AI의 사고과정을 설명할 수 있을까?

2020.11.11

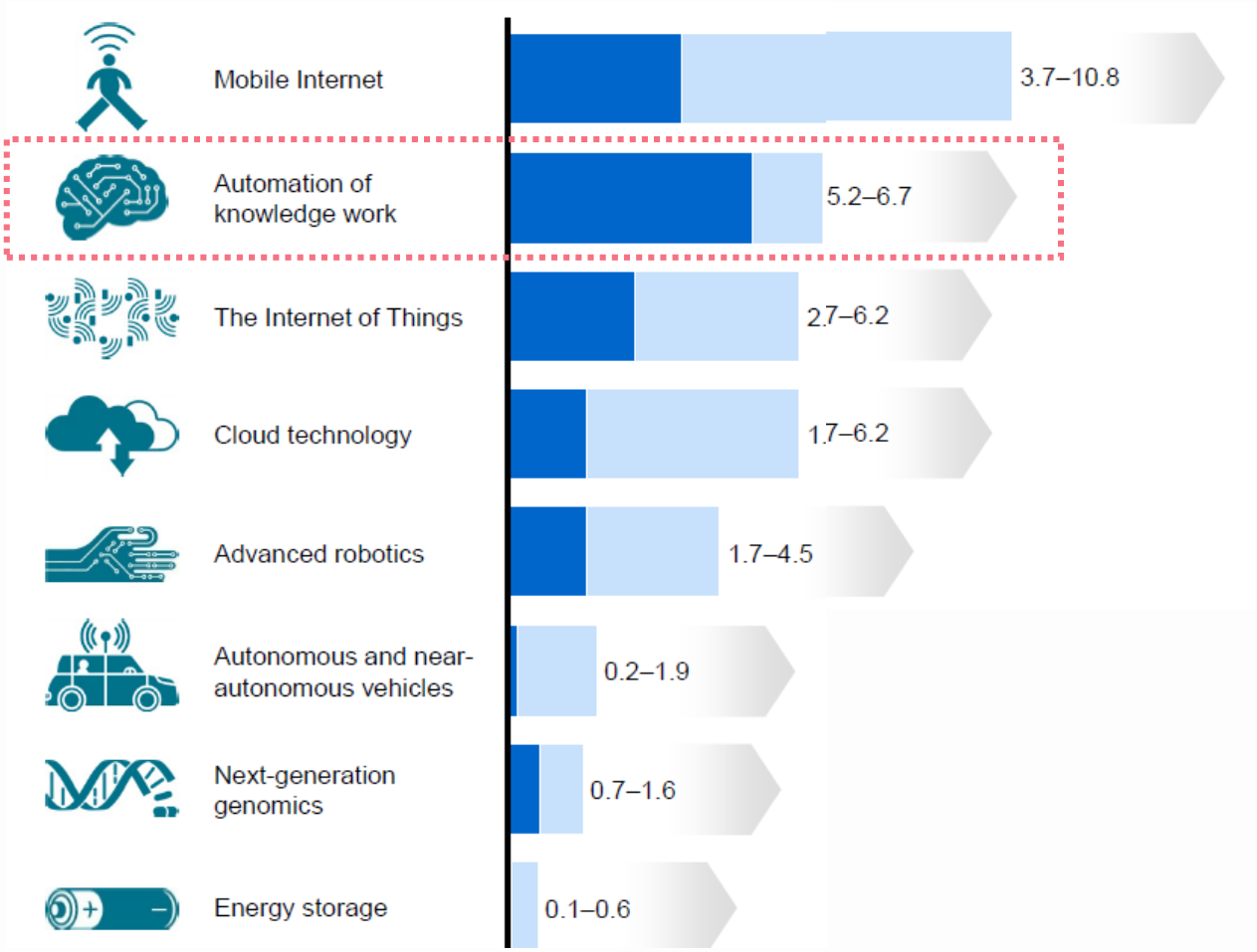
최 재 식

KAIST AI대학원

KAIST 설명가능인공지능 연구센터

AI의 미래

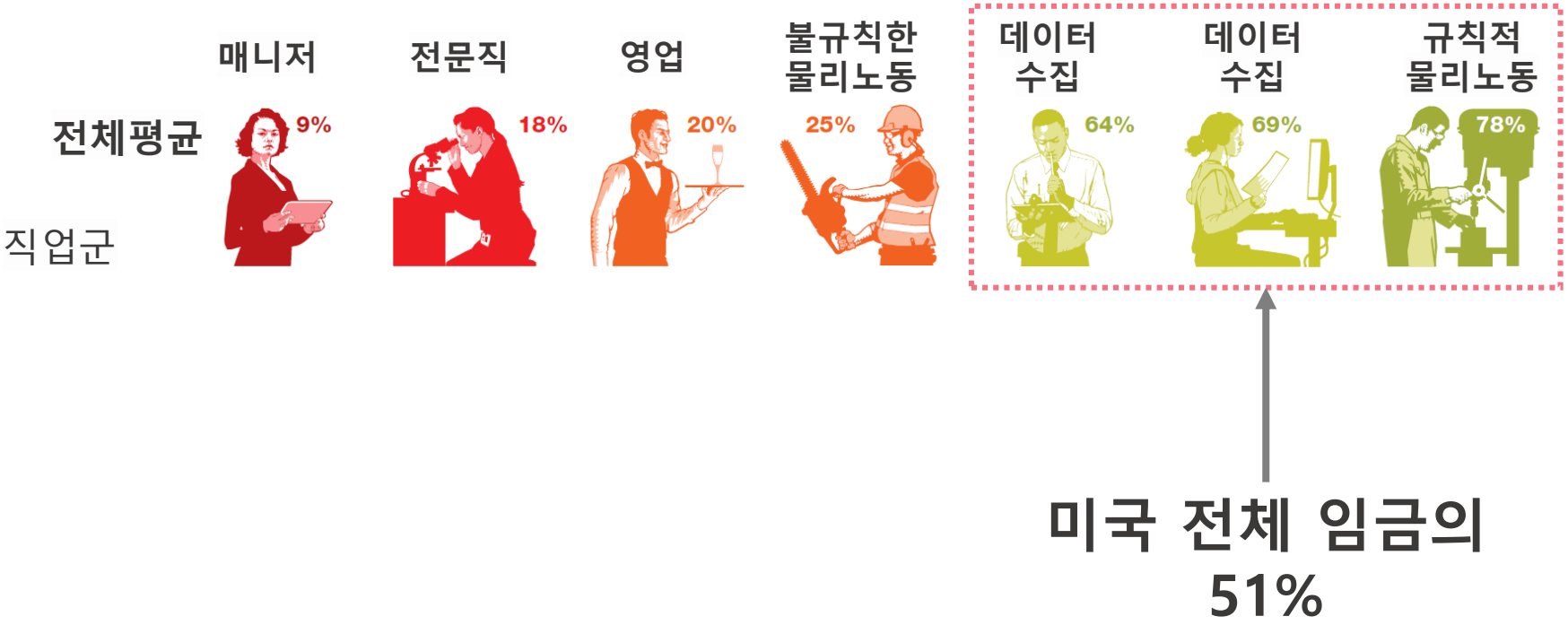
신기술의 발전으로 인한 사회 변화 및 영향



[Estimated Economic Impact of Disruptive Technology, McKinsey, 2013]

AI의 미래

AI 발전으로 인한 각 직업별 자동화 정도 예측



[Estimated Economic Impact of Disruptive Technology, McKinsey, 2013]

우리는 AI를 충분히 이해하고 있을까?

2018년 우버 자율주행차 첫 보행자 사망사고

AI 기반 신용평가 시스템의 문제점

MIT Technology Review

Intelligent Machines

The Financial World Wants to Open AI's Black Boxes

THE WALL STREET JOURNAL.

Capital One Pursues 'Explainable AI' to Guard Against Bias in Models

금융

유럽연합 General Data Protection Regulation 발효(2018년)

- 인공지능 알고리즘에 의해 자동으로 결정된 사안에 대해 회사의 설명을 강제

미국 Equal Credit Opportunity Act/Fair Housing Act

- 신용결정 및 주택 담보 대출 등 주요 금융 결정에 대해서 이유를 제시하도록 강제

현 인공지능 기반 신용 평가의 장단점

장점	단점
세밀한 분석을 통한 정교한 신용 점수 산정으로, **숨어있는 우수 고객을 선발** 할 수 있음	신용 거부의 **이유를 분명히 제시하지 못하여,** 인공지능 모델이 감독기관의 승인을 못 받을 수 있음

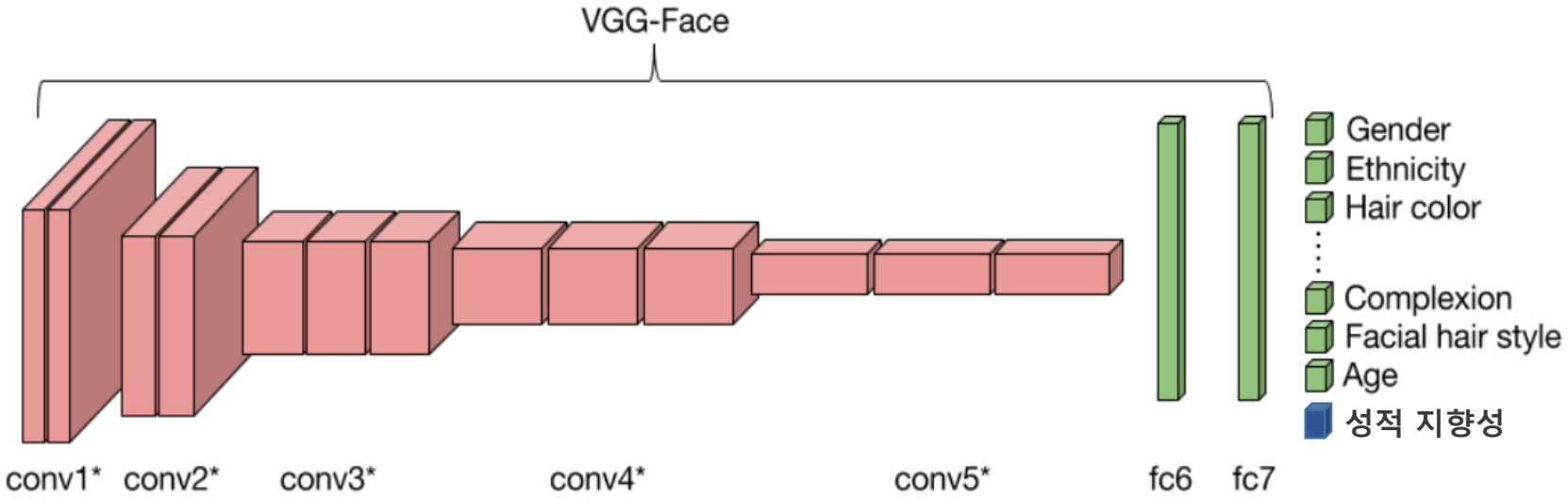
시사점 인공지능 기술을 주요 의사결정에 적용하기 위해서 법적으로 이유 제시 기능이 필요함

