비디오 객체 검출을 위한 이미지 증강 기법의 적용

노시동, 정기석*
한양대학교

sdroh1027@hanyang.ac.kr, *kchung@hanyang.ac.kr

Application of Image Augmentation Techniques for Video Object Detection

Roh, Si-Dong, Chung, Ki-Seok*
Hanyang University, Seoul, Korea

요 약

비디오 객체 검출은 시간상으로 연속된 이미지 시퀀스를 입력으로 받아 객체 검출을 수행하는 작업이다. 기존의 비디오 객체 검출 기법 연구는 대부분 random horizontal flip 등 비교적 단순한 변환만을 적용하였다. 그러나 단일 이미지에서의 객체 검출 연구에서는 성능 향상을 위해 보다 적극적으로 다양한 이미지 증강 기법을 사용하는 것이 일반적이며, 해당 기법들을 큰 변경 없이 비디오 객체 검출 태스크에 적용하는 것이 가능하다. 따라서 본 논문에서는 단일 이미지에서의 객체 검출에 주로 적용되는 이미지 증강 기법을 비디오 객체 검출의 학습 과정에 적용하고 성능 향상을 확인한다. 최신 비디오 객체 검출 모델인 DAFA 에 이미지 증강 기법을 적용 시, ImageNet VID 검증 데이터셋에서 85.8 mAP 의 성능을 보임을 확인하였다.

Ⅰ. 서 론

이미지 증강 기법은 학습용 이미지에 Crop, Flip, 색상의 변경 등을 가하여 의미론적 정보를 유지하면서도 새로운 학습 이미지를 생성하고 이를 학습함으로써 모델의 일반화 성능을 향상시킨다. 이에 따라 SSD [4] 등 대부분의 단일 이미지 객체 검출 모델은 이미지 증강을 적극적으로 적용하여 성능을 높인다.

그러나, MEGA [2], DAFA [3] 등 대부분의 최신 비디오 객체 검출 기법에서는 random horizontal flip 등 비교적 단순한 변환만을 적용한 증강 기법이 사용되어 왔다. 비디오에서의 객체 검출 작업은 시간상 연속된 일련의 이미지들을 입력으로 받는다는 특징을 가지나, 이미지 내 객체의 class 와 bounding box 를 예측한다는 점에서 단일 이미지에서의 객체 검출 과정과 기본적으로 동일하다. 따라서, 기존의 이미지 증강 기법을 비디오 객체 검출 작업에도 동일하게 적용하는 것이 가능하다. 따라서 본 논문에서는 최신 성능을 보이는 비디오 객체 검출 모델인 DAFA 의 학습과정에 이미지 증강 기법을 적용하여 성능 향상 여부를 확인하고 결과를 분석한다.

Ⅱ. 본 론

2.1. 비디오 객체 검출 알고리즘

비디오 객체 검출 모델은 현재 프레임의 정보를 기준으로 다수의 참조 프레임의 정보와의 연관성을 학습하고, 연관성이 높은 영역들의 정보를 융합하여 더욱 향상된 특징을 생성한 뒤, 이를 이용해 객체 검출을 진행한다. 따라서 비디오 객체 검출 모델에 이미지 증강을 적용할 때 현재 프레임과 연관성이 존재하는 다양한 참조 프레임을 생성할 경우 다양한 상황에 대한 학습이 가능하다.

