모션 정보와 spatio-temporal feature aggregation 을 활용한 비디오 물체 검출

김재겸, 고준호, 최준원
한양대학교

jkkim@spa.hanyang.ac.kr, jhkoh@spa.hanyang.ac.kr, junwchoi@hanyang.ac.kr

Video object detection using feature aggregation and motion information

Kim Jaekyum, Koh Junho, Jun Won Choi
Hanyang Univ.

요 약

본 논문은 비디오 이미지에서 모션 정보와 spatio-temporal feature aggregation 기법을 이용하여 물체 검출 기술을 제안하였다. 최근 딥러닝의 상당한 발전으로 물체 검출의 성능이 상당한 발전을 이루었다. 이러한 물체 검출 알고리즘들은 단일 이미지에 대해서만 처리를 하지만, 실제 환경에서는 비디오에서 처리된다. 본 논문에서는 이러한 비디오 정보를 활용하기 위해서 물체의 움직임 정보와 각 시간별 이미지에서 추출된 feature 를 합치는 기술을 제안하였다. 또한 시간 정보를 가지는 feature 를 활용하여 물체 검출 알고리즘의 성능을 높이는 방법을 고안해냈다. 첫 번째로 물체의 움직임 정보를 활용하기 위해서 서로 다른 시간대의 두 인접한 feature 를 correlation 연산을 사용하여 correlation feature 를 만들어 LSTM 의 입력으로 사용한다. 두 번째로 Gated attention 모듈을 사용하여 현재 feature 와 다른 시간의 feature 들을 합쳐주어 spatio-temporal feature 를 만들어준다. 이렇게 다른 특징을 가지는 두 feature 를 동시에 사용하여 시간정보를 이용하여 물체검출을 하여 성능을 높여준다.

Ⅰ. 서 론

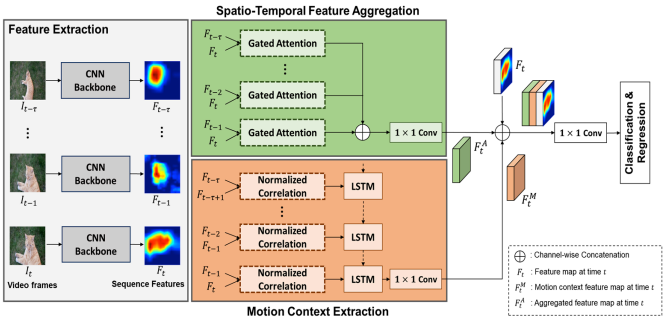
최근 딥러닝의 발달로 물체 검출의 성능이 상당한 발전을 이루었고 다양한 물체 검출 알고리즘들이 소개되었다 [1-6]. 특히 Convolutional Neural Network (CNN) 은 high-level feature 를 만들기 위해서 뼈대 네트워크로 널리 사용되고 있다. 이러한 물체 검출 알고리즘들은 단일 이미지를 활용하여 물체 검출을 시행하기 때문에 비디오 환경에서 시간 정보를 활용하지 않는다. 비디오 데이터셋은 물체의 움직임으로 인한 motion blur, camera defocusing 등의 문제를 가지고 있다. 이러한 문제들을 해결하기 위해서는 feature map 에 물체의 움직임 정보를 포함시켜야 한다. 이러한 문제들을 다루기 위해서 본 논문에서는 비디오 물체 검출 네트워크를 제시하였다.

Ⅱ. 본론

1) 제안하는 네트워크 구조

본 논문에서는 그림 1 과 같이 CNN 을 활용한 비디오 물체 검출기를 제안하였다. 제안하는 구조는 비디오에서의 연속된 이미지를 CNN 으로 각각 처리하며, 특징 지도를 강화 하기 위하여 두 가지 방법을 제안하였다.

