

PyCon Korea 2019

딥러닝 안에서 일어나는 과정을 해석하는 설명가능인공지능 기술

2019년 8월 16일

UNIST 설명가능인공지능 연구센터
허성만, 조소희, 최재식

진행



최재식

UNIST ECE 부교수
UNIST XAIC 센터장



조소희

UNIST XAIC 연구원



허성만

UNIST XAIC 연구원

보조 진행자



차성국

UNIST Sailab



이희찬

UNIST Sailab

CONNECT
THE PYTHONISTAS

목차

발표 1. 설명가능인공지능(Explainable Artificial Intelligence, XAI)이란?

발표 2. 레이어 단위 관련성 전파 모델

2.1 Layer-wise Relevance Propagation: LRP

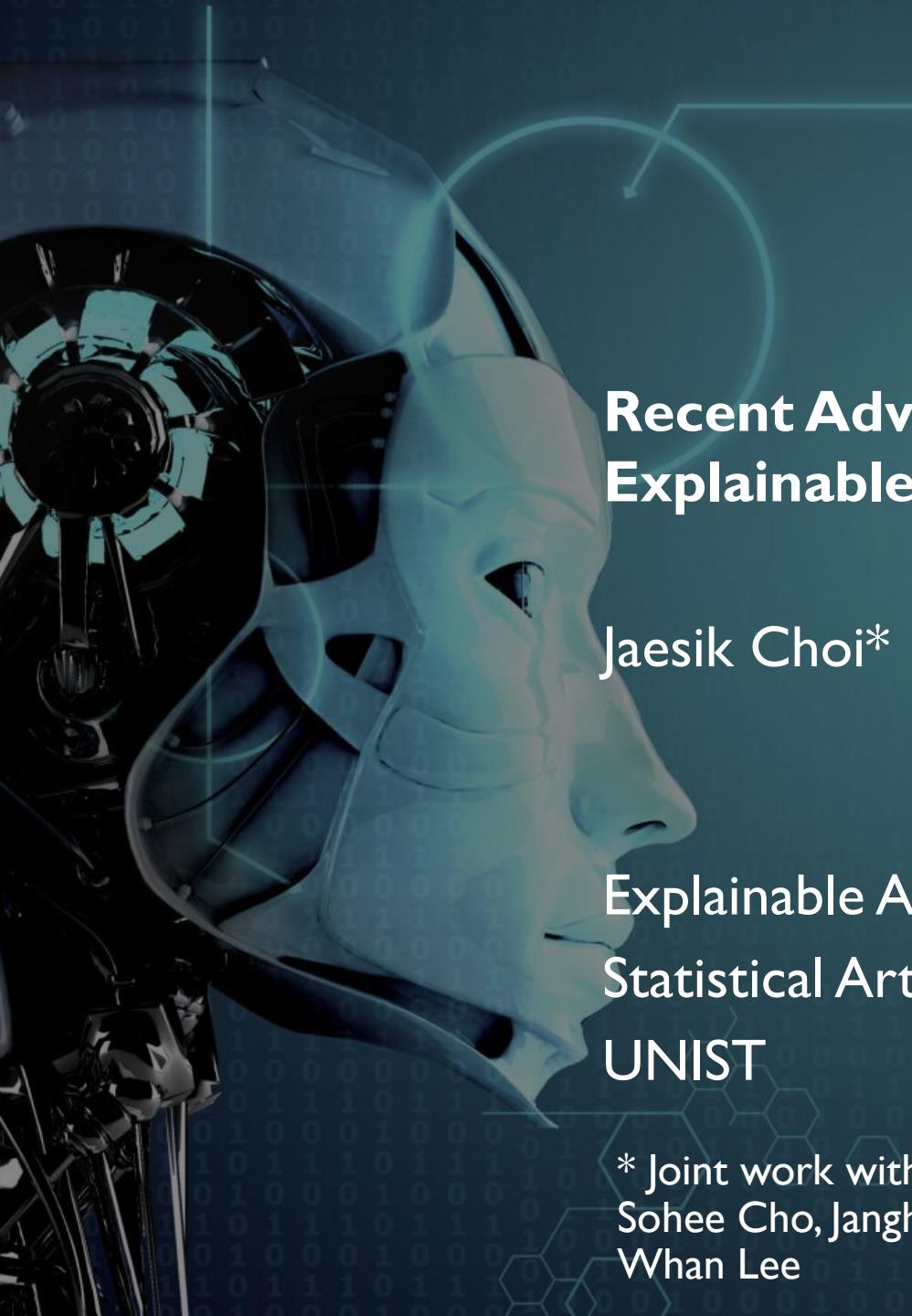
2.1 이미지 데이터 분류 문제

2.2 자연어 감정 분석

발표 3. 모델 불가지론적 방법으로 모델 해석

3.1 Shapley Additive exPlanations: SHAP

3.2 Red Wine Quality 예측 문제



Recent Advances in Explainable Artificial Intelligence

Jaesik Choi*

Explainable Artificial Intelligence Center (XAIC)
Statistical Artificial Intelligence Lab (SAIL)
UNIST

* Joint work with Anh Tong, Yunseong Hwang, Yeeun Chun,
Sohee Cho, Janghoon Ju, Sol-A Kim, Kyowoon Lee, Seong-
Whan Lee

최재식 교수(UNIST 전기전자컴퓨터공학과): 시계열 기계학습 전문가



▪ 주요 경력

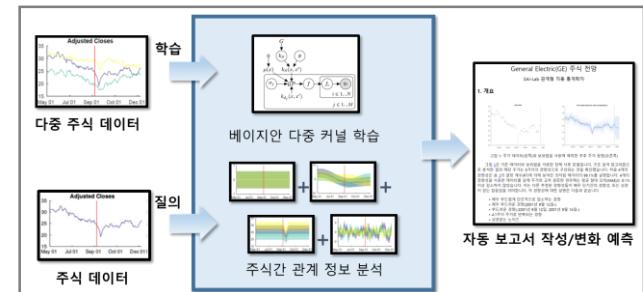
- UNIST 전기전자컴퓨터공학부 조교수(2013~2017), 부교수(2017~)
- 과학기술정보통신부 설명가능인공지능 연구센터, 센터장(2017~)[최연소]
- UNIST 산업인공지능 연구센터, 센터장(2018~)
- POSCO 철강 전문 교수(2017~2020)[최연소]
- LG전자 소재부품연구소, 자문교수(2017~2020)
- 삼성전자 종합기술원, 자문교수(2018~2019)
- 삼성 미래기술연구회(2019~)[최연소]
- Program Committee, ICML AAAI, IJCAI, UAI, NeurIPS(2012~)

▪ 주요 업적

- 세계 최초의 선형 시간 관계형 칼만 필터 기술(IJCAI-2011)
- 다중 시계열 데이터 자동 분석보고서 작성 시스템(ICML-2016, ICML-2018)
- 세계 최초의 딥러닝 기반 고로 노열 온도 예측 – 포스코 포함 제2고로 용선온도 제어 자동화 90%(2018)/국가핵심기술선정(총 69개, 2019)

▪ 수상 실적

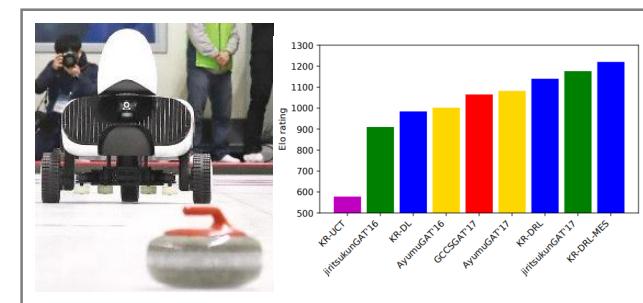
- 국무총리 표창(2019년 4월) [2019년 과학/정보통신의날 수상자중 최연소]
- 젊은 특훈교수(2018, 전임교원중 5%내외)
- 포스코 스마트 혁신상(2018 8월)
- Game AI Tournament, 제 4회 UEC-cup Digital Curling 대회 1위 (2018)
- Game Playing Workshop 주관 Digital Curling 대회 1위 (2017)
- IEEE International Symposium on Biomedical Imaging 주관 2016 영상 기반 피부암 검출 대회 3위
- International Conference on Big Data Intelligence and Computing 최우수논문상(2015)



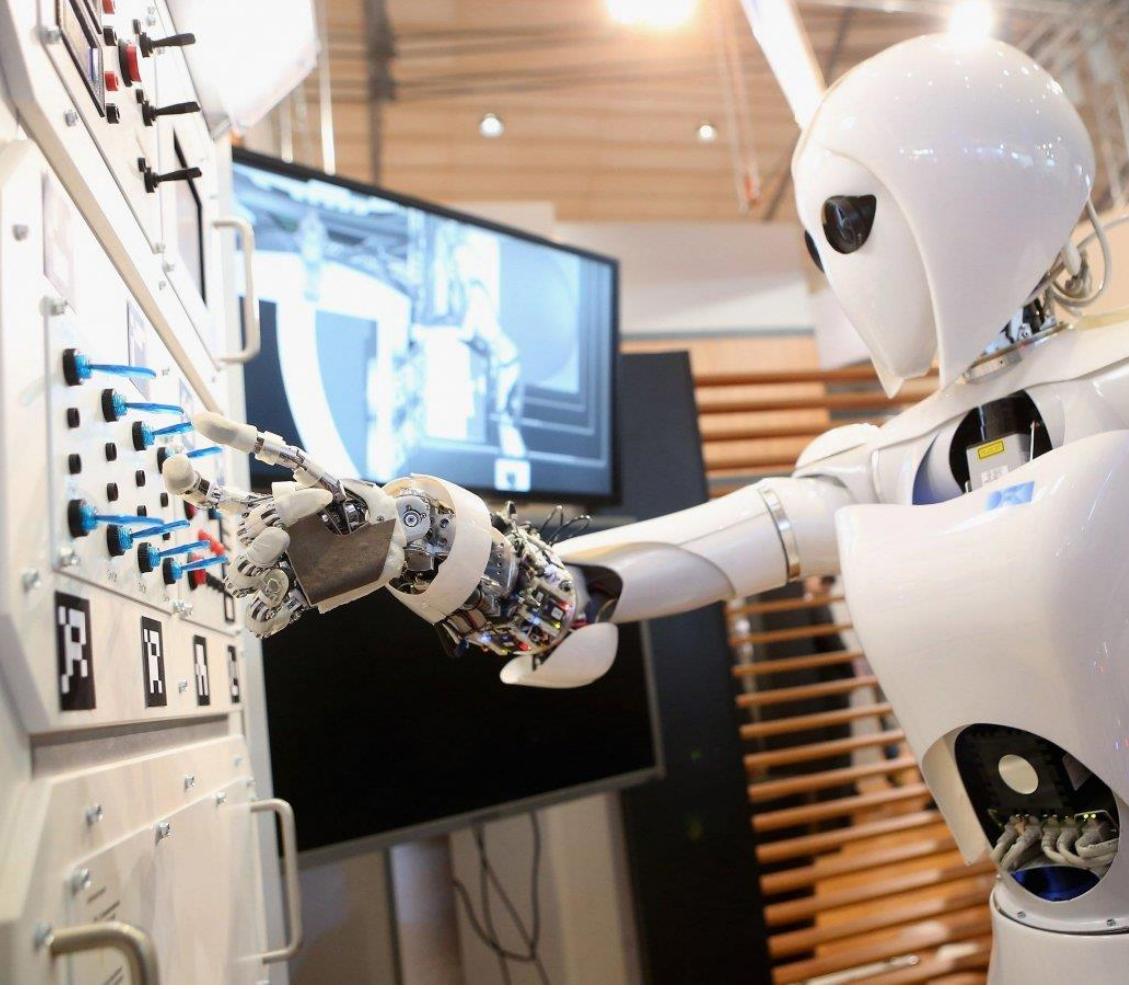
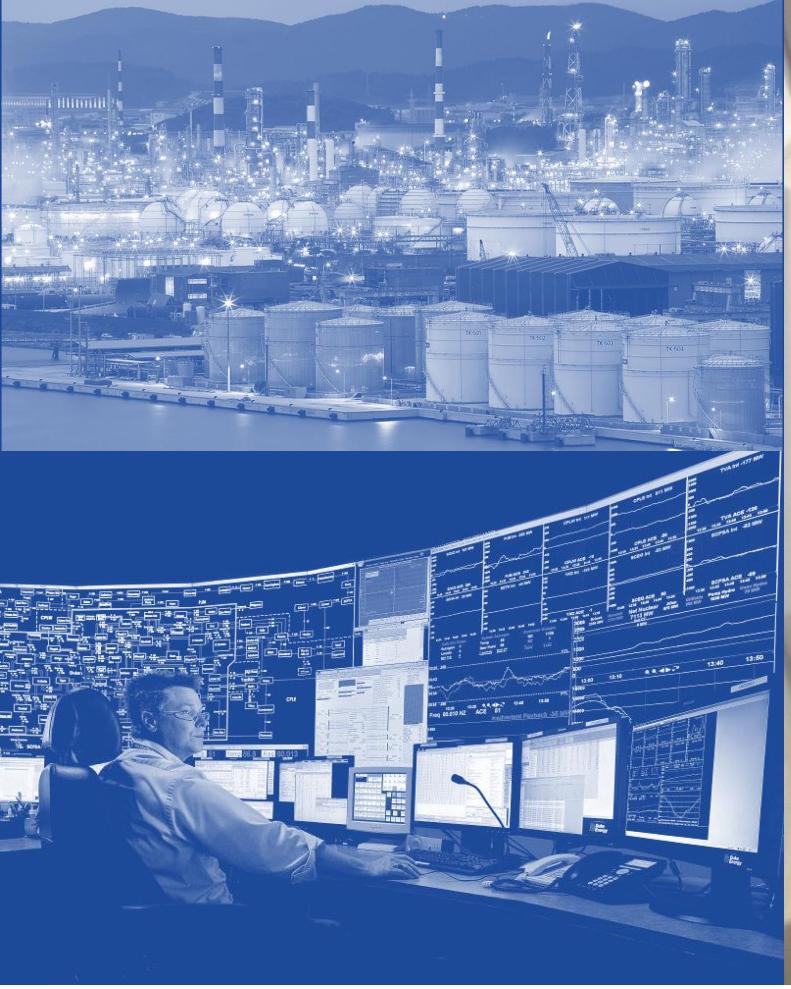
다중 시계열 데이터의 상관관계를 다중 커널 분석 기술로 학습해 변화를 예측하여 자동 보고서를 작성하는 기술 [ICML, 2016]



UNIST 개발 시계열 딥러닝모델(Group CNN) 용광로(고로) 자동조업 AI 적용 [POSCO 현장적용 90% 자동화, 2017]



UNIST 개발 강화학습(KL-DRL-UCT) 국제 디지털 컬링 경진대회 우승 [UEC 2017, GPW 2018, ICML 2018]



인공지능의 미래?



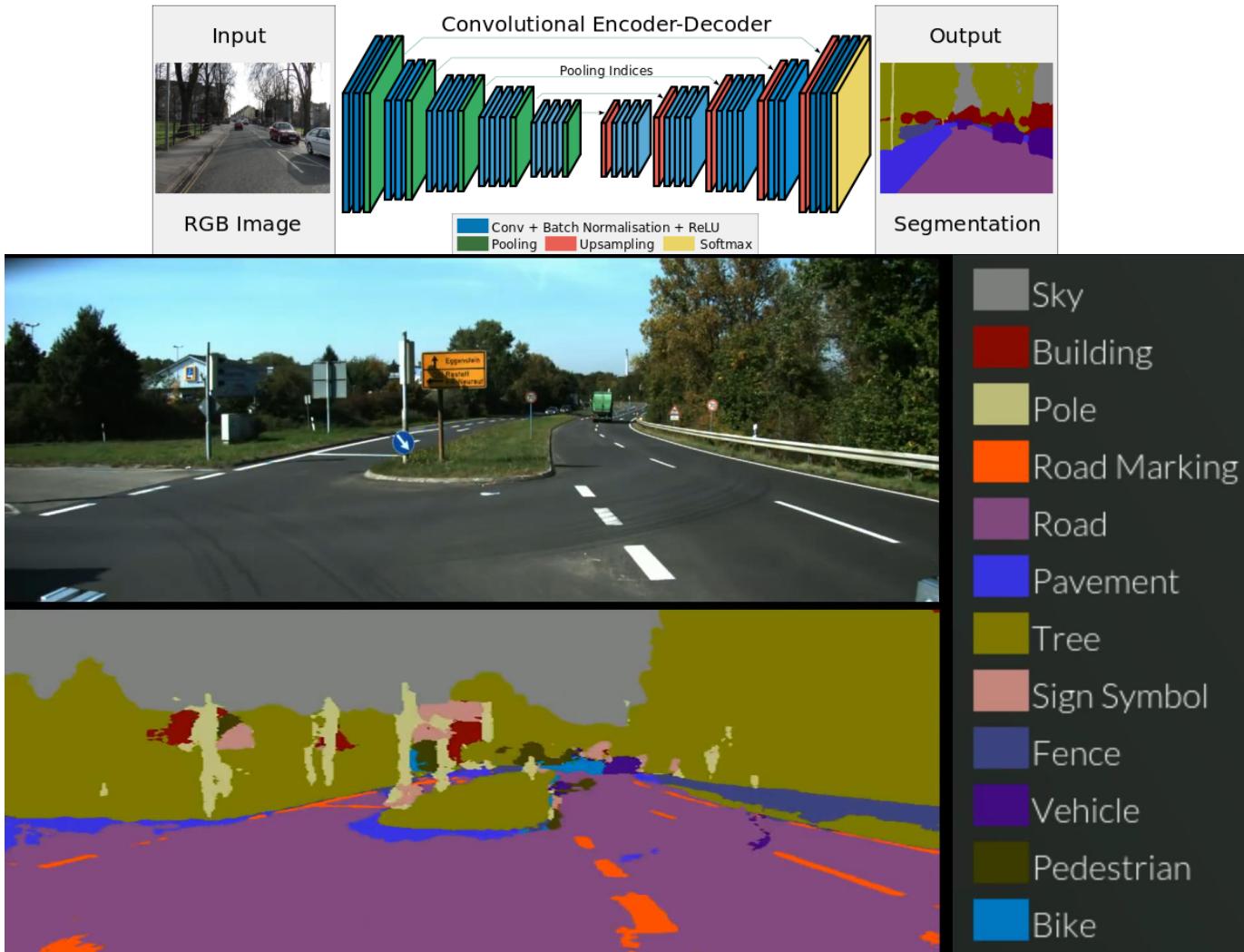
<https://www.youtube.com/watch?v=7a6GrKqOxeU>

