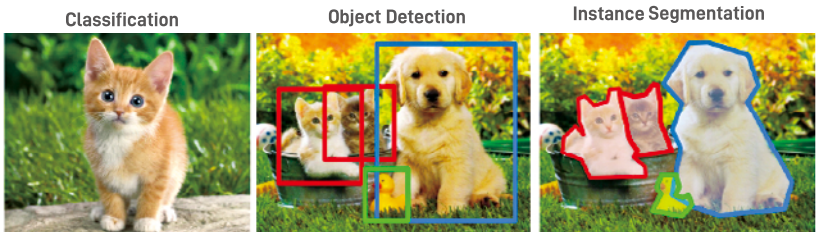
# 인공지능 이미지 인식 기술 동향

**이주열** LG CNS AI빅데이터연구소 연구소장

## 1. 머리말

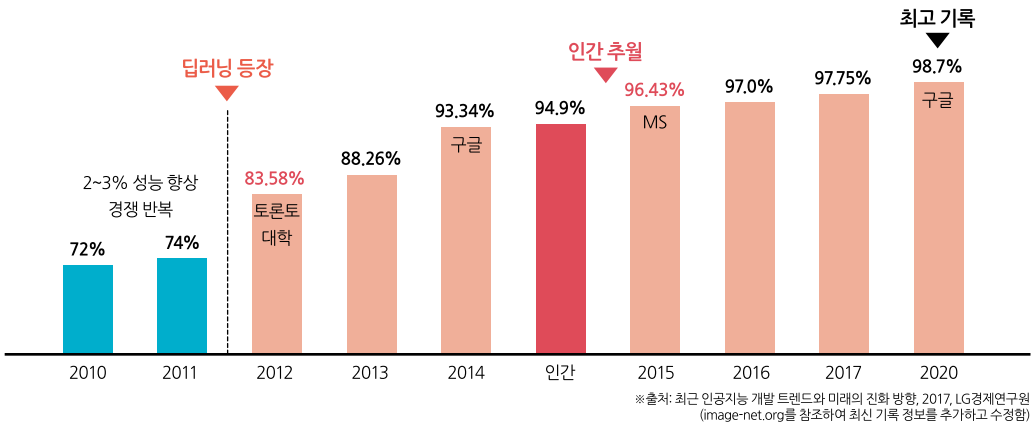
인공지능 이미지 인식은 기계가 마치 사람처럼 사진이나 동영상으로부터 사물을 인식하거나 장면을 이해하는 것으로 정의할 수 있다. 이러한 이미지 인식은 컴퓨터 비전(Computer Vision) 기술 중 하나에 해당한다. 이미지 인식에는 대표적으로 세 가지 태스크(Task)가 존재하는데, [그림 1]과 같이 이미지 내 특정 사물을 분류(Classification)하는 태스크, 여러 사물을 동시에 검출(Detection)하는 태스크, 사물들을 픽셀 단위로 식별하여 분할(Segmentation)하는 태스크 등이 있다.

이러한 이미지 인식 기술에 있어 2012년 혁신적인 연구 결과가 나오게 되는데, 대규모 이미지 인식 경진대회인 ILSVRC(ImageNet Large Scale Visual Recognition Challenge)에서 토론토 대학 연구진이 딥러닝(Deep learning)이라 불리는 새로운 기법을 활용해 기존의 방법론에 대비해 압도적인 성능으로 우승한 것이다. 이 연구는 저명한 신경정보처리시스템(NeurIPS, Neural Information Processing Systems) 학회에 발표되어 지금까지도 인공지능 분야에서 가장 많이 인용되고 있는 논문 중 하나가 되었다[1]. 이를 계기로 딥러닝이 학계, 산업계에 널리 받아들여지게 됨에 따라 딥러닝 또한 폭발적으로 발전하여, 2015년 ILSVRC에서 사람의 인식률(94.90%)을 추월(96.43%)하고, 2020년에는 사람을 한참 뛰어넘는 수준(98.7%)으로 진화했다[2].




