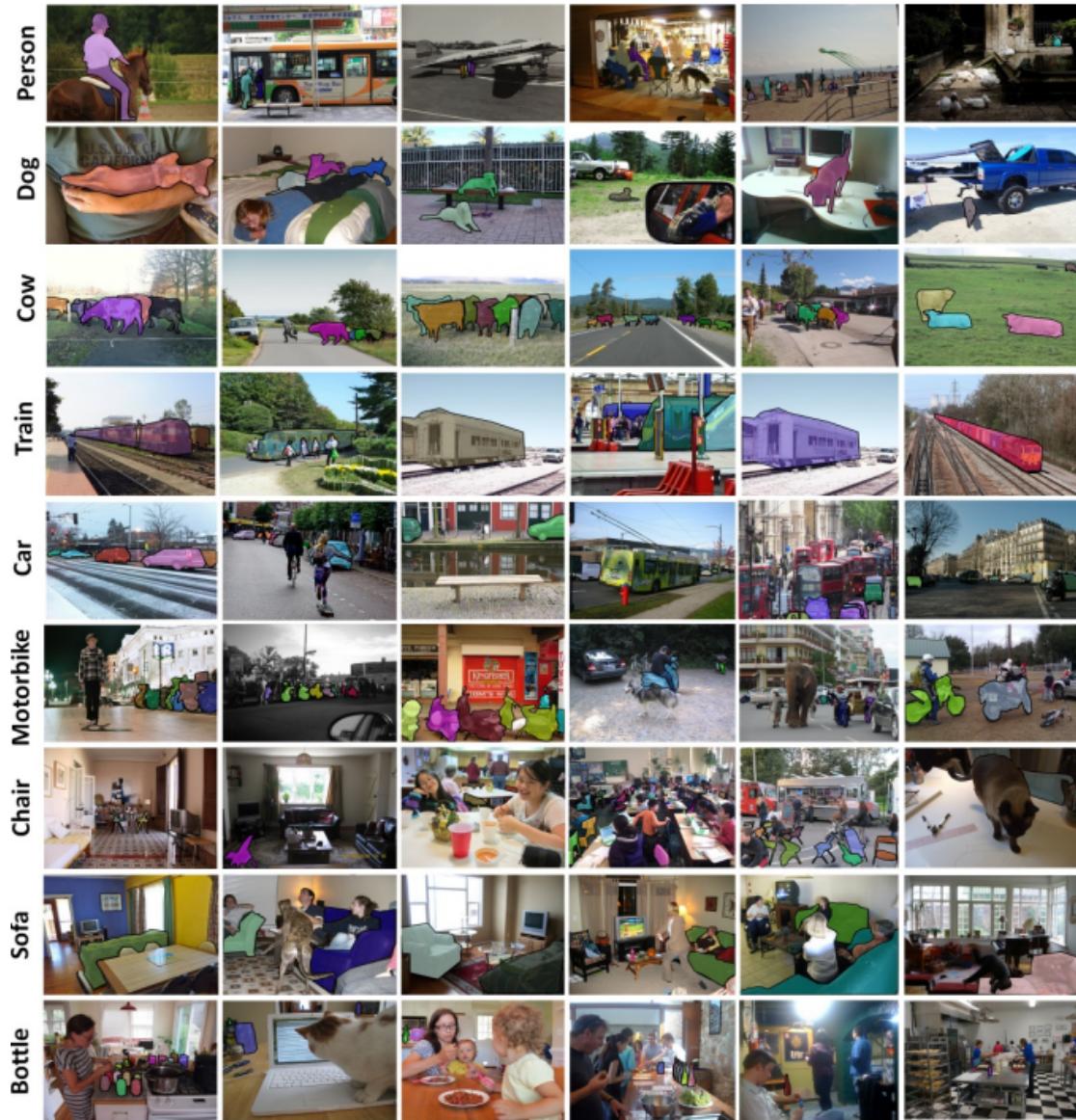


Fig. 6. Samples of annotated images in the MS COCO dataset

Related Work

Object recognition과 관련된 dataset

: **object classification, object detection, semantic scene labeling**

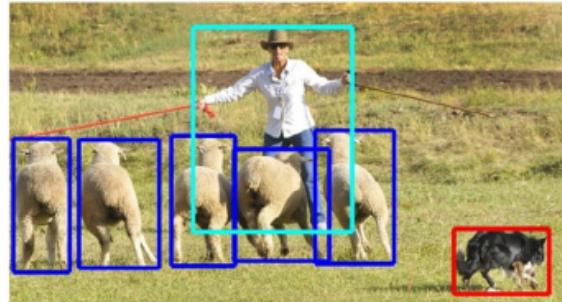
- **Image Classification**
 - image에 object가 있는지 여부부터 나타내는 binary label 필요



(a) Image classification

- **Object Detection**

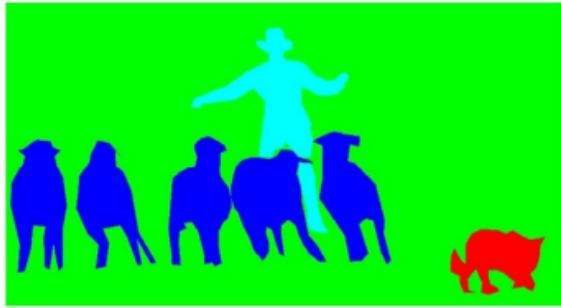
- 지정된 class에 속한 object가 있음을 나타내는 작업과 image에서 object localizing하는 작업 모두 포함
- object 위치는 bounding box로 표시됨
- sunglasses, cell-phone 또는 의자와 같은 많은 object detection 하는 것은 상황 별 정보에 크게 의존하기 때문에 detection dataset이 natural 환경에서 object 포함하는 것이 중요
- bounding box 사용하면 detection algorithm 평가할 수 있는 정확도 제한되기 때문에 연구진들은 정확한 평가 위해 완전 segmentation instance 사용을 제안함



(b) Object localization

- **Semantic Scene Labeling**