DARPA Grand Challenge 2005



<https://www.youtube.com/watch?v=7a6GrKqOxeU>

Say Hello to Waymo 2016



<https://www.youtube.com/watch?v=CrnLINlbFA>

Semantic Segmentation by SegNet 2015

Pyramid Scene Parsing Network

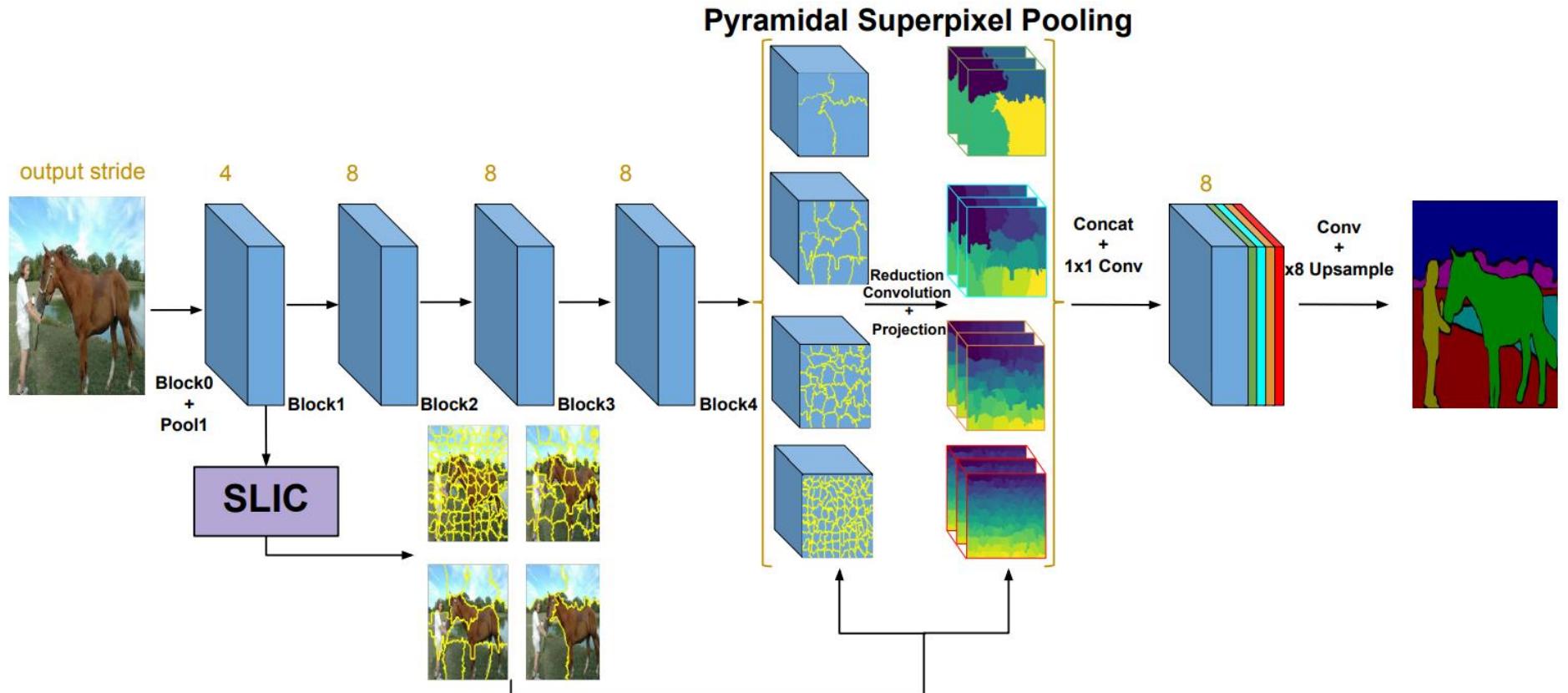
CVPR 2017

Hengshuang Zhao¹ Jianping Shi² Xiaojuan Qi¹ Xiaogang Wang¹ Jiaya Jia¹

¹The Chinese University of Hong Kong ²SenseTime Group Limited

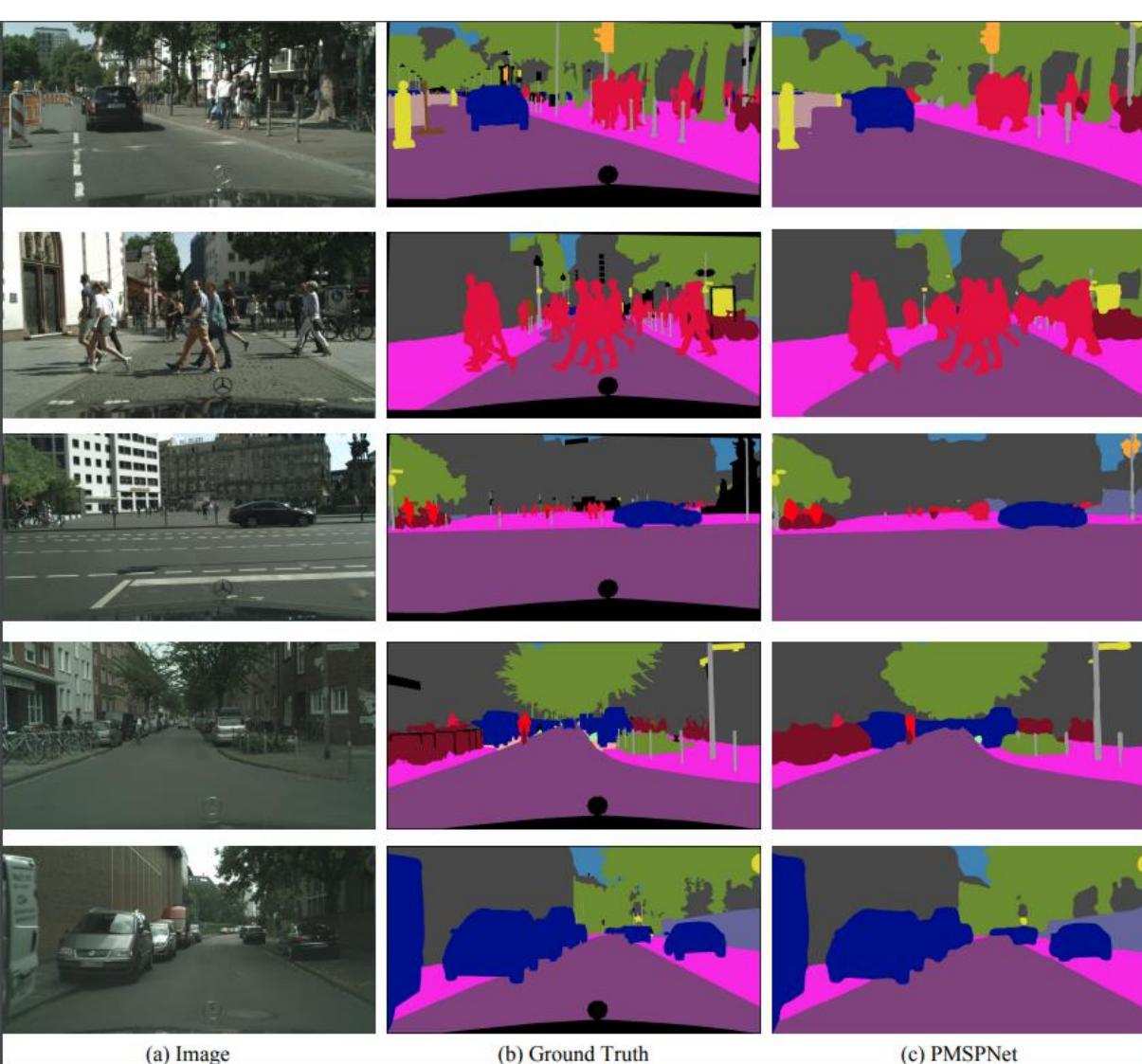
https://www.youtube.com/watch?v=gdAVqJn_J2M

**Semantic Segmentation
by Pyramid Scene Parsing Network 2017**



Ali Tousi and Jaesik Choi, Pyramidal Multiscale Superpixel Pooling Network for Semantic Segmentation, soon will be available at arxiv.

Pyramid Multiscale Superpixel Pooling Network 2019



Method	mIoU
DeepLab-v2 [3]	70.4
LC [14]	71.1
Adelaide_context [16]	71.6
FRRN [22]	71.8
RefineNet[15]	73.6
FoveaNet [13]	74.1
PEARL [9]	75.4
DUC_HDC [29]	77.6
SAC_multiple [37]	78.1
PSPNet [†] [38]	78.4
ResNet-38 [31]	78.5
SegModel [26]	78.5
Multitask Learning [10]	78.5
PSANet [40] [†]	78.6
PSANet [40] [‡]	80.1
PMSPNet [†] (ours)	78.6
PMSPNet [‡] (ours)	80.4

[†] Trained with *fine_train* set only

[‡] Trained with both *fine_train* and *fine_val* set

Pyramid Multiscale Superpixel Pooling Network 2019



우버 자율주행차 첫 보행자 사망사고

Uber's self-driving car killed a pedestrian (Marc 18th, 2018)
The 'safety driver' was watching a TV show (June 22th, 2018)

Do We Understand AI Systems Enough?

COMPAS: Prediction of Crime

Prior Offense	1 attempted burglary	1 resisting arrest without violence
COMPAS' decision		
DYLAN FUGETT		BERNARD PARKER
	LOW RISK	3
		HIGH RISK
		10
Subsequent Offenses	3 drug possessions	None

Do We Understand AI Systems Enough?



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

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

시사점

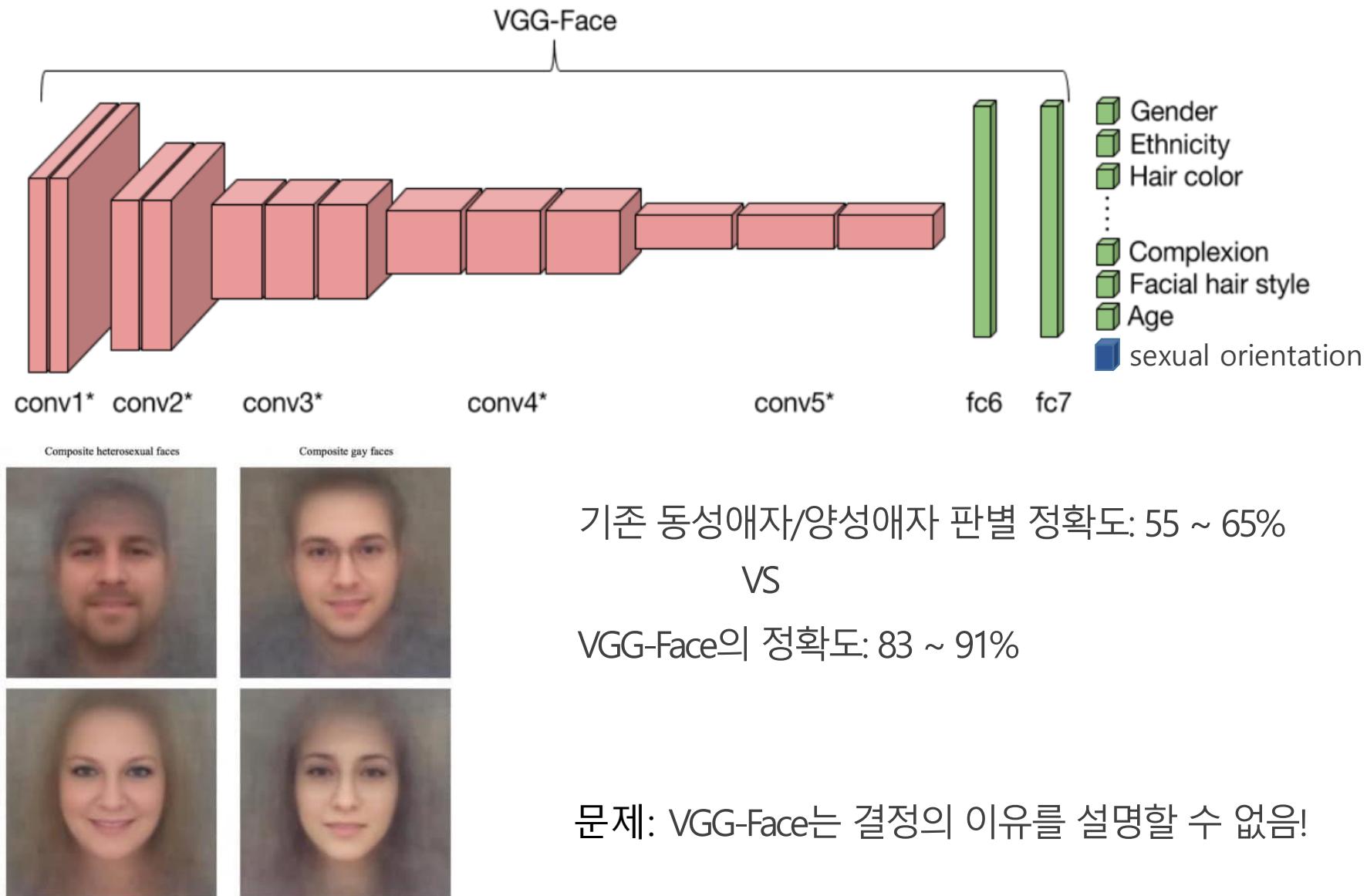
인공지능 기술을 주요 의사결정에 적용하기 위해서 법적으로 이유 제시 기능이 필요함

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

EU의 일반정보보호규정(General Data Protection Regulation)

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

사례 I - 사람의 성적 지향을 판별하는 딥러닝?



DESCRIBE

Handcrafted Knowledge



CATEGORIZE

Statistical Learning

EXPLAIN

Contextual Adaptation

Statistically impressive, but individually unreliable



Inherent flaws can be exploited



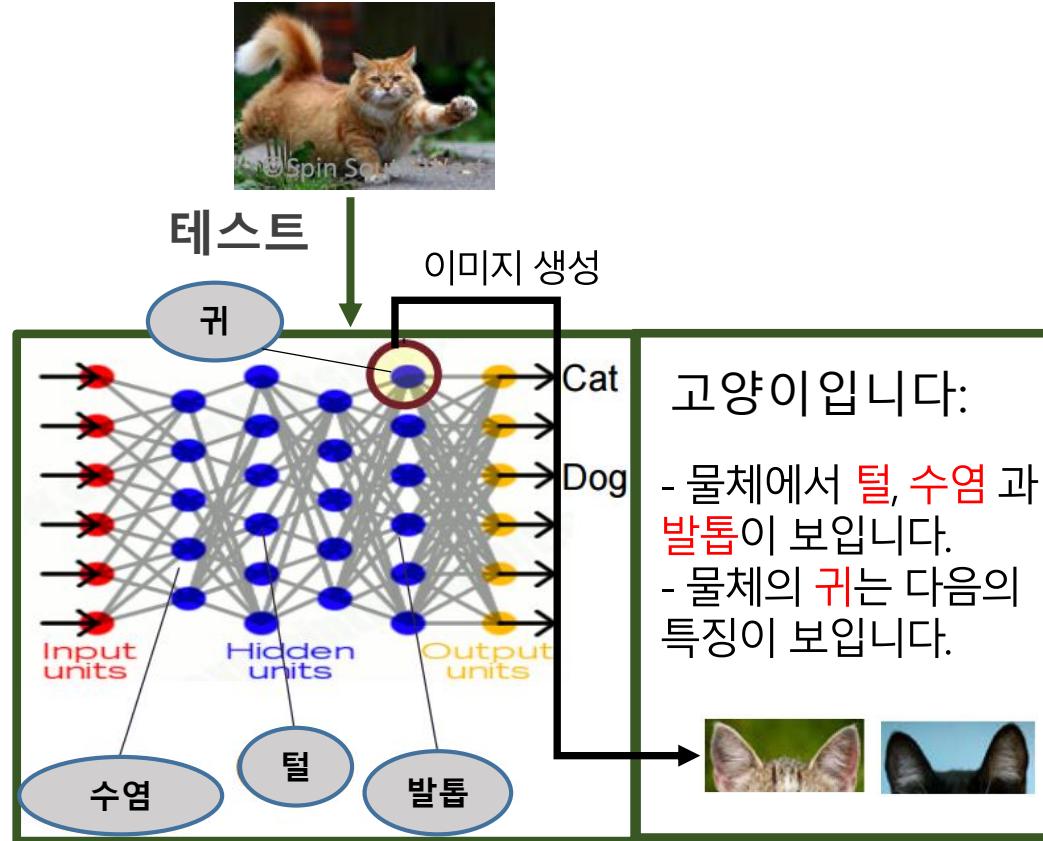
Skewed training data creates Maladaptation

A DARPA Perspective on AI – Three Waves of AI

해결방법 - 설명을 제공하는 인공지능 시스템



학습
과정



<설명 가능한 딥러닝 모델의 예>

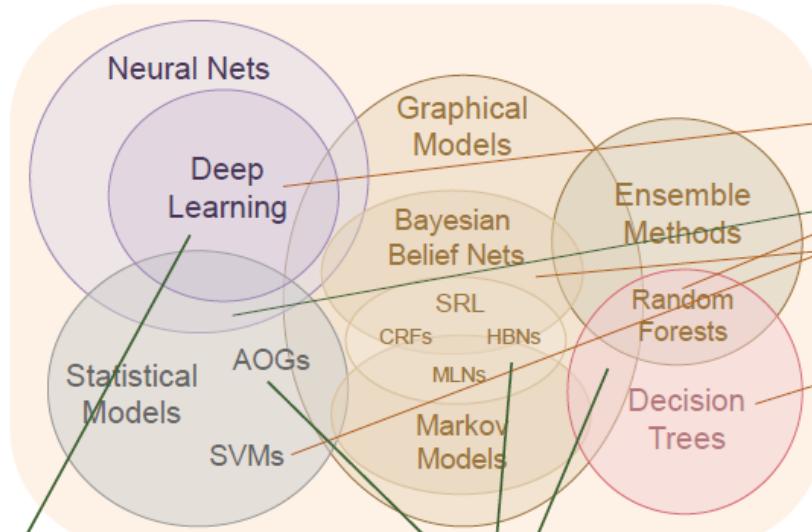


Explainable AI – Performance vs. Explainability

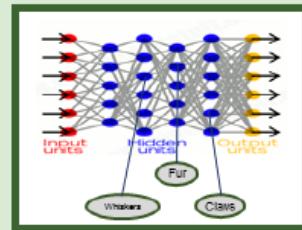
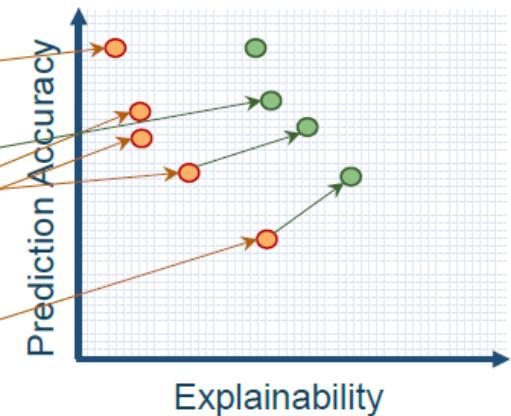
New Approach

Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)

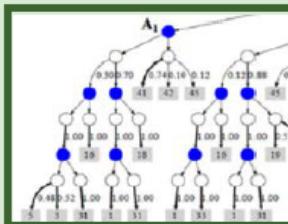


Explainability (notional)



Deep Explanation

Modified deep learning techniques to learn explainable features



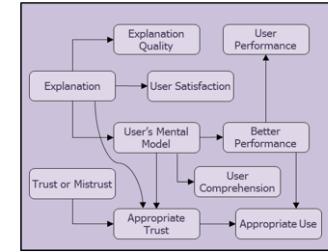
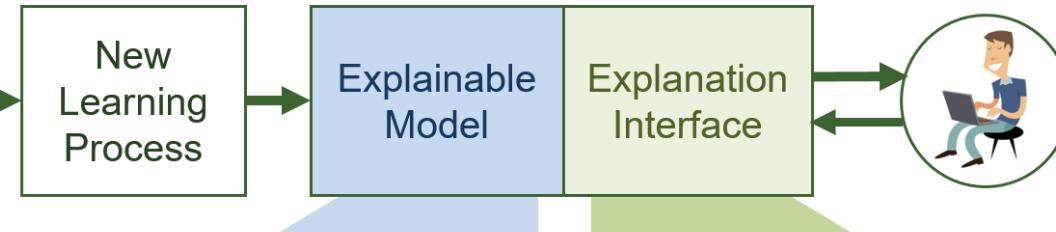
Interpretable Models

Techniques to learn more structured, interpretable, causal models

D. Gunning (DARPA), 2016년



XAI Performers & Technical Approaches

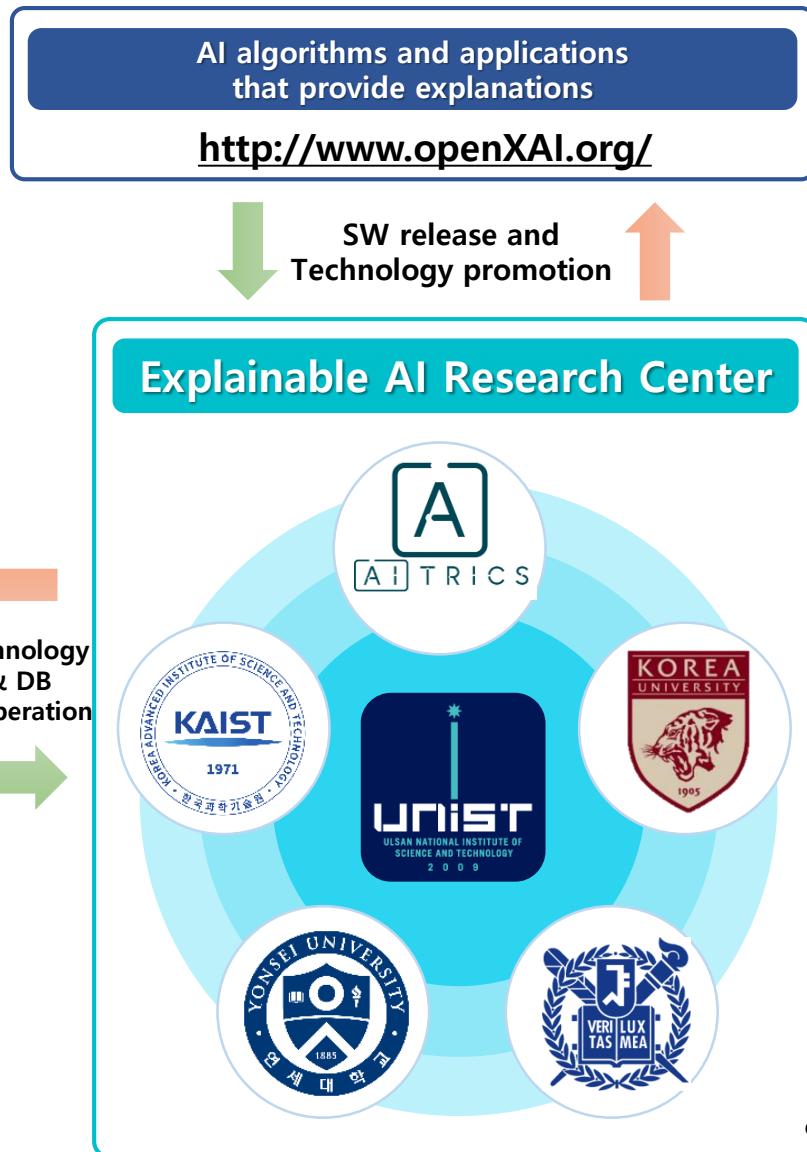


IHMC
Psychological Models
of Explanation

CP	Performer	Explainable Model	Explanation Interface
Both	UC Berkeley	Deep Learning	Reflexive and Rational
	Charles River	Causal Modeling	Narrative Generation
	UCLA	Pattern Theory+	3-level Explanation
Autonomy	Oregon State	Adaptive Programs	Acceptance Testing
	PARC	Cognitive Modeling	Interactive Training
	CMU	Explainable RL (XRL)	XRL Interaction
Analytics	SRI International	Deep Learning	Show and Tell Explanation
	Raytheon BBN	Deep Learning	Argumentation and Pedagogy
	UT Dallas	Probabilistic Logic	Decision Diagrams
	Texas A&M	Mimic Learning	Interactive Visualization
	Rutgers	Model Induction	Bayesian Teaching

UNIST Explainable AI Research Center

Partners



Technology Sharing
→ Competition



Medical application



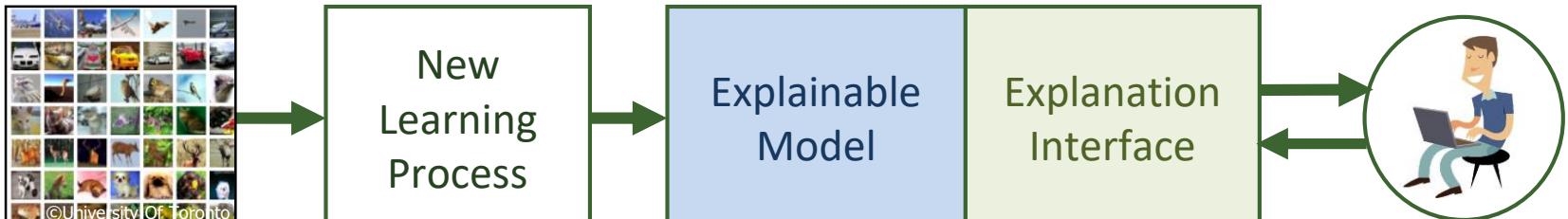
Financial application



Enterprises with intention of technology transfer



XAI Performers & Technical Approaches



Task	Performer	Explainable Model	Explanation Interface	Application
Task 1	KAIST	Deep Learning (RNN)	Explaining Sparse Features	Healthcare
Task 2	KAIST	Medical Imaging	Explainable Computer Aided Diagnosis (XCAD)	Healthcare
	SNU	Visual Q/A	Visual Explanation	Visual Dialog
	Yonsei/ KAIST	Textual Deep Learning	Textual Explanation	Visual Dialog
Task 3	UNIST	Nonparametric Bayesian	Narrative Explanation	Finance
	Korea University	Deep Learning (attribution methods)	Visual Explanation	Healthcare
	KAIST/ UNIST	Explainable RL	Interactive Visualization	Game

Technical Approaches

Explaining Deep Neural Networks

Explaining Deep Reinforcement Learning

Explaining by Combining Explainable Models

Finding Local Explanations (Model Agnostic Methods)