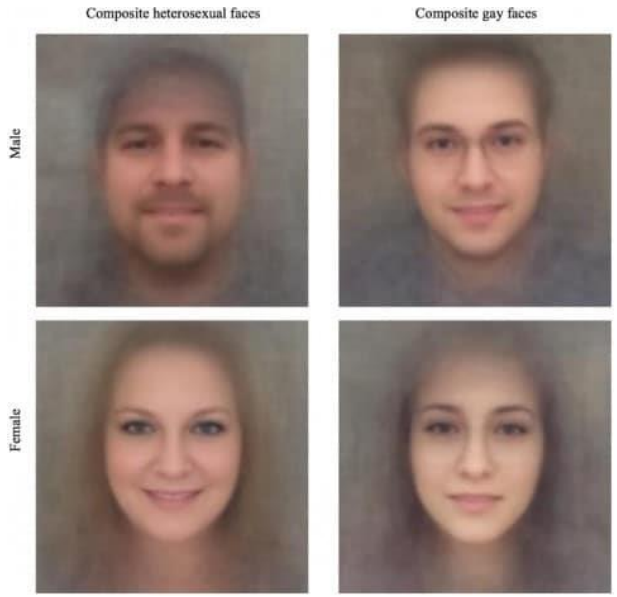
https://www.americanbanker.com/news/is-ai-making-credit-scores-better-or-more-confusing

EU의 일반정보보호규정

항목	내용
잊혀질 권리 **(right to be forgotten)**	제17조 – 정보 주체가 본인의 개인정보 처리를 더 이상 원치 않거나 개인정보를 보유할 법적 근거가 없으면 **해당 정보 삭제**
자동화된 의사결정 제한	제22조 – **자동화된 처리**(프로파일링 포함)에만 근거한 결정의 대상이 **되지 않을 권리**
설명을 요구할 권리 **(right to explanation)**	제13-14조 – 알고리즘에 의해 행해진 결정에 대해 질문하고, **결정에 관여한 논리에 대해 의미있는 설명을 요구할 권리**
EU 집행력	규정 위반시 해당 기업의 **전세계 매출의 최대 4%**까지 벌금 부과
발효	**2018년 5월 28일**

사람의 성적지향을 판별하는 딥러닝?





기존 동성애자/이성애자 판별 정확도: 55 ~ 65%

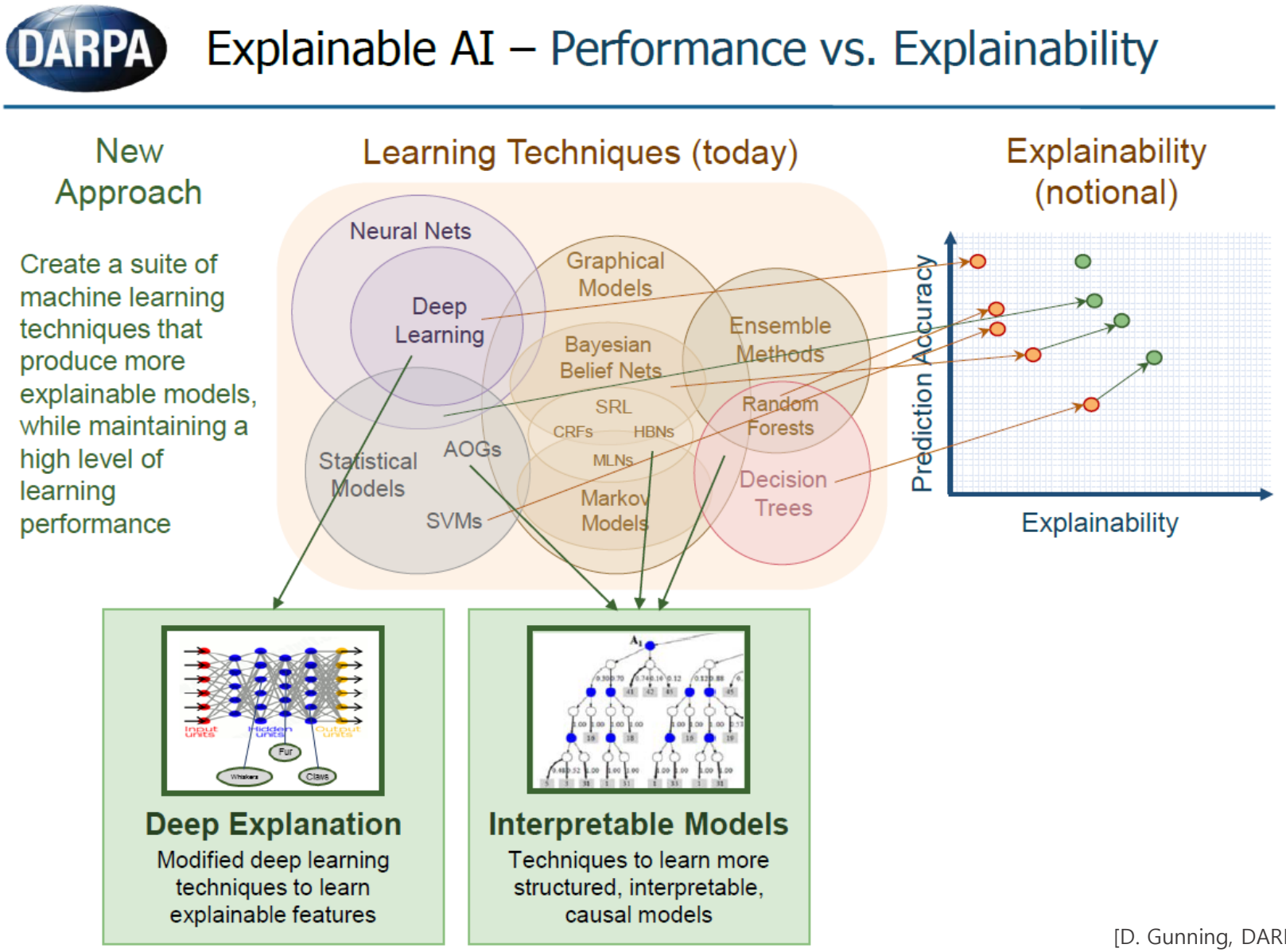
VS

VGG-Face의 정확도: 83 ~ 91%

문제: VGG-Face는 결정의 이유를 설명할 수 없음!

[M. Kosinski 교수 연구팀, Stanford, 2017년]

미국 DARPA 설명가능 인공지능



[D. Gunning, DARPA, 2016년]

설명가능 인공지능에 던지는 질문들

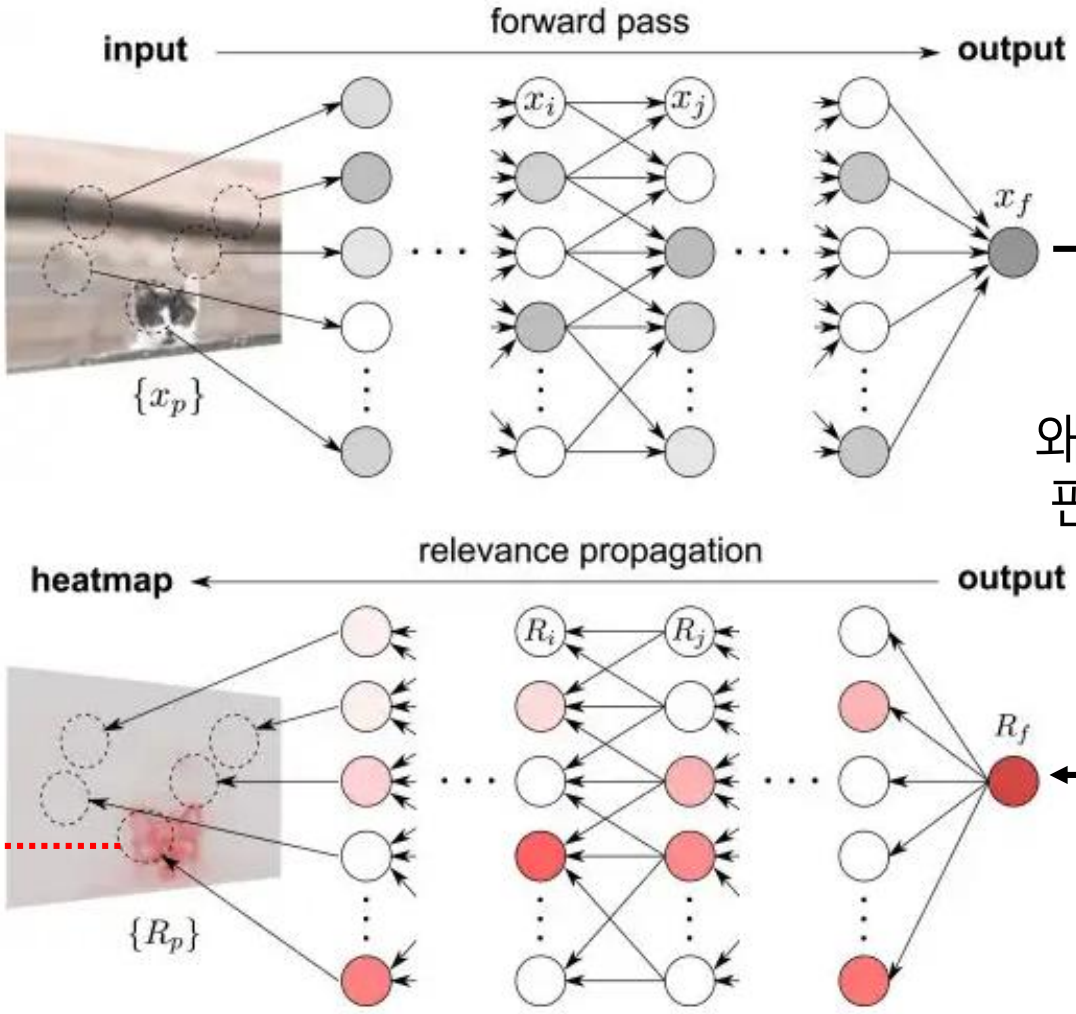
수 천만개의 뉴런이 있는데 딥러닝을 설명하는게 가능한가요?

딥러닝이 아닌 어떤 기계학습 모델도 설명하는게 가능한가요?

설명가능 인공지능이 산업에 응용된 사례가 있나요?

