깊이 예측 백본 네트워크 모델을 사용한 새로운 딥러닝 기반 객체 검출 방법

신은섭 표승우 배성호

경희대학교

Kairos9603@khu.ac.kr, ocn54321@hanmail.net, shbae@khu.ac.kr

A New Deep Learning based Object Detection Method using a Deep Estimation Backbone Network Model

Eunseop Shin, Seoungwoo Pyo, Sung-Ho Bae

Kyung-Hee University

요 약

딥러닝은 객체 검출(object detection)에 활발이 적용되었으며, 괄목할 만한 검출 성능을 보이고 있다. 그러나 기존 딥러닝 기반 객체 검출 방법은 물체가 배경 텍스쳐와 유사하거나, 물체의 크기가 작을 경우 검출율이 낮아지는 문제가 존재했다. 본 논문에서는 원본 영상에 대한 깊이 맵 정보를 객체 검출에 추가로 활용하여 위의 문제를 해결한다. 구체적으로, 기존 깊이 맵 예측 딥러닝 모델을 도입하여 원본 영상으로부터 예측된 깊이 맵을 획득하고, 획득된 깊이 맵을 원본 영상과 함께 객체 검출 방법에 입력으로 사용하여 깊이 맵 정보가 객체 검출 방법에 추가로 활용되도록 딥러닝 모델을 학습했다. 실험 결과, 제안 방법이 기존 객체 검출 방법보다 향상된 검출 성능을 보였다. 결론적으로 본 논문은 깊이 맵이 객체와 배경을 선명하게 구분해 주며 깊이 맵에서 크기가 작은 객체도 배경과 선명하게 구분되는 경우가 많기 때문에, 깊이 맵이 영상 검출 딥러닝 방법에 효과적으로 활용될 수 있다는 점을 발견하였다.

1. 서론

객체 검출은 컴퓨터비전 분야의 가장 전통적인 응용들 중 하나로, 최근 딥러닝 기술을 통해 괄목할 만한 인식 성능을 보였다[10, 14]. 일반적으로, 딥러닝을 통해 학습된 모델이 물체를 인식할 때, 명확하게 구별이 되고, 크기가 큰 물체에 대해서는 딥러닝 모델이 높은 성능을 발휘한다. 하지만 크기가 작거나 배경과 텍스쳐가 비슷한 물체에 대한 인식률은 상대적으로 낮다.

기존 단일 RGB영상 입력을 기반으로 하는 방법들은 물체를 검출할 때 작거나 색이 비슷한 물체에 대해 상대적으로 낮은 검출율을 보이는 문제점이 있었다[10, 11, 12, 14, 15, 16]. 본 논문은, 깊이 정보가 객체 검출에 있어서 중요한 정보를 포함한다는 사실에 기반해서, 깊이 정보를 활용한 객체 검출 딥러닝 모델을 제안한다.

단일 이미지는 RGB값 정보를 가지고 있지만 깊이감에 관한 정보는 가지고 있지 않다. 하지만 최근 연구 결과는 딥러닝을 이용해 RGB 영상으로부터 깊이 맵을 효과적으로 유추할 수 있음을 보였다[7]. 본 논문은 이에 착안하여, 객체 검출을 위한 딥러닝 모델의 입력으로 원본 RGB영상과 유추된 깊이 맵을 사용하여 모델을 학습함으로써 깊이 맵이 제공하는 추가적인 정보를 통해 객체 인식률을 향상시킨다.

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학 사업의 연구결과로 수행되었음"(2017-0-00093)”

2. 관련연구

2.1. Object Detection

딥러닝을 이용한 객체 검출 방법은 모델의 구조에 따라 크게 두가지로 구분할 수 있다.

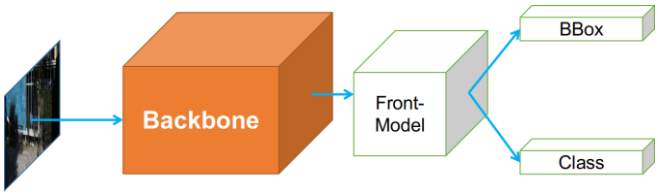
첫 번째 방법은 Two-Stage 모델로, 먼저 영상에서 물체가 있을 만한 부분(Region of Interest, RoI)을 먼저 검출한 후, 그 부분의 물체가 무엇인지 구분하는 방법이다[10, 11, 12, 13]. 이 방법은 딥러닝을 활용한 객체 검출 초기에 생겼던 방법이며, 대표적인 모델로 R-CNN 계열의 R-CNN[10], fast R-CNN[11], faster R-CNN[12], Mask R-CNN[13]등이 있다.