비디오 객체 검출 모델은 주로 2-stage 방식의 단일 이미지 기반 객체 검출 알고리즘인 Faster RCNN [1]에 다양한 시공간 정보를 수집하기 위한 모듈과 이를 이용해 관심 영역의 특징을 강화하는 모듈을 추가하는 방식으로 구성한다. 최근에는 객체 특징(object features) 단위로 연관성을 계산하는 Attention 기반의 방식이 주로 연구되고 있으며, 대표적으로 Diversity-Aware Feature Aggregation (DAFA) [3]가 있다. DAFA 는 참조 프레임을 비디오 전체에서 임의로 선택된 global frame, 현재 프레임 근처에 존재하는 local frame 으로 구분하고, 각 프레임에서 추출된 객체 특징들을 각각 global 및 local memory 에 저장한다. 이후 각 메모리 내에서 관심 객체 특징과 연관성이 높은 후보 정보들을 융합하기 위해 attention 모듈을 이용한다. 이때, global 객체 특징 수집 시 단순히 임의의 프레임에서 추출된 모든 정보를 메모리에 저장할 경우 중복된 정보를 수집하게 되어 성능 향상이 제한될 수 있다. 따라서 DAFA 는 객체 특징 벡터들 간의 Euclidean 거리를 기반으로 global memory 를 업데이트 하는 방법을 제안하였다. 메모리에 포함된 객체 특징들의 집합 S 와 비교하여 새롭게 수집된 객체특징의 다양성 d 는 다음과 같이 정의한다.

$$d_{x,S} = \min_{y \in S} r_{x,y}$$

$r_{x,y}$ 는 객체 특징 벡터 x, y 사이의 Euclidean 거리를 의미한다. DAFA 는 업데이트 시점에 수집된 모든 특징 벡터 x에 대하여 다양성 $d_{x,S}$ 를 계산하고, 이 중 가장 큰 다양성 값을 가지는 x 를 메모리에 추가한다. 이러한 과정을 반복하여 최종적으로 큰 다양성을 가지는

참조 특징벡터 집합을 생성하며, 이를 이용해 관심영역의 특징벡터와의 attention 과정을 거침으로써 더욱 향상된 특징벡터를 생성한다.

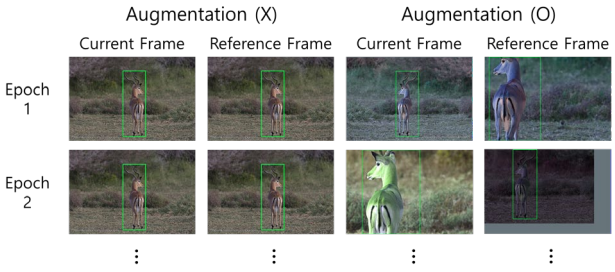

그림 1. 이미지 증강 기법의 적용 전/후 학습과정에서 생성된 현재 프레임 및 참조 프레임 비교

2.2. 이미지 증강 기법의 적용

이미지 증강 기법은 Photometric Distortion 과 Geometric Distortion 으로 구분된다. 먼저 Photometric 방법의 경우 객체의 형태는 유지되고 색상이 변경되는 특징을 가지므로, 적절한 변환 범위를 설정하여 성능을 향상시킬 수 있다. 본 실험에서는 먼저 입력된 RGB 포맷 이미지에 Brightness, Contrast 를 변환하고, HSV 포맷으로 변환한 뒤, Saturation, Hue 의 순서로 수정을 가했다. 그 뒤 다시 RGB 포맷으로 변환하고, 각 RGB 의 채널을 임의로 변경하는 Lightening Noise 를 적용했다. Geometric Distortion 의 경우 일반적인 Random Crop 대신 Random Expand & Crop 을 사용하였다. 해당 방법은 이미지를 축소한 뒤 잘라내므로 더욱 다양한 크기의 객체를 생성할 수 있다. 추가로 좌우 반전을 위해 Random Horizontal Flip 을 적용하였다.

다음으로 앞서 언급한 이미지 증강 기법을 비디오 객체 검출 모델 학습에 적용될 경우 기본적인 이점 외에 추가적으로 발생하는 이점은 다음과 같다. 일반적으로 강건한 모델 학습을 위해서는 비디오가 포함된 ImageNet VID 데이터셋과 단일 이미지 데이터 셋인 ImageNet DET 가 혼합 사용된다. ImageNet VID 의 경우 이미지가 비디오 단위로 존재하므로 현재 프레임에 대한 다수의 참조 프레임을 준비하는 것이 용이하다. 반면 ImageNet DET 는 각 샘플이 독립된 이미지이므로 참조 프레임을 샘플링 하는 것이 불가하여 현재 프레임 이미지를 복사하여 참조 프레임으로 사용한다. 이때 복사 대신 이미지 증강을 이용할 경우 그림 1 과 같이 현재 이미지와 유사한 참조 이미지를 적절하게 생성하는 것이 가능하다.

