You Only Look Once: Unified, Real-Time Object Detection

Euron 10기 Research 유다현

EWHA EURON

#01 Introduction

- 연구 배경 및 분야

☑ 분야 : 컴퓨터 비전(Computer Vision) & 실시간 객체 검출 (Real-Time Object Detection)

☑ 목표 : 이미지 내에서 “무엇이(What), 어디에 (Where) 있는지”를

인간의 시각 시스템처럼 빠르고 정확하게 파악하는 것

☑ 응용 : 자율주행 자동차. 보조 장치. 실시간 로봇 시스템 등

#01 Introduction

- 기존 연구의 한계점

객체 검출을 위해 분류기를 재사용하는 복잡한 파이프라인 가짐

. 1. DPM (Deformable Parts Models)

: 슬라이딩 윈도우 방식을 사용하여 이미지 전체를 균일한 간격으로 훑으며 분류기를 실행함

2. R-CNN 계열
   - Region Proposal : 이미지 내에서 물체가 있을 법한 바운딩 박스 후보군을 먼저 생성함
   - Complex Pipeline : 각 후보 박스에 대해 분류기를 돌리고, 위치를 조정하고, 중복을 제거하는 과정이 각각 독립적으로 진행됨

공통적인 문제점

속도 저하, 최적화의 어려움. 문맥 파악 부족

#01 Introduction

- 논문의 핵심 기여

" YOLO는 검출 프로세스를 단일 신경망으로 통합하여 기존의 패러다임을 바꿈 "

1. 통합 검출 (Unified Detection)

: 객체 검출을 '이미지 픽셀에서 바운딩 박스 좌표와 클래스 확률의 단일 회귀 문제 (Single Regression Problem)' 로 재정의

2. 실시간성 확보 (Extreme Speed)

: 표준 모델은 45 FPS. 경량화 모델 (Fast YOLO)은 150 FPS 이상의 속도를 기록하여 고성능 GPU에서 실시간 스트리밍 비디오 처리가 가능함..

3. 전역적 문맥 학습 (Global Reasoning)

: 학습 및 테스트 시 이미지 전체를 보기 때문에, 클래스의 외형뿐만 아니라 주변 배경과의 관계 (Contextual information)를 암묵적으로 학습함.
→ Fast R-CNN 대비 배경 에러를 절반 이하로 줄임

4. 강력한 일반화 성능 (Generalization)

: 일반 사진으로 학습하더라도 예술 작품과 같은 새로운 도메인에서 다른 모델들보다 훨씬 뛰어난 성능을 보임

#02 Related Work

1. DPM (Deformable Parts Models)

- 방식 : 슬라이딩 윈도우(Sliding Window) 기법을 사용하여 이미지 전체를 균일하게 훑으며 검출

- 특징 : 특징 추출, 분류, 바운딩 박스 예측 등이 독립적인 파이프라인으로 구성

- YOLO와의 차이 : YOLO는 이 모든 단계를 **단일 신경망** 으로 교체하여 더 빠르고 정확한 모델 형성

2. R-CNN 계열 (Region Proposals)

- 방식 : Selective Search를 통해 이미지 내에서 물체가 있을 법한 후보 박스(Region Proposals)를 먼저 생성한 뒤, 각 박스에 대해 CNN 분류기를 실행

- 문제점 : 각 단계를 개별적으로 학습시켜야 하므로 매우 복잡하고 느림
  테스트시 한 장의 이미지를 처리하는 데 수 초가 걸릴 정도로 비효율적

- YOLO와의 차이 : YOLO는 후보 영역을 미리 뽑지 않고, **이미지 전체를 격자로 나눠 한 번에 예측** 하므로 속도 빠름

#02 Related Work

- 기존 접근 방식과의 차별점

☑ 전역적 추론 (Global Reasoning)