설명가능 딥러닝 연구

계층적 기여도 전파 기술 - Layer-wise Relevance Propagation(LRP)

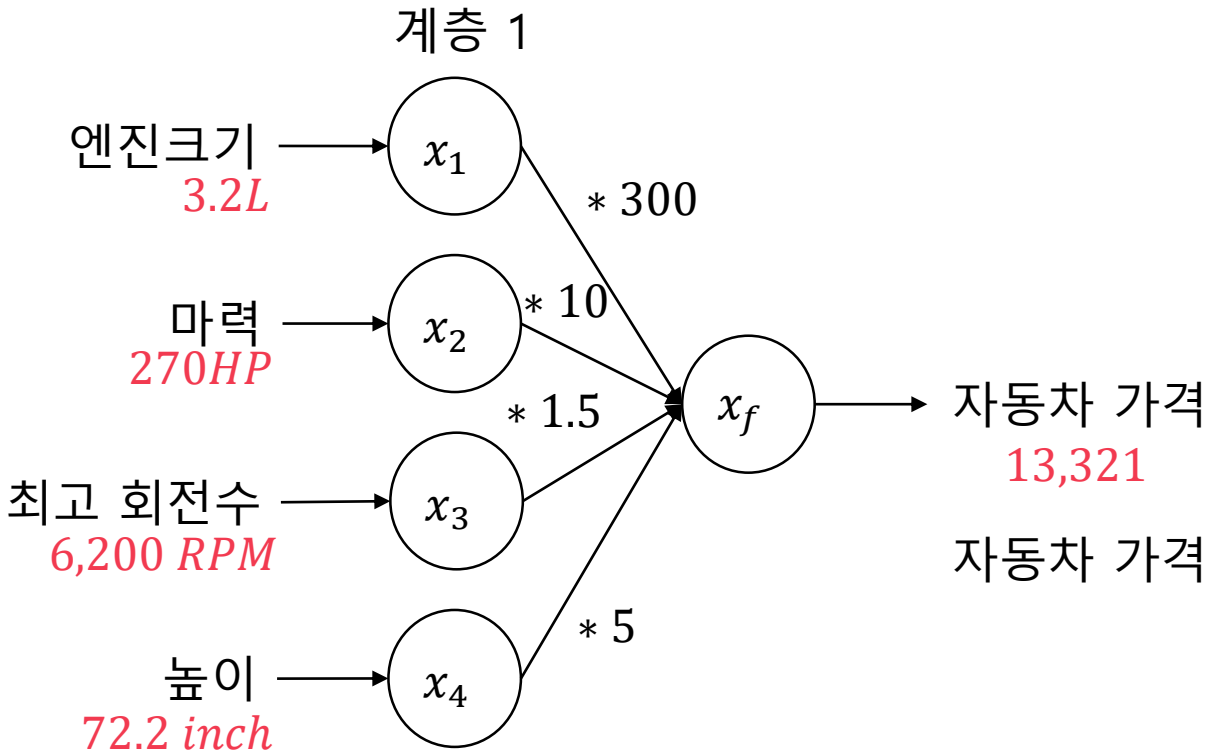


[K.-R. Muller 교수 연구팀, TU Berlin, 2015년]

딥러닝 설명의 원리: 선형 모델

$$\text{자동차의 가격} = 300 * \underset{3.2\ \text{L}}{\text{엔진크기}} + 10 * \underset{270\ \text{HP}}{\text{마력}} + 1.5 * \underset{6{,}200\ \text{RPM}}{\text{최고회전수}} + 5 * \underset{72.2\ \text{Inch}}{\text{차고}}$$

변수	계수
엔진크기	300
마력	10
최고 회전수	1.5
차고	5



$$300 * 3.2 + 10 * 270 + 1.5 * 6200 + 5 * 72.2 = 13{,}321$$

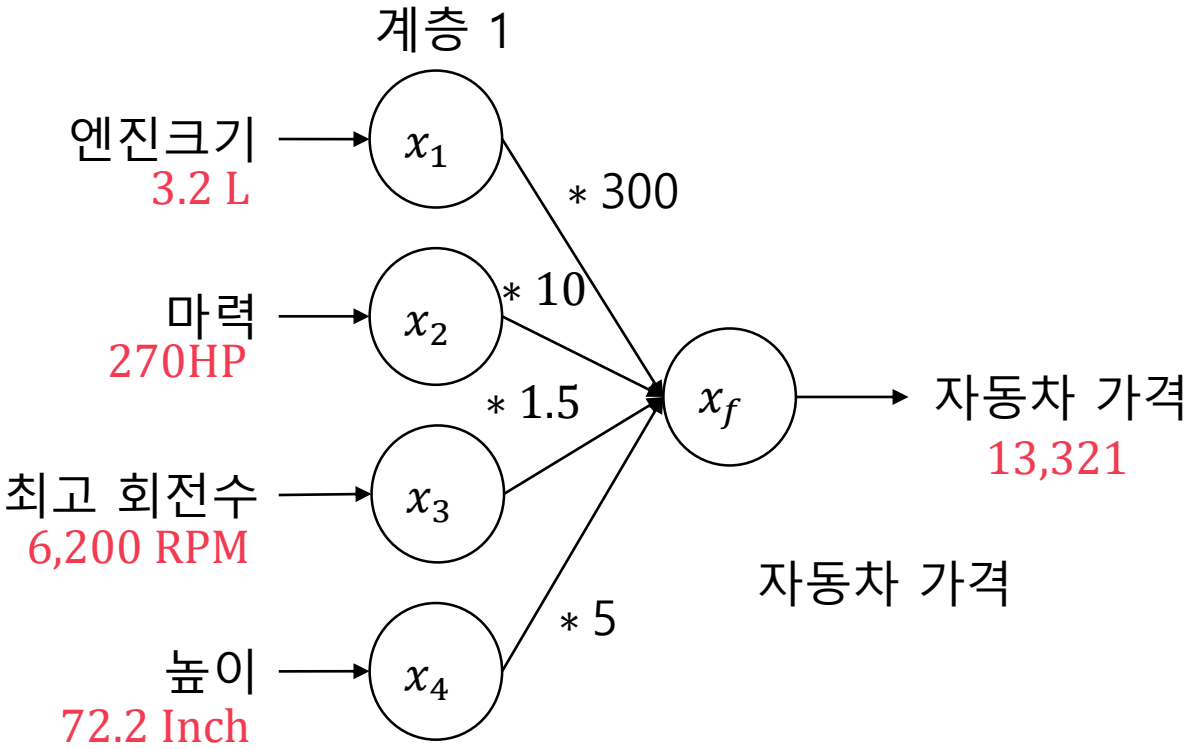
딥러닝 설명의 원리: 선형 모델

3.2 L 270 HP 6,200 RPM 72.2 Inch

$$\text{자동차의 가격} = 300 * \text{엔진크기} + 10 * \text{마력} + 1.5 * \text{최고회전수} + 5 * \text{차고}$$

$$13{,}321\text{원} = \underset{\text{엔진의 기여}}{960} + \underset{\text{마력의 기여}}{2{,}700} + \underset{\text{회선수의 기여}}{9{,}300} + \underset{\text{차고의 기여}}{361}$$

변수	기여도
엔진크기	960
마력	2,700
최고 회전수	9,300
차고	361

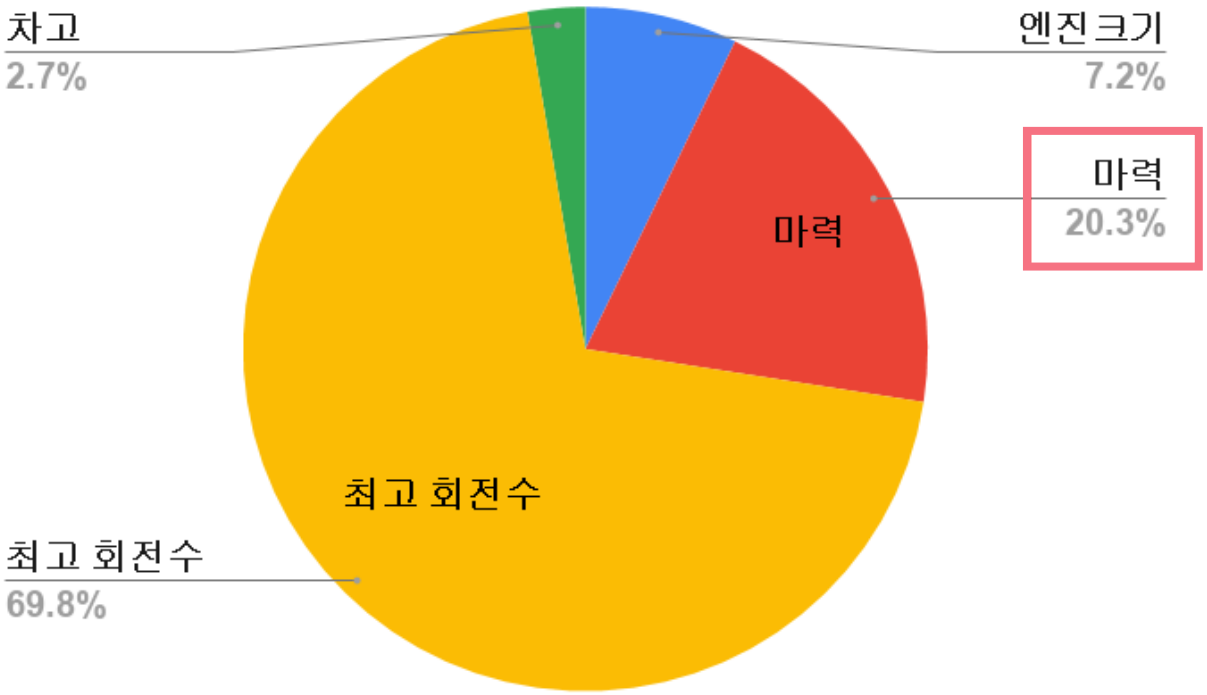


딥러닝 설명의 원리: 선형 모델

3.2 L 270 HP 6,200 RPM 72.2 Inch

$$\text{자동차의 가격} = 300 * \text{엔진크기} + 10 * \text{마력} + 1.5 * \text{최고회전수} + 5 * \text{차고}$$

변수	기여도
엔진크기	960
마력	2,700
최고 회전수	9,300
차고	361

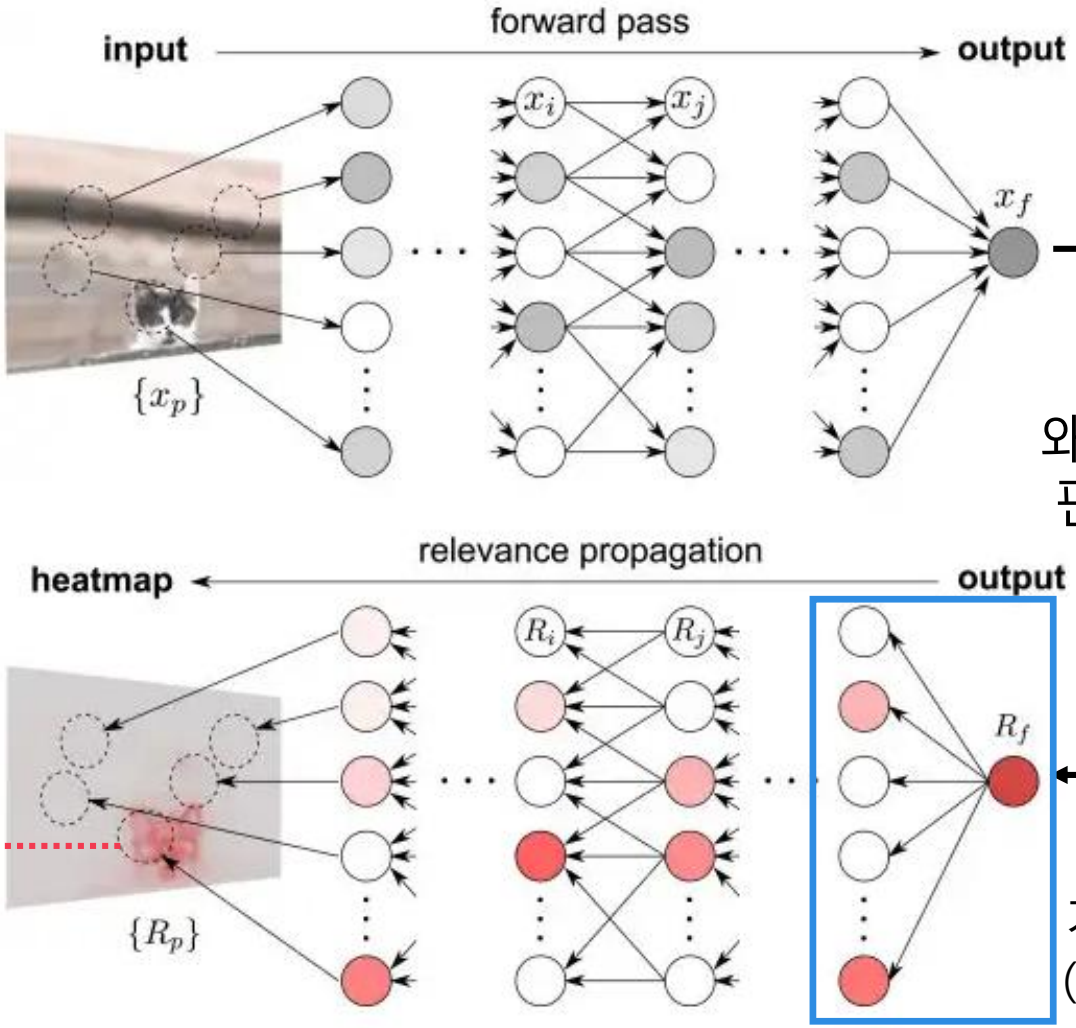


질문: 마력은 자동차의 가격에 얼만큼 기여 했습니까?

