Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography
Vol. 36, No. 6, 469-481, 2018
https://doi.org/10.7848/ksgpc.2018.36.6.469

ISSN 1598-4850(Print)
ISSN 2288-260X(Online)
Original article

딥러닝을 위한 영역기반 합성곱 신경망에 의한 항공영상에서 건물탐지 평가

Evaluation of Building Detection from Aerial Images Using Region-based Convolutional Neural Network for Deep Learning

이대건[1] · 조은지[2] · 이동천[3]
Lee, Dae Geon · Cho, Eun Ji · Lee, Dong-Cheon

Abstract

DL (Deep Learning) is getting popular in various fields to implement artificial intelligence that resembles human learning and cognition. DL based on complicate structure of the ANN (Artificial Neural Network) requires computing power and computation cost. Variety of DL models with improved performance have been developed with powerful computer specification. The main purpose of this paper is to detect buildings from aerial images and evaluate performance of Mask R-CNN (Region-based Convolutional Neural Network) developed by FAIR (Facebook AI Research) team recently. Mask R-CNN is a R-CNN that is evaluated to be one of the best ANN models in terms of performance for semantic segmentation with pixel-level accuracy. The performance of the DL models is determined by training ability as well as architecture of the ANN. In this paper, we characteristics of the Mask R-CNN with various types of the images and evaluate possibility of the generalization which is the ultimate goal of the DL. As for future study, it is expected that reliability and generalization of DL will be improved by using a variety of spatial information data for training of the DL models.

Keywords : Deep Learning, Region-based Convolutional Neural Network, Object Detection, Semantic Segmentation

초 록

딥러닝은 인간의 학습 및 인지능력을 닮은 인공지능을 실현하기 위해 여러 분야에서 활용하고 있으며, 높은 사양의 컴퓨팅 파워가 요구되고 연산 시간이 많이 소요되는 복잡한 구조의 인공신경망에 의한 딥러닝은 컴퓨터 사양이 향상됨에 따라 성능이 개선된 다양한 딥러닝 모델이 개발되고 있다. 본 논문의 주요 목적은 영상의 딥러닝을 위한 합성곱 신경망 중에서 최근에 FAIR (Facebook AI Research)에서 개발한 Mask R-CNN을 이용하여 항공영상에서 건물을 탐지하고 성능을 평가하는 것이다. Mask R-CNN은 영역기반의 합성곱 신경망으로서 픽셀 정확도까지 객체를 의미적으로 분할하기 위한 딥러닝 모델로서 성능이 가장 우수한 것으로 평가받고 있다. 딥러닝 모델의 성능은 신경망 구조뿐 아니라 학습 능력에 의해 결정된다. 이를 위해 본 논문에서는 모델의 학습에 이용한 영상에 다양한 변화를 주어 학습 능력을 분석하였으며, 딥러닝의 궁극적 목표인 범용화의 가능성을 평가하였다. 향후 연구방안으로는 영상에만 의존하지 않고 다양한 공간정보 데이터를 복합적으로 딥러닝 모델의 학습에 이용하여 딥러닝의 신뢰성과 범용화가 향상될 것으로 판단된다.

핵심어 : 딥러닝, 영역기반 합성곱 신경망, 객체탐지, 의미적 분할

Received 2018. 10. 02, Revised 2018. 10. 12, Accepted 2018. 11. 12
1) Member, Dept. of Environment, Energy & Geoinformatics, Sejong University (E-mail: dglee@sju.ac.kr)
2) Dept. of Environment, Energy & Geoinformatics, Sejong University (E-mail: gracee3792@sju.ac.kr)
3) Corresponding Author, member, Dept. of Environment, Energy & Geoinformatics, Sejong University (E-mail: dclee@sejong.ac.kr)

1. 서 론

인간의 인지능력에 근접한 시스템 개발은 오래전부터 관심의 대상이 되었지만, 컴퓨팅 파워가 급속히 향상된 최근에는 다양한 분야에서 인공지능(AI: Artificial Intelligence)의 실현을 위해 인공신경망(ANN: Artificial Neural Network)에 의한 딥러닝(DL: Deep Learning)에 관한 연구와 개발이 활발하게 진행되고 있다. ANN은 "computational brain modeling"이며, 최초의 시도는 1943년 McCulloch and Pitts가 제안한 이진분류(binary classifier)를 위한 neuron 모델이었다. 1957년에는 Rosenblatt가 neuron 모델을 개선하여 지도학습(supervised learning)을 위한 perceptron 알고리즘을 개발하였다(Hertz *et al*., 1991; McCulloch, and Pitts, 1943; Rosenblatt, 1958).

ANN의 중추적인 역할을 담당하는 역전파(backpropagation)에 대한 연구는 1960년대 초부터 시작되어 1986년에 Rumelhart가 다층 신경망(multi-layer network)을 위한 역전파 알고리즘을 체계화하였다(Rumelhart *et al*., 1986). 다층 신경망과 역전파는 2000년 중반부터 본격적으로 시작된 DL의 기반이 되고 있다. 특히 영상 및 컴퓨터비전 분야의 대표적 ANN 경연대회인 ILSVRC (ImageNet Large Scale Visual Recognition Challenge)에서는 DL을 위한 신경망 모델의 영상분류(classification) 성능을 평가하고 우승 모델을 선정하여 분류 오류율(classification error rate)을 발표해오고 있었지만, 최근에는 분류뿐 아니라 분류된 객체들의 위치(localization) 및 객체탐지(object detection) 능력을 평가항목에 추가하여 종합적으로 신경망 모델의 학습 및 범용화 능력 등을 판단하고 있다(Russakovsky *et al*., 2015). 영상매체를 이용한 시각정보 처리를 ANN 기반의 DL 기술로 실현하려는 연구는 많은 관심의 대상이 되고 있으며, 이에 따라 ANN 성능은 급속하게 향상되고 있다(Audebert *et al*., 2018; Ball *et al*., 2017).

국내의 공간정보 분야에서 DL 관련 연구로는, Oh (2010)는 산사태를 발생시키는 인자들의 가중치를 항공영상을 이용한 다층 퍼셉트론(MLP: Multi-Layer Perceptron) 신경망과 GIS 분석으로 추정하였다. Choe and Yom (2017)은 기상 데이터와 위성영상을 효율적으로 MLP 신경망 모델에 적용하여 정확한 지표면 온도를 추정하는 연구를 제안하였다. Chung and Lee (2017)는 복셀(voxel) 형태로 변환한 점군집 데이터를 심층 신경망(DNN: Deep Neural Network) DL 모델에 의한 학습을 통하여 분류하는 연구를 발표하였다. Kim and Bae (2017)는 GNSS (Global Navigation Satellite System) 신호 지연오차로 추정한 가강수량과 기상관측 데이터(기온, 기압, 풍속 및 습도 등)를 RNN (Recurrent Neural Network) 기반의 LSTM (Long Short-Term Memory) 알고리즘에 적용하여 장기적인 강수예측이 가능한 모델을 제안하였다. Baek and Yom (2018)은 점군집 데이터를 사용하여 체적 산정을 위해 CNN (Convolution Neural Network) 기반의 DeMoN (Depth and Motion Network)을 이용한 AI 학습에 의한 방법과 사진측량 방법을 비교하여 서로 유사한 결과를 확인하였다. Lee and Yom (2018)은 머신러닝(ML)의 핵심 요구 사항인 학습에 필요한 데이터의 효율적 확보를 위한 전처리 과정의 자동화 구현을 위해 Python 기반의 라이브러리와 웹기반 공간정보 오픈소스를 활용하여 다수의 사용자들이 접근 가능한 마이크로서비스 시스템 아키텍처 연구와 개발을 수행하였다.

