6.874 Spring 2021

Lecture 05

Interpretable Deep Learning

Prof. Manolis Kellis

Slides by Beomsu Kim, KAIST

Interpretable Deep Learning

1. Intro to Interpretability

1a. Interpretability definition: Convert implicit NN information to human-interpretable information

1b. Motivation: Verify model works as intended; debug classifier; make discoveries; Right to explanation

1c. Ante-hoc (train interpretable model) vs. **Post-hoc** (interpret complex model; degree of "locality")

2. Interpreting Deep Neural Networks

2a. Interpreting Models (macroscopic, understand internals) vs. **decisions** (microscopic, practical applications)

2b. Interpreting Models: Weight visualization, Surrogate model, Activation maximization, Example-based

2c. Interpreting Decisions:

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

3. Evaluating Attribution Methods

3a. Qualitative: Coherence: Attributions should highlight discriminative features / objects of interest

3b. Qualitative: Class Sensitivity: Attributions should be sensitive to class labels

3c. Quantitative: Sensitivity: Removing feature with high attribution ➔ large decrease in class probability

3d. Quantitative: ROAR & KAR. Low class prob cuz image unseen ➔ remove pixels, retrain, measure acc. drop

Slides by Beomsu Kim, KAIST

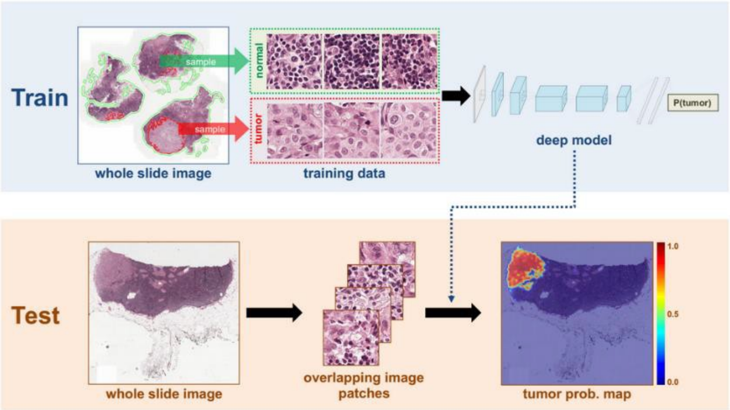
What is Interpretability?

AlphaGo vs. Lee Sedol

ImageNet Challenge

Self-driving Cars

Disease Diagnosis



Neural Machine Translation

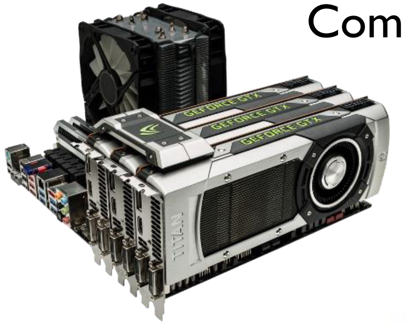
& More to Come!

Slides by Beomsu Kim, KAIST

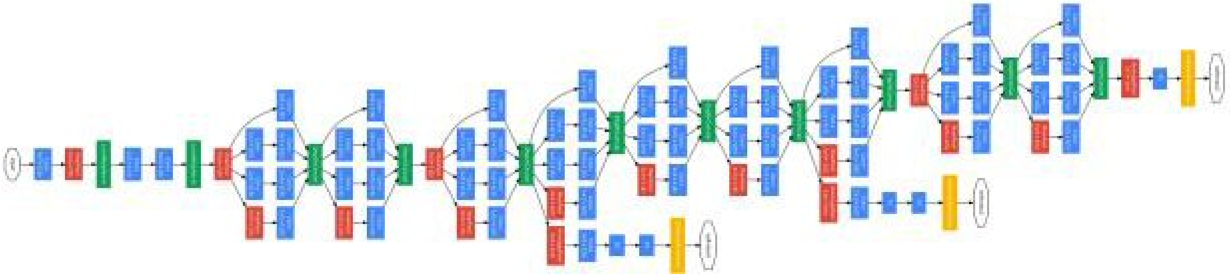
What is Interpretability?

Large Dataset

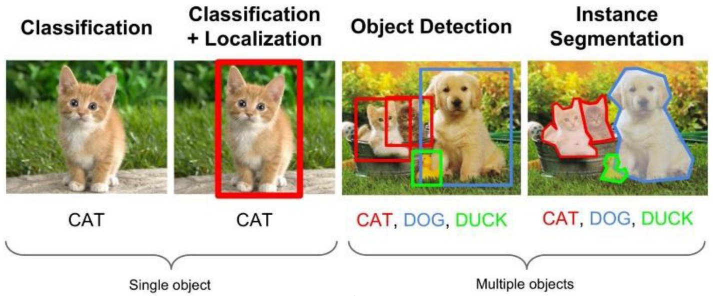
Computing Power

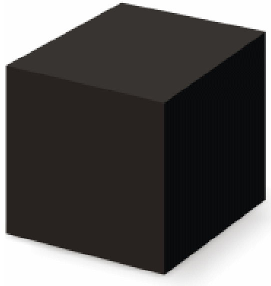


Deep Neural Networks



Task Solving

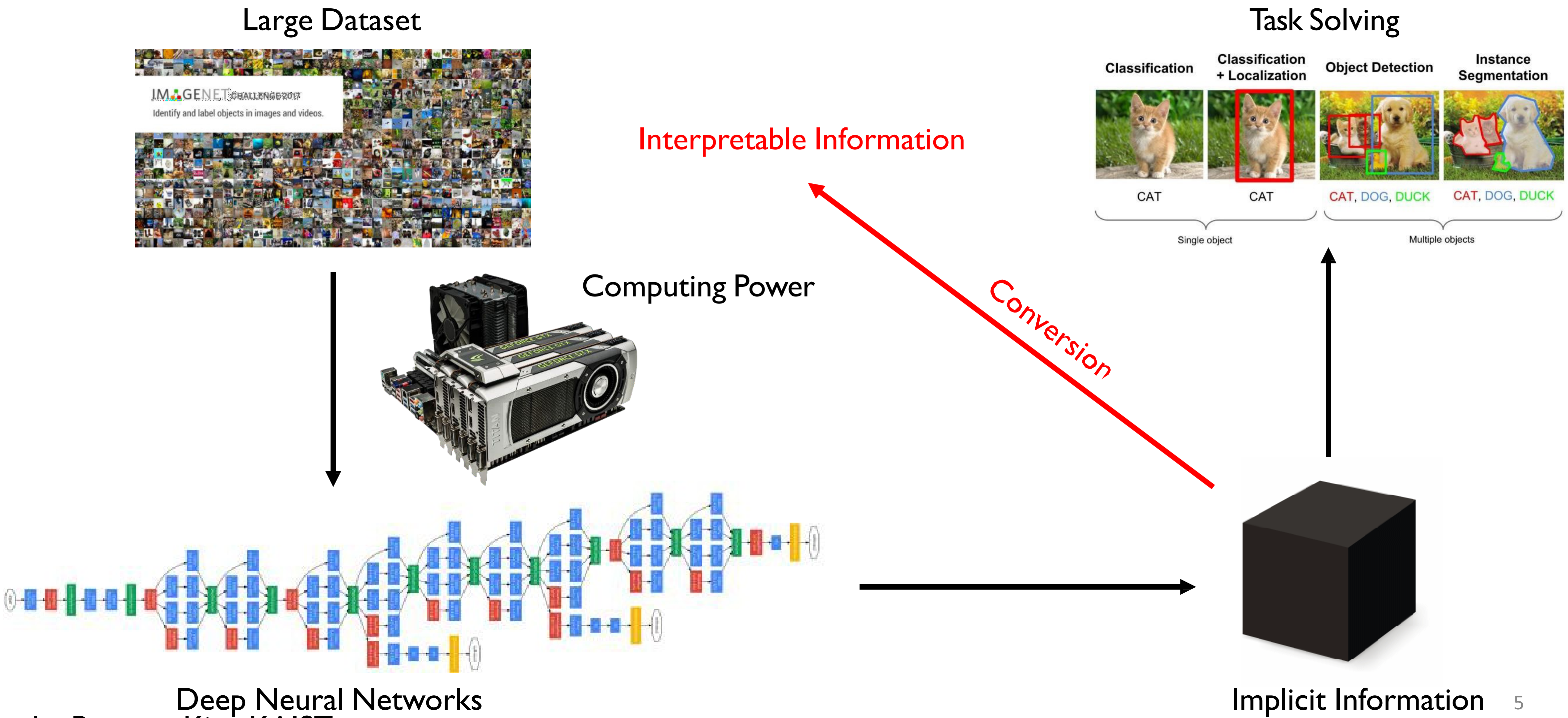




Implicit Information

Slides by Beomsu Kim, KAIST

What is Interpretability?



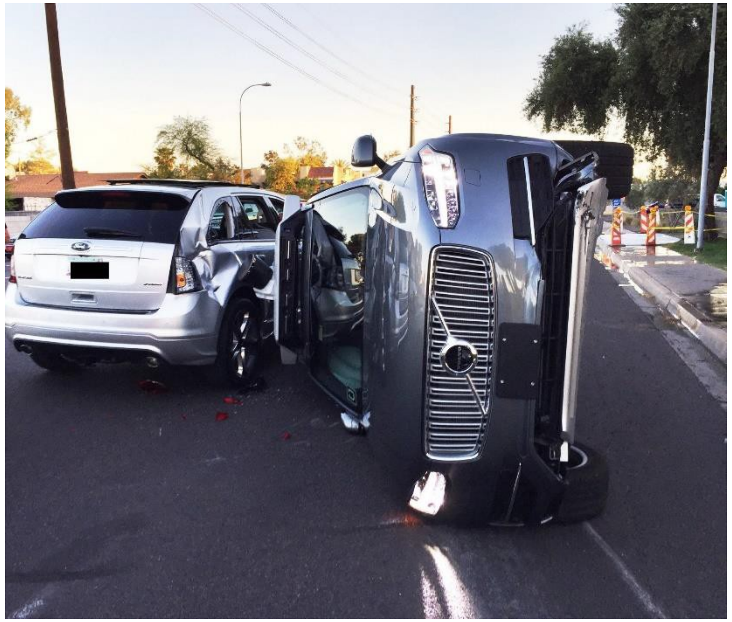
Slides by Beomsu Kim, KAIST

Why Interpretability?

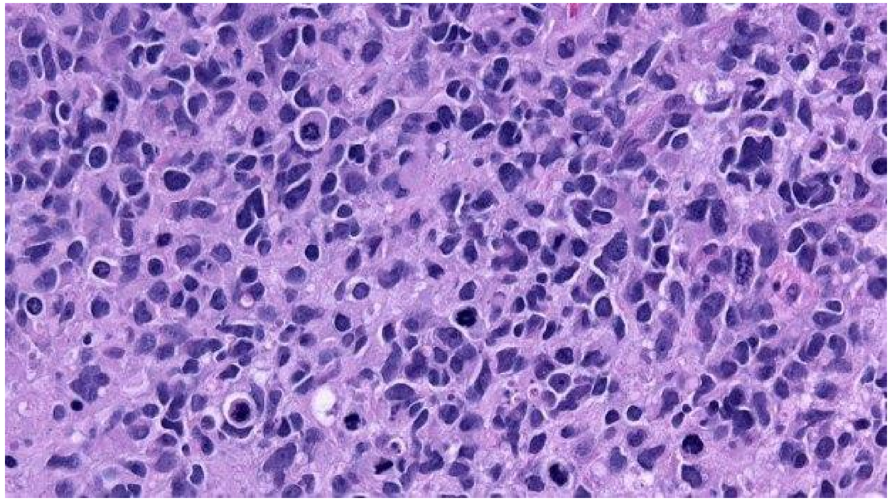
1. Verify that model works as expected

Wrong decisions can be costly and dangerous

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

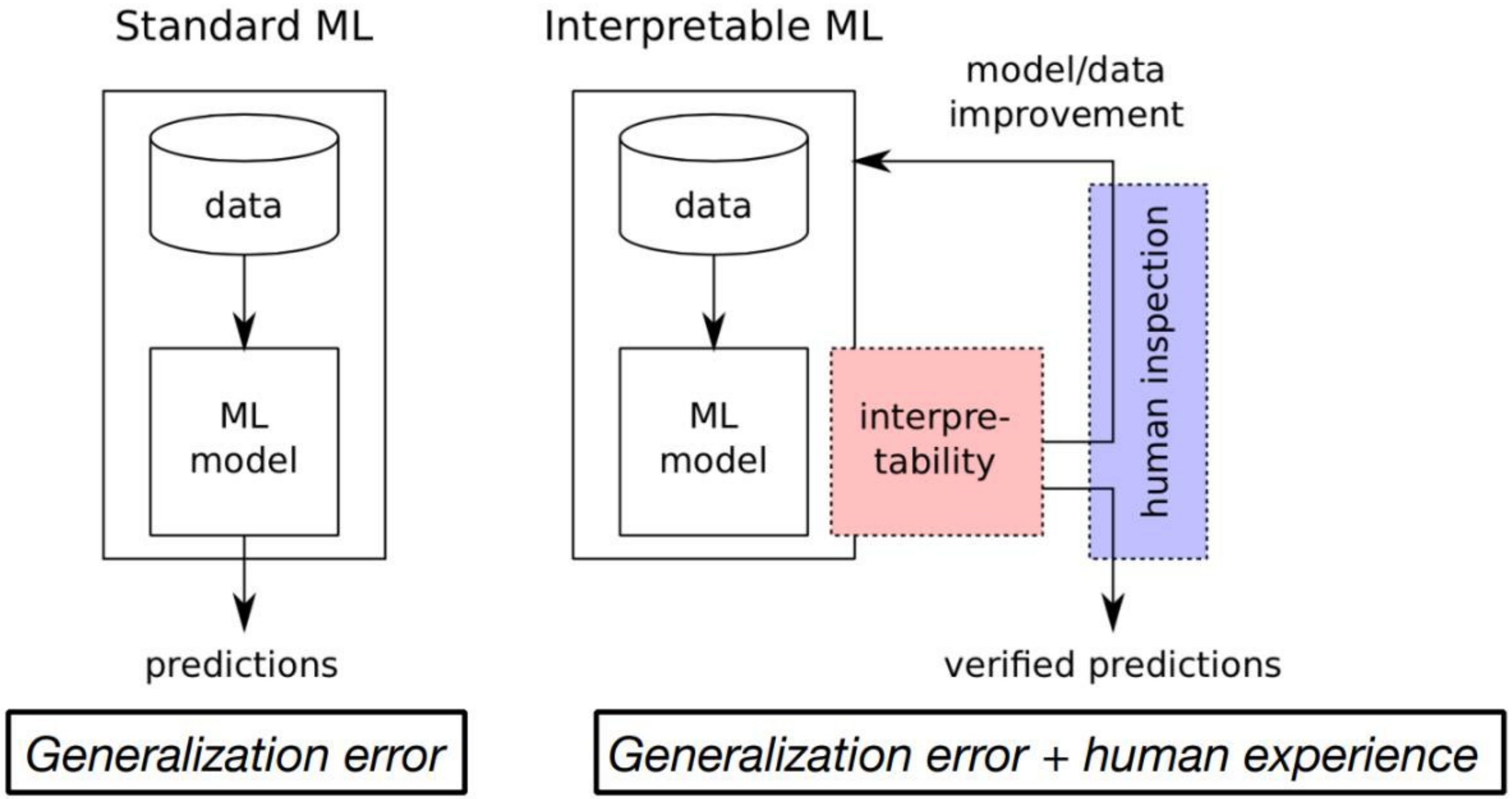


Disease Misclassification



Why Interpretability?

2. Improve / Debug classifier

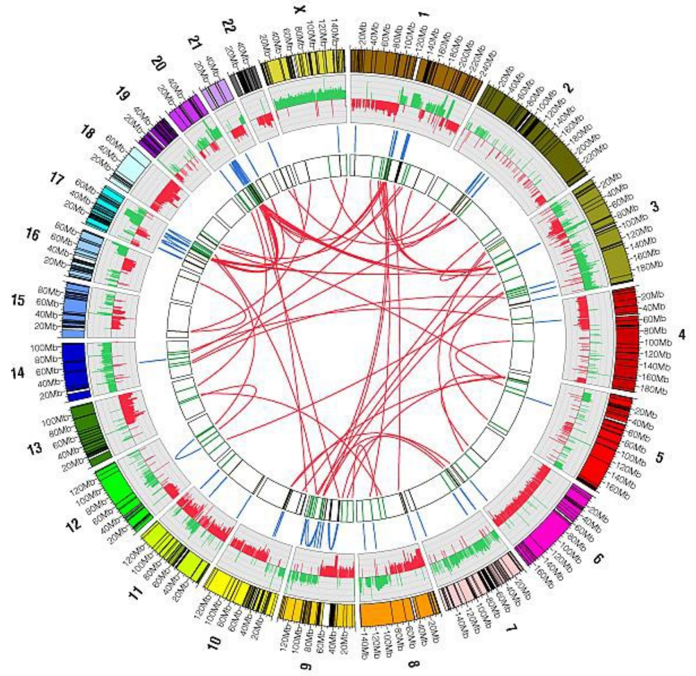


Slides by Beomsu Kim, KAIST

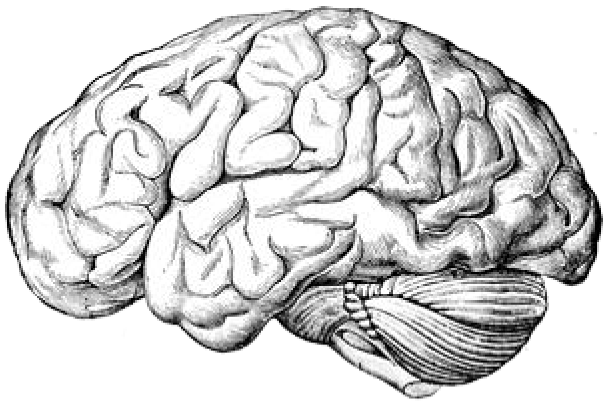
Why Interpretability?

3. Make new discoveries

Learn about the physical / biological / chemical mechanisms



Learn about the human brain



Slides by Beomsu Kim, KAIST

Why Interpretability?

4. Right to explanation

"Right to be given an explanation for an output of the algorithm"

Ex.

- US Equal Credit OpportunityAct
- The European Union General Data Protection Regulation
- France Digital RepublicAct

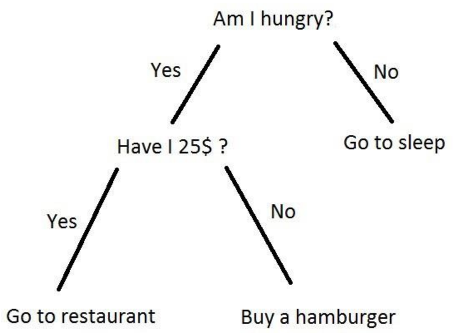
Slides by Beomsu Kim, KAIST

Types of Interpretability in ML

Ante-hoc Interpretability

Choose an interpretable model and train it.

Ex.



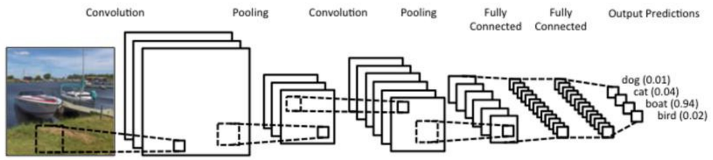
Decision Tree

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.