답변: 20.3%인 2,700원 입니다.

설명가능 딥러닝 연구

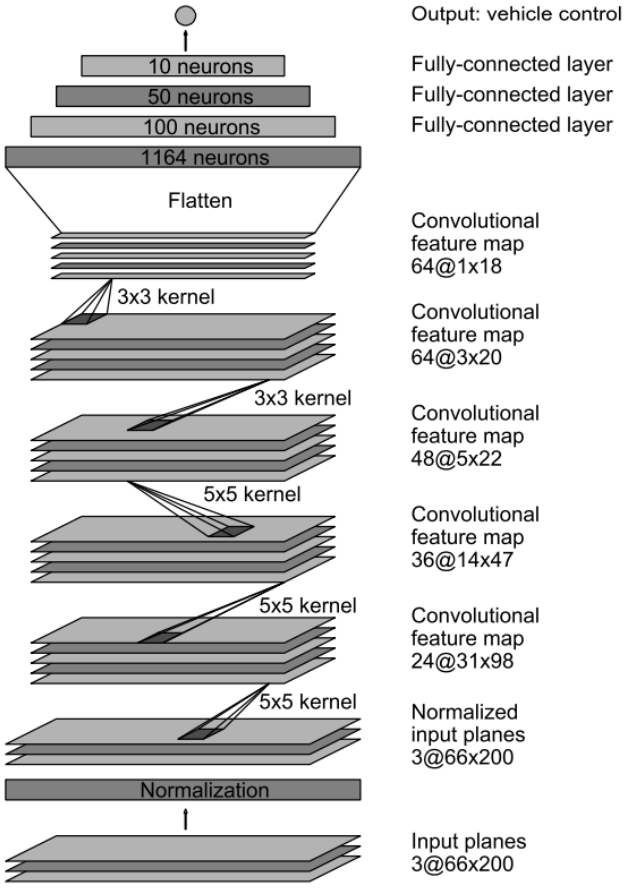
계층적 기여도 전파 기술 - Layer-wise Relevance Propagation(LRP)



[K.-R. Muller 교수 연구팀, TU Berlin, 2015년]

설명가능 딥러닝 연구

NVIDIA의 PilotNet(자율주행 딥러닝)의 결정을 설명하는 딥러닝



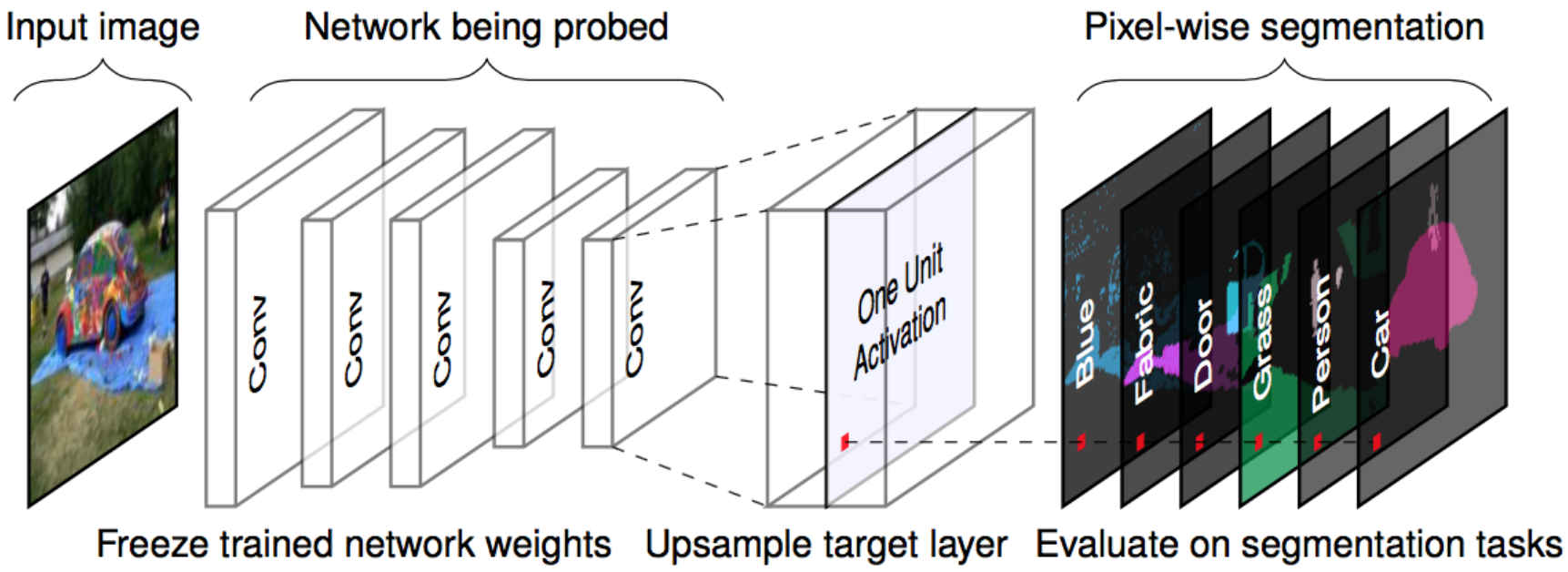
PilotNet 구조도

자율주행의 이유를 설명해주는 딥러닝 기술

[NVIDIA/Google 연구팀, 2017년]

네트워크의 내부를 설명하는 기술

Network Dissection(딥러닝 모델 해부)

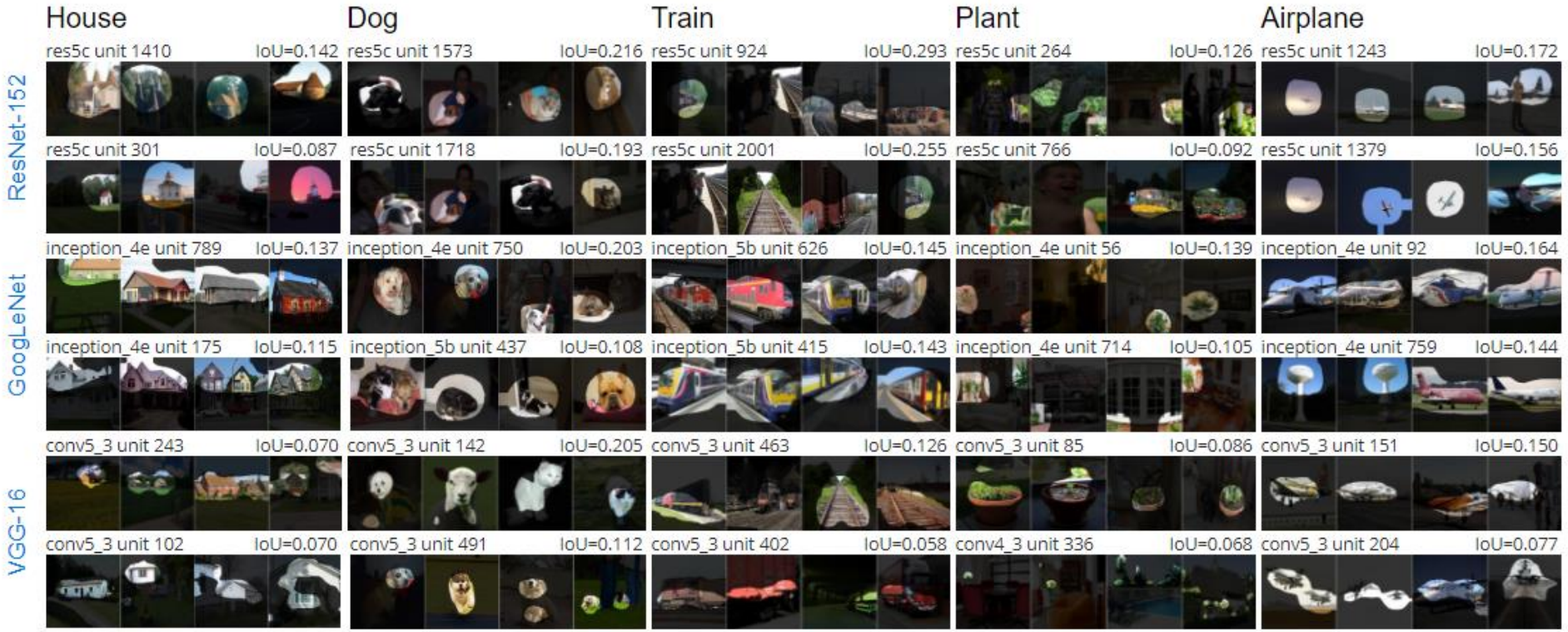


[A. Torralba 교수 연구팀, MIT, 2017]

[D. Bau et. al., CVPR, 2017]

네트워크의 내부를 설명하는 기술

Network Dissection(딥러닝 모델 해부)

