딥러닝을 이용한 객체 인식 및 검출 기술 동향

딥러닝을 이용한 객체 인식 및 검출 기술 동향

최근 딥러닝을 비롯한 인공지능 기술의 활용이 다양한 분야에서 활발해지고 있다. 특히 컴퓨터 비전 분야에서 딥러닝 기술을 기반으로 객체의 인식과 검출을 목적으로 뛰어난 성능을 보이는 알고리즘들이 발표되고 있으며 기존의 방법론으로 해결이 어려운 문제들을 딥러닝 기술을 활용함으로써 쉽게 해결하는 성과를 거두고 있다. 본 기술논문에서는 영상 내 존재하는 여러 객체를 검출하고 각 개체를 인식하기 위해 활용된 다양한 딥러닝 기술에 대하여 소개하고 그 응용 사례들을 살펴보고자 한다.

고광은[1], 심귀보[2] ([1]한국생산기술연구원, [2]중앙대학교 전자전기공학부)

1. 서 론

최근 환경과 사람에 대한 인식을 수행하고 이를 기반으로 서비스를 능동적으로 판단, 제공할 수 있는 인공지능을 개발하기 위한 시각정보 기반의 인간-컴퓨터 상호작용(Human-Computer Interaction, HCI) 연구가 매우 중요하다[1]. HCI는 인공지능, 머신 비전, 신호처리 등 다 학제 간 분야를 아우르는 연구의 융합으로 인간과 인간이 상호작용하는 과정에 대한 기능적 측면에서의 모방에서 시작한다. 이를 위해서는 사람이 수행하는 행동이 무엇인지에 대한 것을 인지하는 것 뿐만 아니라 어떻게 그 행동을 수행해야하는지, 왜 그 행동을 수행해야하는지에 대한 정보 또한 중요해진다. 이러한 측면에서 여러 연구자들은 어포던스(affordance)의 개념에 주목하였다[10,11]. 어포던스는 사람이 속한 환경과 사람이 수행하는 동작 간의 운동정보를 인지적 측면에서 표현되는 관계성으로 해석할 수 있다. 어포던스에 대한 아이디어의 핵심은 행동과 관련된 객체가 무엇인지를 인지하는 것 뿐만 아니라 객체를 사용하는 방법을 인지하기 위해 필요한 정보를 검출하는 기술이다[2]. HCI 과정에서 어포던스가 활용되기 위해서는 입력 영상의 객체에 대한 인식과 위치 검출, 그리고 공간적 상태에 대한 추정이 선행될 필요가 있다. 컴퓨터 비전 분야에서 이에 대한 여러 연구들이 발표되었지만 최근 딥러닝 기술이 급속도로 발전하면서 기존의 방법들을 대체되고 있다.

딥러닝을 비롯한 인공지능에 대한 연구의 시발점은 인간의 지적 능력을 모방하여 기계에 부여하는 인공지능의 개념에 대하여 공식적으로 발표한 1956년 다트머스 컨퍼런스에서 시작되었다. 이 시기 연구된 인공지능 프로그램들은 퍼셉트론 등의 인공신경망과 같이 인간의 학습 능력과 지능을 모방하기 위한 다양한 방법들을 시도하였다. 하지만 70년대에 이르러 퍼셉트론에 근본적인 한계가 있음이 입증되고 인공지능 프로그램으로 해결 가능한 문제의 범위가 극히 제한적이라는 중론으로 인해 인공지능 연구는 암흑기를 맞이하게 되었다. 이후 별다른 진전을 보이지 못하다가 70년대 후반부터 오류 역전파 학습 기법을 통해 학습이 가능한 신경망 모델들이 발표되면서 새로운 돌파구가 열리게 되었다. 특히 1986년도 Rumelhart와 Hinton에 의해 발표된 오류 역전파 알고리즘을 통해 최적의 신경망 파라미터를 찾아낼 수 있음이 증명되었고 이를 계기로 인공신경망이 다시 주목받기 시작했다[3]. 이 때 발표된 신경망 모델들은 하나의 은닉 레이어(이하 통칭 “레이어”) 로 구성된 얕은(shallow) 형태를 갖추고 있었는데 만약 네트워크를 구성하는 레이어의 수가 늘려서 보다 복잡하고 깊은(deep) 구조를 구현할 수 있다면 더욱 복잡한 문제도 풀 수 있을 것이란 전망이 제시되었다. 하지만 당시의 하드웨어의 성능적 한계에 부딪혀 학습 시간의 기하급수적인 증가, 학습 데이터에 대한 과적합(overfitting), 학습 과정에서의 오류 발산 등의 문제들을 직면하게 되었다. 이후 1989년 LeCun 연

구팀이 오류 역전파 알고리즘을 이용하여 손으로 쓴 글씨를 인식하기 위해 여러 개의 레이어로 구성된 컨볼루셔널 신경망(convolutional neural network, CNN)을 발표하면서부터 이를 토대로 다양한 딥러닝 모델들이 등장하기 시작했다[4]. 최근 빅데이터, 사물인터넷(internet of things, IoT), 그리고 병렬처리 성능을 제공하는 고성능 GPU와 같은 하드웨어의 발달은 딥러닝의 유행을 촉발하였다. 본 논문에서는 컴퓨터 비전 분야에서 가장 큰 문제 중 하나인 객체 인식 및 검출을 해결하기 위해 활용된 딥러닝 기술 동향에 대하여 알아보고자 한다.