국외에서의 DL 연구는 오래전부터 활발히 진행되어 왔다. 특히 최근에는 단순한 영상분할 및 분류보다 난이도가 높은 의미적 분할(semantic segmentation) 및 고차원 컴퓨터비전 분야인 영상이해(scene understanding) 등 인간의 인지능력을 모방하는 연구로 발전하고 있다(Tokarczyk *et al*., 2015). 더 나아가 AI의 주요 분야인 컴퓨터비전과 자연어 처리 기능을 융합할 수 있는 복합 신경망 모델을 개발하여 영상으로부터 상황을 인지하고 설명하는 image captioning(또는 scene description)수준까지 구현되고 있으며, 이를 위해서는 지능적 인식에 의한 객체의 의미적 분할이 필수적이다(Shaikh, 2018; You *et al*., 2016; Xu *et al*., 2015). 또한 공간정보 분야에서는 2D 영상에 의존하는 제한적인 학습 방법에서 3D 공간정보 데이터 (RGB-D : Red Green Blue - Depth) 데이터: 광학영상과 depth 정보로 구성된 데이터, LiDAR (Light Detection and Ranging) 데이터, DSM 등)를 복합적으로 학습할 수 있는 DL 모델 개발에 관한 연구가 본격적으로 진행되고 있다(Campos-Taberner *et al*., 2016; Vo *et al*., 2016).

대표적인 DL 연구로는, Marmanis *et al*. (2016)은 Long, *et al*. (2015)에 의해 본격적으로 적용된 심층 완전 CNN (deep FCN)을 기반으로 옥스퍼드 대학에서 개발한 VGG (Visual Geometry Group) - 16 모델로 적외선 항공영상과 DSM을 학습하여 항공영상의 의미적 분류를 수행하였다. Hazirbas *et al*. (2016)은 앞에서 언급한 depth 정보와 영상을 융합하여 영상분할을 위한 DL 모델인 FuseNet을 제안하였다. Audebert *et al*. (2018)은 영상분류를 위해 고해상도 광학영상뿐 아니라 다중분광 영상의 밴드 특성을 이용한 식생지수(NDVI)와 LiDAR 데이터로 생성한 정규화 수치표면모델(nDSM)을 DL 모델 학습을 위한 다중 심층 신경망(multi-modal deep network)을 제안하였다. 특히 CNN기반의 SegNet 모델로 학습시키는 early fusion 방법과 ResNet (Residual Neural Network) 모델로 학습시키기는 late fusion 방법을 비교 분석하였다.

ISPRS (International Society ofr Photogrammetry and Remote Sensing)은 2018년 DL 특별호에 사진측량과 원격탐사를 중심으로 공간정보 분야의 인공신경망과 DL에 관한 다수의 논문을 발간하였다. Kemker *et al.* (2018)은 다중분광영상(MSI)의 분류를 위한 CNN 모델을 제안하였으며, synthetic MSI를 생성하여 학습을 수행한 결과인 사전학습 모델(pre-trained model)에 실제 MSI를 입력한 분류 결과의 정확도를 분석하였다. 실제 영상으로 학습을 수행하지 않고 synthetic 영상으로 학습을 수행하는 장점으로는 학습에 요구되는 다량의 영상 획득과 annotation 데이터를 쉽게 생성할 수 있으며, 학습 효과와 효율성을 높일 수 있다고 제시하고 있다. Paoletti *et al.* (2018)은 MSI보다 밴드가 훨씬 많은 초미세분광영상의 분류를 위한 3D CNN 모델을 개발하여 공간적 파장별 분류를 동시에 수행하였다.

CNN은 주로 영상분류 및 객체인식에 활용되지만, Wang *et al.* (2018)은 위성영상의 registration을 위한 영상정합(image matching)을 수행할 수 있는 CNN 모델을 제안하였으며, 영상정합에 많이 사용하는 SIFT (Scale Invariant Feature Transform)에 의한 결과와 정확도를 비교 분석하였다. Zhang *et al.* (2018)은 영상의 노이즈를 감소시켜 화질을 향상시키기 위한 목적으로 CNN 모델을 적용하였으며, Xing *et al.* (2018)은 MSI의 해상도를 향상시킨 pan-sharpening 영상을 생성하기 위한 DL 모델인 DML (Deep Metric Learning)을 제안하였다. 많은 연구들이 DL 모델에 항공 및 위성영상으로부터 객체인식 방법을 수행하고 있지만, Kang *et al.* (2018)은 지상에서 촬영한 도심지의 street view 영상으로부터 건물을 종류 및 기능별(아파트, 교회, 주택, 창고, 상업용 건물 등)로 인식하고 분류할 수 있는 방법을 제시하였다. 이를 위해 사용한 DL 모델은 기존에 개발된 CNN 모델인 AlexNet, VGG-16 및 ResNet을 사용하였으며, VGG-16 모델이 정확도가 높다는 결론을 제시하였다.

특히 Deng *et al.* (2018)은 본 논문에서 적용한 Mask R-CNN의 근간이 되는 영역기반 CNN (R-CNN)을 이용하여 다중축척 영상에서 객체를 탐지하는 연구를 수행하였다. 그 외에 ASRPS (American Society for Photogrammetry and Remote Sensing)에 발표되는 DL 논문도 지속적으로 증가하고 있다. 특히 IEEE (Institute of Electrical and Electronics Engineers)의 AI, 컴퓨터비전, 영상처리, 원격탐사, 인공신경망 및 학습시스템 등 DL관련 분과(transaction)에서 발행되는 논문은 분야가 방대하고 더욱 확대되고 있다. 많은 연구에서 제시하는 바와 같이 영상을 이용한 대부분의 신경망 모델은 CNN을 모태로 하고 있으며, 향후에는 영상뿐 아니라 다양한 공간정보 데이터를 학습할 수 있는 DL 모델이 주류가 될 것이라고 판단된다.

그러나 원하는 목적을 충족시킬 수 있는 최적의 DL 모델을 개발하고 학습시키고 검증하기 위해서는 많은 경험과 시간 및 노력이 요구되며, 여러 시행착오를 거쳐야 한다. 그러므로 사전학습 모델(pre-trained model)의 전이학습(transfer learning)을 통해 추구하고자 하는 결과를 얻을 수 있다면, DL의 파급효과와 활용은 크게 증대될 것이며, 궁극적으로 AI의 보편화가 실현될 것이다. 그러나 학습에 사용된 데이터와 실제 데이터 사이에 일관성 및 공통적 특성이 없다면, 제한적 조건에서만 사전학습 모델을 적용할 수 있고 원하는 결과는 보장되지 않는다.

본 논문은 객체의 의미적 분할과 탐지를 위한 CNN 기반의 신경망 중에서 2017년에 FAIR (Facebook AI Research)의 He *et al.* (2017)에 의해 개발된 Mask R-CNN을 이용하여 건물탐지결과를 분석하였다. 이를 위해 https://www.crowdai.org/challenges/mapping-challenge/dataset_files에서 제공하는 사전학습 모델을 사용하였다. 사전학습 DL 모델에서 학습에 사용한 영상에 기하학적 변화와 화질 변화를 발생시켜서 DL 신경망의 성능과 한계를 분석하였다. 또한 DL의 최종 목표인 범용화와 신뢰성을 향상시키고 사진측량의 응용 분야에서 DL을 본격적으로 도입하기 위한 다중센서 데이터를 복합적으로 이용할 수 있는 향후 연구방안을 제안하였다.

2. ANN과 DL

ANN 개념은 컴퓨터가 발명되기 이전인 1940년 초에 제안된 이후 발전과 침체기를 거쳐 컴퓨터 성능 및 소프트웨어의 혁신적 향상과 정보통신 기술의 급속한 발전으로 AI 실현을 위한 DL에 관한 많은 연구와 개발이 활발하게 진행되고 있다. 최근에는 공간정보 분야에서도 인공신경망과 DL에 관한 연구들이 발표되고 있다(Audebert *et al*., 2018; Ball *et al*., 2017; Campos-Taberner *et al*., 2016; Marmanis *et al*., 2016).

DL을 위한 ANN 모델 개발을 위해서는 architecture 설계와 학습기법 등 모델 자체 개발에 직접 관련된 사항뿐 아니라 학습과 검증에 필요한 다량의 데이터 획득과 학습 수행에 관련된 여러 hyper-parameter들의 조율(tuning) 등과 같이 많은 실험과 경험이 요구된다. 이와 같이 DL의 높은 진입장벽은 전이학습과 다양한 오픈 소스들을 이용해서 해결할 수 있다. DL에서 가장 중요하고 어려운 과정인 구현을 위한 다양한 tool들이 제공되고 있으므로 특정 목적에 적합하도록 기존 모델들을 customize할 수 있다. 대부분 tool들의 소스가 공개되어 있어 DL의 접근과 활용이 용이하다. 특히 오픈 소스인 Python 기반으로 개발되어 제한 없이 사용할 수 있다. Python 라이브러리로 제공되는 대표적인 DL tool은 Keras, PyTorch, TensorFlow, Theano, Caffe-2 등이 있다(Ball *et al*., 2017).

