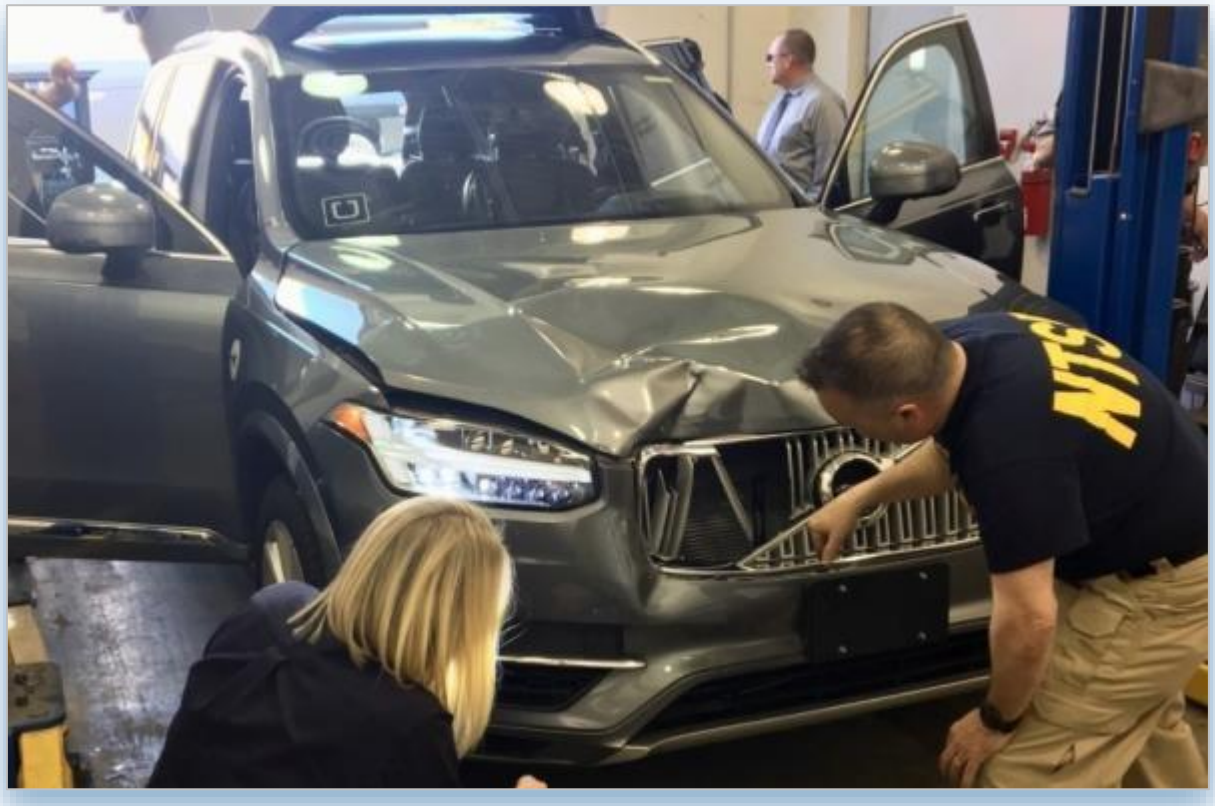
Recent Trends in Explainable Artificial Intelligence (XAI)

KAIST Explainable Artificial Intelligence Center
KAIST Kim Jaechul Graduate School of AI
INEEJI Corporation

Jaesik Choi

Joint work with Gunhyung Kim, Sol-A Kim, Jinsil Lee, Any Tong, Jiyeon Han, Seongun Kim, Wonjoon Jang and Nari Kim

Do We Understand the Internal Mechanism of AI Models?



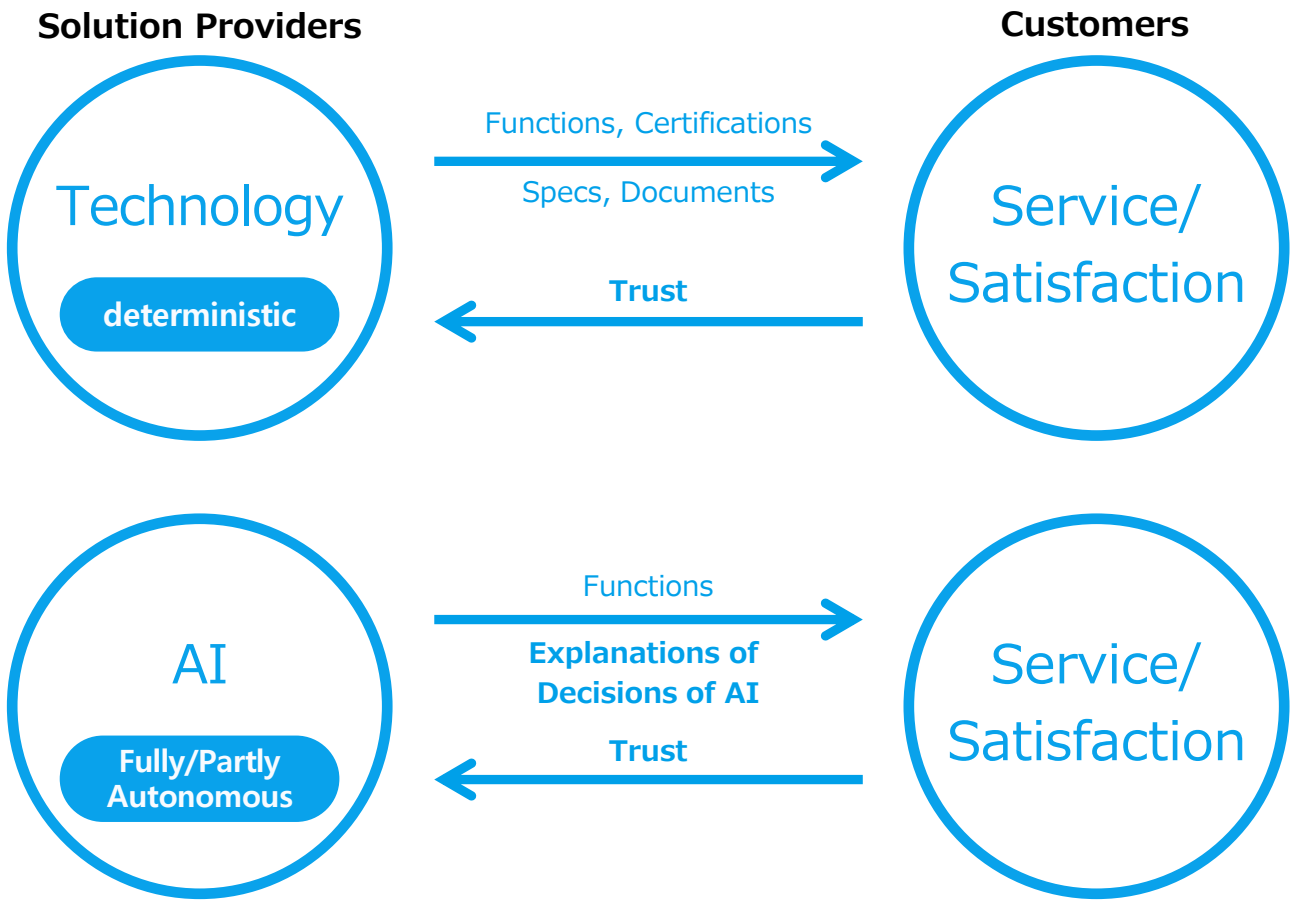
Uber's self-driving car killed a pedestrian (Marc 18th, 2018)

The ‘safety driver’ was watching a TV show (June 22th, 2018)

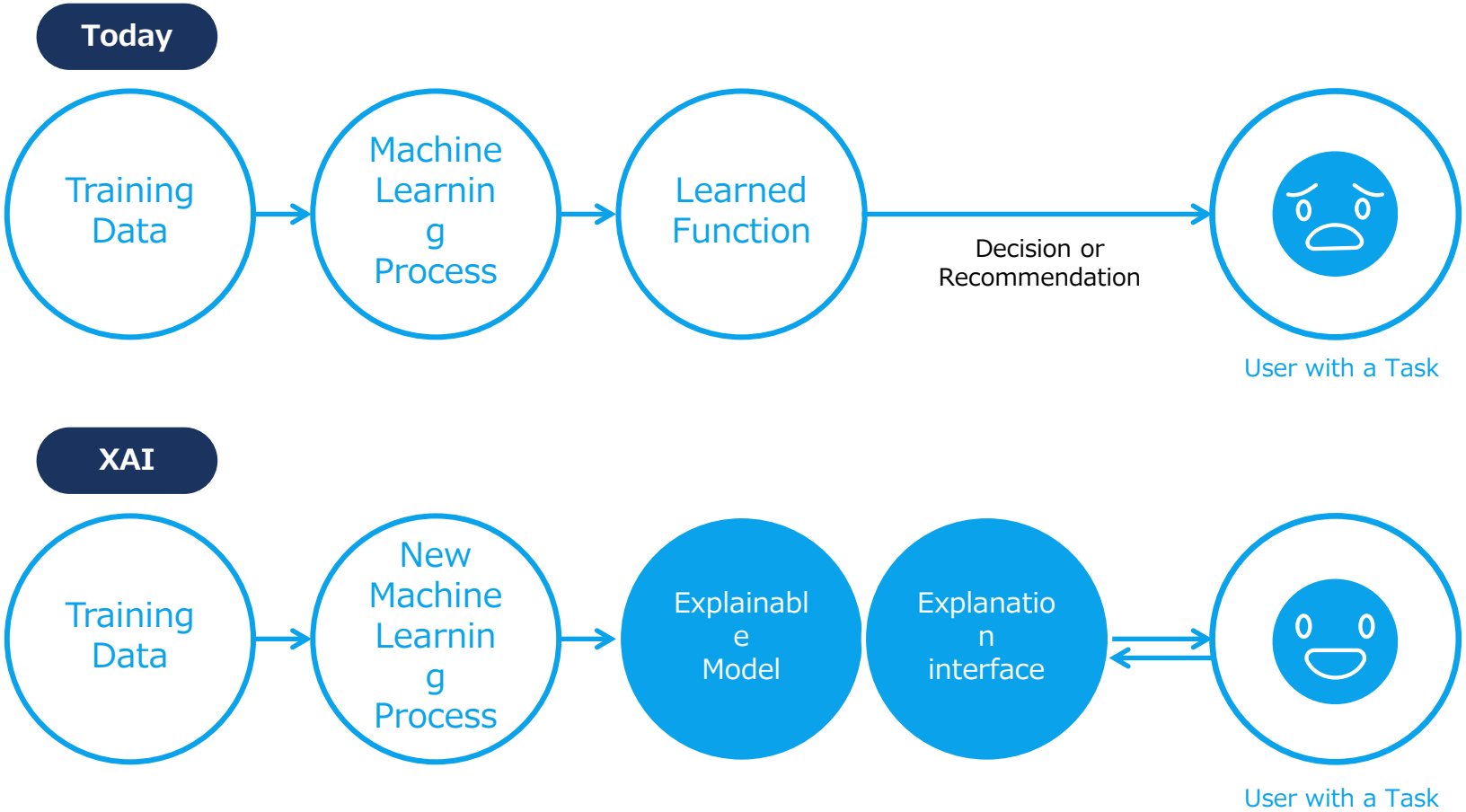
EMPOWERING EXPERTISE
iNEEJi
KAIST

Explainable AI

Existing Services
vs
AI Services



Explainable AI



- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

- I understand why
- I understand why not
- I know when you succeed
- I know when you fail
- I know when to trust you
- I know why you erred

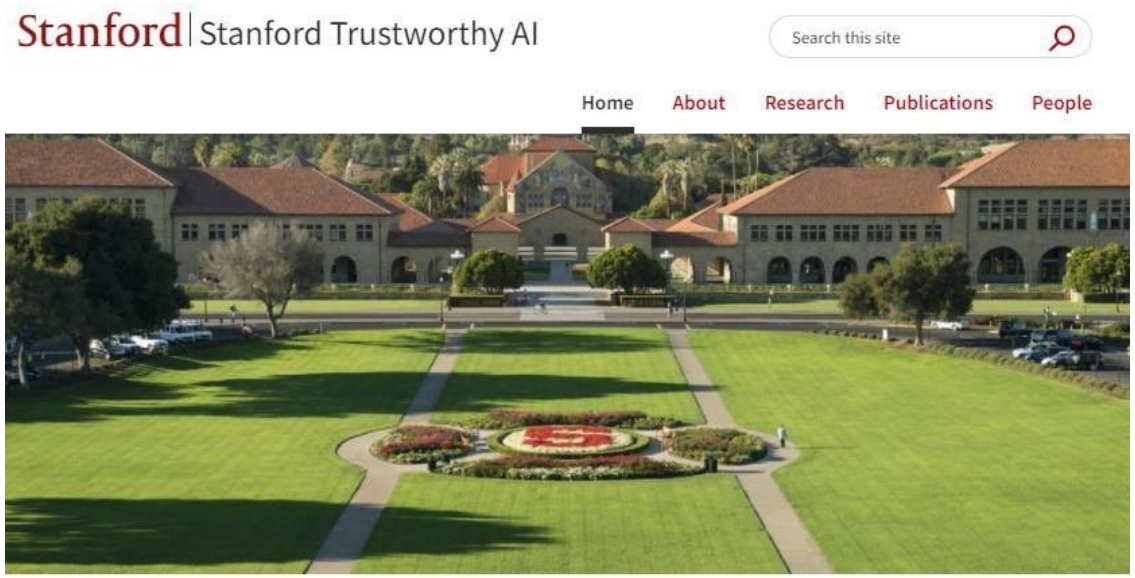
EU General Data Protection Regulations (GDPR)

항목	내용
Right to be forgotten	17 : When customers do not want, the personal contents should be elemen **eliminated**
Llimit of AI decision	22 : Customers have the right not to be handled by **AI algorithm**
Rright to explanation	13-14 : Customers have the right to receive **proper explanations on the decisions made by AI algorithms**
Fines	Up to **4% of total global revenue**
Enact	**2018/05/28**

In the area of high risk AI, the fine will be up to 6% of total global revenue

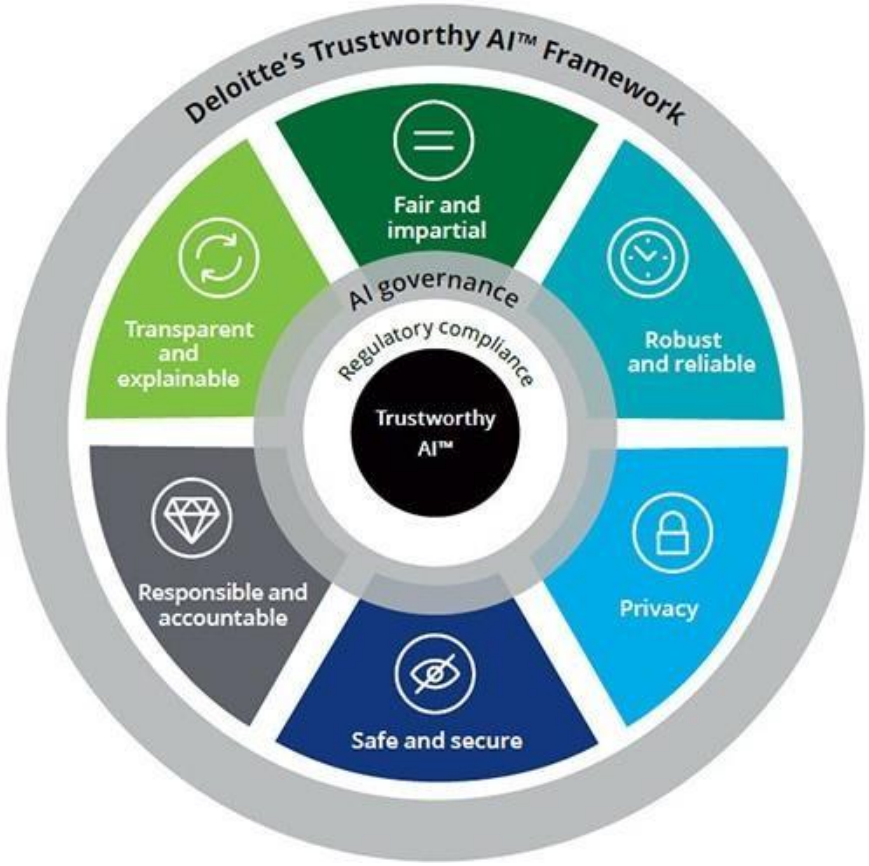
iNEEJI EMPOWERING EXPERTISE
KAIST

Trustworthy AI



Our Mission

Stanford Trustworthy AI aims to **supercharge** innovations in **artificial intelligence** with **human understanding**. We engage in translational research across fairness, explainability, privacy, and robustness, guided by ethics.



EMPOWERING EXPERTISE
iNEEJI
KAIST

Building AI Explanation Capability for the AI-Powered Organization (MIT)

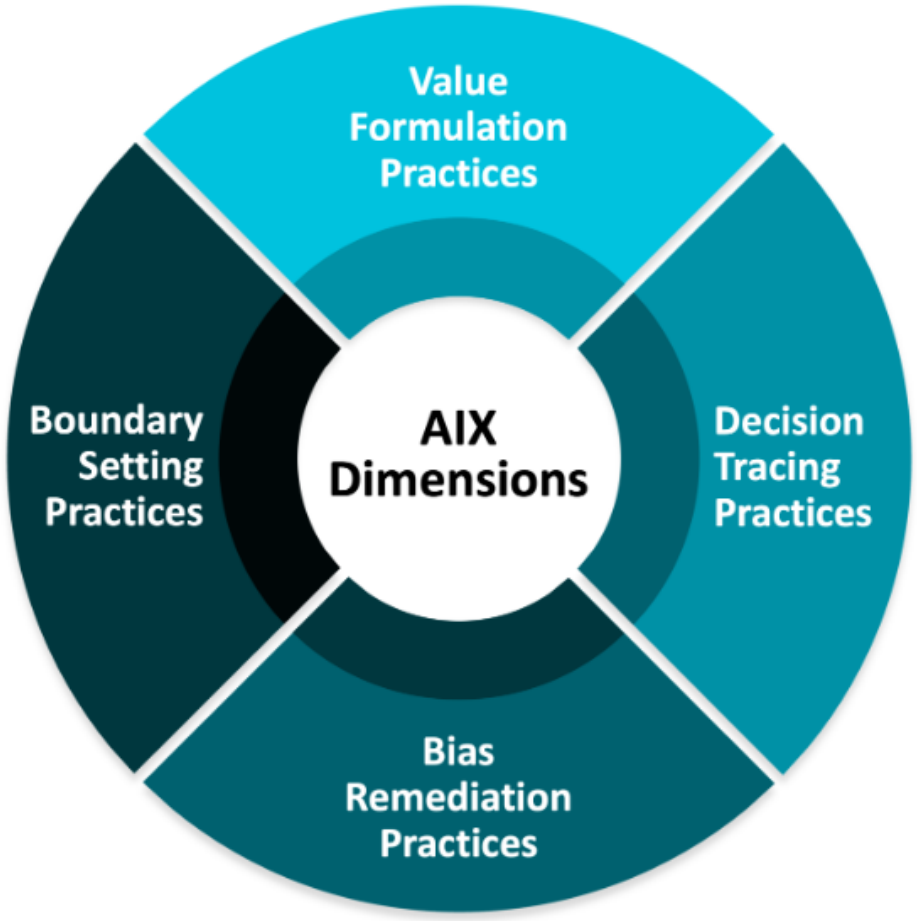


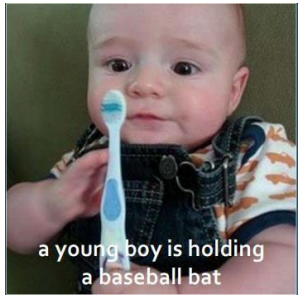
FIGURE 1: FOUR SETS OF PRACTICES THAT BUILD AI EXPLANATION CAPABILITY

- AI 기술의 특성을 4가지로 규정
  - Unproven value: 여전히 시장에서 증명된 유스케이스가 부족하기 때문에, AI 프로젝트의 투자성과에 대한 불확실성이 높음
  - Model opacity: AI 모델의 복잡성으로 인해 모델 예측의 정확성, 적법성, 윤리성을 사전적으로 판단하기 어려움
  - Model drift: 학습데이터에 의존하기 때문에 새로운 데이터에 대한 예측의 정확성이 학습단계의 퍼포먼스와 다름
  - Mindless application: 예측 결과가 항상 오류의 가능성을 가지고 있음
- 설명가능성 확보를 위해 각 특성에 상응하는 practices를 제시
  - Value formulation: 투자성과 입증
  - Decision tracing: 모든 계산 단계 추적
  - Bias remediation
  - Boundary setting

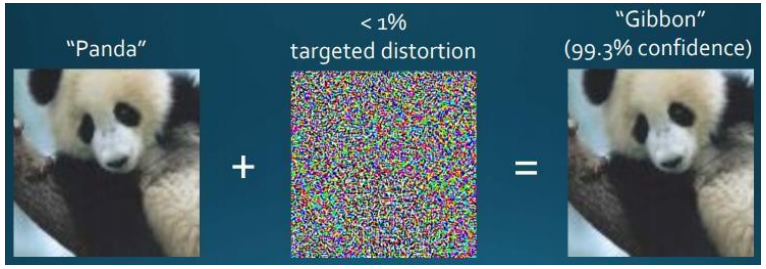
https://cisr.mit.edu/publication/2022_0701_AIX_SomehWixomBeath

A DARPA Perspective on AI – Three Waves of AI

DESCRIBE	CATEGORIZE	EXPLAIN
Handcrafted Knowledge	*Statistical Learning*	*Contextual Adaptation*



Statistically impressive, but individually unreliable

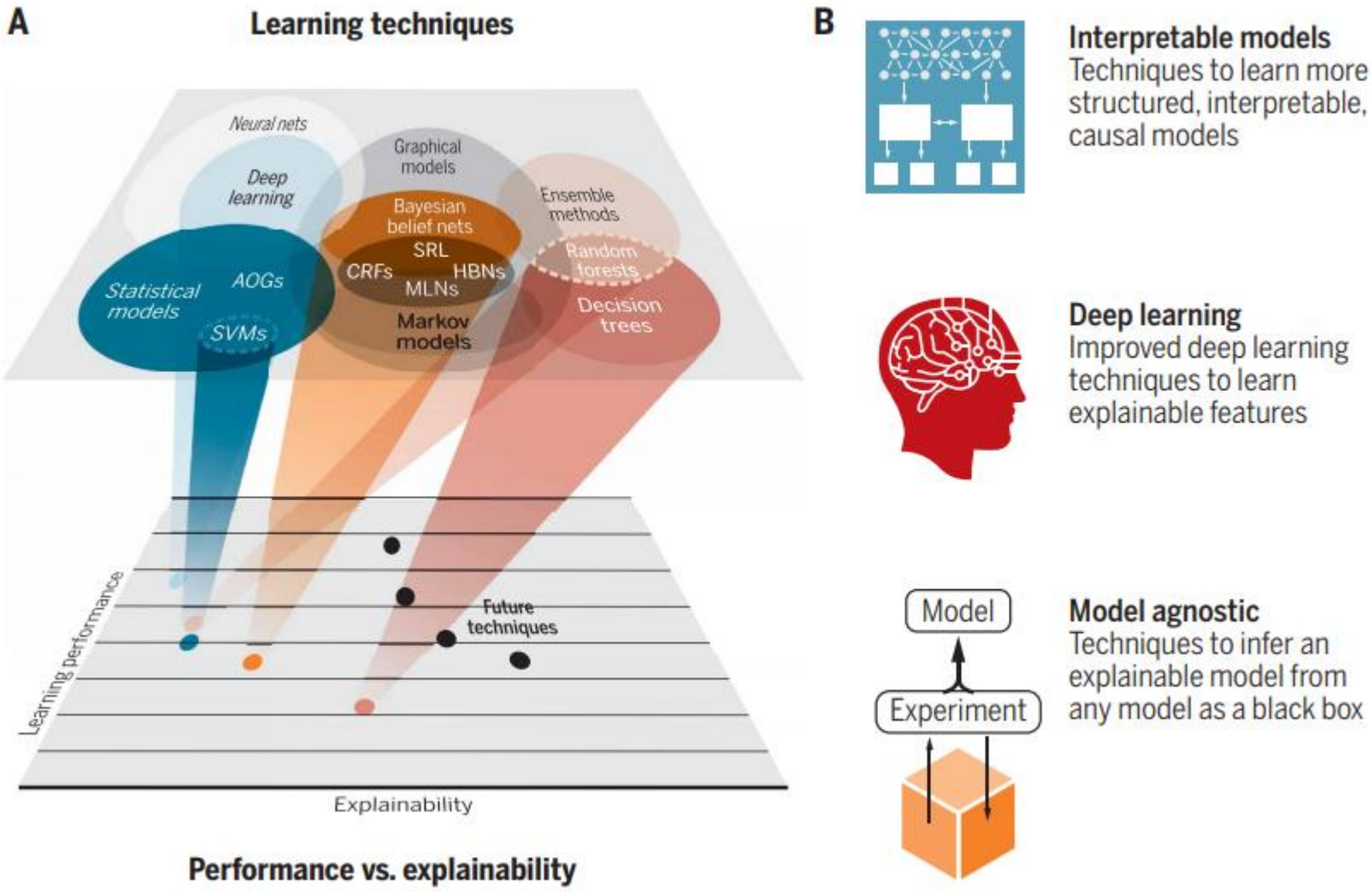


Inherent flaws can be exploited

Skewed training data creates Maladaptation

EMPOWERING EXPERTISE
iNEEJI
KAIST

Methodology in Explainable Artificiail Intelligence

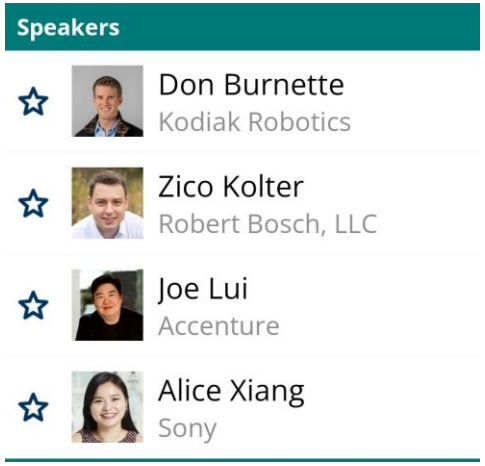


D. Gunning et al., Science Robotics, 2019

EMPOWERING EXPERTISE
iNEEJi
KAIST

The Upcoming Importance of AI Trustworthy@CES2023

- AI Trustworthy is becoming an important topic.
  - IBM: Championing Fairness, Catalyzing Growth
  - Trustworthy AI: Developing Ethical AI Solutions sponsored by Deloitte
  - Leveraging Technology for Social Impact
  - Trustworthy Robots? Building Customer Relationships in Automated Conversational Experience

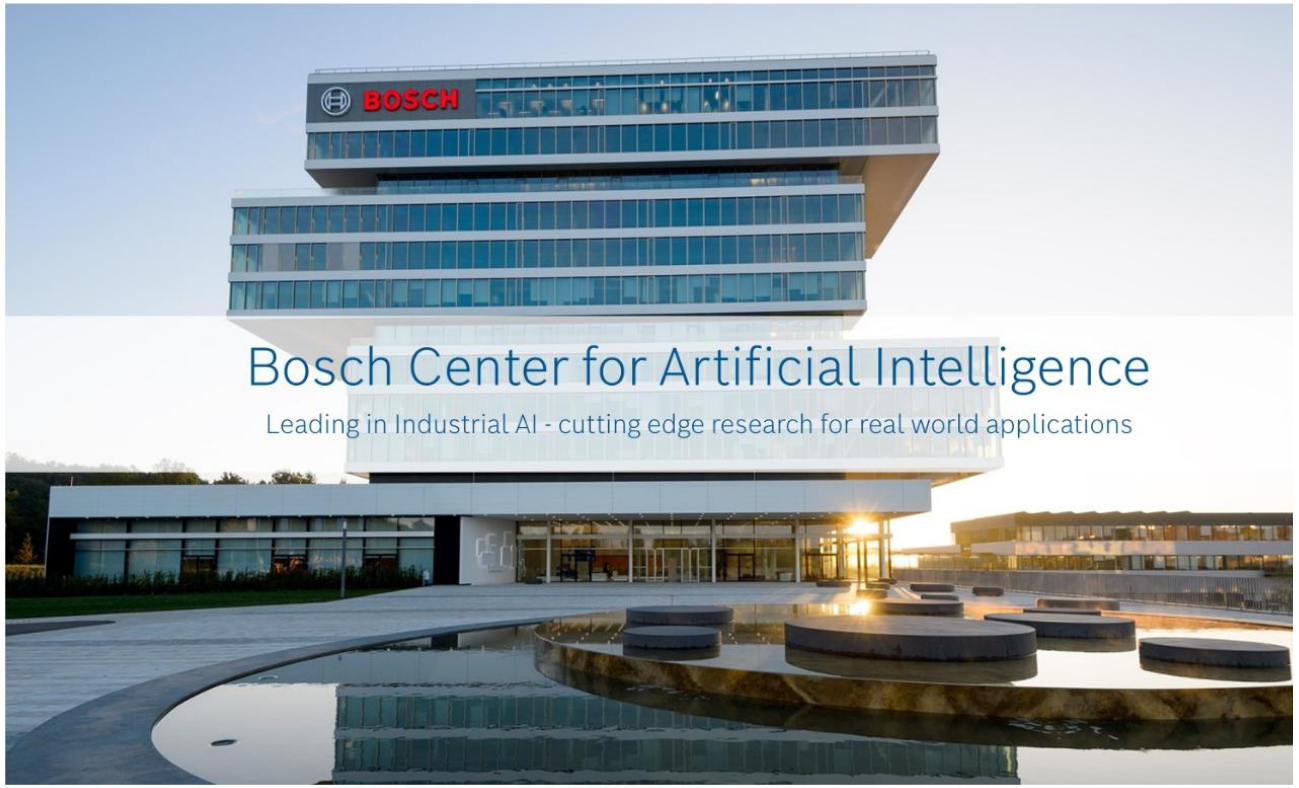


A Explainable Artificial Intelligence (XAI) in EU

(Some) Initiatives: XAI in EU

EMPOWERING EXPERTISE
iNEEJi
KAIST

Bosch Center for Artificial Intelligence



Introduction

The Bosch Center for Artificial Intelligence (BCAI) was created in 2017 out of existing competence centers to deploy cutting-edge AI technologies across Bosch products and services creating solutions that are "Invented for life". We employ roughly 270 associates worldwide, dedicated to over 185 projects within seven locations: Germany, India, United States, Israel and China.

Using data from Bosch's various business divisions, we conduct cutting-edge research that focuses on safe, secure, robust, and explainable AI. We design and implement AI for smart, connected, and autonomous technologies across Bosch business sectors.

By collaborating with international partners, we are committed to fostering scientific exchange. We actively look for opportunities to expand our research network further and to collaborate with industry thought leaders.

Fraunhofer Interpretable AI in predictive maintenance

Fraunhofer CCIT

Research Center Machine Learning

ABOUT US | RESEARCH | PROJECTS | PUBLICATIONS | TALKS | CONTACT

Homepage . Projects . HEAD

Interpretable AI in predictive maintenance

Predictive maintenance plays a central role in modern industry. In the HEAD project, Fraunhofer Institute for Production Technology IPT and Fraunhofer Institute for Algorithms and Scientific Computing SCAI are developing interpretable AI models that predict tool wear in cutting processes from acoustic emission. The objective is to increase both the robustness of the production process and the availability of the machines – while maintaining consistently high production quality. An interpretable AI model also aims to increase engineers' trust in Machine Learning by making model decisions comprehensible.

© pressmaster - adobe.stock.com

Ideally, predictive maintenance increases the

Machine Learning gains popularity in the industry continuously. A prominent field of application is predictive maintenance, in which a system's health and consequently the optimal point in time for maintenance is predicted by Machine Learning (ML) models. Ideally, predictive maintenance increases the uptime and reliability of machines and tools in production environments. Often, however, engineers question the reliability or trustworthiness of the employed ML model itself, the reason being that the model is treated as a black box by engineers having no insight into its inner workings or decision process.

Interpretable and explainable AI: A look into the black box of algorithms

Contact

Dr. Sebastian Mayer
Project Manager

Fraunhofer Institute for Algorithms and Scientific Computing SCAI
Schloss Birlinghoven 1
53757 Sankt Augustin, Germany

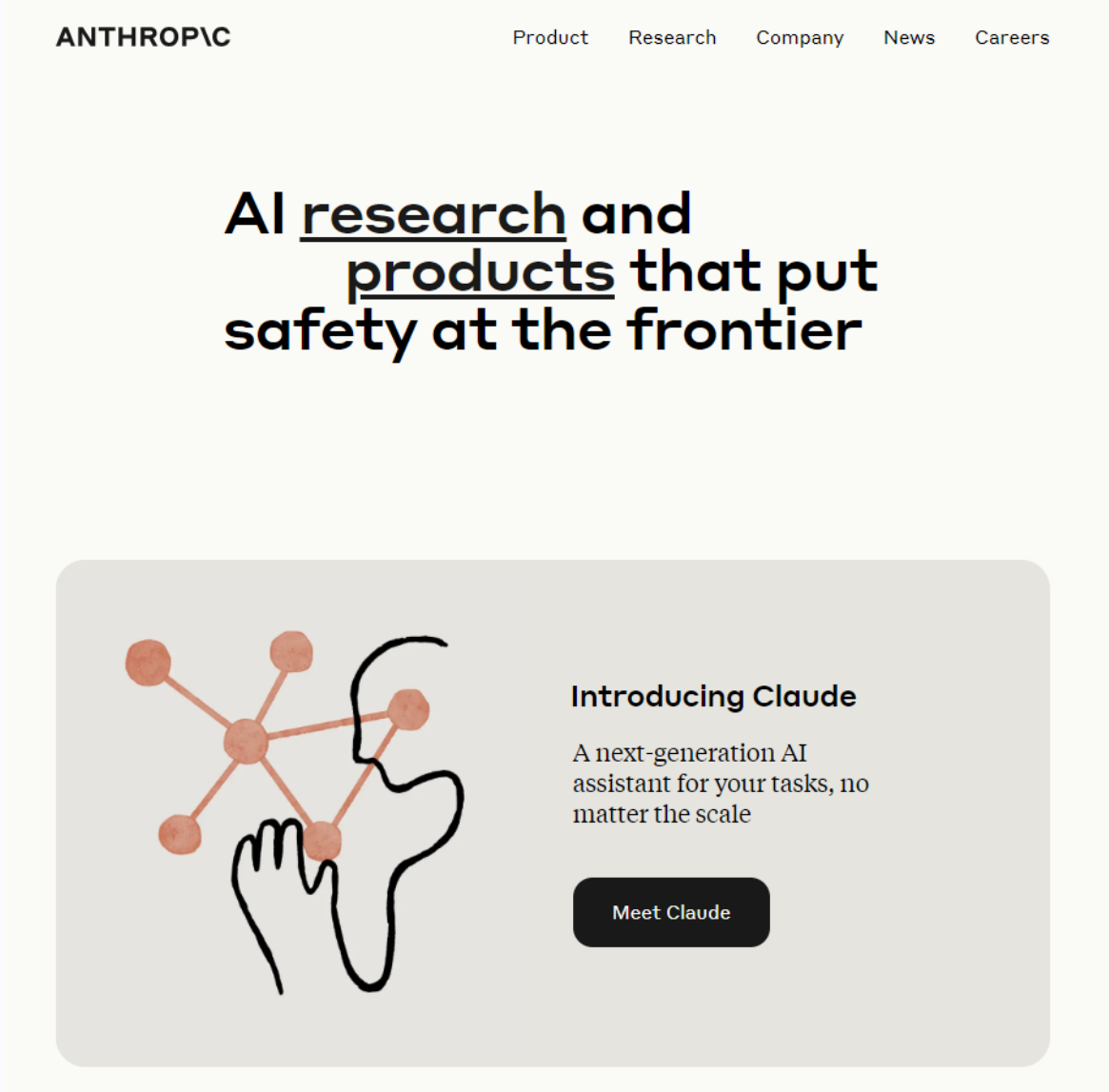
Venture Capital's Investment on XAI

1	미국	Kyndi	KYNDI	$28.5M	9	일본	Skydisc	SKYDISC	¥1.7B
2	미국	Fiddler	Fiddler	$45.2M	10	영국	Intellibonds	intellibonds	500K
3	미국	EquBot	EQUBOT	$2.1M	11	영국	Factmata	FACTMATA	$3.6M
4	미국	Convoy	CONVOY	$675,5M	12	영국	Genie AI	GENIE AI	£2M
5	미국	Zest	ZEST AI	$250M	13	영국	Imandra	IMANDRA REASONING AS A SERVICE	$12.6M
6	미국	Truera	truera	$17.3M	14	프랑스	DreamQuark	DreamQuark	$19.1M
7	미국	Digite	ANTHROPIC	$124M	15	캐나다	FAIRLY	FAIRLY	-
8	일본	Hacarus	HACARUS	$24.3M	16	캐나다	DarwinAI	DARWIN AI	$11.9M

The team analyzes the internal mechanism
of Deep Neural Network

EMPOWERING EXPERTISE
iNEEJI
KAIST

Core Views on AI Safety: When, Why, What, and How(Anthropic)



https://www.anthropic.com/index/core-views-on-ai-safety

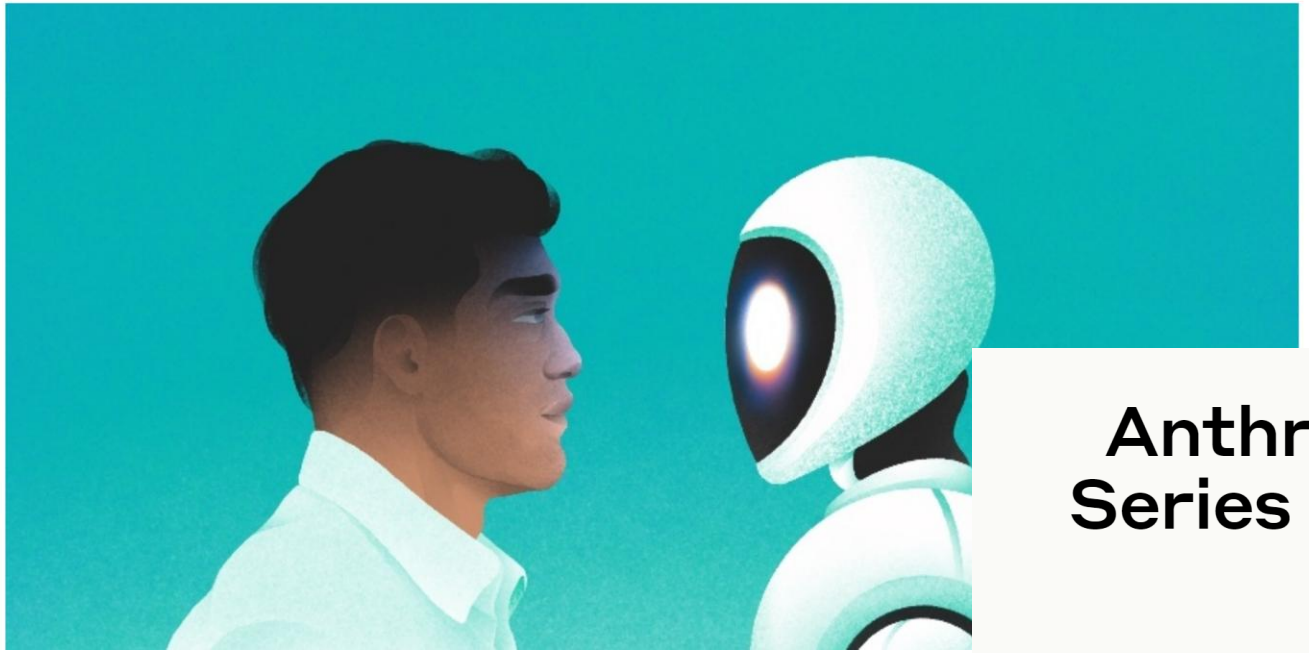
- 2022년 3월 Series B 등 빠른 성장으로 주목받고 있는 테크기업 Anthropic이 제공하는 AI 안전성에 대한 안내글
- **Anthropic의 Claude 챗봇은 윤리원칙을 고려하여 설계되었기 때문에, 윤리적 답변이 요구되는 교육, 의료, 금융 분야 등에 있어 ChatGPT보다 더 높은 활용성을 가질 것으로 기대받고 있음**
- Anthropic이 내세우는 안전성 평가 프로세스의 특징은 '경험적 접근(empirical approach)'
- 모델의 복잡성 등으로 인해 사전에 오작동을 예측하기 어려운 AI기술의 특성을 인정하고, 체계적으로 구성된 다양한 시나리오를 실험해봄으로써 가능한 오작동을 사후적으로 감지해내고 피드백하는 방식을 추구

Venture Capital's Investment on XAI

ARTIFICIAL INTELLIGENCE AND ROBOTICS

AI Startup Anthropic Raising Another $300M At $4.1B Valuation — Report

Chris Metinko March 8, 2023



Anthropic Raises $450 Million in Series C Funding to Scale Reliable AI Products

May 23, 2023 • 2 min read

EMPOWERING EXPERTISE
iNEEJI
KAIST

XAI 시장전망[XAI Statistics]

Explainable AI (XAI) Stats

Leave a Comment / By A.I. / April 21, 2023

Explainable AI (XAI) Statistics 2023

- The global XAI market is expected to grow from $450 million in 2019 to $2.6 billion by 2024.
- The top five use cases for XAI are fraud detection, risk assessment, medical diagnosis, customer service, and autonomous driving.
- The three main types of XAI are rule-based explanation, model-agnostic explanation, and interpretable machine learning.
- The three main challenges of XAI are explainability, interpretability, and trustworthiness.
- The three main benefits of XAI are accountability, fairness, and transparency.
- The three main risks of XAI are bias, explainability, and interpretability.
- XAI's three primary ethical considerations are privacy, safety, and security.
- XAI's three main regulatory considerations are the California Consumer Privacy Act (CCPA), the General Data Protection Regulation (GDPR), and the Algorithmic Accountability Act.
- XAI's three main research areas are interpretable machine learning, explainable AI, and adversarial machine learning.
- The three main tools for XAI are LIME, SHAP, and DeepLIFT.
- The three main challenges for the future of XAI are explainability, interpretability, and trustworthiness.

- **AI Trustworthy 시장은 2019년 $1.45B에서 2024년 $7.2B 수준으로 성장할 것으로 예측**
- **XAI 시장은 2019년 $450M에서 2024년 $2.6B 수준으로 성장할 것으로 예측**
- 테크기업 등 11개 웹사이트(IBM, Google AI, OpenAI, Microsoft 등)의 정보를 바탕으로 XAI 시장 현황 요약
- XAI 기술이 당면한 3가지 주요 challenge 중 하나로 신뢰성을 제시
- XAI가 고려해야할 주요 윤리로 privacy, safety, security를 제시
- 이처럼, 최근 신뢰 가능한 AI를 위해 설명의무 등이 본격적으로 법제화됨에 따라, AI분야에서 선도적인 위치에 있는 기업들 모두 **XAI기술의 안전성, 신뢰성, 윤리성**에 주목하고 있음

https://aistratagems.com/explainable-ai-xai-stats/

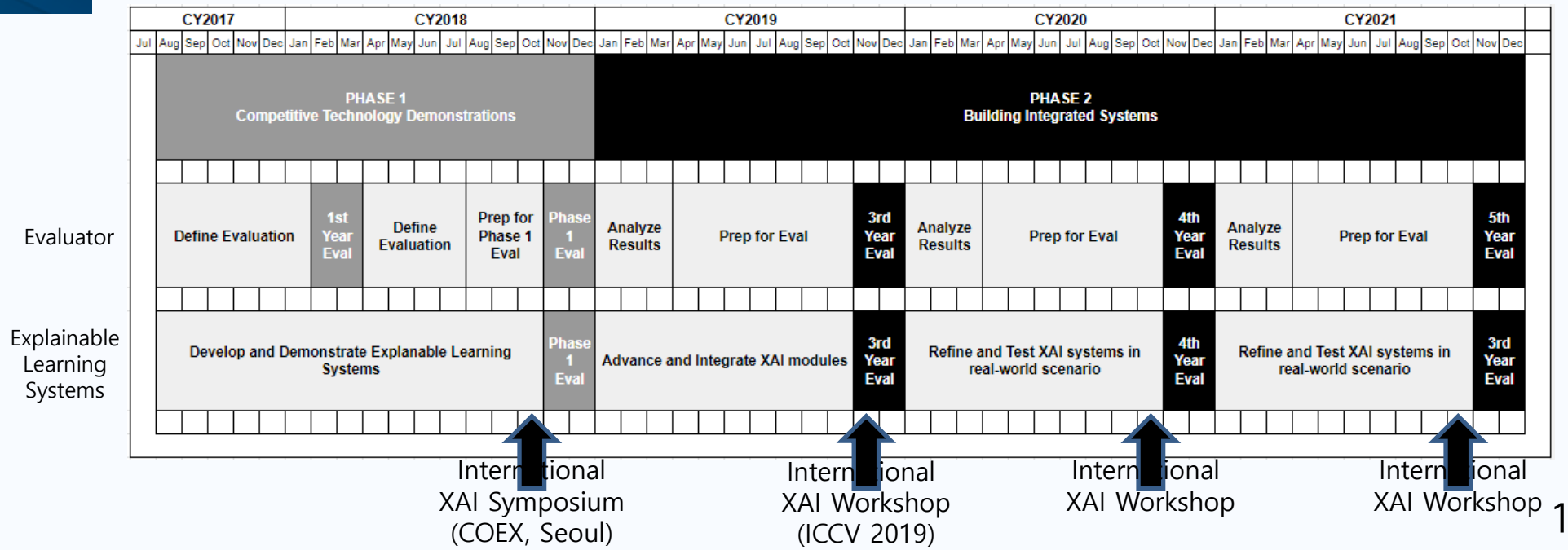
Explainable AI Program in Korea – Part I

Goal → Human-level Learning and Inference to overcome the limitations of Deep Neural Networks

AS-IS (Deep Learning)	TO-BE (Human-level Learning/Inference)
• It is hard to know the decision, so called **Blackbox** model • It does not work well when we **do not have enough training data**	• **Explainable learners** which can provide the reasons of decisions • Learning explainable models even with **data deficient environment**

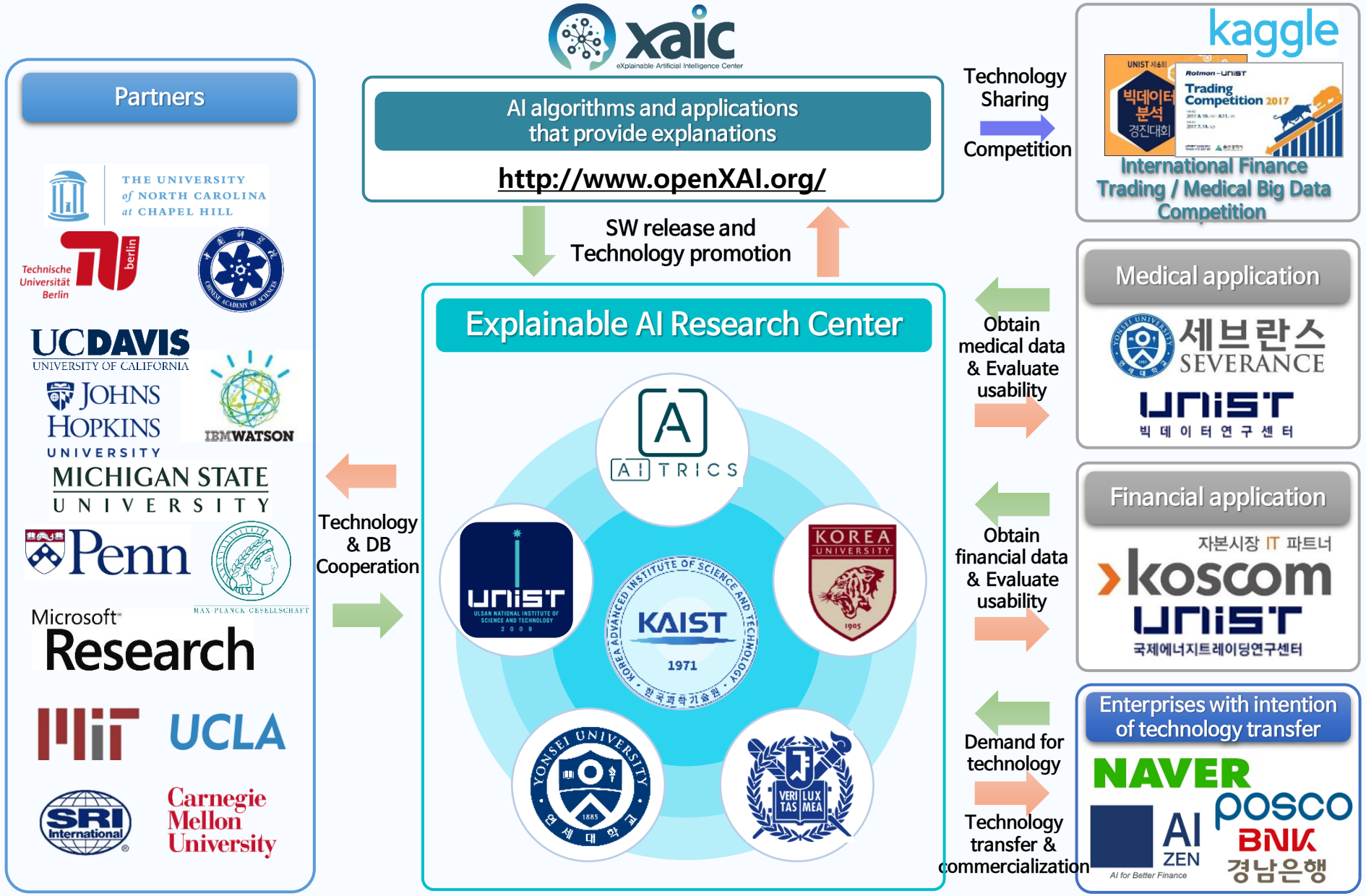
Fund → Institute of Information & Communication Technology Promotion (IITP) under Ministry of Science and ICT (MSICT) as part of Innovative Growth Engine Project

Period → July 2017 ~ December 2021 (54 months)



Explainable AI Program in Korea – Part I

July 2017 ~ December 2021 (54 months)

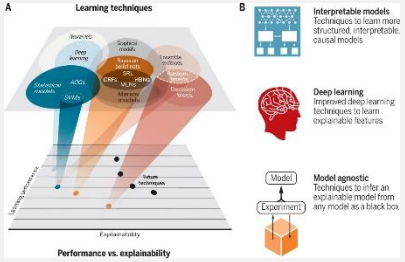


Explainable AI Program in Korea – Part I

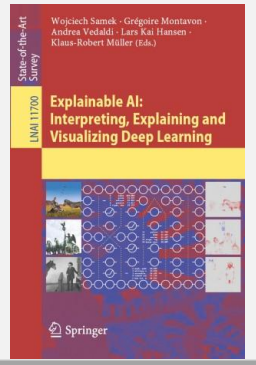
Research Results

AI Top Conference Papers (ICML, NeurIPS, AAAI, …) 87

Top Journal Papers (Science Robotics, …) 45



Book Edited (Explainable AI: Springer)



Technology Transfer

Patents (Registration) 37(2)

Industrial Projects 11

Manufacturing	Healthcare	Finance	Mobile
POSCO Steel; SAMSUNG 삼성전자 Semiconductor	세브란스병원 SEVERANCE HOSPITAL; 강남세브란스병원 GANGNAM SEVERANCE HOSPITAL; 국민건강보험 일산병원 National Health Insurance Service Ilsan Hospital; SNUH 분당서울대학교병원	IBK 기업은행; 신한은행; KEB 하나은행	NAVER; SAMSUNG 삼성전자
Process Explain	ICU monitoring	Credit Rating	Robust Generation

Open Source/Meetings

Open Source Projects 44
github.com/OpenXAIProject

Online Tutorial 31

Open Workshop 10

International Gathering 3

Explainable AI Program in Korea – Part I

2018 International XAI Symposium

KDD2020 Tutorial on

Interpreting and Explaining Deep Neural Networks: A Perspective on Time Series Data

2020 KDD Tutorial http://xai.kaist.ac.kr/Tutorial/2020/

ICCV 2019
Seoul, Korea

2019 ICCV Workshop on

Interpreting and Explaining Visual Artificial Intelligence Models

Saturday, November 2nd, 2019

@ COEX 318AB, Seoul, Korea

2019 International XAI Workshop
http://xai.kaist.ac.kr/workshop/2019/

Trevor Darrell
Professor
UC Berkeley

Wojciech Samek
Head of Machine Learning Group
Fraunhofer Heinrich Hertz Institute

David Bau
PhD student
MIT

Ludwig Schubert
Software Engineer
OpenAI

EMPOWERING EXPERTISE
iNEEJI
KAIST

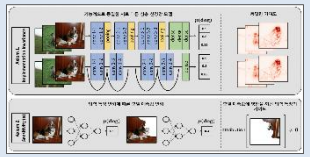
Explainable AI Program in Korea – Part II

April 2022 ~ December 2026 (57 months)

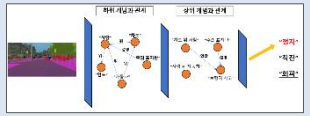
[1] Plug and Play XAI

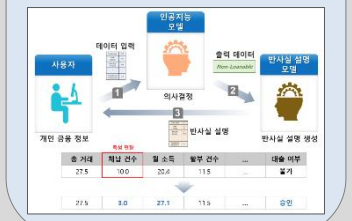
AI Model Explanation

Input Explanation



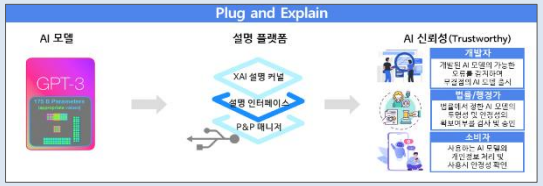
Internal Explanation





<플러그앤플레이 설명가능 인공지능 개발을 위한 기관별 역할>

AI Model Plug and Play XAI



General Method to Explain AI Models

Test Applications

Tech Collaborations

Explaining Internal Decision of AI

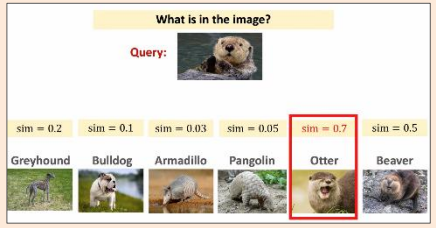
Collaobrators

Industries: POSCO, NAVER, linewalks, SAMSUNG, neurophet, SK telecom, HYUNDAI, SAMSUNG, 삼성서울병원, INEEJI, 하나금융TI, kakaobrain

International Collaboration: EPFL, 上海科技大学 ShanghaiTech University, UCL, PHBS 北京大学汇丰商学院, TUM, VinAI RESEARCH, BIFOLD, Fraunhofer HHI, IBM, UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN, UMKC, University of Nottingham UK | CHINA | MALAYSIA

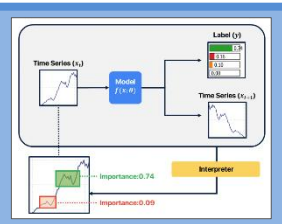
[2] Causal Inference Explanation

Causality Discovery



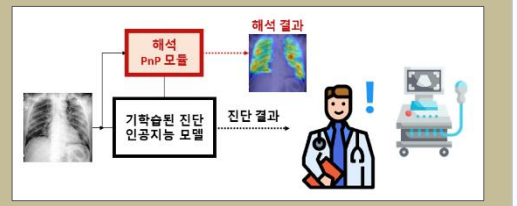
Few Shot based XAI LEarning

[3] XAI for General Audience



Explanations with Commonsense Knowledge for General Public

[4] XAI for Experts

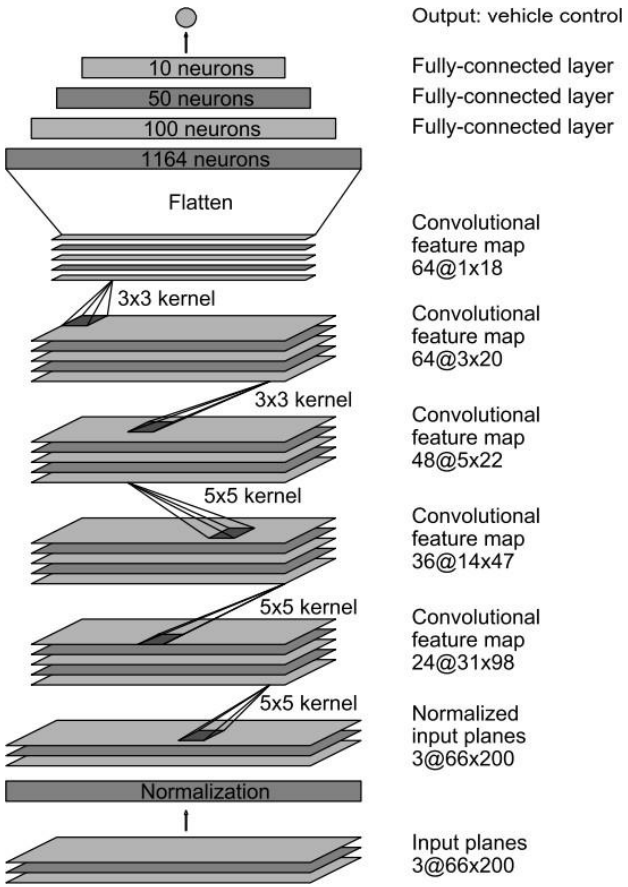


Explanations for Expert to Improve the Work Performance

Agenda

- Motivation
- **Case Studies**
- Social Impact and Regulations
- Trends in XAI Research
- Appendix: What is inside of Deep Neural Networks

Explainable AI for PilotNet of NVIDIA

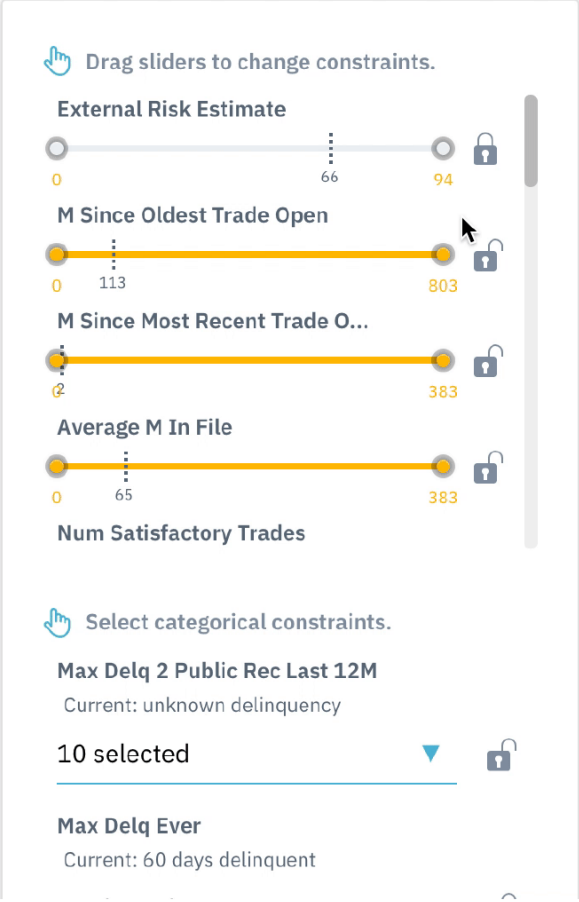


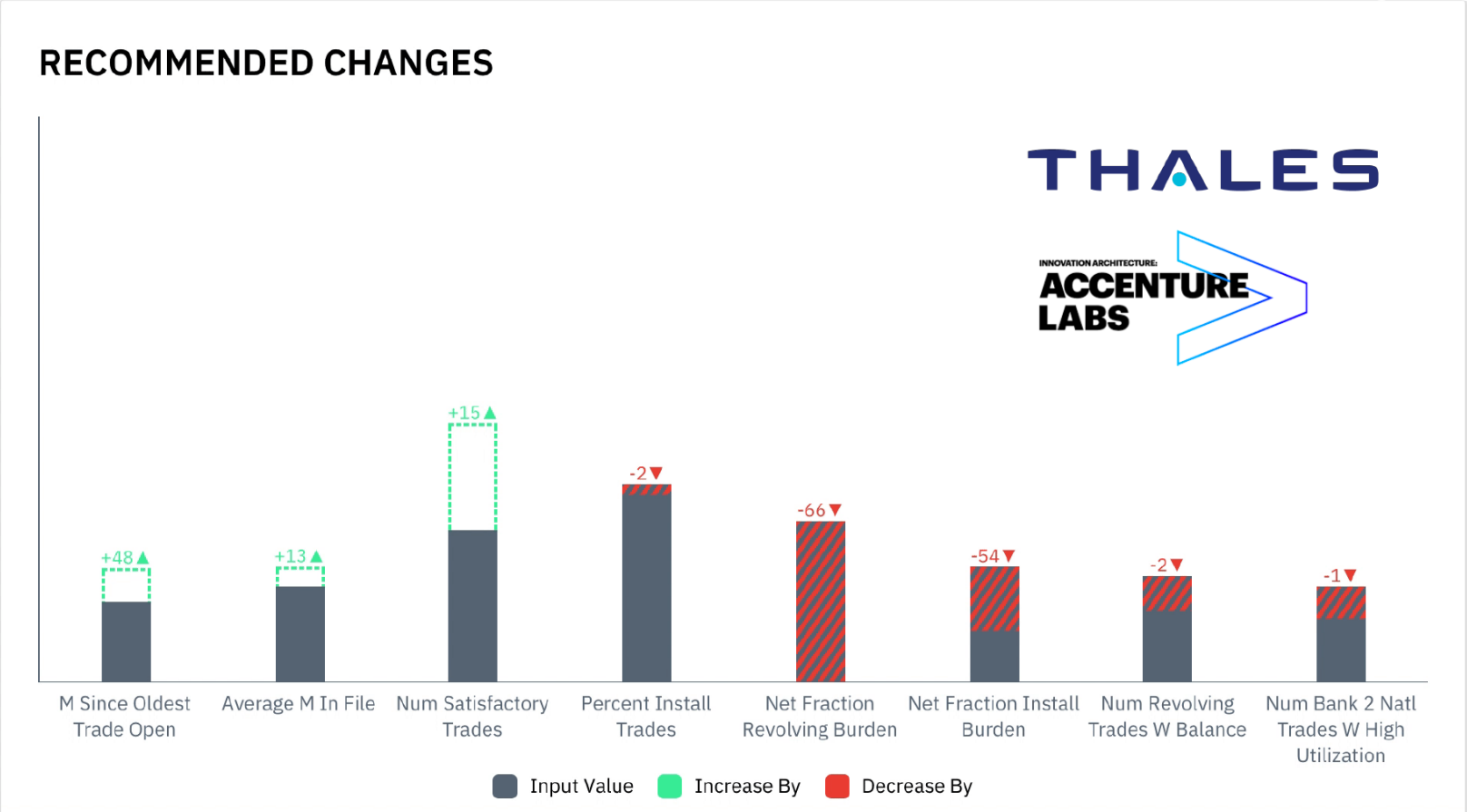
PilotNet 구조도
NVIDIA/Google 연구팀, 2017년

Explaining the decisions of PilotNet

EMPOWERING EXPERTISE
iNEEJi
KAIST

Counterfactual Explanations for Credit Decisions - Finance

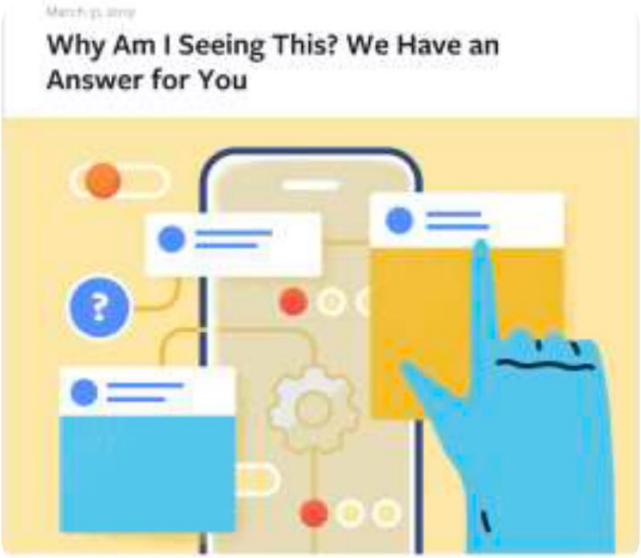


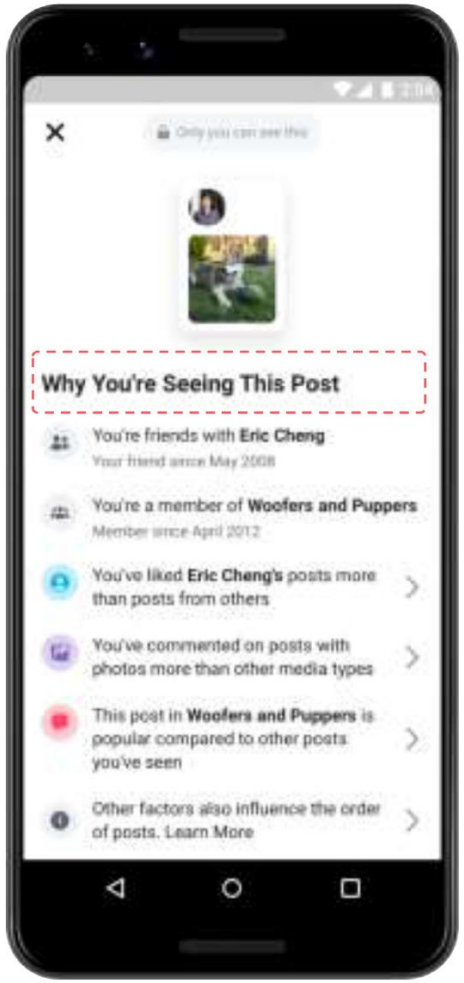


Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, Freddy Lécué: Interpretable Credit Application Predictions With Counterfactual Explanations. FEAP-AI4fin workshop, NeurIPS, 2018.

