Patch-level Augmentation for Object Detection in Aerial Images

ICCV 2019 Workshop paper

Introduction

VisDrone dataset

- 움직이는 드론으로 촬영한 데이터셋
- 영상 속에 존재하는 objec의 크기가 다양함 (tiny / big)
- 영상의 각도가 다양함 (front / side / top view)
- 영상의 품질이 다양함 (motion blur)
- 영상 내 객체 종류 별 수가 불균형 함(class imbalance)
- 영상 내 복잡한 장면들이 많음 (crowd scene)

Introduction

Object Detection's issue : lack of dataset

- Class imbalance
- Hard examples
  - 네트워크가 어려워 할 수 있는 문제들 (pedestrian, person 등)

+

In Aerial view

- High-resolution Image and tiny objects
  - 이미지 사이즈는 크지만 객체는 작고 매우 많음
  - MS COCO 데이터셋과는 확연히 다른 특징을 가짐
- False Positive Rate
  - 배경이어야 하는데 fore-ground object로 오인식 하는 문제

Brief History

- 작고 다양한 크기의 object를 검출하기 위해
- Independent prediction layers of different resolutions [10, 14]
- Multi-scale feature extraction을 하는 network architecture – **FPN**
- 또한 foreground / background 의 class imbalance를 해결하기 위해 sampling이나 focal loss 등과 같은 방법이 등장함

- 최근에는 학습 단계에서 multi-scale object를 다루기 위해 'chip' 이라는 개념이 사용되고 있음
- 'chip-based training' : 전체 이미지가 아닌 object가 들어있는 sub-image(positive chip)만을 가지고 학습하는 것
- 하지만, 위의 경우 false-positive rate 가 증가할 위험이 생기므로 negative chip mining 이라는 것을 도입 – **SNIPER**

이 논문의 전신

- Scale Normalization for Image Pyramids(**SNIP**) , CVPR 2018 -> **SNIPER**, NIPS 2018
- Object Detection 문제에서 multi-scale을 얼마나 효율적으로 다룰 것인가에 대한 논문
- 이미지의 위치에 따라 크고 작은 객체들이 존재하는 상황
- 이미지 안에 GT들을 포함하는 영역들을 추출하고 이를 일정 크기로 조절하여 학습
- background 영역에서 찾아지는 object들에 대해 background label을 지어줘서 negative chip을 구성함

Figure 2: SNIPER negative chip selection. First row: the image and the ground-truth boxes. Bottom row: negative proposals not covered in positive chips (represented by red circles located at the center of each proposal for the clarity) and the generated negative chips based on the proposals (represented by orange rectangles).

이 논문은

- Chip-based training paper
- “Hard chip mining” 이라는 개념을 도입하며 이것이 이 논문의 핵심
- 그리고 foreground class 간의 class imbalance를 해결하고자 함

- **Foreground classes imbalance를 고려한 Hard chip-based training 을 제안**
  ‘Our core insight is to leverage object patches, which are misclassified by models trained from normal chips, as hard examples.’

