산업인공지능

정해동, 이승철*
(울산과학기술원)

1. 머리말

우리는 검색엔진을 이용하거나 휴대폰을 이용하여 사진을 촬영하는 일상적인 작업에서도 인공지능(artificial Intelligence)의 도움을 받고 있다. 인공지능은 기계나 컴퓨터가 인간과 유사하거나 그 이상의 지능을 갖도록 하는 기술로 다양한 방법들을 통해 구현되고 있다. 최근 기계학습(machine learning)의 발전과 딥러닝(deep learning)의 등장은 인공지능의 새로운 시대를 열었고 많은 분야에 적용되어 우수한 성능을 보이고 있다.

대표적으로 영상처리와 음성 인식 분야를 예로 들 수 있다. 영상처리 분야에서 딥러닝은 연구의 패러다임을 바꾸었다. 사람에 의해 정의된 특성인자를 이용한 기존의 사물 인식 방식은 딥러닝 모델 학습과정을 통해 사물 구분을 위한 최적 특성인자를 스스로 찾는 방식으로 변화하였다. 인식 성능 역시 딥러닝을 활용한 모델이 사람의 인식 오차와 데이터 라벨 자체의 오류를 생각했을 때 거의 인간과 같은 수준으로 사물을 인식하고 분류한다고 할 수 있다. 음성인식 분야 또한, 딥러닝을 통해 크게 발전했다. 전통적으로 음성 인식은 음성 데이터의 연속적인 특성을 hidden Markov model(HMM) 등을 통해 학습했으나, 이런 구조는 인간의 음성을 이해하고 표현하기에 한계점이 많았다. 딥러닝은 비선형 활성화 함수를 채택한 뉴런을 선형적으로 결합한 복잡한 구조를 통해 기존 모델의 한계점을 극복하고 대량의 데이터를 통해 인간의 음성을 이해하려고 한다. 애플의 시리, 구글 어시스턴트와 같은 인공지능 기반의 음성 비서가 휴대폰에 탑재되거나 인공지능 스피커를 통해 일상에서 많은 서비스를 제공하고 있다. 그 외에도 인공지능을 활용해서 의료 영상을 분석하거나 자율 주행 자동차의 두뇌 역할을 하는 등 인공지능의 적용 분야는 계속해서 확대되고 있다.

공장이나 발전소와 같이 산업 전반에서 발생하는 데이터 역시 영상 데이터나 음성 데이터와 유사한 형태를 보인다. 그렇다면 이를 위한 산업인공지능(industrial AI)의 개발도 생각해 볼 수 있지 않을까? 인공지능은 산업에 어떻게 적용될 수 있을까?

2. 산업인공지능

인공지능을 이용하여 제품의 생산성을 향상시키거나 생산 설비의 유지보수 비용을 줄이는 등 다양한 성공 사례를 통해 설명하고자 한다.

2.1 제품 품질 검사를 위한 인공지능

대부분의 제조라인에서 조립 품질을 정량화

* E-mail : seunglee@unist.ac.kr

하기 위해 음향 또는 진동 센서 등의 다양한 센서를 부착하여 데이터를 취득하고 이를 분석하고 있다. 일반적으로 분석에는 전문가의 지식과 경험이 이용된다. 예를 들면 기어 결함 주파수와 같이 고장에 관계된 특정 주파수를 감시하여 제품의 품질을 확인한다. 하지만 생산 공정이 복잡해지면서 몇 개의 인자 크기를 단순 비교하는 방법으로는 제품의 품질을 정확하게 예측하는 것이 쉽지 않아졌다. 학습 과정 중 데이터의 패턴을 스스로 분석할 수 있는 인공지능을 적용한다면 품질 예측의 정확도를 증가시킬 수 있을 것이라 기대할 수 있다.

합성곱 신경망은 이미지 분류 분야에서 높은 성능을 보이는 구조로 학습 과정에서 이미지를 분류하기 위한 최적의 커널을 학습한다. 조립 과정 중 음향 또는 진동 데이터를 취득하고 short-time fourier transform(STFT)를 통해 시간-주파수에 대한 이미지를 분석하기 때문에 합성곱 신경망은 제품 품질을 예측하는데 효과적으로 사용될 수 있다. 그림 1은 음향 데이터의 시간-주파수 이미지에 대한 합성곱 신경망 적용 예시이다.

전통적인 분석 방법으로는 시간-주파수 이미지에서 특정 시간대 혹은 특정 주파수대의 비정상적인 패턴을 확인하여 고장의 유무를 확인하기 때문에 전문가의 지식을 통해 특정 주파수나 시간의 위치(특성인자 추출)가 선행적으로 결정되어야 했다. 하지만 예시로 든 딥러닝 모델의 경우, 정상 이미지와 비정상 이미지를 구분하기 위한 커널을 스스로 학습하고 이미지의 정상/비정상을 예측한다. 주파수의 크기와 같은 단순한 형태부터 추상적인 형태의 특성인자까지 다양하게 학습될 수 있으며, 뿐만 아니라 특성인자는 합성곱 신경망의 성질(convolution layer and pooling layer) 때문에 공간 불변성(spatial invariance)도 지닐 수 있어 노이즈에 강건한 예측 결과를 추론할 수 있다. 결과적으로 이미지 분석에 사용하는 합성곱 신경망은 진동 데이터로부터 제품의 품질을 예측하는 산업인공지능으로 탈바꿈한다.

2.2 생산 방식의 변화를 위한 인공지능

포스코에서는 '인공지능 기반 도금량 제어 자동화 솔루션'을 개발하여 자동차 강판 생산에서 인공지능을 통해 도금 작업을 정밀하게 제어하여 도금량 편차를 획기적으로 줄였다고 발표하였다[1]. 이 기술은 도금량 예측 모델과 최적화 기법의 제어모델이 결합하여 실시간으로 도금량을 예측하고 목표 도금량을 정확히 맞추는 자동제어 기술이다. 이처럼 인공지능은 단순히 데이터의 정상, 비정상 여부를 판단할 수 있을 뿐만 아니라, 주어진 데이터로부터 숨겨진 패턴을 추론하고 제어모델과 같은 전통 이론과 연계하여 활용할 수 있다.

철강공정에 딥러닝이 적용되어 생산 설비의 상태 변화를 예측하고 이를 통해 생산 방식의 효율성을 증가 시킨 사례도 있다(그림 2)[2]. 철

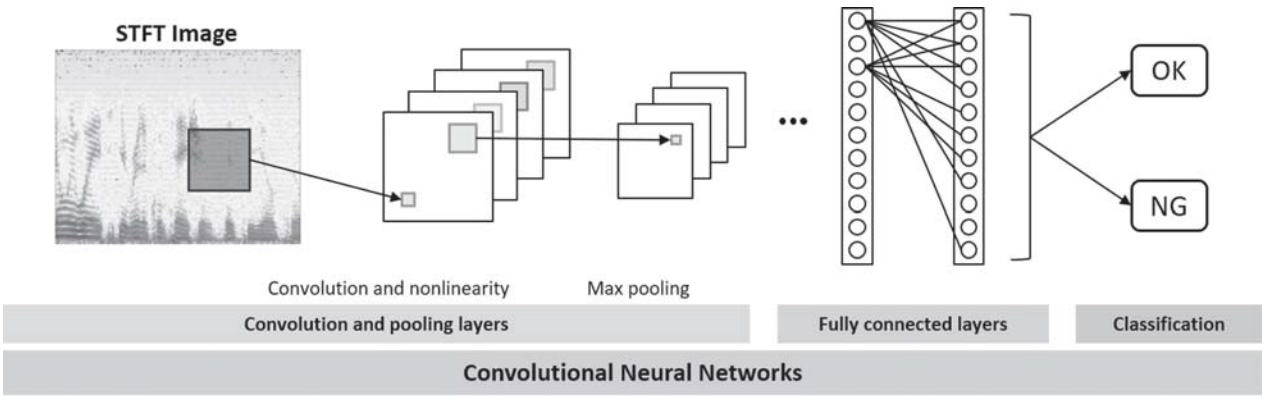


그림 1 음향 데이터에 대한 합성곱 신경망 적용 예시

강공정에는 약 400여종의 시계열 센서 데이터가 발생하는데, 전통적인 데이터 분석 방법을 통해 관계성을 파악하여 고로와 같은 주요 설비의 상태를 예측하기가 쉽지 않다. 비지도 학습을 통해 시계열 센서 데이터를 군집화하는 전처리과정과 딥러닝 모델을 함께 활용하여 철강공정의 발생하는 데이터들의 관계성을 파악하여 고로의 용선온도를 높은 정확도로 예측하였다. 이처럼 딥러닝은 공정에서 발생하는 많은 데이터 간의 관계성을 분석하여 설비 상태 변화를 예측할 때도 좋은 성능을 보이고 있다.

2.3 생산 설비 유지보수를 위한 인공지능

노후화된 생산 설비에서는 예측하기 어려운 고장 및 사고가 발생할 수 있어, 제품 생산성과 안전 문제에 큰 영향을 미칠 수 있다. 산업에서 이행되고 있는 대부분의 설비 유지 보수 방법은 사후 정비이기 때문에 비용 측면에서도 비효율적이다. 이를 개선하고자 최근 산업에서는 설비 상태 진단 및 모니터링 기술을 운영하고자 노력하고 있으나,

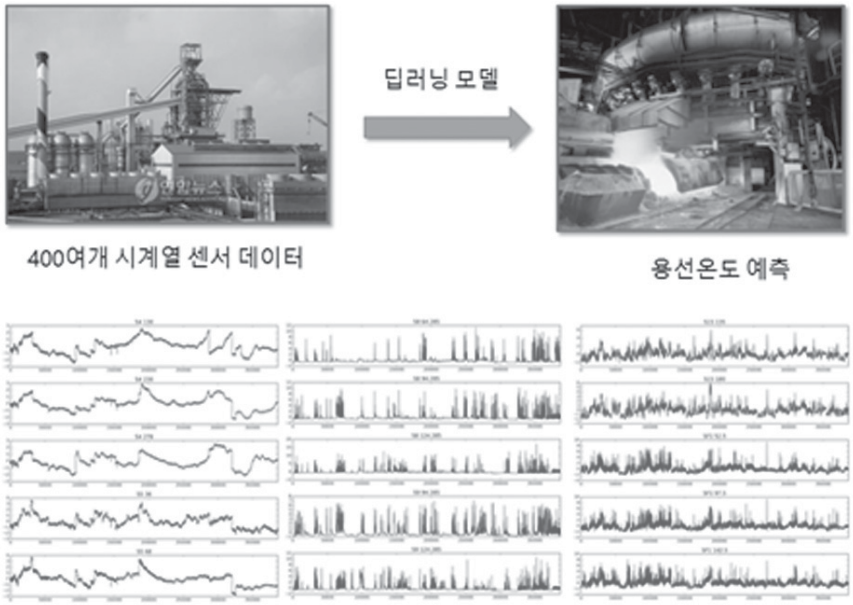


그림 2 철강공정을 위한 대규모 센서 데이터를 처리하는 딥러닝 모델

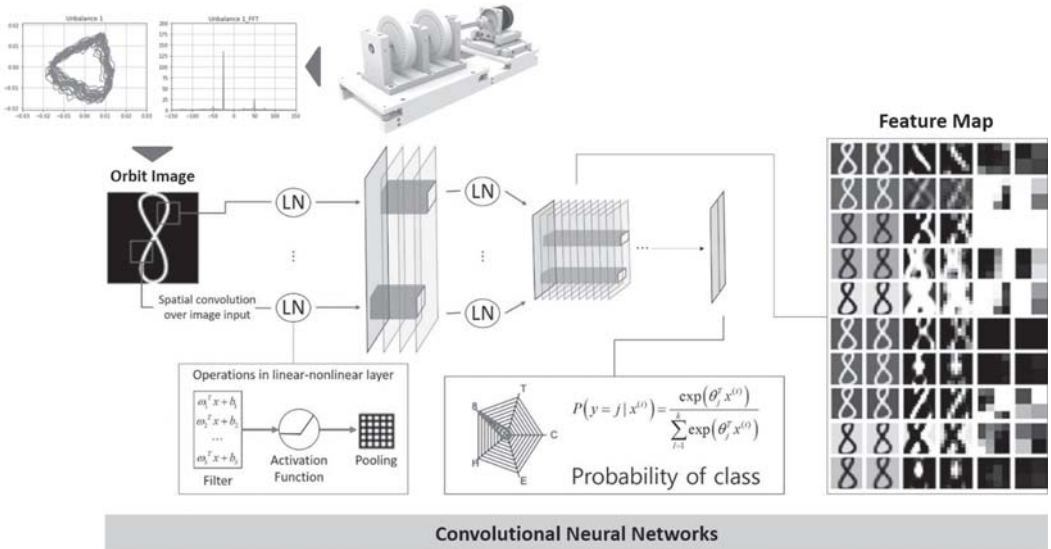


그림 3 인공지능 기반의 발전 설비 상태 진단 기술 개발 사례

생산 설비가 점점 더 복잡해지고, 작동 환경이나 주변 잡음이 복잡하게 연계되어 있기 때문에 강건한 진단 결과를 예측하지 못하고 있다.

