딥러닝을 이용한 영상내 물체 인식 기법

지능형 시스템의 수요가 증가하면서 영상인식의 중요성이 부각되고 있다. 사람이 직접 물체 인식 과정을 모델링하는 방식을 넘어 최근에는 기계학습을 이용하여 이를 자동화하는 방법이 주를 이루고 있다. 그 중 딥러닝은 빅데이터를 활용하는 각종 분야에서 놀라운 성능을 보이며 기계학습 수준을 한 단계 진화시킨 기술로 평가 받고 있으며 영상 인식의 다양한 분야에서 응용되고 있다. 본 글에서는 딥러닝을 이용한 물체 검출 기법의 동향을 살펴보고 이를 차량 전면부 인식에 적용한 사례를 소개한다.

■ 박제강, 박용규, 온한익, 강동중*
(부산대학교 기계공학부)

Ⅰ. 서론

수년 전부터 스스로 판단하고 행동하는 지능형 시스템에 대한 관심이 급격하게 증가하고 있다. 구글, 페이스북과 같은 IT분야의 유명 기업들이 관련 기술을 확보하기 위해 경쟁하고 있으며 기계학습을 아이템으로 하는 스타트업 기업들이 대거 등장하고 있다. 국가적 차원에서도 다양한 방법으로 지능형 시스템의 연구를 장려하고 있다. 이미 미국은 2011년 네바다 등 5개 주에서 자율주행자동차의 일반도로 주행을 부분 허용했으며 최근에는 안전운전 관리자가 동반하지 않는, 즉 완전 무인자동차의 시험 주행도 허용하였다. 국내에서도 ADAS를 탑재한 스마트카가 출시되었으며 자율주행 자동차 경진대회가 매년 개최되고 있다.

지능형 시스템은 다양한 센서를 활용한 주변 상황의 인지에서 시작된다. 전방에 장애물이 있을 때 차량은 이를 피하거나 정지해야 하는데 애초에 장애물의 유무를 인지하지 못하면 정지 명령을 지시할 수 없다. 일반적으로 전방 장애물 감지를 위해 초음파(sonar)센서 또는 레이저(lidar)센서를 사용하는데 이 방법은 센서 앞쪽에 물체가 있다는 사실은 알려주지만 그 물체가 앞서 달리는 차량인지 보행자인지는 구별하기는 쉽지 않다. 반면 카메라 센서는 인간의 시각과 유사한 기능을 수행하므로 시스템이 물체의 종류까지 인식 가능하다. 그림 1은 수백만 장의 영상 인식 문제를 다루는 ILSVRC(ImageNet Large Scale Visual Recognition Challenges)[1]의 영상 샘플을 보여준다. 영상에 포함된 1,000가지 종류의 물체를 인식하는데 같은 종류의 물체라도 촬영 당시 카메라의 각도, 조명, 거리의 변화 또는 물체의 이동, 회전, 변형 등의 조건에 따라 다양한 형상을 가지기 때문에 2010년 당시 최고 인식률은 71.8%였다. 2015년 현재 최고 인식률은 95.2%[2]로 인간의 인식률로 알려진 94.9%[1]를 능가하였다.

딥러닝(deep learning)은 기계학습 알고리즘의 한 종류이다. 전통적 기계학습은 샘플 데이터로부터 인식기의 파라미터를 스스로 학습하는 알고리즘을 의미한다. 딥러닝의 등장은 컴퓨터가 인간의 인식 능력을 능가하게 된 계기로 꼽히며 실제로 딥러닝의 등장을 기점으로 기계학습을 사용하는 시스템들의 성능이 대폭 향상되었다. 최근 영상이나 음성신호 인식에서 인간의 능력을 상회하는 성능을 보이는 학습 알고리즘은 대부분 딥러닝을 기반으로 한다.

본 글에서는 영상 인식 시스템의 기본 개념과 함께 딥러닝을 영상 인식에 활용한 최신 동향 및 실제 문제에 적용한 사례를 소개한다.

