

Explainable AI (XAI)

Interpretability for Deep Learning

Taegyun Jeon (SI Analytics)

Contents

- Explainable AI
- Introduction to Interpretability
- Interpretation of Deep Neural Networks
- Evaluations for Attribution Methods
- Conclusions

SI Analytics

- Profile

Company	SI Analytics Co., Ltd.	
Establishment	July 2 nd , 2018	
CEO	Taegyun Jeon	
Address	Daejeon, South Korea	
Website	www.si-analytics.ai	
Major Products	In-house Analytics Platform	AI analysis platform with independent operation
	LabelEarth™	Data Annotation Tool
	SIDLE™	Inference Engine for Deep Learning

- Satrec Initiative Group



SI Imaging Services (SIIS) is the exclusive worldwide distributor of Korean Government satellites: KOMPSAT-2, KOMPSAT-3, KOMPSAT-3A and KOMPSAT-5. With the unique constellation of very high resolution optical and SAR satellites, SIIS is providing the world-class satellite imagery to Earth observation communities.

SI Analytics (SIA) is a new startup providing artificial intelligence-based geospatial analytics on satellite/aerial imagery. With the cutting-edge deep analytics, SIA is helping Korean Government analyze vast amount of imagery and collect unprecedented intelligence: from military activities to illegal shipping.

Taegyun Jeon



- **Founder and Chief Executive Officer**
 - SI Analytics, Jul 2018 – Present



- **AI Team Leader**
 - Satrec Initiative, Sep 2016 – Jun 2018



- **Google Developer Experts – Machine Learning**
 - Google, Sep 2017 – Present



- **PhD – Machine Learning and Computer Vision**
 - Gwangju Institute of Science and Technology (GIST)



TensorFlow
Korea

- **Co-administrator – TensorFlow User Group Korea**
 - Korea, May 2017 – Oct 2018



Research Area (SI Analytics)



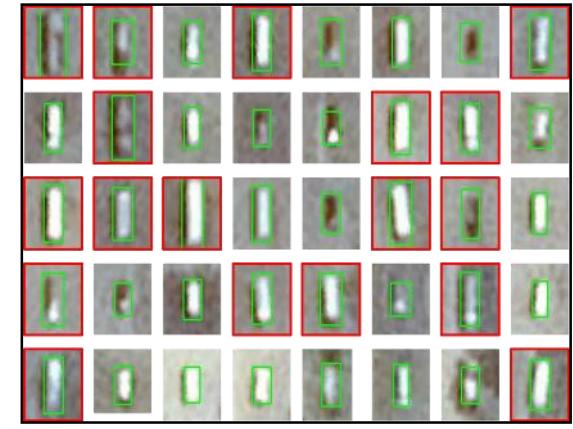
Object Detection (Aircraft & Fighter)



Object Detection (Aerial Imagery)



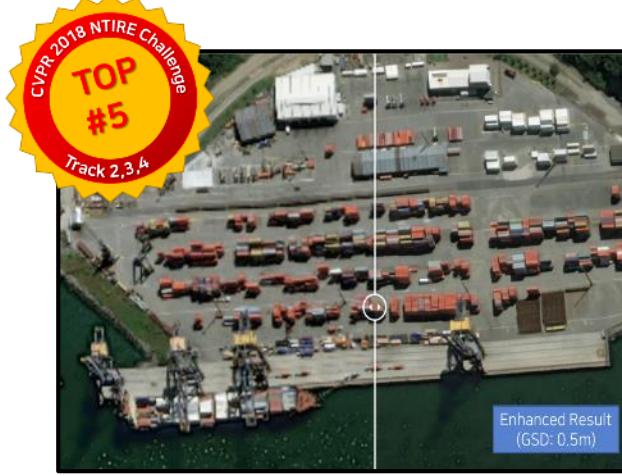
Ship Detection (EO/SAR)



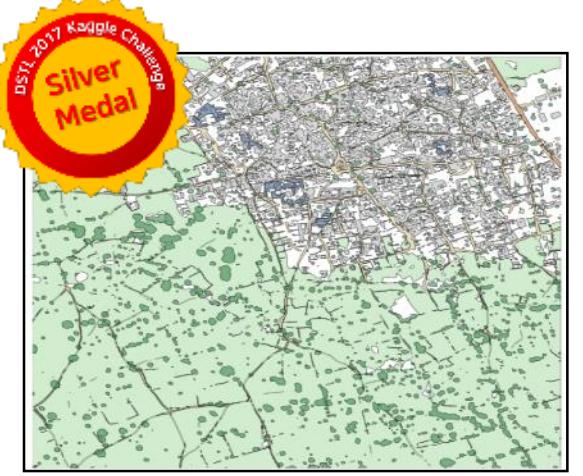
Vehicle Detection & Monitoring



Change Detection



Super-Resolution



Land Cover Classification



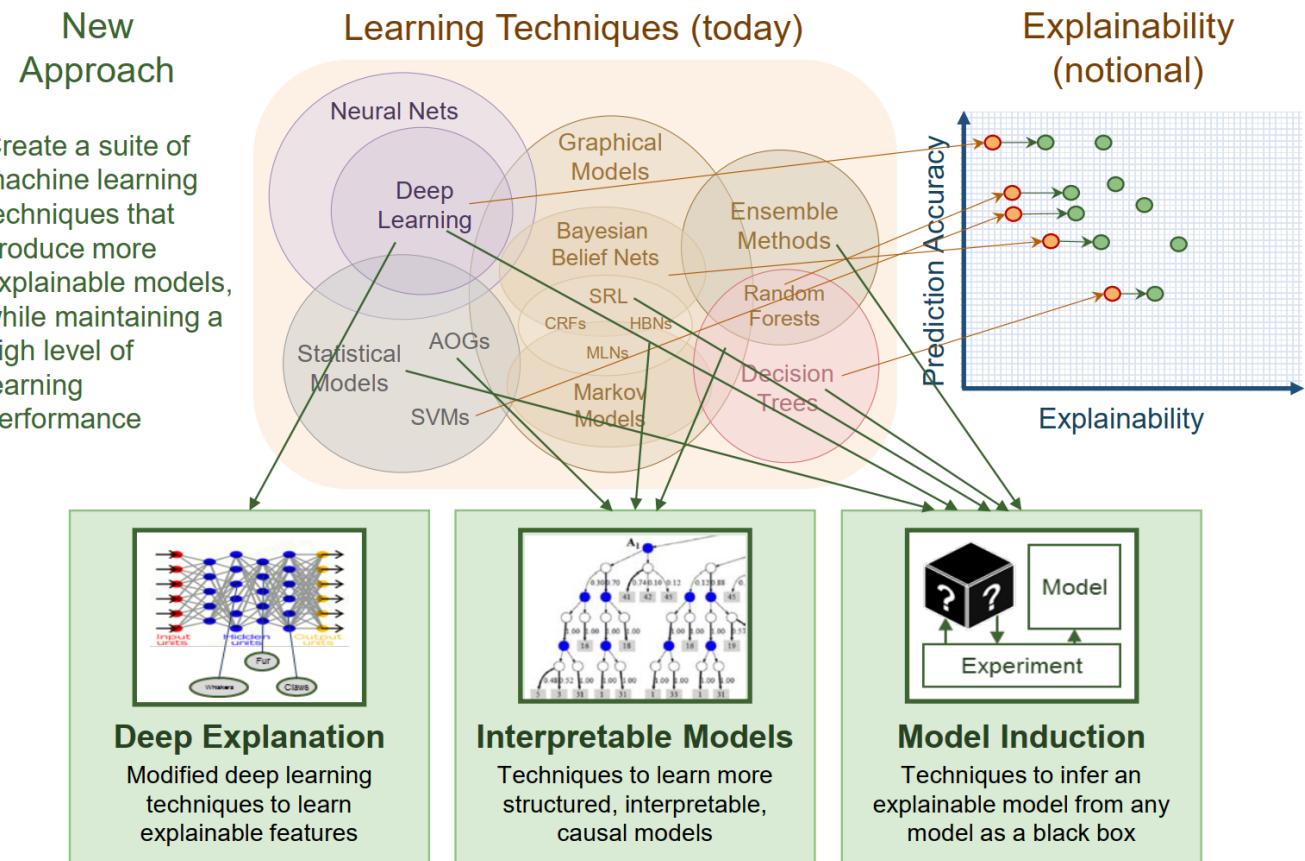
Explainable AI

01

Explainable AI (XAI)

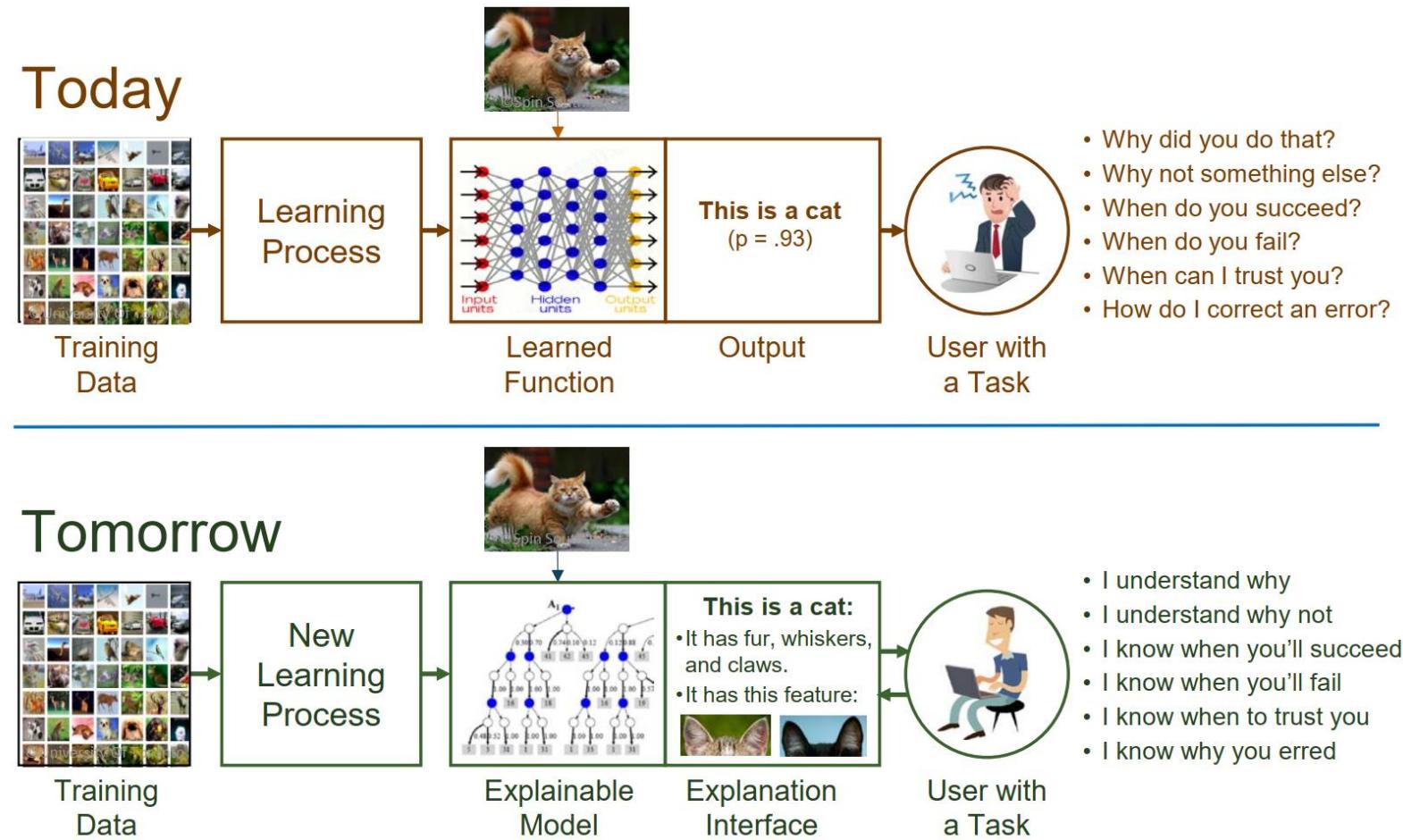
DARPA Research Program: XAI

- 딥러닝의 동작원리 설명
- 사용자가 신뢰할 수 있는 학습/추론 설명 모델 연구
- 모델 분석을 통한 특화된 모델 관리
- Explainable AI (XAI) 기술을 정의하고 2017년 상반기부터 연구 프로젝트 수행 중

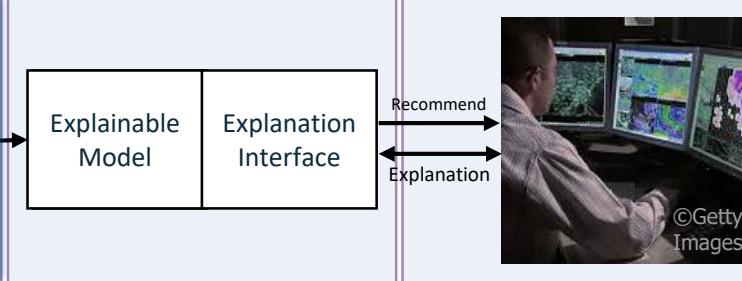
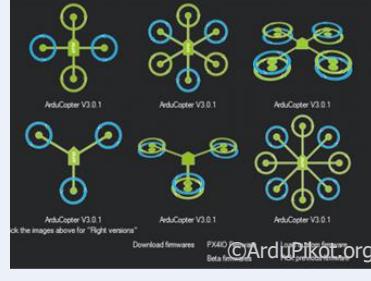


DARPA Research Program: XAI

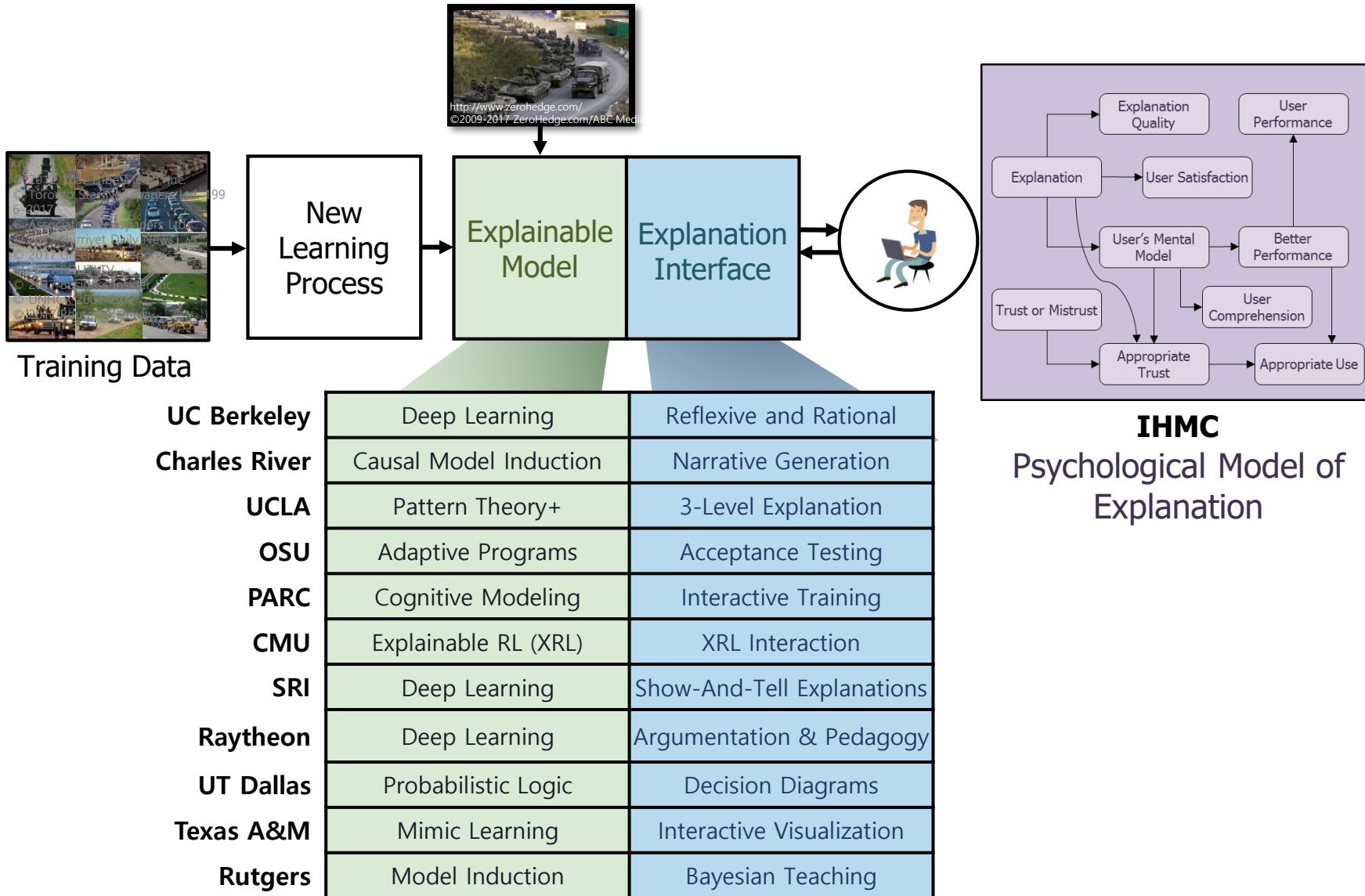
- 딥러닝의 내부동작 이해
 - 입력/출력의 인과관계
 - 경험에 근거한 판단이 중요한 도메인 (“국방”)
 - 학습이 제대로 되지 않은 이유
 - 다양한 지능이나 정보를 융합



Challenge Problems (DARPA)

	Learn a model ↓	Explain decisions ↓	Use the explanation ↓	
Data Analytics Classification Learning Task	 Multimedia Data		An analyst is looking for items of interest in massive multimedia data sets	
Autonomy Reinforcement Learning Task	Classifies items of interest in large data set	Explains why/why not for recommended items	Analyst decides which items to report, pursue	
	 ArduPilot & SITL Simulation		Operator decides which future tasks to delegate	An operator is directing autonomous systems to accomplish a series of missions

XAI Concept and Technical Approaches



02

Interpretability

What is Interpretability?

AlphaGo vs. Lee Sedol



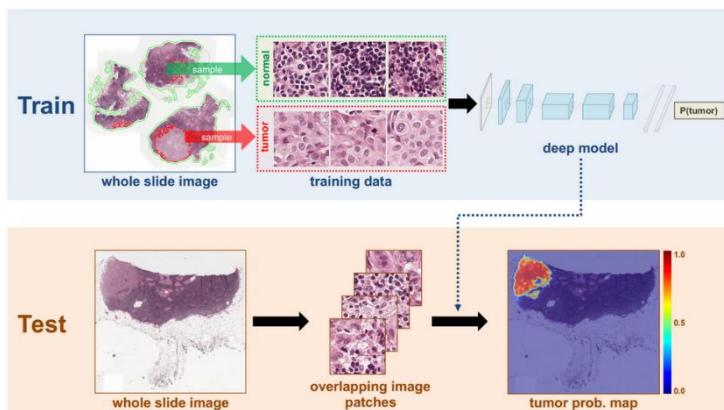
ImageNet Challenge



Self-driving Cars



Disease Diagnosis



Neural Machine Translation



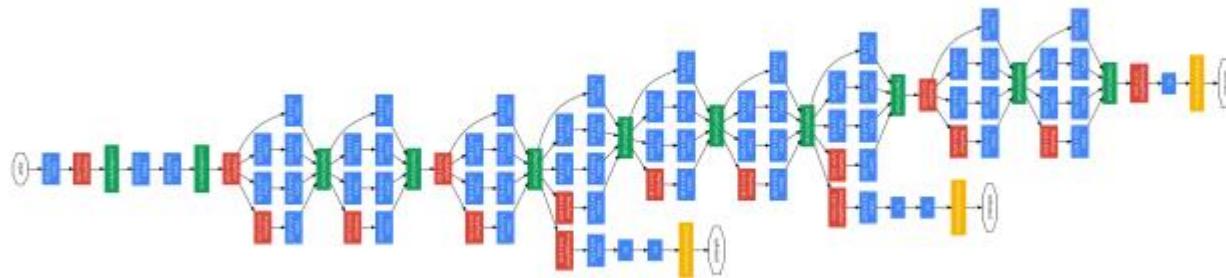
Defense

What is Interpretability?

Large Dataset

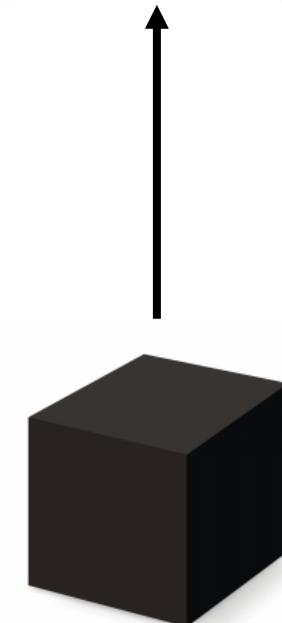
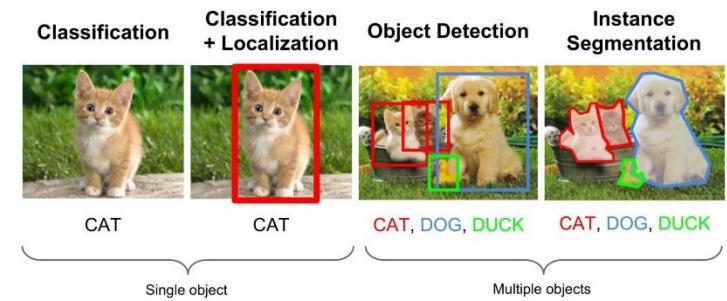


Computing Power



Deep Neural Networks

Task Solving



Implicit Information

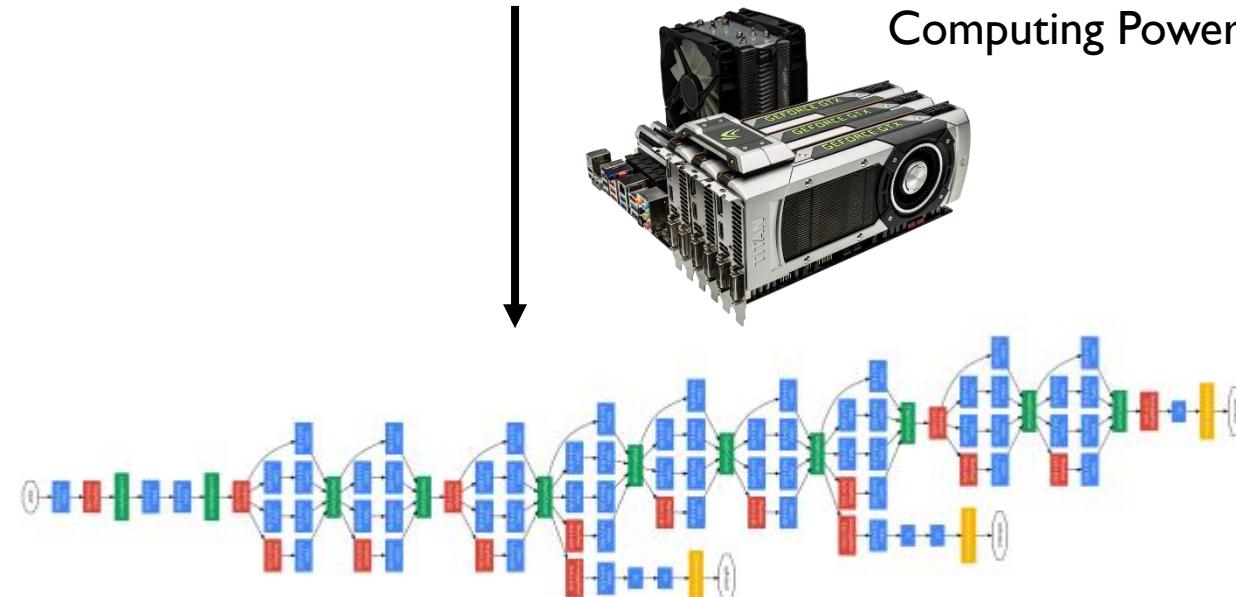
What is Interpretability?

Large Dataset



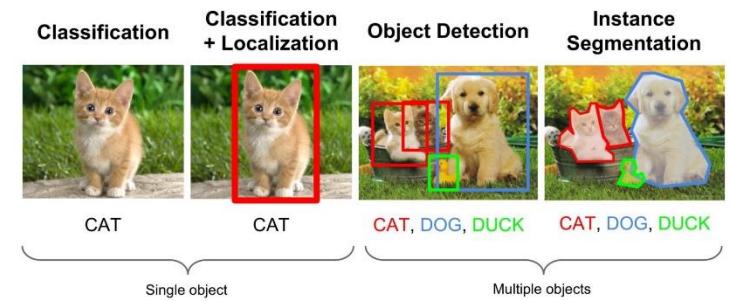
Interpretable Information

Computing Power



Deep Neural Networks

Task Solving



Implicit Information

Why Interpretability?

I. Verify that model works as expected

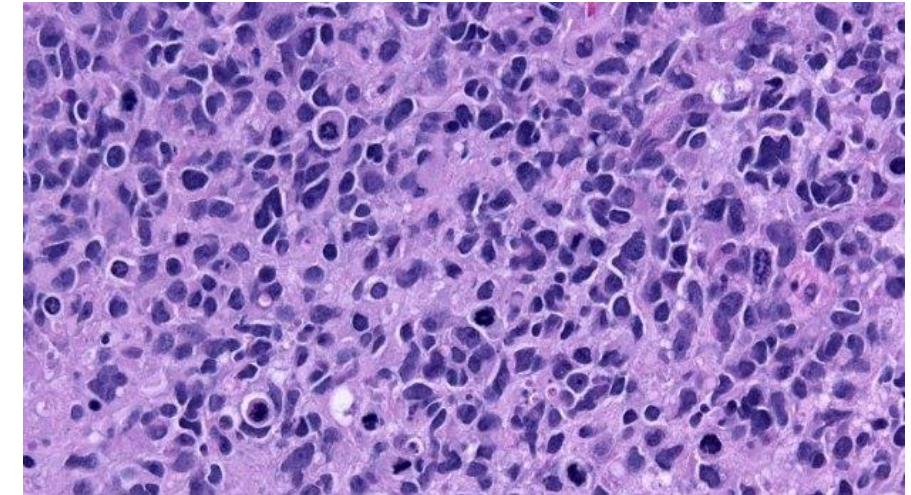
Wrong decisions can be costly and dangerous

Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian



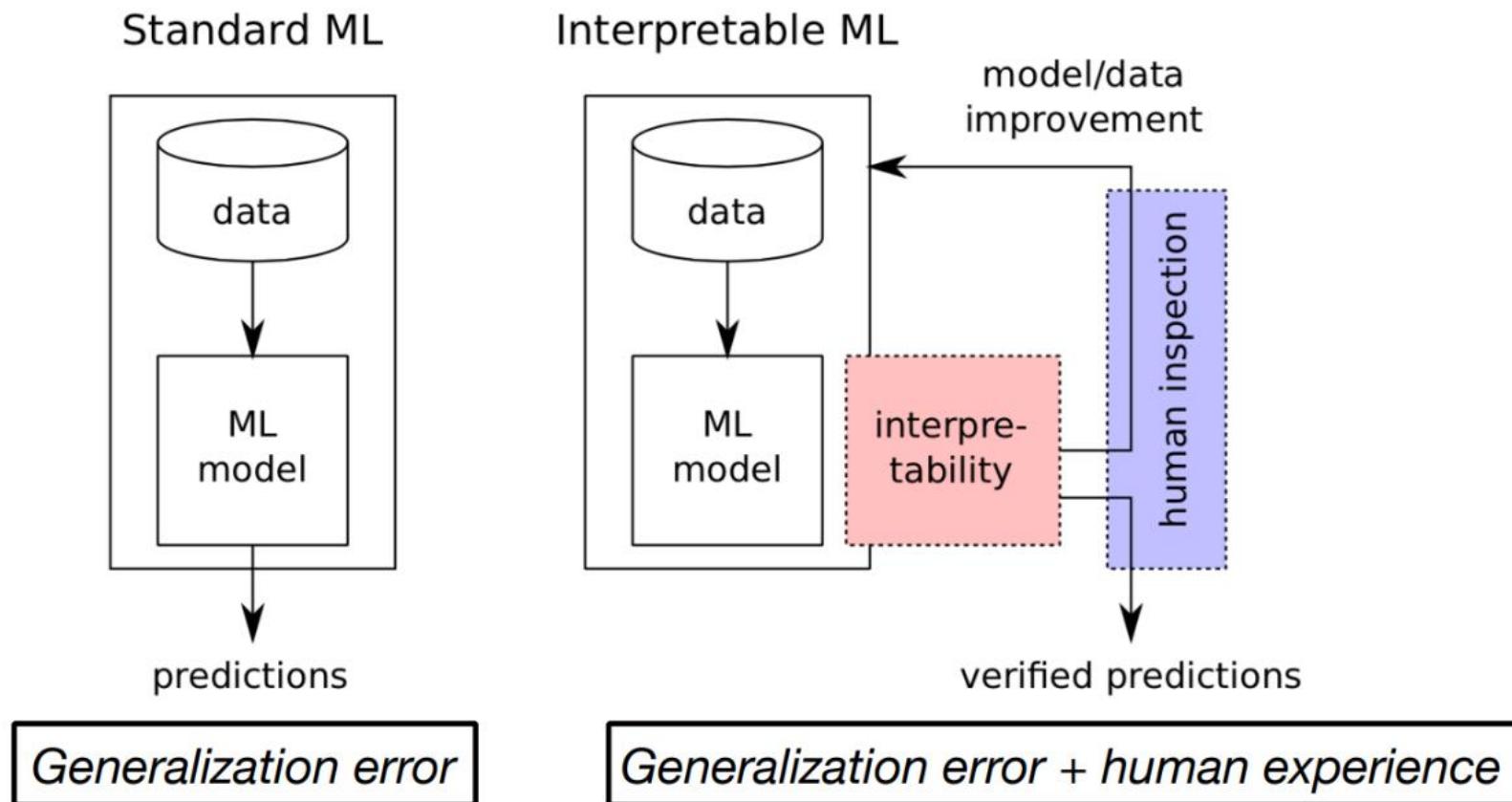
<https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe>

Disease Misclassification



Why Interpretability?