기존 모델들 -> 좁은 영역에 집중
YOLO -> 학습 및 테스트 시 이미 전체 확인

☑ 배경 오류 감소

YOLO : 주변 문맥 정보를 함께 학습하기 때문에, Fast R-CNN 등에 비해
배경을 물체로 착각하는 오류가 절반 이하로 적음

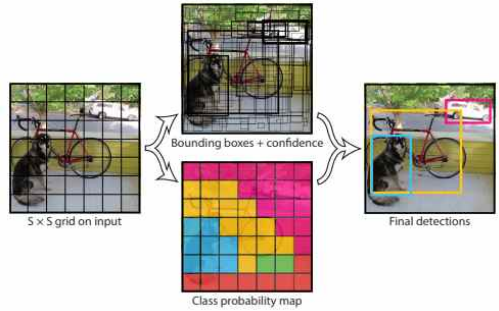
#03 Method

" YOLO는 객체 검출 파이프라인을 단일 신경망으로 통합하여,
이미지 전체를 한 번에 (You Only Look Once) 검출을 수행 "

- Unified Detection (통합 검출)

☑ Grid System : 입력 이미지를 S x S 그리드 (기본 7 x 7)로 나눔

☑ Responsibility :객체의 중심이 특정 그리드 셀에 위치하면,
해당 셀이 그 객체를 검출할 책임을 가짐.



☑ Predictions : 각 그리드 셀은 B개의 바운딩 박스와 각 박스에 대한
Confidence Score, 그리고 C개의Class Probabilities를 동시에 예측함.

☑ Output Tensor : 최종 출력은 S x S x (B x 5 + C) 형태의 단일 텐서이며.
PASCAL VOC 기준 7 x 7 x 30 크기

#03 Method

- Network Design

☑ 구조적 특징 : GoogLeNet의 재해석

Layer 구성 : 24개의 Convolution Layers와 2개의 Fully Connected Layers로 구성
Reduction 레이어 : 1 x 1 Reduction 레이어와 3 x 3 Convolution 레이어를 교차로 배치해 연산 효율 상승

☑ 입력 및 출력의 메커니즘

입력 해상도 (448 x 448) : 일반적인 분류 모델보다 높은 해상도 사용 -> 미세한 위치 정보 중요
최종 출력 텐서의 구조 (7 x 7 x30)

Spatial Grid (7 x 7) = 이미지를 49개의 구역으로 분류
Box Information (2 x 5) = 셀 당 2개의 박스 예측 + 각 박스는 x, y, w, h, confidence 데이터 가짐
Class Probabilities (20) = PASCAL VOC의 20개 클래스에 대한 확률값

☑ 활성화 함수 : (마지막 레이어 제외 모든 레이어) Leaky ReLU 적용 -> 뉴런이 죽는 현상 방지

#03 Method

- Training (학습 전략)

☑ 사전 학습 : 20개의 컨볼루션 레이어를 ImageNet 데이터셋으로 먼저 학습하여 특징 추출 능력을 극대화

☑ Multi-part Loss Function : SSE (Sum-Squared Error) 기반의 통합 손실 함수를 사용

- 가중치 조절 : 위치 오차 ($\lambda coord = 5$)는 중요하게 다루고,
  배경 오차 ($\lambda noobj = 0.5$)는 비중을 낮춰 학습불균형 해결

- 박스 크기 보정 : 큰 박스보다 작은 박스의 오차를 더 민감하게 반영하기 위해 너비와 높이에 제곱근 씌움

☑ 데이터 증강 : 원본 이미지의 20%까지 무작위 크기 조절/이동을 적용 + HSV 색상 공간에서의 노출과 채도 변화
-> 일반화 성능 향상

☑ 과적합 방지 : 첫 번째 전결합 레이어 뒤에 Dropout(0.5)을 적용하여 복잡한 네트워크 의존성 분산

☑ 하이퍼파라미터 : 배치 크기 64, 모멘텀 0.9, 가중치 감쇠 0.0005를 사용하여 135 에포크 동안 학습

#03 Method

- Inference (추론 과정)

☑ Single Pass : 이미지 한 장을 단 한 번의 연산으로 처리하여 실시간성 확보

☑ NMS (Non-Maximum Suppression)

문제) 한 객체에 대해 여러 개의 그리드 셀이 중복으로 박스를 예측하는 경우 발생

💡 신뢰도 (Confidence)가 가장 높은 박스만 남기고 나머지를 제거하여 최종 결과 도출

- Limitations (구조적 한계점)

☑ 공간적 제약 : 좁은 영역에 작은 물체가 밀집된 경우 검출력이 현저히 떨어짐

☑ 일반화 오류 : 학습 시 보지 못한 새로운 가로세로 비율의 물체에 취약함

☑ Loss의 한계 : 큰 박스와 작은 박스의 오차를 동일하게 취급하는 SSE 특성상, 작은 물체의 정확한 위치를 잡는데 한계가 있음

#04 Experiment

- Real-Time Systems 비교 (속도 및 정확도 비교)