2.1 합성곱 ANN (CNN)

Mask R-CNN은 영상의 의미적 분할, 분류, 객체탐지 및 인식을 위해 개발된 DL 모델이다(Krizhevsky, 2012). 일반적인 ANN은 완전결합 층(FCL: Fully Connected Layer)으로 구성되어 1차원 형태의 데이터만 입력이 가능하다(Fig. 1 참조). 반면에 영상은 픽셀의 위치정보와 픽셀에 저장된 밝기값으로 이루어진 3차원 배열이므로 영상을 FCL에 입력하려면 1차원으로 변환시켜야 한다. 이 경우 공간정보가 유실되어 영상으로부터 위치를 보존하면서 특징을 추출할 수 없으므로 학습이 비효율적이다. 이런 문제를 해결하기 위해 영상의 공간적 특성을 유지한 상태로 학습이 가능한 CNN이 제안되었다(Simard *et al*., 2003).

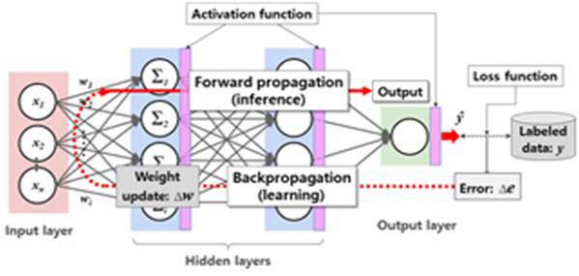


Fig. 1. Learning process in general ANN

CNN은 영상의 학습을 위한 DL 모델로 확고하게 자리 잡아가고 있으며, 학습 효율과 성능 향상을 위해 많은 연구가 진행되어 매년 성능이 개선된 새로운 신경망 모델들이 ILSVRC와 같은 국제적 경진대회에서 발표되고 있다(Russakovsky *et al*., 2015; Pang *et al*., 2018). 1989년에 LeCun *et al*., (1989)에 의해 우편번호 인식으로 널리 알려진 최초로 실용화된 CNN 모델인 LeNet이 개발되었다. LeNet 모델은 합성곱 연산을 수행하여 특성맵(feature map)을 생성하는 "convolution 층"(일반 ANN의 은닉층에 해당), 특성맵을 다음 층으로 전달하기 위한 "activation 층", 데이터 크기를 줄이기 위한 "pooling 층"으로 구성된다. 그리고 출력층 부분은 FCL로 구성되어 class score를 계산하여 객체를 분류한다. 대부분의 CNN 모델들은 LeNet의 아키텍처를 모태로 한다. CNN의 아키텍처는 Fig. 2에서 보여주고 있다.

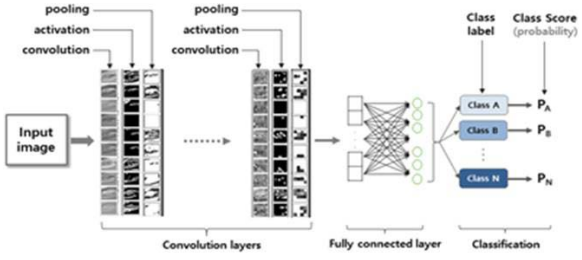


Fig. 2. Architecture of generic CNN model

CNN을 구성하는 층들의 세부적인 역할과 기능은 다음과 같다.

① Convolution(합성곱) 층: 입력된 영상과 필터(또는 kernel) 간에 convolution을 수행하여 특성정보를 추출하고 특성맵을 생성한다. Convolution 연산은 영상처리에서 사용하는 필터링과 같은 Eq. (1)을 사용한다.

$$(I \otimes w)_{ij} = \sum_{m=0}^{k_1-1} \sum_{n=0}^{k_2-1} \sum_{c=1}^{C} (w_{m,n,c} \cdot I_{i+m,j+n,c}) \quad (1)$$

where I and w are input image and filter, respectively. i and j are image size, and k_1 and k_2 are filter size, and C denotes number of channel (or band) of the input image.

CNN의 필터계수(filter coefficient)는 일반 ANN의 가중치에 해당되며, 일반적으로 초기값은 무작위로 부여하고 학습과정에서 반복적으로 업데이트 된다. Convolution을 수행하면 Fig. 3처럼 입력영상의 테두리 부분에서 필터 크기의 반에 해당하는 픽셀들이 소실되어 특성맵의 크기가 작아지므로 입력영상과 같은 크기를 유지하기 위해 소실된 픽셀들을 "0"으로 채우는 zero padding을 수행한다.

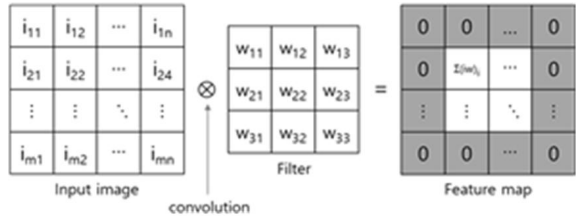


Fig. 3. Zero padding

② Activation(활성화) 층: 중추적인 역할을 하는 층으로서 활성화 또는 전달함수(activation 또는 transfer function)를 사용하여 특성맵의 정보를 출력할 조건과 형태를 판단하며, 임계값을 초과할 경우에 다음 층으로 전달한다. 심층 신경망에 사용

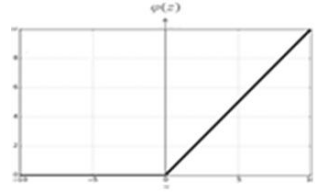


Fig. 4. ReLU function

하는 활성화 함수는 비선형이어야 하며, 대표적인 활성화 함수는 sigmoid (logistic function), tanh (hyperbolic tangent), ReLU (Rectified Linear Unit) 등 이다. 대부분의 CNN 모델에서는 Fig. 4와 같은 ReLU 함수를 사용하며 Eq. (2)로 표현된다.

$$\phi(z) = \max(0, z) \tag{2}$$

where z is input value, and $\phi(z)$ is activation function.

③ Pooling 층: 특성맵의 크기를 줄이는 층으로서 크기가 작아지면, 오히려 영상의 특징이 잘 부각되고 학습이 효율적이며, DL의 문제점인 과적합(overfitting)을 해결할 수 있는 장점이 있다. Fig. 5에서 보여주는 것처럼 대부분의 CNN에서는 2x2 크기를 하나의 단위로 특성맵을 분할하고, 행과 열 방향으로 중복하지 않고 2 픽셀씩 이동하면서(즉 stride=2) 분할된 영역에 속한 4개의 픽셀에 대해 평균값(average), 최대값 (max) 또는 요소들의 제곱의 합에 대한 제곱근(L2-norm)을 취하는 방법이 있다 (Eq. (3), Eq. (4) and Eq. (5)). Pooling은 가중치를 고려하지 않으며, 주로 max pooling 방법을 사용한다.

$$\textit{average pooling}: \text{avg}(fij) \in R_{ij} \tag{3}$$

$$\textit{max pooling}: \quad \max(fij) \in R_{ij} \tag{4}$$

$$\textit{L2-norm pooling}: \text{sqrt}(\sum fij^2) \in R_{ij} \tag{5}$$

where f_{ij} denotes each element of feature map, and R_{ij} represents partitioned region where corresponding feature elements belong to.

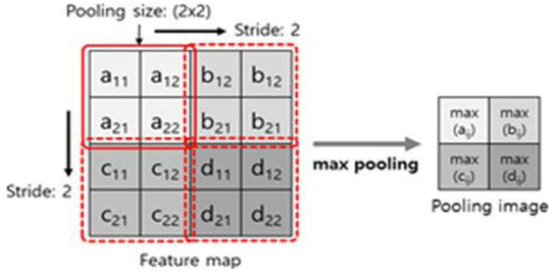


Fig. 5. Resizing feature map by max pooling

④ Fully connected 층(FCL): Pooling 층에서 모든 출력된 값들과 분류 클래스(또는 카테고리)의 모든 요소가 서로 연결된 완전결합 구조로서 FCL 층으로 입력되는 값과 가중치를 곱하여 계산된 값으로 클래스 점수(즉 각각의 클래스에 대한 확률)로 분류를 수행한다.