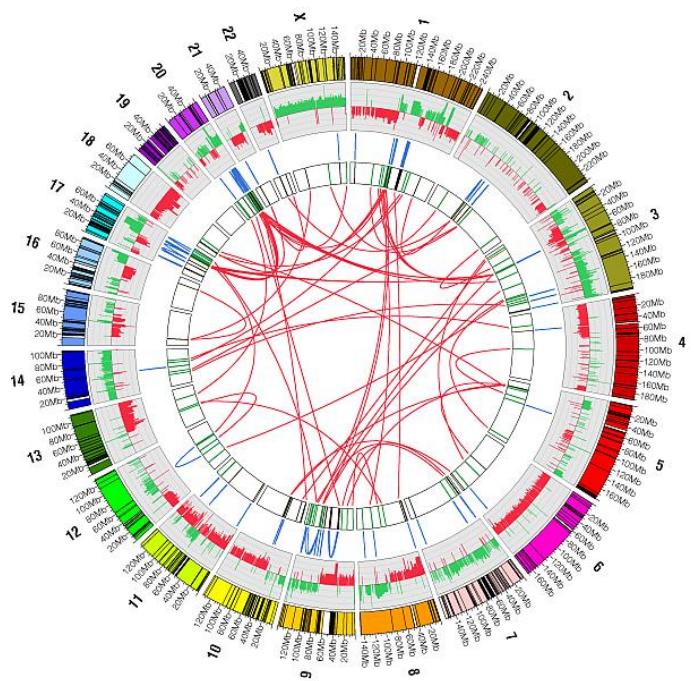
2. Improve / Debug classifier



Why Interpretability?

3. Make new discoveries

Learn about the physical / biological / chemical mechanisms



Learn about the human brain



Why Interpretability?

4. Right to explanation

“Right to be given an explanation for an output of the algorithm”

Ex.

- US Equal Credit Opportunity Act
- The European Union General Data Protection Regulation
- France Digital Republic Act

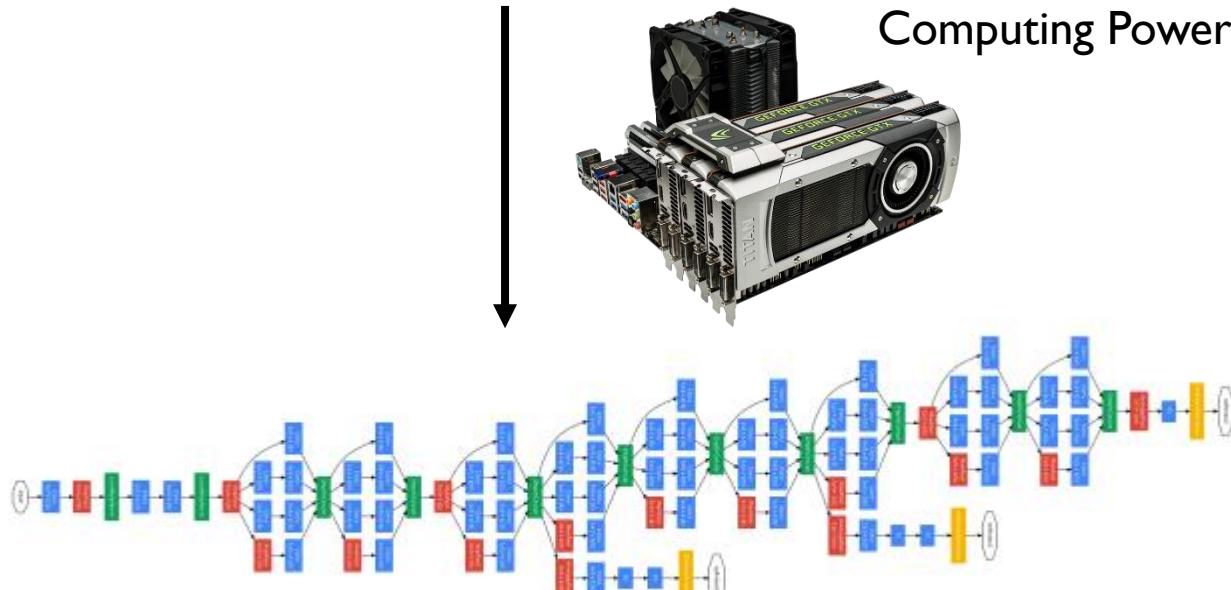
Back to Interpretability!

Large Dataset



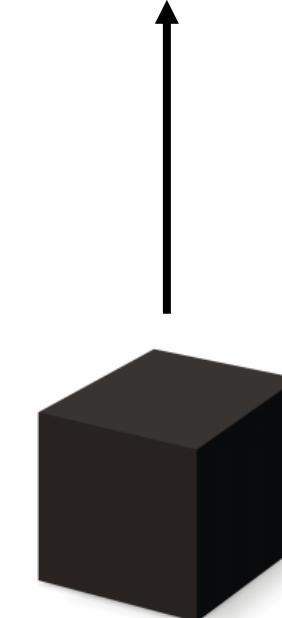
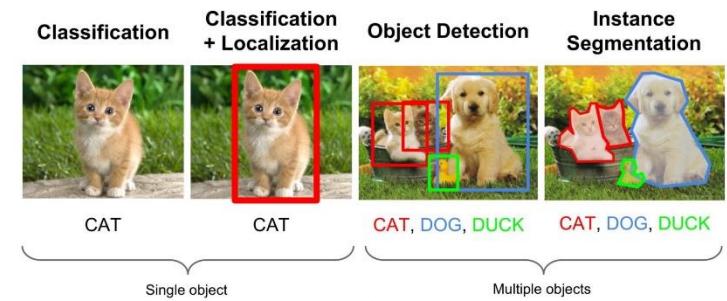
Interpretable Information

Computing Power



Deep Neural Networks

Task Solving



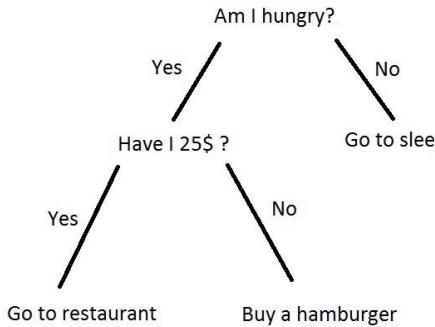
Implicit Information

Types of Interpretability in ML

Ante-hoc Interpretability

Choose an interpretable model and train it.

Ex.



Decision Tree

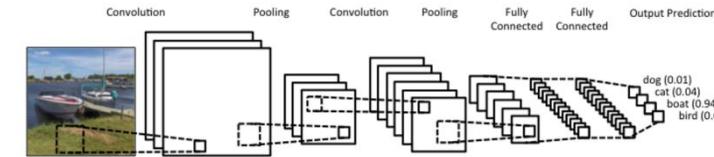
Problem.

Is the model expressive enough to predict the data?

Post-hoc Interpretability

Choose a complex model and develop a special technique to interpret it.

Ex.

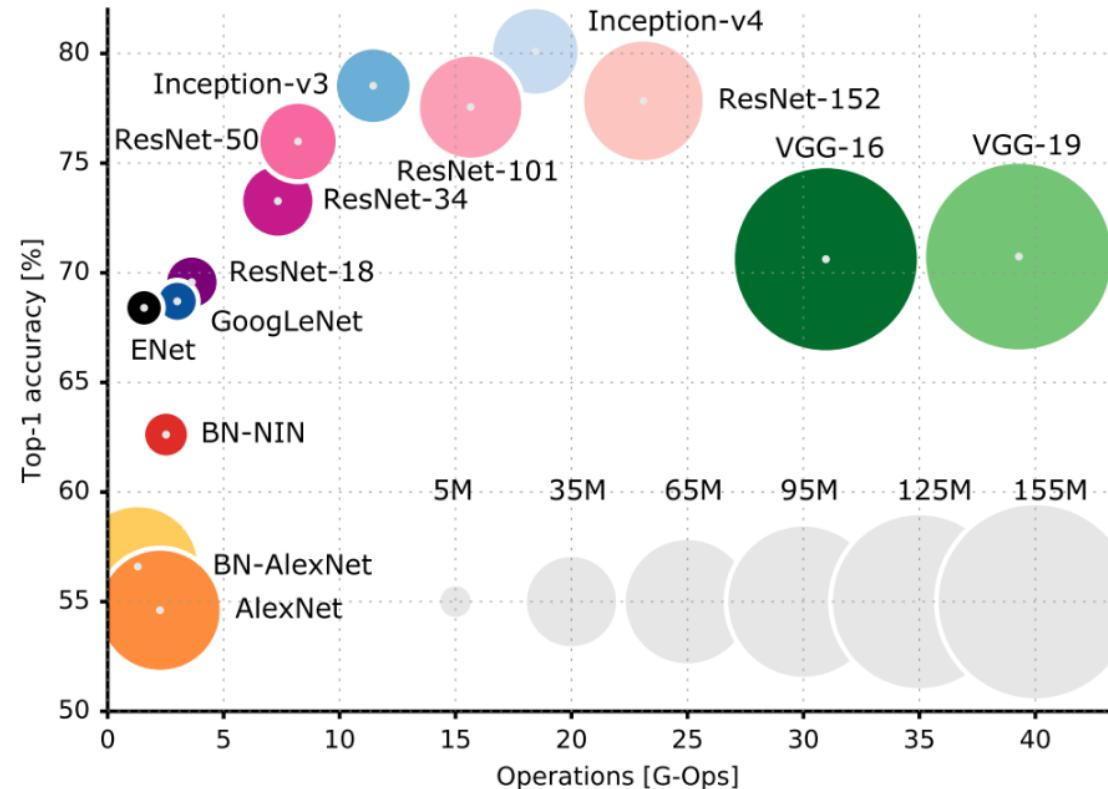


Deep Neural Networks

Problem.

How to interpret millions of parameters?

Types of Interpretability in ML

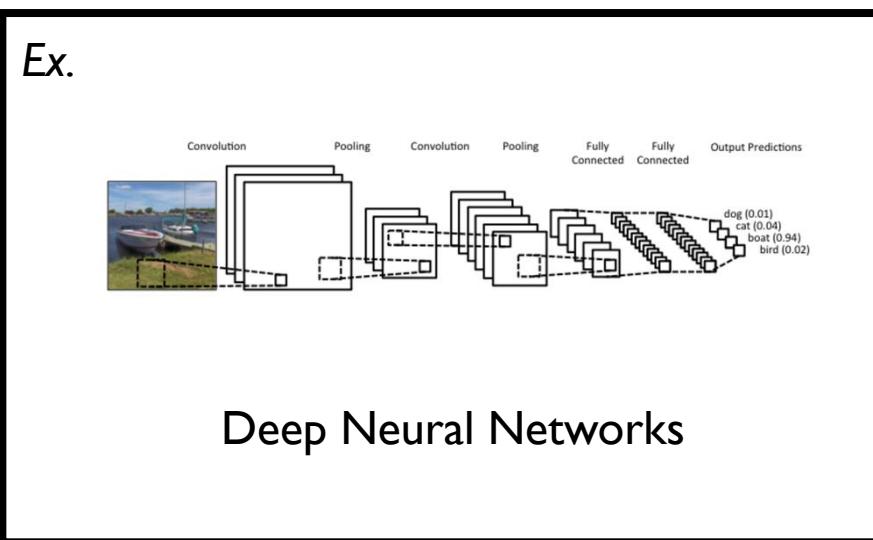


At least 5 million parameters!
(오백만)

Types of Interpretability in ML

Post-hoc Interpretability

Choose a complex model and develop a special technique to interpret it.



Problem. How to interpret millions of parameters?

Types of Post-hoc Interpretability

Post-hoc interpretability techniques can be classified by degree of “locality”

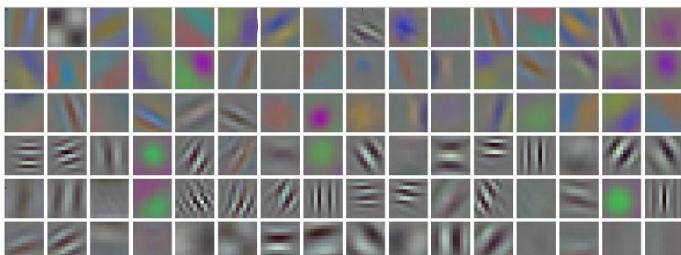


Types of Post-hoc Interpretability

Post-hoc interpretability techniques can be classified by degree of “locality”

Model ← → **Input**

What representations have the DNN learned?



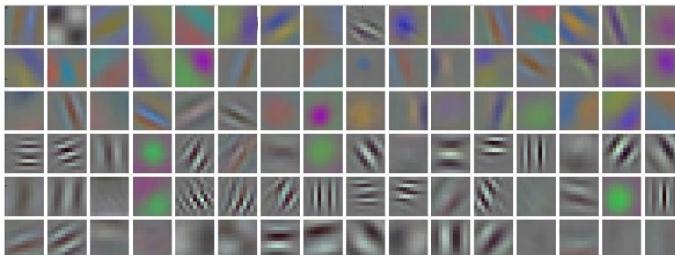
Types of Post-hoc Interpretability

Post-hoc interpretability techniques can be classified by degree of “locality”

Model



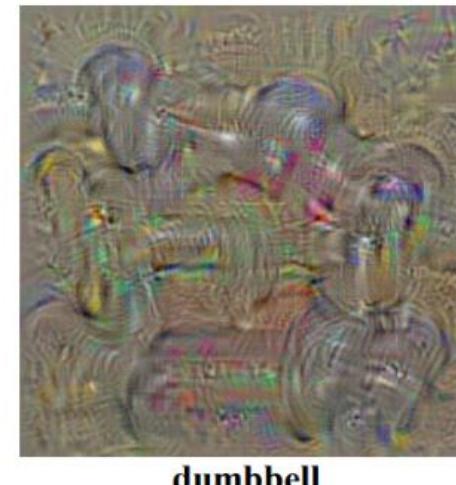
What representations have the DNN learned?



Input



What pattern / image maximally activates a particular neuron?



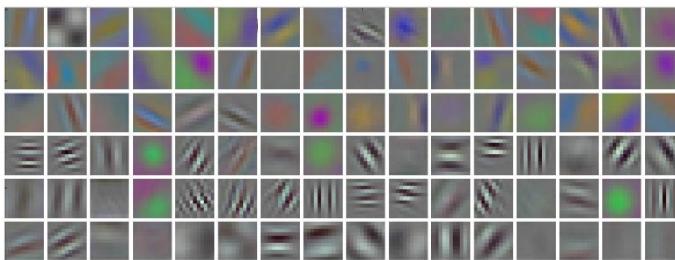
Types of Post-hoc Interpretability

Post-hoc interpretability techniques can be classified by degree of “locality”

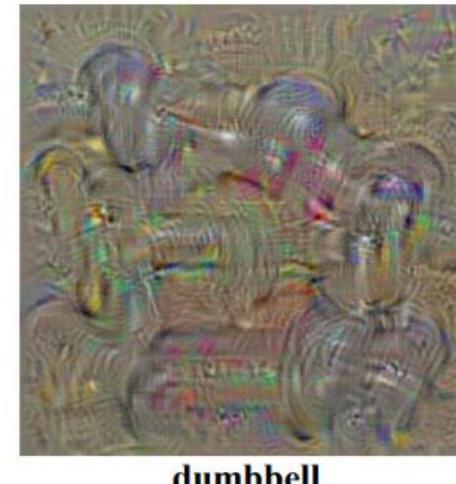
Model



What representations have the DNN learned?

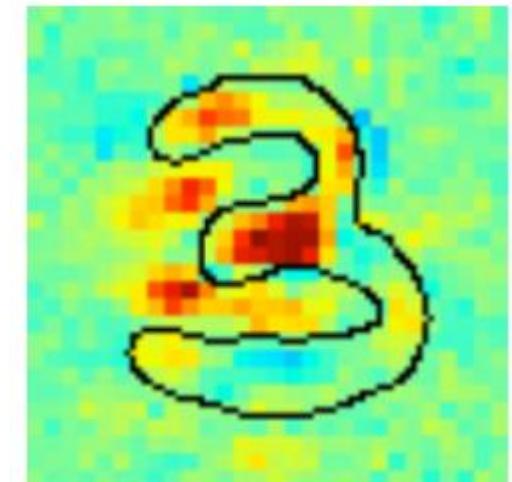


What pattern / image maximally activates a particular neuron?

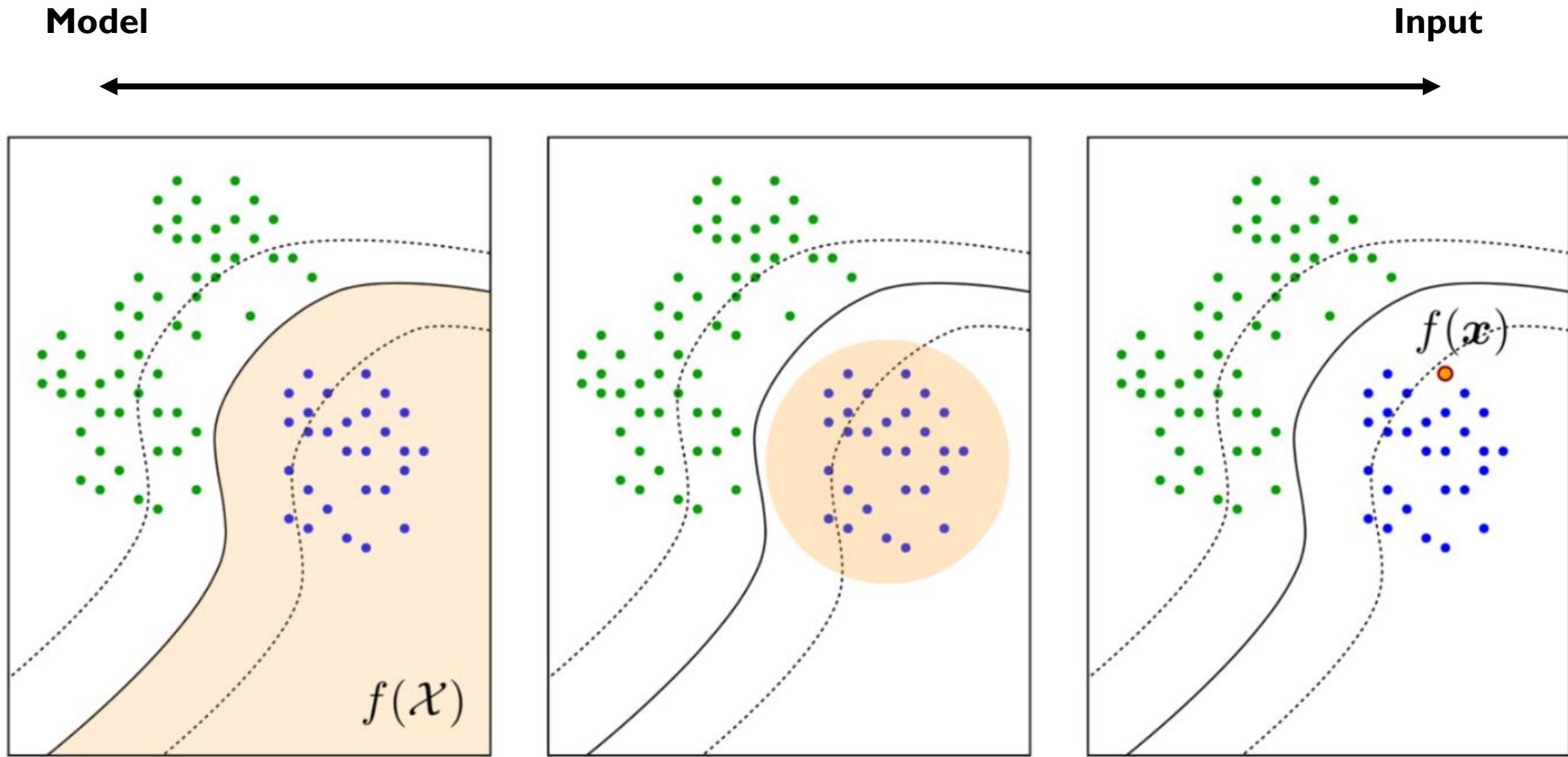


Input

Explain why input x has been classified as $f(x)$.



Types of Post-hoc Interpretability



Summary

1. What is interpretability in Deep Learning?

- Converting implicit information in DNN to (human) interpretable information

2. Why do we need interpretability in Deep Learning?

- Verify model works as intended
- Debug classifier
- Make discoveries
- Right to explanation

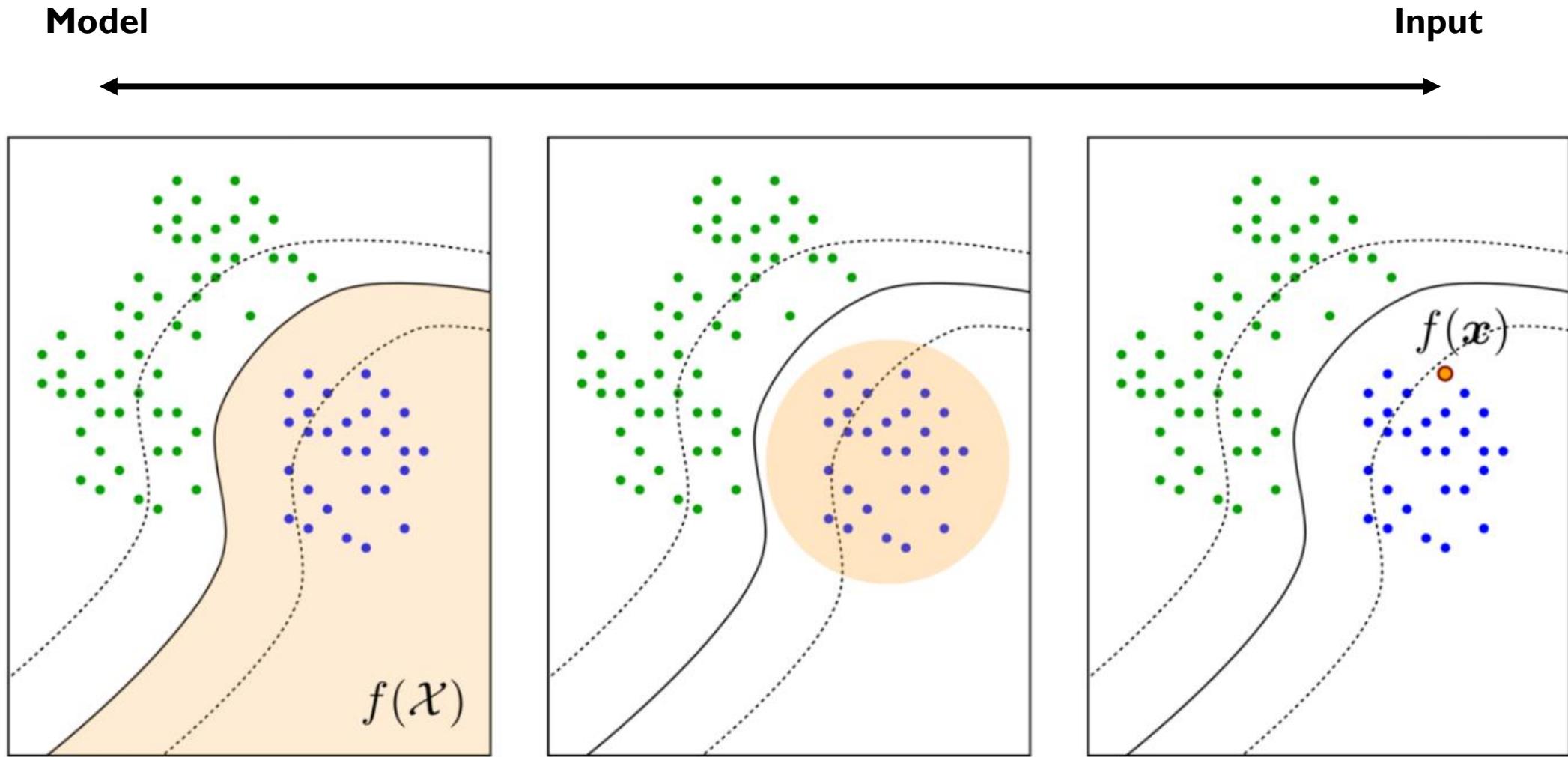
3. Types of Interpretability in ML

- Ante-hoc Interpretability: choose an interpretable model and train it
- Post-hoc Interpretability: choose a complex model and develop a special technique to interpret it
- Post-hoc interpretability techniques can be classified by degree of “locality”

