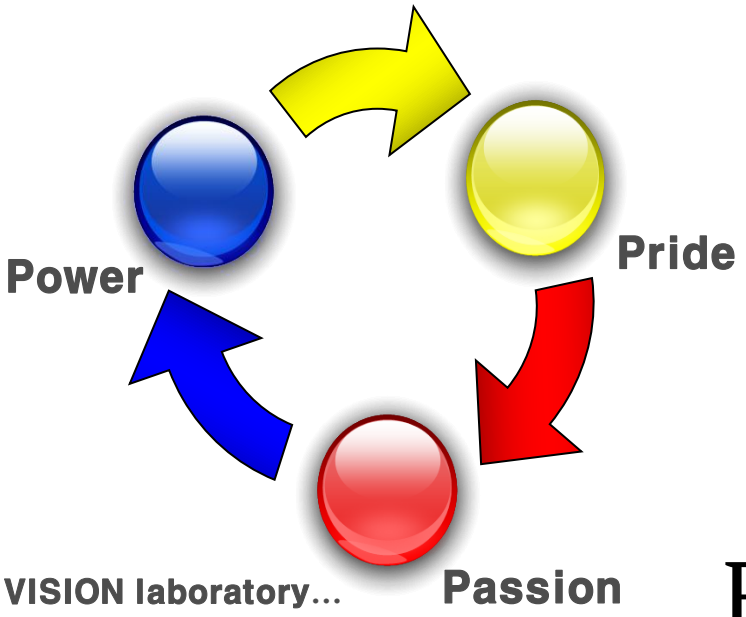


Faster R-CNN

Towards Real-Time Object Detection with Region Proposal Networks

Vision System Lab, Gyumin Park
yywnnaa@gmail.com
Jan 24, 2024

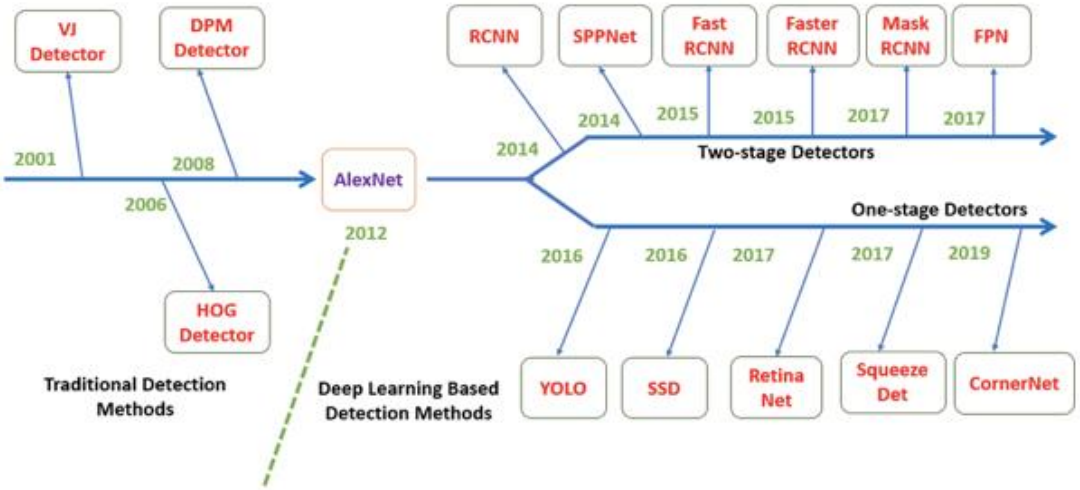
Soongsil University
VISION SYSTEM LABOARTORY

Object Detection

Object Detection이란 한 물체(single object)가 아닌 여러 물체(Multiple objects)에 대해 어떤 물체인지 클래스를 분류하는 **Classification** 문제와, 그 물체가 어디 있는지 박스를 (Bounding box) 통해 위치 정보를 나타내는 **Localization** 문제를 모두 포함

즉 Object Detection = Multiple Object에 대한 Multi-Labeled Classification + Bounding Box Regression(Localization) 라고 정리할 수 있다.

위쪽이 **2-stage Detector** 논문
물체를 식별하는 **Classification** 문제와, 물체의 위치를 찾는 **Localization** 문제
두 가지 task를 동시에 행하는 방법



R-CNN부터 Fast R-CNN,
Faster R-CNN같은 R-CNN 계열이 대표적

아래쪽은 **1-stage Detector** 논문
두 문제를 순차적으로 행하는 방법

YOLO(You Only Look Once)계열과
SSD 계열 등이 포함

1-stage Detector : 비교적 빠르지만 정확도 낮음
2-stage Detector : 비교적 느리지만 정확도 높음

VISION SYSTEM LABORATORY

Abstract

State-of-the-art object detection networks depend on region proposal algorithms to hypothesize object locations.
최첨단 객체 탐지 네트워크는 객체 위치를 가정하기 위해 영역 제안 알고리즘에 의존

Advances like SPPnet and Fast R-CNN have reduced the running time of these detection networks, exposing region proposal computation as a bottleneck.
SPPnet과 Fast R-CNN과 같은 발전은 이러한 탐지 네트워크의 실행 시간을 단축시켰으며,
결과적으로 region proposal영역 제안 계산이 병목 현상으로 드러났음=
: region proposal에서 시간 너무 많이 걸림

RPN : Region Proposal Network 제안, full-image의 합성곱 특징을 detection network와 공유, cost-free region proposals - region proposals의 계산비용/시간 거의 해결

An RPN is a fully convolutional network that simultaneously predicts object bounds and objectness scores at each position.
RPN은 각 위치에서 **객체 경계와 객체성(객체인지 아닌지) 점수를 동시에 예측**하는 완전 컨볼루션 네트워크

The RPN is trained end-to-end to generate high-quality region proposals, which are used by Fast R-CNN for detection.
RPN은 종단간 학습 통해 고품질의 region proposals영역 제안 생성

VISION SYSTEM LABORATORY

Abstract

We further merge RPN and Fast R-CNN into a single network by sharing their convolutional features—
RPN과 Fast R-CNN을 단일 네트워크로 *합쳤음* / 합성곱 특징을 공유

using the recently popular terminology of neural networks with "attention" mechanisms, the RPN component요소 tells the unified통일된 network where to look.
어디에 주목해야 하는지 알려주는 'attention' 메커니즘을 도입

For the very deep VGG-16 model, our detection system has a frame rate of 5fps (including all steps) on a GPU, while achieving state-of-the-art object detection accuracy on PASCAL VOC 2007, 2012, and MS COCO datasets with only 300 proposals per image.
매우 깊은 VGG-16 모델에 대해 / GPU에서 / 모든 단계를 포함하여 / **초당 5프레임의 속도로 작동**하면서 / sota-**최첨단 객체 탐지 정확도** 달성 / PASCAL VOC 2007, 2012, MS COCO 데이터셋에서 / **이미지 당 300개의 proposal**만으로

ILSVRC and COCO 2015 competitions : 1st-place winning entries in several tracks.
ILSVRC와 COCO 2015 competition에서 / 여러 트랙에서 1등을 차지

기존 Fast R-CNN 모델은 여전히 selective search 알고리즘을 통해 region proposal을 추출하기 때문에 학습 및 detection 속도를 향상시키는데 한계 있음,
또한 detection을 위한 과정을 end-to-end로 수행하지 못한다는 문제
이러한 문제를 해결 - Faster R-CNN : 속도와 모델의 완성도 측면에서 더 좋은 모습

VISION SYSTEM LABORATORY

1. Introduction

RPN : Region Proposal Network,영역 제안 네트워크를 도입 - 객체 탐지의 효율성과 정확성을 높임
최첨단 객체 탐지 네트워크와 컨볼루션 계층을 공유 - 테스트 시간에 컨볼루션을 공유함으로써 proposal을 계산하는 데 드는 비용이 적음

convolutional feature가 region proposal을 생성하는 데에도 사용될 수 있다는 것을 관찰
RPN은 region bound(객체의 경계)와 '객체인지 아닌지'를 판별하는 점수를 동시에 예측
- 이로 인해 영역 제안 계산의 비용이 거의 없어짐