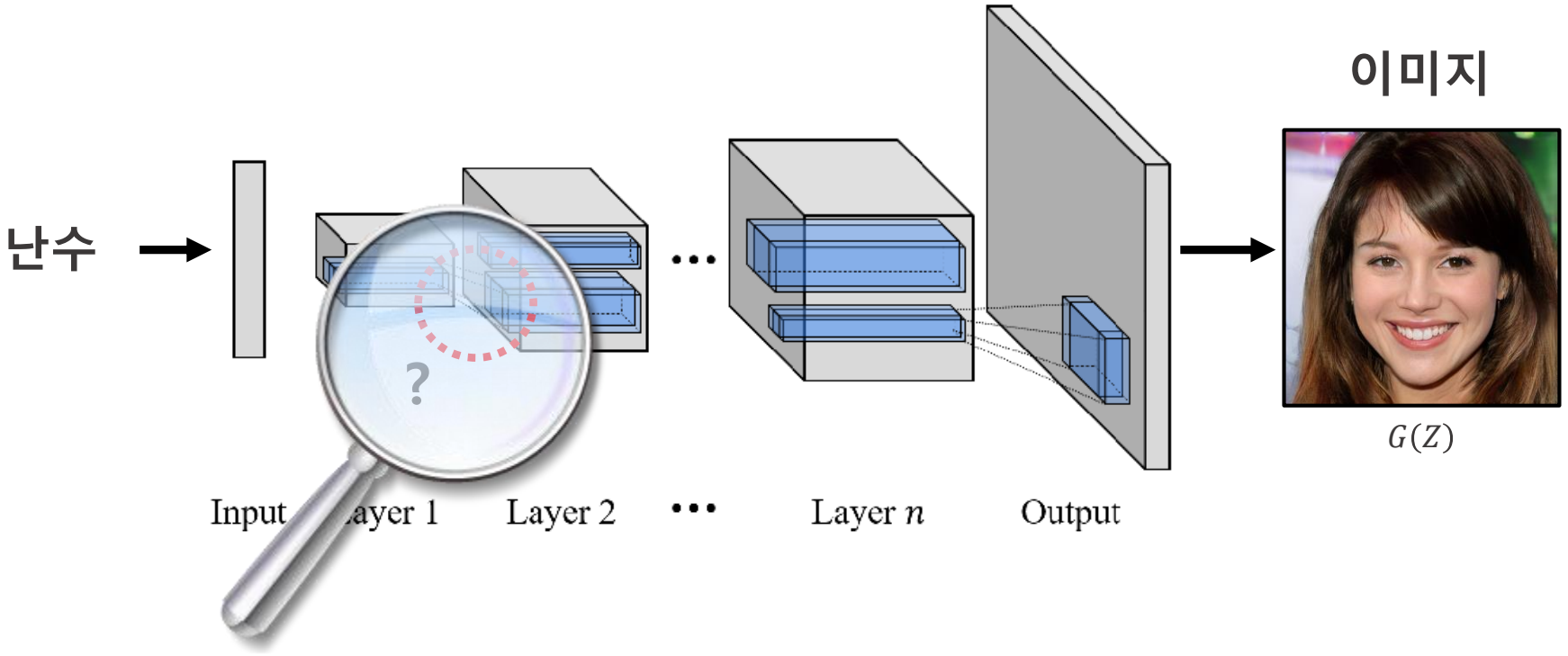


[A. Torralba 교수 연구팀, MIT, 2017]

[D. Bau et. al., CVPR, 2017]

네트워크의 내부를 설명하는 기술

생성 경계를 고려한 딥러닝 돋보기

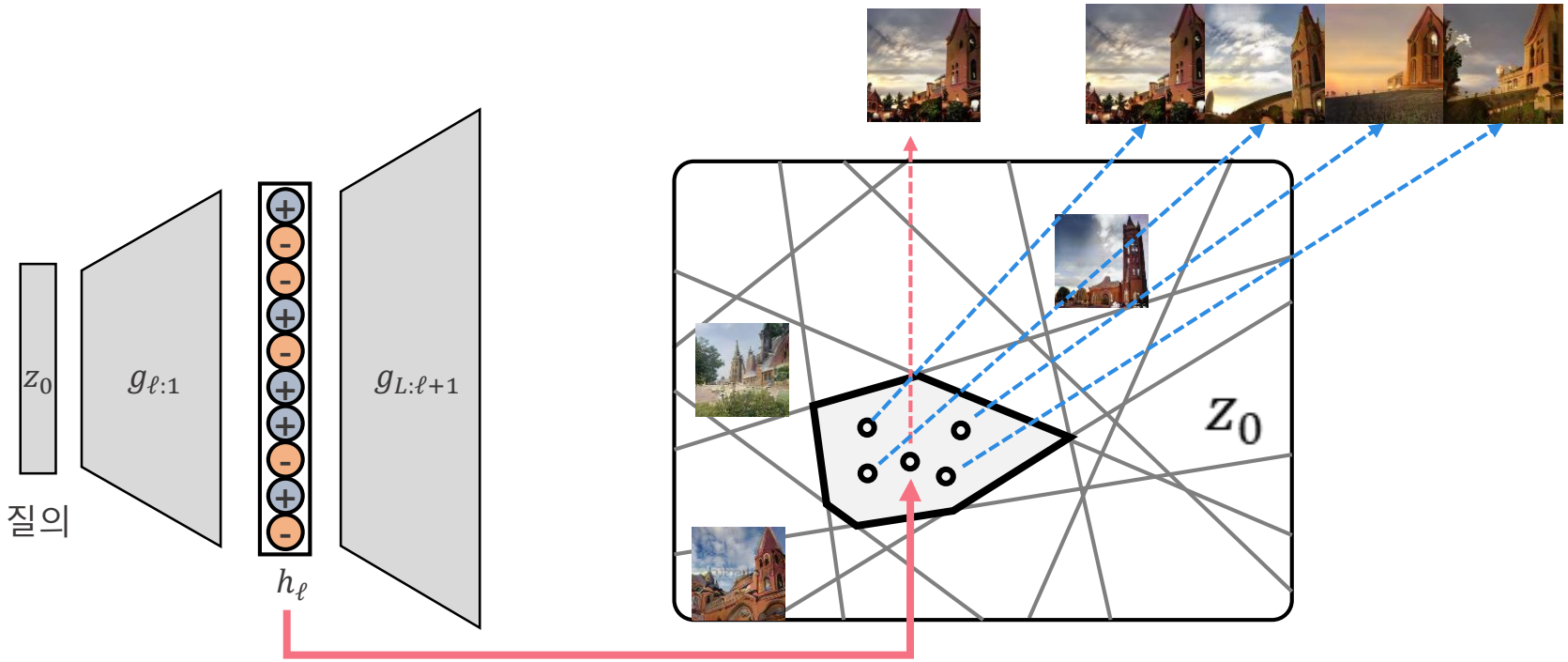


[최재식 교수 연구팀, KAIST/UNIST, 2017]

[G. Jeon et. al., AAAI, 2020]

네트워크의 내부를 설명하는 기술

생성 경계를 고려한 딥러닝 돋보기



딥러닝 생성공간

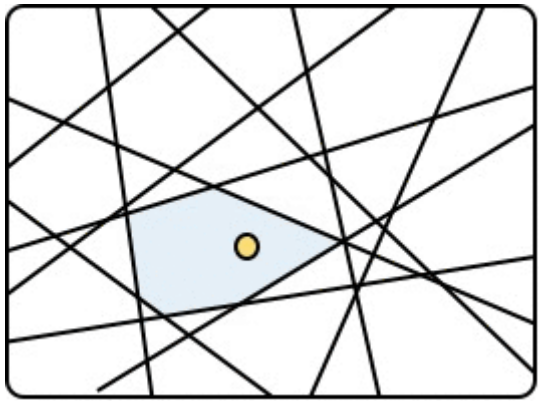
[최재식 교수 연구팀, KAIST/UNIST, 2017]

[G. Jeon et. al., AAAI, 2020]

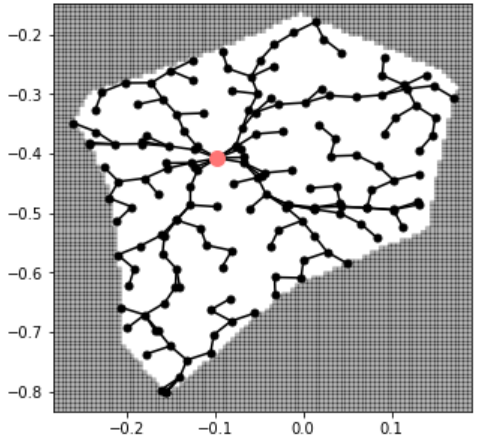
네트워크의 내부를 설명하는 기술

생성 경계를 고려한 딥러닝 돋보기

- 생성 경계를 고려한 빠른 탐색 난수 트리(Rapidly-exploring Random Tree)



Illustrative example



Example in nonconvex region

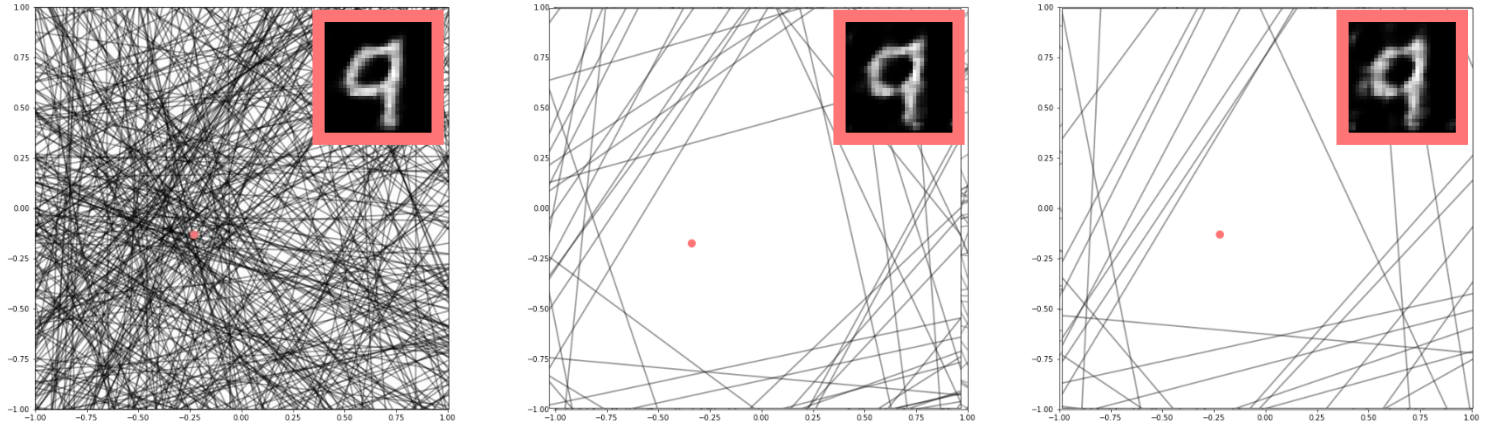
LaValle, Steven M. "Rapidly-exploring random trees: A new tool for path planning". *Technical Report. Computer Science Department, Iowa State University.* 1998.