2. 기술 동향

현재의 딥러닝은 인공신경망을 토대로 하여 2000년대 들어 급격히 발달한 GPU 등의 하드웨어 기술과 기존 신경망 학습 과정에서 발생하는 여러 문제를 해결하기 위한 각종 소프트웨어 공학적 트릭에 대한 연구를 통해 꾸준히 개선한 결과이다. 또한 빅데이터 시대로 들어서면서 방대한 양의 학습 데이터를 활용하여 더 복잡한 문제를 해결 가능한 모델을 구현하기 용이해진 것도 딥러닝 관련 연구가 확산되게 된 배경이다. 딥러닝을 통해 구현된 계산 모델들은 다양한 분야에서 놀라운 성능을 보이는데 특히 컴퓨터 비전에서 좋은 성과를 보여주며 기존의 방식을 대체할 수 있는 새로운 대안으로 여겨지고 있다. 본 고에서는 이와 관련된 기술들의 개요와 발전 동향에 대하여 소개하도록 하겠다.

2.1. 컴퓨터 비전 기반 물체 인식 및 검출 관련 연구 동향

컴퓨터 비전과 영상 처리는 사진이나 비디오 같은 영상 데이터들을 처리하고 분석하여 데이터에 내재되어 있는 정보를 추출하는 일련의 과정을 다룬다. 대표적으로는 사진이나 동영상에 등장하는 물체의 클래스를 분류하거나, 물체의 위치를 검출할 뿐만 아니라, 인공 지능의 다른 갈래인 음성인식, 자연어 처리 등과 결합된 통합 인지 기반 상호작용 문제를 다루는 영역을 포함하기도 한다. 영상은 픽셀이라 불리는 2차원 공간 상의 점들의 분포로 이루어져 있다. 각 점이 단일 채널에서 0~255의 정수 값을 가지거나 3개 채널에서 각각 0~255의 정수 값을 가지는 이 저수준의 데이터는 그 자체만으로 영상이 표현하는 내재 정보를 분석할 수 없다. 기존의 방법들은 2차원 공간 상의 픽셀 간의 상관 관계를 수치화하여 상관성이 높은 영역을 선별하고 이를 특징(feature)이라 명하였다. 대부분의 특징들은 연구자들이 선험적인 지식을 바탕으로 설계하였으며 이와 관련된 연구 분야도 세분화되어 원 영상에서 이러한 특징을 검출하는 별도의 과정에 대한 다양한 연구 사례들이 존재한다.

컴퓨터 비전 기반 물체 인식은 2차원 영상에서 추출된 특징을 기반으로 학습 알고리즘을 사용하여 주어진 객체가 어떤 카테고리의 인스턴스인지 인식하는 분류기를 통해 수행된다. 최근에는 거리 정보를 기반으로 물체를 인식하여 이동 로봇에 적용하는 등 다양한 목적으로 활용되고 있다[5]. 물체 검출은 영상 내 물체의 클래스 뿐만 아니라 물체의 위치 정보도 함께 추정하기 때문에 보다 어려운 문제이다. 특히 보안, 감시 시스템과 관련된 임베디드 머신 비전 분야에서 주목받고 있는 기술로 다중 배경 모델을 활용하여 객체 검출을 수행하는 사례도 존재한다[6]. 머신 비전 알고리즘과 관련하여 가장 널리 쓰이는 물체 검출 방식은 지역 특징 매칭 기반 방법들이다. 일반적인 지역 특징 매칭 과정은 harris corner와 같은 식별이 용이한 특징점들을 개발자가 선택하고, 선택된 특징점 주변의 지역 패치에서의 특징벡터를 추출한다. 이 특징 벡터를 추출하기 위해 SURF와 같은 지역 스케일 불변(local scale-invariant) 특징점 검출 방법이 널리 사용되고 있다[7,8]. 또한, Deformable Part-based Model(DPM) 알고리즘[9] 등을 통해 생성한 물체 모델, 즉 템플릿에서 추출한 특징벡터들과 입력 이미지에서 추출한 특징 벡터 간의 유사도 측정에 의한 매칭을 수행하는 사례도 있다. 이 때 매칭된 쌍들 간의 기하학적 변환 관계를 RANSAC과 같은 방법을 이용하여 추정하는 방식의 템플릿 매칭 방법은 원래 영상의 기하학적 정보를 그대로 유지하며 특징 매칭을 할 수 있지만 대상의 형태나 위치가 조금만 바뀌어도 매칭이 실패하는 한계가 있다. 만약 유효한 변환 관계가 있다면 해당 물체를 발견한 것이고, 그렇지 않다면 물체 검출이 실패한 것 혹은 물체가 없는 것으로 판단할 수 있다. 이러한 방법의 가장 큰 문제는 유효한 물체 모델을 어떻게 구축하는가에 대한 것과 복잡한 특징 벡터 및 매칭 관계 분석 과정이 연결됨으로 인해 계산 속도가 느리기 때문에 실제 환경에서 적용이 어렵다는 점이다.

