합성곱 신경망 기반 객체인식 분류기술 동향

임대성
세종대학교 지능기전공학부 자율운항선박실험실
e-mail : *fd9317@daum.net*

Trends of Object Detection and Object Categorizing Technique on Convolutional Neural Network

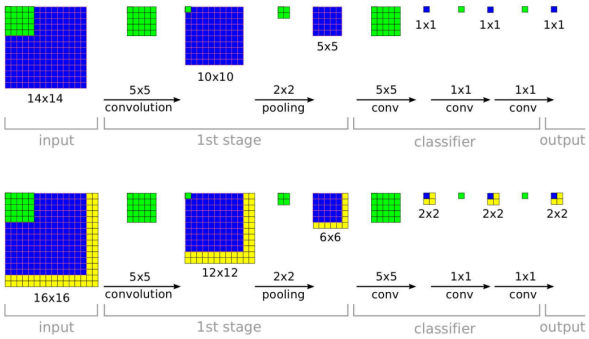
Daeseong Lim
Department of Intelligent Mechatronics Engineering
Autonomous Shipping Lab
Sejong University

I. 서론

자율주행 자동차, 자율운항 선박 등 인간의 개입 없이 스스로 길을 찾고 사고 없이 운행할 수 있도록 하는 자율지능 무인이동체 기술은, 일반인도 기술 산업 발전을 피부로 느낄 수 있을 만큼, 상용화되고 있으며, 동시에 촉망받는 기술이기도 하다. 인간이 길을 걸을 때, 도로를 따라 걷고, 가로수, 행인을 피하는 등의 자연스러운 보행 행위는 대부분 눈으로 본 것을 토대로 하여 수행하게 된다. 이와 마찬가지로, 무인이동체의 경우에도 인간의 눈 역할 즉, 카메라를 통해 들어오는 정보를 주력으로 하여 이동, 회피 행위를 결정하게 된다. 무인이동체에 있어서는, 컴퓨터 비전 기술이 수반되지 않는 무인이동체는 눈을 감고 걷는 인간과 다름이 없다는 것이다. 이러한 중요성을 기반으로 하여, 컴퓨터 비전 기술의 객체 인식 분야 및 객체 분류 분야의 연구 기술 동향에 대해 알아보고자 한다.

과거의 Object Detection 연구는 SIFT(Scale Invariant Feature Transform)[1], SURF(Speeded-Up Robust Feautres)[2], Haar[3], HOG(Histogram of Oriented Gradients)[4] 등과 같이 객체가 가지는 여러 특징, 예를 들어 형상이 가지는 각도, 길이 등을 분석하여 객체 요소를 분리하고 특정하는 방식을 사용했다. 객체 특징을 어떻게 분석하여 영상 내에서 객체를 인식하고 검출할 것인가에 대해 중점을 눈 연구인 것이다. 하지만 합성곱 신경망(CNN:Convolutional Neural Network)이 ImageNet 2012 대회에서 기존 성능을 압도적으로 뛰어넘는 결과를 보여주면서, 딥러닝을 이용한 Object Detection 방법이 주류로 떠오르게 되었다. CNN은 LeCun 교수의 필기체 숫자 인식[5]에서 처음 등장하였는데, 기존의 신경망에서는 픽셀 주위의 지역적인 정보를 표현하지는 못했으나, 합성곱 연산을 통해 이를 극복했다. 이어서, CNN의 인식률을 향상시키기 위해 신경망을 더 깊게 구성하는 방식으로 연구가 진행 되었다.

오늘날 물체 탐지에 주로 사용되는 RADAR(Radio Detection and Ranging), LIDAR(Light Detection and Ranging)와 같은 센서는 비전 센서에 비해 비교적 낮은 정밀도를 가지는 반면, 높은 가격으로 인해 활용적 한계가 존재하며, 비전 센서 기술의 발전에 따라, 비전 센서를 활용한 객체 인식 및 분류 기술의 연구가 활발히 진행 중 에 있다. 하지만 비전기반의 객체 인식 기술은 공통적으로, 각 픽셀에서 들어오는 데이터를 활용도에 맞게 각각 연산해야 하는 탓에 연산량이 높다는 문제점이 있어, 고성능 CPU나 고용량 배터리의 탑재가 어려운 소형 장비에 적용이 어렵다. 따라서 본문에서는 컴퓨터 비전 기반 객체인식 기술의 발전 중에서도, 적은 연산량을 통해서도 일정 수준 이상의 object 검출 성능을 낼 수 있는 성능최적화 이론에 대한 분야를 중점적으로 다루고자 한다.