RPN은 region proposal을 다양한 크기와 가로세로 비율과 함께 효율적으로 예측하도록 설계됨
이미지나 필터의 피라미드를 사용하는 기존 방법과 달리, 이 연구에서는 **다양한 크기와 가로세로 비율을 참조하는 '앵커' 박스를 도입**
이미지 피라미드: 입력 이미지가 주어지면 업샘플링하거나 다운샘플링 - RPN에서는 사용X

RPN과 Fast R-CNN 객체 탐지 네트워크를 통합하기 위해, region proposal영역 제안 작업과 object detection객체 탐지 작업을 번갈아가며 fine-tuning하는 학습 방식을 제안
이 과정에서 proposal제안들은 고정된 채로 유지
이 방식은 빠르게 converge수렴하고(모여들다,만나다) 두 가지 작업 모두에서 공유하는 컨볼루션 특징을 가진 통합 네트워크를 생성
이 방법은 Selective Search의 거의 모든 계산 부담을 해제, proposal제안의 실제 실행 시간은 단지 10밀리초

ILSVRC와 COCO 2015 여러 트랙에서 1등
: 실용적인 사용을 위한 **비용 효율적**인 해결책일 뿐만 아니라
객체 탐지 정확도를 향상시키는 효과적인 방법

VISION SYSTEM LABORATORY

2 RELATED WORK : Object Proposals

Widely used object proposal methods / object proposal으로 널리 사용되는 방법 :

grouping super-pixels 슈퍼픽셀을 그룹화하는 방법 (e.g., Selective Search [4], CPMC [22], MCG [23])

sliding windows 슬라이딩 윈도우를 기반으로 하는 방법 (e.g., objectness in windows [24], EdgeBoxes [6])

Object proposal methods were adopted as external modules independent of the detectors

(e.g., Selective Search object detectors, RCNN, and Fast R-CNN).

객체 제안 방법은 탐지기로부터 독립적인 외부 모듈이 채택됨

무슨말? -> 객체 제안 방법이 객체 탐지 알고리즘과 별개로 작동

Selective Search : 객체 제안 방법 중 하나, 이미지에서 관심 영역을 식별

(슈퍼픽셀을 그룹화하여 관련성이 높은 영역끼리 결합하는 방식으로 작동)

식별된 영역들은 후속 과정에서 객체가 포함될 가능성이 있는 영역으로 간주되어 객체 탐지 알고리즘의 입력으로 사용됨

Fast R-CNN 같은 객체 탐지 알고리즘 : 이렇게 제안된 영역을 받아들이고, 그 영역이 특정 객체를 포함하고 있는지를 판단

이때 Selective Search와 같은 object proposal객체 제안 모듈은 객체 탐지 알고리즘과 독립적으로 작동하며,

그 결과를 객체 탐지 알고리즘에 제공하는 역할

VISION SYSTEM LABORATORY

2 RELATED WORK : Deep Networks for Object Detection

The R-CNN method trains CNNs end-to-end to classify the proposal regions into object categories or background.

R-CNN 방법은 CNN을 종단간으로 학습시켜 제안된 영역을 객체 카테고리나 배경으로 분류

R-CNN mainly plays as a classifier, and it does not predict object bounds (except for refining by bounding box regression).

R-CNN은 주로 분류기로 작동하며, 객체의 경계를 예측하지 않음(바운딩 박스 회귀에 의한 세부 조정 제외)

Its accuracy depends on the performance of the region proposal module

그 정확도는 region proposal영역 제안 모듈의 성능에 의존

: **R-CNN 방법은 CNN을 통해 제안된 영역을 객체 카테고리나 배경으로 분류하는 방법으로,**

주로 분류기 역할을 함

Several papers have proposed ways of using deep networks for predicting object bounding boxes

몇몇 논문들은 객체 바운딩 박스를 예측하기 위해 딥 네트워크를 사용하는 방법을 제안

In the OverFeat method, a fully-connected layer is trained to predict the box coordinates for the localization task that assumes a single object.

OverFeat 방법에서는 완전 연결 계층이 단일 객체를 가정하는 위치 지정 작업에 대한 박스 좌표를 예측하도록 학습됨

2 RELATED WORK : Deep Networks for Object Detection

The MultiBox methods generate region proposals from a network whose last fully-connected layer simultaneously predicts multiple class-agnostic boxes, generalizing the “singlebox” fashion of OverFeat.

MultiBox 방법은 마지막 완전 연결 계층이 여러 클래스에 구애받지 않는 박스를 동시에 예측하는 네트워크에서 영역 제안을 생성

These class-agnostic boxes are used as proposals for R-CNN 이러한 박스는 R-CNN의 제안으로 사용됨

MultiBox does not share features between the proposal and detection networks.

MultiBox는 제안과 탐지 네트워크 간의 특징을 공유하지 않음

OverFeat, MultiBox 방법은 객체의 바운딩 박스를 예측하기 위해 Deep Network를 사용,
이들 방법은 제안과 탐지 네트워크 간의 특징을 공유하지 않음

Concurrent with our work, the DeepMask method [28] is developed for learning segmentation proposals

동시에, DeepMask 방법이 세분화 제안을 학습하는 데 사용되어 발전해 왔음

DeepMask가 동시에 발전했다는 얘기가 갑자기 왜 나오는지? -> DeepMask는 객체 탐지와 관련된 다른 방법론 중 하나로, 이미지 내에서 객체가 있을 가능성이 있는 영역을 학습하여 추출하는 방법을 제공함,

딥마스크가 저자의 작업과 동시에 개발되었다고 언급되어 있는데 / 객체 탐지와 관련된 연구가 다양한 방향으로 동시에 진행되고 있음을 보여주는 예시임

최근에(논문 시점)는 OverFeat, SPP, Fast R-CNN 등의 방법을 통해 **Shared computation**공유된 컨볼루션 특징에 대한 연구가 활발히 이루어지고 있음,

이는 효율적이면서도 정확한 시각 인식을 가능하게 함

VISION SYSTEM LABORATORY

3 FASTER R-CNN

Faster R-CNN, is composed구성된 of two modules. **Faster R-CNN은 두 개의 모듈로 구성**되어 있음

The first module is a deep fully convolutional network that proposes제안 regions,

1 : 영역을 제안하는 deep fully깊은 완전 컨볼루션 네트워크

and the second module is the Fast R-CNN detector that uses the proposed regions.

2 : 제안된 영역을 사용하는 Fast R-CNN 탐지기

The entire system is a single, unified통일된 network for object detection (Figure 2).

전체 시스템은 객체 탐지를 위한 단일 통합 네트워크

Using the recently popular terminology of neural networks with ‘attention’ mechanisms,

the RPN module tells the Fast R-CNN module where to look.

RPN 모듈이 / 'attention' 메커니즘을 사용해 / Fast R-CNN 모듈에게 / 어디를 봐야 할지 알려줌

Section 3.1 : the designs and properties특성 of the network for region proposal

영역 제안을 위한 네트워크의 설계와 특성을 소개

Section 3.2 : we develop개발 algorithms for training both modules with features shared.공유된 특징을 가진

두 모듈을 공유된 특징으로 학습시키는 알고리즘을 개발

VISION SYSTEM LABORATORY

3.1 Region Proposal Networks

RPN - Region Proposal Network : 이미지에서 객체 후보 영역을 제안하는 역할, 다음과 같은 핵심 요소로 구성됨

1. 입력 이미지 처리: RPN은 어떤 크기의 이미지도 입력으로 받아들여 각 객체 후보에 대한 사각형 형태의 제안과 그에 대한 객체 가능성 점수를 출력
2. 완전 합성곱 네트워크 (fully convolutional network) : RPN은 완전 합성곱 네트워크로 모델링되며, 이는 공유된 합성곱 계층을 가진 Fast R-CNN 객체 탐지 네트워크와 연산을 공유하는 것을 목표로 함
3. 공유 가능한 합성곱 계층: 실험에서는 5개의 공유 가능한 합성곱 계층을 가진 ZF 모델과 / 13개의 공유 가능한 합성곱 계층을 가진 VGG-16 모델을 사용
4. 지역 제안 생성: 마지막 공유 합성곱 계층의 출력으로부터 얻은 합성곱 특징 맵 위에 작은 네트워크를 슬라이딩하며 지역 제안을 생성
5. 작은 네트워크 구조: 이 작은 네트워크는 입력 합성곱 특징 맵의 n×n 공간 윈도우를 입력으로 받아들임. 각 슬라이딩 윈도우는 256차원(ZF용) 또는 512차원(VGG용)의 하위 차원 특징으로 매핑되고, ReLU 활성화 함수를 거침

