ETRI

2018
Electronics and
Telecommunications
Trends

Deep Neural Network &
Object Recognition 특집

딥러닝 기반 객체 분류 및 검출 기술 분석 및 동향

Technology Trends and Analysis of Deep Learning Based Object Classification and Detection

이승재 (S.J. Lee, seungjlee@etri.re.kr) 인포콘텐츠기술연구그룹 선임연구원/PL
이근동 (K.D. Lee, zacurr@etri.re.kr) 인포콘텐츠기술연구그룹 선임연구원
이수웅 (S.W. Lee, suwoong@etri.re.kr) 인포콘텐츠기술연그룹 선임연구원
고종국 (J.G. Ko, jgko@etri.re.kr) 인포콘텐츠기술연구그룹 책임연구원/PL
유원영 (W.Y. Yoo, zero2@etri.re.kr) 인포콘텐츠기술연구그룹 책임연구원/그룹장

Object classification and detection are fundamental technologies in computer vision and its applications. Recently, a deep-learning based approach has shown significant improvement in terms of object classification and detection. This report reviews the progress of deep-learning based object classification and detection in views of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), and analyzes recent trends of object classification and detection technology and its applications.

* DOI: 10.22648/ETRI.2018.J.330404

*이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임[No. R0132-15-1005, 온-오프라인에서의 콘텐츠 비주얼 브라우징 기술 개발].

Ⅰ. 서론

객체 분류(Classification)와 검출(Detection)은 영상 분석 및 컴퓨터 비전 분야의 기본 요소 기술로 그동안 많은 연구가 진행되어 왔다. 객체 분류는 미리 정해진 카테고리(Category)에 따라, 영상을 분류하는 것이고, 객체 검출은 정해진 카테고리의 객체들과 위치 정보(Location)를 찾는 문제이다. 객체 분류에 있어서 객체 종류 및 위치를 찾는 것을 객체 분류 및 정확도(Classification with Localization)로 구별하기도 한다. 객체 분류 및 검출은 기술 개발 및 성능의 객관적 비교를 위해 데이터셋이 필수적이며, 그동안 관련 분야 전문가의 노력으로 공개된 PASCAL VOC(Visual Object Classes)[1], ILSVRC(ImageNet Large Scale Visual Recognition Challenge)[2]를 중심으로 기술 개발 및 성능 평가가 이루어져 왔다.

특히 이미지넷(ImageNet)[3]으로 알려진 데이터베이스 구축과 함께 객체 분류 및 검출을 위한 대규모 데이터셋을 공개하고 평가하는 ILSVRC 대회[4]가 진행되면서 기술의 발전에 크게 기여했다. 또한, 딥러닝(Deep Learning) 기술의 적용을 통한 객체 분류 및 검출 성능의 발전은 기존 성능을 개선하며 새로운 발전 가능성을 제시하였고, 현재까지 지속적인 발전을 계속하고 있다. 본 동향 분석에서는 객체 분류 및 검출에 대한 기술 발전 과정을 ILSVRC 대회의 연차별 주요 성과와 제안된 알고리즘을 중심으로 살펴보고, 객체 분류 및 검출 최근 기술 동향과 발전 방향을 검토하고자 한다.

Ⅱ. 이미지넷(ImageNet)

이미지넷(ImageNet)은 WordNet[5]이라는 영어 의미 어휘(Sysnet) 목록을 기반으로 구축된 데이터를 활용하여 구축한 영상 데이터베이스로 의미 어휘 목록은 10만 개가 넘으며, 이 중 대부분은 명사로 이루어진다. ImageNet은 이러한 의미 어휘에 해당되는 이미지를 평균 1,000장 이상 수집한 데이터로 대용량 영상 검색을 위한 데이터베이스 구축에 목적이 있다. 2010년 4월 30일을 기준으로 21,814개의 의미 어휘에 대해서 14,197,122개의 이미지를 구축하였으며, 영상 내 객체에 대해서 객체 위치 정보(Bounding Box Annotation)가 1,304,908개가 구축되었다.

구축된 DB를 활용한 ILSVRC라는 이미지 검색 대회가 2010년부터 열리게 되었으며, 이 대회를 기반으로 영상 검색 관련 기술이 비약적으로 발전하는 계기가 되었다. 다음 절에서는 ILSVRC 대회 및 주요 결과에 대해서 다루고자 한다.

Ⅲ. ILSVRC 대회와 기술 발전 동향

1. ILSVRC 개요

ILSVRC는 대규모 영상에서 영상 분류와 객체 검출 알고리즘을 개발하기 위해서 시작된 대회이다. 이 대회 이전에도 PASCAL VOC Challenge[6]와 같은 영상 검색을 위한 공개 DB를 통한 대회가 있었다. 기존 대회인 Pascal VOC Challenge와의 가장 큰 차이점은 데이터의 규모와 객체의 종류에 있다.

