CCTV 환경에서의 Object Detection 을 위한 효율적인 데이터 설계 방안 연구

정화용 [1], 최정현 [2], 이상민 [3]
[1] 광운대학교 인공지능응용학과 석사과정
[2] 이노뎁㈜ 연구소 선임연구원
[3] 광운대학교 인공지능응용학과 교수

hwayong.jk, smlee5697@gmail.com

Efficient Data Design Approaches for Object Detection in CCTV

Hwa-Yong Jeong[1], Jeong-Hyun Choi[2], Sang-Min Lee[3]
[1]Dept. of Artificial Intelligence Applications, Kwangwoon University
[2]Dept. of New Innovation R&D Division, INNODEP INC.
[3]Dept. of Artificial Intelligence Applications, Kwangwoon University

요 약

최근 computer vision 기술 발달이 가속화되고 있으나, 특정 산업의 경우 산업 적용의 어려움과 데이터적 특성으로 인하여 기술 발전의 속도를 따라가지 못하고 있다. 특히, CCTV 는 대부분 실외 환경에 운영되어 다양한 환경의 변화 및 데이터 고유 특성상 노이즈가 많기 때문에 데이터 산포가 커서 기술의 현장 적용에 어려움이 있다. 본 논문에서는 CCTV 데이터의 특성을 고려하여 CCTV 운용 환경에 강건한 객체탐지(object detector) 학습을 위한 데이터 설계 방안을 제안한다. 제안 기법은 대용량의 CCTV 영상에서 객체탐지에 효과적인 샘플링을 유도하는 방안과 소수의 CCTV 레이블 데이터 외 MS COCO 등 다수 오픈 레이블 데이터를 혼합학습 하여 일반화 성능을 높이는 방안을 제안한다. 다수의 실험을 통해 제안 기법의 우수성을 입증하였으며, 특히 mAP 기준 13.39%의 성능 향상을 꾀할 수 있음을 선보였다.

1 서론

최근 이태원 참사, 연쇄 흉기 난동 사건과 같이 공공안전을 위협하는 비극적인 사건들이 발생함에 따라 CCTV 관제 관련 기술에 대한 수요가 급격히 증가하고 있다. 그러나 기존 많은 computer vision 기술들이 연구된 반면 실제 CCTV 관제 영역에 적용한 연구는 개인정보보호와 관련된 문제로 인해 연구이력이 많지 않으며, 데이터 고유 특성으로 인해 기존 연구를 즉시 적용하기에 많은 어려움이 존재한다[1].

(그림 1) CADP(Car Accident Detection and Prediction) Dataset – Carnegie Mellon University[2]

그러나 일반 공공(야외)에 설치된 CCTV 는 낮, 밤, 날씨, 네트워크와 같이 다양한 원인으로 인해 화질이 떨어지며, 다양한 제조사 및 설치 환경에 따라 CCTV 영상의 품질 산포도 매우 크다. 이러한 데이터로 학습할 경우 충분한 일반화 성능을 보장하지 못하는 상황이 발생한다.

이러한 한계점을 극복하고자 CCTV 데이터 특성을 고려하여 학습 모델을 효율적으로 학습할 수 있는 방안을 모색하고자 한다. 본 논문에서는 객체 탐지 학습에 효율적인 프레임 샘플링 기법과 데이터 혼합 방법을 제안하여 CCTV 환경에 강건한 객체탐지를 위한 데이터 설계 방안을 제시한다.

2 제안 방법

CCTV 영상 데이터는 공공에 설치된 CCTV 로부터 실시간으로 수집되는 영상으로 채널별로 렌즈와 카메라의 설정 및 상태에 따라 조도, 채도, 화각 등 다양성이 존재한다. 또한, CCTV 특성상 거리가 먼 미세객체(tiny object) 또는 객체 간 가림(occlusion)이 발생할 수 있으므로 다양한 상황에 강건한 AI 모델을 학습시키기 위한 전략이 필수적이다.

AI 모델이 좋은 성능을 내기 위해서는 양질의 데이터 설계가 선행되어야 하며, 현재까지 관련 연구에서는 문제 상황과 유사도가 높으며 학습 편향 방지를 위해 다양성을 확보, 가능한 많은 양의 데이터를 제공하여 모델의 정확도 성능 향상을 꾀하였다[3].