Credit: F. Lecue et al., Explainable AI – XAI Tutorial, AAAI 2021

Explaining Recommendation – Social Media





Challenge: How to establish trust between Social Media and their users? Explaining post / news recommendation is crucial for users to engage with content providers.

AI Technology: Artificial Neural Networks.

XAI Technology: Recommendation-based

KDD 2019 Tutorial on Explainable AI in Industry - https://sites.google.com/view/kdd19-explainable-ai-tutorial

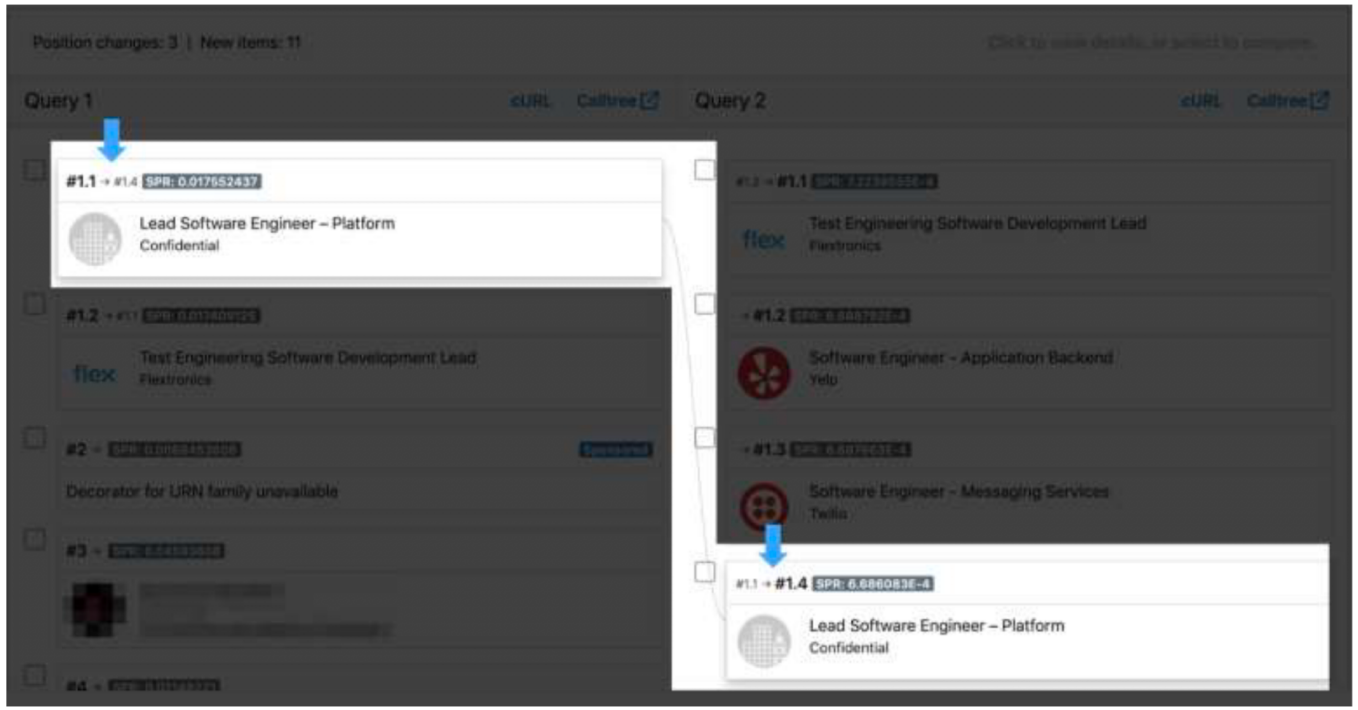
Credit: F. Lecue et al., Explainable AI – XAI Tutorial, AAAI 2021

Relevance Debuggin and Explaining – Industry Agnostic

Challenge: A Machine Learning system can fail in many different points e.g., data features selection, construction, inconsistencies. How to debug bad performance in machine learning models and prediction?

AI Technology: Artificial Neural Networks.

XAI Technology: Model / Prediction comparison



KDD 2019 Tutorial on Explainable AI in Industry - https://sites.google.com/view/kdd19-explainable-ai-tutorial

Credit: F. Lecue et al., Explainable AI – XAI Tutorial, AAAI 2021

Explaining Breast Cancer Survival Rate Prediction - Health

predict
breast cancer

Age at diagnosis: – 69 +
Age must be between 25 and 85

Post Menopausal? Yes | No | Unknown

ER status: Positive | Negative

HER2 status: Positive | Negative | Unknown

Ki-67 status: Positive | Negative | Unknown
Positive means more than 10%

Tumour size (mm): – 7 +

Tumour grade: 1 | 2 | 3

Detected by: Screening | Symptoms | Unknown

Positive nodes: – 2 +

Micrometastases: Yes | No | Unknown
Enabled when positive nodes is zero

Results

Table | Curves | Chart | Texts | Icons

New recording

These results are for women who have already had surgery. This table shows the percentage of women who survive at least 5 | 10 | 15 years after surgery, based on the information you have provided.

Treatment	Additional Benefit	Overall Survival %
Surgery only	-	72%
+ Hormone therapy	0%	72%

If death from breast cancer were excluded, 82% would survive at least 10 years.

Show ranges? Yes | No

Challenge: Predict is an online tool that helps patients and clinicians see how different treatments for early invasive breast cancer might improve survival rates after surgery.

AI Technology: competing risk analysis

XAI Technology: Interactive explanations, Multiple representations.

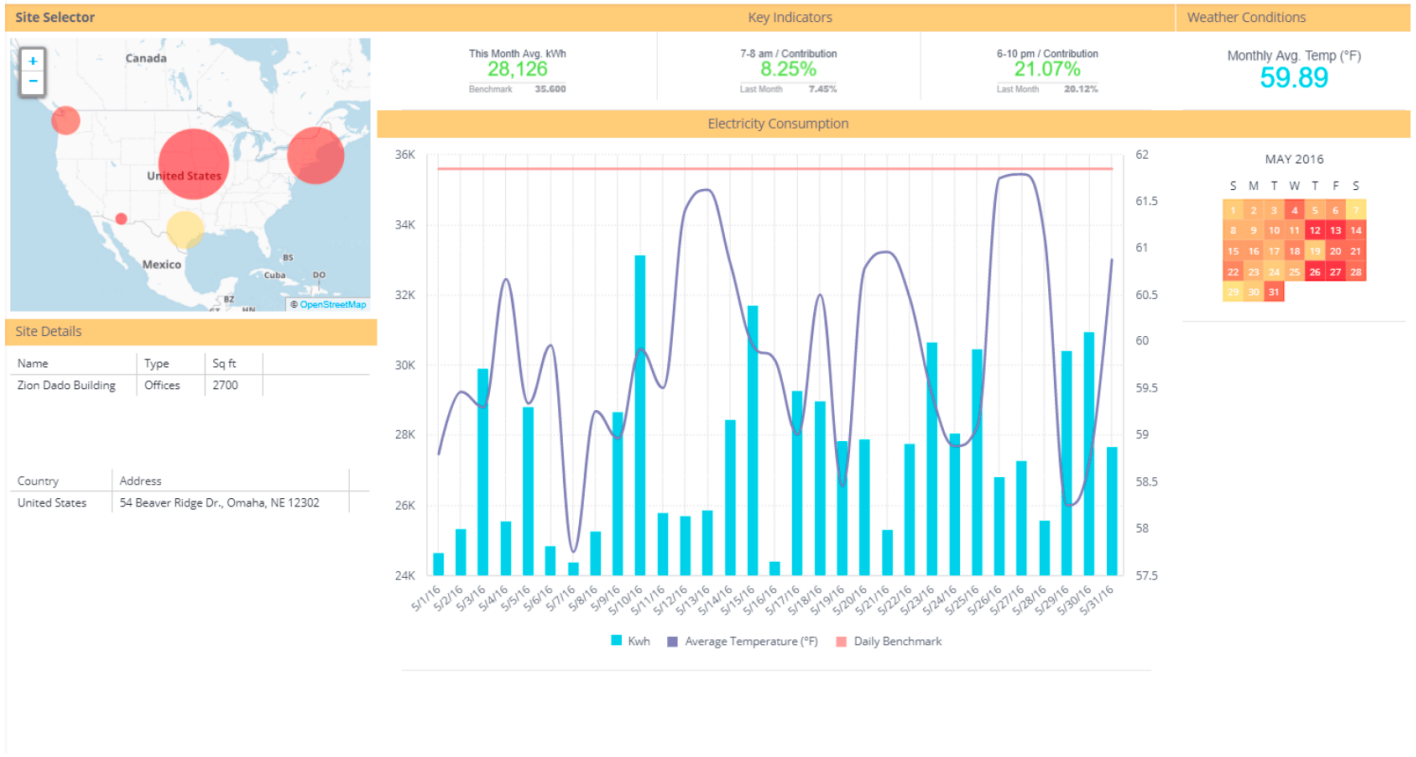
David Spiegelhalter, Making Algorithms trustworthy, NeurIPS 2018 Keynote

predict.nhs.uk/tool

Credit: F. Lecue et al., Explainable AI – XAI Tutorial, AAAI 2021

EMPOWERING EXPERTISE
iNEEJI
KAIST

Explaining Energy Consumption – A Global Perspective



Challenge: Predicting energy consumption is crucial to satisfy high-demand. However some demands might be difficult to forecast, particularly in case of abnormal events. How to augment energy consumption data with open / event data to reach better accuracy and explainability of out-of-distribution demand.

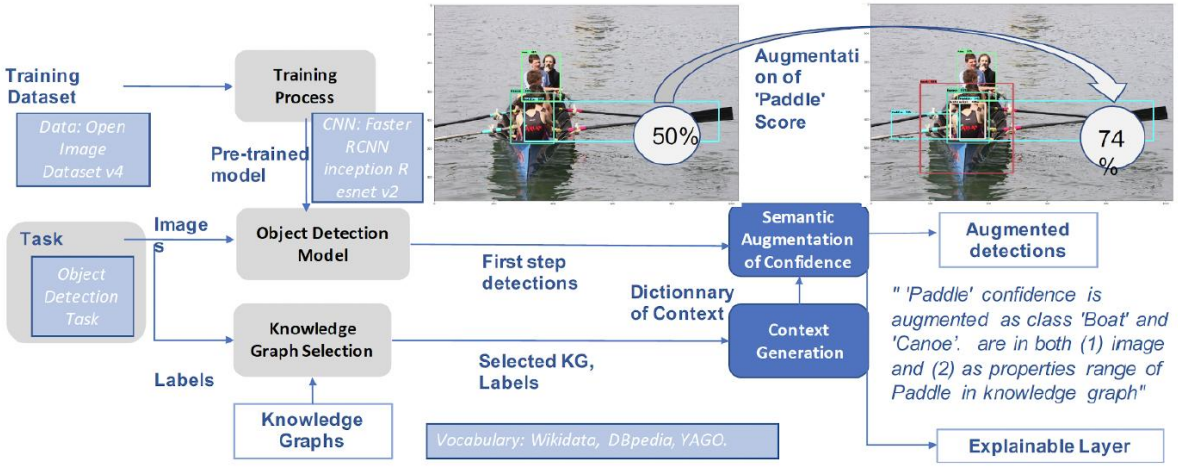
AI Technology: Artificial Neural Network

XAI Technology: Artificial Neural Network, Data Augmentation, Knowledge Graphs

Credit: F. Lecue et al., Explainable AI – XAI Tutorial, AAAI 2021

EMPOWERING EXPERTISE
INEEJI
KAIST

Explainable Boosted Object Detection – Industry Agnostic



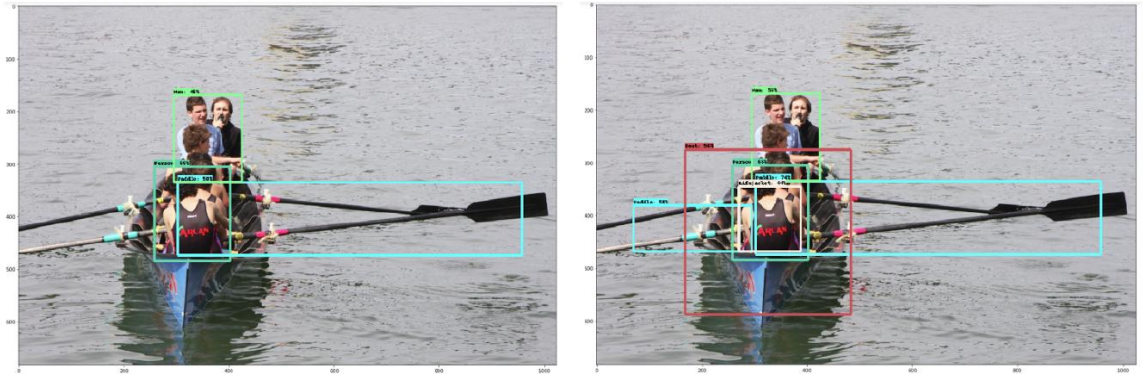


Fig. 2. Left image: results from baseline Faster RCNN: Paddle: 50% confidence, Person: 66%, Man: 46%. Right image: results from the semantic augmentation: **Paddle: 74%** confidence, Person: 66%, **Man: 56%**, **Boat: 58% with explanation: Person, Paddle, Water as part of the context in the image and knowledge graph of concept Boat**. (color print).

Challenge: Object detection is usually performed from a large portfolio of Artificial Neural Networks (ANNs) architectures trained on large amount of labelled data. Explaining object detections is rather difficult due to the high complexity of the most accurate ANNs.

AI Technology: Integration of AI related technologies i.e., Machine Learning (Deep Learning / CNNs), and knowledge graphs / linked open data.

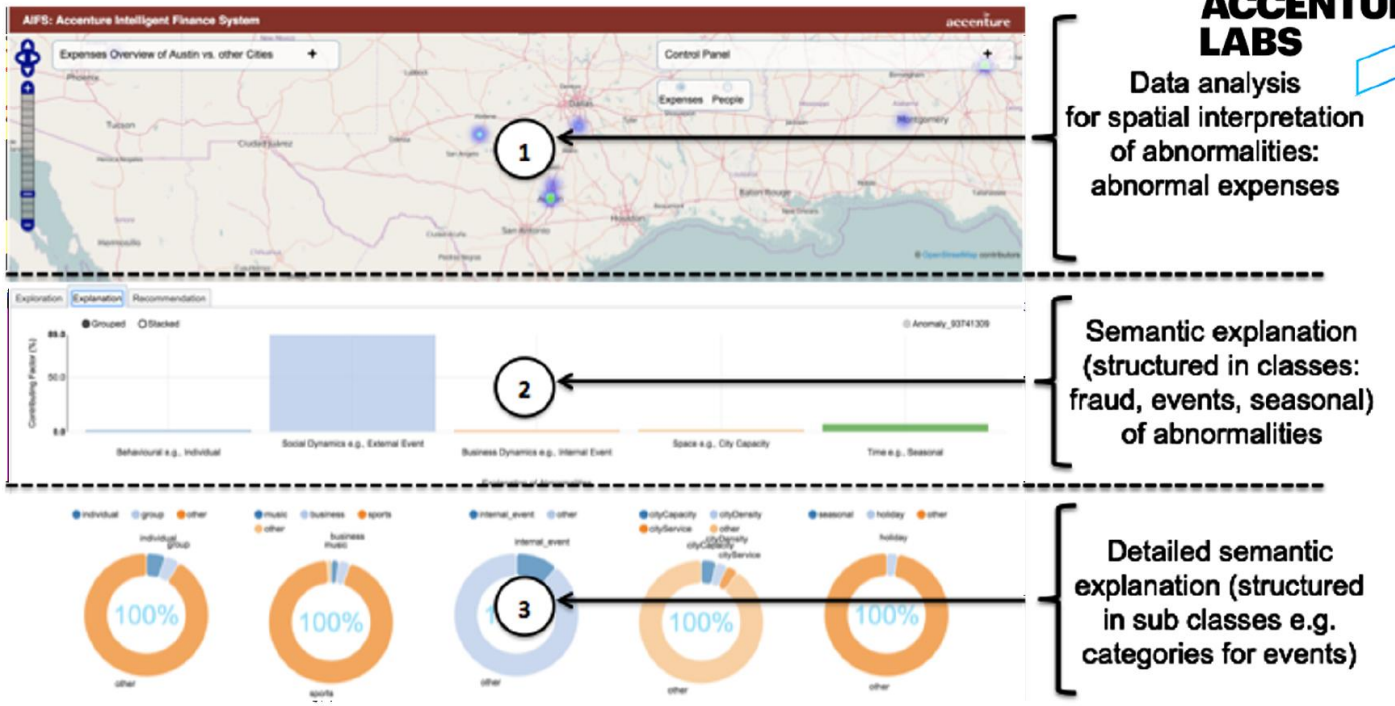
XAI Technology: Knowledge graphs and Artificial Neural Networks

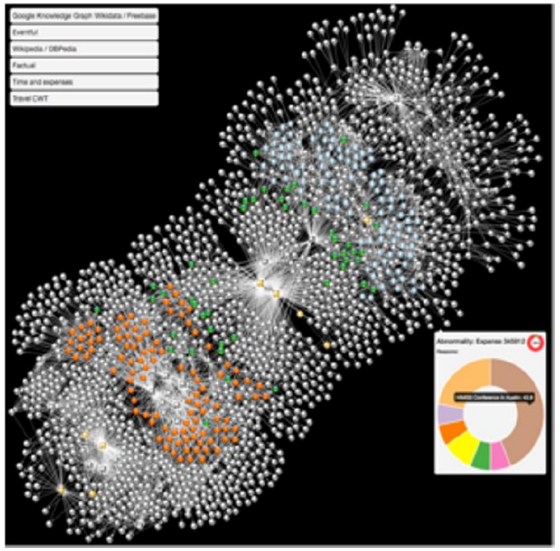
THALES

Credit: F. Lecue et al., Explainable AI – XAI Tutorial, AAAI 2021

EMPOWERING EXPERTISE
iNEEJI
KAIST

Explainable Anomaly Detection - Finance





Freddy Lécué, Jiewen Wu: Explaining and predicting abnormal expenses at large scale using knowledge graph based reasoning. J. Web Sem. 44: 89-103 (2017)

Challenge: Predicting and explaining abnormally employee expenses (as high accommodation price in 1000+ cities).

AI Technology: Various techniques have been matured over the last two decades to achieve excellent results. However most methods address the problem from a statistic and pure data-centric angle, which in turn limit any interpretation. We elaborated a web application running live with real data from (i) travel and expenses from Accenture, (ii) external data from third party such as Google Knowledge Graph, DBPedia (relational DataBase version of Wikipedia) and social events from Eventful, for explaining abnormalities.

XAI Technology: Knowledge graph embedded Ensemble Learning . **Video**: https://www.dropbox.com/s/sst232gu0yeqy21/IUI-2017-Final.mp4?dl=0

Credit: F. Lecue et al., Explainable AI – XAI Tutorial, AAAI 2021

iNEEJI EMPOWERING EXPERTISE
KAIST

Explaining Visual Question Answering – Industry Agnostic

Tabular QA

Rank	Nation	Gold	Silver	Bronze	Total
1	India	102	58	37	197
2	Nepal	32	10	24	65
3	Sri Lanka	16	42	62	120
4	Pakistan	10	36	30	76
5	Bangladesh	2	10	35	47
6	Bhutan	1	6	7	14
7	Maldives	0	0	4	4

Q: How many medals did India win?
A: 197

Neural Programmer (2017) model
33.5% accuracy on WikiTableQuestions

Visual QA

Q: How symmetrical are the white bricks on either side of the building?
A: very

Kazemi and Elqursh (2017) model.
61.1% on VQA 1.0 dataset
(state of the art = 66.7%)

Reading Comprehension

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager

Q: Name of the quarterback who was 38 in Super Bowl XXXIII?
A: John Elway

Yu et al (2018) model.
84.6 F-1 score on SQuAD (state of the art)

Challenge: What is the robustness of Visual Question Answering models? What is the impact of semantics?

AI Technology: Artificial Neural Networks.

XAI Technology: Integrated Gradients

Google AI

Q: How symmetrical are the white bricks on either side of the building?
A: very

Q: How asymmetrical are the white bricks on either side of the building?
A: very

Q: How big are the white bricks on either side of the building?
A: very

Q: How fast are the bricks speaking on either side of the building?
A: very

What is the **man** doing? → What is the **tweet** doing?
How many **children** are there? → How many **tweet** are there?

VQA model's response remains the same 75.6% of the time on questions that it originally answered correctly

Source: Explainable AI in Industry. KDD 2019 Tutorial. Ankur Taly, Mukund Sundararajan, Kedar Dhamdhere, Pramod Mudrakarta

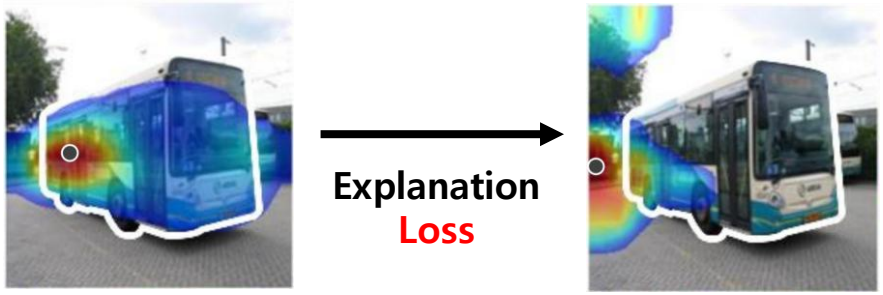
Credit: F. Lecue et al., Explainable AI – XAI Tutorial, AAAI 2021

EMPOWERING EXPERTISE
iNEEJI
KAIST

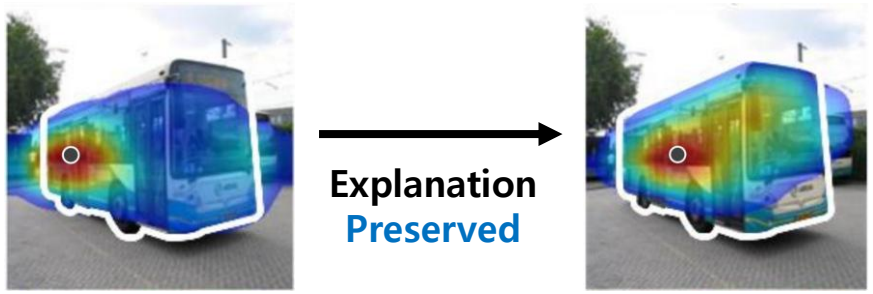
Attribution Interpretability Preservation in Network Compression

Maintaining explainability while compressing neural networks

Prior art [UMD/NEC Labs, ICLR 2017] → **SOTA [KAIST, NeurIPS 2020]**



Before — **Explanation Loss** — After

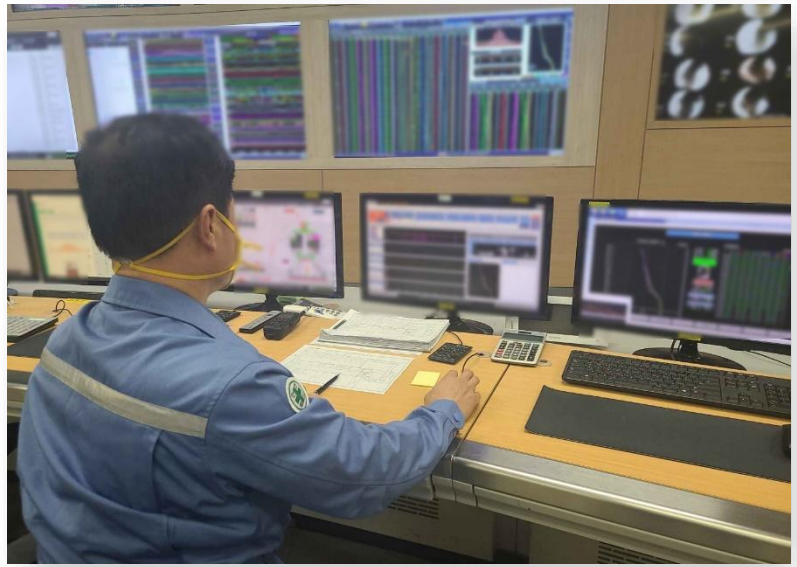


Before — **Explanation Preserved** — After

[1] H. Li et al., "Pruning filters for efficient convnets." *ICLR,* 2017.
[2] G. Park et al., "Attribution Preservation in Network Compression for Reliable Network Interpretation." *NeurIPS*, 2020.

EMPOWERING EXPERTISE
iNEEJi
KAIST

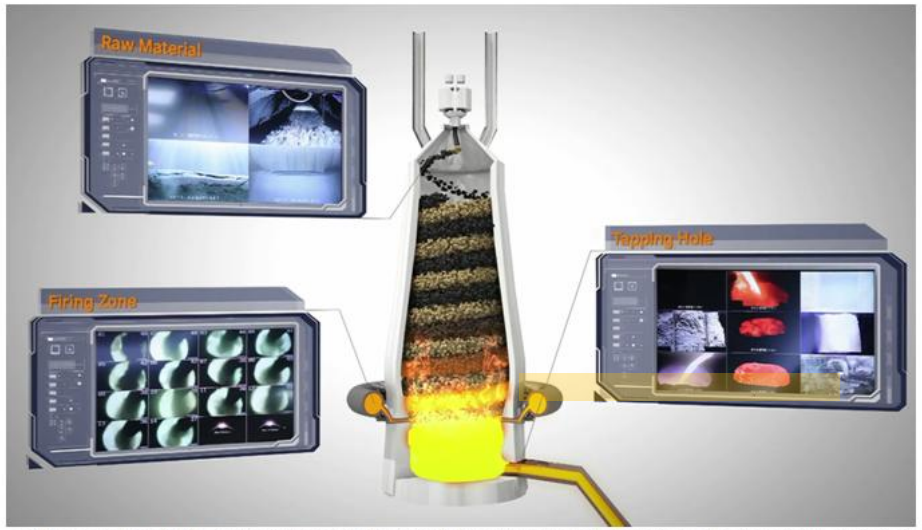
POSCO Smart Blast Furnace



포스코 포항제철소 고로에 '설명 가능 인공지능' 기술 적용한다

2고로와 3고로에 XAI 확대 적용해 품질, 생산성 향상 기대

김재광 기자(=경북) stmkjki@polinews.co.kr 등록 2020.02.10 16:16:16



▲ 스마트 고로 내부 일러스트. 이미지 빅데이터 기반 인공지능으로 고로 상부에서는 통기성, 중부에서는 연소성, 하부에서는 용선온도를 스스로 제어해 쇳물을 생산한다. <사진제공=포스코>

포스코가 '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.

The First International Standard on XAI initiated from Korea

ISO/IEC JTC 1/SC 42 **N 782**

ISO/IEC JTC 1/SC 42 "Artificial intelligence"
Secretariat: **ANSI**
Committee Manager: **Benko Heather Ms.**

Official Form 4 - NP - Information technology -- Artificial intelligence -- Objectives and methods for explainability of ML models and AI systems

Document type	Related content	Document date	Expected action
Ballot / Reference document	Project: ISO/IEC NP TS 6254 Ballot: ISO/IEC NP TS 6254 (restricted access)	2020-11-16	VOTE by 2021-02-09

Description

SC 42 N 782 is a NP for ballot to approve the proposal "Information technology -- Artificial intelligence -- Objectives and methods for explainability of ML models and AI systems" and has also been issued via the electronic balloting procedure with the ballot opening on 17 November 2020. SC 42 N 711 is the Draft Document related to the Form 4 contained in SC 42 N 782. Votes should be submitted by 9 February 2021. Any comments submitted with votes should be provided in the standard format.

韓, 인공지능 국제표준화 주도권 강화한다

조명의 기자 | 승인 2020.11.04 17:45

[테크월드=조명의 기자]

인공지능(AI) 국제 표준화 회의에서 우리나라는 한국판 뉴딜 정책의 핵심인 인공지능의 표준화를 위해 인공지능 데이터의 프레임워크와 서비스 생태계, 머신러닝 데이터 품질, 인공지능 신뢰성 등에서 국제표준 논의를 주도했다.

국립전파연구원
National Radio Research Agency

과학기술정보통신부 국립전파연구원과 산업통상자원부 국가기술표준원은 10월 20일부터 30일까지 온라인으로 개최된 '제6차 인공지능 국제 표준화 회의(ISO/IEC JTC1/SC42)'에 산·학·연·관 전문가 33명이 우리나라 대표단으로 참가한 결과, ▲인공지능 서비스 생태계 표준화를 위한 신규 특별작업반 설립, ▲설명가능한 인공지능(XAI)의 신규 국제표준 제안(NP), ▲신러닝 데이터 품질 신규 국제표준안 작업 지속 등의 성과를 거뒀다고 밝혔다.

아울러 우리 대표단은 지난 1년 동안 우리나라가 주도한 인공지능 데이터 특별작업반 운영 결과를 공유했으며, '인공지능 데이터 프레임워크'에 관한 신규 국제표준안 제안을 위해 논의를 이어갈 계획이다.

이외에도 '설명가능한 인공지능 시스템 개발 지침'을 신규 표준화 과제로 제안(서울시립대 이재호 교수)했으며, 2021년 초에 신규 과제로 최종 채택될 예정이다.

설명가능한 인공지능은 예를 들면, 인공지능을 이용한 금융대출 심사결과에 대해 그 결정 과정과 이유 등을 소비자에게 설명해 주는 등 인공지능의 신뢰성을 높이는 기술이다.

Agenda

- Motivation
- Case Studies
- **Social Impact and Regulations**
- Trends in XAI Research
- Appendix: What is inside of Deep Neural Networks

주요국 인공지능 전략 분석(NIA)

□ 참고 : 분석 국가별 주요 AI 이니셔티브

국가	주요 AI 이니셔티브	분야
중국	('17) 차세대 AI 발전계획	종합
	('17) 차세대 AI 개발을 위한 3개년 실행계획	종합
	('18) 대학 AI 혁신 행동계획	인재·교육
	('21) 차세대 AI 윤리 규범	신뢰가능 AI
	('22) 인터넷 정보서비스 알고리즘 추천 관리 규정	신뢰가능 AI
EU	('18) EU를 위한 AI	종합
	('18) AI에 관한 협력 계획	종합
	('20) 데이터법 초안	인프라
	('21) 핵심 원자재에 관한 실행 계획	인프라
	('21) AI에 관한 협력 계획 업데이트	종합
	('21) AI법 초안	신뢰가능 AI
	('21) EU 반도체법 초안	인프라
호주	('18) 호주 AI 윤리 프레임워크	신뢰가능 AI
	('19) AI 로드맵	종합
	('21) 호주 AI 실행계획	종합
	('21) 호주 데이터 전략	인프라
	('21) 호주 AI 윤리 원칙	신뢰가능 AI
일본	('17) AI 기술 전략	혁신·연구개발
	('19) AI 전략 2019	종합
	('19) 인간 중심의 AI 사회원칙	신뢰가능 AI
	('22) AI 전략 2022	종합
	('22) 반도체 산업 부활을 위한 기본전략	인프라
인도	('18) AI 국가전략	종합
	('20) 국가 교육 정책 2020	인재·교육
	('21) 책임 있는 AI를 위한 원칙	신뢰가능 AI
	('21) 인도-미국 AI 이니셔티브	국제협력

- 미국 외 중국, EU 회원국, 호주, 일본, 인도 등 선도적 위치에 있는 국가들 모두 신뢰가능 AI를 위한 원칙을 세우고 법제화하고 있음
- 중국: '신뢰할 수 있는 AI에 관한 백서'(`21.7), '인터넷 정보서비스 알고리즘 추천 관리 규정'(`22.3) 등
- EU: 'AI법'(`21.4), 'AI민사책임지침'(`22.9) 등
- 호주: '호주의 AI 윤리 프레임워크에 관한 논의 보고서'(`18). '호주 AI 실행계획-8대 AI 윤리원칙'(`22.6)
- 일본: '인간 중심의 AI 사회원칙'(`19.3), 'AI 원칙 이행을 위한 거버넌스 가이드라인 1.0'(`21.7)
- 인도: '책임있는 AI를 위한 원칙'(`21)

https://www.nia.or.kr/site/nia_kor/ex/bbs/View.do?cbIdx=25932&bcIdx=25130&parentSeq=25130

Three Pillars of America's AI Strategy(Stanford)

Table 1. Summary of Implementation of AI Leadership Order, AI in Government Order, and AI in Government Act of 2020

	Implemented	Unknown Implementation	Not Implemented
Total	**27%**	**58%**	**16%**
AI Leadership Order: 24 Requirements	39%	57%	4%
AI in Government Order: 16 Requirements*	13%	75%	13%
AI in Government Act: 6 Requirements	17%	17%	67%

*Requirement in section 5(c)(ii) has not been implemented, but the deadline for implementation has not yet passed, so it is not classified as implemented, not implemented, or not known. Therefore section 5(c)(ii)'s requirement has been excluded from this calculation so that the percentages will add up to 100 percent. Note: Section 5(c)(ii) was included in Table 3's calculations.

- The three pillars of America's AI Strategy
  - the AI Leadership Order (2019-02)
  - AI in Government Order (2020-12)
  - The AI in Government Act (2020-12)
- 상기 3개 법령이 규정하는 미국의 신뢰가능한 AI 전략사업의 진척도를 평가한 보고서
- 2022년 12월 현재 3개 법령의 45개 법률적 요구사항에 대해 40% 미만의 달성률로 미진함
- 그 원인으로 입법지원 등 실행자원 부족, 리더쉽 부재, 미션과 세부목표 사이의 미스매치(capacity gap)가 지목됨
- 체크리스트 수준에 불과한 법률적 요구사항들을 재구성하여 목표 달성을 위한 전략적 계획으로 만들고, 그에 맞는 적절한 입법지원이 필요함을 강조

https://hai.stanford.edu/white-paper-implementation-challenges-three-pillars-americas-ai-strategy

미국 - 백악관 주요테크기업 대상 AI 신뢰성/안전성 강조

구글·MS·오픈AI 등 AI제품 알고리즘 대중에 공개된다

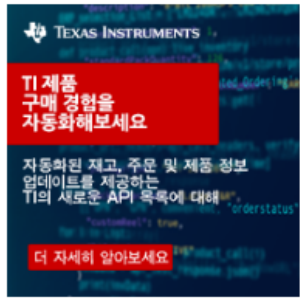
| 미 정부 "7개 테크기업 참여" 발표...AI 연구시설 7곳 추가 설립도

디지털경제 | 입력 :2023/05/06 15:37 수정: 2023/05/06 15:56

방은주 기자 | 기자페이지 구독 기자의 다른기사 보기

[이벤트] Microsoft Surface 특별 할인 프로모션 | Surface PC 15대를 구매하면 추가 5대 무료

챗GPT 등장으로 인공지능(AI)제품과 서비스의 신뢰성과 안정성 문제가 크게 부각한 가운데 미국 정부가 구글, 마이크로소프트(MS), 오픈AI, 앤쓰로픽(Anthropic) 등 테크기업 4개사 최고경영자(CEO)를 백악관으로 불러 환담하며 AI제품의 안정성과 신뢰성 확보를 강력히 촉구했다.



4일(미국시각) 외신에 따르면 환담에 앞서 백악관은 ▲7개 하이테크 기업의 AI 서비스에 사용한 데이터 공중(公衆) 평가 ▲AI 연구개발에 1억4000만 달러(약 1860억 원) 투입 및 관련 기관 7곳 추가 신설 ▲AI위험은 낮추고 기회는 높이는 방향의 새 정책 마련 등 3개 항의 새 AI정책도 발표했다.

백악관은 이번 조치로 구글, MS, 오픈AI, 엔비디아, 앤쓰로픽, 스태빌리티AI(Stability AI),허깅 페이스(Hugging Face) 등 7개 테크기업이 자사의 AI시스템에 사용한 학습된 데이터를 공중에 공개, 미국 정부가 제정한 지침에 부합하는 지를 평가받는다고 밝혔다. 평가는 이 분야 민간 전문기업인 스케일AI가 주관하며 오는 8월 10일~13일 라스베이가스에서 열리는 세계적 해커 행사 '데프콘 2023'의 부대행사인 'AI빌리지'에서 이뤄질 전

https://zdnet.co.kr/view/?no=20230506152252

- 미국 정부 차원에서 AI 관련 주요 테크기업들의 학습데이터에 대한 공중 평가 개최 예정(데프콘 2023, 라스베이거스, 2023-08-10~2023-08-13)
- AI 제품이 'AI 권리장전 청사진(Blueprint for an AI Bill of Rights)', 'AI 위험관리 프레임워크(AI Risk Management FRAMEwork)', 'AI 연구자원 로드맵' 등 지침에 부합하는지 평가
- 공중 평가 개최와 더불어, AI R&D에 1억 4000만 달러 투입 및 관련기관 7곳 추가신설과 AI 위험을 낮추기 위한 새 AI정책 마련 등 발표
- 미국에서 AI 신뢰성 확보를 위해 정부차원의 정책적 개입 및 투자가 적극적으로 이루어지고 있음을 확인할 수 있음

Artificial intelligence risks: Biden meets top tech CEOs, asks them to mitigate AI perils

Washington • Edited By: Sneha Swaminathan • Updated: May 05, 2023, 11:49 AM IST

Alphabet CEO Sundar Pichai, left, and OpenAI CEO Sam Altman arrive to the White House for a meeting with Vice President Kamala Harris on artificial intelligence, on May 4, 2023, in Washington. Photograph:(AP)

FOLLOW US

EU - A European approach to Excellence and Trust in AI

EUROPEAN
COMMISSION

Brussels, 19.2.2020
COM(2020) 65 final

WHITE PAPER

On Artificial Intelligence - A European approach to excellence and trust

- AI분야에서 글로벌 리더가 되기 위한 EU의 청사진
- AI분야 R&D 투자 효과 극대화를 위한 법률적 지원의 형태와 시기를 규정
- EC가 제정한 '신뢰할 수 있는 AI를 위한 윤리 가이드라인'
- **2019년 '신뢰가능한 AI를 위한 윤리 가이드라인'에 기반하여, 제도적 프레임워크를 구체화하기 위한 정책적 방향 제시**
- 또한, 제도적 프레임워크를 완성하기 위해 신뢰 가능한 AI에 도달하기 위한 요구사항들을 세분화 및 구조화
- 신뢰할 수 있는 AI의 세 가지 구성요소*: △적법성, △윤리성, △강건성
- 신뢰할 수 있는 AI의 7가지 요건*: △인간 행위자와 감독, △기술적 견고성과 안전성, △프라이버시와 데이터 거버넌스, △투명성, △다양성, 차별금지 및 공정성, △사회적 환경적 웰빙, △책임성

https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0065&from=FR

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

EU – 세계 최초 'AI규제법안'

유럽의회, 세계 최초 'AI규제법안' 가결..."생성AI, 출처 밝혀야"

머니투데이 | 변휘 기자 2023.06.15 20:42

https://news.mt.co.kr/mtview.php?no=2023061518205145062&type=1 기사주소 복사

EU 집행위 협상 남겨둬...타결되면 2026년 시행

벨기에 브뤼셀 유럽연합(EU) 집행위원회 본부 밖에서 EU 깃발이 휘날리고 있다. ⓒ 로이터=뉴스1 Copyright (C) 뉴스1. All

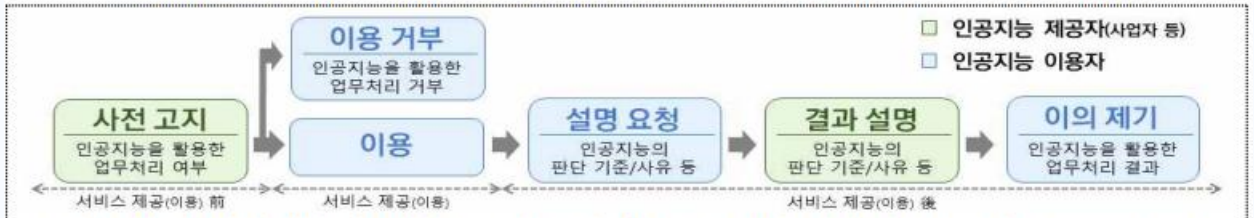
- EU는 인공지능 법에 따라 모든 AI는 최소 위험부터 허용할 수 없는 위험까지 단계로 분류된다. 예측적 치안(프로파일링, 위치 또는 과거 범죄 행위 기반) 도구, 얼굴 인식 데이터베이스 구축을 위해 인터넷이나 CCTV 영상에서 얼굴 이미지를 목적 없이 스크랩하는 행위(인권 및 사생활 침해)와 같은 기술은 금지
- 어린이, 기타 취약 계층 및 채용 관행에 초점을 맞춘 AI는 더 엄격한 규제.
- 인공지능으로 사람을 판단하고 개인의 사회적 행위 등을 평가 또는 점수화 하는 소셜 스코어링(Social Scoring), 민감한 특성(성별, 인종, 민족, 시민권 상태, 종교, 정치적 성향)을 사용하는 생체 인식 분류(Biometric Categorisation) 및 감정 인식(Emotion Recognition)을 위한 AI 시스템 금지
- 챗GPT, 바드, 미드저니 등 생성 AI(Generative AI) 및 대형언어모델(LLM)로 생성된 콘텐츠가 인간이 구현한 것이 아님을 명시하는 것과 학습에 사용된 데이터(자료)를 모두 공개
- 출처 : 인공지능신문(https://www.aitimes.kr)

https://artificialintelligenceact.eu/

한국 - 신뢰할 수 있는 인공지능 실현 전략(과기부)

비전	"누구나 신뢰할 수 있는 인공지능, 모두가 누릴 수 있는 인공지능 구현" - Trustworthy AI for Everyone -		
목표 (~'25)	책임있는 인공지능 활용 세계 5위	신뢰 있는 사회 세계 10위	안전한 사이버국가 세계 3위
추진 전략	신뢰 가능한 인공지능 구현 환경 조성	안전한 인공지능 활용을 위한 기반 마련	사회 전반 건전한 인공지능 의식 확산
	① 인공지능 제품·서비스 신뢰 확보 체계 마련 ② 민간 신뢰성 확보 지원 ③ 인공지능신뢰성원천기술개발	① 학습용 데이터 신뢰성 제고 ② 고위험 인공지능 신뢰 확보 ③ 인공지능 영향평가 추진 ④ 신뢰 강화 제도 개선	① 인공지능 윤리 교육 강화 ② 주제별 체크리스트 마련 ③ 인공지능윤리정책플랫폼운영

< 고위험 인공지능 서비스의 운용/이용단계에 따른 주체별 의무/권리(안) >

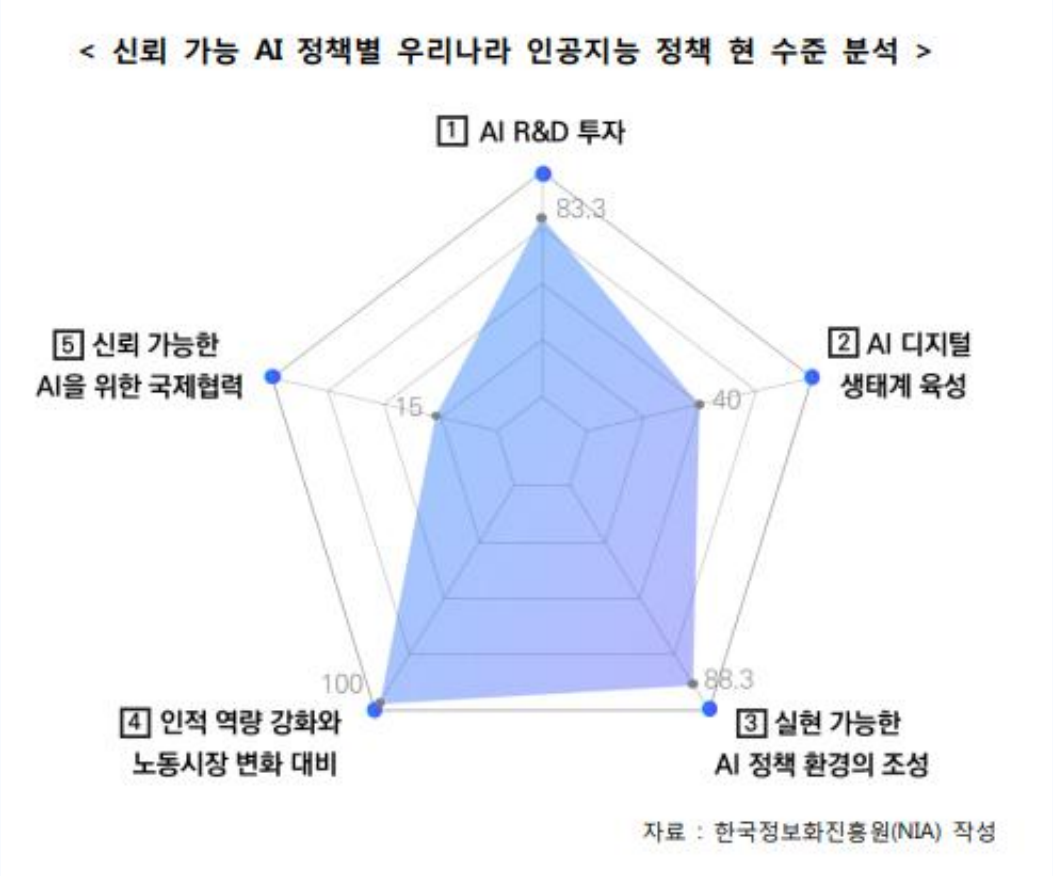


※ '자동화된 의사결정'에 대한 동 의무/권리는 △「유럽 개인정보보호법」(GDPR, '18.5~)은 기 제도화, △우리나라는 「개인정보보호법」('21.1.6일 입법예고)에서 제도화 추진 중

- AI 신뢰성 확보를 위해 과학기술정보통신부에서 정책적 방향성과 구체적 방안을 제시
- **"누구나 신뢰할 수 있는 인공지능, 모두가 누릴 수 있는 인공지능 구현"을 목표로 2025년까지 △신뢰가능한 인공지능 구현 환경 조성, △안전한 인공지능 활용을 위한 기반 마련, △사회 전반 건전한 인공지능 의식 확산을 전략**
- 환경조성을 위해 △AI 제품 및 서비스 개발의 각 단계에 신뢰성을 확보할 수 있는 체계적 프로세스를 확립, △민간분야 참여자들이 확립된 프로세스를 활용할 수 있는 플랫폼 구축 및 운영, △AI 신뢰성 확보를 위한 원천 기술을 개발
- 기반 마련을 위해 △신뢰성 높은 데이터를 위한 표준 마련, △고위험 AI 분류를 위한 기준 설정, △AI 영향평가 추진, △AI 제도 개선
- 윤리 보급을 위해 △교육 강화, △체크리스트 보급 등

https://www.korea.kr/common/download.do?fileId=195009613&tblKey=GMN

신뢰 가능 AI 구현을 위한 정책 방향(NIA)



- OECD AI 권고안(2019-05)에 따른 국가 정책 방향 제시
- OECD는 권고안을 통해 신뢰 가능한 AI 구현을 위한 5개 원칙을 제시함: △포용성장, 지속가능 발전, 복지증진, △인간중심 가치 지향, 공정성 지향, △투명성 확보, 설명가능성 확보, △견고성 확보, 보안 및 안전성 확보, △책임완수
- 이에 따라, 향후 국가 정책 방향으로 △AI R&D 투자 확대, △AI 디지털 생태계 육성, △실현 가능한 AI 정책 환경 조성, △인적역량 강화와 노동시장 변화 대비, △신뢰 가능한 AI 구현을 위한 국제협력을 제안 및 현황 진단

https://www.nia.or.kr/site/nia_kor/ex/bbs/View.do?cbIdx=25932&bcIdx=21258&parentSeq=21258

미국 - NIST AI Risk Management Framework

AI Risk Management Framework

The AI Risk Management Framework (AI RMF) is intended for voluntary use and to improve the ability to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems.

As a consensus resource, the AI RMF was developed in an open, transparent, multidisciplinary, and multistakeholder manner over an 18-month time period and in collaboration with more than 240 contributing organizations from private industry, academia, civil society, and government. Feedback received during the development of the AI RMF is publicly available on the NIST website.

Download the framework

1 Framing risk

Framing risk includes information on:

- Understanding and Addressing Risks, Impacts, and Harms
- Challenges for AI Risk Management

2 Audience

Identifying and managing AI risks and potential impacts requires a broad set of perspectives and actors across the AI lifecycle. The Audience section describes AI actors and the AI lifecycle.

3 AI Risks and Trustworthiness

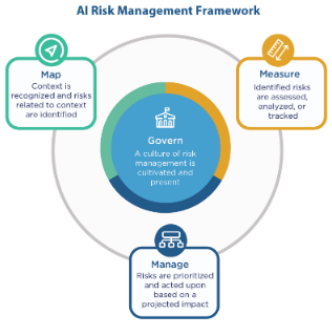
For AI systems to be trustworthy, they often need to be responsive to a multiplicity of criteria that are of value to interested parties. Approaches which enhance AI trustworthiness can reduce negative AI risks. The AI Risks and Trustworthiness section articulates the characteristics of trustworthy AI and offers guidance for addressing them.

4 Effectiveness of the AI RMF

The Effectiveness section describes expected benefits for users of the framework.

5 AI RMF Core

The AI RMF Core provides outcomes and actions that enable dialogue, understanding, and activities to manage AI risks and responsibility develop trustworthy AI systems. This is operationalized through four functions: Govern, Map, Measure, and Manage.



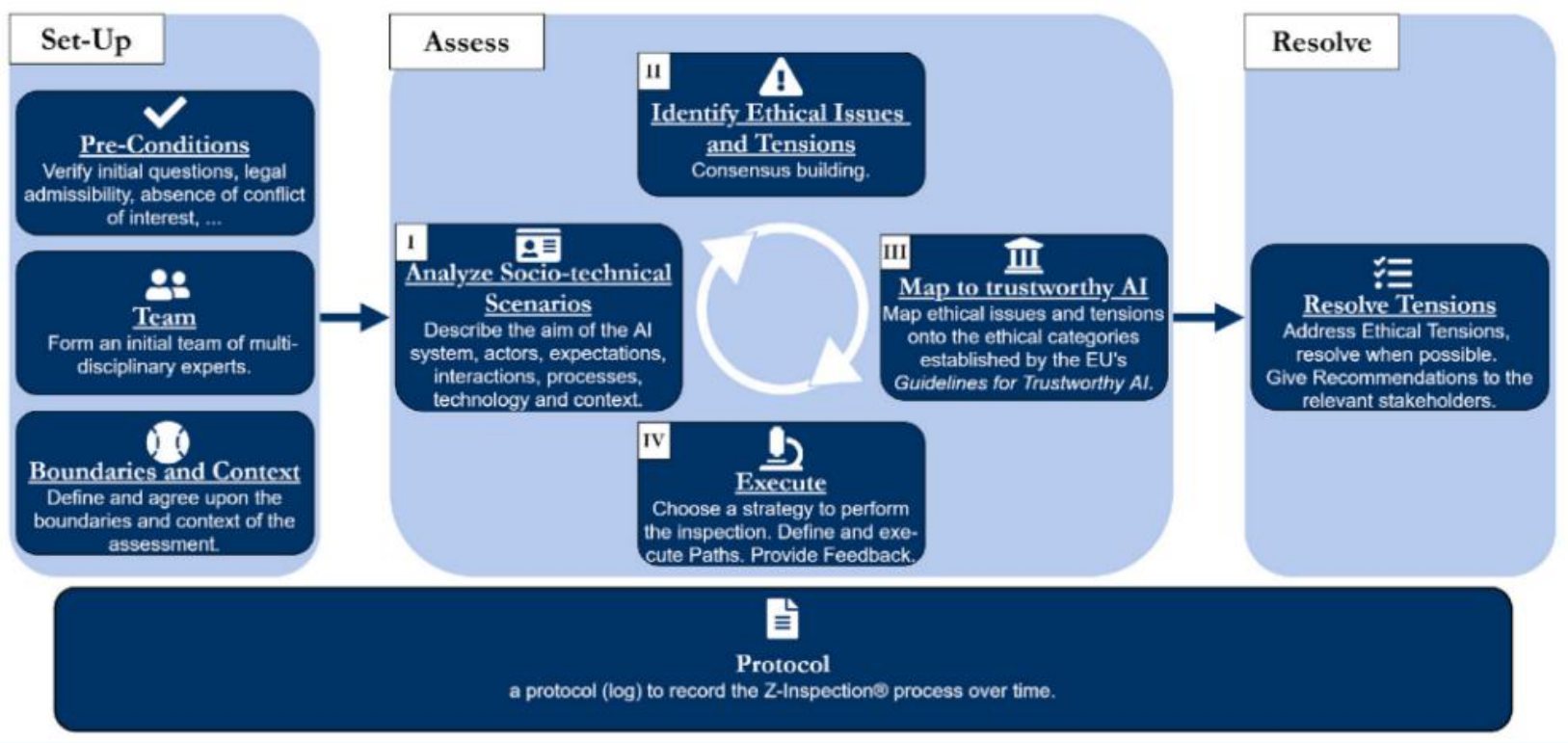
6 AI RMF Profiles

The use-case Profiles are implementations of the AI RMF functions, categories, and subcategories for a specific setting or application based on the requirements, risk tolerance, and resources of the Framework user.