〈표 1〉은 두 대회의 객체 검출을 위한 데이터 규모를 비교한 것으로 데이터 규모와 수가 매우 큰 차이를 보임을 알 수 있다. PASCAL VOC가 20개의 객체, 약 1만장

〈표 1〉 ILSVRC와 PASCAL VOC 데이터셋 비교

		PASCAL VOC(2012)	ILSVRC(2013)
객체의 수		20	200
Train	이미지수	5,717	395,909
	객체수	13,609	345,854
Val	이미지수	5,823	20,121
	객체수	13,841	55,502
Test	이미지수	10,991	40,152
	객체수	비공개	비공개

[출처] ILSVRC[4] 및 PASCAL VOC 2012대회[6] 재구성

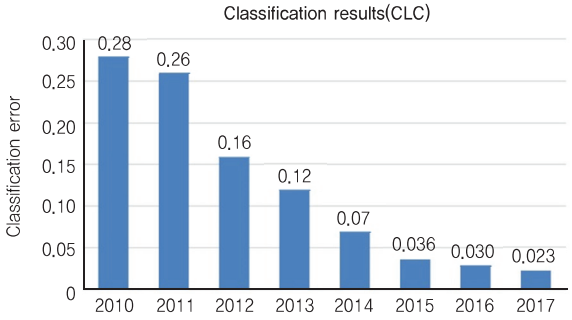
규모의 데이터셋인 반면에 ILSVRC는 200개의 객체, 약 40만 장 규모의 데이터셋이다. 이 데이터는 연차별 진행에 따라 일부가 추가되거나 정제되었다.

최초 ILSVRC 대회는 2010년 PASCAL VOC 대회와 함께 1,000개의 객체 분류 문제로 시작되었으며, 2011년에 객체 분류와 객체 위치 문제(Localization), 2013년에 객체 검출(Detection) 문제를 추가하였다. 2015년에는 비디오 내의 객체 검출(Object Detection From Video)과 장면 분류(Scene Classification) 문제를 같이 평가하였고, 2016년에는 장면 분할(Scene Parsing)을 추가해서 다루었다. 2017년에 대회가 종료되면서 Kaggle[7]에서 추후 대회를 관리하게 되었다.

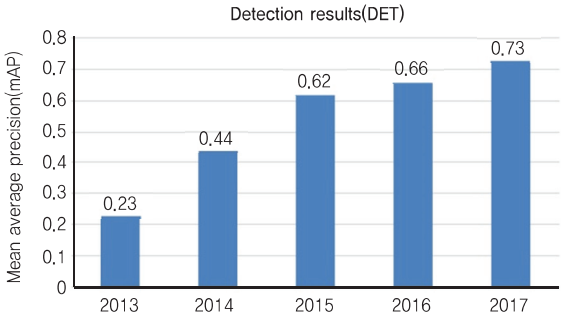
ILSVRC 대회의 진행과 함께 딥러닝 기술의 적용 및 발전은 영상 분류 및 객체 검출 분야에 성능 향상과 기술적 발전을 가져오게 된다. 다음 절에서는 연차별 대회의 주요 특징과 알고리즘을 살펴본다.

2. ILSVRC 대회 및 기술 발전 방향

ILSVRC(ImageNet Large Scale Visual Recognition Challenge)는 대회 진행 과정에서 평가 부분의 일부 변화가 있었으나 가장 주요한 것은 객체 분류와 검출이었다. 객체 분류 분야에서는 알고리즘이 예측한 5개의 정답에 대해서 오분류율(Top-5 Error)을 평가하며, 위치를 같이 평가하는 객체 분류 및 정확도(Object Classification With Localization)는 객체의 분류 정보와 객체의 위치 정보를 Bounding Box값 기반으로 평가한다. 위치 정보는 객체의 최외각 사각형 box 좌표값과 예측된 사각형 Box 좌표값을 이용하여 두 사각형의 합집합 면적 대비 교집합 면적의 비로 계산하고, 0.5 이상인 경우에 정답으로 한다. 객체 검출(Object Detection)은 영상 내에서 정의된 카테고리 내의 모든 객체를 찾고, 그 위치를 예측된 사각형 Box와 정답 Box간의 overlap 기준으로 평가한다. 객체 분류 및 위치 정확도와 같이



(그림 1) ILSVRC 객체 분류 성능 변화



(그림 2) ILSVRC 객체 검출 성능 변화

0.5 이상인 경우를 정답으로 평가하고, 평균 검출 정확도(mAP: mean Average Precision)으로 평가한다.

2010년부터 2017년까지 객체 분류와 검출의 성능 변화는 아래의 그림과 같다. (그림 1)과 (그림 2)에서 보듯이 객체 분류의 경우 평가 Top-5 오분류율이 0.28에서 0.023까지 떨어져 인간의 오분류율인 0.05를 넘어섰으며, 2013년 대회부터 시작된 객체 검출은 0.23에서 0.72까지 3배 이상의 성능 향상이 있었다.

가. 2010~2011년 대회

2010년에는 처음 대회가 시작되었으며, 1,000개 종류의 영상을 분류하는 문제를 평가하였다. 120만 장의 Training 영상과 5만 장의 Validation 영상, 15만 장의 test 영상이 활용되었다. 총 11개 팀이 참가해서 35가지의 알고리즘을 제출했다. NEC와 UIUC(University of Illinois Urbana-Champaign)팀이 1위를 제록스(XEROX 연구소, 現 Naver Labs Europe)이 2위를 기록

했다. 2010년 대회에서는 기존의 지역 서술자(Local Descriptor)를 활용하여 효율적으로 특징을 표현하는 Descriptor Coding[8] 및 Fisher Kernel[9]과 SVM (Support Vector Machine)[10]을 활용하는 방법이 주요 기술적 흐름이었다. 2011년 대회에서도 유사한 기술적 흐름을 유지하였으며, 객체 분류에서는 제록스 연구소가 Compressed Fisher Vectors[11]와 Product Quantization[12]을 기반으로 1위를 차지했다. 처음 평가한 객체 분류 및 정확도(Classification with Localization)에서는 University of Amsterdam과 University of Trento가 1위를 기록했다. 이들은 Selective Search [13]를 기반으로 한 후보영역 선별과 SVM 기반의 Regression을 활용하였고, 특징으로는 BoW[14], 픽셀 값 및 SIFT[15] 등을 융합하여 사용하였다.