빅데이터를 기반으로 숨겨진 패턴을 파악하는 인공지능을 활용한다면 설비 가동 데이터 (소음이나 진동, 온도 등)와 설비 상태 간의 관계성을 학습과정에서 스스로 추론하고 주변 환경으로부터 강건한 진단 결과를 도출할 수 있다. 그림 3은 회전 설비에서 취득한 진동 데이터를 딥러닝에 적용한 개발 사례이다[3]. 이 개발 사례에서는 회전 설비의 고장 유형별로 진동 데이터를 취득하고 이를 궤도 이미지 형태로 변환하여 합성곱 신경망에 적용함으로써 궤도 형상을 자동으로 구분할 수 있는 시스템을 개발하였다.

2.4 잔존수명예측을 위한 인공지능

인공지능은 앞선 예시와 같은 분류 문제뿐만 아니라 설비의 현재 상태를 반영하여 남은 수명을 예측하기 위해서도 사용할 수 있다. 소음, 진동 데이터와 같은 순차적인 데이터를 이용하여 분석하는 대표적인 딥러닝 구조로 순환 신경망(recurrent neural networks, RNN)을 예로 들 수 있다. 순환 신경망으로 예측된 수명은 사용자 경험에 의한 단순 역치(threshold) 기반이

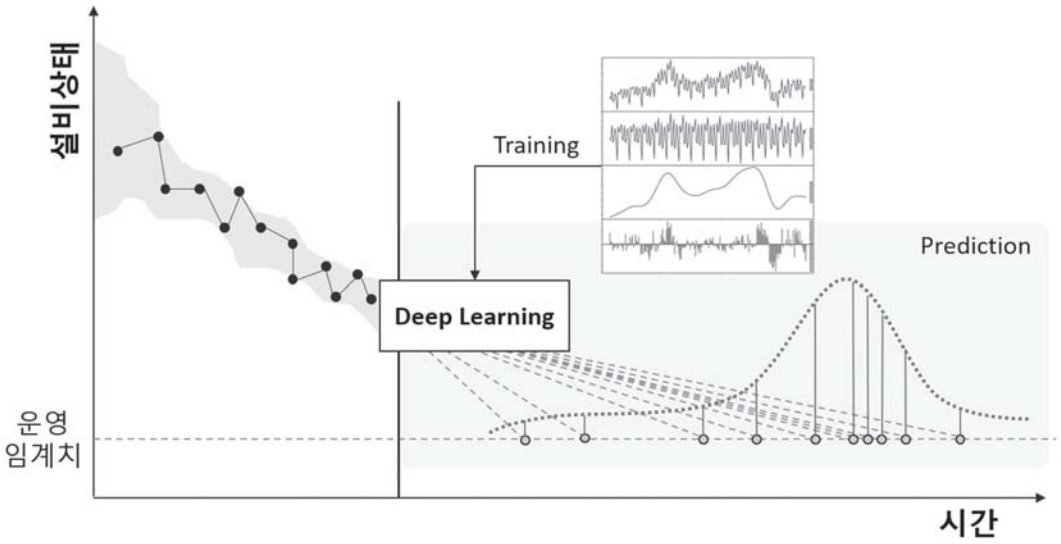


그림 4 인공지능 기반의 잔존수명예측 예시

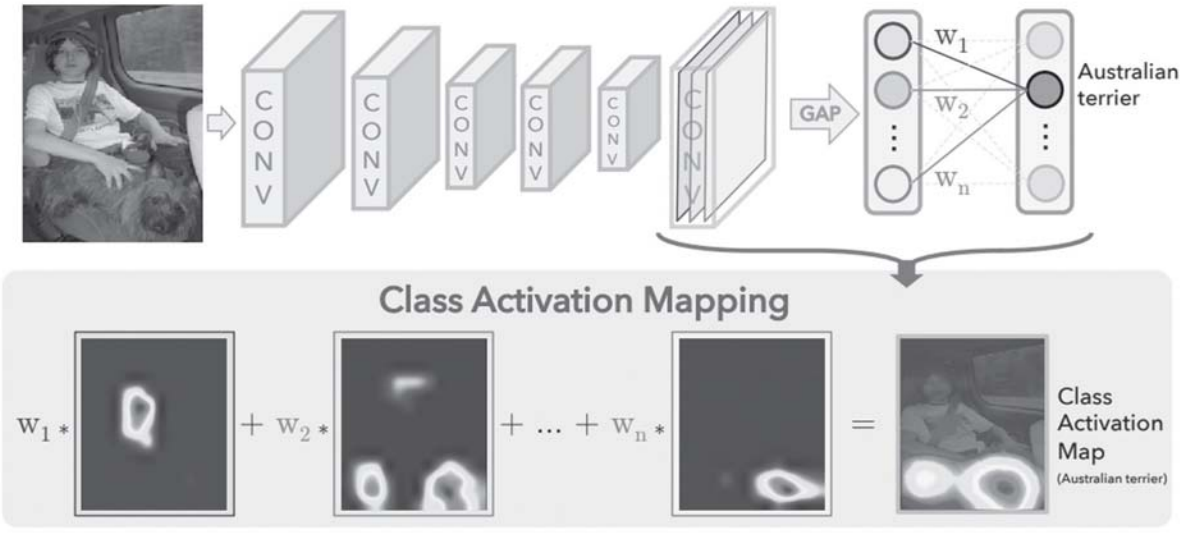


그림 5 Attention model

아닌, 시계열 데이터의 동적 특성이나 패턴을 반영한 형태이기 때문에 보다 정확한 설비 정비시기를 결정하는 데 적합하다. 이를 통해 설비 유지보수 비용을 획기적으로 줄일 수 있다.

3. 해석 가능한 인공지능

앞선 사례를 보면 데이터를 충분히 준비하여 인공지능을 학습하면 모든 문제를 풀 수 있을 것처럼 보인다. 하지만 인공지능에도 맹점은 존재한다. 인공지능을 학습하는 방법론은 대부분 데이터 기반으로 이루어지기 때문에 특정 문제를 잘 해결할 수 있지만, 왜 우수한 성능을 내는가에 대해서는 복잡한 구조 때문에 물리적으로 설명하기 어렵다. 이러한 이유 때문에 의사결정이 필요한 실제 산업에 적용 가능성이 떨어질 수 있다는 의견도 많다.

하지만 최근에는 주목모델(attention model)이라는 새로운 딥러닝 구조를 제시함으로써 낮은 설명력을 해결하면서 딥러닝의 높은 성능을 유지하려는 연구가 이루어지고 있다(그림 4)[4]. 이 구조는 학습 과정에서 딥러닝 모델이 특정한 결과를 예측하는 이유까지 학습에 포함하고 있다. 합성곱 신경망 말단에 레이어를 추가하여 이미지 분류를 위한 가중치를 학습함과 동시에 주목가중치를 학습하는 방식이다. 예를 들어 강아지 이미지를 인식하는 모델이 강아지 이미지의 어떤 부분을 보고 강아지라고 추론하는지 추적할 수 있다.

인공지능이 지금 바로 적용되어 산업의 모든 문제를 해결할 수는 없다. 하지만 주목모델과 같이 인공지능은 지금도 계속 발전하고 있다. 산업인공지능도 다량의 데이터를 바탕으로 연구가 진행된다면 많은 문제를 효과적으로 해결할 수 있는 인공지능으로 변화할 것이다.

4. 맺음말

누구나 쉽게 인공지능 서비스를 이용할 수 있는 시대가 되었다. 지금까지 인공지능의 유용함과 성능에 대하여 실제 사례와 연구결과를 통해 확인하였으며, 산업 데이터를 이용한 인공지능 활용 가능성에 대해서도 알아보았다. 산업인공지능의 활용은 그동안 정체되어 있던 제조 산업에 활력을 가져와 생산성 향상, 비용 절감 등 다양한 형태의 부가가치를 창출할 수 있을 것으로 본다. 소음과 진동, 온도 등 다양한 종류의 데이터가 쏟아지는 산업에서 인공지능은 산업 혁신을 위한 핵심 요소가 될 것이라는 확신으로 이 글을 마친다. KSNVE

참고문헌

(1) News, http://blog.posco.com/2736
(2) Statistical Artificial Intelligence Lab at UNIST, http://sail.unist.ac.kr
(3) Jeong, H., Park, S., Woo, S. and Lee, S., 2016, Rotating Machinery Diagnostics Using Deep Learning on Orbit Plot Images, Procedia Manufacturing, Vol. 5, pp. 1107~1118.
(4) Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. and Torralba, A., 2016, Learning Deep Features for Discriminative Localization, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921~2929.

소음진동과 산업인공지능

설명가능 인공지능

한지연, 최재식*
(울산과학기술원)

1. 머리말

최근 인공지능 분야는 딥러닝 등 기계학습 기술의 발전으로 다양한 응용분야에서 활용성을 높이고 있다. 다수의 학습 데이터를 바탕으로 모델을 학습해서 새로운 데이터에 적용하는 기계학습 기술은 이미지 분류, 자연어 처리, 시계열 데이터 분석 및 예측 등 인공지능의 다양한 분야에 큰 성능 향상을 가져왔다. 이제 인공지능은 자동주행, 의료수술 보조, 공장 자동화 등 우리 삶에 깊게 관여하고 있다. 인공지능 시스템이 제공해주는 서비스가 많아지면서, 과연 이런 인공지능 시스템의 의사결정 및 그 의사결정에 대한 추천을 우리가 얼마나 신뢰할 수 있는지 판별하는 것은 매우 중요한 문제이다. 이 글에서는 인공지능이 제시하는 결과와 그 과정에 대해서 사람이 이해할 수 있는 방식으로 설명하는 설명가능 인공지능의 개념을 소개하고, 현재 국내외 관련 연구 동향을 공유하고자 한다.