※출처: http://cs231n.stanford.edu/

[그림 1] 이미지 인식의 대표적 태스크




[그림 2] ILSVRC 연도별 정확도 향상[3]

사람 수준을 초월한 인공지능 이미지 인식 기술은 자율주행, 의료, 제조 등의 산업에 활용되는 단계로 진입하고 있다. 이러한 성과를 통해 딥러닝은 현재 인공지능 기술의 핵심이라고 할 수 있다. 따라서, 본고에서는 딥러닝을 중심으로 인공지능 이미지 인식 기술의 동향을 조망하고자 한다.

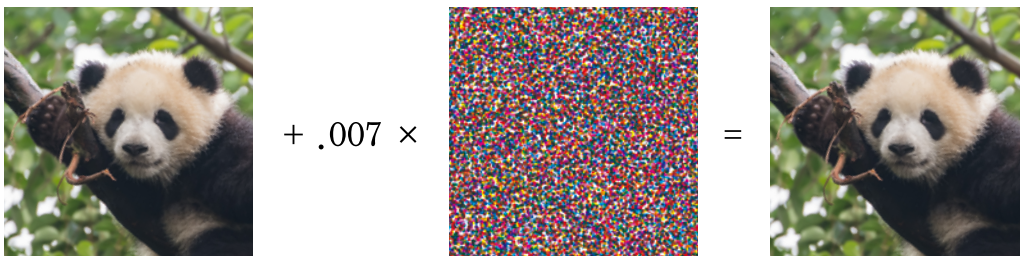
## 2. 딥러닝 기반 이미지 인식 기술 동향

### 2.1 안전하고 신뢰성 있는 이미지 인식

2018년 우버와 테슬라 차량이 자율주행 중 사망 사고를 냈다. 각기 보행자 인식과 차선 인식의 오류로 촉발된 사고였다. 자율주행에 탑재된 인공지능 이미지 인식 기술이 실험실을 벗어나자 사고를 낸 것이다. 인공지능의 활용도가 높아질수록 기술에 대한 안전성과 신뢰성을 확보하는 것은 선택이 아닌 필수가 된다.

특히 딥러닝은 [그림 3]처럼 적대적 예제(Adversarial examples)라 불리는 이미지를 엉뚱하게 판단하는데, 악의적 노이즈(Adversarial noise)를 이미지에 주입해서 적대적 예제를 만들 수 있다.

이 적대적 예제는 사람이 판단하기에는 문제가 없으나 인공신경망의 판정을 교란시킬 수 있다. 이를 악용할 경우 문제가 될 수 있는데, 적대적 예제를 응용하여 인공신경망이 교통 표지판이나 사람을 인식하지 못하게 만드는 사례가 등장했다[5][6]. 이러한 악의적 이미지 인식 교란에 대해서 올바른 판



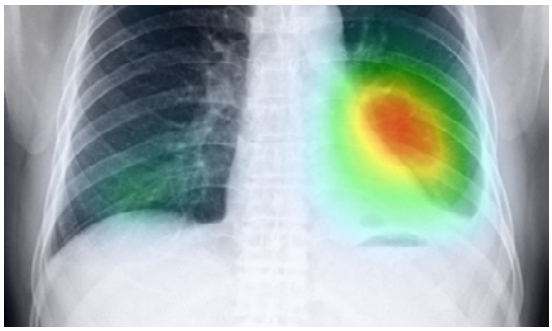

[그림 3] 적대적 예제 사례[4]

정을 할 수 있도록 적대적 예제 자체를 학습하는 적대적 학습(Adversarial training) 방법이나 노이즈를 제거·완화하는 노이즈 감쇄기(Denoiser) 방식 등이 연구되고 있다.

적대적 예제처럼 인공의 노이즈에 대응하는 기술 외에도 이미지 인식의 강건함(Robustness) 자체를 개선하기 위한 연구가 활발한데, 블러링(Blurring), 포깅(Fogging) 등 현실에서 흔하게 발생할 수 있는 노이즈가 적용된 이미지 벤치마킹 데이터세트 ImageNet-C, ImageNet-P 등이 공개되었다[7]. 이를 통해서 각종 노이즈에 대해 안정성과 신뢰성을 갖춘 강건한 이미지 인식 기술이 발전할 것으로 기대된다.