나. 2012~2013년 대회

2012년 대회에서는 University of Toronto에서 제출한 딥러닝 기반의 AlxeNet[16]이 1위를 기록했다. 객체 분류에 있어서 오분류율을 0.16을 기록했으며, 객체 분류 및 정확도는 0.33을 기록했다. 이 기록은 2위 팀의 오분류율 0.26과 객체 분류 및 정확도 0.5와 매우 큰 차이를 나타낸 것으로 기술적 흐름을 딥러닝 기반으로 변경하게 되는 중요한 계기가 된다. AlexNet에서는 ReLU, Dropout 및 GPU를 활용하여, 딥러닝의 가능성을 성능으로 증명하였다.

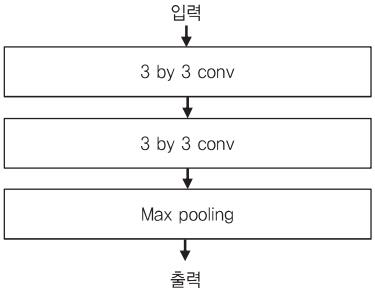
2013년에는 객체 검출 부분이 새로이 추가되어 진행되었으며, 2012년 대회에서의 딥러닝 기반 성능 향상에 따라 관심도가 급증하여 총 81개 팀이 참가하고, 대부분 딥러닝 기반 알고리즘을 제출하였다. 객체 분류에 있어서는 Clarifai[17]가 0.117로 1위를 객체 분류 및 정확도에서는 New York대학이 0.298로 1위를 각각 기록했다. 처음 실시된 객체 검출 분야에서는 University of Amsterdam이 0.226으로 1위를 기록했다. 객체 분류 1위인 Clarifai[18]는 2012년 Alexnet을 기반으로 성능을 개선했으며, 객체 분류 및 정확도 1위인 New York 대학은 객체의 위치 검출하기 위해 후보 위치들을 기존에 사용되는 제거(Suppression)방식에서 후보들의 값을 활용하는 결합 방식(Voting)[19]을 제안했다. 객체 검출에서는 University of Amsterdam이 SVM 기반의 Regression을 학습한 방법과 객체 종류에 기반한 사전 정보를 활용하여 성능을 개선하였다. 한편, 객체 분류에서 2위를 기록한 국립싱가폴대학(NUS)은 NIN(Network in Network)[20] 구조를 제안하고, 1 by 1 conv 기반 레이어 구성은 많은 후속 방법들에 영향을 주게 된다.

다. 2014년 대회

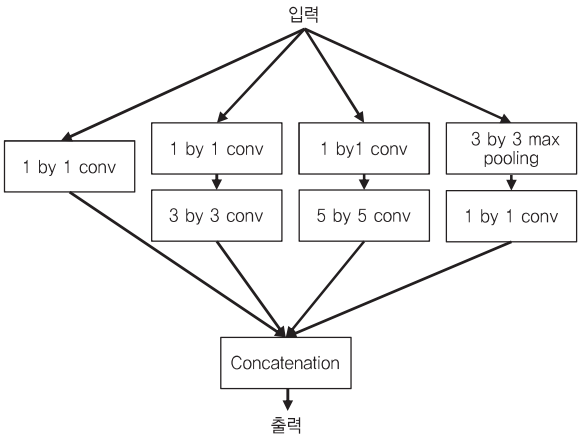
2014년 대회에서는 현재 딥러닝 기술 개발에 활용되고 있는 주요한 구조가 많이 제안되었다. 특히 Oxford 대학에서 제안한 VGG Net[21]은 3 by 3 conv을 활용한 단순한 구조와 성능으로 주목을 받았고, 현재까지 계속 활용되고 있다[(그림 3) 참조].

구글은 Inception 모듈에 기반한 구조를 제안하였다. Inception 모듈에서는 3 by 3 conv외에 1 by 1 conv, 5 by 5 conv를 활용하여 다양한 특징을 연결한다. 구글은 이 구조를 계속 변형 발전시켜 여러 버전의 Inception 모듈[22]-[24]을 제안하였다[(그림 4) 참조].

Microsoft Research Asia는 SPP-net[25]을 제안하였다. SPP-net은 이전에 제안된 R-CNN[26] 방법이 CNN(Convolutional Neural Networks)의 입력 과정에서 발생하는 영상 변화(Cropping and Warping)로 인한



(그림 3) VGG Net[21]의 기본 구조



(그림 4) Inception 모듈 기본 구조[22]

정보 손실 문제 및 속도를 개선하기 위해서 CNN을 거친 후에 SPP(Spatial Pyramid Pooling) 기반 영역별 조합을 통해 속도 및 성능을 개선하였다. 향후 이 방법은 Fast-RCNN[27], Faster-RCNN[28]으로 개선된다.