- 미국 NIST(국립표준기술연구소)는 2023년 3월 Thrustworthy and Responsible AI Resource Center(AIRC, https://airc.nist.gov/Home)를 통해 **AI 개발 기관이 책임감 있게 서비스를 지원하는 체계 구축**
- **AI Risk Management Framework(AI RMF로 약칭)은 AI 제품, 서비스, 시스템의 설계, 구현, 사용과 평가에 '신뢰성 (Trustworthiness)'을 포함하고 평가**하도록 함
- 18개월 동안 240개 이상의 기관(기업, 학계, 시민단체, 정부)이 협업하여 투명하고 개방적인 방식으로 합의를 도출함

https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF

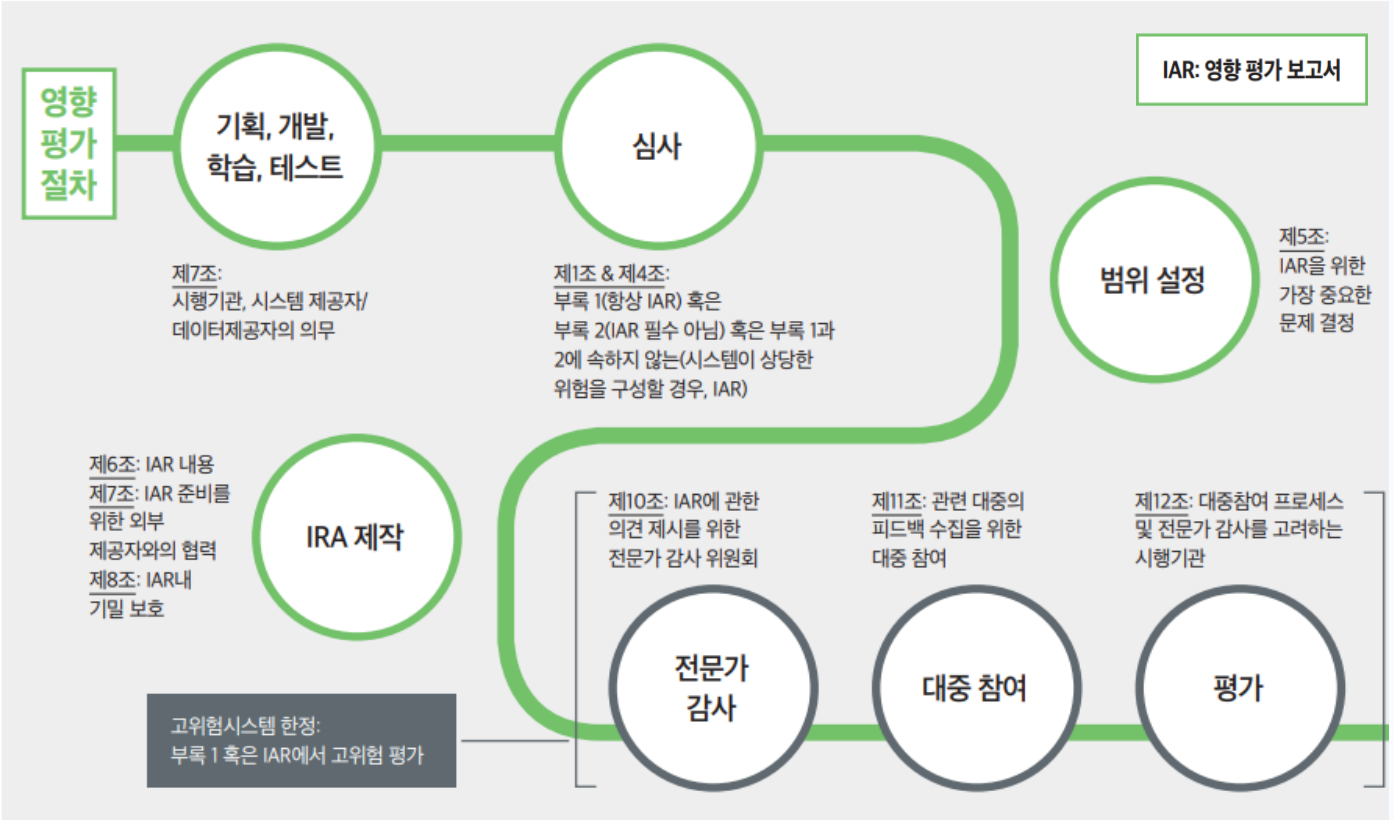
EU - Z-inspection(Frankfurt Big Data Lab)



- **EU '신뢰할 수 있는 AI를 위한 윤리 가이드라인'의 원칙에 맞게 Frankfurt Big Data Lab에서 개발한 윤리적 AI 평가 프로세스**
- 윤리 원칙에 맞는 실질적 practices의 연구 및 개발에 주력
- 심장질환진단 AI에 대한 실제 유스케이스를 가지고 있음: Cardis.io

https://futurium.ec.europa.eu/en/european-ai-alliance/best-practices/z-inspection-towards-process-assess-ethical-ai-0

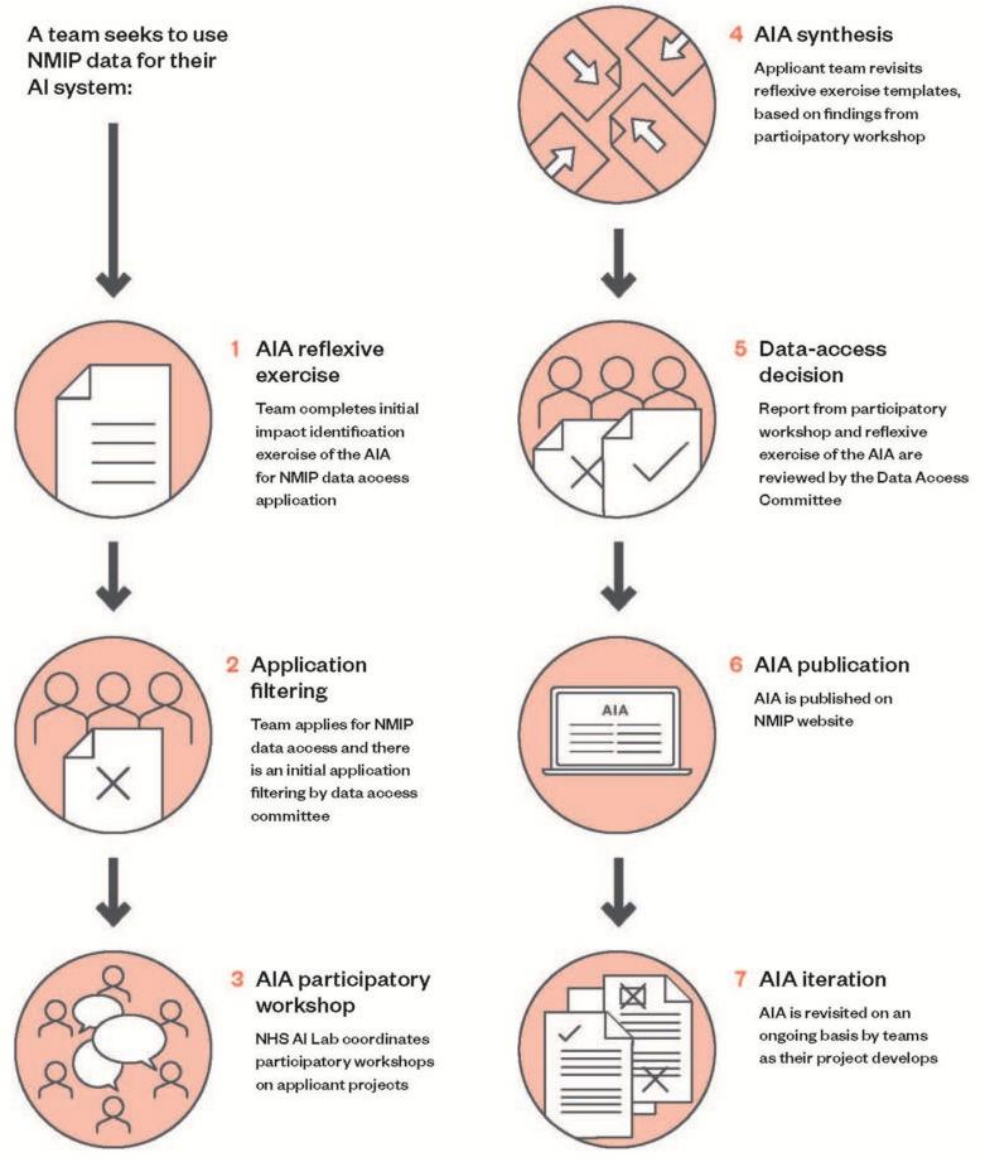
한국 – AI 서비스와 AI 관련법의 영향평가[한국법제연구원]



- **영향평가: AI 시스템의 개인, 사회 및 환경 등에 미치는 잠재적 영향을 평가하는 절차**
- **입법영향평가: 제안된 인공지능 관련 법률 또는 규제가 개인, 사회 및 경제, 법체계 등에 미치는 잠재적 영향을 평가**하는 절차
- 잠재적 위험요소를 폭넓게 고려한 체계적인 영향평가 절차 수립의 필요성 강조
- 영향평가 및 입법영향평가를 위한 AI의 활용 안전성 제시

https://www.kipa.re.kr/cmm/fms/FileDown.do?atchFileId=FILE_000000000016581&fileSn=4

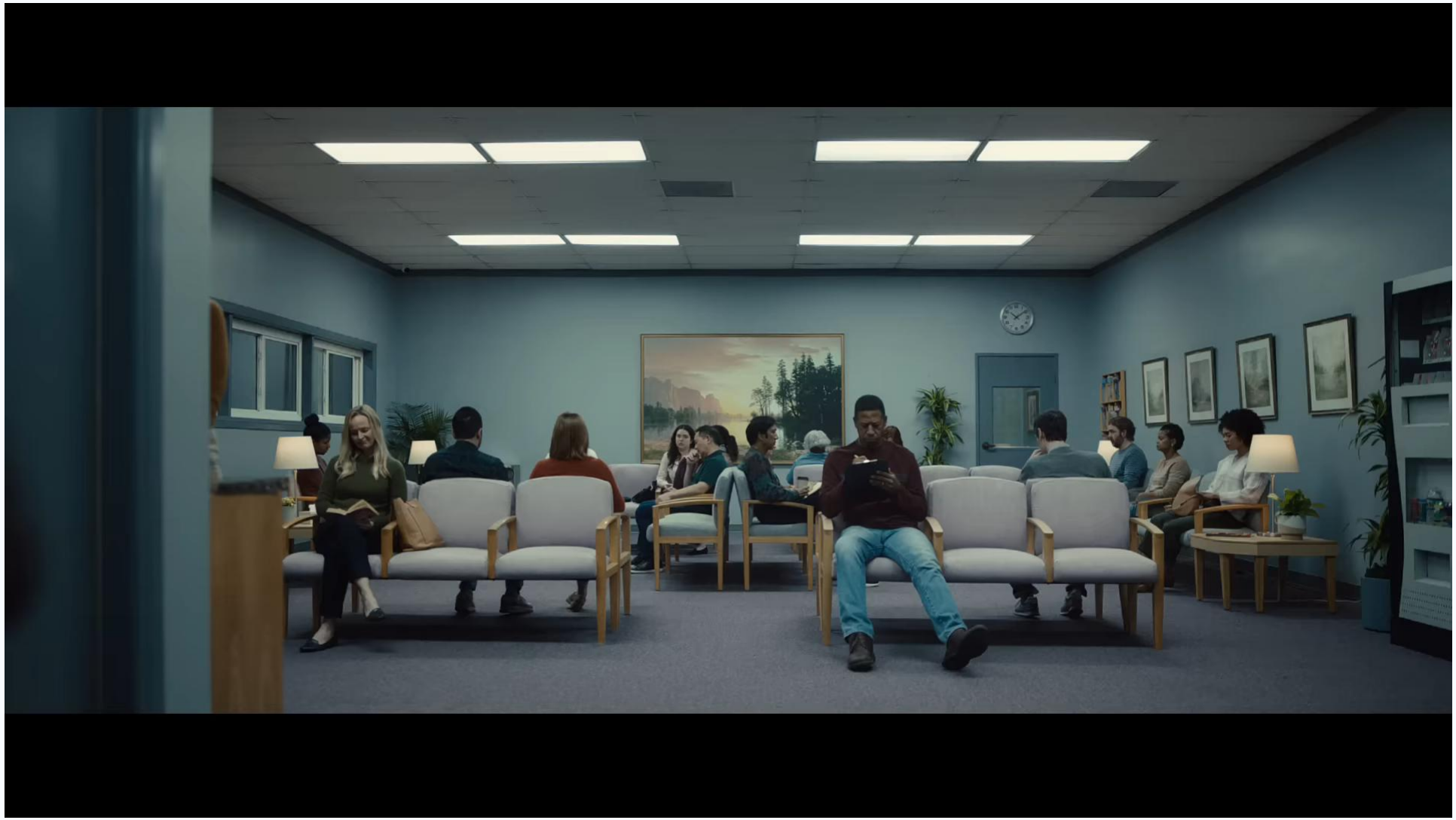
영국 - 알고리즘 영향평가 모델 사례[헬스케어 분야]



- **Ada Lovelace Institute이 개발한 AI 알고리즘에 대한 영향평가** (Algorithmic Impact Assesment, 이하 AIA) 모델 을 영국 국민건강서비스(NHS)에서 시범적용 결정(이하 NHS NMIP AIA)
- NHS의 **NMIP 데이터에 액세스하여 개발을 진행하고자 하는 모든 회사는 AIA 프로세스를 수행**해야 하며, NHS는 이를 검토하여 최종적으로 누가 NMIP 데이터에 액세스할 수 있는지 결정
- **NHS NMIP AIA 프로세스는 AI 기술의 신뢰성을 확보하기 위해 책임성(accountability)와 투명성(transparency) 등을 목적 가치로 제시**하고 있음

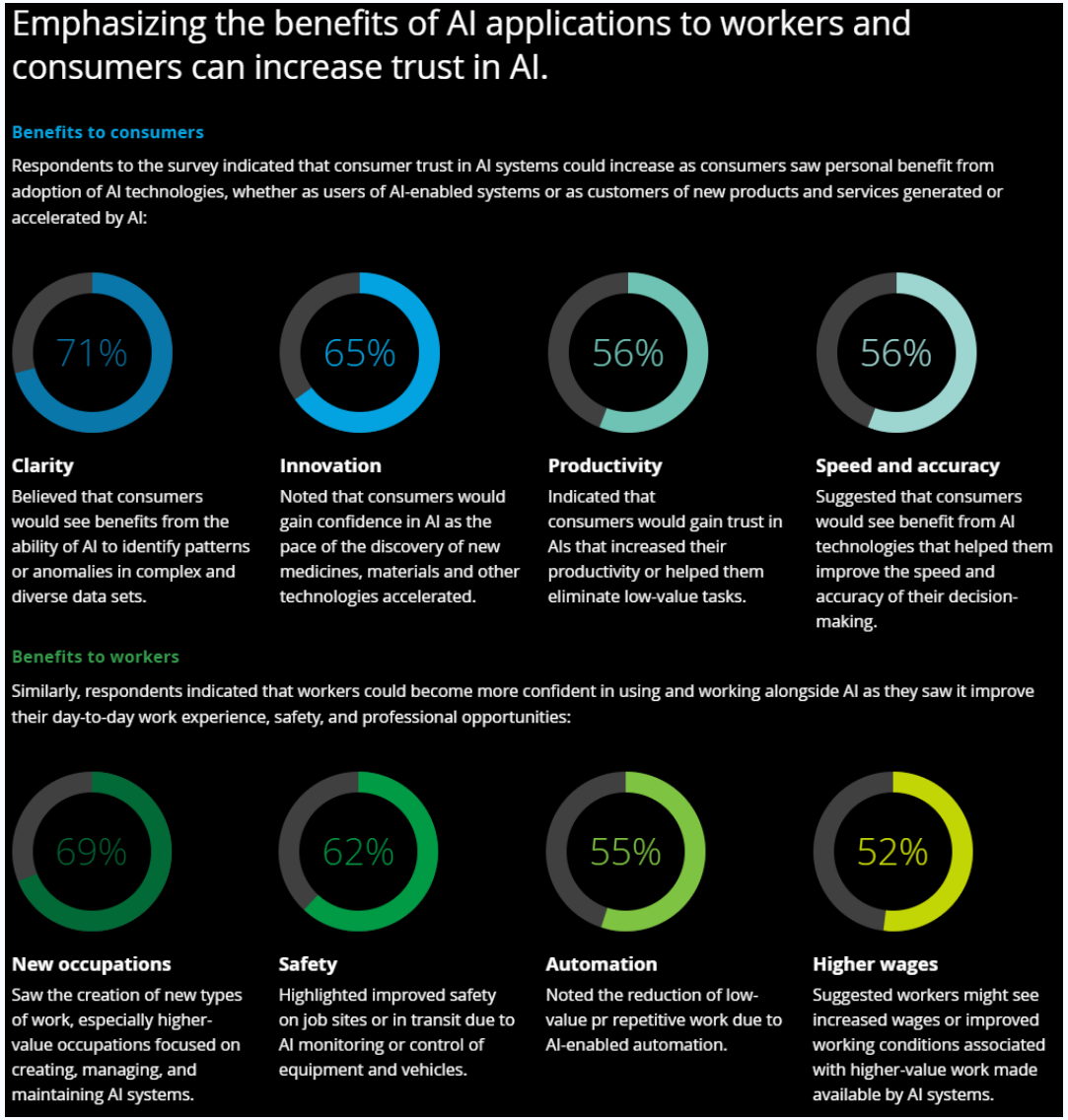
https://www.koraia.org/chtml/board.php?template=base&com_board_basic=file_download&com_board_id=12&com_board_idx=363&com_board_file_seq=0

애플의 개인정보보호 노력[헬스케어 분야]



https://www.koraia.org/chtml/board.php?template=base&com_board_basic=file_download&com_board_id=12&com_board_idx=363&com_board_file_seq=0

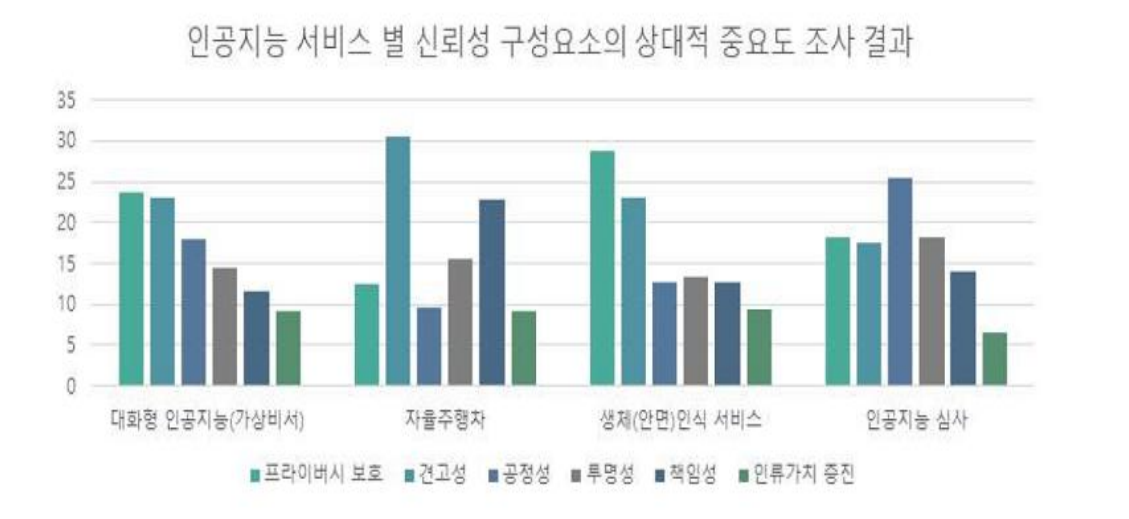
AI Trustworthiness의 경제적 효과(Deloitte)



- Deloitte사에서 자체적으로 시행한 서베이를 바탕으로 AI 신뢰성을 확보할 수 있는 정책적 방향 제시
- 서베이에 따르면, **AI의 △ 편향, △ 책임성(accountability)의 부족, △ 설명성(explainability)의 부족** 등이 AI 신뢰성을 떨어뜨리는 주요 요인으로 지목됨
- 또한, AI 신뢰성을 개선할 수 있는 새로운 공공정책 및 투자에 대한 요구가 높은 것으로 나타났으며, 이러한 정부차원의 개입 및 투자가 AI 기술의 경제적 효과를 증대시킬 것이란 기대가 확인되었음
- 세부적으로, 기초연구분야 뿐만 아니라 AI 신뢰성 확보를 위한 표준 제정 등 표준화 분야에 대한 정부의 적극적 개입이 요구되고 있음

https://www2.deloitte.com/us/en/pages/consulting/articles/investing-in-ai-trust.html

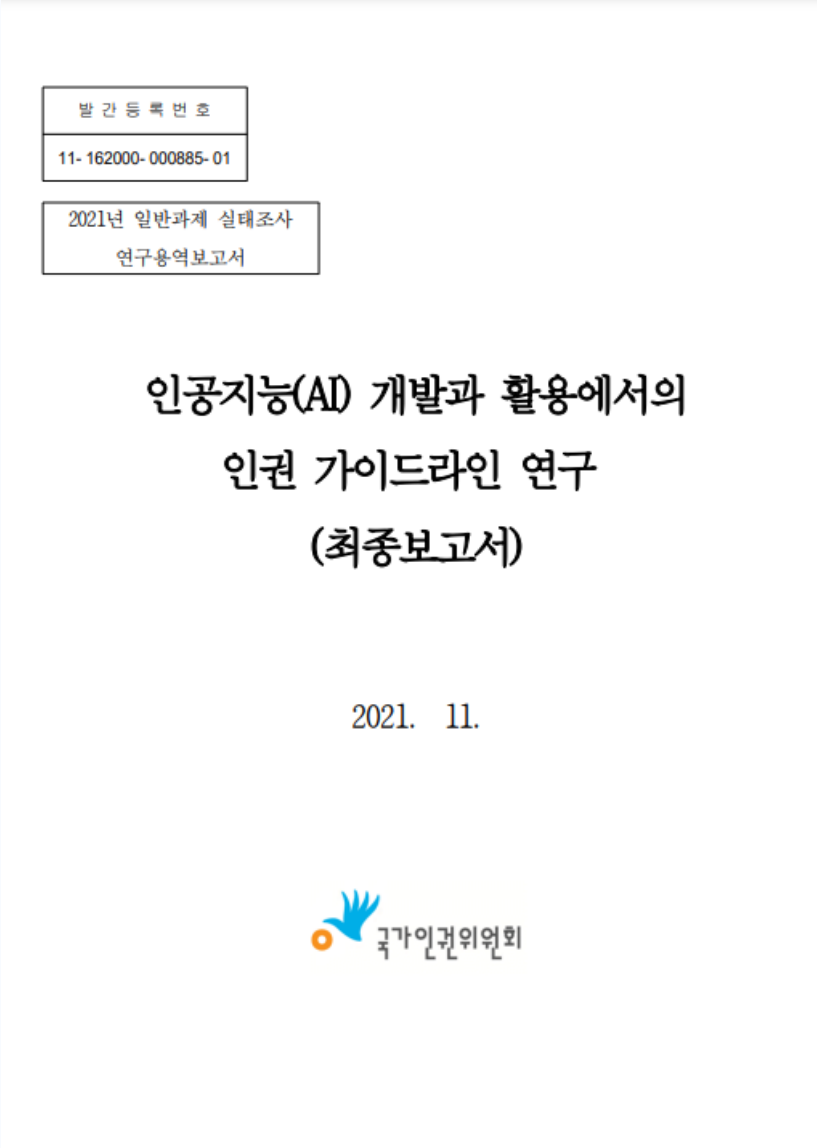
인공지능 신뢰체계 정립방안 연구(소프트웨어정책연구소)



- AI 신뢰성 구성요소 도출: △프라이버시 보호, △견고성, △공정성, △투명성, △책임성, △인류가치 증진 (총 6개)
- **국내 전문가를 대상으로 한 AI 신뢰성 요구수준 분석: 견고성 > 프라이버시 보호> 투명성 > 책임성 > 공정성 > 인류가치 증진 순으로 중요도가 조사됨**
- AI 신뢰성 기업준비도 조사: 2021년 초 '이루다' 사건을 계기로 AI 신뢰성에 대한 인식이 업계 전반에 크게 확산되었으며, 신뢰성 확보 분야가 새로운 사업 기회로 인식되고 있음
- 인공지능 신뢰성 기술도구 분석: AI 개발 과정에서 기술적으로 확보할 수 있는 신뢰성은 매우 제한적이며, 비기술적 혹은 제도적 접근이 요구됨

https://spri.kr/posts/view/23438?code=research

AI 개발과 활용에서의 인권 가이드라인 연구(국가인권위원회)



발 간 등 록 번 호
11- 162000- 000885- 01

2021년 일반과제 실태조사
연구용역보고서

인공지능(AI) 개발과 활용에서의
인권 가이드라인 연구
(최종보고서)

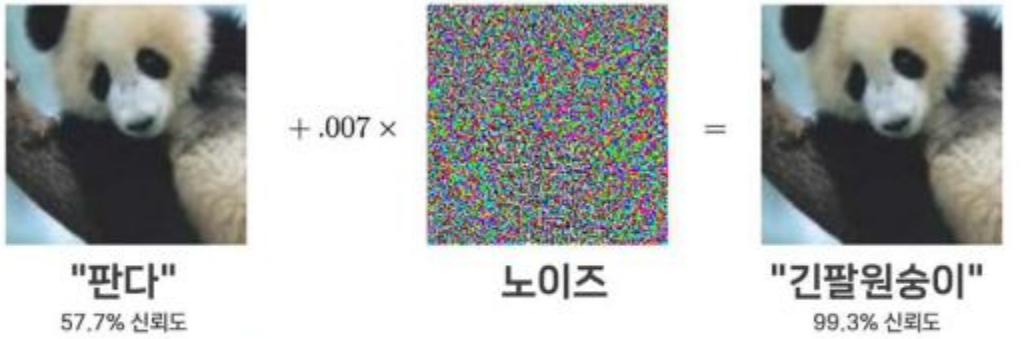
2021. 11.

국가인권위원회

- 국가인권위원회에서, 공공부문에서 인공지능 도입 현황과 국제규범 등을 조사하고, 인권 가이드라인을 제시
- AI 기술 활용시 침해받을 수 있는 인간의 존엄성, 알 권리 등 인권 사항들을 다시 확인하고, 투명성, 설명의무 등 인권 보장을 위해 AI 기술이 갖추어야할 성격들을 명문화하고 있음
- AI가 일으킬 수 있는 인권 문제와 관련하여 국가정책이 나아가야할 방향을 인권위가 종합적으로 제시했다는 데 의의가 있음 (인공지능의 법적 이슈와 전망, 톰슨로이터코리아, 2022)

https://library.humanrights.go.kr/search/detail/CATTOT000000053138

생성형 AI - ‘가짜’ 데이터가 만드는 ‘진짜’ 인공지능 시대(NIA)



< 적대적 공격 사례 : FGSM(fast qradient sign method) >

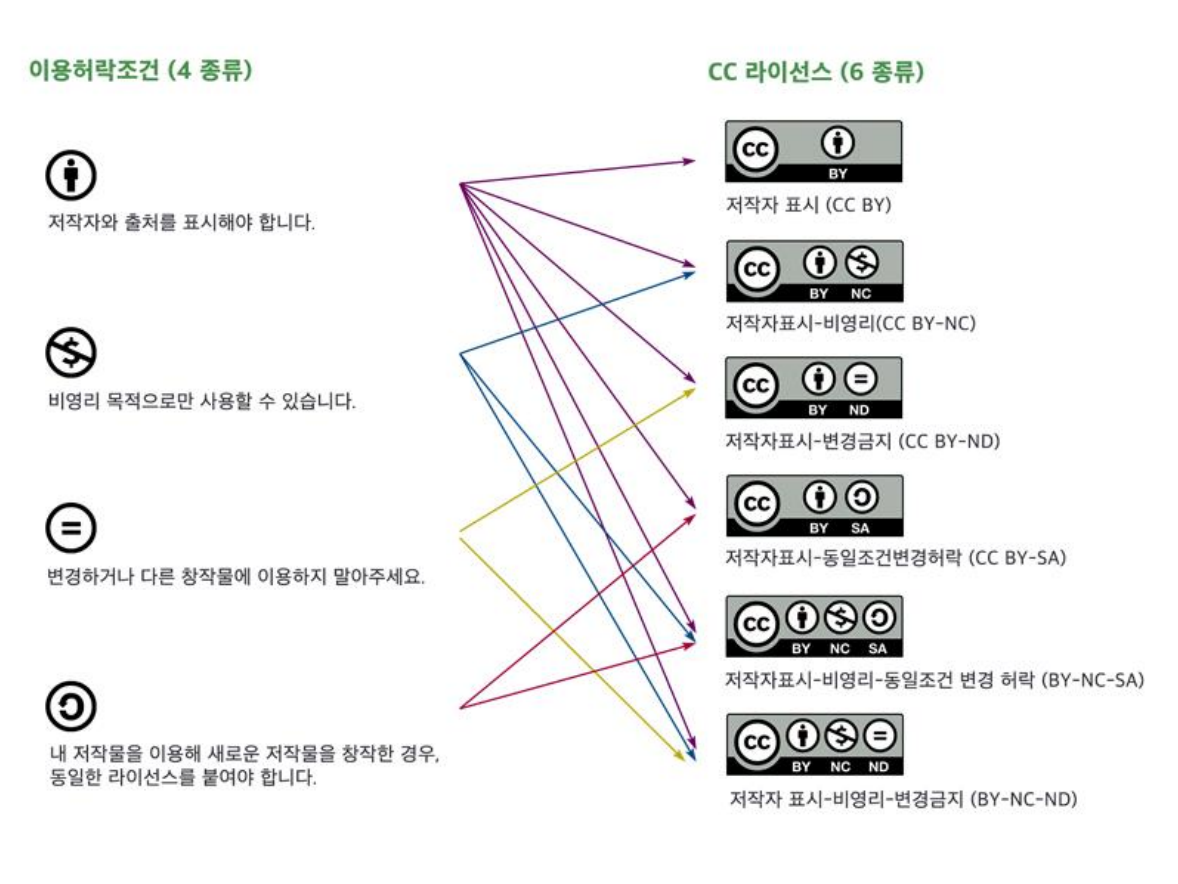
※ 출처 : Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy(2015), 'Explaining and harnessing adversarial examples'

- Labelling cost, 보안, 저작권, 프라이버시, 데이터 편향성 등으로 인해 양질 데이터 확보가 어려우며, 그 대안으로 합성데이터(synthetic data, 제목상 ‘가짜’ 데이터를 지칭)의 중요성이 강조됨
- 2030년까지 국내외 합성 데이터 시장의 예상 성장률은 34.8%로 매우 빠르게 성장하고 있음
- 다만, 합성데이터의 데이터 편향성 등으로 인해 잘못된 예측 결과를 제공하여 사회적 이슈를 야기할 우려
- 이에 따라, 합성데이터의 올바른 사용을 위한 기준 마련이 요구됨

https://www.nia.or.kr/site/nia_kor/ex/bbs/View.do?cbIdx=25932&bcIdx=25112&parentSeq=25112

생성형 AI - ChatGPT 현황과 저작권 이슈[한국저작권위원회]

CCL 이용허락 조건 및 종류
(출처: CCL Korea)



https://www.copyright.or.kr/information-materials/trend/the-copyright/viewPress.do?brdctsno=51647

- ChatGPT의 저작권 관련 법적 윤리적 쟁점은 크게 3가지로 특징 지어짐

1. **학습자료 이용의 저작권 문제:** 인터넷에 공표된 저작물 중 상당 부분이 저작권법상 보호받는 저작물이기 때문에 해당 저작물의 수집, 저장, 학습과정에서 발생하는 전송 등의 과정을 고려해볼 때 학습자료에 대한 저작권 문제가 발생할 수 있음
2. **CCL과 오픈소스 문제:** ChatGPT가 학습한 소스코드나 학습한 소스코드를 통해 창작한 소스코드가 오픈소스 라이선스 저작물을 사용하였고, 오픈소스 라이선스의 사용조건을 지키지 않았다면 저작권 침해 문제가 발생할 수 있음
3. **ChatGPT 생성물의 저작물성:** 현행 저작권법상 저작자는 인간을 전제로 하기 때문에 ChatGPT의 생성물은 저작물로 인정받기 어려움

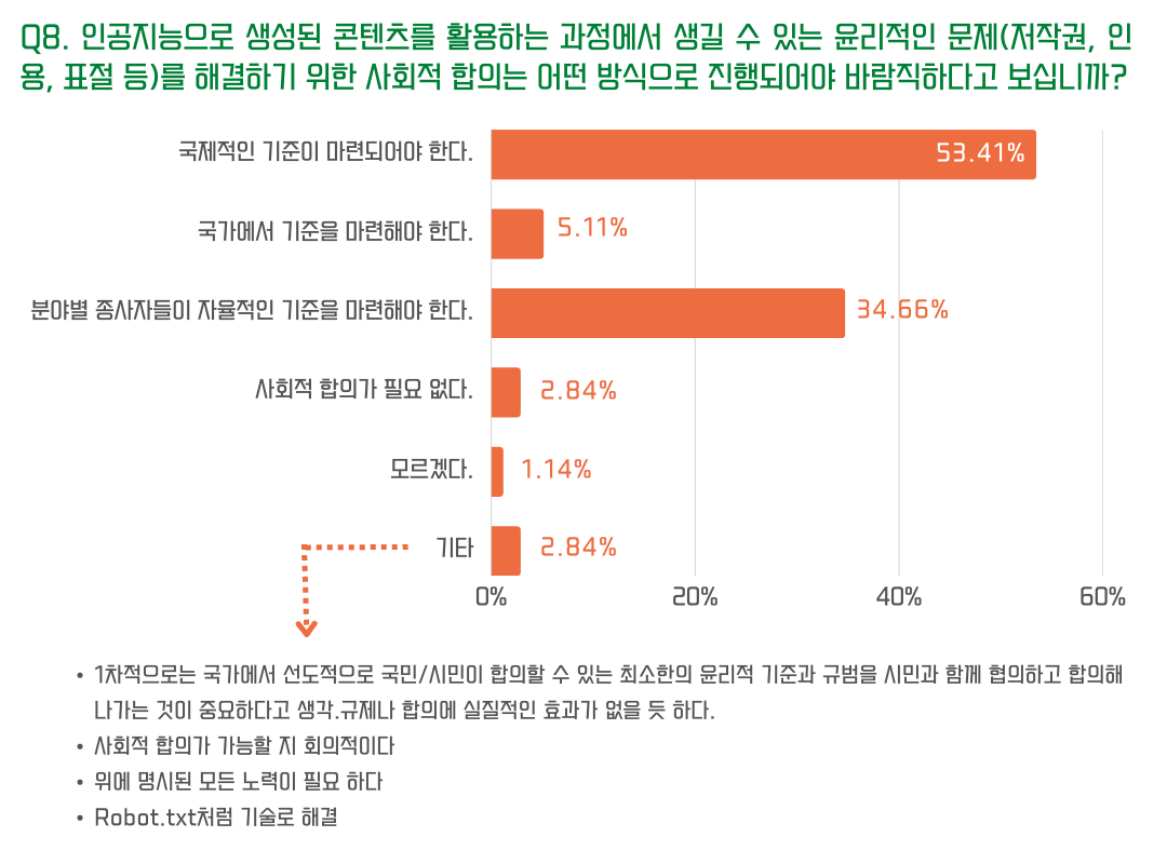
생성형 AI - ChatGPT 등장과 법제도 이슈[NIA]

https://investingnews.com/invest-in-openai-chatgpt/

- ChatGPT와 관련된 법적, 윤리적 쟁점은 크게 4가지로 특징지어짐: △**저작권**, △**개인정보**, △**오용**, △**책임**
- **저작권**: 학습 데이터 및 생성결과에 대한 저작권 문제
- **개인정보**: 학습 데이터 및 생성결과에 포함될 수 있는 개인정보 문제
- **오용**: 결과 생성을 위해 민감 정보를 입력하거나, 생성결과를 범죄행위에 활용하는 문제
- **책임**: 서비스 제공자에겐 일반적, 포괄적으로 면책되는 반면, 이용자에게 이용책임이 과도하게 부과되는 문제
- AI기술로 인한 새로운 법적 문제와 사회적 영향력을 주의깊게 모니터링하고, 국제적 협력을 통해 법제 개선 방향의 지향점을 통일시켜 나가는 노력이 필요

https://www.nia.or.kr/site/nia_kor/ex/bbs/View.do?cbIdx=82618&bcIdx=25371&parentSeq=25371

ChatGPT 사례로 본 AI 생산 저작물 활용에 대한 윤리인식[숲사이]



- 일반시민, 학생, 과학기술자 등으로 구성된 176명을 대상으로 ChatGPT가 생성해낸 자료의 저작물 활용에 관한 윤리의식 서베이 진행
- 응답자의 34.04%가 ChatGPT가 생성해낸 자료를 본인 저작물에 '인용할 수 없다'고 응답했으며, 52.27%가 '인용에 대한 합의가 이루어져야만 가능하다'고 응답
- **응답자의 70% 이상이 원작자 동의여부와 상관없이 학습데이터를 인공지능 기술에 활용하고 있는 ChatGPT의 구조에 대해 '문제가 있다'**('문제가 많다' 혹은 '문제가 어느정도 있다')고 응답
- 응답자의 53.41%는 AI의 저작권 관련 윤리문제에 대해 '국제 기준이 마련되어야 한다'고 응답한 반면, 34.66%는 '분야별 종사자들이 자율적 기준을 마련해야한다'고 응답
- 다시 말해, ChatGPT 다수 사용자들이 저작권 이슈를 중요하게 인식하고 있으며, 이에 대한 제도적 기준을 요구하고 있음

https://soopsci.com/Survey/?idx=13911052&bmode=view

AI의 법적 이슈와 전망[톰슨로이터코리아]

- **ChatGPT와 같은 초거대 AI 모델의 출력물에 대한 저작권 문제에 대해, 기존의 출처 기재 방식을 기대하기 어려움**
- 인공지능 개발자의 방어논리인 '공정이용'에 대해, 학습데이터의 수집 및 이용이 저작권법상 공정이용에 해당하는지 명확한 법적 판단이 내려진 바가 없음
- 현재 AI 관련 저작권 문제는 오픈소스 소프트웨어 업계에 한정되어 법적 분쟁으로 이어지고 있으나, 더 광범위한 분야에서 유사한 문제가 발생할 것으로 예상되며, 사회적 합의와 규범이 갖추어지기 전까지 지속적으로 논란의 중심이 될 것으로 예상됨

https://servicedesign.tistory.com/388?fbclid=IwAR1C_h9FlZ2uYB65zXX1q7-Iw6LK-8q0kh2nZEW_u45YZVL7NxzM1ITKJtA&mibextid=kdkkhi#gsc.tab=0

ChatGPT의 표절[이동원 교수팀]

WWW '23, May 1–5, 2023, Austin, TX, USA Lee et al.

Type	Machine-Written Text	Training Text
Verbatim	*** is the second amendment columnist for Breitbart news and host of bullets with ***, a Breitbart news podcast. [...] (*Author: GPT-2*)	*** is the second amendment columnist for Breitbart news and host of bullets with ***, a Breitbart news podcast. [...]
Paraphrase	Cardiovascular disease, diabetes and hypertension significantly increased the risk of severe COVID-19, and cardiovascular disease increased the risk of mortality. (*Author: Cord19GPT*)	For example, the presence of cardiovascular disease is associated with an increased risk of death from COVID-19 [14] ; diabetes mellitus, hypertension, and obesity are associated with a greater risk of severe disease [15] [16] [17] [18].
Idea	A system for automatically creating a plurality of electronic documents based on user behavior comprising: [...] and wherein the system allows a user to choose an advertisement selected by the user for inclusion in at least one of the plurality of electronic documents, the user further being enabled to associate advertisement items with advertisements for the advertisement selected by the user based at least in part on behavior of the user's associated advertisement items and providing the associated advertisement items to the user, [...] . (*Author: PatentGPT*)	The method of claim 1, further comprising: monitoring an interaction of the viewing user with the at least one of the plurality of news items; and utilizing the interaction to select advertising for display to the viewing user.

Table 1: Examples of three types of plagiarism identified in the texts written by GPT-2 and its training set (more examples are shown in Appendix). Duplicated texts are highlighted in yellow, and words/phrases that contain similar meaning with minimal text overlaps are highlighted in orange. [...] indicates the texts omitted for brevity. Personally identifiable information (PII) was masked as *.**

오픈AI 언어모델 'GPT-2'의 교육자료 표절 예

[펜실베이니아주립대 이동원 교수팀 논문 캡처. 재판매 및 DB 금지]

- 오픈AI의 'GPT-2'가 생성한 텍스트를 훈련에 사용된 자료와 비교 분석한 결과 다양한 표절이 발견
  1. 복사해 붙여넣기(vervatim)
  2. 출처 인용 없이 문장 바꾸기(paraphrase)
  3. 아이디어 도용(idea plagiarism) 같은 다양한 표절
- 연구팀은 이 연구에서 복사해 붙이기와 출처 인용 없이 문장 바꾸기, 출처 명시 없는 아이디어 도용하기 같은 표절을 자동 감지하는 프로그램을 개발, GPT-2가 생성한 21만 건의 텍스트를 언어모델 훈련에 사용된 800만 건의 문서와 비교

https://www.yna.co.kr/view/AKR20230217111500009

Agenda

- Motivation
- Case Studies
- Social Impact and Regulations
- **Trends in XAI Research**
- Appendix: What is inside of Deep Neural Networks

Language Models (Mostly) Know What They Know

국외

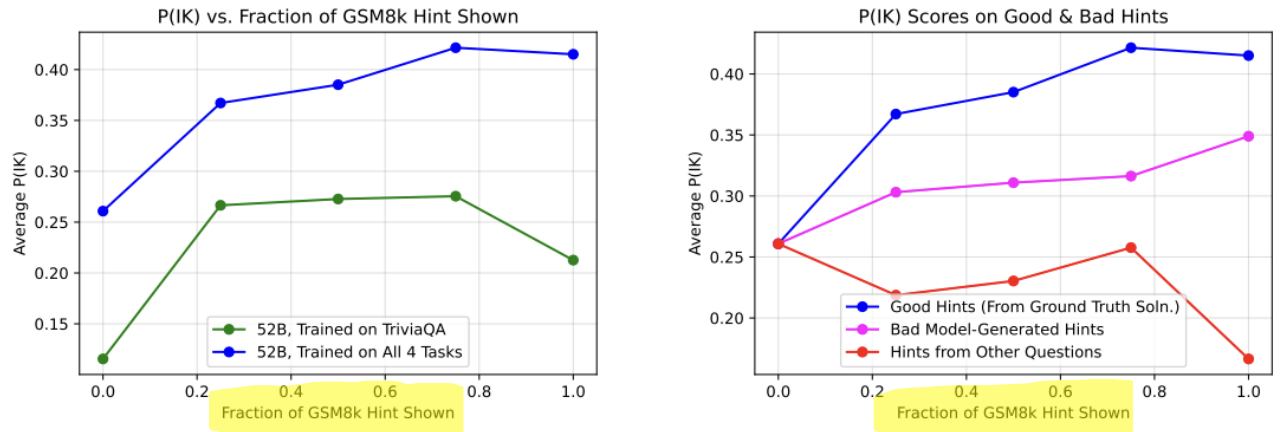


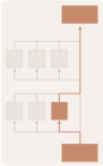
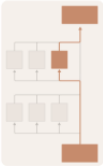
Figure 19 Effect of hints on P(IK) applied to GSM8k – all of these results represent generalization, as the models were not trained on GSM8k. Left: We see that showing more of the GSM8k hint results in higher P(IK). The effect is more consistent for the model trained on all 4 tasks (TriviaQA, LAMBADA, Arithmetic, and Python Function Synthesis), rather than the one trained on just TriviaQA. Right: We evaluate the model trained on all 4 other tasks on various hints. We see lower P(IK) scores for bad hints (though the models are partially fooled), and actual decreases in the P(IK) score when the hints are irrelevant because they come from other questions.

- Saurav Kadavath et al., Anthropic (2022.11)
- 본 논문은 언어 모델(language models)이 자신의 주장에 대한 타당성을 평가하고 올바른 답을 정확하게 예측할 수 있는지 연구
- 오픈엔드 샘플링(open-ended sampling tasks) 작업에서 모델이 먼저 답변을 제안한 후, 스스로 자신의 답변에 대해 정답일 확률 "P(True)"를 평가하는데, 평가에 있어 여러 샘플을 고려하면 평가 확률이 높아지는 것을 확인
- 모델이 특정한 답변을 참조하지 않고 "나는 알고 있다(I know)"는 문제에 대한 답변을 예측하는 확률 "P(IK)"도 문맥과 관련된 소스가 제공될 경우 확률이 증가
- 두 관찰 결과를 통해 정직한 모델을 훈련시키는데 기초 마련, 모델의 정직성의 일반화 조사에 기여

https://arxiv.org/pdf/2207.05221

A Mathematical Framework for Transformer Circuits

국외

Type	Example	Marginal Loss Reduction	Notes
Layer 1 Attention Heads		**- 0.05 nats** -0.004 nats/head relative to direct path + layer 2 **- 1.3 nats** -0.1 nats/head relative to direct path only	Relatively small effect, but keep in mind these heads also contribute to layer 2 QK circuits.
Layer 2 Attention Heads		**- 4.0 nats** -0.3 nats/head relative to direct path + layer 1 **- 5.2 nats** -0.4 nats/head relative to direct path only	**We'll focus on these.** Much larger effect. These heads are a lot more sophisticated than the layer 1 heads, since they can use layer 1 heads in their QK circuits.

- Nelson Elhage et al., Anthropic (2021.12)
- 트랜스포머 기반의 모델 (GPT-3, LaMDA, Codex, Meena, Gopher 등) 크기가 거대해짐에 따라 예측하지 못했던 결과가 계속해서 발견되고 있음 (unexpected and sometimes harmful behaviors)
- 이러한 문제의 원인을 밝히기 위해 어텐션 블록만을 가진 간단한 트랜스포머 모델인 **'transformers with two layers or less which have only attention blocks'**을 리버스 엔지니어링함
- Findings: **"induction heads"라고 명명된 특정 어텐션 헤드가 이런 간단한 트랜스포머 모델 내에서 in-context learning을 설명할 수 있으며, 이러한 어텐션 헤드는 적어도 2개 이상 어텐션 레이어가 있는 모델에서 발견됨**

https://transformer-circuits.pub/2021/framework/index.html#technical-details

iNEEJI EMPOWERING EXPERTISE
KAIST

In-context Learning and Induction Heads

국외

- **Argument 1** *(Macroscopic co-occurence)*: Transformer language models undergo a "phase change" early in training, during which induction heads form and simultaneously in-context learning improves dramatically.
- **Argument 2** *(Macroscopic co-perturbation):* When we change the transformer architecture in a way that shifts whether induction heads can form (and when), the dramatic improvement in in-context learning shifts in a precisely matching way.
- **Argument 3** *(Direct ablation)*: When we directly "knock out" induction heads at test-time in small models, the amount of in-context learning greatly decreases.
- **Argument 4** (*Specific examples of induction head generality*): Although we define induction heads very narrowly in terms of copying literal sequences, we empirically observe that these same heads also appear to implement more sophisticated types of in-context learning, including highly abstract behaviors, making it plausible they explain a large fraction of in-context learning.
- **Argument 5** (*Mechanistic plausibility of induction head generality):* For small models, we can explain mechanistically how induction heads work, and can show they contribute to in-context learning. Furthermore, the actual mechanism of operation suggests natural ways in which it could be re-purposed to perform more general in-context learning.
- **Argument 6** (*Continuity from small to large models*): In the previous 5 arguments, the case for induction heads explaining in-context learning is stronger for small models than for large ones. However, many behaviors and data related to both induction heads and in-context learning are smoothly continuous from small to large models, suggesting the simplest explanation is that mechanisms are the same.

https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html

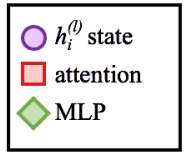
- Catherine Olsson, Nelson Elhage, Neel Nanda et al., Anthropic (2022.8)
- Transformer 생성모델에서 기계적 해석 가능성(Mechanistic interpretability)은 모델이 수행하는 상세한 계산을 역공학함으로 안전 문제에 대응할 수 있음
- 이 논문은 induction heads가 전체적으로 transformer 모델의 대다수 in-context 학습의 메커니즘을 제공하는 주요한 역할을 할 수 있다는 주장에 대해 논의
- 총 6가지 argument를 훈련 과정에서 34개의 transformer 모델의 분석 결과 및 50,000개 이상의 attention head 제거 실험 데이터를 바탕으로 증거를 제시

EMPOWERING EXPERTISE
iNEEJI
KAIST

ChatGPT의 개발 관련 – ChatGPT는 의식이 있을까?

Locating and Editing Factual Associations in GPT

국외



- Findings
  - 1. Factual associations can be localized along three dimensions, to (1) MLP module parameters (2) at a range of middle layers and (3) specifically during processing of the last token of the subject.

https://rome.baulab.info/

EMPOWERING EXPERTISE
iNEEJI
KAIST

Knowledge Neurons in Pretrained Transformers

국외

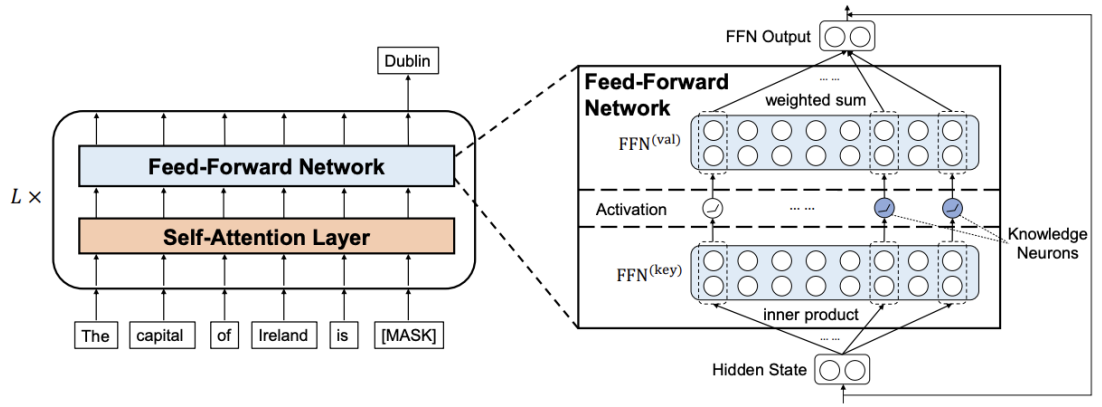


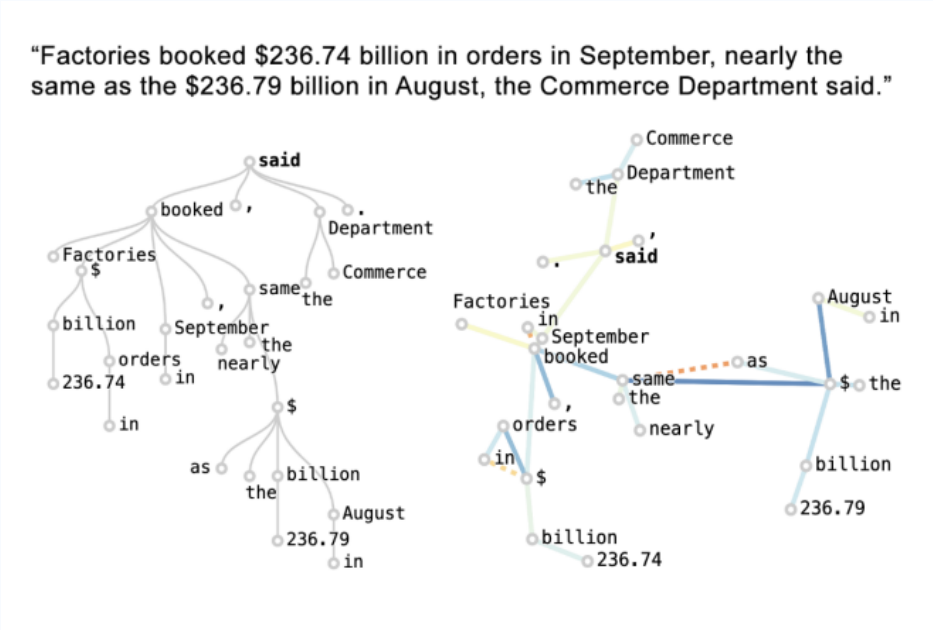
Figure 2: Illustration of how an FFN module in a Transformer block works as a key-value memory. The first linear layer $\mathrm{FFN}^{(\mathrm{key})}$ computes intermediate neurons through inner product. Taking the activation of these neurons as weights, the second linear layer $\mathrm{FFN}^{(\mathrm{val})}$ integrates value vectors through weighted sum. We hypothesize that knowledge neurons in the FFN module are responsible for expressing factual knowledge.

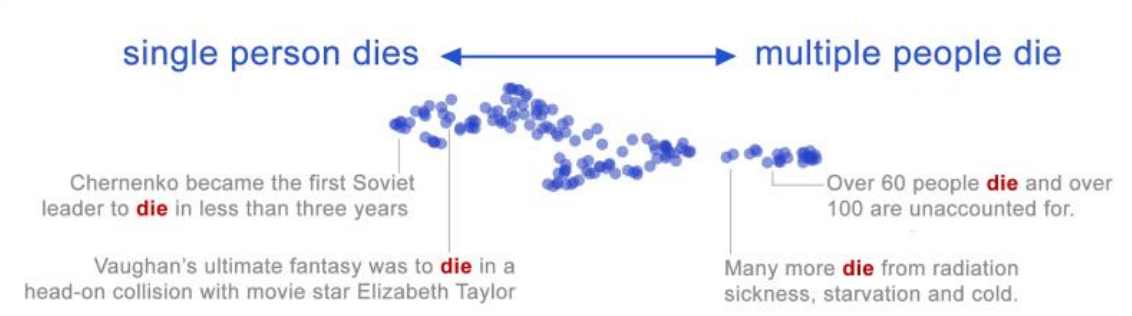
- Damai Dai et al. (2022.03)
- 본 논문에서는 사전 학습된 Transformers model 내에서 사실적인 지식이 어떻게 저장되는지에 대한 초기 연구 결과를 제시하기 위해 지식 뉴런(knowledge neurons) 개념 제시
- BERT를 위한 문장 맥락 완성(cloze) 작업에서 우리는 관계형 사실을 주어졌을 때 해당 사실을 표현하는 뉴런들을 식별하기 위한 knowledge attribution method 제안
- 지식 뉴런들의 활성화가 해당 사실의 표현과 긍정적으로 상관관계가 있다는 것을 발견하였고, 사전 학습된 모델을 세세하게 튜닝하지 않고도 지식 뉴런을 활용하여 특정 사실적인 지식을 편집(업데이트, 삭제 등)함으로 지식 뉴런이 모델 내 지식을 어떻게 저장되고 이해하는지 연구

https://arxiv.org/pdf/2104.08696

Visualizing and Measuring the Geometry of BERT [Coenen et al. 2019]

국외





- The paper focuses on investigating the internal representation of linguistic information in BERT.
- Linguistic features in BERT appear to be represented in separate semantic and syntactic subspace.
- Fine-grained geometric representation of word senses within the model.
- The paper provides empirical descriptions of syntactic representations found in attention matrices and individual word embeddings.

https://proceedings.neurips.cc/paper_files/paper/2019/file/77d2afcb31f6493e350fca61764efb9a-Paper.pdf

EMPOWERING EXPERTISE
INEEJI
KAIST

A General Language Assistant as a Laboratory for Alignment

국외

E.1 How the HHH criteria relate to alignment

At a very high level, alignment can be thought of as the degree of overlap between the way two agents rank different outcomes. For example, if agent A completely internalizes the desires of agent B—i.e. the only desire A has is to see B's desires satisfied—we could say that agent A is maximally aligned with agent B.[20]

We believe it is difficult for an AI assistant to always be helpful, honest, and harmless towards an agent or group without also being highly aligned with that agent or group according to this definition of alignment. To see why, suppose we want the AI assistant to be aligned with a specific group of humans. Here is what each of the three conditions implies about the assistant:

- **Helpfulness:** the assistant will always try to do what is in the humans' best interests
- **Honesty:** the assistant will always try to convey accurate information to the humans and will always try to avoid deceiving them[21]
- **Harmlessness:** the assistant will always try to avoid doing anything that harms the humans

An AI assistant that is always helpful, honest, and harmless towards a group of humans will always try to act in a way that satisfies the interests of this group, including their interest not to be harmed or be misled. It is therefore likely to be highly aligned with the interests of that group of humans.

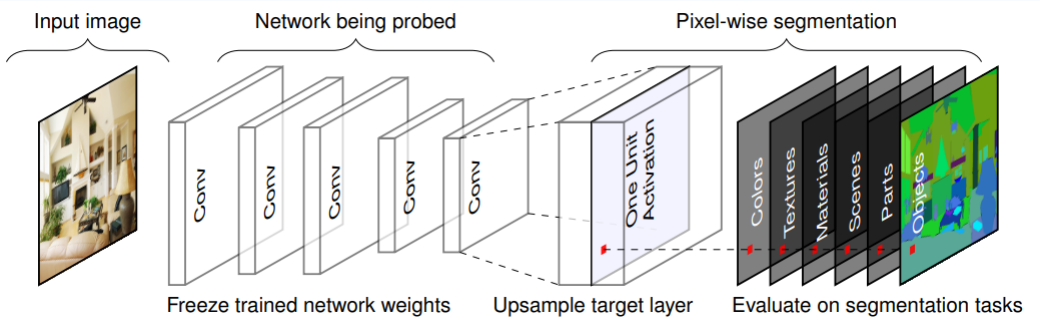
This account of alignment is still vague and leaves many open questions. In particular, it does not tell us:

- What kinds of outcome orderings are most relevant for AI alignment (preferences, idealized preferences, wellbeing, ethical rankings, etc.)
- The degree to which these outcome orderings are objective or subjective
- Which agents the AI systems should be aligned to (users, developers, humanity, etc.)
- How AI systems can or should aggregate different outcome orderings if they are aligned to more than one agent
- What is the precise formulation of "overlap between outcome rankings"
- How large or small the space of maximally aligned agents is, given the above

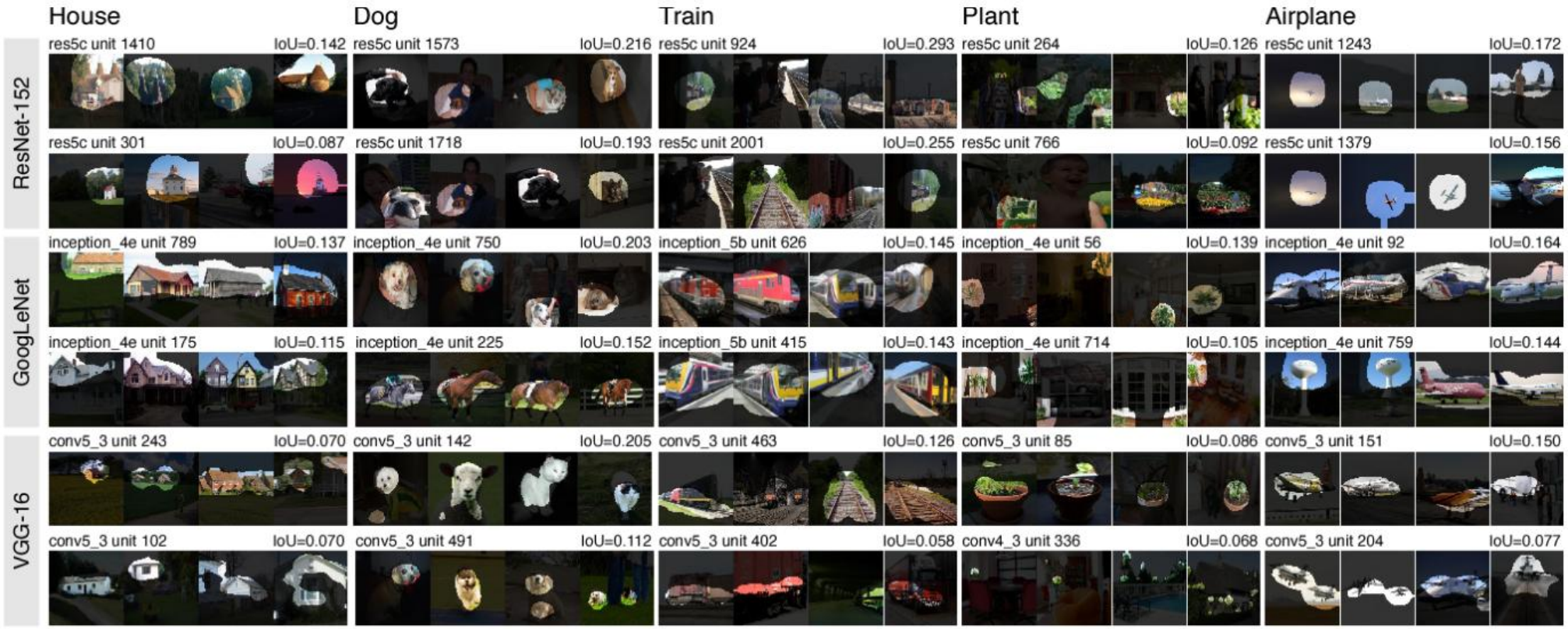
https://arxiv.org/abs/2112.00861

- Amanda Askell Elhage et al., Anthropic (2021.12)
- large language models을 개발할 때 human values(helpful, honest, and harmless)와 일치하는 a general-purpose, text-based assistant를 개발 필요
- 이 논문은 초기 탐구로, prompting와 같은 간단한 기준 기술과 평가를 연구
- large language models을 개발할 때 Modest intervention을 했을 경우, 대형 모델의 성능을 저하시키지 않으면서 다양한 alignment 평가에 일반화 연구, alignment와 관련된 여러 훈련 목표의 크기에 따른 동향 연구, 선호도에 대한 세부 조정 시 sample 효율성을 향상시키기 위한 'preference model pre-training' stage of training 연구

Network Dissection: Quantifying Interpretability of Deep Visual Representations



- The paper introduces a framework for assessing the interpretability of latent representations in CNNs.
- The proposed method evaluates the alignment between individual hidden units and a set of semantic concepts by leveraging a diverse dataset of visual concepts.
- Hidden units are assigned semantic labels related to objects, parts, scenes, textures.