두 번째 방법은 One-Stage 모델로, 영상에서 물체가 있을 부분과 그 물체가 무엇일 지를 동시에 예측하는 방식이다[14,15,16,17]. 이 방법은 Two-Stage 모델보다 속도가 빠르다는 장점이 있으나, 정확도는 조금 떨어졌는데 최근에는 정확도로 향상시킨 모델이 나오면서 주류가 되어 가고 있다[17].

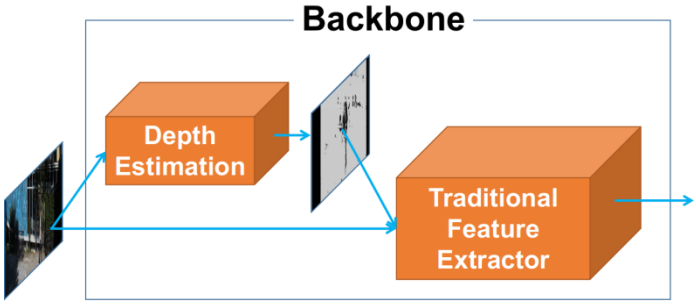
객체 검출의 성능하락의 원인은 주로 작은 물체와, 색이 비슷하거나 가려진 물체, 비슷한 물체에 대해 객체 검출 네트워크가 객체를 잘 판별하지 못하기 때문이다. 이 중 작은 물체를 잘 인식하기 위해, 최근에 멀티 스케일을 사용한 모델이 많이 등장하고 있다[17].

2.2. 깊이 예측

기존의 추정된 깊이 맵의 경우 왜곡되고 객체의 경계가 모호하게 구성이 된다. 더 높은 해상도를 갖는 깊이 맵을 추정하기 위해 최근 모델은 다음 두 가지의



[그림1] 최근 Object Detection 모델의 일반적인 구조. Classification 과업에서 좋은 성능을 보인 ResNet[9], DenseNet[8]을 Feature Extractor(backbone)로 사용한다.



[그림2] 본 논문에서 제안하는 Backbone네트워크의 구조. RGB이미지를 이용하여 Depth Estimation을 수행한 뒤 기존의 Feature Extractor에 RGB와 같이 넘겨준다.

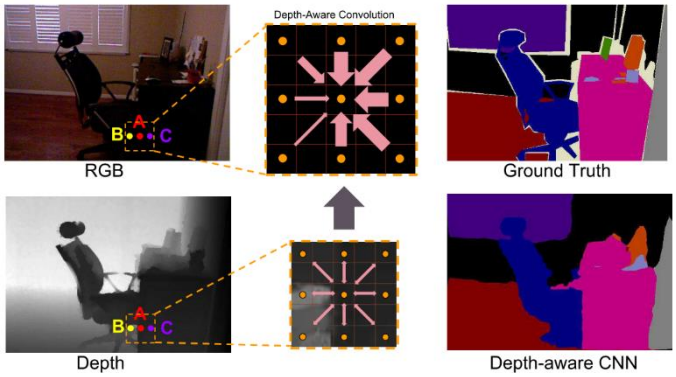
방법을 이용하였다. 첫번째는 서로 다른 스케일로 추출된 특징을 융합하는 방법이며, 두번째는 모델 학습 시 사용하는 손실 함수를 개선하는 것이다.

깊이, 기울기, 표면의 법선의 오차를 각각 측정하는 세 가지 손실 조건이 보완 방식의 정확도 향상에 도움이 된다. 서로 다른 크기에서 추출된 특징을 인코더, 디코더, 멀티 스케일 융합 모듈, 정제 모듈로 구성된 네트워크 구조를 이용하여 깊이 맵을 예측한다[1].