첫 번째는 spatio-temporal feature aggregation (STFA) 방식으로 현재 feature 와 다른 시간대의 feature 들을 gated attention 방식을 이용하여 각각 합쳐주는 module 을 소개하고 있다. 이렇게 feature aggregation 을 하면 motion blur 가 존재하여 물체검출이 잘 되지 않은 상황에 대해서는 더 좋은 feature 를 추출할 수 있지만, 실제 물체의 움직임을 직접 사용하지는 않았다.



두 번째 방식은 motion context extractor (MCE)로 실제 움직임을 직접 추출할 수 있는 correlation operation 을 사용하여 첫 번째 방식의 단점을 보완해주는 방식이다. 두 인접한 frame 에서 추출된 feature 들을 correlation operation 으로 연산하여 LSTM 의 입력으로 사용한다. 이렇게 추출된 feature 는 물체의 움직임 정보를 가지고 있지만, high level feature 는 아니기 때문에 Gated attention 으로 합쳐진 feature 와 현재 feature 들을 동시에 사용하여 최종 물체 검출에 사용한다.

2) 실험 및 데이터 셋

본 논문에서 실험은 ImageNet [12] VID 데이터를 사용하였다. 우리는 DET 데이터를 같이 사용했으며 DET 데이터는 VID 와는 다르게 단일 이미지 물체 검출에 이용되는 데이터이다. DET 데이터는 200 개의 class 를 가지며 VID 데이터가 가지는 30 개의 class 를 포함하고 있다. 우리가 제안한 네트워크를 실험하기 위해서 우선은 DET 데이터를 사용하여 pretraining 을 하였으며, DET 에서

VID 데이터가 가지는 30 개의 class 를 포함하는 데이터와 VID 를 동시에 사용하여 최종 실험을 진행한다.

표 1 에서와 같이 우리는 3 가지 방식을 나누어 하나씩 추가하며 실험을 진행하였다. 첫 번째는 STFA 만 추가하여 진행한 실험으로 2.34% 향상하였다. MCE 만을 추가 하였을 때는 2.3% 향상하였지만, 두 가지 모두 추가하였을 때는 4.33% 향상하는 것을 볼 수 있다. 그림 4 와 같이 물체의 feature map 은 STFA 만 사용하였을 때 occlusion 이 되는 상황에 대해서는 한계가 있었지만, MCE 를 같이 사용하였을 때는 occlusion 에도 강력한 모습을 보였다.

	mAP (%) (Slow)	mAP (%) (Medium)	mAP (%) (Fast)	mAP (%)
SSD	77.45	64.38	41.52	66.70
SSD + STFA	78.91	67.54	44.34	69.04
SSD + MCE	80.27	66.36	43.18	69.00
Proposed	82.08	67.99	46.13	71.03

Ⅲ. 결론

이 논문에서 우리는 spatio-temporal feature aggregation 과 motion context 추출을 사용하여 새로운 one-stage 비디오 객체 검출기를 개발하였다. Attention 모듈을 사용하여 시공간 측면에서 feature 를 aggregation 하고 correlation 연산으로 global motion feature 를 추출하였다. 우리의 방법은 end-to-end 학습 framework 이며 모든 one-stage 검출기에 적용 할 수 있다. 또한, 우리는 우리의 방법이 물체의 움직임 및 가중치 map 을 추출하는 데 효율적으로 작동함을 보여주었다. SSD 베이스 라인 검출기가 있는 ImageNet VID 데이터 세트에 대해 실험을 수행하였다. 우리의 방법은 기준에 비해 막대한 성능 향상을 보이며 기존의 one stage 비디오 물체 검출 방법보다 성능이 뛰어남을 증명하였다..

.

ACKNOWLEDGMENT

이 논문은 2020 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (2016-0-00564, 사용자의 의도와 맥락을 이해하는 지능형 인터랙션 기술 연구개발)

참 고 문 헌

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, " Rich feature hierarchies for accurate object detection and semantic segmenatation," in CVPR (2014).

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, " Ssd: Single shot multi box detector," in ECCV (2016).