노이즈 대응 외에 현재 딥러닝 기반 이미지 인식 기술의 근본적인 불안감 중 하나는 학습하지 못한 패턴에 대한 처리이다. 이는 딥러닝과 같은 데이터 학습 기반의 기계학습(Machine learning) 기술들이 가진 한계이다. 이 한계를 극복하기 위해서는 인공지능에 입력된 이미지가 학습된 확률 분포(Probability distribution)의 데이터인지, 아닌지를 식별하는 것이 중요한데, 이것을 '학습 외 분포 데이터 탐지(Out-of-Distribution Detection)'라고 한다. 학습 외 분포 데이터 탐지를 통해 인공신경망이 판단하기 어려운 이미지를 걸러 내거나 예외 처리하여 안전성과 신뢰성을 높일 수 있다. 학습 외 분포 데이터 탐지를 위해서 딥러닝이 판정에 대해 얼마나 확신(Confidence)하는지를 나타내는 확률 값을 보정(Calibration)하거나, 학습 외 분포 데이터를 생성적 대립 신경망(GAN, Generative Adversarial Network)으로 생성하고 학습하여 탐지 정확도를 높이는 방법이 있다[8].

신뢰성 확보에 있어 또 다른 접근은 '설명 가능 인공지능(XAI, eXplainable AI)'이다. XAI는 인공지능의 동작 또는 판단을 사람이 이해할 수 있는 형태로 설명하는 기술을 의미한다. 예를 들어 인공지능이 고양이 이미지를 분류할 경우 판단 결과만을 제공하는 것이 아니라, 고양이라고 판단한 근거(수염, 뾰족한 귀 등)까지 제공한다. 이와 같이 인공지능의 판단에 대해 왜 그렇게 작동하는지를 이해할 수 있다면, 판단 결과에 대하여 신뢰 여부를 결정할 수 있다. 특히 의료 영상 분석과 같이 법적 책임 또는 규정 준수가 엄밀히 요구되는 분야에 인공지능을 도입할 수 있는 안전 장치가 될 수 있다. XAI는 신뢰성 외에 인공지능 자체의 성능 향상에도 기여할 수 있는 기술이지만, 아직 초기 연구 수준에 머물러 있다. 그러나, 인공지능 판단에 크게 기여하는 특징(Feature) 정보를 알려 주는 CAM(Class Activation Map)과 같은 기술들[9][10]은 이미 활용되고 있다[11].



※출처: lunit.io

[그림 4] XAI 사례
(폐렴 판정 흉부 X-Ray에서 오른쪽 중간에 증상을 표시)

정확도 측면에서 인공지능 이미지 인식 기술이 사람 이상의 수준으로 진화되었으나 현실 세계에 적용하기에는 안전성 및 신뢰성에 대한 추가적 보완이 요구되고 있는 실정이다. 앞서 설명한 바와 같이 인공지능 이미지 인식 기술은 품질 수준을 개선하는 방향으로 이제 막 발돋움을 하고 있는 단계이다.

### 2.2 인공지능 학습의 한계 극복

딥러닝으로 인공지능이 빠르게 발전하고 있지만, 딥러닝은 학습 과정에서 대규모의 데이터와 컴퓨팅 파워를 요구한다. 또한 인공지능 인재 확보 전쟁이


[그림 5] **어노테이션 예시**(수작업으로 이미지 내 각 사물을 식별하고 표기함. 태스크에 따라 어노테이션 방법이 상이)