이를 CCTV 환경에 적용시키자면, CCTV 로부터 수집된 영상 데이터를 활용해야 하며, 다양한 객체가 등장해야 하고, 수집된 영상 데이터를 프레임(frame) 단위로 나눠 가능한 모든 프레임(frame)을 학습에 사용해야 한다. 하지만, 대다수 야외 운용중인 CCTV 는 고정된 위치에서 촬영되어 등장 객체의 종류가 한정적이며, 동일 채널에서 촬영된 영상 데이터는 유사한 정보를 담고 있어 데이터 편향이 불가피하다.

또한, CCTV 영상에서 촬영되는 객체는 전체 화면 대비 미세객체로 등장하는 경우가 많으므로, crop 된 해당 객체 이미지로부터 획득할 수 있는 정보가 제한적이다.

따라서, 이러한 2 가지 한계점을 해결하기 위해 영상 데이터로부터 객체탐지 학습모델 구축에 효율적인 프레임 샘플링 방법과, 관심 객체(사람, 차량, 등)의 특성을 충분히 표현학습(representation learning)할 수 있는 데이터 혼합(data fusion) 방안을 제안한다.

3 실험 설계

CCTV 관제시스템에서는 동시에 다채널을 처리하며 실시간에 가까운 추론능력을 요구하기 때문에 경량화된 모델을 선정해야 한다. 객체탐지 아키텍처에는 1-stage detector 및 2-stage detector 방식이 존재하는데, 여기에서는 추론속도를 고려하여 1-stage detector 를 선택하였으며 이중 GFLOPs 대비 가장 우수한 성능을 보이는 YOLO 계열 모델로 실험을 진행하였다. 데이터는 실제 CCTV 데이터(actual data)와, 공개된 MS COCO, BDD100K 를 활용했다.

실제 수집된 영상 데이터는 총 2,387 개의 CCTV 채널로부터 수집된 영상 데이터이며, 프레임 기준 총 2,132,457 장의 이미지로 구성되어 있다. 등장 객체 수는 약 22,589,583 개로 CCTV 에서의 주요 관심 객체인 사람, 일반차량, 사물의 비율이 약 73%를 차지한다. 데이터에 등장하는 개별 객체의 클래스 및 포함 비율은 아래 <표 1>과 같다.

<표 1> Actual Data class 별 분포

Class	Count	Ratio(%)
Person	3,159,985	13.40
Ground Vehicle	8,932,320	37.90
Accessory	5,733,100	24.30
Animal	62.378	0.3
Person Head	2,659,821	11.30
License Plate	3,041,979	12.90
Total	**23,589,583**	**100.00**

실험 진행에 앞서, 2,387 개의 채널 중 유사도가 높은 채널은 동일 채널로 병합하고, 동일 채널 내 밤/낮에 따라 다른 채널로 분할하는 작업을 통해 정지상태로 지속 등장하는 객체를 학습/검증/테스트 데이터셋에 모두 포함되어 성능이 높게 나오는 것을 방지하였다. 전처리 방식으로는 2,387 개 채널의 1 번째 이미지를 clustering 하는 방식을 통해 진행했다.

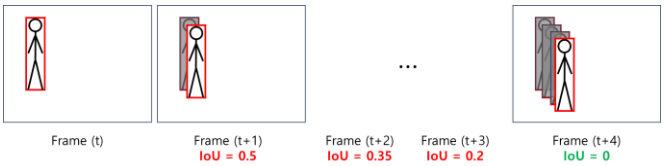
프레임 샘플링(frame sampling)과 데이터 혼합(data fusion)의 충분한 실험적 검증을 위해서 각 실험에서 총 4 개 데이터셋을 구축하였고, 2 가지 실험에 사용될 가장 기초적인 데이터셋은 프레임 샘플링(frame sampling)에서 제안한 3 가지 샘플링 방식 중 2 가지를 적용한 데이터를 활용해 실험을 진행했다.

이 때 성능비교 실험에 활용된 개별 모델은 실제 CCTV 영상 데이터를 기준으로 별도 테스트 데이터셋을 구축하여 공정하게 실험을 진행하였다.