2. 설명가능 인공지능이란

인공지능이 삶에 깊숙이 관여함에 따라, 인공지능의 작동원리, 판단근거를 분석하는 연구에 대한 중요성이 대두되고 있다. 딥러닝 등 빅데이터를 기반으로한 기계학습 기술의 발달로 인해 각종 질환 진단 및 예측 기술이 발전하였지만, 어려운 문제일수록 모델의 복잡도가 증가하여 모델의 의사 결정에 대한 설명이 어렵기 때문에 주가 예측, 군사 작전 수행, 질병 진단과 같이 인공지능의 의사 결정에 대한 신뢰도가 보장되어야 하는 주요 분야에 대한 적용이 제한되어 왔다. 최근에는 사회적으로도 인공지능의 의사 결정 이유를 제시하도록 제도화 하고 있다. 유럽연합은 2018년부터 General Data Protection Regulation에 의거하여 알고리듬에 의해 자동으로 결정된 사안에 대해서는 회사에서 설명을 제공하도록 강제하고 있고, 미국에서는 회사가 내린 신용카드 발급, 주택 담보 대출 등의 주요 금융 결정에 대해서 이유를 제시하도록 법적으로 강제하고 있다. 이러한 필요성에 따라 설명가능 인공지능(explainable artifical intelligence)에 대한 관심이 높아지고 있고, 미국에서는 방위고등연구계획국(DARPA: Defense Advanced Research Projects Agency)의 주도하에 활발한 연구가 진행되고 있다.

설명가능 인공지능이란 주어진 데이터에 대해서 분류 · 예측할 뿐만 아니라 결정에 대한 인과관계를 분석하여 적절한 근거를 찾고 ICT, 심리, 언어 분야 등의 학제 간 융 · 복합적 기술 개발을 통해 인공지능 모델의 의사결

* E-mail : jaesik@unist.ac.kr

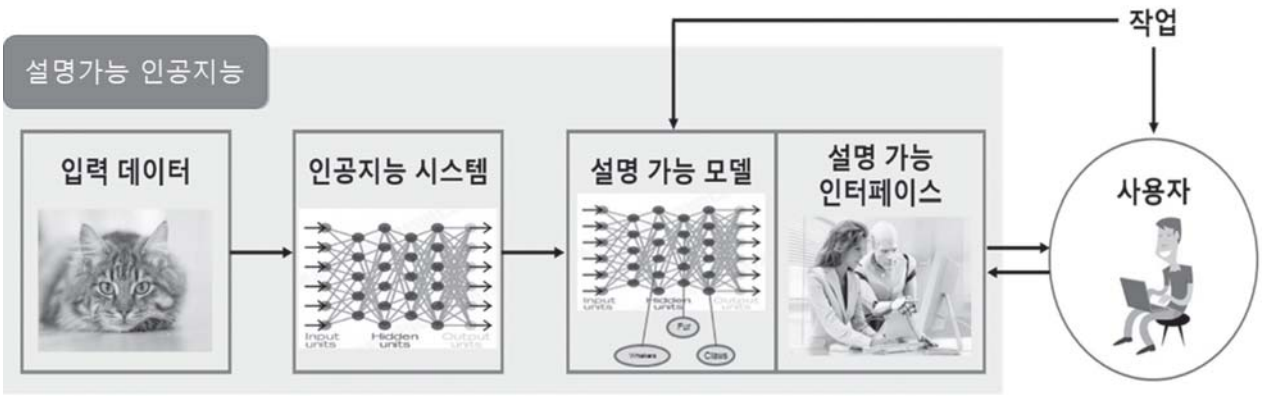


그림 1 DARPA에서 정의한 설명가능 인공지능

정 결과를 사용자 레벨에서 설명하는 인공지능 기술이다. 설명가능 인공지능을 통해 인간과 인공지능 상호간의 신뢰할 수 있는 의사결정을 가능케 함으로서 의료, 법률, 금융 등의 다양한 분야에 인공지능이 도움이 되고 사람들의 인공지능에 대한 불안감을 해소할 수 있는 계기가 될 것으로 기대한다. 다음으로 현재까지 국내외 설명가능 인공지능의 연구동향을 소개하겠다.

3. 사람처럼 배우고 생각하는 인공지능

현재의 인공지능을 설명하기 힘든 이유 중 하나는 사람과 인공지능의 의사결정 과정이 다르기 때문이다. 예를 들어, 사람과 인공지능이 개와 고양이 이미지를 구분하는 작업을 한다고 하자. 사람은 개와 고양이의 이미지 열 장 정도만 보더라도 쉽게 구분할 수 있는 반면 인공지능이 개와 고양이의 사진을 구분하기 위해서는 훨씬 더 많은 양의 이미지와 학습 시간이 필요하다. 이렇게 사람과 인공지능의 생각하는 방법이 다른 이유는 무엇일까? 사람은 어떻게 그렇게 빨리 학습하고, 추론하고, 새로운 분야에 적용할 수 있는 것일까?

첫 번째로, 사람은 발달 과정에서 두 가지 직관을 체득한다. 하나는 물리적 직관(intuitive physics)이고, 다른 하나는 심리적 직관(intuitive psychology)이다. 물리적 직관은 다양한 물리 현상들을 머릿속으로 시뮬레이션 할 수 있게 되는 것이다. 어린 아이들은 물리 법칙의 수식을 몰라도 손에 들고 있는 컵을 놓았을 때 컵이 낙하해서 깨질 것이라는 것을 알 수 있다. 또, 날아오는 공이 1초 뒤에 어디에 있을 지도 예측할 수 있다. 심리적 직관은 객체는 목표가 있으며 그 목표를 이루기 위해서 효율적으로 행동할 것이라고 생각하는 것이다. 더불어 객체의 행동을 보고 객체들 간의 관계가 대립적인지 우호적인지도 알 수 있다. 이는 학습에도 도움이 되는데, 사람은 게임을 배울 때 게임을 잘하는 사람의 플레이를 보고 어떻게 해야 목표를 달성 할 수 있을지, 어떤 걸 피해야 하고 어떤 걸 잡아야 하는지 직접 게임을 해보지 않아도 배울 수 있다.

두 번째로, 사람은 학습 모델을 빠르게 구축할 수 있다. 그 비법은 세 가지가 있는데, 하나는 문제를 더 기초적인 문제로 조합해서 학습하는 것이고, 또 하나는 인과관계를 통해 추론하는 것이고, 다른 하나는 이미 배운 것을 새로운 분야에 적용하는 것이다.

세 번째로, 사람은 빠르게 생각할 수 있다. 사람은 가설을 세우고, 이미 알고 있는 지식을 이용해서 또는 실험적으로 검증하여 정확하지 않더라도 빠르게 근사하여 추론할 수 있다.

이렇게 사람이 생각하는 방식을 인공지능 기술에도 적용할 수 있다면 더욱 복잡한 과업을 빠르게 학습할 수 있을 뿐 아니라 인공지능의 의사결정 과정을 이해하는 데에도 도움이 될 것이다.

4. DARPA XAI

DARPA의 설명가능 인공지능(XAI: explainable artificial intelligence) 연구는 크게 설명 가능한 모델 구축, 이유 설명 인터페이스로 이루어진다. 설명 가능한 모델 구축은 다시 심층 설명 학습, 해석 가능한 모델, 모델 귀납으로 나눌 수 있다.

심층 설명 학습은 변형된 딥러닝 기술로서 심층 신경망이 설명 가능한 특징들을 학습하도록 하는 기술이다. 다시 말하면 은닉계층의 노드가 의미 있는 속성을 나타내도록 학습한다[1]. 예를 들어 개와 고양이의 이미지를 구분하는 모델을 학습한다면, 각 은닉 노드가 귀 모양, 꼬리 모양, 입의 위치 등을 나타내도록 학습해서 모델이 어떤 이미지를 개라고 판단했을 때 활성화 된 은닉 노드를 통해 판단의 근거를 알 수 있다. 이러한 판단의 근거는 RNN 등의 자연어 생성 모델을 통해 언어적으로 나타낼 수도 있고, 이미지에 근거가 되는 부분을 표시해 시각적으로 나타낼 수도 있다.

해석 가능한 모델은 구조화된, 해석 가능한 인과관계 모델을 구축하는 연구이다. 한 가지 방법은 BPL(bayesian program learning)[2]을 이용한 것으로 모델을 작은 조각들의 조합으로 표현하도록 학습하는 방법이다. 예를 들어 글자를 생성하는 모델을 학습할 때 글자를 획으로 나누어서 가장 그럴 듯한 획의 조합으로 생성하도록 한다. 해석 가능한 모델을 구축하는 또 다른 방법은 And-Or-Graph[3]를 이용하는 것이다. And-Or-Graph는 어떤 객체를 특성들의 논리곱(and)과 논리합(or) 조합으로 나타낸 그래프로, 모델이 구조화 되어있어 모델의 결정과정을 논리적으로 설명하기 쉬운 장점이 있다.

모델 귀납은 임의의 블랙박스 모델을 설명 가능한 모델로 추론하는 기술이다. LIME (local interpretable model-agnostic explanations)[4]는 모델 귀납의 한 방법으로, 임의의 블랙박스 모델을 이미 설명이 가능한 데이터 주변에서 희소 선형 결합을 통해 국부적으로 설명 가능하게 만드는 방법이다. 예를 들어, 이미지를 분류하는 블랙박스 모델이 어떤 이미지를 개라고 판단했다면 이미 설명 가능한 다른 모델의 개에 대한 설명 즉, 개를 표현하는 픽셀들을 주어진 이미지와 대조하여 어느 부분이 개라고 판단한 근거인지 제시할 수 있다. 모델을 일련의 if-then 조건문으로 표현하는 BRL(bayesian rule lists)[5]도 모델 귀납의 다른 방법이다. BRL은 고차원, 다변수인 특징공간을 간단하고 이미 해석 가능한 조건문으로 나누어 복잡한 모델을 이해할 수 있게 한다.

이유 설명 인터페이스는 모델의 의사결정에 대한 설명을 사용자가 이해할 수 있는 방식으로 표현하는 연구이다. 이유 설명 인터페이스는 두 가지 원칙이 충족되어야 하는데, 설명가능성과 정정가능성이다. 설명가능성의 구체적 항목으로는 제시한 설명이 반복적일 것, 필요한 설명을 모두 포함하고 있을 것, 불필요한 설명을 포함하지 않을 것, 양이 적절할 것 등이 있다. 정정가능성은 설명이 유동적일 것, 사용자의 피드백을 존중할 것, 점진적인 변화를 주시할 것 등이 포함된다. 이렇게 제시된 설명에 대해서 사용자에게 설명의 명확도와 활용도 등에 대한 피드백을 받아 이유 설명 인터페이스의 효과를 평가하고 발전시킨다.

5. 인과관계 모델

인공지능의 의사결정 과정을 설명하기 위해서는 입력 데이터에 존재하는 여러 변수들 간의 복잡한 인과관계 및 연관관계를 찾아야 한다. 변수들 간의 연관관계를 학습하기 위해 딥러닝과 마르코브 랜덤 필드를 결합하는 모델이 제시되었다[6]. 먼저 학습 데이터로부터 심층 마르코브 랜덤 필드 모델의 확률 분포를 모델링하고, 확률 변수들 사이의 조건부 독립성을 나타내는 마르코브 랜덤 필드의 구조를 학습한다. 구조가 학습된 마르코브 랜덤 필드의 잠재

함수를 심층 신경망으로 추론하여 입력 변수의 수가 증가함에 따라 잠재 함수에 필요한 매개변수의 수가 기하급수적으로 증가하는 문제를 완화하고, 변수 연관관계에 대한 제약 없이 복잡한 연관관계를 학습할 수 있다.