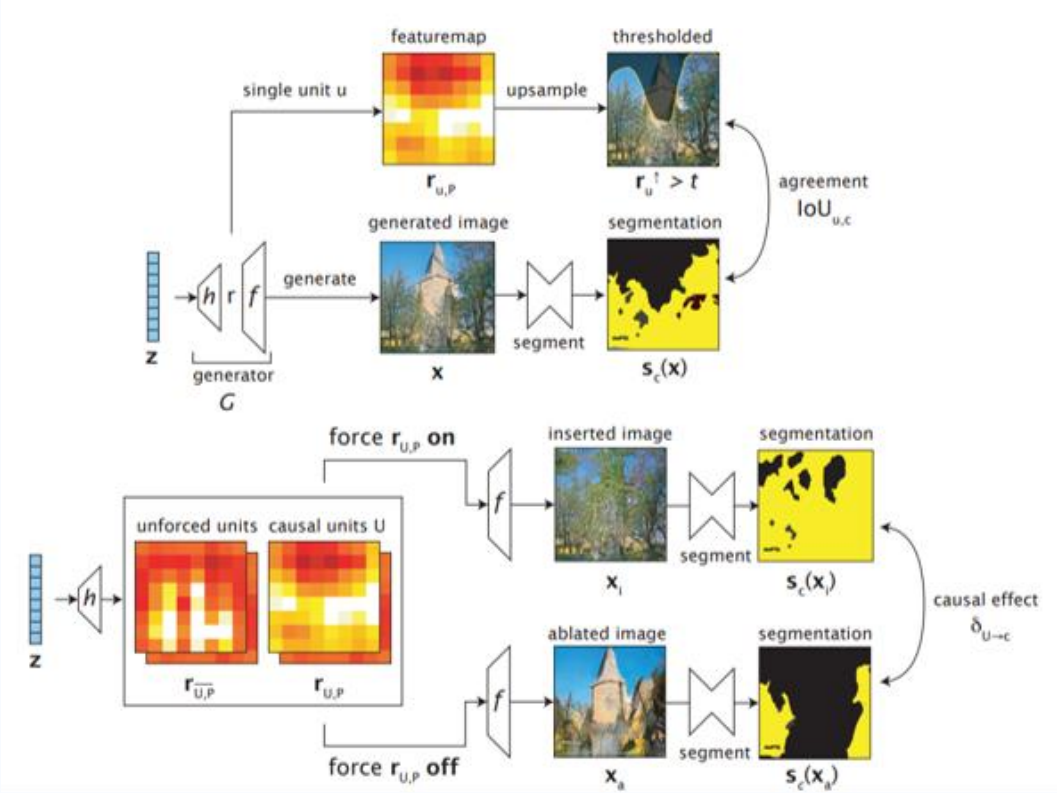


https://arxiv.org/pdf/1704.05796.pdf

EMPOWERING EXPERTISE
iNEEJI
KAIST

GAN Dissection: Visualizing and Understanding Generative Adversarial Networks

국외

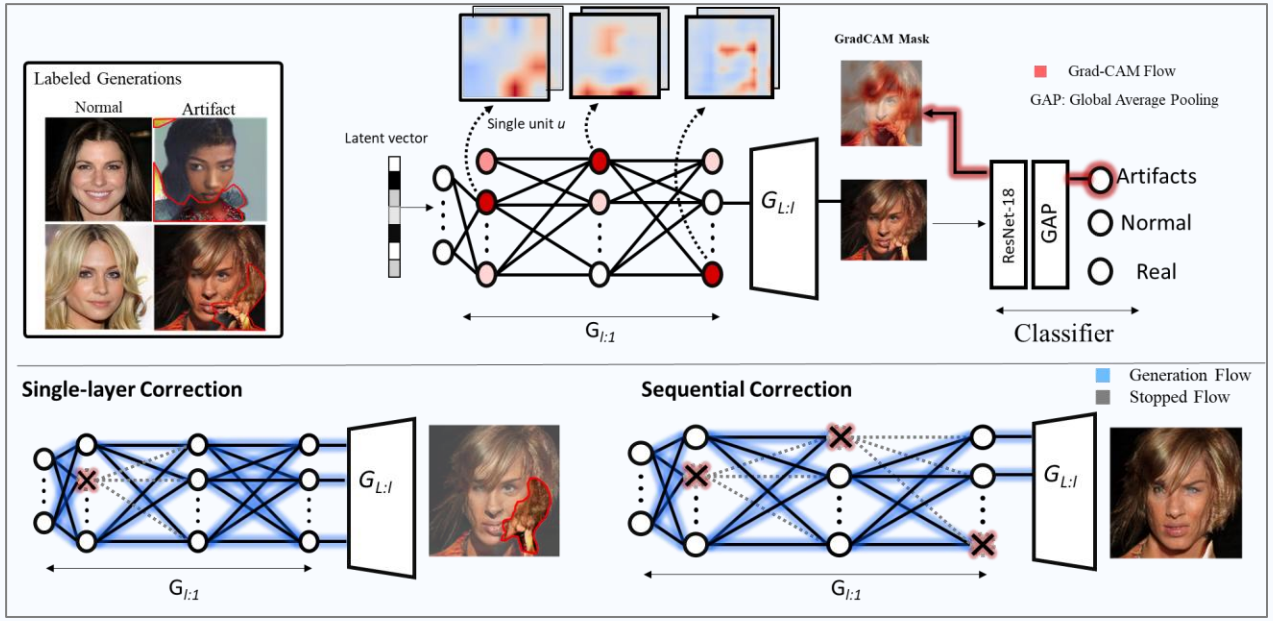


- GANs have achieved impressive results in real-world applications, but their internal representation and artifacts are not well understood.
- Provide an analytic framework to visualize and understand GAN at different levels.
- The causal effect of interpretable units is quantified by measuring interventions' ability of control object in the output.
- Practical applications of the framework including comparing internal representations, improving GANs by removing artifact-causing units, and interactively manipulating objects in a scene.

https://openreview.net/pdf?id=Hyg_X2C5FX

Automatic Correction of Internal Units in Generative Neural Networks

국내

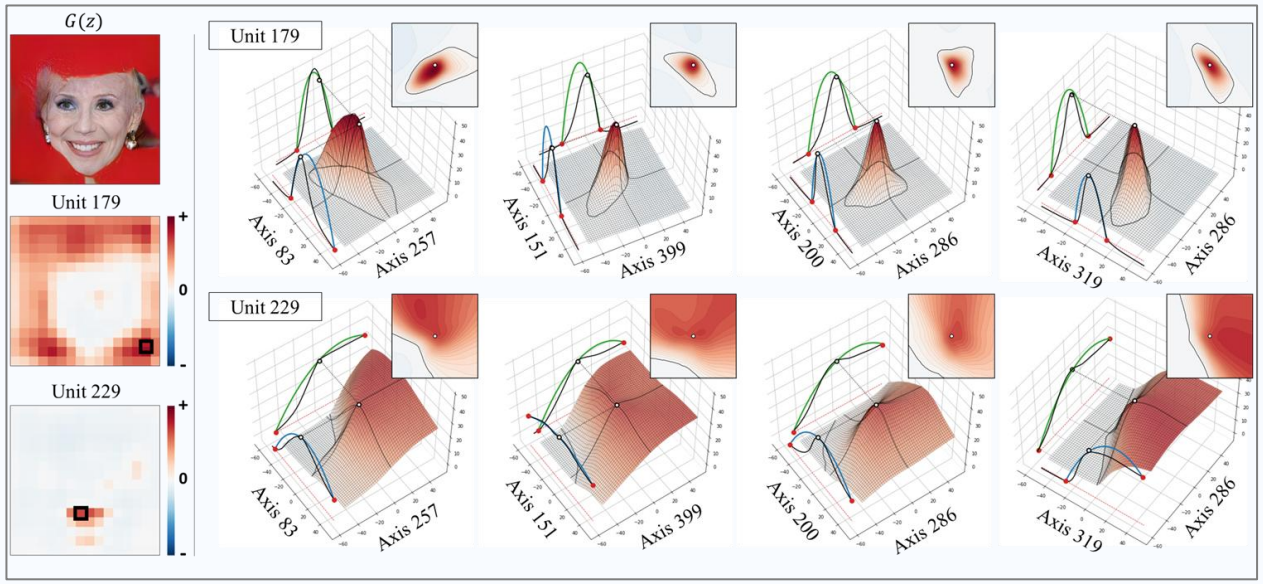


https://openaccess.thecvf.com/content/CVPR2021/papers/Tousi_Automatic_Correction_of_Internal_Units_in_Generative_Neural_Networks_CVPR_2021_paper.pdf

- Tousi and Jeong et al., CVPR (2021.07)
- 생성모델의 비약적 발전에도 불구하고 자연스럽지 않은 이상생성들이 존재함.
- 적대적 생성 신경망에서 이상생성을 유발하는 특징맵을 자동으로 찾고 수정하여, 정상적인 이미지를 생성하도록 하는 연구.
- 이미지를 직접 정상/이상 생성으로 분류한 데이터를 바탕으로 분류기를 학습하고, 해당 분류기에 XAI 기법 중 하나인 Grad-CAM을 적용하여 이미지에서 이상 생성으로 분류하도록 하는 부분을 찾아, 해당 이상생성 마스크와 비슷한 특징맵을 타겟 적대적 생성 신경망에서 찾아 수정하는 알고리즘.
- 여러 계층에 연속적으로 적용해, 이상생성이 제거되는 효과를 확인 함.

An Unsupervised Way to Understand Artifact Generating Internal Units in Generative Neural Networks

국내



https://ojs.aaai.org/index.php/AAAI/article/view/19989

- Jeong et al., AAAI (2022.06)
- 생성모델의 비약적 발전에도 불구하고 자연스럽지 않은 이상생성들이 존재함.
- 적대적 생성 신경망에서 이상생성과 관련된 뉴런을 비지도 방식으로 자동으로 찾고, 이를 이용하여 이성생성을 자동으로 판별하는 연구.
- 이상생성부분에 위치한 뉴런들이 정상생성부분에 위치한 뉴런들에 비해 공통적으로 국소적으로 활성화되는 경향성을 발견하여, 활성화도의 곡률을 바탕으로 이상생성의 정도(국소활성화도)를 측정함.
- 국소활성화도의 높고 낮음을 통해 이상생성과 정상생성을 효과적으로 판별할 수 있음을 보임.

Can We Find Neurons that Cause Unrealistic Images in Deep Generative Networks?

국내

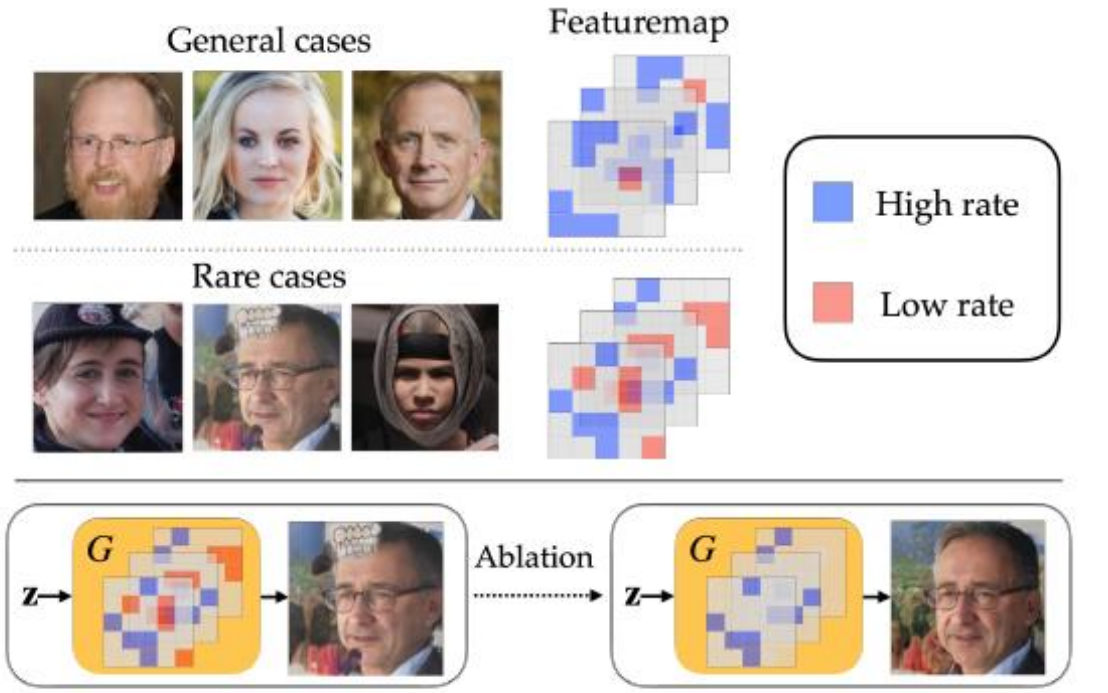


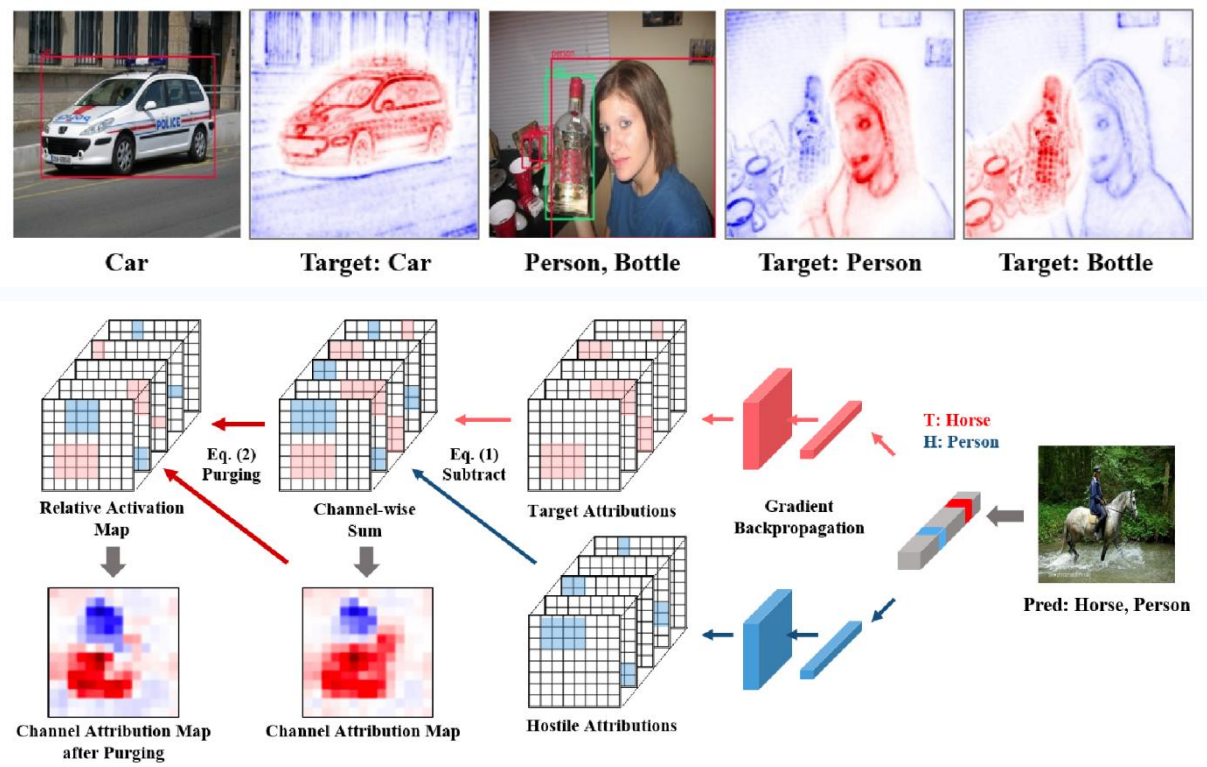
Figure 3: (Top) Comparison of feature maps between general samples and rare samples and (Bottom) our proposed ablation process. The red color denotes rarely activated neurons and the blue color denotes frequently activated neurons.

- Hwanil Choi et al. (IJCAI 2022)
- 적대적 생성 네트워크(Generative Adversarial Networks, GANs)는 고화질 이미지 생성에서 놀라운 성능을 보여주고 있지만, 항상 실제 사진같은 생성을 보장하지는 않으며. 때때로 아티팩트(artifact)라 불리는 결함이 있거나 부자연스러운 이미지를 생성
- 본 논문에서는 생성 모델 내부 뉴런들에 대한 통계치와 역할을 분석함으로써, 드물게 활성화되는 뉴런들이 이미지의 다양한 특징을 생성하다 실패한 것이 아티팩트와 연관이 있다는 것을 실험적으로 보임
- 앞의 관측을 통하여 생성 이미지에서 결함이 있는 부분을 수리하여 정상적인 이미지를 생성하는 '연쇄적 제거' 방법을 제안

https://arxiv.org/abs/2201.06346

Interpreting Deep Neural Networks with Relative Sectional Propagation by Analyzing Comparative Gradients and Hostile Activations

국내



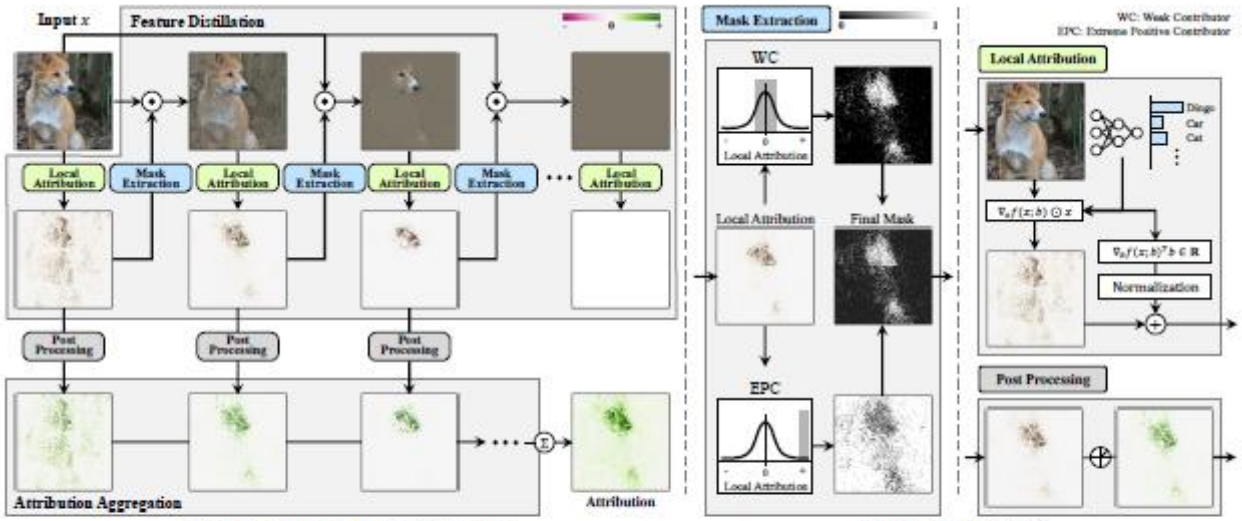
https://ojs.aaai.org/index.php/AAAI/article/view/17380

- Nam et al., AAAI (2021.02)
- 다중 클래스 이미지 분류기에 적용 가능한 입력 기여도 방법을 제안함
- 기존의 입력 기여도 방법들은 다중 클래스 이미지 분류기를 설명하는데 있어서 1) 설명의 디테일함이 부족하거나 2) 타겟 클래스별로 설명을 구분하지 못하는 문제가 있었음
- 다중 클래스 이미지 분류기가 예측한 클래스들 중 설명을 원하는 target class와 그외의 hostile class간의 기여도를 분리하여 계산하고 두 기여도가 상충되는 위치의 기여도를 제거함으로써 타겟 클래스에 해당하는 입력 기여도를 계산함

iNEEJI EMPOWERING EXPERTISE
KAIST

Distilled Gradient Aggregation: Purify Features for Input Attribution in the Deep Neural Network

국내



(a) Distilled Gradient Aggregation (DGA) (b) Modules (detailed)

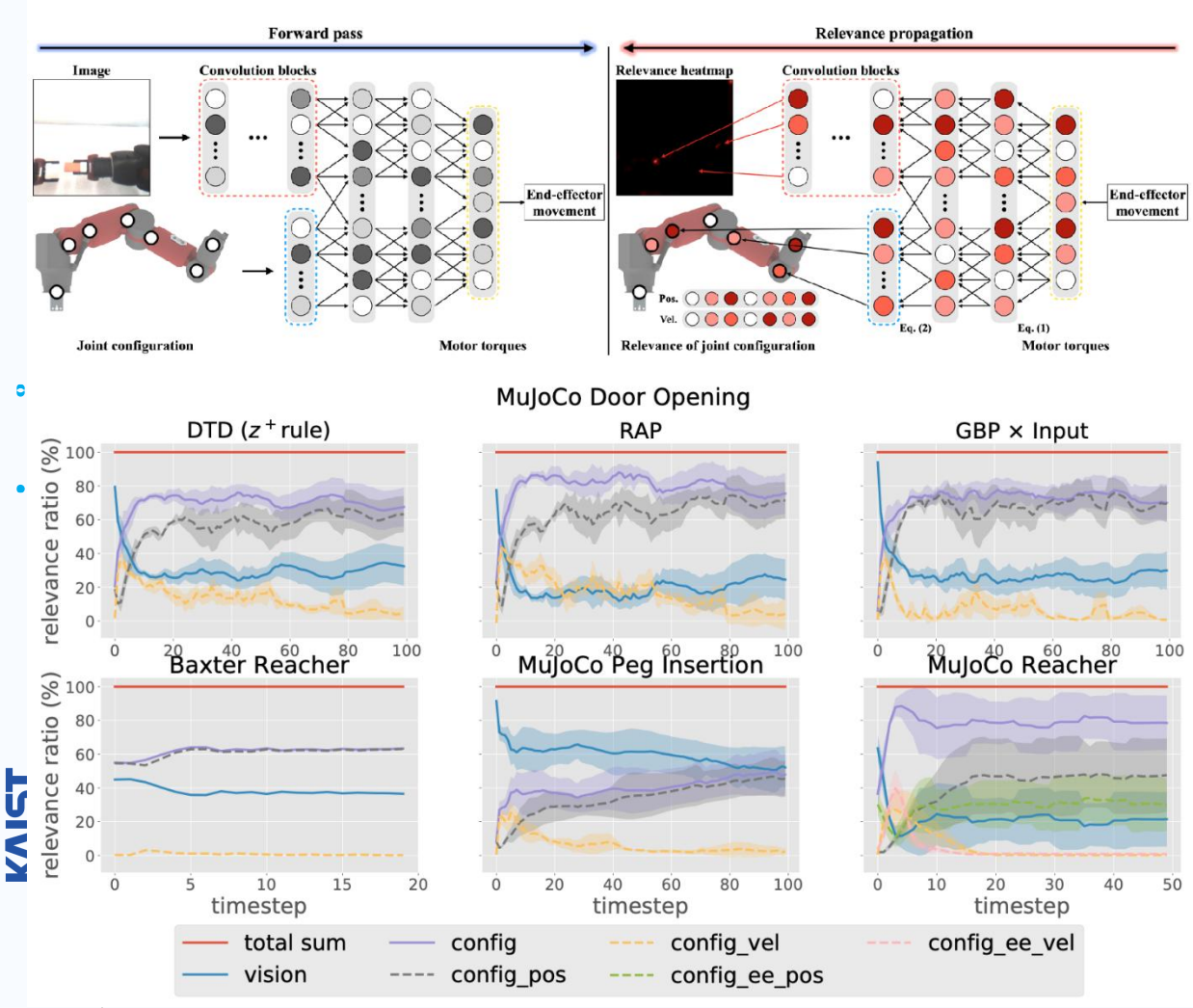
Figure 1: An illustration of Distilled Gradient Aggregation (DGA) and details of modules included in the computation of DGA. DGA generates the discrete sequence of anchor points by distilling the input features. We identify that aggregating the local attributions along the distillation sequence provides the class (*Dingo*) oriented attribution heatmap. The details are described in the Section 3.2

- Jeon et al., NeurIPs (2022.11)
- 대표적인 입력 기여도(input attribution) 계산 방법인 Integrated Gradient(IG)는 입력 노이즈에 취약하다는 단점이 존재하며, 이는 IG 계산 과정에서 사용되는 입력 경로에 따라 딥러닝 모델 내부의 선형 영역(linear region)이 달라지기 때문임을 보임
- 본 논문은 딥러닝 내부 영역에 따라 국소적 입력 경사를 누적하여 입력의 기여도를 측정하는 Distilled Gradient Aggregation(DGA) 기술 개발
- DGA를 통해 이미지 분류 딥러닝 모델 설명의 잡음에 강건성을 부여하고, 기존 IG, FullGrad 등의 입력 기여도 방식보다 더 뛰어난 정량적 결과를 보임

https://openreview.net/pdf?id=OkLee4SfLKh

Explaining the Decisions of Deep Policy Networks for Robotic Manipulations

국내

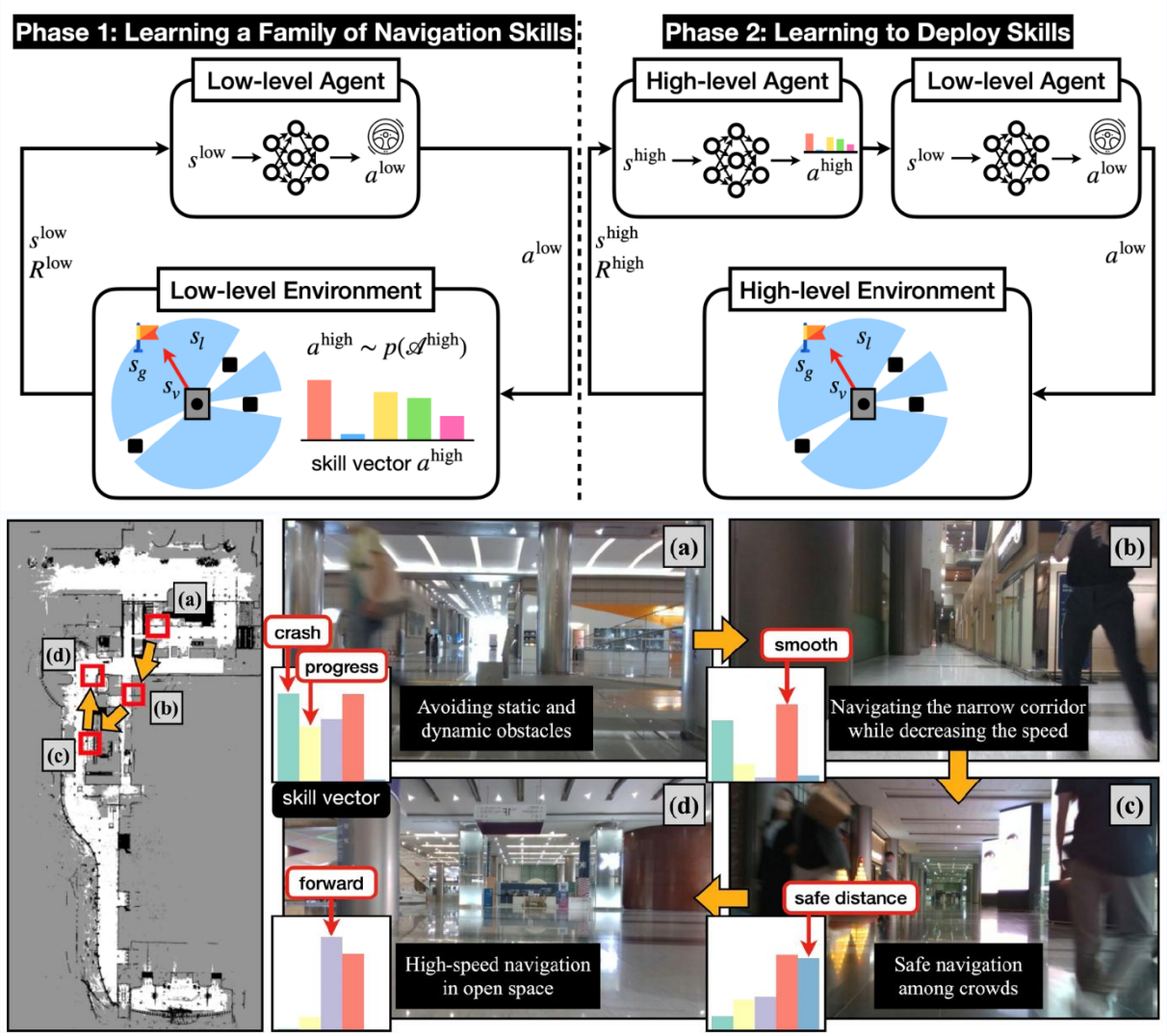


- Kim et al., IROS (2021.09)
- 이미지 분류기 모델을 설명하기 위해 주로 사용되었던 입력 기여도 방법을 로봇 매니퓰레이션을 위한 policy network로 확장하여 적용할 수 있음을 보임
- 모든 입력 및 출력 값이 양수인 이미지 분류 작업과 달리 robot policy network는 입력 및 출력 값이 음수를 포함할 수 있는데, 기존의 입력 기여도 방법을 적용할 경우 입력의 기여도가 정확히 측정되지 않는 문제를 발견
- 제안한 기여도 전파 방법을 통해 음수를 포함하는 입력 및 출력층의 기여도를 측정함
- Robot policy를 rollout하며 시간에 따른 각 입력 특징들의 기여도를 정량적으로 분석함으로써, 각 시간대별 로봇팔을 조작하는데 있어서 공통적으로 중요했던 특징들을 발견함

https://ieeexplore.ieee.org/document/9636594/

Adaptive and Explainable Deployment of Navigation Skills via Hierarchical Deep Reinforcement Learning

국내



- Lee and Kim et al., ICRA (2023.05)
- 강화학습 기반의 로봇 자율주행 기술을 해석할 수 있는 계층형 모델구조를 제안함
- 단일 policy model을 사용하여 LiDAR 혹은 카메라 이미지 등의 raw sensory observation으로부터 모바일 로봇의 선속도/각속도를 제어하던 기존의 강화학습 방법들과 달리, 로봇 제어를 상위/하위 계층의 policy model로 분리함
- High-level policy가 raw sensory observation으로부터 출력하는 skill vector를 통해 현재 로봇이 처한 환경의 semantic을 이해하고, 이 semantic에 적합한 low-level policy를 자동으로 선택하여 로봇을 제어함
- 장애물이 많은 환경, 좁은 복도 등 각 주행 시나리오별로 high-level policy가 환경에 적합한 skill vector를 선택하는 것을 보임으로써 제안한 계층형 모델구조가 로봇이 처한 환경의 semantic을 정확히 이해하고 있는 것을 확인함

https://arxiv.org/abs/2305.19746

Examples are not Enough, Learn to Criticize! Criticism for Interpretability

국외

MMD-critic can also be applied to intermediate data representations that have passed through an embedding stage or a part of a larger model. Here, a single ImageNet category is visualized after passing through an image embedding. In this representational space, full-color frontal images are prototypical of this category, whereas black-and-white images and odd viewing angles are exceptions.

- Been Kim et al. (NIPS 2016)
- Presents the MMD-critic (maximum mean discrepancy-critic), a scalable framework for prototype and criticism selection to improve the interpretability of ML methods.
- MMD-critic uses the MMD statistic as a measure of similarity between points and potential prototypes to select prototypes that maximize the statistic.
- Together with prototypes, MMD-critic selects criticism samples i.e. samples that are not well-explained by the prototypes using a regularized witness function score.
- A human subject pilot study shows that including the criticism together with prototypes is helpful for an end-task.

https://papers.nips.cc/paper_files/paper/2016/file/5680522b8e2bb01943234bce7bf84534-Paper.pdf

EMPOWERING EXPERTISE
iNEEJI
KAIST

Interpreting Internal Activation Patterns in Deep Temporal Neural Networks by Finding Prototypes

국내

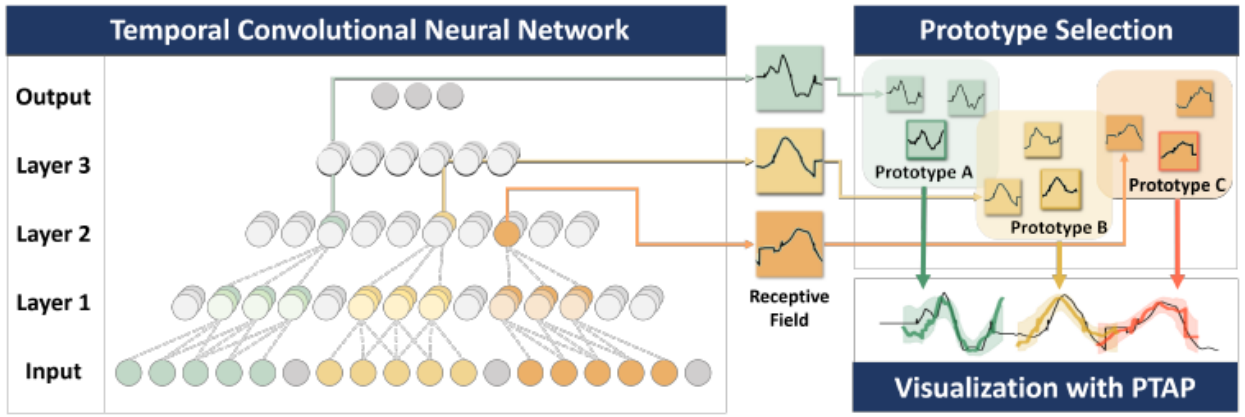


Figure 1: Our proposed framework, PTAP (Prototypes of Temporally Activated Patterns), extracts highly activated periods from activation maps of a temporal convolutional neural network, and categorizes highly activated periods using prototype selection method, and then visualizes those patterns on the original inputs. By doing so, it identifies temporal patterns activating each channel most and makes users to easily verify which patterns are commonly observed from trained data.

- S. Cho et al., KDD (2021.08)
- 딥러닝 분류 모델은 다양한 시계열 데이터에도 널리 쓰이고 있지만, 시계열의 어떤 패턴을 보고 모델이 판단한 것인지 알기 어려움
- 본 논문은 입력 시계열 내에서 모델 판단에 주로 쓰인 대표적인 패턴(Prototypes)을 분포 근사 기법을 통해 찾아내는 알고리즘을 개발
- 추가적인 외부 정보 없이 딥러닝 내부의 feature map을 이용한 알고리즘을 개발하여, 모델이 학습한 feature를 더 정확하게 찾아낼 수 있음

https://dl.acm.org/doi/abs/10.1145/3447548.3467346

Understanding Black-box Predictions via Influence Functions [PW Kol et al. 2017]

국외

$$\mathcal{I}_{\text{pert,loss}}(z, z_{\text{test}}) = -\nabla_\theta L(z_{\text{test}}, \hat{\theta})^\top H_{\hat{\theta}}^{-1} \nabla_x \nabla_\theta L(z, \hat{\theta}).$$

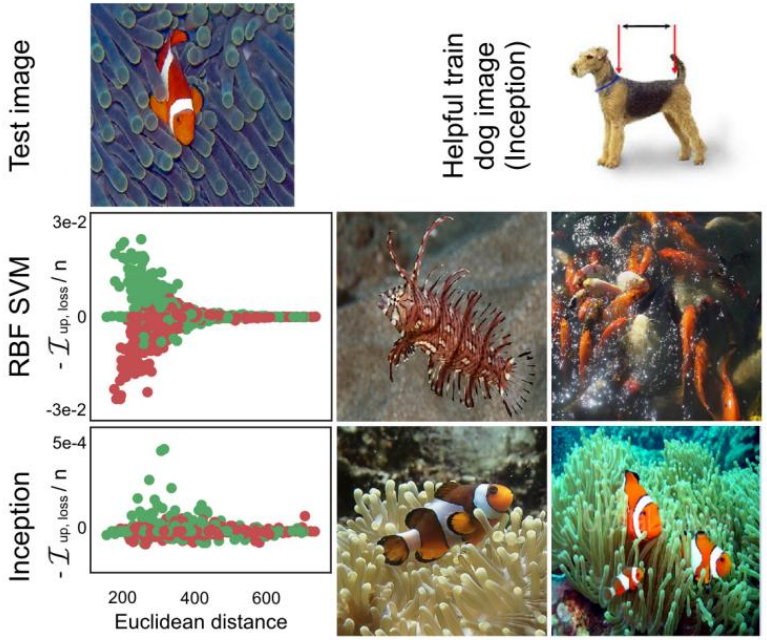


Figure 4. **Inception vs. RBF SVM. Bottom left:** $-\mathcal{I}_{\text{up,loss}}(z, z_{\text{test}})$ vs. $\|z - z_{\text{test}}\|_2^2$. Green dots are fish and red dots are dogs. **Bottom right:** The two most helpful training images, for each model, on the test. **Top right:** An image of a dog in the training set that helped the Inception model correctly classify the test image as a fish.

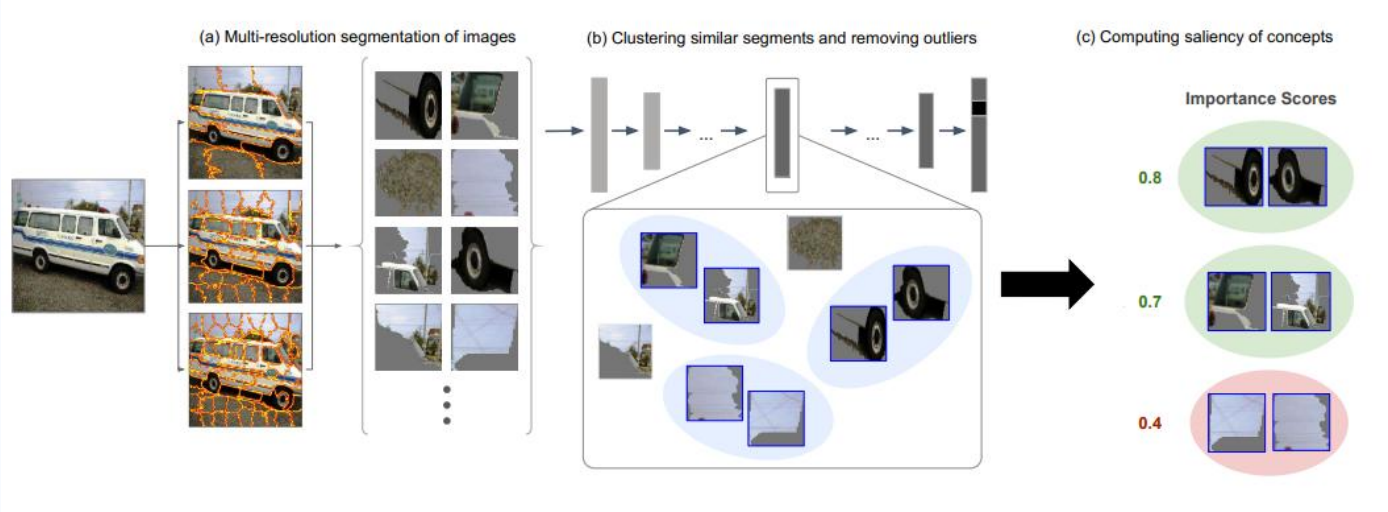
- The paper addresses the challenge of explaining prediction made by black-box models.
- Influence function are used to trace a model's prediction back to its training data, allowing identification of the training points that have the most impact on a given prediction.
- Experimental results conducted on linear models and convolutional neural networks show the utility of influence functions for understanding model behavior, debugging models, detect dataset errors and even creating visually-indistinguishable training-set attacks.

https://proceedings.mlr.press/v70/koh17a/koh17a.pdf

EMPOWERING EXPERTISE
iNEEJI
KAIST

Towards Automatic Concept-based Explanations [Ghorbani et al. 2019]

국외



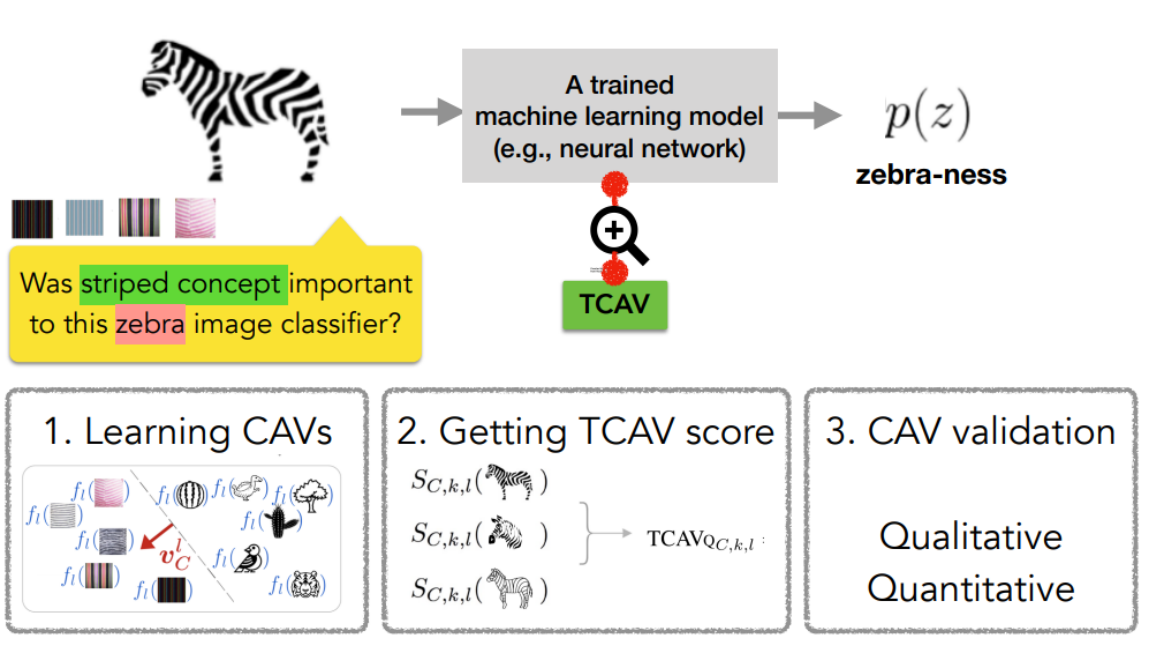
https://proceedings.neurips.cc/paper_files/paper/2019/file/77d2afcb31f6493e350fca61764efb9a-Paper.pdf

- Interpretability is a crucial area of research due to the widespread deployment of machine learning models and their impact on decision-making.
- Current explanation methods focus on feature importance scores of individual inputs but lack a systematic approach to summarize and interpret these scores.
- The paper introduces the concept-based explanation, aiming identify higher-level of human-understandable concepts that apply to the entire dataset.

iNEEJI EMPOWERING EXPERTISE
KAIST

Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vector

국내



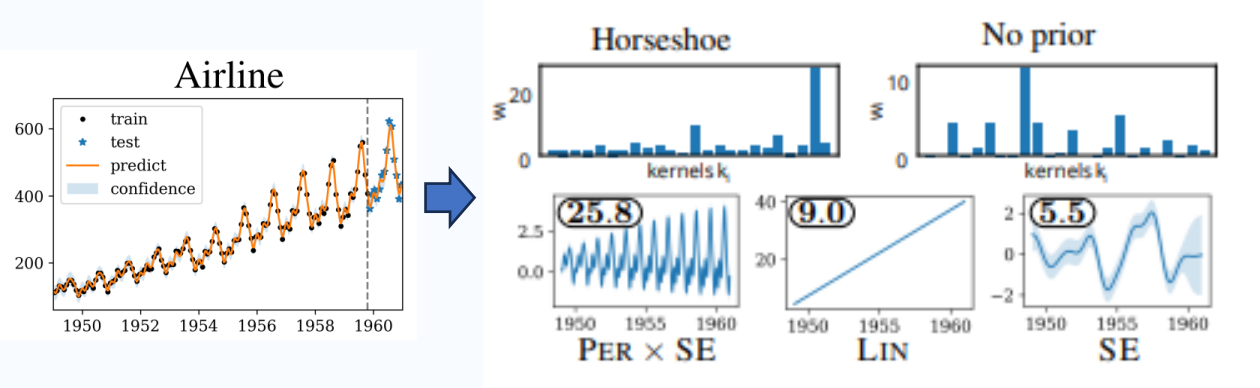
- The paper aims to provide a method to explain DNN at high-level interpretations.
- It introduces Concept Activation Vectors to interpret a neural network's internal state using human-friendly concepts.
- Testing with CAVs is proposed to quantify the importance of user-defined concept to a classification result by using directional derivatives.
- It can help determine the sensitivity of a prediction to the presence of specific features such as stripes in the case of a "zebra" prediction.

https://arxiv.org/pdf/1711.11279.pdf

EMPOWERING EXPERTISE
iNEEJi
KAIST

Learning Compositional Sparse Gaussian Process with a Shrinkage Prior [Tong et al. 2021]

국내

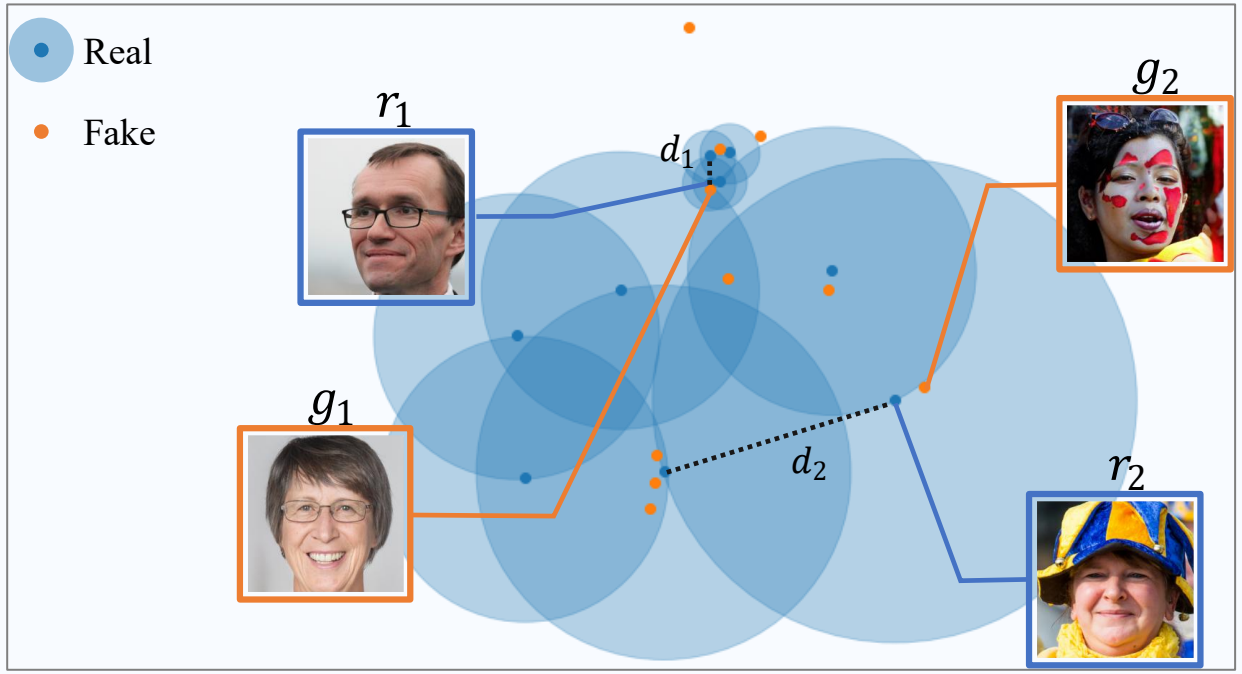


https://ojs.aaai.org/index.php/AAAI/article/view/17190/16997

- Automatic kernel composition methods have emerged to provide accurate predictions and interpretable models through search-based techniques.
- The paper introduces a method to construct groups of inducing points associated with individual additive kernels in compositional kernels.
- Instead of using a search-based approach, a novel probabilistic algorithm based on Horseshoe prior is used to learn kernel composition.
- The proposed model demonstrate the ability to capture interpretable decompositions of time series, significantly reducing computational time, and achieving competitive regression performance.

Rarity Score: A New Metric to Evaluate the Uncommonness of Synthesized Images

국내



https://openreview.net/forum?id=JTGimap_-F

- Han et al., ICLR (2023.05)
- 최근 생성모델이 활발히 연구되고, 대중에게 이용되고 있으나 생성 모델의 개별 생성의 개성에 대한 평가지표는 이전에 연구되지 않음.
- 생성 모델의 개별 생성의 개성을 평가하기 위해, 모델 학습에 쓰인 데이터셋을 바탕으로 K-Nearest Neighbor를 적용해 학습데이터의 분포를 근사하고, 개별생성의 분포 내 밀도(nearest neighbor distance)로 개별생성의 개성을 측정함.
- 해당 측정법은 모델과 상관없이 적용이 가능해, 적대적 생성신경망 뿐만 아니라 VAE 또는 Diffusion model에도 적용이 가능하며, 데이터셋에서 흔한 대표 이미지와 특징이 두드러지는 개성있는 이미지를 효과적으로 구분할 수 있음을 보임.

Confidential-PROFITT: Confidential PROof of FaIr Training of Trees

국외

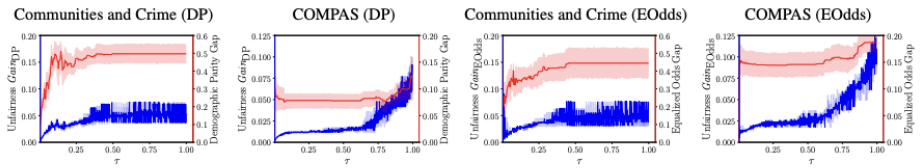


Figure 3: Relationship between fair splitting criteria and unfairness definitions. As the bound on the unfairness-based information gain decreases , the unfairness of trained decision trees decreases.

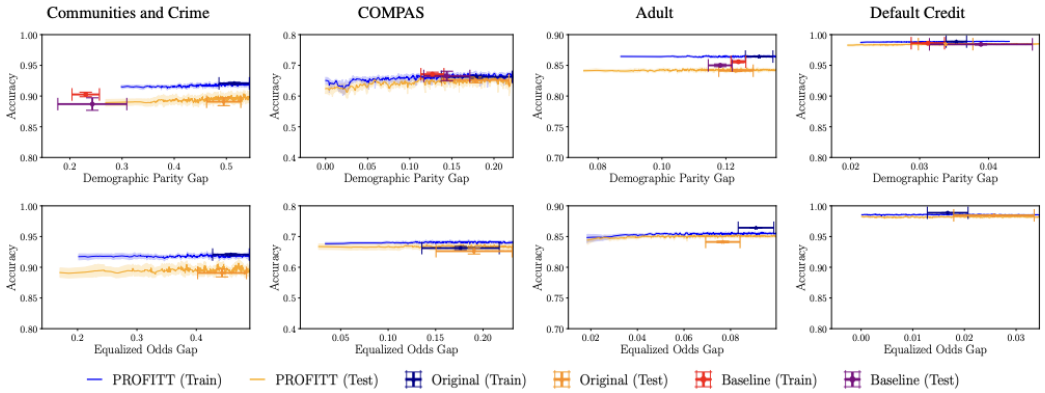


Figure 4: Effectiveness of Confidential-PROFITT in terms of demographic parity (first row) and equalized odds (second row) of random forests. Confidential-PROFITT decreases the unfairness of the random forest while keeping the accuracy close that of the original random forest trained without fairness constraints. Original indicates random forest trained without any fairness constraint.

COMPAS		Communities and Crime		Default Credit		Adult	
Eodds	Dem. Parity	Eodds	Dem. Parity	Eodds	Dem. Parity	Eodds	Dem. Parity
108.73s	73.79s	73.49s	51.88s	652.9s	372.55s	908.63s	537.59s

Table 2: Confidential-PROFITT executes the proof of fair random forests training in practical time.

- Ali Shahin Shamsabadi et al., ICLR 2023
- **모델 공정성의 사후 감사(Post hoc auditing of model fairness)는 (1) 선택된 테스트 샘플에 매우 민감하다는 것과 (2) 모델 및/또는 그 훈련 데이터를 감사인과 공유함으로 기밀성이 깨짐**
- 이 논문은 의사결정 트리의 문맥에서 훈련의 기밀성 증명을 제공하기 위한 방법인 Confidential-PROFITT를 소개
- Fair decision tree learning algorithms along with customized zero-knowledge proof protocols을 통해 기밀성을 보장하고 제 3자가 공정성을 증명
- 각 노드에 대한 정보 이득을 민감한 속성과 관련하여 제한함으로써 최종 트리의 불공정성을 감소시킬 수 있다는 것을 보여줌으로써 기업이 Confidential-PROFITT을 사용하여 의사결정 트리의 공정성 검증
- Confidential-PROFITT을 트리 앙상블에 적용할 수 있도록 확장

https://openreview.net/forum?id=iIfDQVyuFD

"Why Should I Trust You?" Explaining the Predictions of Any Classifier

국외

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

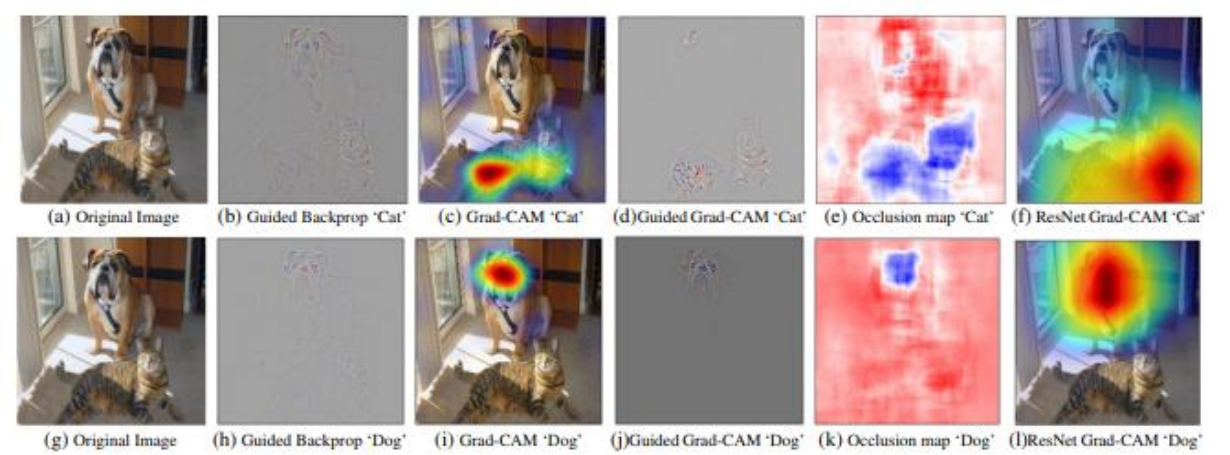
- Marco Tulio Ribeiro et al. (KDD 2016)
- Proposes LIME, an algorithm that can explain the predictions of any classifier or regressor in a faithful way, by approximating it locally with an interpretable model.
- Comprehensive evaluation on various scenarios that requires trust: deciding if one should trust a prediction, choosing between models, improving an untrustworthy classifier.
- Also shows how understanding the predictions of a neural network on images helps practitioners know when and why they should not trust a model.

https://arxiv.org/abs/1602.04938

iNEEJI EMPOWERING EXPERTISE
KAIST

Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

국외



- The paper introduces a technique for producing visual explanations that make CNN-based models more transparent.
- It uses gradient flowing into the final convolution layer to generate a coarse localization map, highlight important regions in the image for predicting a target concept.
- For classification tasks, Grad-CAM provides insights into failure modes, robustness against adversarial images.
- Also give a method to identify important neurons and combine it with neuron names to provide textual explanations for model decisions.

https://arxiv.org/pdf/1610.02391.pdf

Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities

국외

Recent research published in 2020 and 2021 has focused on most of these challenges and research directions

Challenges and research directions	2017	2018	2019	2020	2021	Total
Towards more formalism	1	4	4	11	11	31
Explanations and the nature of user experience and expertise	1	1	2	6	2	12
XAI for trustworthiness AI	1	1	0	4	5	11
Multidisciplinary research collaborations	0	1	1	3	5	10
Interpretability vs. performance trade-off	0	1	1	3	2	7
XAI for non-image, non-text, and heterogeneous data	0	0	0	3	3	6
Explainability methods composition	0	3	0	1	1	5
Causal explanations	0	1	0	3	1	5
Challenges in the existing XAI models/methods	0	1	1	1	2	5
Contrastive and counterfactual explanations	0	0	1	0	2	3
Communicating uncertainties	0	0	0	2	0	2
Time constraints	1	0	0	1	0	2
Natural language generation	0	0	0	1	0	1
Analyzing assumption-free black-box models, not assumption-based data models	0	0	1	0	0	1
Reproducibility	0	0	0	1	0	1
The economics of explanations	0	1	0	0	0	1

- Waddah Saeed, Christian Omlin (2023)
- Explainable AI (XAI) has been proposed to make AI more transparent and thus advance the adoption of AI in critical domains.
- A systematic meta-survey of challenges and future research directions in XAI organized in two themes:
  (1) general challenges and research directions of XAI
  (2) challenges and research directions of XAI based on machine learning life cycle's phases: design, development, and deployment.

https://www.sciencedirect.com/science/article/pii/S0950705123000230

EMPOWERING EXPERTISE
iNEEJI
KAIST

Robust Physical-World Attacks on Deep Learning Visual Classification

국외

Table 1: Sample of physical adversarial examples against LISA-CNN and GTSRB-CNN.

Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5′ 0°					
5′ 15°					
10′ 0°					
10′ 30°					
40′ 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%

- Kevin Eykholt et al. (2017)
- 최근 연구들에 따르면 DNN은 아주 작은 입력값의 변화에도 취약하다는 점이 밝혀짐. (small-magnitude perturbations added to the input)
- 다양한 물리적 환경에서 강건한 도로 표지판 분류기를 만들기 위한 적대적 예제를 생성하는 알고리즘인 Robust Physical Perturbations (RP2)을 제안
- Contribution : Understanding the space of physical adversarial examples when the objects themselves are physically perturbed.

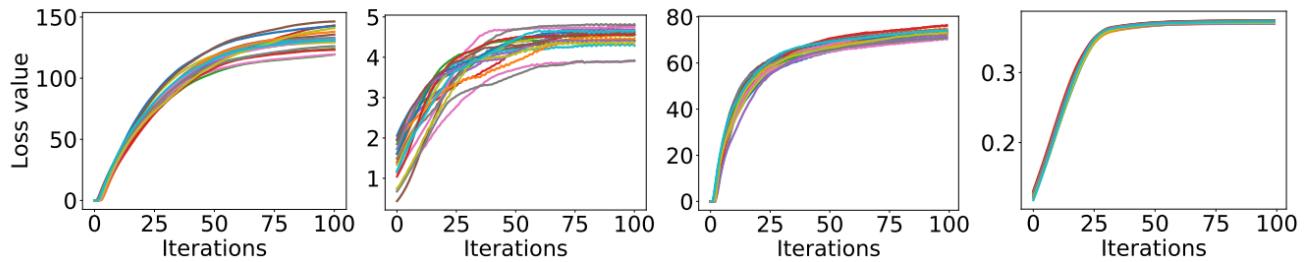
https://openaccess.thecvf.com/content_cvpr_2018/papers/Eykholt_Robust_Physical-World_Attacks_CVPR_2018_paper.pdf

EMPOWERING EXPERTISE
iNEEJI
KAIST

Towards Deep Learning Models Resistant to Adversarial Attacks

국외

We observe that the loss achieved by the adversary increases in a fairly consistent way and plateaus rapidly when performing projected ℓ_∞ gradient descent for randomly chosen starting points inside $x + \mathcal{S}$ (see Figure 1).



https://arxiv.org/pdf/1706.06083.pdf

- Aleksander Madry et al. (2019)
- 최근 연구들에 따르면 적대적 공격에 취약하다는 점이 딥러닝 모델의 근본적 약점이라고 알려짐
- 기존 연구들은 제한된 타입의 적대적 공격만을 가정하고 그에 대한 특정 방어 메커니즘만을 제시하였으나, 본 연구는 "Robust Optimization" 이론에 기반하여 다양한 적대적 예제 타입을 정의하고, 이런 적대적 공격에 강건한 Universal Robust Classifiers를 제안함
- Projected Gradient Descent (PGD)를 "first-order adversary"로 하여 실험한 결과 MNIST, CIFAR10 모델에 대해 각각 89%, 46% accuracy를 달성함. (기존의 black box/transfer 타입 공격에 대해서는 MNIST 95%, CIFAR10 64% accuracy로 최고 수준의 성능을 보임)

To Trust Or Not To Trust A Classifier

국외

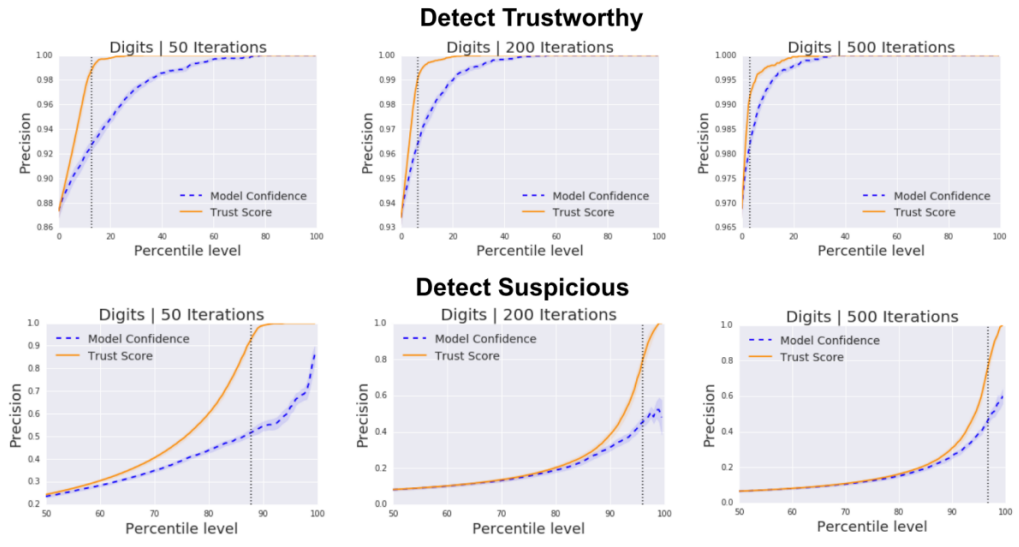


Figure 2: We show the performance of trust score on the Digits dataset for a neural network as we increase the accuracy. As we go from left to right, we train the network with more iterations (each with batch size 50) thus increasing the accuracy indicated by the dotted vertical lines. While the trust score still performs better than model confidence, the amount of improvement diminishes.

https://arxiv.org/abs/1805.11783

- Heinrich Jiang et al. (2018)
- 기계학습 분야에서 대부분의 연구는 분류 모델의 성능을 개선하는 데에 집중되고 있으나, 분류기의 예측 결과를 신뢰할 것인지 여부를 판별하는 것에 대한 연구 역시 중요함.
- 분류기 결과를 신뢰할지 여부를 나타내는 척도로서 Trust Score를 제안: Measures the agreement between the classifier and a modified nearest-neighbor classifier on the testing example.
- High (low) trust scores produce surprisingly high precision at identifying correctly (incorrectly) classified examples.

EMPOWERING EXPERTISE
iNEEJI
KAIST

Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples

국외

Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples

Anish Athalye [*1] Nicholas Carlini [*2] David Wagner [2]

Abstract

We identify obfuscated gradients, a kind of gradient masking, as a phenomenon that leads to a false sense of security in defenses against adversarial examples. While defenses that cause obfuscated gradients appear to defeat iterative optimization-based attacks, we find defenses relying on this effect can be circumvented. We describe characteristic behaviors of defenses exhibiting the effect, and for each of the three types of obfuscated gradients we discover, we develop attack techniques to overcome it. In a case study, examining non-certified white-box-secure defenses at ICLR 2018, we find obfuscated gradients are a common occurrence, with 7 of 9 defenses relying on obfuscated gradients. Our new attacks successfully circumvent 6 completely, and 1 partially, in the original threat model each paper considers.

apparent robustness against iterative optimization attacks: *obfuscated gradients*, a term we define as a special case of gradient masking (Papernot et al., 2017). Without a good gradient, where following the gradient does not successfully optimize the loss, iterative optimization-based methods cannot succeed. We identify three types of obfuscated gradients: *shattered gradients* are nonexistent or incorrect gradients caused either intentionally through non-differentiable operations or unintentionally through numerical instability; *stochastic gradients* depend on test-time randomness; and *vanishing/exploding gradients* in very deep computation result in an unusable gradient.