VISION SYSTEM LABORATORY

3.1 Region Proposal Networks

6. 완전 연결 계층: 이 특징은 두 개의 sibling 완전 연결 계층으로 전달되는데, 하나는 박스 회귀(reg) 계층이고 다른 하나는 박스 분류(cls) 계층임

   - 박스 회귀(reg): 각 후보 영역의 위치와 크기를 조정하는 값들을 예측. 이는 실제 객체가 존재할 것으로 예상되는 영역의 경계를 더 정확하게 만드는 데 사용됨

   - 박스 분류(cls): 각 후보 영역에 대해 객체가 존재할 확률을 계산. 이 값은 객체의 '있음' 또는 '없음'을 나타내는 점수로, 객체 탐지의 신뢰도를 표현

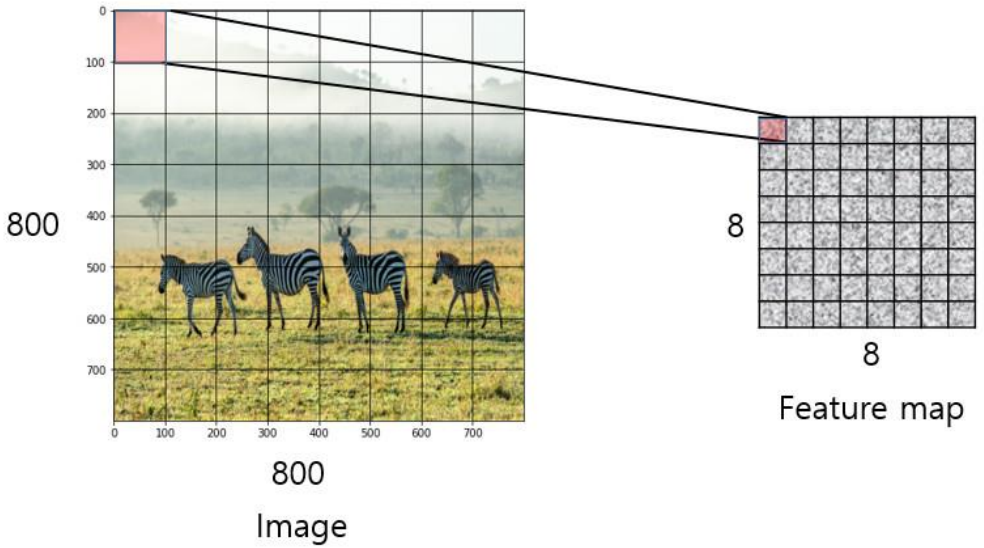
7. 네트워크 아키텍처: 이 아키텍처는 n×n 합성곱 계층 다음에 두 개의 형제 1×1 합성곱 계층을 두어 구현되며, 이는 reg와 cls에 각각 해당함
8. 효과적인 수용 필드: n = 3을 사용하며, 이는 입력 이미지에 대해 큰 효과적 수용 필드(171 픽셀 for ZF, 228 픽셀 for VGG)를 가짐
9. 공간 위치에서의 계층 공유: 미니 네트워크는 슬라이딩 윈도우 방식으로 작동하기 때문에, 완전 연결 계층은 모든 공간 위치에 걸쳐 공유됨

이러한 구조를 통해 RPN은 효율적으로 이미지 내에서 객체의 후보 영역을 식별하고,

각 영역에 대한 객체 가능성 점수를 할당함

VISION SYSTEM LABORATORY

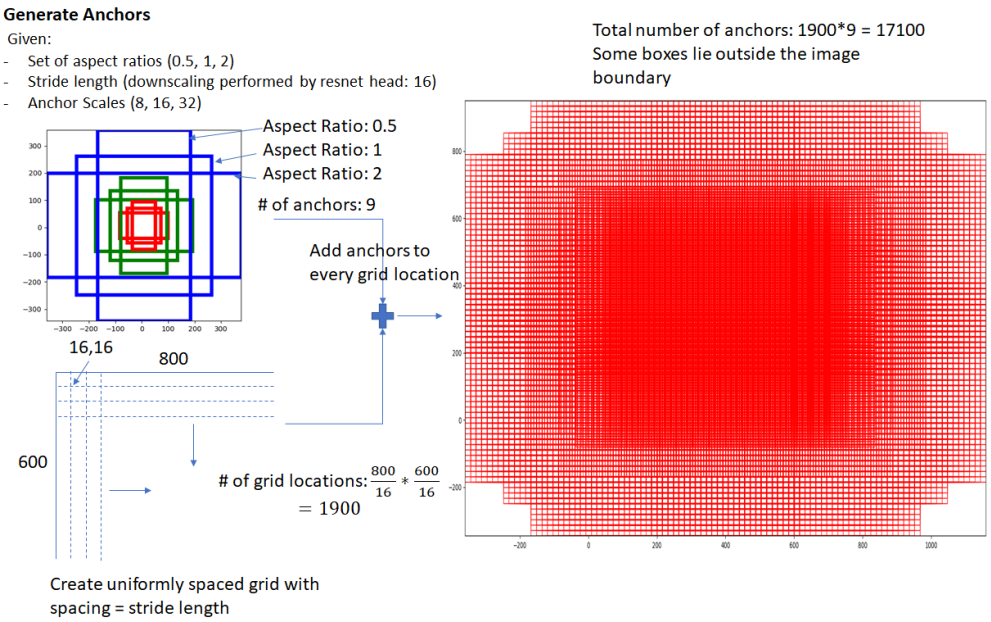
3.1.1 Anchors



- Selective search를 통해 region proposal을 추출하지 않을 경우, 원본 이미지를 일정 간격의 grid로 나눠 각 grid cell을 bounding box로 간주하여 feature map에 encode하는 Dense Sampling 방식을 사용. 이같은 경우 sub-sampling ratio를 기준으로 grid를 나누게 됨. 가령 원본 이미지의 크기가 800x800이며, sub-sampling ratio가 1/100이라고 할 때, CNN 모델에 입력시켜 얻은 최종 feature map의 크기는 8x8(800x1/100)가 됨. 여기서 feature map의 각 cell은 원본 이미지의 100x100만큼의 영역에 대한 정보를 함축하고 있다고 할 수 있음. 원본 이미지에서는 8x8개만큼의 bounding box가 생성된다고 볼 수 있음

- 이처럼 고정된 크기(fixed size)의 bounding box를 사용할 경우, 다양한 크기의 객체를 포착하지 못할 수 있다는 문제. 본 논문에서는 이러한 문제를 해결하고자 지정한 위치에 사전에 정의한 서로 다른 크기(scale)와 가로세로비(aspect ratio)를 가지는 bounding box인 Anchor box를 생성하여 다양한 크기의 객체를 포착하는 방법을 제시. 논문에서 3가지 scale([128, 256, 512])과 3가지 aspect ratio([1:1, 1:2, 2:1])를 가지는 총 9개의 서로 다른 anchor box를 사전에 정의(pre-define)

3.1.1 Anchors



anchor box는 원본 이미지의 각 grid cell의 중심을 기준으로 생성

본 이미지에서 sub-sampling ratio를 기준으로 anchor box를 생성하는 기준점인 anchor를 고정

이 anchor를 기준으로 사전에 정의한 anchor box 9개를 생성

위의 그림에서 원본 이미지의 크기는 600x800이며, sub-sampling ratio=1/16

이 때 anchor가 생성되는 수는 1900(=600/16 x 800/16)이며, anchor box는 총 17100(=1900 x 9)개가 생성

이같은 방식을 사용할 경우, 기존에 고정된 크기의 bounding box를 사용할 때보다 9배 많은 bounding box를 생성하며, 보다 다양한 크기의 객체를 포착하는 것이 가능

VISION SYSTEM LABORATORY

3.1.1 Anchors

At each sliding-window location, we simultaneously동시에 predict multiple region proposals,

각 슬라이딩 윈도우 위치에서, 동시에 여러 region proposal을 예측

where the number of maximum possible proposals for each location is denoted as k.

k : 각 위치에서 가능한 최대 proposal 수

So the reg layer has 4k outputs encoding the coordinates of k boxes,

따라서 reg 레이어는 k 상자의 좌표를 인코딩하는 4k 출력을 가지고 있고,

and the cls layer outputs 2k scores that estimate probability of object or not object for each proposal

cls 레이어는 각 제안에 대해 객체 또는 비객체일 확률을 추정하는 2k 점수를 출력

The k proposals are parameterized relative to k reference boxes, which we call anchors.