Data Sets and Applications

Technical Approaches

Explaining Deep Neural Networks

Input Attribution Method (TU Berlin)

XCAD: Explainable Computer Aided Diagnosis (KAIST)

Dissecting Deep Neural Networks (MIT)

Generating Examples inside of Deep Neural Networks (UNIST)

Explaining Deep Reinforcement Learning

Explaining by Combining Explainable Models

Finding Local Explanations (Model Agnostic Methods)

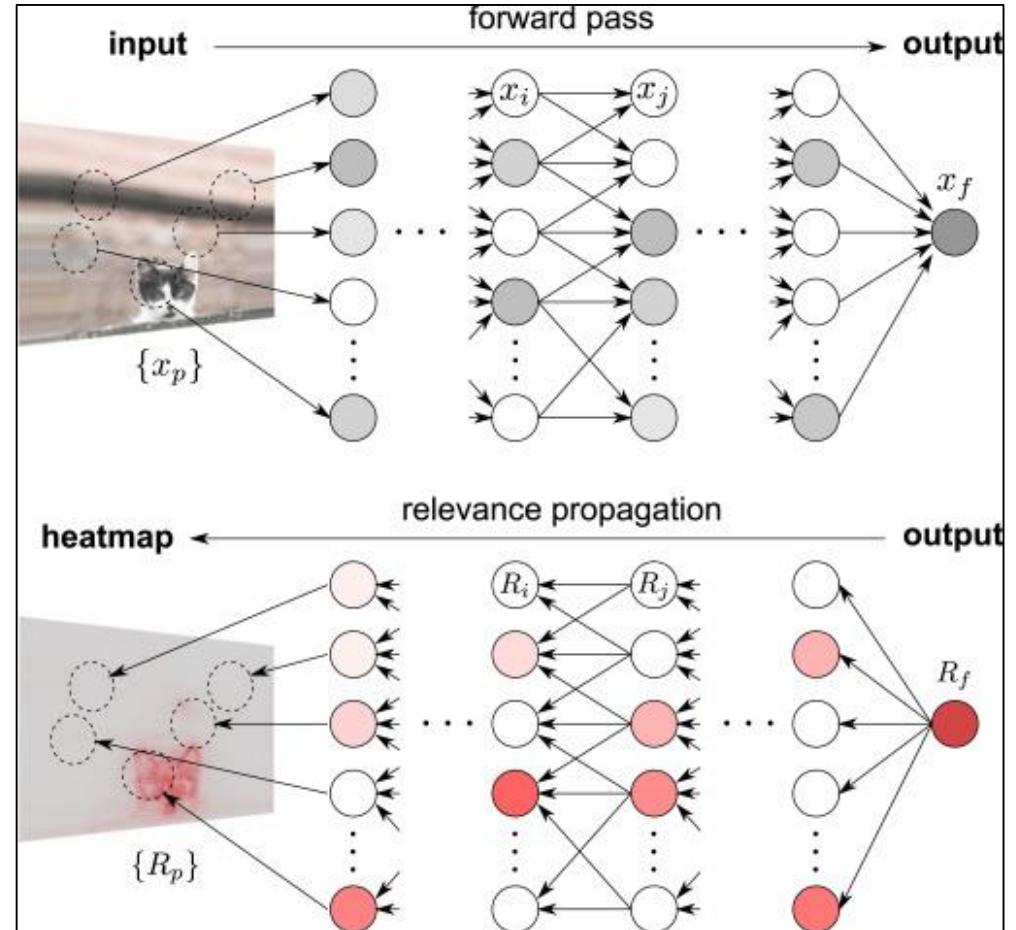
Data Sets and Applications

Input Attribution Methods

Layer-wise Relevance Propagation(LRP)



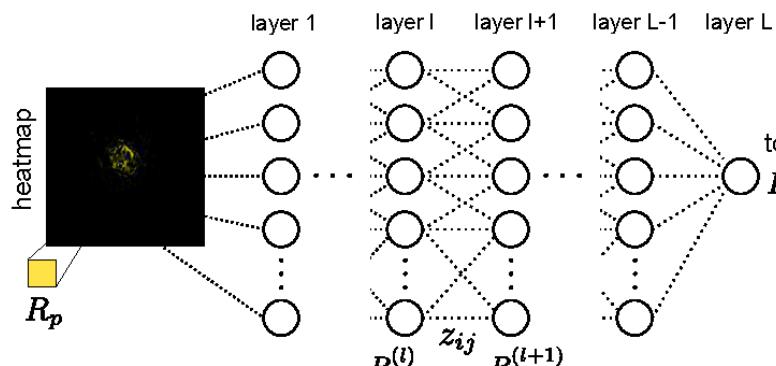
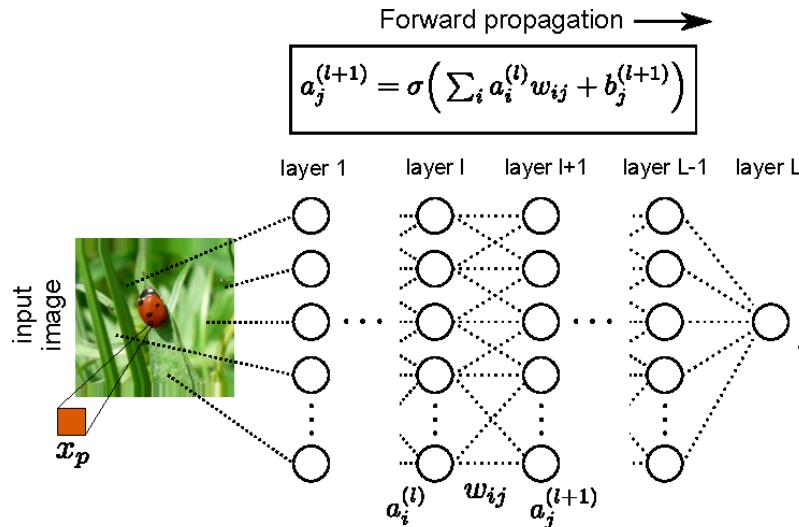
딥러닝 모델의 예측 결과에 대해 시각적 해석을 돋는 공개 SW를 Heatmapping.org에 공개



딥러닝 예측 결과 이유 설명을 위한 LRP 알고리즘 [CVPR, 2015]

Input Attribution Methods

Layer-wise Relevance Propagation (LRP)



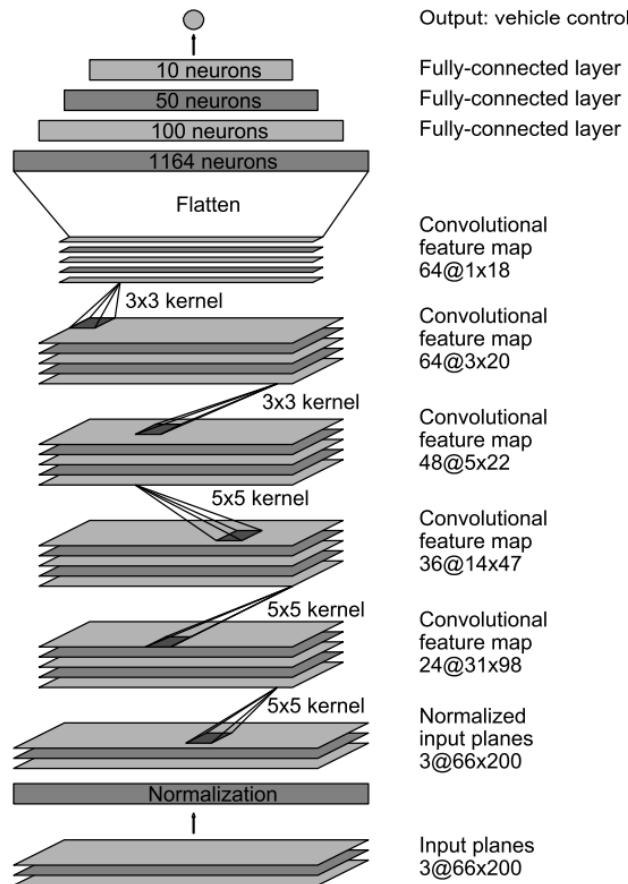
$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_{i'} z_{i'j}} R_j^{(l+1)}$$

← Layer-wise relevance propagation

$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_{i'} z_{i'j}} R_j^{(l+1)}$$

Input Attribution Methods

NVIDIA의 PilotNet(자율주행 딥러닝)의 결정을 설명하는 딥러닝



PilotNet 구조도



자율주행의 이유를 설명해주는 딥러닝 기술

Input Attribution Methods

Saliency Maps

Simonyan et al. 2015

Integrated Gradients

Sundararajan et al. 2017

DeepLIFT

Shrikumar et al. 2017

LIME

Ribeiro et al. 2016

Gradient * Input

Shrikumar et al. 2016

Layer-wise Relevance Propagation (LRP)

Bach et al. 2015

Guided Backpropagation

Springenberg et al. 2014

Grad-CAM

Selvaraju et al. 2016

Simple occlusion

Zeiler et al. 2014

Meaningful Perturbation

Fong et al. 2017

Prediction Difference Analysis

Zintgraf et al. 2017

KernelSHAP/DeepSHAP

Lundberg et al., 2017

Slides courtesy of [Marco Ancona, et. al., Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Values Approximation, ICML 2019]

Input Attribution Methods

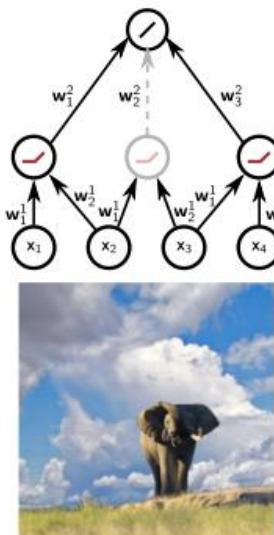
PatternNet – Learning to Explain Neural Net – 1/2



Klaus-Robert Müller (TU Berlin/Korea University)

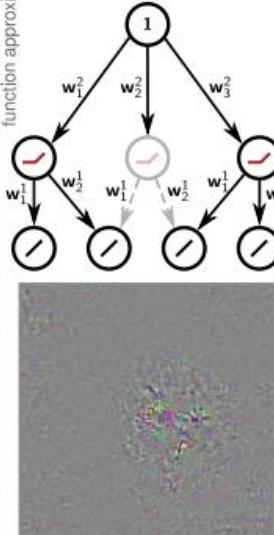
Learning (Optimizing) Attributions to Reduce Signal Noise

Forward pass



Example
VGG-16 classification

Gradient (Baehrens et al, Simonyan et al)

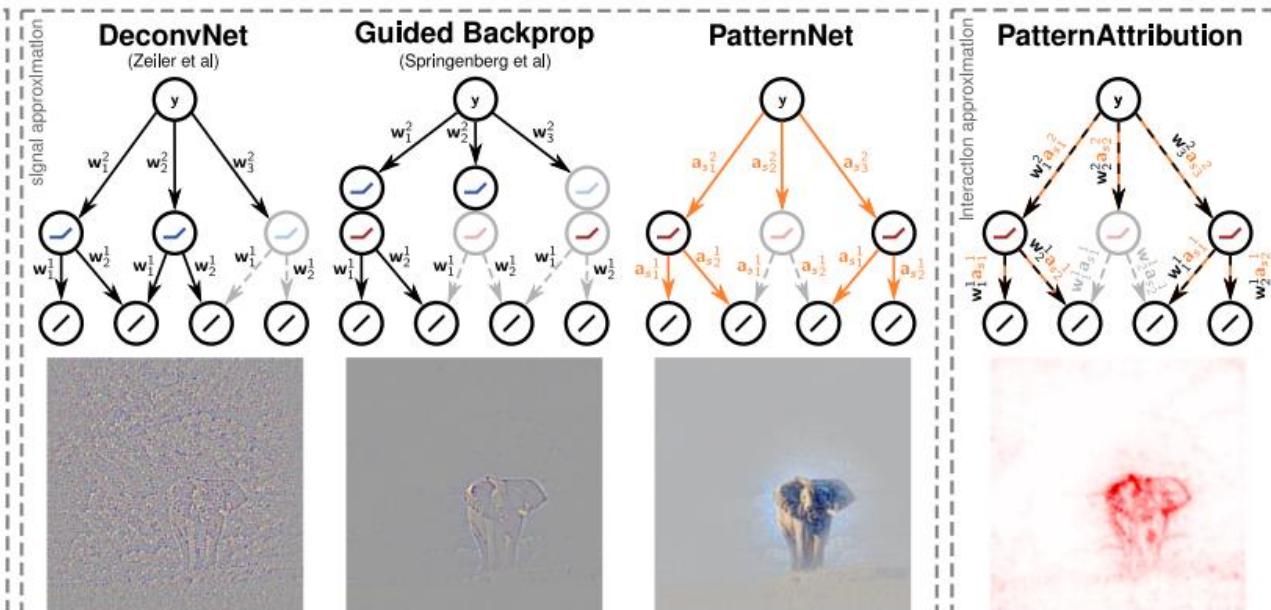


DeconvNet (Zeiler et al)

Guided Backprop (Springenberg et al)

PatternNet

PatternAttribution



Forward ReLU
(activated)

Forward ReLU
(deactivated)

Backward ReLU
(activated)

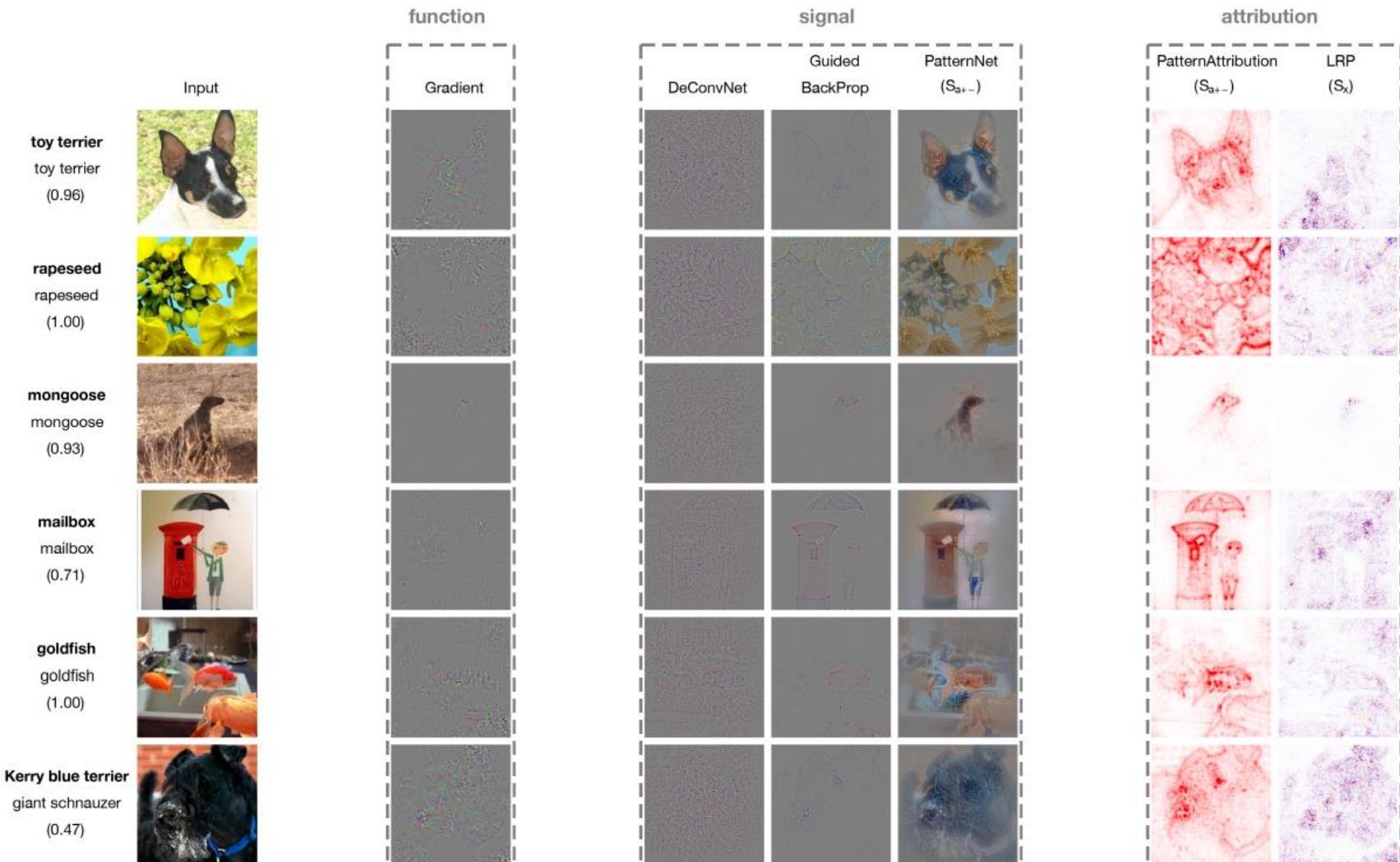
Backward ReLU
(deactivated)

Linear Neuron

Input Attribution Methods

PatternNet – Learning to Explain Neural Net – 2/2

Klaus-Robert Müller (TU Berlin/Korea University)

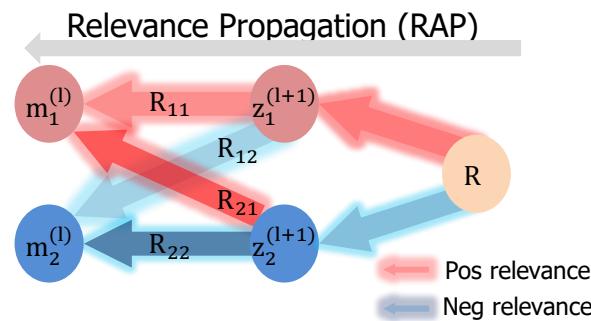
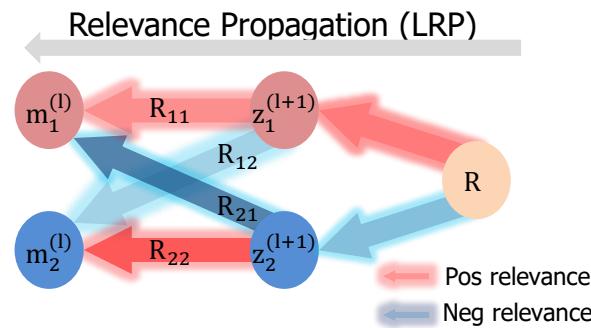
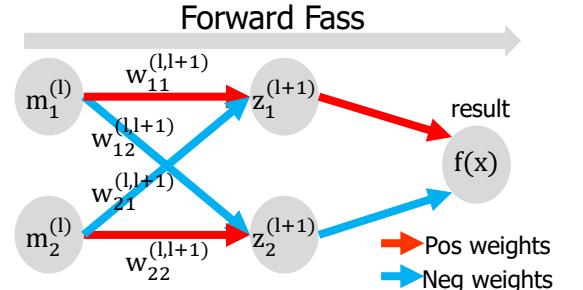


Pieter-Jan Kindermans, Kristof T. Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne, "Learning How to Explain Neural Networks: PatternNet and PatternAttribution", ICLR 2018

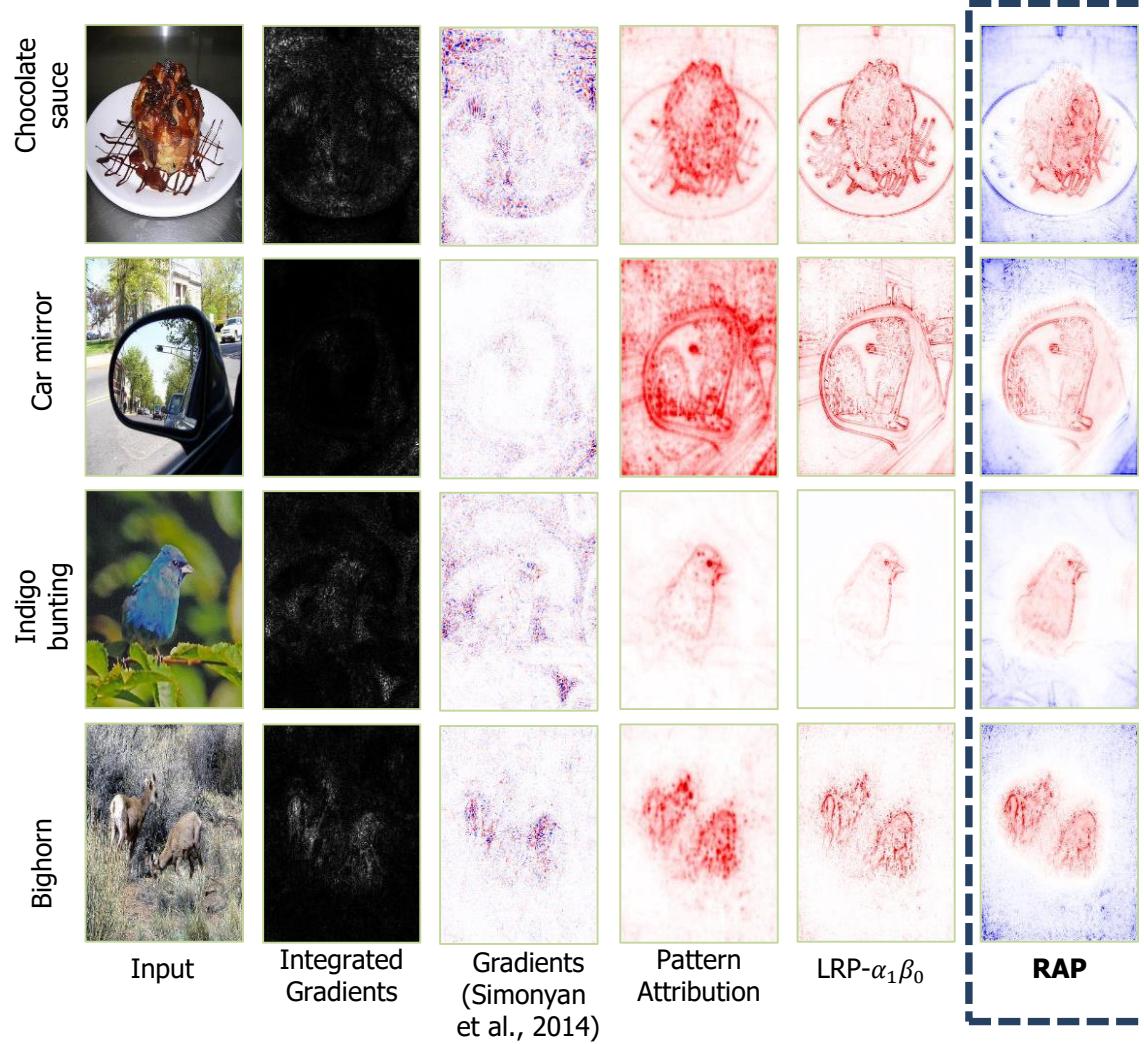
Input Attribution Methods

RAP - Relative Attribution Propagation

Seong-Whan Lee (Korea University)
Jaesik Choi (UNIST)