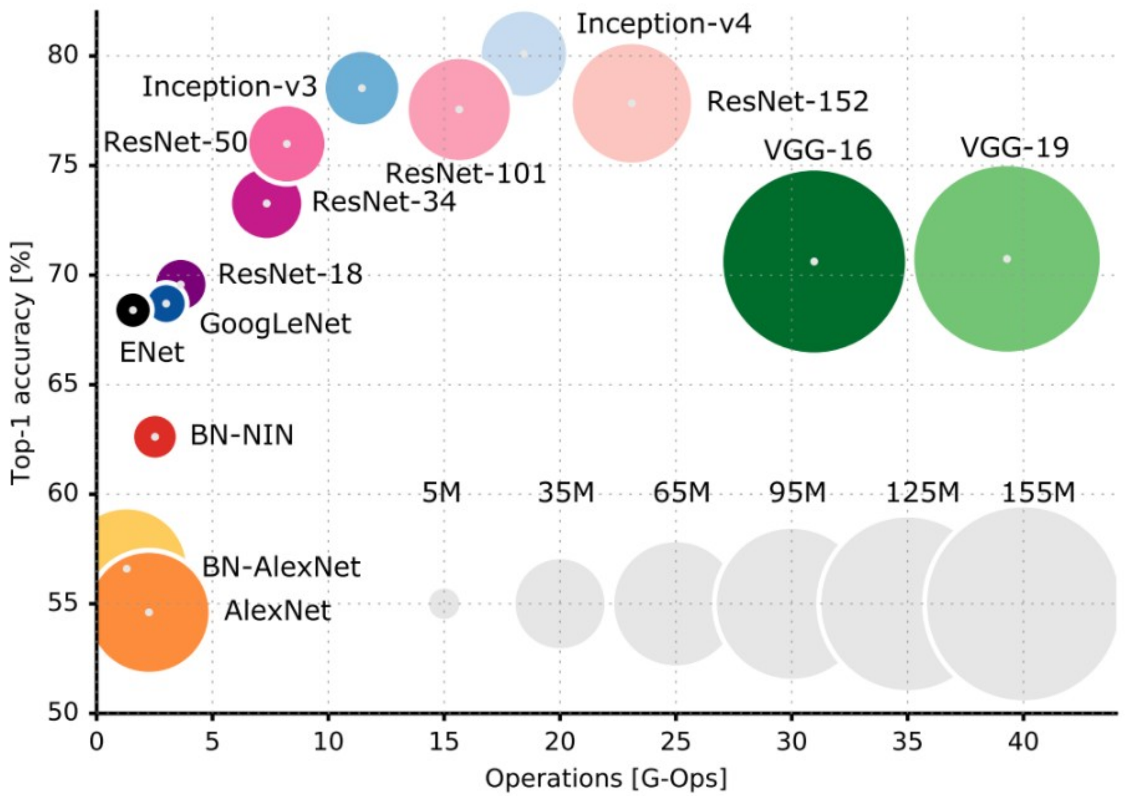


Deep Neural Networks

Problem. How to interpret millions of parameters?

Slides by Beomsu Kim, KAIST

Types of Interpretability in ML



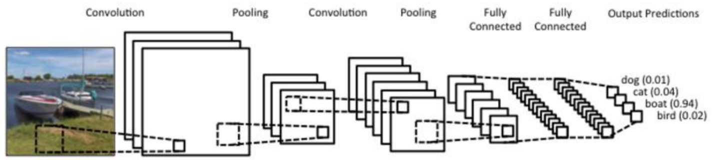
At least 5 million parameters!

“An Analysis of Deep Neural Models for Practical Applications”, https://arxiv.org/pdf/1605.07678.pdf

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?

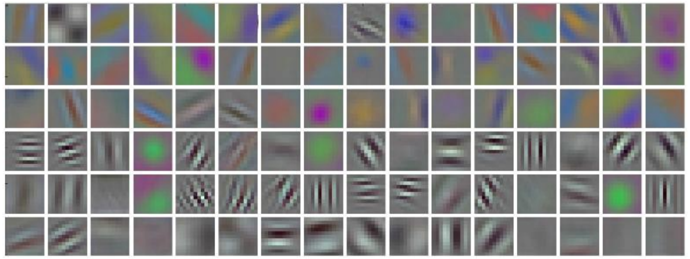
Slides by Beomsu Kim, KAIST

Types of Post-hoc Interpretability

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

Model ⟷ **Input**

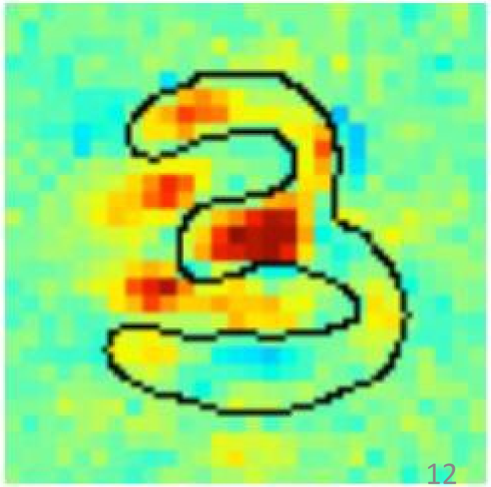
What representations have the DNN learned?



What pattern / image maximally activates a particular neuron?

dumbbell

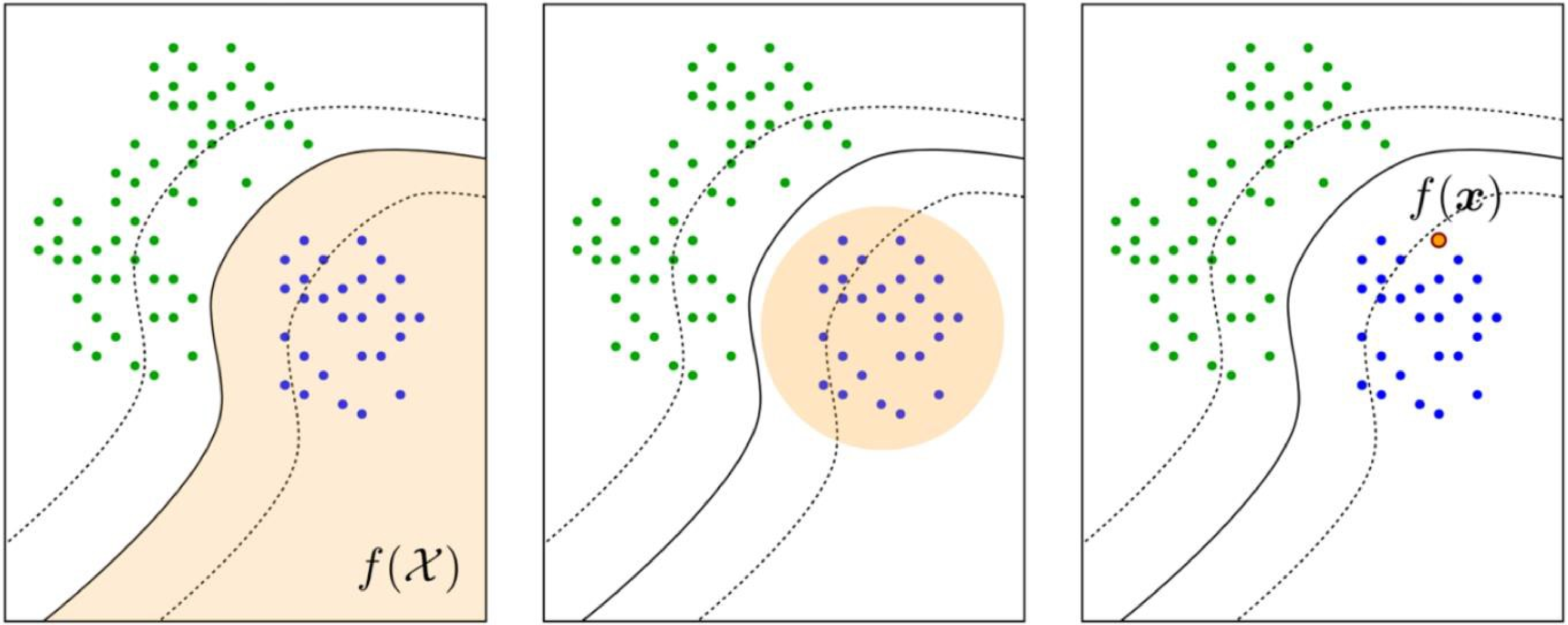
Explain why input x has been classified as $f(x)$.



Slides by Beomsu Kim, KAIST

Types of Post-hoc Interpretability

Model ⟷ **Input**



Slides by Beomsu Kim, KAIST

Interpretable Deep Learning

1. Intro to Interpretability

1a. Interpretability definition: Convert implicit NN information to human-interpretable information

1b. Motivation: Verify model works as intended; debug classifier; make discoveries; Right to explanation

1c. Ante-hoc (train interpretable model) vs. **Post-hoc** (interpret complex model; degree of “locality”)

2. Interpreting Deep Neural Networks

2a. Interpreting Models (macroscopic, understand internals) vs. **decisions** (microscopic, practical applications)

2b. Interpreting Models: Weight visualization, Surrogate model, Activation maximization, Example-based

2c. Interpreting Decisions:

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

3. Evaluating Attribution Methods

3a. Qualitative: Coherence: Attributions should highlight discriminative features / objects of interest

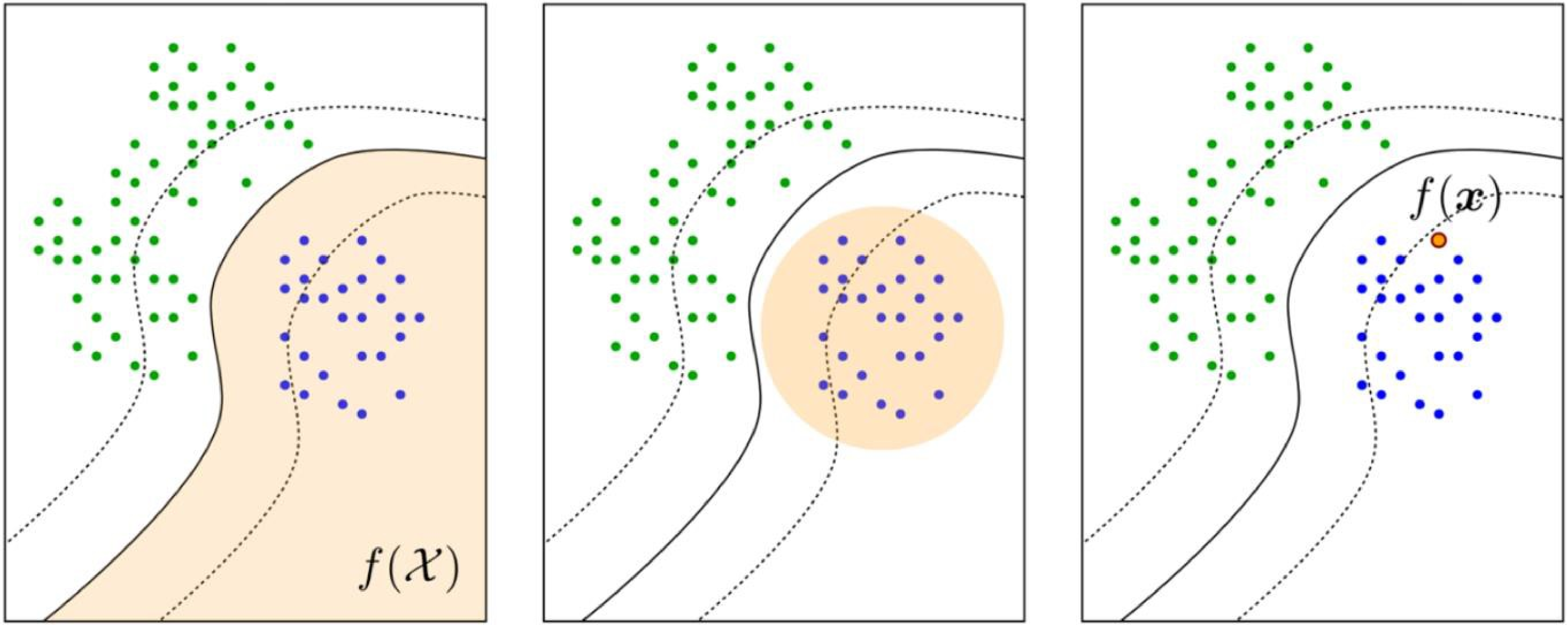
3b. Qualitative: Class Sensitivity: Attributions should be sensitive to class labels

3c. Quantitative: Sensitivity: Removing feature with high attribution ➔ large decrease in class probability

3d. Quantitative: ROAR & KAR. Low class prob cuz image unseen ➔ remove pixels, retrain, measure acc. drop

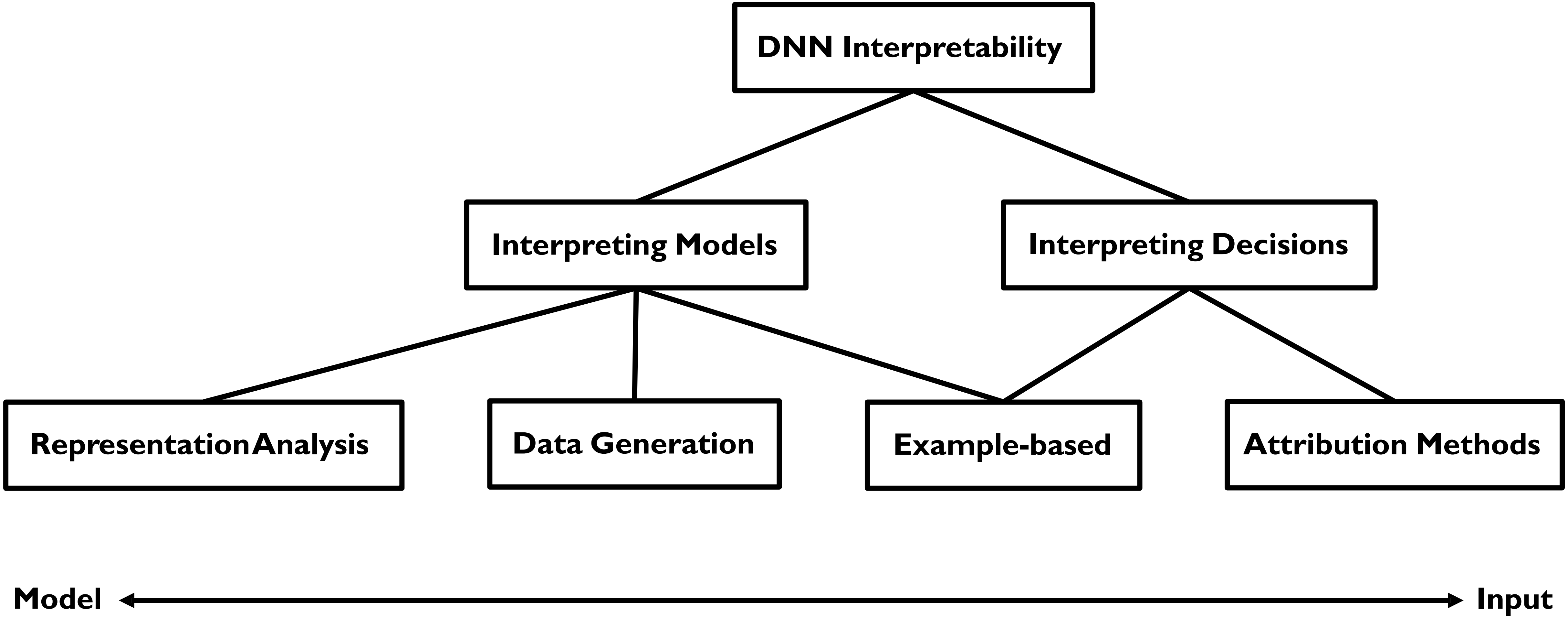
Types of Post-hoc Interpretability

Model **Input**



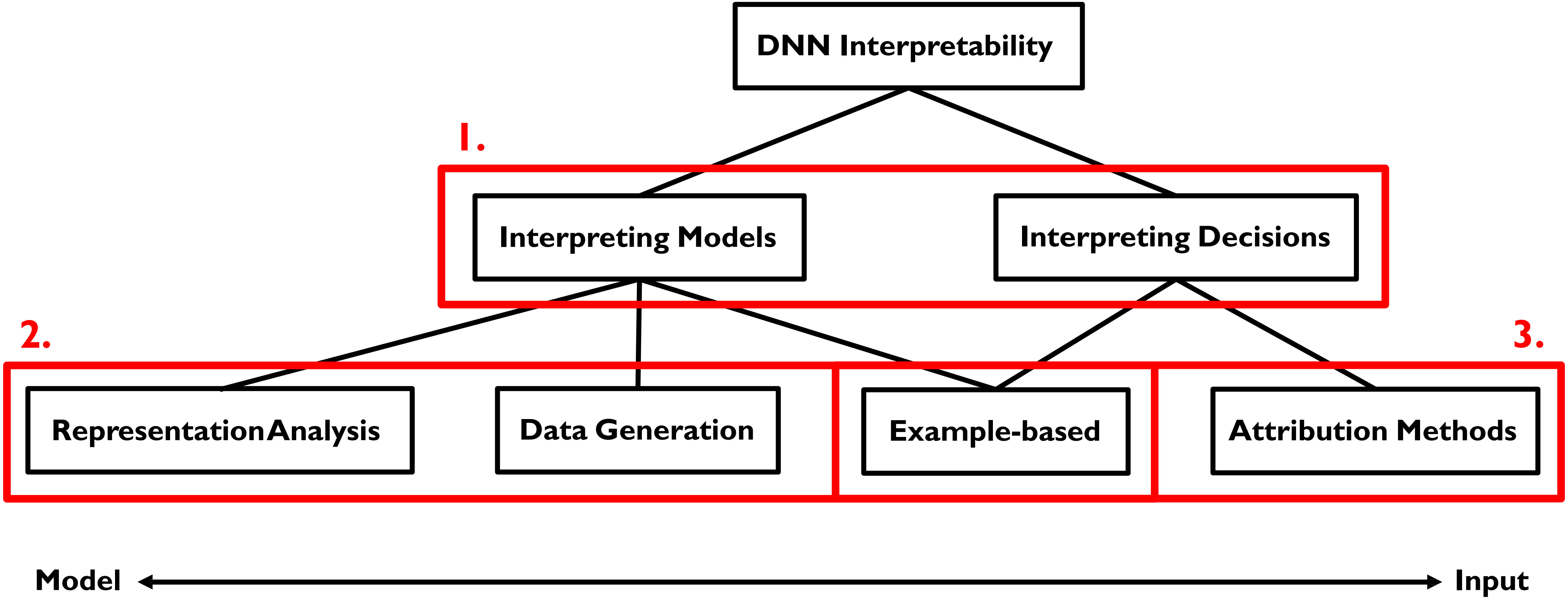
Slides by Beomsu Kim, KAIST

Types of DNN Interpretability

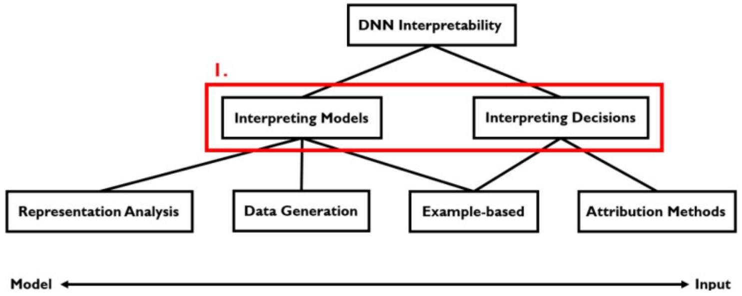


Slides by Beomsu Kim, KAIST

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

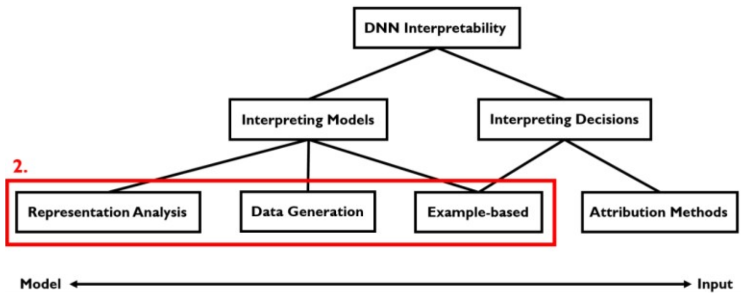
Interpreting Decisions
(Microscopic)