그림 1. 수백만 장의 영상을 학습하고 분류하는 문제를 도전하는 ILSVRC[1].

Ⅱ. 관련 연구

2.1 영상인식

영상인식(image recognition)은 영상 데이터에서 의미 있는 정보(영상에 포함된 특정 물체의 종류, 위치, 자세 등)를 추출하는 문제를 뜻한다. 앞서 언급했듯이 영상은 조명 변화나 주변 환경에 영향을 많이 받으며 물체는 관찰각에 따라 형상이 다양하기 때문에 이를 수학적으로 모델링 하기 쉽지 않다. 일반적으로 물체를 표현하는 특징을 추출하기 위해 경계(edge) 혹은 경계의 히스토그램(histogram)을 계산하거나 물체의 요소(e.g.사람의 경우 머리, 몸통, 팔, 다리 등)를 개별적으로 모델링하고 요소 사이의 관계를 정의하는 파트 모델 등이 사용된다. 추출된 특징을 모델과 비교하여 물체 여부를 판단하는 역할은 기계학습 알고리즘이 담당한다.

기존 연구의 방향은 물체를 표현하는 더 좋은 모델을 찾는 것이었다. 전통적 기계학습 알고리즘은 복잡한 영상을 그대로 학습할 수 없었고 물체를 잘 표현하는 특징 벡터를 사용자가 정의하고 추출해야 학습이 가능했다. 때문에 특징 벡터를 추출하는 다양한 학습모델이 제안되었고 딥러닝이 등장하기 전까지 인식률이 높은 대표적인 방법으로는 HOG 특징[3]과 물체를 구성요소로 나누어 SVM 분류기를 기반으로 모델링하는 파트 모델[4](part model) 등이 있다.

2.2 딥러닝

딥러닝은 깊은 신경망(DNN: Deep Neural Network) 알고리즘과 이를 학습하는 방법을 의미한다. 깊은 신경망은 입력층과 출력층을 제외한 은닉층이 2개 이상인 구조의 신경망으로 1980년대에 처음 제안되었으나 학습에 오랜 시간이 걸리고 학습 데이터에 과적합(overfitting)되는 단점 때문에 일반적인 문제에 사용할 수 없었다. 이러한 문제들은 2000년대 이후 병렬 연산이 가능한 GPU (Graphics Processing Unit)의 대거 등장과 과적합을 방지할 수 있는 기법[5][6]이 제안되며 해결되었다. 이후 딥러닝은 급속도로 발전하였고 현재는 음성인식, 장면인식, 영상복원 등의 다양한 분야에서 딥러닝이 사용되고 있다.

그림 2는 딥러닝의 한 종류로 영상인식에 주로 사용되는 컨볼루션 신경망[11](Convolutional Neural Network: 이하 CNN)의 구조를 나타낸다. 층 사이의 노드 쌍들 중 일부만 연결하는 컨볼루션 층(convolution layer)과 다운샘플링 층(pooling layer)이 전체 신경망의 앞쪽에 위치하며 층 사이의 노드를 모두 연결하는 완전연결 층(fully connected layer)이 신경망의 뒤쪽에 위치한다. 신경망을 학습할 때는 출력 층 뒤에 손실 층(loss layer)이 추가로 위치하여 신경망의 파라미터들을 학습시킨다. 그림 내에서 층과 층 사이의 연결선들은 CNN이 학습 시에 결정해야 하는 내부의 파라메터들로 노드 연결 가중치(weight) 값들이다. 가중치 값으로 마스크를 만들어 컨볼루션 연산을 수행한다. 이때 CNN을 분류기(classifier)로 학습시킬 때는 로그 손실(log loss)을 사용하며 회귀(regressor)로 학습시킬 때는 제곱 오차(euclidean loss)를 사용한다.