인과관계를 학습하기 위해서는 희소 선형 결합을 다중작업 형태로 내장하여 사람에게 학습된 결과를 설명하고 피드백을 제안할 수 있는 모델이 제시되었다[7]. 구체적으로 클래스 분류 문제를 보조태스크인 속성, 슈퍼카테고리와 같이 학습한 후, 출력 단계에서 선형 결합하여 효과적인 표현이 가능한 모델이 개발 중이다. 예시로 이미지를 보고 어떤 동물인지 판별하는 모델을 학습한다면 이 동물이 털이 있는지, 점박이인지, 줄무늬가 있는지, 발이 네 개인지 등의 속성과 이 동물이 개과인지 고양이과인지와 같은 슈퍼카테고리를 함께 선형 결합할 수 있도록 학습하여 모델의 예측 결과에 대해 논리적으로 인과관계를 설명할 수 있다. 또한 인과관계가 정확히 학습되었는지 사람이 확인하고 피드백을 주어 수정할 수 있도록 하는 상호작용 학습 알고리듬이 개발 중에 있다.

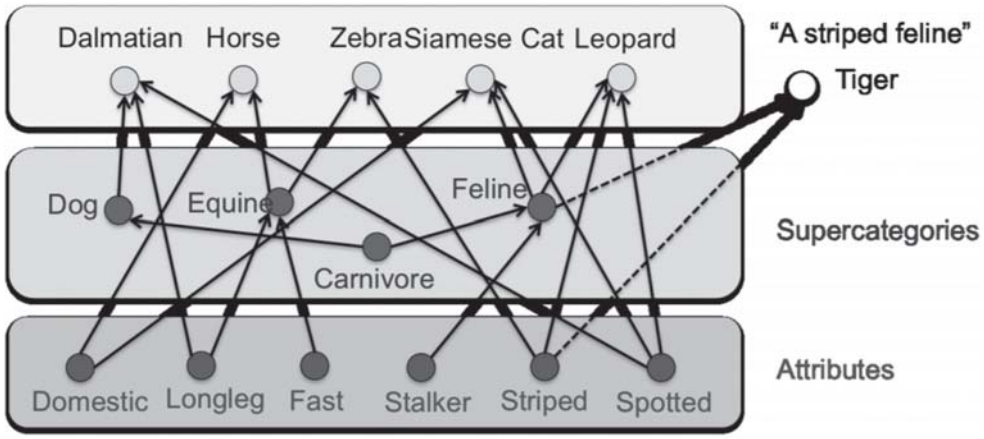


그림 2 의사결정을 설명하는 선형결합 모델

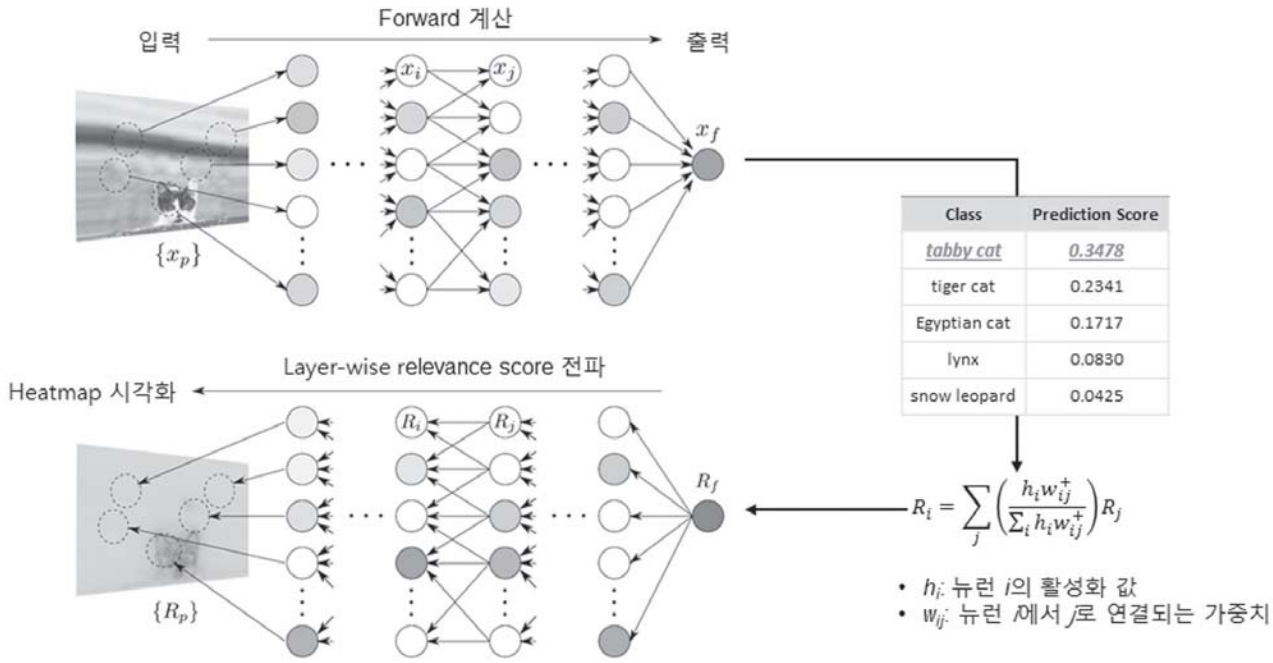


그림 3 LRP에 의한 입력 시각화의 예

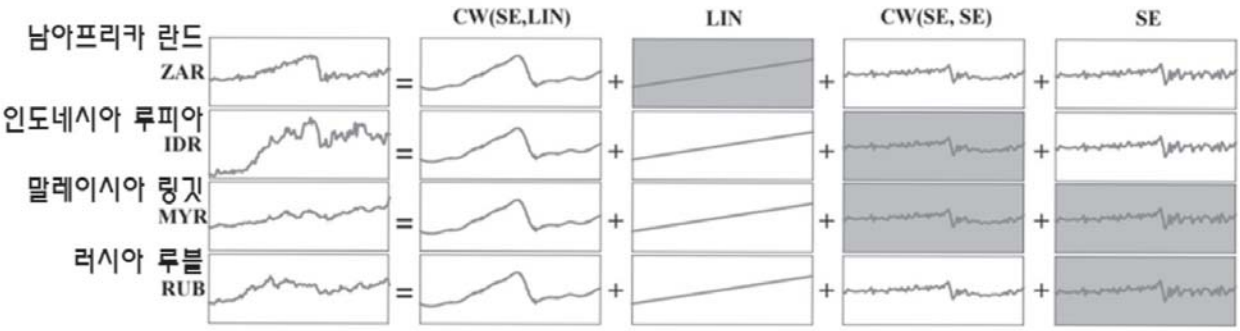
6. 계층적 상관성 전파를 통한 입력 시각화

국내에서는 신경망의 모델에서 예측 결과로부터 역전파 형태로 신경망의 각 계층 별 기여도를 측정할 수 있는 계층적 상관성 전파(LRP) 기술[8]이 개발되었다. 각 계층의 기여도를 히트맵 형태로 시각화하여 직관적으로 이해할 수 있다. 그림 3과 같이 고양이 이미지를 보고 고양이의 종을 분류하는 신경망을 학습했다고 하면, 학습된 가중치를 바탕으로 이전 계층의 각 노드에 기여도를 역전파 한다. 이 기여도를 열지도로 표현하면, 입력 이미지의 어떤 부분이 신경망의 결과 도출에 큰 영향을 미쳤는지 직관적으로 알 수 있다. LRP 기술은 특히 질병 진단 등에 유용하게 쓰일 수 있는데 MR과 같은 의료 영상을 보고 질병을 진단할 때, 어떤 부분이 그 질병으로 진단하는 데에 근거가 되었는지 의료 영상에 표시해 진단의사가 참고할 수 있도록 하여 진단의 효율을 높일 수 있다.

7. 자동통계학자 시스템을 통한 시계열 분석 보고서 작성

시계열 데이터 분석은 금융데이터 분석, 공장 자동화 등과 같은 주요 분야에 필수적인 중요한 문제이다. 그러나 인공지능이 시계열 데이터를 분석 및 예측하더라도 그럴듯한 근거를 제시하지 못한다면 맹목적으로 믿고 적용하기에는 위험요소가 크다. 시계열 데이터를 분석하고 설명하기 위해서 가우시안 프로세스 커널의 조합을 학습하는 베이지안 추론 알고리듬[9]이 개발되었다. 가우시안 프로세스란 부드러운 시계열 함수를 다양한 커널을 바탕으로 다변수 가우시안으로 회귀분석 하는 기술이다. 가우시안 프로세스에서 어떤 커널을 쓸 것인지는 매우 중요한 문제인데, 위 연구에서는 이 커널을 표현하는 최적의 커널 조합을 학습하여 주어진 시계열 데이터를 위에서 찾은 커널 조합을 바탕으로 설명할 수 있다.

더 나아가서 여러 개의 시계열 데이터가 있을 때에도 공통적으로 표현되는 커널 및 각 시계열 데이터의 특성을 표현하는 커널의 조합을 학습하여 여러 개의 시계열 데이터에서 공통적으로 나타나는 특징을 설명할 수 있는 알고리듬[10]을 개발하였다. 시계열 데이터 분석 모델을 통해 찾은 커널의 조합을 자연어로 작성하여 자동으로 보고서를 생성할 수 있다. 예를 들어 그림 4와 같이 다수의 환율 데이터를 분석하여 각 환율 데이터가 언제, 어떻게 움직이는지를 자동으로 자연어로 표현할 수 있다.



❖ CW(SE, LIN): 〔남아프리카 란드(ZAR)〕, 〔인도네시아 루피아(IDR)〕, 〔말레이시아 링킷(MYR)〕, 〔러시아 루블(RUB)〕은 2015년 9월 15일까지 부드러운 변화를 보임(단위길이: 6.4일).
❖ LIN: 〔인도네시아 루피아(IDR)〕, 〔말레이시아 링킷(MYR)〕 및 〔러시아 루블(RUB)〕은 전체 기간에서 선형적으로 증가하는 경향을 보임.
❖ CW(SE, SE): 〔남아프리카 란드(ZAR)〕와 〔러시아 루블(RUB)〕은 2015년 10월 23일부터 2015년 11월 9일을 제외하고, 급격한 변화를 보임(단위길이: 29.6시간).
❖ SE는 전 기간에 걸쳐, 급격한 변화를 보임(단위길이: 28.1시간).

그림 4 관계형자동통계학자 시스템에서 생성한 자동 보고서

8. 맺음말

앞서 살펴본 바와 같이 설명가능 인공지능 기술은 딥러닝 모델의 의사결정 과정을 입력의 입장에서 시각화하고, 의사결정에 과정에 관여한 딥러닝 모델의 구성요소를 희소선형결합법을 통하여 설명할 수 있다. 또한, 복잡한 모델의 상관관계를 분석하여 원인요소와 결과요소로 나누어 설명할 수 있다. 더불어 이는 사람이 이해할 수 있는 자동 보고서의 형식으로 작성되어 데이터를 분석한 결과뿐만 아니라 이유를 제공하여 인공지능 시스템이 보다 밀접하게 인간과 상호 작용하여 발전할 수 있게 해 줄 것으로 기대한다. KSNVE

참고문헌

(1) Liu, J., Cheng, H., Javed, O. et al., 2013, SRI-Sarnoff AURORA System at TRECVID 2013 Multimedia Event Detection and Recounting, TREC Video Retrieval Evaluation: Notebook Papers and Slides.

(2) Lake, B. M., Salakhutdinov, R. and Tenenbaum, J. B., 2015, Human-level Concept Learning through Probabilistic Program Induction, Science, Vol. 350, No. 6266, pp. 1332~1338.

(3) Si, Z. Z. and Zhu, S. C., 2013, Learning and-or Templates for Object Recognition and Detection, IEEE transactions on Pattern Analysis and Machine Intelligence, Vol. 35, No. 9, pp. 2189~2205.

(4) Ribeiro, M. T., Singh, S. and Guestrin, C., 2016, Why Should I Trust You?: Explaining the Predictions of Any Classifier, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM.