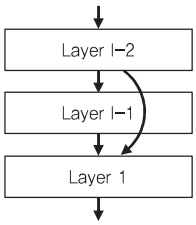
국립싱가폴대학(NUS)는 NIN구조를 활용하여 R-CNN을 적용하고, 딥러닝 학습에 의한 특징이 아닌 사람이 디자인한 Handcraft 특징을 Global Context로 활용하여 성능을 개선하였다. 홍콩 중문대학(CUHK)도 R-CNN 기본 구조에서 객체 검출을 위해서 pooling layer에서 물체의 형태에 따른 레이어를 별도 학습하여 전체 결과에 반영하였다. 이러한 Global Context나 객체 주변 특징의 연관 정보를 활용하는 기법은 이후 대회에서 일반적인 기법으로 활용되었다.

라. 2015년 대회

2015년에는 MS COCO Challenge[29]와 같이 진행되었으며, COCO[30]는 ImageNet 데이터베이스가 다루고 있는 객체 검출 문제 외에도 객체 분할(Object Segmentation), 객체 특징 검출(Keypoints Detection) 등의 문제를 다루면서 다양성을 추구하고 있다. 객체 검출을 위해서 80개 객체 대한 데이터베이스를 제공하고 있다. ILSVRC의 객체 분류 및 정확도, 객체 검출 분야에서 MSRA(Microsoft Research Asia)가 1위를 기록하

〈표 2〉 ILSVRC 2015 대회 결과

분야	MSRA(1위)	2위
ILSVRC localization	0.09	0.12
ILSVRC detection	0.621	0.536



(그림 5) ResNet[31] 기본 구조

[출처] https://commons.wikimedia.org/wiki/File:ResNets.svg, CC BY-SA 4.0.

였고, MS COCO Challenge에서도 MSRA가 객체 검출 및 분할 부분에서 1위를 기록하였다[〈표 2〉 참조].

이러한 성능의 개선은 깊은 네트워크 학습을 가능하게 한 ResNet[31]의 구조에 있다. ResNet은 Identify Mapping을 활용하여 기존 딥러닝의 문제점인 Vanishing Gradient 문제를 대응하여 아주 깊은 네트워크의 학습이 가능함을 확인시켜 주었다. 그림에서 보는 것처럼 기본 구조의 출력에 다시 입력을 더해서 다음 레이어로 넘어가도록 되어 있고, 이를 통해서 에러를 역전파하는 과정에서 미분값이 상수로 남아서 계속 더해져서 값이 작아지지 현상을 방지해 준다. 이러한 구조를 통해서 설계된 네트워크로 MSRA는 새로운 특징(Feature)을 추출하여 기존의 객체 분류나 검출의 구조의 특별한 변화 없이 높은 성능 개선을 보였다[(그림 5) 참조].

마. 2016~2017년 대회

2015년 MSRA의 ResNet을 통한 성능개선 이후로 실질적인 성능 개선이 어려울 것이라는 우려와 달리 2016년에도 객체 분류 및 정확도는 중국 제3 공안 연구소[32]가 1위로 성능을 0.0299 및 0.0771로 각각 개선하였으며, 객체 검출 정확도는 홍콩중문대학(CUHK)와 SenseTime[33]이 0.663으로 1위를 중국의 CCTV 업체

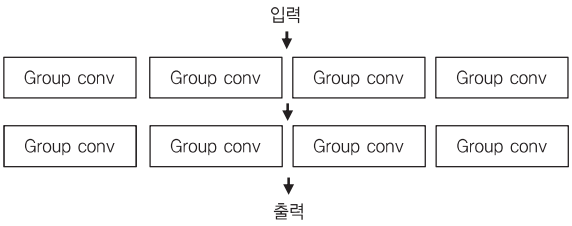
Hikvision[34]이 0.653으로 2위를 기록하였다. ILSVRC 대회가 진행되면서 중국의 딥러닝 관련 대학과 기업의 참가가 계속되었고, 세계적 수준에 근접하고 있음을 확인할 수 있다. 한편, 2015년에 이어 동시에 진행된 MS COCO 대회에서는 객체 검출 분야는 구글이, 객체 분할 분야에서는 MSRA이 각각 1위를 기록했다.

객체 분류 및 정확도 분야에서는 중국 제3 공안 연구소가 기존 모델들의 앙상블(Ensemble)을 통해서 성능을 향상하였다. 객체 검출에서는 홍콩중문대학(CUHK) 팀이 특징을 추출하는 영역 선별 및 학습 방법을 제안하였고, 성능 향상을 위해서는 마찬가지로 앙상블 기법을 활용하였다. 2위 팀인 Hikvision에서도 성능 개선을 위해서 앙상블 기법이 활용되어 이전 대회와 같은 근본적인 성능 개선 방법이 제시되지는 못했다.