Propagating flipped attributions of negative relevance

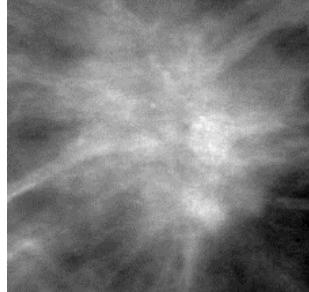
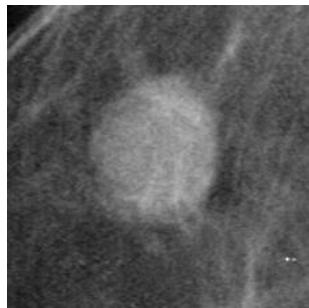
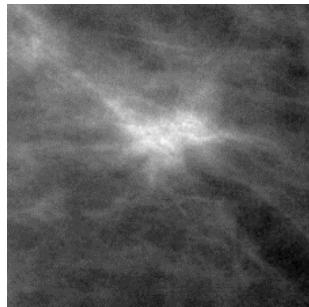


XCAD: Explainable Computer Aided Diagnosis – I/2

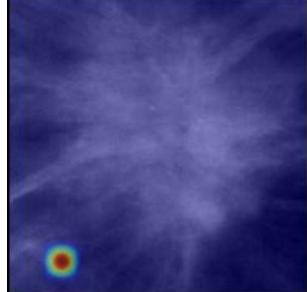
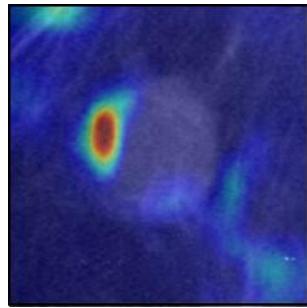
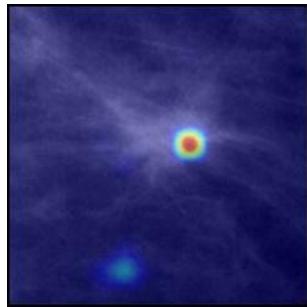


Yong Man Ro (KAIST)

Test image



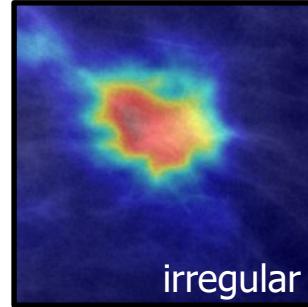
grad-CAM



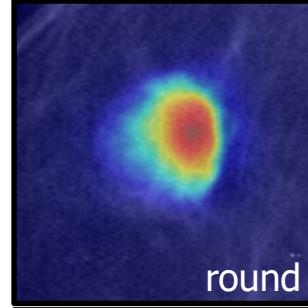
Difficult for medical doctors

Explanations of XCAD guided by BI-RADS*

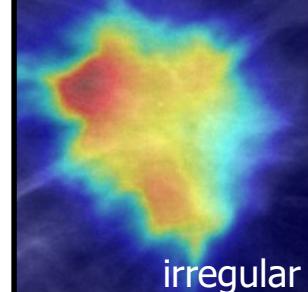
Shape



irregular

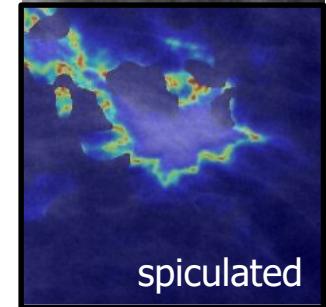


round

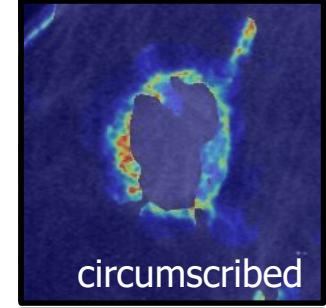


irregular

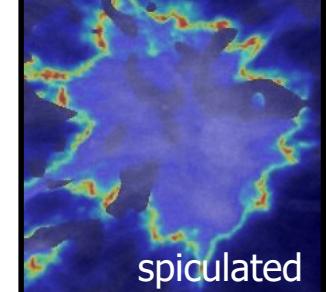
Margin



spiculated



circumscribed



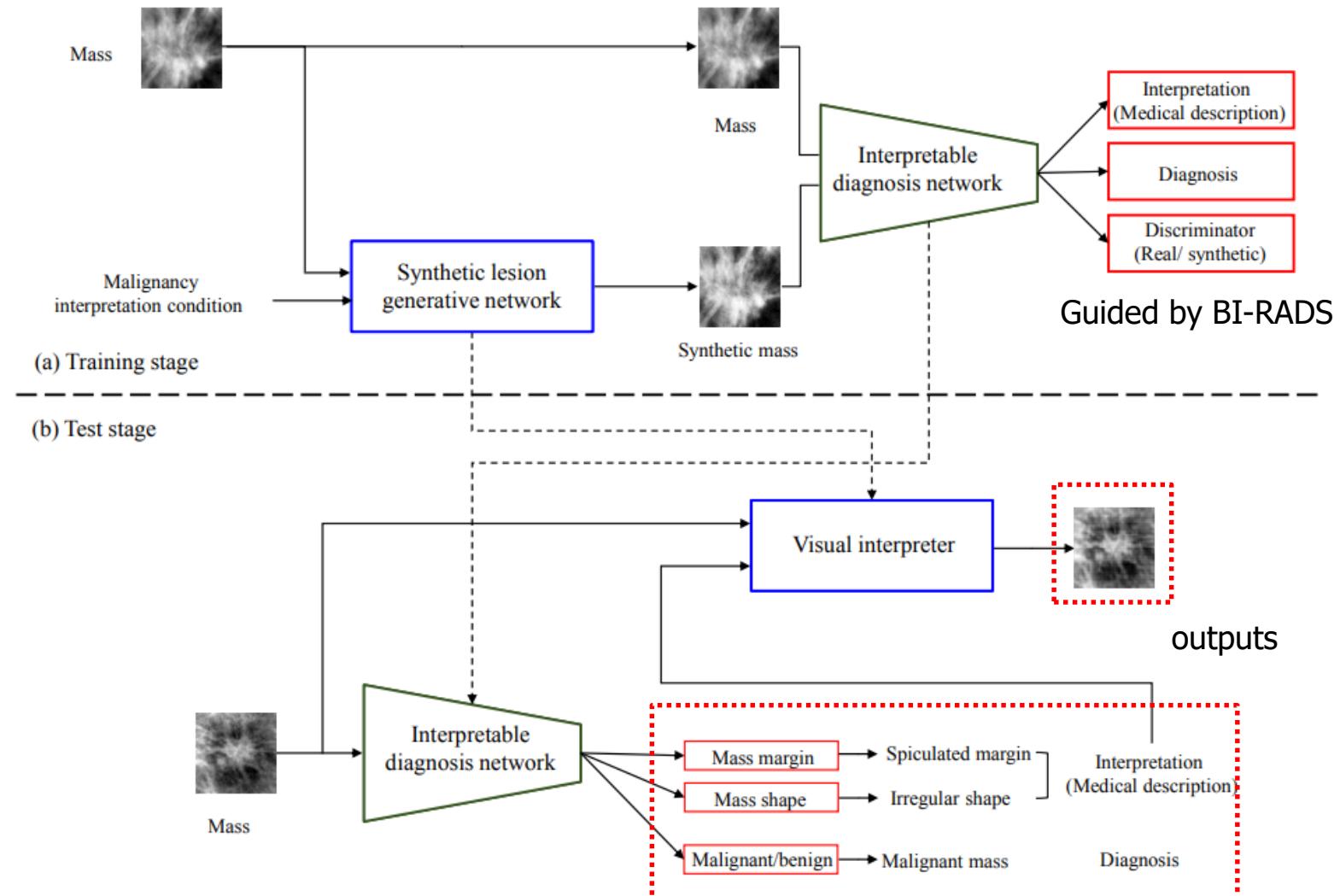
spiculated

Explanations that are familiar to doctors

*BI-RADS: Breast Imaging Reporting and Data System

XCAD: Explainable Computer Aided Diagnosis – 2/2

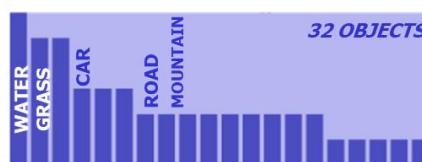
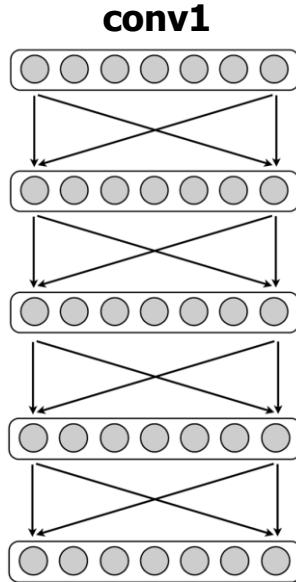
Yong Man Ro (KAIST)



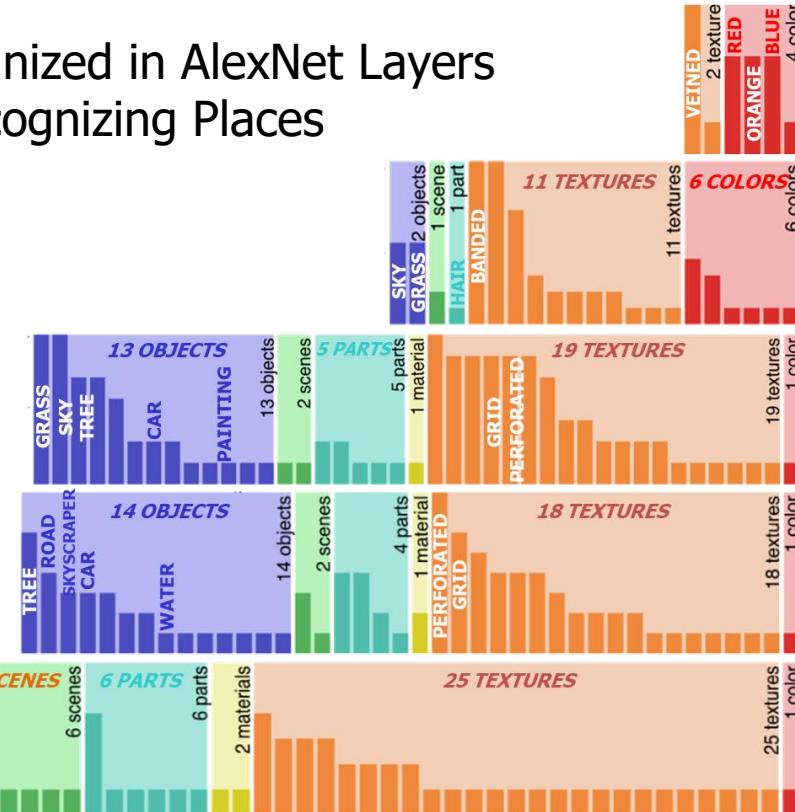
Dissecting Deep Neural Networks

Network Dissection

David Bau et. al., 2017 (A. Torralba, MIT)



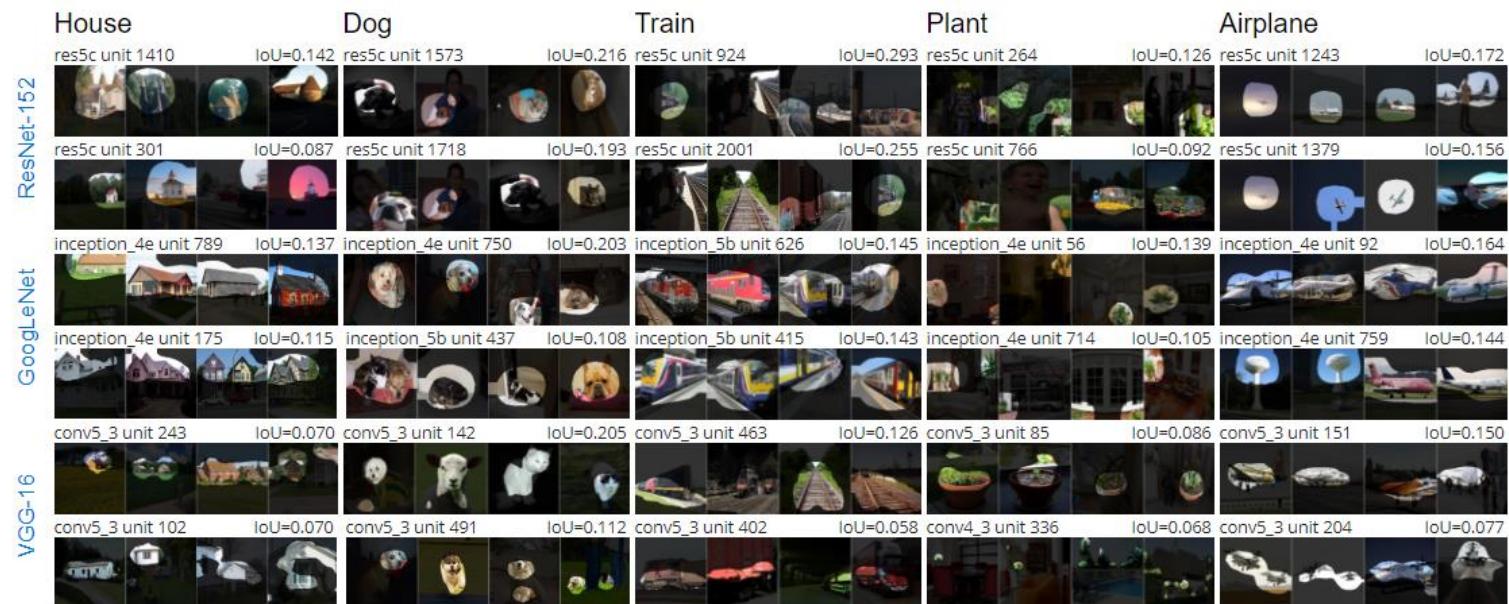
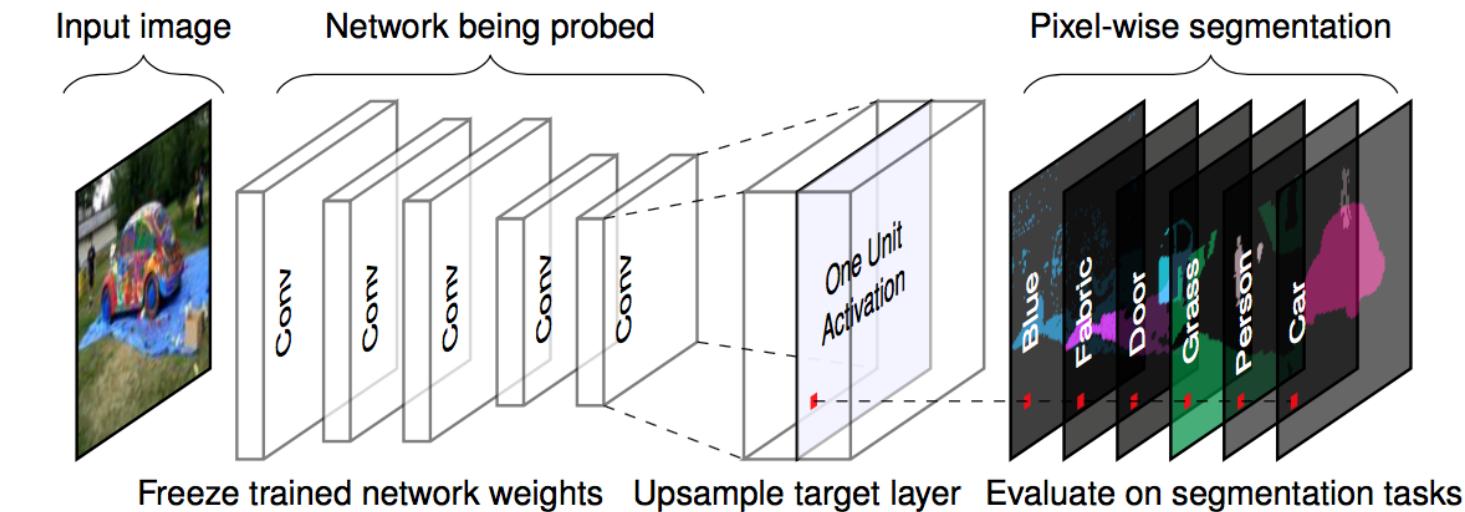
Concepts Recognized in AlexNet Layers
for Recognizing Places



Dissecting Deep Neural Networks

Network Dissection

David Bau et. al., 2017 (A. Torralba, MIT)



Dissecting Deep Neural Networks

Network Dissection

David Bau et. al., 2017 (A. Torralba, MIT)

Buildings

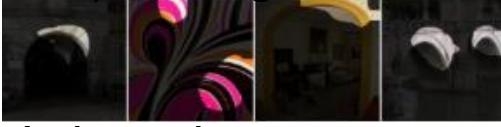
56) building



8) bridge

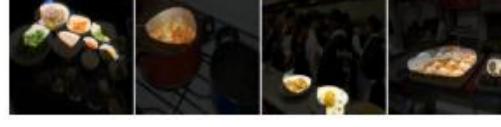


123) building



Indoor objects

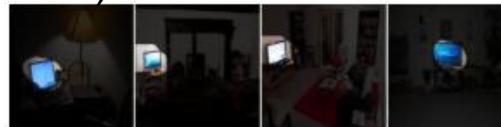
182) food



46) painting

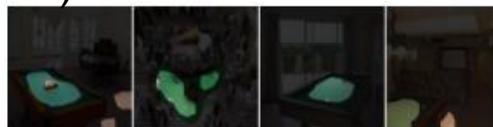


106) screen

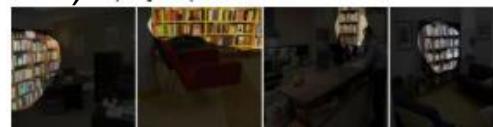


Furniture

18) Billiard table



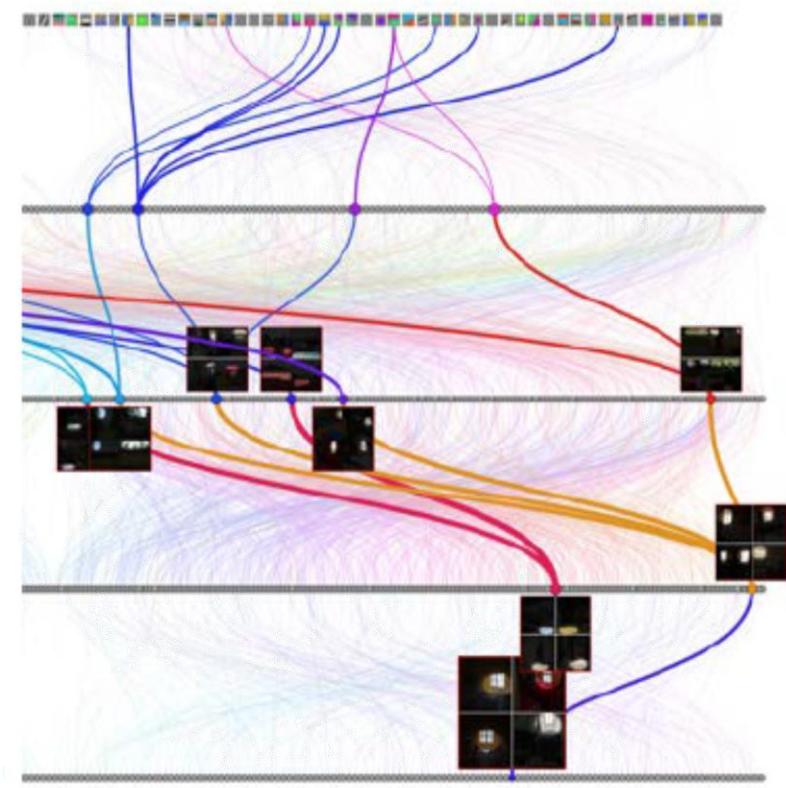
155) bookcase



116) bed



Audit trail: for a particular output unit, the drawing shows the most strongly activated path



Interpretation of several units in pool5 of AlexNet trained for place recognition

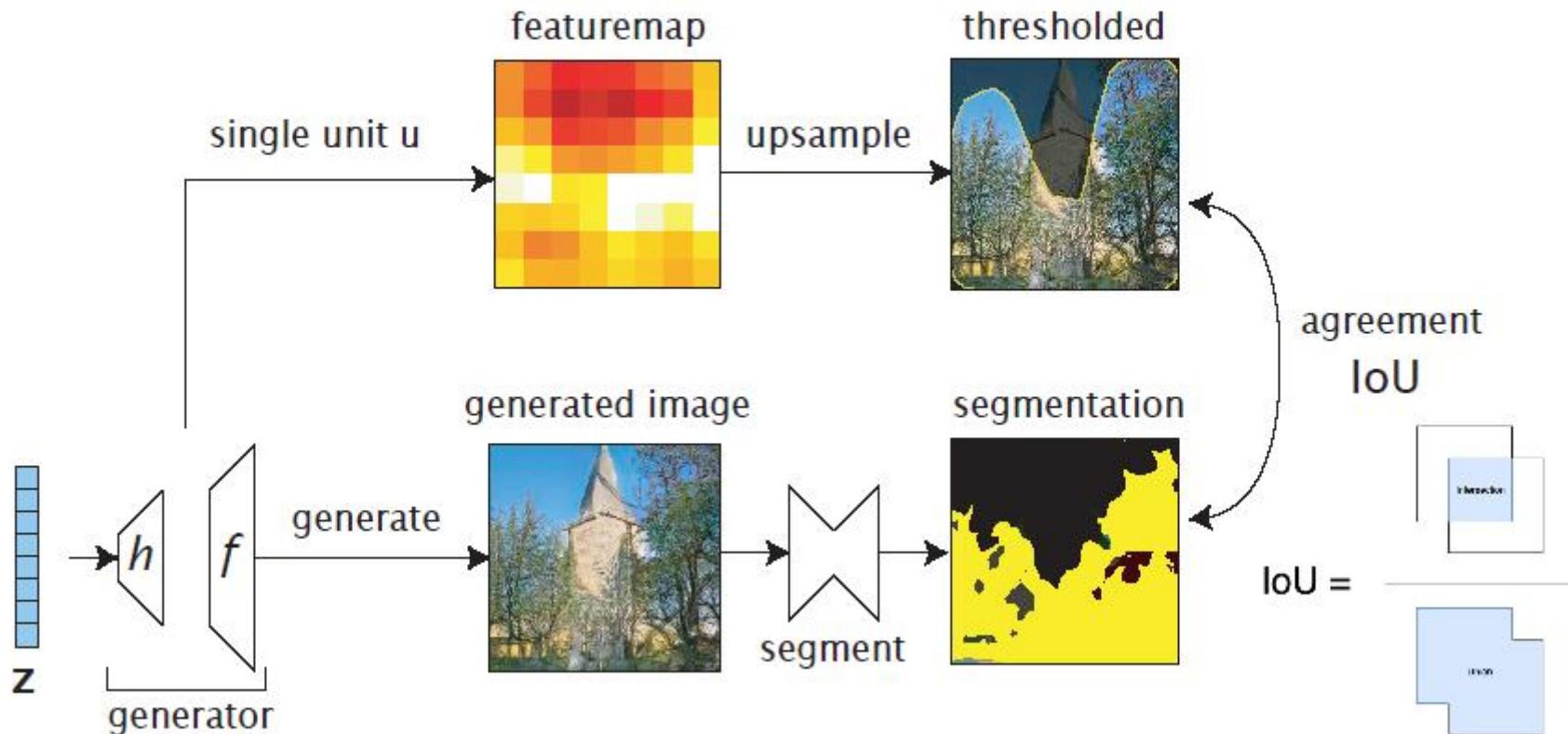
A. Torralba (MIT) et. al., 2017

Dissecting Deep Neural Networks

GAN Dissection

David Bau et. al., 2019 (A. Torralba, MIT)

Dissecting explainable units in a GAN



Dissecting Deep Neural Networks

GAN Dissection

David Bau et. al., 2019 (A. Torralba, MIT)

Do units correlate to an object class?