인간이 모든 도메인의 영상에 대하여 적절한 특징을 설계하기도 어려울 뿐 아니라 전체 성능적으로나 효율성 측면에서 볼 때 영상의 적절한 특징을 데이터로부터 기계가 스스로 이끌어 낼 수 있는 대체 방안이 요구되기 시작했다. 최근의 급성장한 딥러닝 기술이 바로 이러한 특징 검출에 관련된 문제들을 대부분 대체할 수 있음이 증명되면서 새로운 전기를 맞이하게 되었다. 자동화된 공장 시스템에서의 불량 검출과 같은 미세 정밀 작업에서부터 인간과 로봇 간의 감정 인식과 같은

영상 정보에 기반한 포괄적인 상호작용 작업에 이르기까지 다양한 분야에서 수요가 폭발적으로 증가하게 되었다.

2.2. 딥러닝 기반 객체 인식 기술 발전 동향

본 고에서는 영상 내 객체 분류 등 컴퓨터 비전 분야의 주요 과제에서 뛰어난 성능을 보였던 딥러닝 모델의 대표적인 사례들을 소개한다.

LeNet-5: 영상 인식에서 활용되는 딥러닝 기술의 근간이 되는 모델은 네트워크 내 하나 이상의 은닉 레이어를 포함하고 레이어의 모든 노드들이 완전히 연결된(Fully-Connected) 형태를 갖춘 FCNN(Fully-Connected Neural Network) 이다. 하지만 FCNN을 학습하는 과정에서 최적의 레이어 수(깊이) 와 레이어 내 노드 혹은 유닛의 수(넓이), 학습 진도율, 감쇄율, 모멘텀과 같이 선험적으로 구해야 하는 하이퍼 파라미터를 적절히 설정하지 못한다면 성능 저하의 주요한 요인이 된다. LeCun 연구팀이 1989년 발표한 CNN은 FCNN 내부의 유닛 간의 연결 형태를 개선하여 local receptive field, shared weights, 그리고 sub sampling 을 적용한 결과이다. 이후 개선을 거듭하여 LeNet-5 모델로 완성되었다(그림 1).

LeNet-5는 0~9까지 영상으로 주어지는 10개의 숫자를 분류하는 목적으로 설계되었지만 학습 목적에 따라 RGB 컬러 영상을 인식도 성공했다. 하지만 FCNN의 경우와 마찬가지로 모델의 학습을 위해 3일이 소요되고 적절한 하이퍼 파라미터의 설정이 선행되지 않는다면 모델 loss 함수가 학습과정에서 수렴하지 않거나, 학습에 활용된 데이터에 대하여 과적합 되는 현실에서 풀어야 하는 복잡한 문제에 적용하기에는 제약조건이 많았다.

AlexNet: ILSVRC(ImageNet Large Scale Visual Recognition Competition)은 대용량의 영상 데이터를 기반으로 영상 내 객체를 인식하거나 위치를 추정하는 등의 과제를 제시하고 이를 해결할 수 있는 머신 러닝 알고리즘을 개발하여 성능을 겨루는 대회이다[10]. 이 대회에서 발표되는 기술들을 통해 당대의 컴퓨터 비전 및 머신 러닝 관련 연구 동향을 살펴볼 수 있다. 2012년도 이전까지는 대부분의 연구팀들이 SVM 기반의 방법들을 활용했다. 하지만 ILSVRC 2012 객체 클래스 분류 과제 부문에서 Hinton 연구팀이 발표한 Deep CNN 모델 AlexNet이 등장하여 16%의 평균 오분류율을 보이며 우승을 차지하였다[11]. 이 결과는 기존의 방법들이 최소 25%의 평균 오분류율을 보인 것과 대비되어 컴퓨터 비전 연구 분야에 큰 충격을 주었으며 이후 대부분의 연구팀들이 딥러닝 기반 알고리즘을 활용하게 만들었다. AlexNet의 주요 특징으로 큰 틀에서 LeNet-5과 유사하지만 보다 깊은 구조(5 컨볼루션 레이어, 2 FC 레이어)를 설계하였으며 학습과정에서 나타나는 vanishing gradient 문제를 해결하기 위해 sigmoid나 tanh 함수 대신 Rectified Linear Unit(ReLU) 함수를 활성함수로 사용했다. 또한 모든 레이어의 각 유닛마다 Local Response Normalization을 수행하여 네트워크의 일반화(Generalization) 오류를 줄이기 위한 시도를 하였으며 서브 샘플링 레이어에서 stride를 receptive field의 크기보다 크게 하여 오버래핑된 pooling을 수행하도록 하였다. 학습과정에서의 과적합을 방지하기 위해 학습용 데이터에 대한 데이터 증강(augmentation)과 dropout 기법을 활용하였다. AlexNet을 구성하는 주요 요인들은 다른 딥러닝 모델에서도 널리 활용되고 있다.