CNN의 주목할 점은 입력 층으로 특징 벡터가 아닌 영상 데이터가 그대로 입력된다는 점이다. 입력된 영상은 컨볼루션 층을 통과하며 특징이 추출되고 완전연결 층을 통해 분류된다. 즉 물체의 형상을 직접 모델링 하지 않아도 CNN이 알아서 특징 추출과 분류 역할을 모두 수행하는 것이다. 때문에 딥러닝이 본격적으로 알려지기 시작하면서 특징 분석과 모델링에 시간을 투자하던 기존의 연구 방식이 주어진 데이터를 더 잘 학습하는 신경

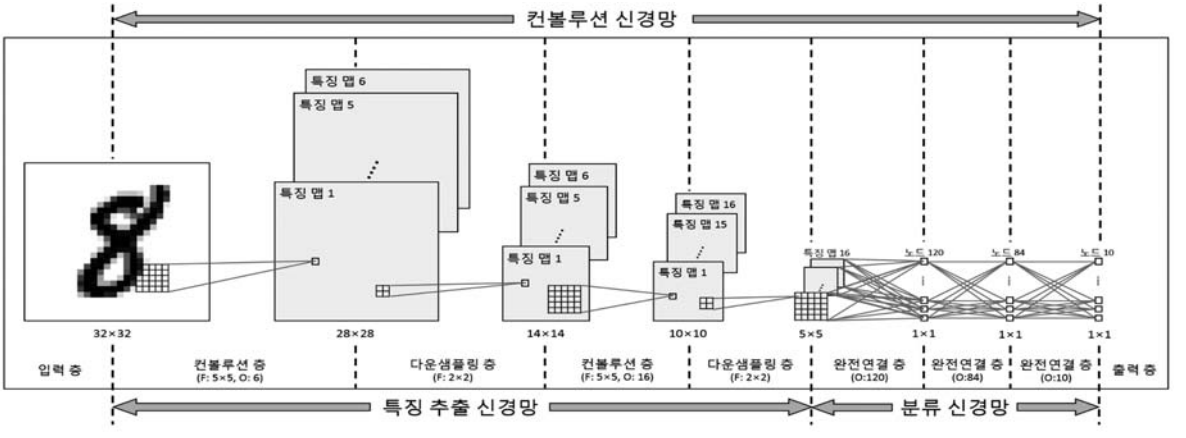

그림 2. CNN의 기본 구조[11]. F는 마스크의 크기, O는 출력 노드의 개수, $h \times w$는 각각 특징 벡터의 높이와 너비를 의미한다.

망의 구조를 연구하는 방향으로 변화하였다.

2.3 물체검출(object detection)

검출은 인식의 한 범주에 속하며 영상에 포함된 특정 물체의 위치와 종류를 추정하는 문제를 의미한다. 영상 내에서 물체는 임의의 위치에 존재하므로 물체의 위치를 추정하는 탐색과정이 먼저 필요하다. 가장 정확한 방법은 슬라이딩 윈도우(sliding window) 방식으로 분류기로 영상을 처음부터 끝까지 쭉 훑으면서 각 부분을 배경과 물체로 분류하는 것이다. 이 방법은 전역탐색(exhaustive search)이라고 부르며 정확도가 높으나 계산량이 많아 시간이 오래 걸리는 단점이 있다. 다른 방법으로는 영역분할법 등을 이용하여 먼저 물체가 있을 법한 위치를 추정 후 해당 위치만 분류기로 분류하는 방법[8]이 있다.

영상의 전역탐색을 CNN의 구조적 특성을 이용해 효과적으로 수행하는 방법[9]이 있다. 일반적으로 CNN은 고정된 크기의 입력 데이터를 사용한다. CNN의 후반부에 위치한 완전연결 층이 고정된 크기의 입력을 받기 때문이다. 반면 컨볼루션 층은 입력 데이터의 크기에 따라 출력 데이터의 크기가 결정된다. 때문에 완전연결 층을 마스크 크기가 1인 컨볼루션 층으로 대체하면 입력 데이터의 크기 제한에서 자유로워진 CNN을 학습할 수 있다. 이러한 구조의 CNN은 학습에 사용한 영상보다 큰 영상을 입력하면 영상을 전역 탐색한 결과와 정확하게 동일한 결과를 출력한다.