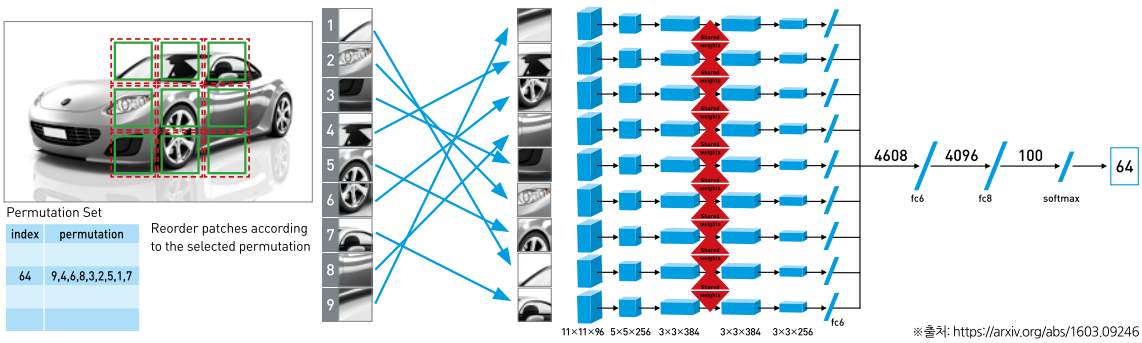
라는 표현이 나올 만큼 인적 자원도 제한적이다. 최근에는 이러한 한계를 극복하기 위해 크게 세 가지 정도의 동향이 두드러지고 있다.

첫 번째는 학습 데이터 절감이다. 앞서 언급한 바와 같이 딥러닝에는 엄청난 양의 데이터가 필요한데, 특히 이미지 인식에 있어서 대부분 지도 학습(Supervised learning)을 실시하는 만큼 출력값이 있는 데이터가 필요하다. 즉, 출력값에 해당하는 라벨(Label) 또는 어노테이션(Annotation)을 준비해야 하는데, 이미지에 대한 어노테이션은 [그림 5]와 같이 인식하고자 하는 출력값에 따라 이미지 내 각종 사물을 일일이 구분하여 경계선을 그리거나 해당 사물이 무엇인지 기록해야 한다.

이미지 어노테이션은 규모와 난이도 등에 따라 많은 비용이 소요되는데, 앞서 언급되었던 ILSVRC에 활용되는 ImageNet 데이터 구축의 경우 1400만 장의 이미지 어노테이션을 크라우드소싱을 통해 약 4만9천여 명이 4년에 걸쳐 작업했다[12]. 더욱이 의료 영상과 같은 전문 영역일수록 어노테이션 비용은 크게 증가할 수밖에 없다.

이러한 어노테이션 비용 문제를 극복하는 최근 동향 중 하나는 '자기 지도 학습(Self-supervised learning)'을 활용하여 학습에 필요한 어노테이션을 줄이는 것이다. 자기 지도 학습은 어노테이션 없이 데이터를 학습하는 방법에서 비지도 학습(Unsupervised learning)과 유사하지만 데이터를 표현(Representation)하는 방법을 학습하는 것에 중점을 두고 있다. [그림 6]의 예시처럼 자기 문제를 해결하기 위해서 스스로 필요한 특징을 찾아 적절하게 표현(Feature representation)하는 방법을 학습하는 것이다.

자기 지도 학습 후에는 이른바 미세 조정(Fine




[그림 6] **자기 지도 학습 예시**(직소(Jigsaw) 퍼즐처럼 이미지 조각을 끼워 맞추어 올바른 이미지가 완성되도록 학습[13])

tuning) 또는 다운스트림 태스크(Downstream task)에 해당하는 소수의 어노테이션 데이터로 지도 학습 과정을 거쳐 최종적으로 목적한 이미지 인식을 완성하게 된다. 이와 같은 자기 지도 학습 방식의 장점은 고비용에 해당하는 어노테이션 데이터를 절약할 수 있다는 것과 추후 설명할 전이 학습(Transfer learning) 등에 활용되는 사전 학습(Pre-trained) 모델을 확보할 수 있다는 것이다.