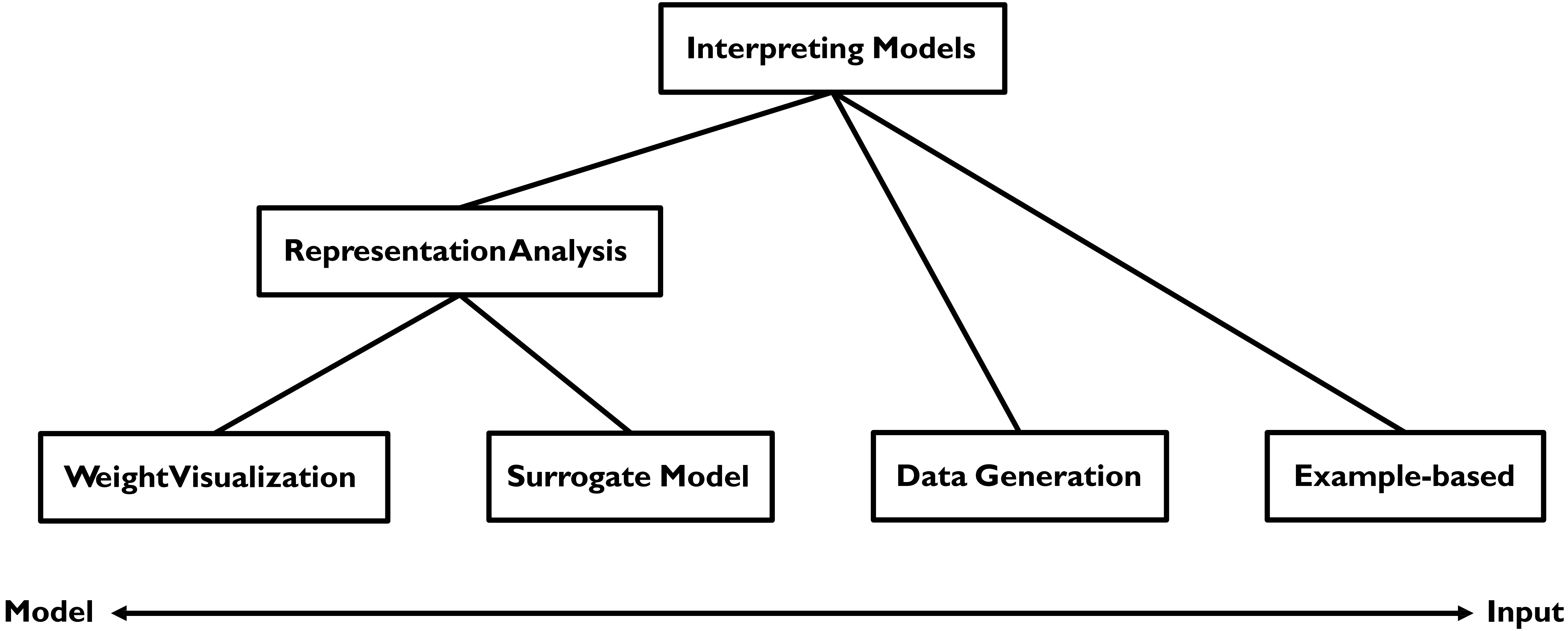
- Why did DNN make this decision
- Verify that model behaves as expected
- Find evidence for decision

Important for practical applications

Slides by Beomsu Kim, KAIST

Types of DNN Interpretability



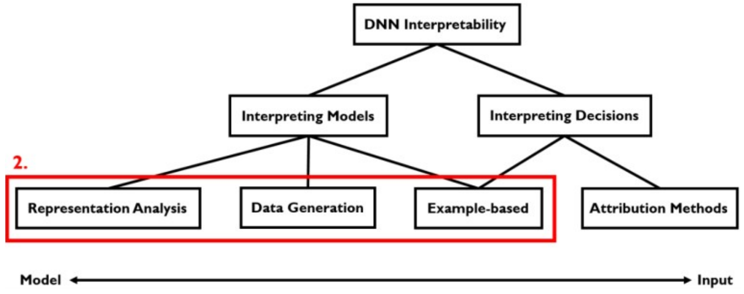


Slides by Beomsu Kim, KAIST

2b – Interpreting models:

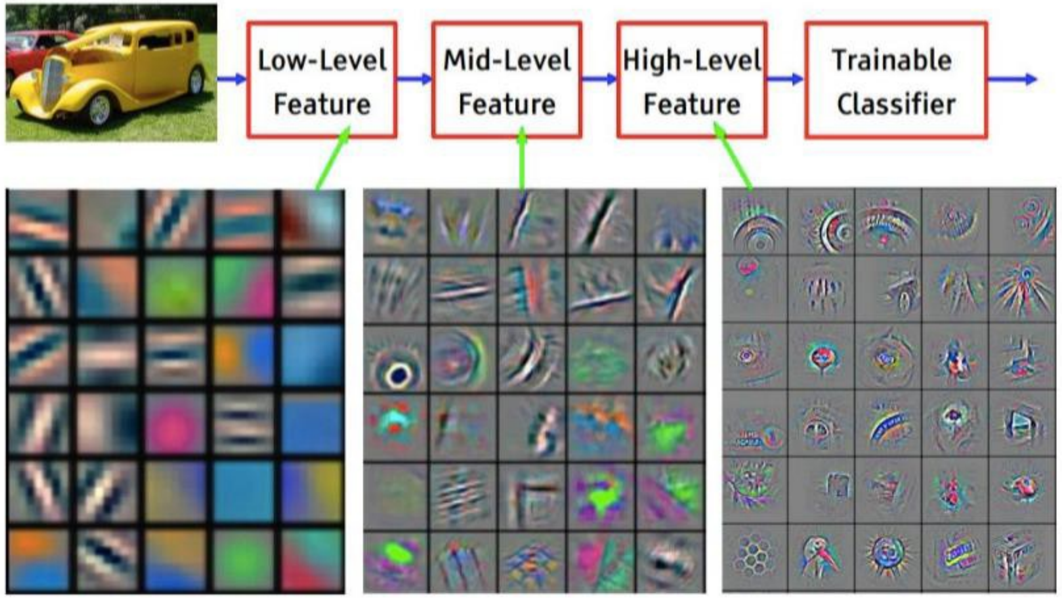
(i) Representation analysis: Weight Visualization
(ii) Representation analysis: Surrogate Model
(iii) Data Generation / Activation Maximization
(iv) Example based

Types of DNN Interpretability



WeightVisualization | **Surrogate Model** | **Data Generation** | **Example-based**

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



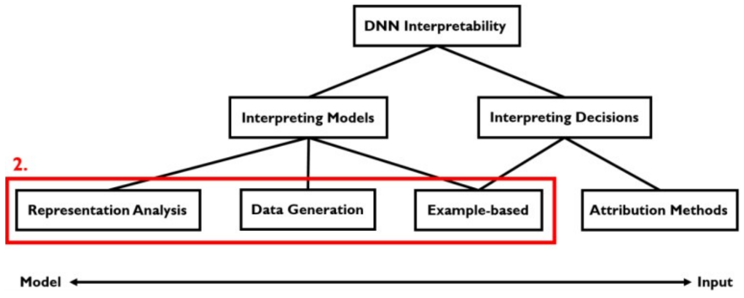
Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

“Deep Inside Convolutional Networks: Visualizing Image Classification Models and Saliency Maps”, https://arxiv.org/pdf/1312.6034.pdf

2b – Interpreting models:

(i) Weight Visualization
(ii) Surrogate Model
(iii) Data Generation / Activation Maximization
(iv) Example based

Slides by Beomsu Kim, KAIST



Types of DNN Interpretability

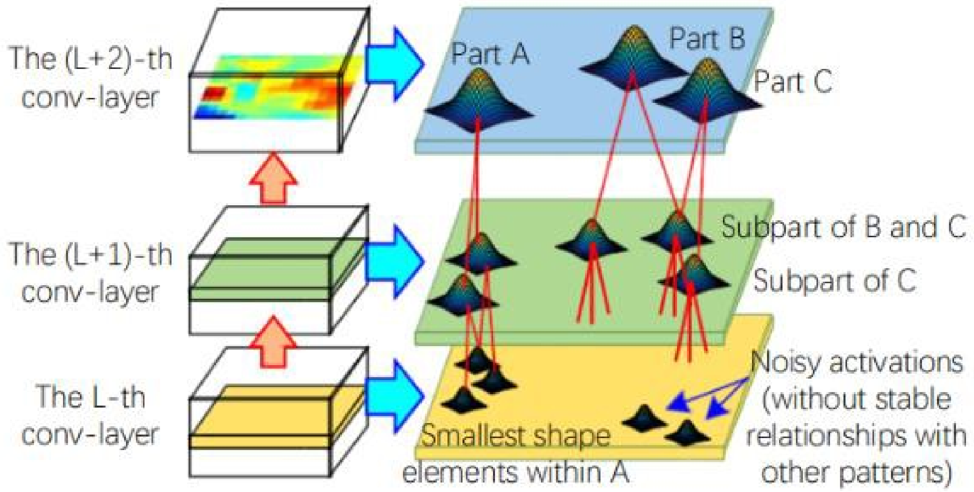
WeightVisualization | **Surrogate Model** | **Data Generation** | **Example-based**

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

(aka. Meta-model, approximation model, response surface model, emulator…)

Idea: Train an *<u>Interpretable</u>* Machine Learning model on the Outputs of our "Black Box" model with the specific goal of interpreting it. Not exact, but close and interpretable.

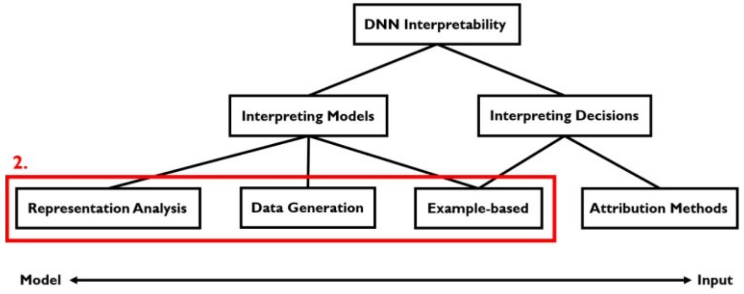
Model agnostic. Approximate the predictions, not the actual real world.



"Explaining CNN Knowledge via an Explanatory Graph", https://arxiv.org/pdf/1605.07678.pdf

2b – Interpreting models:

(i) Weight Visualization
(ii) Surrogate Model
(iii) Data Generation / Activation Maximization
(iv) Example based

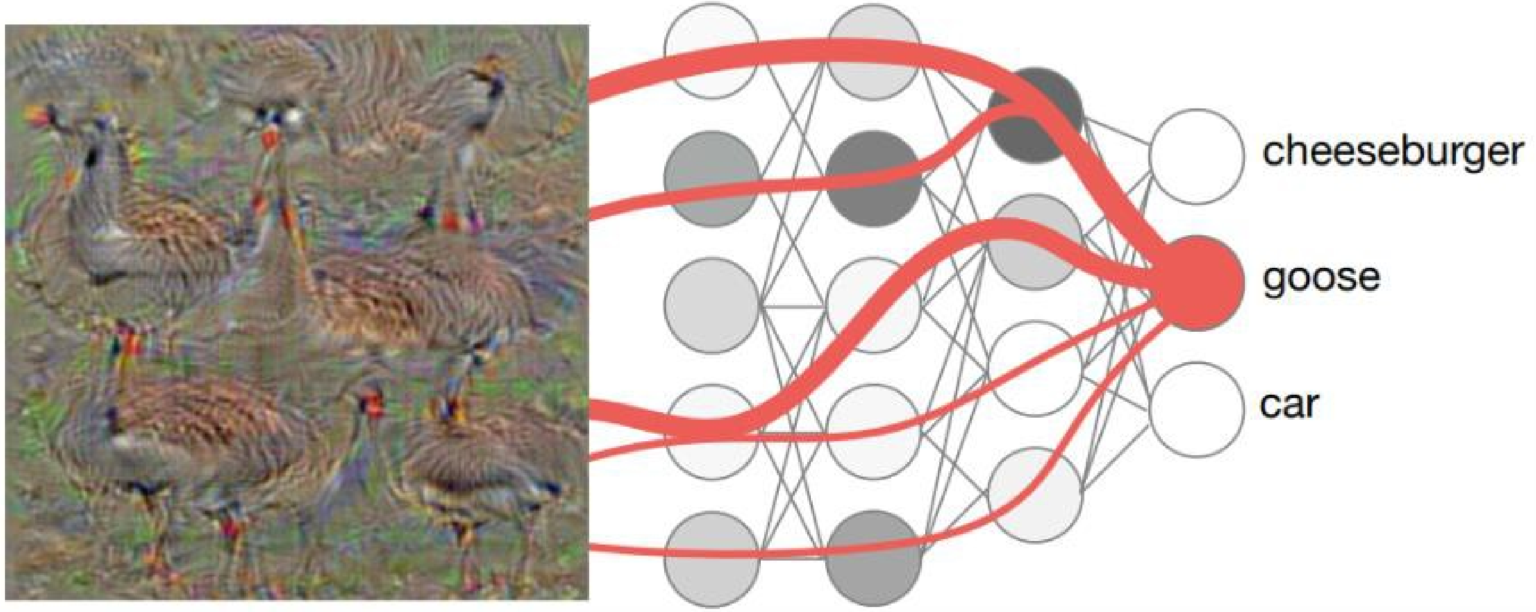


Types of DNN Interpretability

WeightVisualization | **Surrogate Model** | **Data Generation** | **Example-based**

Activation Maximization

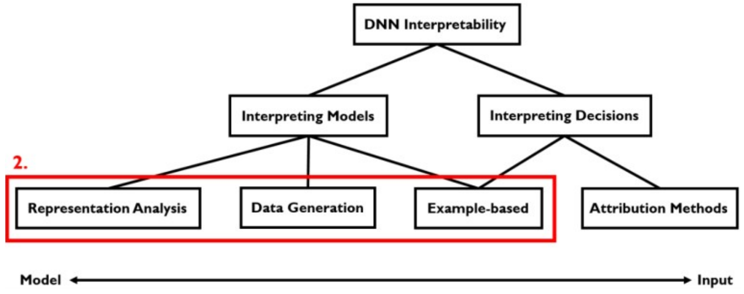
- Find pattern maximizing activation of a neuron



"Deep Inside Convolutional Networks: Visualizing Image Classification Models and Saliency Maps", https://arxiv.org/pdf/1312.6034.pdf

Slides by Beomsu Kim, KAIST

Types of DNN Interpretability



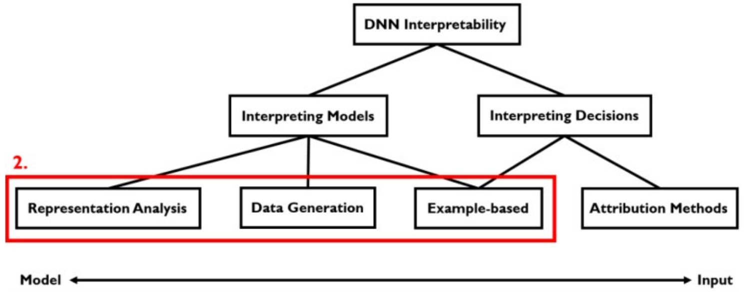
WeightVisualization | **Surrogate Model** | **Data Generation** | **Example-based**

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

Class Probability

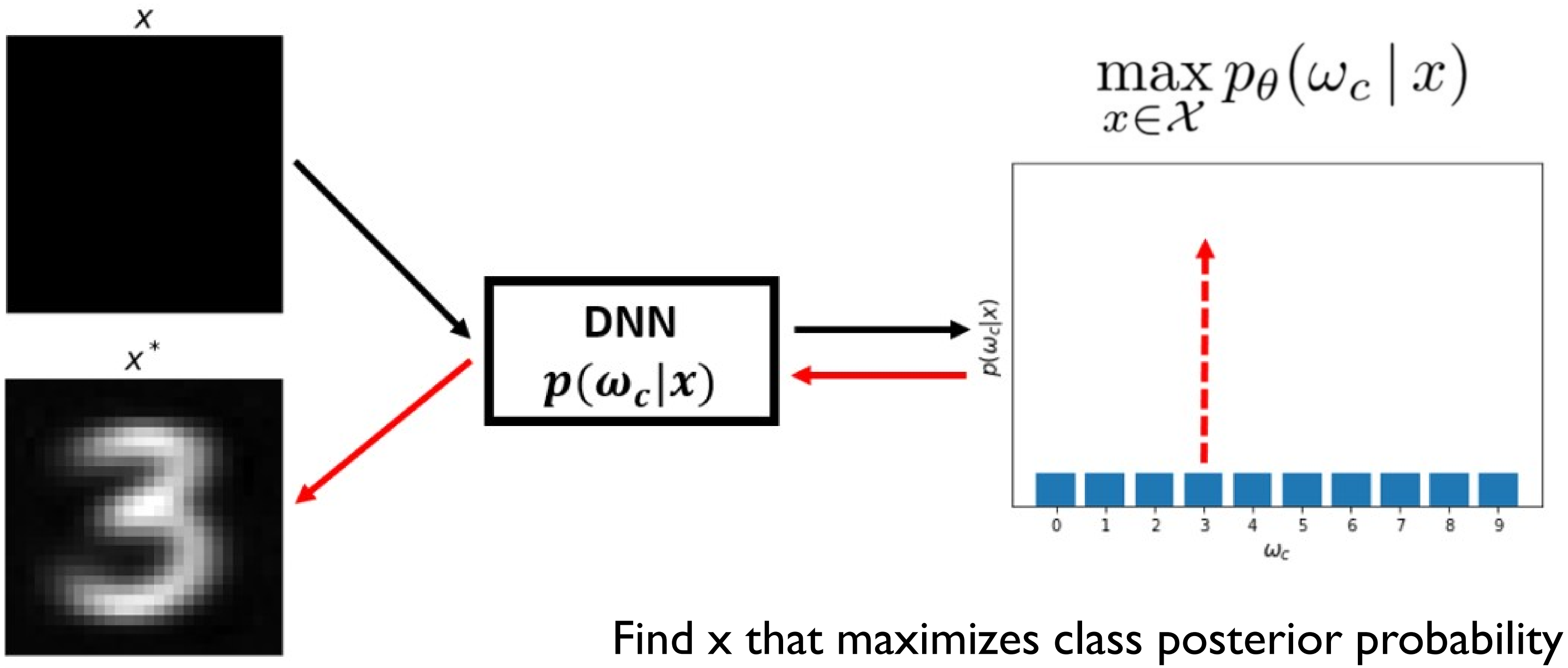
Regularization Term

Slides by Beomsu Kim, KAIST



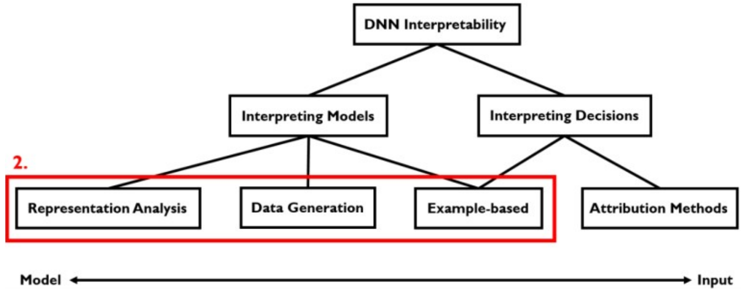
Types of DNN Interpretability

WeightVisualization | **Surrogate Model** | **Data Generation** | **Example-based**