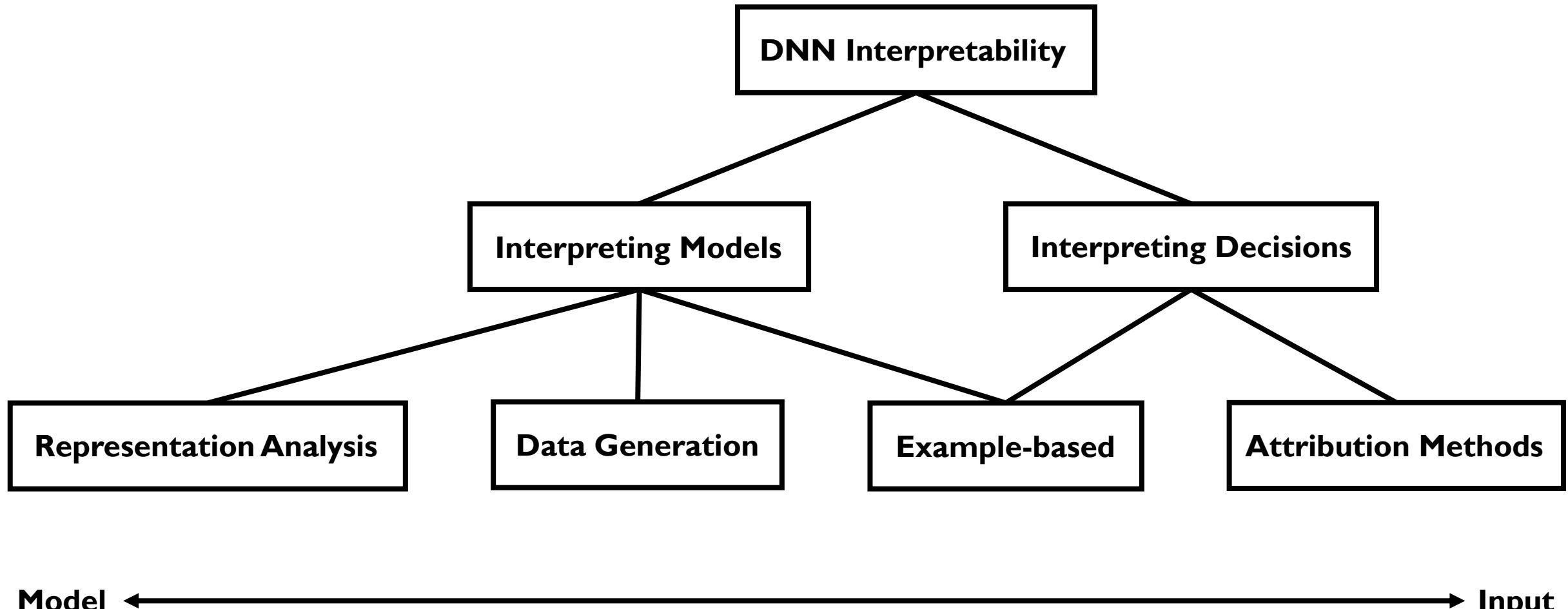
03

Interpretation of DNN

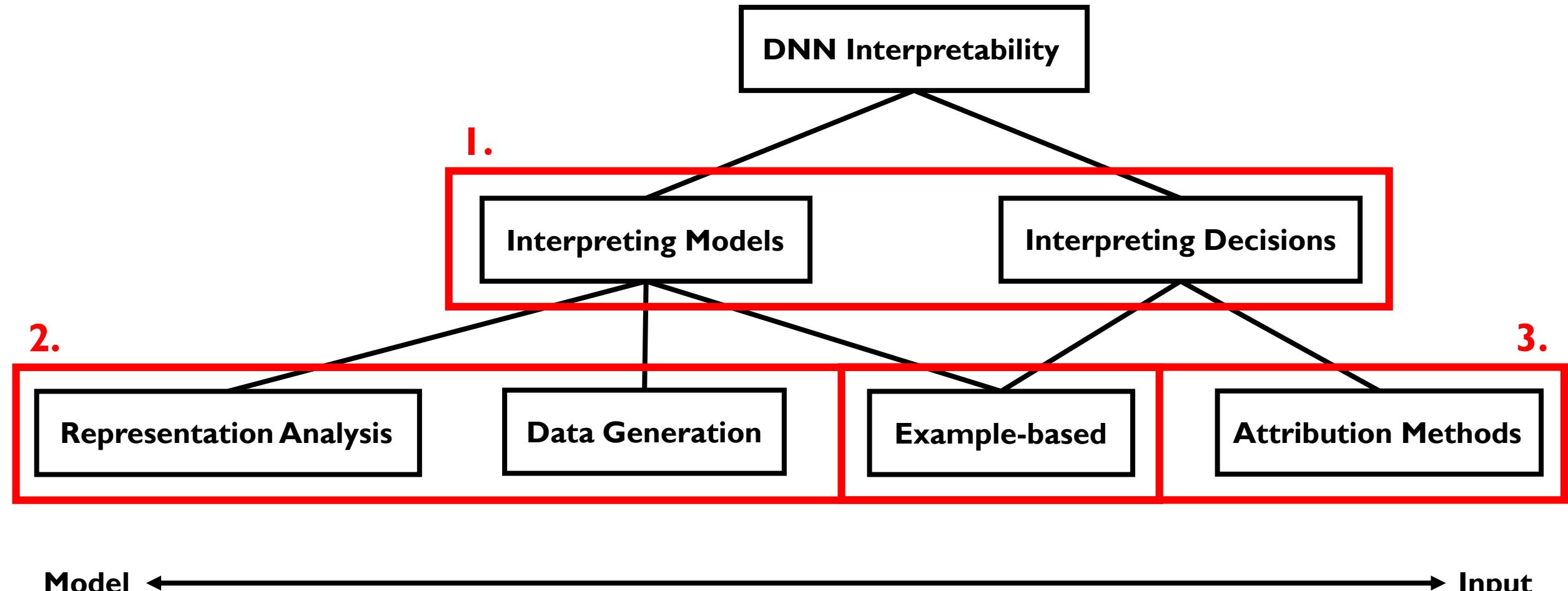
Types of Post-hoc Interpretability



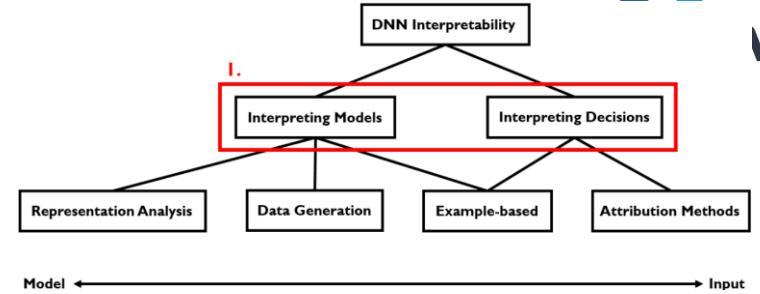
Types of DNN Interpretability



Types of DNN Interpretability



Types of DNN Interpretability



Interpreting Models (Macroscopic)

- “Summarize” DNN with a simpler model (e.g. decision tree)
- Find prototypical example of a category
- Find pattern maximizing activation of a neuron

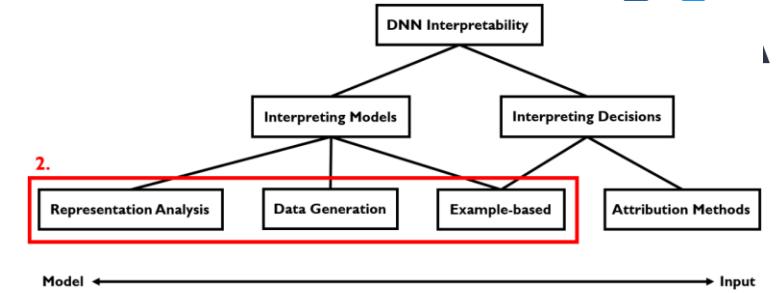
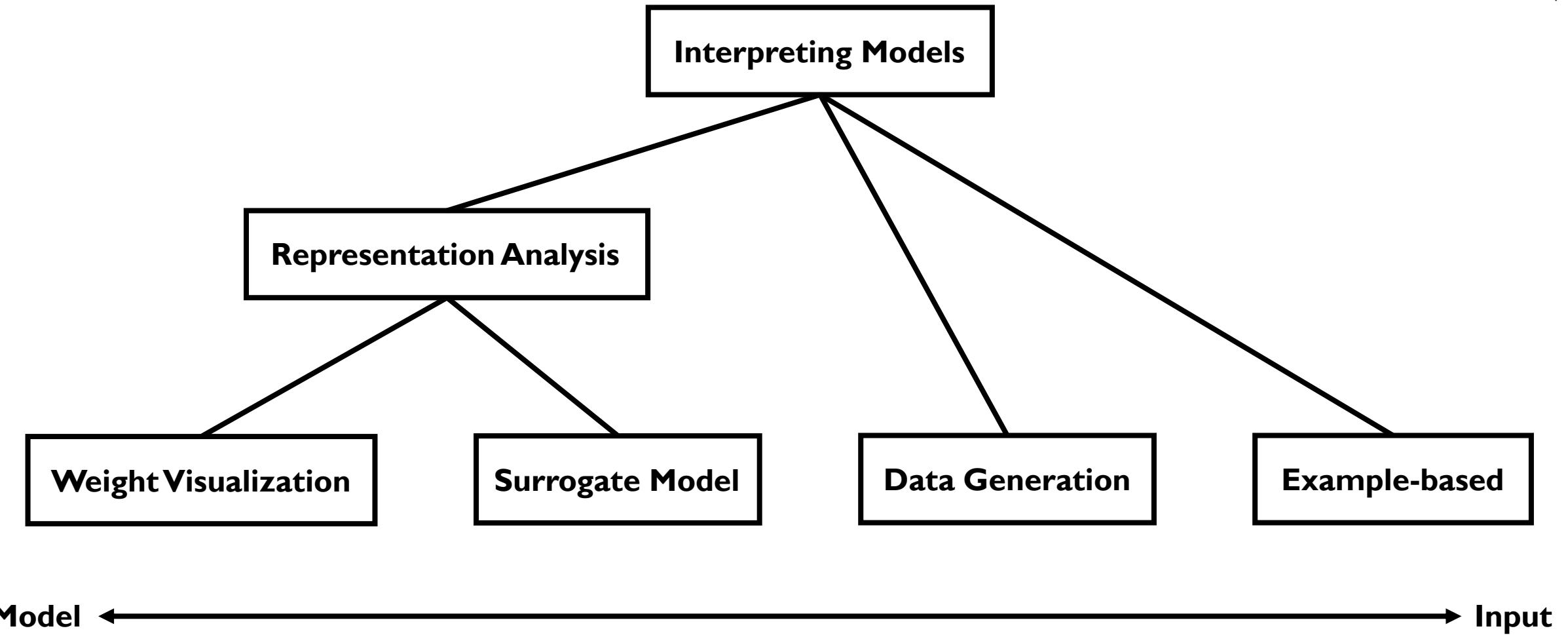
Better understand internal representations

Interpreting Decisions (Microscopic)

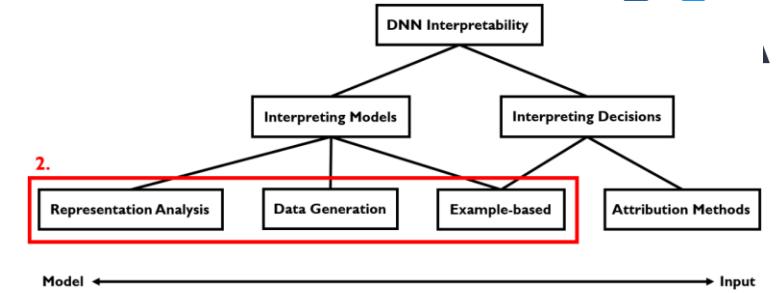
- Why did DNN make this decision?
- Verify that model behaves as expected
- Find evidence for decision

Important for practical applications

Types of DNN Interpretability



Types of DNN Interpretability



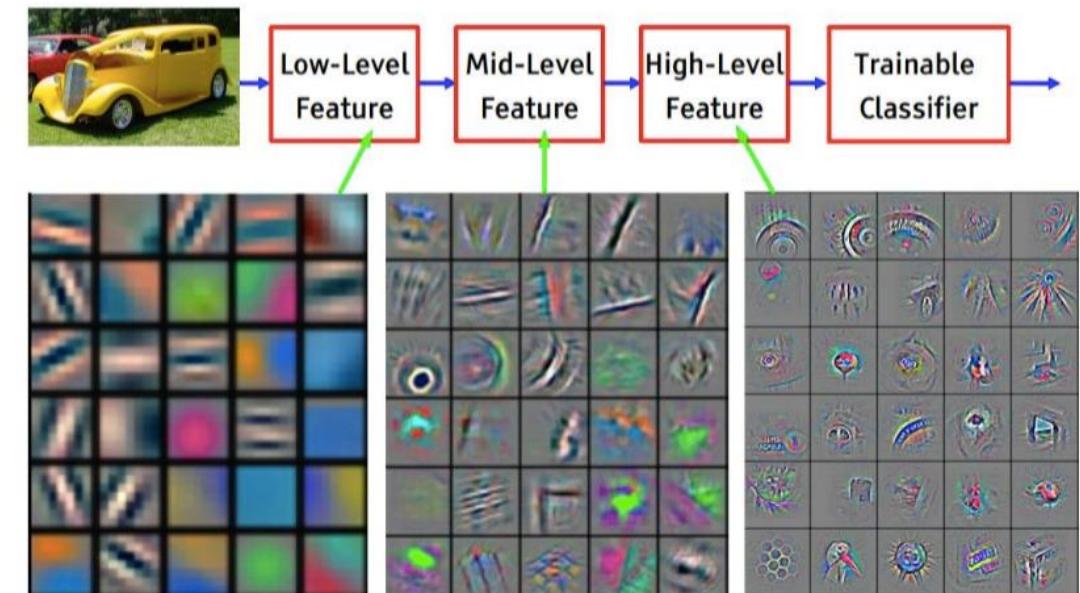
Weight Visualization

- Filter visualization in Convolutional Neural Networks
- Can understand what kind of features CNN has learned
- Still too many filters!

Surrogate Model

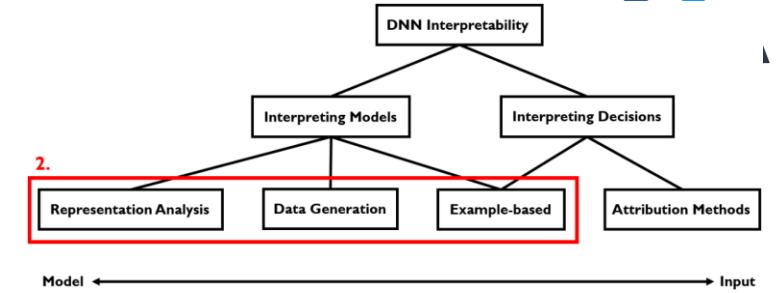
Data Generation

Example-based



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Types of DNN Interpretability



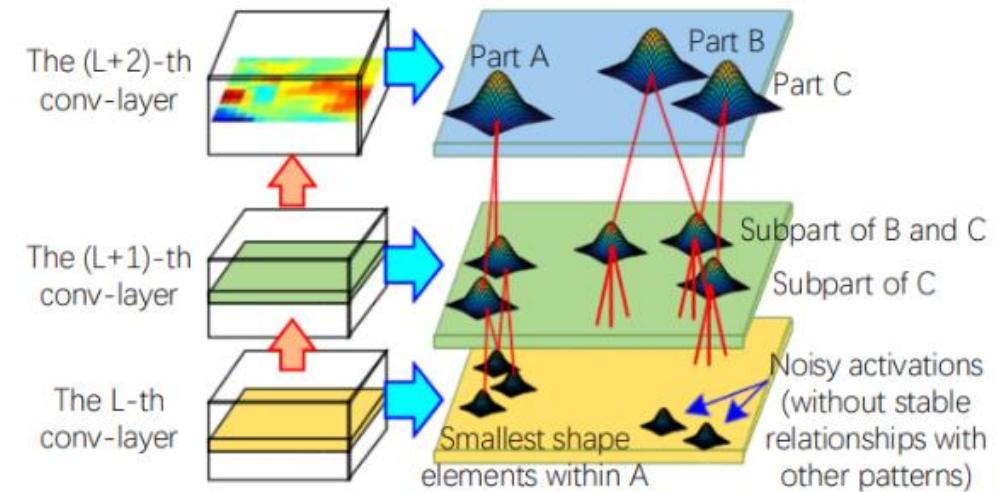
Weight Visualization

- “Summarize” DNN with a simpler model
- E.g. Decision trees, graphs or linear models

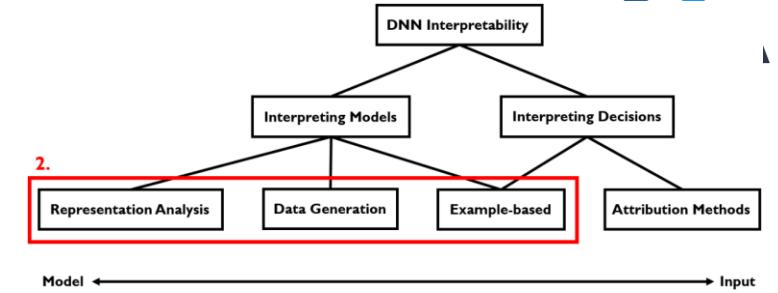
Surrogate Model

Data Generation

Example-based



Types of DNN Interpretability



Weight Visualization

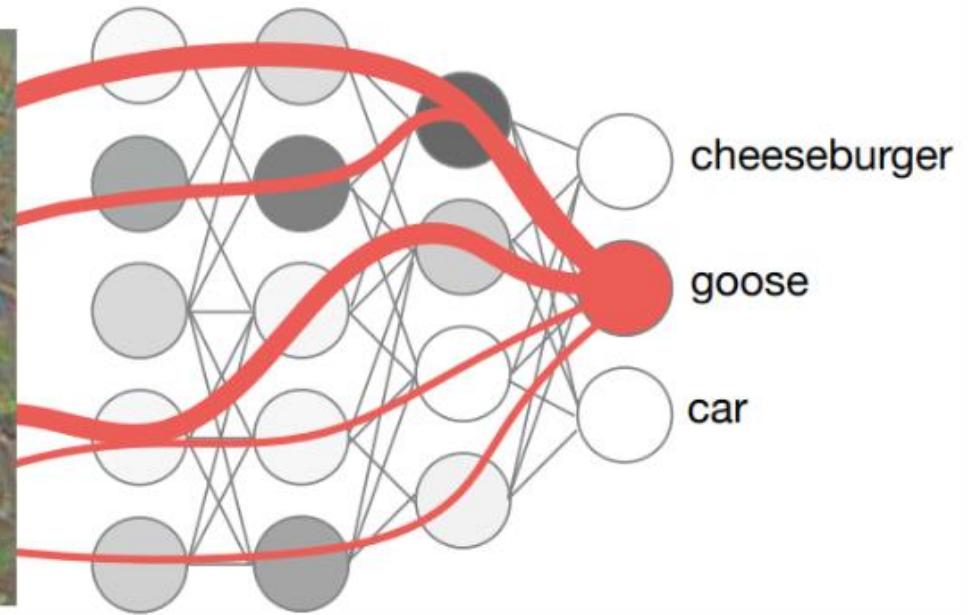
Surrogate Model

Data Generation

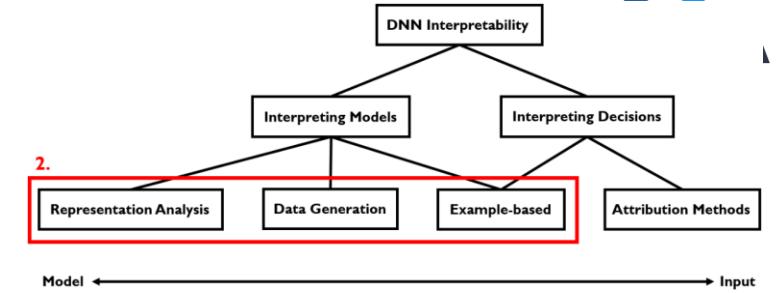
Example-based

Activation Maximization

- Find pattern maximizing activation of a neuron



Types of DNN Interpretability



Weight Visualization

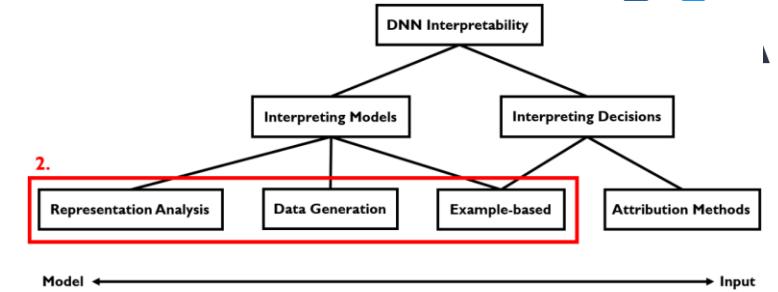
Surrogate Model

Data Generation

Example-based

$$\max_{x \in \mathcal{X}} p_\theta(\omega_c | x) + \lambda \Omega(x)$$

Types of DNN Interpretability



Weight Visualization

Surrogate Model

Data Generation

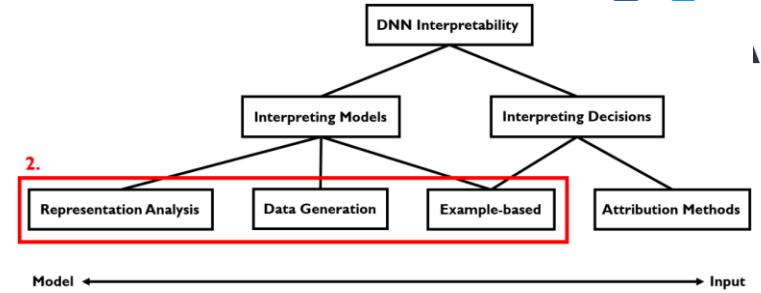
Example-based

$$\max_{x \in \mathcal{X}} p_{\theta}(\omega_c | x) + \lambda \Omega(x)$$

Class Probability Regularization Term

Two red arrows point from the highlighted terms in the equation to the corresponding labels below it.