- Semantic Scene Labeling은 image에서 의미 있는 object에 label 붙이는 작업으로, image의 각 pixel이 하늘, 의자, 바닥, 거리 같은 범주에 속하는 것으로 label 붙여야 함
- detection 작업과는 달리 개별 instance segmentation 하지 않아도 됨
→ 잔디, 거리, 벽과 같이 개별 instance 구별하는 것 목표로 해 각 object의 범위 확실히 이해해야 함



(c) Semantic segmentation

Image collection

Object categories 및 Candidate images 선택되는 방법 설명

- **Common object categories**

- categories는 모든 categories의 대표 세트 형성하고 실제 적용에 관련되어야 하며, 대규모 dataset 수집할 수 있을 만큼 충분히 빈번하게 발생해야 함
- 중요한 결정은 'thing'과 'stuff' 모두 포함할지 여부와, 세분화된 categories와 object 부분 categories 포함 할지 여부
- 'thing' categories에는 사람, 의자, 자동차와 같은 쉽게 label 붙일 수 있는 object가 포함되는 반면, 'stuff' categories에는 하늘, 거리, 잔디와 같은 명확한 경계가 없는 object 포함됨
- 해당 논문에서는 object instance의 정확한 object instance의 localization에 관심 있기 때문에 'stuff'가 아닌 'thing' 범주만 포함하기로 함
- 연구진들은 'thing'의 보급형 object categories 수집을 위해 여러 소스 사용함
PASCAL VOC의 categories와 시각적으로 식별 가능한 대상을 나타내는 자주 사용되는 1200개 단어 중 일부를 결합해 categories list 작성함