어노테이션 데이터 사용을 효율화하는 또 다른 방법은 액티브러닝(Active learning)이다. 액티브러닝의 핵심은 '어떤 데이터를 먼저 어노테이션해서 학습에 사용할 것인가'이다. 즉, 무작위로 많은 데이터를 어노테이션한 후 학습하는 것이 아니라, 학습에 크게 기여할 수 있는 데이터를 먼저 선별하여 어노테이션하고 학습한 후 점진적으로 목표한 정확도를 달성할 때까지 어노테이션 데이터를 늘려가며 학습하는 방식이다. 동일한 최고 정확도를 달성할 때 무작위로 데이터를 어노테이션한 경우보다 액티브러닝을 활용할 경우 40%가량 어노테이션 데이터를 절감하고, 같은 규모의 어노테이션 데이터를 학습한 경우라도 액티브러닝으로 최고 정확도를 3.4% 높인 연구 사례가 있다[14]. 이 외에도 생성적 대립 신경망(GAN)을 활용하여 가상의 데이터를 생성하는 방식으로 어노테이션 데이터를 만들어내는 시도가 있다[15].

두 번째 동향은 부족한 인공지능 전문가를 대신할 학습 자동화이다. 자동화된 기계학습(AutoML, Automated Machine Learning) 기술은 학습 데이터 전처리, 딥러닝의 심층 신경망 구조 탐색(NAS, Neural Architecture Search), 학습 최적화를 위한 하이퍼파라미터(Hyper-parameter) 조정, 최종 모델 선택(Model selection) 등과 같은 딥러닝 학습 과정의 각 단계를 자동화할 수 있다. AutoML은 인공지능 전문가 부족에 대한 극복 방안이자 인공지능 민주화(AI Democratization)를 위한 기술이라고도 할 수 있다. 또한, 최근에는 AutoML이 최고 수준의 인공지능 이미지 인식 성능을 달성하기 위한 방법으로도 활용되고 있는데, 자동화된 신경망 구조 탐색(NAS)으로 찾은 심층 신경망 모델을 인공지능 전문가가 미세 조정하여 최고의 성능을 달성하는 방식이다[16][2].

세 번째 동향은 학습 데이터와 컴퓨팅 파워 절감을 위한 전이 학습(Transfer learning) 고도화다. 전이 학습은 원천 도메인(Source domain)으로부터 목표 도메인(Target domain)을 학습시키는 방법의 총칭이며, 딥러닝 이미지 인식에서는 원천 도메인에서 학습된 심층 신경망 모델을 목표 도메인 데이터로 추가 학습하는 방식으로 구현된다. 전이 학습은 손쉽게 학습 데이터와 컴퓨팅 파워 또는 학습 시간을 줄일 수 있는 방법이기에 딥러닝 기반 이미지 인식에 있어 기본처럼 활용되고 있다. 그러나, 최근에 전이 학습이 더욱 주목받는 이유는 앞서 언급한 것처럼 자기 지도 학습(Self-supervised learning), AutoML 기술 등이 고도화되고 이를 적용한 사전 학습 모델(Pre-trained model) 자체의 성능도 고도화되면서 전이 학습을 다양한 태스크에 적용할 수 있기 때문이다[17]. 또한, 전이 학습의 성능을 더 높이기 위한 조건 등이 연구되면서 의료 영상, 제조 비전 검사 등의 전문 도메인 영역에 특화된 사전 학습 모델들이 개발되고 있다[18][19].

### 2.3 온 디바이스(On-Device) 인공지능 이미지 인식

구글이 만드는 스마트폰의 카메라 기술이나 이미지 인식 기술에 딥러닝이 사용되고 있는 것은 익히 알려진 사실이다. 스마트폰에서 딥러닝으로 인물사진을 더욱 또렷하게 만들거나[20] 상품이나 사물 인식 등을 처리한다[21]. 이와 같이 모바일 디바이스, 경량 디바이스 등에서 인공지능 이미지 인식

기술 적용 사례가 등장하고 있으며, 이러한 추세에 따라 경량(Lightweight) 딥러닝 연구와 하드웨어 가속화 기술 연구가 진행되고 있다[22].