Ⅲ. 차량 전면부 검출

본 글에서는 DNN의 일종인 CNN을 영상 내 차량 검출에 적용한 사례를 살펴본다. 이를 통해 사용자의 특징 추출기, 분류기 설계 등의 상세한 과정이 없이도 입력 영상만으로 직접 물체 검출이 가능함을 보여준다[10].

차량 검출 시스템은 주정차 감시시스템, 주차 위치 알림 시스템과 같은 여러 분야에서 활용 가능하며 특히 차량의 전면부 검출은 번호판 인식을 위한 후보 영역 제공과 같은 전처리 시스템으로도 활용할 수 있다.

3.1 CNN 구조

그림 3은 차량 전면부 검출에 사용한 CNN의 구조를 보여주며 Alexnet[7]과 OverFeat[9]을 참고하여 구성하였다. 먼저 첫 번째에서 세 번째 컨볼루션 층은 입력 영상의 해상도를 줄임과 동시에 차량 전면부를 나타내는 주요 특징을 추출한다. 때문에 11×11 또는 5×5와 같은 비교적 크기가 큰 마스크를 사용한다. 네 번째에서 여섯 번째 컨볼루션 층은 회전에 불변한 특징을 추출한다. 어파인(affine) 변환행렬 역할을 하는 3×3크기의 마스크를 사용하며 다수의 층을 사용할수록 더 복잡한 변환을 수행할 수 있다. 차량 전면부는 큰 변형이 발생하지 않는 강체(rigid body)이므로 3개의 컨볼루션 층을 사용한다. 지금까지 언급한 6개의 컨볼루션 층은 입력 영상에서 특징을 추출하게 되며 특징 추출 신경망이라 부른다. 그림 3의 구조에서 특징 추출 신경망은 9×4×128개의 특징 벡터를 출력한다.

특징 추출 신경망 이후에는 분류 신경망과 위치 추정 신경망이 함께 위치하며 그림 3의 전단부에서 특징 벡터를 추출한 이후에 두 갈래로 나뉘는 구조를 가진다. 특징 추출 신경망 이후 위쪽이 차량 전면부와 배경을 구분하는 분류 신경망이며 아래쪽이 탐색된 윈도우의 위치를 보정하는 위치 추정 신경망이다. 두 신경망은 거의 동일한 구조를 가지며 완전연결 층 역할을 하는 1×1크기의 마스크를 사용하는 컨볼루션 층으로 구성되어

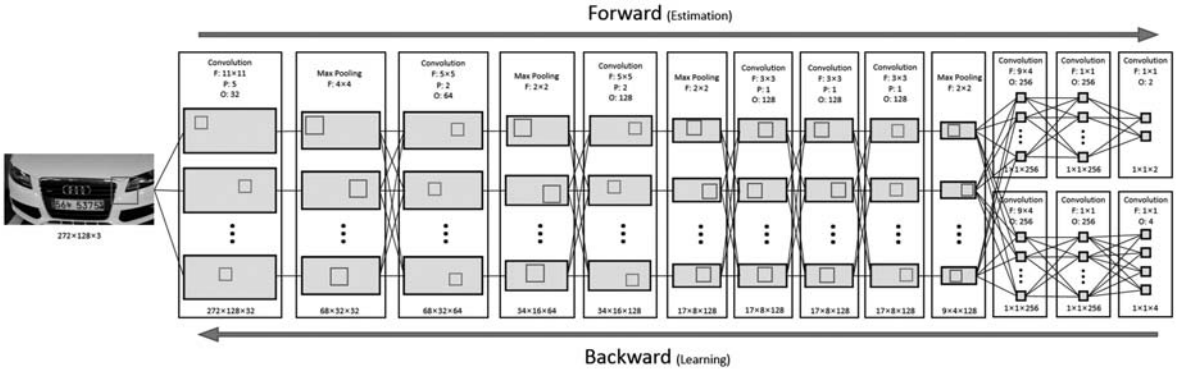

그림 3. 차량 전면부 검출에 사용한 CNN의 세부 구조 및 파라미터(10). F는 마스크의 크기, P는 패딩(padding), O는 출력 노드의 개수, $h \times w \times c$는 각각 특징 벡터의 높이, 너비, 채널을 의미한다.