Church samples



Unit #119

Tree



Unit #32

Dome



Dissecting Deep Neural Networks

GAN Dissection

David Bau et. al., 2019 (A. Torralba, MIT)

Do units correlate to an object class?

Dining room samples



Unit #139
Window



Unit #65
Table



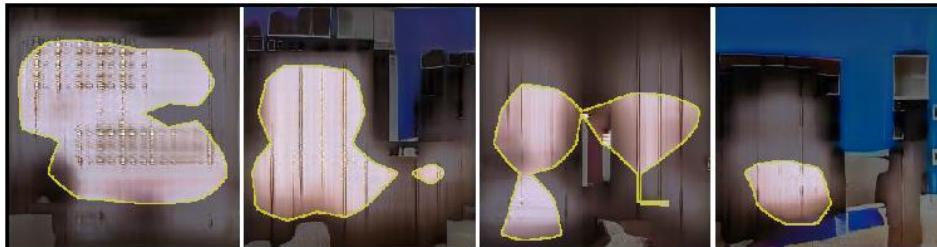
Dissecting Deep Neural Networks

GAN Dissection

David Bau et. al., 2019 (A. Torralba, MIT)

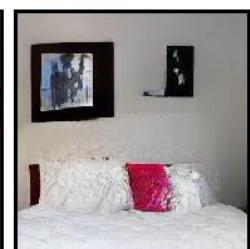
Debugging and Improving GANs

Unit #63



Bedroom images with artifacts

Unit #231



Example artifact-causing units

Ablating “artifact” units improves results

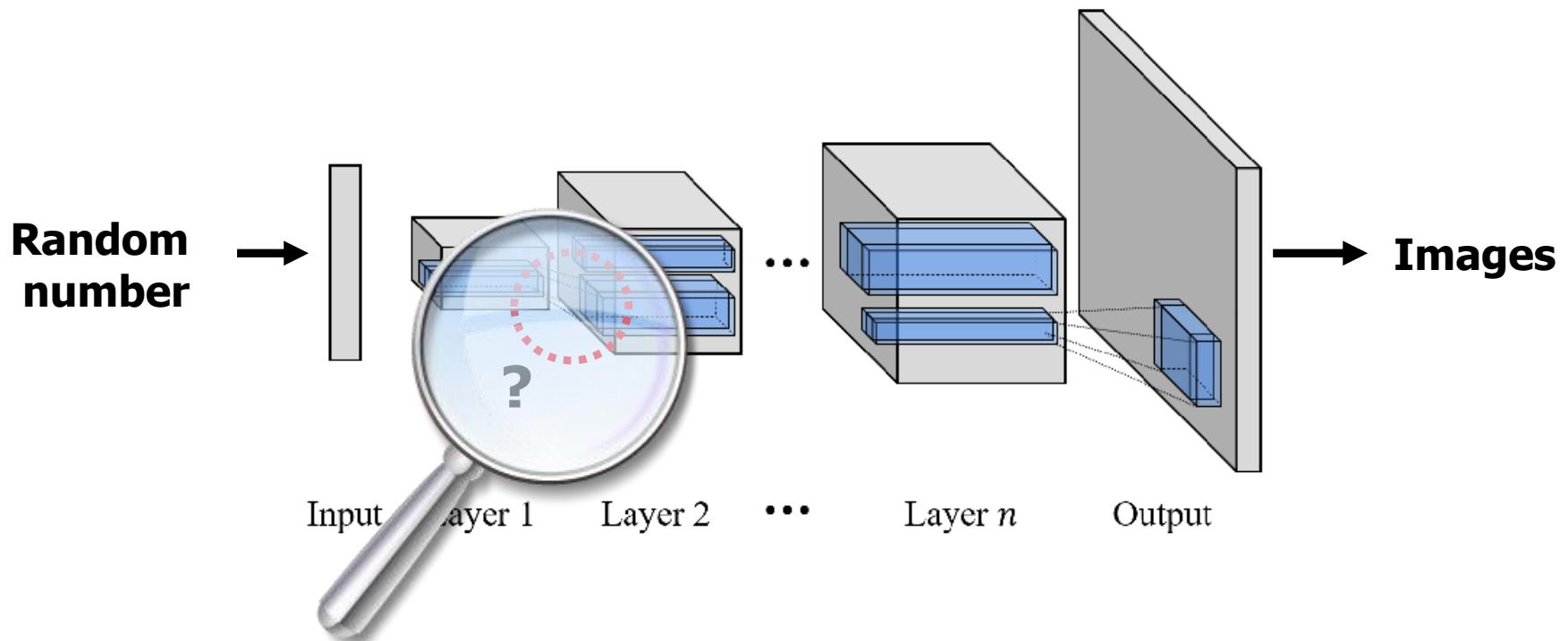
Dissecting Deep Neural Networks

Generative Boundary Aware Sampling

Giyoung Jeon et. al., 2019 (J. Choi, UNIST)



A Generative Neural Network (GNN)



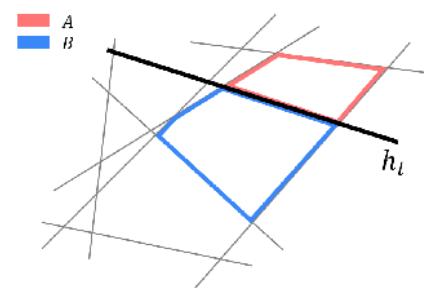
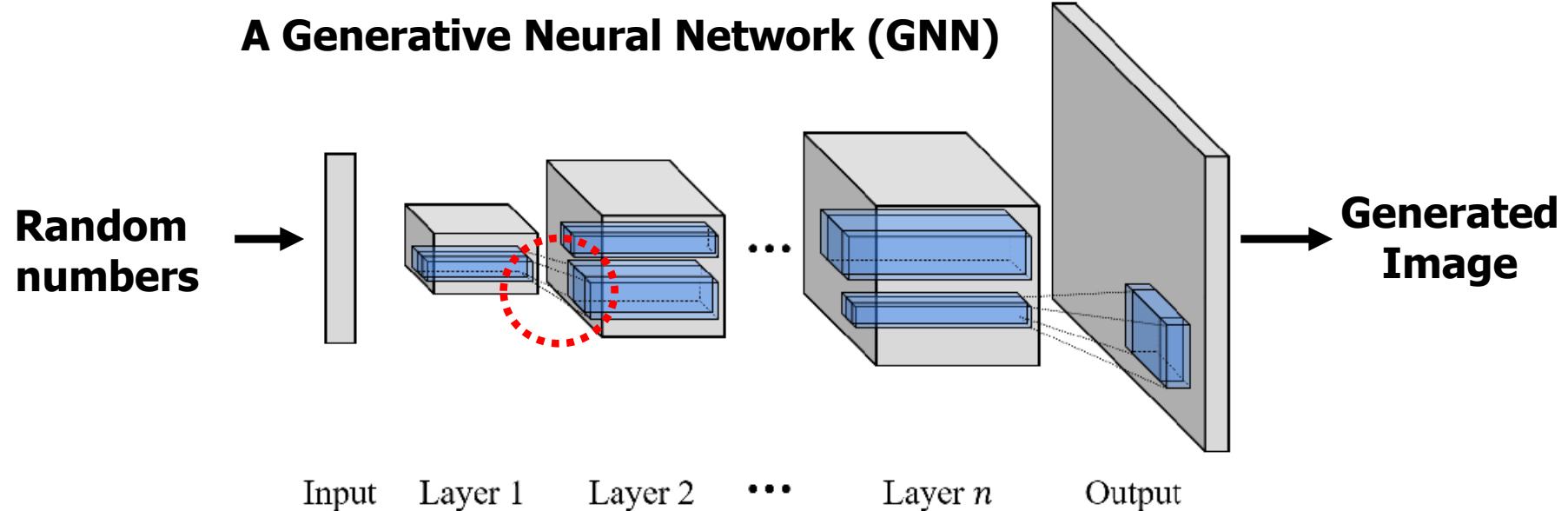
Dissecting Deep Neural Networks

Generative Boundary Aware Sampling

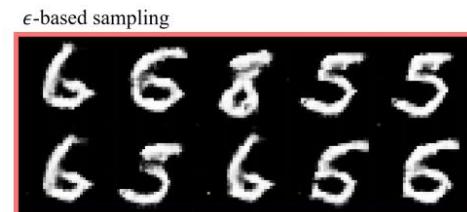
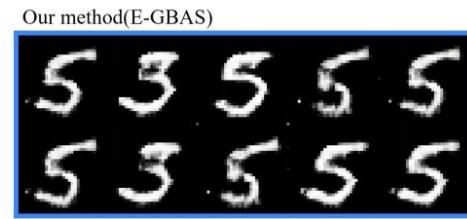
Giyoung Jeon et. al., 2019 (J. Choi, UNIST)



A Generative Neural Network (GNN)



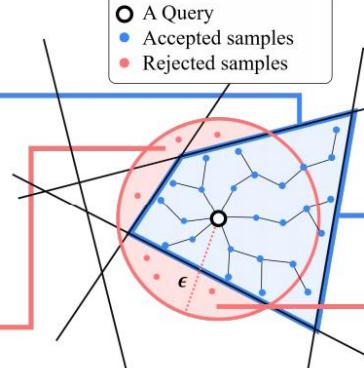
Generative boundaries



DCGAN on MNIST

Layer n Output

- A Query
- Accepted samples
- Rejected samples



Our method(E-GBAS)



ϵ -based sampling

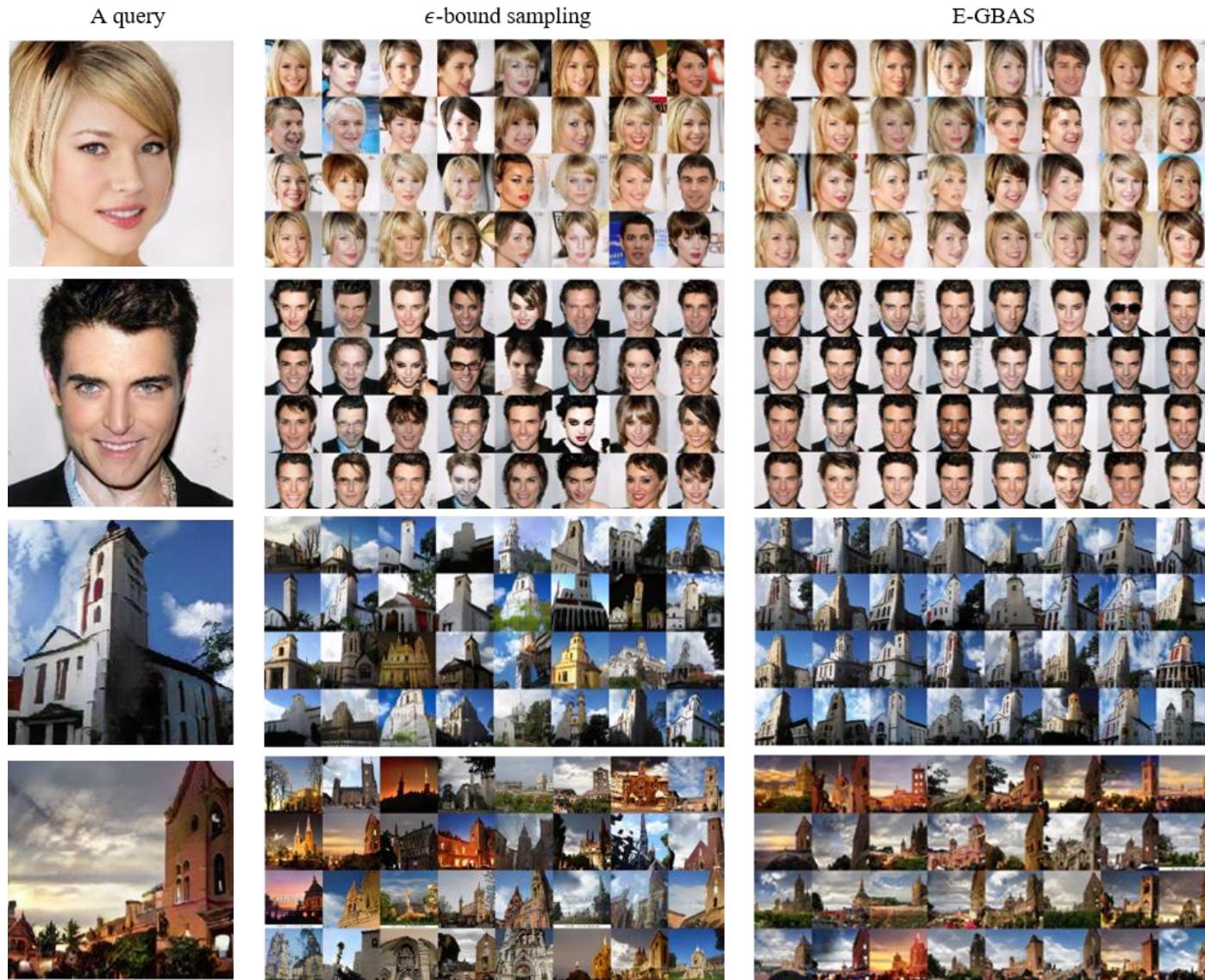


PGGAN on LSUN Church

Dissecting Deep Neural Networks

Generative Boundary Aware Sampling

Giyoung Jeon et. al., 2019 (J. Choi, UNIST)



Technical Approaches

Explaining Deep Neural Networks

Explaining Deep Reinforcement Learning (DRL)

Extract State Representation of DRL (Oregon)

Modular DRL (Berkeley)

Explaining by Combining Explainable Models

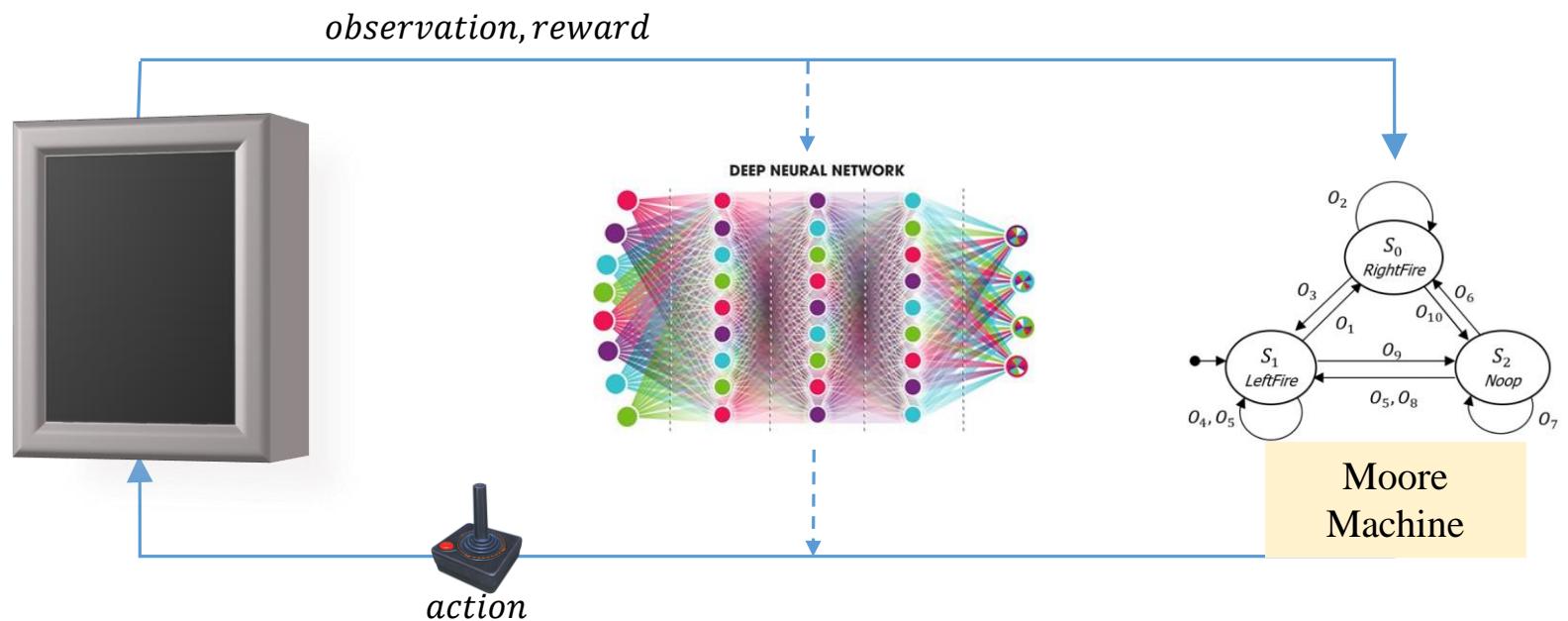
Finding Local Explanations (Model Agnostic Methods)

Data Sets and Applications

Extract State Representation of DRL

Alan Fern (Oregon State)

Objective



Learning Finite State Representations of Recurrent Policy Networks

Anurag Koul, Alan Fern, and Sam Greydanus

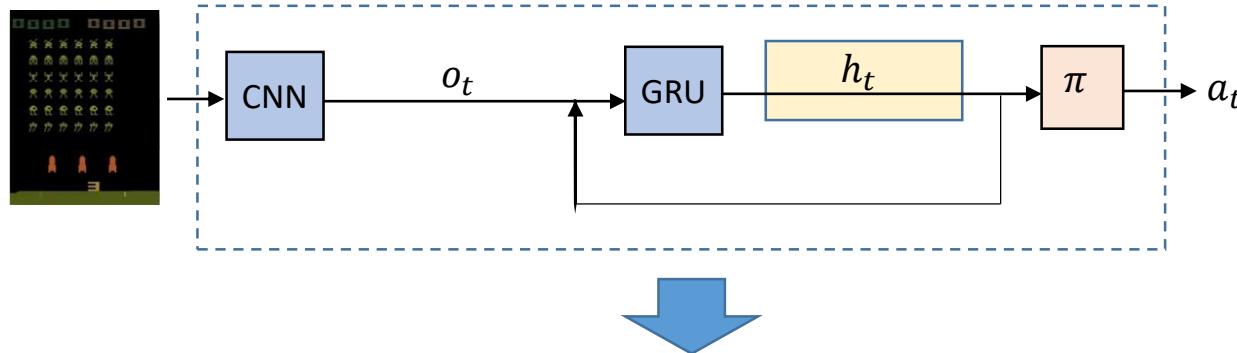
International Conference on Learning Representations (ICLR-2019).