VGGNet: VGGNet은 ILSVRC 2014에서 가장 많은 주목을 받은 네트워크이다. 객체 분류 성능에서는 GoogLeNet보다 다소 떨어지지만 구현이 용이한 단순 구조이면서 동시에 단일 네트워크에서는 더 좋은 성능을 보이는 특징으로 다양한 응용 시스템에서 폭넓게 활용되고 있다. VGGNet의 구조는 AlexNet와 비슷한 형태를 가지며 입력 영상을 포함하는 각 레이어에 대하여 상대적으로 고정된 작은 크기의(3×3, 1×1) receptive field를 사용하고 버전에 따라 컨볼루션 레이어의 수를 11개에서 19개까지 다양하게 설정하며 실험을 수행했다[12]. VGGNet

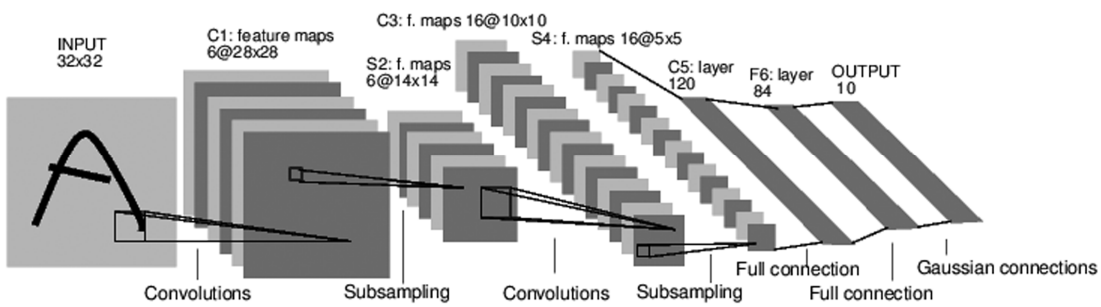

그림 1. LeNet-5 아키텍처[4].

과 AlexNet의 가장 큰 차이는 작은 크기의 receptive field를 사용함으로써 학습해야하는 파라미터의 수를 큰 폭으로 감소시켰으며 입력 영상에 대하여 더 많은 ReLU non-linearity를 적용하게 함으로써 네트워크의 성능을 보다 향상 시킬 수 있었다는 점이다. 또한 네트워크의 초기 파라미터를 AlexNet과 같은 초기 모델로 선행 학습 시키고 이를 fine-tuning 하여 학습 시간과 메모리 소모를 줄였다.

GoogLeNet: VGGNet이 ILSVRC 2014에서 많은 관심을 받았지만 실제 영상 인식에서 우승을 차지한 모델은 Google에서 발표한 GoogLeNet 이다. 딥러닝 네트워크가 깊어지고 레이어가 넓어질 수록 좋은 성능을 기대할 수 있지만 학습과정에서의 과적합과 vanishing gradient의 문제가 발생한다. 이러한 문제는 하이퍼 파라미터나 레이어의 가중치 및 바이어스와 같은 파라미터의 초기값 설정을 적절히 하는 것만으로 해결하기에는 한계가 있기 때문에 GoogLeNet 연구팀은 네트워크의 구조 개선 측면에서 접근하였다. 영상 내 객체 인식 등을 진행하는 과정에서 하드웨어 메모리를 적게 사용하면서 동시에 학습 성능을 극대화하기 위해 네트워크의 깊이와 레이어의 넓이를 더욱 깊고 넓게 하는 구조를 설계하는데 그 목표를 두었다. 이렇게 설계된 GoogLeNet의 핵심은 1 × 1 컨볼루션과 Network-In-Network(NIN)으로 이루어진 Inception 모듈이다. 1 × 1, 3 × 3, 5 × 5 컨볼루션, 그리고 3 × 3 pooling 레이어가 연계된 Inception 모듈 구조를 바탕으로 하나의 컨볼루션 레이어를 표현하는 네트워크를 구축하였다. 이 때, 1 × 1 컨볼루션은 연산량을 크게 경감시키고 단일 컨볼루션 레이어 안의 여러 스케일의 receptive field를 가지는 컨볼루션 레이어가 중첩된 NIN 구조를 통해 깊고 넓은 네트워크 구조를 구현함으로써 VGGNet 보다 우수한 인식률을 보였다[13].

ResNet: 딥러닝 네트워크 구조가 깊어질수록 vanishing gradient 문제 뿐만 아니라 학습 자체의 난이도가 높아져 학습과정에서 오류가 발산하는 경우가 생기는 문제도 존재한다. ResNet 연구팀은 네트워크의 깊은 구조에 비례하는 학습 성능을 얻기 위한 방법을 모색하여 Residual Learning이라 이름 붙인 딥러닝 학습 방법을 제안했다. 일반적으로 딥러닝 모델의 각 레이어의 출력은 그 자체에 대한 오류를 최소화하는 방향으로 학습이 진행된다. 하지만 네트워크의 깊이에 따라 학습 성능를 향상시키기 위해서 ResNet 연구팀은 레이어의 출력과 입력 간의 차이에 대한 오류를 최소화하는 방향으로 학습하는 방법을 고안하였다. 그림 2와 같이 두 개의 레이어로 구성된 네트워크 블록을 가정할 때, 입력과 출력 간의 차를 학습 대상으로 설정하기 위해 출력 단과 입력 단을 연결하는 identity shortcut을 구성했다. 단순한 변형이지만 이로 인해 150개 이상의 레이어로 구성된 네트워크를 구축할 수 있게 되었으며, 늘어난 깊이에 비례하여 인식 정확도도 개선되는 효과를 거둘 수 있게 되었다[14]. ResNet은 Residual Learning 블록을 VGGNet과 유사한 구조와 CNN에 적용한 결과로 볼 수 있으며, 이 네트워크는 ILSVRC 15의 객체 분류 부문에서 우승을 차지하였다.