있다. 이는 영상의 전역 탐색을 효율적으로 수행하기 위함이며 2.3절에서 언급하였다. 두 신경망의 구조적 차이점은 출력 노드의 개수이다. 분류 신경망은 입력 영상이 차량 전면부와 배경일 확률을 각 노드로 출력하기 때문에 출력 노드의 개수가 2개이다. 반면 위치 추정 신경망은 그림 4와 같이 탐색 윈도우 모서리의 오프셋 벡터를 출력해야 하므로 4개의 출력 노드를 가진다.

3.2 CNN 학습

학습해야 할 신경망은 총 3개로 각각 특징 추출, 분류, 위치 추정 신경망이다. 여기서 분류 신경망과 위치 추정 신경망이 공유하는 특징 추출 신경망은 분류 신경망과 함께 학습하며 입력 영상에 대해 배경과 차량 전면부 영상을 이진 분류하는 문제에 최적화된 특징을 학습하게 된다. 첫 번째 CNN은 특징 추출 신경망과 분류 신경망으로 구성되며 학습을 위한 목적 함수는 다음과 같다.

$$E = -\sum_j t_j \log\left(y_j\right) \tag{1}$$

$$y_j = \frac{e^{x_j}}{\sum_i e^{x_i}} \tag{2}$$

(2)에서 x_i는 출력 노드의 값이며 i와 j는 출력 노드의 인덱스를 의미한다. y_j는 소프트맥스(softmax) 혹은 정규화된 지수함수(normalized exponential function)라고 하며 부류에 대한 확률 분포를 정의한다. t_j는 교사값(teaching value)으로 학습에 사용되는 각 샘플의 부류 정보를 나타내며 j는 학습에 사용한 샘플의 인덱스를 의미한다. (1)은 학습 샘플을 오분류한 경우 손실을 증가시키는 의미를 가진다. 때문에 (1)이 최소값을 가지도록 신경망 파라미터를 학습해야 한다. 학습에는 오차 역전파(error back propagation)알고리즘을 사용한다. 오차 역전파 알고리즘은 입력 데이터가 출력값으로 계산되는 방향을 전방(feed forward)이라고 했을 때 그 반대 방향(feed backward)으로 손실을 전달하며 전체 신경망 파라미터를 학습하는 방법이다.

두 번째 CNN은 특징 추출 신경망과 위치 추정 신경망으로 구성된다. 앞선 학습에서 특징 추출 신경망에는 차량 전면부를 표현하는 최적의 특징이 학습되어 있고 위치 추정 신경망의 목적은 추출된 특징의 위치를 추정하는 것이므로 특징 추출 신경망은 더 이상 학습하지 않고 앞서 학습된 파라미터를 그대로 사용한다. 때문에 두 번째 CNN은 위치 추정 신경망만 학습한다. 위치 추정 문제는 회귀분석(regression) 문제로 모델링 할 수 있으며 따라서 목적 함수는 최소 제곱 오차로 정의한다.

$$E = \frac{1}{2}\left(t_j - y_j\right)^2 \tag{3}$$

(3)에서 t_j와 y_j는 크기가 4인 벡터로 t_j는 신경망의 추정값을 나타내고 y_j는 현재 탐색된 윈도우와 실제 물체 위치의 차이 벡터로 그림 4의 tl과 br벡터를 의미한다. 즉 (3)이 최소값을 가지도록 신경망을 학습하면 탐색 윈도우의 위치를 보정하여 차량 전면부의 정확한 위치를 추정할 수 있다.

