hutom(killerchef)

Abstract
CNN기반 학습 객체 검출기에서 바운딩 박스에 대한 anchor-based training은 종종 현재 훈련 클래스에 다른 클래스의 객체를 포함시키는 문제 발생
->anchorless object detector 사용
->a semi-supervised learning-based object detection using a single object tracker 사용. (단일 객체 추적기를 사용하는 detector)
The proposed technique performs single object tracking by using the sparsely annotated bounding box as an anchor in the temporal domain for successive frames.
2가지 문제 해결법이 있었고, 위 팀은 2번째 방법을 택하였다.

1. Introduction
CNN의 빠른 발전과 그로인한 object detector 의 빠른 발전으로 평가를 위한 dataset도 PASCAL VOC 과 같이 복잡도가 낮은 데이터셋에서 시작하여 MS-COCO와 같이 복잡도가 높은 데이터셋으로 개발되었다.
에픽키친 데이터셋은 다른 데이터셋과 다른 다음과 같은 특징이 있다.
1. 학습 검출기용 이미지는 원본 비디오에서 수집하고 해당 프레임 시퀀스를 제공한다.
2. 훈련 이미지에서 훈련 가능한 객체 중 일부만 희소하게 주석 처리된다.
3. 소수의 샷 클래스와 많은 샷 클래스 간의 주석 양의 차이는 훈련 데이터셋의 객체 모양 분포에 따라 크다.
에픽키친 데이터셋에서는 anchor-based training을 수행할 때 학습해야하는 객체이지만 레이블 정보가 누락되어 훈련 성능이 저하된다.

위 이유에 따라 에픽키친의 주석은 기존 데이터셋과 다른 방식으로 제공되기때문에, 기존 객체 감지 모델을 훈련시키기는 방법을 그대로 적용하기 어려운 특성이 있다.

figure2
f는 N개의 프레임으로 구성된 액션 클립을 나타내고, o는 액션 클립에 존재하는 총 M개의 객체를 나타낸다.
액션 클립에 있는 모든 학습 가능한 객체에 대한 조밀한 레이블을 얻기 위해 양방향 추적을 통해 반지도 학습을 수행했다.

일반적인 anchor-based training을 수행하는 detector나 RPN의 구조로 학습시키는 detector는 효과적인 학습을 위해 bounding box와 함께 IoU를 고려하여 일괄 샘플링을 수행한다.
**RPN(RPN은 fully convolutional network로서 동시에 물체의 경계를 예측하고, 물체의 위치에 대한 점수를 예측한다. RPN은 고품질의 region proposals를 만들기 위**해서 end-to-end로 학습된다. Fast R-CNN에서 쓰였던 것과 같다.)
그러나 위와같이 학습을 할 경우 bounding box 근처의 객체 분포로 인해 학습의 효율성이 저하된다.

이 문제를 해결하기 위해 두가지 접근으로 객체를 훈련시켰다.
1. anchorless object detector 사용
- FCOS(Fully Convolutional One-Stage Object Detection) 네트워크를 활용하여 앵커 기반 샘플링의 영향을 받는 희소 주석이 달린 훈련 이미지의 영향을 최소화할 수 있었다.
2. a semi-supervised learning-based object detection using a single object tracker 사용. (단일 객체 추적기를 사용하는 detector)
- 특정 프레임에 존재하는 bounding box label을 시간 영역에서 초기 bounding box로 설정하고 단일 개체 추적기의 입력으로 사용한다.
그 후, 추적기의 임계값을 초과하는 예측 출력은 pesudo 주석으로 가정하고 학습 가능한 모든 객체에 대한 레이블이 훈련을 위해 이미지에 제공되었다.

Subsequently, joint NMS-based ensemble was performed for FCOS models with trained inhomogeneous backbones.

2. Related Work
-Object detection
CNN기반 모델은 1단계 모델과 2단계 모델로 나누어진다.
1단계 모델에서는 객체의 클래스와 위치를 예측하는 과정이 하나의 구조로 이루어지며, 예로는 YOLO, SSD, RetinaNet 등이 있다.
일반적으로 분류와 회귀가 하나의 구조에서 수행되기 때문에 2단계 모델보다 회귀 정확도가 낮은 것으로 알려져 있다.