(5) Letham, B., Rudin, C., McCormick, T. H. and Madigan, D., 2015, Interpretable Classifiers using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model, The Annals of Applied Statistics, Vol. 9, No. 3, pp. 1350~1371.

(6) Suggala, A. S., Yang, E. H. and Ravikumar, P., 2017, Ordinal Graphical Models: A Tale of Two Approaches, International Conference on Machine Learning.

(7) Hwang, S. J. and Sigal, L., 2014, A Unified Semantic Embedding: Relating Taxonomies and Attributes, Advances in Neural Information Processing Systems.

(8) Bach, S., Binder, A., Montavon, G. et al., 2015, On Pixel-wise Explanations for Non-linear Classifier Decisions by Layer-wise Relevance Propagation, PloS one, Vol. 10, No. 7, e0130140.

(9) Lloyd, J. R., Duvenaud, D. and Grosse, R., 2014, Automatic Construction and Natural -language Description of Nonparametric Regression Models, AAAI.

(10) Hwang, Y. S., Tong, A. and Choi, J., 2015, The Automatic Statistician: A Relational Perspective,arXiv preprint arXiv:1511.08343.

생성적 적대 신경망과 소음진동공학

박승태, 이승철*
(울산과학기술원)

1. 머리말

최근 인공지능 분야에서 비지도 학습에 대한 논문들이 많은 비중을 차지하고 있다. 이전의 논문들이 이미지, 오디오, 텍스트를 분류하기 위한 인공지능 모델의 구조에 관심이 많았다면, 최근에는 이러한 데이터의 성질을 인공지능이 이해하고, 데이터를 스스로 생성하는 것에 많은 관심이 있다. 다시 말해, 분류모델 보다는 생성모델에 대한 연구가 더 많이 진행되고 있다. 이러한 흐름을 이해하기 위해서는 지도 학습의 한계점에 대해서 살펴볼 필요가 있다. 예를 들어, 동그라미와 네모를 구분하는 문제에 대하여 생각해보자. 분류모델은 수많은 데이터를 기반으로 학습되는데, 학습에 이용되는 도형은 동그라미인지 네모인지 함께 제시된다(지도학습). 따라서 지도학습을 통하여 학습된 분류모델은 학습과정에 사용된 데이터를 기반으로 분류하기 때문에, 데이터의 유사성을 따져본다고 할 수 있다. 하지만, 이상적인 분류모델은 동그라미와 네모가 가지는 기하학적 성질을 이해하고 습득된 지식을 기반으로 분류하는 모델일 것이다. 만약, 인공지능이 데이터의 유사성만을 따져서 분류를 하는 것이라면, 도형들이 가지는 기하학적 성질을 이해하지 못하였기 때문에 스스로 다양한 동그라미와 네모를 그려낼 수는 없을 것이다. 따라서, 데이터를 생성하는 것은 데이터를 구분하는 것보다 더 포괄적인 개념이라고 할 수 있다. 그렇기 때문에 인공지능을 학습시킬 때, 도형만 제시할 뿐, 제시된 도형이 동그라미인지 네모인지에 대한 정보를 제공하지 않은 채 인공지능이 도형을 스스로 생성할 수 있게 하는 학습법에 대하여 연구가 이루어지고 있다(비지도학습). 이는 '학습'이라는 개념에 대한 일반화가 이루어지고 있는 것이며, 이러한 흐름이 인공지능을 학습할 때 제시되는 데이터의 형태(비지도 학습은 도형만을 제시함)에도 드러나는 것이라고 해석할 수 있다.

이 글에서는 최근 인공지능 분야에서 제시된 비지도 학습방법 중에서 가장 각광받고 있는 generative adversarial network(GAN)에 대해 소개하려고 한다.

2. 오토인코더

우선, 기존에 주로 이용된 비지도학습 모델인 오토인코더(autoencoder)에 대하여 간략히 살펴보자. 앞서 살펴보았듯이 데이터를 생성하기 위해선 데이터의 성질을 이해한 후 생성할 필요가 있는데, 오토인코더는 이러한 아이디어를 구조적으로 구현한 것이다. 데이터의 성질을

* E-mail : seunglee@unist.ac.kr

$$h = f(x) \quad : \text{Encoder}$$
$$r = g(h) \quad : \text{Decoder}$$
$$g(f(x)) = x \quad : \text{Reconstruction}$$

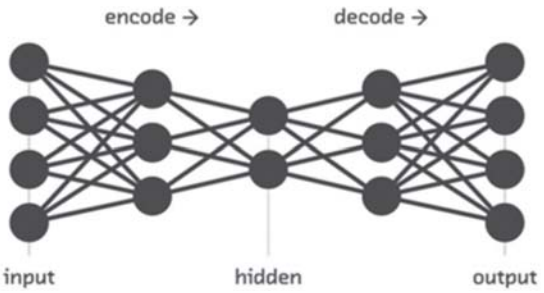


그림 1 오토인코더 구조

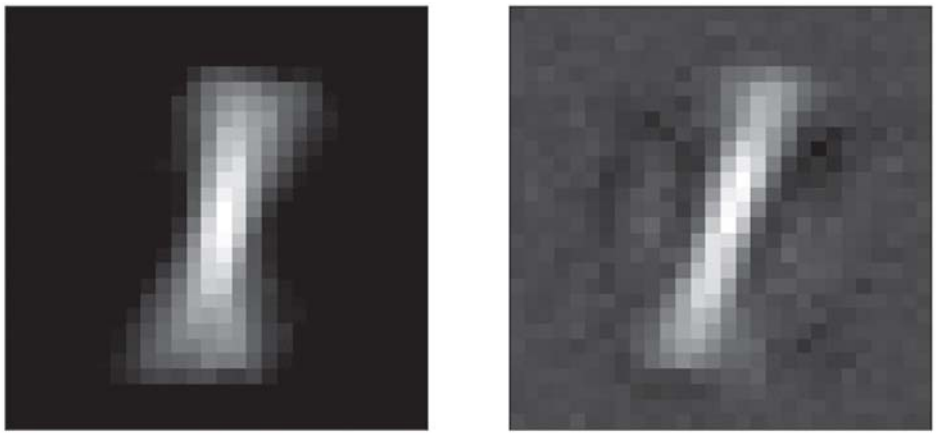


그림 2 평균 이미지(왼쪽)와 오토인코더를 이용해서 생성된 이미지(오른쪽)

이해하는 인코더, 그리고 그것을 바탕으로 데이터를 생성하는 디코더로 이루어져 있다. 개념적으로, 데이터의 성질은 복잡한 구조를 간단히 설명할 수 있는 것을 의미하기 때문에 인코더는 데이터의 차원을 축소시킨다(그림 1). 뿐만 아니라, 잠재변수의 개념을 활용하여 오토인코더를 확률 그래프 모델을 통하여 해석하는 연구도 진행되었다(variational autoencoder)[1].

하지만 오토인코더를 통하여 생성된 데이터 대부분 흐릿한 형태를 가지고 있다. 오토인코더가 학습할 때 전체 데이터를 대표하는 성질만을 학습하려는 성향을 가지고 있기 때문이다. 다시 말해, 데이터의 평균에 해당하는 성질만 학습하기 때문에 학습데이터를 평균한 것과 비슷한 데이터를 생성한다(그림 2). 따라서, 오토인코더를 통하여 생성된 데이터는 학습데이터에 대한 대표성은 가지고 있으나, 다양성이 존재 하지 않는다. 이러한 문제를 해결하기 위해서는 데이터의 성질을 좀 더 면밀히 살피는 학습방법론이 요구된다.

3. 생성적 적대 신경망

GAN(generative adversarial network)을 직역해보면 생성하는(generative)것과 적대적인(adversarial)것에 관한 네트워크이다. 데이터를 생성하기 때문에 generative라는 수식어가 붙어있다. 하지만 무엇에 대하여 적대적인가? GAN은 생성자(generator)와 구분자(discriminator)로 이루어져 있다. 생성자는 데이터를 생성하는 역할을 하며, 구분자는 생성해 낸 데이터를 실제 데이터와 구별해내는 역할을 한다. 따라서 생성자와 구분자가 경쟁하기 때문에, 생성자는 더욱 실제와 가까운 데이터를 생성하게 된다(그림 3). GAN이 가지고 있는 경쟁의 원리는 위조지폐범과 경찰에 비유되곤 한다. 경찰이 위조지폐를 감별하는 능력이 뛰어날수록 위조지폐범이 정밀한 위조지폐를 만들어 낸다는 것

이다. 반대로, 위조지폐범의 능력이 뛰어날수록 경찰의 능력 또한 증가할 것이다. GAN은 생성자(위조지폐범)와 구분자(경찰)라는 인공지능을 경쟁시킴으로써 서로의 능력을 증대 시킨다.

생성자(G)와 구분자(D)는 서로 상반되는 목적을 가지고 최적화 문제를 풀어낸다. 생성자는 구분자의 정확도를 최소화 시켜야 하지만, 구분자는 자신의 정확도를 최대화해야 한다. 따라서 아래와 같은 minimax 최적화 문제로 나타낼 수 있다.

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1-D(G(z)))] \quad (1)$$

여기서 구분자는 입력된 데이터에 대한 판단을 실제데이터의 진위여부에 대한 확률로 나타내기 때문에 0과 1 사이의 값을 가진다. 우변에서 첫 번째 항은 구분자가 실제데이터(x)를 입력 받았을 때, 올바르게 판단할 경우(D=1) 최대화되며, 두 번째 항은 가상데이터(z)를 입력받았을 때, 올바르게 구분할 경우(D=0) 최대화 된다. 반대로, 생성자의 경우 두 번째 항을 최소화해야 한다. 수학적으로 봤을 때 실제데이터와 가상데이터에 대한 엔트로피를 최소화하는 것과 같다. 위와 같은 최적화과정이 올바르게 이루어졌을 경우, 생성자는 데이터 이면에 존재하는 확률 분포를 완벽히 모사할 것이며, 따라서 구분자는 더 이상 가상데이터와 실제데이터를 구분할 수 없기 때문에 균일 분포(uniform distribution)로 수렴할 것이다. 그림 4에서 파란색은 구분자의 확률분포를 나타내며, 검은색은 데이터의 확률분포, 초록색은 생성자의 확률분포를 나타낸다.