CNN 모델의 종류와 특성은 convolution 층의 수, 필터의 수 및 크기, 활성화 함수의 종류 및 pooling 방식과 층들의 배열 형태에 의해서 결정된다. 또한 동일한 구조의 신경망 모델도 학습 데이터의 특성과 학습 수행에 관련된 hyper-parameter(backpropagation 방법, epoch, iteration, 학습률 등)에 따라 결과는 다를 수 있다. 최근에는 단순히 영상분류 성능 이외에도 분류된 객체의 위치 및 탐지/인식 정확도 등 여러 측면을 고려한 신경망 모델의 평가 방법을 적용하고 있으며, DL의 궁극적 목표인 일반화 및 범용화를 실현하기 위해서 학습과 검증에 사용된 특정 영상 이외에 전혀 학습에 관여하지 않은 새로운 영상(즉 new, unseen or future image)을 사전에 학습된 모델에 입력하고 결과를 평가하여 범용성을 판단하고 있다.

본 연구에서는 현재 많은 연구가 진행되고 있으며 관심의 대상인 R-CNN 모델의 발전 과정을 파악하고 가장 최근에 발표된 모델인 Mask R-CNN의 전이학습을 통해 실험을 수행하고 결과를 분석하여 문제점과 향후 연구방향을 제시하였다.

2.2 영역기반 CNN 모델의 개요

R-CNN은 영상으로부터 객체를 탐지하고 객체들을 종류(또는 특성)별로 분류하고, 객체들의 위치(또는 영역)를 결정하는 것을 주요 목적으로 하는 DL 모델로서 인간의 객체인식 능력을 구현하기 위해 개발되었다. 신경망 학습에 의해 수행되는 객체탐지는 다음 두 가지 방식으로 가능하다.

- Sliding window: 객체가 존재할 수 있는 모든 크기의 탐색 영역(search window)을 설정하고 모든 픽셀에 대해 분류를 수행하는 방식으로 탐색해야 할 영역의 수가 많으므로 연산시간이 많이 소요되어 비효율적이다.
- Region proposal: 비효율적인 sliding window 방식을 개선한 방법으로, 모든 픽셀에 대해 탐색하지 않고 객체가 존재할 가능성이 높은 영역에 대해서 탐색하는 방식으로서 성능과 속도를 향상시킬 수 있다.

위의 방법을 수치사진측량에서 사용하는 기법과 비교하면, sliding window 방식은 입체영상에서 공액점(conjugate point)을 자동으로 탐색하기 위해 모든 픽셀을 중심으로 좌우영상에서 탐색 영역(search window)간의 영역기반 영상정합(area-based image matching) 방법과 유사하다. 또한 region proposal 방식은 입체영상에서 공액점이 될 가능성이 높은 관심점 또는 특이점(interest points), 즉 공액점 후보 픽셀(candidate pixel)을 추출하고 모든 픽셀에 적용하지 않고 후보 픽셀에서만 영상정합을 수행하는 방법과 유사하다(Schenk, 1999).

R-CNN은 2013년에 Girshick가 처음 제안하였으며, 2015년에 “Fast R-CNN”, 또한 2015년에 “Faster R-CNN”으로 발전되

어 속도와 성능이 향상되었다. 그리고 가장 최근인 2017년에는 기존 모델의 문제점을 해결하여 한층 더 발전된 형태인 Mask R-CNN이 개발되어 학습 속도뿐 아니라 효율성과 객체탐지 정확도가 개선되어 탐지된 객체의 윤곽선 묘사를 픽셀 단위까지 가능하도록 향상되었다.

2.2.1 R-CNN

영상에서 객체들을 구별하여 추출하는 과정은 단순히 영상을 분류하는 과정보다 난이도가 높다. R-CNN은 2013년에 FAIR에서 개발하였으며, 객체가 존재할 가능성이 있는 영역을 결정하기 위한 region proposal(또는 bounding box) 방법을 적용하고 있다. 즉 객체가 존재하는 영역에 사각형을 형성하는 과정을 적용하고 있으며, 이런 영역을 탐지하기 위해서 선택적 탐색(selective search) 알고리즘이 제안되었다(Girshick *et al*., 2016).

Selective search는 영상의 색조 및 밝기값 등 특성이 유사한 인접 픽셀들을 그룹핑하는 방법이다. 이는 기존의 영상분할 기법 중 하나인 영역확장(region growing)과 비슷한 개념이다. Selective search 결과를 CNN 모델에 입력하게 된다. 사용한 CNN 모델은 2012년 ILSVRC에서 우승한 AlexNet에 기반한 신경망 모델이다(Krizhevsk, 2015; Garcia-Garcia, *et al*., 2017)이며, 객체가 존재하는 영역을 나타내는 bounding box의 위치 정확도를 향상시키기 위해서 선형회귀를 수행한다.

2.2.2 Fast R-CNN

2015년 Microsoft에서 개발한 Fast R-CNN은 모델의 이름이 의미하는 것처럼 기존 R-CNN의 학습 속도를 향상시킨 모델이다. 복잡하고 시간이 많이 소요되는 학습과 검증과정을 통합하여 속도와 정확도를 향상시켰다. 이를 위해 R-CNN에서 객체가 존재하는 영역인 모든 bounding box를 모델에 입력하고 분류해야 하는 학습의 비효율성을 해결하기 위해 제안된 모델이다. 서로 근접한 위치의 객체들에 형성된 bounding box들은 겹쳐질 수 있다. 이와 같이 중복된 bounding box들을 개별적으로 학습시키는 대신에 RoIPool (Region of Interest Pooling) 기법을 도입하여 bounding box 정보를 모델에 입력하여 생성된 특성맵으로부터 해당 영역을 추출하여 pooling하는 방법이다. 그러므로 Fast R-CNN은 CNN, classifier와 bounding box regressor를 단일 네크워크로 구성한 통합 학습 체계(joint training framework)이다. 이 모델은 bounding box를 모델에 입력하는 R-CNN 보다 학습에 소요되는 시간을 단축할 수 있는 장점이 있다(Girshick, 2015).

또한 R-CNN과 다른 점은 선형회귀 방법 대신에 최종 출력층에 softmax 함수를 배치하여 영상을 분류하는데 사용하고 있다. softmax 함수는 Eq. (6)으로 표현되며, 지수함수(exponential function)를 사용하여 입력값들을 정규화하여 출력하므로 normalized exponential function이라고도 한다. 입력된 값을 0과 1사이의 값으로 변환(즉 정규화)하고 출력값들의 합이 1이 되도록 하는 함수이다. 특히 softmax 함수에 의한 정규화는 "softmax"는 "최대값을 완화하다"라는 의미처럼 극단적인 값이나 과대오차를 제거하지 않아도 이들의 영향을 감소시키는 장점이 있다.

$$f(z)_j = \frac{e^{z_j}}{\sum_{k=0}^{K} e^{z_k}} \text{ for } j = 1, \ldots, K \tag{6}$$

where z is input value, and K denotes number of inputs.

2.2.3 Faster R-CNN

Faster R-CNN은 Fast R-CNN의 속도를 향상시키기 위해 제안한 모델로서 2015년에 Microsoft에서 개발한 대표적인 컴퓨터비전 연구 성과이다. 이전 모델인 Fast R-CNN에서 사용한 bounding box 생성 방법은 시간이 많이 소요되므로 이를 개선한 방식인 RPN (Region Proposal Network)을 모델 내부에 통합하여 속도를 더욱 향상시킴으로써 실시간에 근접한 객체탐지 가능성을 제시하고 있다. 기본적으로 FCL 구조를 가지고 있는 RPN의 핵심 역할은 입력영상에 사각형의 객체의 존재를 표시하는 bounding box인 region proposal(또는 anchor box)의 위치와 객체 존재에 대한 순위를 부여하여 어떤 영역이 객체가 존재할 확률이 높은지 결정한다. 즉 RPN은 sliding window를 설정하여 탐색 영역을 이동시키면서 객체가 존재할 가능성이 높은 위치, 즉 bounding box가 생성될 후보 지역에 anchor box들을 생성한다. 요약적으로 설명하면, convolution에 의해 생성된 특성맵에 sliding window를 적용하여 anchor 지점의 좌표를 중심으로 여러 크기의 bounding box를 결정하고 각각의 bounding box에 대해 우선 순위를 계산한다(Ren *et al*., 2017).