[최재식 교수 연구팀, KAIST/UNIST, 2017]

[G. Jeon et. al., AAAI, 2020]

네트워크의 내부를 설명하는 기술

생성 경계를 고려한 딥러닝 돋보기

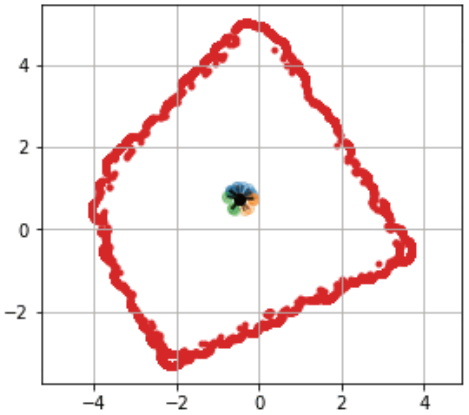


[최재식 교수 연구팀, KAIST/UNIST, 2017]

[Giyoung Jeon et. al., AAAI, 2020]

네트워크의 내부를 설명하는 기술

생성 경계를 고려한 딥러닝 돋보기



- Accepted Cluster 1
- Accepted Cluster 2
- Accepted Cluster 3
- Rejected Sample

[최재식 교수 연구팀, KAIST/UNIST, 2017]

[Giyoung Jeon et. al., AAAI, 2020]

설명가능 인공지능에 던지는 질문들

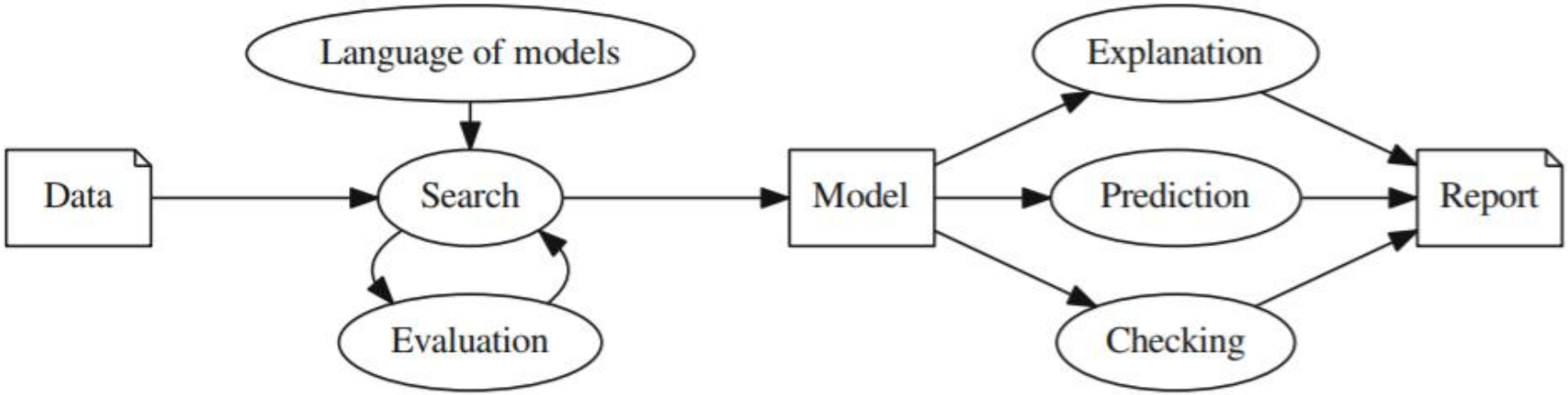
수 천만개의 뉴런이 있는데 딥러닝을 설명하는게 가능한가요?

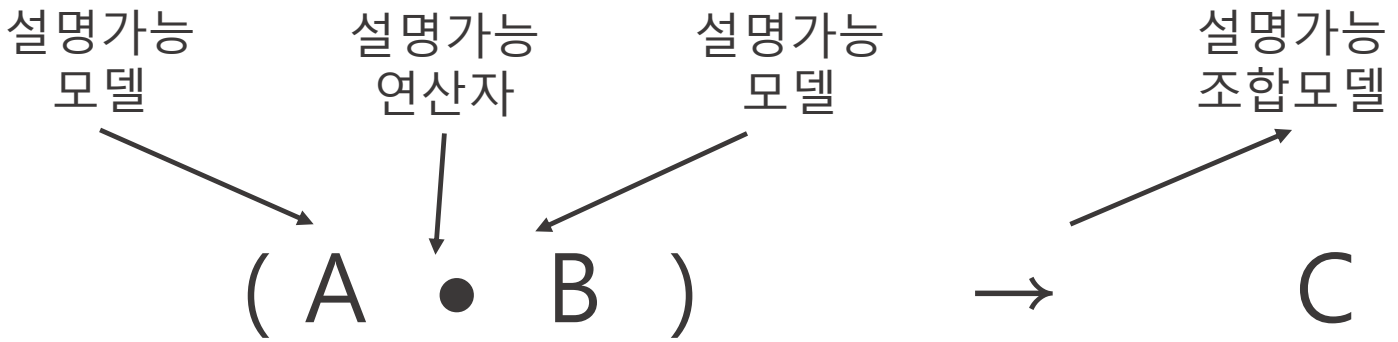
딥러닝이 아닌 어떤 기계학습 모델도 설명하는게 가능한가요?

설명가능 인공지능이 산업에 응용된 사례가 있나요?

설명가능한 베이지안 모델

자동 통계학자



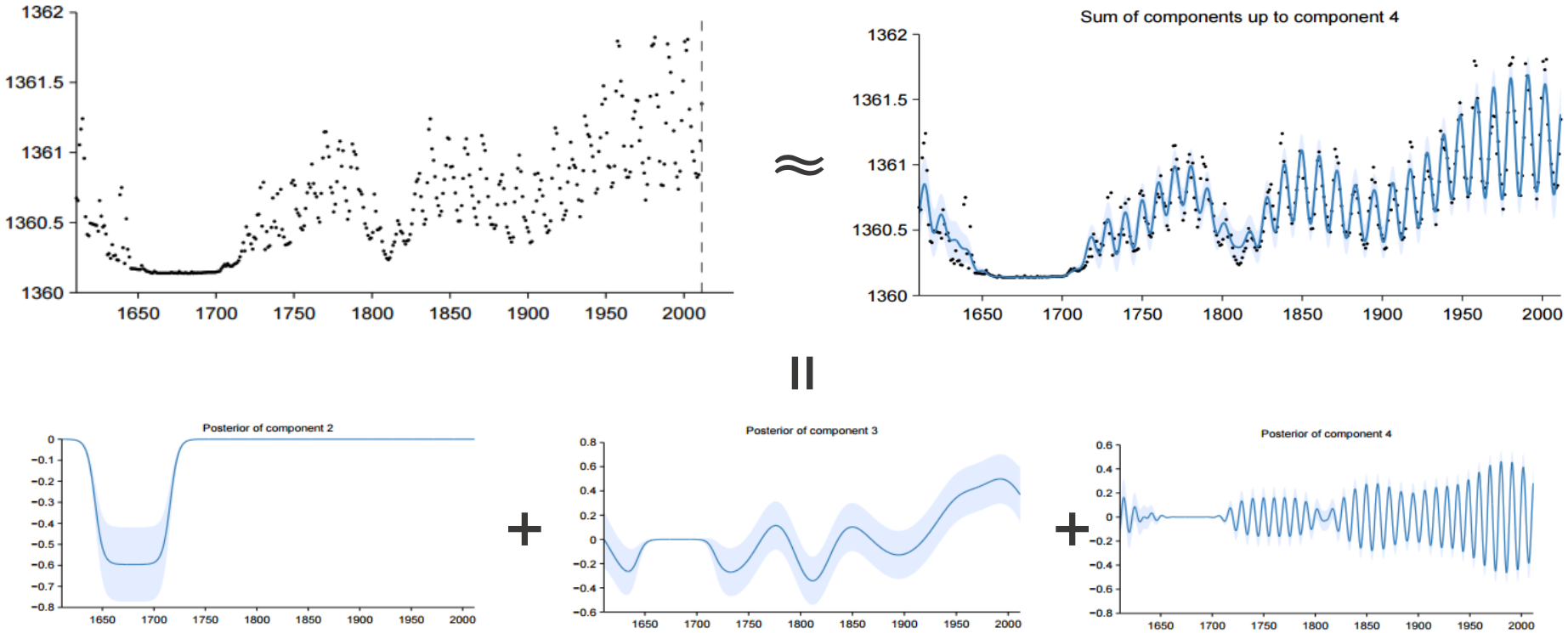


[Z. Ghahramani 교수 연구팀, Cambridge, 2016]

설명가능한 베이지안 모델

자동 통계학자

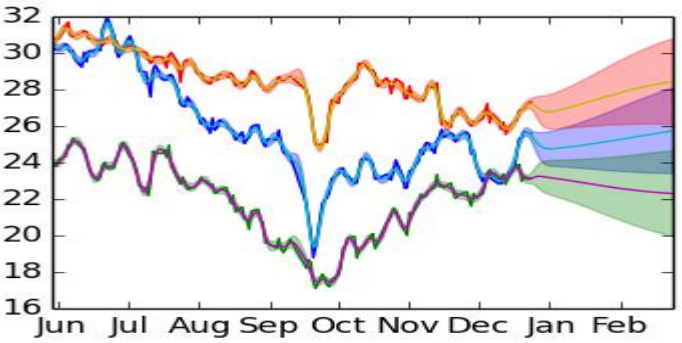
태양의 흑점 활동 데이터

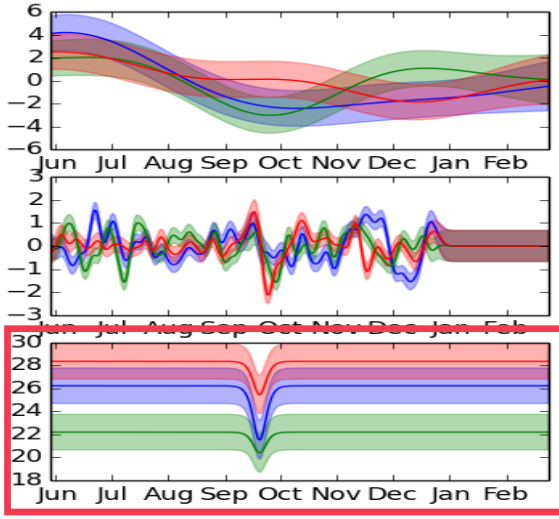


[Z. Ghahramani 교수 연구팀, Cambridge, 2016]

설명가능한 베이지안 모델

관계형 자동 통계학자





Smooth function
Length scale: y weeks

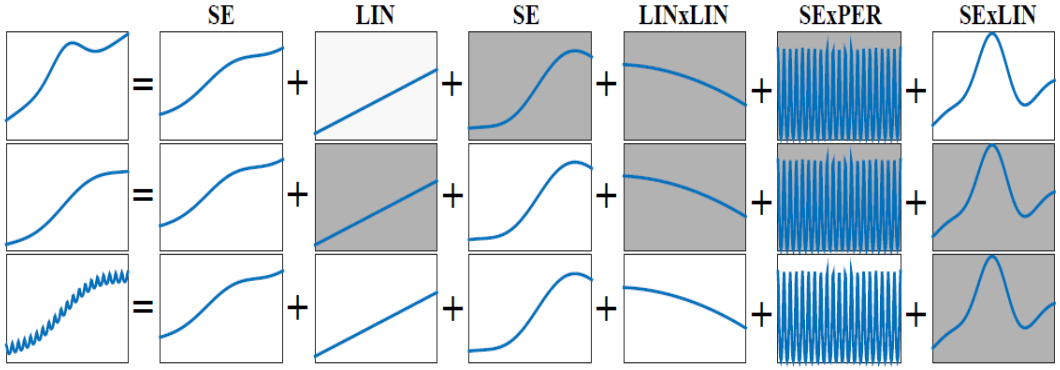
Rapidly varying
smooth function
Length scale: z hours

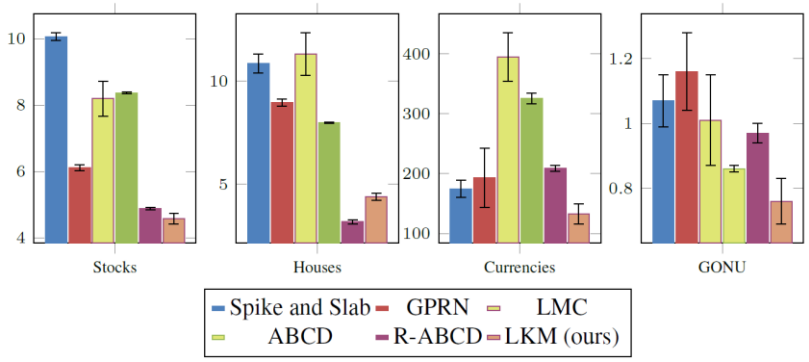
Constant function
Sudden drop btw
9/12/01 ~ 9/15/01

[Y. Hwang et. al., ICML, 2016]

Anh Tong and Jaesik Choi, "Discovering Relational Covariance Structures for Explaining Multiple Time Series", ICML 2019

Latent Kernel Model (Tong and Choi, ICML 2019)





[최재식 교수 연구팀, KAIST/UNIST, 2019]

[Y. Hwang et. al., ICML, 2016]

[A. Tong and J. Choi, ICML, 2019]

설명가능한 베이지안 모델

관계형 자동 통계학자

- Gold, Oil, NASDAQ, USD index share the following property:

This component is periodic with a period of 1.4 years but with varying amplitude. The amplitude of the function increases linearly away from Apr 2017. The shape of this function within each period has a typical lengthscale of 4.9 days.