앵커라고 부르는 k개의 레퍼런스 박스에 상대적으로 설정되어 있다

An anchor is centered at the sliding window in question, and is associated with a scale and aspect ratio (Figure 3, left).

By default we use 3 scales and 3 aspect ratios종횡비, yielding생성 k = 9 anchors at each sliding position.

For a convolutional feature map of a size W × H (typically일반적으로 ~2,400), there are WHk anchors in total.

CNN을 통해 뽑아낸 피쳐 맵을 입력으로 받음

VISION
SYSTEM
LABORATORY

Translation-Invariant Anchors

- Translation-Invariant 이동 불변성: 접근 방식은 앵커와 앵커에 대해 제안을 계산하는 함수 모두에 있어서 이동 불변성을 가짐. 즉, 이미지 내 객체가 이동하더라도, 제안된 영역도 같이 이동해야 하고, 동일한 함수가 어느 위치에서든 제안을 예측할 수 있어야 함
- MultiBox와의 비교: MultiBox 방법은 800개의 앵커를 생성하기 위해 k-평균 클러스터링을 사용하는데, 이는 이동 불변성이 없음. 따라서 객체가 이동했을 때 같은 제안을 보장하지 않음
- 모델 크기 감소: 이동 불변성은 모델 크기를 줄이는 데에도 도움이 됨. MultiBox는 (4 + 1) × 800 차원의 완전 연결 출력 계층을 가지는 반면, 제안된 방식은 k=9 앵커의 경우 (4 + 2) × 9 차원의 합성곱 출력 계층을 가짐
- 파라미터 수: 결과적으로 제안된 방식의 출력 계층은 VGG-16을 사용할 때 2.8 × 10^4개의 파라미터를 가지며, 이는 MultiBox의 출력 계층이 가진 6.1 × 10^6개의 파라미터보다 훨씬 적음
- 과적합 위험 감소: 더 적은 수의 파라미터를 가진 제안된 방식은 작은 데이터셋에서 과적합의 위험이 더 적을 것으로 예상됨. 예를 들어 PASCAL VOC와 같은 데이터셋에서 더 효과적일 수 있음

- 계산 효율성을 높이고 모델의 복잡성을 줄이며, 작은 데이터셋에서도 잘 동작할 수 있는 구조를 제공

VISION SYSTEM LABORATORY

Multi-Scale Anchors as Regression References

다양한 스케일과 종횡비를 다루는 새로운 앵커(anchor) 설계에 대해 설명

멀티스케일 예측을 위한 기존 방법들:

첫 번째 방법 : 이미지/특성 피라미드에 기반. 여러 스케일에서 이미지를 재조정하고 각 스케일에 대해 특성 맵(예: HOG 또는 심층 합성곱 특성)을 계산. 이 방법은 유용하지만 시간이 많이 소요

두 번째 방법 : 특성 맵 위에서 다양한 스케일(and/or aspect ratios)의 슬라이딩 윈도우를 사용. DPM과 같은 모델에서는 다른 종횡비를 가진 모델을 다른 필터 크기를 사용하여 별도로 훈련

앵커 기반 방법의 비교:

앵커 기반 방법은 앵커의 피라미드 위에 구축되며, 비용 효율적

이 방법은 단일 스케일의 이미지와 특성 맵에 의존하고, 단일 크기의 필터(특성 맵 위의 슬라이딩 윈도우)를 사용

다양한 스케일의 앵커 디자인:

다양한 스케일과 종횡비의 앵커 박스에 대한 참조를 통해 경계 상자를 분류하고 회귀

이 멀티스케일 앵커 설계는 추가적인 비용 없이 스케일을 처리하기 위해 특성을 공유하는 핵심 구성 요소

실험을 통해 이 설계가 다양한 스케일과 크기를 다루는 효과를 확인

따라서, 이 앵커 기반 접근법은 다양한 크기와 종횡비를 효율적으로 다루면서, Fast R-CNN 탐지기와 같이 단일 스케일 이미지에서 계산된 합성곱 특성을 활용할 수 있는 방법을 제공

VISION SYSTEM LABORATORY

3.1.2 Loss Function

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i, t_i^*)$$

rpn_loss = rpn_cls_loss + (rpn_lambda * rpn_loc_loss)

- i : mini-batch 내의 anchor의 index
- p_i : anchor i에 객체가 포함되어 있을 예측 확률
- p_i^* : anchor가 양성일 경우 1, 음성일 경우 0을 나타내는 index parameter
- t_i : 예측 bounding box의 파라미터화된 좌표(coefficient)
- t_i^* : ground truth box의 파라미터화된 좌표
- L_{cls} : Loss loss
- L_{reg} : Smooth L1 loss

2400제한이유 : object보다 배경으로 검출되는 개수 더 많음
제한 안 두면 배경 위주로 학습, 배경의 개수를 제한하기 위한 목적

- N_{cls} : mini-batch의 크기(논문에서는 256으로 지정)
- N_{reg} : anchor 위치의 수
- λ : balancing parameter(default=10)

VISION SYSTEM LABORATORY

3.1.3 Training RPNs

- 훈련 방법: RPN은 역전파(backpropagation)와 확률적 경사 하강법(SGD)을 통해 종단간(end-to-end) 훈련될 수 있음
- 샘플링 전략: '이미지 중심' 샘플링 전략을 따르며, 각 미니 배치는 많은 긍정 및 부정 예제 앵커를 포함하는 단일 이미지에서 생성됨
- 앵커 샘플링: 모든 앵커에 대한 손실 함수를 최적화할 수 있지만, 부정적인 샘플이 지배적이기 때문에 편향될 수 있음. 대신, 손실을 계산하기 위해 이미지에서 무작위로 256개의 앵커를 샘플링하고, 이 중 긍정과 부정 샘플의 비율은 최대 1:1, 긍정 샘플이 128개 미만인 경우(보통 객체가 배경보다 적기때문에) 부정 샘플로 미니 배치를 채움
- 초기화: 새로운 계층은 표준편차가 0.01인 zero-mean 가우시안 분포(분포의 평균값이 0)에서 추출한 가중치로 무작위 초기화되며, 다른 모든 계층(공유 합성곱 계층)은 ImageNet 분류를 위한 모델 사전 훈련을 통해 초기화됨
- 학습률 설정: PASCAL VOC 데이터셋에서는 처음 60k 미니 배치에 대해 학습률을 0.001로 설정하고, 이후 20k 미니 배치에 대해서는 0.0001로 조정
- 모멘텀과 가중치 감소: 모멘텀은 0.9로, 가중치 감소는 0.0005로 설정
- 구현: Caffe 프레임워크를 사용

VISION SYSTEM LABORATORY

3.2 Sharing Features for RPN and Fast R-CNN

RPN과 Fast R-CNN을 공유 합성곱 계층을 가지고 통합적으로 학습하는 방법에 대해 설명

세 가지 주요 훈련 방법 제시:

1. 교대 훈련(Alternating training): 먼저 RPN을 훈련하고, 그 제안을 사용하여 Fast R-CNN을 훈련. Fast R-CNN에 의해 조정된 네트워크가 다시 RPN 초기화에 사용되고, 이 과정을 반복. 이 방법은 이 논문의 모든 실험에서 사용