그러므로 Faster R-CNN 모델에서는 anchor의 역할이 매우 중요하다. 또한 이 모델의 장점은 영상 또는 특성맵에 대해 피라미드를 생성할 필요가 없으며, 필터의 크기도 변경하지 않아도 되므로 효율적이다. Fig. 6에서 보여주는 것처럼 3가지 크기(128x128, 256x256 및 512x512)와 3개의 비율(1:1, 2:1 및 1:2)의 anchor box들을 사전에 정의하고 총 9개의 anchor box들을 적용하여 객체를 탐지한다. Fig. 7은 anchor box 예시를 보여주고 있다.

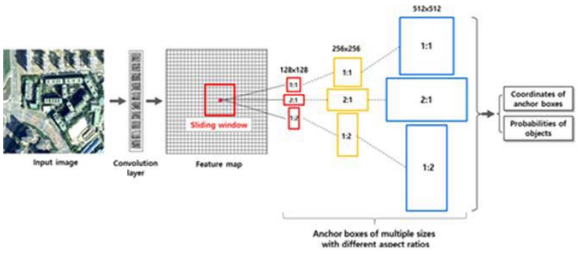


Fig. 6. Anchor boxes for object detection

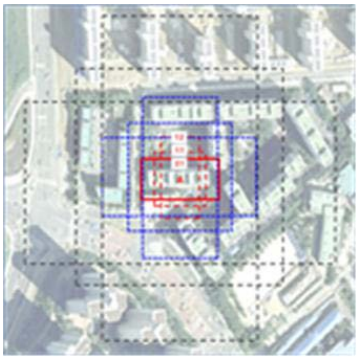


Fig. 7. Demonstration of nine possible anchor boxes

2.2.4 Mask R-CNN

Mask R-CNN은 2017년에 Facebook AI 팀이 개발한 DL 모델로서 분할된 영상을 masking하는 방법이 핵심이고, 픽셀 단위 수준까지 분할하기 위해 Faster R-CNN을 확장 개선한 모델이다. 즉 객체를 탐지하여 객체의 영역을 정의하는 bounding box의 위치뿐 아니라 bounding box내에 존재하는 객체를 픽셀 수준까지 정확하게 결정하는 것이 목적이다. Mask R-CNN의 개념은 단순하면서도 자연스러운 직관적인 발상으로 제안된 모델이다. 이전 모델인 Faster R-CNN에서는 각 후보 객체에 대해 class label과 bounding box를 출력하지만, Mask R-CNN은 object mask를 생성하는 과정이 추가되었다(He *et al*., 2017).

Object mask에서 출력되는 값은 class와 bounding box이며 이를 통해 Fast R-CNN보다 정교하고 객체의 윤곽이 추출하는 역할을 한다. 그러므로 Fast R-CNN과 Faster R-CNN에서는 수행하지 못했던 픽셀 단위까지 조정을 가능케 한다. 이를 위해 Mask R-CNN에는 각각의 픽셀이 객체에 해당하는지를 판단하는 binary mask를 생성하는 과정이 추가되었으며, 픽셀의 정확한 위치를 추출하기 위해 RoIAlign를 제안하였다. RoIAlign은 Fast R-CNN의 RoIPool 기법을 개선하여 영역의 위치에 발생하는 오차를 2D 선형보간법을 적용하여 감소시키는 과정이다(Parthasarathy, 2017). Fig. 8은 Mask R-CNN 모델의 구조를 보여주고 있다.

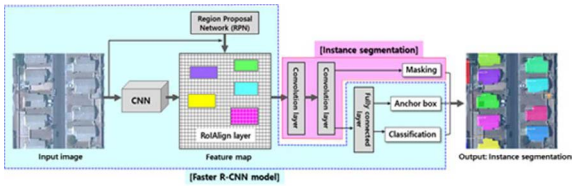
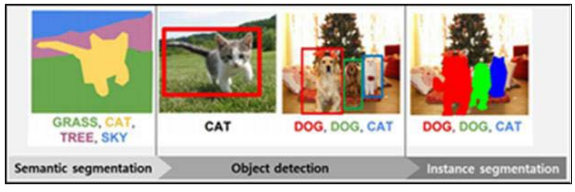


Fig. 8. Mask R-CNN model architecture

이와 같이 CNN은 R-CNN, Fast R-CNN, Faster R-CNN 그리고 Mask R-CNN으로 진화화면서 Fig. 9에서 보여주는 예시처럼 단순 영상분할에서 객체영역 탐지, 의미적 분할, 그리고 픽셀 레벨의 위치정확도를 유지하면서 개별 객체를 구분할 수 있는 단계로 발전하고 있다.



(Source:cseweb.ucsd.edu/classes/sp18/cse252C-a/CSE252C_20180509.pdf)

Fig. 9. Progress of CNN: From object detection to instance segmentation

3. 실험방법

본 연구는 Python Keras 라이브러리를 기반으로 구현된 Mask R-CNN 모델과 학습에 사용한 영상에 다양한 변화를 발생시켜서 건물을 추출하였다. 학습을 위한 데이터에 대한 자세한 내용은 “Mapping Challenge: Building Missing Maps with Machine Learning” (www.crowdai.org/challenges/ mapping-challenge)에 설명되어 있으며, 광학 위성영상과 건물을 인식하여 표시한 annotation 데이터(label 데이터 또는 ground truth)를 제공하고 있다(Fig. 10 참조). 학습 데이터 셋은 300 x 300 화소의 280,741개의 RGB 영상과 이에 대응하는 annotation으로 구성되어 있으며, 검증용 영상은 같은 화소수의 60,317개 사용을 권장하고 있다. Annotation 데이터는 객체탐지, 분할 및 영상이해를 위한 captioning을 구현하기 위해 구축한 마이크로소프트의 데이터 셋인 COCO (Common Object in Context) 포맷을 사용하고 있다.

(a) Training images

(b) Annotation data

Fig. 10. Examples of RGB image and corresponding annotation data

Fig. 11은 학습을 위해 사용된 다수의 항공영상 중에서 하나의 영상을 보여주고 있으며, 또한 학습에 관여하지 않은 새로운 지역의 영상에서 건물을 탐지하였다. 지상에서 촬영된 영상은 여러 객체들이 혼재되어 있거나 뒤에 있는 객체는 앞에 있는 객체에 의해 부분적으로 폐색되는 경우가 많지만, 항공영상은 이런 경우가 드물다. 특히 엄밀정사영상(true orthoimage)을 사용할 경우에는 폐색지역이 존재하지 않으므로 학습효과가 높아 원하는 결과를 얻을 수 있는 장점이 있다.

다양한 형태의 건물이 밀집된 주거지 영상을 사용하여 실험을 수행하였으며, 다양한 조건으로 영상을 인위적으로 변화시켜서 결과를 분석하였다. 영상 크기 및 해상도 변경, 영상 분리(partition)와 회전과 같은 기하학적 변화를 주었으며, 또한 화질에 변화를 발생시키기 위해 밝기 및 대조비 변화, 영상의 RGB 배열순서 변경과 서로 다른 레벨의 노이즈를 첨가하여 각 경우에 대한 결과를 분석하였다. 이와 같이 영상의 변화 요소가 결과에 미치는 영향을 분석하는 것은 DL 모델의 신뢰성과 범용성 향상에 중요한 자료로 활용될 수 있으며 새로운 모델 개발 및 검증에 고려해야 할 사항을 제시할 수 있다.

영상을 회전할 경우 Fig. 12에서 보여주는 것처럼 빈 픽셀(검은색 부분)이 발생하게 된다. 이와 같은 상태로 영상을 입력하면, 빈 픽셀 부분도 객체로 인식될 수 있으므로 회전하기 전에 원래 영상을 패딩하여 빈 픽셀(empty pixel)이 발생하지 않도록 하였으며, Fig. 13과 같이 연속된 자연스러운 영상이 되기 위해 mirror 패딩을 적용하였다.