3.1 프레임 샘플링(frame sampling) 방향성

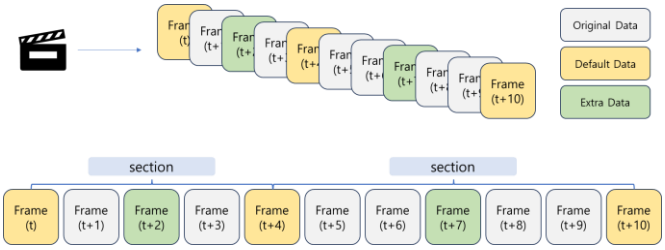
본 논문에서 정의하는 Sampling 은 전처리된 데이터셋을 활용해 2 가지의 고정된 조건과 1 가지의 비교조건을 적용하여 총 4 가지 데이터셋을 생성해 실험을 진행했다.

2 가지 고정 조건은 사전에 처리된 데이터를 활용해 CCTV 영상으로부터 탐지할 대상 객체가 최소 1 개 이상 등장하도록 설계했으며, 특정 프레임에서 등장한 객체가 이전 프레임 대비 IoU(Intersection of Union)가 0 인 프레임만 추출, 즉 객체의 겹침이 전혀 발생하지 않은 경우에만 사용하는 방식으로 기본 데이터셋을 구성했다. CCTV 의 경우 고정된 장면만 촬영하며, 객체탐지에서 관심 객체인 사람과 차량이 비교적 작은 크기로 등장하기 때문에 모델에게 중복된 정보는 최소한으로 학습시키며, 유의미한 정보를 학습시키기 위해 이러한 2 가지 고정 조건으로 데이터를 추출했다. 이 과정에서 연속된 프레임으로부터 고정되지 않은 간격(section)으로 프레임이 추출될 수 있었다.



(그림 2) IoU Thresholding 예시

또 다른 조건은 2 가지 고정 조건을 적용한 샘플링의 효과를 검증하기 위한 단계로, 총 4 가지 로직을 통해 4 가지 데이터셋을 구성했으며, 비교 실험을 통해 CCTV 영상에서 유효한 정보를 추출하기 위해 어떤 방식으로 데이터를 분할해야 하는지 실험했다. 2 가지 고정된 조건인 'IoU Thresholding'과 '객체 등장 여부'로 추출된 데이터의 프레임과 프레임 사이, 즉 데이터와 데이터 사이인 section 에서 추가 프레임을 추출해 section 으로부터 1 개, 2 개, 5 개의 추가 프레임을 추출한 데이터셋을 설계했다.



(그림 3) Sampling Case 2 예시

- S-Case1: Default Data
- S-Case2: Default Data + 1 additional frame from each section
- S-Case3: Default Data + 2 additional frames from each section
- S-Case4: Default Data + 5 additional frames from each section

3.2 데이터 혼합(data fusion) 방향성

Fusion 성능 검증을 위해서는 Default Data 와 함께 공개된 MS COCO[5]와 BDD100K[6] 데이터를 조합해 실험을 진행했다.

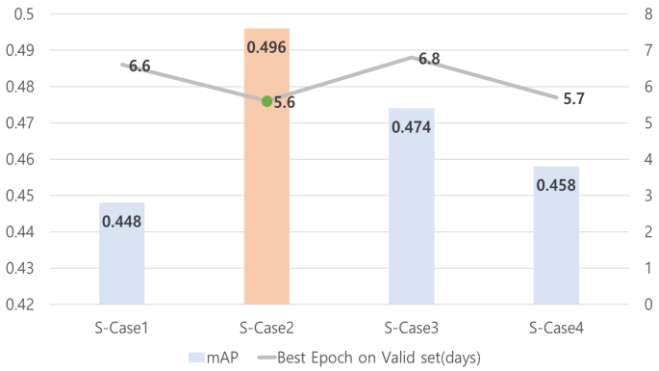
MS COCO 는 일반적인 화각에서 촬영된 80 개 클래스에 대한 각 118,000/5,000 장의 학습/검증 이미지로 구성된 데이터셋이며, BDD100K 는 자율주행 데이터로 차량의 화각에서 약 100 만 대의 차량과 13 만 명의 사람이 촬영된 데이터셋이다.