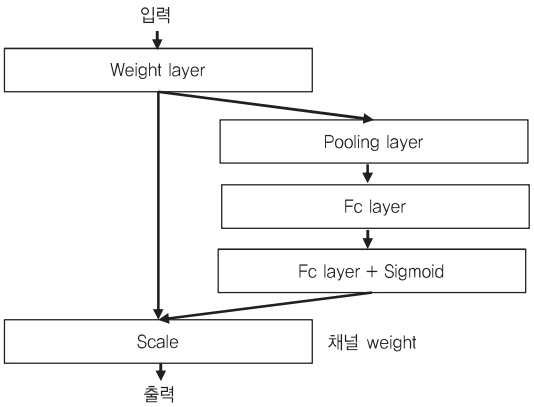
한편, 구글은 MS COCO 대회의 객체 검출을 위해서 기존에 제안했던 Inception 모듈과 ResNet을 결합하여 Faster RCNN 구조에 적용하였다. 구글 역시 앙상블 기법 및 Test Augmentation기법을 적용하여 최적의 성능을 얻기 위한 많은 노력을 하였다. MSRA는 객체 분할을 위해서 FCN(Fully Convolutional Network) 기반의 구조[35]를 제안하였다. 기존의 객체 분류를 위한 네트워크가 위치 변화에 불변한 특징을 가지는 반면 객체 검출이나 분할의 경우 위치에 따른 변화를 고려하는 것이 필요하고 이러한 점을 반영할 수 있는 네트워크 구조를 제안하였다. 최종 성능을 위해서는 마찬가지로 Test Augmentation 기법과 앙상블기법을 적용하였다.

객체 분류 부분에서는 Facebook이 ResNext[36]라는 새로운 구조를 제안하였으며, 1위와 근소한 차이인 0.03을 달성하였다. ResNext는 기존 ResNet의 conv 레이어를 group conv로 변경하여 성능을 개선한 구조로 현재 많이 활용되고 있다[(그림 6) 참조].

2017년 ILSVRC는 마지막 대회로 객체 분류 및 검출에 대한 기술 경쟁을 통해 기술의 비약적 발전을 이끌어



(그림 6) ResNext[36] 기본 구조



(그림 7) Squeeze and Excitation[41] 기본 구조

낸 중요한 역할을 마무리하는 대회였다. 이후 대회는 Kaggle에서 관리 및 운영될 예정이다. ILSVRC 대회의 시작과 성공으로 MS COCO를 비롯하여 다양한 데이터셋[37]-[39]의 공개와 함께 기술 경쟁 대회가 이루어지고 있으며, 이는 기술의 발전속도를 가속화 하고 있다. 2017년 객체 분류 및 정확도의 객체 분류 부분에서는 중국 자율 주행 업체인 Momenta와 Oxford대학이 0.0225로 1위를, 위치 정확도에서는 국립 싱가폴대학(NUS)과 Qihoo360[40]이 0.0623으로 1위를 기록했다. 객체 검출에서는 중국 난징공대(NUIST)가 0.732로 1위를 기록했다.

객체 분류에서는 Momenta가 Squeeze and Excitation 구조[41], [(그림 7) 참조]을 제안하였다. 기존 네트워크 학습 과정에 특징의 채널 정보에 대한 가중치를 학습하여 성능을 개선 시킨 방법이다. 객체 분류 정확도에서는 국립싱가폴대학(NUS)가 Dual Path 네트워크[42]를 제안하였다. 이 네트워크는 기존의 ResNet[31]과

DenseNet[43]을 하나의 네트워크로 결합한 것으로 둘 사이의 연관관계를 이용해서 하나의 네트워크로 설명하였다. 단일 모델 기준으로 기존의 ResNext와 Poly-Net[44] 보다 개선된 성능을 보인다고 주장하였다.

객체 검출 분야에서는 난징 공대(NUIST)에서 기존에 제안된 모델을 앙상블하고, Test Augmentation, 학습 데이터 정제를 통한 영역 선별 및 NMS(Non Maximum Suppression)의 반복 적용 등을 활용하여 성능을 개선하였다.

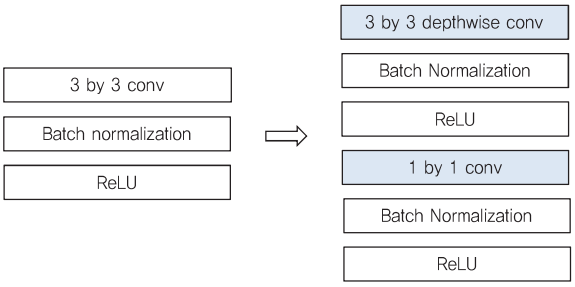
ILSVRC는 8년간 대회를 진행하면서 객체 분류에서는 10배의 성능 향상과 객체 검출에서는 3배의 성능 향상이 있었고, 다양한 네트워크 구조와 성능 개선을 위한 기법이 제안되었다. 이와 함께 객체 분류와 검출 기술을 실제적인 서비스에 적용하기 위한 다양한 노력이 진행되고 있다. 다음 절에서는 객체 분류 및 검출 분야의 기술적 변화와 최근 동향에 대해서 살펴보겠다.