CRF(Conditional Random Field)기법이 객체 검출을 위해 활발히 활용되어 왔다[6]. 왼쪽-오른쪽의 일관성을 갖는 통제되지 않은 하나의 이미지의 깊이를 추정하는 방법이다. 이때 두 개의 이미지를 사용한다. 한쪽의 이미지 뷰가 주어지면 다른 쪽의 이미지와의 불일치 맵을 재구성하여 깊이 맵을 구성한다. 외관의 매칭 loss, 불균형 평탄도 loss 및 좌우 불일치 일관성 loss 이 방법은 교육 도중 깊이 맵을 입력할 필요가 없어 완전히 통제되지 않는다[2].

두 개의 심층 네트워크 스택을 사용하여 작업을 처리한다. 하나는 전체 이미지를 기반으로 한 거시적인 예측(global view), 다른 하나는 전체적인 이미지를 기반으로 예측된 정보를 세부적(local view)으로 나누어 여러 개의 스케일에 대해 정리한다[3]. CNN으로 교육할 때 가장 자주 사용되는 loss 함수는 L2 loss 함수이다.

하지만 depth를 예측하는데 L2 loss함수는 충분한 성능을 발휘하지 못한다. 여기서 gradient based loss 함수를 사용한다. 이를 통해 서로 다른 맵에서의 경사도의 차이를 줄일 수 있다. 또 추가적으로 Normalized Cross Correlation(NCC) based loss함수를 이용한다. Gradient based loss 함수와 NCC based loss함수를 더한 새로운 loss함수를 사용한다면 정확도를 향상시킬 수 있다[4].



[그림3] RGB 이미지에서는 구분이 안가는 물체의 경계를 깊이 맵을 이용해서 구분할 수 있다[3].

2.3. 깊이 맵을 사용한 객체 검출

객체 검출에 깊이 맵을 이용하는 다양한 시도가 있었다.

논문 S. Hou et al.[1]에서는 RGB와 Depth 이미지를 이용해서 넣어 색상, 경계, 좌우 영상차이, 높이, 각도를 추출, 이를 새로운 입력으로 넣는 방식을 사용하였다. 하지만 이 방법은 직접적인 깊이 맵이 없는 경우 사용할 수 없는 단점이 있다.

Cao et al. [2] 에서는 이 단점을 보완하여 RGB영상만을 입력 받아 직접적으로 Depth이미지를 생성한다. RGB영상과 생성된 깊이 맵을 구조가 동일한 두개의 네트워크에 각각 넣어 객체 검출을 수행한다. 그러나 이 방법은 두개의 네트워크를 각각 학습시켜야 해서 학습시간이 오래 걸리고 모델이 무겁다는 단점이 있다.

3. 제안하는 아이디어

기존의 객체 검출 모델들은 깊이 정보를 고려하지 않은 모델이 대부분이었다. 이러한 논문들은 Feature를 뽑아 내기 위해 단순히 객체 분류에서 높은 성능을 보인 ResNet[9], DenseNet[8] 등을 Backbone으로 사용하였다. 본 논문에서는 Feature Extractor로서 사용되는 Backbone 네트워크를, 기존 RGB 영상뿐만 아니라 깊이 맵 정보도 함께 입력으로 받아 학습시킨 네트워크로 대체한다. 본 논문은 Depth 정보를 함께 고려하여, 객체 검출의 성능을 향상시키는 방법을 제안한다.

3.1. Geometric 정보

깊이 맵은 RGB 데이터에는 없는 새로운 Geometric 정보를 가지며, 이 정보는 RGB색상 정보에서는 유추할 수 없다. Geometric 정보는 특히 물체의 경계부분을 구분하는데 중요한 정보로 사용된다. [그림 3]은 RGB에서 구분할 수 없는 경계를, 깊이 맵을 이용해서 구분한 그림이다.

3.2. 제안 모델

본 논문에서는 모든 모델에 적용 가능한 깊이 특징을 추출하는 네트워크를 제안한다. [그림 2]은 본 논문에서 제안하는 네트워크의 전체 구조이다. Backbone 네트워크는 Depth Estimator와 Feature Extractor 두 부분으로 구성된다.

Depth Estimator는 입력으로 받은 RGB이미지를 이용하여 Depth를 예측하는 부분이다. 입력으로 들어온

[표1] Baseline 모델과 제안하는 모델의 실험 결과

	Total Loss	Regression Loss	Classification Loss	mAP
Baseline	0.3071	0.2720	0.0351	**0.4036**
Ours	**0.1002**	**0.0973**	**0.0029**	0.4006

RGB 이미지와 Depth Estimator에서 예측된 깊이 맵을 합쳐서 Feature Extractor의 입력으로 넘긴다. Feature Extractor는 RGBD이미지를 입력 받아 유용한 특징을 생성해내는 모델이다. 제안 모델의 configuration은 4. 실험 Section에 상세히 기술하였다.