표 1. 딥러닝 모델의 영상 내 객체 인식 성능 발전 동향

ILSVRC	Classification	Localization	CNN
2012	SuperVision: 0.15315	SuperVision: 0.335463	AlexNet
2013	Clarifai: 0.11197	OverFeat-NYU: 0.298772	
2014	GoogLeNet: 0.06656	VGGNet: 0.253231	GoogLeNet VGGNet
2015	MSRA: 0.03567	Trimps-Soushen: 0.122285	ResNet
2016	Trimps-Soushen: 0.02991	Trimps-Soushen: 0.077377	

지금까지 살펴본 모델들은 영상 내 객체의 클래스를 분류, 인식하는 과제에서 뛰어난 성과를 거둔 모델들로 성능을 ILSVRC의 결과를 토대로 평가해보면 다음 표 1과 같이 정리할 수 있다.

2.3. 딥러닝 기반 객체 검출 기술 발전 동향

컴퓨터 비전 분야 연구자들의 다음 목표는 정지 영상과 동영상에서 나타나는 여러 객체의 위치를 검출(detection) 하는

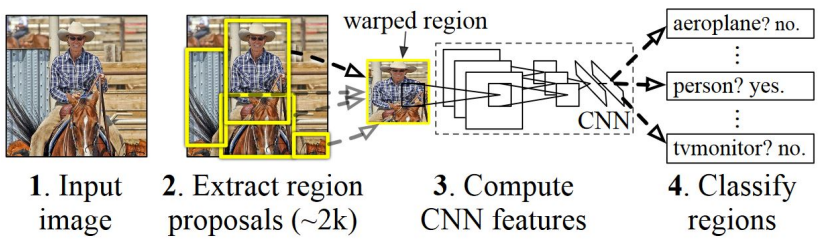

그림 2. R-CNN 파이프라인[15].

문제를 해결하는 것이다. 객체 검출은 객체의 위치뿐만 아니라 클래스에 대한 분류까지 포함한다. 만약 이러한 문제를 해결할 수 있다면 지능형 감시 시스템, 공장 자동화 불량 검사 등 기존의 영상 처리 알고리즘을 기반으로 구축된 응용 사례에서 매우 유용하게 쓰일 수 있다. 이를 위해 시도된 최근의 방법들은 다음과 같다.

R-CNN: 기존의 객체 검출 연구에서는 저수준(low-level) 특징에 기반한 SIFT(Scale-Invariant Feature Transform), SURF (Speeded-Up Robust Feature), HOG(Histogram of Oriented Gradient)와 같은 방법들을 활용하였다. 이러한 접근법으로는 성능 향상에 한계에 부딪히게 되었고 2012년 이후 CNN을 활용하여 객체 검출에 활용해보고자 하는 다양한 시도들이 발생했다. Girshick 연구팀이 제안한 R-CNN은 Selective Search 알고리즘을 기반으로 입력 영상 내 객체 후보 영역(region proposal)을 제안한다. 각 영역에 대하여 CNN을 거쳐 SVM 분류기를 통한 영역 내 객체 패턴의 클래스를 분류한다[15]. 이 CNN은 AlexNet을 기반으로 하며 객체 클래스 분류에 대하여 선행 학습된 상태의 모델을 가져온다. 일련의 과정은 그림 2와 같다.

분류 결과에 대한 오류 값을 기반으로 후보 영역에 대한 경계 상자 회귀(bounding box regression)를 수행함으로써 영상 내 객체의 위치를 표현하는 경계 상자의 좌표를 구한다. R-CNN이 기존의 저수준 특징 기반 방법과 비교하여 좋은 결과를 보인 것은 사실이지만 입력 영상에서 객체 후보 영역을 검출하기 위한 별도의 프로세스(e.g. Selective Search 알고리즘) 가 필요하며, 이 프로세스의 결과로 2,000개의 후보 영역과 각 영역에 대한 CNN 연산이 수행된다. 즉, CNN 연산만 2,000회 수행되는 셈이므로 많은 시간과 메모리가 소모되기 때문에 그에 대한 개선이 필요하다.

SPPNet, Fast R-CNN: R-CNN의 문제점은 입력 영상이 CNN에 적용되는 과정에서 일정 크기로 warping 혹은 crop 되야 한다는 것과 객체의 위치 검출이 결과적으로 Selective Search와 같은 별도의 프로세스에 의존하게 된다는 사실이다. 입력 영상이 CNN으로 적용될 때 일정 크기로 crop 되는 이유는 CNN의 분류기에 해당하는 FC(fully-connected) 레이어가 일정한 크기의 입력 벡터를 수용해야만 학습이 가능하기 때문이다. SPPNet에서는 이로 인해 왜곡되는 정보가 객체 검출 정확도에 영향을 준다고 보았다. 이를 해결하기 위해 입력 영상에서 추정한 객체 후보 영역 경계 상자 내 영상에 대하여 개별적으로 CNN을 적용하는 대신, 입력 영상 자체를 CNN에 적용하여 생성된 마지막 컨볼루션 특징 맵에서 후보 영역을 검출한다. R-CNN과 마찬가지로 입력 영상을 Selective Search 알고리즘에 적용하여 추정된 객체 후보 영역 경계 상자를 CNN의 마지막 컨볼루션 특징 맵에 사영시키고 각 영역에 대하여 서로 다른 스케일을 가지는 3종의 max pooling 커널을 적용시킨다. 이러한 pooling 레이어의 출력이 고정된 크기의 특징 벡터가 되도록 사전에 정의하여 pooling 과정의 stride만 가변적으로 조정하는 방식을 SPP(spatial pyramid pooling) 이라 부른다. 각 피라미드를 통해 나온 특징 벡터는 하나의 특징 벡터로 연결되어 FC 레이어로 전달된다[16]. 결과적으로 SPPNet은 영상 크기에 영향을 받지 않고 CNN에서 가장 많은 시간이 소모되는 컨볼루션 반복 연산을 한번만 수행하기 때문에 R-CNN과 비교하여 100배 이상 빠른 검출 속도를 보인다.