2.3. 실험 환경 및 실험 결과

실험 환경은 DAFA 와 유사한 설정으로 진행되었다. 실험을 위해 학습을 위한 3862 개의 비디오와 검증을 위한 555 개의 비디오로 구성된 ImageNet VID 를 사용한다. 그러나 ImageNet VID 만으로는 학습데이터가 충분하지 않기 때문에, ImageNet DET 데이터 셋을 추가로 혼합하였다. 이때, ImageNet DET 데이터 세트를 구성하는 200 개의 카테고리 중 ImageNet VID 와 겹치는 30 개의 카테고리를 사용한다. DAFA 에서 제시한 모델은 DAFA_F 와 DAFA_G 의 두가지가 있는데, DAFA_F 는 local 과 global memory 를 모두 활용하고, DAFA_G 는 global memory 만을 활용하는 모델이다. 본 실험에서는 이 중 DAFA_G 모델을 선택하였으며, Backbone 으로 ResNet-101 을 사용한다. SGD optimizer 와 0.001 의 learning rate 를 사용하여 훈련하였으며, 전체 학습의 3/2 지점에서 1/10 으로 LR decay 가 발생한다. 모델의 물체 감지 성능을 평가하는 지표로서 IoU > 0.5 의 임계값을 기준으로 하는 mean Average Precision (mAP)을 사용한다.

실험 결과는 표 1 과 같다. 이미지 증강을 하지 않을 경우 최고의 성능을 보이는 120,000 iteration 동안 학습하였으며, 이미지 증강을 적용할 경우 epoch 가 다를 때마다 다른 데이터로 학습이 가능한 장점을 활용하기 위해 기존보다 3 배 많은 iteration 동안 학습하였다. 이미지 증강을 적용하지 않았을 때, DAFA_G 는 [3]에서 보고된 바에 따르면 83.5 mAP 의 성능을 보였다. 반면 이미지 증강을 적용할 경우 기존보다 2.3 mAP 증가한 85.8 mAP 의 성능을 달성하였다. 이는 이미지 증강 시 학습이 진행됨에 따라 매번 다른 현재 및 참조 이미지를 얻게 되므로 모델이 보다 다양한 객체 특징 간의 관계를 학습하기 용이해지기 때문인 것으로 보인다.

표 1. ImageNet VID 검증 데이터셋에서 이미지 증강 여부에 따른 DAFA_G 모델의 성능 비교

모델	학습 iteration	mAP	증가
DAFA_G	120,000	83.5	-
DAFA_G w/ Augmentation	360,000	85.8	+ 2.3

Ⅲ. 결론

본 논문에서는 단일 이미지에서의 객체 검출에 주로 적용되는 이미지 증강 기법을 비디오 객체 검출 작업에 적용하고 성능 향상을 확인하였다. 비디오 객체 검출 모델인 DAFA_G 의 학습과정에 이미지 증강 기법을 적용시, ImageNet-VID 검증 데이터셋에서 85.8 mAP 의 성능을 보임을 확인하였다.

ACKNOWLEDGMENT

이 논문은 2020 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. 2020-0-01304, 모바일 자가 학습 가능 재귀 뉴럴 네트워크 프로세서 기술 개발).

참 고 문 헌

[1] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. “Faster rcnn: Towards real-time object detection with region proposal networks.” Advances in Neural Information Processing Systems, 2015.

[2] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. “Memory enhanced global-local aggregation for video object detection.” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[3] Si-Dong Roh, and Ki-Seok Chung. "DAFA: Diversity-Aware Feature Aggregation for Attention-Based Video Object Detection." IEEE Access, 2022.

[4] Liu, Wei, et al. "SSD: Single shot multibox detector." European conference on computer vision, 2016.