

Faster R-CNN의 중심 Convolution Neural Network 모델이 훈련 및 추론에 미치는 영향에 대한 분석

신수진, 박준건, 김윤영, 장준호, 문일철
한국과학기술원 산업및시스템공학과

Abstract

객체 탐지는 많은 인공지능 연구자들의 이목을 집중시키는 컴퓨터 비전 응용 기술이다. 로봇, 자율 주행, 검색 시스템 등의 영역에서 활용될 수 있는 주요한 기술로서, 주어진 이미지 데이터에서 훈련 대상이 된 객체를 탐지하는 기술이다. 최근 딥 러닝 기반 모델의 등장으로 객체 탐지에 특화된 다양한 모델들이 제안되었고, 그 중 우리는 Faster R-CNN이라는 모델을 분석 대상으로 선정하였다. Faster R-CNN은 기존의 R-CNN이라고 하는 초기 모델을 가속화하고 외부 툴에의 의존성을 제거하여 개량된 최신 모델이다. 본 논문에서는 Faster R-CNN의 중심 convolution neural network로서 사용할 수 있는 AlexNet과 ResNet을 이용해 훈련한 결과를 비교 분석하고, 뿐만 아니라 훈련에 미치는 다양한 요인들, 예컨대 네트워크 구조 및 크기 등에 관한 이슈에 대해 논의하고자 한다.

1. 서론

Faster Region-based Convolutional Neural Network (Faster R-CNN) [1]은 객체 탐지 과업을 수행하는 딥러닝 기반 최신 모델 중 하나이다. 초기에 관련 모델로 Region-based Convolutional Neural Network (R-CNN) [2]이 제안되었으나, 이 모델은 훈련 및 추론 속도가 매우 느리고 외부 툴에의 의존성이 심해서 multi-stage pipeline 구조를 띄는 단점이 있었다. 이후 이를 개선한 Fast Region-based Convolutional Neural Network (Fast R-CNN) [3] 모델이 제안되었고 이 모델은 훈련 및 추론 속도가 가속화되었을 뿐만 아니라 외부 툴에의 의존성을 감소시켜 two-stage pipeline을 구

현하였다. 이 Fast R-CNN을 한번 더 개선한 것이 본 논문에서 다루게 될 Faster R-CNN이다. Faster R-CNN은 외부 툴을 별도로 필요로 하지 않고 단독으로 설계된 네트워크로 구동이 가능하며 (single pipeline), 실시간 추론을 목표로 제안되었기 때문에 매우 빠른 추론이 가능하다.

Faster R-CNN은 중심에 위치한 Convolution Neural Network (CNN)와 이 CNN의 output인 final feature map을 공유하는 두 개의 sub-network로 구성된다. 두 sub-network는 각각 Region Proposal Network (RPN)과 Object Detector에 해당하며, 전자는 입력 이미지에 대해 후보 region을 제안하는 네트워크이고 후자는 제안된 region들 각각에 대해 어떤 object인지 탐지하는 네트워크이다. 본 논문은 Faster R-CNN의 중심부 CNN을 AlexNet [4], ResNet [5]이라고 하는 저명한 두 모델을 사용해보고, 이 중심부 CNN이 훈련 및 추론에 미치는 영향에 대해 분석해보고자 한다. 이와 더불어 네트워크 구조 및 크기 등에 대한 이슈에 대해 논의할 것이다.

2. Faster R-CNN

Faster R-CNN은 크게 세 가지 모듈로 구성되는데 그 구조를 요약하면 다음 그림 1과 같다.

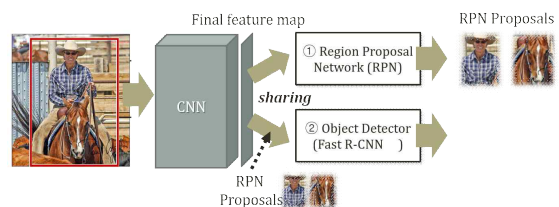


그림 1. Faster R-CNN

Faster R-CNN의 첫 번째 모듈은 CNN으로, 입력 이미지의 중요한 정보가 압축된 feature map을 계산한다. 두 번째 모듈은 Region Proposal Network (RPN)으로, 입력 이미지에서 object가 있을만한 region을 제안하고 마지막 모듈은 Fast R-CNN으로, 제안된 각 region에 대해 어떤 object일지 판별한다.