Types of DNN Interpretability

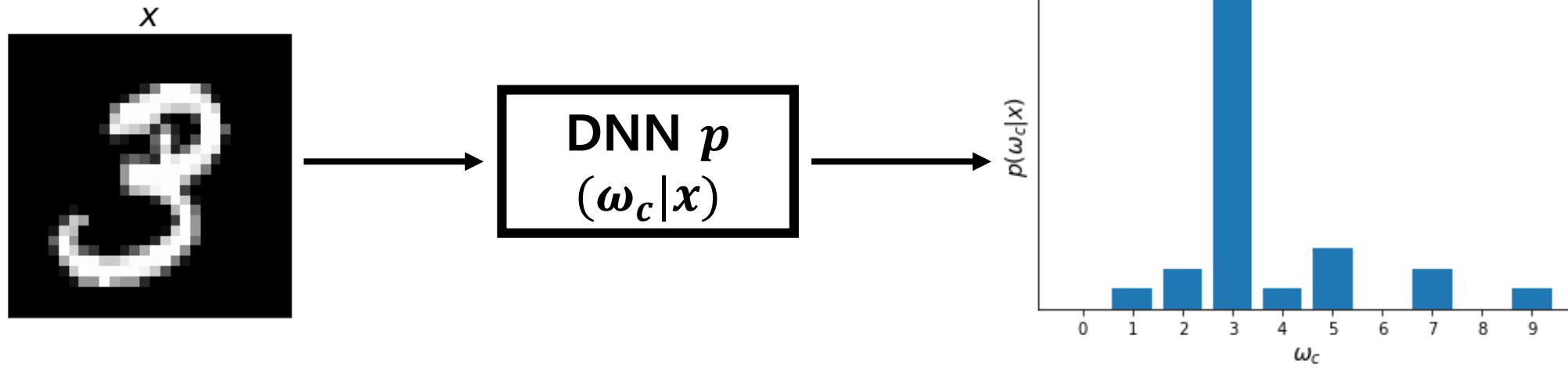


Weight Visualization

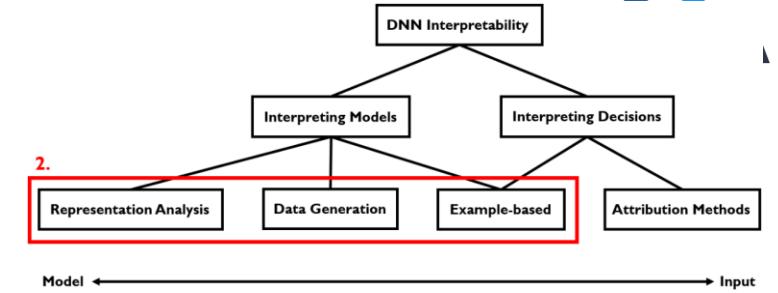
Surrogate Model

Data Generation

Example-based



Types of DNN Interpretability

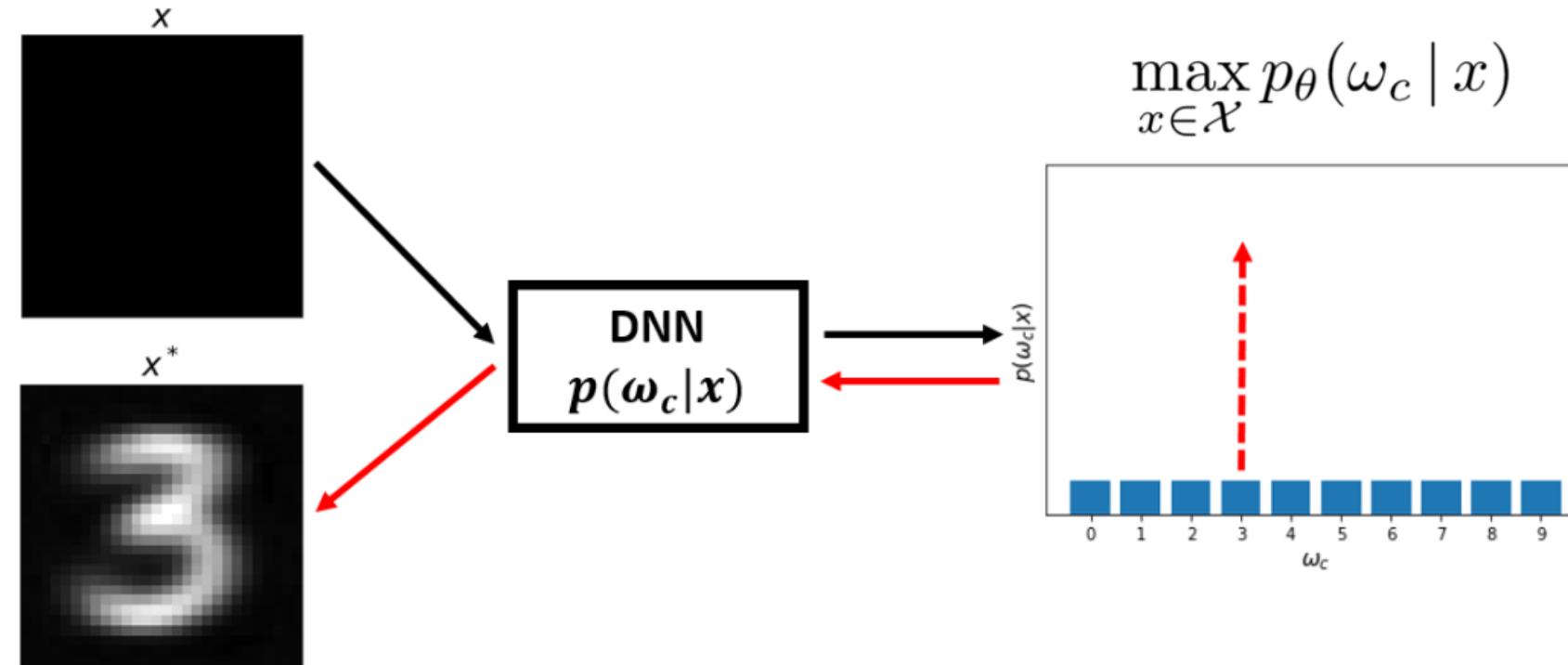


Weight Visualization

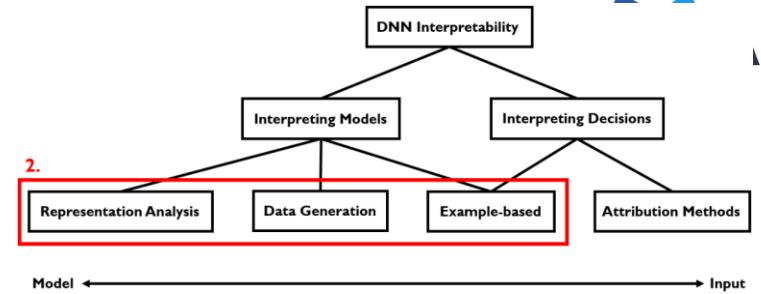
Surrogate Model

Data Generation

Example-based



Types of DNN Interpretability

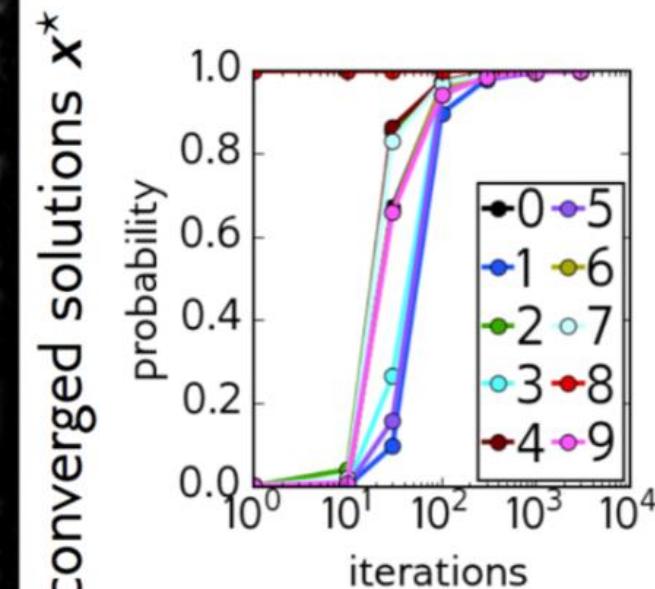
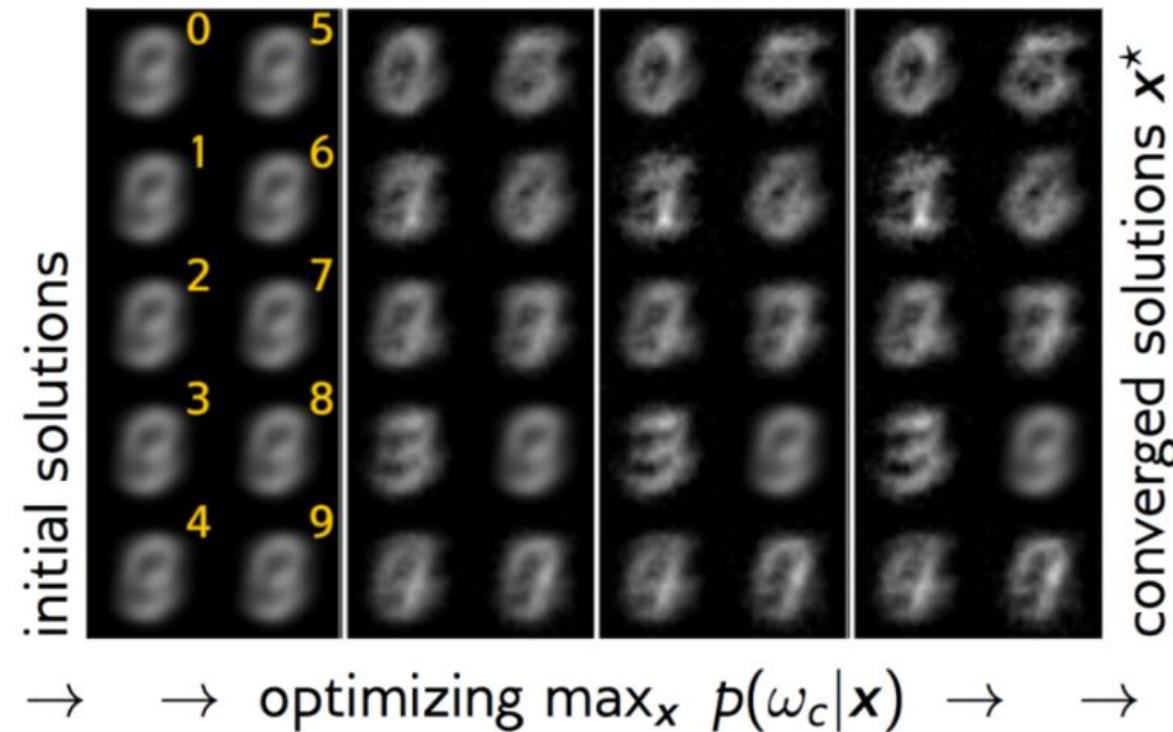


Weight Visualization

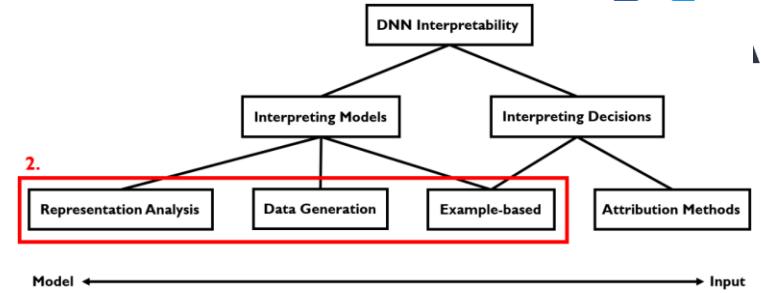
Surrogate Model

Data Generation

Example-based



Types of DNN Interpretability



Weight Visualization

Surrogate Model

Data Generation

Example-based

goose

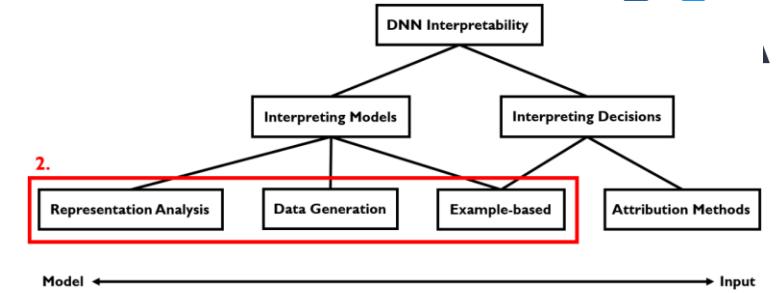


ostrich



Images from **Simonyan et al. 2013** "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps"

Types of DNN Interpretability



Weight Visualization

Surrogate Model

Data Generation

Example-based

Advantages

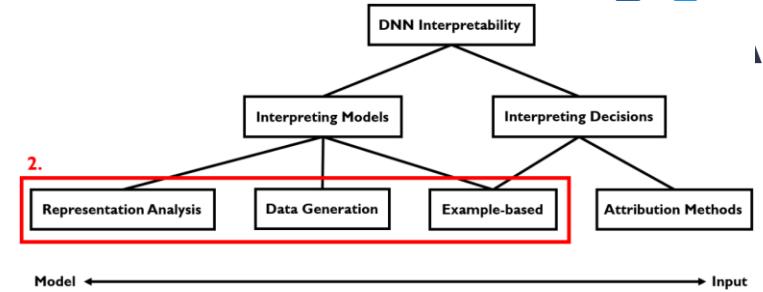
- AM builds typical patterns for given classes (e.g. beaks, legs)
- Unrelated background objects are not present in the image

Disadvantages

- Does not resemble class-related patterns
- Lowers the quality of the interpretation for given classes

Redefine optimization problem!

Types of DNN Interpretability



Weight Visualization

Surrogate Model

Data Generation

Example-based

- Does not resemble class-related patterns
- Lowers the quality of the interpretation for given classes

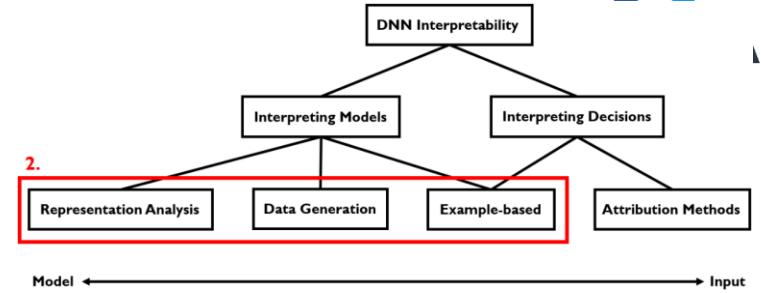
Redefine optimization problem!

Force the generated data x^* to match the data more closely

Find the input pattern that maximizes class probability

Find the most likely input pattern for a given class

Types of DNN Interpretability



Weight Visualization

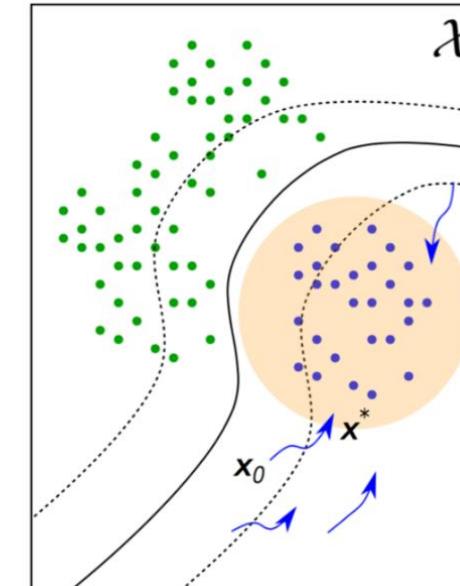
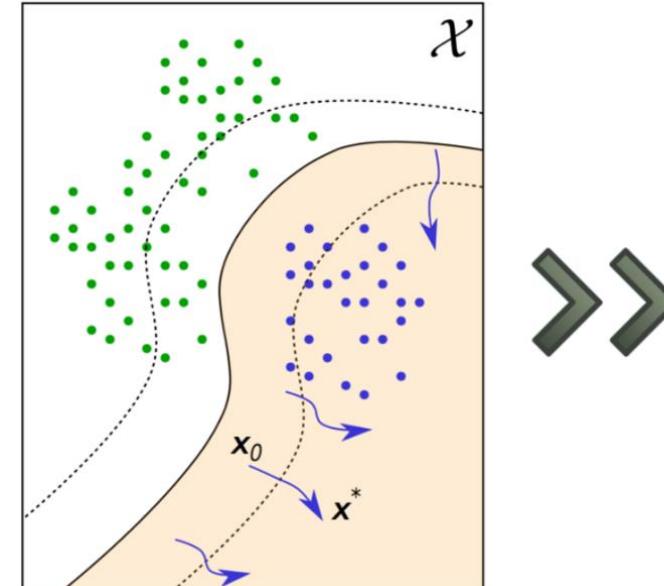
Surrogate Model

Data Generation

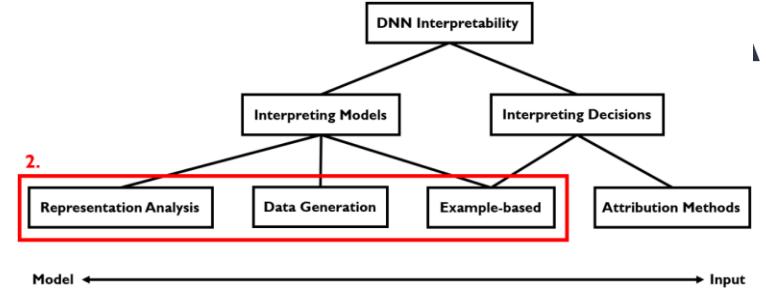
Example-based

Find the input pattern that maximizes class probability

Find the most likely input pattern for a given class



Types of DNN Interpretability



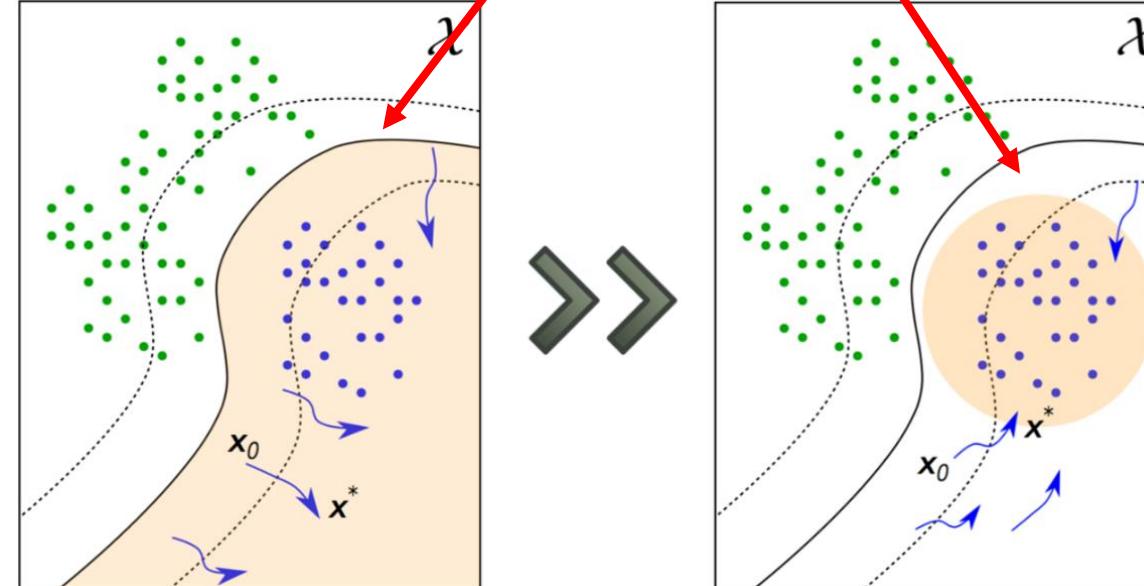
Weight Visualization

Surrogate Model

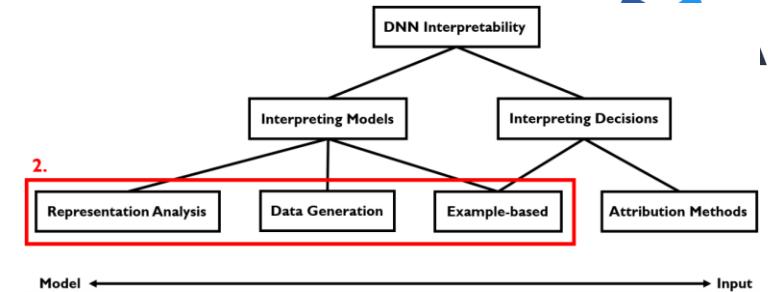
Data Generation

Example-based

$$\max_{x \in \mathcal{X}} p_\theta(\omega_c | x) + \lambda \Omega(x)$$



Types of DNN Interpretability



Weight Visualization

Surrogate Model

Data Generation

Example-based

Find the input pattern that maximizes class probability

Find the most likely input pattern for a given class

Activation Maximization with *Expert*

$$p(x|\omega_c) \propto p(\omega_c|x) \cdot p(x)$$

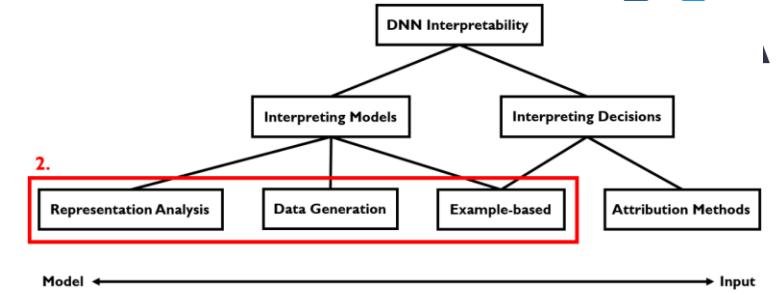
original

Activation Maximization in *Code Space*

$$\max_{\mathbf{z} \in \mathcal{Z}} p(\omega_c | \underbrace{g(\mathbf{z})}_{\mathbf{x}}) + \lambda \|\mathbf{z}\|^2 \quad \mathbf{x}^* = g(\mathbf{z}^*)$$

These two techniques require an **unsupervised model of the data**, either a density model $p(x)$ or a generator $g(z)$

Types of DNN Interpretability



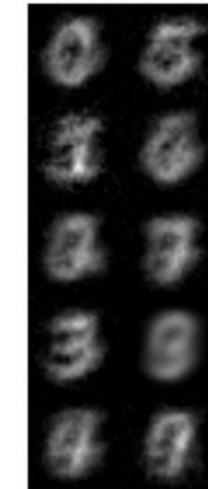
Weight Visualization

Surrogate Model

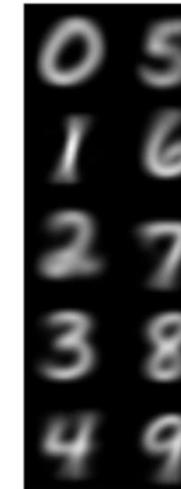
Data Generation

Example-based

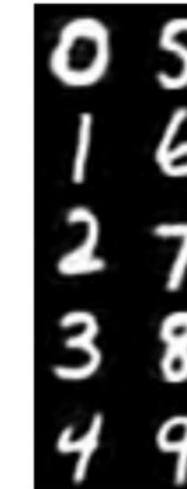
simple AM
(initialized
to mean)



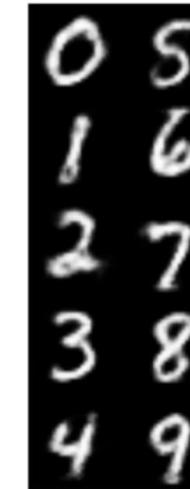
simple AM
(init. to
class
means)



AM-density
(init. to
class
means)

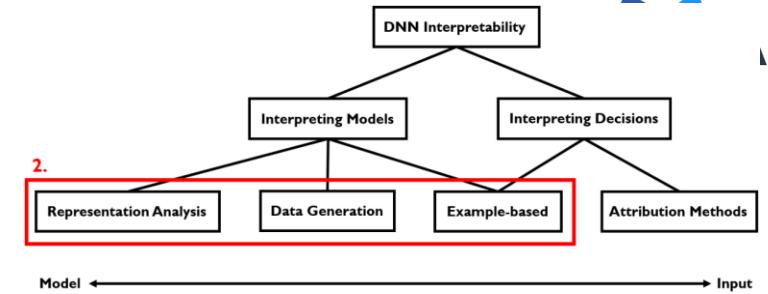


AM-gen
(init. to
class
means)



Observation: Connecting to the **data** leads to **sharper** visualizations.

Types of DNN Interpretability



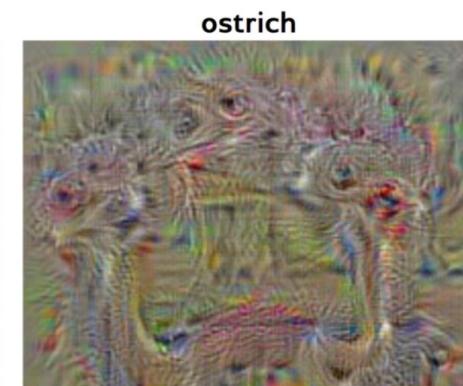
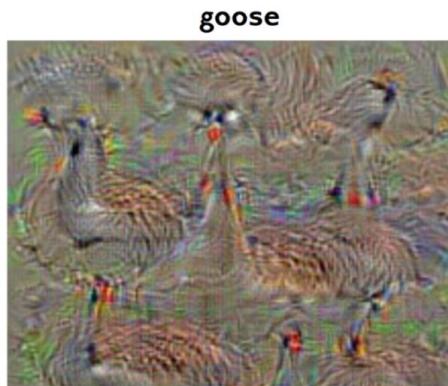
Weight Visualization

Surrogate Model

Data Generation

Example-based

Activation Maximization



Images from Simonyan et al. 2013 "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps"

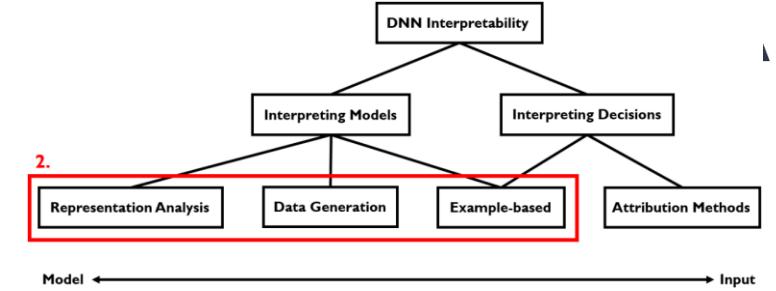
Activation Maximization in Code Space

Images from Nguyen et al. 2016. "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks"



Observation: Connecting to the **data** leads to **sharper** visualizations.

Types of DNN Interpretability



Weight Visualization

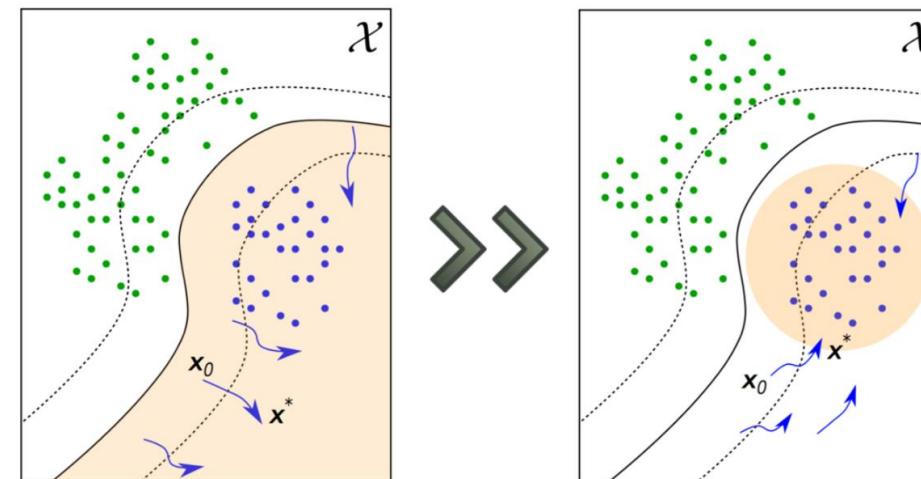
Surrogate Model

Data Generation

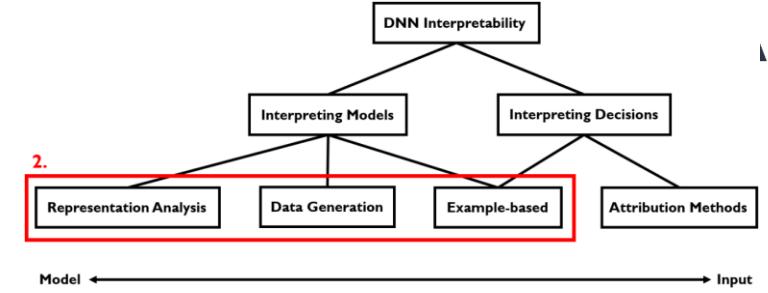
Example-based

Summary

- DNNs can be interpreted by finding input patterns that maximize a certain output quantity.
- Connecting to the data improves the interpretability of the visualization.



Types of DNN Interpretability



Weight Visualization

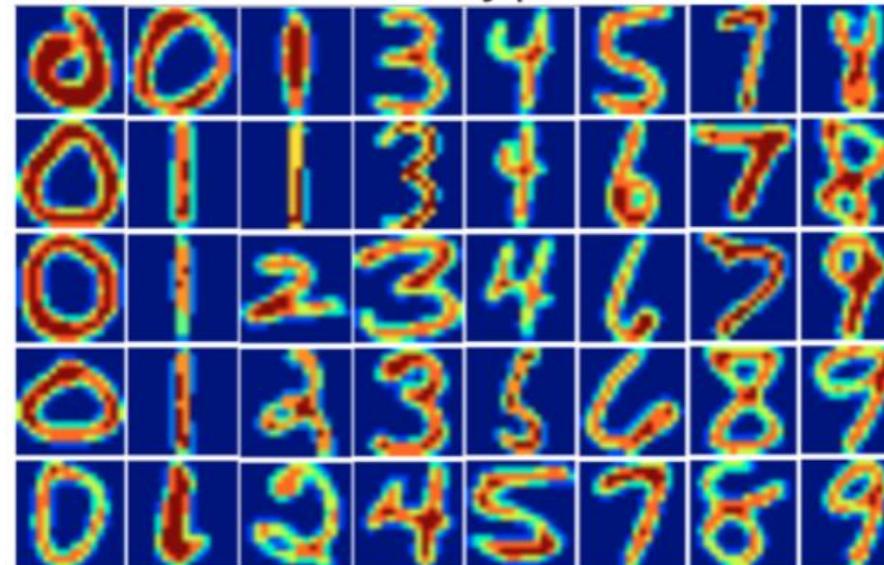
Surrogate Model

Data Generation

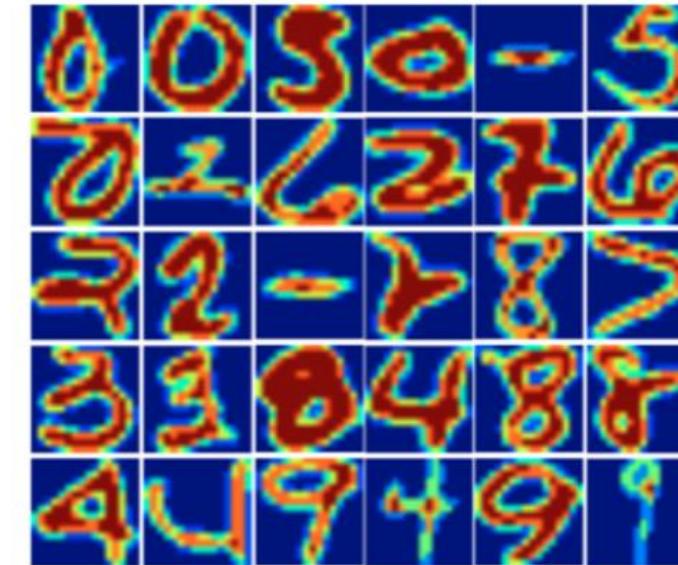
Example-based

- Find image instances that represent / do not represent the image class

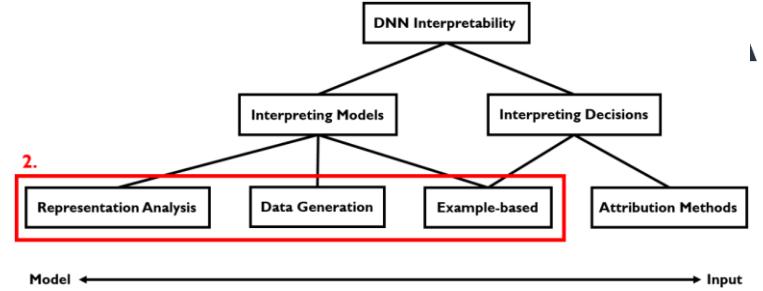
Prototypes



Criticisms



Types of DNN Interpretability



Weight Visualization

Surrogate Model

Data Generation

Example-based

- Find image instances that represent / do not represent the image class

Prototypes



Criticisms



Prototypes

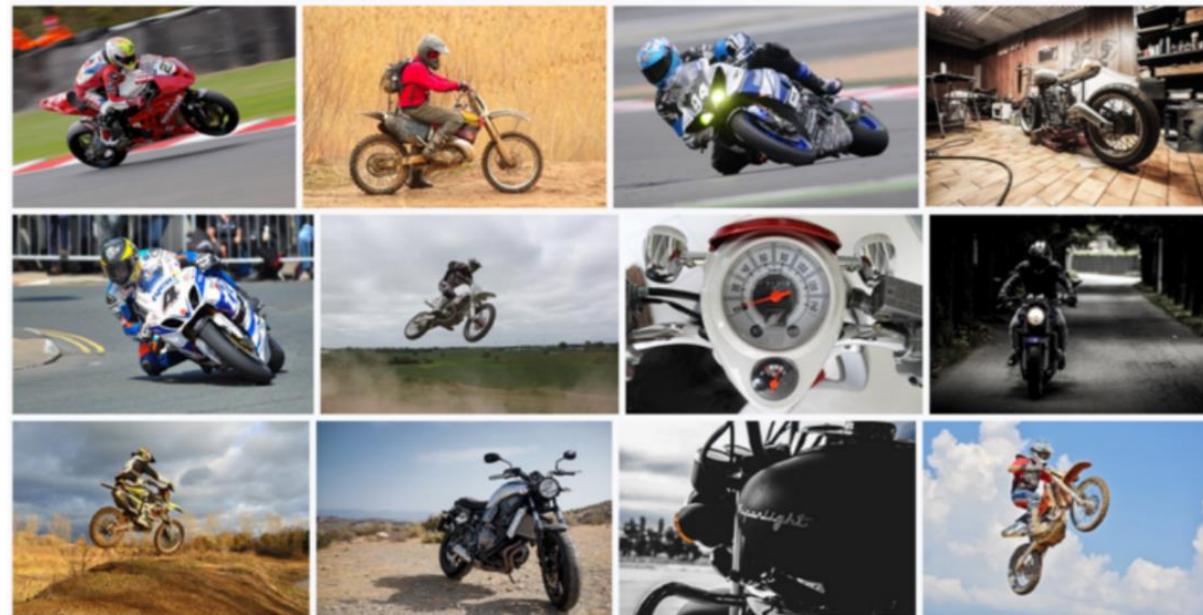


Criticisms



Limitation of Model Interpretations

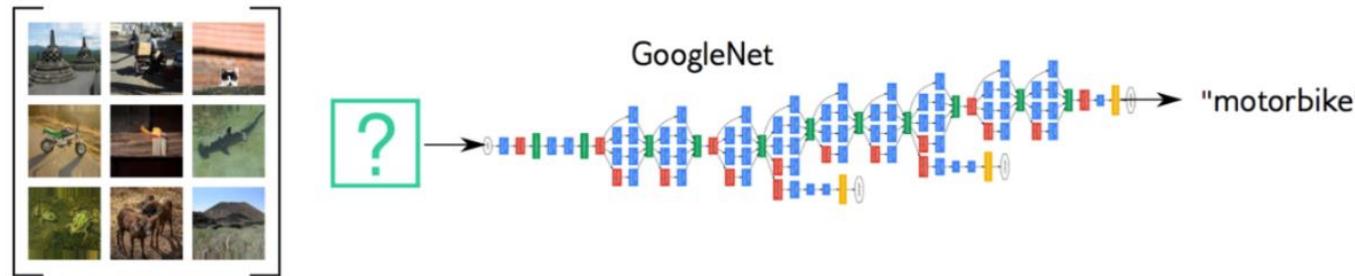
Question: What would be the best image to interpret the class “motorcycle”?