We propose new techniques to overcome obfuscated gradients caused by these three phenomena. We address gradient shattering with a new attack technique we call Backward Pass Differentiable Approximation, where we approximate derivatives by computing the forward pass normally and computing the backward pass using a differentiable approx-

- Anish Athalye et al. (2018)
- 기존에 제안되었던 대부분의 방어기법들이 실질적으로 방어효과가 없었고, 현존하는 방어기법들을 다시금 공격이 가능하다는 것을 보여줌.
- 기존의 방어기법들의 문제점을 유형별로 정리하여 분석
- Adversarial training을 제외한 사실상 대부분의 방어기법들을 뚫을 수 있음
- 2018년도 ICLR에서 제시된 whitebox defense 논문들을 면밀히 분석해서 9개중 7개가 공통적으로 문제가 있음을 지적하고, 그중 6개를 완전히 뚫을 수 있음을 보여줌.
- 미분이 불가능한 연산을 우회하여 뚫는 Backward Pass Differentiable Approximation (BPDA)를 제안

https://arxiv.org/pdf/1802.00420.pdf

EMPOWERING EXPERTISE
iNEEJI
KAIST

Adversarial Examples are not Bugs, They Are Features

국외

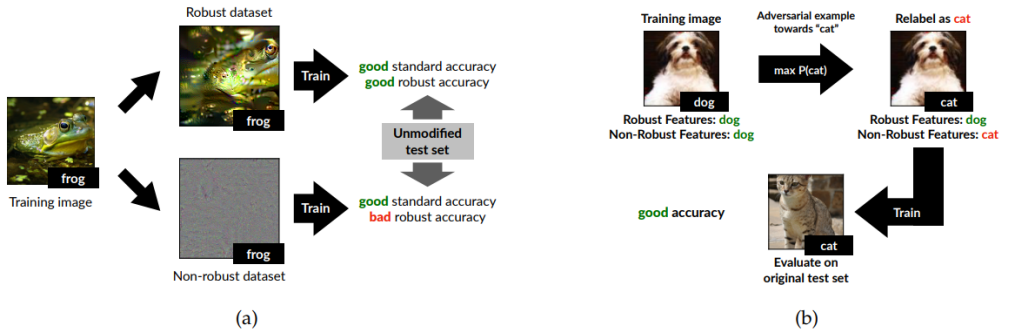


Figure 1: A conceptual diagram of the experiments of Section 3. In (a) we disentangle features into combinations of robust/non-robust features (Section 3.1). In (b) we construct a dataset which appears mislabeled to humans (via adversarial examples) but results in good accuracy on the original test set (Section 3.2).

- Andrew Ilyas et al. (2019)
- 대부분의 기존 논문들은 adversarial example을 일종의 버그처럼 생각을 하고 그러한 버그를 어떻게 해결할 수 있을지의 관점에서 바라보았는데, 본 논문에서는 adversarial example을 일종의 feature로서 바라본다는 새로운 시각을 제시.
- Adversarial perturbation이 non-robust feature, 즉 쉽게 noise가 섞이면서 변경될 수 있는 feature 라고 설명함.
- 모델이 data를 학습할 때, 인간이 인지할 수 없지만 모델에는 큰 영향을 미치는 작은 features 까지 학습하게 되는데, adversarial perturbation이 해당 특성을 갖는 non-robust feature 의 일종이라 볼 수 있음. (adversarial example 의 특징 중에 하나인 transferability를 feature 로서 직관적으로 설명할 수 있음.)
- 데이터에서 robust와 non-robust features 를 구분할 수 있음을 보여줌.

https://transformer-circuits.pub/2021/framework/index.html#technical-details

Adversarial Training and Robustness for Multiple Perturbations

국외

Adversarial Training and Robustness for Multiple Perturbations

Florian Tramèr
Stanford University

Dan Boneh
Stanford University

Abstract

Defenses against adversarial examples, such as adversarial training, are typically tailored to a single perturbation type (e.g., small ℓ_∞-noise). For other perturbations, these defenses offer no guarantees and, at times, even increase the model's vulnerability. Our aim is to understand the reasons underlying this robustness trade-off, and to train models that are simultaneously robust to multiple perturbation types.

We prove that a trade-off in robustness to different types of ℓ_p-bounded and spatial perturbations must exist in a natural and simple statistical setting. We corroborate our formal analysis by demonstrating similar robustness trade-offs on MNIST and CIFAR10. We propose new multi-perturbation adversarial training schemes, as well as an efficient attack for the ℓ_1-norm, and use these to show that models trained against multiple attacks fail to achieve robustness competitive with that of models trained on each attack individually. In particular, we find that adversarial training with first-order ℓ_∞, ℓ_1 and ℓ_2 attacks on MNIST achieves merely 50% robust accuracy, partly because of gradient-masking. Finally, we propose *affine attacks* that linearly interpolate between perturbation types and further degrade the accuracy of adversarially trained models.

- Florian Tramèr, Dan Boneh (2019)
- Adversarial training은 보통 하나의 perturbation에 대서만 국한되어 있어, 다른 perturbation에 대해서는 방어를 보장할 수 없고 오히려 모델의 취약성을 높일 수 있음.
- 위의 robustness tradeoff 의 원인을 파악하고, 여러 perturbation 에 robust한 multi-perturbation adversarial training 방법을 제시.
- Proved that a trade-off in robustness to different types of `p-bounded and spatial perturbations must exist in a natural and simple statistical setting.

https://proceedings.neurips.cc/paper_files/paper/2019/file/5d4ae76f053f8f2516ad12961ef7fe97-Paper.pdf

Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-free Attacks

국외

Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-free Attacks

Francesco Croce [1] Matthias Hein [1]

Abstract

The field of defense strategies against adversarial attacks has significantly grown over the last years, but progress is hampered as the evaluation of adversarial defenses is often insufficient and thus gives a wrong impression of robustness. Many promising defenses could be broken later on, making it difficult to identify the state-of-the-art. Frequent pitfalls in the evaluation are improper tuning of hyperparameters of the attacks, gradient obfuscation or masking. In this paper we first propose two extensions of the PGD-attack overcoming failures due to suboptimal step size and problems of the objective function. We then combine our novel attacks with two complementary existing ones to form a parameter-free, computationally affordable and user-independent ensemble of attacks to test adversarial robustness. We apply our ensemble to over 50 models from papers published at recent top machine learning and computer vision venues. In all except one of the cases we achieve lower robust test accuracy than reported in these papers, often by more than 10%, identifying several broken defenses.

variations are using other losses (Zhang et al., 2019b) and boost robustness via generation of additional training data (Carmon et al., 2019; Alayrac et al., 2019) or pre-training (Hendrycks et al., 2019). Another line of work are provable defenses, either deterministic (Wong et al., 2018; Croce et al., 2019a; Mirman et al., 2018; Gowal et al., 2019) or based on randomized smoothing (Li et al., 2019; Lecuyer et al., 2019; Cohen et al., 2019). However, these are not yet competitive with the empirical robustness of adversarial training for datasets like CIFAR-10 with large perturbations.

Due to the many broken defenses, the field is currently in a state where it is very difficult to judge the value of a new defense without an independent test. This limits the progress as it is not clear how to distinguish bad from good ideas. A seminal work to mitigate this issue are the guidelines for assessing adversarial defenses by (Carlini et al., 2019). However, as we see in our experiments, even papers trying to follow these guidelines can fail in obtaining a proper evaluation. In our opinion the reason is that at the moment there is no protocol which works reliably and autonomously, and does not need the fine-tuning of parameters for every new defense. Such protocol is what we aim at in this work.

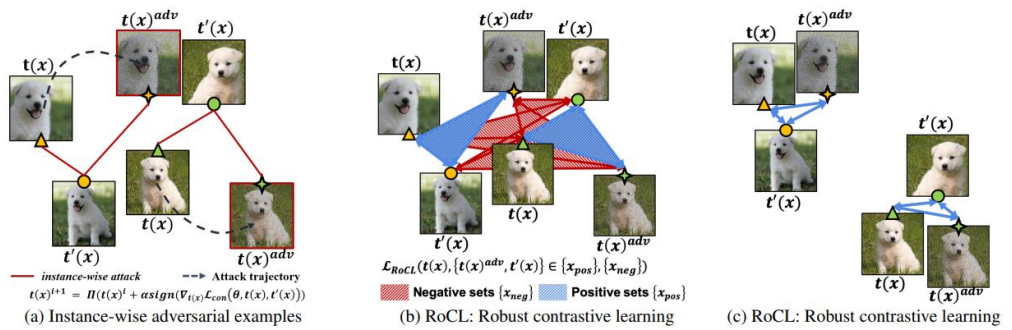
The most popular method to test adversarial robustness is the PGD (Projected Gradient Descent) attack (Madry et al.,

http://proceedings.mlr.press/v119/croce20b/croce20b.pdf

- Francesco Croce, Matthias Hein (2020)
- 이전의 저명한 방어기법들이 이후에 파훼되어 SOTA기법이라 하기 어려운 점을 지적.
- Evaluation 단계에서 Hyperparameters 의 부적절한 튜닝, gradient obfuscation or masking 에 의한 pitfalls 가 빈번하게 발생함을 지적.
- Suboptimal stepsize 와 objective function 의 문제에 대해 확장된 PGD(Projected Gradient Descent)-attack 두가지의 기법을 제시하고, parameter-free, computationally affordable and user-independent ensemble of attacks 를 위한 기존의 방법들을 결합하여 adversarial robustness를 증명함.

Adversarial Self-Supervised Contrastive Learning

국내



(a) Instance-wise adversarial examples (b) RoCL: Robust contrastive learning (c) RoCL: Robust contrastive learning

Figure 1: **Overview of our adversarial contrastive self-supervised learning.** (a) We generate instance-wise adversarial examples from an image transformed using a stochastic augmentation, which makes the model confuse the instance-level identity of the perturbed sample. (b) We then maximize the similarity between each transformed sample and their instance-wise adversaries using contrastive learning. (c) After training, each sample will have significantly reduced adversarial vulnerability in the latent representation space.

https://proceedings.neurips.cc/paper/2020/file/1f1baa5b8edac74eb4eaa329f14a0361-Paper.pdf

- Minseon Kim, Jihoon Tack, Sung Ju Hwang (2020)
- (2020년 기준) 현존하는 적대적 학습(adversarial learning) 방법들은 특히, semi-supervised adversarial learning 또한 여전히 class label을 필요로 함.
- 본 연구는 self-supervised contrastive learning framework를 제안하는데, 이로부터 labeled-data 없이 적대적 학습을 진행해 DNN의 robustness를 강화할 수 있게 됨.
- 구체적으로는, 특정 데이터 샘플의 augmented-version 과 (instance-level에서의) adversarial sample과 유사도를 늘리는 방향으로 학습을 하게 됨.
- supervised adversarial learning 방법에 필적할 만한 성능과, fine-tuning을 사용할 경우 더 좋은 robustness 보장하는 상당한 결과를 보여줌.

Poisoning and Backdooring Contrastive Learning

국외

Published as a conference paper at ICLR 2022

POISONING AND BACKDOORING CONTRASTIVE LEARNING

Nicholas Carlini
Google

Andreas Terzis
Google

ABSTRACT

Multimodal contrastive learning methods like CLIP train on noisy and uncurated training datasets. This is cheaper than labeling datasets manually, and even improves out-of-distribution robustness. We show that this practice makes *backdoor* and *poisoning* attacks a significant threat. By poisoning just 0.01% of a dataset (e.g., just 300 images of the 3 million-example Conceptual Captions dataset), we can cause the model to misclassify test images by overlaying a small patch. Targeted poisoning attacks, whereby the model misclassifies a particular test input with an adversarially-desired label, are even easier requiring control of 0.0001% of the dataset (e.g., just three out of the 3 million images). Our attacks call into question whether training on noisy and uncurated Internet scrapes is desirable.

1 INTRODUCTION

Contrastive learning (Chopra et al., 2005; Hadsell et al., 2006) trains a model that projects a data distribution onto a lower-dimensional embedding space such that similar objects in the origin space are closer together in the embedding space than dissimilar objects (Chechik et al., 2010; Sohn, 2016; Oord et al., 2018; Wu et al., 2018). Significant advances over the last years have enabled self-supervised classifiers to achieve state of the art accuracy by training on noisy and uncurated datasets (Radford et al., 2021; Tian et al., 2021), which brings two significant benefits.

- Nicholas Carlini, Adreas Terzis (2022)
- noisy and uncurated 데이터를 학습하는 CLIP과 같은 multimodal contrastive learning 의 위험성을 보여줌.
- 인터넷으로부터 스크랩한 데이터와 같은 noisy and uncurated 데이터가 multimodal contrastive model 로 하여금 targeted poisoning 과 backdoor 와 같은 공격을 할 수 있음을 보여줌.
- 기존의 연구에서 backdooring 공격을 하는데 평균 1%의 poisoning 이 필요하다는 것에 반해, multimodal contrastive model 공격에는 오직 0.01%의 감염이면 충분(backdoor attack)하고, poisoning attack 은 0.0001% 면 가능함.

https://arxiv.org/pdf/2106.09667.pdf

Robustness of Rotation-Equivariant Networks to Adversarial Perturbations

국외

ROBUSTNESS OF ROTATION-EQUIVARIANT NETWORKS TO ADVERSARIAL PERTURBATIONS

Béranger Dumont, Simona Maggio & Pablo Montalvo *†
Rakuten Institute of Technology Paris
{beranger.dumont, simona.maggio}@rakuten.com,
pablo.montalvo.leroux@gmail.com

ABSTRACT

Deep neural networks have been shown to be vulnerable to adversarial examples: very small perturbations of the input having a dramatic impact on the predictions. A wealth of adversarial attacks and distance metrics to quantify the similarity between natural and adversarial images have been proposed, recently enlarging the scope of adversarial examples with geometric transformations beyond pixel-wise attacks. In this context, we investigate the robustness to adversarial attacks of new Convolutional Neural Network architectures providing equivariance to rotations. We found that rotation-equivariant networks are significantly less vulnerable to geometric-based attacks than regular networks on the MNIST, CIFAR-10, and ImageNet datasets.

- Beranger Dumont et al. (2018)
- Adversarial examples 에 대한 deep learning models 의 취약점이 대두되고 있는데, 몇몇의 rotation-equivariant architectures 가 adversarial perturbations of a geometric kind에 robust 하다는 것을 발견하고 보여줌.
- CNN, H-Net, G-CNN, DREN, OR-Net 모델을 조사하여 H-Net을 제외하고는 rotation or translation 에 의한 adversarial samples 에 대해 robust함을 실험적으로 보여줌. H-Net은 spatial transformer-based adversarial attack 에 robust함을 보여줌.

https://arxiv.org/pdf/1802.06627.pdf

EMPOWERING EXPERTISE
iNEEJI
KAIST

The Adversarial Attack and Detection under the Fisher Information Metric

국외

The Adversarial Attack and Detection under the Fisher Information Metric

Chenxiao Zhao[*1] **P. Thomas Fletcher**[2] **Mixue Yu**[1] **Yaxin Peng**[3,4] **Guixu Zhang**[1] **Chaomin Shen**[†1,4]
[1]Department of Computer Science, East China Normal University, Shanghai, China
[2]Department of Electrical and Computer Engineering, and Department of Computer Science, University of Virginia, Virginia, USA
[3]Department of Mathematics, Shanghai University, Shanghai, China
[4]Westlake Institute for Brain-Like Science and Technology, Zhejiang, China

Abstract

Many deep learning models are vulnerable to the adversarial attack, i.e., imperceptible but intentionally-designed perturbations to the input can cause incorrect output of the networks. In this paper, using information geometry, we provide a reasonable explanation for the vulnerability of deep learning models. By considering the data space as a non-linear space with the Fisher information metric induced from a neural network, we first propose an adversarial attack algorithm termed one-step spectral attack (OSSA). The method is described by a constrained quadratic form of the Fisher information matrix, where the optimal adversarial perturbation is given by the first eigenvector, and the vulnerability is reflected by the eigenvalues. The larger an eigenvalue is, the more vulnerable the model is to be attacked by the corresponding eigenvector. Taking advantage of the property, we also propose an adversarial detection method with the eigenvalues serving as characteristics. Both our attack and detection algorithms are numerically optimized to work efficiently on large datasets. Our evaluations show superior performance compared with other methods, implying that the Fisher information is a promising approach to investigate the adversarial attacks and defenses.

sensitive to the attacks. The existence of adversarial examples has raised severe challenges for deep learning models in security-critical computer vision applications.

Understanding the mechanism of adversarial examples is a fundamental problem for defending against the attacks. Many explanations have been proposed from different facets. (Szegedy et al. 2013) first observes the existence of adversarial examples, and suggests it is due to the excessive nonlinearity of the neural networks. On the contrary, (Goodfellow, Shlens, and Szegedy 2014) suggests that the vulnerability results from the models being too linear. Despite its contradiction to the general impression, the explanation is supported by numbers of experimental results (Krotov and Hopfield 2017; Tabacof and Valle 2016; Tanay and Griffin 2016; Tramèr et al. 2017). On the other hand, by approximating the vertical direction of the decision boundary in the sample space, (Moosavidezfooli, Fawzi, and Frossard 2016) proposes to find the closest adversarial examples to the input with an iterative algorithm. (Moosavidezfooli et al. 2017b) further studies the existence of universal perturbations in state-of-the-art deep neural networks. They suggest the phenomenon is resulted from the high curvature regions on the decision

https://dl.acm.org/doi/abs/10.1609/aaai.v33i01.33015869

- Chenxiao Zhao et al. (AAAI 2019)
- Adversarial attack 에 대한 deep learning models 의 취약성에 대해 information geometry 를 이용하여 설명함.
- Neural network로부터 Fisher information metric (FIM)을 계산하고 이를 이용한 새로운 attach 알고리즘(one-step spectral attack, OSSA)을 제시함.
- FIM의 eigenvalues 를 일종의 feature로 사용하여, adversarial attack을 발견할 수 있음.
- Deep learning의 취약성에 대해 새로운 geometrical interpretation을 제시하여 OSSA의 optimality를 증명하고, 이를 deep learning model 의 취약성에 대한 basis로 쓸 수 있음을 말함.

Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples

국내

Published as a conference paper at ICLR 2018

TRAINING CONFIDENCE-CALIBRATED CLASSIFIERS FOR DETECTING OUT-OF-DISTRIBUTION SAMPLES

Kimin Lee* **Honglak Lee**[§,†] **Kibok Lee**[†] **Jinwoo Shin***
*Korea Advanced Institute of Science and Technology, Daejeon, Korea
†University of Michigan, Ann Arbor, MI 48109
§Google Brain, Mountain View, CA 94043

ABSTRACT

The problem of detecting whether a test sample is from in-distribution (i.e., training distribution by a classifier) or out-of-distribution sufficiently different from it arises in many real-world machine learning applications. However, the state-of-art deep neural networks are known to be highly overconfident in their predictions, i.e., do not distinguish in- and out-of-distributions. Recently, to handle this issue, several threshold-based detectors have been proposed given pre-trained neural classifiers. However, the performance of prior works highly depends on how to train the classifiers since they only focus on improving inference procedures. In this paper, we develop a novel training method for classifiers so that such inference algorithms can work better. In particular, we suggest two additional terms

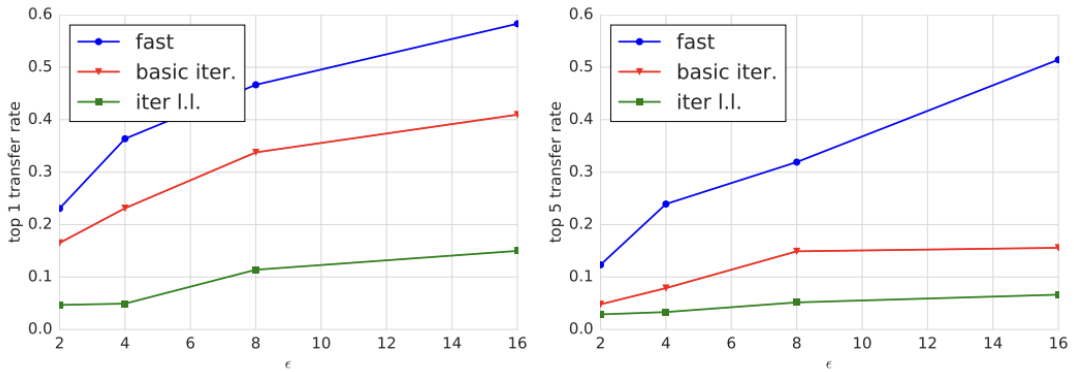
https://arxiv.org/abs/1711.09325

- Kimin Lee, Honglak Lee, Kibok Lee, Jinwoo Shin (ICLR 2018)
- 일반적으로 DNN 모델에 대한 연구는 inference 성능에만 집중하는 경향이 있으나 실세계에서 기계학습모델을 활용하기 위해서는 테스트 샘플이 in-distribution인지 (training distribution by a classifier), out-of-distribution인지를 감지하는 문제가 매우 중요함.
- 본 연구에서는 테스트 샘플이 in-distribution인지 out-of-distribution인지를 이진 분류할 수 있는 비교적 간단하고 실용적인 기법을 제시하고 AlexNet, VGGNet을 사용하는 이미지 분류 실험을 통해 제안된 기법의 효과를 보임
- out-of-distribution인지 여부를 감지하는 모델과 out-of-distribution 데이터를 생성하는 모델을 동시에 학습시킴

Adversarial Machine Learning at Scale

국외

	source model	FGSM target model A	B	C	D	basic iter. target model A	B	C	D	iter l.l. target model A	B	C	D
top 1	A (v3)	100	56	58	47	100	46	45	33	100	13	13	9
	B (v3)	58	100	59	51	41	100	40	30	15	100	13	10
	C (v3 ELU)	56	58	100	52	44	44	100	32	12	11	100	9
	D (v4)	50	54	52	100	35	39	37	100	12	13	13	100
top 5	A (v3)	100	50	50	36	100	15	17	11	100	8	7	5
	B (v3)	51	100	50	37	16	100	14	10	7	100	5	4
	C (v3 ELU)	44	45	100	37	16	18	100	13	6	6	100	4
	D (v4)	42	38	46	100	11	15	15	100	6	6	6	100



Top 1 transferability. Top 5 transferability.

Figure 2: Influence of the size of adversarial perturbation on transfer rate of adversarial examples. Transfer rate was computed using two Inception v3 models with different random intializations. As could be seen from these plots, increase of ϵ leads to increase of transfer rate. It should be noted that transfer rate is a ratio of number of transferred adversarial examples to number of successful adversarial examples for source network. Both numerator and denominator of this ratio are increasing with increase of ϵ, however we observed that numerator (i.e. number of transferred examples) is increasing much faster compared to increase of denominator. For example when ϵ increases from 8 to 16 relative increase of denominator is less than 1% for each of the considered methods, at the same time relative increase of numerator is more than 20%.

- Alexey Kurakin et al., ICLR (2017.02)
- Adversarial training은 Adversarial examples을 명시적으로 사용하여 모델을 공격에 견고하게 만들거나 깨끗한 입력에 대한 테스트 오류를 줄이기 위해 모델을 훈련하는 과정으로 지금까지는 small problem에 활용되었으나 본 연구에서는Adversarial training을 ImageNet 데이터셋에 적용
- Contributions:
  (1) 적대적 훈련을 대규모 모델과 데이터셋에 성공적으로 확장하기 위한 recommendations
  (2) 적대적 훈련이 단계별 공격 방법에 견고성을 제공하는 것을 관찰
  (3) 다단계 공격 방법은 단계별 공격 방법보다 전이성이 약간 떨어지므로 블랙박스 공격에는 단계별 공격이 가장 좋다는 발견
  (4) 적대적으로 훈련된 모델이 깨끗한 예제보다 적대적 예제에서 더 좋은 성능을 내는 "레이블 누출" 효과

https://arxiv.org/abs/1611.01236

Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles

국외

Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles

Balaji Lakshminarayanan **Alexander Pritzel** **Charles Blundell**
DeepMind
{balajiln,apritzel,cblundell}@google.com

Abstract

Deep neural networks (NNs) are powerful black box predictors that have recently achieved impressive performance on a wide spectrum of tasks. Quantifying predictive uncertainty in NNs is a challenging and yet unsolved problem. Bayesian NNs, which learn a distribution over weights, are currently the state-of-the-art for estimating predictive uncertainty; however these require significant modifications to the training procedure and are computationally expensive compared to standard (non-Bayesian) NNs. We propose an alternative to Bayesian NNs that is simple to implement, readily parallelizable, requires very little hyperparameter tuning, and yields high quality predictive uncertainty estimates. Through a series of experiments on classification and regression benchmarks, we demonstrate that our method produces well-calibrated uncertainty estimates which are as good or better than approximate Bayesian NNs. To assess robustness to dataset shift, we evaluate the predictive uncertainty on test examples from known and unknown distributions, and show that our method is able to express higher uncertainty on out-of-distribution examples. We demonstrate the scalability of our method by evaluating predictive uncertainty estimates on ImageNet.

- Balaji Lakshminarayanan et al. (2017)
- 기존의 uncertainty 를 구하는 방법들은 대부분 베이지안과 관련이 되어 있었는데, 이로 인한 몇몇 문제점들을 지적하고 새로운 방법을 제시함.
- 베이지안 방법들은 훈련과정에서 많은 수정이 필요하여 computationally expensive 하며 여러 번 sampling 하는 과정이 필요함. 그래서 논문에서는 하이퍼 파라미터 튜닝이 별로 필요하지 않고 병렬적으로 실행할 수 있으면서도 uncertainty 추정을 잘하는 non-bayesian방법을 제시함.
- 베이지안 방법들은 여러 결과를 함께 고려한다는 점이 Ensemble과 비슷하게 보임. 하지만 Bayesian model averaging(soft model selection)의 경우는 posterior를 잘 근사하였다고 가정하기 때문에 그렇지 않을 경우 결과가 좋지 못함. 반면 Ensemble(model combination)은 posterior를 잘 근사하지 못해도 좋은 결과를 내며 더욱 robust 한 모델을 얻을 수 있어 논문에서는 이 방법을 사용함.

https://proceedings.neurips.cc/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf

Enhancing the Reliability of Out-of-distribution Image Detection in Neural Networks

국외

Published as a conference paper at ICLR 2018

Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks

Shiyu Liang
Coordinated Science Lab, Department of ECE
University of Illinois at Urbana-Champaign
sliang26@illinois.edu

Yixuan Li
University of Wisconsin-Madison*
sharonli@cs.wisc.edu

R. Srikant
Coordinated Science Lab, Department of ECE
University of Illinois at Urbana-Champaign
rsrikant@illinois.edu

Abstract

We consider the problem of detecting *out-of-distribution* images in neural networks. We propose *ODIN*, a simple and effective method that does not require any change to a pre-trained neural network. Our method is based on the observation that using temperature scaling and adding small perturbations to the input can separate the softmax score distributions between in- and out-of-distribution images, allowing for more effective detection. We show in a series of experiments that *ODIN*

https://arxiv.org/pdf/1706.02690.pdf

- Shiyu Liang et al. (2020)
- Out of distribution samples 를 detect하기 위한 새로운 방법을 제시함.
- 첫째로, 새로운 이미지가 classifier 로 들어올 때, image 를 highest softmax output의 방향으로 push하기 위해 작은 perturbation을 더하고, 해당 softmax에 temperature를 더함.
- 위의 output에 대해 주어진 threshold 만족 유무에 따라 in-distribution 과 out-distribution을 결정.
- CIFAR-10/100 과 몇몇의 out of distribution 데이터셋에 대해 bassline을 significant margins 으로 outperform 함을 보여줌.

Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer

국외

Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer

David Madras, Toniann Pitassi and Richard Zemel

University of Toronto, Vector Institute

Abstract

In many machine learning applications, there are multiple decision-makers involved, both automated and human. The interaction between these agents often goes unaddressed in algorithmic development. In this work, we explore a simple version of this interaction with a two-stage framework containing an automated model and an external decision-maker. The model can choose to say PASS, and pass the decision downstream, as explored in rejection learning. We extend this concept by proposing *learning to defer*, which generalizes rejection learning by considering the effect of other agents in the decision-making process. We propose

https://arxiv.org/abs/1711.06664

- David Madras et al. (2018)
- 기계학습 모델이 활용되는 실제 응용에서는 예측결과를 출력하는 모델과 의사결정권자인 사람이 interaction을 하면서 task를 수행하는 환경임
- Decision Maker (이하 DM으로 표기)의 타입 (High-Accuracy DM, Highly-Biased DM, Inconsistent DM 등)에 맞추어 PASS decisions을 내리는 'Adaptive Reject Learning (Learning to Defer)' 기법을 제안하여accuracy와 fairness를 균형적으로 모두 개선하였음.

※ Learning to Reject (Reject Learning): **machine learning with a reject (PASS) option. (i.e.** Enables machine learning models to abstain from making a prediction when likely to make a mistake)

A Simple Baseline for Bayesian Uncertainty in Deep Learning

국외

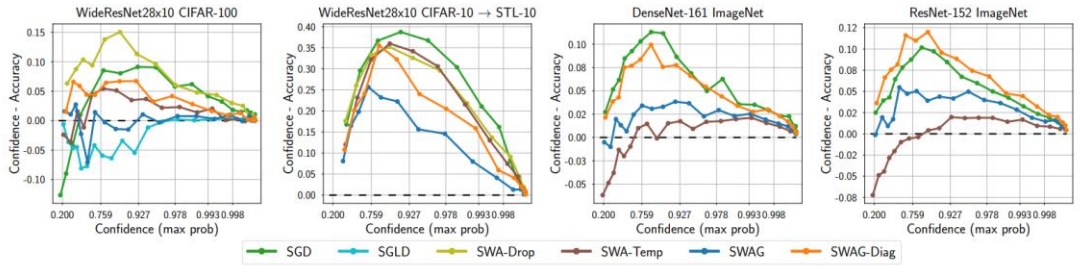


Figure 3: Reliability diagrams for WideResNet28x10 on CIFAR-100 and transfer task; ResNet-152 and DenseNet-161 on ImageNet. Confidence is the value of the max softmax output. A perfectly calibrated network has no difference between confidence and accuracy, represented by a dashed black line. Points below this line correspond to under-confident predictions, whereas points above the line are overconfident predictions. SWAG is able to substantially improve calibration over standard training (SGD), as well as SWA. Additionally, SWAG significantly outperforms temperature scaling for transfer learning (CIFAR-10 to STL), where the target data are not from the same distribution as the training data.

- Wesley J. Maddox et al. (2019)
- There has been a great desire to apply Bayesian methods in deep learning due to their theoretical properties and past success with small neural networks.
- Developed Stochastic Weight Averaging-Gaussian (SWAG) for approximate Bayesian inference in deep learning.
- SWAG: Computes the first moment of stochastic gradient descent (SGD) and iterates with a modified learning rate schedule.
- SWAG는 기존에 널리 알려진 방법론인 MC dropout, KFAC Laplace, SGLD, temperature scaling 등에 비해 개선된 성능을 보임 (sample detection, calibration, and transfer learning에서)

https://arxiv.org/abs/1902.02476

EMPOWERING EXPERTISE
iNEEJI
KAIST

Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift

국외

Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift

Yaniv Ovadia[*]
Google Research
yovadia@google.com

Emily Fertig[*†]
Google Research
emilyaf@google.com

Jie Ren[†]
Google Research
jjren@google.com

Zachary Nado
Google Research
znado@google.com

D Sculley
Google Research
dsculley@google.com

Sebastian Nowozin
Google Research
nowozin@google.com

Joshua V. Dillon
Google Research
jvdillon@google.com

Balaji Lakshminarayanan[‡]
DeepMind
balajiln@google.com

Jasper Snoek[‡]
Google Research
jsnoek@google.com

Abstract

Modern machine learning methods including deep learning have achieved great success in predictive accuracy for supervised learning tasks, but may still fall short in giving useful estimates of their predictive *uncertainty*. Quantifying uncertainty

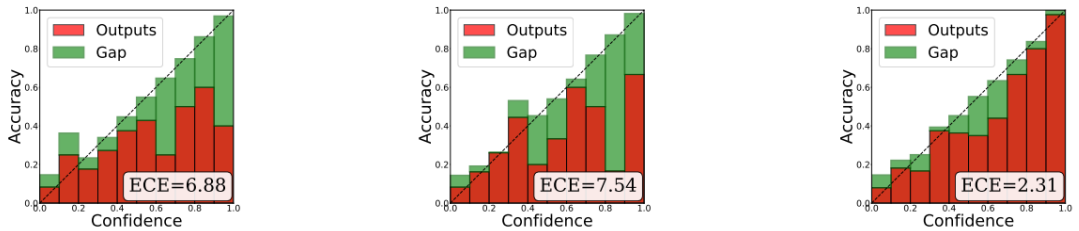
https://arxiv.org/abs/1906.02530

- Yaniv Ovadia et al. (2019)
- Deep learning have achieved great success in predictive accuracy for supervised learning tasks, but may still fall short in giving useful estimates of their predictive uncertainty.
- predictive uncertainty(예측 불확실성)을 정량화하기 위한 Bayesian-, non-Bayesian 기법들이 제안되었으나 data shift 상황을 고려한 대규모 비교실험은 없었음.
- A large-scale benchmark of existing state-of-the-art methods on classification problems and investigate the effect of dataset shift on accuracy and calibration.

※ Data shift: input distributions that are shifted from the training distribution due to a variety of factors including sample bias and non-stationarity

Uncertainty-Aware Attention for Reliable Interpretation and Prediction

국내



(a) Deterministic Attention [3] (b) Stochastic Attention [28] (c) Uncertainty-aware Attention (Ours)

Figure 1: Reliability diagrams [9] which shows the accuracy as a function of model confidence, generated from RNNs trained for mortality risk analysis from ICU records (PhysioNet-Mortality). ECE [22] in (8) denotes Expected Calibration Error, which is the weighted-average gap between model confidence and actual accuracy. (Gap is shown in green bars.) Conventional attention models result in poorly calibrated networks while our UA yields a well-calibrated one. Such accurately calibrated networks allow us to perform reliable prediction by leveraging prediction confidence to decide whether to predict or defer prediction.

- The attention mechanism is effective for focusing deep learning models on relevant feature and interpreting them. However, it can become unreliable when generated by weakly-supervised trained networks.
- The Uncertainty-aware Attention (UA) mechanism generates attention with vary degrees of noise for each feature based on the input, learning larger variance for uncertainty instances.
- Evaluation of the uncertainty calibration and prediction performance with an "I don't know" decision shows that UA produces networks with high reliability.

https://arxiv.org/pdf/1711.11279.pdf

AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty

국외

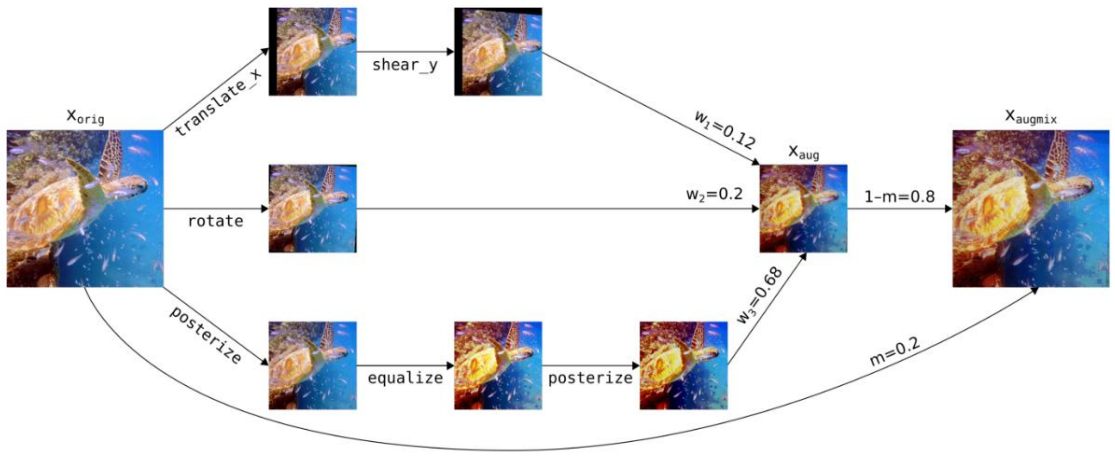


Figure 4: A realization of AUGMIX. Augmentation operations such as translate_x and weights such as m are randomly sampled. Randomly sampled operations and their compositions allow us to explore the semantically equivalent input space around an image. Mixing these images together produces a new image without veering too far from the original.

- Dan Hendrycks et al. (ICLR 2020)
- Currently there are few techniques that improve robustness to unforeseen data shifts encountered during deployment.
- Proposes AugMix, which utilizes stochasticity and diverse augmentations, a Jensen-Shannon Divergence consistency loss, and a formulation to mix multiple augmented images to achieve state-of-the-art performance.
- On ImageNet, AUGMIX also achieves state-of-the-art corruption robustness and decreases perturbation instability from 57.2% to 37.4%

https://arxiv.org/abs/1912.02781

iNEEJI EMPOWERING EXPERTISE
KAIST

Can Autonomous Vehicles Identify, Recover From, and Adapt to Distribution Shifts?

국외

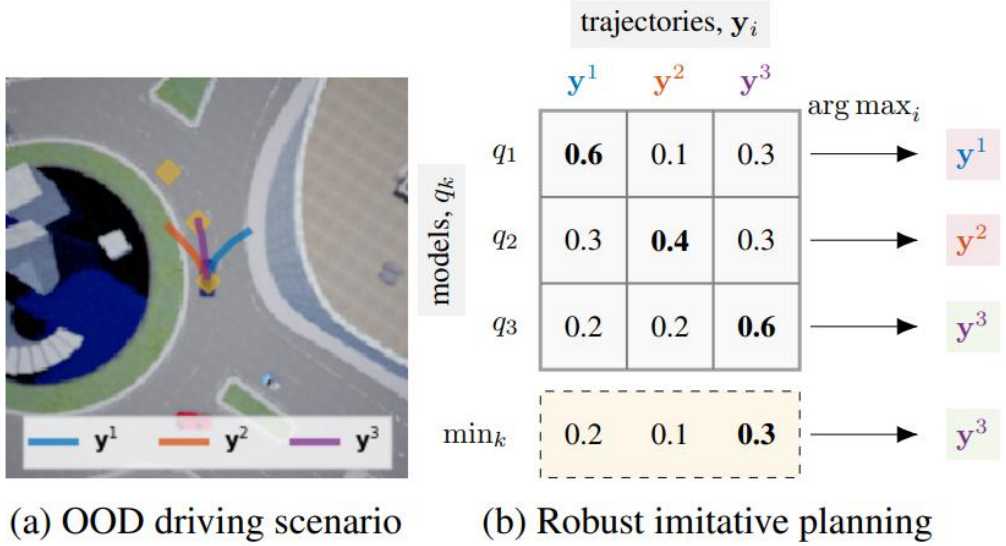


Figure 1. Didactic example: (a) in a novel, out-of-training distribution (OOD) driving scenario, candidate plans/trajectories $\mathbf{y}^1, \mathbf{y}^2, \mathbf{y}^3$ are (b) evaluated (row-wise) by an ensemble of expert-likelihood models q_1, q_2, q_3. Under models q_1 and q_2 the best plans are the catastrophic trajectories $\mathbf{y}^1$ and $\mathbf{y}^2$ respectively. Our epistemic uncertainty-aware robust (RIP) planning method aggregates the evaluations of the ensemble and proposes the safe plan $\mathbf{y}^3$. RIP considers the disagreement between the models and avoid overconfident but catastrophic extrapolations in OOD tasks.

- Angelos Filos et al. (2020)
- Out-of-training-distribution (OOD) scenarios are a common challenge of learning agents at deployment, typically leading to arbitrary deductions and poorly-informed decisions.
- Highlights the limitations of current approaches to novel driving scenes and proposes an epistemic uncertainty-aware planning method, called Robust Imitative Planning (RIP).
- Introduces an autonomous car novel-scene benchmark, CARNOVEL, to evaluate the robustness of driving agents to a suite of tasks with distribution shifts, where the proposed methods outperform all the baselines.

https://arxiv.org/abs/2006.14911

EMPOWERING EXPERTISE
iNEEJi
KAIST

It is Not “Accuracy vs. Explainability”—We Need Both for Trustworthy AI Systems

국외

It is Not “Accuracy vs. Explainability”—We Need Both for Trustworthy AI Systems

Publisher: IEEE Cite This PDF

Dragutin Petkovic **All Authors**

167
Full
Text Views

Abstract

Document Sections

I. Introduction

II. Brief Summary of XAI Advances Today

Abstract:
We are witnessing the emergence of an “AI economy and society” where AI technologies and applications are increasingly impacting health care, business, transportation, defense and many aspects of everyday life. Many successes have been reported where AI systems even surpassed the accuracy of human experts. However, AI systems may produce errors, can exhibit bias, may be sensitive to noise in the data, and often lack technical and judicial transparency resulting in reduction in trust and challenges to their adoption. These recent shortcomings and concerns have been documented in both the scientific and general press

https://ieeexplore.ieee.org/document/10029927
https://arxiv.org/ftp/arxiv/papers/2212/2212.11136.pdf

- Dragutin Petkovic (IEEE Transactions on Technology and Society, vol.4 Issue.1, 2023)
- Challenges the recent arguments that present “accuracy vs. explainability” as being mutually exclusive and for focusing mainly on deep learning with its limited XAI capabilities.
- Claims that XAI analysis can play a critical role in delivery of trustworthy and cost effective AI systems of any type (deep learning or feature based) in all of their lifecycle components:
- a) development including internal testing
- b) external verification, audit, certification
- c) trustworthy production and maintenance

EMPOWERING EXPERTISE
iNEEJi
KAIST

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

jaesik.choi@kaist.ac.kr

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

Do you wish to Change the world with

World-leading XAI Technology and Experience!

INEEJI EMPOWERING EXPERTISE

Copyright©2019 by INEEJI.

Appendix

Agenda

- Motivation
- Social Impact and Regulations
- XAI Research Trends
- **Appendix: What is inside of Deep Neural Networks**

Q1: Can we find the role of internal units in deep neural networks?

*Equal contribution

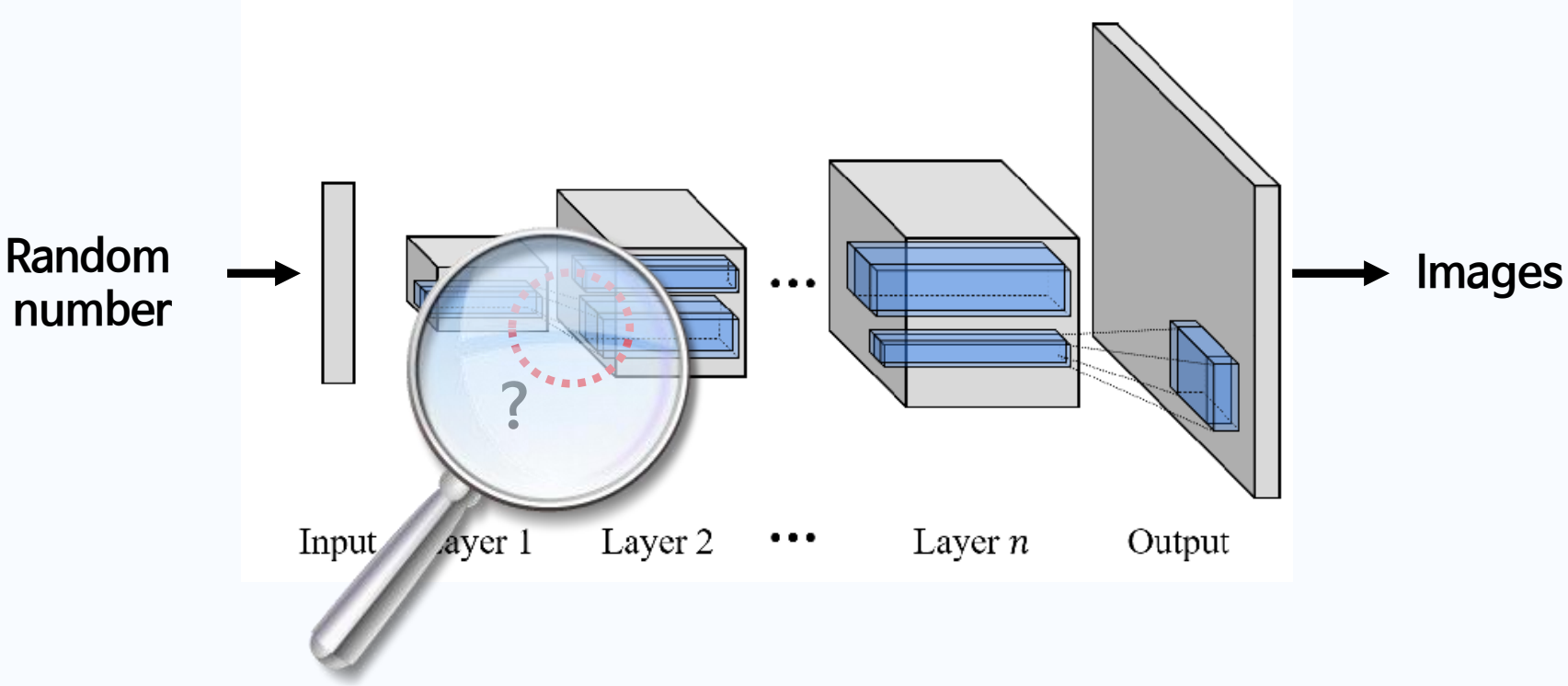
Explorative Generative Boundary Aware Sampling

[G. Jeon*, H. Jeong* and J. Choi, An Efficient Explorative Sampling Considering the Generative Boundaries of Deep Generative Neural Networks, AAAI-20]

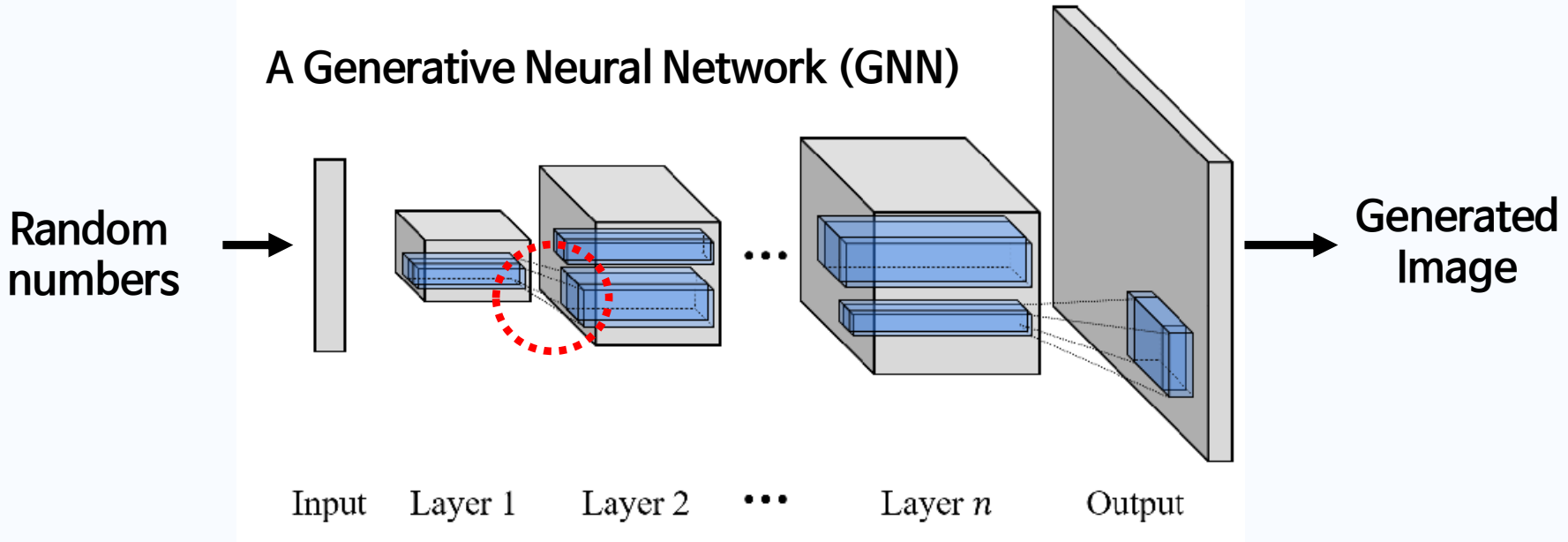
*Equal contribution

Analyzing Inside of Deep Neural Networks

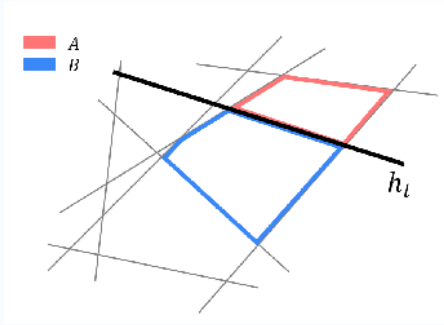
A Generative Neural Network (GNN)



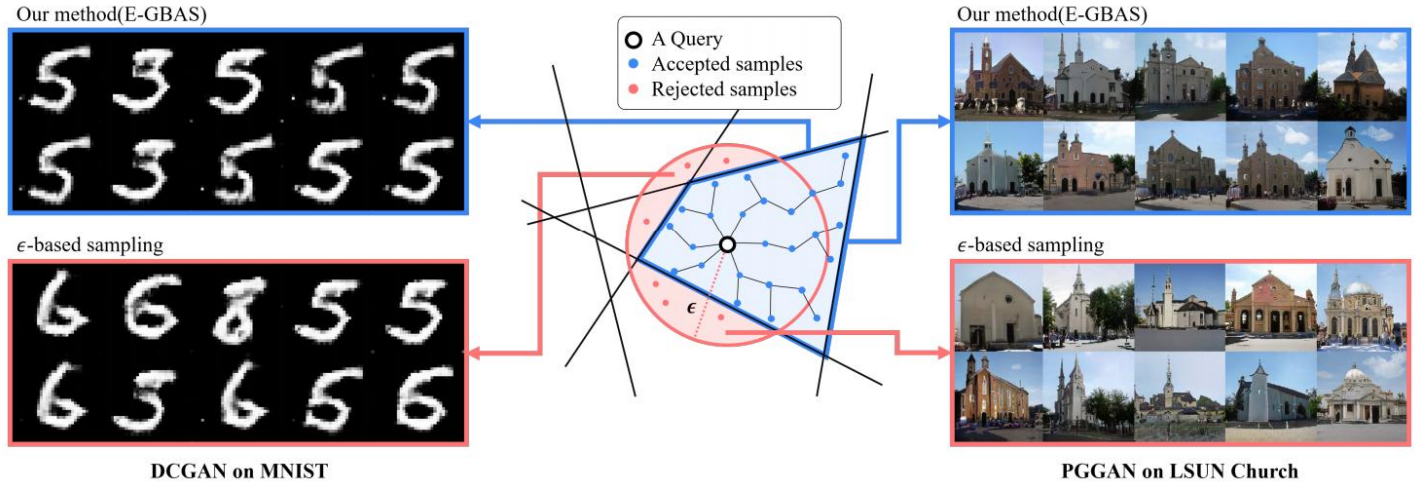
Analyzing Inside of Deep Neural Networks



Generative Boundary Aware Sampling

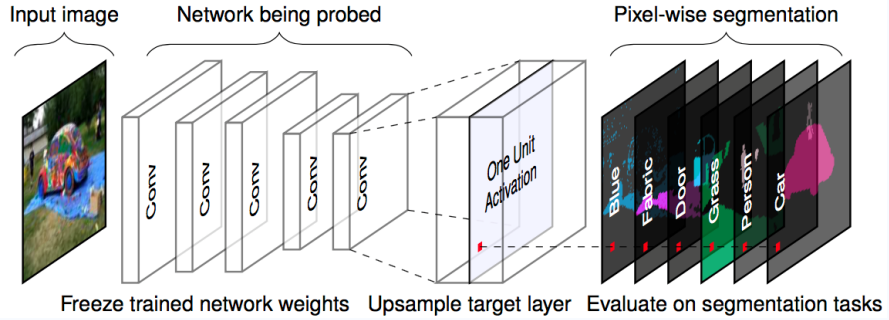


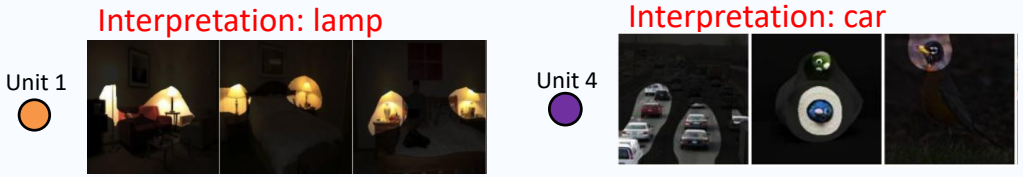
Generative boundaries



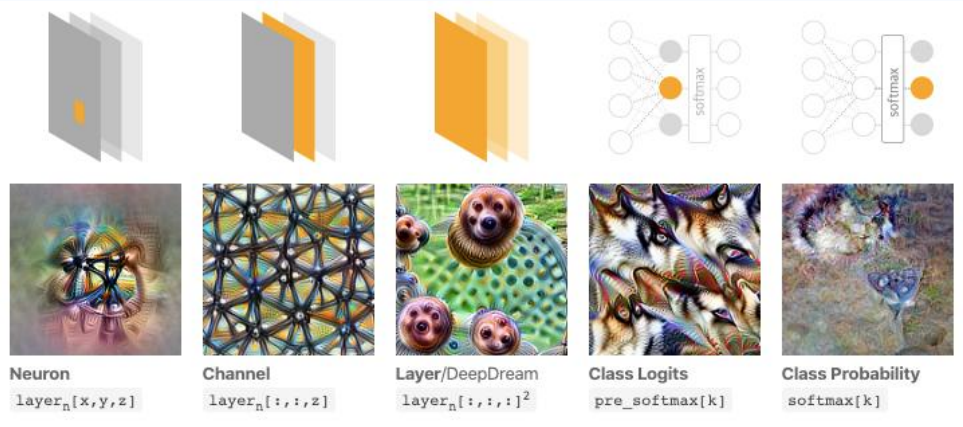
Related Work to Analyzing Inside of Deep Neural Networks

Network dissection (Bau et al., 2017)

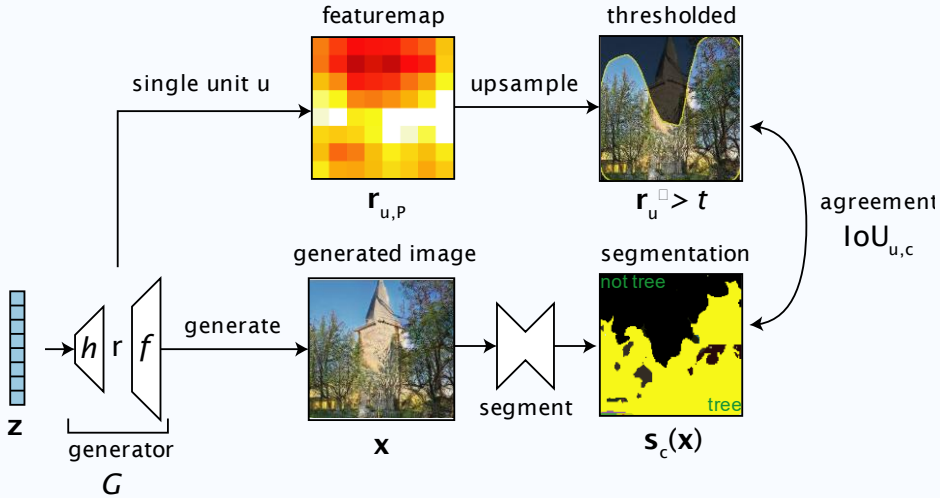




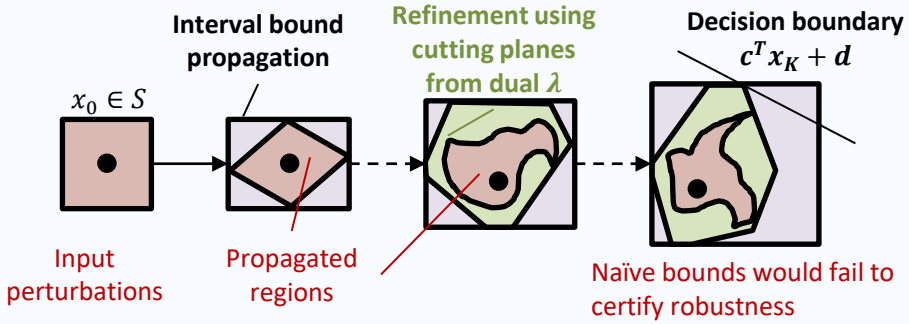
Google Deep Dream (Mordvintsev et al., 2015)



GAN dissection (Bau et al., 2019)



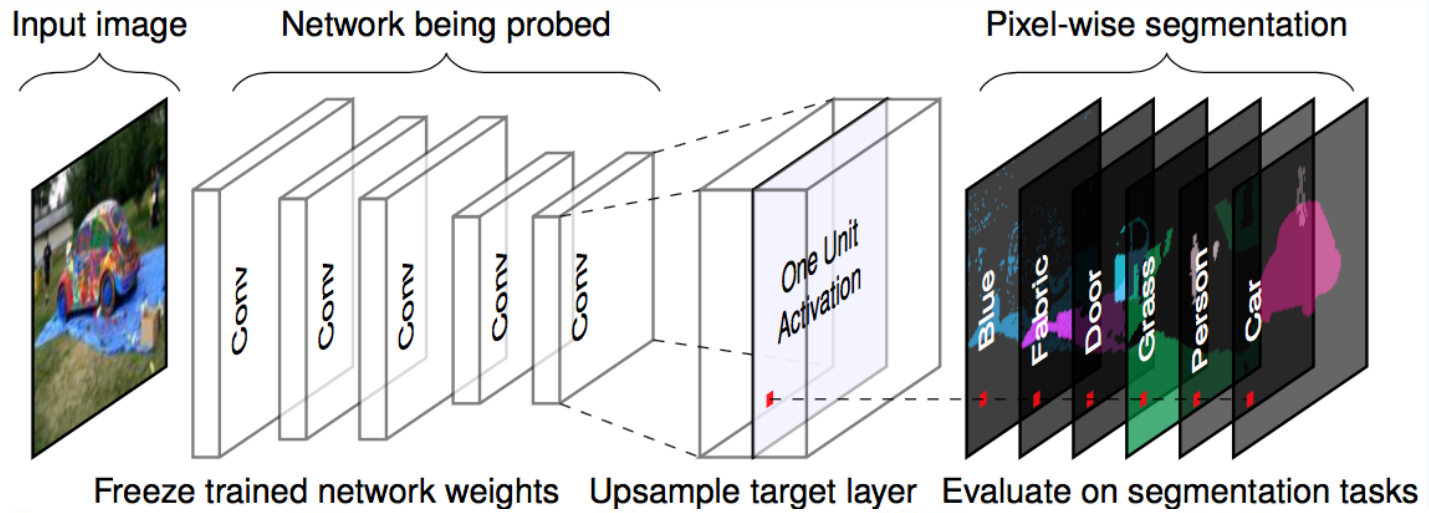
Lagrangian relaxed decision boundary (Dvijotham et al., 2018)

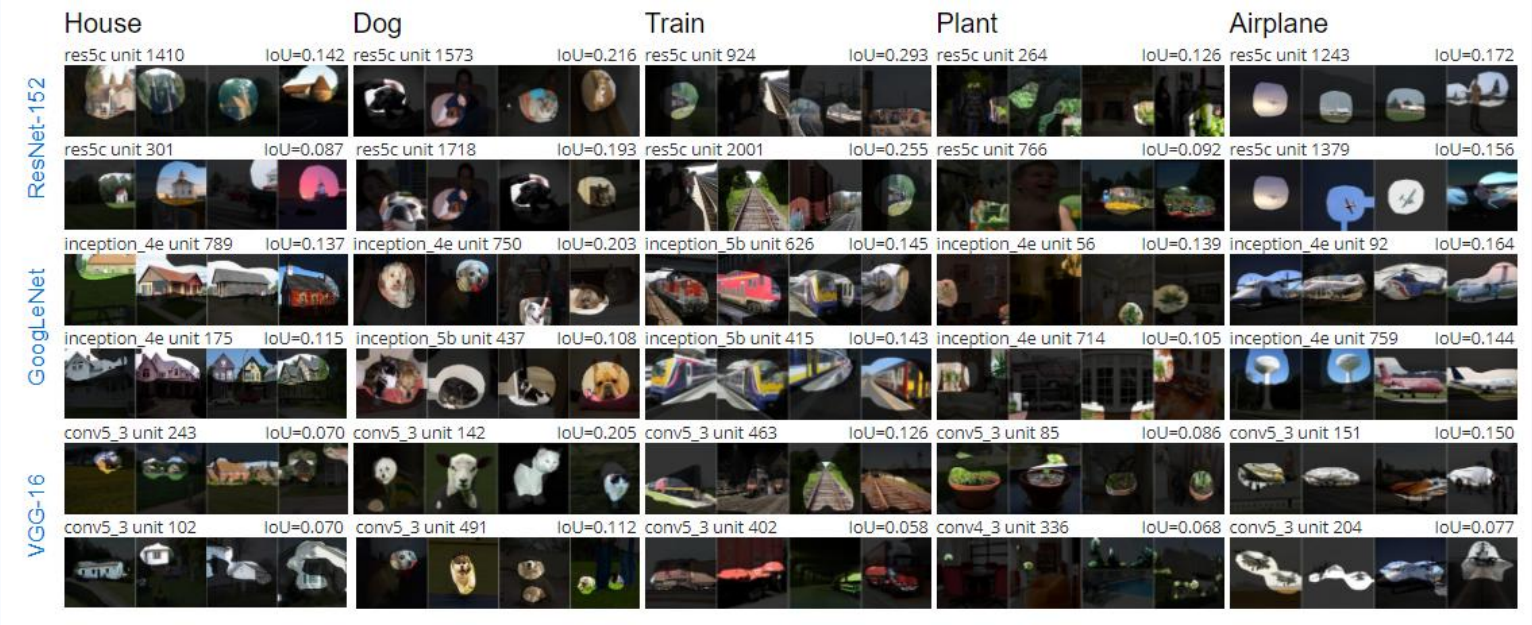


D. Bau et al., "Network Dissection: Quantifying Interpretability of Deep Visual Representations." CVPR, 2017.
D. Bau et al., "GAN Dissection: Visualizing and Understanding Generative Adversarial Networks", *ICLR*, 2019.
A. Mordvintsev et al., "Inceptionism: Going deeper into neural networks", 2015.
K. Dvijotham et al., "A Dual Approach to Scalable Verification of Deep Networks." *UAI*, 2018.