2단계 모델의 경우 탐지기의 하위 네트워크인 RPN(Region Proposal Network)에서 물체의 위치에 대한 사전 지식을 추정한다.
RPN은 클래스에 구애받지 않는 서브넷으로 객체성을 판별하고 후속 헤드 구조를 통해 클래스 인식 감지를 수행한다.
Faster R-CNN, R-FCN, Cascade R-CNN, Cascade RPN, etc.
다양한 헤드 구조를 대표하며 상대적으로 회귀 정확도가 높은 것으로 알려져 있다.

1단계 및 2단계 모델의 철학을 결합한 RefineDet과 같은 모델도 제안되었으며, FCOS와 같은 구조적 이점보다 경계 상자 회귀에 대한 다른 매개변수화를 활용하는 검출기도 제안되었다.

-Semi-supervised learning for object detection.
semi-supervised learning을 사용한 객체 감지는 학습할 충분한 수의 주석을 수동으로 획득하기 어려운 상황 또는 상대적으로 많은 수의 라벨이 지정되지 않은 데이터에서 의사 라벨을 획득해야 하는 경우에 사용된다.
연속 프레임에서 좋은 의사 레이블을 얻기 위해 사전 훈련된 객체 감지기와 강력한 추적기를 사용하여 의사 레이블을 평가하고 재훈련하기 위한 반복 프레임워크를 제안했다.

제안하는 single object tracker-based semi-supervised learning은 추적기를 사용한다는 점에서 유사하지만 기존의 lean annotation information를 이용하여 특정 영상에 대한 조밀한 주석 정보를 얻는 데 차이가 있다.
동시에 추적을 위한 초기 입력으로 객체 검출기를 사용하지 않기 때문에 반복 훈련 시나리오로 훈련을 적용하지 않는다.

-Single object visual tracking.
단일 객체 추적 네트워크를 사용하여 드물게 표기된 데이터 세트에 대한 의사 레이블을 생성했다.
단일 개체 추적에서 the Siamese network-based visual tracker는 다양한 데이터 세트에서 균형 잡힌 정확도와 속도를 보여준다.
Siamese network-based visual tracker는 기본적으로 대상 이미지에 대한 CNN 기능과 추적을 위한 입력 이미지의 유사성으로 학습한다.
SiamMask를 단일 개체 추적기로 사용했다.

3. Fully Convolutional One-Stage Object Detection (FCOS)
FCOS 모델을 사용하여 detector training에서 앵커가 있는 좋은 긍정적인 예를 선택하기 위한 계산 프로세스를 제외했다.
FCOS는 bounding box의 회귀에 대한 매개 변수를 다르게 정의하고 앵커가 없는 검출기를 제시했다.

4. Semi-Supervised Learning with Single Object Tracker
Epic-Kitchens 데이터셋은 특정 개체에 대한 bounding box 라벨과 개체가 나타나는 액션 클립에 대한 시퀀스 프레임을 동시에 제공한다.
개체의 bounding box는 모든 프레임에 조밀하게 지정되지 않고 동작 시퀀스에서 드물게 지정된다.
다양한 단일 개체 추적기 중 SiamMask를 사용하여 추적 정확도와 속도에서 균형 잡힌 성능을 보여준다.
각 bounding box를 단일 객체의 초기 값으로 사용하여 DAVIS 데이터셋에서 훈련된 SiamMask 모델을 사용하여 양방향 추적을 수행했다.

하나의 액션 클립 입력에 대해 SiamMask를 사용한 정방향 추적에 대한 자세한 내용은 알고리즘 1에 설명되어 있다.
알고리즘 1은 양방향 추적을 완료하기 위해 역방향 추적에 동일한 방식으로 사용된다.

그림 3의 파란색 상자로 표시된 프레임은 추적이 시작된 이후 동일한 개체로 추적된 프레임이고, 빨간색 상자로 표시된 프레임은 알고리즘 1의 종료 조건으로 인해 추적이 종료된 프레임이다.

그림 4에서 semi-supervised learning에 사용할 최종 주석의 예는 빨간색 점선으로 표시된 훈련 이미지에 표시됩니다.

그림 3은 단일 객체에 대한 알고리즘 1에 따른 추적의 시작과 끝의 예를 보여주고 있으며, 그림 4는 객체를 추적한 후 생성된 유사 레이블이 있는 훈련 이미지를 보여준다.

5. Epic-Kitchens Object Detection Results

-Training details.