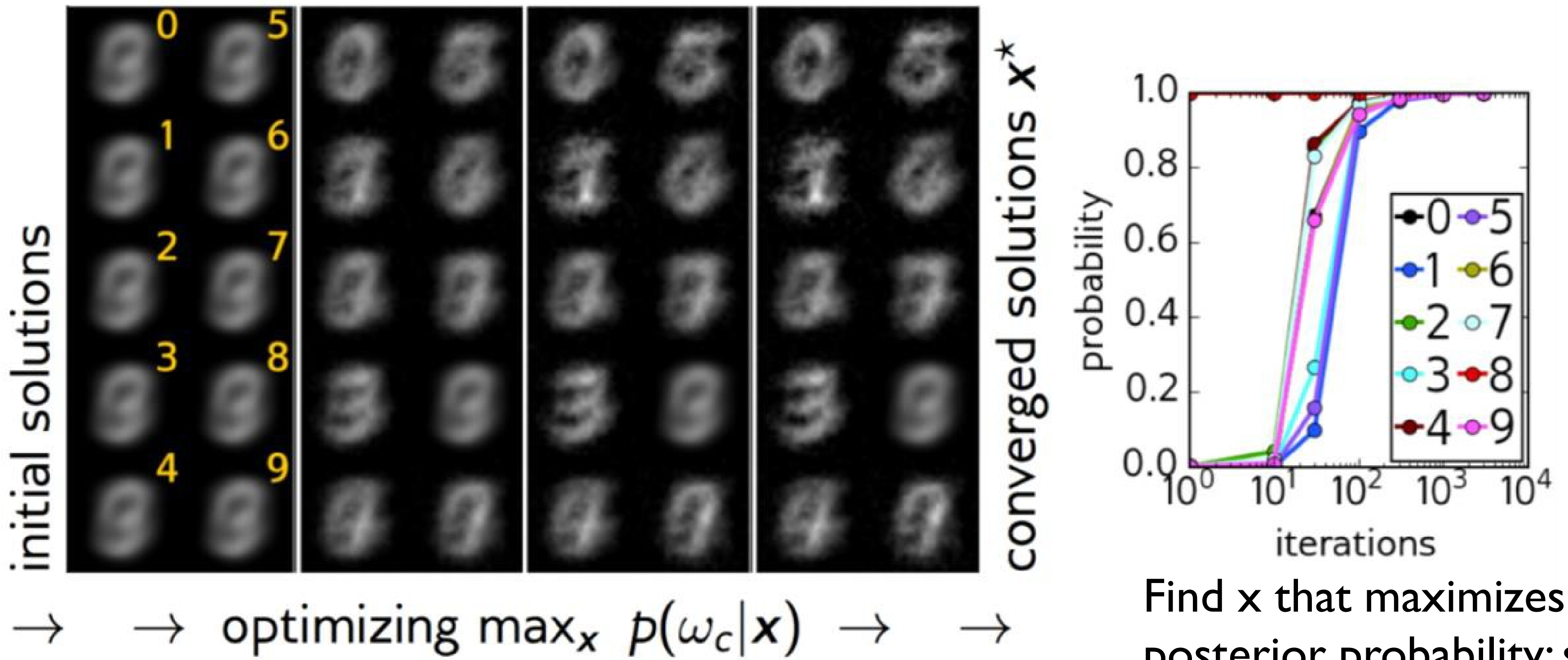
Find x that maximizes class posterior probability

Slides by Beomsu Kim, KAIST

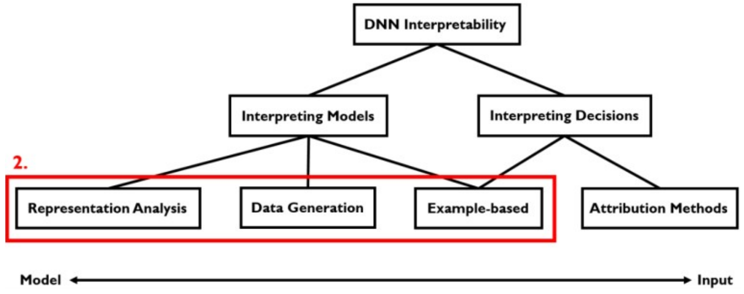


Types of DNN Interpretability

WeightVisualization | **Surrogate Model** | **Data Generation** | **Example-based**

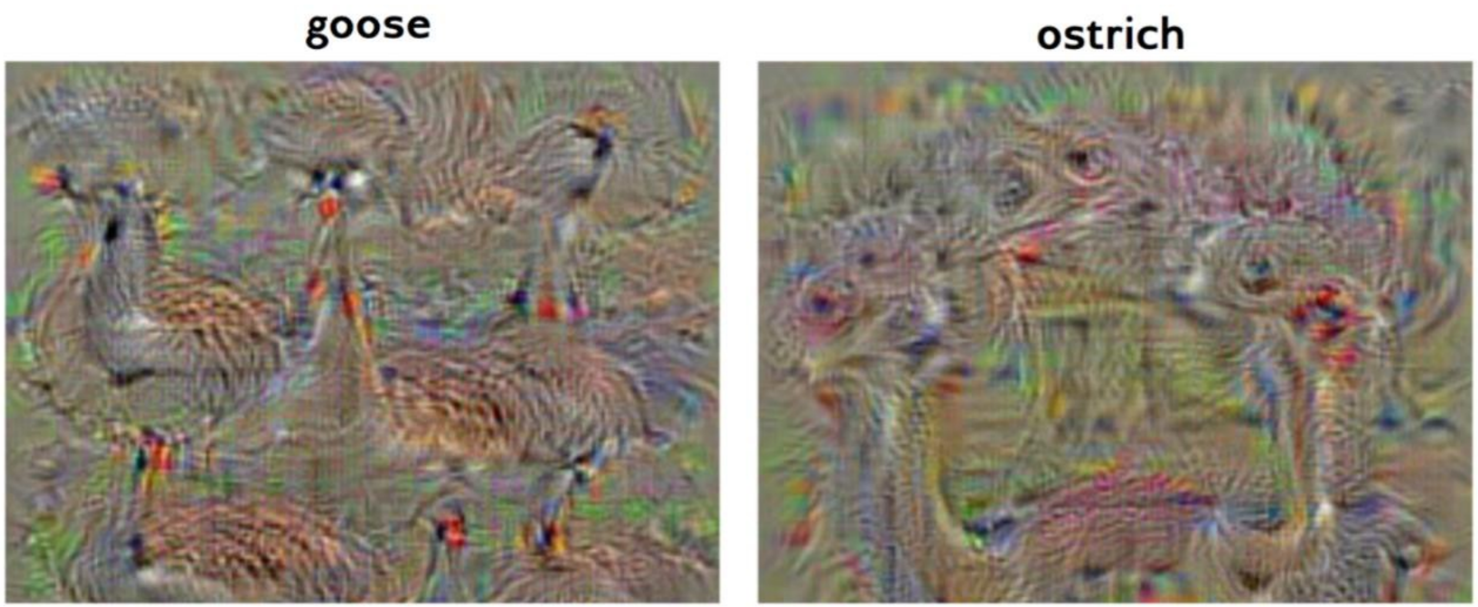


Find x that maximizes class posterior probability: search



Types of DNN Interpretability

WeightVisualization | **Surrogate Model** | **Data Generation** | **Example-based**

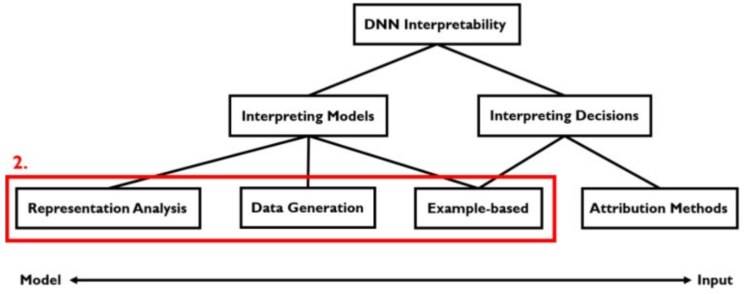


Images from **Simonyan et al. 2013** "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps"

"Deep Inside Convolutional Networks: Visualizing Image Classification Models and Saliency Maps", https://arxiv.org/pdf/1312.6034.pdf

Slides by Beomsu Kim, KAIST

Types of DNN Interpretability



WeightVisualization | **Surrogate Model** | **Data Generation** | **Example-based**

Advantages

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

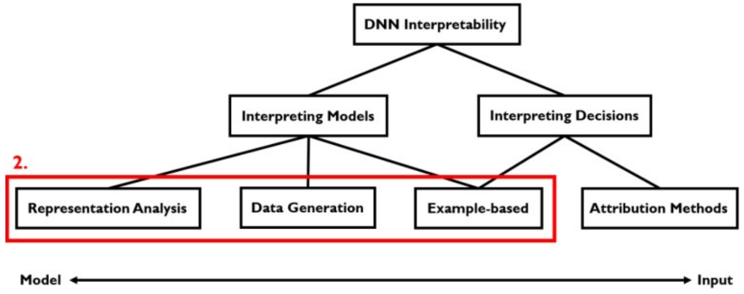
Disadvantages

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

Redefine optimization problem!

Slides by Beomsu Kim, KAIST

Types of DNN Interpretability



WeightVisualization | **Surrogate Model** | **Data Generation** | **Example-based**

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

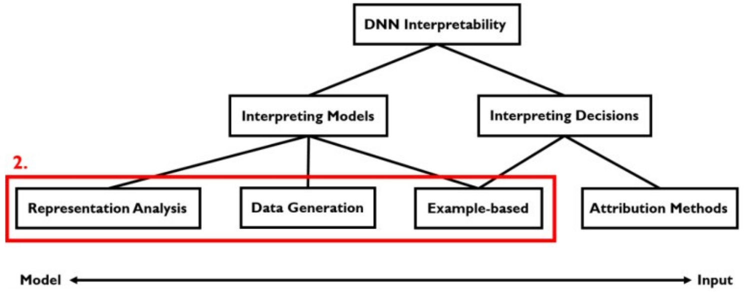
Redefine optimization problem!

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

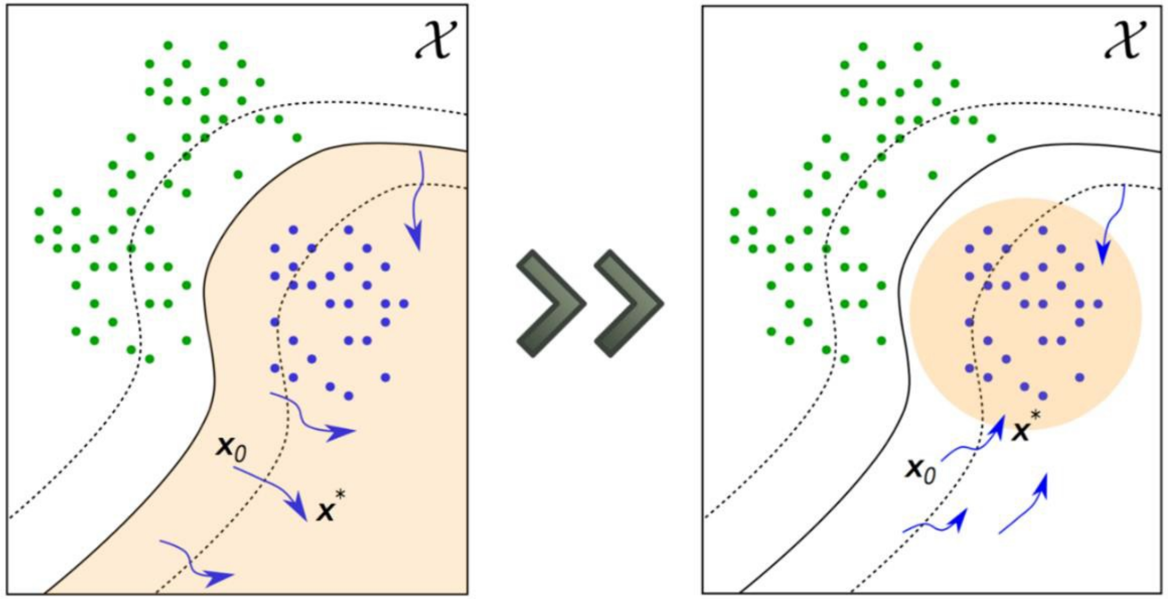
Slides by Beomsu Kim, KAIST

Types of DNN Interpretability



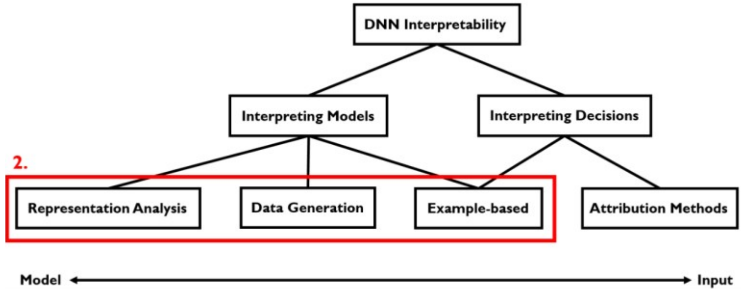
WeightVisualization | **Surrogate Model** | **Data Generation** | **Example-based**

Find the input pattern that maximizes class probability → Find the most likely input pattern for a given class



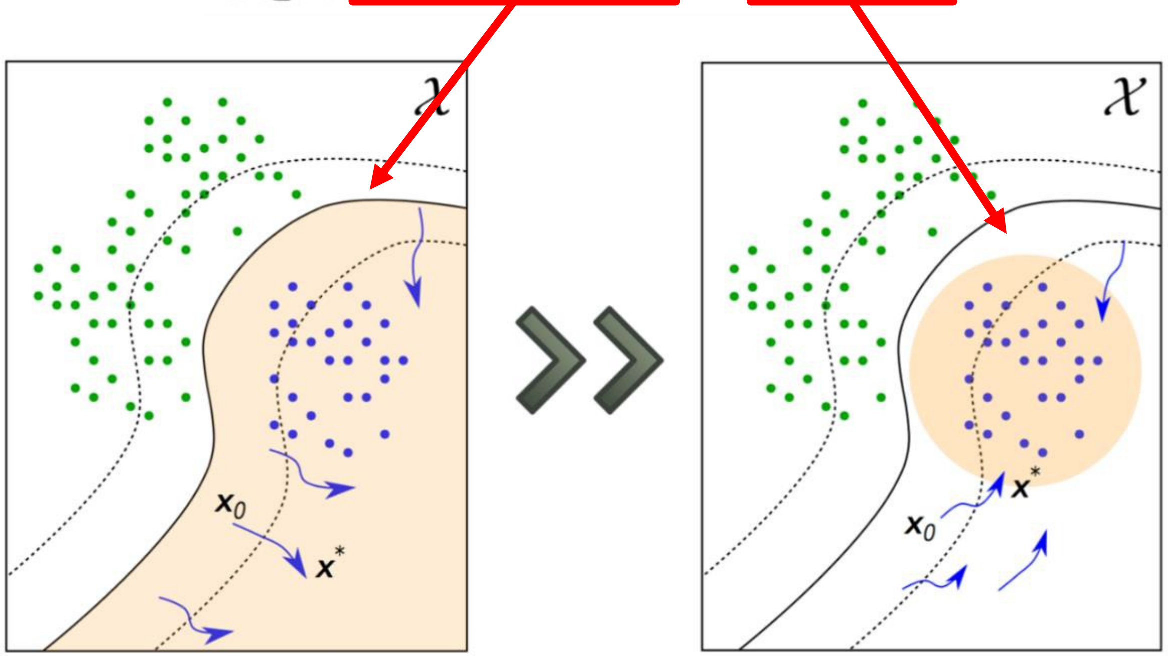
Slides by Beomsu Kim, KAIST

Types of DNN Interpretability

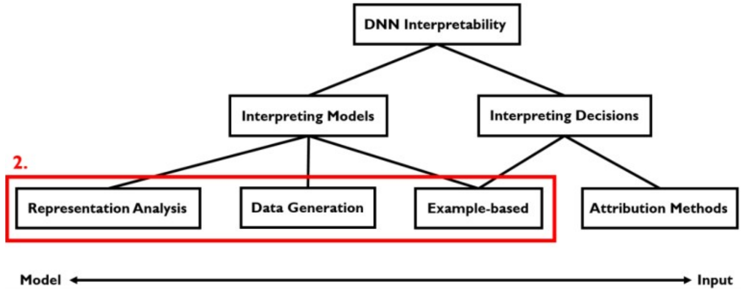


WeightVisualization | Surrogate Model | **Data Generation** | Example-based

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



Slides by Beomsu Kim, KAIST



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

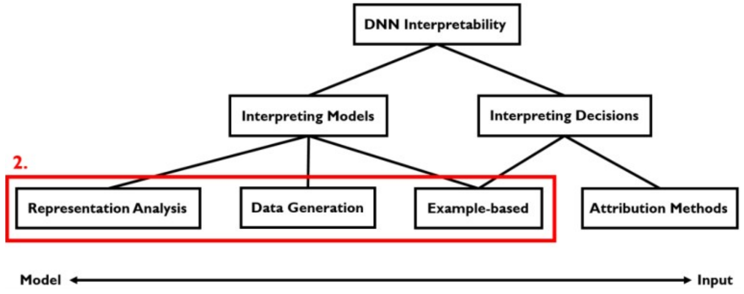
$$p(\boldsymbol{x}|\omega_c) \propto \underbrace{\boxed{p(\omega_c|\boldsymbol{x})}}_{\text{original}} \cdot p(\boldsymbol{x})$$

Activation Maximization in Code Space

$$\max_{\boldsymbol{z} \in \mathcal{Z}} p(\omega_c | \underbrace{g(\boldsymbol{z})}_{\boldsymbol{x}}) + \lambda \|\boldsymbol{z}\|^2 \qquad \boldsymbol{x}^\star = g(\boldsymbol{z}^\star)$$

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

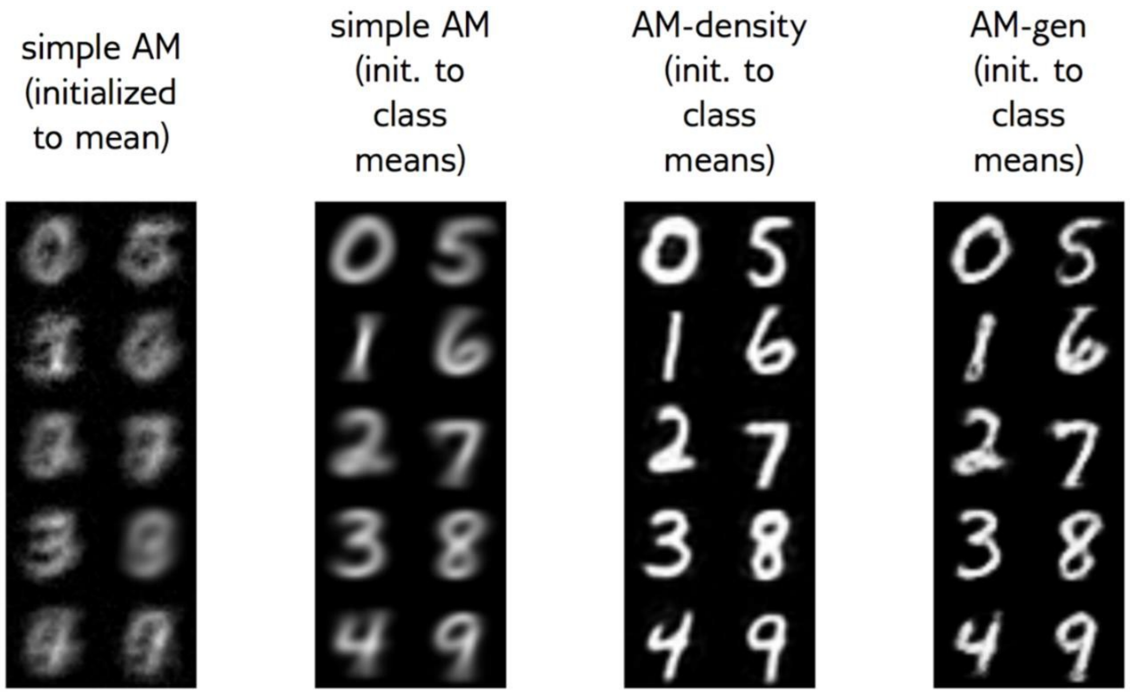
Slides by Beomsu Kim, KAIST



Types of DNN Interpretability

WeightVisualization | **Surrogate Model** | **Data Generation** | **Example-based**

Original Activation Maximization

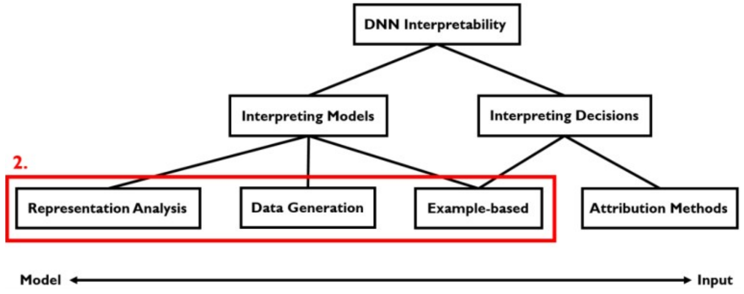


Constrained Activation Maximization

Observation: Connecting to the data leads to sharper visualizations.