3.3 실험 결과

실제로 CNN을 학습할 때는 배치 크기(batch size), 학습율(learning rate), 모멘텀(momentum), 가중치 감소율(weight decay)와 같은 학습 파라미터를 설정해야 한다. 일반적으로 Alexnet[7]의 설정값을 기준으로 자신의 학습 데이터의 종류 및 특성에 따라 적절하게 조절하여 사용한다. 소개하는 사례의 샘플 영상 및 학습 환경, 파라미터에 대한 정보는 차량 전면부 검출[10]을 참고한다.

그림 5는 CNN을 이용한 차량 전면부 검출 결과를 나타낸다. 왼쪽 영상은 차량 전면부로 분류된 윈도우의 중심점을 나타내며 오른쪽 영상은 분류결과에 위치 추정 신경망의 결과를 더해 최종 위치를 추정한 결과이다. 차량 전면부를 65%이상 포함하는 샘플로 학습하였고 CNN이 이동에 불변한 특징도 추출하기 때문에 차량 전면부 근처에서 다수의 윈도우가 검출된다. 그 외 검출된 윈도우는 거짓 긍정(false positive)오류이며 배경을 차량

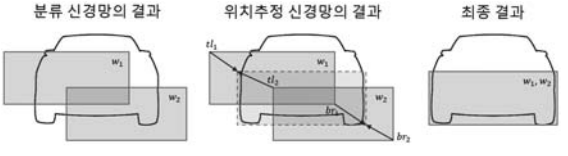

그림 4. 분류 CNN과 위치추정 CNN 각각의 출력 결과와 출력을 합친 최종 결과. w는 차량 전면부로 분류된 탐색 윈도우이며 tl (top-left)과 br (bottom-right)은 각각 왼쪽 위와 오른쪽 아래 모서리의 오프셋 벡터를 의미한다.

그림 5. CNN을 이용한 차량 전면부 검출 결과. 왼쪽은 분류 신경망의 결과(탐색된 윈도우의 중심점만 표시), 오른쪽은 위치 추정 신경망의 결과.

그림 6. 거짓 긍정 오류를 제거한 최종 추정 위치. (a)와 (b)는 각각 지하 주차장과 야외 환경에서 실험한 결과.

그림 7. CNN을 이용한 교통 표지판 검출 결과.

전면부로 잘못 분류한 경우이다.

딥러닝은 모든 파라미터가 샘플에 맞춰 학습되는데 배경은 일관적이지 않고 패턴이 없기 때문에 정확한 학습을 위해서는 많은 샘플이 필요하다. 차량 전면부 검출에는 직접 수집한 샘플 영상을 사용했으며 원본 샘플의 수가 적기 때문에 거짓 긍정 오류가 발생하였다. 군집화 알고리즘과 간단한 후처리 알고리즘[10]으로 거짓 긍정 오류를 제거하고 차량 전면부의 최종 위치를 추정한 결과는 그림 6과 같다. 실제 샘플은 그림 6(a)인 지하 주차장에서 수집하였으나 (b)의 야외 환경에서도 차량 전면부를 정확하게 인식하였다.

그림 7은 같은 구조의 CNN으로 교통 표지판을 검출한 결과이다. 위치 추정 신경망은 사용하지 않고 분류 신경망으로 탐색과 분류만 수행하였다. 40여 종류의 표지판이 다양한 형상과 색상을 가지기 때문에 차량 전면부에 비해 검출이 더 어려운 문제이다. 그러나 표지판 샘플이 약 8만장, 배경 샘플이 약 80만장으로 충분한 수의 샘플로 학습하였기 때문에 고해상도 영상에서도 거짓 긍정 오류가 거의 발생하지 않았으며 표지판을 정확하게 검출하였다.