Extract State Representation of DRL

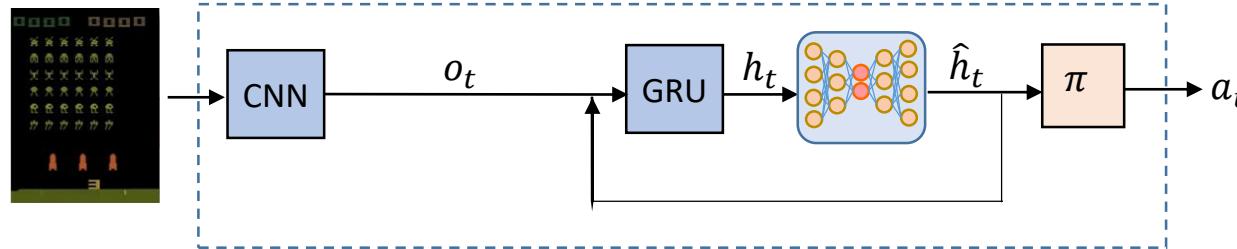
Alan Fern (Oregon State)

Quantized Bottleneck Insertion

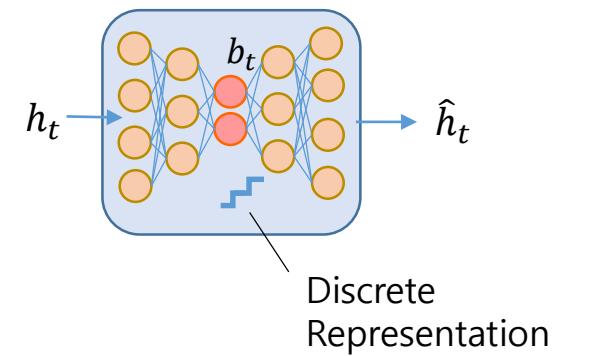
Original Continuous RNN



Discrete Memory



Quantized Bottleneck Network



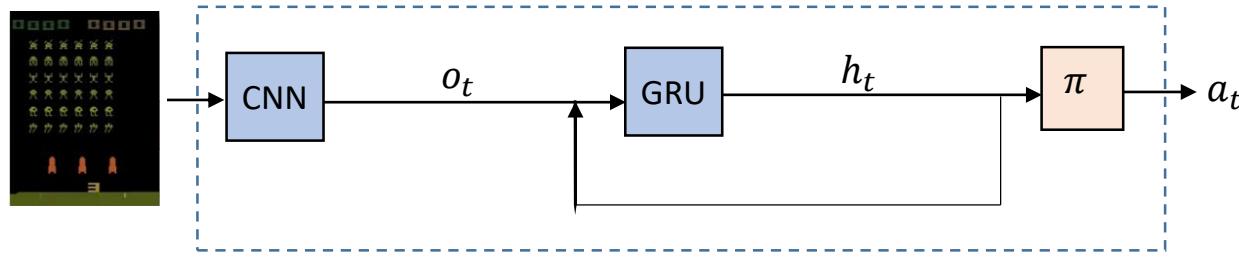
Discrete Representation

Extract State Representation of DRL

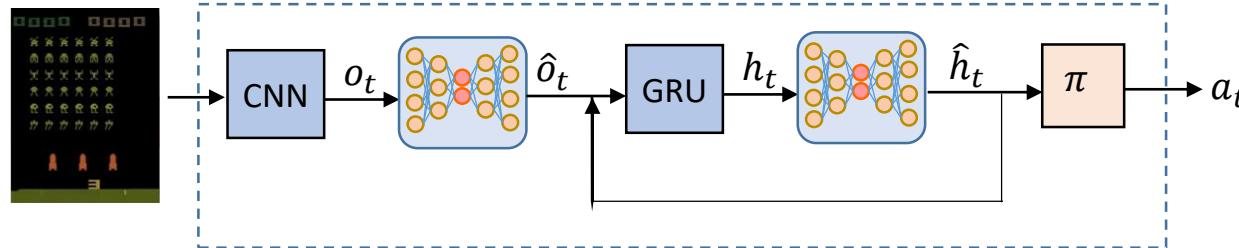
Alan Fern (Oregon State)

Quantized Bottleneck Insertion

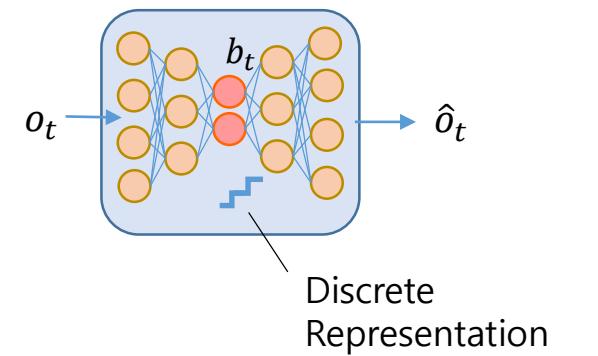
Original Continuous RNN



Discrete Memory + Discrete Input



Quantized Bottleneck Network

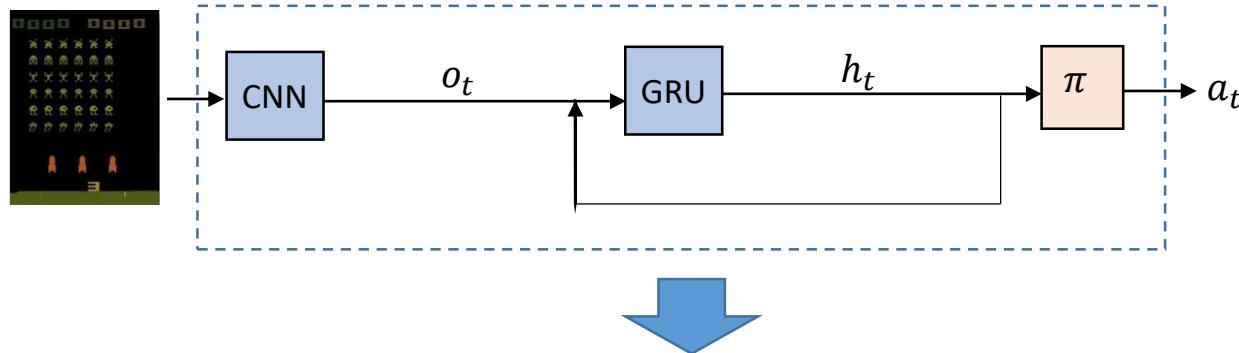


Extract State Representation of DRL

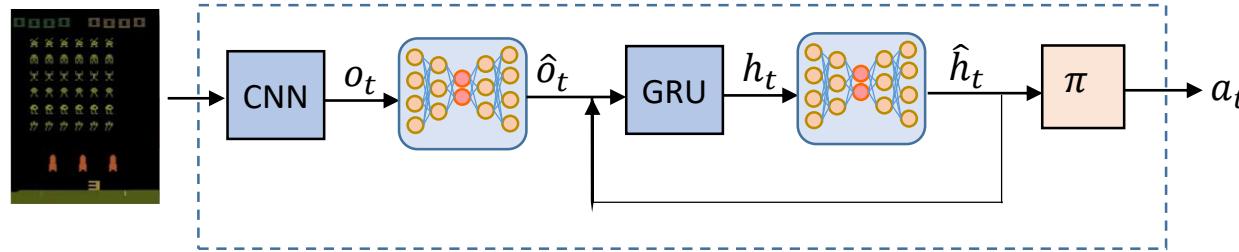
Alan Fern (Oregon State)

Quantized Bottleneck Insertion

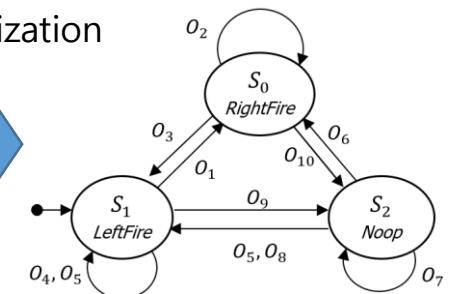
Original Continuous RNN



Discrete Memory + Discrete Input



Finite State Machine
Minimization

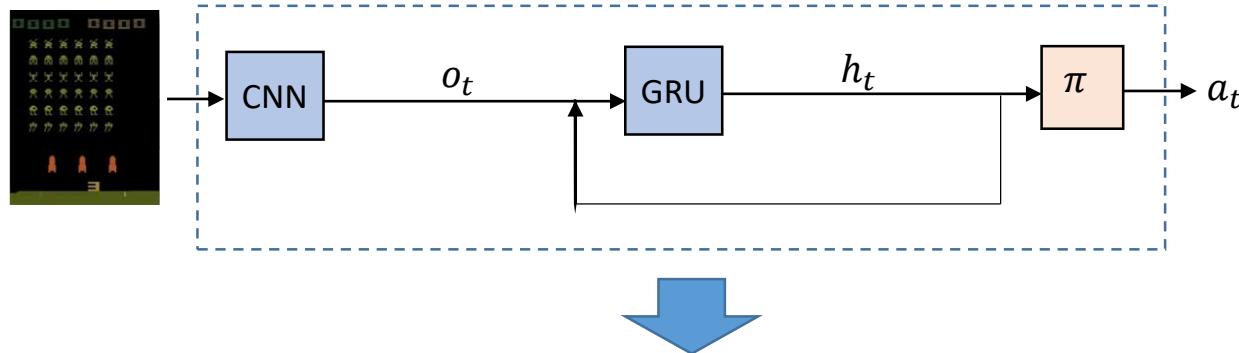


Extract State Representation of DRL

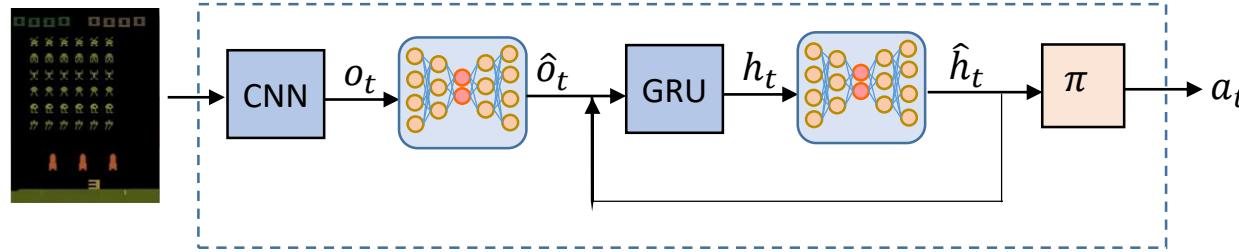
Alan Fern (Oregon State)

Quantized Bottleneck Insertion

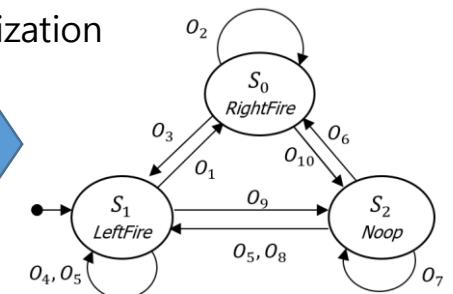
Original Continuous RNN



Discrete Memory + Discrete Input



Finite State Machine
Minimization

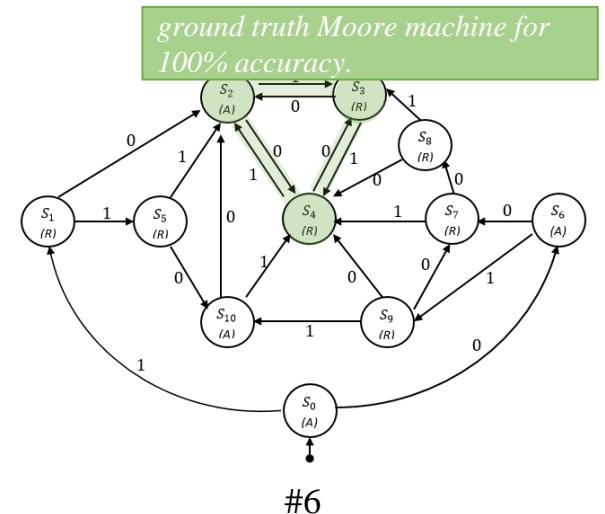
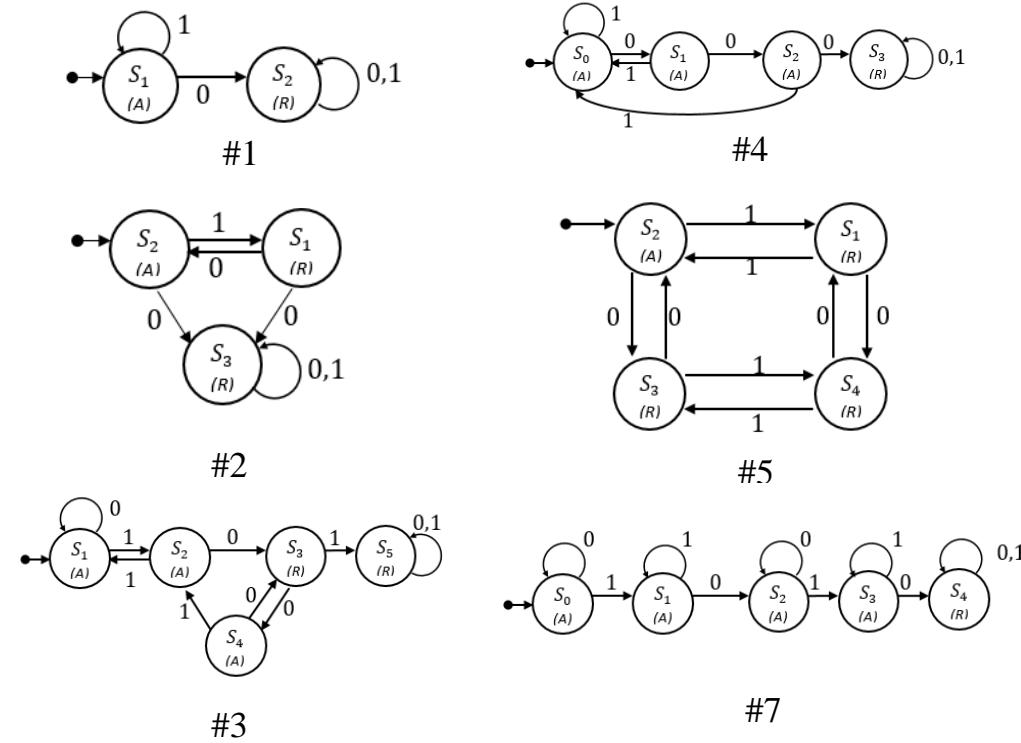


Extract State Representation of DRL

Alan Fern (Oregon State)

Grammar Learning Benchmarks: Tomita Grammars

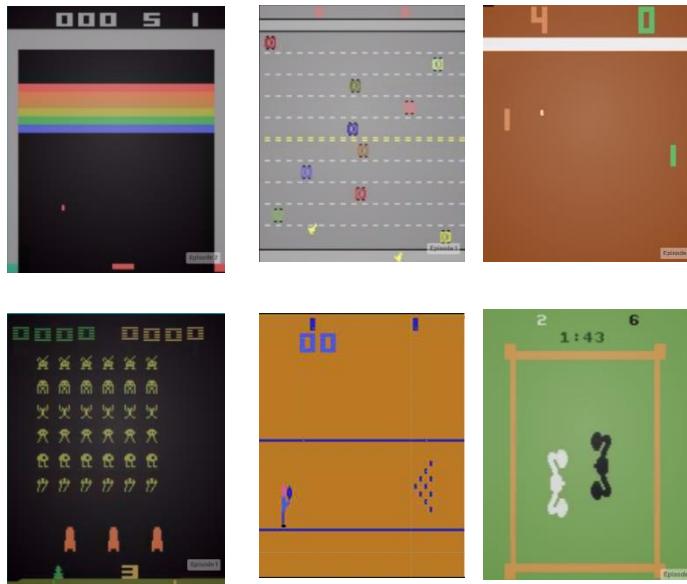
- 'A' and 'R' are acceptor and rejecter states
- Other than #6, all are 100% accurate. #6 is 99% accurate.



Extract State Representation of DRL

Alan Fern (Oregon State)

Atari – 6 Games



Game (# actions)	Stage - 4 Moore Machine					
	Before Minimization			After Minimization		
	H	O	Score	H	O	Score
Pong (3)	373	372	21	3	10	21
Freeway (3)	1	1	21	1	1	21
Breakout (4)	1888	1871	415	8	30	415
Space Invaders (4)	1625	1620	1235	12	29	1235
Bowling (6)	49	1	60	33	1	60
Boxing (18)	2621	2605	100	14	119	100

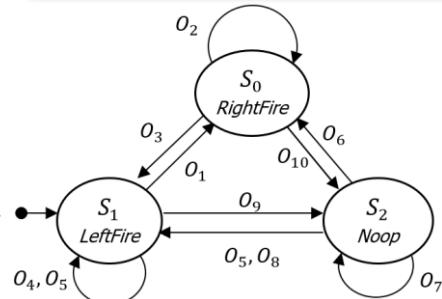
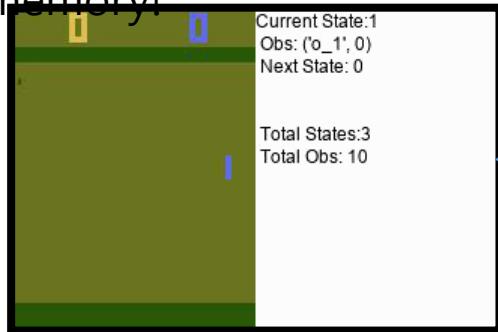
Extract State Representation of DRL

Alan Fern (Oregon State)

Atari – Demonstration (3 of our 6 Atari Games)

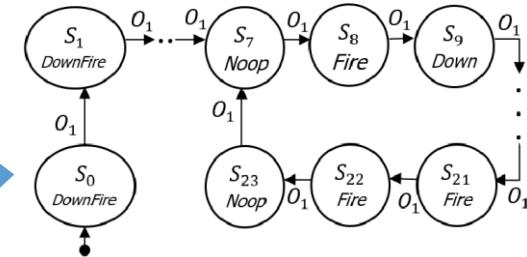
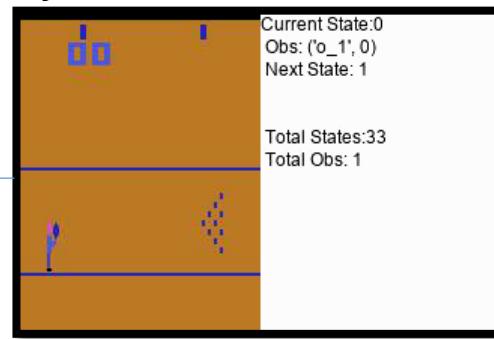
Pong

Policy does not use memory!



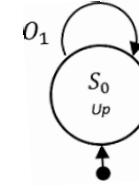
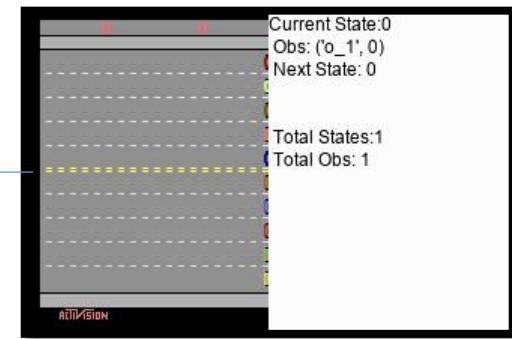
Bowling

Policy does not use observations



Freeway

No memory nor observation



Modular DRL

Pieter Abbeel (Berkeley)

- A popular real-time strategy game by Blizzard (2010 - now).
- Academic challenges: sparse rewards, large state & action spaces.
- Jan, 2019: DeepMind's AlphaStar beats pro players.



Collect Resources



Construct Buildings



Produce Units



Attack Enemies

[Dennis Lee*, Haoran Tang*, Jeffrey O Zhang, Huazhe Xu, Trevor Darrell, Pieter Abbeel AIIDE 2018]

Modular DRL

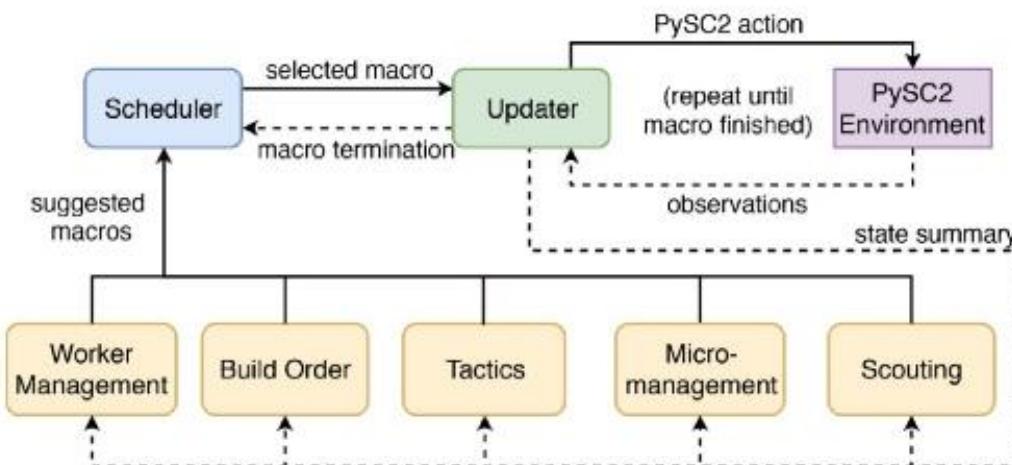
Pieter Abbeel (Berkeley)

How to simplify the complex decision process?

- Split responsibilities between five modules
- Routine modules are scripted
- Hard-to-script modules (Build Order, Tactics) are trained with deep reinforcement learning + self-play

Benefits

- Efficient Learning
- Explainable: automatically explains its decisions in text. e.g. (06:10) [Micro] Attack enemy armies



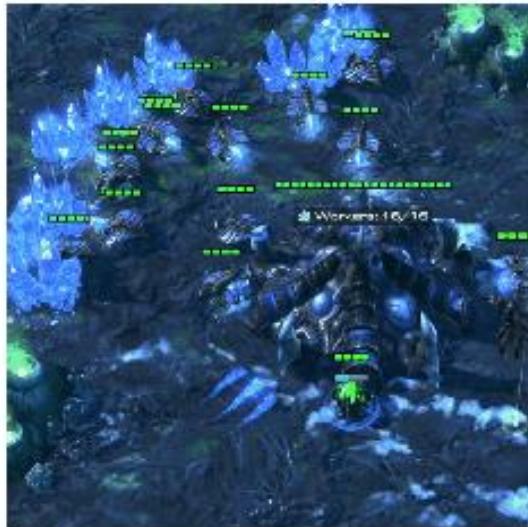
[Dennis Lee*, Haoran Tang*, Jeffrey O Zhang, Huazhe Xu, Trevor Darrell, Pieter Abbeel AIIDE 2018]

Modular DRL

Pieter Abbeel (Berkeley)

The experts are provided 3 examples of explaining certain behavior.