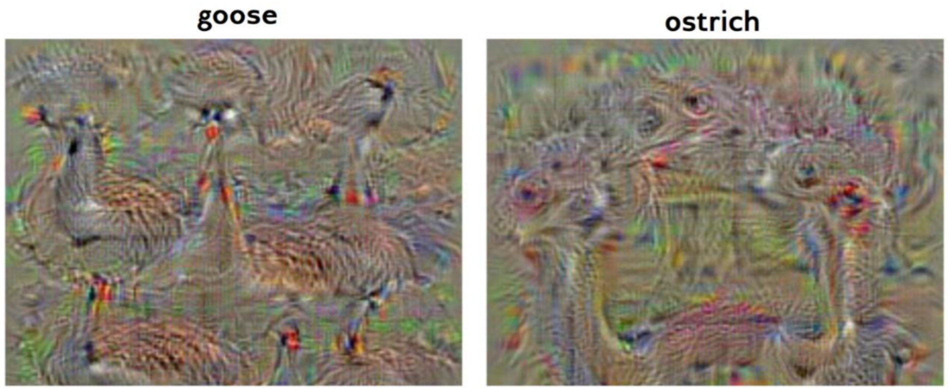
Slides by Beomsu Kim, KAIST

Types of DNN Interpretability



Weight Visualization | **Surrogate Model** | **Data Generation** | **Example-based**

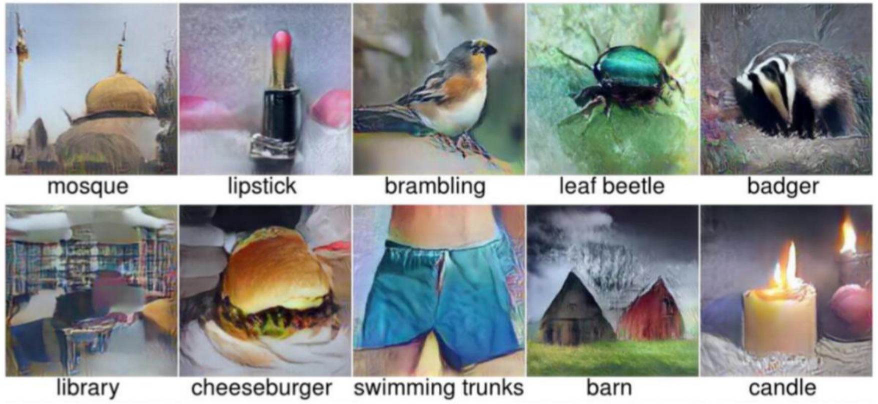
Activation Maximization



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

Activation Maximization in Code Space

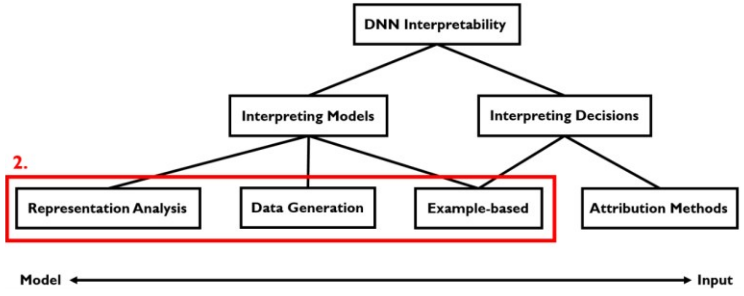
Images from **Nguyen et al.** 2016. "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks"



Observation: Connecting to the data leads to sharper visualizations.

Slides by Beomsu Kim, KAIST

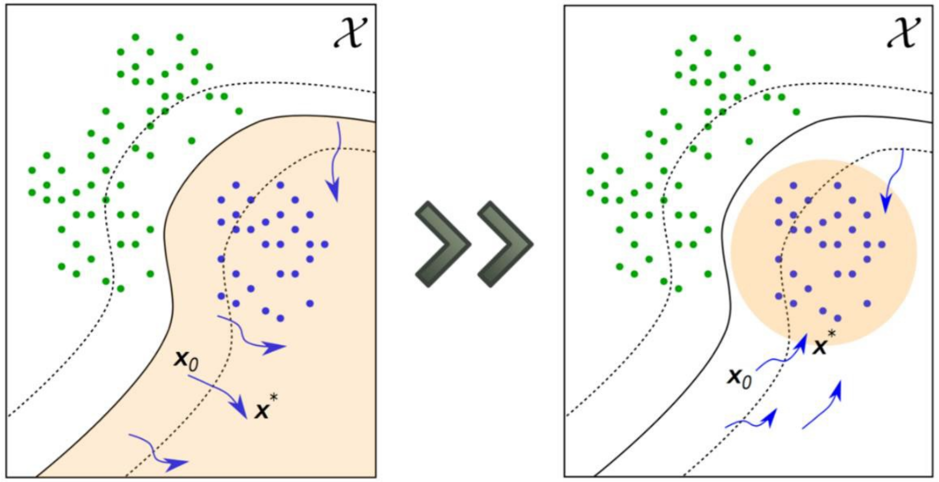
Types of DNN Interpretability



WeightVisualization | **Surrogate Model** | **Data Generation** | **Example-based**

Summary

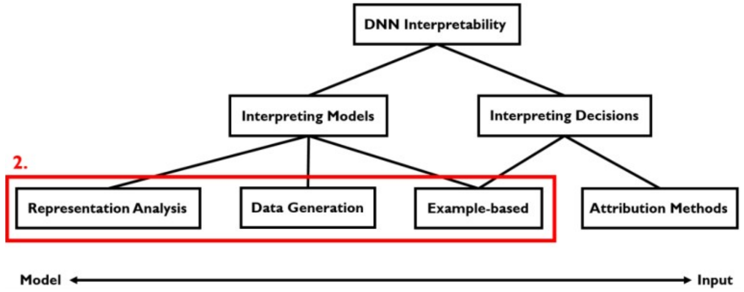
- DNNs can be interpreted by finding input patterns that maximize a certain output quantity.
- Connecting to the data improves the interpretability of the visualization.



2b – Interpreting models:

(i) Weight Visualization
(ii) Surrogate Model
(iii) Data Generation / Activation Maximization
(iv) Example based

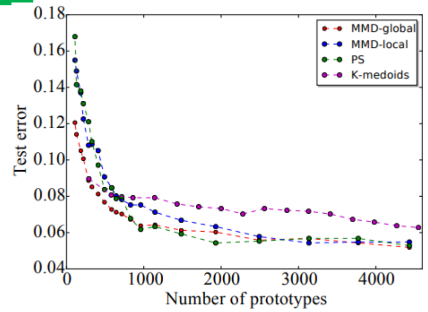
Slides by Beomsu Kim, KAIST

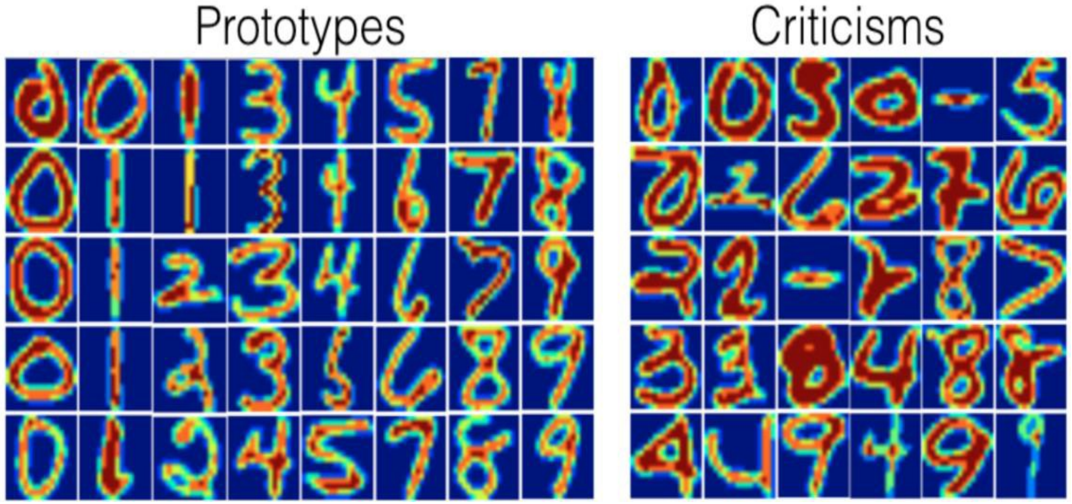


Types of DNN Interpretability

WeightVisualization | **Surrogate Model** | **Data Generation** | **Example-based**

- Find image instances that represent / do not represent the image class
- Maximum Mean Discrepancy – MMD-critic, efficient prototype selection
- Nearest prototype classifier
- Example: digits; Classifying proximal dog breeds; dogs in costumes misclassified





"Examples are not Enough, Learn to Criticize! Criticism for Interpretability", https://people.csail.mit.edu/beenkim/papers/KIM2016NIPS_MMD.pdf

Interpretable Deep Learning

1. Intro to Interpretability

1a. Interpretability definition: Convert implicit NN information to human-interpretable information

1b. Motivation: Verify model works as intended; debug classifier; make discoveries; Right to explanation

1c. Ante-hoc (train interpretable model) vs. **Post-hoc** (interpret complex model; degree of "locality")

2. Interpreting Deep Neural Networks

2a. Interpreting Models (macroscopic, understand internals) vs. **decisions** (microscopic, practical applications)

2b. Interpreting Models: Weight visualization, Surrogate model, Activation maximization, Example-based

2c. Interpreting Decisions:

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

3. Evaluating Attribution Methods

3a. Qualitative: Coherence: Attributions should highlight discriminative features / objects of interest

3b. Qualitative: Class Sensitivity: Attributions should be sensitive to class labels

3c. Quantitative: Sensitivity: Removing feature with high attribution ➔ large decrease in class probability

3d. Quantitative: ROAR & KAR. Low class prob cuz image unseen ➔ remove pixels, retrain, measure acc. drop

Limitation of Model Interpretations

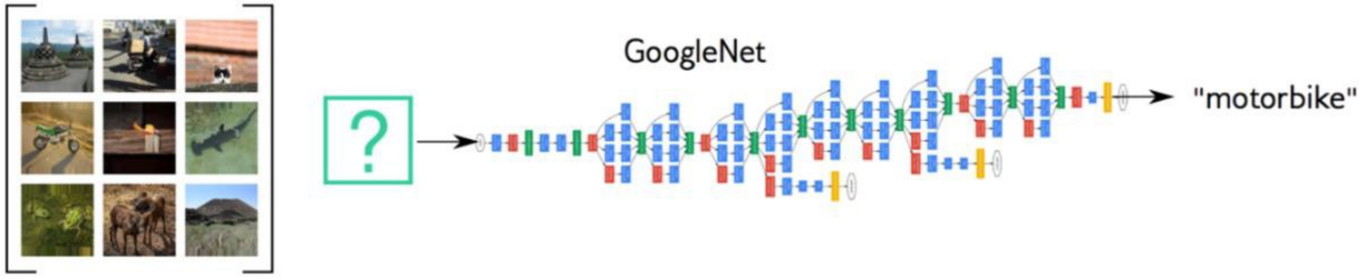
Question: What would be the best image to interpret the class "motorcycle"?

- 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.

Slides by Beomsu Kim, KAIST

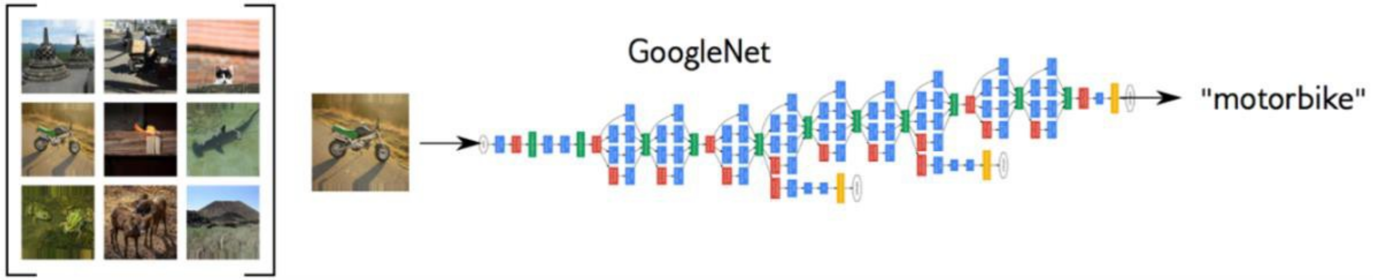
Limitation of Model Interpretations

Finding a prototype:



Question: How does a "motorbike" **typically** look like?

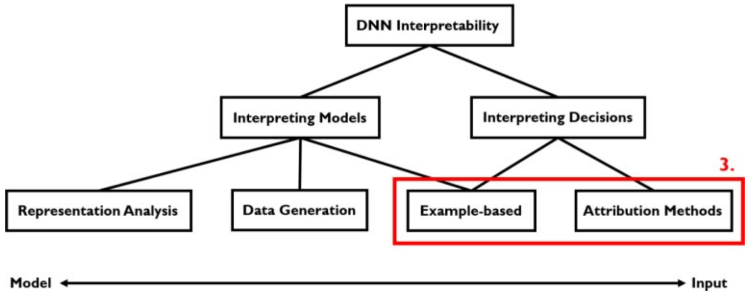
Decision explanation:

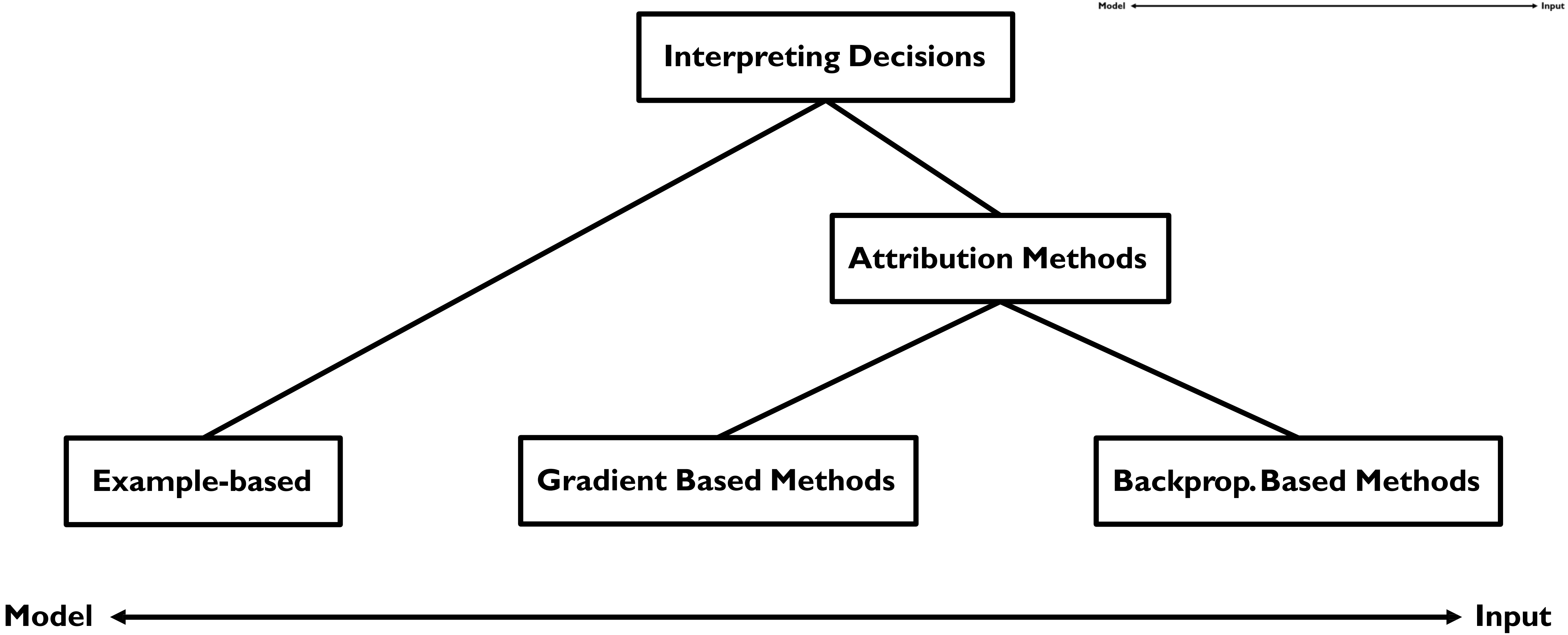


Question: Why is <u>**this example**</u> classified as a motorbike?