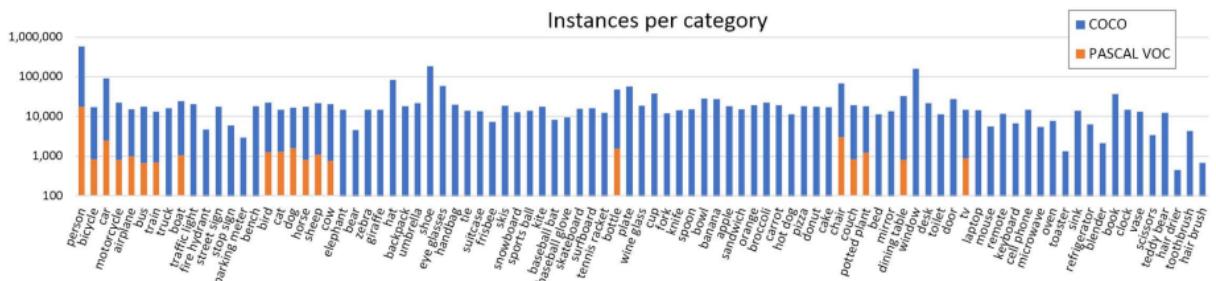


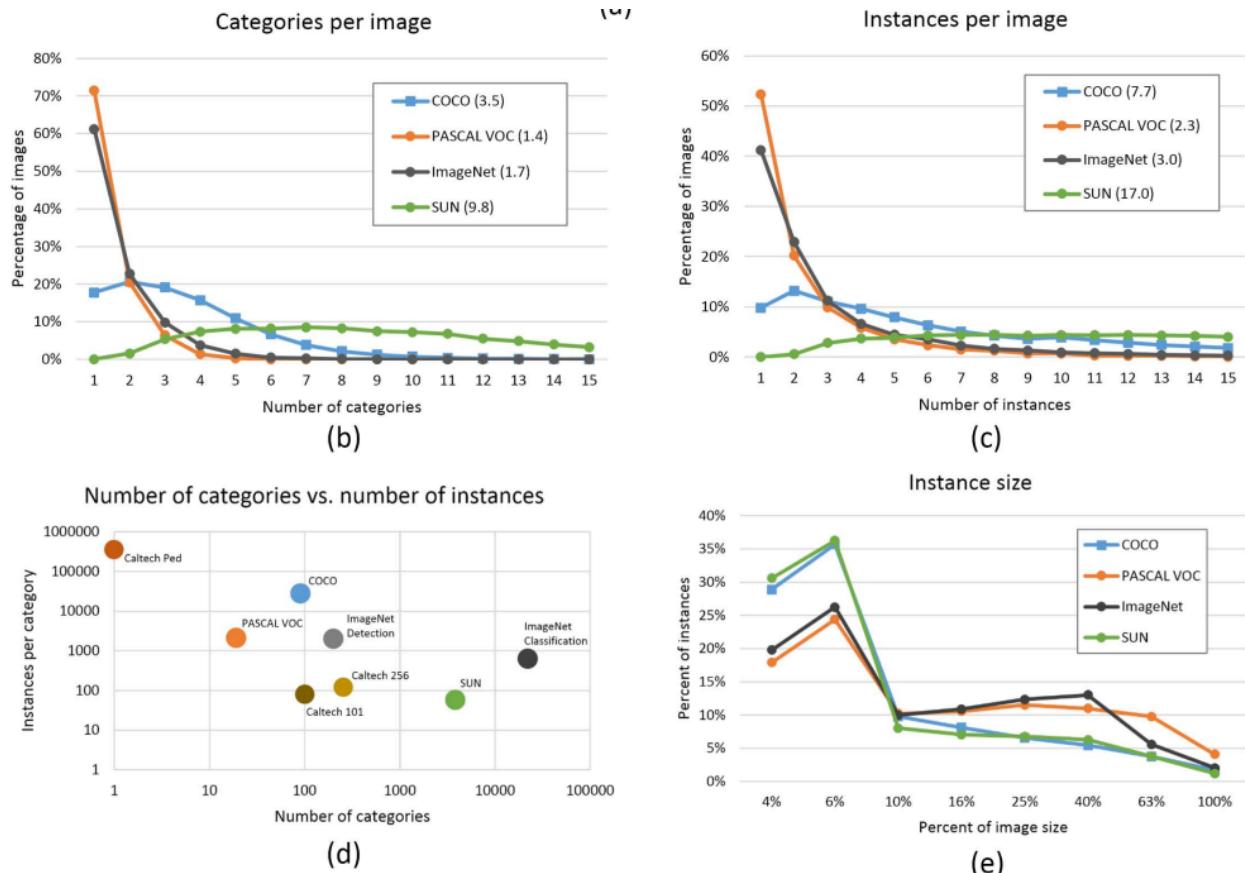
Fig. 2. Example of (a) iconic object images, (b) iconic scene images, and (c) non-iconic images. In this work we focus on challenging non-iconic images.