2. 근사적 결합 훈련(Approximate joint training): RPN과 Fast R-CNN을 하나의 네트워크로 합쳐서 훈련. 각 SGD(Stochastic Gradient Descent) 반복에서, 순방향 패스는 영역 제안을 생성하고, 이를 Fast R-CNN 탐지기를 훈련할 때와 같이 고정된, 사전 계산된 제안처럼 다룸. 역전파는 공유 계층에 대해 RPN 손실과 Fast R-CNN 손실의 신호를 결합하여 일반적으로 진행됨. 이 방법은 구현이 쉽지만, 제안된 박스의 좌표에 대한 미분을 무시하기 때문에 근사적임. 실험에서는 이 방법이 **비슷한 결과**를 내면서 교대 훈련 대비 훈련 **시간을 25-50% 절약**한다는 것을 발견

3. 비근사적 결합 훈련(Non-approximate joint training): RPN에 의해 예측된 경계 상자bounding box도 입력의 함수(also functions of the input)이므로, 이론적으로 유효한 역전파 solver는 박스 좌표에 대한 그래디언트를 포함해야 함(위의 근사적 결합 훈련에서는 이 그래디언트를 무시함) 비근사적 결합 훈련에서는 박스 좌표에 대해 미분 가능한 RoI 풀링 계층이 필요. 이는 복잡한 문제이며, "RoI warping" 계층을 통해 해결할 수 있지만, 이 논문의 범위를 벗어남(beyond the scope)

solver : 주어진 수학적 모델의 파라미터를 최적화하기 위해 사용되는 알고리즘을 지칭

4-Step Alternating Training

RPN과 Fast R-CNN을 번갈아가며 학습시키는 Alternating Training 방법을 사용

1) 먼저 Anchor generation layer에서 생성된 anchor box와 원본 이미지의 ground truth box를 사용하여 Anchor target layer에서 RPN을 학습시킬 positive/negative 데이터셋을 구성함. 이를 활용하여 **RPN을 학습**시킴. 이 과정에서 pre-trained된 VGG16 역시 학습됨

2) Anchor generation layer에서 생성한 anchor box와 학습된 RPN에 원본 이미지를 입력하여 얻은 feature maps를 사용하여 proposals layer에서 region proposals를 추출함. 이를 Proposal target layer에 전달하여 Fast R-CNN 모델을 학습시킬 positive/negative 데이터셋을 구성함. 이를 활용하여 **Fast R-CNN을 학습**시킴. 이 때 pre-trained된 VGG16 역시 학습됨

3) 앞서 학습시킨 RPN과 Fast R-CNN에서 **RPN에 해당하는 부분만 학습(fine tune)**시킴. 세부적인 학습 과정은 1)과 같음. 이 과정에서 두 네트워크끼리 공유하는 convolutional layer, 즉 **pre-trained된 VGG16은 고정(freeze)**함

4) 학습시킨 RPN(3)번 과정)을 활용하여 추출한 region proposals를 활용하여 **Fast R-CNN을 학습(fine tune)**시킴. 이 때 **RPN과 pre-trained된 VGG16은 고정(freeze)**

실제 학습 절차가 상당히 복잡하여 이후 두 네트워크를 병합하여 학습시키는 Approximate Joint Training 방법으로 대체됨

3.3 Implementation Details

- 이미지 스케일링: 학습과 테스트 모두 단일 스케일의 이미지를 사용. 이미지는 짧은 쪽이 600픽셀이 되도록 재조정
- 멀티스케일 특성 추출: 이미지 피라미드를 사용한 다양한 스케일의 특성 추출은 정확도를 향상시킬 수 있지만, 속도와 정확도 사이에서 좋은 트레이드오프를 보여주지 않음
  : 정확도는 향상시킬 수 있지만, 그 과정에서 소요되는 시간이 많아져서 전체적인 성능(속도 대비 정확도)이 크게 향상되지 않음
- 총 보폭 : 재조정된 이미지에서 ZF와 VGG 네트워크의 마지막 합성곱 계층의 총 보폭 16픽셀. 이는 일반적인 PASCAL 이미지(대략 500×375)에서 재조정 전 약 10픽셀에 해당
- 앵커 박스: 3가지 스케일(128^2, 256^2, 512^2 픽셀의 박스 영역)과 3가지 종횡비(1:1, 1:2, 2:1)를 사용하는 앵커 박스를 사용
- 하이퍼파라미터: 이 하이퍼파라미터들은 특정 데이터셋에 대해 not carefully chosen세심하게 선택되지 않았으며, 그 영향에 대한 ablation 실험을 다음 섹션에서 제공
  대신, 이러한 하이퍼파라미터 설정이 다양한 데이터셋에 대해 광범위하게 적용 가능하며, 이에 대한 영향을 연구하기 위한 Ablation 실험을 수행했다고 언급 : 하이퍼파라미터가 너무 특정한 경우에만 유효한 것이 아니라 일반적인 상황에서도 충분히 좋은 성능을 낼 수 있음을 보여주고자 함
- 이미지 피라미드 불필요: 여러 스케일의 영역을 예측하기 위해 이미지 피라미드나 필터 피라미드가 필요하지 않으므로 상당한 실행 시간을 절약
- 앵커 박스 처리: 학습 중에는 경계를 넘는 모든 앵커를 무시하여 손실에 기여하지 않게 함
- NMS(Non-Maximum Suppression): 서로 많이 겹치는 RPN 제안들을 줄이기 위해, cls 점수를 기준으로 제안 영역에 NMS를 적용. NMS의 IoU 임계값은 0.7로 설정되며, 이미지 당 약 2000개의 제안 영역을 남김
- RPN 제안 사용: NMS 이후에는 탐지를 위해 상위 N개의 제안 영역을 사용. Fast R-CNN 학습에는 2000개의 RPN 제안을 사용하지만, 테스트 시에는 다른 수의 제안을 평가함

4.1 Experiments on PASCAL VOC

PASCAL VOC 2007 detection benchmark : 20개의 객체 카테고리에 걸쳐 약 5,000개의 trainval 이미지와 약 5,000개의 테스트 이미지로 구성

We also provide results on the PASCAL VOC 2012 benchmark for a few models. For the ImageNet pre-trained network, we use the “fast” version of ZF net [32] that has 5 convolutional layers and 3 fully-connected layers, and the public VGG-16 model7 [3] that has 13 convolutional layers and 3 fully-connected layers.

ImageNet 사전 훈련된 네트워크에 대해서는 ZF net의 "빠른" 버전과 VGG-16 모델을 사용

We primarily evaluate detection mean Average Precision (mAP), because this is the actual metric for object detection (rather than focusing on object proposal proxy metrics). 평가 지표로는 객체 검출의 실제 메트릭스인 mAP를 사용

Fast R-CNN 프레임워크 내에서 다양한 지역 제안 방법을 활용한 실험 결과를 통해, RPN이 적은 수의 제안(최대 300개)을 사용하면서도 경쟁력 있는 mAP인 59.9%를 달성

RPN을 사용하면 공유된 컨볼루션 계산 덕분에 Selective Search나 EdgeBoxes를 사용하는 시스템보다 훨씬 빠른 탐지 시스템을 구현할 수 있으며, 적은 수의 제안은 지역별 완전 연결 계층의 비용도 줄여준다

VISION SYSTEM LABORATORY

Ablation Experiments on RPN

Ablation : 실험적으로 어떤 부분이나 구성요소를 제거하거나 비활성화하여 시스템이나 모델의 동작이 어떻게 변하는지를 조사하는 실험적인 방법

Ablation study(Experiment)의 단계

1. 원래 모델: 실험의 기본이 되는 원래 모델이나 시스템을 설정
2. 특정 부분 제거: 특정 부분, 특성, 레이어, 또는 구성요소를 제거하거나 비활성화
3. 평가: 수정된 모델을 사용하여 원래 목표를 평가하고 성능을 측정
4. 결과 비교: 수정된 모델의 결과를 원래 모델의 결과와 비교하여 특정 부분이나 기능이 모델의 성능에 미치는 영향을 이해