[그림]OverFeat[10] 신경망 시각화 모델

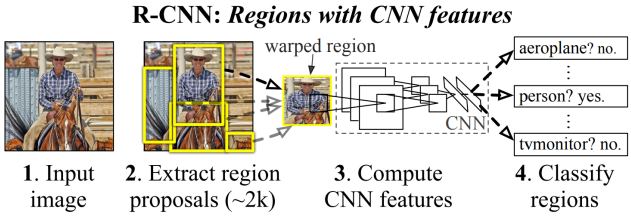
II. CNN

합성곱 신경망(CNN)은 흔히 알려진 딥러닝 아키텍쳐로, 동물의 시각정보 인식 과정으로부터 영감을 받아 설계되었다. 1959년 Hubel, Wiesel는 고등 포유동물의 시각 피질의 뉴런이 시각적 형태를 어떻게 부호화하는지에 대해 많은 정보를 제공했고[6], 신경과학 연구에서 영감을 받아 Neocognitron 이라는 원형 CNN이 고안되었다[7]. 이후 CNN을 이용하기 위한 많은 연구들이 수행되었으나, CNN과 딥러닝은 학습 데이터셋의 부족과 당시 연산처리장치의 성능 한계로 인해 오랜 시간동안 주목받지 못했던 것이 현실이다. 이어 2000년대에 들어서 GPU를 비롯한 병렬 연산처리장치의 발전으로 인해 CNN은 다시 주목을 끌기 시작했다. 특히 2010년대에 들어서, ImageNet과 같은, 대규모 데이터셋의 공개와 함께 CNN은 특정한 규칙, 법칙을 기반으로 한 인식기법들과는 대비되게, 데이터 간의 연동성을 알고리즘을 통해 스스로 학습하게 하는 과정을 거쳐, 크게 향상된 성능 결과를 나타내며 학계와 산업계를 넘어 대중들에게까지 큰 관심을 받게 되었다. 이를 무인이동체 제어기술에 접목하기 위해, 성능최적화 관점에서의 합성곱 신경망의 빠른 처리에 대한 연구 동향에 대해 알아보고자 한다.

2.1 객체인식

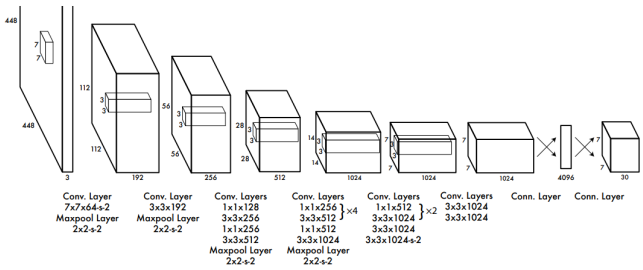
합성곱 신경망이 ImageNet 챌린지에서 큰 성공을 거둔 2012년 이후로, CNN은 객체인식 분야로 크게 성장했으며, 객체인식을 위한 대규모 데이터셋 Pascal VOC[8] MSCOCO[9]도 공개되었다.

객체인식 기술은 CNN기술개발 초기의 비효율적 구조를 개선하여, 중복연산을 줄이면서도, 각 데이터가 가지는 특성은 보존할 수 있도록 발전해 왔다.



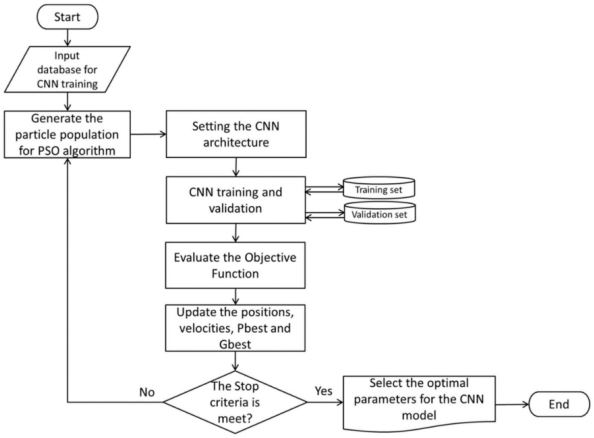
[그림]R-CNN[11] CNN features

대표적인 객체인식의 방식으로는 OverFeat, R-CNN 두 가지의 객체인식 방식을 예로 들 수 있는데, OverFeat는, 영상 피라미드를 CNN을 이용하여 특징점을 추출하는 방식으로 객체의 위치를 인식하는 방식[10]이며, 모든 레이어가 컨볼루션 방식으로 적용되어 특징점을 추출하는 방식이다. 반면 R-CNN은 별도의 객체영역 제안 알고리즘으로, selective search를 이용하여 제안된 영역들을 각각 CNN을 사용하여 영상분류 하는 방식이다[11]. 이 두가지 방식을 절충하여 영상을 CNN입력으로 넣어 얻어 낸 특징맵 상에서 다시 selective search에서 얻어진 영역을 추출하여 영상 피라미드를 이용해 영역 크기를 동일하게 제한하고, fully-connected layer를 통과하는 방식을 제안하여[12] 계산속도 측면에서 큰 발전을 이루어 냈다. 하지만, 여전히 실시간에 가까운 처리 속도를 필요로 하는 로봇, 자율주행 등의 응용 분야에 적용하기에는 처리 속도가 충분하지 못했고, 이러한 속도 문제를 해결하기 위해 객체 인식의 모든 과정을 하나의 Deep Learning Network로 구성하는 방법인 YOLO(You Only Look Once)방법이 제안되었다[13]. 최근에는 모바일 기기에서도 동작 가능한 정도의 빠른 검출속도를 보이는 방법인 SSD(Single Shot Multibox Detector) 등도 제안되고 있다[14].