Ⅳ. 결론

본 글에서는 지능형 시스템과 영상 인식 분야의 최근 트렌드와 발전 방향을 언급하였으며 특징 설계가 필요한 트리기반 분류기나 SVM(support vector machine)과 같은 얕은 학습(shallow learning)과 비교하여 딥러닝이 다시 주목 받게 된 이유를 설명하였다. 또한 영상 인식 분야에 주로 사용되는 CNN의 기본 구조와 특성을 설명하였으며 물체 검출 문제에 실제로 적용한 사례를 소개하였다.

딥러닝은 실제 데이터를 원하는 출력값으로 한번에 변환하는 종단 학습(end to end learning)이 가능한 장점 때문에 언뜻 보기에 어렵지 않고 인공지능과 같이 무엇이든 학습할 수 있을 거라는 환상에 빠지기 쉽다. 그러나 새로운 문제에 딥러닝을 적용하기 위해서는 신경망의 구조부터 학습 파라미터, 학습 과정까지 기계학습의 기본적인 내용에 대한 많은 지식이 필요하다. 또한 딥러닝이 기계학습의 패러다임을 한차례 바꾼 것은 사실이나 매스컴이나 군중심리에 의해 일부 과장된 부분이 있으며 이를 우려의 시선으로 바라보는 전문가들이 많다. 따라서 딥러닝을 수용하되 지나친 환상에 빠지지 않고 기초 이론을 바탕으로 냉철하게 판단하고 분석할 수 있는 능력이 필요하다.

참고 문헌

[1] Russakovsky, Olga, et. al., "Imagenet large scale visual recognition challenge." *International Journal of Computer Vision* pp. 1 42, 2014.

[2] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." arXiv preprint arXiv:1502.03167, 2015.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Proc. Computer Vision and Pattern Recognition*, pp. 886 893, 2005.

[4] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan,

"Object detection with discriminatively trained part based models," *Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627 1645, 2010.

[5] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, pp. 1527 1554, 2006.

[6] Srivastava, Nitish, et. al., "Dropout: A simple way to prevent neural networks from overfitting." *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929 1958, 2014.

[7] A. Krizhevsky, I. Sutskever, and G. Hinton. "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, pp. 1097 1105, 2012.

[8] Girshick, Ross, et. al., "Rich feature hierarchies for accurate object detection and semantic segmentation." *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. IEEE*, 2014.

[9] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y.LeCun, "OverFeat: Integrated recognition, localization and detection using convolutional networks," *Proc. of International Conference on Learning Representations*, 2014.

[10] Y. K. Park, J. K. Park, H. I. On, and D. J. Kang, "Convolutional Neural Network based System for Vehicle Front Side Detection", *Journal of Institute of Control, Robotics and Systems*, vol. 21, no. 11, pp. 1008 1016, 2015.

[11] LeCun, Yann, et al. "Gradient based learning applied to document recognition." *Proceedings of the IEEE*, pp. 2278 2324, 1998.

저 자 약 력

박 제 강

· 2012년 부산대학교 기계공학부 졸업.
· 2012년~현재 부산대학교 대학원 기계공학부 석 · 박통합 과정.
· 관심분야 : 패턴인식, 신경망, 머신러닝.

박 용 규

· 2014년 동아대학교 기계공학과 졸업.
· 2014년~현재 부산대학교 대학원 기계공학부 석사과정.
· 삼성 S/W 멤버십 회원.
· 관심분야 : 머신비전, 패턴인식, 신경망.

온 한 익

· 2014년 부산대학교 기계공학부 졸업.
· 2014년~현재 부산대학교 대학원 기계공학부 석 · 박통합 과정.
· 관심분야 : 머신비전, 패턴인식.

강 동 중

· 1988년 부산대 정밀기계공학과 졸업.
· 1990년 한국과학기술원 기계공학과 석사.
· 1999년 동대학원 자동화및설계공학과 공학박사.
· 2006년~현재 부산대학교 기계공학부 교수.
· 관심분야 : 머신비전, 이동로봇, 영상기반 검사시스템 개발.