4. 실험

4.1. 모델 Configuration

실험에 사용한 모델은 다음과 같이 구성하였다. Depth Estimator로 KITTI Dataset에서 좋은 성능을 보인 Gordard et al.[5]의 모델을 사용하였고, Feature Extractor 모델은 DenseNet을 마지막으로 Front-Model은 RetinaNet을 사용하였다.

4.2. 데이터 셋

KITTI 객체 검출 데이터 셋에서 Left RGB이미지와 Bounding Box와 Class가 기록된 Annotation데이터를 사용하였고, 추가적으로 깊이 맵을 만들기 위해 Right RGB 이미지를 사용하였다. KITTI 객체 검출 task의 데이터 셋에는 직접적인 깊이 맵을 제공해 주지 않는다. 때문에 Left이미지와 Right이미지를 입력으로 넣어 OpenCV를 이용하여 Disparity를 구하여 깊이 맵 Label로 사용하였다.

4.3. 결과

[표1]은 실험 결과를 보인다. 여기서 Loss는 Regression Loss와 Classification Loss를 합친 Loss이다. Classification Loss는 RoI의 물체의 Class를 잘 맞추었는지에 대한 Loss이고, Regression Loss는 Bbox를 얼마나 잘 맞췄는지에 대한 Loss이다. 마지막으로 mAP는 Mean Average Precision으로 예측한 박스가 실제박스와 얼마나 일치하는지에 대한 mAP를 나타낸다. Baseline은 본 논문에서 Front Model로 사용하였던 RetinaNet의 원본 모델이다.

[표1]에서 보이듯이, 제안 방법이 기존 방법보다 Loss가 감소하였음을 볼 수 있다. mAP에서는 제안하는 방법이 기존 방법보다 차이가 거의 없음을 볼 수 있다.

5. 결론

본 논문에서는 Feature Extractor로서 사용되고 있는 백본 네트워크의 새로운 구조를 제안하였다. 백본 네트워크 내부에서 깊이 맵을 예측하여 깊이 맵과 관련 깊은 새로운 특징을 추출하여 객체 검출에서 좋은 성능을 보였다.

참 고 문 헌

[1] S. Hou et al., “Object detection via deeply exploiting depth information”, Neurocomputing, Volume 286, Pages 58-66, 2018.

[2] Y. Cao et al., "Exploiting Depth From Single Monocular Images for Object Detection and Semantic Segmentation," in IEEE Transactions on Image Processing, vol. 26, no. 2, pp. 836-846, Feb. 2017.

[3] L. Porzi et al., "Depth-aware convolutional neural networks for accurate 3D pose estimation in RGB-D images," 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages. 5777-5783, 2017.

[4] R. Mukhometzianov and J. Carrillo. “Capsnet comparative performance evaluation for image classification.”, arXiv preprintarXiv:1805.11195, 2018.

[5] C. Godard et al., "Unsupervised Monocular Depth Estimation with Left-Right Consistency," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages. 6602-6611, 2017.

[6] D. Xu et al., “Monocular depth estimation using multi-scale continuous crfs as sequential deep networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018.

[7] Koch, Tobias, et al. "Evaluation of CNN-based single-image depth estimation methods." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[8] G. Huang et al., "Densely Connected Convolutional Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, pages 2261-2269, 2017.

[9] K. He et al., "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770-778, 2016.

[10] R. Girshick et al., "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 580-587, 2014.

[11] R. Girshick, "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1440-1448, 2015.

[12] Ren, S et al. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39, 1137-1149, 2015.

[13] K. He et al., "Mask R-CNN," 2017 IEEE International Conference on Computer Vision (ICCV), pages. 2980-2988, 2017.

[14] J. Redmon et al., "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages. 779-788, 2016.

[15] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages. 6517-6525, 2017

[16] Liu, W et al., “Ssd: Single shot multibox detector”, European conference on computer vision, pages. 21-37, 2016.

[17] Lin, T. Y et al., “Feature pyramid networks for object detection”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages. 2117-2125, 2017.