(0:53) [Build Order] Build
the second base



(01:24) [Worker Management]
Move 3 workers to mine gas



(06:10) [Micro] Attack enemy
armies



Modular DRL

Pieter Abbeel (Berkeley)

Without explanations



With explanations



Technical Approaches

Explaining Deep Neural Networks

Explaining Deep Reinforcement Learning (DRL)

Explaining by Combining Explainable Models

Automatic Statistician (MIT/Cambridge)

Relational Automatic Statistician (UNIST)

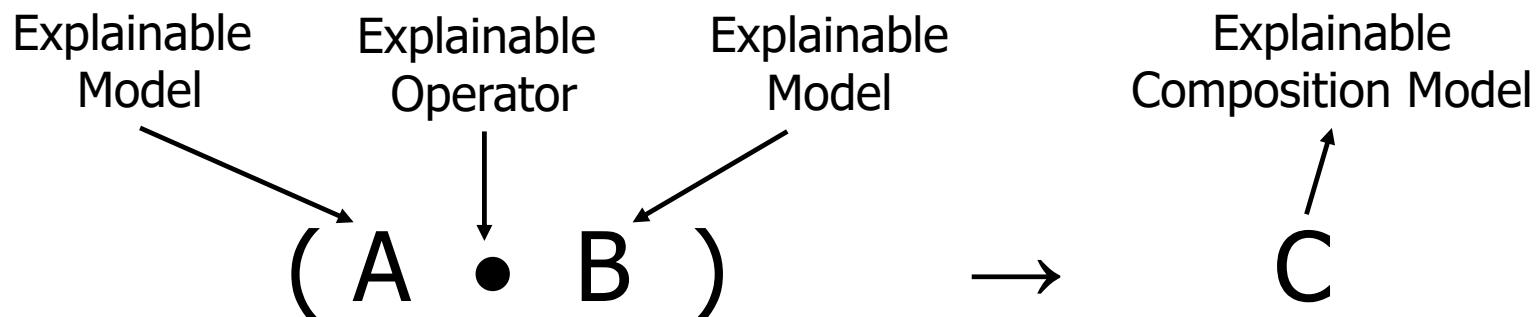
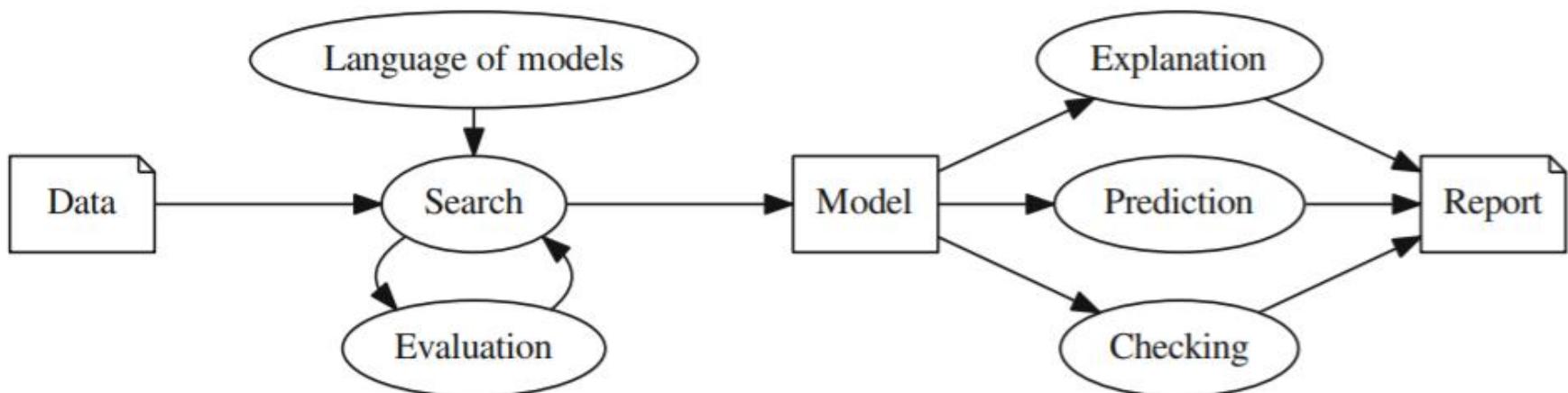
Finding Local Explanations (Model Agnostic Methods)

Data Sets and Applications

Automatic Statistician

Z. Ghahramani (Cambridge)

J. Tenenbaum (MIT)



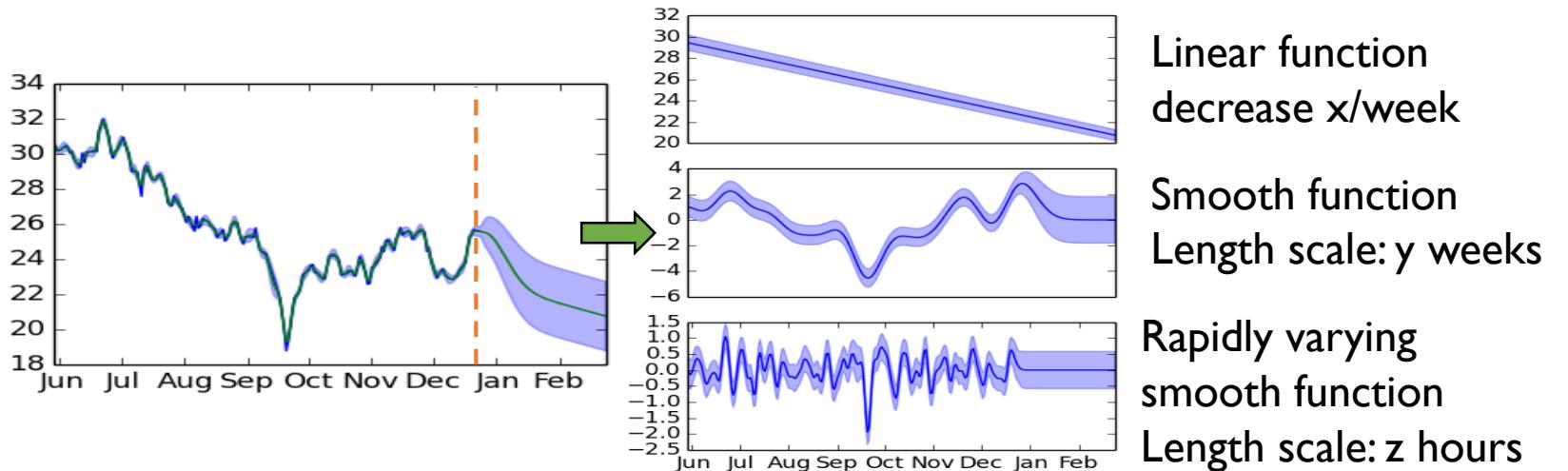
Automatic Statistician

Z. Ghahramani (Cambridge)

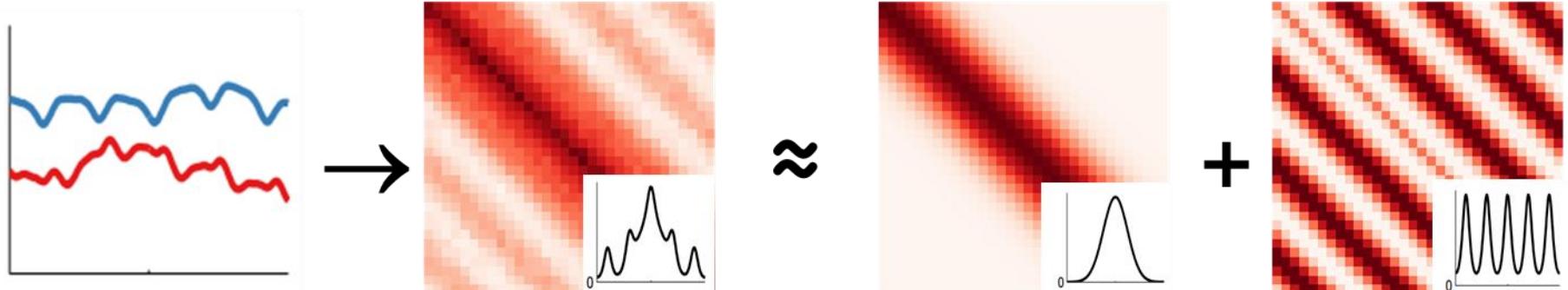
J. Tenenbaum (MIT)

Explaining Multivariate Time Series in Financial Prediction

The Automatic Statistician System (Lloyd et al, 2014)



Covariance Decomposition: Learn Explainable Models from Data

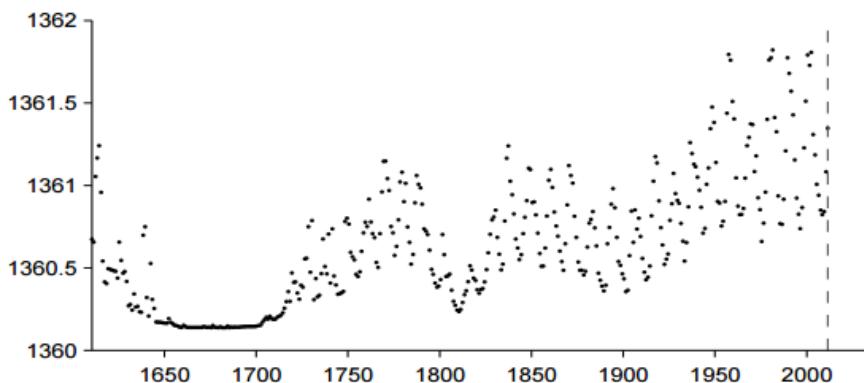


Automatic Statistician

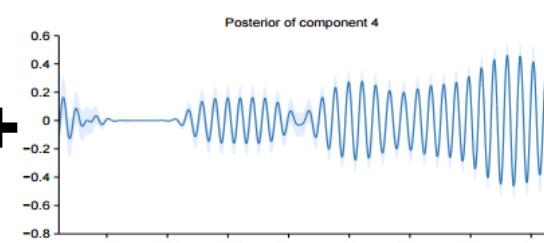
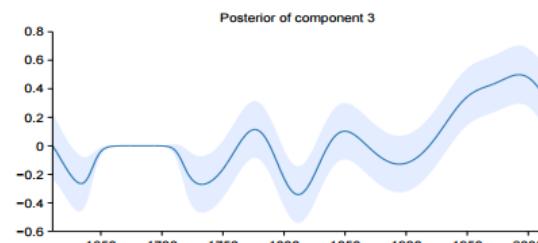
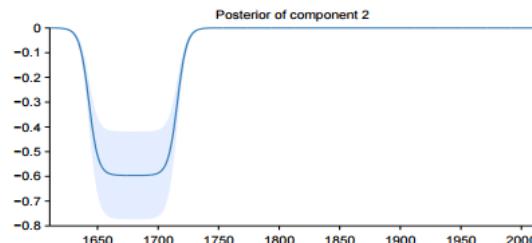
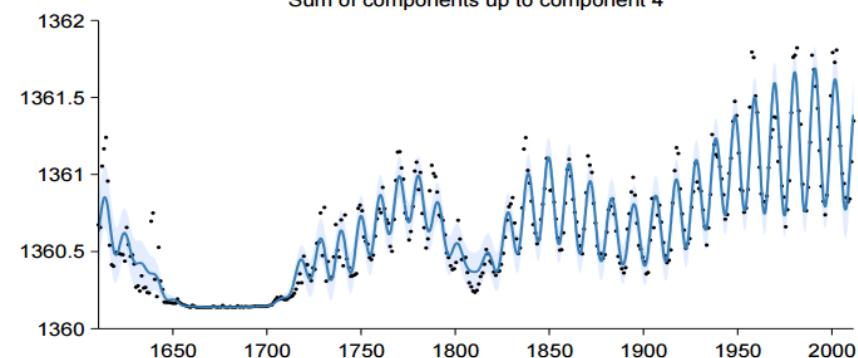
Z. Ghahramani (Cambridge)

J. Tenenbaum (MIT)

태양의 흑점 활동 데이터



Sum of components up to component 4



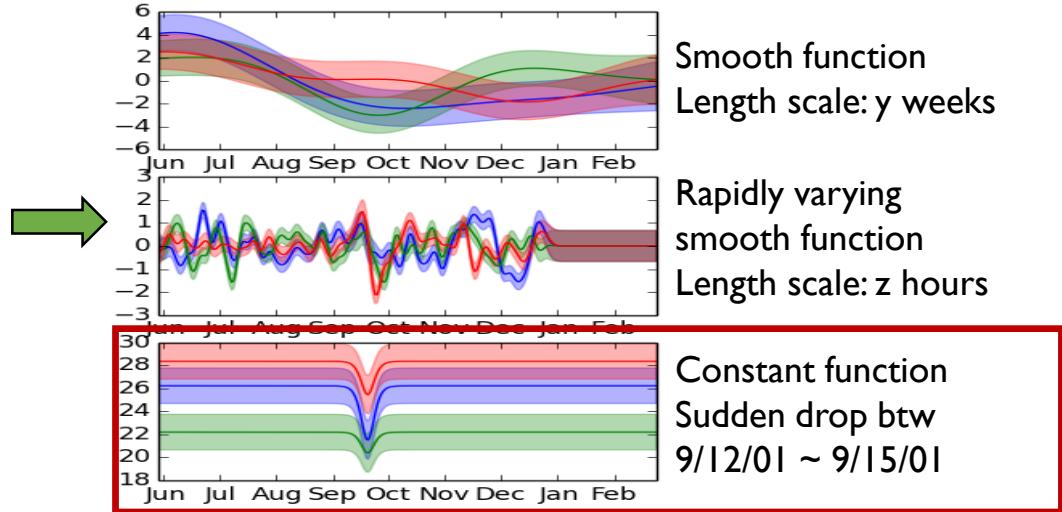
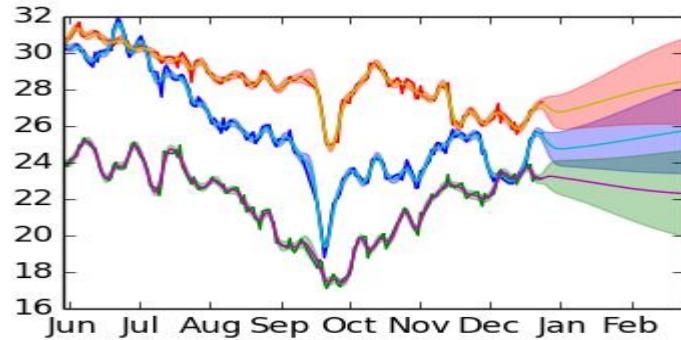
Relational Automatic Statistician

Jaesik Choi (UNIST)

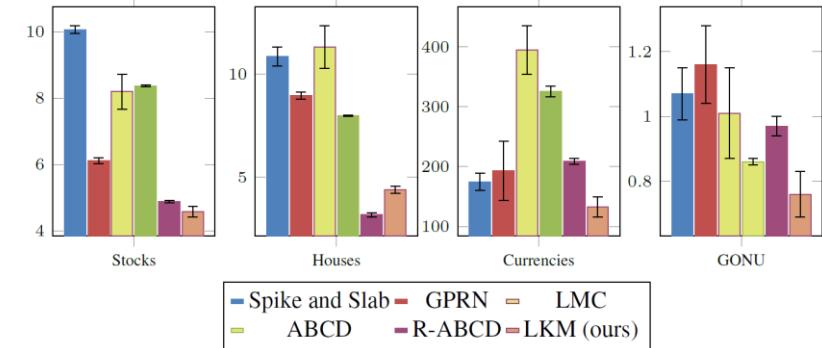
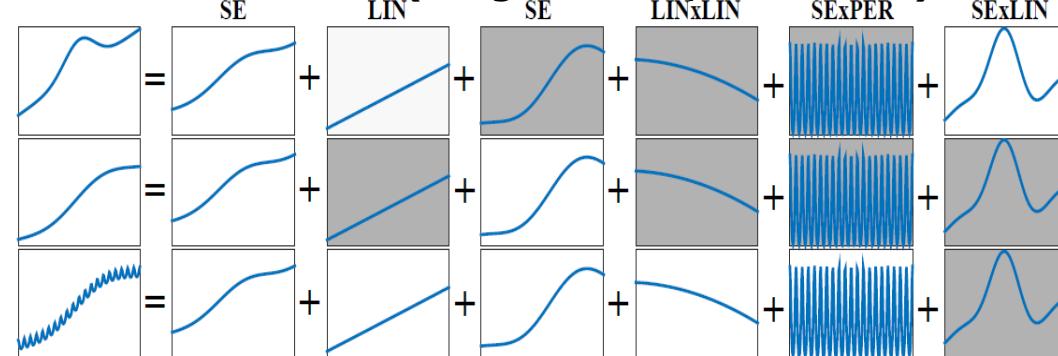


Explaining Multivariate Time Series in Financial Prediction

Relational Automatic Statistician System (Hwang et al, ICML 2016)



Latent Kernel Model (Tong and Choi, ICML 2019)



Yunseong Hwang, Anh Tong and Jaesik Choi, *Automatic Construction of Nonparametric Relational Regression Models for Multiple Time Series*, ICML 2016

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

Comparisons of RMSE

Relational Automatic Statistician

Jaesik Choi (UNIST)



Explaining Multivariate Time Series in Financial Prediction

An example of analysis of Gold, Oil, NASDAQ and USD

- 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.

The Automatic Statistician

An artificial intelligence for data science

<http://www.automaticstatistician.com/>

Research

Below is a selection of research articles relating to the Automatic Statistician project.

Scaling up the Automatic Statistician: Scalable Structure Discovery using Gaussian Processes

Hyunjik Kim, Yee Whye Teh

Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (2018)

[pdf](#) | [supplementary pdf](#) | [arxiv](#) | [bibtex](#)

The Automatic Statistician: A Relational Perspective

Yunseong Hwang, Anh Tong, Jaesik Choi

ICML 2016: Proceedings of the 33rd International Conference on Machine Learning (2016).

[pdf](#) | [supplementary pdf](#) | [arxiv](#) | [slides](#) | [bibtex-1](#) | [bibtex-2](#)

Statistical Model Criticism using Kernel Two Sample Tests

James Robert Lloyd, Zoubin Ghahramani

Advances in Neural Information Processing Systems 28 (2015)

[pdf](#) | [code](#) | [bibtex](#)

Automatic Construction and Natural-Language Description of Nonparametric Regression Models

James Robert Lloyd, David Duvenaud, Roger Grosse, Joshua B. Tenenbaum, Zoubin Ghahramani

Association for the Advancement of Artificial Intelligence (AAAI) Conference, 2014

[pdf](#) | [code](#) | [examples](#) | [bibtex](#)

관계형 자동 통계학자(UNIST)
The Relational Automatic Statistician

Explaining Financial Prediction - Demo



Jaesik Choi & Byunggi Seo (UNIST)



HOME ABOUT NEWS&INFO RESEARCH OPEN SOURCE SYMPOSIUM CONTACT

Automatic News

Company Stock Prediction

No.	Company Name
1	아모레퍼시픽
2	현대모비스
3	현대차
4	한국전력
5	LG화학
6	LG생활건강
7	삼성물산
8	삼성전자
9	삼성에스디에스
10	SK하이닉스

Raw Material Prediction

No.	Raw Material Name
1	WTI유
2	브렌트유
3	금
4	온
5	구리
6	알루미늄
7	천연가스
8	미국커피
9	미국옥수수
10	미국코코아

XAIC

© 2018 eXplainable Artificial Intelligence Center. All right reserved.

Yunseong Hwang, Anh Tong and Jaesik Choi, Automatic Construction of Nonparametric Relational Regression Models for Multiple Time Series, ICML 2016

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

Technical Approaches

Explaining Deep Neural Networks

Explaining Deep Reinforcement Learning (DRL)

Explaining by Combining Explainable Models

Finding Local Explanations (Model Agnostic Methods)

Counterfactual Generation (Toronto)

SHapley Additive exPlanation (SHAP)

Datasets and Applications

Technical Approaches

Explaining Deep Neural Networks

Explaining Deep Reinforcement Learning (DRL)

Explaining by Combining Explainable Models

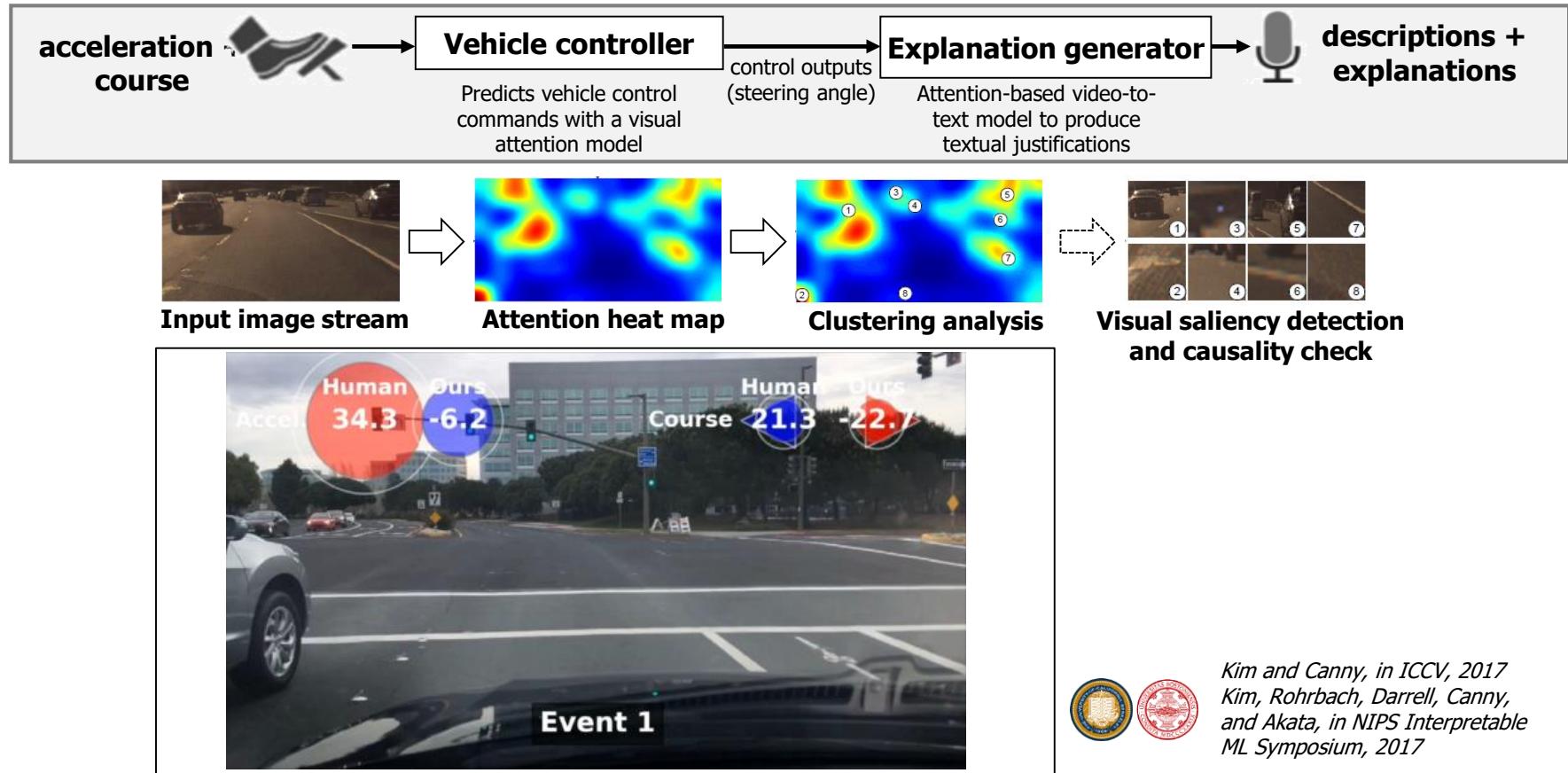
Finding Local Explanations (Model Agnostic Methods)

Data Sets and Applications

DeepDrive (Berkeley)

Explaining Visual Causes (Yonsei)

Distill (Google, OpenAI)





Visual Choice of Plausible Alternatives

Seung-won Hwang (Yonsei) **A new dataset for Visual Causal Reasoning**

190 Dev set
190 Test set



<https://github.com/antest1/VCOPA-Dataset>



Premise (cause)



Alternative 1



Alternative 2

(a) Visual disambiguation



Premise (cause)



Alternative 1



Alternative 2

(b) Temporal disambiguation



Premise (cause)



Alternative 1



Alternative 2

(c) Fine-grained object recognition



Premise (cause)



Alternative 1



Alternative 2

(d) Event recognition



Premise (effect)



Alternative 1



Alternative 2

(e) Inter-event relationship



Premise (effect)



Alternative 1



Alternative 2

(f) Event-sentiment relationship



Premise (effect)



Alternative 1



Alternative 2

(g) Inter-sentiment relationship



Premise (cause)



Alternative 1



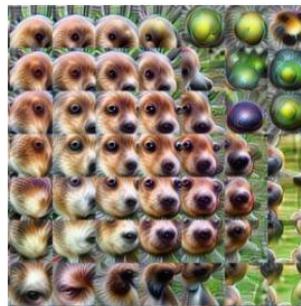
Alternative 2

(h) Commonsense knowledge

Distill for clear explanations of machine learning

OpenAI + Google

Machine Learning Research
Should Be Clear, Dynamic and Vivid.
Distill Is Here to Help.