Network Dissection

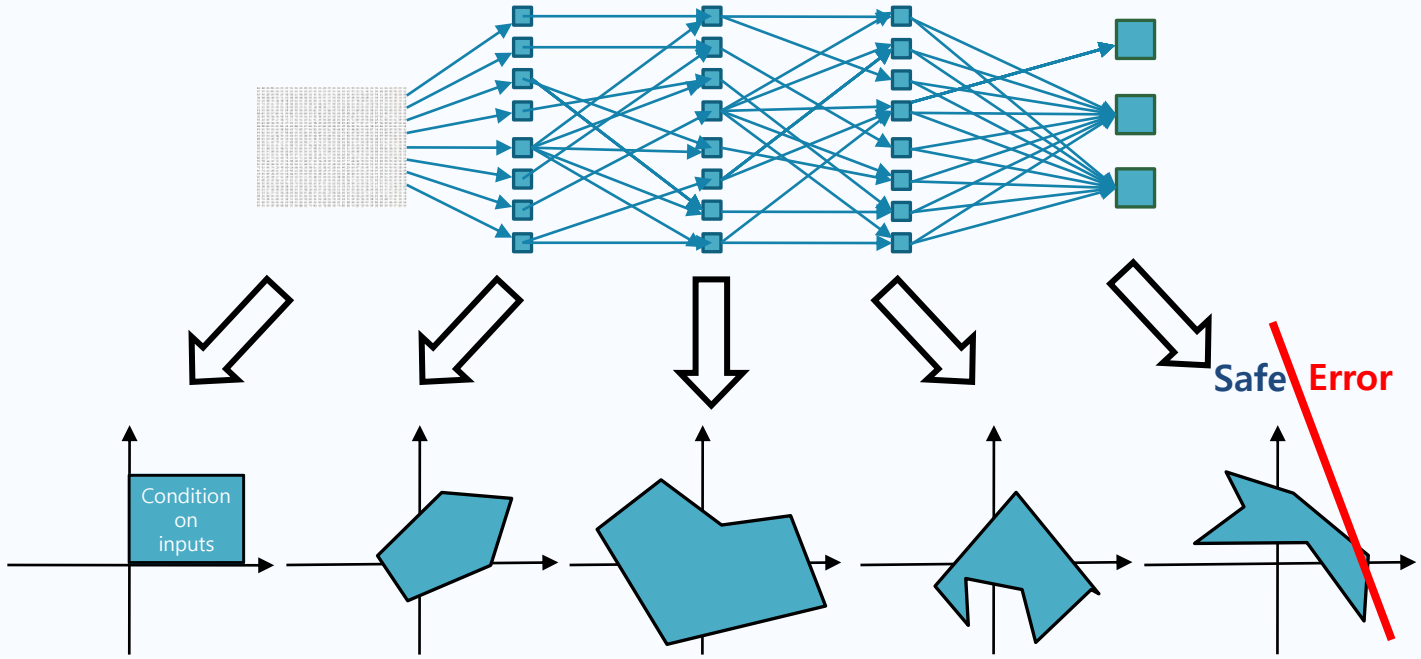
David Bau et. al., 2017 (A. Torralba, MIT)





Neural Networks Verification – Robust Deep Learning

Is there an erroneous output?



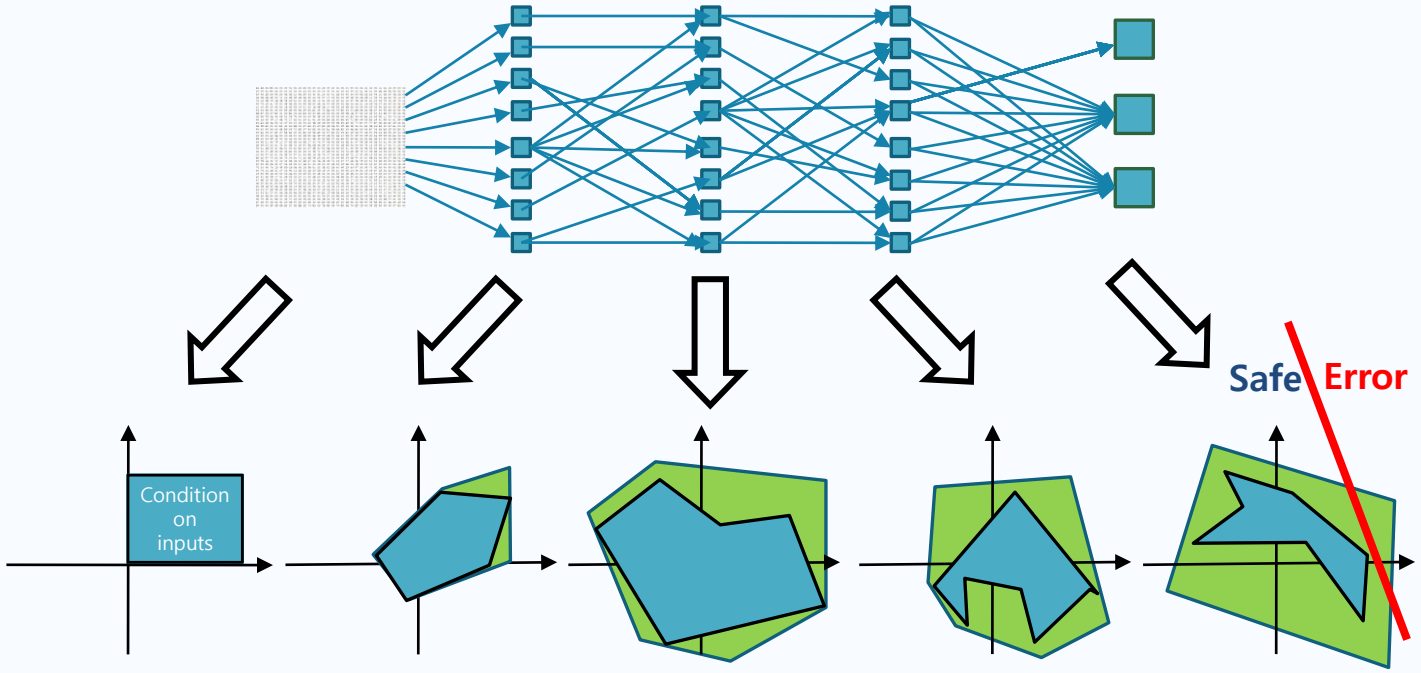
Non-convexity makes the problem NP-hard

[Slide courtesy of M. Pawan Kumar]

EMPOWERING EXPERTISE
iNEEJI
KAIST

Neural Networks Verification – Robust Deep Learning

Is there an erroneous output?



Convex superset might give incorrect answer

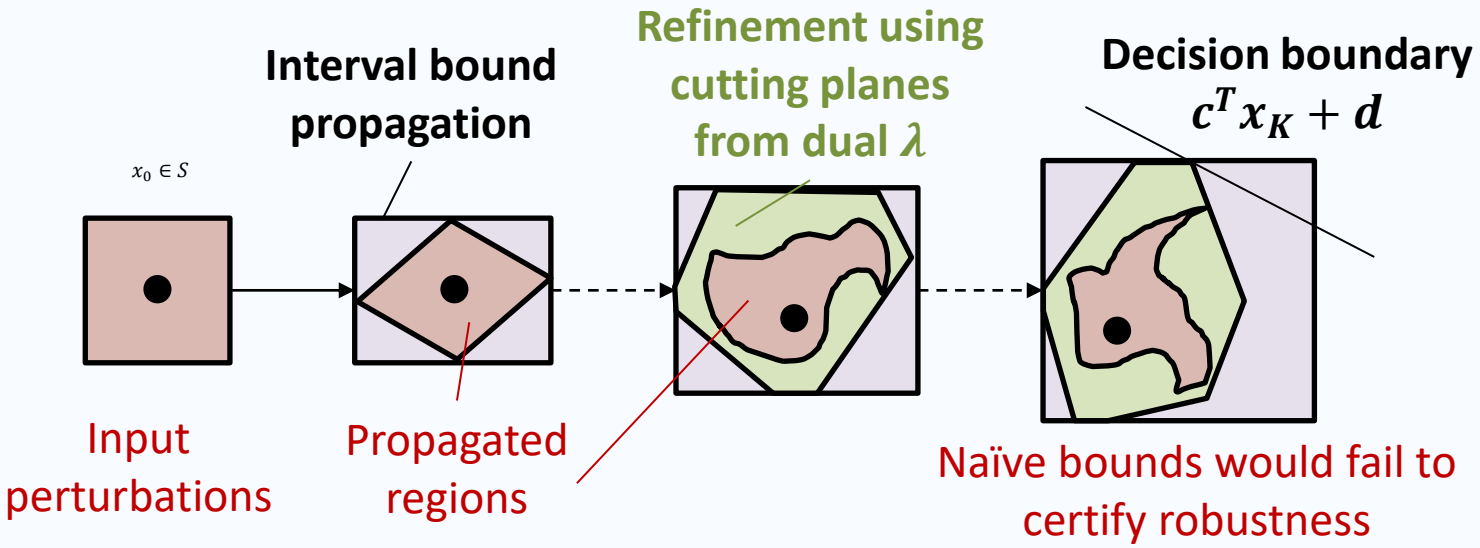
[Slide courtesy of M. Pawan Kumar]

EMPOWERING EXPERTISE
iNEEJi
KAIST

Neural Networks Verification – An Incomplete Method

[Dvijotham et al., 2018]

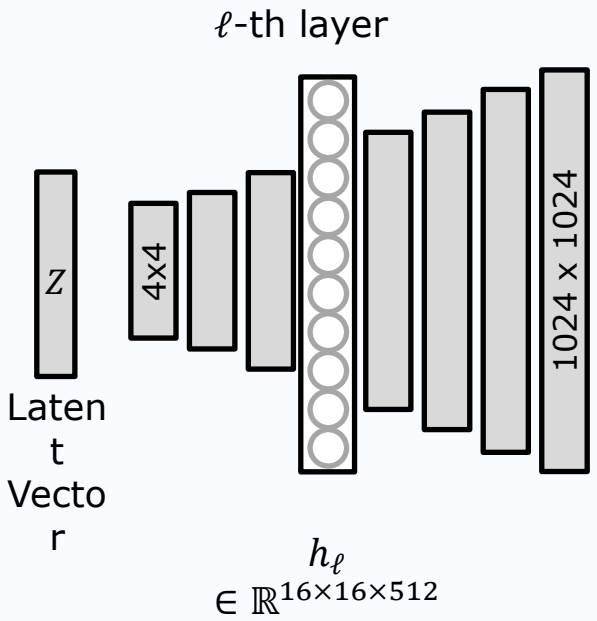
Finding a tight convex bound with Lagrangian relaxed decision boundary



Analyzing Inside of Deep Neural Networks – E-GBAS

Motivation

The generative process is not well understood yet.
We wish to give example-based explanation on the generative process.



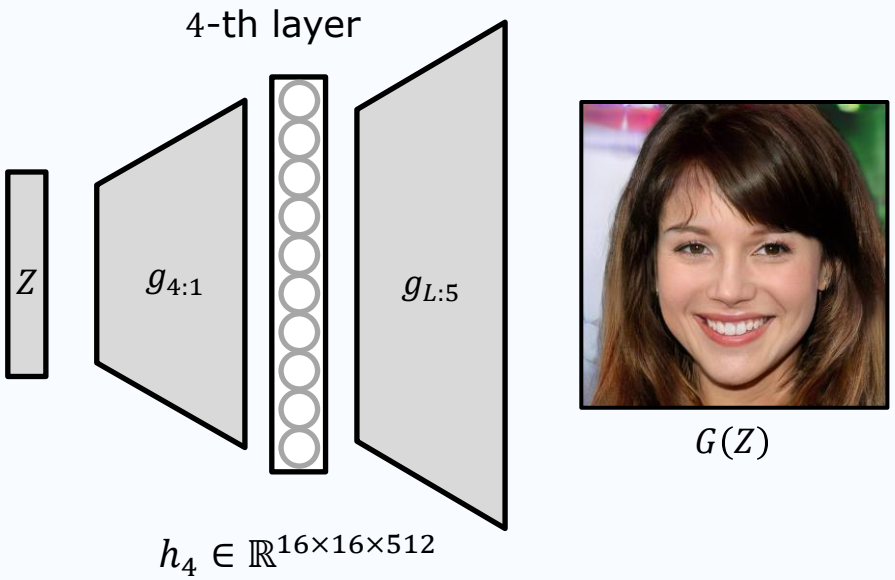
LSUN dataset

CelebA dataset

Analyzing Inside of Deep Neural Networks – E-GBAS

Definitions

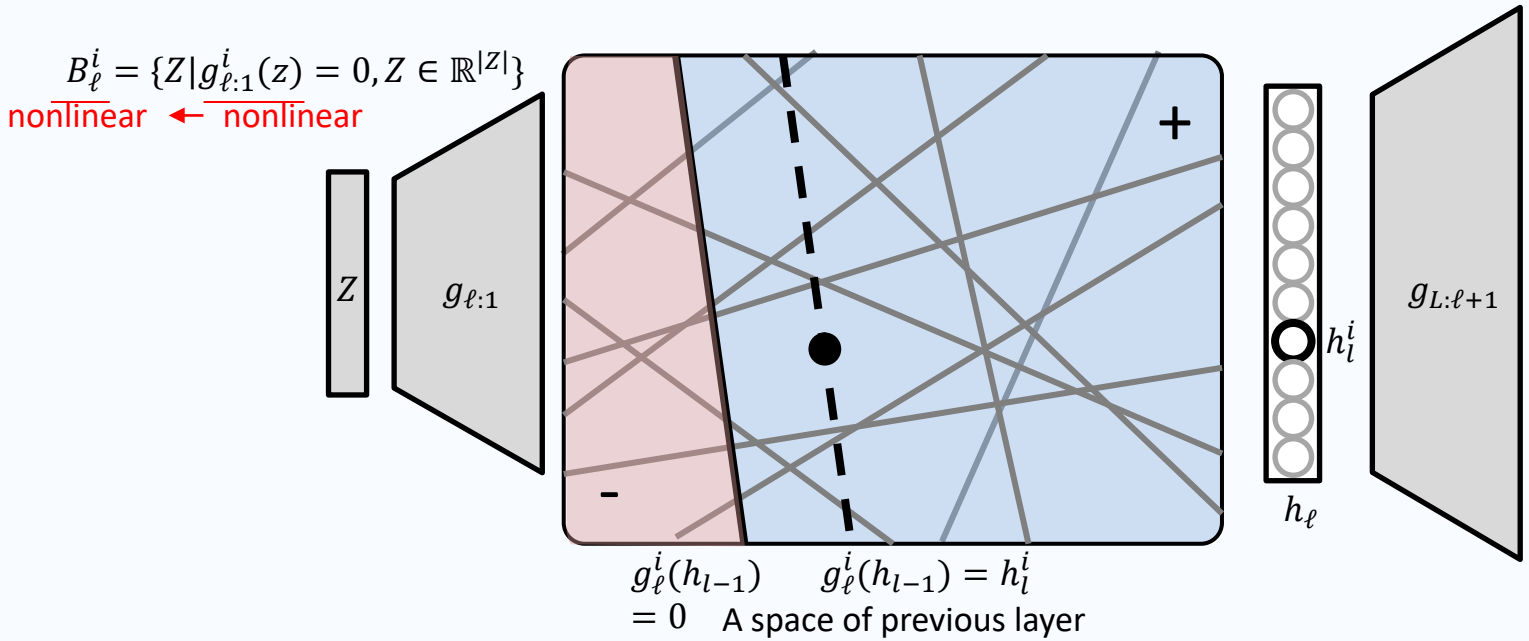
Generator $G(Z)$: a generated image from Z
Hidden nodes h_ℓ: a neural representation of ℓ−th layer
Partial generation $g_{j:i}: \mathbb{R}^{|h_i|} \rightarrow \mathbb{R}^{|h_j|}$: a generative function from layer i to layer j



EMPOWERING EXPERTISE
iNEEJI
KAIST

Generative Boundary

A value of h_ℓ is determined by the linear hyperplane in the space of the previous layer, $h_{\ell-1}$
Stacking of layers toward input makes highly non-linear and non-convex shape
We want to see only feasible regions which constructed from the input to the target.
Trained to fool the discriminator in GANs



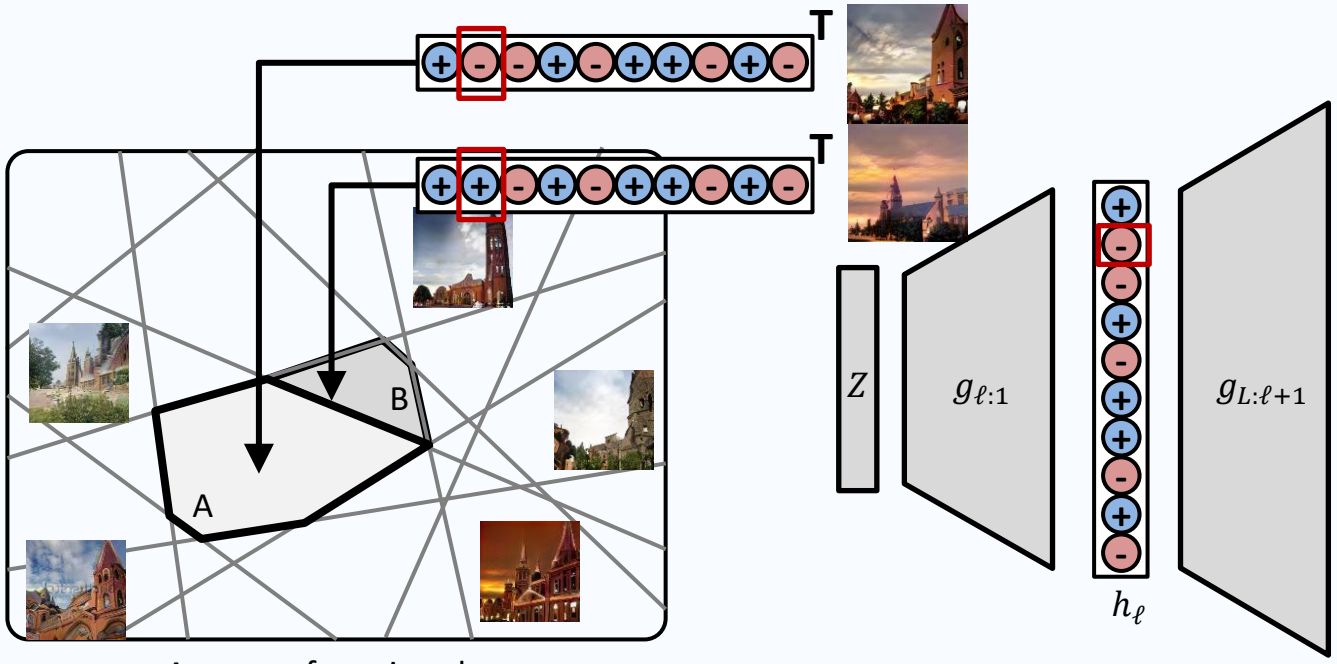
Our approach: Go Back to the Basic of Generative Boundaries!

Generative Region

In the ℓ-th layer, a space (S_ℓ) which is surrounded by a set of generative boundaries.

In the input space, a set of equivalent class of Z w.r.t S_ℓ.

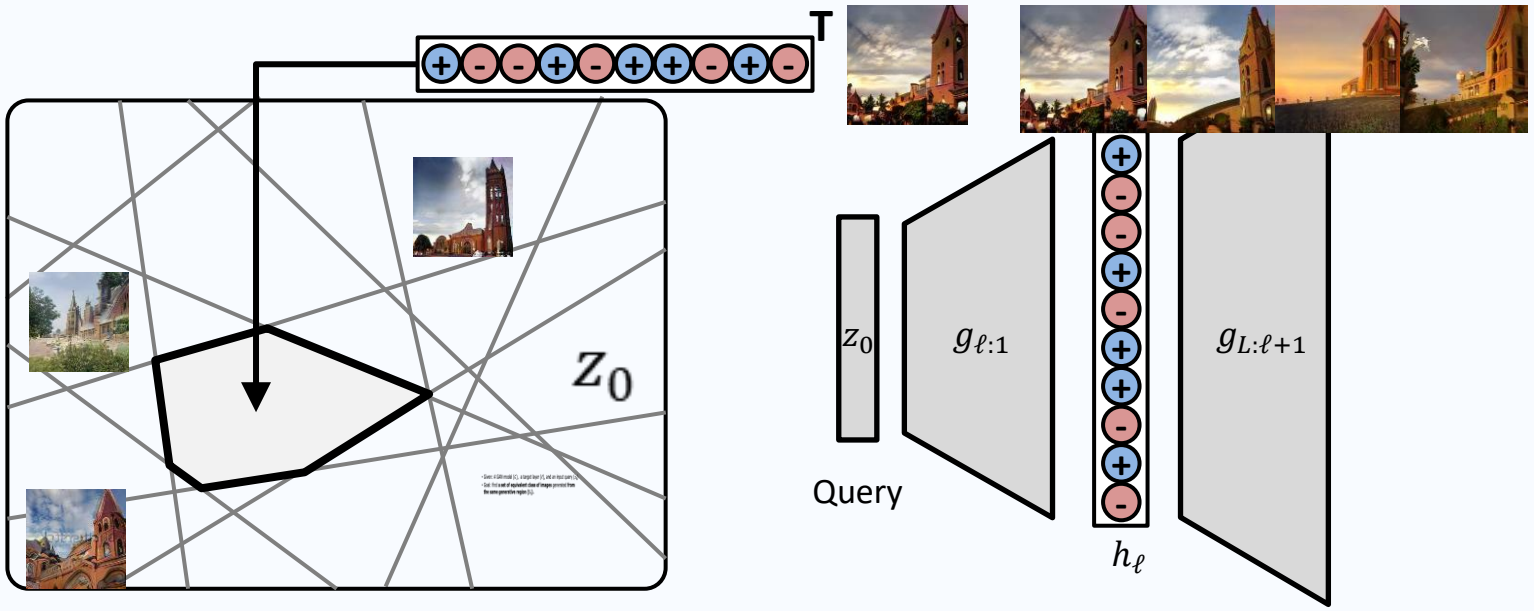
In the image space, a set of equivalent class of image w.r.t. S_ℓ.



EMPOWERING EXPERTISE
iNEEJI
KAIST

Problem Definition: Explorative sampling in a generative region

Given: A GAN model (G), a target layer (ℓ), and an input query (z_0)
Goal: find a set of equivalent class of images generated from the same generative region (S_ℓ).

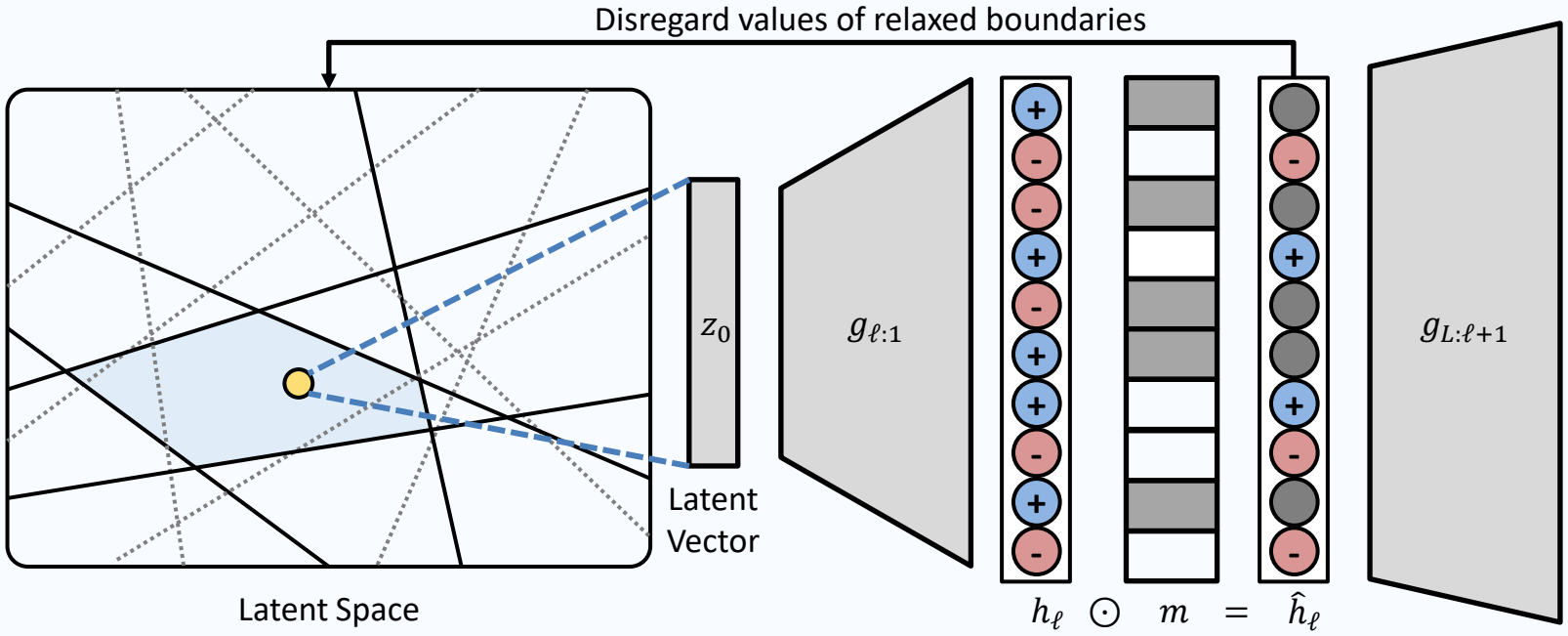


A space of previous layer

EMPOWERING EXPERTISE
iNEEJI
KAIST

Challenges of Handling too many Generative Boundaries

Using all the boundaries, constraints get **too tight** and **computationally expensive**.
We observe not all the boundaries affects equally on the output.



EMPOWERING EXPERTISE
iNEEJI
KAIST

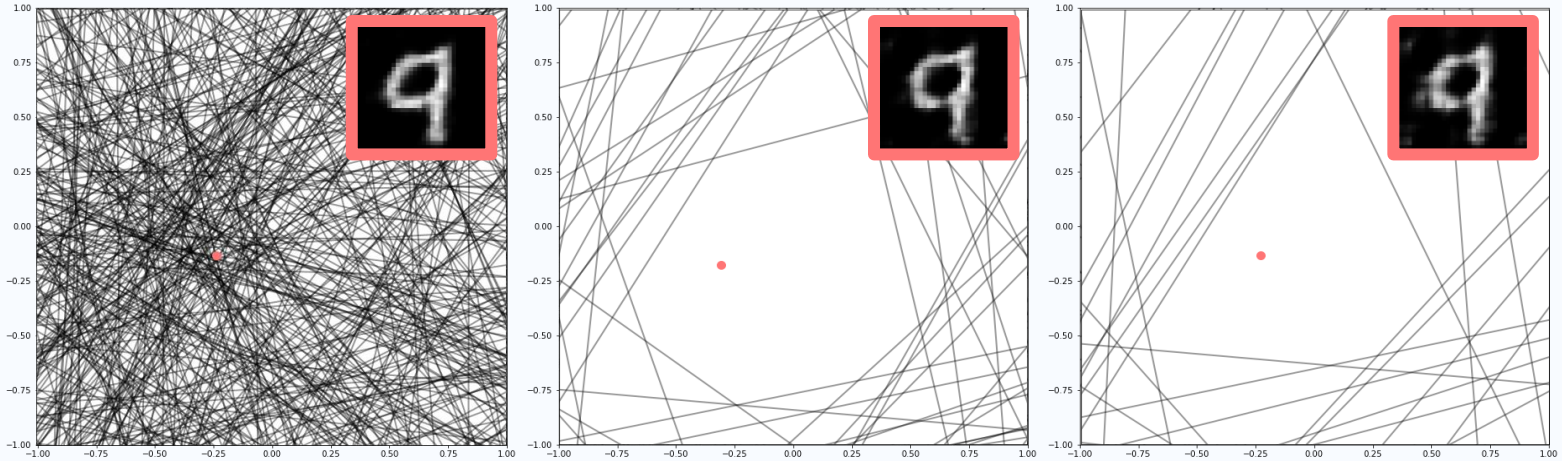
Analyzing Inside of Deep Neural Networks – E-GBAS

Our Solution: Smallest Supporting Generative Boundary Set

Apply **Bernoulli mask optimization** to relax boundaries but maintain the output.

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}(z_0, \ell, \theta)$$
$$= \underset{\theta}{\operatorname{argmin}} \|g_{L:\ell+1}(g_{l:1}(z_0) \odot m) - G(z_0)\| + \lambda\|\theta\|_1 \qquad \text{where } m \sim Ber(\theta)$$

Masked image reconstruction error Mask l1 regularizer

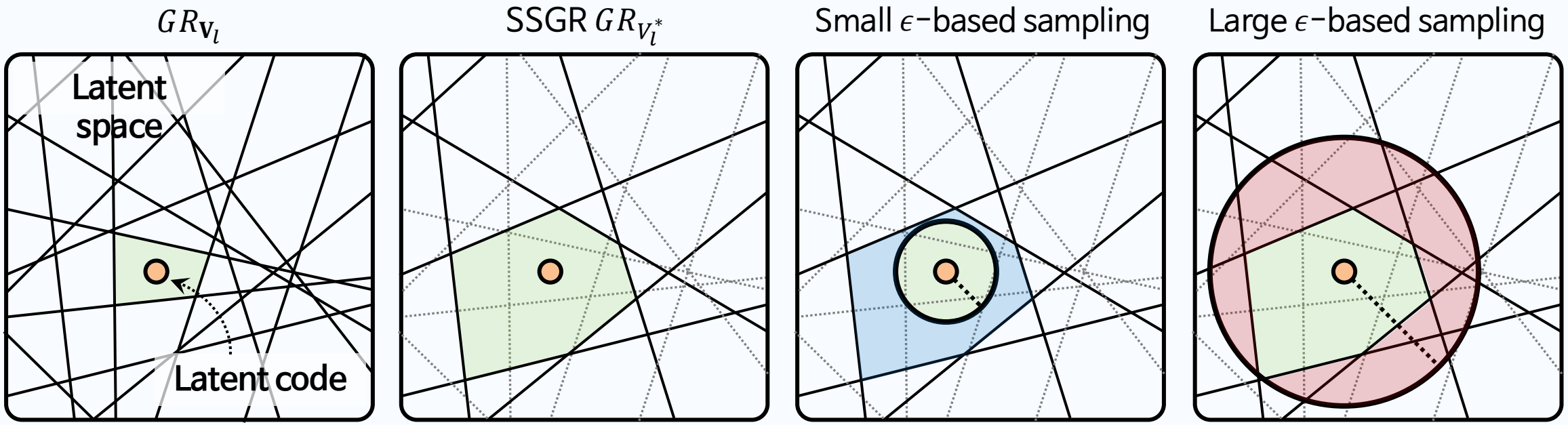


Entire boundaries Using 10% Using 5%

Chang, Chun-Hao, et al. "Explaining image classifiers by adaptive dropout and generative in-filling." *International Conference on Learning Representations (ICLR)*. 2018.

Challenges of Sampling in a Generative Region

- The **dimension** of latent space and **a lot of hyperplanes** are hard to handle in practice. (E.g., 4th layer in PGGAN: $\mathbb{R}^{512} \rightarrow \mathbb{R}^{8192}$)
- Typically, generative region is **nonconvex** in higher layer due to nonlinear activations.

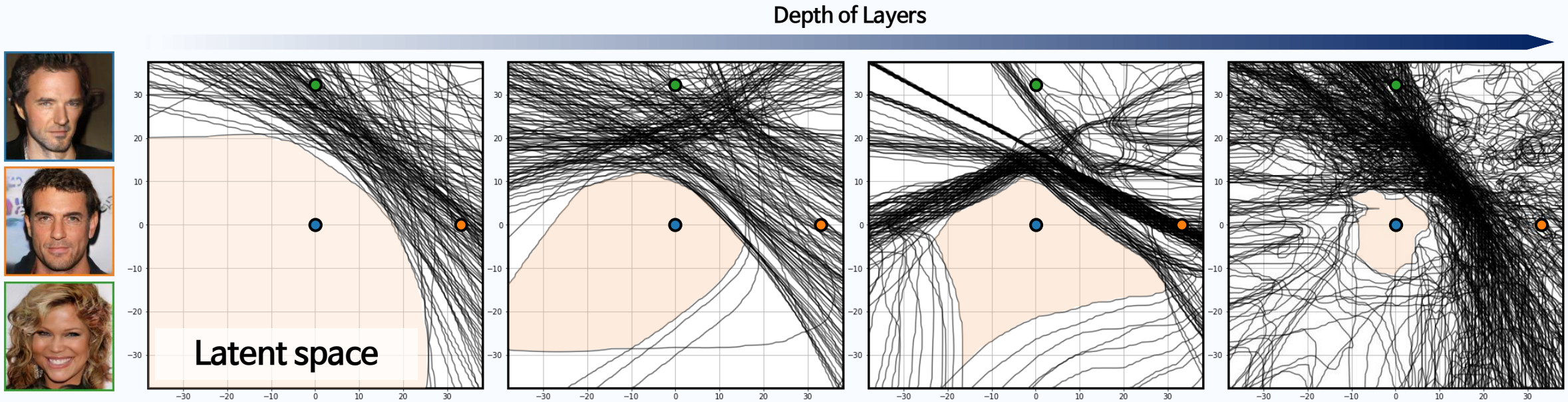


Feasible area Blind area Out-of-region area

- All samples in the region
- Exist blind area

- Cover the region
- Out-of-region samples

Challenges of Sampling in a Generative Region

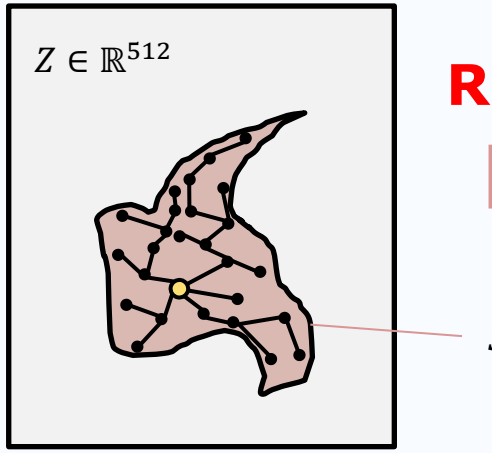


*We visualize only 200 GBs (with high activation corresponding the blue dot) for clear visualization

EMPOWERING EXPERTISE
iNEEJI
KAIST

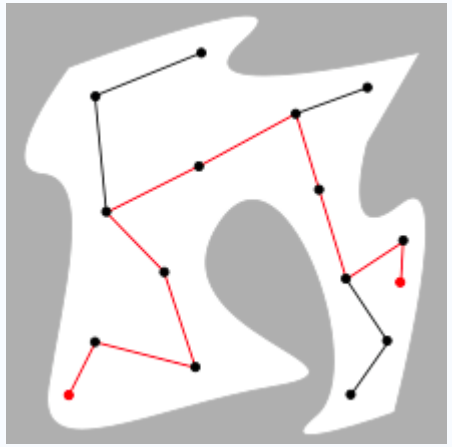
Reduction to the Robot Planning Problem

Exploring a Generative Region Problem



Reduction

Robot Planning Problem



- Searching samples in non-convex space
- High dimensional explorative space

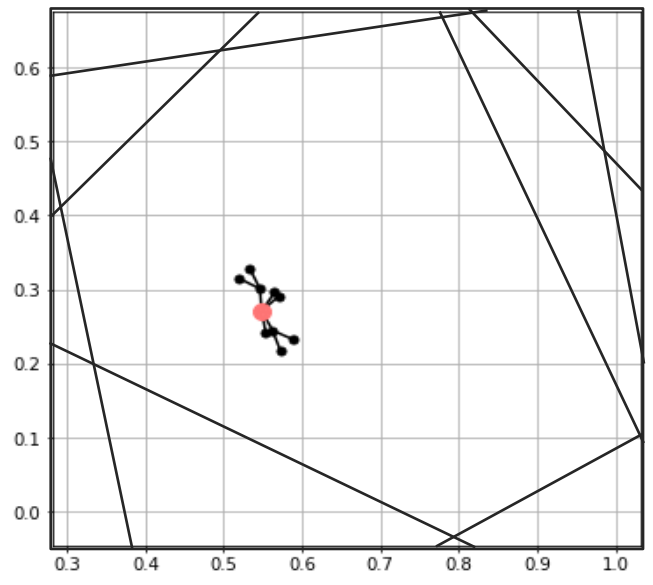
- Searching a path in non-convex space
- High degree of freedom of robot joint

We reduce our sampling problem into robot-planning problem.

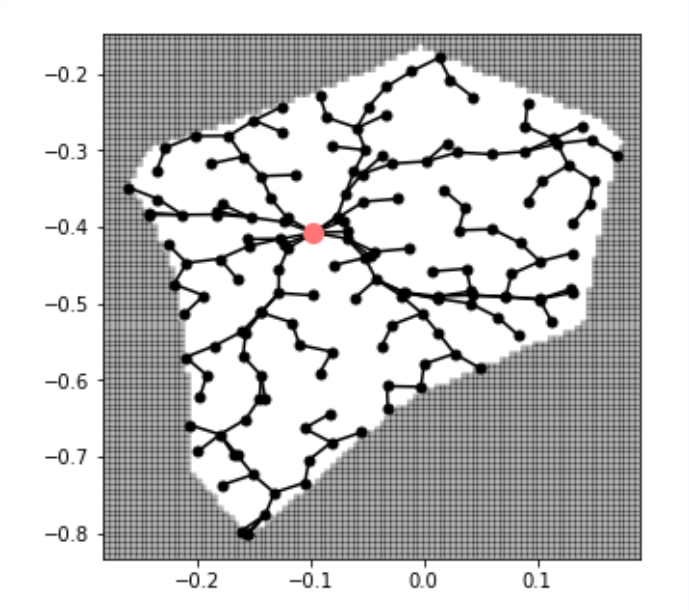
EMPOWERING EXPERTISE
iNEEJI
KAIST

Generative Boundary constrained Rapidly-exploring Random Tree (GB-RRT)

Given generative boundary as constraints,
RRT is gives solution to search over the generative region.
This explorative sampling always guarantee acceptance inside the region



Illustrative example

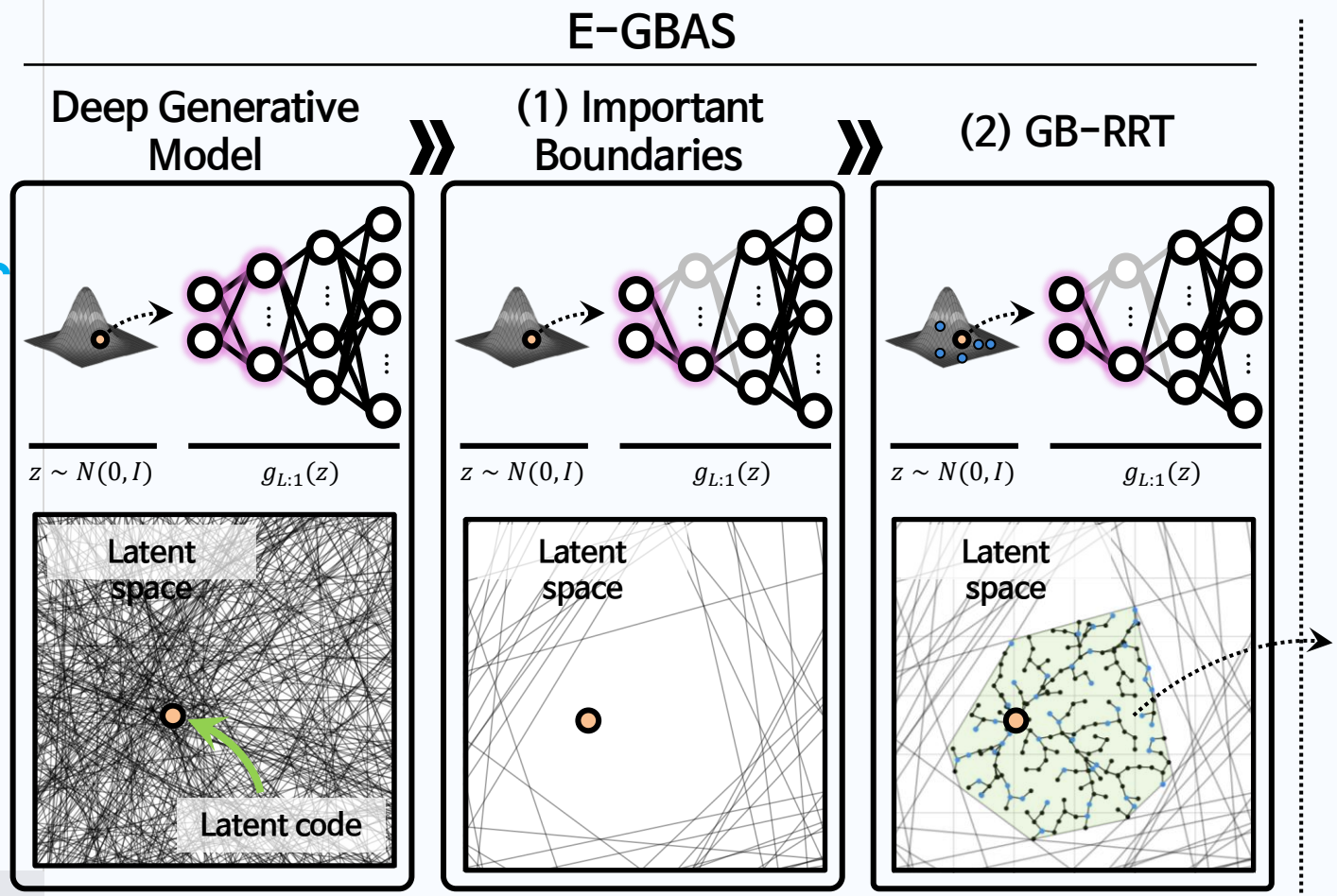


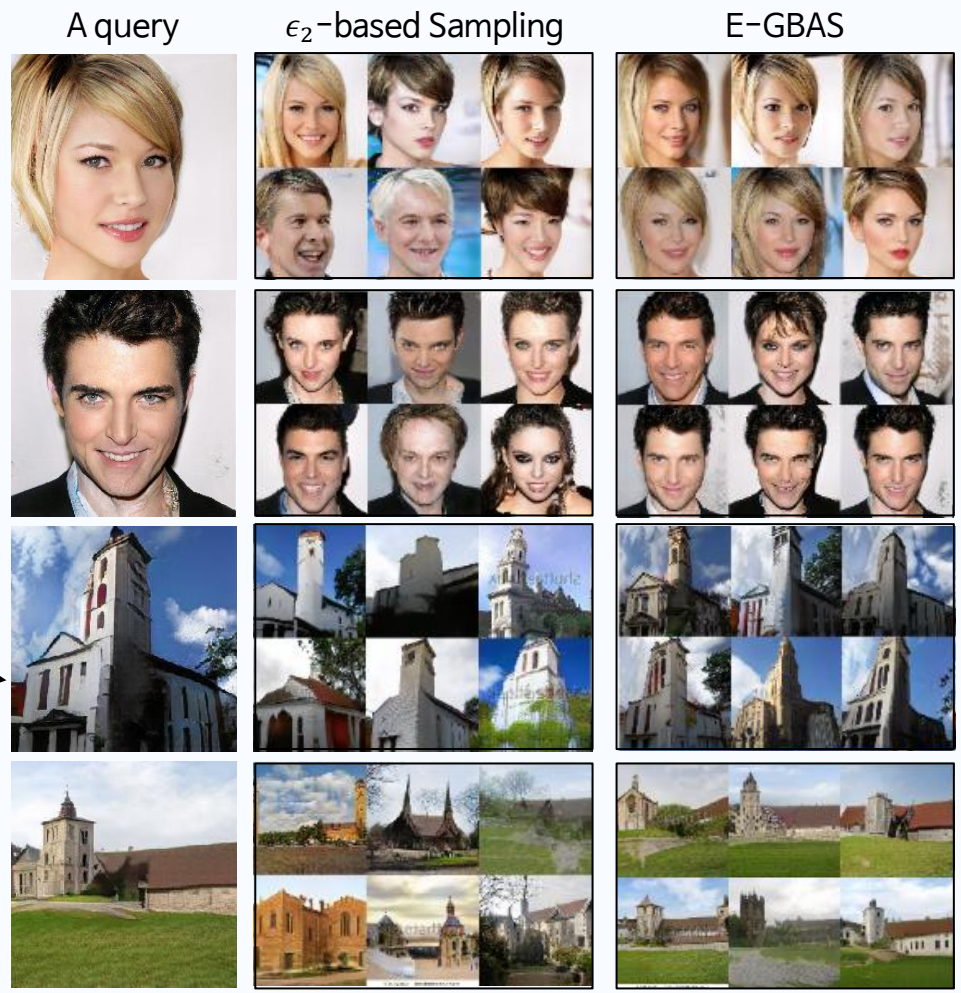
Example in nonconvex region

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

Analyzing Inside of Deep Neural Networks – E-GBAS

Proposed Algorithm





Analyzing Inside of Deep Neural Networks – E-GBAS

Explorative Generative Boundary Aware Sampling

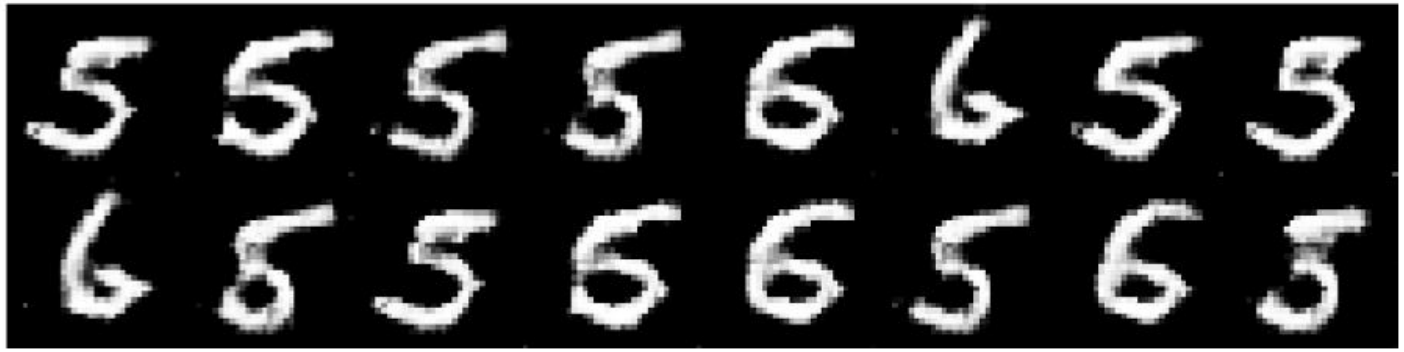
- Accepted Cluster 1
- Accepted Cluster 2
- Accepted Cluster 3
- Rejected Sample

EMPOWERING EXPERTISE
iNEEJI
KAIST

Experiment : DCGAN-MNIST

Query

ϵ-based sampling



Our method

EMPOWERING EXPERTISE
iNEEJI
KAIST

Experiment : PGGAN-LSUN-church

Query

ϵ-based sampling

Our method

iNEEJI EMPOWERING EXPERTISE
KAIST

Experiment : PGGAN-celebA

Query

ϵ-based sampling

Our method

EMPOWERING EXPERTISE
INEEJI
KAIST

Experiment : According to the portion of activate mask

iNEEJI EMPOWERING EXPERTISE
KAIST

Q2: Can we find internal units which behaves badly (e.g., making artifacts) in deep neural networks?

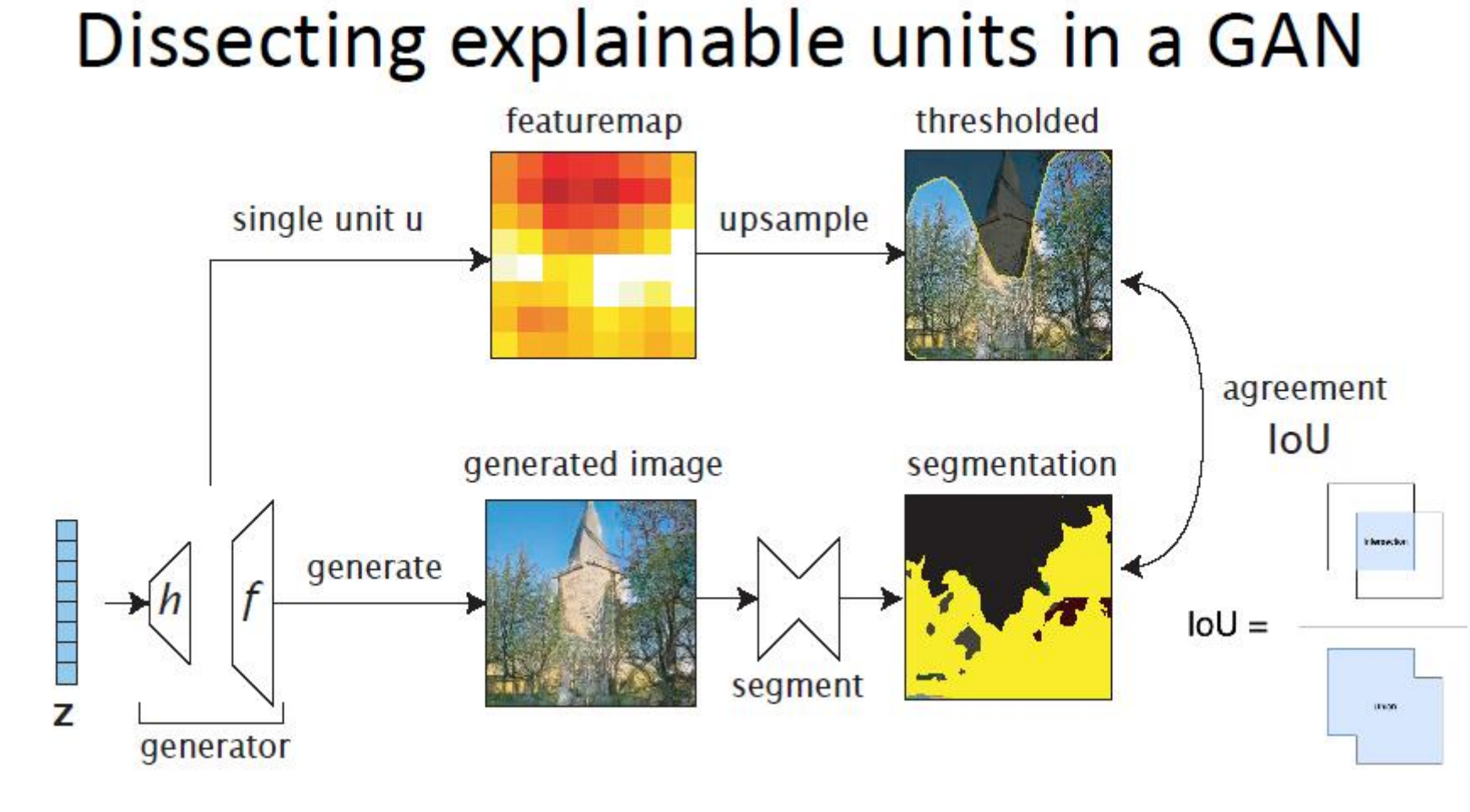
Q3: Can we control internal units which behaves badly in deep neural networks?

*Equal contribution

Related Work - Dissecting Deep Generative Neural Networks

GAN Dissection

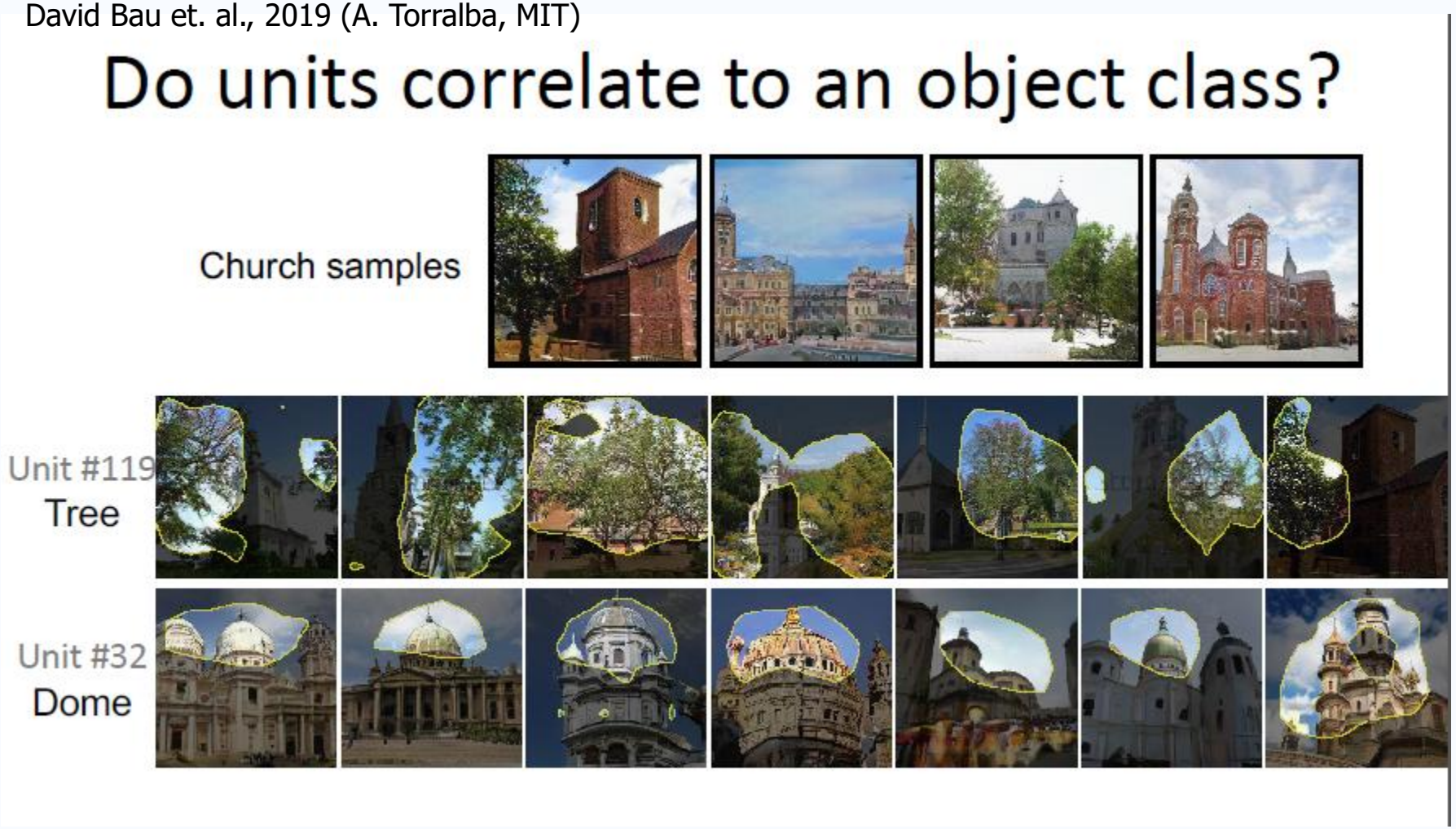
David Bau et. al., 2019 (A. Torralba, MIT)



Related Work - Dissecting Deep Generative Neural Networks

GAN Dissection

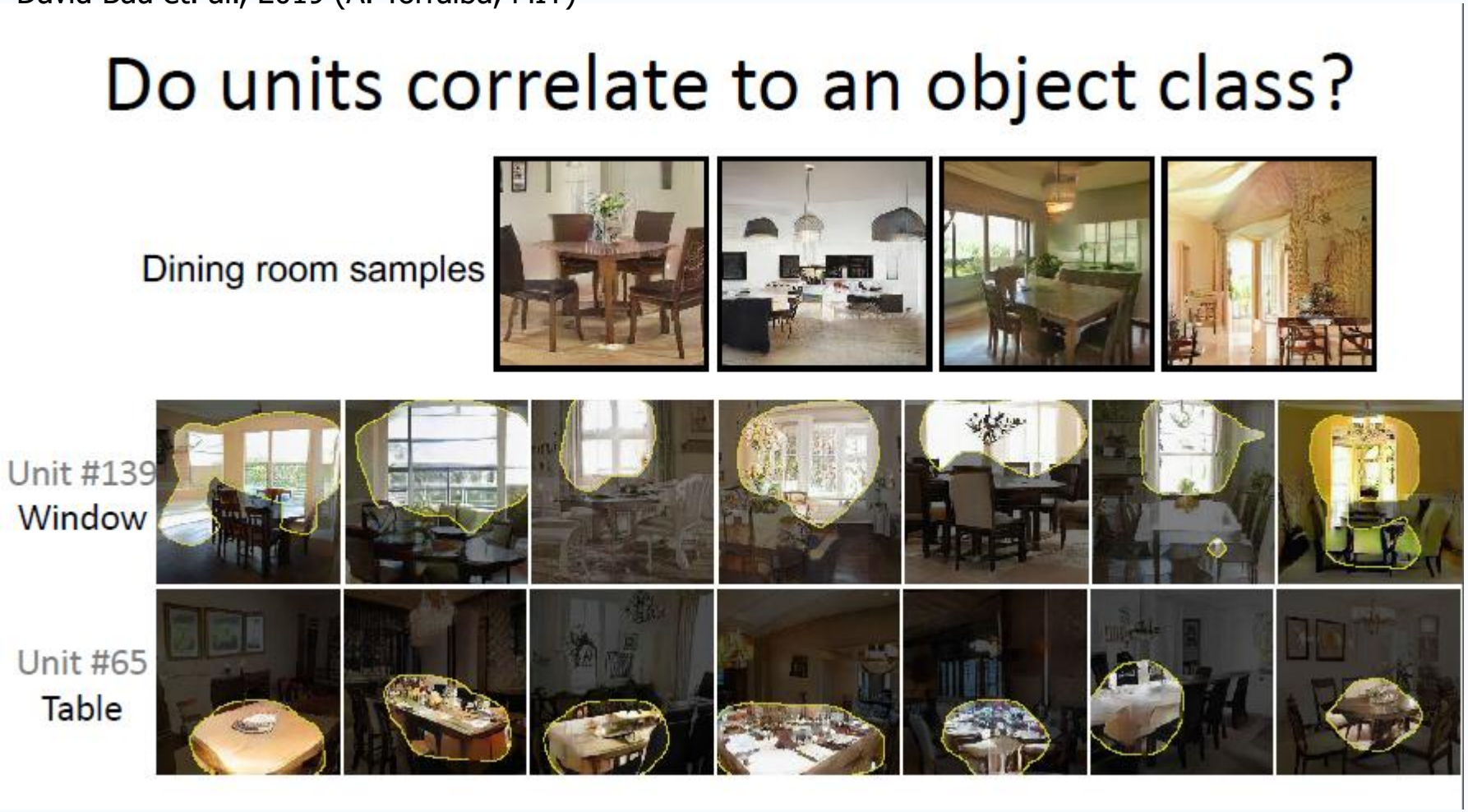
David Bau et. al., 2019 (A. Torralba, MIT)



Related Work - Dissecting Deep Generative Neural Networks

GAN Dissection

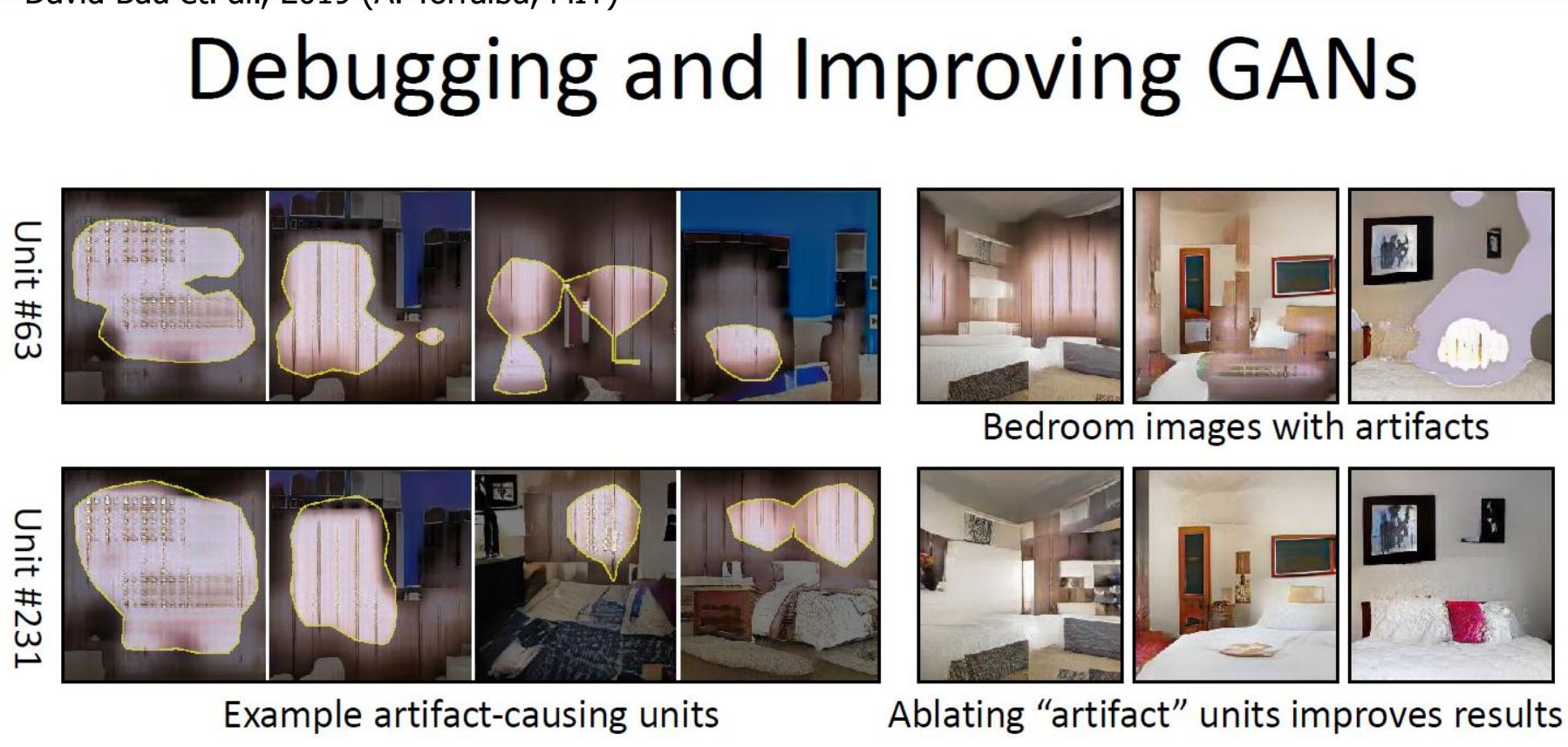
David Bau et. al., 2019 (A. Torralba, MIT)