[그림]YOLO[13] Convolution layer figure

이어서 최근에는 군집 최적화 알고리즘 PSO(Particle swarm optimization algorithm)을 이용한 신경망 아키텍처를 설계하는 접근 방식이 제안되었다. PSO알고리즘을 사용하여, 컨볼루션 레이어의 수, 컨볼루션 과정에 사용한 필터의 크기, 컨볼루션 신경망의 최적 매개 변수를 찾는 것을 목적으로 한 알고리즘이며 최적화된 알고리즘은 멕시코, 미국 수화알파벳, MNIST 세가지 데이터베이스에서 99%이상의 인식률을 나타냄으로 다른 접근법들에 비해서도 호의적인 결과를 가져온 것으로 나타났다[15].



[그림]PSO를 사용한 CNN 최적화 프로세스 순서도[15]

III. 결론

현재의 객체 인식 기술은 CNN을 기반으로 하여 충분히 만족스러운 객체 인식 정확도 성과를 내고있는 것이 사실이다. 하지만 일반적인 객체 인식의 문제에 있어서는 아직 좋은 결과를 보이지 못하는데, 이는 CNN 구조 자체의 문제로도 볼 수 있는데, 정확도를 높이기 위해 더 깊은 CNN모델을 구현하면, 과적합이나 Gradient 소실 등의 문제가 발생하기 때문이다. 또한 CNN 내부의 하이퍼 파라미터들은 내부적 의존성을 가지고 있어 튜닝하기 쉽지 않은데, 하이퍼파라미터를 조절하기 위해 PSO알고리즘을 사용하는 연구에 대해서도 알아볼 수 있었다[15].

마지막으로 여전히 CNN은 벤치마크에서 뛰어난 정확성과 높은 성능을 보여주고 있지만, 어떻게 이런 높은 성능을 낼 수 있었는지에 대해 기반할 수 있는 이론연구는 아직 미흡한 단계에 있다. CNN에 대한 이해를 높이고 여러 연구 분야에서 CNN이 다양하게 활용될 수 있도록, 합성곱 신경망 모델 자체에 대한 수학적 이해와 수학적 해석을 통한 신경망 개념의 확장 및 정의가 필요할 것으로 보인다.

참고문헌

[1] Lowe, D. G. (2004). "Distinctive image features from scale-invariant keypoints." International journal of computer vision 60(2): 91-110.

[2] Bay, H., et al. (2008). "Speeded-up robust features (SURF)." Computer vision and image understanding 110(3): 346-359.

[3] Viola, P. and M. Jones (2001). Rapid object detection using a boosted cascade of simple features. Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001, Ieee.

[4] Dalal, N. and B. Triggs (2005). Histograms of oriented gradients for human detection. 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), Ieee.

[5] LeCun, Y., et al. (1998). "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86(11): 2278-2324.

[6] Hubel, D. H. and T. N. Wiesel (1968). "Receptive fields and functional architecture of monkey striate cortex." The Journal of physiology 195(1): 215-243.

[7] Fukushima, K. and S. Miyake (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. Competition and cooperation in neural nets, Springer: 267-285.

[8] Everingham, M., et al. (2015). "The pascal visual object classes challenge: A retrospective." International journal of computer vision 111(1): 98-136.

[9] Lin, T.-Y., et al. (2014). Microsoft coco: Common objects in context. European conference on computer vision, Springer.

[10] Sermanet, P., et al. (2013). "Overfeat: Integrated recognition, localization and detection using convolutional networks." arXiv preprint arXiv:1312.6229.

[11] Girshick, R., et al. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition.

[12] He, K., et al. (2015). "Spatial pyramid pooling in

deep convolutional networks for visual recognition." IEEE transactions on pattern analysis and machine intelligence 37(9): 1904-1916.

[13] Redmon, J., et al. (2016). You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition.

[14] Liu, W., et al. (2016). Ssd: Single shot multibox detector. European conference on computer vision, Springer.

[15] Fregoso, J., et al. (2021). "Optimization of Convolutional Neural Networks Architectures Using PSO for Sign Language Recognition." Axioms 10(3): 139.