Ⅳ. 영상 분류 및 객체 기술 동향

Ⅲ장에서 살펴본 바와 같이 영상 분류 및 객체검출기술은 ILSVRC 대회를 중심으로 발전해왔으며, 주요 알고리즘과 성능의 검증에 따라 성능 중심의 연구에서 실제 서비스를 위한 연구로 변화가 진행되고 있다. 특히 객체 분류 및 검출 기술을 스마트 기기 등 휴대용 장비에서 활용하려는 시도를 중심으로 모델의 경량화 및 고속화 시도가 활발히 이루어지고 있다. 또한, 성능의 개선에 따라 관련 응용서비스의 선제적 적용 및 다양한 응용기술의 융합을 통한 새로운 시도들이 이루어지고 있다.

1. 모델 경량화 및 고속화

기존 제안된 객체 분류 및 검출 모델들은 학습 및 테스트를 위해서 많은 연산량이 소요되어 GPU 등의 고가 장비가 필수적이다. 모델 크기도 상대적으로 수백 MB에 이르며, 테스트에도 상당한 시간이 소요되어 실시간 서비스를 위해서는 개선이 필수적이다. 이러한 문제를 해결하기 위해서 딥러닝 모델을 압축하거나 계산량을 줄이려는 다양한 시도가 이루어지고 있다.



(그림 8) MobileNet-v1[51] 기본 구조

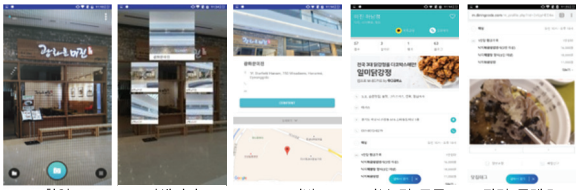
연산량과 모델 사이즈를 줄이기 위해서 기존 모델의 파라미터를 Low-rank로 근사화[45]-[48]하거나 양자화[49], pruning[50] 기법들이 제안되었다. 또한, 기존 모델 내의 기본 모듈인 conv을 변경하여 파라미터와 연산량을 줄이는 방법이 MobileNet-v1[(그림 8) 참조], [51], MobileNet-v2[52], ShuffleNet[53] SqueezeNet [54]으로 제안되었다. 이 네트워크들은 기존 네트워크의 성능을 유지하면서 메모리와 연산량을 줄이는데 목적이 있으며, 상당한 진전을 나타내고 있다. 아래 그림은 구글의 MobileNet-v1로 1 by 1 conv과 depthwise conv을 활용하여 연산량을 줄인 구조이다.

MobileNet-v2에서는 기존의 bottleneck 구조와 다르게 압축(Compression) 후 확장(Expansion)하는 구조에서 벗어나 선 확장(Expansion)하는 방향으로 변경하여 연산량 및 메모리를 개선하였다. MobileNet-v1[51], MobileNet-v2[52]은 기존 모델과의 성능과 유사한 성능을 내면서 메모리 및 처리속도에 강점을 가진다. 이와 같이 최근 연구들은 기존 모델의 성능을 유지하면서 모델을 경량화 방향으로 연구되고 있다. 이러한 추세와 연계되어 최근에는 저전력 이미지 검색 대회[55]도 열리고 있으며, 딥러닝을 위한 개발 도구에서도 모바일 환경을 지원하기 위한 Tensorflow Lite[56], caffe2[57], CoreML[58] 등이 공개되고 있다.

2. 사업화 및 기술 융합

영상 내 객체 검색 서비스는 다양한 영상 분야의 검색을 위한 핵심 기술이다. 관련 기술을 확보한 ICT 글로벌 기업들은 일차적으로 클라우드 서비스를 통해 관련 API를 제공하고 있으며, 구글[59], 아마존[60], MS[61] 등이 대표적이다. 또한, 검색 기술을 기반으로 그림과 같이 실·내외에서 상점을 검색하거나 음식을 검색하는 등의 다양한 시도가 이루어지고 있으며, 스마트폰과 연계한 다양한 비주얼 검색 서비스가 이루어지기 시작했다[(그림 9) 참조].

객체 분류 및 검출 기술은 기존 언어 처리 기술 및 AI 기술과 결합하여 스마트 도시를 위한 지능형 감시[62], 스마트 쇼핑[63] 등 다양한 서비스 활용 가능성을 보여주고 있다. 또한, 객체의 분류, 검출 기술 개발을 확장하여 객체 분할(Object Segmentation), 비정형 요소 분할(Stuff Segmentation) 특징점(Keypoints) 추출 및 객체 간의 관계를 파악하려는 시도(Relationship)[64], [65]가 이루어지고 있다.



(그림 9) 비주얼 검색을 이용한 상점 정보 검색의 예

V. 결론 및 시사점

객체 분류 및 검출 기술은 영상 분석을 위한 핵심 요소 기술로 지속적으로 연구되어 왔다. 최근 이미지넷 데이터베이스를 활용한 ILSVRC 대회와 딥러닝 기술의 적용을 통해서 객체 분류 및 검출 기술의 성능이 비약적으로 발전하였다. 딥러닝 기반 객체 분류 및 검출 기술은 이제 가능성을 넘어서 실제적인 서비스 및 응용을 위한 성능의 지속적인 개선과 함께, 모델의 경량화, 고속화 등의 문제를 해결하기 위해 노력하고 있으며, 객체 분할, 객체관의 관계를 파악하기 위한 시도가 이루어지고 있다.

용어해설

Ensemble 딥러닝으로 학습된 서로 다른 모델의 결과를 결합하여 성능을 향상하는 방법.