- Summarizing a concept or a category like “motorcycle” into a single image is difficult.
- A good interpretation would grow as large as the diversity of the concept to interpret.

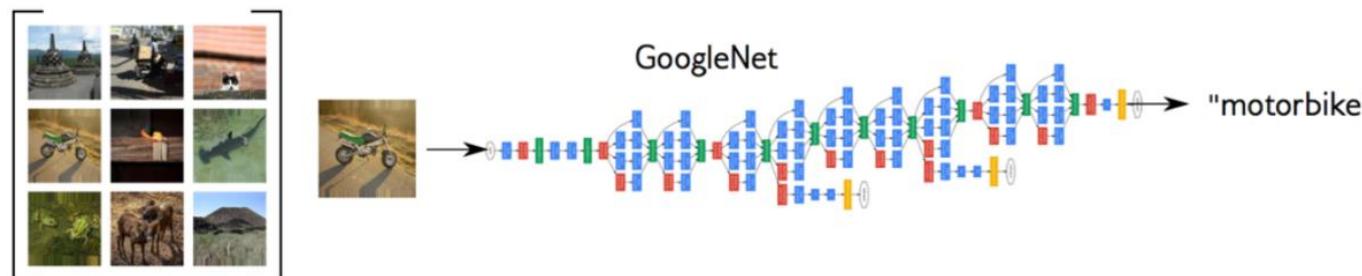
Limitation of Model Interpretations

Finding a prototype:



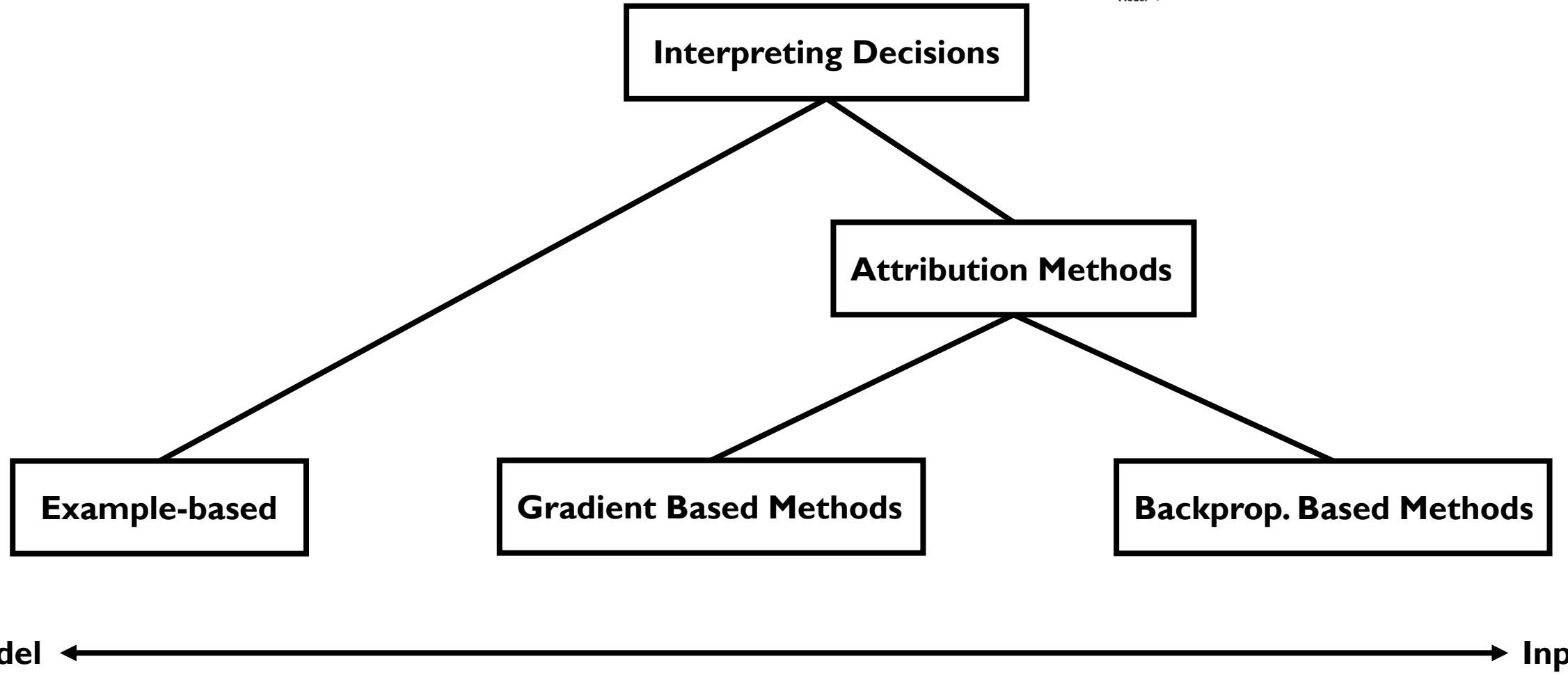
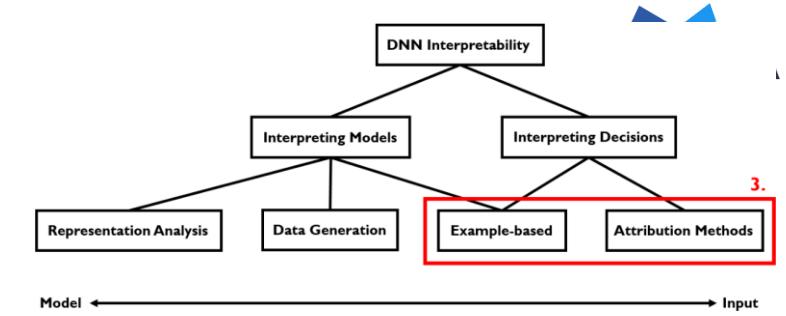
Question: How does a “motorbike” typically look like?

Decision explanation:

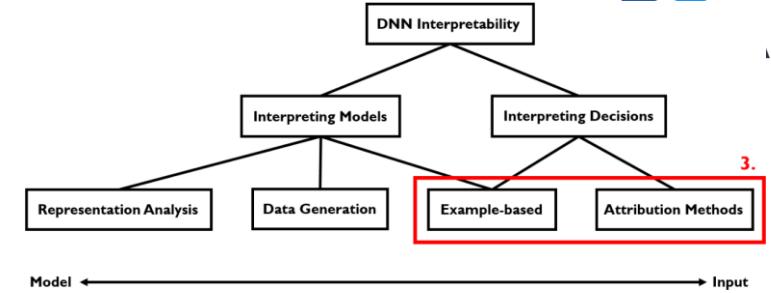


Question: Why is this example classified as a motorbike?

Types of DNN Interpretability



Types of DNN Interpretability



Example-based

Attribution Methods

Gradient Based

Backprop. Based

- Which training instance influenced the decision most?
- Still does not **specifically highlight** which features were important.

'Sunflower': 59.2% conf.

Original



Influence: 0.09



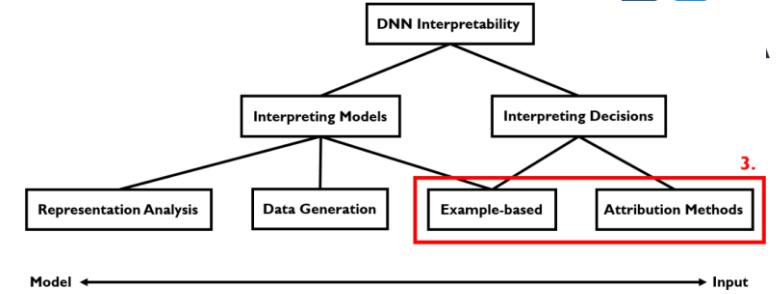
Influence: 0.14



Influence: 0.42



Types of DNN Interpretability



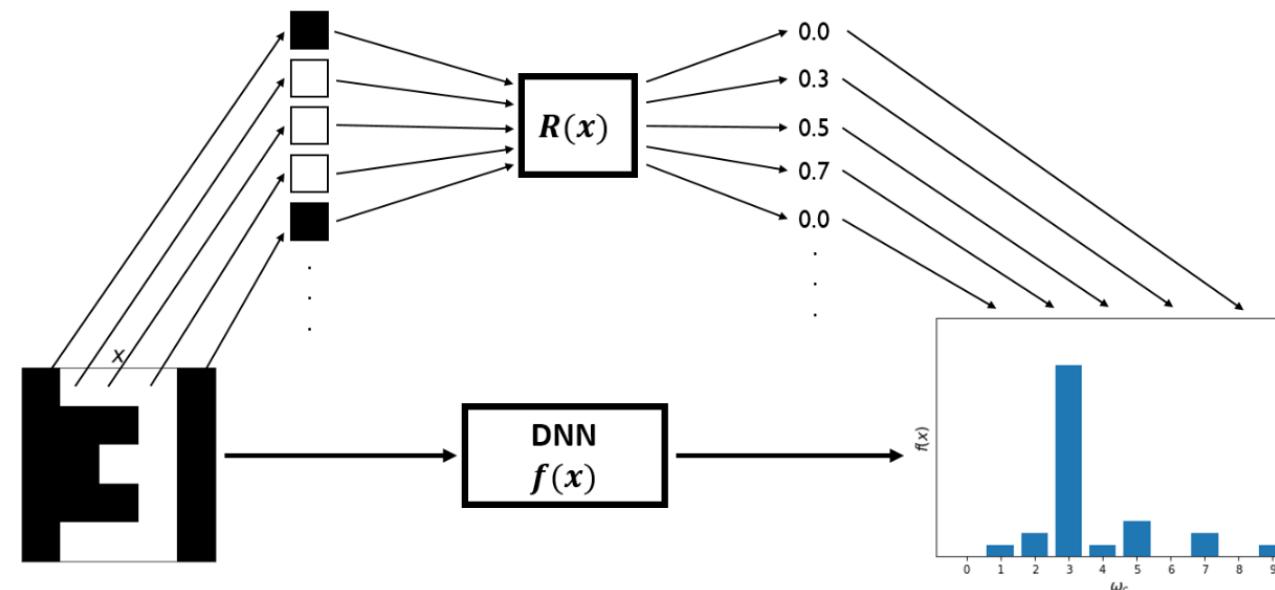
Example-based

Attribution Methods

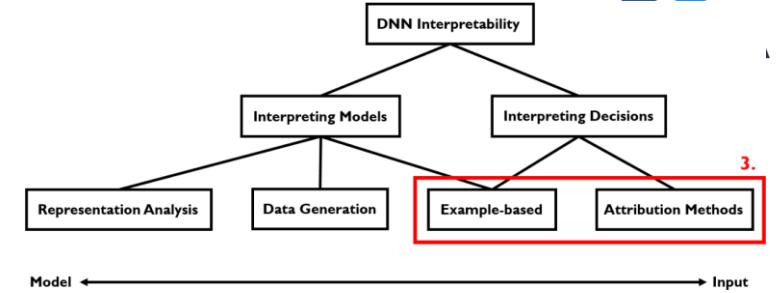
Gradient Based

Backprop. Based

Given an image $x \in \mathbb{R}^n$ and a decision $f(x)$,
assign to each pixel x_1, x_2, \dots, x_n **attribution values** $R_1(x), R_2(x), \dots, R_n(x)$.



Types of DNN Interpretability



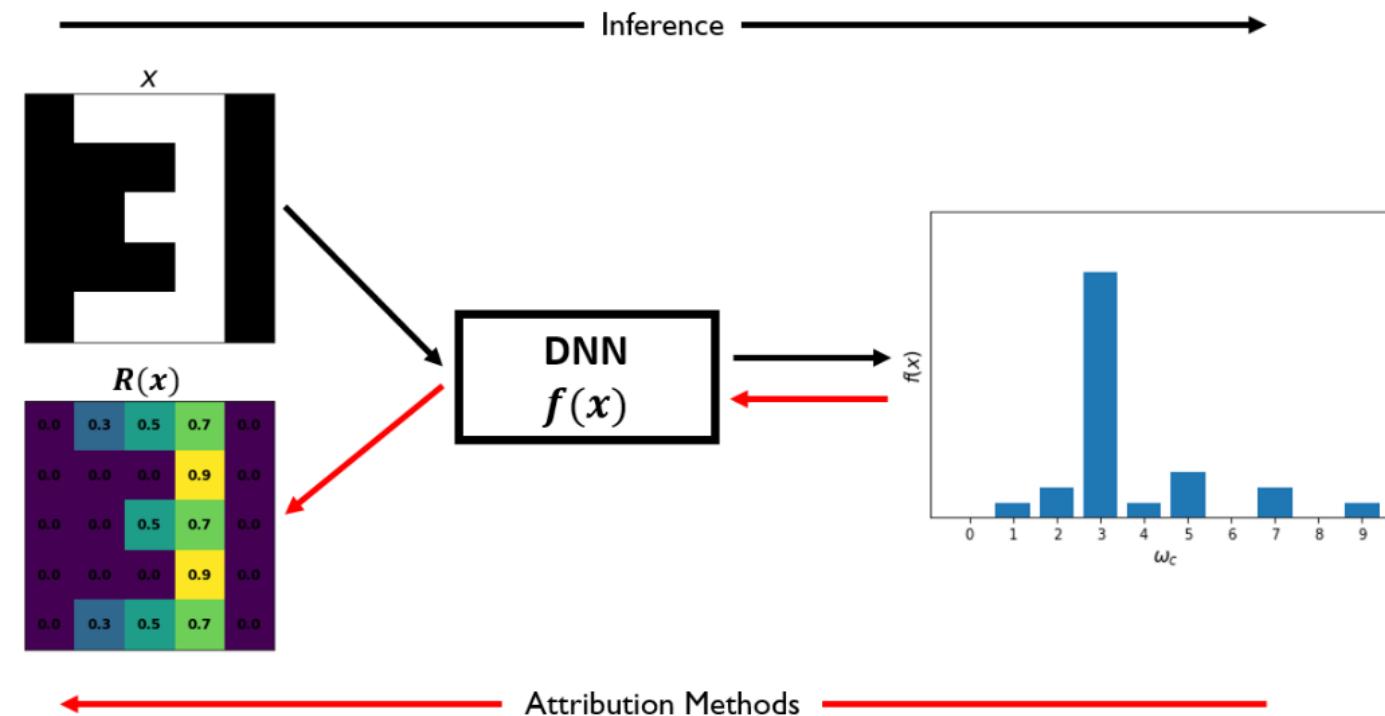
Example-based

Attribution Methods

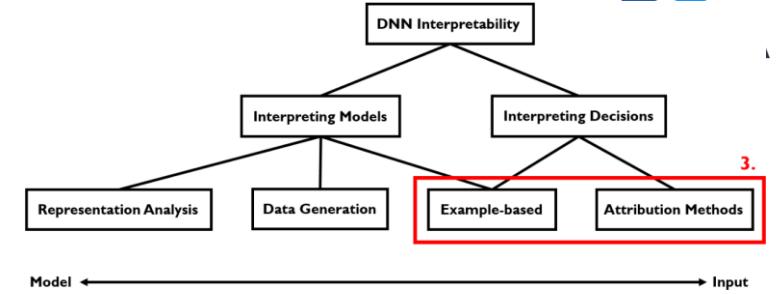
Gradient Based

Backprop. Based

Usually visualized as heatmaps



Types of DNN Interpretability



Example-based

Attribution Methods

Gradient Based

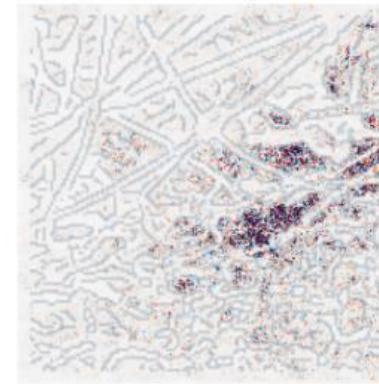
Backprop. Based

Usually visualized as **heatmaps**

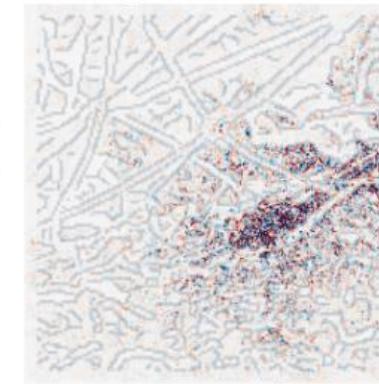
Original (label: "garter snake")



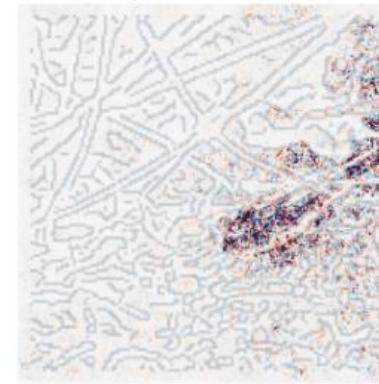
Grad * Input



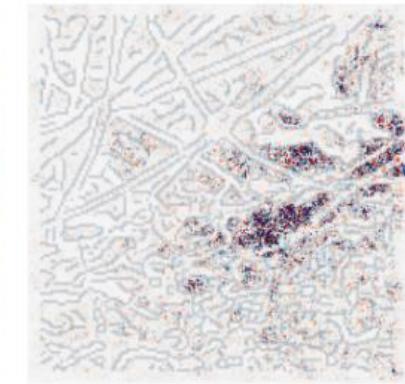
Integrated Gradients



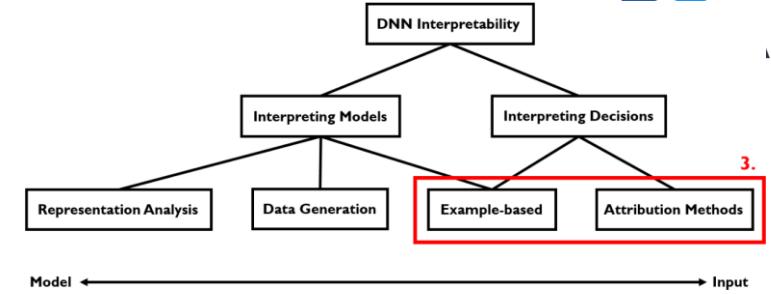
DeepLIFT (Rescale)



ϵ -LRP



Types of DNN Interpretability



Example-based

Attribution Methods

Gradient Based

Backprop. Based

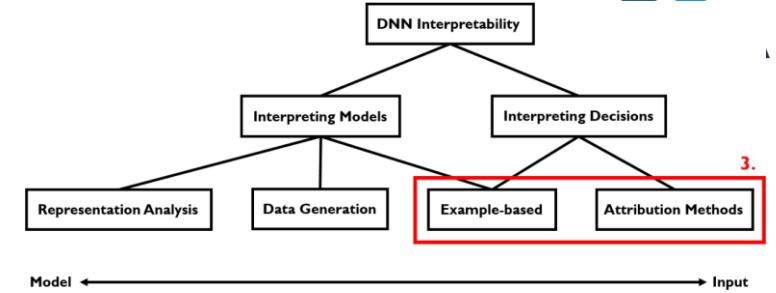
The Baseline Attribution Method **Saliency Map**

- Gradient of the decision $f(x)$ with respect to the input image x :

$$\text{Saliency}(x) := \nabla_x f(x) = \frac{\partial f(x)}{\partial x}$$

- Can be calculated through **backpropagation**.

Types of DNN Interpretability



Example-based

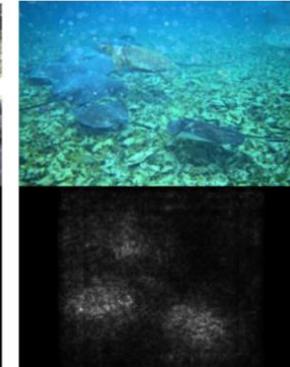
Attribution Methods

Gradient Based

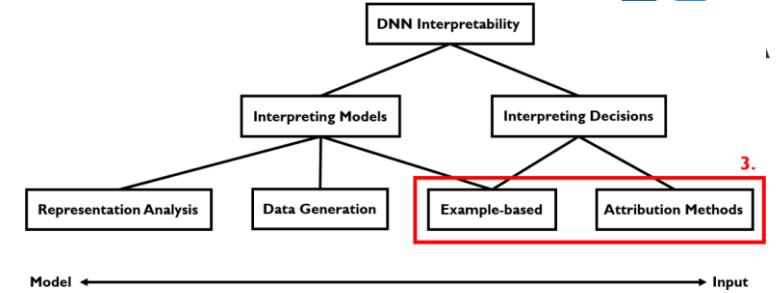
Backprop. Based

The Baseline Attribution Method **Saliency Map**

- Saliency maps are very **noisy!**
- Only roughly correlated with the object(s) of interest.



Types of DNN Interpretability



Example-based

Attribution Methods

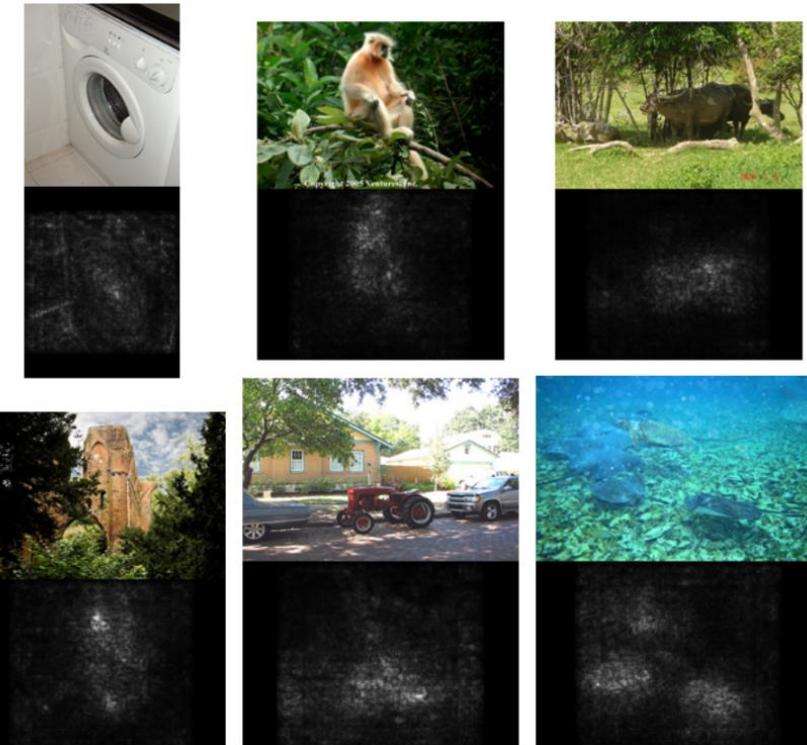
Gradient Based

Backprop. Based

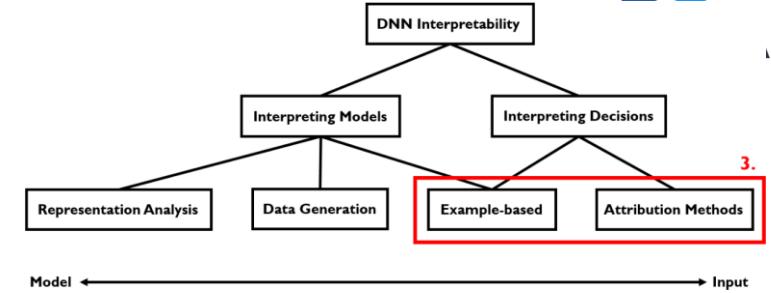
The Baseline Attribution Method **Saliency Map**

- Saliency maps are very **noisy!**
- Only roughly correlated with the object(s) of interest.

Question: How to improve saliency maps?



Types of DNN Interpretability



Example-based

Attribution Methods

Gradient Based

Backprop. Based

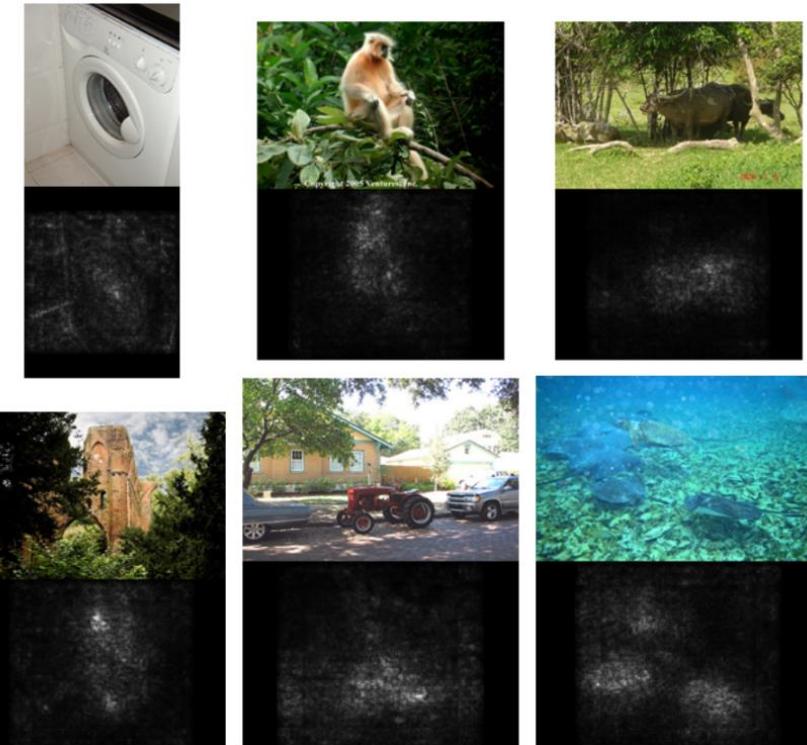
The Baseline Attribution Method **Saliency Map**

- Saliency maps are very **noisy!**
- Only roughly correlated with the object(s) of interest.