Fast R-CNN은 SPPNet과 거의 유사한 형태이나 마지막 컨볼루션 특징 맵에서의 객체 후보 영역 검출 시 여러 스케일의 pooling을 수행하는 SPP 레이어 대신 RoI(Region of Interest) pooling 레이어를 적용한다는 점에서 차이가 있다[17]. SPP 레이어에서 3개 피라미드의 개별 특징 벡터를 생성하고 이를 연결하는 단계 때문에 SPPNet 자체에서 CNN의 컨볼루션 레이어들에 대한 미세 조정이 불가능했다. 따라서 ImageNet 등의 다른 대규모 영상 데이터베이스를 기반으로 선행 학습된 모델을 기반으로 네트워크를 구축할 수밖에 없었다. Fast R-CNN의 핵심이 되는 RoI pooling은 단일 스케일의 커널을 적용하여 stride 값만 조정함으로써 max pooling으로 고정된 크기의 특징 벡터 추출하는 역할을 수행한다. 따라서 특징 벡터를 연결할 필요가 없고 이는 Fast R-CNN 자체에서 CNN의 컨볼루션 레이어들에 대한 미세 조정이 가능해졌음을 의미한다. 이러한 변화에 의해 Fast R-CNN은 SPPNet 보다 학습과정에서 최소 2.7배 이상, 테스트 과정에서 최소 7배 이상 빨라졌으며 평균 인식률 또한 개선되는 효과를 보였다.

Faster R-CNN: Fast R-CNN을 이용한 영상 객체 검출 속도와 정확도가 기존의 방법과 비교하여 상대적으로 개선되었지만 실시간에서 활용하기에는 속도 측면에서 부족한 측면을 보인다. Fast R-CNN에서 전체 파이프라인에 통합되지 않은 부분은 객체 후보 영역 생성에 관련 부분이다. R-CNN에서부터 Fast R-CNN에 이르기까지 Selective Search 알고리즘 등의 프로세스를 통해 입력 영상 내 객체 후보 영역을 1차로 추정하였는데 이 프로세스는 CNN의 학습과 별도로 수행된다. 이를 네트워크로 통합시켜 객체 검출 과정에서 시간 소모를 줄이고자 Faster R-CNN이 제안되었다. Faster R-CNN은 객체 후보 영역을 추정하기 위해 RPN(region proposal network)를 고안하였다. 입력 영상을 CNN에 적용하여 구한 마지막 컨볼루션 특

징 맵을 다시 RPN의 입력으로 적용하여 입력 영상 내 객체 후보 영역의 경계 상자를 추정한다. 추정된 결과는 Fast R-CNN의 RoI pooling 레이어로 전달되며 이후 과정은 동일하다[18]. RPN을 적용하여 통합된 네트워크 모델을 통해 객체 검출을 수행할 경우 인식률이 소폭 상승하면서 검출 속도는 최소 10배 이상 빨라지는 효과를 볼 수 있었다.

YOLO: 지금까지 대부분의 딥러닝 기반 물체 인식 및 검출은 물체의 위치 영역을 추정하는 프로세스와 물체의 클래스를 분류하는 프로세스를 별도의 CNN 기반의 딥러닝 모델을 이용하여 구성하는 방식으로 이를 해결했다. R-CNN 계열의 객체 검출 모델들은 영상 내 객체 후보 영역을 먼저 추정한 후 이를 기반으로 클래스 분류 및 객체 경계 상자를 찾는다. 이 과정에서 추정된 후보 영역의 갯수가 많고 그로 인한 오버헤드가 크기 때문에 검출 속도 측면에서 실제 보안 시스템이나 로봇 원격 제어 등의 현장에서 활용하기 위한 응용 프로그램으로서 그 성능이 못 미치는 한계가 존재한다. 때문에 객체 인식률을 유지하면서 동시에 검출 속도를 향상 시키려는 다양한 연구들이 시도되었다. 그 중 최근 가장 주목을 받은 방법인 YOLO (You Only Look Once) 네트워크는 최종 출력단에서 경계 상자 검출과 클래스 분류가 동시에 수행되도록 설계되었다. R-CNN 계열 모델 중 가장 빠르고 정확한 Faster R-CNN에서 네트워크를 구성하는 3개의 모듈이 특징 검출, 경계상자 생성, 클래스 분류를 각각 담당하는 것과 대조적으로 YOLO는 모든 단계가 단일 네트워크 안에서 이루어지므로 간단하고 빠르다. YOLO 네트워크의 최종 출력단은 영상 내 객체 클래스 및 경계 상자의 가능성이 있는 모든 후보를 표현하는 특징 텐서(tensor)이다. 이 텐서는 입력 영상을 일정 크기의 그리드로 분할하고 각 그리드에서 생성된 경계 상자가 목적 객체의 경계 상자가 될 스코어와 해당 객체의 클래스가 무엇인가에 대한 사후 확률을 표현한다. 이를 기반으로 가장 유력한 객체 클래스 및 객체 경계 상자를 구하는 후처리로 NMS(Non-Maximum Suppression) 알고리즘이 적용된다[19,20]. 그 결과 YOLO 네트워크를 이용한 객체 검출 시 평균 45fps의 놀라운 속도를 보였다. 정확도를 측정한 결과 Faster R-CNN에 비교하여 다소 낮은 결과를 보였지만 속도 측면에서 압도적으로 빠르기 때문에 이를 상쇄한다.