경량 딥러닝 기술은 정확도를 유지하면서 모델의 크기를 줄이거나 연산을 간소화하여 작은 디바이스 등에 탑재할 수준으로 경량화하는 것이다. 특히 이미지 인식에 주로 사용되는 콘볼루션 신경망(CNN, Convolutional Neural Network)의 경우 콘볼루션 필터를 변형하여 연산 차원을 축소(Reduction)하거나 큰 영향이 없는 신경망의 가중치(Weight)를 삭제하는 가지치기(Pruning), 가중치 값의 부동 소수점을 줄여 연산을 간소화하는 양자화(Quantization) 등의 기법이 있다. 최근에는 지식증류(Knowledge Distillation) 활용도가 높아지고 있는데, 미리 잘 학습시킨 큰 신경망의 출력을 작은 신경망이 모방 학습하여 상대적으로 경량화 되면서도 정확도를 유지하는 기술이다[23].

하드웨어 가속화 기술 연구는 주요 IT 기업 주도로 인공지능의 판정 연산 가속화를 위한 전용 칩세트 개발 중심으로 진행되고 있다. 대표적으로 구글의 EdgeTPU, 인텔의 Movidius, 엔비디아 Jetson 등이 있고 퀄컴, 애플 등 모바일 AP(Application Processor) 개발사들은 NPU(Neural Processing Unit)를 표방하며, 딥러닝 처리를 모바일 AP에 통합해 구현하고 있다.

## 3. 맺음말

딥러닝의 발전으로 인공지능 이미지 인식 수준은 정확도 측면에서 사람의 수준을 뛰어넘기도 했지만, 의료산업, 자율주행, 제조산업, 안전산업 영역 등에 실제 적용되면서부터 한편으로는 한계에 부닥치기 시작했다. 그렇기 때문에 최근 인공지능 이미지 인식 기술 동향은 이러한 한계를 돌파하기 위한 방향으로 진행되고 있다. 즉, 앞서 상술한 내용을 요약하자면, 인공지능 판정의 안전성과 신뢰성을 높이기 위해 악의적 공격, 자연 발생적 노이즈 그리고 예외 상황에 대해서도 강건하고 투명하게 처리하고, 인공지능 학습에 필요한 막대한 비용과 자원을 최소화하며, 일상 모든 곳에 인공지능 적용이 가능하도록 가볍게 만드는 것이 최근의 인공지능 이미지 인식 기술 동향이다. 이는 결국 현재의 딥러닝 기술 동향과 동일한데, 특히 이미지 인식에 있어 중요한 기술을 중심으로 지금까지 살펴봤다.

범용 인공지능(Artificial General Intelligence)처럼 인공지능의 근본적인 한계를 해결하기 위한 연구 동향도 있지만, 각 산업 영역에 인공지능 기술을 적용하기 위한 현실적 문제를 해결하는 것 역시 주요 기술 동향이라고 할 수 있다. 이렇게 수많은 현실적 문제를 차례 차례 돌파하다 보면 인류의 꿈과 같은 범용 인공지능도 어느덧 우리 곁에 현실로 다가올 수 있지 않을까 기대한다.