Real-Time Detectors	Train	mAP	FPS
100Hz DPM [31]	2007	16.0	100
30Hz DPM [31]	2007	26.1	30
Fast YOLO	2007+2012	52.7	**155**
YOLO	2007+2012	**63.4**	45
Less Than Real-Time			
Fastest DPM [38]	2007	30.4	15
R-CNN Minus R [20]	2007	53.5	6
Fast R-CNN [14]	2007+2012	70.0	0.5
Faster R-CNN VGG-16[28]	2007+2012	73.2	7
Faster R-CNN ZF [28]	2007+2012	62.1	18
YOLO VGG-16	2007+2012	66.4	21

☑ 압도적인 실시간성

* Fast YOLO : 155 FPS의 속도로 현존 모델 중 가장 빠르며, mAP 52.7 기록

* YOLO (Base) :45 FPS로 실시간 검출을 유지하면서 mAP 63.4를 달성

#04 Experiment

- Error Analysis : YOLO vs Fast R-CNN

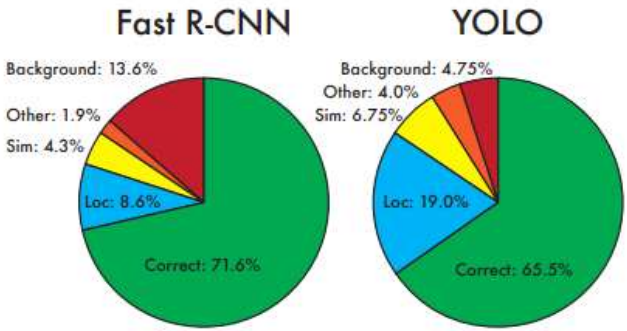


Figure 4: Error Analysis: Fast R-CNN vs. YOLO These charts show the percentage of localization and background errors in the top N detections for various categories (N = # objects in that category).

☑ Localization Error (위치 오차)

YOLO(19.0%)가 Fast R-CNN(8.6%)보다 높음

원인) SEE 손실 함수와 그리드 제약으로 인한 정교한 박스 조정의 한계

☑ Background Error (배경 오차)

YOLO(4.75%)가 Fast R-CNN(13.6%)보다 약 3배 낮음.

원인) 이미지 전체를 한 번에 보기 때문에 배경을 물체로 착각하는 실수가 훨씬 적음 (Global Reasoning의 증거)

#04 Experiment

- 모델 결합 (Ensemble : YOLO + Fast R-CNN)

	mAP	Combined	Gain
Fast R-CNN	71.8	-	-
Fast R-CNN (2007 data)	**66.9**	72.4	.6
Fast R-CNN (VGG-M)	59.2	72.4	.6
Fast R-CNN (CaffeNet)	57.1	72.1	.3
YOLO	63.4	**75.0**	**3.2**

Table 2: Model combination experiments on VOC 2007. We examine the effect of combining various models with the best version of Fast R-CNN. Other versions of Fast R-CNN provide only a small benefit while YOLO provides a significant performance boost.

☑ 시너지 효과

Fast R-CNN의 정교한 위치 예측 능력과
YOLO의 낮은 배경 오차를 결합

☑ 성능 향상

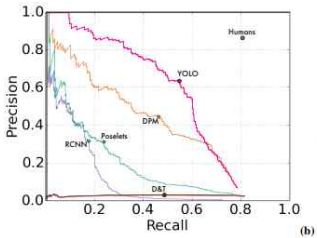
Fast R-CNN 단독((71.8 mAP)보다
3.2% 향상된 75.0 mAP를 기록

-> YOLO가 Fast R-CNN이 잡지 못한 배경 노이즈를 필터링하는 역할을 수행

#04 Experiment

- Generalization: 예술 작품 데이터셋 (Picasso, People-Art)

☑ 일반화 능력 입증 : 자연 이미지로 학습한 모델을 피카소나 모네의 회화에 적용



(a) Picasso Dataset precision-recall curves.