- Gold, Oil, USD index share the following property:

This component is a smooth function with a typical lengthscale of 2.7 weeks.

- NASDAQ has the following property:

This component is a linear function.

[최재식 교수 연구팀, KAIST/UNIST, 2019]

[A. Tong and J. Choi, ICML, 2019]

의사결정 트리를 설명하는 공개SW

xgboost library의 'save_model' 함수를 활용하여 저장한 파일을 업로드 해주세요.

(학습된 XGBoost model 파일을 업로드 하지 않고 제출시 타이타닉 예제가 실행됩니다.)

파일 선택 선택된 파일 없음 제출

[최재식 교수 연구팀, KAIST, 2020]

[H. Kim et. Al., Github, 2020]

설명가능 인공지능에 던지는 질문들

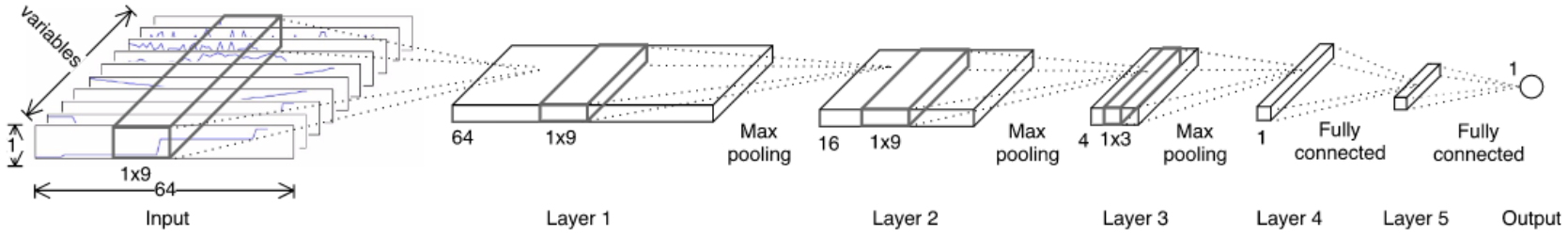
수 천만개의 뉴런이 있는데 딥러닝을 설명하는게 가능한가요?

딥러닝이 아닌 어떤 기계학습 모델도 설명하는게 가능한가요?

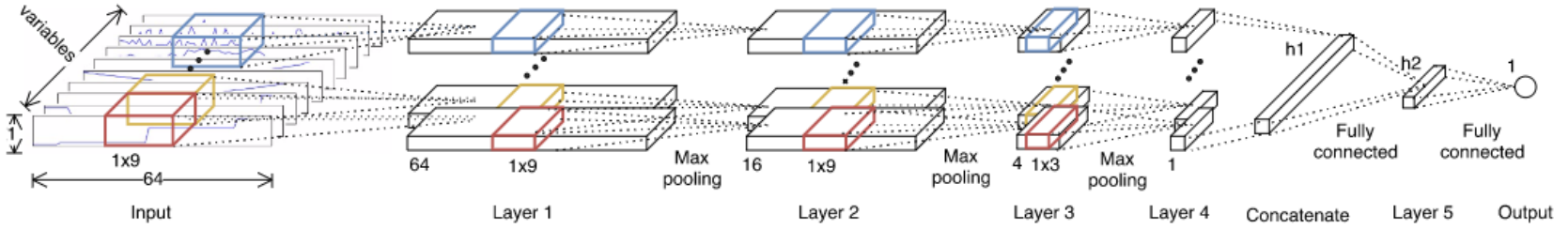
설명가능 인공지능이 산업에 응용된 사례가 있나요?

AI의 산업응용

스마트고로를 제어하는 AI



(a) Convolutional Neural Network (CNN).



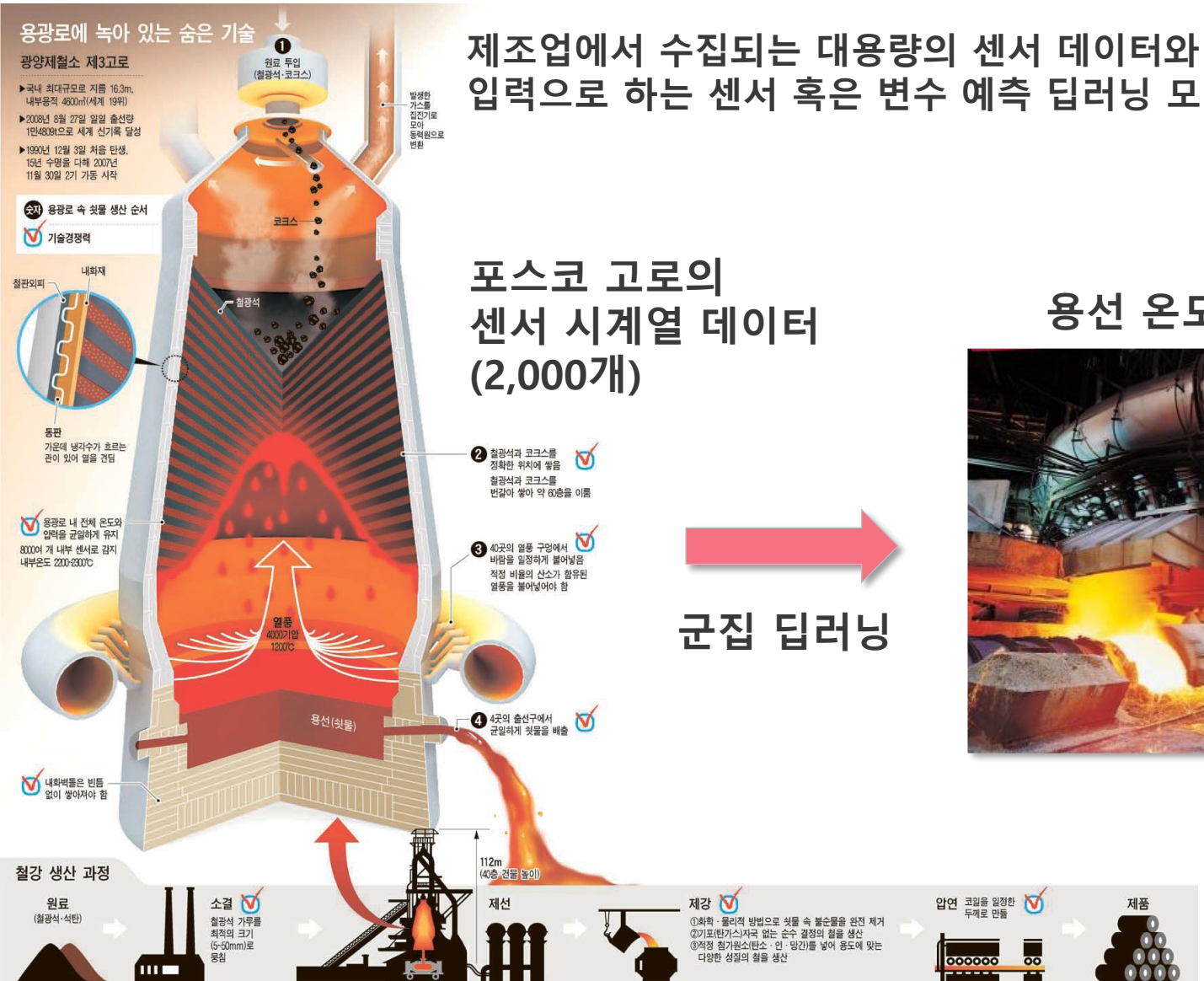
(b) CNN with grouped convolutional layers.

[최재식 교수 연구팀, UNIST, 2017]

[S. Yi et al., 2017]

AI의 산업응용

스마트고로를 예측하고 제어하는 AI



제조업에서 수집되는 대용량의 센서 데이터와 파생 변수를 입력으로 하는 센서 혹은 변수 예측 딥러닝 모델

포스코 고로의
센서 시계열 데이터
(2,000개)

군집 딥러닝

용선 온도 변화 예측

AI의 산업응용

스마트고로를 예측하고 제어하는 AI

국가핵심기술 선정 – 2019년 7월

포스코, 인공지능 고로조업 기술·스마트 수(水) 냉각기술..국가핵심기술로 지정

김인규 기자 | 승인 2019.07.23 17:11

철강 전체 국가핵심기술 9건 중 6건 보유... 절반이 스마트기술 기반

▲ 포스코가 자체개발한 스마트 팩토리 기술로 수집·분석한 정보를 활용해 조업하는 모습을 구현한 이미지 사진

'딥러닝 인공지능 기반의 고로 조업 자동제어기술'은 데이터를 분류하고 예측하는 딥러닝(Deep Learning) 기술을 활용해 고로 내부 상태를 자동으로 제어하는 기술이다. 그동안에는 작업자가 2시간마다 노열(爐熱)을 수동으로 측정해야 했지만, 이제는 고로 하부에 설치된 센서가 쇳물의 온도를 실시간 측정하고, 노내 열 수준을 예측해 용선 온도를 자동제어한다. 또한 풍구에 설치된 카메라로 노내 상태를 평가하고 철광석과 코크스 장입을 자동제어한다.포항제철소 2고로는 해당 기술을 적용해 연평균 생산량은 5% 개선되고 연료량은 1% 절감하는 효과를 냈다.

AI의 산업응용

스마트고로를 예측하고 제어하는 AI

국가핵심기술 선정 – 2019년 7월

딥러닝 인공지능 기반의 고로 조업 자동제어 기술

국가핵심기술이란?

국내외 시장에서 차지하는 **기술적·경제적 가치가 높거나 관련 산업의 성장 잠재력이 높아** 해외로 유출될 경우에 국가의 안전보장 및 국민경제의 발전에 중대한 악영향을 줄 우려가 있는 산업기술로서 「산업기술의 유출방지 및 보호에 관한 법률(이하 "법률")」 제9조에 따라 지정된 산업기술

반도체/디스플레이(10기), 전기전자(3개), 자동차/철도(9개), 철강(9개), 조선(7건), 원자력(5개), 정보통신(9개), 우주(4개), 생명공학(3개), 기계(6개), 로봇(3개) 총 69건

AI의 산업응용

POSCO 국내최초 세계제조포럼(WEF) 등대공장 선정 2019년 7월

한국일보

WEF서 세계 제조업 미래 이끌 '등대공장'으로 포스코 선정

입력 2019.07.29 22:00

포스코가 구현하고 있는 스마트 팩토리 콘셉트.