Slides by Beomsu Kim, KAIST

Types of DNN Interpretability





Slides by Beomsu Kim, KAIST

2c – Interpreting decisions:

(i) Example based

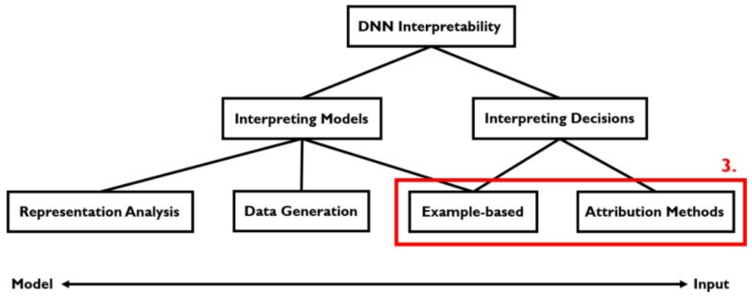
(ii) Attribution Methods: why are gradients noisy?

(iii) Gradient-based Attribution: SmoothGrad, Interior Gradient

(iv) Backprop-based Attribution: Deconvolution, Guided Backprop

Slides by Beomsu Kim, KAIST

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
- Influence functions for interpreting black-box methods. Fragility of NN model interpretation.

"Understanding Black-box Predictions via Influence Functions", https://arxiv.org/pdf/1703.04730.pdf
"Interpretation of Neural Networks is Fragile", https://arxiv.org/pdf/1710.10547.pdf

2c – Interpreting decisions:

(i) Example based

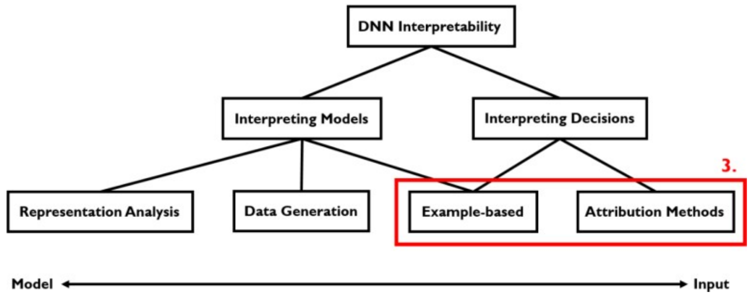
(ii) Attribution Methods: why are gradients noisy?

(iii) Gradient-based Attribution: SmoothGrad, Interior Gradient

(iv) Backprop-based Attribution: Deconvolution, Guided Backprop

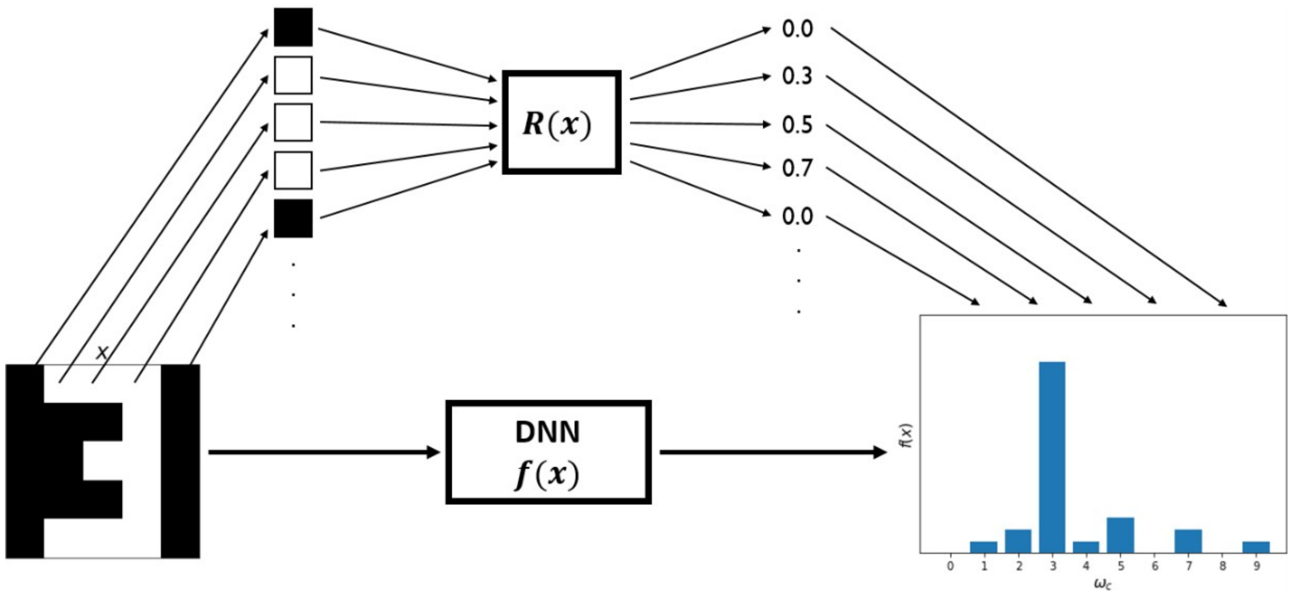
Slides by Beomsu Kim, KAIST

Types of DNN Interpretability



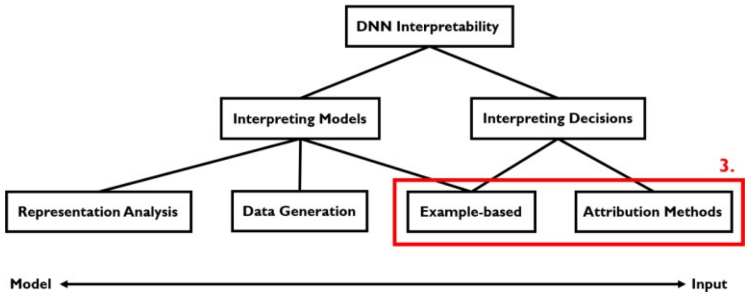
Example-based | **Attribution Methods** | **Gradient Based** | **Backprop. Based**

Given an image $\boldsymbol{x} \in \mathbb{R}^n$ and a decision $f(\boldsymbol{x})$,
assign to each pixel $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ attribution values $R_1(\boldsymbol{x}), R_2(\boldsymbol{x}), \ldots, R_n(\boldsymbol{x})$.



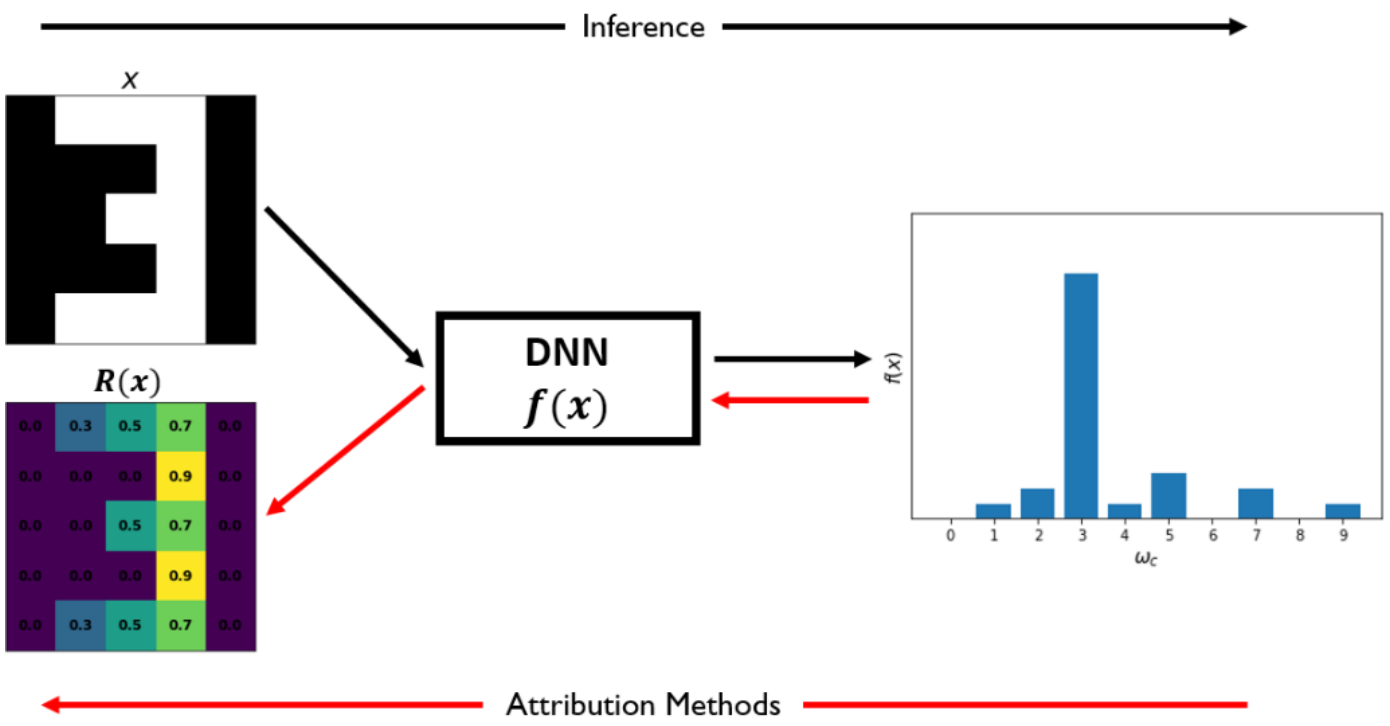
Slides by Beomsu Kim, KAIST

Types of DNN Interpretability

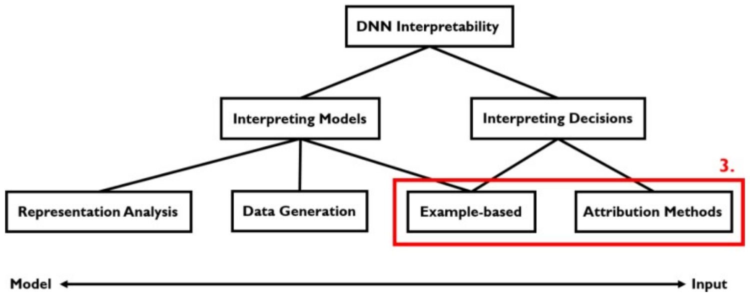


Example-based | **Attribution Methods** | **Gradient Based** | **Backprop. Based**

Attributions visualized as heatmaps

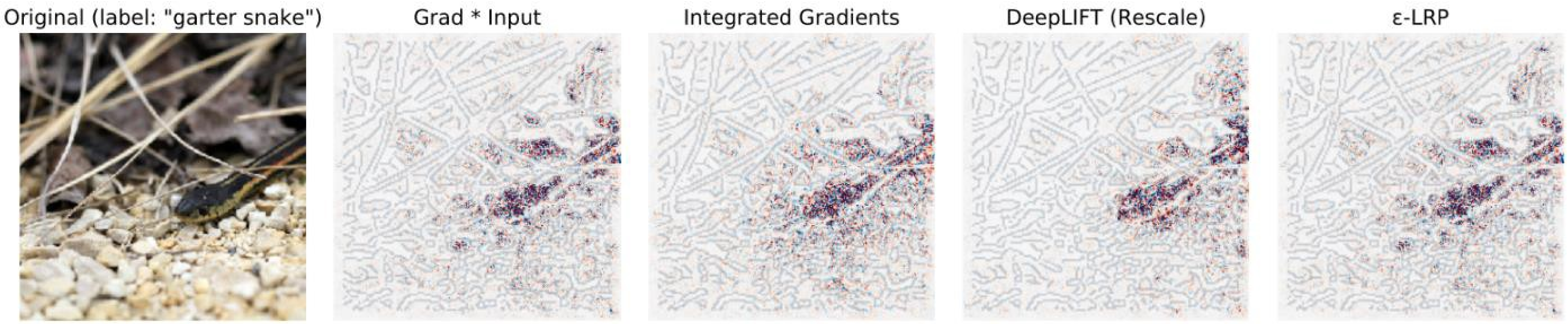


Types of DNN Interpretability



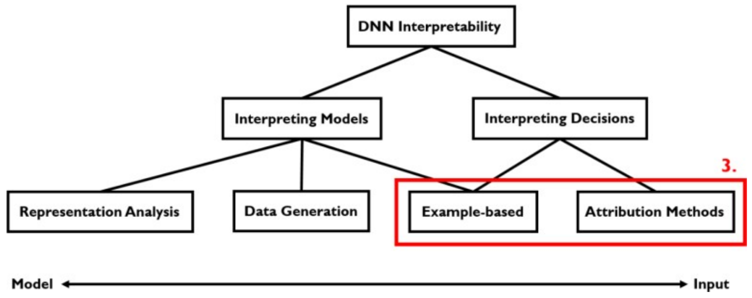
Example-based | **Attribution Methods** | **Gradient Based** | **Backprop. Based**

Attributions visualized as heatmaps



"Towards better understanding of gradient-based attribution methods for Deep Neural Networks", https://arxiv.org/pdf/1711.06104.pdf

Types of DNN Interpretability



Example-based | **Attribution Methods** | **Gradient Based** | **Backprop. Based**

The Baseline Attribution Method Saliency Map

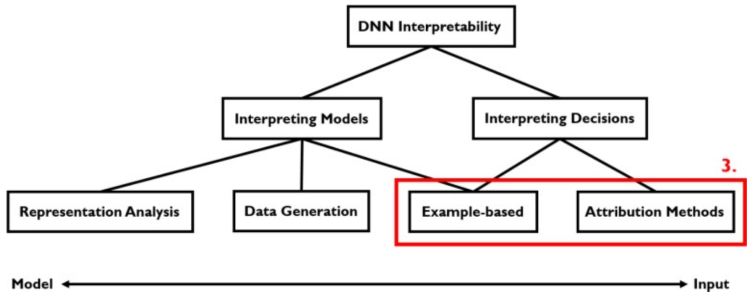
- Gradient of the decision $f(\boldsymbol{x})$ with respect to each **input image $\boldsymbol{x}$:**

$$Saliency(\boldsymbol{x}) \coloneqq \nabla_{x} f(\boldsymbol{x}) = \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}$$

- Can be calculated through backpropagation.

"Deep Inside Convolutional Networks: Visualizing Image Classification Models and Saliency Maps", https://arxiv.org/pdf/1312.6034.pdf

Slides by Beomsu Kim, KAIST



Types of DNN Interpretability

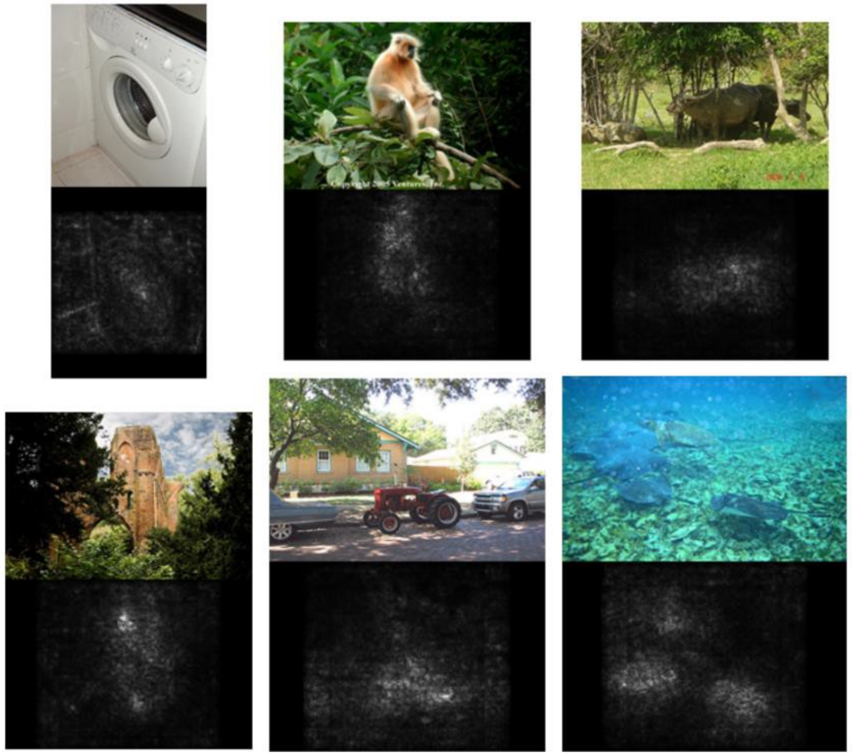
Example-based | **Attribution Methods** | **Gradient Based** | **Backprop. Based**

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?

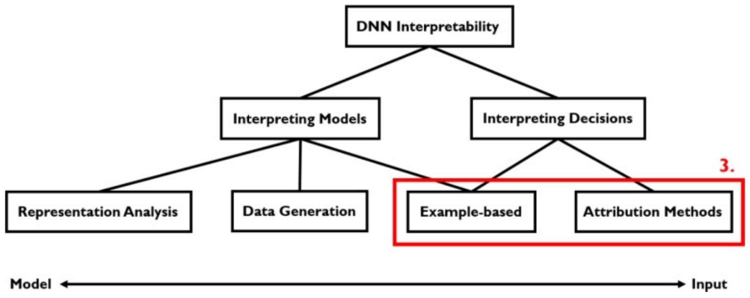
Question: Why are saliency maps noisy?



“Deep Inside Convolutional Networks: Visualizing Image Classification Models and Saliency Maps”, https://arxiv.org/pdf/1312.6034.pdf

Slides by Beomsu Kim, KAIST

Types of DNN Interpretability

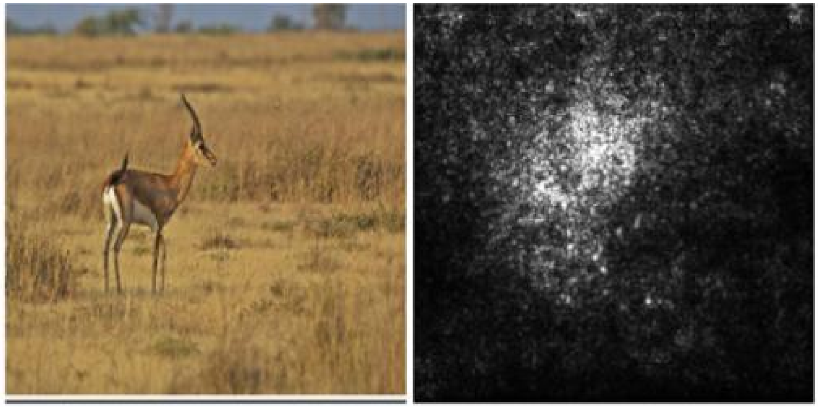


Example-based | **Attribution Methods** | **Gradient Based** | **Backprop. Based**

Question: Why are saliency maps noisy?