Fig. 11. A sample of training image

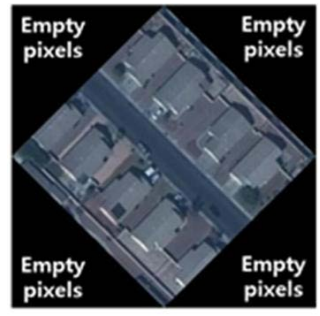


Fig. 12. Rotated image without padding

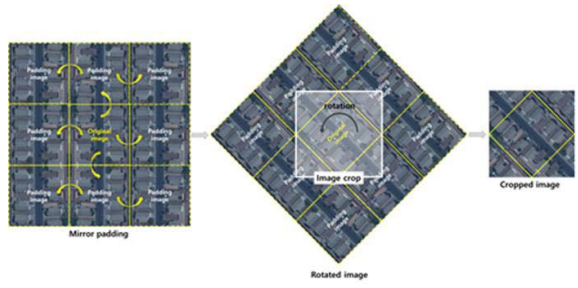


Fig. 13. Mirror padding for image rotation

4. 결과 및 분석

4.1 학습영상의 건물탐지

일반적으로 검증용 데이터(validation data) 및 시험용 데이터(test data)로 모델의 성능을 평가하지만, 본 연구는 기존의 방식과는 다르게 모델의 성능과 범용성을 다양한 측면에서 평가하는 것이 목적이므로 학습에 사용된 영상을 기하학적 그리고 화질을 변형하여 모델에 입력한 결과를 분석하였다. 탐지된 건물은 서로 다른 색으로 구분하여 표시되었다.

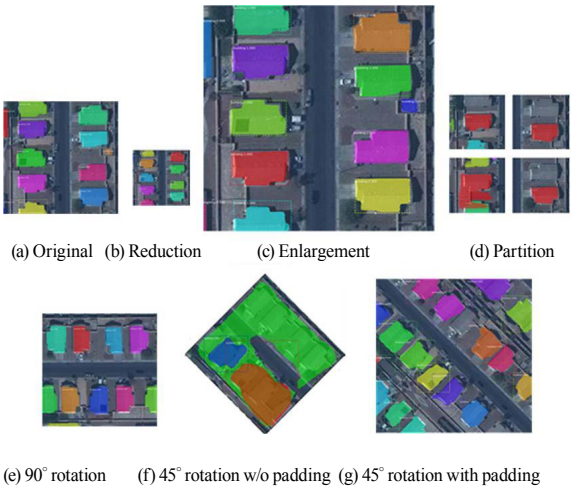


Fig. 14. Building detection with geometrically transformed images

Fig. 14는 기하학적으로 변형된 영상들에 대한 결과이다. Fig. 14(a)는 원래 학습에 사용된 영상에서 건물을 탐지한 결과이므로 모든 건물들이 실제 형태대로 탐지되었음을 알 수 있다. Figs. 14(b)와 (c)는 크기를 각각 1/2과 2배로 변경한 영상이며, 학습 영상의 결과와 거의 동일하다. 그러나 극단적으로 축소 또는 확대한 경우에는 결과가 올바르게 나오지 않을 것으로 추측되며, 이러한 판단은 다음 절(4.2 새로운 영상의 건물탐지)의 실험 결과로부터 추측이 가능하다. Fig. 14(d)는 원래 영상을 4개 부분으로 분할한 후 각각의 영상에 대한 결과를 보여주고 있으며, 탐지되지 않은 건물과 부정확하게 탐지된 건물들이 있다.

Figs. 14(e), (f)와 (g)는 90°와 45°로 회전한 영상의 결과이며, 90° 회전 영상에서는 학습데이터와 유사성에 변화가 없기 때문에 건물이 정확하게 탐지되었다. 반면에 객체의 형태 변화가 최대로 발생하는 45° 회전 영상 중에서 Fig. 15(f)는 패딩을 하지 않고 회전한 영상이므로 회전 후 발생하는 빈 픽셀 영역이 객체로 인식되어 올바를 결과를 얻을 수 없었다. Fig. 14(g)는 패딩한 후 회전한 영상이므로 건물들이 탐지되었으나 원래 영상과 비교하면 건물 형태가 다소 정확하지 않게 탐지되었다. 그러므로 회전에 영향을 받는다는 것을 알 수 있다.

Fig. 15는 영상의 화질을 변경한 영상들에 대한 결과를 보여주고 있다. Figs. 15(a), (b)와 (c)는 RGB 순서를 변경한 영상들이며, 전체적인 결과는 RGB 순서에는 큰 영향을 받지 않지만, 다소 다른 결과임을 알 수 있다. 그러므로 데이터 배열순서도 고려해야 할 사항이라고 사료된다. Figs. 15(d)와 (e)는 밝기값을 변경한 결과이며, 서로 같은 결과는 아니지만, 극단적인 경우를 제외하면 밝기 변화에는 큰 영향이 없는 것으로 판단된다.

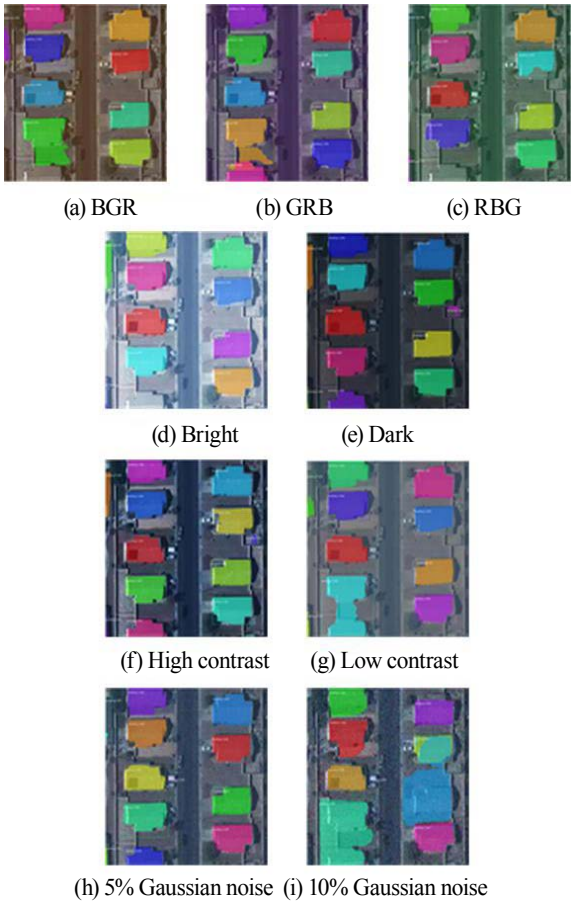


Fig. 15. Building detection with radiometrically degraded images

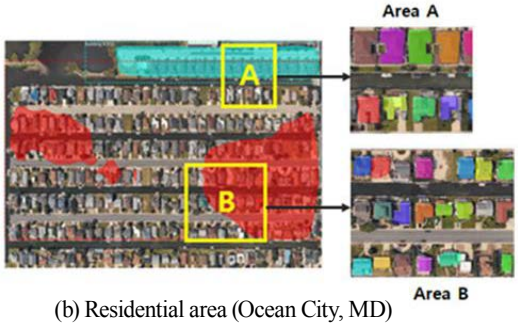
Figs. 15(f)와 (g)는 대조비를 변경한 결과이며, 밝기 변화와 마찬가지로 많은 차이는 발생하지 않았지만, 밝기값 변화와 대조비 변화 중에서는 대조비를 향상시킨 결과가 우수하다. 그러므로 좋은 결과를 얻기 위해서 대조비를 향상시키는 것도 고려해야 한다. Figs. 15(h)와 (i)는 노이즈 레벨을 다르게 부여하여 화질이 저하된 영상의 결과를 보여주고 있으며, 사람은 객체 식별이 가능한 노이즈 레벨에서도 만족할만한 결과는 아니다. 그러므로 노이즈가 심한 영상은 사전에 노이즈를 제거할 필요가 있다.

4.2 새로운 영상의 건물탐지

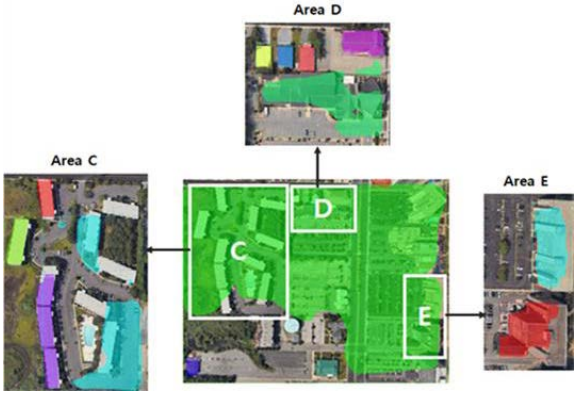
Fig. 16은 모델의 학습에 관여하지 않은 전혀 다른 2개 지역의 영상에서 건물을 추출한 결과를 보여주고 있다. 대상지역은 서울의 주택지와 미국의 주택지 및 도심지이며, 추출된 건물들은 서로 다른 색으로 표시되었다.