GAN은 데이터를 생성하는 역할만 존재하는 오토인코더의 개념에 경쟁이라는 개념을 도입하여 생성자가 더 면밀히, 더욱 실제와 가까운

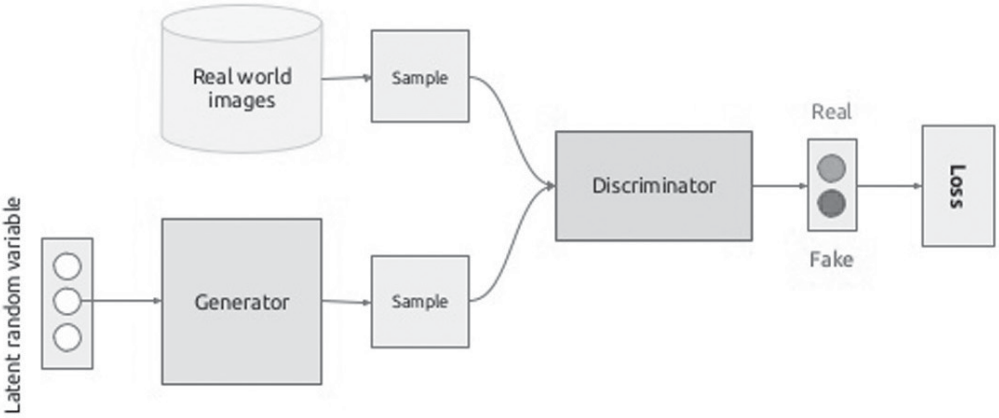


그림 3 생성적 적대 신경망(GAN) 구조

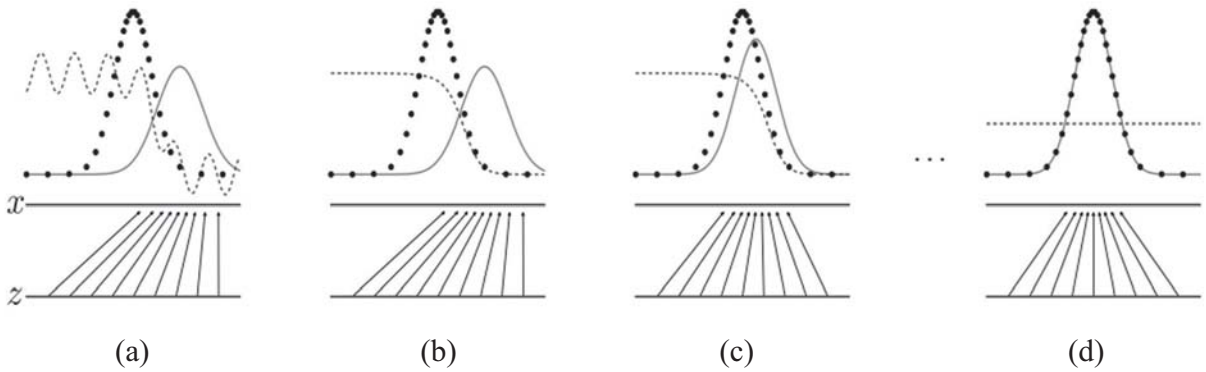


그림 4 GAN의 학습과정

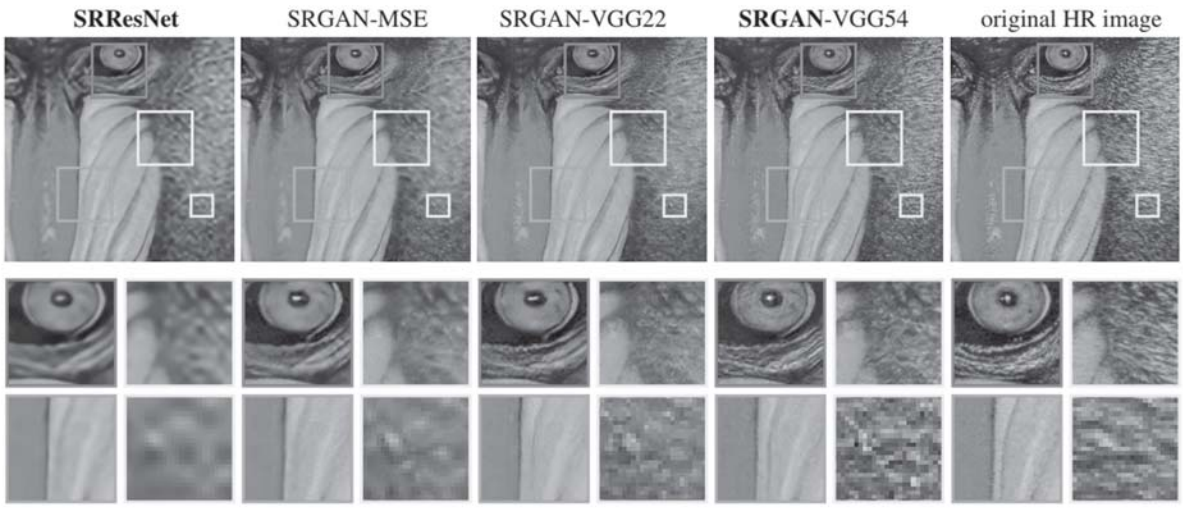


그림 5 Super-resolution 예시

데이터를 생성하게 유도하는 학습방법이다. 따라서 오토인코더가 가진 흐릿한 데이터에 관한 문제를 해결할 수 있다. GAN의 이러한 성질을 이용하여 흐릿한 이미지를 복원하는 기법이 연구되고 있다(그림 5)[4].

하지만 GAN 또한 단점이 존재한다. 첫째로, GAN은 학습하기가 어렵다. 학습초기 단계에 구분자는 엉성하게 생성된 데이터를 실제데이터와 구분하는 것이 쉬우나, 생성자는 실제와 같은 데이터를 학습초반부터 생성하기는 힘들다. 따라서 모델이 학습되지 않는 현상이 발생한다[5]. 이러한 문제를 해결하기 위해서 다양한 학습법들이 연구되고 있다[6,7]. 또한, GAN은 차원축소의 개념이 존재하지 않는다. 오토인코더가 주어진 데이터의 잠재변수를 추론할 수 있는 것과 대조적이다. GAN은 양질의 데이터를 생성할 수 있으나 잠재변수를 파악할 수 없고 오토인코더는 양질의 데이터를 생성할 수 없으나 잠재변수를 파악할 수 있다. 이는 복잡한 데이터를 간단히 표현하고 파악하고자 하는 문제에 적절하지 않다. 따라서, 최근 오토인코더와 GAN의 성질을 동시에 사용가능한 모델구조가 연구되고 있다[8,9].

4. 소음진동에서의 가능성

GAN은 인공지능만을 위한 기술은 아니다. GAN을 소리나 진동데이터를 기반으로 학습할 경우 여러 가지 효과를 기대할 수 있다. 영상처리 분야에서 노이즈가 추가된 이미지를 복원하는 것(super resolution)과 같이 소음진동 분야에서도 소음이 들어간 소리를 복원하는 문제는 중요한 화두이다. 이와 관련하여 사람의 목소리를 소음이 존재하는 환경에서 GAN을 통하여 복원하는 기술(speech enhancement)이 소개되었다[10]. 또한 스스로 소리를 생성하는 모델 또한 소개되었다. 여러 클래식 음악을 학습한 후, 스스로 작곡을 하는 GAN이 개발되었다[11]. 양질의 데이터 생성은 고장진단분야에서 중요한 의미를 가진다. 기계시스템의 고장은 그 확률이 매우 낮으나 고장이 발생할 경우 막대한 비용을 초래한다. 따라서, 고장데이터의 생성은 모니터링 시스템 구축에 큰 기여를 할 수 있다. 예를 들어, 정상데이터와 고장데이터 숫자의 불균형을 극복하기 위한 해결책으로서 GAN을 활용할 수 있다.

GAN과 더불어 인공지능 분야에서 개발된 비지도학습법은 소음진동분야에 여러 가지 가능성을 제시한다. 앞서 언급한 오토인코더 같은 경우, 기존에 잠재변수를 파악하기 위해 주로 이용되었던 PCA/SVD 알고리듬을 뛰어넘었다. 이전에는 데이터와 선형적인 관계를 가진 잠재변수만 파악이 가능했던 반면에, 인공지능을 통하여 비선형 관계를 가진 잠재변수

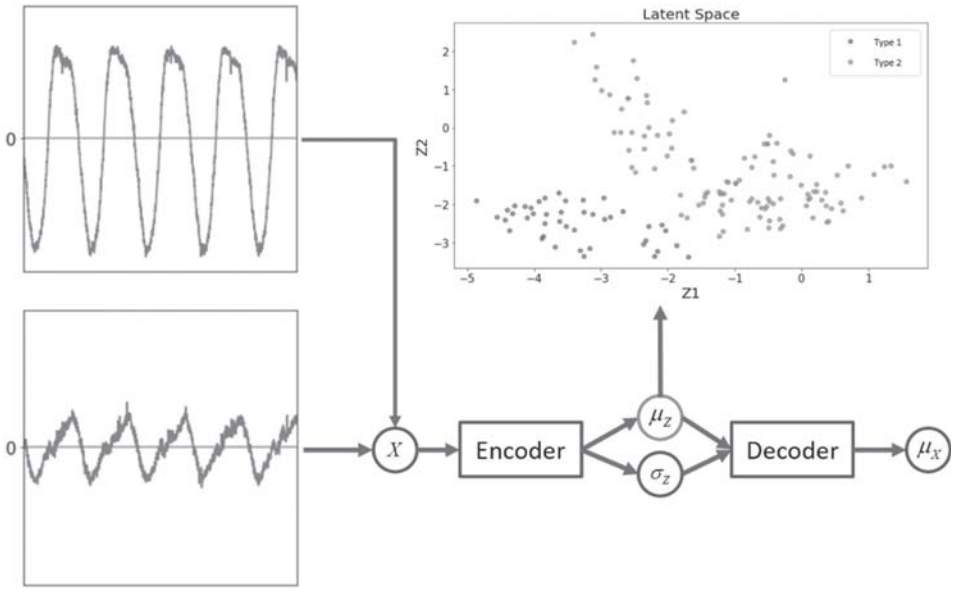


그림 6 신호 차원축소

또한 추론이 가능해진 것이다. 이는 취득 주파수에 따라 높은 차원을 가진 소리 및 진동데이터를 효과적으로 차원축소 할 수 있음을 의미한다. 따라서 다양한 소리와 진동 신호들의 위상적 관계를 밝혀내고, 분류나 군집화와 같은 분석을 효과적으로 수행할 수 있다(그림 6). 효과적인 차원축소 기술은 고장진단분야의 핵심이다. 가스터빈이나 반도체 생산시설과 같은 기계시스템에서 우리는 적어도 수백 가지의 데이터를 취득하는데, 이렇게 취득된 데이터들을 효과적으로 합성하여 핵심이 되는 정보만 제시하는 것은 시스템 상태파악의 기초가 되기 때문이다.

5. 맺음말

이처럼 GAN이나 오토인코더와 같이 인공지능 분야에서 개발된 비지도학습법은 기계시스템 이면에 존재하는 성질을 쉽게 이해 가능하게 하며, 이러한 장점에 기반하여 인공지능이 능동적으로 기계시스템을 생성 및 분석 가능하게 한다. 인공지능은 결코 소음진동공학과 동떨어져 있는 학문이 아니다. 최근에 다양한 인공지능 알고리듬이 개발이 되었다. 하지만 이는 알고리듬일 뿐, 우리 눈앞에 움직이고 소통하는 것이 아니다. '지능'에 대한 기술이 발전되었을 뿐이다. 인공지능을 세상밖으로 꺼내어 줄 다리와 팔을 함께 만들어가는 것이 현시대의 기계공학자에게 필요한 것이 아닐까 생각한다. 이를 위해서 인공지능에 대한 깊은 이해가 이젠 기계공학자에게도 필요 할 것이다. 이 글이 기계공학자들에게 지능적 기계시스템에 대한 연구의 단초를 마련하길 기대한다.

참고문헌

(1) Kingma, D. P. and Welling, M., 2013, Auto-encoding Variational Bayes, arXiv preprint arXiv:1312.6114.
(2) Hou, X., Shen, L., Sun, K. and Qiu, G., 2017, Deep Feature Consistent Variational Autoencoder, In Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on, pp. 1133~1141, IEEE.
(3) Mao, X. J., Shen, C. and Yang, Y. B., 2016, Image Restoration using Convolutional Auto-encoders with Symmetric Skip Connections, arXiv preprint arXiv:1606.08921.
(4) Ledig, C., Theis, L., Huszár, F. et al., 2016,

Photo-realistic Single Image Super-resolution using a Generative Adversarial Network, arXiv preprint arXiv:1609.04802.
(5) Goodfellow, I., Pouget-Abadie, J., Mirza, M. et al., 2014, Generative Adversarial Nets, In Advances in Neural Information Processing Systems, pp. 2672~2680.
(6) Salimans, T., Goodfellow, I., Zaremba, W. et al., 2016, Improved Techniques for Training GANs, In Advances in Neural Information Processing Systems, pp. 2234~2242.
(7) Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. and Courville, A., 2017, Improved Training of Wasserstein GANs, arXiv preprint arXiv:1704.00028.
(8) Makhzani, A., Shlens, J., Jaitly, N. et al., 2015, Adversarial Autoencoders, arXiv preprint arXiv:1511.05644.
(9) Mescheder, L., Nowozin, S. and Geiger, A., 2017, Adversarial Variational Bayes: Unifying Variational Autoencoders andGgenerative Adversarial Networks, arXiv preprint arXiv:1701.04722.
(10) Pascual, S., Bonafonte, A. and Serrà, J., 2017, SEGAN: Speech Enhancement Generative Adversarial Network, arXiv preprint arXiv:1703.09452.
(11) Yang, L. C., Chou, S. Y. and Yang, Y. H., 2017, MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation, In Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR' 2017), Suzhou, China.