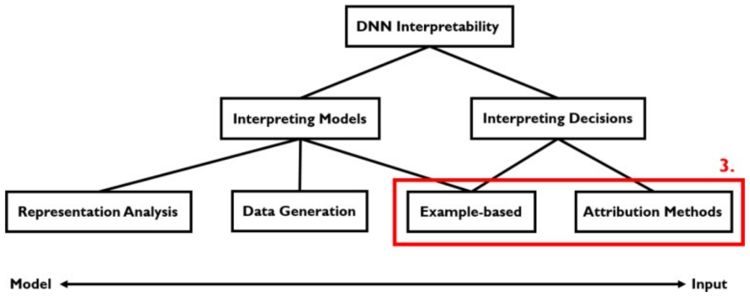
Hypothesis 1 – Saliency maps are truthful

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



"SmoothGrad: removing noisy by adding noise", https://arxiv.org/pdf/1706.03825.pdf

Slides by Beomsu Kim, KAIST



Types of DNN Interpretability

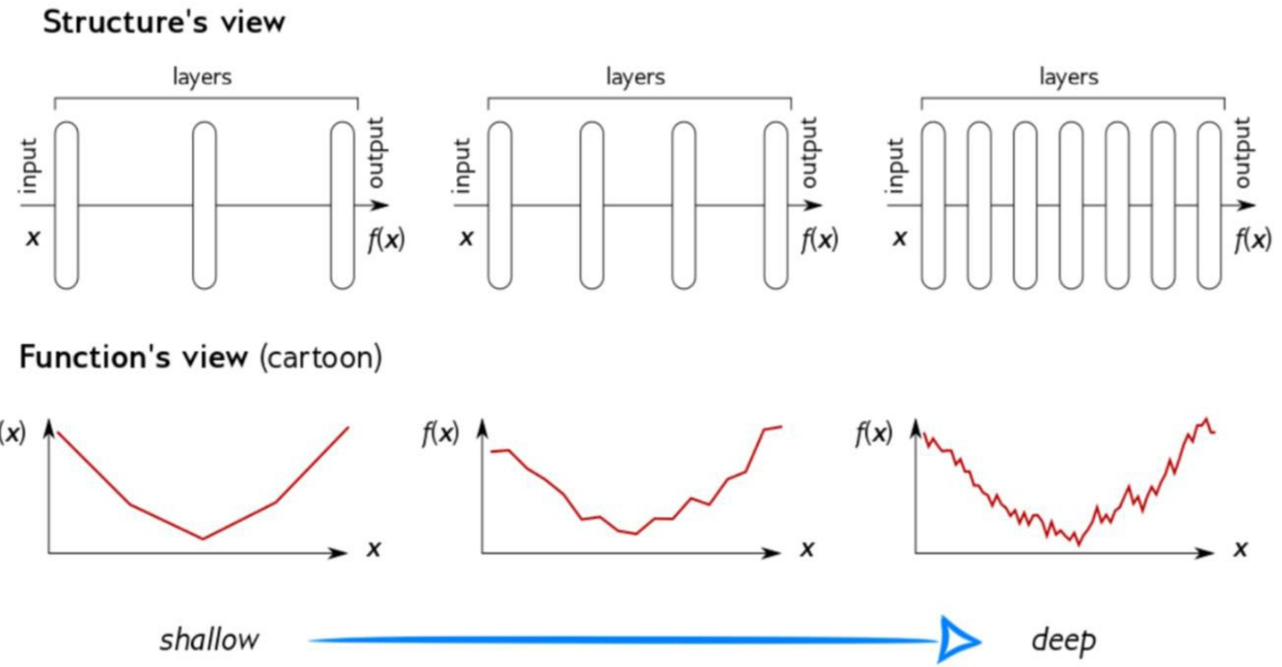
Example-based | **Attribution Methods** | **Gradient Based** | **Backprop. Based**

Question: Why are saliency maps noisy?

Hypothesis 2 – Gradients are discontinuous

- DNN uses piecewise-linear functions (ReLU activation, max-pooling, etc.).

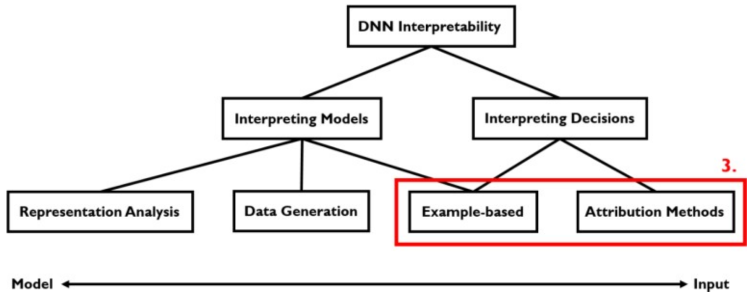
-Sudden jumps in the importance score over infinitesimal changes in the input.



“SmoothGrad: removing noisy by adding noise”, https://arxiv.org/pdf/1706.03825.pdf

Slides by Beomsu Kim, KAIST

Types of DNN Interpretability



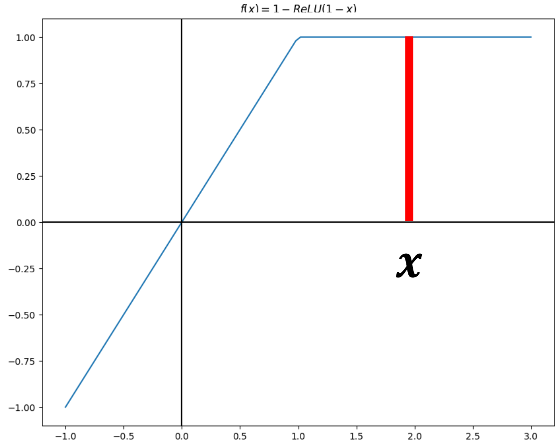
Example-based | **Attribution Methods** | **Gradient Based** | **Backprop. Based**

Question: Why are saliency maps noisy?

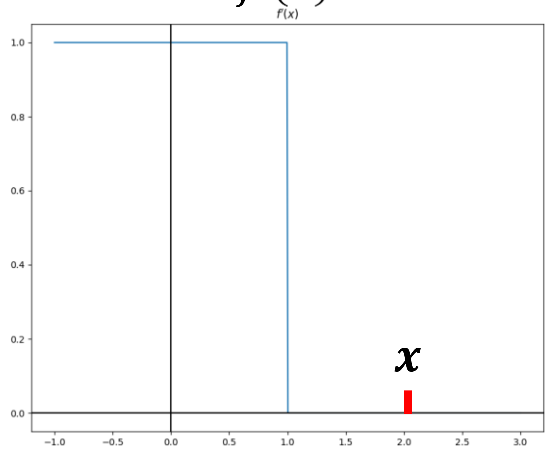
Hypothesis 3 – $f(x)$ saturates

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

$f(x)$



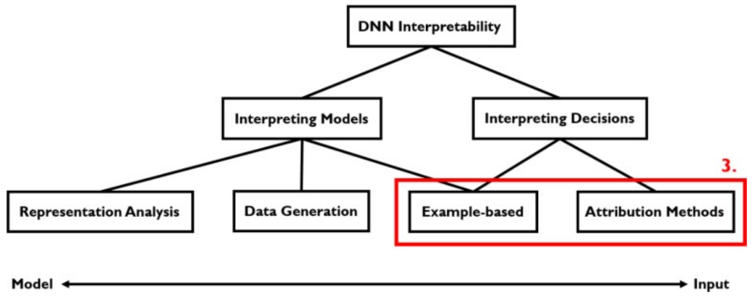
$f'(x)$



“Axiomatic Attribution for Deep Networks”, https://arxiv.org/pdf/1703.01365.pdf

Slides by Beomsu Kim, KAIST

Types of DNN Interpretability



Example-based | **Attribution Methods** | **Gradient Based** | **Backprop. Based**

Question: How to improve saliency maps?

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

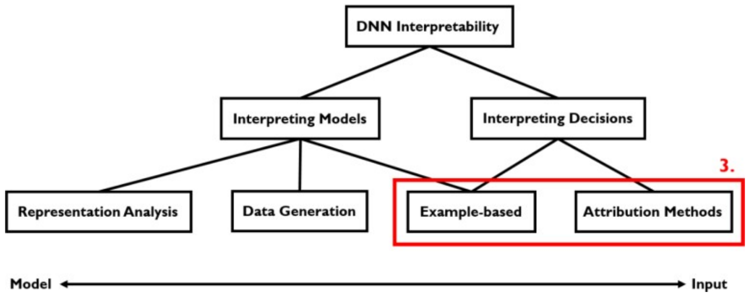
Gradient-based Methods

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

Backprop-based Methods

- Modify the backpropagation algorithm.

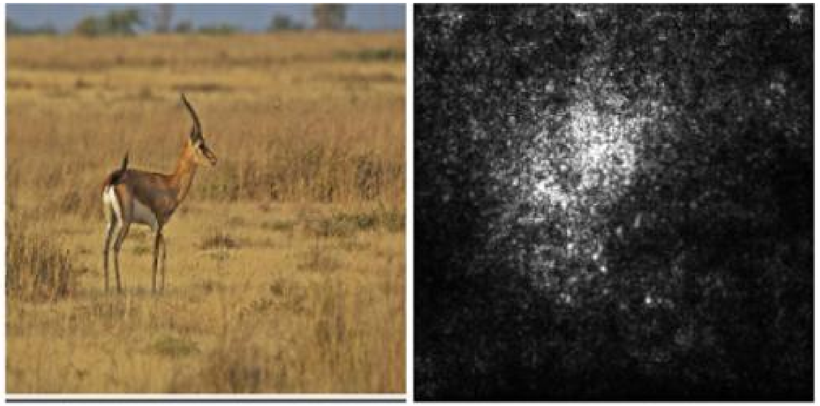
Slides by Beomsu Kim, KAIST



Types of DNN Interpretability

Example-based | **Attribution Methods** | **Gradient Based** | **Backprop. Based**

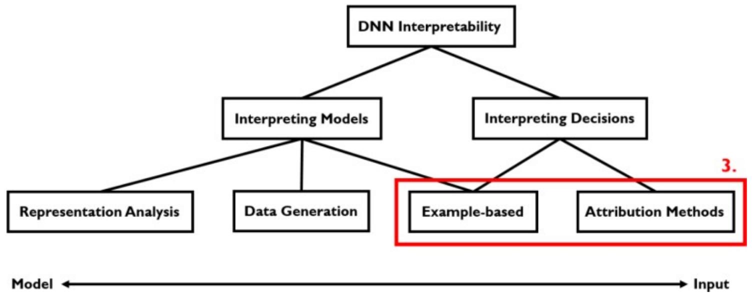
Summary



- 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.

Slides by Beomsu Kim, KAIST

Types of DNN Interpretability



Example-based | **Attribution Methods** | **Gradient Based** | **Backprop. Based**

Other Attribution Methods

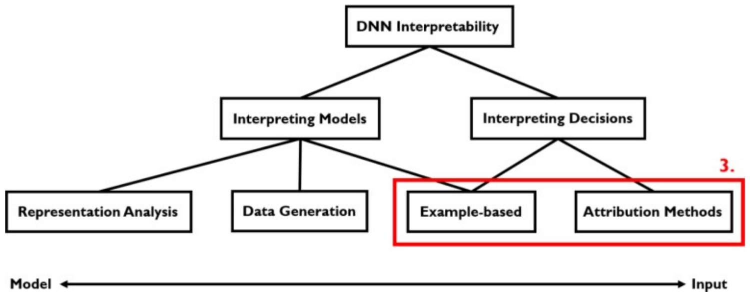
- Gradient * Input – https://arxiv.org/pdf/1704.02685.pdf
- Integrated Gradient – https://arxiv.org/pdf/1703.01365.pdf
- Layer-wise Relevance Propagation – https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130140
- Deep Taylor Decomposition – https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130140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

2c – Interpreting decisions:

(i) Example based
(ii) Attribution Methods: why are gradients noisy?
(iii) Gradient-based Attribution: SmoothGrad, Interior Gradient
(iv) Backprop-based Attribution: Deconvolution, Guided Backprop

Slides by Beomsu Kim, KAIST

Types of DNN Interpretability

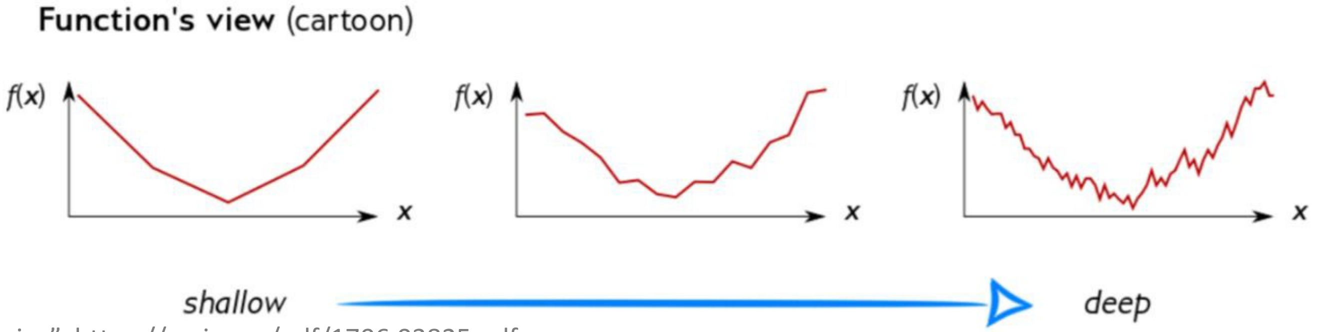


Example-based | **Attribution Methods** | **Gradient Based** | **Backprop. Based**

1. SmoothGrad *Hypothesis 2 – Gradients are discontinuous: Just smooth the gradient!*

$$SmoothGrad(\boldsymbol{x}) \coloneqq \frac{1}{n}\int_{1}^{n} \frac{\partial f(\boldsymbol{x}^*)}{\partial \boldsymbol{x}^*}, \quad \boldsymbol{x}^* = \boldsymbol{x} + \mathcal{N}(0, \sigma^2)$$

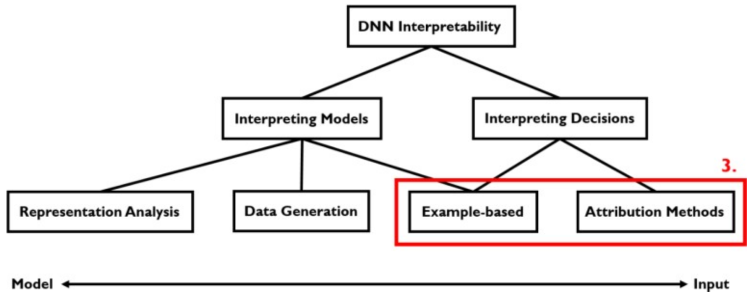
Gaussian smoothing



“SmoothGrad: removing noisy by adding noise”, https://arxiv.org/pdf/1706.03825.pdf

Slides by Beomsu Kim, KAIST

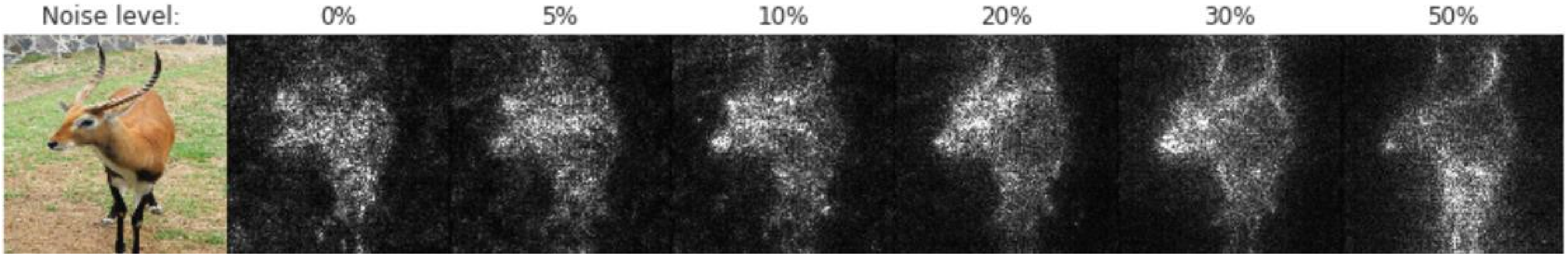
Types of DNN Interpretability



Example-based | **Attribution Methods** | **Gradient Based** | **Backprop. Based**

1. SmoothGrad *Hypothesis 2 – Gradients are discontinuous: Just smooth the gradient!*

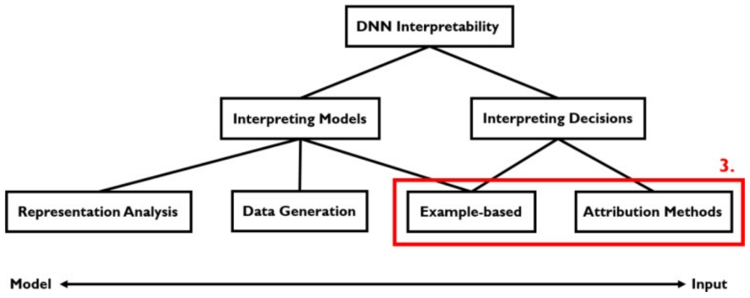
$$SmoothGrad(\boldsymbol{x}) \coloneqq \frac{1}{n}\int_{1}^{n} \frac{\partial f(\boldsymbol{x}^*)}{\partial \boldsymbol{x}^*}, \quad \boldsymbol{x}^* = \boldsymbol{x} + \mathcal{N}(0, \sigma^2)$$



"SmoothGrad: removing noisy by adding noise", https://arxiv.org/pdf/1706.03825.pdf

Slides by Beomsu Kim, KAIST

Types of DNN Interpretability

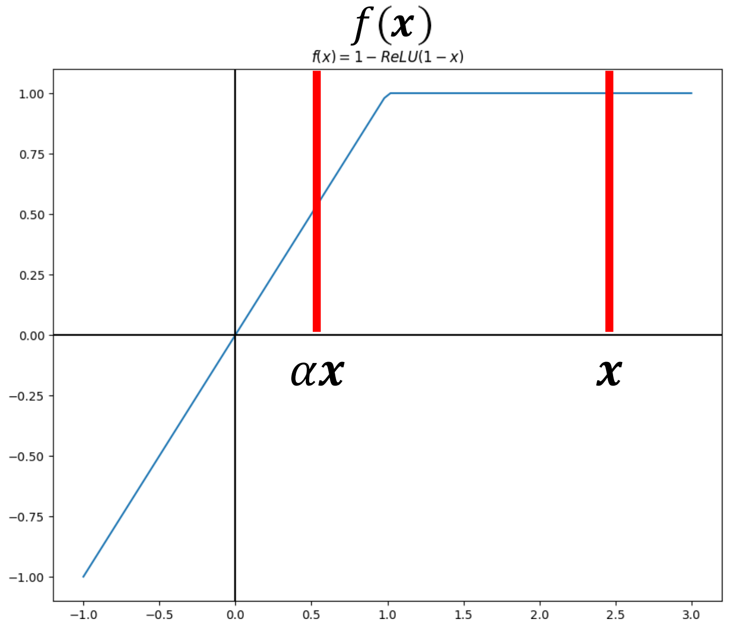


Example-based | **Attribution Methods** | **Gradient Based** | **Backprop. Based**

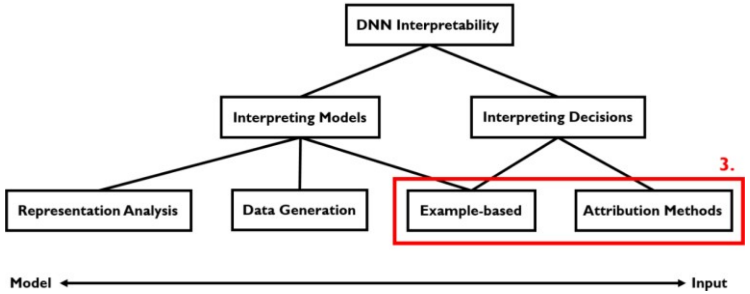
2. Interior Gradient *Hypothesis 3 – $f(\boldsymbol{x})$ saturates*