- F-Case1: Default Data_only
- F-Case2: Default Data + MS COCO
- F-Case3: Default Data + BDD100K
- F-Case4: Default Data + MS COCO + BDD100K

4 실험 결과

4.1 프레임 샘플링(frame sampling) 실험 결과

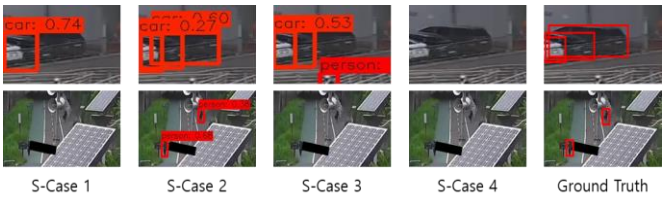
샘플링 실험 결과 S-Case2, 즉 전 프레임 대비 객체의 겹침이 없고 객체가 등장한 프레임으로만 구성되어 있는 데이터셋에서 추출된 각 프레임 사이 1 개 프레임을 추가로 추출한 경우가 학습 시간 및 성능 측면 모두에서 더 좋았음을 확인할 수 있었다.



(그림 4) Sampling Cases test results(mAP)

<표 2> Sampling Cases test results(class)

Case	Person	Bicycle	Car	Motor cycle	Bus	Truck
S-Case1	0.548	**0.166**	**0.663**	0.312	0.616	0.385
S-Case2	**0.568**	**0.170**	**0.691**	**0.484**	**0.634**	**0.431**
S-Case3	0.558	**0.166**	0.642	**0.414**	**0.630**	**0.432**
S-Case4	**0.565**	0.153	0.655	0.374	0.60	0.403



(그림 5) Sampling Cases test results visualize

IoU 와 객체 등장 여부를 고려하여 영상으로부터 데이터를 추출한 S-Case1 대비 S-Case2, 3 과 같이 추가적 프레임을 더 추출하여 학습 데이터를 구성할 필요성이 있음을 확인했다. 다만, S-Case3, S-Case4 에서 확인할 수 있는 것처럼 영상 데이터 즉 연속된 프레임은 매우 높은 유사도를 가지고 있어, 정보량과 내용 측면에서 동일한 이미지로 식별되기 때문에, 학습에 끼치는 영향이 적은 것을 의미한다.

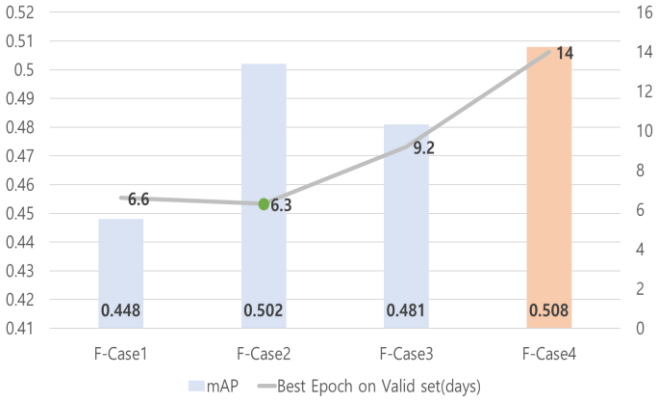
결론적으로 연속된 전체 프레임을 사용할 시 오히려 성능 저하가 발생할 수 있으며, 본 실험에서 고정 조건으로 적용한 'IoU Thresholding'과 '객체 등장 여부'의 효과를 확인할 수 있었다. 다만, S-Case1 보다 S-Case2 에서 성능 개선이 이루어진 것을 통해 추가적인 로직 설계를 통해 샘플링 시 유의미한 정보를 추가로 고민하는 방안에 대한 연구 필요성을 확인했다.

4.2 데이터 혼합(data fusion) 실험 결과

데이터 혼합 실험 결과 MS COCO 와 BDD100K 를 모두 활용한 F-Case4 가 가장 높은 mAP 를 달성했다. 비록,

F-Case2 와 비교 시 성능측면에서 0.006 의 mAP 개선을 이루었으나, 검증 데이터셋을 기준으로 최대 에폭(max epoch)에서 수렴까지 학습 소요 시간이 약 1.75 배로 BDD100K 데이터셋이 모델 성능에 미치는 영향보다 학습 소요 시간에 끼치는 영향이 더 크기 때문에 전체적인 효율 측면에서 떨어진다는 것을 확인했다.

또한, 2 가지 공개 데이터에 대한 비교 결과, 일반 이미지에 해당하는 MS COCO 는 성능 개선에 많은 영향을 주었으며, 학습 시간에는 큰 영향을 끼치지 않는 것을 확인할 수 있었으며, 차량의 블랙박스 화각에 해당하는 BDD100K 는 MS COCO 대비 비교적 CCTV 와 성격이 유사해 성능 개선에 큰 영향을 끼치지 못한 것을 확인할 수 있었다.