이 논문에서

- 공유된 합성곱 계층의 영향: RPN과 Fast R-CNN 탐지 네트워크 간에 합성곱 계층을 공유할 때와 공유하지 않을 때의 성능 차이를 조사
- 훈련된 탐지 네트워크에서 RPN의 영향: RPN이 탐지 네트워크 훈련에 어떻게 영향을 미치는지 파악하기 위해, 테스트 시 다른 제안 영역들을 사용한 탐지 mAP를 측정
- cls와 reg 출력의 역할: RPN이 생성하는 제안의 품질에 대해 cls(분류) 레이어와 reg(회귀, 즉 위치 조정) 레이어가 각각 어떤 역할을 하는지 분석하기 위해 테스트 시 이들 중 하나를 비활성화
- 강력한 네트워크의 영향: 더 성능이 좋은 VGG-16 네트워크를 사용하여 RPN을 훈련시키고, 이로 인한 제안 품질의 변화를 조사

VISION SYSTEM LABORATORY

Ablation Experiments on RPN

Ablation 결과

1. 공유된 합성곱 계층: RPN과 Fast R-CNN이 합성곱 계층을 공유할 때, 더 빠른 속도와 개선된 탐지 성능을 보여줌 : 중복 계산을 줄이고, 학습된 특징을 효율적으로 재사용함으로써 얻어진 결과
2. 훈련된 탐지 네트워크에서 RPN의 영향: RPN으로부터의 제안을 사용할 때 탐지 네트워크가 더 높은 mAP를 달성 - RPN이 더 정확한 영역 제안을 생성해내기 때문
3. cls와 reg 출력의 역할: 분류와 회귀 출력은 제안의 품질에 중요한 역할을 함. 분류는 객체가 있는지 없는지를 판단하는 반면, 회귀는 객체의 정확한 위치를 조정함. 둘 중 하나가 비활성화되면 성능이 저하됨
4. 강력한 네트워크의 영향: VGG-16과 같은 더 강력한 네트워크를 사용하면, 제안의 품질이 향상되고, 결과적으로 탐지 성능이 개선됨

Performance of VGG-16

Table 3: Detection results on **PASCAL VOC 2007 test set**. The detector is Fast R-CNN and VGG-16. Training data: "07": VOC 2007 trainval, "07+12": union set of VOC 2007 trainval and VOC 2012 trainval. For RPN, the train-time proposals for Fast R-CNN are 2000. †: this number was reported in [2]; using the repository provided by this paper, this result is higher (68.1).

method	# proposals	data	mAP (%)
SS	2000	07	66.9†
SS	2000	07+12	70.0
RPN+VGG, unshared	300	07	68.5
RPN+VGG, shared	300	07	69.9
RPN+VGG, shared	300	07+12	**73.2**
RPN+VGG, shared	300	COCO+07+12	**78.8**

표 3은 VGG-16을 사용하여 제안과 탐지 모두에 대한 결과

RPN과 VGG를 결합한 방식(RPN+VGG)은 unshared features공유되지 않은 특징에 대해 68.5%의 결과를 보여주며, 이는 Selective Search(SS) 베이스라인보다 약간 높은 성능 - RPN+VGG에 의해 생성된 제안들이 SS보다 더 정확하기 때문

사전에 정의된 SS와 달리, RPN은 활발히 훈련되며 benefits from better network더 나은 네트워크로부터 혜택을 받음
무슨뜻?->지속적으로 데이터로부터 학습하고, 이 과정에서 성능 향상을 위해 최신의 더 발전된 신경망 구조를 활용한다는 의미

특징이 공유된 변형에 대해서는 69.9%의 결과를 보여주며, 이는 SS 베이스라인보다 뛰어나고 거의 추가 비용 없이 제안을 생성
또한, RPN과 탐지 네트워크를 PASCAL VOC 2007 trainval과 2012 trainval의 합집합에 대해 추가로 훈련시키면 mAP는 73.2%

RPN과 VGG-16 네트워크를 결합한 객체 탐지 방법이 기존의 Selective Search방법보다 더 나은 성능
이 방법은 공유되지 않은 특징에 대해 68.5%, 공유된 특징에 대해서는 69.9%의 mAP를 달성하며,
추가적인 훈련을 통해 더 높은 mAP를 얻을 수 있음을 보여줌

VISION SYSTEM LABORATORY

Performance of VGG-16

Table 5: **Timing** (ms) on a K40 GPU, except SS proposal is evaluated in a CPU. "Region-wise" includes NMS, pooling, fully-connected, and softmax layers. See our released code for the profiling of running time.

model	system	conv	proposal	region-wise	total	rate
VGG	SS + Fast R-CNN	146	1510	174	1830	0.5 fps
VGG	RPN + Fast R-CNN	141	**10**	47	**198**	**5 fps**
ZF	RPN + Fast R-CNN	31	**3**	25	**59**	**17 fps**

표 5 : 객체 탐지 시스템 전체의 실행 시간을 요약

Selective Search(SS)는 내용에 따라 1-2초가 걸리며(평균적으로 약 1.5초),
Fast R-CNN에 VGG-16을 사용할 경우 2000개의 SS 제안에 대해 320ms가 걸림
(완전 연결 계층에 SVD를 사용하면 223ms).
VGG-16을 적용한 우리 시스템은 제안과 탐지를 합쳐 총 198ms가 걸리며,
합성곱 특징이 공유될 때 RPN만으로 추가 계층을 계산하는 데 10ms만 소요
또한, 이미지 당 제안 수가 적은(300개) 덕분에 지역별 계산 시간도 줄어든다
ZF 네트워크를 사용했을 때 우리 시스템은 초당 17 프레임의 처리 속도

논문의 객체 탐지 시스템이 Selective Search를 사용하는 기존 시스템보다 훨씬 빠른 처리 시간 가진다
이 시스템은 제안 생성과 객체 탐지를 결합하여 198ms 만에 처리할 수 있으며,
이는 공유된 합성곱 특징과 적은 수의 제안 덕분에 가능
결과적으로, 이 시스템은 ZF 네트워크를 사용할 때 초당 17 프레임의 속도로 작동
(Fast R-CNN 모델은 0.5fps)

VISION SYSTEM LABORATORY

Sensitivities to Hyper-parameters

앵커 설정에 대한 조사 (Table 8):
anchor의 다양한 설정이 모델의 성능에 미치는 영향에 관한 내용

By default we use 3 scales and 3 aspect ratios (69.9% mAP in Table 8)
기본적으로 3개의 스케일과 3개의 종횡비를
사용할 때의 결과는 69.9%의 mAP

If using just one anchor at each position,
the mAP drops by a considerable margin of 3-4%.
각 위치에 하나의 앵커만 사용하면 mAP가
3-4% 정도 크게 감소

Table 8: Detection results of Faster R-CNN on PASCAL VOC 2007 test set using **different settings of anchors**. The network is VGG-16. The training data is VOC 2007 trainval. The default setting of using 3 scales and 3 aspect ratios (69.9%) is the same as that in Table 3.

settings	anchor scales	aspect ratios	mAP (%)
1 scale, 1 ratio	128^2	1:1	65.8
	256^2	1:1	66.7
1 scale, 3 ratios	128^2	{2:1, 1:1, 1:2}	68.8
	256^2	{2:1, 1:1, 1:2}	67.9
3 scales, 1 ratio	$\{128^2, 256^2, 512^2\}$	1:1	**69.8**
3 scales, 3 ratios	$\{128^2, 256^2, 512^2\}$	{2:1, 1:1, 1:2}	**69.9**

The mAP is higher if using 3 scales (with 1 aspect ratio) or 3 aspect ratios (with 1 scale), demonstrating that using anchors of multiple sizes as the regression references is an effective solution.
3개의 스케일 (1개의 종횡비) 또는 3개의 종횡비 (1개의 스케일)를 사용하는 경우 mAP가 더 높아지며,
이는 여러 크기의 앵커를 regression references회귀 참조로 사용하는 것이 효과적인 해결책임을 보여줌

Using just 3 scales with 1 aspect ratio (69.8%) is as good as using 3 scales with 3 aspect ratios on this dataset, suggesting that scales and aspect ratios are not disentangled dimensions for the detection accuracy. But we still adopt these two dimensions in our designs to keep our system flexible.
3개의 스케일과 1개의 종횡비를 사용하는 경우 (69.8%) 해당 데이터셋에서는 3개의 스케일과 3개의 종횡비를 사용하는 것과 거의 동일한 결과, 이는 스케일과 종횡비가 감지 정확도에 대해 not disentangled dimensions별개의 차원이 아님을 나타냄 : **스케일과 종횡비가 서로 독립적이지 않다**
그러나 이러한 두 가지 차원은 시스템을 유연하게 유지하기 위해 여전히 채택됨

VISION SYSTEM LABORATORY

Sensitivities to Hyper-parameters

Table 9: Detection results of Faster R-CNN on PASCAL VOC 2007 test set using **different values of** λ in Equation (1). The network is VGG-16. The training data is VOC 2007 trainval. The default setting of using $\lambda = 10$ (69.9%) is the same as that in Table 3.