소음진동과 산업인공지능

시각화 기반 딥러닝 분석 기술

이재성, 김기환, 이충기, 고성안*
(울산과학기술원)

1. 머리말

오늘날 딥러닝 기술은 다양한 종류의 패턴 인식 문제에서 인간에 필적하는, 또는 때때로 능가하는 모습을 보여주고 있다. 기존 기계학습 모형과 달리 딥러닝은 데이터에서 복잡한 특성을 반복된 학습을 통하여 자동적으로 추출하고 이를 모델의 의사결정 수단으로 사용한다. 딥러닝은 자율주행, 음성인식, 기계번역 등 여러 분야에서 이미 성능을 입증하였고 산업계에서도 점차 이용범위를 확대하고 있다. 그러나 딥러닝에도 단점이 존재하는데, 그것은 딥러닝 기술 내부 각 계층의 의사결정 과정을 인간이 이해하기 어렵다는 것이다. 이를 해결하기 위하여, 딥러닝 프로세싱 과정을 해석하기 위한 연구가 여러 학계와 산업계에서 다방면으로 진행되고 있다. 이 글에서는 지난 10월에 열린 IEEE VIS 시각화 기술 학회에서 발표된 딥러닝 시각화 및 해석 기술들 중, CNN, RNN을 위한 시각화 기술과, FACEBOOK이 자체 개발한 딥러닝 해석용 시각화 기술을 소개하고자 한다.

2. CNN에 관한 시각적 해석

CNN(convolutional neural network)은 합성곱 연산을 이용하여 2차원 혹은 3차원 공간의 이미지의 특징을 뽑아내어 분류하는데 사용되는 딥러닝 모델의 한 종류로, 이미지의 특징을 뽑는 방식 및 학습 과정의 해석이 매우 어렵다. 이는 곧 모델 간 비교, 수정, 개선에 어려움이 있음을 의미한다. 이를 극복하기 위한 방법으로 Pezzotti와 그의 동료들은 중간 은닉 계층들의 상태를 시각화하는 DeepEyes 시각화 시스템을 제안하였다[1]. DeepEyes는 은닉계층들과 필터, 네트워크로부터 추출된 학습 데이터의 패턴들과 학습에 필요한 또는 불필요한 계층들을 시각화를 통하여 나타내어준다. 따라서 이 시스템을 활용하면, 딥러닝의 학습이 어떻게 진행되고 있는지, 어떠한 특징들이 모델로부터 추출되었는지, 그리고 모델에 필요한 혹은 불필요한 계층이 어느 곳인지가 명확하게 드러나, 모델을 이해하고 개선하는데 도움이 된다.

그림 1은 DeepEyes의 구성을 나타내는 그림이다. 여기서 (a)는 모형 평가에 사용되는 loss와 accuracy 그래프, (b)는 각 층의 활성화 값들을 확률로 변환한 히스토그램, (c)는 필요 없는 필터 찾기 위한 활성화 빈도 히트맵, (d)는 입력 이미지에 대한 필터의 활성화 값 시각화, (e)는 어떤 한 층에서의 필터들 간의 관계를 시각화 한 것이다. 이와 같은 정보들을 통하여 학습 과정에서의 각 계층들의 세부 정보를 알

* E-mail : sako@unist.ac.kr

수 있고 이를 모델 평가와 개선을 위하여 사용할 수 있다. 효용성을 보여주기 위하여, 해당 툴을 사용하여 fine tuning을 수행 내용을 간략히 소개한다.

Fine tuning은 딥러닝 모델의 성능을 개선시키기 위한 하나의 방법으로써 먼저 학습된 모델의 특정한 계층을 추가적으로 학습시키는 방법이다. Fine tuning은 어렵고 시행착오가 많이 필요한 작업인데 시각화를 이용하면 어느 부분에서의 tuning이 필요한지 직관적으로 알 수 있다. 그림 2는 DeeyEyes를 이용하여 AlexNet이라는 1000여 개의 사물 구분 필터를 가진 모

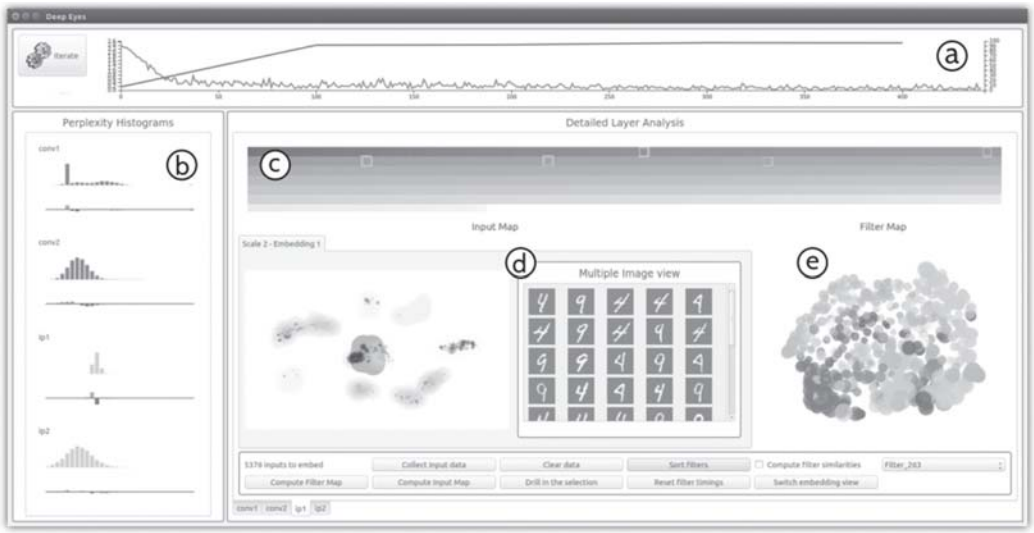


그림 1 DeepEyes 시각화 모델

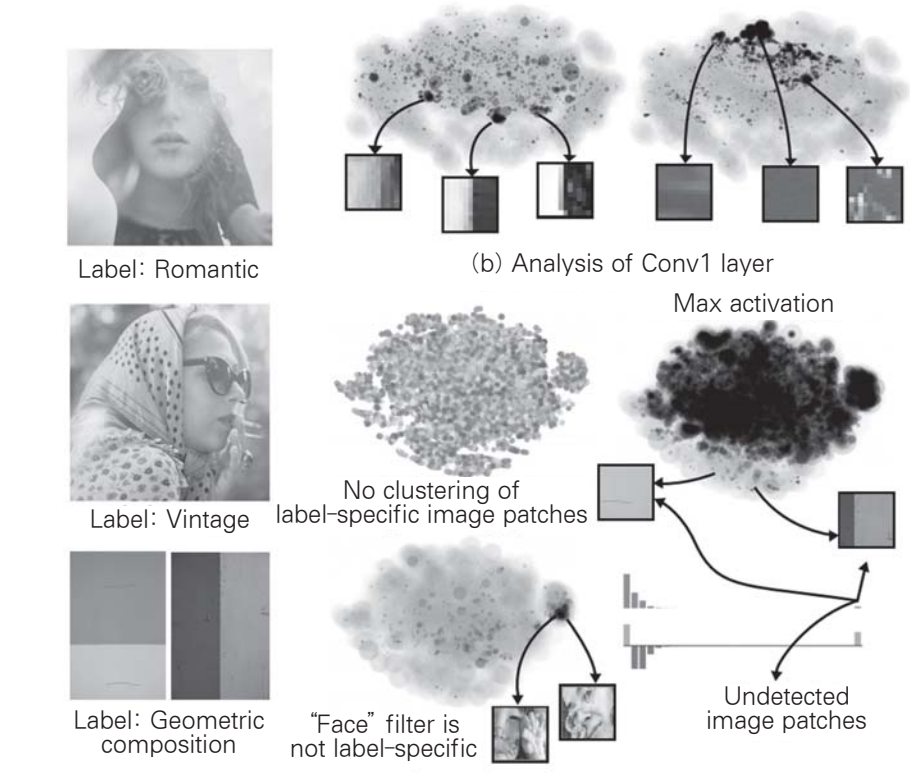


그림 2 DeepEyes를 이용하여 중요 필터 확인 가능한 fine tuning 과정

형을 fine tuning을 통하여, 그림 2(a)의 romantic, vintage, geometric composition 등의 20개의 스타일로 구분 할 수 있는 필터를 지닌 모형으로 바꾸는 과정을 나타낸다. (b)에서 설명하는 바는, 첫 번째 층인 conv1은 주로 색과 모서리 특성을 잘 학습하고 있다는 것을 알 수 있는데, 이는 인터랙션을 통한 시각화에서, 필터가 파란색 필터끼리, 세로축 필터끼리 군집한 것을 알 수 있기 때문이다. (c)의 시각화가 함축하는 바는 여러 가지 인데, 이 모델은 레이블 별로 군집화가 되지 않는다는 것과 geometric composition 이미지들은 활성화를 거의 일으키지 않는다는 사실이다. 그리고 전혀 다른 스타일을 가진 두 가지 사람 얼굴 사진이 filter map에서 서로 가까이에 위치한 것이 시각화 되는데, 이는 원래 AlexNet이 사물을 분류하는 모형이란 점에서 야기된 것으로 판단된다. 위에서의 같이 시각화 시스템을 활용하여, conv5층은 모형의 성능에 전혀 기여하지 못하기 때문에, 삭제하는 것이 학습 속도 향상을 위해서라도 좋고, geometric composition 데이터가 더욱 많이 필요하다는 것과 인물 사진 데이터는 이 모델에서 오류가 발생하기 쉽다는 것 또한 알 수 있다.

3. RNN에 관한 시각적 해석

RNN(recurrent neural network)은 자연어 처리와 시계열 데이터 처리에 강점을 보이는 있는 딥러닝 모델이며, 이 모델 역시 내부의 은닉계층의 데이터 처리, 학습 과정 및 결과 예측 과정을 판단하기에 어려움이 있다. RNN 해석을 위하여, Ming과 그의 동료는 여러 RNN 모델의 은닉계층의 정보를 시각화하여 모델의 상태를 진단하고 구조 설계에 대한 이해를 돕기 위하여, RNNVis라는 새로운 시각화 툴을 제안하고[2], 해당 툴의 유용성을 RNN이 자연어를 처리 어떻게 하는지, 툴이 어떻게 이해도를 높여줄 수 있는지 보여줌으로써 증명한다. 예를 들어, RNN에서 단어 단위로 처리되는 은닉 계층의 메카니즘, 각 단어가 입력될 때, 전체 단어 분포 및 단어와 문장과의 관계에 기반을 둔 모델 학습 과정을 비교 설명한다.