포스코가 국내 기업 최초로 세계 제조업의 미래를 선도할 '등대공장'으로 선정됐다.

포스코는 세계경제포럼(WEF, World Economic Forum, 다보스포럼)이 중국 다롄에서 열린 '2019 세계경제포럼'에서 포스코를 세계의 '등대공장(Lighthouse factory)'으로 선정했다고 지난 3일 밝혔다.

AI의 산업응용

스마트고로의 의사결정을 설명하는 기술

포스코, 포항제철소 고로에 '설명 가능 AI' 기술 적용

박재철 기자 | 승인 2020.02.10 17:23 | 댓글 0

> 2고로와 3고로에 XAI 확대 적용해 품질 및 생산성 향상 기대

포스코(회장 최정우)가 '인공지능 용광로'로 불리는 포항제철소 2고로와 3고로에 '설명 가능 인공지능'(XAI·Explainable AI) 기술을 적용한다.

XAI는 인공지능이 의사결정을 내린 이유를 설명해주는 시스템으로 인공지능의 활용성을 높일 수 있는 차세대 AI기술이다.

기존 인공지능 시스템은 주어진 자료를 정확히 분석하고 예측할 수 있으나 그 결과에 대한 원인을 알기 어려운 단점이 있었다. 그러나 XAI는 결과에 대한 핵심 원인을 파악 할 수 있어 인공지능에 대한 신뢰성을 획기적으로 높일 수 있다.

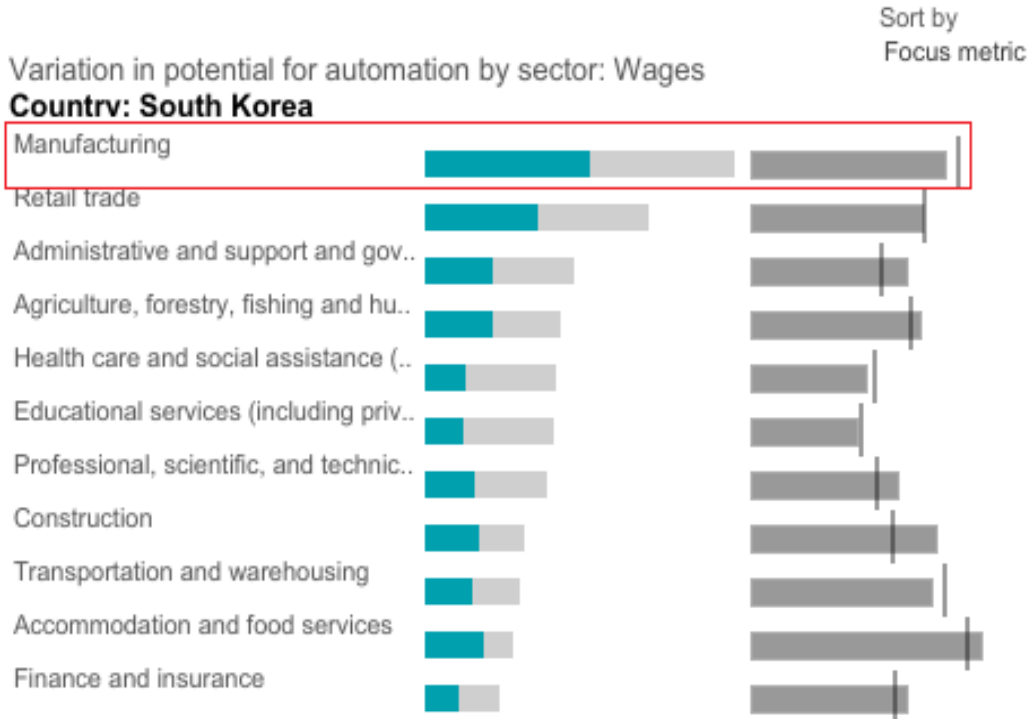
포스코 기술연구원은 최재식 한국과학기술원(KAIST) 교수팀과 협업해 포스코 고로에 XAI 기술 적용을 추진 중이다.

철강금속신문

AI의 산업응용

2030년까지 인공지능 기술의 국내 경제적 효과

- 의료(최대 109조원), 제조(최대 92조원), 금융(최대 47조원)순으로 효과가 기대
- 특히, 6대 주력산업(전자/자동차/석유화학/기계장치/철강/조선)의 AI 기술 발전에 따라 최대 25조원의 경제효과가 기대



- AI 기반 국내 제조업의 자동화 가능성은 60%로 의료(35%), 금융(48%)에 비하여 높아

,

인공지능의 미래는?

글을 읽고 예측하며 설명하는 AI

당신은 어떻게 장기 투자를 잘 하십니까?

I read annual reports of the company I'm looking at and I read the annual reports of the competitors – that is the main source of material.

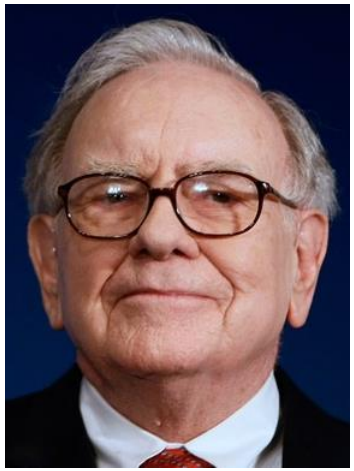
저는 그 회사의 연차 보고서를 읽고, 경쟁회사의 연차보고서를 읽습니다. 그것이 제 의사결정의 주요한 소재입니다.

글을 읽고 예측하며 설명하는 AI

당신은 어떻게 장기 투자를 잘 하십니까?

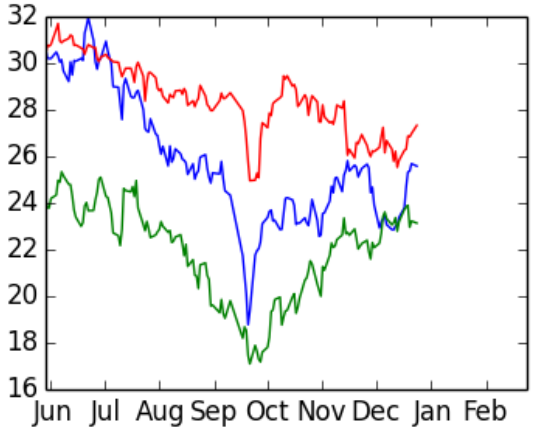
I read annual reports of the company I'm looking at and I read the annual reports of the competitors – that is the main source of material.

저는 그 회사의 연차 보고서를 읽고, 경쟁회사의 연차보고서를 읽습니다. 그것이 제 의사결정의 주요한 소재입니다.



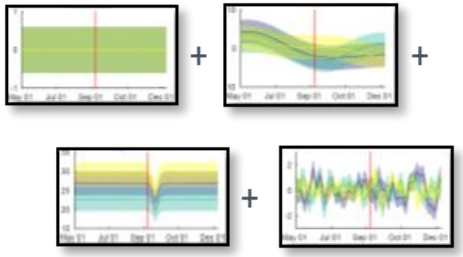
워렌 버핏
Warren Buffett

글을 읽고 예측하며 설명하는 AI

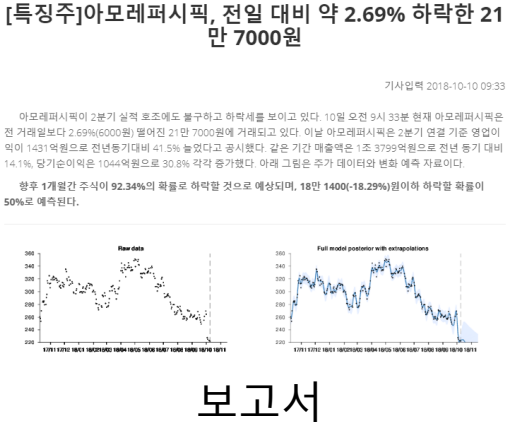


시계열 데이터

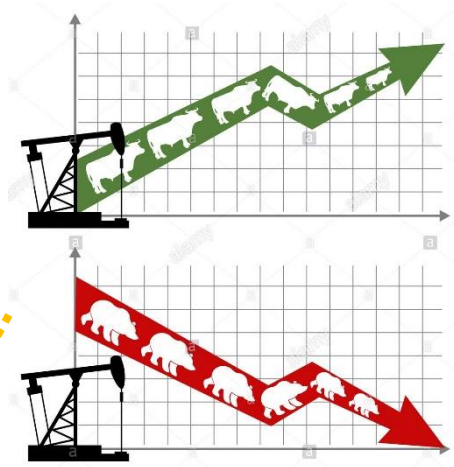
연차보고서



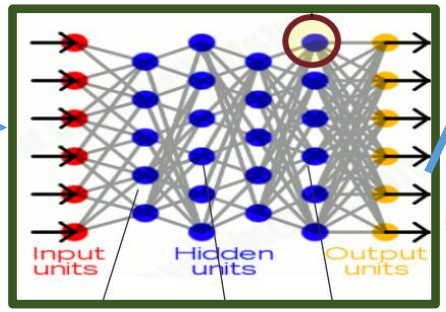
베이지안 학습



보고서



예측



딥러닝

설명 탐색

보고서와 데이터를 읽고 설명하는 AI

Learn, Practice and Generate Knowledge to
Solve Some of the World's Greatest Problems in AI.

감사합니다.

jaesik.choi@kaist.ac.kr

http://xai.kaist.ac.kr
http://sailab.kaist.ac.kr

참고문헌

- S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation, PLOS ONE, 2015.
- J. Manyika, M. Chui, J. Bughin, R. Dobbs, P. Bisson, and A. Marrs, Disruptive technologies: Advances that will transform life, business, and the global economy, McKinsey Global Institute, 2013.
- M. W. Kosinski, Y. Wang, Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images, Journal of Personality and Social Psychology, 2017.
- D. Gunning, Explainable Artificial Intelligence, DARPA, 2016.
- M. Bojarski, L. Jackel, B. Firner and U. Muller, Explaining How a Deep Neural Network Trained with End-to-End Learning Steers a Car, Arxiv:1704.07911, 2017.
- D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, Network Dissection: Quantifying Interpretability of Deep Visual Representations, IEEE Conference on Computer Vision and Pattern Recognition, 2017.
- G. Jeon, H. Jeong and J. Choi, An Efficient Explorative Sampling Considering the Generative Boundaries of Deep Generative Neural Networks, AAAI Conference on Artificial Intelligence, 2020.
- Z Ghahramani, Probabilistic machine learning and artificial intelligence, Nature, 2015.
- Y. Hwang, A. Tong and J. Choi, Automatic Construction of Nonparametric Relational Regression Models for Multiple Time Series, International Conference on Machine Learning, 2016.
- A. Tong and J. Choi, Discovering Explainable Latent Covariance Structure for Multiple Time Series, International Conference on Machine Learning, 2019.
- S. Yi, J. Ju, M.-K. Yoon and J. Choi, Grouped Convolutional Neural Networks for Multivariate Time Series, arXiv 1703.09938, 2017.
- H. Kim, S. Cho and J. Choi, Visualizing Extream Gradient Boost Models, https://github.com/OpenXAIProject/ExplainableXGBoost, 2020