λ	0.1	1	10	100
mAP (%)	67.2	68.9	69.9	69.1

In Table 9 we compare different values of λ in Equation (1).

λ에 대한 여러 가지 값을 비교

By default we use λ = 10 which makes the two terms in Equation (1) roughly equally weighted after normalization.

기본값으로는 λ = 10을 사용하며, 이는 정규화 후 방정식 (1)의 두 항을 대략 동일한 가중치로 만든다

Table 9 shows that our result is impacted just marginally (by ~ 1%) when λ is within a scale of about two orders of magnitude (1 to 100).

표 9의 결과는 λ가 약 두 개 정도의 크기 (1에서 100) 내에서 변할 때 약 1% 정도의 영향만을 미친다는 것을 보여줌

This demonstrates that the result is insensitive to λ in a wide range.

λ의 변화에 대해 모델의 성능이 민감하지 않음

VISION SYSTEM LABORATORY

Analysis of Recall-to-IoU

recal리콜? -> 정확하게 탐지해야 할 전체 대상 중 얼마나 많은 대상을 실제로 탐지했는지를 나타내는 비율

$$Recall(True\ Positive\ Rate) = \frac{TP}{TP + FN}$$

True Positive, TP: 모델이 정확하게 긍정으로 예측하고 실제로도 긍정인 경우의 수

False Negative, FN: 모델이 부정으로 예측했지만 실제로는 긍정인 경우의 수

It is noteworthy주목할만한 that the Recall-to-IoU metric is just loosely related to the ultimate최종 detection accuracy.

It is more appropriate적절한 to use this metric to diagnose진단 the proposal method than to evaluate it.

리콜 대 IoU 지표가 최종적인 탐지 정확도와는 느슨한 연관성이 있음을 지적하며,

이 지표로 proposal method 제안 방법을 평가하기보다는 / diagnose진단하는 데 더 적절함

loosely relate느슨한 연관? -> 어느 정도 관련성은 있으나 이 관계가 원인과 결과를 명확하게 설명하진 못함

제안 방법을 "진단" ? -> 해당 방법이 얼마나 잘 작동하는지, 즉 region proposal이 실제 객체와 얼마나 잘 일치하는지를 파악하는 데 사용한다는 의미

제안 방법의 효율성을 평가하는 것보다, 그 방법이 실제로 유용한 결과를 얼마나 잘 생성하는지를 이해하는 데 더 중점을 둠

다시 말해, 최종적인 탐지 정확도(어떤 객체가 실제로 탐지되었는지의 정확성)를 측정하는 것이 아니라, 제안 방법이 얼마나 잠재적으로 유용한 region proposal을 생성하는지를 확인하는 데 사용한다는 것

VISION SYSTEM LABORATORY

Analysis of Recall-to-IoU

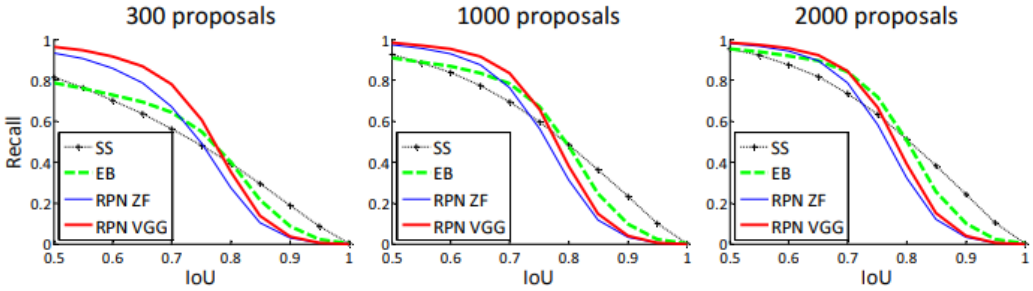


Figure 4: Recall *vs.* IoU overlap ratio on the PASCAL VOC 2007 test set.

In Figure 4, we show the results of using 300, 1000, and 2000 proposals. We compare with SS and EB, and the N proposals are the top-N ranked ones based on the confidence generated by these methods.

300, 1000, 2000개의 제안을 사용한 결과를 그림 4에 나타내고 있으며, Selective Search(SS)와 EdgeBoxes(EB) 방법과 비교

The plots show that the RPN method behaves gracefully when the number of proposals drops from 2000 to 300. This explains why the RPN has a good ultimate detection mAP when using as few as 300 proposals.

RPN 방법은 제안의 수가 2000개에서 300개로 줄어들 때도 성능이 gracefully우아하게 유지되는 것을 보여주며, 이는 RPN이 단지 300개의 제안을 사용했을 때에도 좋은 최종 탐지 mAP(mean Average Precision)를 얻을 수 있는 이유를 설명해 줌.

이유 : **고품질의 proposal제안을 생성하기 때문**

제안의 수가 300개로 줄어들었을 때에도, 이 중 대부분이 실제 객체와 높은 IoU를 갖게 되므로, 실제로 객체가 존재하는 지역을 놓치지 않고 잘 탐지할 수 있음

As we analyzed before, this property is mainly attributed to the cls term of the RPN. The recall of SS and EB drops more quickly than RPN when the proposals are fewer. 이 능력은 이 'cls term', 즉 분류 기능에 주로 기인

이는 Selective Search(SS)와 EdgeBoxes(EB)와 같은 다른 제안 방법들이 제안의 수가 적을 때 리콜이 더 빠르게 감소하는 것과 대비되는데, 이는 RPN이 상대적으로 더 효율적으로 고품질의 제안을 생성한다는 것을 시사

VISION SYSTEM LABORATORY

One-Stage Detection vs. Two-Stage Proposal + Detection

two-stage 시스템인 Faster R-CNN과 / one-stage 시스템인 OverFeat 간의 비교

OverFeat는 컨볼루션 특징 맵 위의 슬라이딩 윈도우에 분류기와 회귀 분석기를 사용하는 탐지 방법을 제안하고, 클래스별 탐지 파이프라인이며,

Faster R-CNN은 클래스에 무관한 제안과 클래스별 탐지로 구성된 두 단계의 cascade (단계로 구성된 프로세스)

OverFeat와 RPN은 모두 슬라이딩 윈도우를 사용하지만, RPN에서는 이러한 제안을 더 세밀하게 수정하기 위해 Fast R-CNN 검출기가 제안에 주의를 기울임

one-stage와 두 단계 시스템을 비교하기 위해, one-stage Fast R-CNN으로 OverFeat 시스템을 에뮬레이션함. 이 시스템에서는 "proposals"이 3개의 스케일(128, 256, 512)과 3개의 종횡비(1:1, 1:2, 2:1)의 밀집한 슬라이딩 윈도우

에뮬레이션(emulation) : 하나의 시스템이 다른 시스템처럼 행동하게 만드는 것

OverFeat 시스템이 사용하는 이미지 피라미드 방식이나, 특정한 스케일과 종횡비를 가진 밀집 슬라이딩 윈도우 방식을 Fast R-CNN 시스템에서 사용함으로써, OverFeat의 방식을 '에뮬레이트'

구현상의 다른 차이점들을 뛰어넘어 두 시스템을 더 공정하게 비교할 수 있게 됨

One-Stage Detection vs. Two-Stage Proposal + Detection

Table 10: **One-Stage Detection *vs.* Two-Stage Proposal + Detection**. Detection results are on the PASCAL VOC 2007 test set using the ZF model and Fast R-CNN. RPN uses unshared features.