2.1 Region Proposal Network

RPN은 입력 이미지 상에서 object가 있을만한 구역을 region으로서 제안하는 sub-network이다. RPN을 훈련시키기 위해서는 다음 과정이 필요하다: 1) anchor라고 하는 candidate region에 대해 scale과 aspect ratio를 규명한다. 여기서 scale이란 region의 넓이를, aspect ratio는 종횡 비율을 의미한다. 예를 들어, scale을 128*128, 256*256, 512*512, aspect ratio를 1:1, 2:1, 1:2로 지정했을 때 총 9종류의 다양한 anchor가 정의될 수 있다. 2) 정의된 각 anchor에 대해 supervision을 부여한다. 즉, 각 anchor별로 가장 가까운 groundtruth를 찾고 그 groundtruth와 일정 비율 이상 겹치면 positive, 일정 비율 이하 겹치면 negative로 labeling을 한다.

RPN은 그림 2와 같은 구조로 구성된 네트워크이며, 앞서 구축한 anchor 데이터셋으로 훈련이 이루어진다. RPN은 CNN에서 생성한 feature map 위에서 $n \times n$ 의 window가 sliding하면서 convolution 연산을 수행하고 해당 중심에 위치한 anchor들에 대해서 object가 있는지 없는지 여부를 판단하고 anchor의 세부 모양을 보정한다.

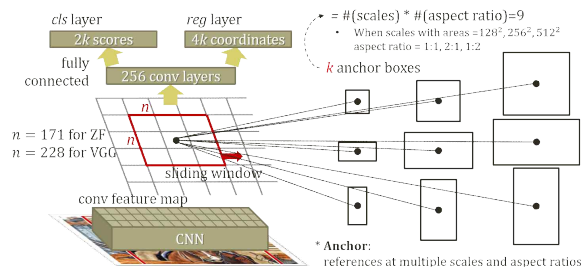


그림 2. Region Proposal Network

2.2 Object Detector

Faster R-CNN은 object detector로 기존의 Fast R-CNN [3] 모델을 적용했다. 앞서 RPN에서 제안된 각 region에 대해서 Fast R-CNN은 어떤 object인지 판단하는 역할을 수행한다. Fast R-CNN의 구조는 개략적으로 다음과 같다: 1) 중심부 CNN으로부터 final feature map을 얻는다, 2) RPN으로부터 제안된 크기가 각기 다른 region을 RoI_Pooling layer를 통과하여 동일한 크기의 region으로 pooling한다, 3) 두 fully connected layer를 거친다, 4) 다른 두 개의 branch에서 각각 또 한번의 fully connected layer를 통과하여 어떤 object가 있는지를 판단하고 해당 region의 세부 모양을 보정한다.

3. Convolution Neural Network

3.1 AlexNet

AlexNet [4]은 2012년 Krizhevsky *et al.*에 의해서 제안된 네트워크로, CNN을 활용해 이미지 분류 (Image Classification) 과업을 수행한 초기 모델이라 할 수 있다. Pooling layer, normalization layer를 제외한 architecture는 5개의 convolution layer와 3개의 fully connected layer로 총 8개의 layer가 쌓여 있는 형태를 띤다. 그 동안 깊은 신경망은 훈련 관련 overfitting 문제를 안고 있었고, 이를 극복하기 위해 generalization 성능을 높이기 위한 기술이 필요로 했다. 본 논문은 처음으로 Rectified Linear Unit (ReLU) 라고 하는 연산을 도입하여 gradient vanishing 문제를 완화하고 dropout 연산을 도입함으로써 신경망 훈련의 혁신을 이루어냈다. 해당 모델은 이미지 분류 대회에서 처음으로 CNN 기반의 모델을 이용해 우승을 하였다.