(i) Residential houses (ii) Apartments

(a) Residential area (Seoul, Korea)



(b) Residential area (Ocean City, MD)



(C) Urban area (Ocean City, MD)

Fig. 16. Building detection from unseen images

Fig. 16(a)는 서울의 빌라 건물과 아파트가 밀집된 지역으로 건물이 회전된 형태로 배열되어 있으며, 부분만 촬영된 건물은 제외하고 빌라지역은 총 18개의 건물 중 9개 건물이 탐지되었다. 아파트 단지는 10개 건물 중 9개의 건물의 대부분이 형태가 정확히 탐지되었다. Figs. 16(b)와 (c)는 미국 메릴랜드주 Ocean City의 주택지와 도심지 영상이다. Fig. 16(b)는 단독주택이 밀집된 거주지역의 영상 전체에서 건물을 추출한 결과와 두 개 지역(Area A와 Area B)을 선정하여 각 지역에 대한 결과를 보여주고 있다. 전체 영상의 경우는 건물을 탐지하지 못하였으나, 영상을 분할한 주택지역의 경우에는 일부 건물을 제외하고 대부분 정확하게 건물을 탐지하였다.

Fig. 16(c)는 도심지역 영상이며, 주택지와 마찬가지로 영상 전체와 지역을 선정하여 결과를 평가하였다. 전체 영상의 경우는 건물을 추출하지 못했다. 지역별로 나누어서 수행한 결과는 지역별로 차이는 있지만, 결과는 만족할 수준은 아니다. Area C는 일부 건물은 정확하게 탐지되었지만, 나머지 건물들은 개별적 건물로 탐지되지 않고 하나의 건물로 인식되거나 주변의 주차장까지 건물로 탐지되었다. 단독주택이 포함된 Area D지역에서는 작은 규모의 건물은 정확하게 탐지하였으나 다양한 형태의 건물들이 조합된 큰 건물(녹색으로 표시된 건물)은 올바르게 인식하지 못하였다. Area E는 고층건물이 포함된 지역의 결과이며, 건물 지붕은 탐지하지 못하고 벽면을 탐지하였다. 그러나 정사영상에서는 이런 문제는 발생하지 않는다.

사전학습 모델은 DL의 특성상 제한적인 조건에서 사용 가능하고, 학습 영상과 유사하지 않은 영상에서는 원하는 결과를 얻기 힘들다. DL의 일반화와 범용화를 실현하기 위해서 다양한 특성이 복합적으로 포함된 대량의 영상으로 학습을 수행하는 것은 효율적이지 못하다. 그러므로 지역적 및 객체의 특성을 고려하여 다양한 목적지향 사전학습 DL 모델베이스를 구축할 필요가 있다. 예를 들면, 영상을 입력하면 영상의 특성을 파악하여 최적의 모델을 자율적으로 결정할 수 있도록 신경망 내에 여러 특정 목적의 신경망을 구성하여(즉 ANN in ANN) 보다 성능과 기능이 강화된 DL 시스템이 필요하다.

4.3 향후 연구방안

CNN은 주로 광학영상을 학습시켜 DL을 수행하기 위해 제안된 신경망 모델이다. 그러므로 영상에 의한 DL의 학습 결과는 영상이 가지고 있는 특성에 영향을 받을 수밖에 없다. 더욱 신뢰성 높은 DL 모델이 되기 위해서는 광학영상에만 의존하지 않고 다양한 공간정보 데이터 (LiDAR 데이터, DSM (Digital Surface Model), 수치지형도 등)를 복합적으로 학습 데이터로 사용하는 것이 필요하다. 특히 높은 정확도의 3차원 공간정보를 제공하는 라이다 데이터는 객체인식 및 3D 객체 모델링에 중요한 역할을 할 수 있으므로 영상과 함께 사용하면 상호보완적으로 DL의 효율성과 성능을 향상시킬 수 있다고 판단된다. 향후에는 영상이외에도 다양한 센서에서 획득한 공간정보 데이터에 특화된 학습 알고리즘을 적용하여 범용성이 획기적으로 향상된 DL 시스템을 구현할 수 있을 것으로 사료된다. 현재 수준의 “DL 모델” 단계를 넘어서 속도 및 성능의 개선뿐 아니라 공간정보 분야에 전문화된 “공간정보 DL 시스템” 등장이 기대된다. 이를 실현하기 위해서는 DL 모델의 일부 기능 개선과 같은 좁은 의미의 성능 향상이 아니라 다양한 종류의 데이터와 정량적 및 정성적 정보를 학습할 수 있는 DL 시스템으로부터 구축된 전이학습에 의한 DL의 범용화를 실현하기 위한 연구와 개발이 필요하다.

5. 결론 및 토의

본 논문에서 적용한 Mask R-CNN은 현재까지 가장 우수한 DL 모델 중 하나라고 평가되고 있지만, 모델 학습에 사용된 영상도 변화에 따라 결과가 다르거나 정확하지 못한 경우가 발생한다. 이는 학습에 사용한 데이터 이외의 데이터를 적용하면 원하는 결과를 얻지 못할 수 있으며, 현재 DL의 한계점을 보여주고 있는 것이다. 특히 지도학습의 경우, 가중치를 업데이트하기 위해 기준이 되는 label 데이터를 생성하거나 획득하는 것도 쉬운 과정이 아니다.

제한된 데이터로 DL 모델의 성능을 평가하는 것은 한계가 있지만, 기하학적 변화에 민감하며 특히 노이즈에 취약하다. 모델을 학습시키기 위해 사용되는 데이터의 특성이 결과에 미치는 영향이 크다. 즉 학습 데이터에 종속되어 불가피하게 발생하게 되는 과적합 문제를 해결하는 것이 중요하다. DL의 목표인 일반화와 범용성을 위해 학습 데이터는 일관성과 다양성을 동시에 만족할 수 있어야 하는데, 이러한 특성의 데이터를 다수 획득하고 학습 시켜서 신뢰성 높은 결과를 얻는 것은 어려운 일이다.

최근에 DL은 많은 관심과 기대를 갖고 집중적으로 연구가 진행되는 분야이다. 인간의 신경망구조와 메커니즘을 모방한 ANN을 기반으로 하는 DL의 개념적 모델과 실제 구현된 DL 모델의 능력과 성능은 아직 많은 차이가 있다. 이를 극복하기 위해 성능을 향상시키려는 노력과 연구가 지속적으로 수행되어 여러 종류의 DL 모델이 제시되고 있다.

가중치(CNN의 경우 필터계수)를 업데이트하는 것으로 학습과정을 수행하는 것도 한계가 있다. 즉 역전파에 의한 가중치 업데이트 과정이 인간의 신경망의 원리이고 뇌의 학습과정이라는 현재의 DL은 능력면에서 다소 과대포장된 측면이 있다. 아직까지 표준이 되는 DL 모델은 존재하지 않고 대부분의 경우 시행착오를 통한 임시방편적인 방법에 의존하고 있으므로 DL은 한계에 도달할 수 있다. 그러므로 학습방법이 더욱 적응적이고 융통성이 강화된 범용적으로 활용할 수 있는 DL 모델에 대한 연구가 필요하다. 최근에는 뇌과학 분야와 연계한 ANN과 DL의 연구가 활발하기 시작하고 있으므로 더욱 진화된 DL 시스템이 개발될 것으로 기대한다.

감사의 글

이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2018R1D1A1B07048732)

References

Audebert, N., Le Saux, B., and Lefèvre, S. (2018), Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks, *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 140, pp. 20-32.

Back, C.S. and Yom, J.H. (2018), Comparison of point cloud volume calculated by artificial intelligence learning method and photogrammetric method, *Proceedings of Korean Society of Surveying, Geodesy, Photogrammetry and Cartography*, 19-20 April, Yongin, Korea, pp. 227-230.