Related Work - Dissecting Deep Generative Neural Networks

GAN Dissection

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



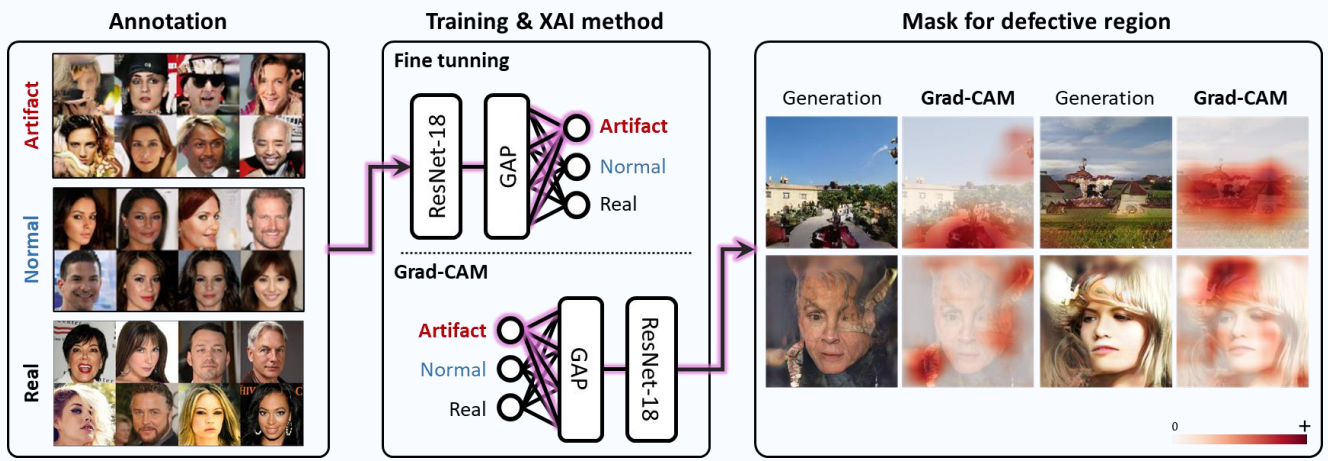
Automatic Correction of Artifact Generating Internal Units

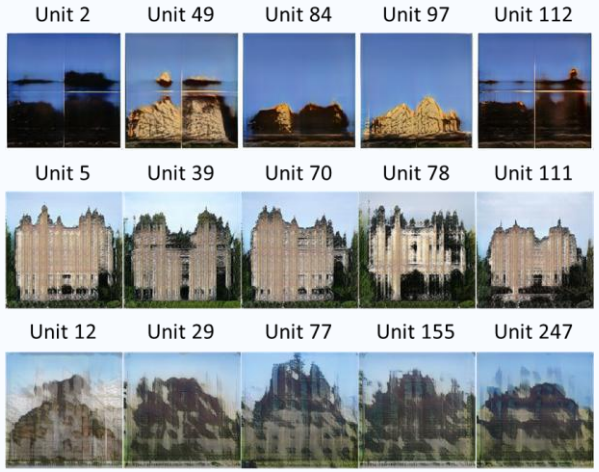
[A. Tousi*, H. Jeong*, J. Han, H. Choi and J. Choi, Automatic correction of internal units in generative neural networks, CVPR-21]

*Equal contribution

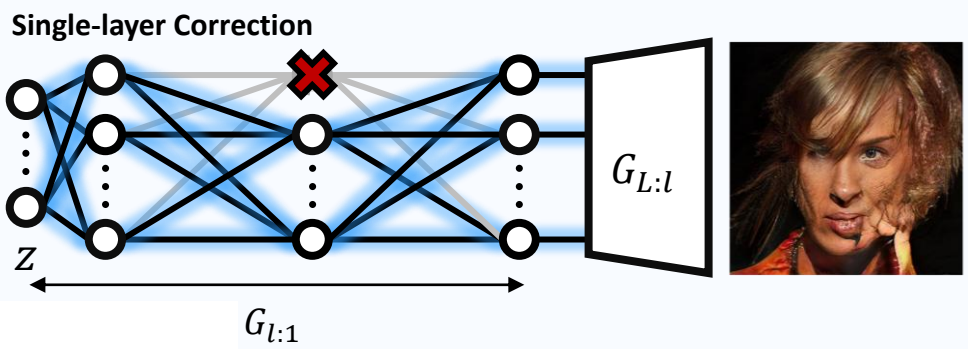
Automated Correction of Deep Neural Networks - SeqC

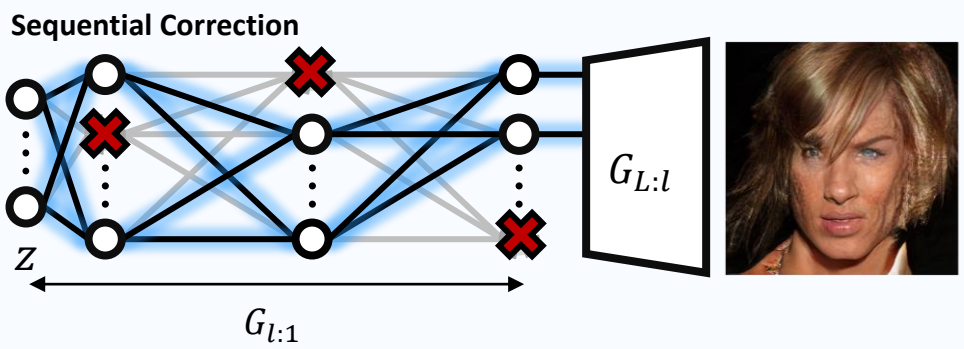
- **Finding unstable internal nodes with a XAI technique (Grad-CAM)**





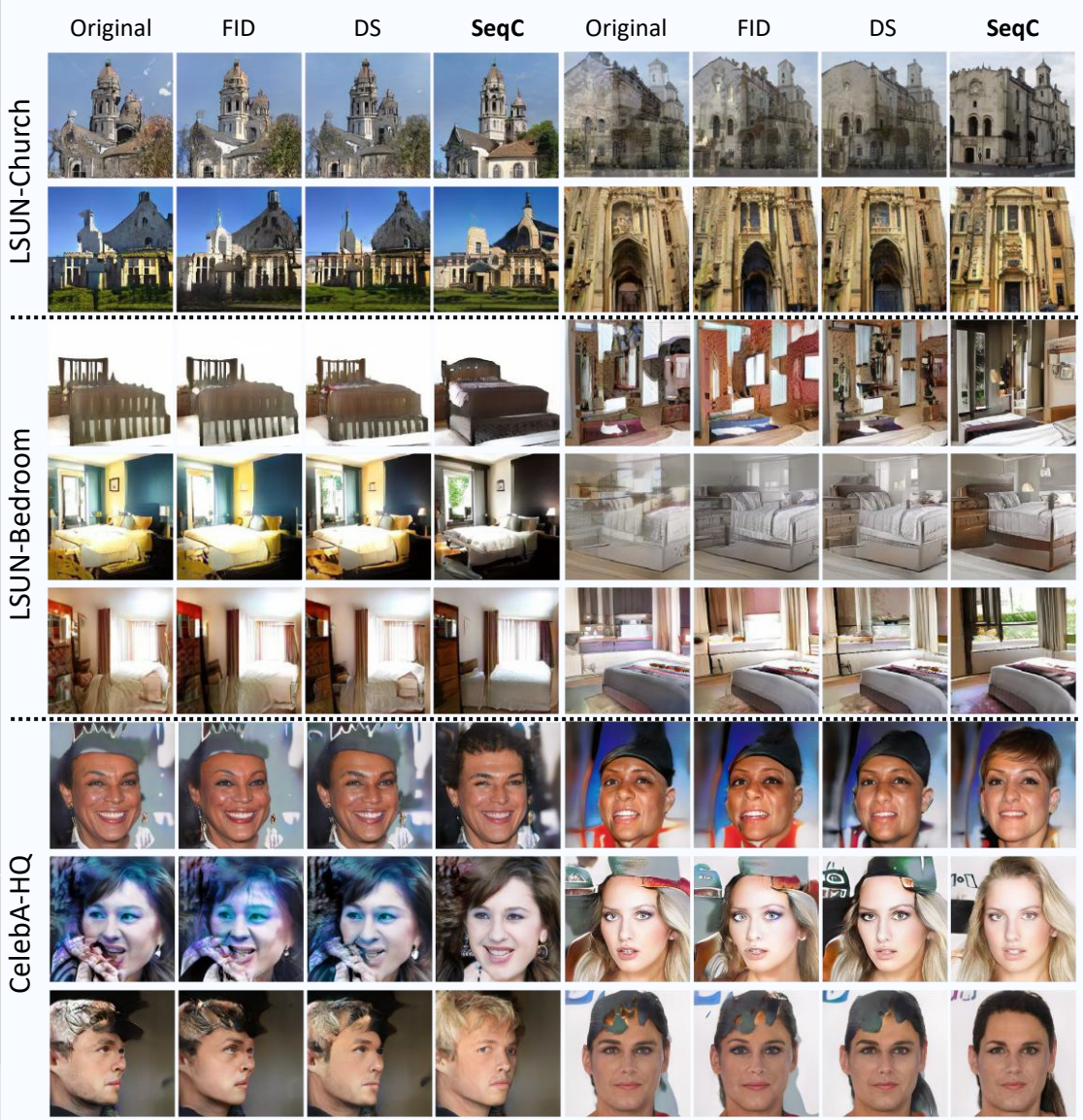
- **Deactivate unstable nodes over multiple layers to remove artifacts**

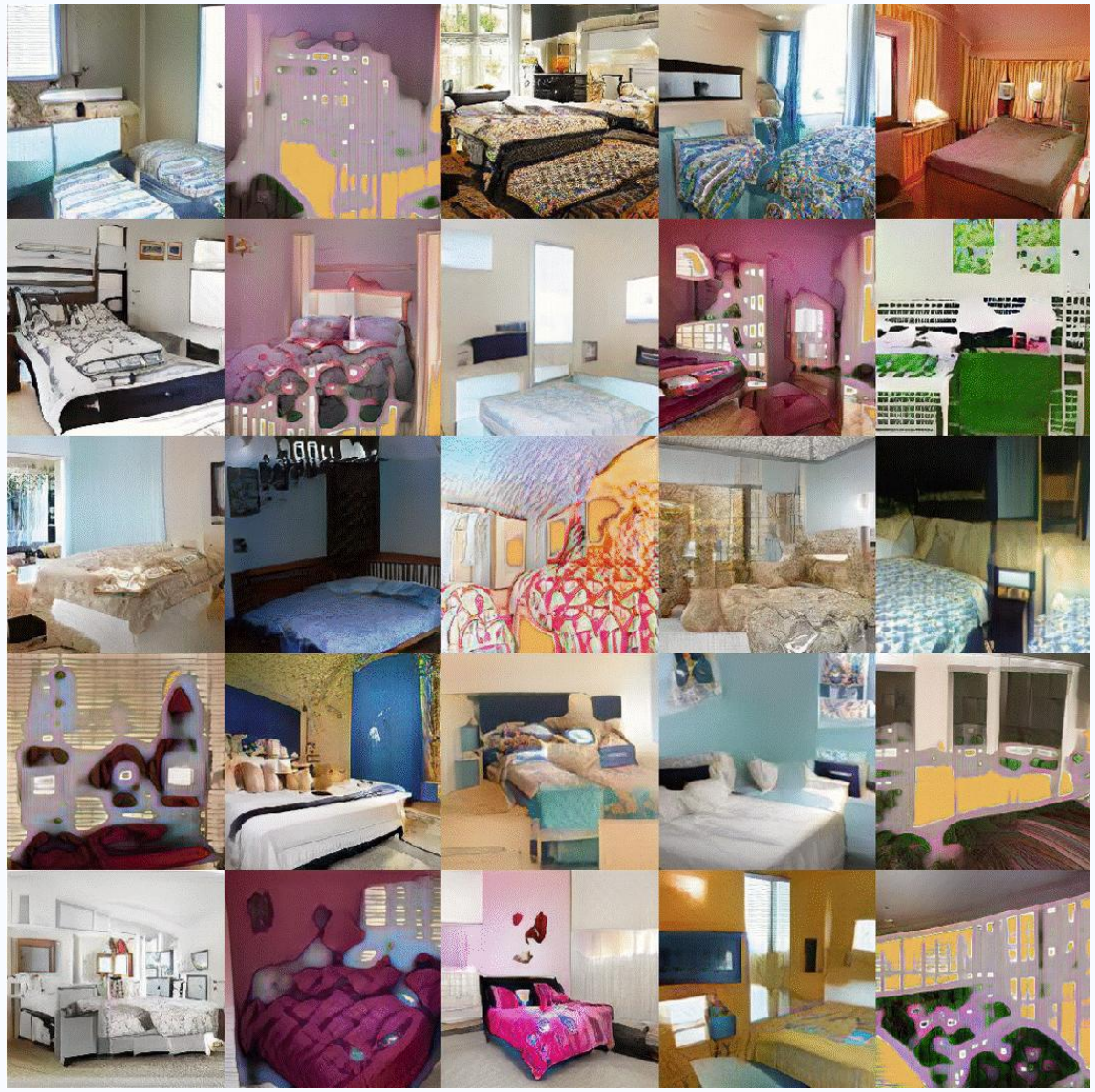




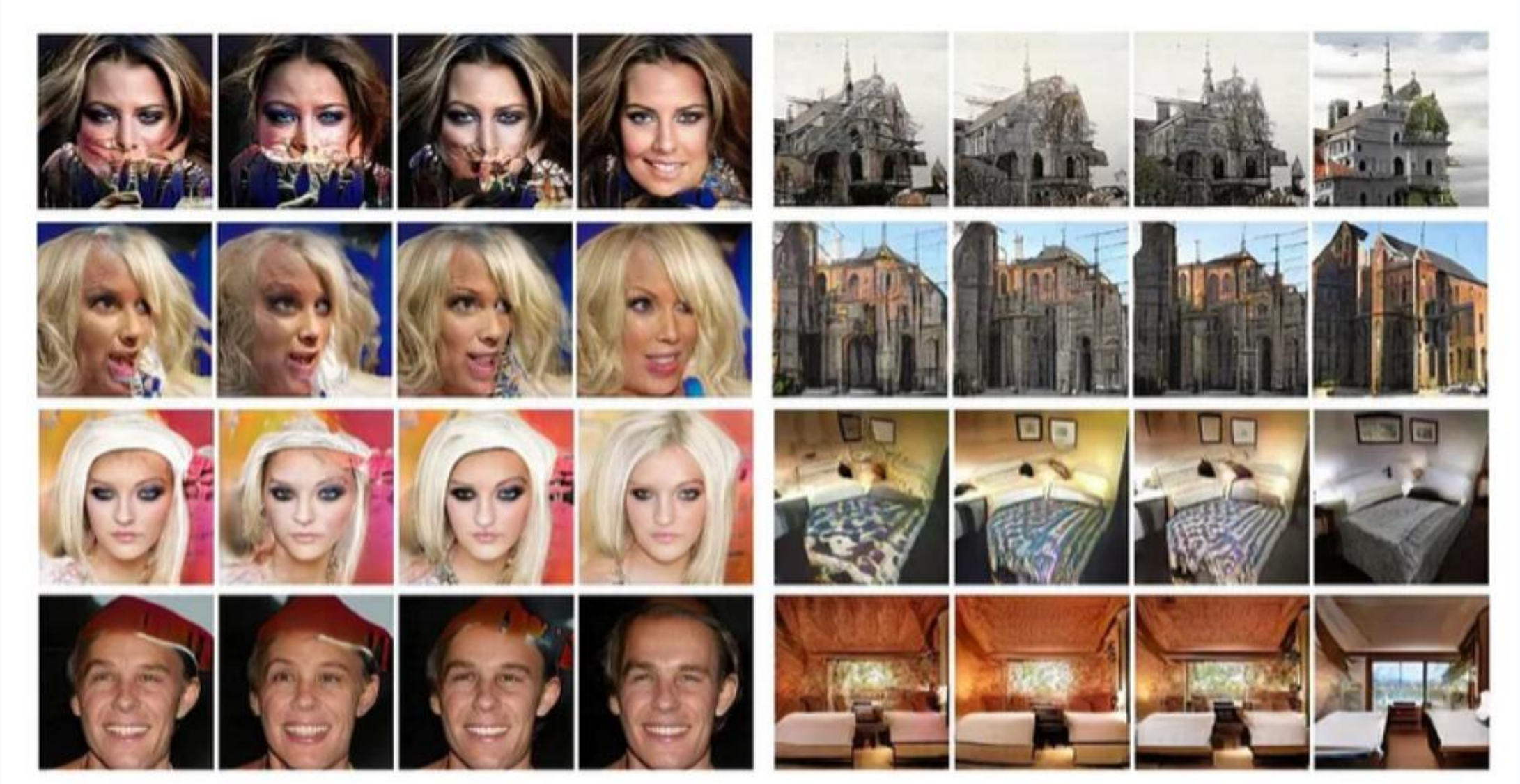
iNEEJI EMPOWERING EXPERTISE
KAIST

Automated Correction of Deep Neural Networks - SeqC





Automated Correction of Deep Neural Networks - SeqC



EMPOWERING EXPERTISE

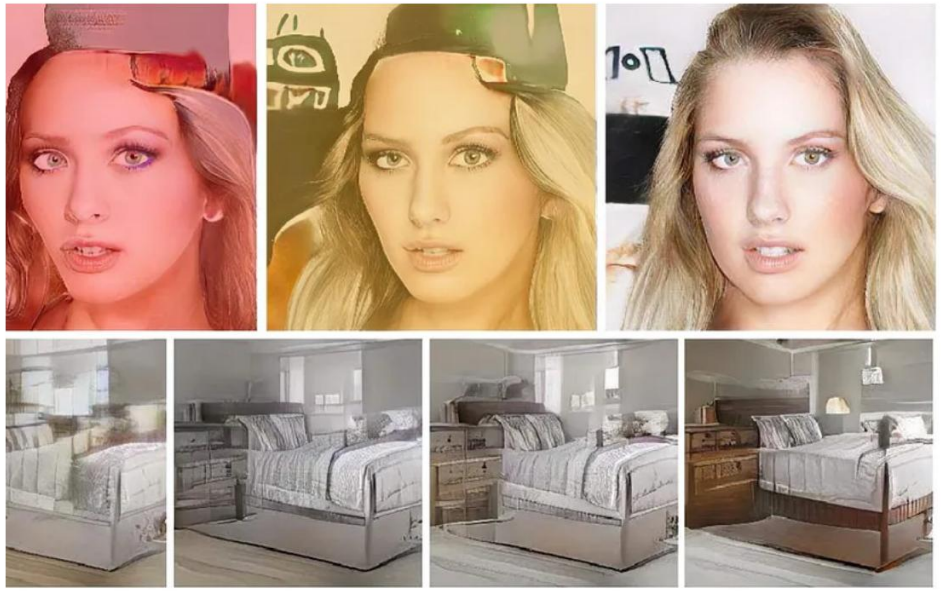
Automated Correction of Deep Neural Networks - SeqC

UNITE.AI NEWS AI 101 CERTIFICATIONS CONFERENCES INTERVIEWS THOUGHT LEADERS MEET THE TEAM CONTACT

ARTIFICIAL NEURAL NETWORKS

Plastic Surgery For GAN-Generated Faces

Updated 1 month ago on April 14, 2021
By Martin Anderson



New research out of South Korea promises to improve the quality of synthetic face data created by Generative Adversarial Networks (GANs).

The system is capable of identifying image artifacts produced by GAN processes, and remediating them, even to the point of replacing hair that was obscured by a cap, replacing parts of the face entirely missing in the original, and removing occlusions such as hands and sunglasses, and also works well on scenic and architectural output.

SEARCH

RECENT POSTS

Dr. Judith Bishop, Senior Director of AI Specialists at Appen – Interview Series

AI Enables Smartphones To Detect Ecological Change

Muddu Sudhakar, CEO and Co-Founder of Aisera – Interview Series

AI Model Predicts How We Type on Touchscreens

New AI Detects Sarcasm in Social Media

INTERVIEWS

Zayd Enam, Co-founder and CEO of Cresta

Zack Dvey-Aharon, Ph.D., CEO and Co-Founder of AEYE Health

Yohan Lee, Chief Strategy Officer at Riiid Labs

Yi Zou, Senior Director of Engineering, ASML Silicon Valley

Yasser Khan, CEO of ONE Tech

Wilson Pang, Co-Author of Real World AI

https://www.unite.ai/improving-gan-generated-faces-synthetic-data/

EMPOWERING EXPERTISE
iNEEJI
KAIST

Automated Correction of Deep Neural Networks - SeqC

The Webzine is optimized for viewing in Chrome

KAIST Breakthroughs

Biannual Engineering Research Webzine

50 INNOVERSARY

Search | Subscribe | Editorial Board

Automatic Correction of Internal Units in Generative Neural Networks

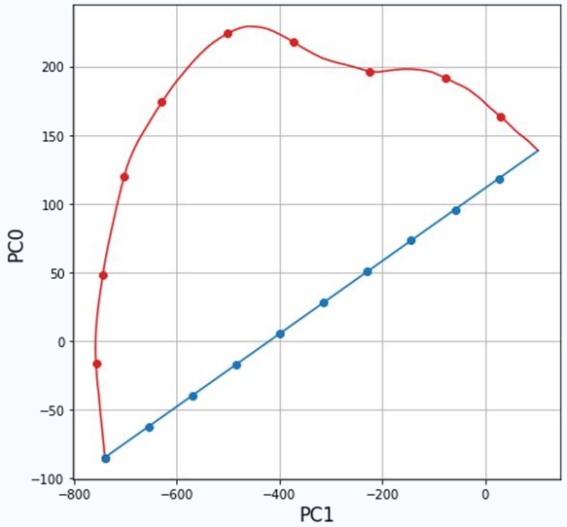
Automatic correction of internal units is a retraining-free framework to improve the quality and reliability of the deep generative neural networks which are models trained to generate various images. This method is a crucial element of the generative model for mission-critical applications.

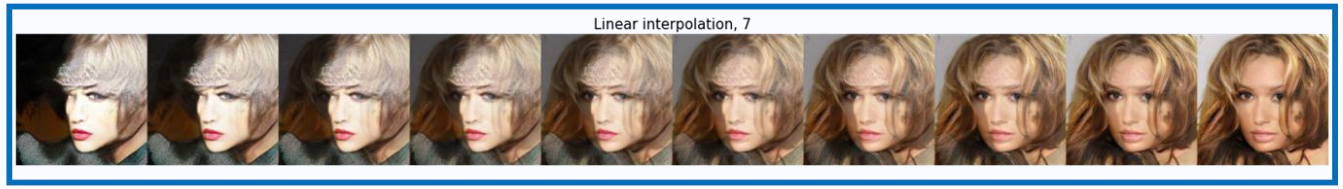
Article | Fall 2021

Prof. Jaesik Choi's research team developed a framework for the automatic correction of internal units in generative neural networks. The project was led by researchers of the KAIST AI graduate school Ali Tousi and Haedong Jeong (advisor: Jaesik Choi). This research was published under the title "Automatic Correction of Internal Units in Generative Neural Networks" at the Conference on Computer Vision and Pattern Recognition (CVPR) 2021, a top conference in computer vision.

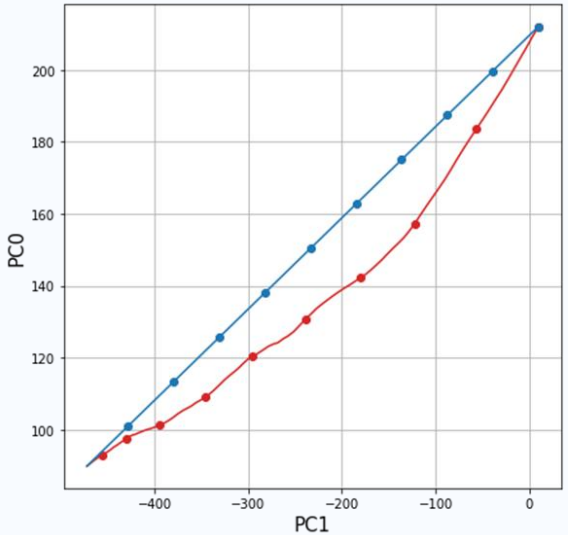
read more >>

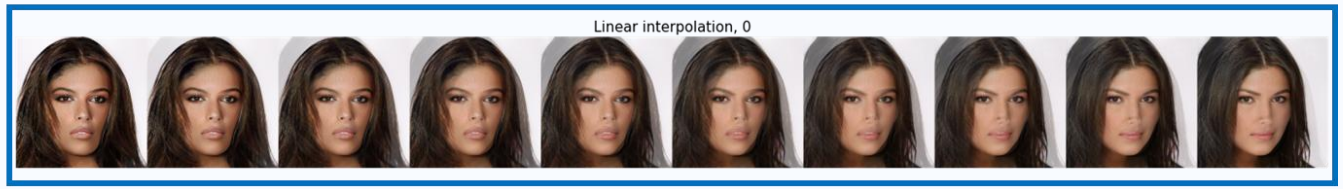
Traveling over the generative boundaries

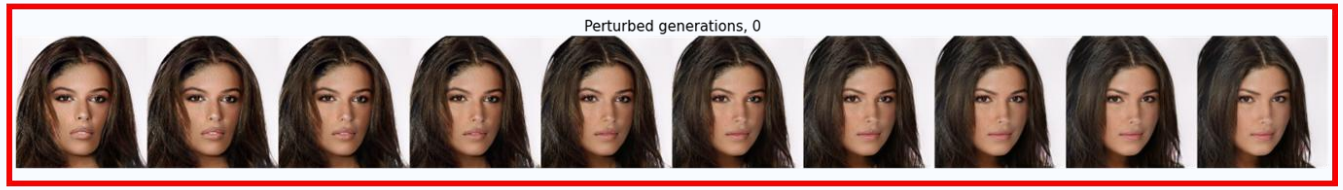




Artifact





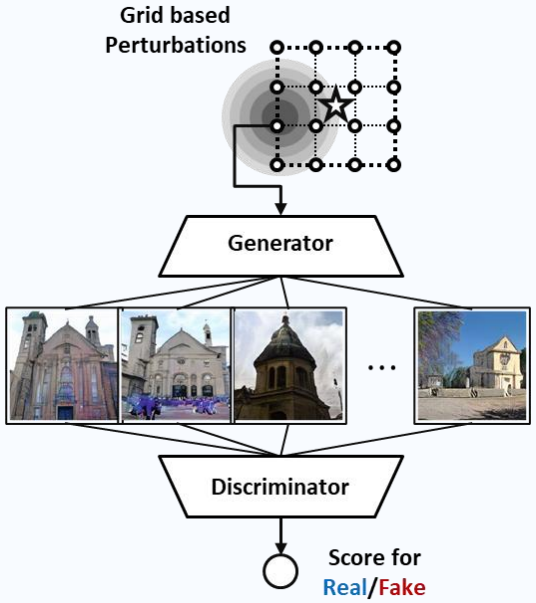


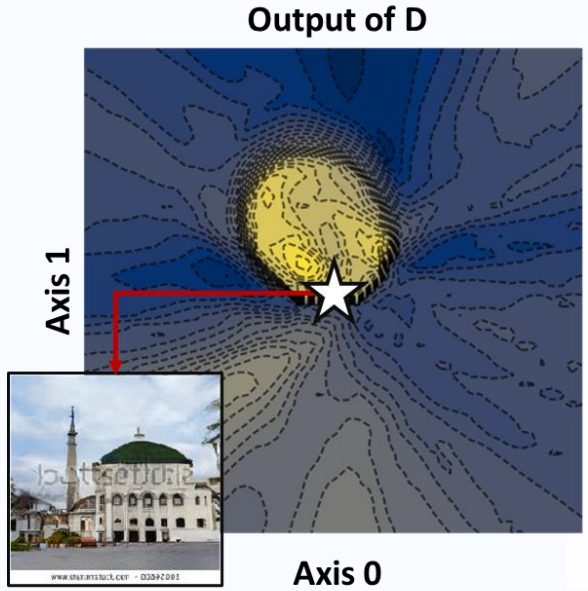
EMPOWERING EXPERTISE
iNEEJI
KAIST

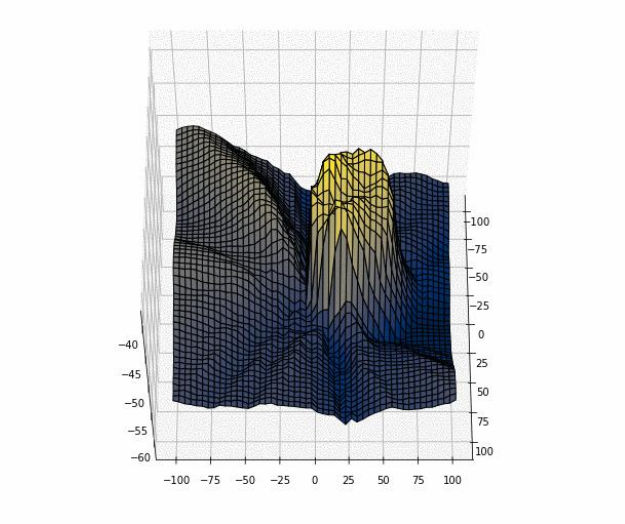
Traveling over the generative boundaries

For perturbations (grids) in the latent space

- Obtain the manifolds between the latent space and output of discriminator (D)
- For the given latent vector
  - **Although the perturbations are smooth, a pillar emerges rapidly**
  - Output of D is changed rapidly

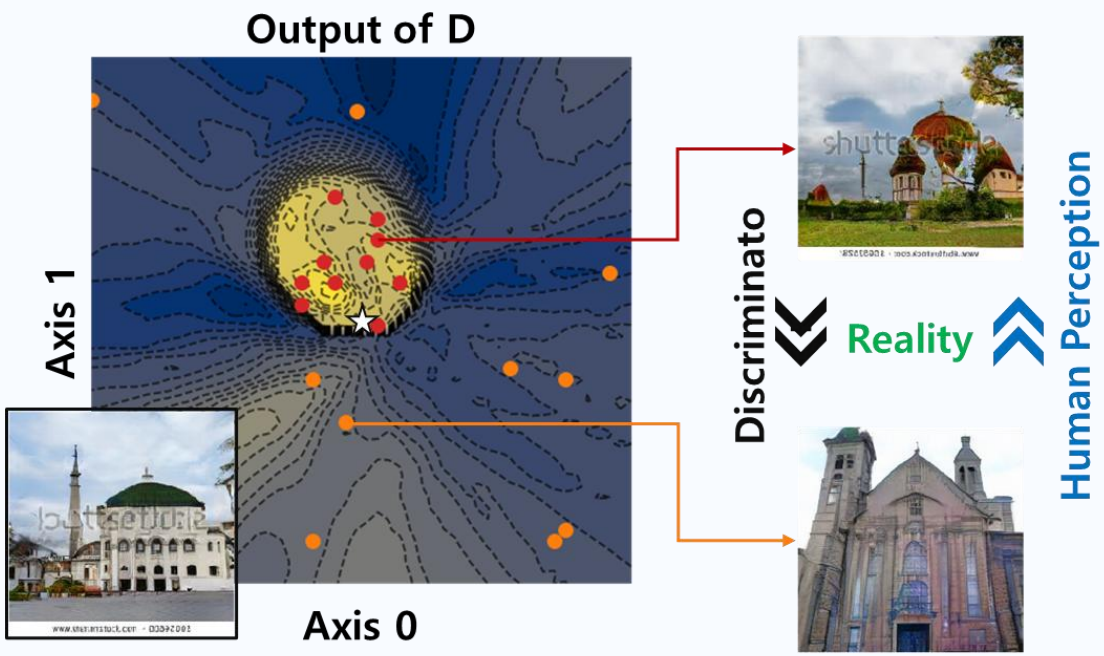






Traveling over the generative boundaries

The discriminator has high confidence to determine reality based on **the watermark** not **features** for churches

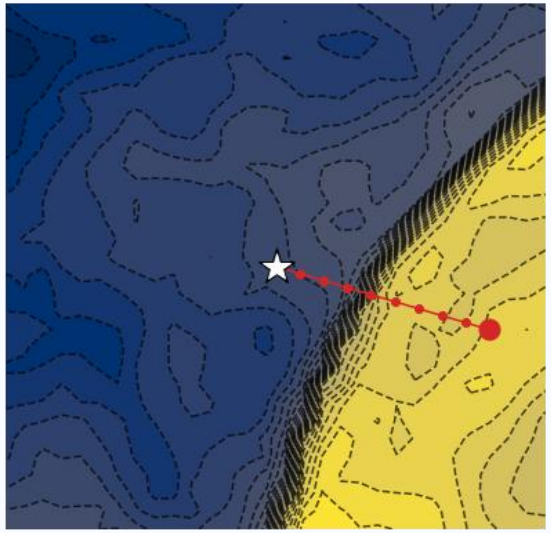


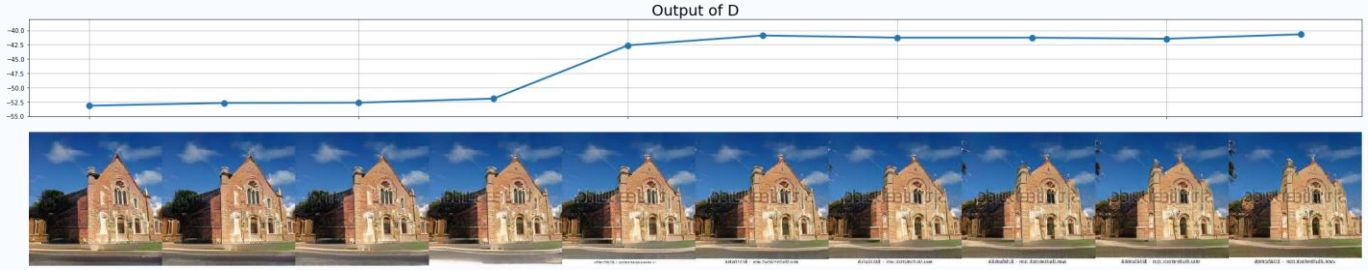
EMPOWERING EXPERTISE
iNEEJI
KAIST

Traveling over the generative boundaries

- When the watermark emerges
  - Although the almost semantic information is maintained,

  →The output of D increases rapidly





EMPOWERING EXPERTISE
iNEEJi
KAIST

Q4: Can we correct internal units which behaves badly in deep neural networks in an unsupervised way?

*Equal contribution

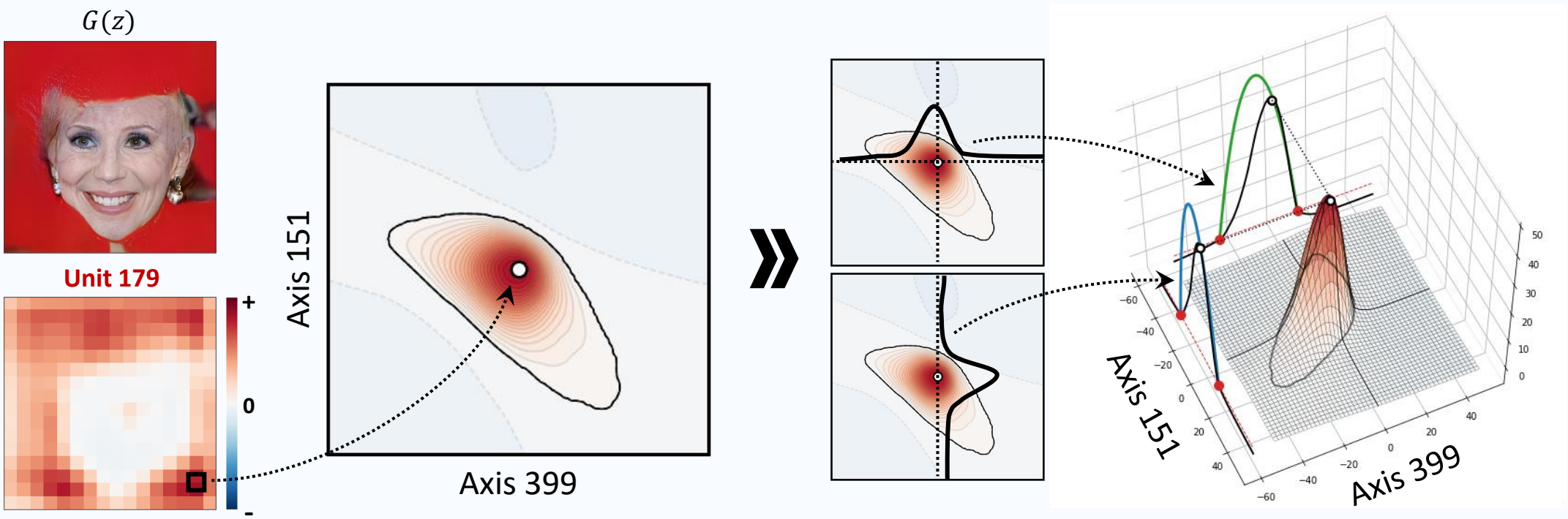
An Unsupervised Way to Understand Artifact Generating Internal Units

[H. Jeong, J. Han, and J. Choi, An Unsupervised Way to Understand Artifact Generating Internal Units in Generative Neural Networks, AAAI-22]

Automated Correction of Deep Neural Networks - Hypothesis

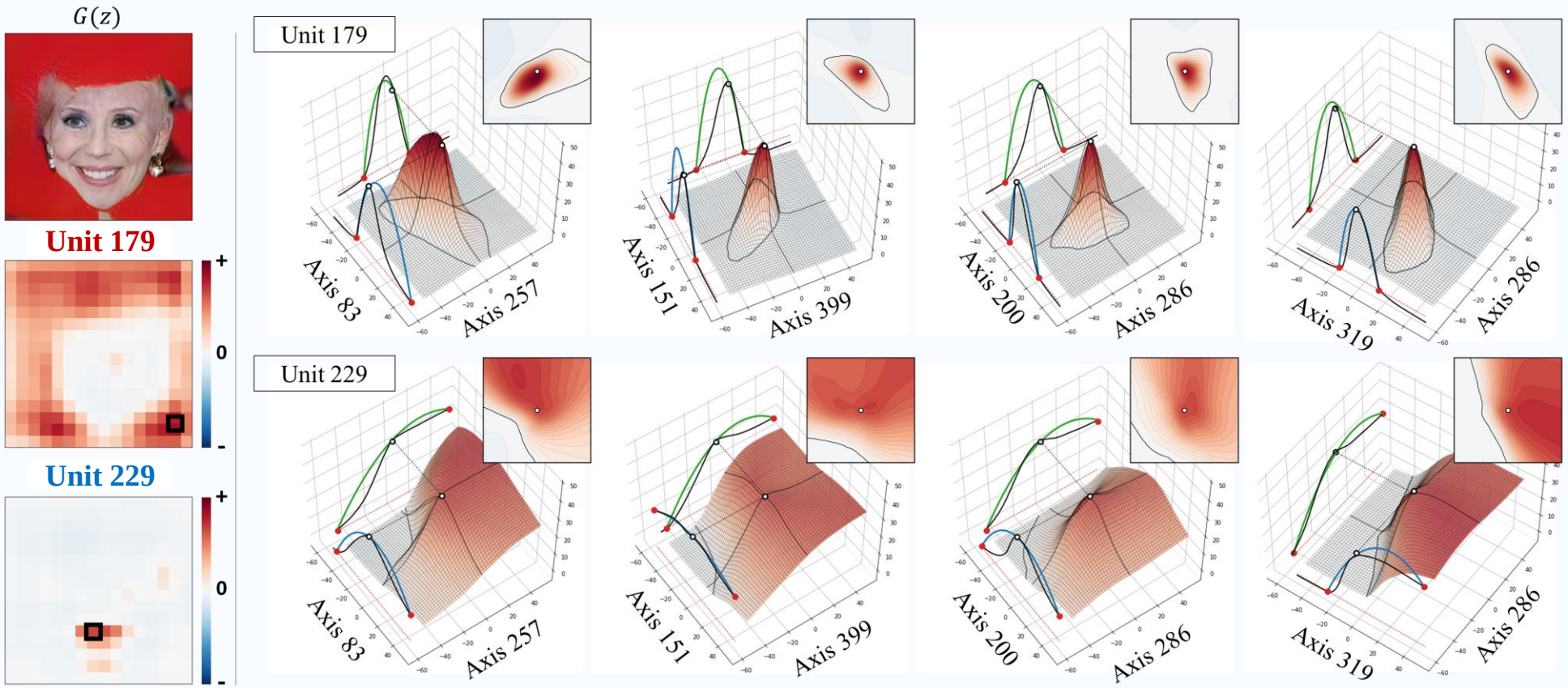
- **Hypothesis**: Locally activated neurons (internal units) may be unstable to generate artifacts.

- Here, locally activated neurons refer to neurons which strongly (sharply) are activated for a small spatial region in the laten space.



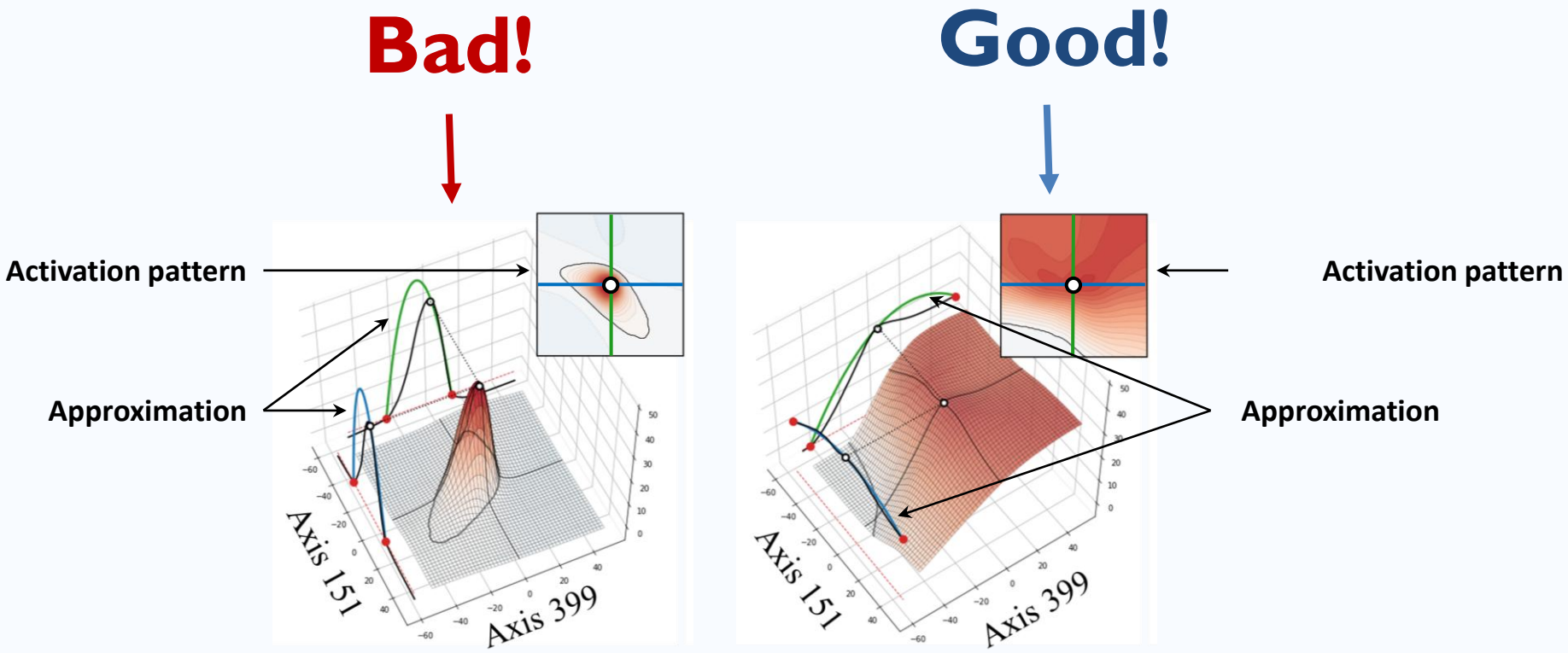
Automated Correction of Deep Neural Networks - Hypothesis

Let's compare the background (Unit 179) and the lip (Unit 229)



Automated Correction of Deep Neural Networks - Hypothesis

- Our hypothesis

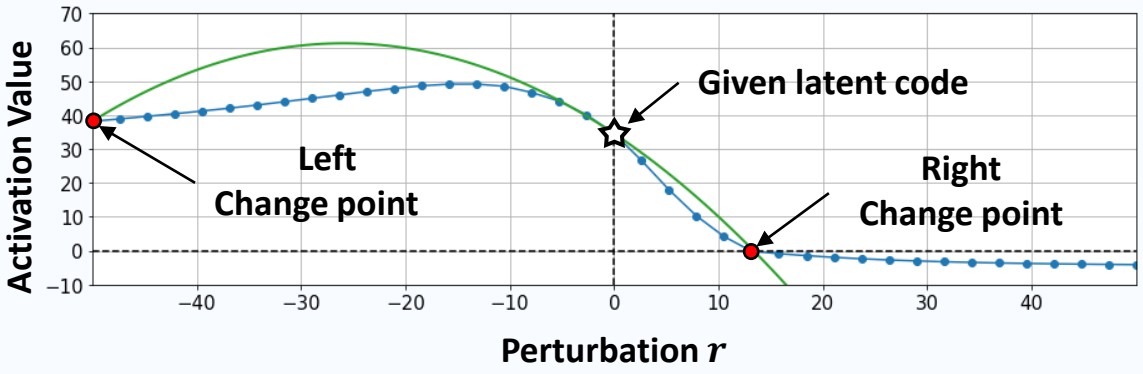


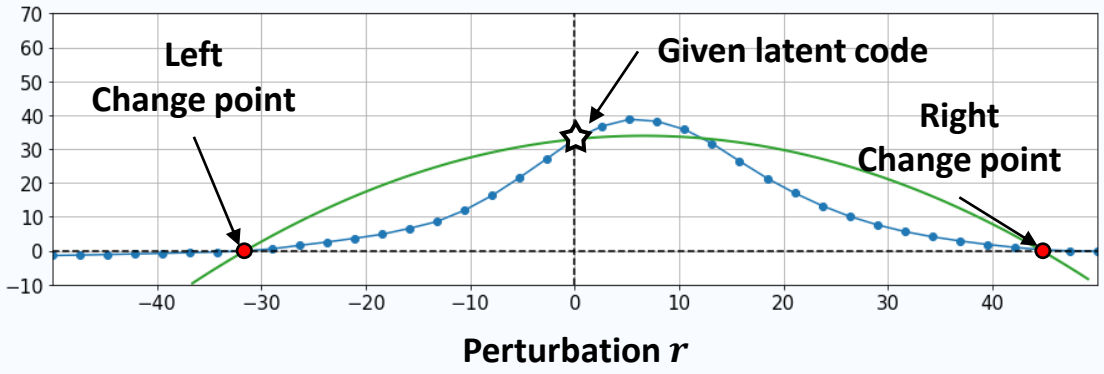
Automated Correction of Deep Neural Networks - Local Activations

- We define curvature of local activation (CLA)
  - The quadratic approximation of three points (left/right CPs and the given latent code)
  - Average the approximate CLA of each axis to represent CLA for i-th neuron

Definition (Curvature of Local Activation (CLA)) *Let the given latent code z_0, the left/right change points be ${}^{l}p$ and ${}^{r}p$ respectively as in Definition of change point. The left slope is defined as ${}^{l}s = \left(g^{i}_{l:1}(z_0 + {}^{l}p \cdot e_d) - g^{i}_{l:1}(z_0)\right) / {}^{l}p$ and the right slope is ${}^{r}s = \left(g^{i}_{l:1}(z_0 + {}^{r}p \cdot e_d) - g^{i}_{l:1}(z_0)\right) / {}^{r}p$ for latent dimension d. With $C_{i,l}(d, z_0) = ({}^{r}s - {}^{l}s)/({}^{r}p - {}^{l}p)$, the curvature of local activation for the i-t h neuron in the l-t h layer around the given latent code z_0 is defined as,*

$$\bar{C}_{i,l}(z_0) = \frac{1}{D_z} \sum_{d=1}^{D_z} C_{i,l}(d, z_0)$$



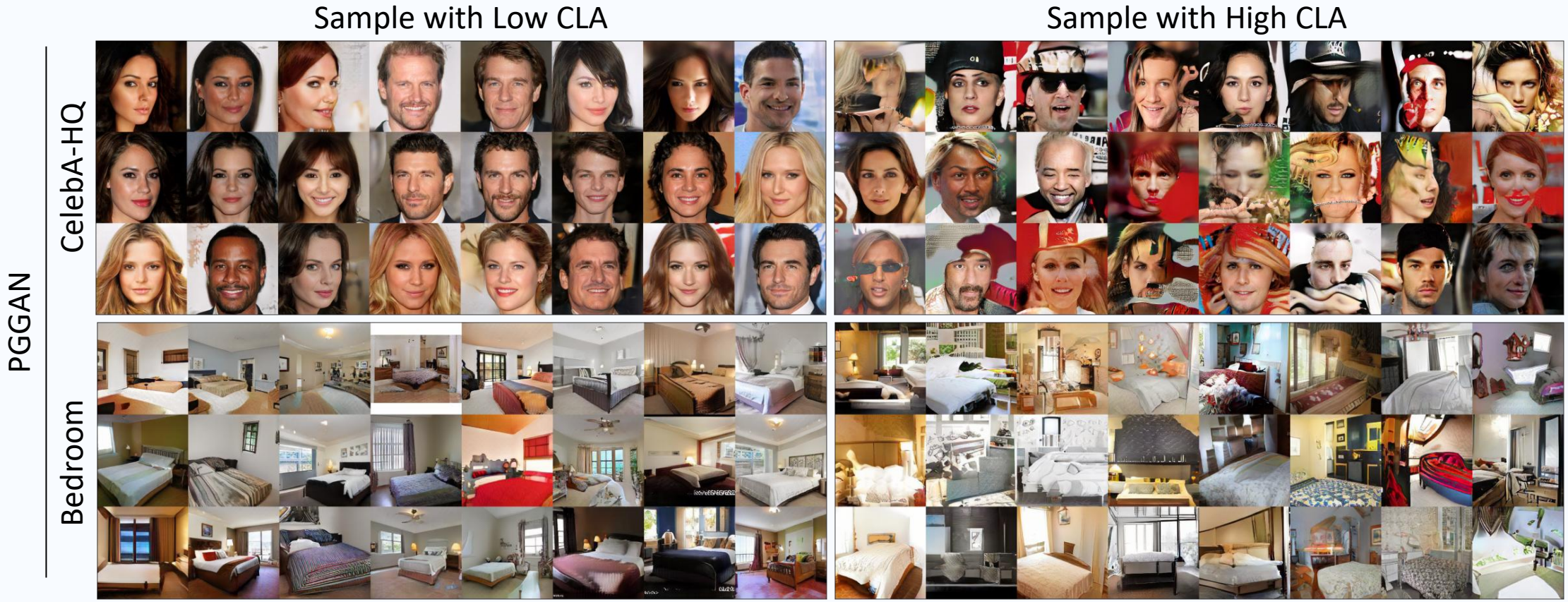


Qualitative Evaluations

We randomly select 10k latent codes without truncation for each GAN and calculate the score $S_l(z)$ on layer $l \in \mathbb{R}^{512\times8\times8}$ for each generation

We divide top/bottom 1k samples as High/Low CLA groups based on the score, respectively

Artifact Detection



EMPOWERING EXPERTISE
iNEEJI
KAIST

Qualitative Evaluations

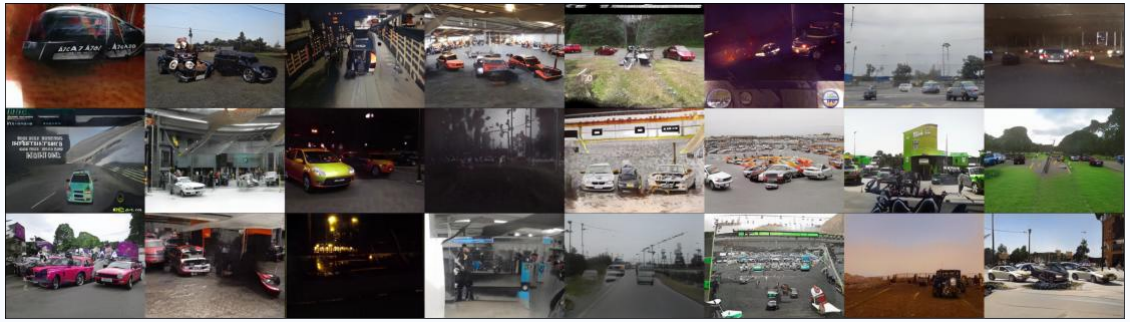
We randomly select 10k latent codes without truncation for each GAN and calculate the score $S_l(z)$ on layer $l \in \mathbb{R}^{512\times8\times8}$ for each generation
We divide top/bottom 1k samples as High/Low CLA groups based on the score, respectively.

Artifact Detection

StyleGAN2

Sample with Low CLA | Sample with High CLA

Car



Cat

Quantitative Evaluations

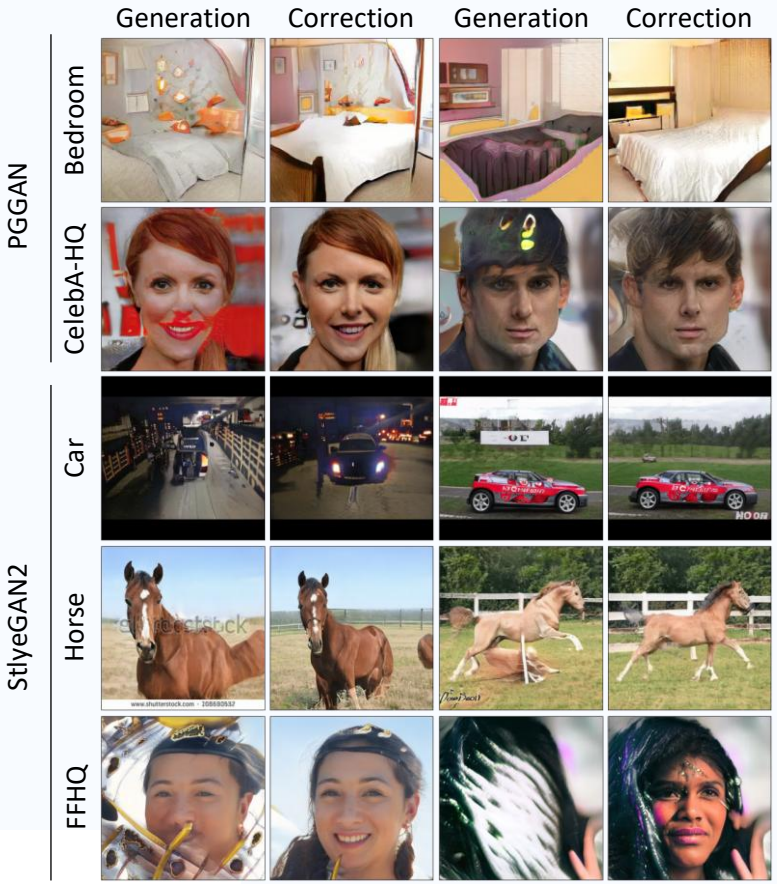
- We measure the RS [1] and PPL [2] to evaluate the average of visual fidelity for each group
  - RS needs **train dataset** and **pre-trained network** to construct feature manifold
  - PPL needs a **pre-trained network** to calculate the perceptual difference
- We make random group for baseline (30 trials)

Metric	Model	Dataset	Random	Low CLA	High CLA
RS (↑ is better)	PGGAN	LSUN-Bedroom	1.028±0.003	**1.042**	1.017
		LSUN-Chruch	1.036±0.004	**1.059**	1.012
		CelebA-HQ	1.076±0.004	**1.132**	1.011
	StyleGAN2	LSUN-Car	1.066±0.004	**1.084**	1.044
		LSUN-Cat	1.048±0.004	**1.071**	1.027
		LSUN-Horse	**1.056±0.004**	1.046	1.053
		FFHQ	1.075±0.004	**1.077**	1.069
PPL (↓ is better)	PGGAN	LSUN-Bedroom	423.8±7.1	**243.9**	683.3
		LSUN-Chruch	356.4±7.9	**213.7**	558.0
		CelebA-HQ	243.1±12.9	**114.9**	443.7
	StyleGAN2	LSUN-Car	1472.6±29.3	**920.7**	1938.9
		LSUN-Cat	1501.3±27.7	**1053.2**	2060.4
		LSUN-Horse	1207.4±21.5	**885.5**	1552.5
		FFHQ	484.9±19.2	**377.7**	596.2

[1] Kynk¨a¨anniemi et al. "Improved Precision and Recall Metric for Assessing Generative Models." *NeurIPS 2019*
[2] *Karras et al. "A Style-based Generator Architecture for Generative Adversarial Networks." CVPR 2019*

Quantitative Evaluations

- We can perform internal artifact generating unit identification
  - To validate the detection procedure, we apply sequential correction with detected unit for high CLA group
- Eliminating High CLA units improve the fidelity of individual samples in many datasets



Metric	Model	Dataset	High CLA	Correction
RS (↑ is better)	PGGAN	LSUN-Bedroom	1.017	1.002
		LSUN-Chruch	1.012	1.000
		CelebA-HQ	1.011	**1.018**
	StyleGAN2	LSUN-Car	1.044	**1.061**
		LSUN-Cat	1.027	**1.047**
		LSUN-Horse	1.053	**1.054**
		FFHQ	1.069	**1.097**

Can We Find Neurons that Cause Unrealistic Images in Deep Generative Networks?

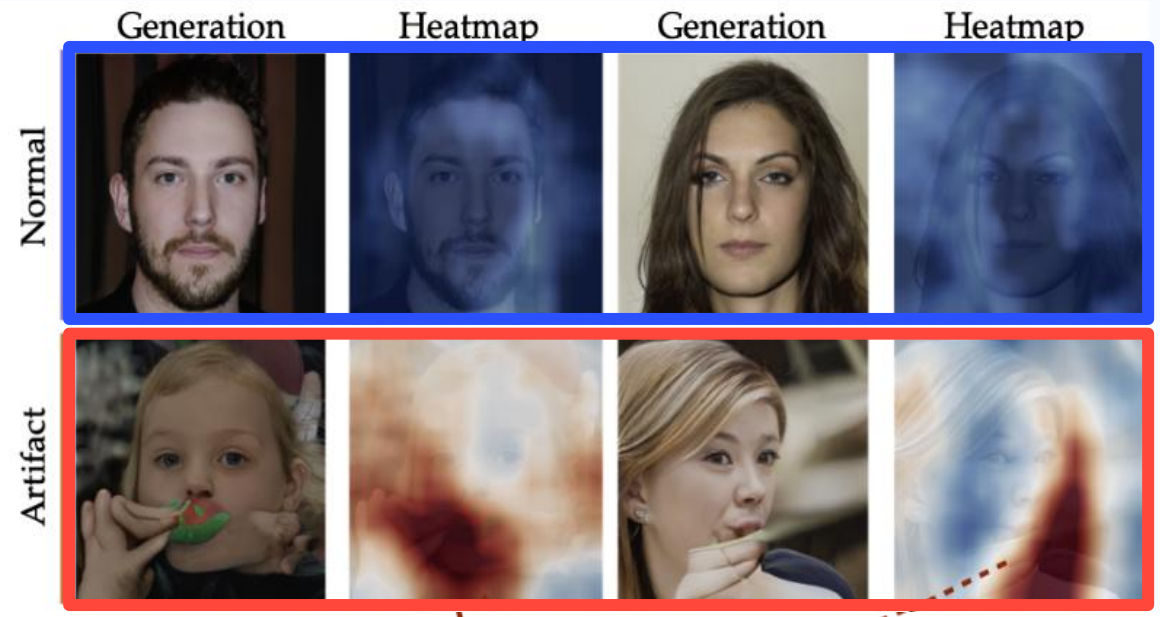
[H. Choi, W. Chang and J. Choi, Can We Find Neurons that Cause Unrealistic Images in Deep Generative Networks?, IJCAI-22]

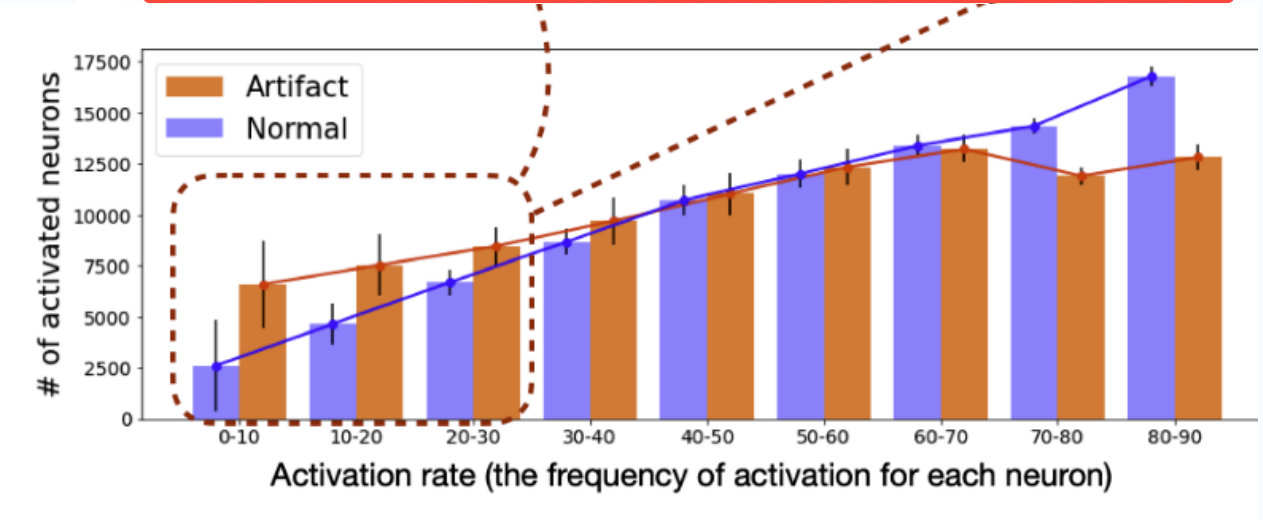
Automated Correction of Deep Neural Networks – Hypothesis II

Hypothesis

- **Rarely activated neurons (LR-neurons) are used to create diverse components including failures.**
- **Frequently activated neurons (HR-neurons) are used to construct general components** that are necessary for realistic images.