3.2 ResNet

ResNet [5]은 2015년 Kaiming He *et al.*에 의해서 제안된 네트워크로, 아주 깊은 layer로 구성되어 있음에도 불구하고 gradient vanishing 문제를 해결한 혁신적인 모델이라 할 수 있다. 해당 연구의 저자들은 새로운 학습 메커니즘인 Residual Learning을 제안하였는데, 이는 두 개의 3x3 convolution layer와 ReLU layer로 이루어

어진 residual block 단위에서 정의될 수 있다. Residual learning의 이 기본 블록은 입력에서 출력으로 이어지는 shortcut 개념이 있어, 이전 연구와 다른 점은 shortcut에 의해 입력 값이 출력 값에서 보존됨으로써 원래의 출력 값에서 입력 값을 제한 나머지만 residual을 0으로 하는 방향으로 학습이 이루어진다는 것이다. ResNet은 이러한 residual block의 stack 형태의 구조를 띄며 이 block을 몇 개 쌓았느냐에 따라 총 34, 50, 101 혹은 152개의 layer의 여러 ResNet을 제안하였다.

4. 실험

4.1 실험 환경

본 실험에서 쓰인 Faster R-CNN은 Google의 TensorFlow 프레임워크 하에 자체 개발된 모델이다. Faster R-CNN의 중심부 CNN은 AlexNet [4]과 ResNet [5]을 기용하였으며, ResNet은 Google 측에서 미리 훈련해 게시한 152개의 layer로 구성된 ResNet을 그대로 로딩하여 이후 fine-tuning 작업을 거쳤다. 해당 모델은 1)에서 얻을 수 있다. 훈련은 Nvidia Titan X의 single GPU 하에서 이루어졌다.

4.2 실험 데이터

본 연구에서는 고고도 이미지에서 자동차를 탐지하는 과업을 수행하였다. 해당 고고도 이미지는 공개된 데이터 셋인 Cars Overhead With Context (COWC) [6] 데이터를 사용했다. COWC는 positive, negative region supervision으로 제공되며 1000x1000의 규격으로 crop하여 전처리하였다. 데이터 관련 정보는 표 1에 요약되어 있다.

표 1. Data Information

	COWC
훈련 이미지 수	996
테스트 이미지 수	1,074
총 이미지 수	2,060
Anchor Specification	'1:1', '48*48'

1)

http://download.tensorflow.org/models/resnet_v2_152_2017_04_14.tar.gz

4.3 실험 결과 및 분석

훈련 관련 사항을 비교하되, 네트워크의 특성과 결부시켜 분석을 하고자 한다. 네트워크 구조 상 AlexNet과 ResNet은 깊이에서 큰 차이를 보이며, 따라서 학습 대상 파라미터 수도 다르다. 훈련에 걸린 시간, 테스트 성능 등 다양한 기준에 근거해서 측정된 통계치는 다음의 표 2와 같다. 표 2에서 작성된 학습 대상 파라미터 수에 대한 계수는 네트워크 단독이 아닌 Faster R-CNN 전체에 대한 계수이다.


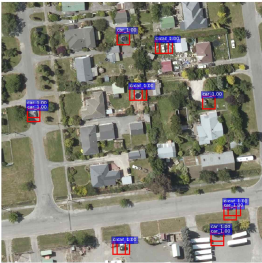


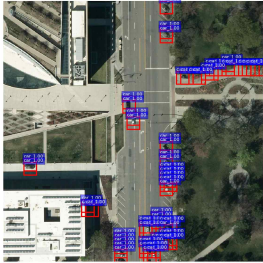
표 2. AlexNet [4], ResNet [5]을 사용한 R-CNN 간 비교 (파라미터 수에 대한 계수는 네트워크 단독이 아닌 R-CNN 전체에 대한 계수이다.)