	proposals		detector	mAP (%)
Two-Stage	RPN + ZF, unshared	300	Fast R-CNN + ZF, 1 scale	58.7
One-Stage	dense, 3 scales, 3 aspect ratios	20000	Fast R-CNN + ZF, 1 scale	53.8
One-Stage	dense, 3 scales, 3 aspect ratios	20000	Fast R-CNN + ZF, 5 scales	53.9

- Using the ZF model, the one-stage system has an mAP of 53.9%. This is lower than the two-stage system (58.7%) by 4.8%.
  -> 표현 오류? 두 값의 차이는 4.8%가 아니라 / 4.8 퍼센트 포인트 pencentage point (53.9+4.8=58.7)
- 53.9라는 숫자에서 4.8% 증가시키면 53.9*1.048=56.5
- %로 표현하려면 8.9%임 (58.7/53.9=1.089, 53.9*1.089=58.69)
- 비슷한 사례 (수능 영어 오류)

- 이 결과는 지역 제안의 효과와 객체 탐지의 연속성을 증명
- one-stage 시스템은 더 많은 제안을 처리해야 하므로 더 느림
- 이런 내용들을 통해, two-stage 시스템인 Faster R-CNN이 one-stage인 OverFeat에 비해 **더 높은 성능**을 보이며, 특히 객체 탐지 작업에서 더 정확하다는 점을 강조

VISION SYSTEM LABORATORY

4.2 Experiments on MS COCO

Fast R-CNN과 Faster R-CNN 모델이 Microsoft COCO 객체 탐지 데이터셋에서 어떻게 성능을 발휘하는지, 그리고 Faster R-CNN이 Fast R-CNN에 비해 어떤 개선을 보여주는지

- Fast R-CNN과 Faster R-CNN 모델은 훈련, 검증, 테스트-개발 세트에서 평가됨
- 모델 훈련에는 8-GPU 구현이 사용되었고, 학습률과 미니 배치 크기 등 여러 파라미터가 조정됨
- Fast R-CNN과 Faster R-CNN 모두에서 음성 샘플의 정의를 변경하여 성능을 향상

음성 샘플의 정의 변경 : 이전에는 음성 샘플을 ground truth와의 IoU(Intersection over Union)가 [0.1, 0.5)인 경우로 정의

-> ground truth와의 IoU가 [0, 0.5)인 경우로 변경

SVM(서포트 벡터 머신) 단계에서는 hard-negative mining이라는 과정을 통해 음성 샘플들을 검토함. 이 과정에서는 분류기가 잘못 분류한(즉, 양성으로 잘못 판단한) 음성 샘플들을 찾아내어 분류기의 성능을 개선하는 역할을 함. 이 때문에 SVM 단계에서는 [0, 0.5) 범위의 모든 음성 샘플들이 검토 대상이 될 수 있음. 그러나 Fast R-CNN은 이 SVM 단계를 포기하였고, 대신 네트워크 fine-tuning 과정에서 [0.1, 0.5) 범위의 음성 샘플만을 사용, 이로 인해 [0, 0.1) 범위의 음성 샘플들은 절대 방문되지 않게 됨

이러한 문제를 해결하기 위해, Fast R-CNN 단계에서 음성 샘플의 정의를 ground truth와의 IoU가 [0, 0.5)인 경우로 변경.

이렇게 하면 [0, 0.1) 범위의 음성 샘플들도 포함되어 학습 과정에 참여하게 되어 모델의 성능을 개선하는 데 도움이 됨

- Faster R-CNN은 Fast R-CNN에 비해 더 높은 mAP를 보였고, 특히 더 높은 IoU 임계값에서의 위치 정확도를 향상시키는 데 탁월

VISION SYSTEM LABORATORY

Faster R-CNN in ILSVRC & COCO 2015 competitions

- Faster R-CNN은 RPN이 신경망을 통해 영역을 제안하는 방법을 완전히 학습하므로, 더 좋은 특징을 활용하여 더 큰 이점을 얻을 수 있음

RPN completely learns : 완전히라는게 무슨 의미? ->

"RPN이 완전히 학습한다"는 표현은 RPN이 신경망을 통해 영역 제안 방법을 스스로 학습한다는 뜻

"완전히"라는 단어는 이 과정이 RPN 내부의 신경망에 의해 독립적으로 이루어진다는 것을 강조하는 말

사전에 정의된 휴리스틱이나 수동으로 설정된 규칙 없이도, 학습 데이터를 통해 어떤 영역이 객체를 포함할 가능성이 높은지 판단하는 방법을 스스로 학습함 (데이터로부터 복잡 패턴을 추출하고 이해하는 딥 뉴럴 네트워크의 능력에 기반)

- VGG-16을 101층의 잔차 네트워크(ResNet-101)로 교체하면, Faster R-CNN 시스템은 COCO val 세트에서 mAP를 상당히 향상시킬 수 있음
- Faster R-CNN은 COCO 2015 객체 탐지 대회 1위, ILSVRC 2015 객체 탐지 대회 1위
- RPN은 ILSVRC 2015 위치 지정 및 COCO 2015 분할 대회 1위를 차지한 항목의 구성 요소

Faster R-CNN은 객체 탐지 분야에서 높은 성능을 달성할 수 있는 효과적인 방법, 여러 대회에서 우수 성과

VISION SYSTEM LABORATORY

4.3 From MS COCO to PASCAL VOC

대규모 데이터는 딥 뉴럴 네트워크를 개선하는 데 중요함
이를 확인하기 위해 MS COCO 데이터셋이 PASCAL VOC의 탐지 성능에 어떻게 도움이 되는지 조사 :

COCO 탐지 모델을 PASCAL VOC 데이터셋에서 직접 평가한 결과,
PASCAL VOC 데이터에 세부 조정을 하지 않아도 76.1%의 mAP를 달성
이 결과는 PASCAL VOC 데이터를 활용하지 않았음에도 VOC07+12에서 훈련된 결과보다 우수
대규모 데이터셋인 MS COCO를 활용한 딥 뉴럴 네트워크 모델이 객체 탐지 성능을 획기적으로 향상

COCO 탐지 모델을 VOC 데이터셋에서 세부 조정(fine-tuning)한 결과, PASCAL VOC 2007 테스트 세트에서 78.8%의 mAP
이는 COCO 세트에서 추가 데이터를 활용함으로써 mAP가 5.6% 증가한 결과
MS COCO 데이터셋에 대해 학습된 모델을 PASCAL VOC 데이터셋에서 직접 평가했을 때도 이미 좋은 성능,
이후에 PASCAL VOC 데이터셋에 대해 추가로 세부 조정(fine-tuning)을 진행하니 성능이 더욱 향상

대규모 데이터셋에서 학습된 모델이 다른 데이터셋에도 잘 적용될 수 있음을 보여주며,
추가적인 세부 조정을 통해 성능을 더욱 개선할 수 있음

COCO+VOC에서 훈련된 모델은 PASCAL VOC 2007에서 모든 개별 카테고리에 대해 최고의 AP 보임
PASCAL VOC 2012 테스트 세트에서도 유사한 개선이 관찰

이런 강력한 결과를 얻는 데 테스트 시간의 속도는 이미지 당 약 200ms
: 대규모 데이터를 활용한 딥 뉴럴 네트워크의 효율성과 성능을 보여줌

VISION SYSTEM LABORATORY

5. Conclusion

- We have presented RPNs for efficient효율적 and accurate정확한 region proposal generation.

- By **sharing convolutional features**합성곱 특징 with the down-stream detection network,
- the region proposal step is nearly cost-free. / region proposal 단계는 거의 비용이 들지 않음

- Our method enables a unified통합된, deep-learning-based object detection system to run at near real-time frame rates.
- 통합된 딥러닝 기반 객체 탐지 시스템이 거의 실시간 프레임 속도로 실행될 수 있게 함

- The learned RPN also improves region proposal quality and thus the overall종합적인 object detection accuracy.
- 학습된 RPN은 또한 region proposal의 품질을 향상시키고, 종합적인 object detection 정확도를 개선함

VISION SYSTEM LABORATORY