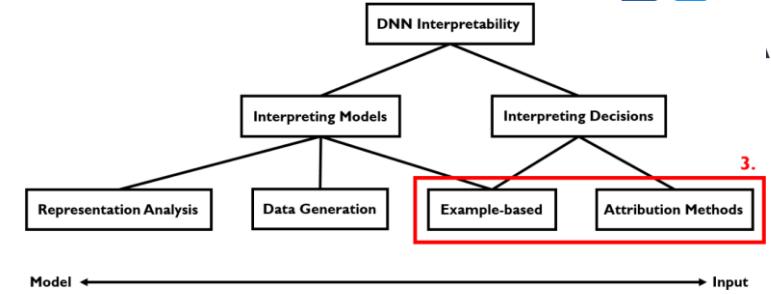
Question: How to improve saliency maps?



Question: Why are saliency maps noisy?



Types of DNN Interpretability



Example-based

Attribution Methods

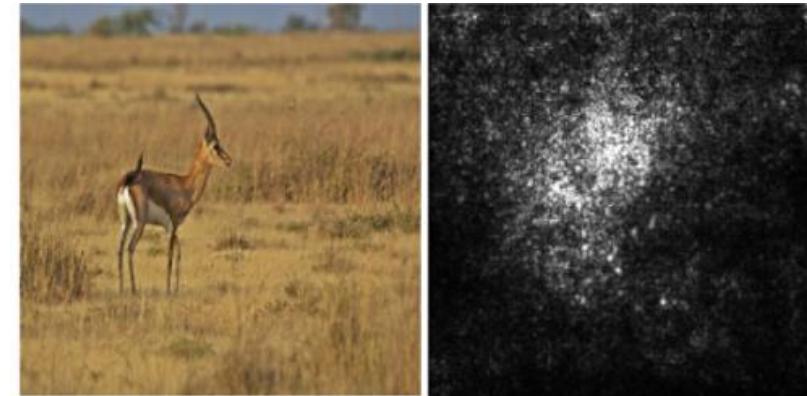
Gradient Based

Backprop. Based

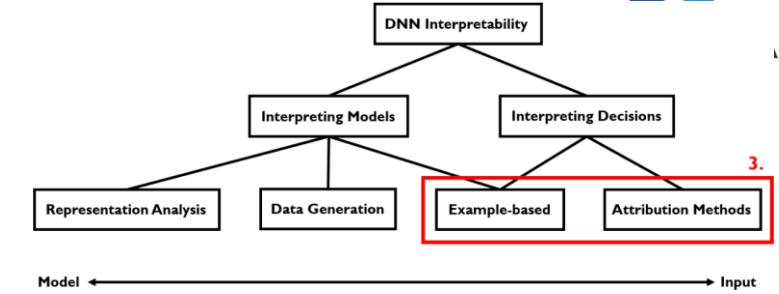
Question: Why are saliency maps noisy?

Hypothesis 1 – Saliency maps are truthful

- Certain pixels scattered randomly across the image are central to how the network is making a decision.
- Noise is important!



Types of DNN Interpretability



Example-based

Attribution Methods

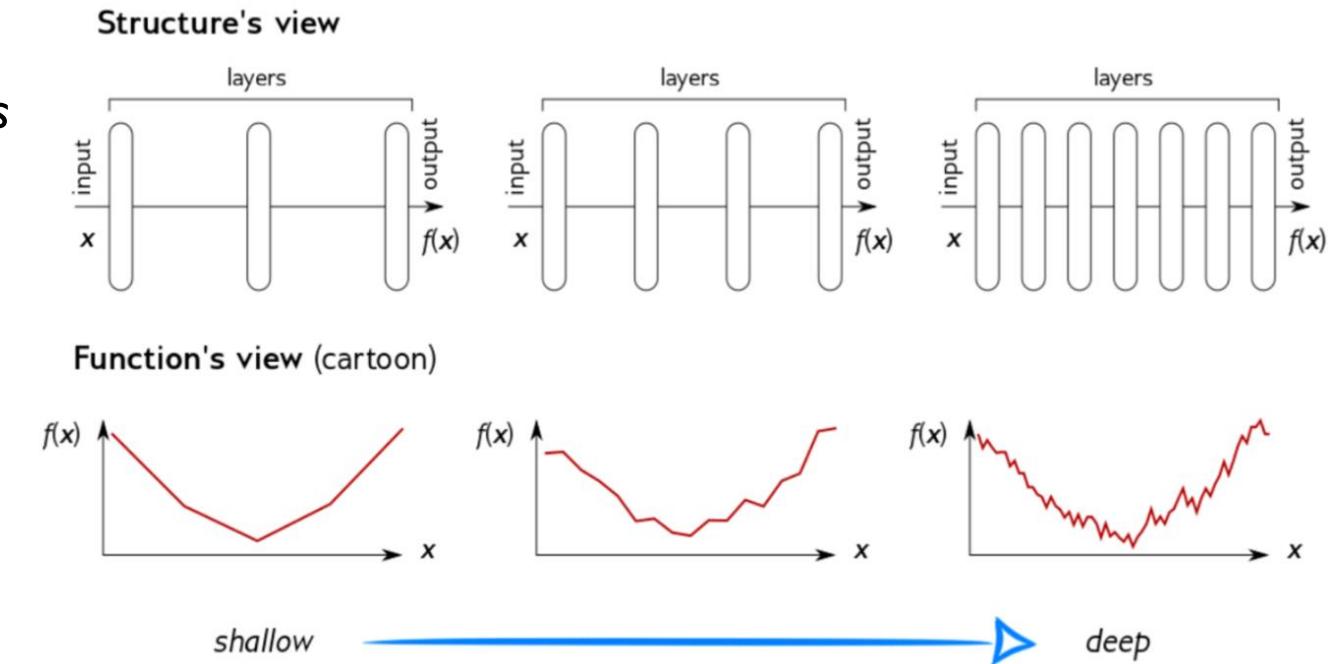
Gradient Based

Backprop. Based

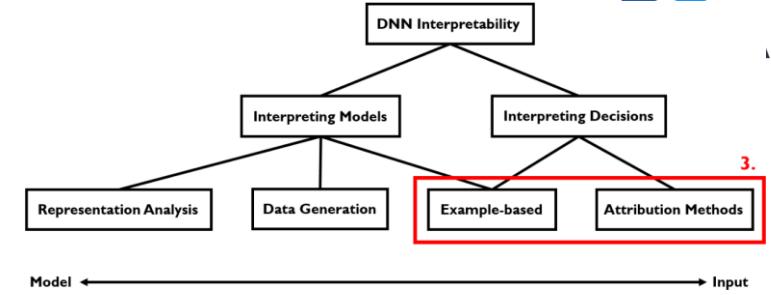
Question: Why are saliency maps noisy?

Hypothesis 2 – Gradients are discontinuous

- DNN uses piecewise-linear functions (ReLU activation, max-pooling, etc.).
- Sudden jumps in the importance score over infinitesimal changes in the input.



Types of DNN Interpretability



Example-based

Attribution Methods

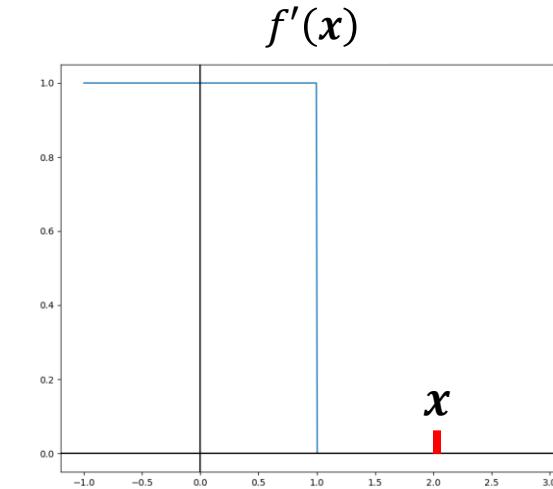
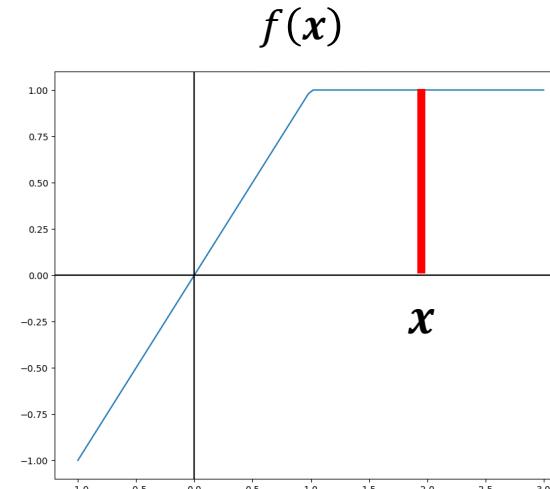
Gradient Based

Backprop. Based

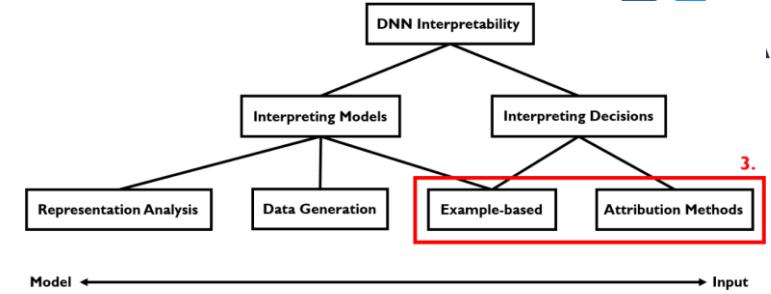
Question: Why are saliency maps noisy?

Hypothesis 3 – $f(x)$ saturates

- A feature may have a strong effect globally, but with a small derivative locally.



Types of DNN Interpretability



Example-based

Attribution Methods

Gradient Based

Backprop. Based

Question: How to improve saliency maps?

$$\text{Saliency}(\mathbf{x}) := \nabla_{\mathbf{x}} f(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$$

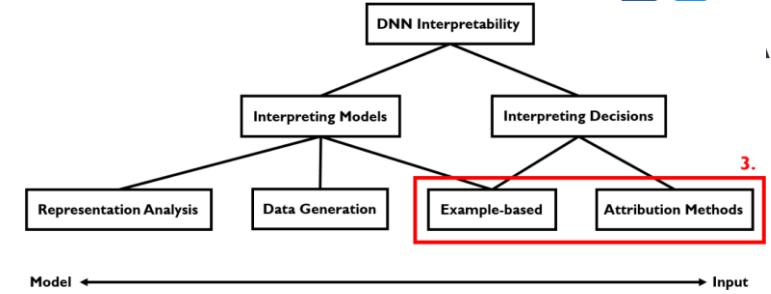
Gradient-based Methods

- Perturb the input \mathbf{x} to \mathbf{x}^* and use $\nabla_{\mathbf{x}^*} f(\mathbf{x}^*)$.
- Some methods take the average over the perturbation set $\{\mathbf{x}_1^*, \mathbf{x}_2^*, \dots, \mathbf{x}_n^*\}$.

Backprop-based Methods

- Modify the backpropagation algorithm.

Types of DNN Interpretability



Example-based

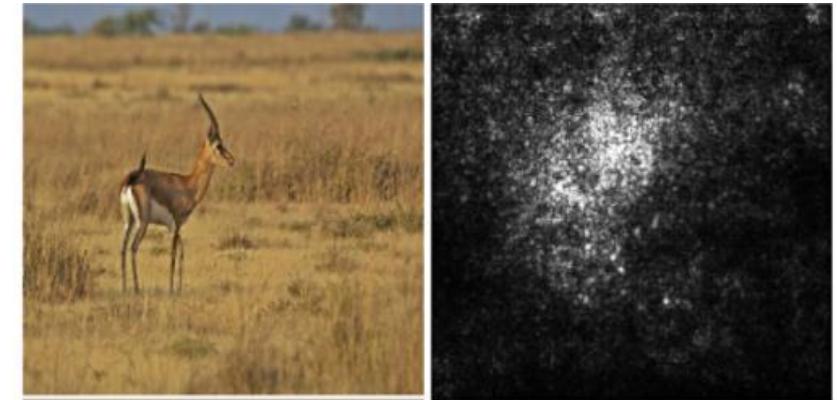
Attribution Methods

Gradient Based

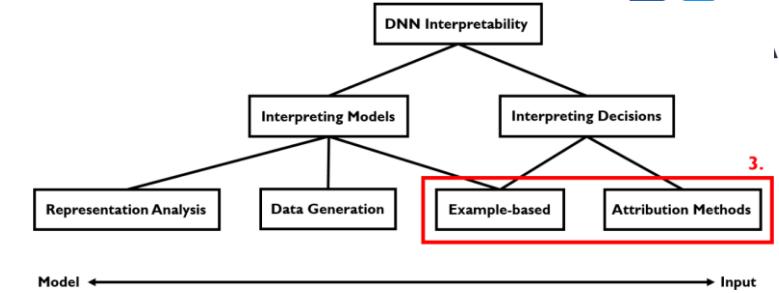
Backprop. Based

Summary

- Attribution method assigns “attribution score” to each input pixel.
- Baseline attribution method Saliency Map is noisy.
- Hypothesis 1: Saliency maps are truthful.
- Hypothesis 2: Gradients are discontinuous.
- Hypothesis 3: $f(x)$ saturates.
- Two solution approaches: Gradient based method and Backprop. based method.



Types of DNN Interpretability



Example-based

Attribution Methods

Gradient Based

Backprop. Based

I. SmoothGrad Hypothesis 2 – Gradients are discontinuous

$$\text{SmoothGrad}(\mathbf{x}) := \frac{1}{n} \sum_1^n \frac{\partial f(\mathbf{x}^*)}{\partial \mathbf{x}^*}, \quad \mathbf{x}^* = \mathbf{x} + \mathcal{N}(0, \sigma^2)$$

Gaussian smoothing

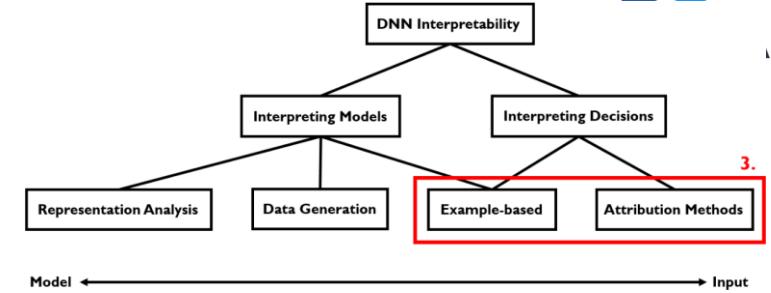
Function's view (cartoon)



shallow

deep

Types of DNN Interpretability



Example-based

Attribution Methods

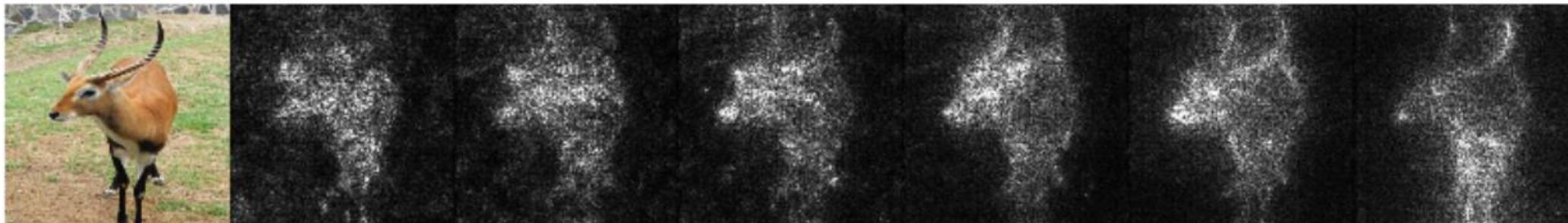
Gradient Based

Backprop. Based

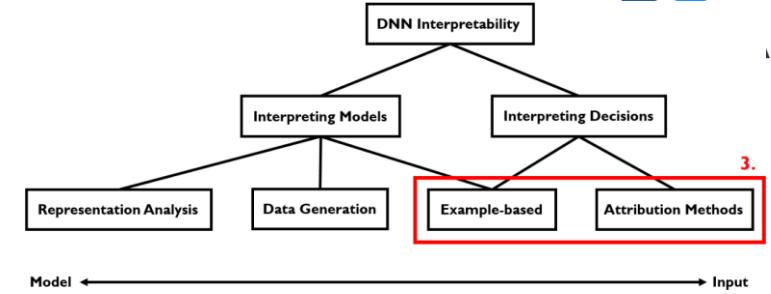
I. SmoothGrad *Hypothesis 2 – Gradients are discontinuous*

$$\text{SmoothGrad}(\mathbf{x}) := \frac{1}{n} \sum_1^n \frac{\partial f(\mathbf{x}^*)}{\partial \mathbf{x}^*}, \quad \mathbf{x}^* = \mathbf{x} + \mathcal{N}(0, \sigma^2)$$

Noise level:



Types of DNN Interpretability



Example-based

Attribution Methods

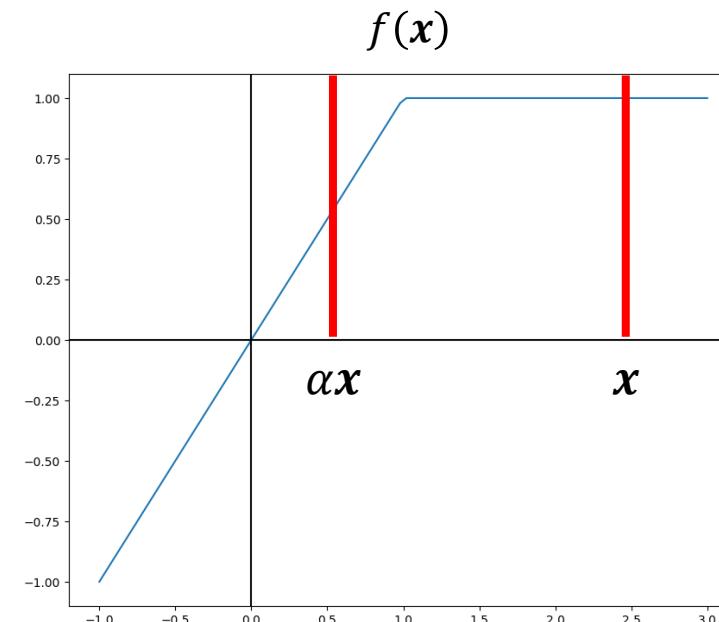
Gradient Based

Backprop. Based

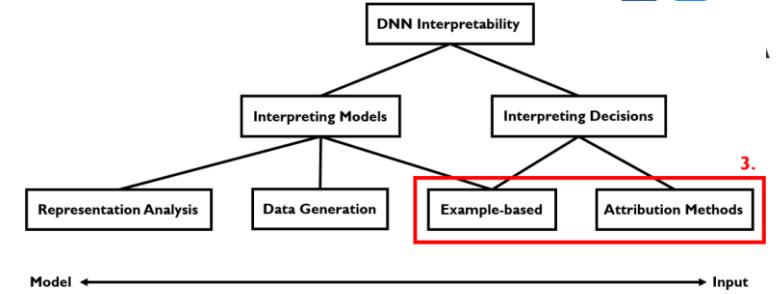
2. Interior Gradient Hypothesis 3 – $f(x)$ saturates

$$IntGrad(x) := \frac{\partial f(x^*)}{\partial x^*}, \quad x^* = \alpha x, \quad 0 < \alpha \leq 1$$

- Appropriate α will trigger informative activation functions



Types of DNN Interpretability



Example-based

Attribution Methods

Gradient Based

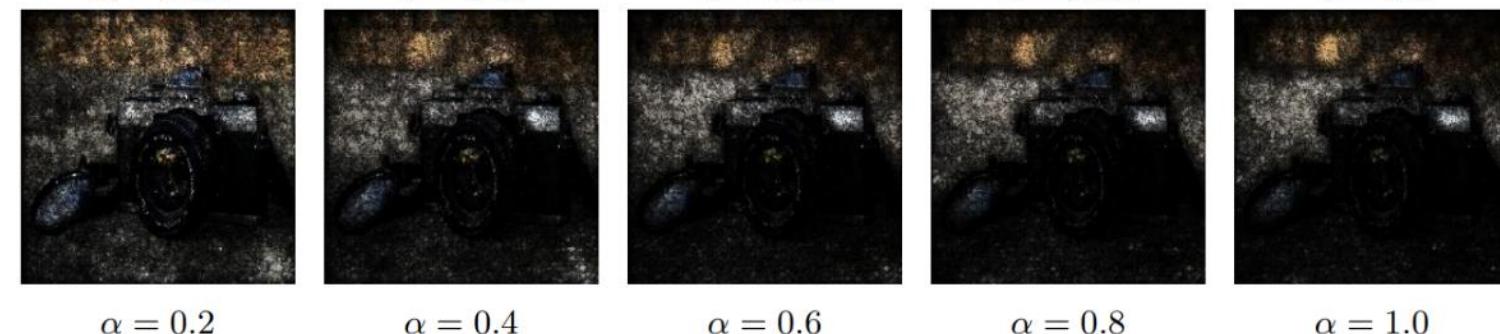
Backprop. Based

2. Interior Gradient

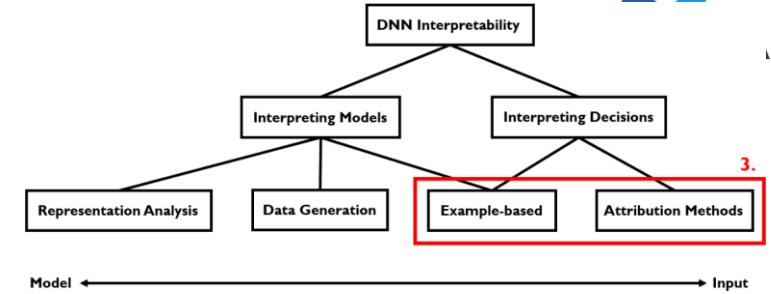
$$IntGrad(x) := \frac{\partial f(x^*)}{\partial x^*},$$

$$x^* = \alpha x,$$

$$0 < \alpha \leq 1$$



Types of DNN Interpretability



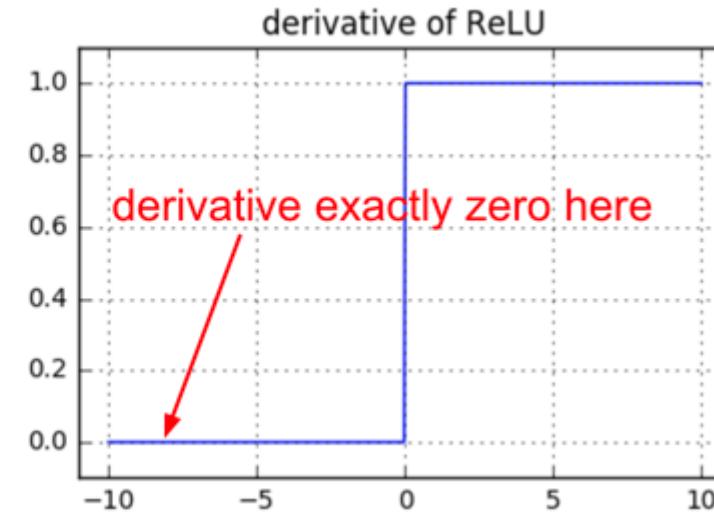
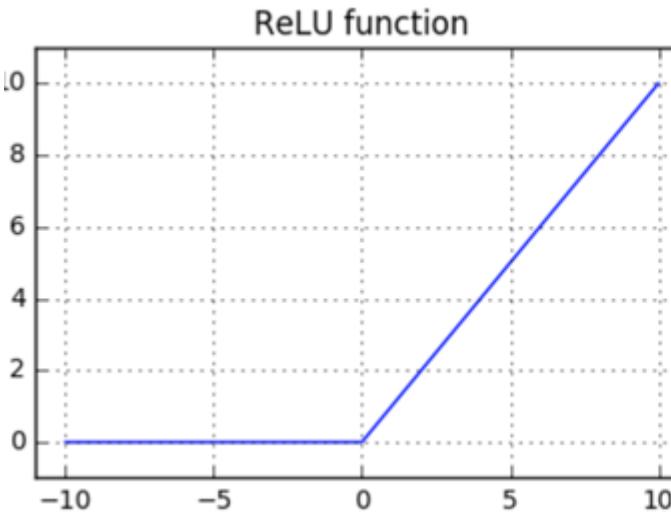
Example-based

Attribution Methods

Gradient Based

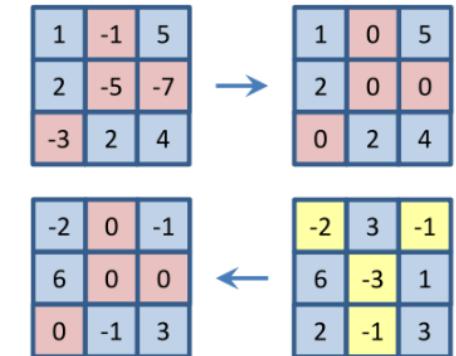
Backprop. Based

Review: Backpropagation at ReLU

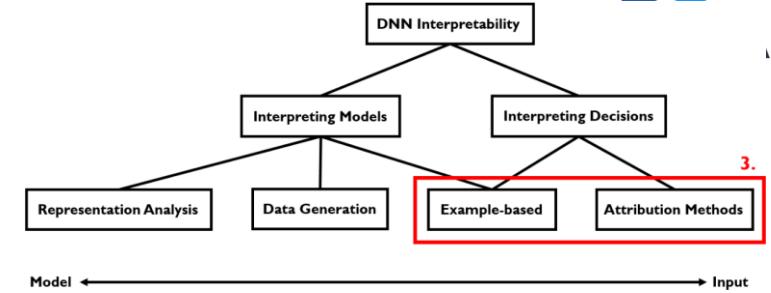


Forward pass

Backward pass:
backpropagation



Types of DNN Interpretability



Example-based

Attribution Methods

Gradient Based

Backprop. Based

I. Deconvnet

- Maps feature pattern to input space (image reconstruction)
- To obtain valid feature reconstruction, pass the reconstructed signal through ReLUs
- Removing noise by removing negative gradient

Forward pass

1	-1	5
2	-5	-7
-3	2	4

→

1	0	5
2	0	0
0	2	4

Backward pass:
backpropagation

-2	0	-1
6	0	0
0	-1	3

←

-2	3	-1
6	-3	1
2	-1	3

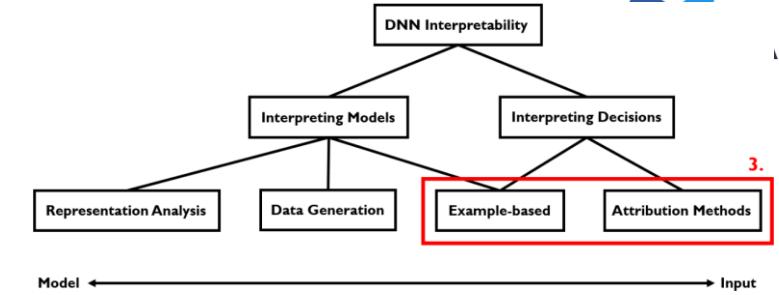
Backward pass:
“deconvnet”