Faster R-CNN과 Cascade R-CNN을 anchor-based detector로 사용하고 FCOS를 anchorless detector로 사용하여 EpicKitchens 객체 감지 데이터 세트의 성능을 비교했다.
검출기 훈련을 위한 백본 CNN은 ImageNet으로 사전 훈련된 ResNet-50, ResNet-101, ResNeXt-101, HRNet-V2p-W32를 사용하였으며, 각각의 백본과 헤드 구조 조합에 대한 훈련 내용은 Table 1과 같다.
모든 실험은 MMDetection 라이브러리를 사용하여 수행되었다.

-Anchor-based vs. anchorless detector

Table 2는 anchor-based detector와 anchorless detector의 단일 모델에서의 훈련 성능을 보여준다.
Table 2에 따르면 훈련의 기본 성능에서 anchorless detector의 성능이 우수하고 안정적인 학습 결과를 보여주고 있음을 알 수 있다.
그림 5는 anchor-based detector와 anchorless detector를 훈련하는 동안의 손실 변화를 보여준다.
anchorless detector는 상대적으로 안정적인 손실 곡선을 보여준다.
동시에 Table 2는 다른 백본에 따라 FCOS 모델의 성능 변화를 보여준다.
단일 모델의 경우 ResNet-101 백본을 활용한 FCOS 모델이 Seen 집합에서 최고의 일반화 성능을, HRNet 백본 모델이 Unseen 집합에서 가장 좋은 성능을 보이는 것을 확인하였다.

-Semi-supervised learning.

DAVIS 데이터 세트에서 사전 훈련된 SiamMask 모델을 사용하여 고밀도 주석으로 FCOS 모델을 훈련하는 고밀도 레이블을 생성했다.
Table 2는 IoU > 0.5에서 FCOS 모델의 일반화 성능이 단일 객체 추적을 기반으로 하는 semi-supervised learning을 사용할 때 지속적으로 향상됨을 보여준다.

-Inhomogeneous backbone ensemble.

훈련된 검출기로부터 최상의 검출 성능을 달성하기 위해 공동 NMS 기술을 사용하여 앙상블을 수행했습니다. 여기에서 모든 검출기에서 얻은 최대 300개의 높은 예측 점수를 가진 bounding box에 NMS를 적용하여 앙상블을 달성할 수 있다.

-Visualizations.

inhomogeneous backbone ensemble의 효과를 확인하기 위해 각 모델의 추론 결과를 시각화했다.
그림 7은 다양한 백본을 가진 FCOS 모델의 추론 bounding box의 시각화를 보여준다.
그림 7에서 볼 수 있듯이 detector는 동일한 구조를 가지고 있지만 다른 백본은 추론 결과가

매우 다른 유형을 획득하기 위해 사용될 수있다.
이를 통해 앙상블 모델이 단일 모델에 비해 매우 큰 성능 향상을 달성할 수 있음을 확인하였다.

6. Conclusion
제안한 기법의 효용성을 검증하기 위해 Epic-Kitchens 객체 탐지 데이터셋을 사용하였으며, 제안된 semi-supervised learning은 단일 모델뿐만 아니라 앙상블에서도 좋은 성능을 보였다.
그러나 anchor-based 모델의 장단점에 대해서는 semi-supervised learning으로 좀 더 면밀히 분석할 필요가 있으며, pseudo label을 얻기 위해 간단한 규칙 기반 엔진을 사용한다는 한계가 있다.

위 paper에서는 기본적으로 영상에 대한 객체검출을 수행하기 때문에 이미지에서도 똑같이 적용될지는 의문이다.
추적이 frame 단위로 수행되는듯한 느낌이 들어서 만약 frame이 없는 이미지에도 똑같이 적용되지 않는다면 이것은 사용할 수 없는 방법이 된다.
또한 해당 팀이 사용한 방법이 이론적으로만 설명되있을뿐, 실습의 경우 우리가 직접 경험하며 습득하는 방법으로 진행하여 시간이 걸릴것으로 생각된다.
해당 팀의 방법을 사용하기 위해서는 추가적으로 detector FCOS 모델, semi-supervised learning의 SiamMask 모델 및 방법, ResNet-101 백본을 활용하는 방법 및 NMS기술을 사용하여 앙상블을 수행하는방법에 대해 학습이 필요할 것으로 보인다.
성능은 약 44.48로 낮아보이지만 챌린지에서 1등한 팀인만큼 해당 수치도 상당히 높다고 판단된다.