- **Non-iconic image collection**

- object categories list 고려할 때 다음 목표는 candidate image set 수집하는 것
- 일반적인 iconic object image는 fig.2 (a)에 해당, image 중앙에 하나의 객체 가지고 있음
- iconic scene image는 fig.2 (b)에 해당, 사람이라는 object가 부족함
- iconic image들은 고품질의 object instance 제공하지만 연구진들의 목표는 대부분의 image가 non-iconic 하도록 dataset 수집하는 것
- non-iconic image는 fig.2 (c)에 해당, object categories 따로 검색하지 않고 Flickr에서 image 수집

Dataset Statistics





ImageNet, PASCAL VOC 2012, and SUN datasets과 비교

- **ImageNet**

- 많은 수의 object categories capture하기 위해 생성되었으며, 그 중 많은 categories가 세분화되어 있음
- Training data에는 label 지정된 단일 object만 있기 때문에 object detection validation set floating함
- 평균 이미지 당 2개 미만의 category와 3개의 instance 가짐

- **SUN**

- scene 유형과 scene에서 일반적으로 발생하는 object에 label 지정하는데 초점 맞춤
- scene 기반으로 가상 상황에 맞는 정보 가짐

- **PASCAL VOC**

- 주요 application은 natural image에서의 object detection
- 평균 이미지 당 2개 미만의 category와 3개의 instance 가짐

- **MS COCO**

- 자연스러운 맥락에서 발생하는 object 감지하고 세분화하도록 설계됨
- MS COCO는 ImageNet과 SUN보다 category수 적지만 더 많은 instance 가짐
- MS COCO는 PASCAL VOC와 비교해 더 많은 category와 instance 가짐
- 평균적으로 dataset에는 image 당 3.5개 category와 7.7개 instance 포함됨

dataset의 평균 object 크기를 분석해 보면 일반적으로 크기가 작은 object는 인식 어렵고, 인식하기 위해 더 많은 상황 추론 필요

객체 평균 크기는 MS COCO와 SUN 모두 작음

Algorithmic Analysis

구체적인 benchmark 설립 위해 dataset을 training, validation, test data로 나눔

164,000개의 training set image와 각각 82,000개의 validation image, test image

Discussion

- 해당 논문은 일상 생활에서 발견되는 object들을 detection하고 segmentation 하기 위한 새로운 dataset에 대해 설명함
- natural 환경과 다양한 관점에 있는 object의 non-iconic한 image 찾는데 중점 둠
- 해당 연구진들은 'things'에만 label 붙였지만 'stuff' labeling은 detection에 도움 줄 수 있는 중요한 상황 정보 제공 가능

Reference

COCO - Common Objects in Context

 <https://cocodataset.org/>



Papers with Code - COCO Dataset

The MS COCO (Microsoft Common Objects in Context) dataset is a large-scale object detection, segmentation, key-point detection, and captioning dataset. The dataset consists of 328K images. Splits: The first version of MS COCO dataset was released in

 <https://paperswithcode.com/dataset/coco>