이 논문의 방법

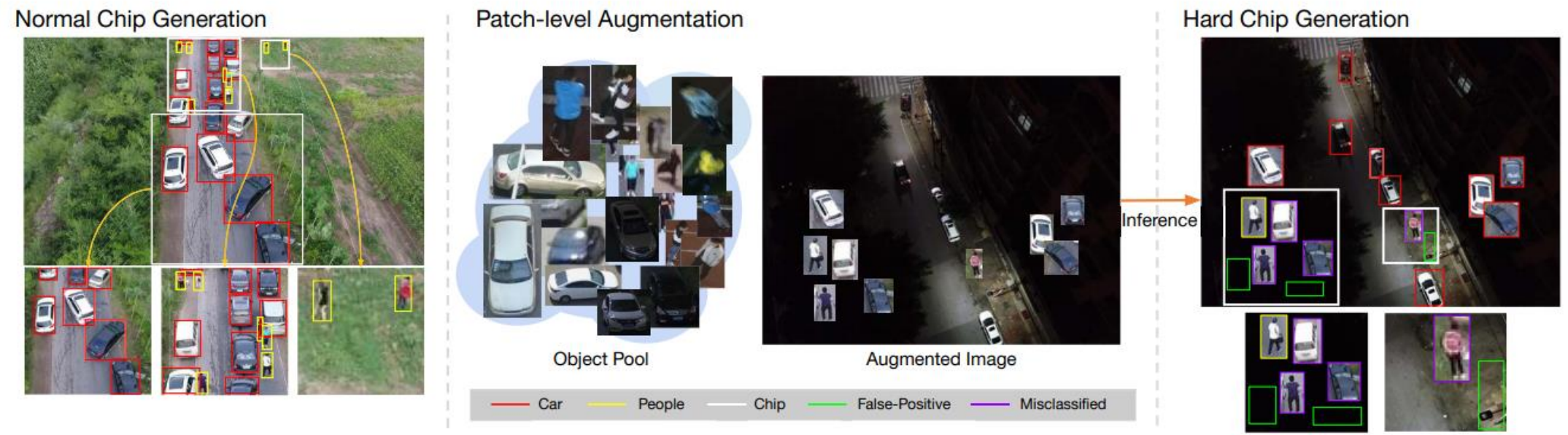

Figure 1. Overview of the proposed method. First, we train our detection model with positive (i.e., normal) chips. We then cut all object instances in the dataset and paste them into the existing randomly selected images considering class imbalance and geometric shape information. We also use an external dataset to reduce the false-positive rate and apply instance masking to resolve heterogeneity between datasets efficiently. Given patch-level augmented images, we make an inference using models trained on normal chips; and then generate hard chips containing misclassified instances. Finally, we retrain our detection model with hard chips as well as normal chips.

이 논문의 방법

1. **Network Architecture**
   - 1-stage / 2-stage architectures
   - 2018 VisDrone challenge : SSD / Faster R-CNN / R-FCN / FPN
   - **FPN-based Faster R-CNN in which the backbone is ResNet-101**
2. **Normal chip Mining**
   - Normal chip – ROIs containing ground truth instances in an image
   - Normal chip을 이용하여 1차 학습
3. **Hard Chip Mining**

이 논문의 방법

Figure 2. Patch-level augmented images. (top) augmented images from VisDrone-DET dataset (bottom) augmented images from DOTA dataset.

3. Hard Chip Mining

- Object 들을 patch 단위로 뽑아 object pool이란 곳에 다 모음
- Object pool에서 patch를 임의로 뽑아 이미지(canvas image)에 붙임
- Canvas image 로 외부 데이터셋 이미지도 활용함
  - 이 때 외부 데이터셋 이미지에 보여지는 conflicting classes object 들은 가림
- Normal chip으로 학습한 model을 이용하여 1차 inference 진행
- Misclassified foreground, background region 을 뽑아서 hard chip을 생성

실험

Method	AP[%]	AP_{50}[%]	AP_{75}[%]	AR_1[%]	AR_{10}[%]	AR_{100}[%]	AR_{500}[%]
Ours (multi-scale)	**37.15**	**65.54**	**36.56**	0.32	1.47	7.28	**53.78**
Ours (single-scale)	35.64	63.96	34.27	0.4	2.74	17.52	50.19
FPN	32.88	60.66	30.86	**0.43**	**2.77**	14.38	47.72
FPN (Default setting on MS COCO)	29.1	52.84	28.12	0.42	**2.77**	**26.41**	41.34

Table 1. Ablation studies on validation set of VisDrone2019-DET dataset.

Method	AP[%]	AP_{50}[%]	AP_{75}[%]	AR_1[%]	AR_{10}[%]	AR_{100}[%]	AR_{500}[%]
DPNet-ensemble	**29.62**	54.00	**28.70**	0.58	3.69	17.10	42.37
RRNet	29.13	**55.82**	27.23	**1.02**	**8.50**	**35.19**	46.05
Ours	29.13	54.07	27.38	0.32	1.48	9.46	44.53
S+D	28.59	50.97	28.29	0.50	3.38	15.95	42.72
BetterFPN	28.55	53.63	26.68	0.86	7.56	33.81	44.02
HRDet	28.39	54.53	26.06	0.11	0.94	12.95	43.34
CN-DhVaSa	27.83	50.73	26.77	0.00	0.18	7.78	**46.81**
SGE-cascade R-CNN	27.33	49.56	26.55	0.48	3.19	11.01	45.23
EHR-RetinaNet	26.46	48.34	25.38	0.87	7.87	32.06	38.42
CNAnet	26.35	47.98	25.45	0.94	7.69	32.98	42.28
CornerNet*	17.41	34.12	15.78	0.39	3.32	24.37	26.11
Light-RCNN*	16.53	32.78	15.13	0.35	3.16	23.09	25.07
FPN*	16.51	32.20	14.91	0.33	3.03	20.72	24.93
Cascade R-CNN*	16.09	31.91	15.01	0.28	2.79	21.37	28.43
DetNet59*	15.26	29.23	14.34	0.26	2.57	20.87	22.28
RefineDet*	14.90	28.76	14.08	0.24	2.41	18.13	25.69
RetinaNet*	11.81	21.37	11.62	0.21	1.21	5.31	19.29

Table 2. Top 10 comparisons results in the VisDrone-DET2019 challenge. * indicates that the baseline algorithm submitted by committee. More details can be found on the VisDrone homepage (http://aiskyeye.com/)