Ball, J., Anderson, D., and Chan, C. (2017), A comprehensive survey of deep learning in remote sensing: Theories, tools and challenges for the community, *Journal of Applied Remote Sensing*, Vol. 11. No. 4, pp. 1-54.

Campos-Taberner, M., Romero-Soriano, A., Gatta, C., Camps-Valls, G., Lagrange, A., Le Saux, B., Beaupere, A., Boulch, A., Chan-Hon-Tong, A., Herbin, S., Randrianarivo, H., Ferecatu, M., Shimoni, M., Moser, G., and Tuia, D. (2016), Processing of extremely high-resolution LiDAR and RGB data: Outcome of the 2015 IEEE GRSS data fusion contest–Part A: 2-D contest, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 9, No. 12, pp. 5547-5559.

Choe, Y.J. and Yom, J.H. (2017), Downscaling of MODIS land surface temperature to LANDSAT scale using multi-layer perceptron, *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography*, Vol. 35, No. 4, pp. 313-318. (in Korean with English abstract)

Chung, D. and Lee, I. (2017), Point cloud classification base on deep learning, *Proceedings of Korean Society of Surveying, Geodesy, Photogrammetry, and Cartography*, Yeosu, Korea, pp. 110-113. (in Korean with English abstract)

Deng, Z., Sun, H., Zhou, S., Zhao, Lei, L., and Zou, H. (2018), Multi-scale object detection in remote sensing imagery with convolutional neural networks, *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 145, pp. 3-22.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., and Garcia-Rodriguez, J. (2017), A review on deep learning techniques applied to semantic

segmentation, arXiv:1704.06857.

Girshick, R. (2015), Fast R-CNN, *IEEE International Conference on Computer Vision*, *ICCV 2015*, 13-16 December, Santiago, Chile, pp. 1440-1448.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2016), Region-based convolutional networks for accurate object detection and segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 38, No. 1, pp. 1-16.

Hazirbas, C., Ma, L., Domokos, C., and Cremers, D. (2016), FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture, *Proceedings of the Asian Conference on Computer Vision*, Vol. 2, 20-24 November, Taipei, Taiwan.

He, k., Gkioxari, G., Dollár, p., and Girshick, R. (2017), Mask R-CNN, *Proceedings of IEEE International Conference on Computer Vision (ICCV) 2017*, 22-29 October, Venice, Italy, pp. 2980-2988.

Hertz, J., Krogh, A., and Palmer, R. (1991), *Introduction to the Theory of Neural Computation*, Addison-Wesley, Reading, MA, 327p.

Kang, J., Körner, M., Wang, Y., Taubenböck, H., and Zhu, X. (2018), Building instance classification using street view images, *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 145, pp. 44-59.

Kemker, R., Salvaggio, C., and Kanan, C. (2018), Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning, *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 145, pp. 60-77.

Kim, H. and Bae, T., (2017), Preliminary study of deep learning-based precipitation prediction, *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry, and Cartography*, Vol. 35, No. 5, 423-430.

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012), ImageNet classification with deep convolutional neural networks, *Proceedings of the 25th International Conference on Neural Information Processing Systems*, Vol. 1, 3-8 December, Lake Tahoe, Nevada, pp. 1097-1105.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R. Hubbard, W., and Jackel, L. (1989), Backpropagation applied to handwritten zip code recognition. *Neural Computation*, No. 1, Vol. 4, pp. 541-551.

Lee, G. and Yom, J.H. (2018), Design and implementation of web-based automatic preprocessing system of remote sensing imagery for machine learning modeling, *Journal of the Korean Society for Geospatial Information Science*, Vol. 26 No. 1, pp. 61-67. (in Korean with English abstract)

Long, J., Shelhamer, E., and Darrell, T. (2015), Fully convolutional networks for semantic segmentation, *Proceedings of IEEE Conference on Computer Vision and Patton Recognition*, 7-12 June, Boston, MA, pp. 3431–3440.

Marmanis, D., Wegner, J., Galliani, S., Schindler, K., Datcu, M., and Stilla, U. (2016), Semantic segmentation of aerial images with an ensemble of CNNS, *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. 3-3, XXIII ISPRS Congress*, 12-19 July, Prague, Czech Republic, pp. 473-480.

Maturana, D. and Scherer, S. (2015), 3D Convolutional neural networks for landing zone detection from LiDAR, *IEEE International Conference on Robotics and Automation*, Seattle, Washington, 26-30 May, pp. 3471-3478.

McCulloch, W. and Pitts, W. (1943), A logical calculus of the ideas immanent in nervous activity, *Bulletin of Mathematical Biophysics*, Vol. 7, pp. 115-133.

Oh, H. (2010), Landslide detection and landslide susceptibility mapping using aerial photos and artificial neural networks, *Korean Journal of Remote Sensing*, Vol. 26, No. 1, pp. 47-57. (in Korean with English abstract)

Pang, Y., Sun, M., Jiang, X., and Li, X. (2018), Convolution in convolution for network in network, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 29, No. 5, pp. 1587-1597.

Parthasarathy, D. (2017), A brief history of CNNs in image segmentation: From R-CNN to Mask R-CNN, https://blog.athelas.com/a-brief-history-of-cnns-in-image-segmentation-from-r-cnn-to-mask-r-cnn-34ea83205de4 (last date accessed: 6 September 2018).

Ren, S., He, K., Girshick, R., and Sun, J. (2017), Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 6, pp. 1137-1149.

Rosenblatt, F. (1958), The perceptron: A probabilistic model for information storage and organization in the brain, *Psychological Review*, Vol. 65, No. 6, pp. 386-408.

Rumelhart, D., Hinton, G., and Williams, R. (1986), Learning internal representations by back-propagating errors, *Nature*, Vol. 323, No. 9, pp. 533-536.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang Z., Karpathy, A., Khosla, A., Bernstein, M., and Berg, A. (2015), Imagenet large scale visual recognition challenge, *International Journal of Computer Vision*, Vol. 115, No. 3, pp. 211-252.

Schenk, T. (1999), *Digital Photogrammetry: Volume 1*, TerraScience, Laurelville, OH, 428p.

Shaikh, F. (2018), Automatic image captioning using deep learning (CNN and LSTM) in PyTorch, *Analytics vidhya,* https://www.analyticsvidhya.com/blog/2018/04/solving-an-image-captioning-task-using-deep-learning/ (last date accessed: 31 October 2018).

Simard, P., Steinkraus, D., and Platt, J. (2003), Best practices for convolutional neural networks applied to visual document analysis, *Proceedings of the Seventh International Conference on Document Analysis and Recognition, ICDAR 2003*, 3-6 August, Vol. 2, pp. 958–962.

Tokarczyk, P., Wegner, J., Walk, S., and Schindler, K. (2015), Features, color spaces, and boosting: new insights on semantic classification of remote sensing images, *IEEE Transactions on Geoscience And Remote Sensing*, Vol. 53, No. 1, pp. 280-295.

You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. (2016), Image captioning with semantic attention, *IEEE Conference on Computer Vision and Pattern Recognition*, 26 June-1 July, Las Vegas, Nevada, pp. 4651-4659.

Vo, A.V., Truong-Hong, L., Laefer, D., Tiede, D., d'Oleire-Oltmanns, S., Baraldi, A., Shimoni, M., Moser, G., and Tuia, D. (2016), Processing of extremely high resolution LiDAR and RGB Data: Outcome of the 2015 IEEE GRSS data fusion contest—Part B: 3-D Contest, *IEEE Journal of Selected Topics In Applied Earth Observations And Remote Sensing*, Vol. 9, No. 12, pp. 5560-5575.

Wang, S., Quan, D., Liang, X., Ning, M., Guo, Y., and Jiao, L. (2018), A deep learning framework for remote sensing image registration, *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 145, pp. 148-164.

Xing, Y., Wang, M., Yang, S., and Jiao, L. (2018), Pan-sharpening via deep metric learning, *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 145, pp. 165-183.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. (2015), Show, attend and tell: Neural image caption generation with visual attention, *International Conference on Machine Learning*, 6-11 July, Lille, France, pp. 2048-2057.

Zhang, B., Gu, J., Chen, C., Han, J., Su, X., Cao, X., and Liu, J. (2018), One-two-one networks for compression artifacts in remote sensing, *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 145, pp. 184-196.