SSD: YOLO는 영상을 일정 크기(e.g. 7×7) 그리드로 분할하여 각 셀을 중심으로 하는 객체의 경계 상자 후보를 고정된 숫자(e.g. 2개) 만큼 예측한다. 이를 토대로 실제 경계 상자에 대한 검출 및 인식을 수행하기 때문에 객체가 겹쳐 있거나 작은 객체에 대한 검출 성능은 떨어지는 경향을 보인다. 이러한 단점으로 인해 YOLO의 빠른 속도와 일정 수준의 정확도에도 불구하고 실제 응용 시스템에는 널리 활용되지 못한다. SSD (Single Shot multibox Detector)는 VGG-16 네트워크를 기반으로 하며 각 컨볼루션 레이어마다 YOLO의 최종 출력 텐서에 해당하는 특징 맵을 생성하도록 한다. 이를 통해 예측하는 경계 상자 후보가 YOLO보다 훨씬 많고 다양한 스케일의 객체 후보 경계 상자를 추정 가능해지므로 인식 정확도가 개선되는 효과를 보인다[21]. 또한 단일 네트워크 구조를 사용하면서 네트워크의 각 레이어가 객체 검출을 분할하여 수행하는 과정이기 때문에 검출 속도는 거의 유사한 수준을 유지할 수 있다.

2.4. 딥러닝 기반 객체 인식 및 검출 응용

딥러닝을 포함하는 인공지능 관련 산업 영역의 급속도로 확대 추세를 보이고 있으며 그 중에서도 딥러닝 관련 시장의 성장세가 가장 크게 두드러지고 있다. 딥러닝을 활용한 영상 인식 기술은 다양한 시스템으로 적용되고 있다. 예를 들어, 자율주행 자동차에 딥러닝 기술의 적용을 위해 완성차 업계 뿐만 아니라 구글, 애플 등의 글로벌 IT 기업에서 딥러닝 기술에 대한 역량 강화에 나서는 추세가 두드러지고 있다. 또한, 페이스북은 14년도부터 DeepFace라 명명된 얼굴 인식 서비스를 제공하며 이미 사람의 얼굴 인식 수준보다 높은 성능을 보이고 있으며 구글의 구글포토는 클라우드에 업로드된 사진 속 여러 객체를 각각의 클래스에 따라 세분화하여 인식하는 서비스를 제공하고 있다.

딥러닝 기반의 객체 인식 및 검출 기술은 공장 자동화 관련 산업현장에서도 다양하게 활용된다. 예를 들어, 효율적인 철강 제조 프로세스를 위해 강판 정보에 대한 자동 식별을 목적으로 Deep CNN을 이용하여 강판식별번호의 위치를 검출하거나 식별번호 자체에 대한 인식하는 알고리즘이 발표되었다[22,23]. 단순히 영상의 객체에 대한 분류 뿐만 아니라, 객체의 물리적 상태 별로 세분화된 패턴을 분할할 수 있는 기능도 공장 자동화에서 매우 유용하게 쓰일 수 있다. 예를 들어, 항공기 엔진의 여러 부품이 조립된 복잡한 영상에서 불량 상태 부품의 위치를 파악한 결과를 영상 분할하여 도시할 수 있는 다중 스케일 기반 Fully CNN에 대한 연구도 있다[25]. 실제 공정 상에서 발생하는 제품의 하자나 결함에 대한 원인은 이미지의 상태에 대한 인식이나 영상 분할만으로 추론하기 어렵다. 이는 공정 모니터링 과정에서 수집되는 다양한 센서 신호를 시간 축 상에서 분석함으로써 해결 가능하다. 최근에는 CNN을 활용하여 이러한 FDC(Fault Detection & Classification) 문제의 솔루션을 제안한 사례도 존재한다[26].

3. 결 론

딥러닝을 이용하여 영상 내 물체를 인식하고 검출하는 기술은 여러 선행 연구들을 통해 그 성능이 입증되었다. 본 기술논문에서는 이러한 객체 인식 및 검출을 수행하는데 최적화된 CNN 기반의 최신 알고리즘들과 그 응용 사례에 대하여 소개하였다. 지금까지 살펴본 바와 같이 딥러닝 기술의 발전이 가속화함에 따라 컴퓨터 비전 분야 뿐만 아니라 음성 인식, 자연어 처리 분야에서도 본 논문에서 소개하지 못한 여러 가지 뛰어난 모델과 알고리즘들이 존재하며 이를 활용한 다양한 어플리케이션들이 시장에 선보여지고 있다. 하지만 일부 연구자들은 딥러닝 기술의 과도한 관심에 우려를 표하기도 한다. 딥러닝을 이루는 근간 기술들이 최근 새롭게 등장한 개념도 아닐 뿐더러 딥러닝을 비롯한 인공 신경망 등의 연구에서 꾸준히 지적되고 있는 블랙박스 문제, 1956년 다트머스 컨퍼런스 이후 인공지능 연구가 중흥과 몰락을 반복해 오던 역사 등이 그 원인이다. 그럼에도 불구하고 향후 4차 산업혁명 시대에 진입하면서 딥러닝에 밀접한 연관성을 가진 빅데이터, IoT 등 첨단 산업의 발전과 더불어 세계 유수의 뛰어난 연구자들이 가장 활발히 활동하고 있는 연구 분야이기 때문에 딥러닝 기술의 혁신적인 발전을 기대한다.