Artifact images have many Rarely Activated Neurons (LR-neurons) in their heatmaps.





***Activation rate (frequency)** : probability that a neuron will be activated.

Automated Correction of Deep Neural Networks – Hypothesis II

EXAMPLE: Generated Cat Images in PGGAN

few LR-neurons

many LR-neurons

EMPOWERING EXPERTISE
iNEEJI
KAIST

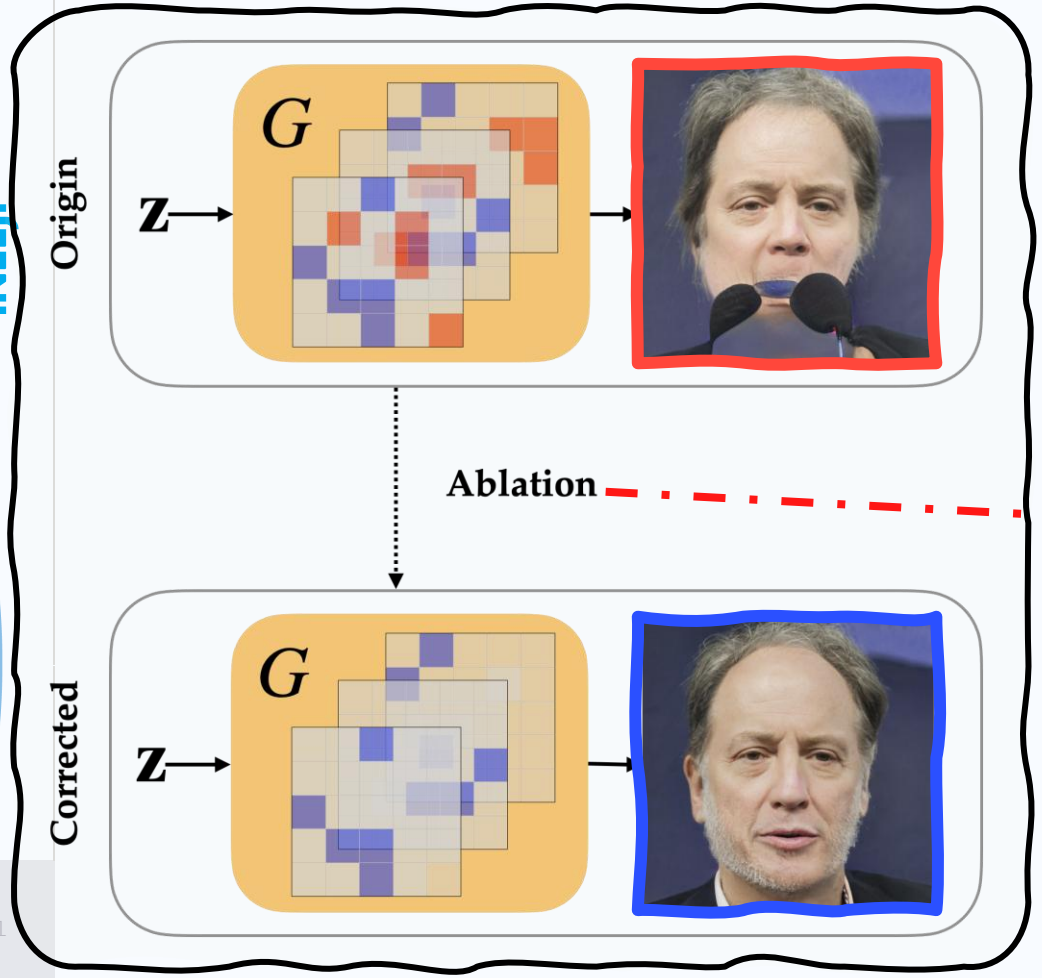
Automated Correction of Deep Neural Networks – Approach II

How can we correct artifact parts in the generated images?

It will be corrected **by turning off the activated LR-neurons! - Sequential Ablation**

G: Generator
$\mathbf{z}$: latent variables

: HR-Neuron
: LR-Neuron



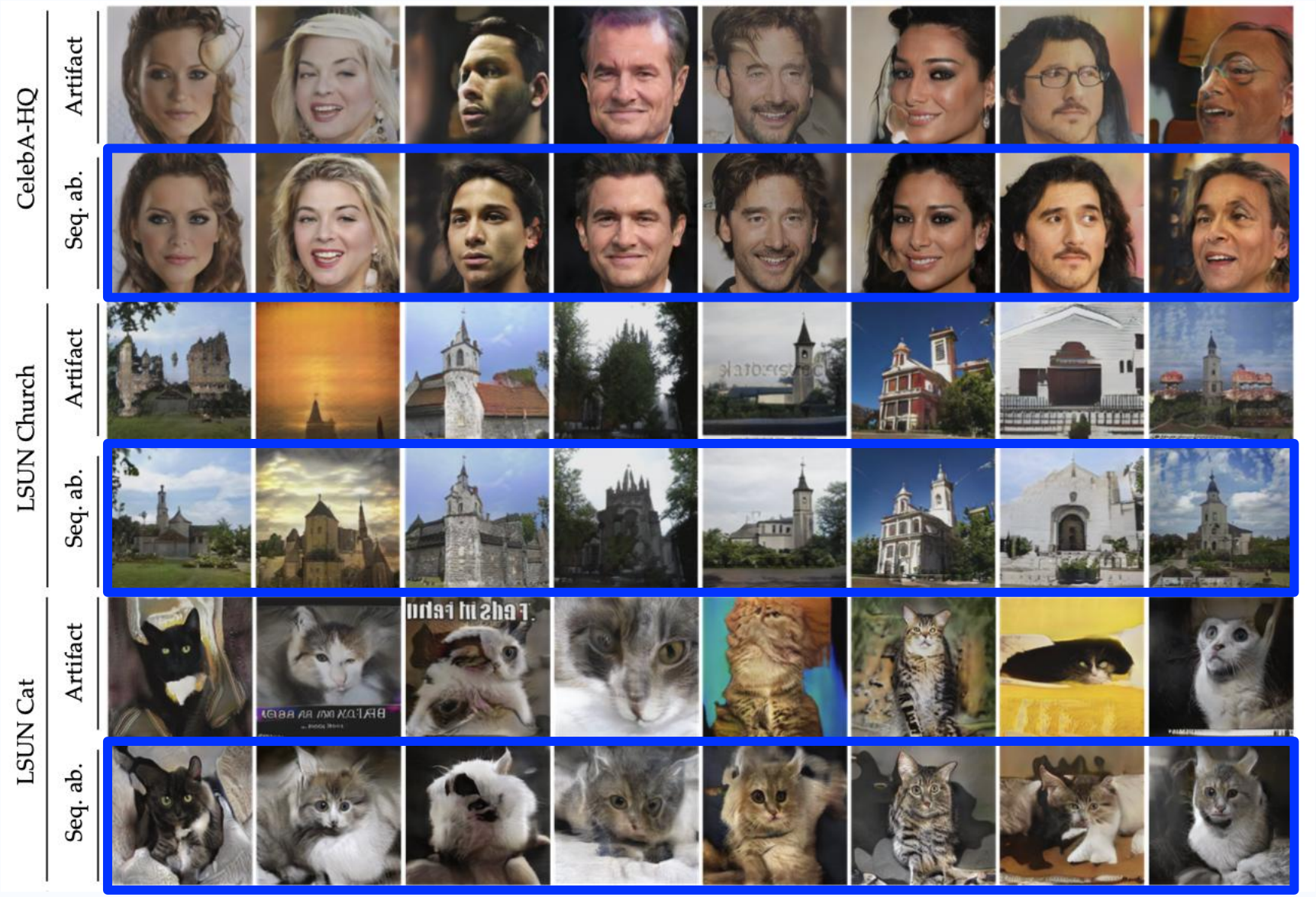
$\underline{f}^{i}_{\cdot,R}$: A set of neurons with a rate under R. (**LR-neurons**)

Algorithm 1 Sequential Ablation

Input: Generator G, A latent code $\mathbf{z} \in \mathcal{Z}$
Parameter: All Layer index list **L**, Layer index list **M** for ablation, Activation rate R
Output: Corrected Image I_{corr}

Let $I \leftarrow \mathbf{z}$.
for i in **L** **do**
  $f^i \leftarrow G^{(i-1):i}(I)$
  if i in **M** **then**
    $[\underline{f}^{i}_{\cdot,R}(I)]^{+} \leftarrow 0$
  else
    do not anything
  end if
  $I \leftarrow f^i$
end for
$I_{corr} \leftarrow I$
return I_{corr}

Automated Correction of Deep Neural Networks – Results



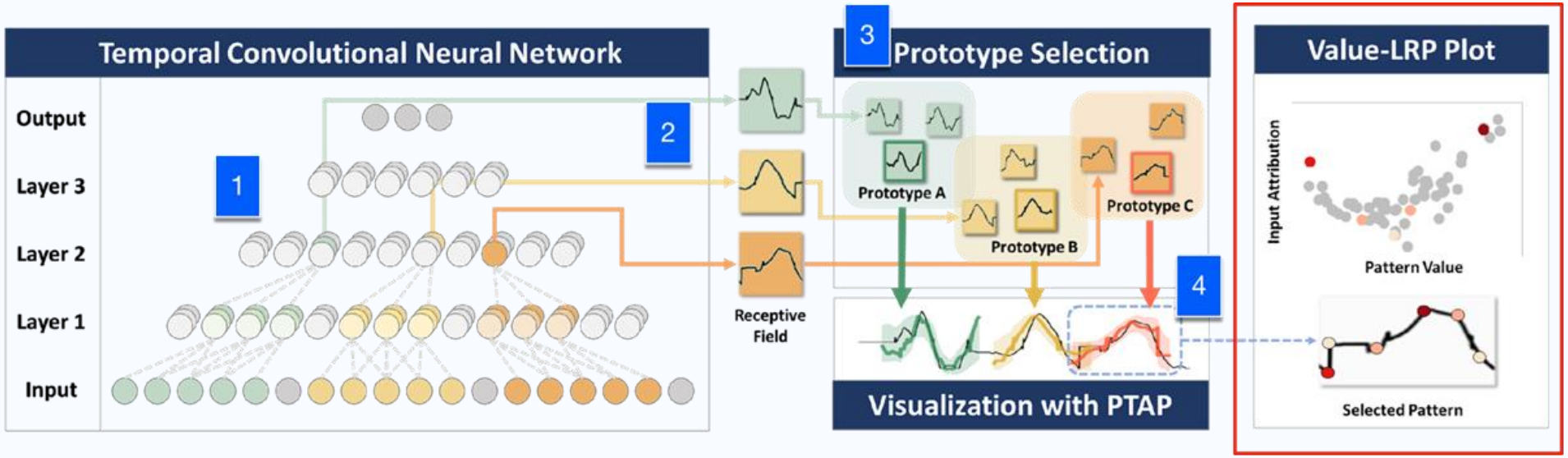
Defective parts in Artifact images are well corrected by Sequential Ablation!

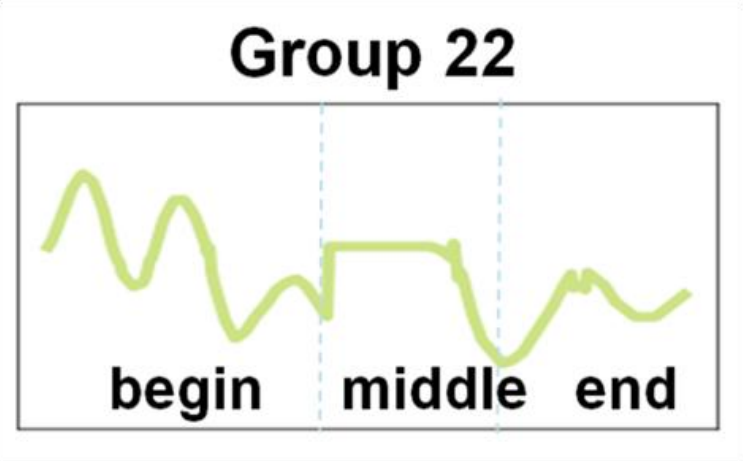
EMPOWERING EXPERTISE
iNEEJI
KAIST

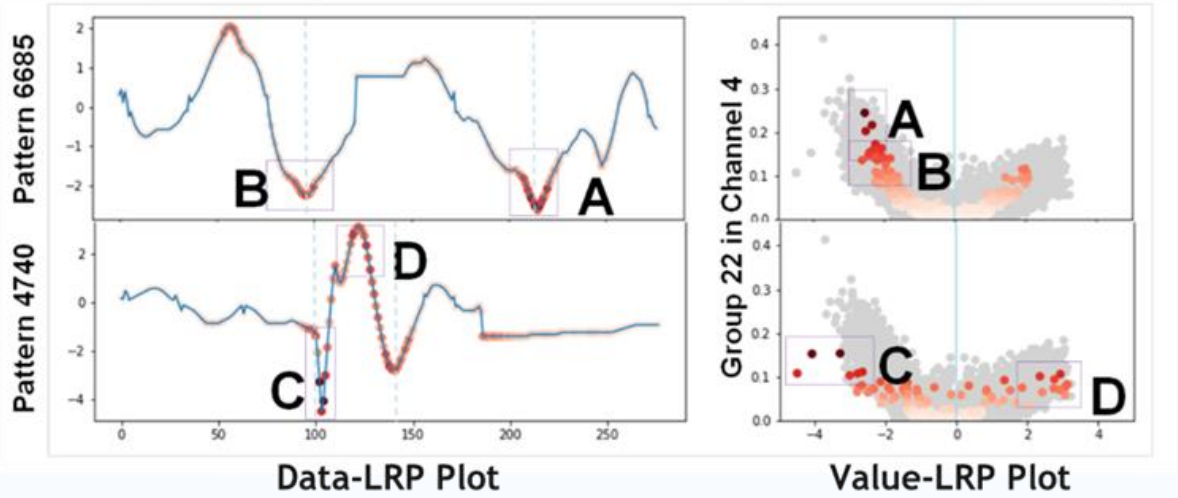
Q5: How can we analyze deep temporal neural network models?

*Equal contribution

Understanding of Deep Temporal Neural Networks







S. Cho*, W. Chang*, G. Lee and J. Choi, Interpreting Internal Activation Patterns in Deep Temporal Neural Networks by Finding Prototypes, KDD-21. * Equal contribution

Q6: Can we improve the input attribution method by finding the better path?

*Equal contribution

Distilled Gradient Aggregation:
Purify Features for Input Attribution in the Deep Neural Network

[G. Jeon*, H. Jeong* and J. Choi, Distilled Gradient Aggregation: Purify Features for Input Attribution in the Deep Neural Network, NeurIPS-22]

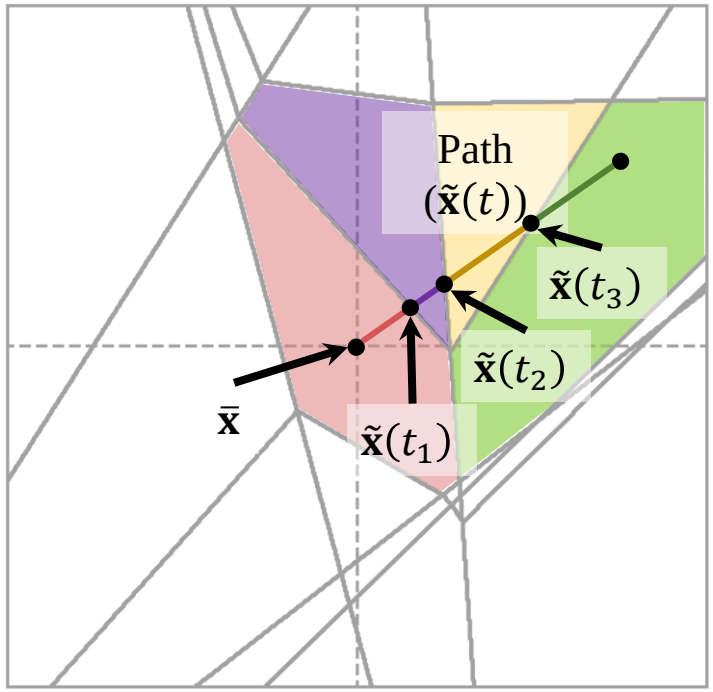
*Equal contribution

Interpreting Path Integration as Aggregation

- Observing that the path of IG crosses several linear regions, the integration can b e rewritten as the sum of integrals in each crossing region, R_i,

$$(\mathbf{x}-\bar{\mathbf{x}}) \odot \int_0^1 \nabla_{\tilde{\mathbf{x}}(t)} F(\tilde{\mathbf{x}}(t))dt = (\mathbf{x}-\bar{\mathbf{x}}) \odot \left(\int_0^{t_1} \mathbf{w}_1\, dt + \int_{t_1}^{t_2} \mathbf{w}_2\, dt + \int_{t_2}^{t_3} \mathbf{w}_3\, dt + \int_{t_3}^{1} \mathbf{w}_4\, dt \right)$$

where each term only depends on local function.



- We reformulate the global attribution by aggregating the local attribution over the sequence of regions $\mathcal{R}$

$$\phi^{global}(\mathbf{x}) = \sum_{R_i \in \mathcal{R}} \phi^{R_i}(\mathbf{x})$$

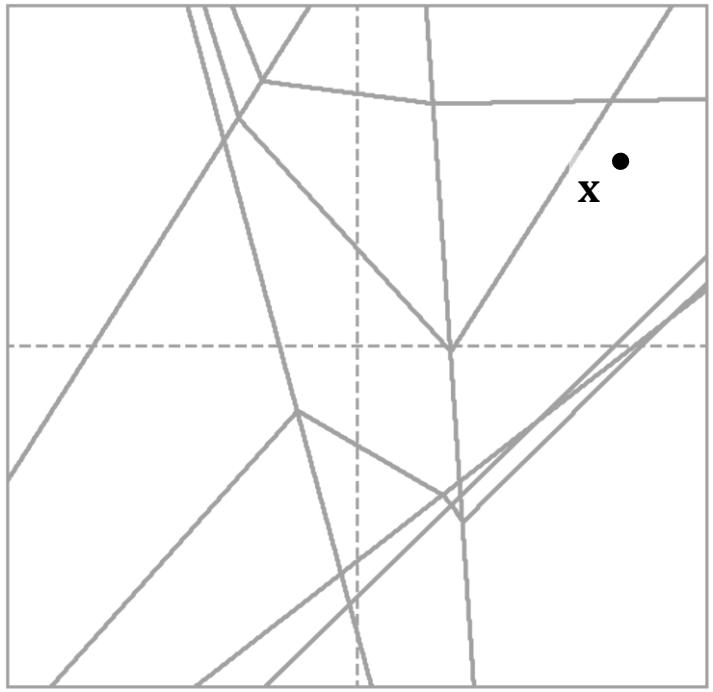
where ϕ^{R_i} is local attribution at region R_i.

- Then this aggregation consists of
  1) Selecting the **local attribution method** ⇒ Input*grad, FullGrad
  2) Collecting **the sequence of regions** $\mathcal{R}$

Not cumulating through path, but collecting regions to interpret

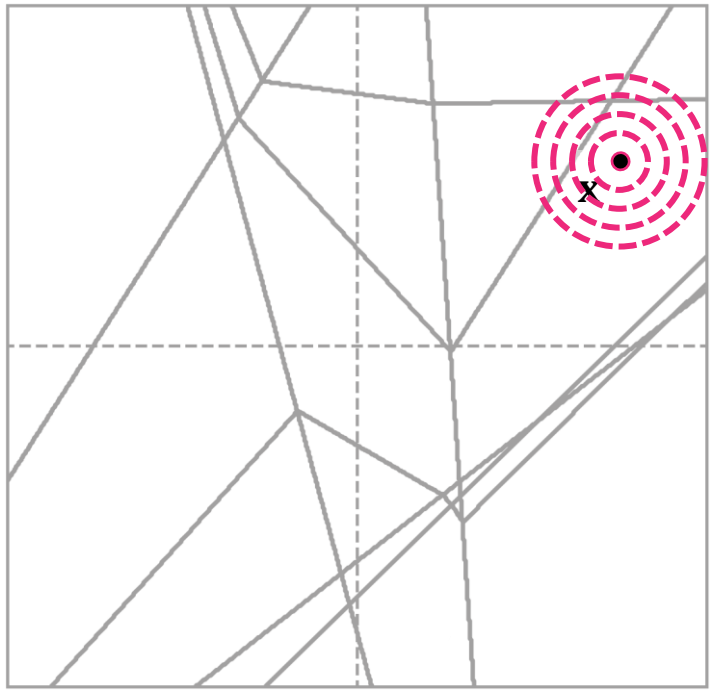
Vulnerability with local attributions

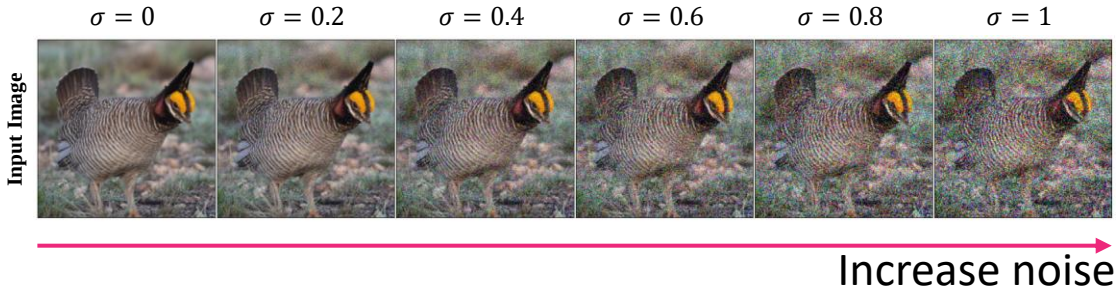
- Computing attribution only using the local information is vulnerable in that **small perturbation on input may cause big difference** in the attribution.



Vulnerability with local attributions

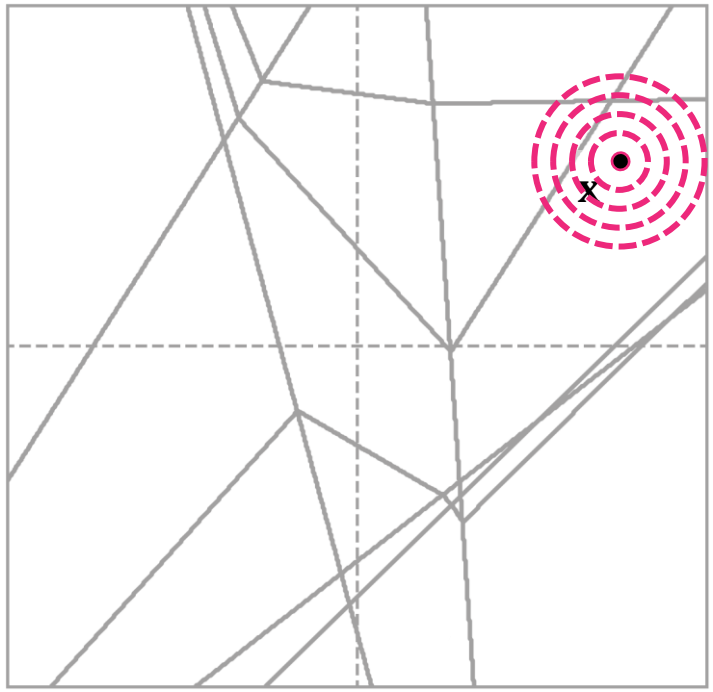
- Computing attribution only using the local information is vulnerable in that **small perturbation on input may cause big difference** in the attribution.
- For example, if we add Gaussian noise $\boldsymbol{\epsilon} \sim N(0, \sigma\mathbf{I})$ to the input where the prediction is maintained, $F(\mathbf{x}) = F(\mathbf{x} + \boldsymbol{\epsilon})$, it highly affects the local attribution.

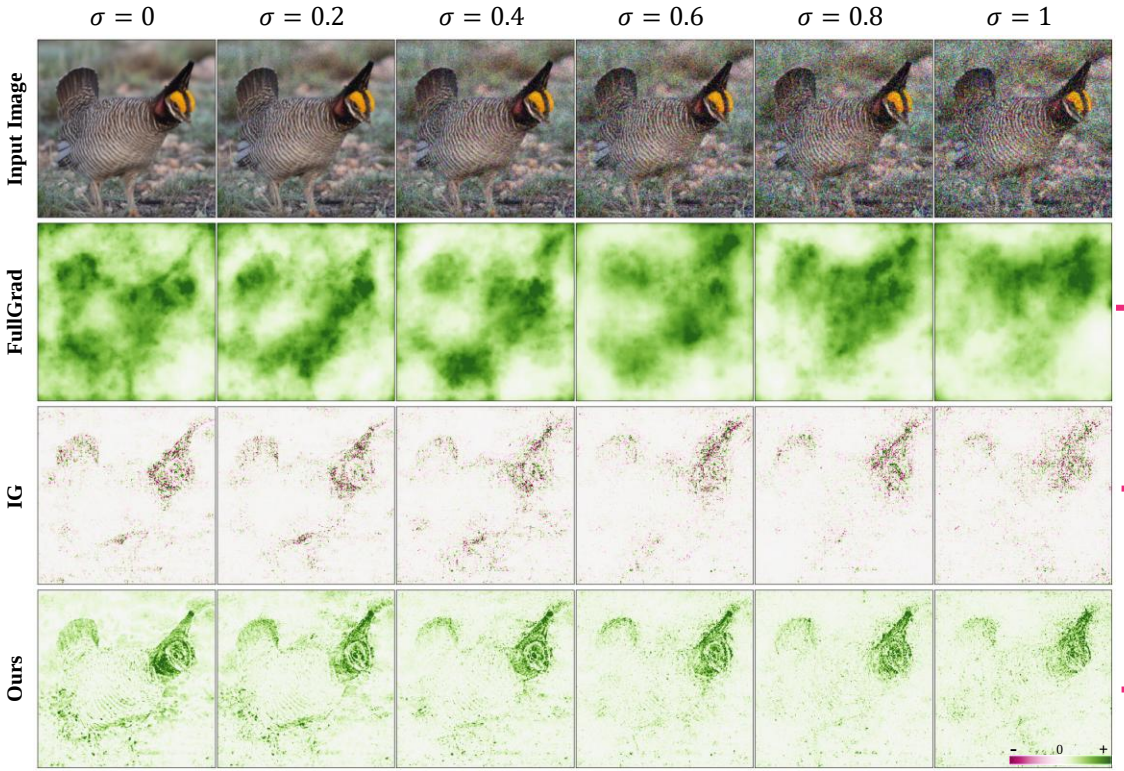




Vulnerability with local attributions

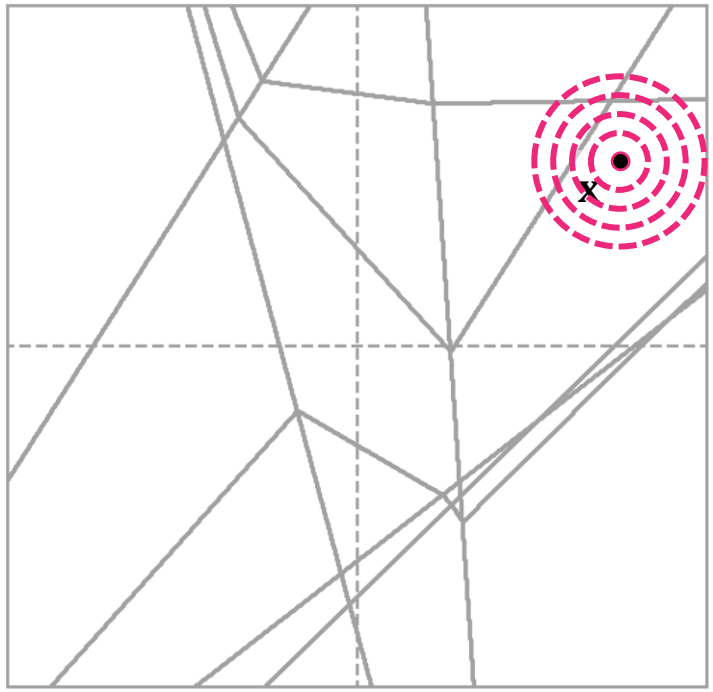
- Computing attribution only using the local information is vulnerable in that **small perturbation on input may cause big difference** in the attribution.
- For example, if we add Gaussian noise $\boldsymbol{\epsilon} \sim N(0, \sigma\mathbf{I})$ to the input where the prediction is maintained, $F(\mathbf{x}) = F(\mathbf{x} + \boldsymbol{\epsilon})$, it highly affects the local attribution.

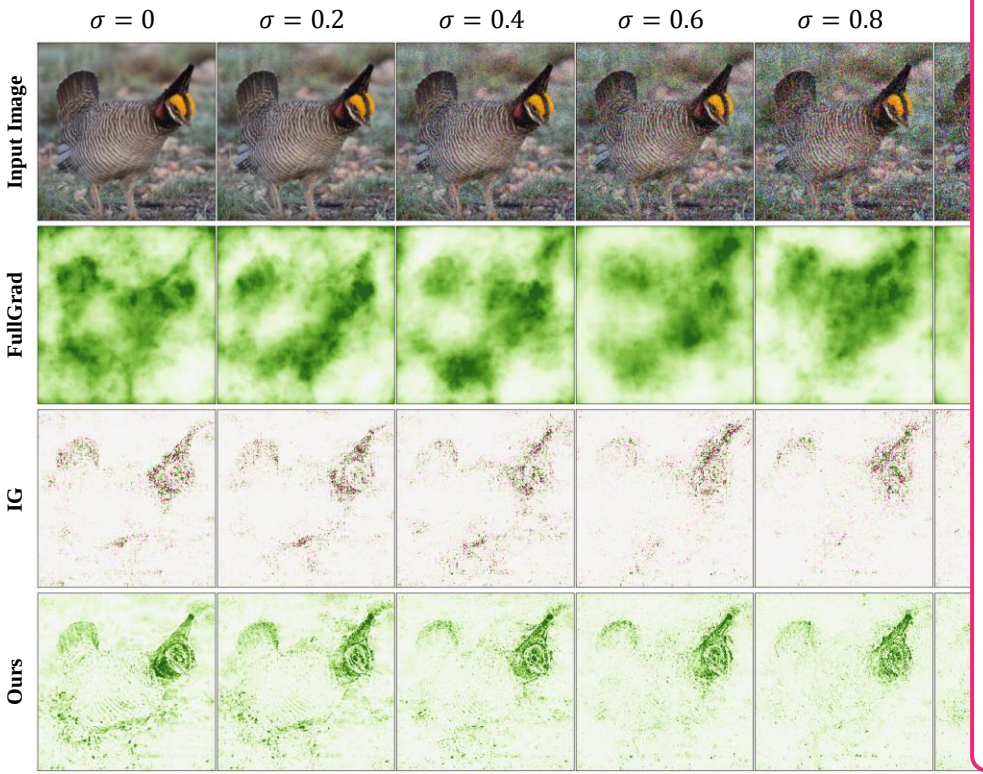


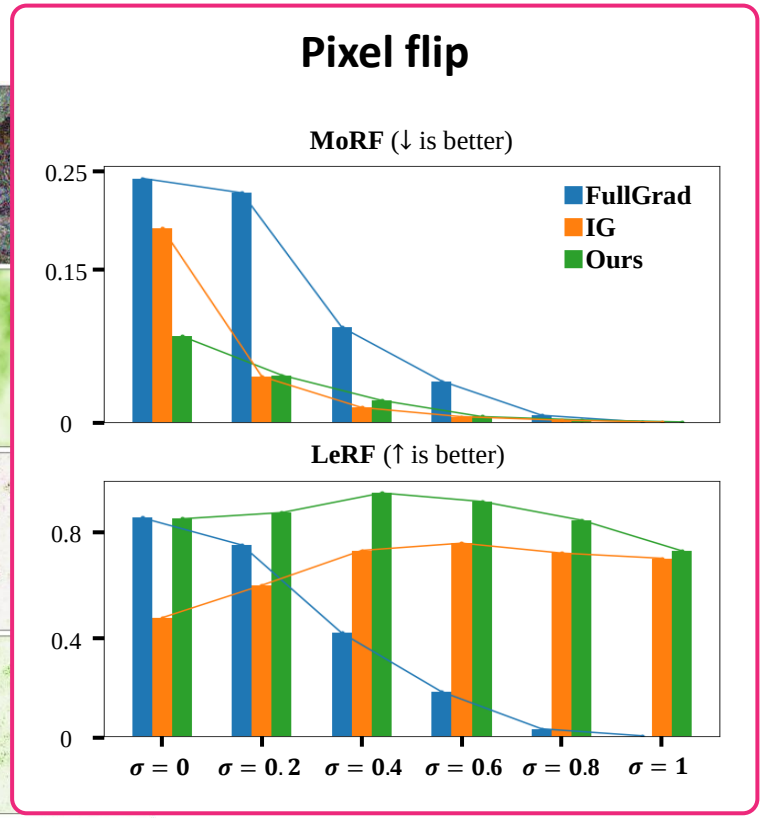


Vulnerability with local attributions

- Computing attribution only using the local information is vulnerable in that **small perturbation on input may cause big difference** in the attribution.
- For example, if we add Gaussian noise $\boldsymbol{\epsilon} \sim N(0, \sigma\mathbf{I})$ to the input where the prediction is maintained, $F(\mathbf{x}) = F(\mathbf{x} + \boldsymbol{\epsilon})$, it highly affects the local attribution.







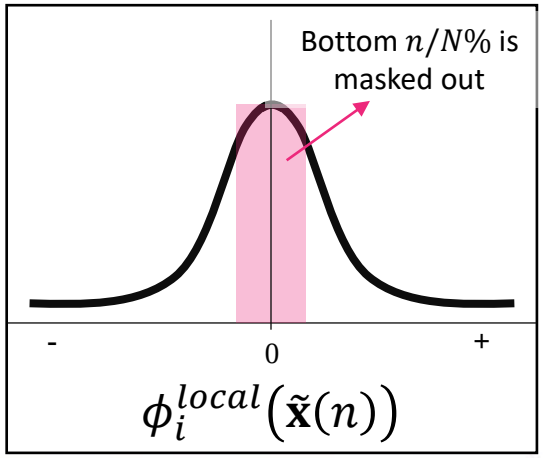
Distilling with Weak Contributors (WC) mask

- We propose to gradually **distill the irrelevant features** from input. By removing impurities, we **encourage more attribution to be assigned to important features**.

- The value of WC mask $\mathcal{M}_i^{WC}(\tilde{\mathbf{x}}(n), n)$ is set to 0 if $\left|\phi_i^{local}\left(\tilde{\mathbf{x}}(n)\right)\right|$ is less than bottom $n/N\%$.

- Distillation step with WC mask then defined as

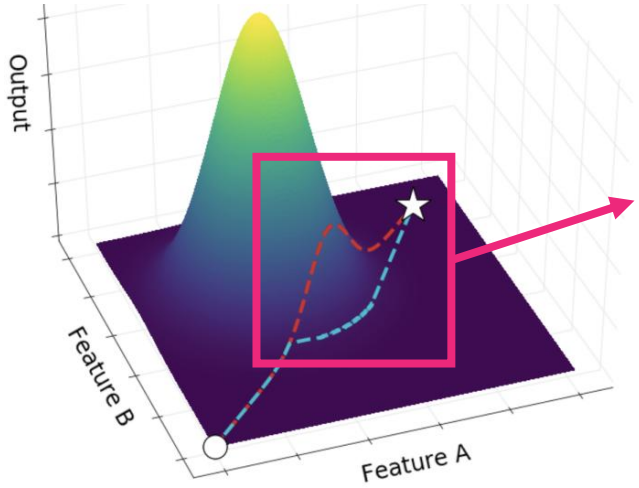
$$\tilde{\mathbf{x}}(n+1) = \mathcal{M}^{WC}(\tilde{\mathbf{x}}(n), n) \odot \mathbf{x}$$



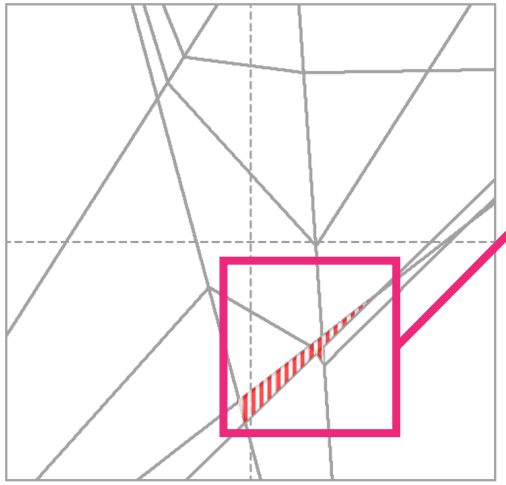
Generated distillation sequence with WC

Regularizing distillation with Extreme Positive Contributors (EPC)

- Distillation sequence using only WC mask may fail to distill the irrelevant features and **saturates when high local attribution is accidently assigned**.



Need to avoid transient gradients



Usually observed in narrow linear regions

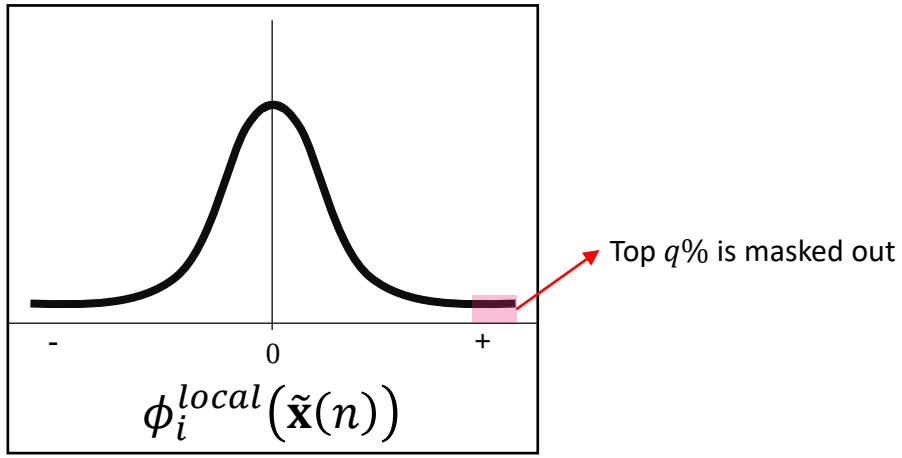
Using only WC

Regularizing distillation with Extreme Positive Contributors (EPC)

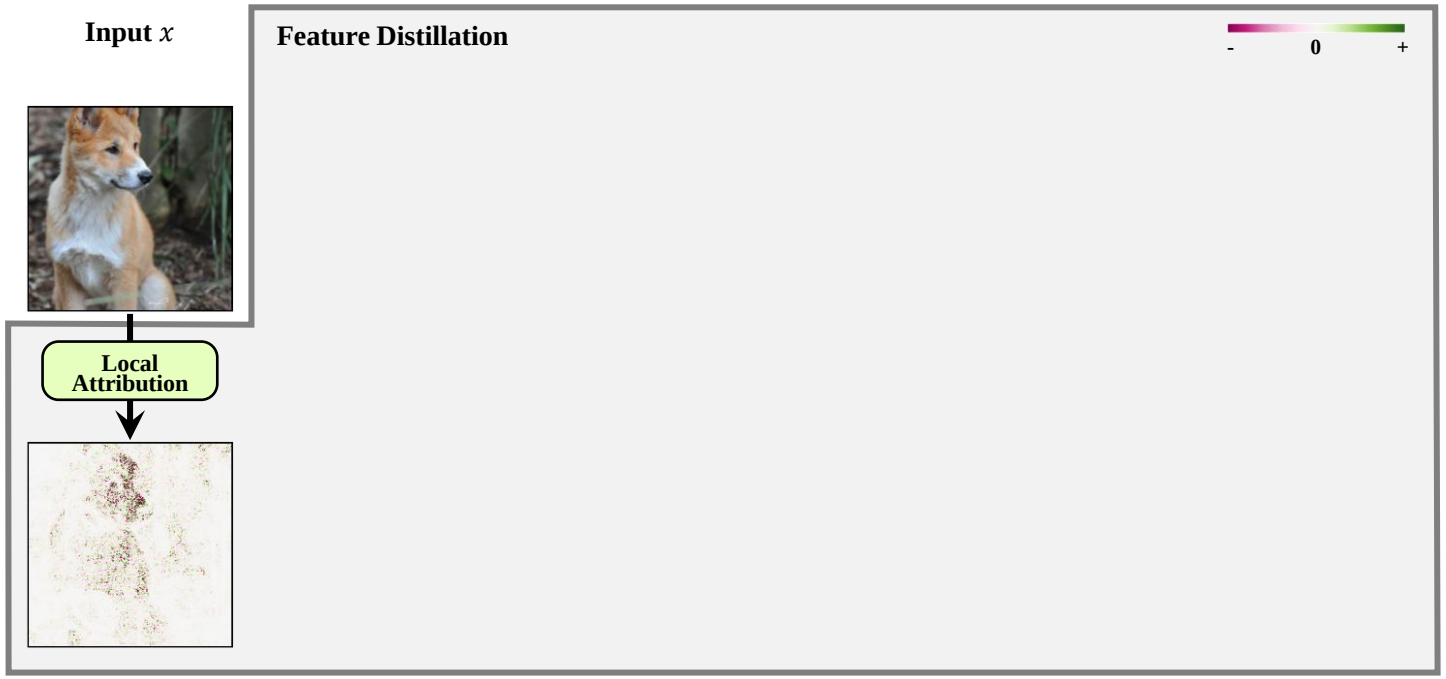
- Distillation sequence using only WC mask may fail to distill the irrelevant features and **saturates when high local attribution is accidently assigned**.
- We avoid the saturation by regularizing the **Extreme Positive Contributor**.
- The value of EPC mask $\mathcal{M}_i^{EPC}(\tilde{\mathbf{x}}(n), q)$ is set to 0 if $\phi_i^{local}\big(\tilde{\mathbf{x}}(n)\big)$ exceeds top $q\%$
- Combined with WC mask, we propose following distillation step,

$$\tilde{\mathbf{x}}(n+1) = \left(\frac{n}{N}\mathcal{M}^{WC}(\tilde{\mathbf{x}}(n), n) + \left(1 - \frac{n}{N}\right)\mathcal{M}^{EPC}(\tilde{\mathbf{x}}(n), q)\right) \odot \mathbf{x}$$

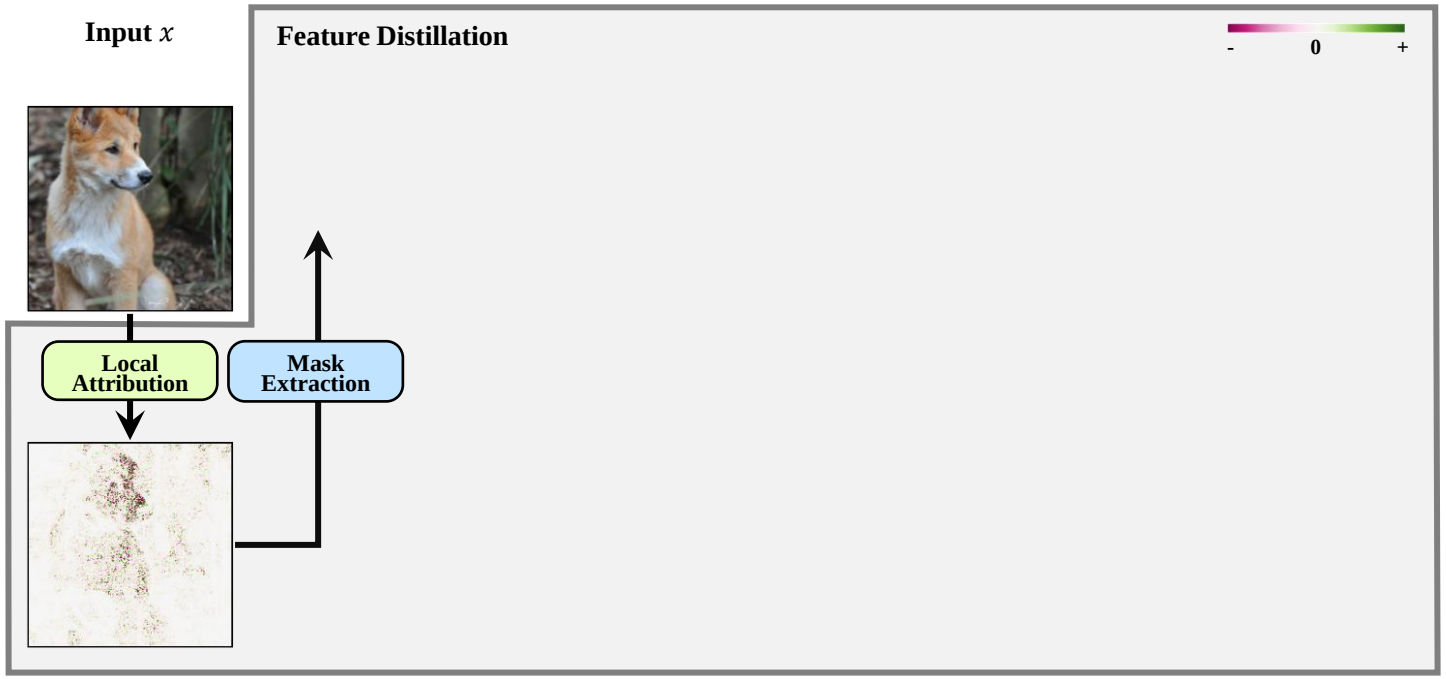
Using only WC

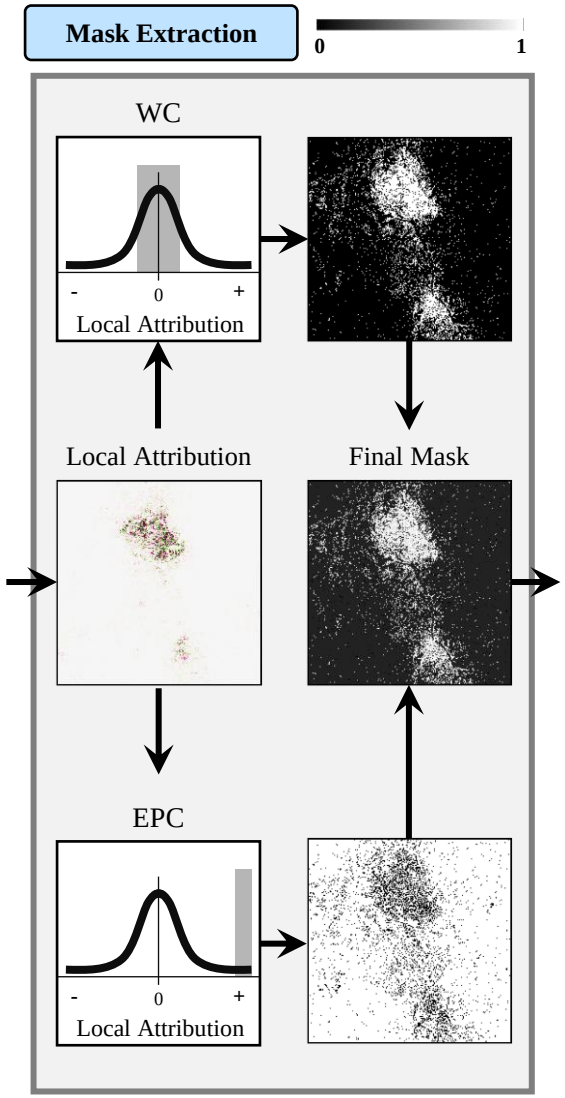


Overall structure of DGA

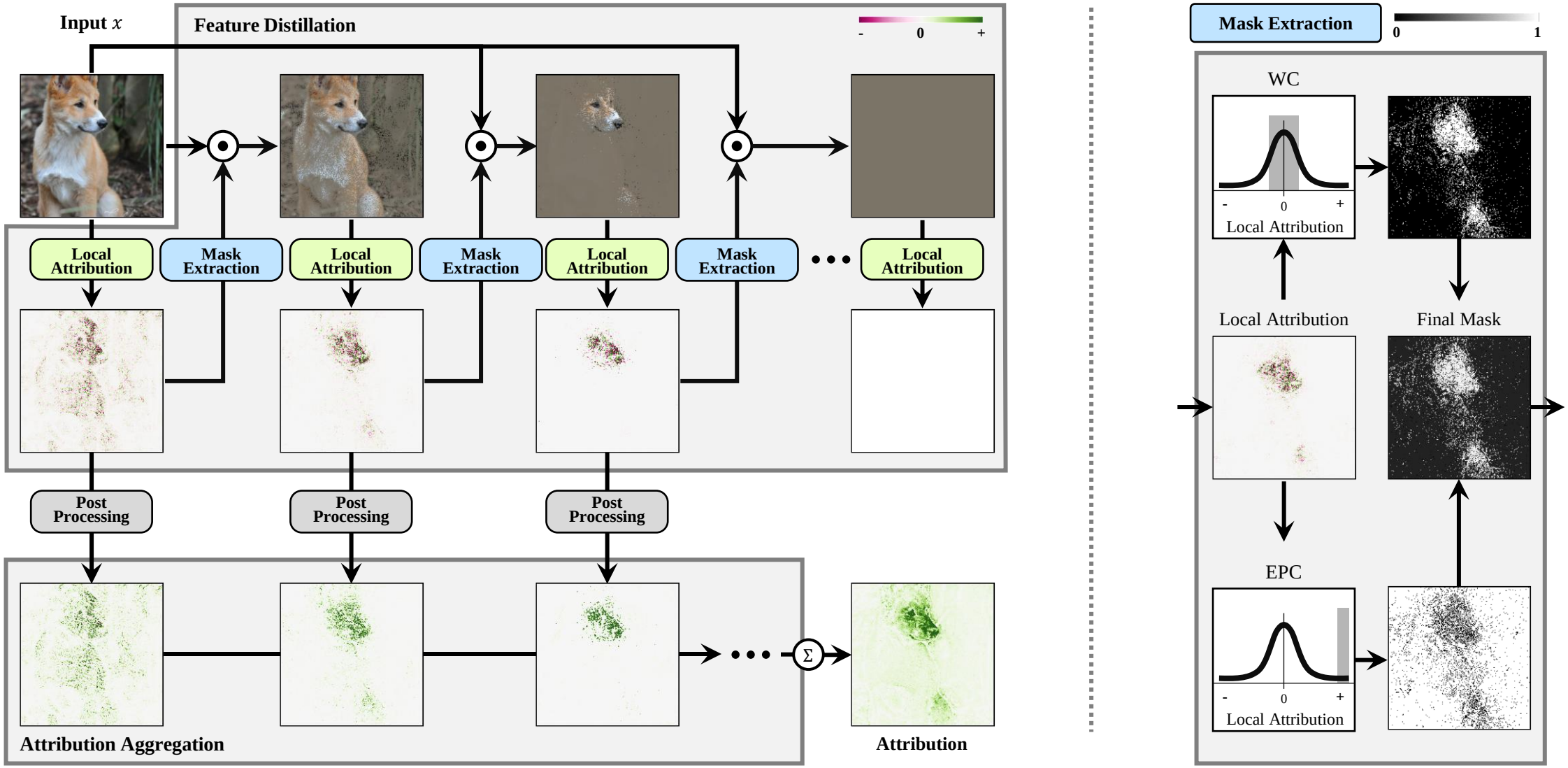


Overall structure of DGA





Overall structure of DGA

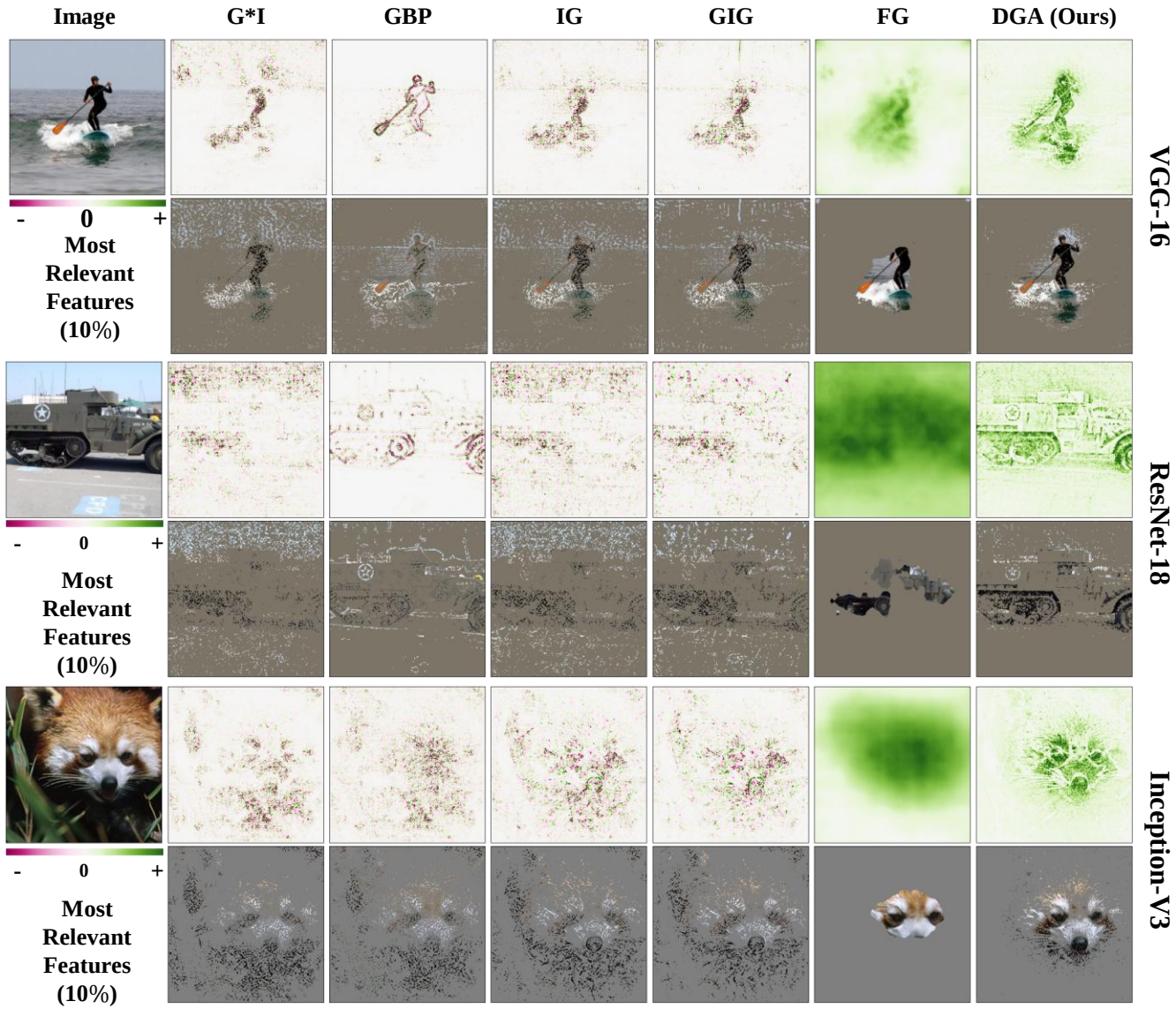


Qualitative comparison with existing methods

- We compare our method with several gradient-based methods,
  - Grad*Input (G*I)
  - GuidedBackPropagation (GBP)
  - Integrated Gradient (IG)
  - Guided IG (GIG)
  - FullGrad (FG)

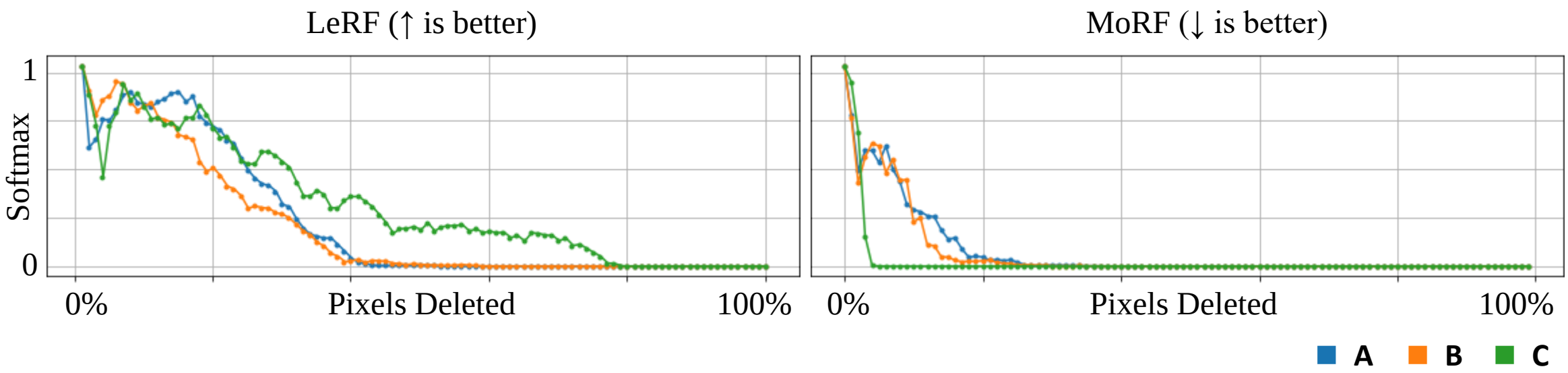
- For visualization
  1) the attribution heatmap
  2) top 10% input features

" DGA shows more **object aligned, less noisy and high-definition** results. "



Quantitative evaluation – Pixel Flip

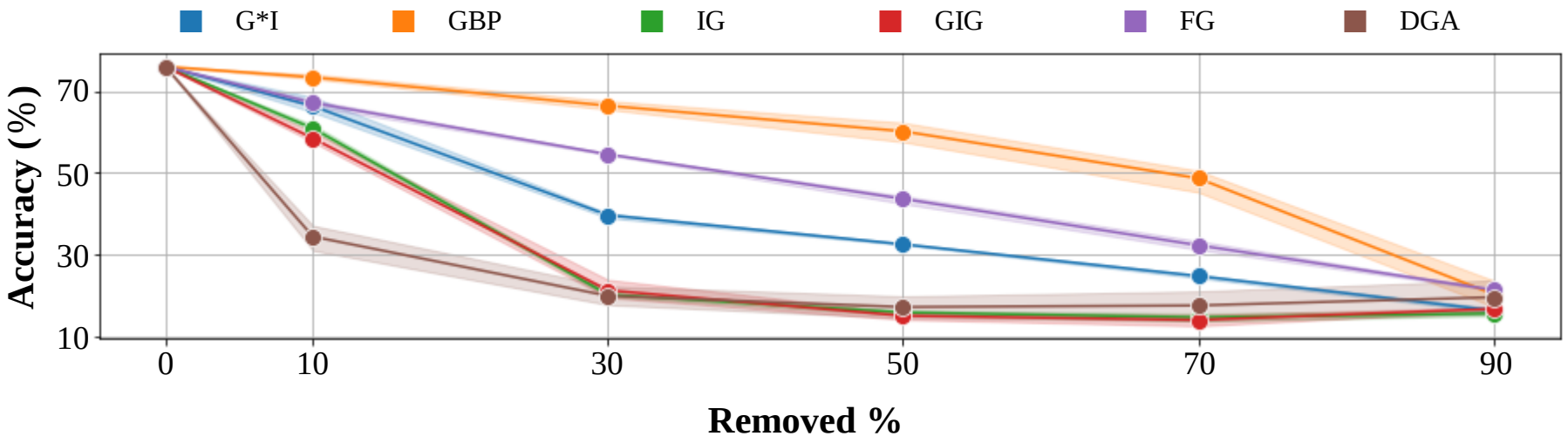
- Pixel Flip **evaluates** the **relationship** between the **attributions** and the **predictions**.
- Measure Area-Under-Curve (AUC) while removing pixel in order of attributions.
  - Least-Relevant First (LeRF) removes low to high attribution value
    Prediction should decrease less ⇒ **AUC should be large** ↑
  - Most-Relevant First (MoRF) removes high to low attribution value
    Prediction should decrease more ⇒ **AUC should be small** ↓



Quantitative evaluation – Pixel Flip

- Pixel Flip **evaluates** the **relationship** between the **attributions** and the **predictions**.
- Measure Area-Under-Curve (AUC) while removing pixel in order of attributions.
  - Least-Relevant First (LeRF) removes low to high attribution value
    Prediction should decrease less ⇒ **AUC should be large** ↑
  - Most-Relevant First (MoRF) removes high to low attribution value
    Prediction should decrease more ⇒ **AUC should be small** ↓

		Grad*Input	GBP	IG	FullGrad	GuidedIG	**DGA**
LeRF (↑ is better)	VGG-16	0.078	0.113	0.096	0.415	0.110	**0.434**
	ResNet-18	0.114	0.145	0.158	0.448	0.185	**0.533**
	Inception-V3	0.171	0.162	0.243	0.558	0.255	**0.691**
MoRF (↓ is better)	VGG-16	0.045	0.094	0.036	0.110	0.029	**0.023**
	ResNet-18	0.050	0.124	0.038	0.131	0.029	**0.019**
	Inception-V3	0.105	0.145	0.066	0.175	0.061	**0.041**

Quantitative evaluation – RemOve-And-Retrain (ROAR)

- ROAR **evaluates** if the attribution captures the important features in model training.
- ROAR modifies the training set by removing pixels with high attribution and retrain.
- When the same portion of image is removed, performance should drop much.