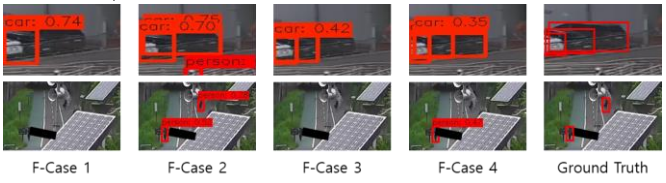


(그림 6) Fusion Cases test results(mAP)

<표 3> Fusion Cases test results(class)

Case	Person	Bicycle	Car	Motor cycle	Bus	Truck
F-Case1	0.548	0.166	0.663	0.312	0.616	0.385
F-Case2	**0.558**	**0.209**	0.682	**0.506**	0.596	**0.459**
F-Case3	0.540	**0.210**	**0.684**	0.414	0.597	0.439
F-Case4	**0.575**	0.180	**0.711**	**0.473**	**0.654**	**0.440**

(그림 7) Fusion Cases test results visualize



이를 통해 CCTV 영상에서 등장하는 객체는 화면상 매우 작은 사이즈로 등장에 클래스 식별에 영향을 미치는 정보를 충분히 습득하지 못할 수 있음을 확인했다. 특히 CCTV 영상에 대한 객체탐지 시 객체 종류에 따라 빈도수가 매우 상이한 클래스 불균형 문제가 발생할 수 있어 단순한 focal loss 만으로 이를 해결할 수 없으므로, 본 연구에서는 학습데이터를 추가하는 방향으로 두 가지 문제를 해결하고자 하였다. 비록 공개된 데이터셋을 혼합하여 활용하는 것이 데이터의 분포적 특징이 상이한 데이터를 활용하게 하여 모델의 수렴성을 저해할 수 있지만, 일반화 성능 향상을 꾀할 수 있다는 점을 실험을 통해 입증하였다.

5 결론

본 논문에서는 CCTV 환경에서 강건한 객체탐지를 위한 데이터 설계 방안을 제안하였다. 제안 방안으로는 CCTV 영상 데이터에서 유효한 정보 추출을 위한 샘플링 방식과 적은 CCTV 레이블 데이터와 공개된 객체탐지 학습데이터를 혼합 학습하여 학습모델의 작은 객체 탐지율(%) 향상 및 클래스 불균형에 따른 소수 클래스(minority class) 객체 인지율 향상을 꾀하였다.

이번 실험에서는 샘플링 방식에서 객체의 겹침 유무와, 객체의 등장 여부만을 고려하여 샘플링을 수행했기 때문에 향후에는 최적의 겹침 정도를 연구할 필요가 있으며, 겹침과 등장 외에도 다른 알고리즘적 제안을 통해 성능 개선이 있을 것으로 기대된다. 추가로, 데이터 혼합 시 MS COCO 와 BDD100K 외 다른 특성을 가진 데이터가 추가되었을 때의 비교 실험 진행이 필요하며, CCTV 데이터와 혼합 데이터의 적정 비율에 대한 비교 실험을 추가로 진행할 계획이다.

참고문헌

[1] Feng, Weitao, Deyi Ji, Yiru Wang, Shuorong Chang, Hansheng Ren, and Weihao Gan. “Challenges on Large Scale Surveillance Video Analysis.” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018.

[2] Shah, Ankit, Jean Baptiste Lamare, Tuan Nguyen Anh, and Alexander Hauptmann. “CADP: A Novel Dataset for CCTV Traffic Camera Based Accident Analysis.” arXiv.org, November 16, 2018.

[3] Zendel, Oliver, Katrin Honauer, Markus Murschitz, Martin Humenberger, and Gustavo Fernandez Dominguez. “Analyzing Computer Vision Data — the Good, the Bad and the Ugly.” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[4] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. “Yolov4: Optimal Speed and Accuracy of Object Detection.” arXiv.org, April 23, 2020.

[5] Lin, Tsung-Yi, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. “Microsoft Coco: Common Objects in Context.” arXiv.org, February 21, 2015.

[6] Yu, Fisher, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. “BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning.” Papers With Code. Accessed September 24, 2023.