Test Augmentation 모델 테스트 시에 성능을 개선하기 위한 여러 기법(Multi Cropping/Scale, Dense Estimation).

NMS(Non Maximum Suppression) 객체 검출 알고리즘의 결과는 알고리즘에 따라 중복된 결과를 포함하고 있어, 위치 정보(Bounding Box)와 신뢰도를 기반으로 정답 후보 결과를 정제하는 방법.

약어 정리

CNN	Convolutional Neural Network
FCN	Fully Convolutional Network
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
mAP	mean Average Precision
MSRA	Microsoft Research Asia
NIN	Network in Network
SPP	Spatial Pyramid Pooling
SVM	Support Vector Machine
UIUC	University of Illinois Urbana-Champaign
VOC	Visual Object Classes

참고문헌

[1] M. Everingham et al., "The Pascal Visual Object Classes (VOC) Challenge - A Retrospective," *IJCV*, vol. 111, no. 1, 2014, pp. 98-136.

[2] O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, vol. 115, no. 3, 2015, pp. 211-252.

[3]Standford University and Princeton University, "ImageNet," Accessed 2018. http://image-net.org

[4] Standford University, "ILSVRC" Accessed 2018. http://image-net.org/challenges/LSVRC/

[5] Princeton University, "WordNet," Accessed 2018. https://wordnet.princeton.edu/

[6] M. Everingham et al., "PASCAL Visual Object Classes Challenge (VOC) 2005-2012," Accessed 2018. http://

host.robots.ox.ac.uk/pascal/VOC/voc2012/index.html

[7] Kaggle, “ImageNet Challenge (Kaggle),” Accessed 2018. https://www.kaggle.com/image-net

[8] K. Yu, T. Zhang, and Y. Gong, “Nonlinear Learning Using Local Coordinate Coding,” in *NIPS*, Vancouver, Canada, Dec. 2009, pp. 2223-2231.

[9] F. Perronnin and C. Dance, “Fisher Kernels on Visual Vocabularies for Image Categorization,” *CVPR*, Minneapolis, MN, USA, June 17-22, 2007, pp. 1-8.

[10] M.A. Hearst et al., “Support Vector Machines,” *IEEE Intell. Syst. Their Applicat.*, vol. 13, no. 4, 1998, pp. 18-28.

[11] F. Perronnin et al., “Large-Scale Image Retrieval with Compressed Fisher Vectors,” *CVPR*, Sna Francisco, CA, USA, June 13-18, 2010, pp. 3384-3391.

[12] J. Sánchez and F. Perronnin, “High-Dimensional Signature Compression for large-Scale Image Classification,” *CVPR*, Colorado Springs, CO, USA, June 20-25, 2011, pp. 1665-1672.

[13] K.E. Van de Sande et al., “Segmentation as Selective Search for Object Recognition,” *ICCV*, Barcelona, Spain, Nov. 6-13, 2011, pp. 1879-1886.

[14] J. Sivic and A. Zisserman, “Video Google: A Text Retrieval Approach to Object Matching in Videos,” *ICCV*, Nice, France, Oct. 14-17, 2003, pp. 1470-1478.

[15] D.G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *IJCV*, vol. 60, no. 2, 2004, pp. 91-110.

[16] A. Krizhevsky, I. Sutskever, and G.E. Hinton, “Imagenet Classification with deep Convolutional Neural Networks,” *NIPS*, Lake Tahoe, CA, USA, Dec. 3-8, 2012, pp. 1097-1105.

[17] Clarifai, “Clarifai Website,” Accessed 2018. https://clarifai.com/

[18] M.D. Zeiler, and R. Fergus, “Visualizing and Understanding Convolutional Networks,” *ECCV*, Zurich, Swiss, Sept. 6-12, 2014, pp. 818-833.

[19] P. Sermanet et al., “Overfeat: Integrated Recognition, Localization and Detection Using Convolutional Networks,” arXiv preprint arXiv: 1312.6229, 2013.

[20] M. Lin, Q. Chen, and S. Yan, “Network in Network,” arXiv preprint arXiv: 1312.4400, 2013.

[21] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition,” arXiv preprint arXiv:1409.1556, 2014.

[22] C. Szegedy et al., “Going Deeper with Convolutions", *CVPR*, Boston, MA, USA, June 8-10, 2015.

[23] C. Szegedy et al., “Rethinking the Inception Architecture for Computer Vision,” *CVPR*, Las Vegas, NV, USA, June 26-July 1, 2016, pp. 2818-2826

[24] C. Szegedy et al., “Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning,” *AAAI*, San Francisco, CA, USA, Feb. 4-9, 2017, pp. 4278-4284.

[25] K. He et al., “Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition,” *IEEE Trans. Pattern Anal Mach. Intell.*, vol. 37, no. 9, 2015, pp. 1904-1916.

[26] R. Girshick et al., “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation,” *CVPR*, Columbus, OH, USA, June 24-27, 2014, pp. 580-587.

[27] R. Girshick, “Fast R-CNN,” *CVPR*, Boston, MA, USA, June 8-10, 2015. pp. 1440-1448.

[28] S. Ren et al., “Faster R-CNN: Towards real-Time Object Detection with Region Proposal Networks,” *NIPS*, Montreal, Canada, Dec. 7-12, 2015, pp. 91-99.