	AlexNet	ResNet
CNN의 총 layer 수	8	152
CNN의 커널 크기	11, 5, 3	7, 1, 3, 5
CNN의 fully connected layer의 차원	4096, 4096, 1000	1000
총 parameter 수	18M	74M
Training		
훈련에 소요된 총 시간	4일 22시간	16일 16시간
Inference		
이미지 당 테스트에 소요되는 시간	약 0.07초	약 0.4초
Accuracy	0.9902	0.9980
Precision	0.3838	0.5293
Recall	0.5448	0.8720

비록 ResNet이 152 layer로 무려 AlexNet의 8 layer에 비해 약 20배에 달하지만, residual learning이 training & testing time efficiency를 가져다주는 것으로 확인되었다. 또한 ResNet은 그 깊이가 깊지만 3x3의 커널을 사용하고 fully connected layer의 사용도 최소화하였기 때문에 AlexNet에 비해 파라미터 수는 약 4배로 layer 수에 비하면 비교적 적게 측정되었다. 훈련에 소요된 시간은 마찬가지로 파라미터 수에 비례하게 ResNet이 4배였다. 테스트의 경우 더 긴 5 ~ 6배의 시간이 소요되었다.

테스트 성능은 F-measure로 측정하였다. Recall, precision 모두 ResNet의 Faster R-CNN이 앞섰다. 표 3은 특정 example을 살펴봄으로써 탐지 비교 결과를 구체적으로 제시한다.

표 3. 두 네트워크의 탐지 결과 비교

선정 이유	Groundtruth	AlexNet	ResNet
랜덤 선정 예제1			
랜덤 선정 예제2			
AlexNet에서 precision이 가장 저조했던 예제			
AlexNet에서 recall이 가장 저조했던 예제			
ResNet에서 precision이 가장 저조했던 예제			
ResNet에서 recall이 가장 저조했던 예제			

COWC 데이터 셋의 supervision 특징은 그림에서 가정할 수 있는 모든 positive/negative region을 데이터베이스화하고 있지는 않다는 것이다. 즉, positive/negative region들을 샘플링하여 제공하기 때문에 표 3의 groundtruth에는 육안으로 관찰했을 때 자동차인 부분에 표식이 없을 수 있다. 하지만, AlexNet과 ResNet을 기반으로 한 Faster R-CNN은 훈련되어진 대로 자동차에 해당하는 부분은 적절히 탐지하고 있는 것을 확인할 수 있다.

AlexNet에 비해 상대적으로 ResNet을 중심부 CNN으로 기용했을 경우, COWC 데이터 셋에 대해서 precision은 38%, recall은 60% 향상된 결과를 얻을 수 있었다.

5. 결론

본 연구는 최신 딥 러닝 모델 기반의 객체 탐지 모델 Faster R-CNN에 대해서 중심부 CNN을 달리하여 그 비교 분석하였다. ResNet의 경우 깊지만 overfitting 문제를 완화하여 아주 좋은 성능을 보였다. 깊지만 계산 편의로 복잡도와 효율성 간 trade-off의 균형이 적절히 이루어진 모델임을 파악할 수 있었다. 같은 Faster R-CNN이지만, 그 중심부 CNN이 어떤 네트워크냐에 따라 성능이 좌지우지될 수 있으므로 중심부 CNN을 과업에 따라 적절히 선택하는 것이 중요하다.

6. Acknowledgments

본 연구는 방위사업청과 국방과학연구소의 지원으로 한국과학기술원 초고속비행체특화센터에서 수행되었습니다.

7. 참조 문헌

- [1] Ren, Shaoqing, et al. "Faster R-CNN: Towards real-time object detection with region proposal networks." Advances in neural information processing systems. 2015.
- [2] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[3] Girshick, Ross. "Fast r-cnn." Proceedings of the IEEE international conference on computer vision. 2015.

[4] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[5] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[6] Mundhenk, T. Nathan, et al. "A large contextual dataset for classification, detection and counting of cars with deep learning." European Conference on Computer Vision. Springer International Publishing, 2016.