0	3	0
6	0	1
2	0	3

←

-2	3	-1
6	-3	1
2	-1	3

Types of DNN Interpretability



Example-based

Attribution Methods

Gradient Based

Backprop. Based

2. Guided Backpropagation

- Combine Deconvnet with Backpropagation
- Removing negative gradient + consider forward activations

Forward pass

$$\begin{matrix} 1 & -1 & 5 \\ 2 & -5 & -7 \\ -3 & 2 & 4 \end{matrix} \rightarrow \begin{matrix} 1 & 0 & 5 \\ 2 & 0 & 0 \\ 0 & 2 & 4 \end{matrix}$$

Backward pass:
backpropagation

$$\begin{matrix} -2 & 0 & -1 \\ 6 & 0 & 0 \\ 0 & -1 & 3 \end{matrix} \leftarrow \begin{matrix} -2 & 3 & -1 \\ 6 & -3 & 1 \\ 2 & -1 & 3 \end{matrix}$$

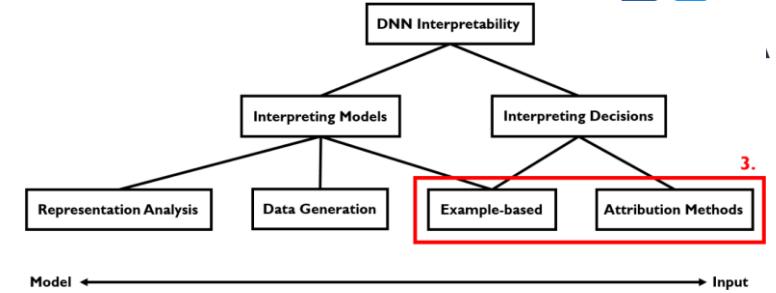
Backward pass:
“deconvnet”

$$\begin{matrix} 0 & 3 & 0 \\ 6 & 0 & 1 \\ 2 & 0 & 3 \end{matrix} \leftarrow \begin{matrix} -2 & 3 & -1 \\ 6 & -3 & 1 \\ 2 & -1 & 3 \end{matrix}$$

Backward pass:
*guided
backpropagation*

$$\begin{matrix} 0 & 0 & 0 \\ 6 & 0 & 0 \\ 0 & 0 & 3 \end{matrix} \leftarrow \begin{matrix} -2 & 3 & -1 \\ 6 & -3 & 1 \\ 2 & -1 & 3 \end{matrix}$$

Types of DNN Interpretability



Example-based

Attribution Methods

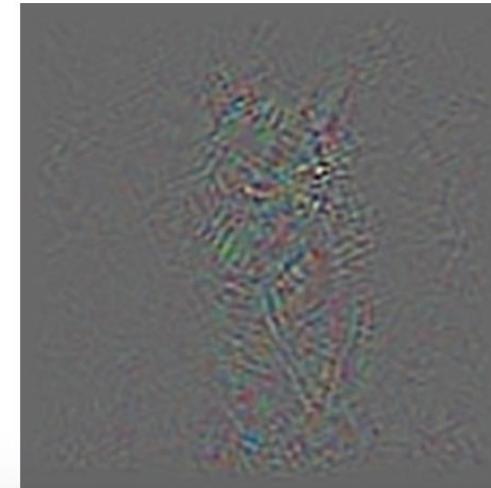
Gradient Based

Backprop. Based

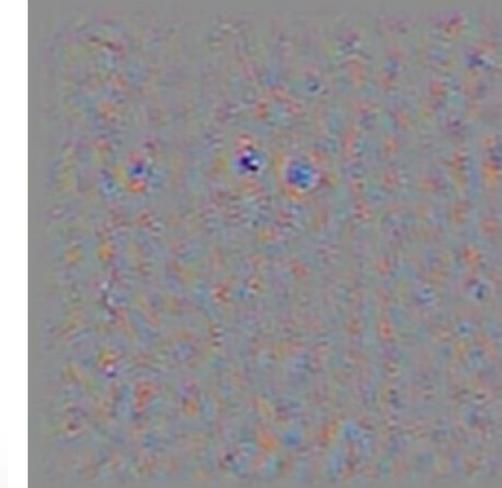
Input image



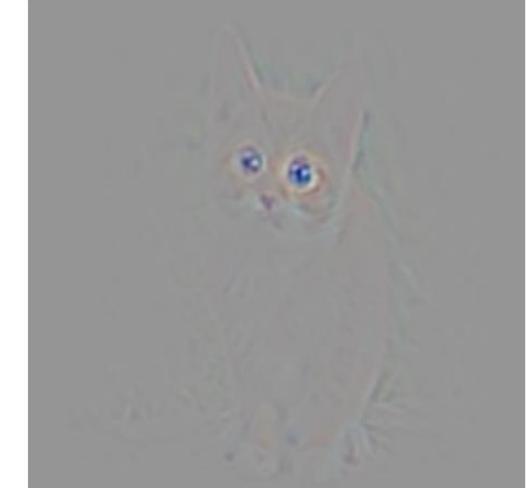
Backpropagation



Deconvolution

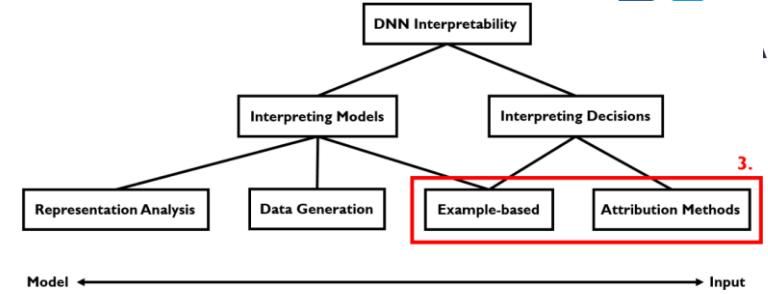


Guided Backprop



Observation: Removing **more** gradient leads to **sharper** visualizations

Types of DNN Interpretability



Example-based

Attribution Methods

Gradient Based

Backprop. Based

Other Attribution Methods

- Gradient * Input – <https://arxiv.org/pdf/1704.02685.pdf>
- Integrated Gradient – <https://arxiv.org/pdf/1703.01365.pdf>
- Layer-wise Relevance Propagation – <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130140>
- Deep Taylor Decomposition – <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130140>
- DeepLIFT – <https://arxiv.org/pdf/1704.02685.pdf>
- PatternNet and PatternAttribution – <https://arxiv.org/pdf/1705.05598.pdf>

Summary

1. Interpreting Models vs. Interpreting Decisions

- Interpreting models: macroscopic view, better understand internal representations
- Interpreting decision: microscopic view, important for practical applications

2. Interpreting Models

- Weight visualization
- Surrogate model
- Activation maximization
- Example-based

3. Interpreting Decisions

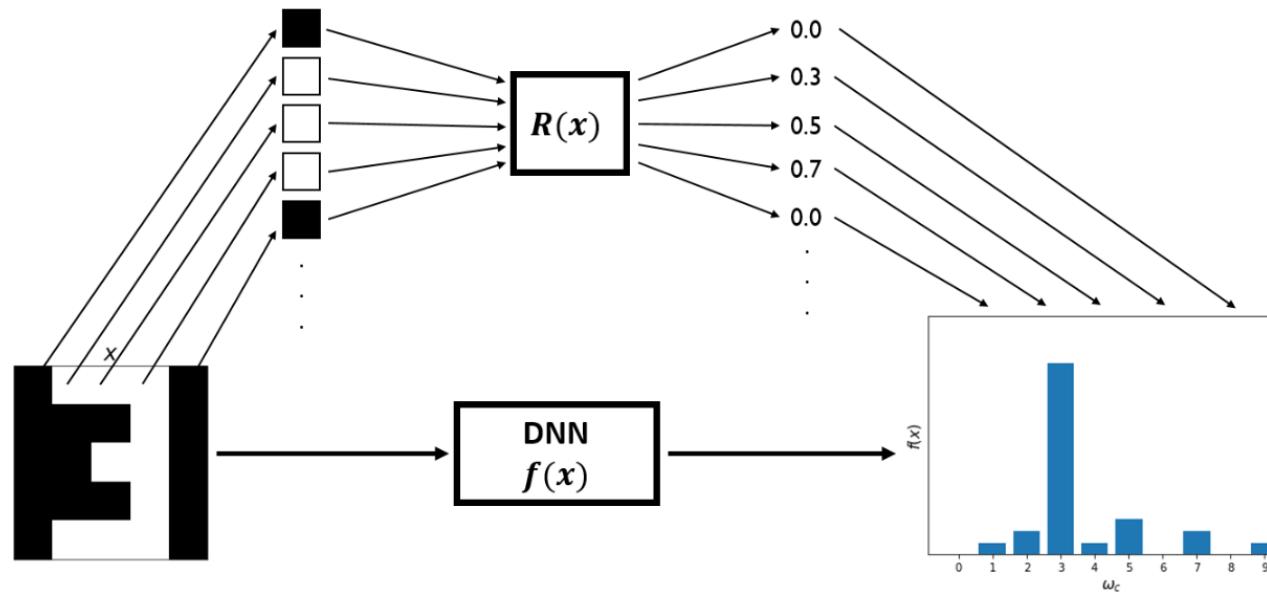
- Example-based
- Attribution Methods: why are gradients noisy?
- Gradient based Attribution Methods: SmoothGrad, Interior Gradient
- Backprop. based Attribution Methods: Deconvolution, Guided Backpropagation

04

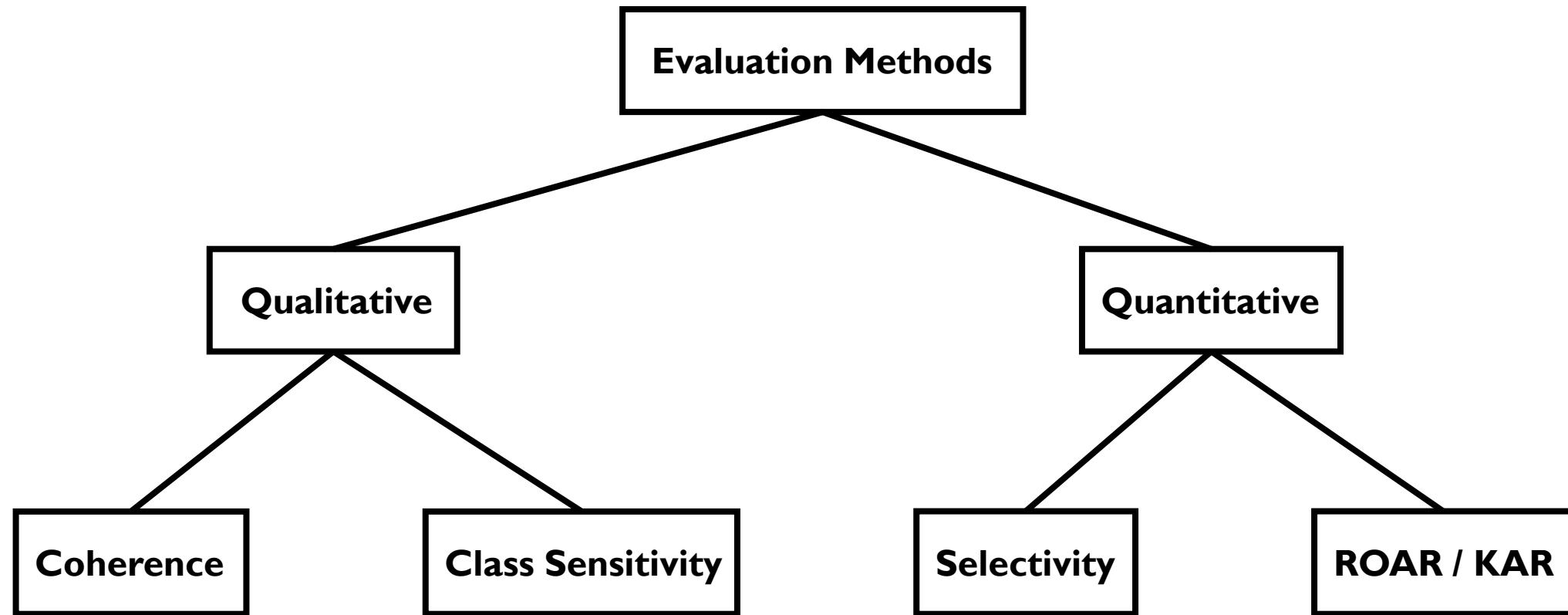
Evaluation for Attribution Methods

Attribution Method Review

Given an image $x \in \mathbb{R}^n$ and a decision $f(x)$,
assign to each pixel x_1, x_2, \dots, x_n **attribution values** $R_1(x), R_2(x), \dots, R_n(x)$.



Evaluation for Attribution Methods



Evaluation for Attribution Methods

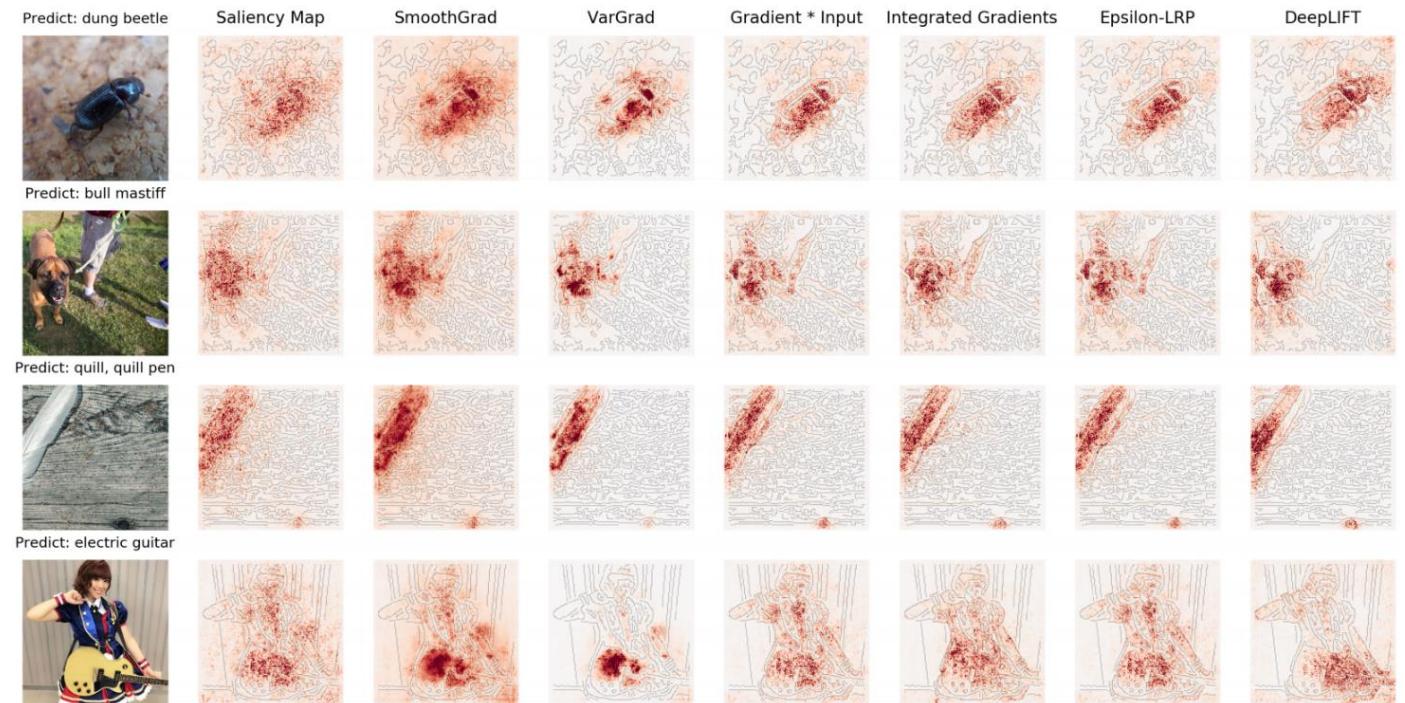
Coherence

- Attributions should fall on discriminative features (e.g. the object of interest)

Class Sensitivity

Selectivity

ROAR / KAR



Evaluation for Attribution Methods

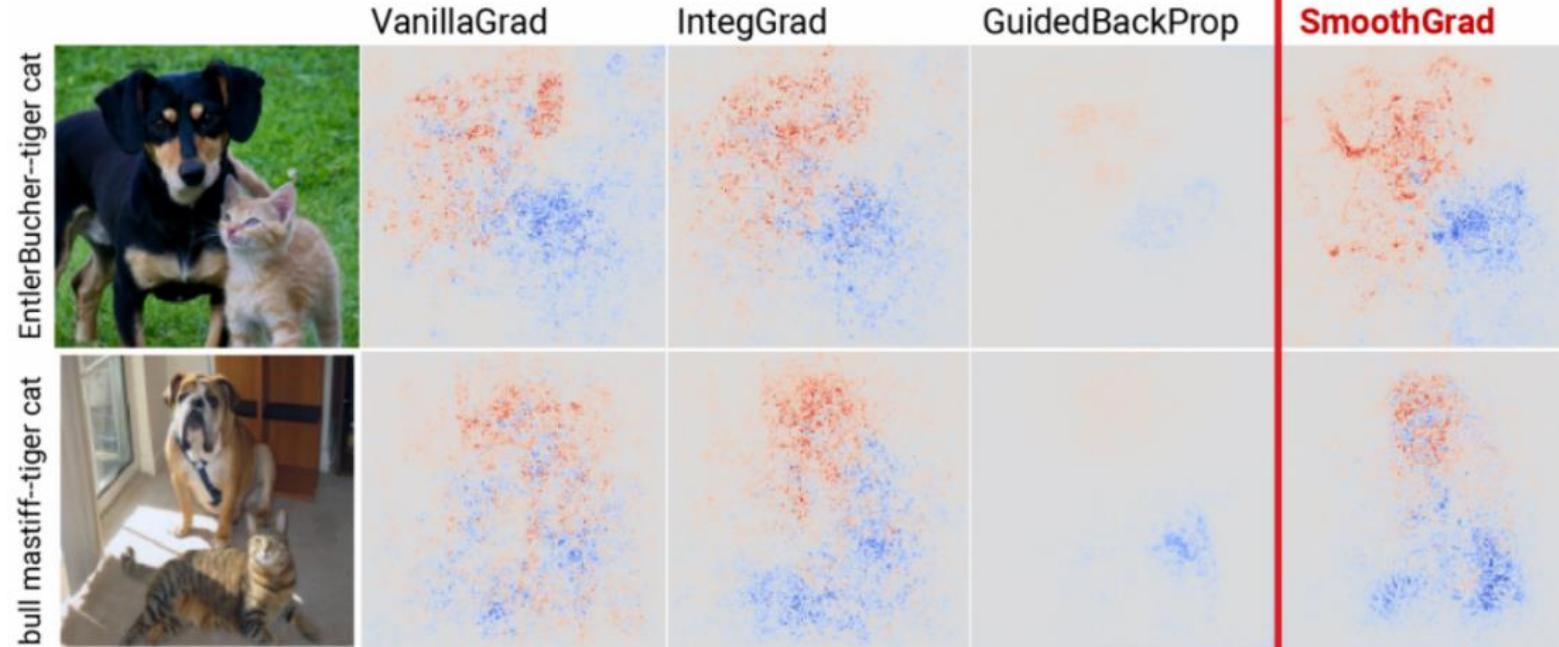
Coherence

Class Sensitivity

Selectivity

ROAR / KAR

- Attributions should be sensitive to class labels



Evaluation for Attribution Methods

Coherence

Class Sensitivity

Selectivity

ROAR / KAR

- Removing feature with high attribution should cause large decrease in class probability

Algorithm

Sort pixel attribution values $R_i(x)$

Iterate:

 Remove pixels

 Evaluate $f(x)$

Measure decrease of $f(x)$

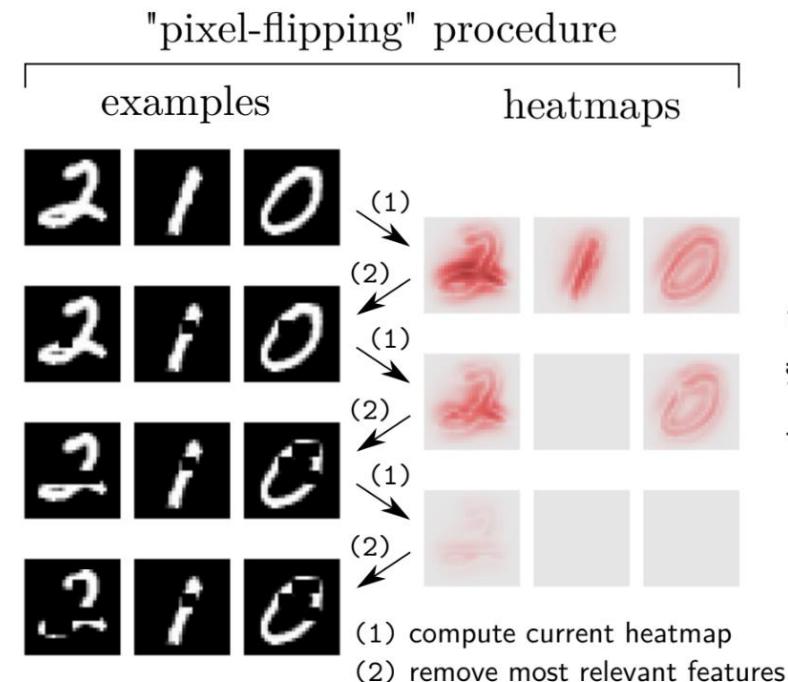
Evaluating Attribution Methods

Coherence

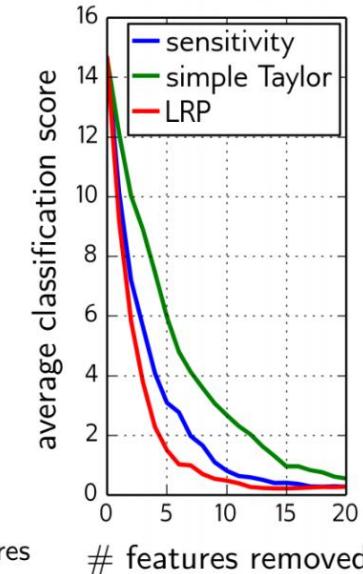
Class Sensitivity

Selectivity

ROAR / KAR

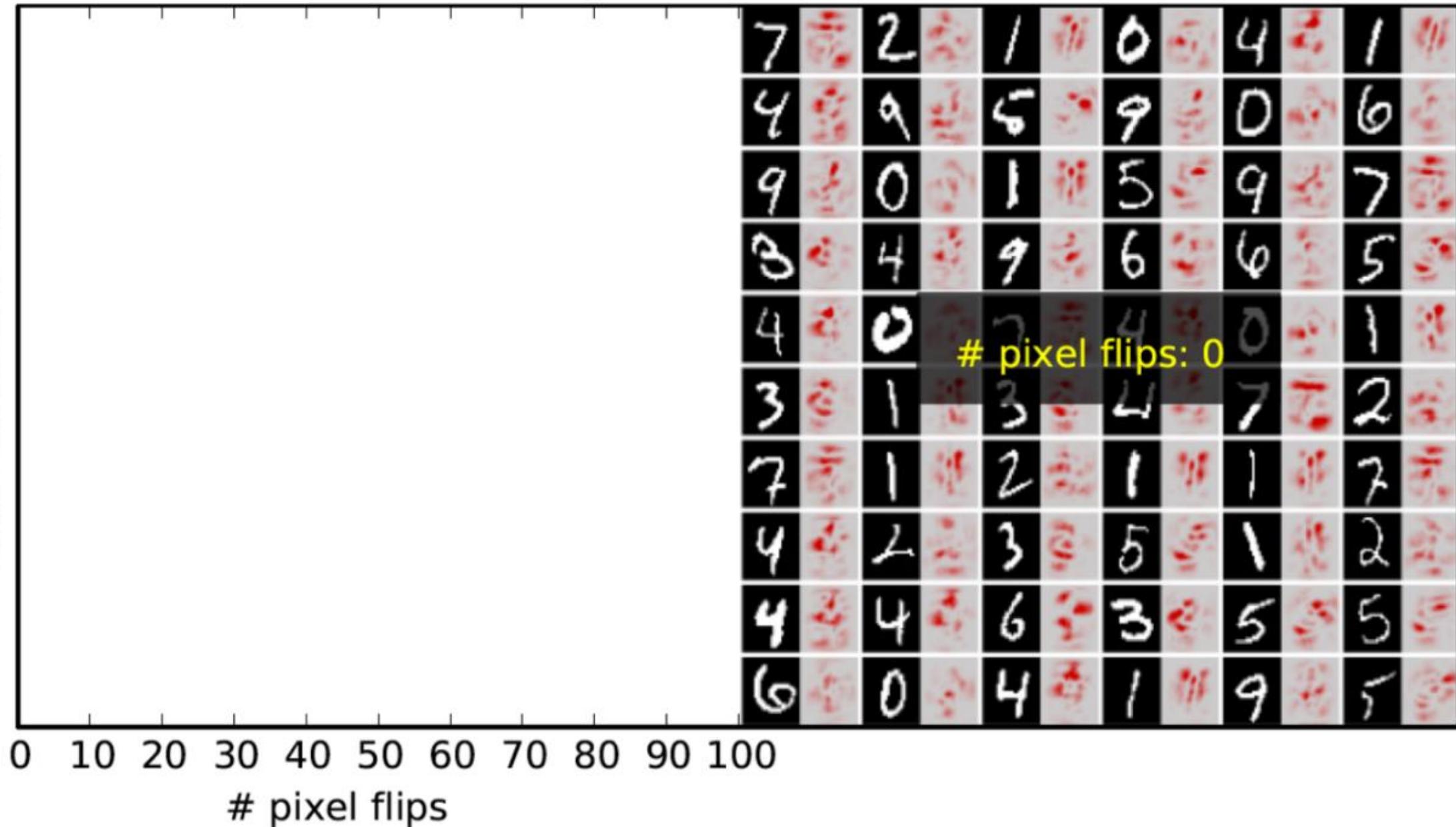


comparing explanation techniques



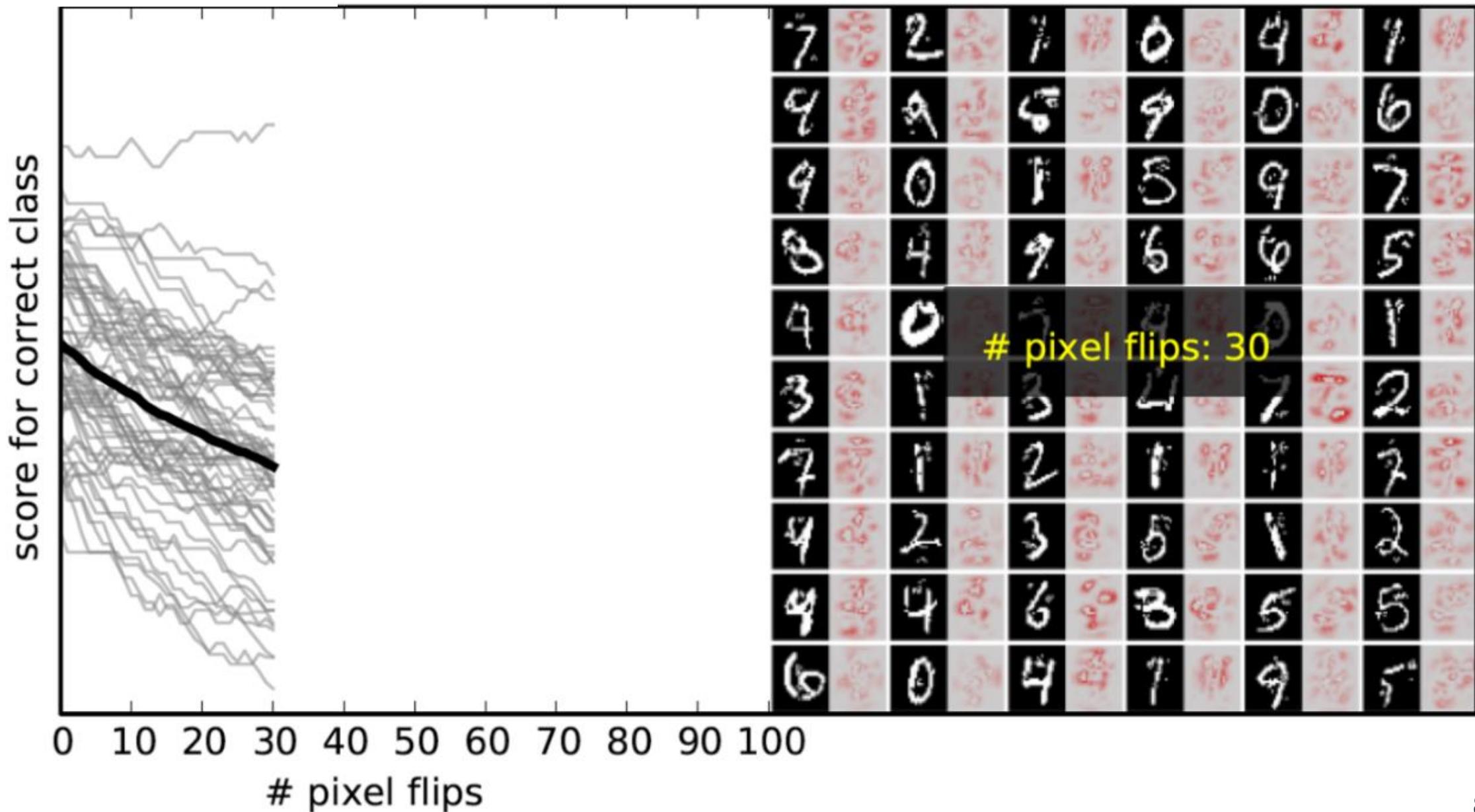
Selectivity on Saliency Map

score for correct class



Selectivity on Saliency Map

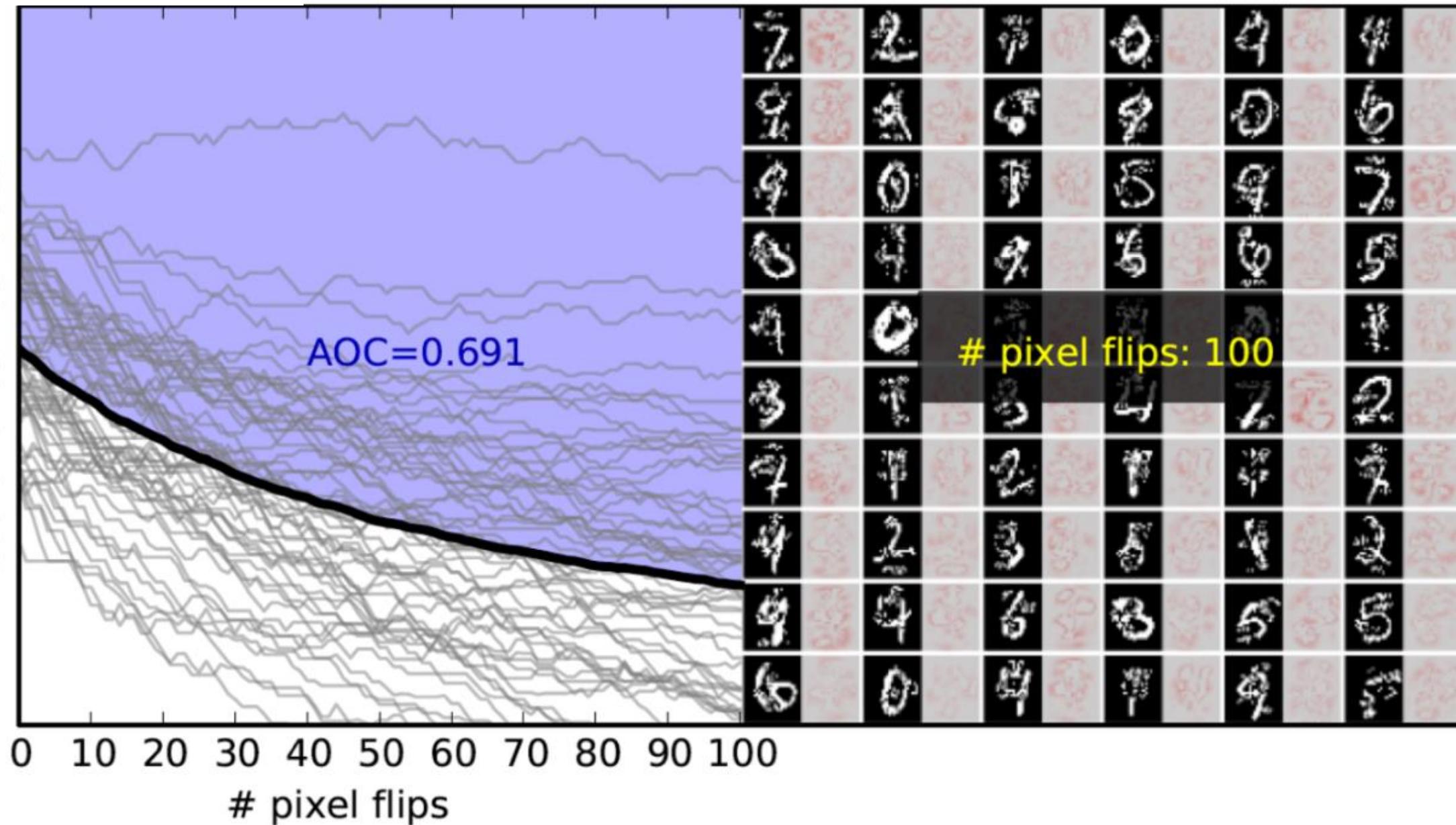
SIA



Selectivity on Saliency Map

SIA

score for correct class



Evaluating Attribution Methods

Coherence

Class Sensitivity

Selectivity

ROAR / KAR

- Sensitivity may not be accurate
- Class probability may decrease because the DNN has never seen such image

Remove and Retrain (ROAR) / Keep and Retrain (KAR)

Measure how the performance of the classifier changes as features are removed based on the attribution method

- ROAR: replace $N\%$ of pixels estimated to be *most* important
- KAR: replace $N\%$ of pixels estimated to be *least* important
- Retrain DNN and measure change in test accuracy

Evaluating Attribution Methods

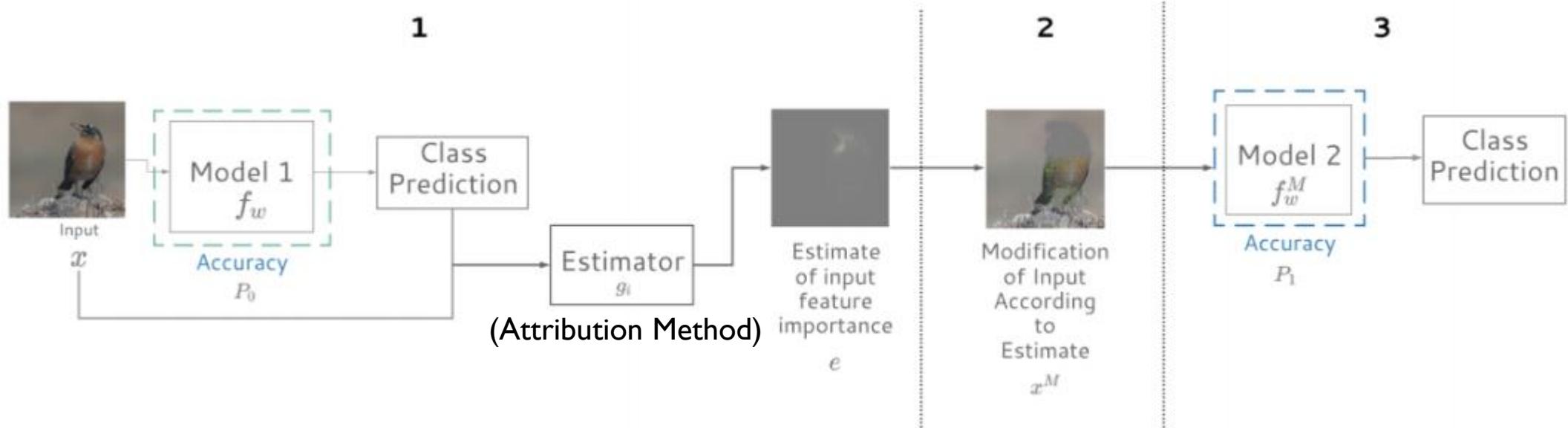
Coherence

Class Sensitivity

Selectivity

ROAR / KAR

ROAR – RemOve And Retrain



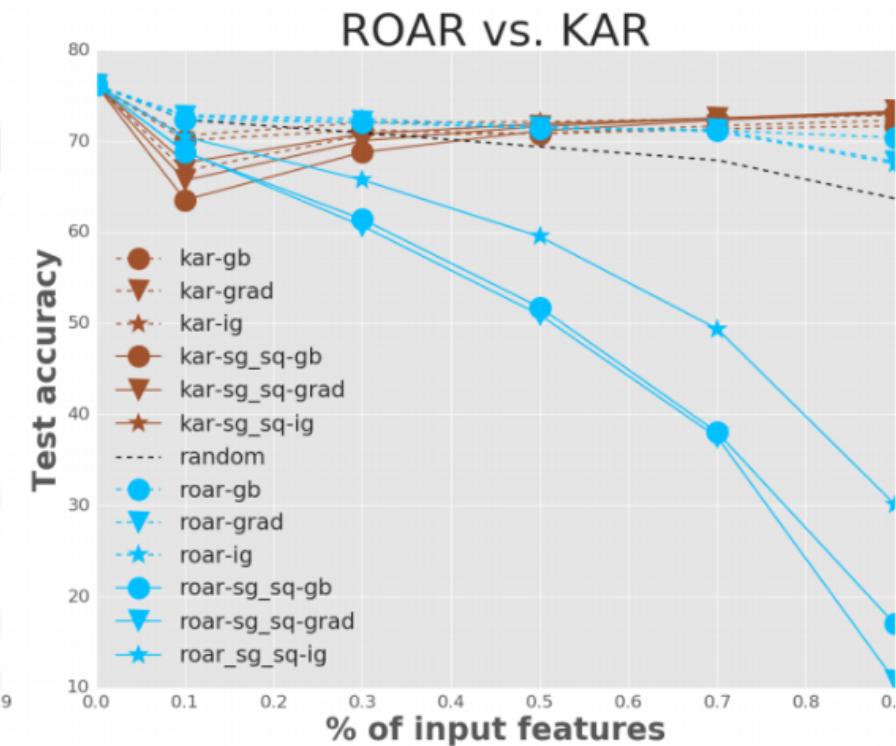
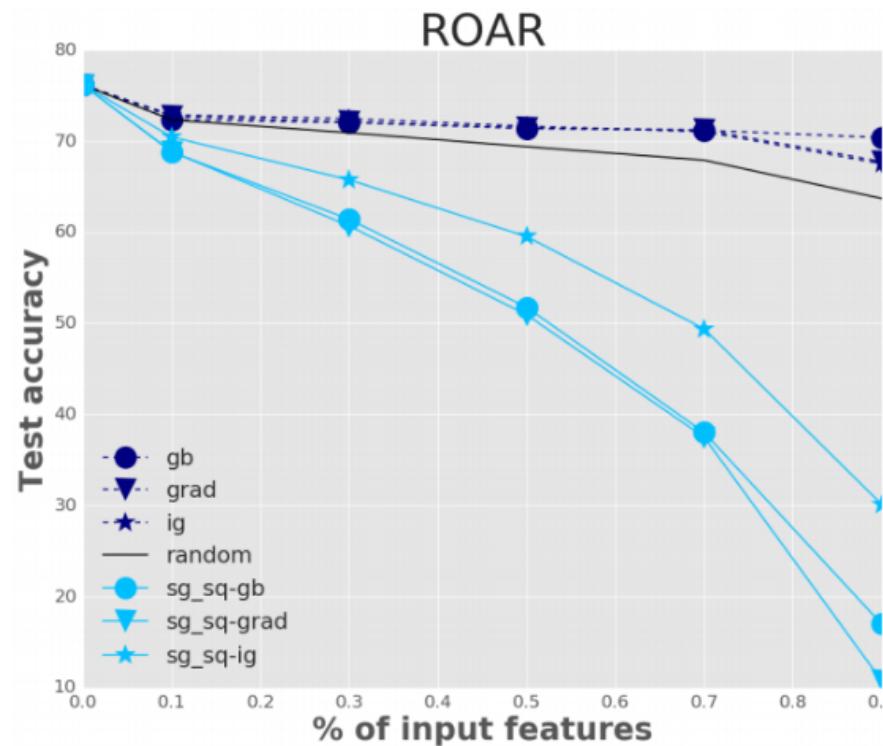
Evaluating Attribution Methods

Coherence

Class Sensitivity

Selectivity

ROAR / KAR



Part 3 Summary

1. Qualitative: Coherence

- Attributions should highlight discriminative features / objects of interest

2. Qualitative: Class Sensitivity

- Attributions should be sensitive to class labels

3. Quantitative: Sensitivity

- Removing feature with high attribution should cause large decrease in class probability

4. Quantitative: ROAR & KAR

- Problem: class probability may decrease because the DNN has never seen such image
- Solution: remove pixels, retrain and measure drop in accuracy

05

Conclusions

Summary

- **I. Introduction to Interpretability**

- Interpretability is converting implicit information in DNN to (human) interpretable information
- Ante-hoc Interpretability vs. Post-hoc Interpretability
- Post-hoc interpretability techniques can be classified by degree of “locality”

- **2. Interpretation of Deep Neural Networks**

- Interpreting Models vs. Interpreting Decisions
- Interpreting Models: weight visualization, surrogate model, activation maximization, example-based
- Interpreting Decisions: example-based, attribution methods

- **3. Evaluation for Attribution Methods**

- Qualitative Evaluation Methods: coherence, class sensitivity
- Quantitative Evaluation Methods: Sensitivity, ROAR & KAR

Rectified Gradient (RectGrad)

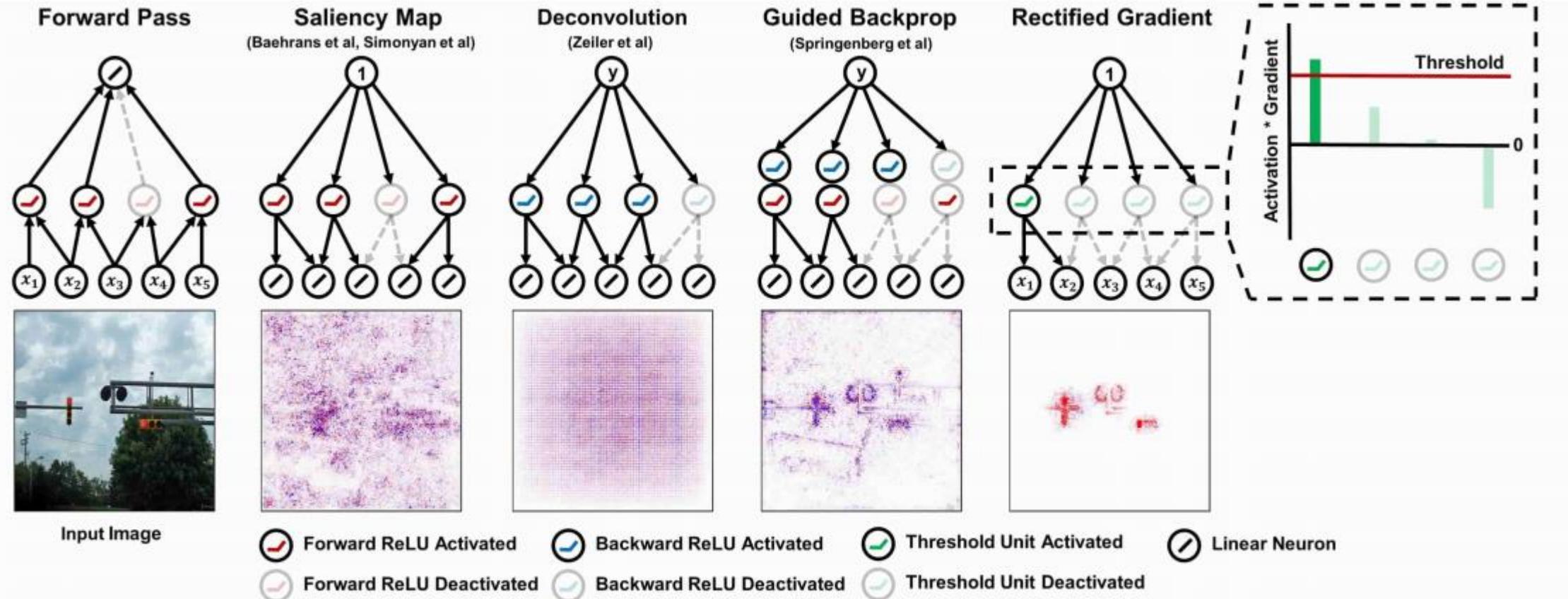
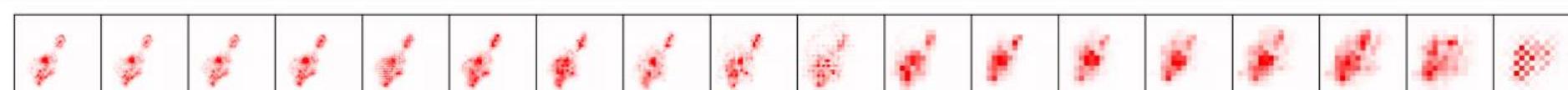
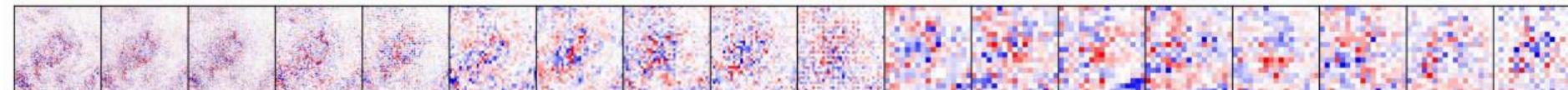
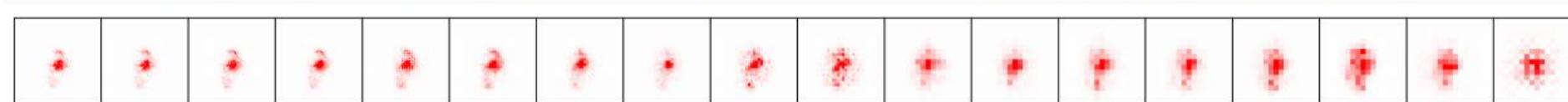
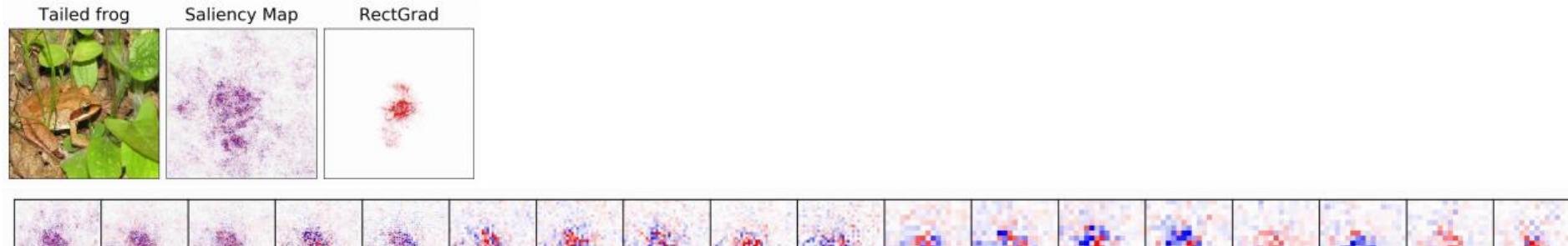


Figure 1. Comparison of attribution methods. See Appendix F.1 for details on the visualization.

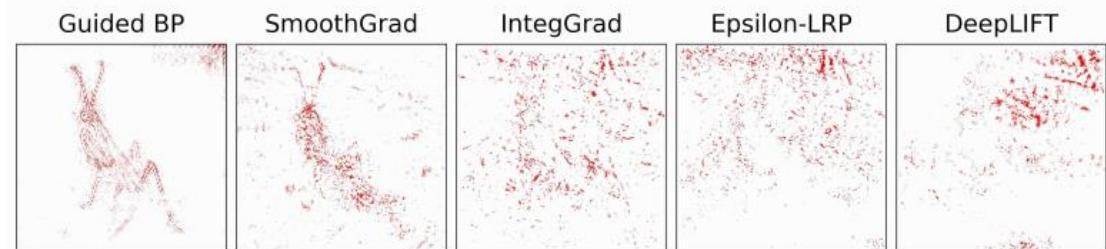
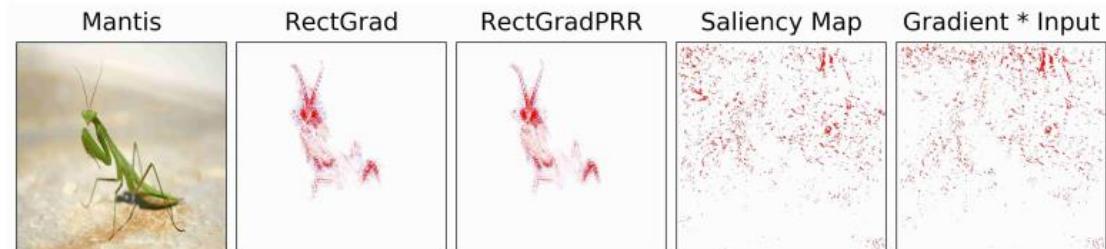
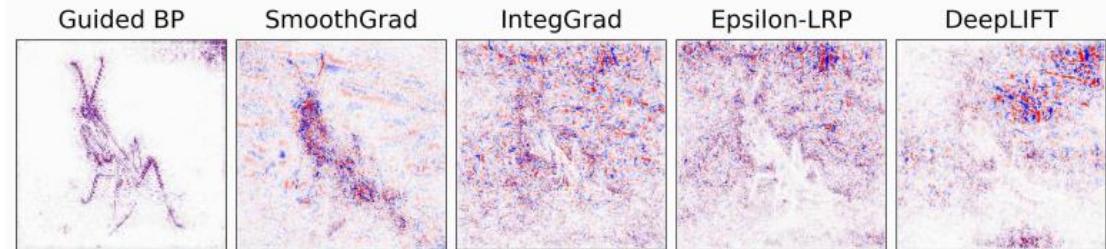
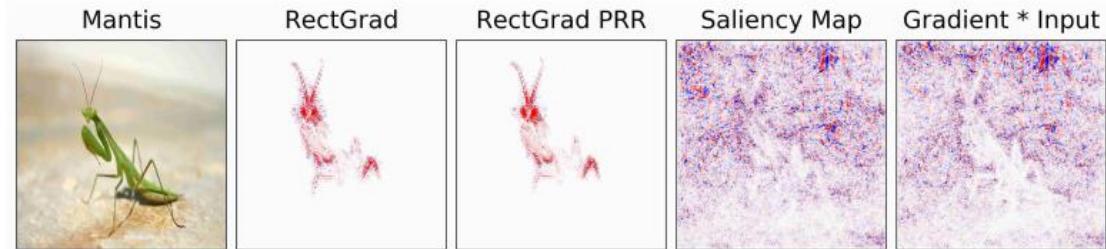
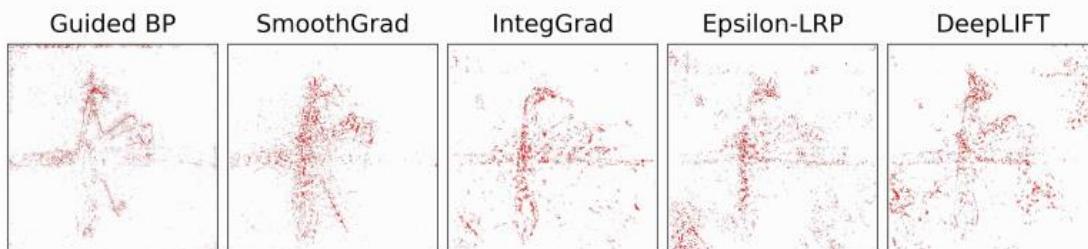
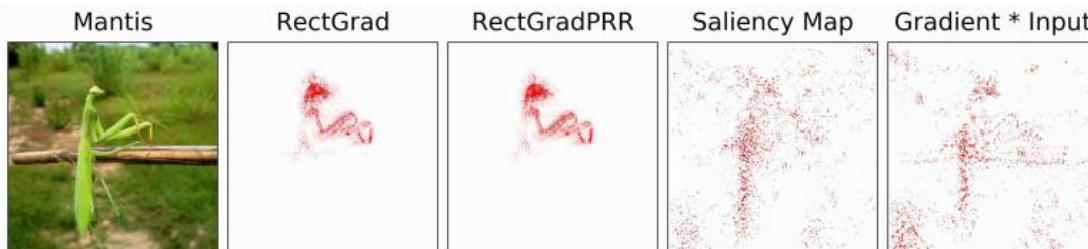
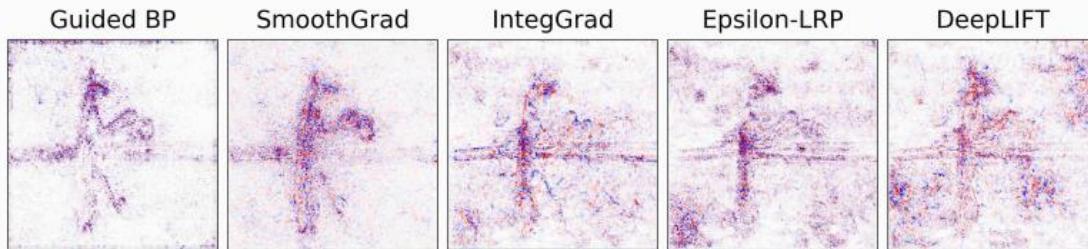
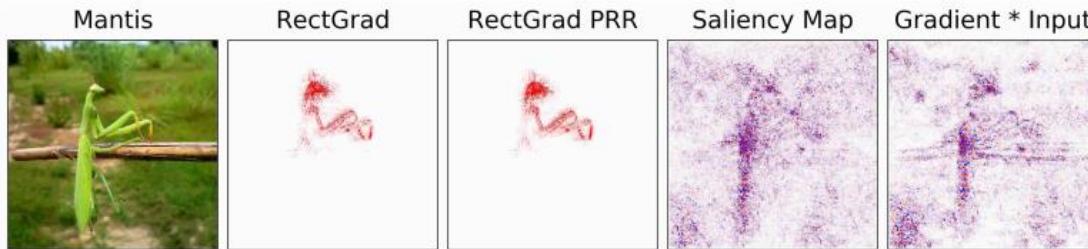
[Why are Saliency Maps Noisy? Cause of and Solution to Noisy Saliency Maps \(arXiv:1902.04893\)](#)

Rectified Gradient (RectGrad)



Why are Saliency Maps Noisy? Cause of and Solution to Noisy Saliency Maps (arXiv:1902.04893)

Rectified Gradient (RectGrad)





주식회사 에스아이에이
SI ANALYTICS

Thank you for attention!

SI Analytics Co., Ltd. (Satrec Initiative Group)
441 Expo-ro, Yuseong-gu, Daejeon, 34051, Korea

tgjeon@si-analytics.ai
+82-42-330-6800 / +82-10-6584-7321

www.si-analytics.ai