## 주요 용어 풀이

- 컴퓨터 비전: 사람이나 동물 시각 체계의 기능을 컴퓨터로 구현하는 것
- 기계학습: 컴퓨터 프로그램이 데이터와 처리 경험을 이용한 학습을 통해 정보 처리 능력을 향상시키는 것
- 지도 학습: 기계학습 중 컴퓨터가 입력 값과 그에 따른 출력 값이 있는 데이터를 이용하여 주어진 입력에 맞는 출력을 찾는 학습 방법
- 비지도 학습: 기계학습 중 컴퓨터가 입력 값만 있는 훈련 데이터를 이용하여 입력들의 규칙성을 찾는 학습 방법
- 인공 신경망: 사람 또는 동물 두뇌의 신경망에 착안하여 구현된 컴퓨팅 시스템의 총칭
- 심층 신경망: 입력층(input layer)과 출력층(output layer) 사이에 다중의 은닉층(hidden layer)을 포함하는 인공 신경망
- 딥러닝: 일반적인 기계학습 모델보다 더 깊은 신경망 계층 구조를 이용하는 기계학습
- 생성적 대립 신경망: 생성모델과 판별모델이 경쟁하면서 실제와 가까운 이미지, 동영상, 음성 등을 자동으로 만들어 내는 기계학습 방식의 하나
- 콘볼루션 신경망: 심층 신경망의 한 종류로 하나 또는 여러 개의 콘볼루션 계층(convolutional layer)과 통합 계층(pooling layer), 완전하게 연결된 계층(fully connected layer)들로 구성된 신경망
- 범용 인공지능: 특정 문제뿐 아니라 주어진 모든 상황에서 생각과 학습을 하고 창작할 수 있는 능력이 있는 인공지능
- 제조 비전 검사(Vision Inspection): 제조 공장 등에서 생산되는 제품 및 부품에 대한 외관 검사를 컴퓨터 비전 기술로 처리하는 것

## 참고문헌


[1] Krizhevsky, A. et al., 'ImageNet Classification with Deep Convolutional Neural Networks', NeurIPS 2012.

[2] Qizhe Xie. et al., 'Self-training with Noisy Student improves ImageNet classification', arXiv:1911.04252v2, 2019. https://arxiv.org/abs/1911.04252v2

[3] 최근 인공지능 개발 트렌드와 미래의 진화 방향(2017, LG경제연구원)

[4] Ian J. Goodfellow. et al., 'Explaining and Harnessing Adversarial Examples', ICLR 2015.

[5] Kevin Eykholt. et al., 'Robust Physical-World Attacks on Deep Learning Models', CVPR 2018.

[6] Simen Thys. et al., 'Fooling automated surveillance cameras: adversarial patches to attack person detection', arXiv:1904.08653, 2019. https://arxiv.org/abs/1904.08653

[7] Dan Hendrycks. et al., 'Benchmarking Neural Network Robustness to Common Corruptions and Perturbations', ICLR 2019.

[8] Kimin Lee. et al., 'Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples', ICLR 2018.

[9] Bolei Zhou. et al., 'Learning Deep Features for Discriminative Localization', CVPR 2016.

[10] Selvaraju, R. R. et al., 'Grad-CAM: Visual explanations from deep networks via gradient-based localization', CVPR 2017.

[11] Han Liu. et al., 'SDFN: Segmentation-based Deep Fusion Network for Thoracic Disease Classification in Chest X-ray Images', arXiv:1810.12959, 2018. https://arxiv.org/abs/1711.05225

[12] http://image-net.org/

[13] Mehdi Noroozi. et al., 'Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles', ECCV 2016.

[14] Donggeun Yoo. et al., 'Learning Loss for Active Learning', CVPR 2019.

[15] Wenyuan Li. et al., 'Semi-supervised learning based on generative adversarial network: a comparison between good GAN and bad GAN approach', arXiv:1905.06484, 2019. https://arxiv.org/abs/1905.06484

[16] Mingxing Tan. et al., 'EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks', ICML 2019.

[17] Alexander Kolesnikov. et al., 'Large Scale Learning of General Visual Representations for Transfer', arXiv:1912.11370, 2019. https://arxiv.org/abs/1912.11370

[18] Weifeng Ge. et al. 'Borrowing Treasures from the Wealthy: Deep Transfer Learning through Selective Joint Fine-tuning', CVPR 2017.

## 참고문헌


[19] Maithra Raghu. et al., 'Transfusion: Understanding Transfer Learning for Medical Imaging', NeurIPS 2019.

[20] Improvements to Portrait Mode on the Google Pixel 4 and Pixel 4 XL, https://ai.googleblog.com/2019/12/improvements-to-portrait-mode-on-google.html

[21] https://lens.google.com/

[22] 경량 딥러닝 기술 동향(2019, ETRI)

[23] G. Hinton. et al., 'Distilling the Knowledge in a Neural Network', arXiv:1503.02531, 2015. https://arxiv.org/abs/1503.02531