RNNVis 크게 세 부분으로 구성되어 있다(그림 3). 왼쪽은 모델 매니저로 RNN 모델과 데이터셋을 관리한다. 가운데 evaluator 모듈은 모델 매니저로부터 받은 데이터를 가지고 처리하는 과정에서 RNN 은닉계층의 각 파라미터를 시간에 따라서 해석하고 클러스터링 한다. 오른쪽 visualization 모듈은 평가 부분에서 처리된 데이터를 받아 시각화 하는 부분이다. 이 시각화 모델은 은닉계층으로부터 받은 정보를 시각화를 통하여 시간단위로 해석할 수 있다.

그림 4는 RNN 모델중 하나인 LSTM의 마지막 계층을 클러스터링을 이용하여 시각화한 모습이다. A는 각 품사를 색깔로 대응시켜서 각각의 단어와 계층을 이어 합동 군집화(co-clustering) 시각화이다. B는 특정 단어에 대한 모델 응답 횟수를 나타낸 분포도이고(ex. when, where 응답 분포), C는 48번째와 123번째의 계층의 상태를 단어가 나오는 빈도를 이용하여 해석한 그래프이다. 위의 시각화 툴을 활용하여, 사용자는 어느 계층에서 어느 단어가 학습되고 출력되는지에 대한 경향성을 알 수 있고, 같은 품사의 비슷한 단어가 어떤 빈도로 사용되는지 확인 할 수 있다. 이는 곧 특정 모델이 데이터에 적합한지 확인하는데 기여를 하며, 모델의 개선 및 구현에 큰 도움이 된다.

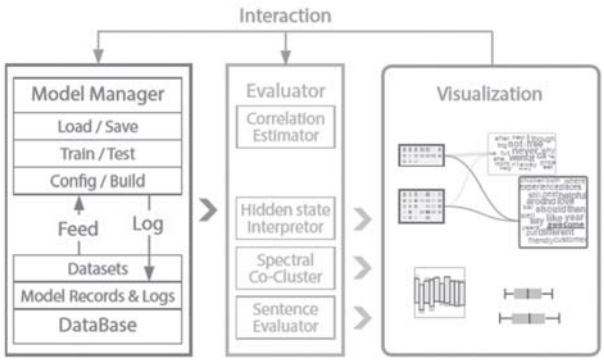


그림 3 RNNVis의 model manager, evaluator, visualization

4. 산업계에서의 딥러닝 모델의 시각화

딥러닝은 다양한 산업 분야의 문제를 해결하는 돌파구로 각광을 받고 있으며, 컴퓨터 비전, 자연어 처리, 음성 인식, 자율 주행 등 많은 분야의 문제를 해결하기 위해 구글, 바이두, 인텔, 애플, 페이스북 등의 회사에서는 딥러닝에 대한 연구가 활발히 진행되고 있다. 이전 섹션과 같은 맥락으로, 산업계에서도 활용에 더하여, 딥러닝 모델이 어떻게 학습되고 결론이 도출되는지에 대해 이해하기 위해 노력하고 있으며, 그 가운데에 시각화 툴이 큰 역할을 하고 있다(ex. 구글 텐서플로[3]). 여기서는 페이스북에서 개발한 시각화 시스템인 ActiVis를 소개한다[4].

페이스북은 그동안 기존 프레임워크 활용한 코드기반 환경(command line interface)에서 딥러닝 작업을 수행했으나, 곧 이러한 방식은 직

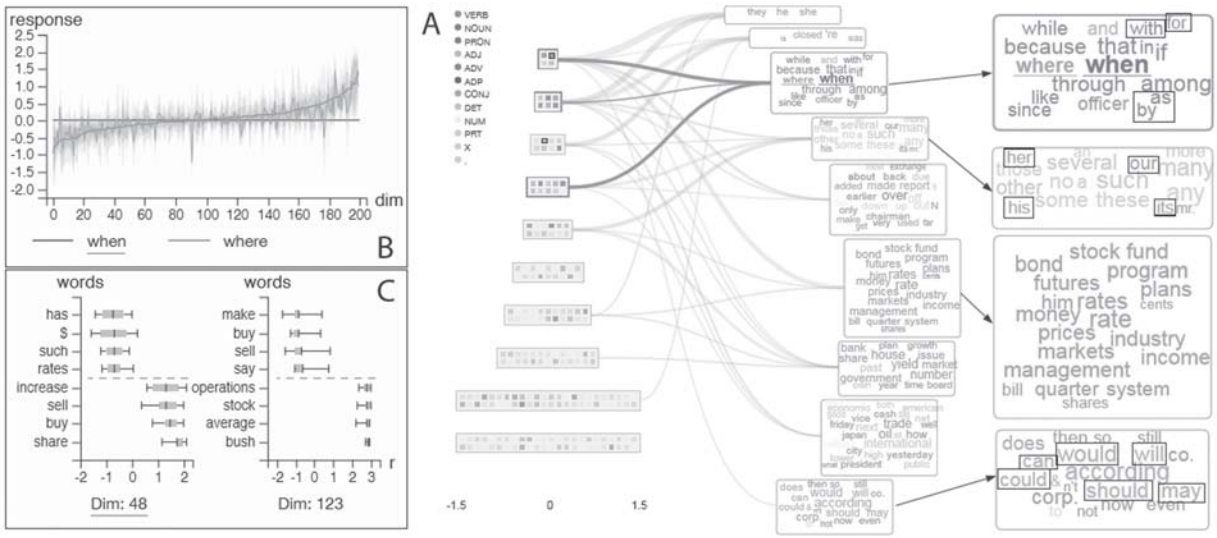


그림 4 LSTM-small의 마지막 계층 시각화 예

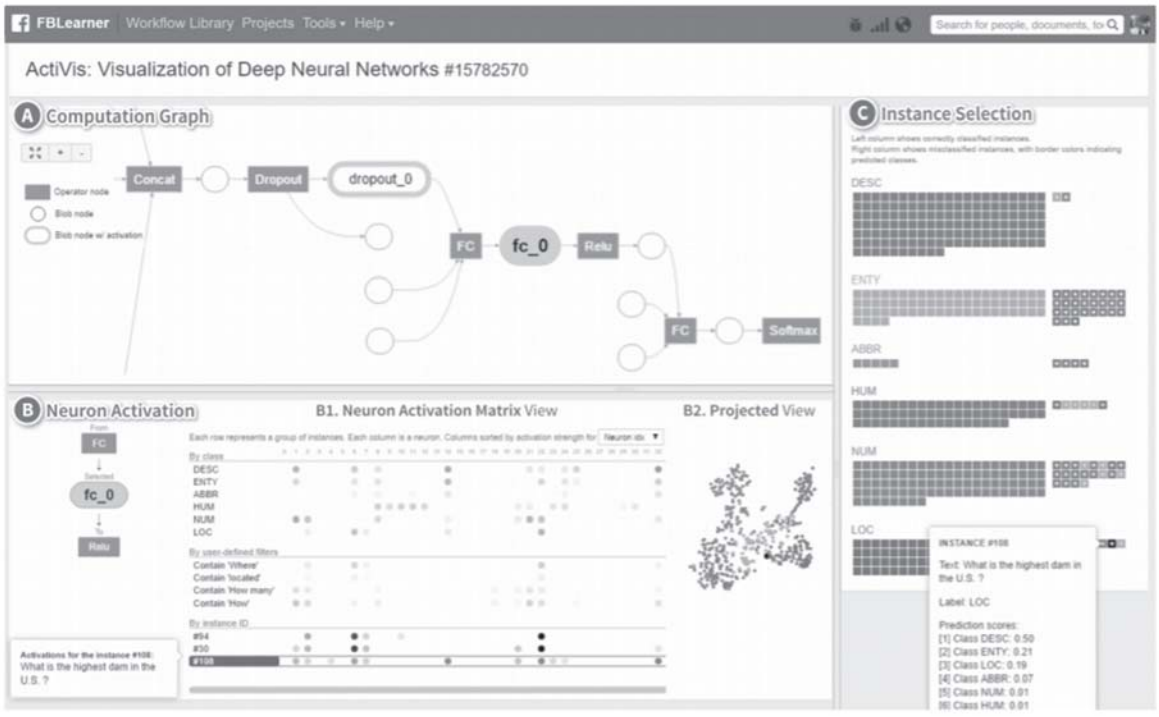


그림 5 ActiVis 시각화 모델

관적으로 모델을 만들거나 이해하는데 어려움이 있다는 것을 인식했다. 따라서 여러 연구자들과 개발자, 그리고 데이터 분석가들이 모델을 개발하고 해석하기 위한 업무 프로세스를 개선하고자 하였으며, 그 결과로 ActiVis라는 딥러닝 시각화 툴을 개발하였다[4]. ActiVis는 개발 환경을 코드 기반인 CLI에서 GUI (graphic user interface)로 변화시켜, 딥러닝 전문가가 아닌 개발자들이나 분석가들도 모델을 유연하고 편리하게 만들거나 수정할 수 있게 하였고, 시각화를 통하여 학습 프로세스를 보여줌으로써 전체 모델이 어떻게 작동하는지 이해하는 데 큰 도움을 주도록 하였다. 뿐만 아니라 은닉계층들의 정보를 시각화를 이용하여 계층화하여 표현함으로써 인스턴스 기반과 하위 집합 기반의 분석이 가능하게 하였다.

그림 5는 ActiVis 시각화 시스템으로, 여기서 A는 모델 전체의 구조를 보여주는 그래프 시각화로, 각각의 구조를 탐색하기 위해서 구현되었다. B는 뉴런 활성화 패널 행렬 뷰로 활성화된 인스턴스로 하위 세트 및 클래스의 활성화 정도를 나타낸다. C는 인스턴스 선택 패널로 인스턴스와 그 분류 결과를 표시한다. 왼쪽에는 올바른 그리고 오른쪽에는 잘못 분류된 인스턴스를 시각화 한다.

5. 맺음말

딥러닝의 뛰어난 성능은 여러 분야에서 기존 문제를 해결할 돌파구로 인식되지만, 내부 프로세싱이 black-box라는 근본적인 문제로 인하여, 딥러닝이 도출한 결론을 사용자가 해석하고 믿기 어렵다는 문제가 존재한다. 이 문제를 시각화 기술을 통해 해결하고자 하는 노력이 인공지능 및 시각화 기술 학계에서 시작되고 있으며, 이 글에서 최근 발표된 시각화 시스템 및 기술에 대하여 살펴보았다. 아직 초기 단계이지만 시각화 기술을 활용하면 딥러닝 해석에 도움이 된다는 것이 여러 사례 분석을 통하여 나타나고 있으며, 점차 발전하는 시각화 기술은 딥러닝 결과 및 해석을 신속, 정확하게 도와줄 기술로 기계 분야에 딥러닝 적용 및 해석을 위하여 함께 고려해야 할 기술이다. KSNVE

참고문헌

(1) Pezzotti, N., Höllt, T., Gemert, J. V. et al., 2017, DeepEyes: Progressive Visual Analytics for Designing Deep Neural Networks, IEEE Transactions on Visualization and Computer Graphics (IEEE VIS 17), Vol. 24, No. 1, pp. 98~108.

(2) Ming, Y., Cao, S., Zhang, R. et al., 2017, Understanding Hidden Memories of Recurrent Neural Networks, In Proceedings of IEEE Visual Analytics Science and Technology (IEEE VIS 17).

(3) Wongsuphasawat, K., Smilkov, D., Wexler, J. et al., 2017, Visualizing Dataflow Graphs of Deep Learning Models in TensorFlow, IEEE Transactions on Visualization and Computer Graphics (IEEE VIS 17), Vol. 24, No. 1, pp. 1~12.

(4) Kahng, M., Andrews, P., Kalro, A. and Chau, D. H., 2017, ActiVis: Visual Exploration of Industry-scale Deep Neural Network Models, In Proceedings of IEEE Visual Analytics Science and Technology (IEEE VIS 17), Vol. 24, No. 1, pp. 88~97.