[29] ICCV 2015, “ImageNet and MS COCO Visual Recognition Challenges Joint Workshop,” Accessed 2018. http://image-net.org/challenges/ilsvrc+ mscoco2015

[30] COCO, “COCO (Common Objects in Context,” Accessed 2018. http://cocodataset.org/

[31] K. He et al., “Identity Mappings in Deep Residual Networks,” *ECCV*, Amsterdam, Netherlands, Oct. 8-16, 2016, pp. 630-645.

[32] The Third Research Institute of the Ministry of Public Security, P.R. China, “TRIMPS,” Accessed 2018. http://hr.trimps.ac.cn/

[33] SenseTime, “SenseTime,” Accessed 2018. https://www.sensetime.com/

[34] Hikvision, “Hikvision,” Accessed 2018. http://en.hikrobotics.com/welcome.htm

[35] Y. Li, et al., “Fully Convolutional Instance-Aware Semantic Segmentation,” *CVPR*, Honolulu, HI, USA, July 21-26, 2017, pp. 4438-4446.

[36] S. Xie et al., “Aggregated Residual Transformations for Deep Neural Networks,” *CVPR*, Honolulu, HI, USA, July 21-26, 2017, pp. 5987-5995.

[37] Cityscapes Team, “Cityscapes Dataset,” Accessed 2018. https://www.cityscapes-dataset.com/

[38] Google, “Youtube8M Dataset,” Accessed 2018. https://research.google.com/youtube8m/

[39] Google, "Open Images Dataset v4," Accessed 2018. https://storage.googleapis.com/openimages/web/index.html

[40] Qihoo 360, "About Qihoo 360," Accessed 2018. https://www.360totalsecurity.com/en/about/

[41] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," arXiv preprint arXiv:1709.01507, 2017.

[42] Y. Chen et al., "Dual Path Networks," *NIPS*, Long beach, CA, USA, Dec. 4-9, 2017, pp. 4470-4478.

[43] G. Huang et al., "Densely Connected Convolutional Networks," *CVPR*, Honolulu, HI, USA, July 21-26, 2017, pp. 4700-4708.

[44] X. Zhang et al., "PolyNet: A Pursuit of Structural Diversity in Very Deep Networks," *CVPR*, Honolulu, HI, USA, July 21-26, 2017, pp. 718-726.

[45] M. Denil et al., "Predicting Parameters in Deep Learning," *NIPS*, Lake Tehoe, CA, USA, Dec. 5-10, 2013.

[46] M. Jaderberg, A. Vedaldi, and A. Zis-serman, "Speeding up Convolutional Neural Networks with Low Rank Expansions," *BMVC*, Nottingham, UK, 2014.

[47] E. Denton et al., "Exploiting Linear Structure within convolutional Networks for Efficient Evaluation," *NIPS*, Montreal, Canada, Dec. 8-13, 2014.

[48] V. Lebedev et al., "Speeding-up Convolutional Neural Networks Using Fine-Tuned CP-Decomposition," *ICLR*, San Diego, CA, USA, May 7-9, 2015.

[49] Y. Gong, L. Liu, and L. Bourdev, "Compressing Deep Convolutional Networks Using Vector Quantization," arXiv preprint arXiv:1412.6115v1, Dec. 2014.

[50] S. Hang et al., "Learning Both Weight and CONNEctions for Efficient Neural Networks," *NIPS*, Montreal, Canada, Dec. 7-12, 2015.

[51] A.G. Howard et al., "Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv preprint arXiv:1704.04861, 2017.

[52] M. Sandler et al., "Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation," arXiv preprint arXiv:1801.04381, 2018.

[53] X. Zhang et al., "Shufflenet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," arXiv preprint arXiv:1707.01083, 2017.

[54] F.N. Iandola et al., "SqueezeNet: AlexNet-Level Accuracy with 50x Fewer Parameters and < 0.5 MB Model Size," arXiv preprint arXiv:1602.07360, 2016.

[55] IEEE Rebooting Computing, "Low-Power Image Recognition Challenge," Accessed 2018. https://rebootingcomputing.ieee.org/lpirc

[56] Google, "Tensorflow Lite," Accessed 2018. https://www.tensorflow.org/mobile/tflite/

[57] Facebook, "Caffe2," Accessed 2018. https://caffe2.ai/

[58] Apple, "CoreML," Accessed 2018. https://developer.apple.com/documentation/coreml

[59] Google, "Google Cloud AI," Accessed 2018. https://cloud.google.com/products/machine-learning/

[60] Amazon, "AWS 기반 기계 학습," Accessed 2018. https://aws.amazon.com/ko/machine-learning/

[61] Microsoft, "Microsoft AI," Accessed 2018. https://www.microsoft.com/en-us/ai/

[62] Nokia, "Nokia Smart City," Accessed 2018. https://networks.nokia.com/smart-city

[63] Amazon, "Amazon Go," Accessed 2018. https://www.amazon.com/b?ie=UTF8&node=16008589011

[64] C. Lu et al., "Visual Relationship Detection with Language Priors," *ECCV*, Amsterdam, Netherlands, Oct. 8-16, 2016, pp. 852-869.

[65] A. Santoro et al., "A Simple Neural Network Module for Relational Reasoning," arXiv preprint arXiv:1706.01427, 2017.