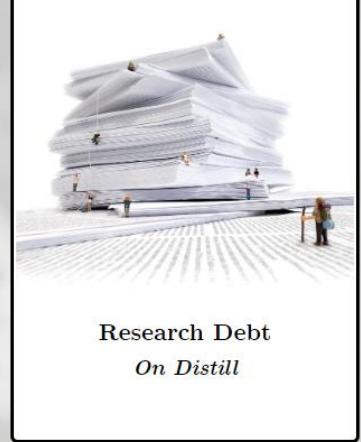
The Building Blocks of Interpretability

On Distill



Feature Visualization
How neural networks build up their understanding of images

On Distill



Research Debt
On Distill

The people behind Distill

EDITORS



Shan Carter
Google Brain Team



Chris Olah
OpenAI



Arvind Satyanarayanan
MIT CSAIL

STAFF



Ludwig Schubert
OpenAI

STEERING COMMITTEE



Yoshua Bengio
Université de Montréal



Mike Bostock
Data-Driven Documents



Amanda Cox
The New York Times



Ian Goodfellow
Google Brain Team



Andrej Karpathy
Tesla



Shakir Mohamed
DeepMind



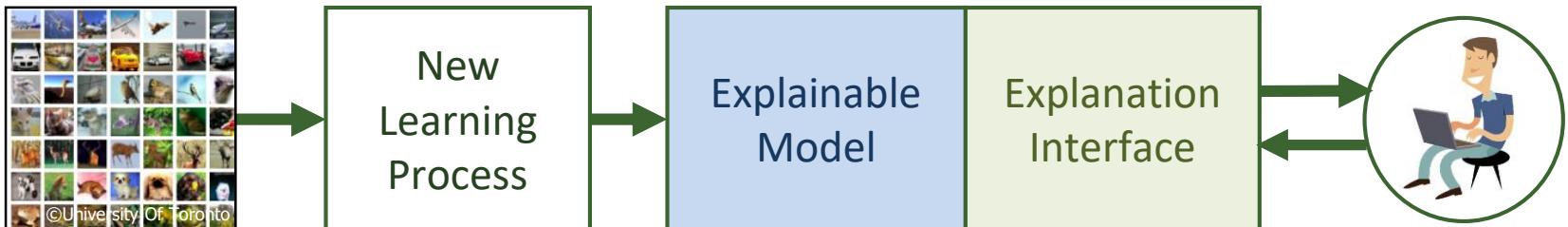
Michael Nielsen
YC Research



Fernanda Viégas
Google Big Picture

Future Plan?

Future Plans



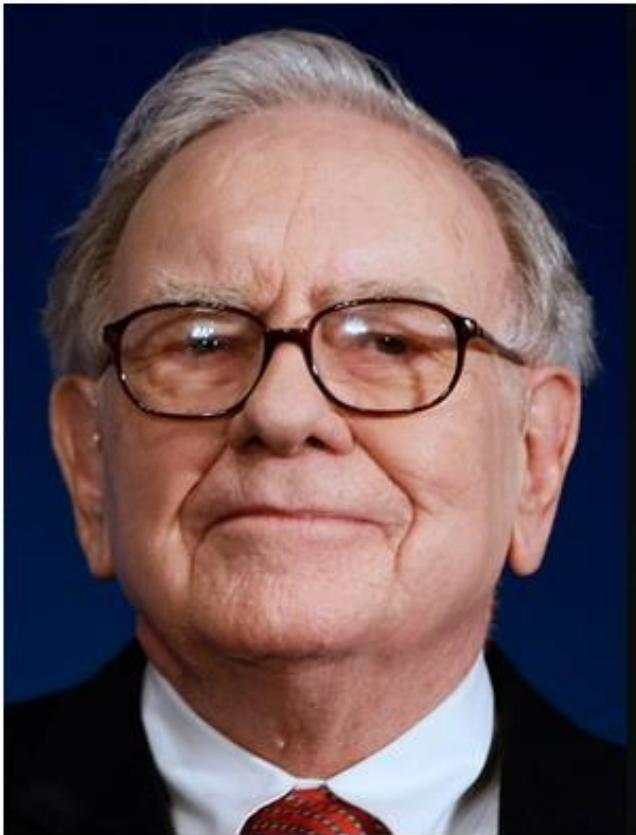
Task	Performer	Explainable Model	Explanation Interface	Application
Task 1	KAIST	Deep Learning (RNN)	Explaining Sparse Features	Healthcare
Task 2	KAIST	Medical Imaging	Explainable Computer Aided Diagnosis (XCAD)	Healthcare
	SNU	Visual Q/A	Visual Explanation	Visual Dialog
	Yonsei/ KAIST	Textual Deep Learning	Textual Explanation	Visual Dialog
Task 3	UNIST	Nonparametric Bayesian	Narrative Explanation	Finance
	Korea*	Deep Learning (attribution methods)	Visual Explanation	Healthcare
	KAIST/ UNIST	Explainable RL	Interactive Visualization	Game

* Korea: Korea University

Future: Toward Reading/Explaining Reports

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.

Future: Toward Reading/Explaining Reports

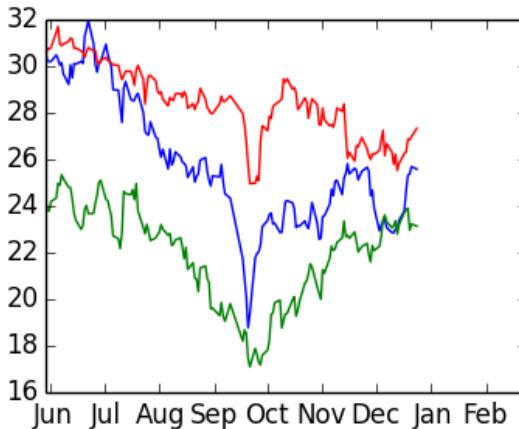
A portrait photograph of Warren Buffett, an elderly man with white hair and glasses, wearing a suit and tie, looking slightly to the right.

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.

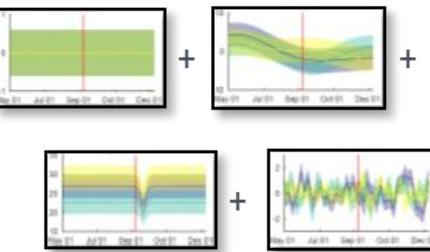
— *Warren Buffett* —

AZ QUOTES

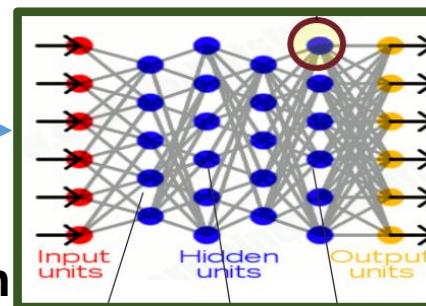
Future: Toward Reading/Explaining Reports



Find
Explanation



Bayesian Learning



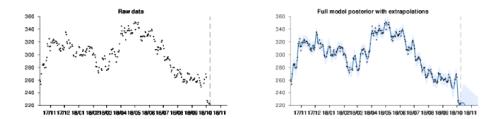
Deep Learning

[특징주]아모레퍼시픽, 전일 대비 약 2.69% 하락한 21만 7000원

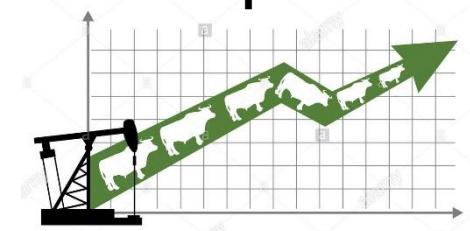
기사입력 2018-10-10 09:33

아모레퍼시픽이 2분기 실적 호조에도 불구하고 하락세를 보이고 있다. 10월 모전 9시 33분 현재 아모레퍼시픽은 전 거래일보다 2.69%(6000원) 떨어진 21만 7000원에 거래되고 있다. 이날 아모레퍼시픽은 2분기 연결 기준 영업이익이 1431억원으로 전년동기대비 41.5% 늘었지만 공시했다. 같은 기간 매출액은 1조 3793억원으로 전년 동기 대비 14.1%, 당기순이익은 1044억원으로 30.6% 각각 증가했다. 아래 그림은 주가 데이터와 변화 예측 자료이다.

향후 1개월간 주식이 92.34%의 확률로 하락할 것으로 예상되며, 18만 1400(-18.29%)원이 하락할 확률이 50%로 예측된다.



Report



Prediction



Annual Reports

Read the Reports and Data and Explain It

International Meetings



2018 International XAI Symposium



Trevor Darrell

 Professor
 UC Berkeley




Wojciech Samek

 Head of Machine Learning Group
 Fraunhofer Heinrich Hertz Institute




2019 ICCV Workshop on Interpreting and Explaining Visual Artificial Intelligence Models

Saturday, November 2nd, 2019

@ COEX 318AB, Seoul, Korea

2019 International XAI Workshop



David Bau

 PhD student
 MIT




Ludwig Schubert

 Software Engineer
 OpenAI


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

Thank you

jaesik@unist.ac.kr



Reference : site

- Part1
 - <https://www.americanbanker.com/news/is-ai-making-credit-scores-better-or-more-confusing>
 - <https://www.youtube.com/watch?v=7a6GrKqOxeU>
 - <https://www.youtube.com/watch?v=uHbMt6WDhQ8>
 - <https://www.youtube.com/watch?v=CrnLINIbfFA>
 - https://www.youtube.com/watch?v=gdAVqJn_J2M
- Part2
 - <https://www.imdb.com/interfaces/>
 - <https://medium.com/mlreview/understanding-lstm-and-its-diagrams-37e2f46f1714>
 - <https://wikidocs.net/32105>
 - https://cyc1am3n.github.io/2018/11/10/classifying_korean_movie_review.html
 - <http://xai.unist.ac.kr/opensource/relatedproject/>
- Part3
 - Spring 2019 session (<https://mlcourse.ai/tutorials>) - ML interpretability by @Christophe Rigon

Reference : publication

- Marco Ancona, et. al., "Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Values Approximation", ICML 2019
- Pieter-Jan Kindermans, Kristof T. Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne, "Learning How to Explain Neural Networks: PatternNet and PatternAttribution", ICLR 2018
- Pieter-Jan Kindermans, Kristof T. Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne, "Learning How to Explain Neural Networks: PatternNet and PatternAttribution", ICLR 2018
- Woo-Jeong Nam, Jaesik Choi, and Seong-Whan Lee, "Relative Attribution Propagation", arXiv 1904.00605, 2019
- Seong Tae Kim, Hakmin Lee, Hak Gu Kim, and Yong Man Ro, "ICADx: Interpretable computer aided diagnosis of breast masses", SPIE Medical Imaging 2019 (Best Student Paper Award)
- David Bau et. al., "Network Dissection", CVPR, 2017 (A. Torralba, MIT)
- David Bau et. al., "GAN Dissection", 2019
- Kyowoon Lee, Sol-A Kim, Jaesik Choi, and Seong-Whan Lee, "Deep Reinforcement Learning in Continuous Action Spaces: a Case Study in the Game of Simulated Curling", ICML 2018
- Yunseong Hwang, Anh Tong and Jaesik Choi, "Automatic Construction of Nonparametric Relational Regression Models for Multiple Time Series", ICML, 2016
- Sebastian Bach et all, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation", ICANN, 2016

XAI센터의 연구결과를 더 알고싶다면

1. XAI 센터 공식홈페이지

<http://xai.unist.ac.kr/>

2. XAI 센터 깃허브

<https://github.com/openxaiProject>

3. XAI 센터 Youtube 채널

https://www.youtube.com/channel/UCGxsflsOry_LdBaPSet2p7g

실습 환경 구축

1. 튜토리얼 안내 페이지 접속

<http://xai.unist.ac.kr/> >event>2019 PYCON TUTORIAL

2. 구글에 로그인

colab_setting.ipynb 파일의 주소를 복사하여,

https://github.com/OpenXAIProject/PyConKorea2019-Tutorials/blob/master/colab_setting.ipynb

OpenXAIProject / PyConKorea2019-Tutorials

Watch 11 Star 0 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Security Insights Settings

Tutorials about Layer-wise Relevance Propagation (LRP) and SHapley Additive exPlanations (SHAP)

Edit

Manage topics

14 commits 1 branch 0 releases 2 contributors Apache-2.0

Branch: master New pull request Create new file Upload files Find File Clone or download

kirarenctaon	Delete image_example.ipynb	Latest commit e5e7a5c 9 minutes ago
LRP	Delete image_example.ipynb	9 minutes ago
SHAP-Tutorial	Add files via upload	9 hours ago
LICENSE	Initial commit	17 hours ago
README.md	Initial commit	17 hours ago
colab_setting.ipynb	Add files via upload	2 hours ago

OpenXAIProject / PyConKorea2019-Tutorials

Watch ▾ 11 ⭐ Star 0 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Security Insights Settings

Tutorials about Layer-wise Relevance Propagation (LRP) and SHapley Additive

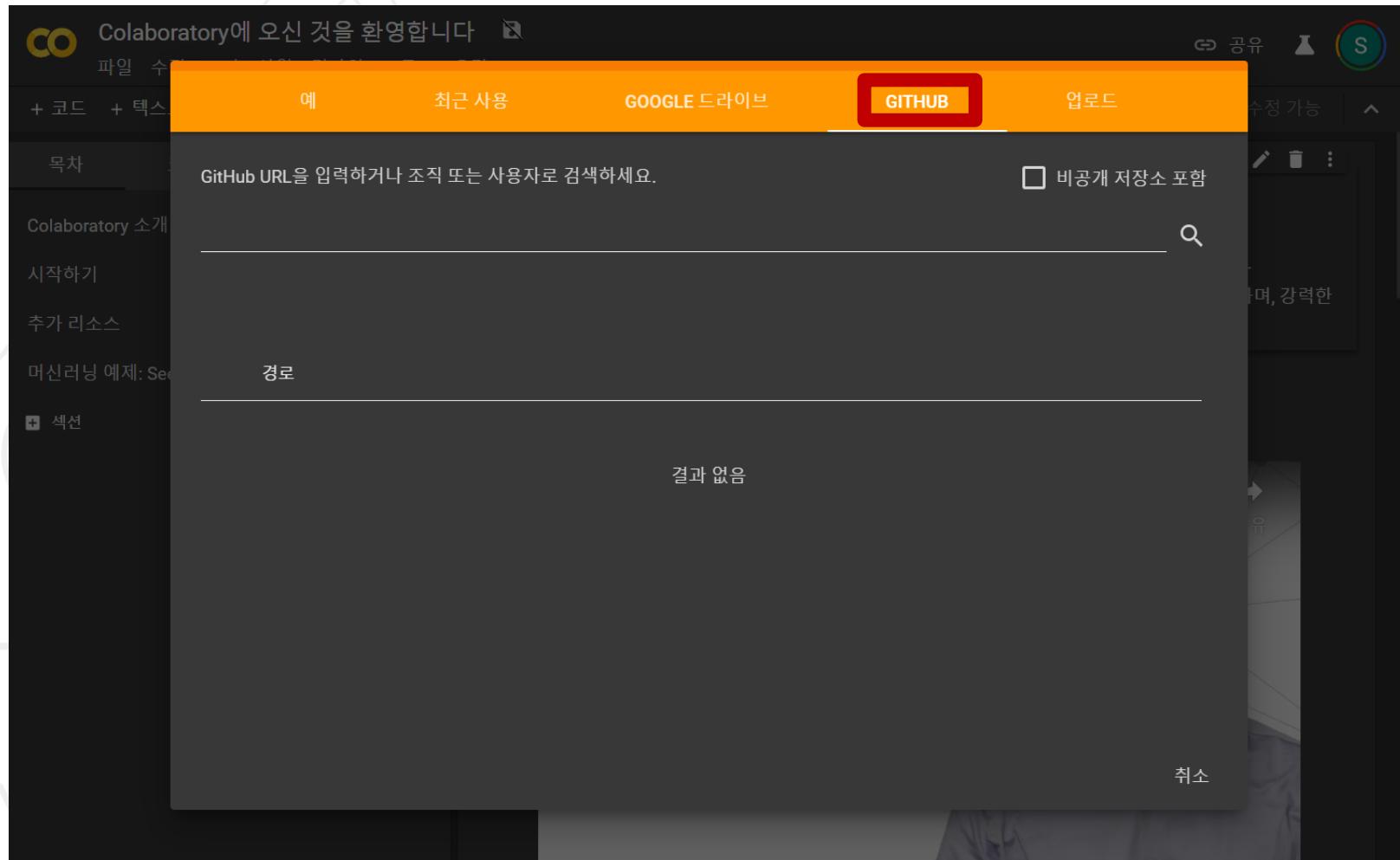
Edit Manage topics

14 commits 1 branch 0 releases Apache-2.0

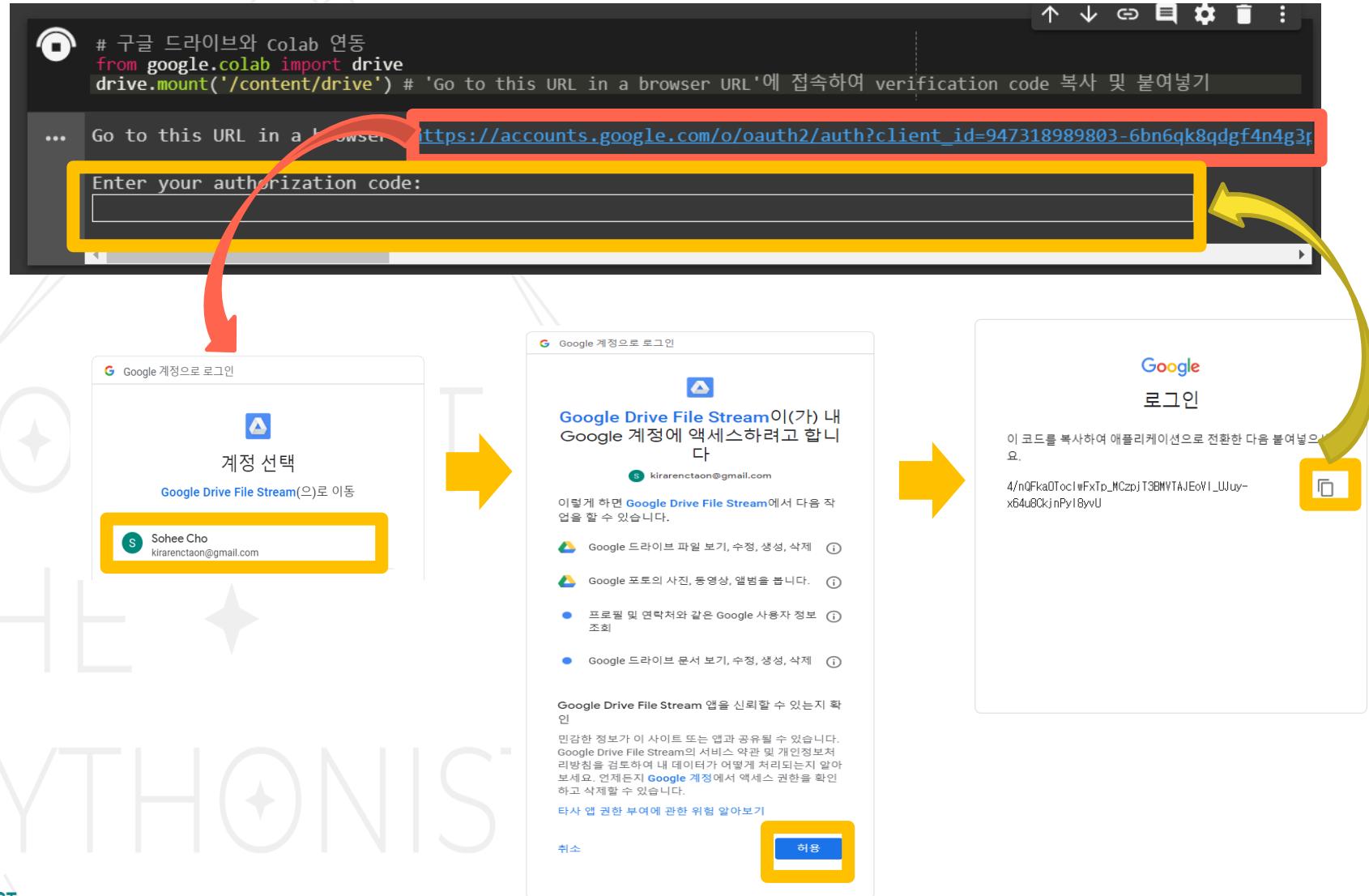
Branch: master New pull request Clone or download

 kirarenctaon	Delete image_example.ipynb	Latest commit e5e7a5c 9 minutes ago
 LRP	Delete image_example.ipynb	9 minutes ago
 SHAP-Tutorial	Add files via upload	9 hours ago
 LICENSE	Initial commit	17 hours ago
 README.md	Initial commit	17 hours ago
 colab_setting.ipynb	Add files via upload	2 hours ago

구글 코랩 GITHUB 탭에 복사해 온 주소를 입력합니다.



구글 드라이브에 마운팅하기 위한 인증 실행해주세요.



만약 아래와 같은 안내가 나온다면 ...

```
[116] # 구글 드라이브와 colab 연동
from google.colab import drive
drive.mount('/content/drive') # 'Go to this URL in a browser URL'에 접속하여 이용동의 후 인증코드 복사해서 붙여넣기

↳ Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

이미 연동된 것으로 그대로 진행하면 됩니다.

연동 후에는 아래 코드를 모두 실행해
오늘 사용할 코드를 구글 드라이브에 클론합니다.