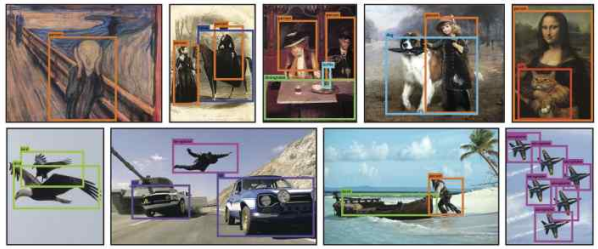
	VOC 2007	Picasso		People-Art
	AP	AP	Best F_1	AP
YOLO	**59.2**	**53.3**	**0.590**	**45**
R-CNN	54.2	10.4	0.226	26
DPM	43.2	37.8	0.458	32
Poselets [2]	36.5	17.8	0.271	
D&T [4]	-	1.9	0.051	

(b) Quantitative results on the VOC 2007, Picasso, and People-Art Datasets. The Picasso Dataset evaluates on both AP and best F_1 score.

Figure 5: Generalization results on Picasso and People-Art datasets.

결과) R-CNN이나 DPM은 스타일이 바뀐 이미지에서 성능이 급격히 하락하지만, YOLO는 높은 성능을 유지

-> 객체의 외형적 픽셀 정보 뿐만 아니라 전체적인 문맥과 형태를 학습했음을 의미함



#05 Conclustion

YOLO는 객체 검출 분야에서 기존의 복잡한 파이프라인을 탈피해
단일 통합 모델을 구현 했다는 점에서 큰 의의를 가짐.

- 핵심 기여

1. 간결한 구조
   신경망을 단일 구조로 통합하여 구성이 매우 단순
   이미지를 전체 직접 학습시키는 효율적 방식 채택

2. 압도적 실시간성
   Fast YOLO는 현존하는 가장 빠른 범용 객체 검출기
   YOLO는 실시간성을 유지하면서 높은 정확도 확보

3. 강력한 일반화 능력
   특정 데이터셋에 과적합되지 않고 새로운 도메인에서도 안정적인 성능을 보여, 다양한 실무 환경에 적용 가능

- 실무적 가치

1. 통합 파이프라인
   복잡한 제안 단계 없이 이미지 전체를 직접 연산하여 추론 속도가 배우 빠름

2. 범용성 증명
   단순 웹캠 연결만으로도 실시간 객체 추적 시스템으로 즉시 활용 가능함을 입증함

EWHA EURON

#05 Conclustion

YOLO는 객체 검출 분야에서 기존의 복잡한 파이프라인을 탈피해
단일 통합 모델을 구현 했다는 점에서 큰 의의를 가짐.

- 핵심 기여

1. 간결한 구조
   신경망을 단일 구조로 통합하여 구성이 매우 단순
   이미지를 전체 직접 학습시키는 효율적 방식 채택

2. 압도적 실시간성
   Fast YOLO는 현존하는 가장 빠른 범용 객체 검출기
   YOLO는 실시간성을 유지하면서 높은 정확도 확보

3. 강력한 일반화 능력
   특정 데이터셋에 과적합되지 않고 새로운 도메인에서도 안정적인 성능을 보여, 다양한 실무 환경에 적용 가능

- 실무적 가치

1. 통합 파이프라인
   복잡한 제안 단계 없이 이미지 전체를 직접 연산하여 추론 속도가 배우 빠름

2. 범용성 증명
   단순 웹캠 연결만으로도 실시간 객체 추적 시스템으로 즉시 활용 가능함을 입증함

EWHA EURON

#06 Discussion

1. YOLO는 실시간성을 위해 그리드 기반의 근사화(Approximation) 방식을 택하면서 결과적으로 Localization(위치 추정) 정확도에 손해를 보았습니다. 단순히 네트워크를 더 깊게 쌓는 것 외에, 속도를 유지하면서도 작은 물체에 대한 Localization 정확도를 높일 수 있는 구조적 대안은 무엇이 있을까요?

2. 논문에서는 YOLO가 이미지 전체를 바라봄으로써 배경 오차를 줄인다는 점을 강점으로 내세웠습니다. 그렇다면 데이터셋의 도메인이 극단적으로 제한적이거나(예: 의료 영상, 위성 사진), 물체가 매우 작고 밀집된 환경에서도 이러한 Global Context 접근법이 R-CNN 계열의 Regional Proposal 방식보다 여전히 유효한 전략일까요?

THANK YOU