REFERENCES

[1] R. Kelley, A. Tavakkoli, C. King, M. Nicolescu, and M. Nicolescu, "Understanding activities and intentions for human-robot interaction," *Human-Robot Interaction*, 2010: InTech.

[2] E. J. Gibson, K. Adolph, and M. A. Eppler, "Affordances," *The MIT encyclopedia of the cognitive sciences*: MIT Press, 1999.

[3] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst for Cognitive Science1985.

[4] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE,* vol. 86, no. 11, pp. 2278-2324, 1998.

[5] 박정길 and 박재병, "실내 이동로봇을 위한 거리 정보 기반 물체 인식 방법," 제어로봇시스템학회 논문지, vol. 21, no. 10, pp. 958-964, 2015.

[6] 박수인 and 김민영, "고정형 임베디드 감시 카메라 시스템을 위한 다중 배경모델기반 객체검출," 제어로봇시스템학회 논문지, vol. 21, no. 11, pp. 989-995, 2015.

[7] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Computer vision–ECCV 2006,* pp. 404-417, 2006.

[8] D. G. Lowe, "Object recognition from local scale-invariant features," *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, 1999, vol. 2, pp. 1150-1157: Ieee.

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence,* vol. 32, no. 9, pp. 1627-1645, 2010.

[10] O. Russakovsky et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision,* vol. 115, no. 3, pp. 211-252, 2015.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.

[13] C. Szegedy et al., "Going deeper with convolutions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1-9.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580-587.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling deep convolutional networks for visual recognition," *European Conference on Computer Vision*, 2014, pp. 346-361: Springer.

[17] R. Girshick, "Fast r-cnn," *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440-1448.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, 2015, pp. 91-99.

[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779-788.

[20] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," *arXiv preprint arXiv:1612.08242,* 2016.

[21] W. Liu et al., "Ssd: Single shot multibox detector," *European conference on computer vision*, 2016, pp. 21-37: Springer.

[22] S. J. Lee and S. W. Kim, "Recognition of Slab Identification Numbers Using a Deep Convolutional Neural Network," *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*, 2016, pp. 718-721: IEEE.

[23] S. J. Lee, J. Ban, H. Choi, and S. W. Kim, "Localization of slab identification numbers using deep learning," *Control, Automation and Systems (ICCAS), 2016 16th International Conference on*, 2016, pp. 1174-1176: IEEE.

[24] R. Ren, T. Hung, and K. C. Tan, "A Generic Deep-Learning-Based Approach for Automated Surface Inspection," *IEEE Transactions on Cybernetics,* 2017.

[25] X. Bian, S. N. Lim, and N. Zhou, "Multiscale fully convolutional network with application to industrial inspection," *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, 2016, pp. 1-8: IEEE.

[26] K. B. Lee, S. Cheon, and C. O. Kim, "A Convolutional Neural Network for Fault Classification and Diagnosis in Semiconductor Manufacturing Processes," *IEEE Transactions on Semiconductor Manufacturing,* 2017.

저자약력

고 광 은

- 2007년 중앙대학교 전자전기공학부(공학사)
- 2017년 중앙대학교 대학원 전자전기공학부 (공학박사)
- 현재 한국생산기술연구원 포스트닥터
- 관심분야 : Deep learning, Human-Robot Interaction, Human Intention Recognition 등

심 귀 보

- 1984년 중앙대학교 전자공학과(공학사)
- 1986년 중앙대학교 전자공학과(공학석사)
- 1990년 The University of Tokyo 전자공학과 (공학박사)
- 1991년~현재 중앙대학교 전자전기공학부 교수
- 2002년~현재 중앙대학교 중소기업산학협력센터 센터장
- 2006년~2007년 한국지능시스템학회 회장
- 2007년~2013년 (사)한국산학연협회 서울지역협회 회장
- 2009년~2010년 중앙대학교 중앙도서관장 및 박물관장
- 2011년~현재 중앙대학교 스마트지능로봇연구센터 센터장
- 관심분야 : 뇌-컴퓨터 인터페이스, 의도 인식, 감성 인식, 유비쿼터스 지능형 로봇, 지능 시스템, 컴퓨테이셔널 인텔리젼스, 지능형 홈 및 홈 네트워크, 유비쿼터스 컴퓨팅 및 센서 네트워크, 소프트 컴퓨팅 (신경망, 퍼지, 진화연산), 다개체 및 자율분산로봇시스템, 인공 면역 시스템, 지능형 감시 시스템, 사물인터넷(IoT), Deep Learning, 빅데이터 등, ICROS Fellow