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

- Appropriate α will trigger informative activation functions



"Gradients of Counterfactuals", https://arxiv.org/pdf/1611.02639.pdf



Types of DNN Interpretability

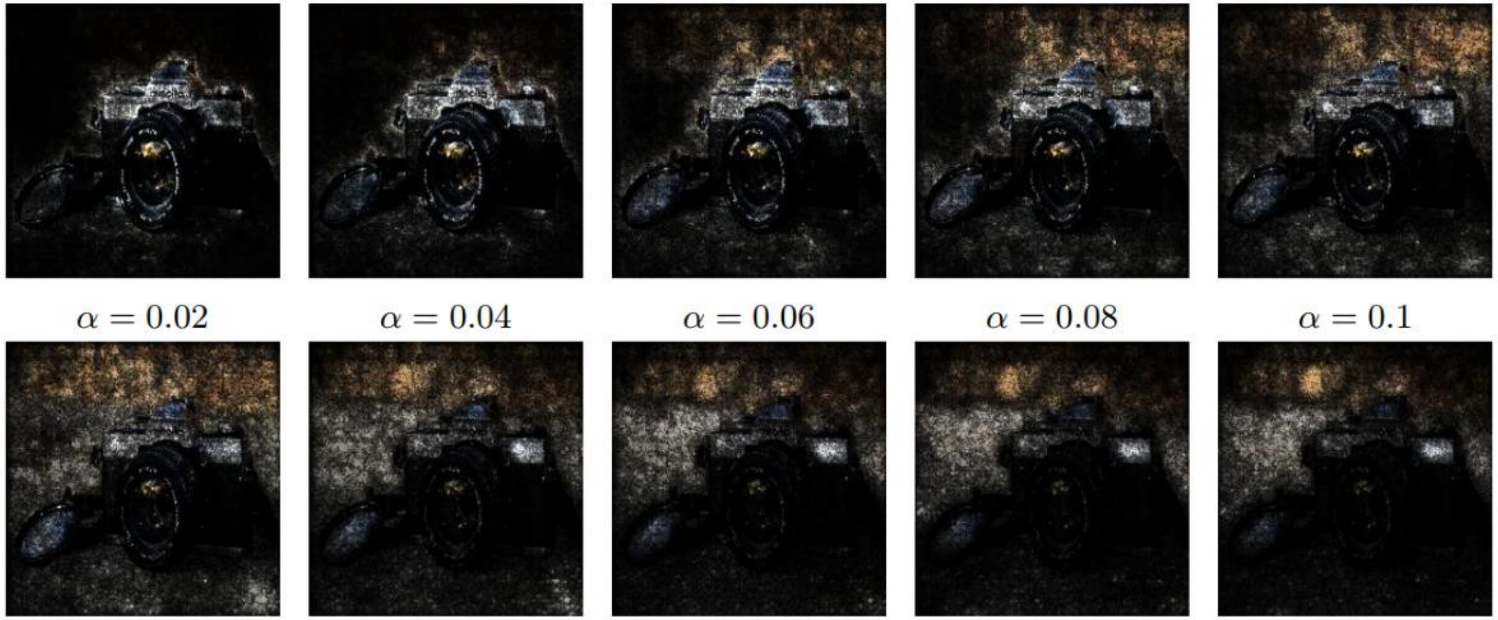
Example-based | **Attribution Methods** | **Gradient Based** | **Backprop. Based**

2. Interior Gradient

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

$$\boldsymbol{x}^* = \alpha \boldsymbol{x},$$

$$0 < \alpha \leq 1$$


$\alpha = 0.02$ $\alpha = 0.04$ $\alpha = 0.06$ $\alpha = 0.08$ $\alpha = 0.1$
$\alpha = 0.2$ $\alpha = 0.4$ $\alpha = 0.6$ $\alpha = 0.8$ $\alpha = 1.0$

"Gradients of Counterfactuals", https://arxiv.org/pdf/1611.02639.pdf

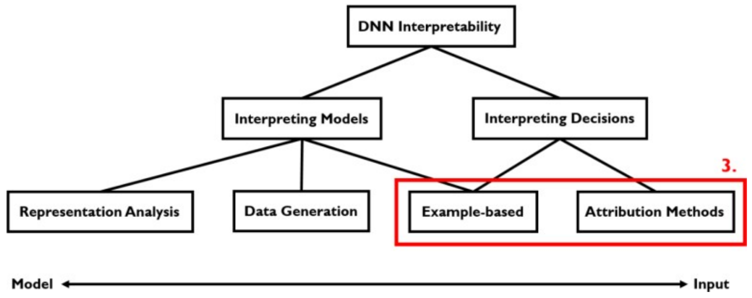
2c – Interpreting decisions:

(i) Example based

(ii) Attribution Methods: why are gradients noisy?

(iii) Gradient-based Attribution: SmoothGrad, Interior Gradient

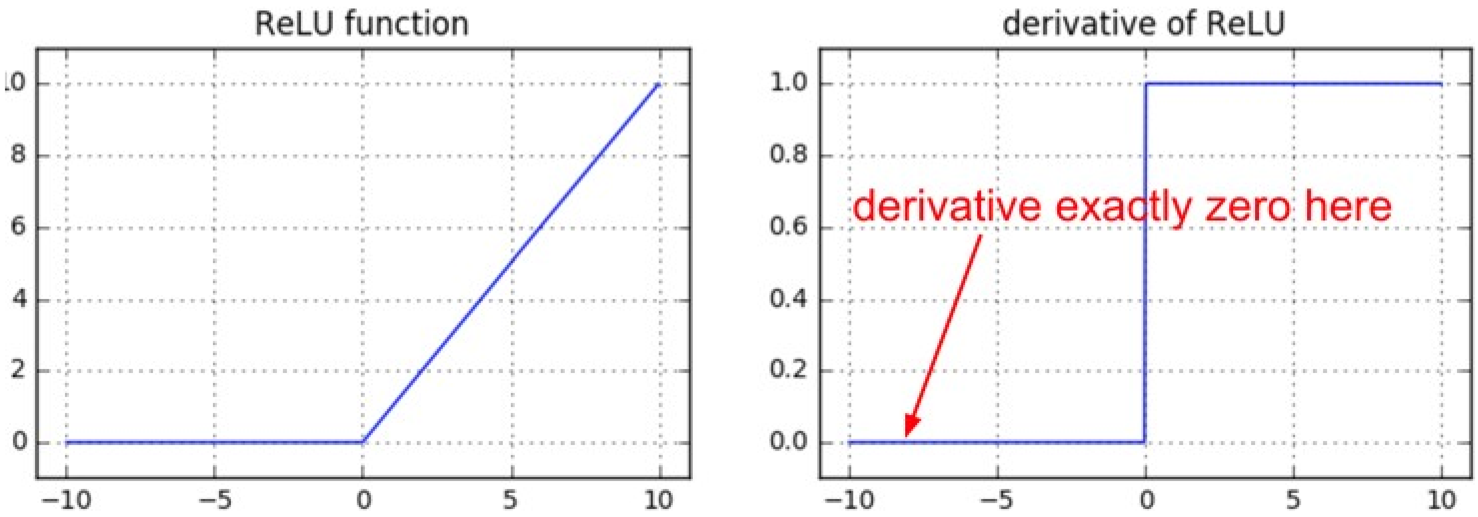
(iv) Backprop-based Attribution: Deconvolution, Guided Backprop



Types of DNN Interpretability

Example-based | **Attribution Methods** | **Gradient Based** | **Backprop. Based**

Review: Backpropagation at ReLU

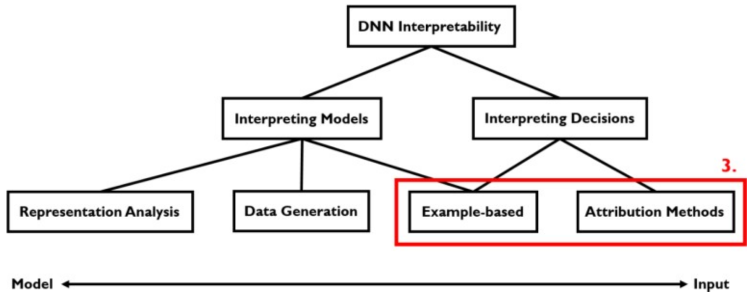


$$ReLU(z) = \max(0, z)$$

$$ReLU'^{(z)} = (z > 0)$$

“Striving for Simplicity: The All Convolutional Net”, https://arxiv.org/pdf/1412.6806.pdf

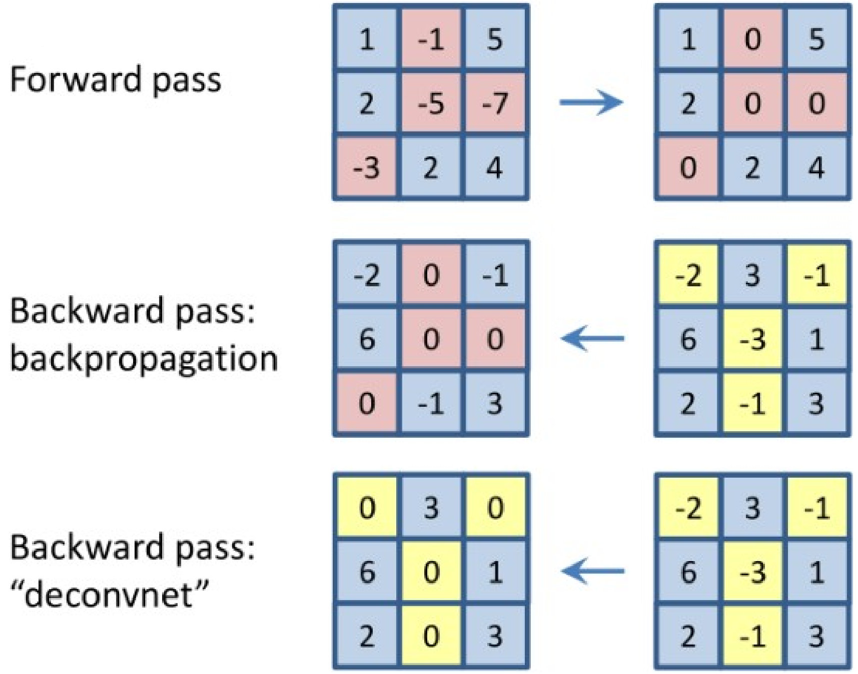
Types of DNN Interpretability



Example-based | **Attribution Methods** | **Gradient Based** | **Backprop. Based**

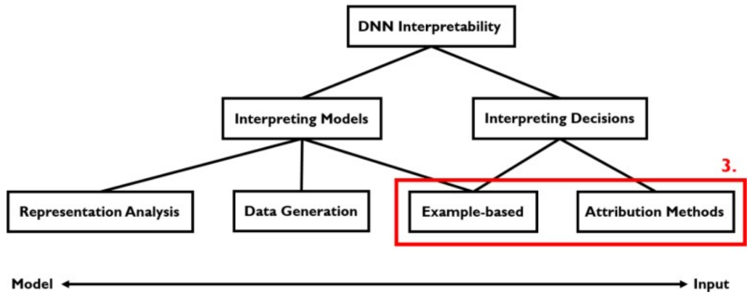
1. Deconvnet

- 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



"Visualizing and Understanding Convolutional Networks", https://arxiv.org/pdf/1311.2901.pdf

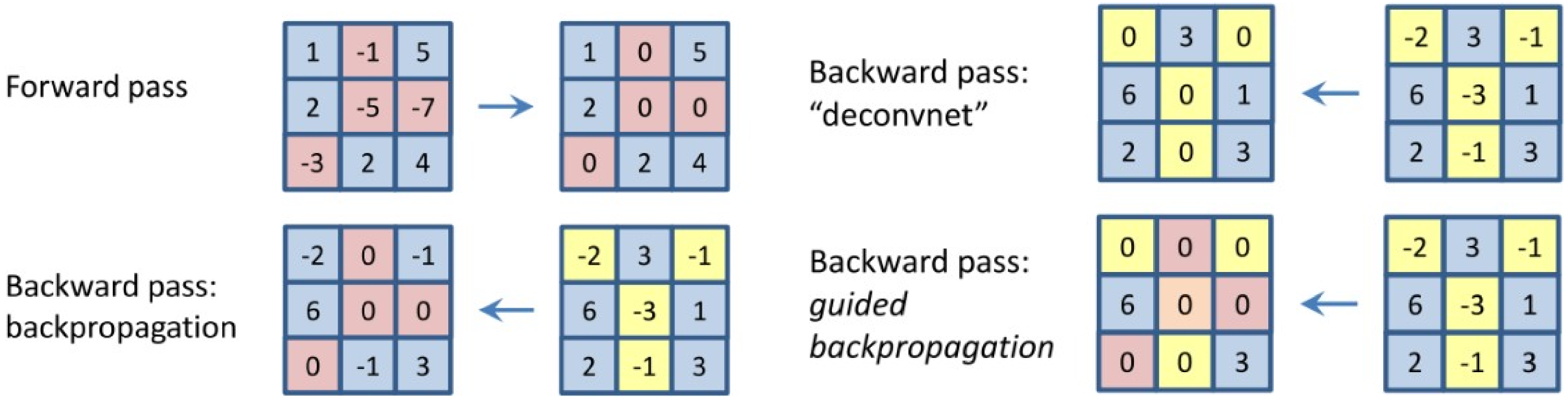
Types of DNN Interpretability



Example-based | **Attribution Methods** | **Gradient Based** | **Backprop. Based**

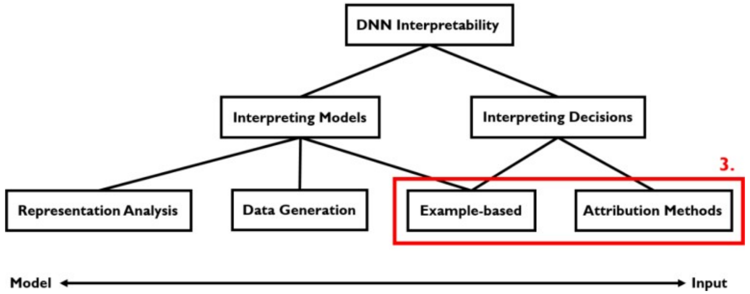
2. Guided Backpropagation

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



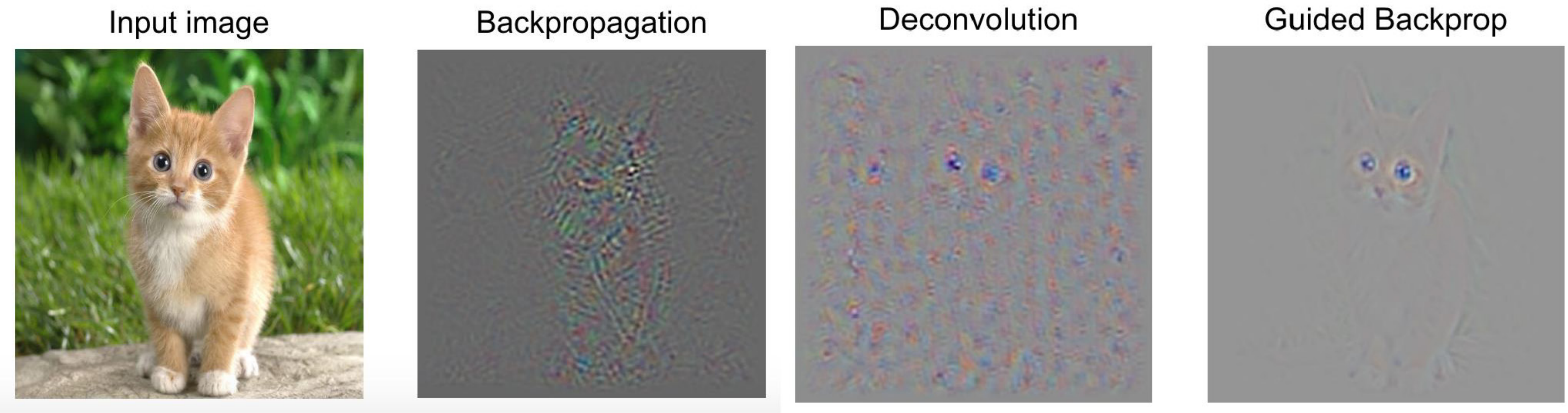
"Striving for Simplicity: The All Convolutional Net", https://arxiv.org/pdf/1412.6806.pdf

Slides by Beomsu Kim, KAIST



Types of DNN Interpretability

Example-based | **Attribution Methods** | **Gradient Based** | **Backprop. Based**



Observation: Removing more gradient leads to sharper visualizations

Interpretable Deep Learning

1. Intro to Interpretability

1a. Interpretability definition: Convert implicit NN information to human-interpretable information

1b. Motivation: Verify model works as intended; debug classifier; make discoveries; Right to explanation

1c. Ante-hoc (train interpretable model) vs. **Post-hoc** (interpret complex model; degree of “locality”)

2. Interpreting Deep Neural Networks

2a. Interpreting Models (macroscopic, understand internals) vs. **decisions** (microscopic, practical applications)

2b. Interpreting Models: Weight visualization, Surrogate model, Activation maximization, Example-based

2c. Interpreting Decisions:

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

3. Evaluating Attribution Methods

3a. Qualitative: Coherence: Attributions should highlight discriminative features / objects of interest

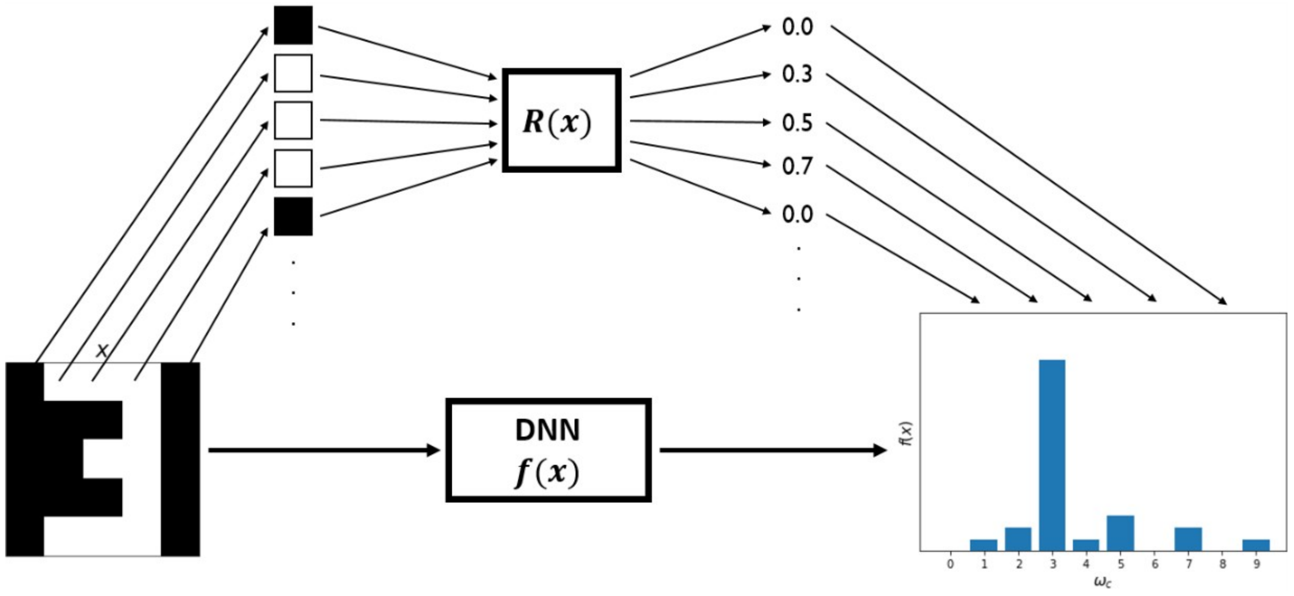
3b. Qualitative: Class Sensitivity: Attributions should be sensitive to class labels

3c. Quantitative: Sensitivity: Removing feature with high attribution ➔ large decrease in class probability

3d. Quantitative: ROAR & KAR. Low class prob cuz image unseen ➔ remove pixels, retrain, measure acc. drop

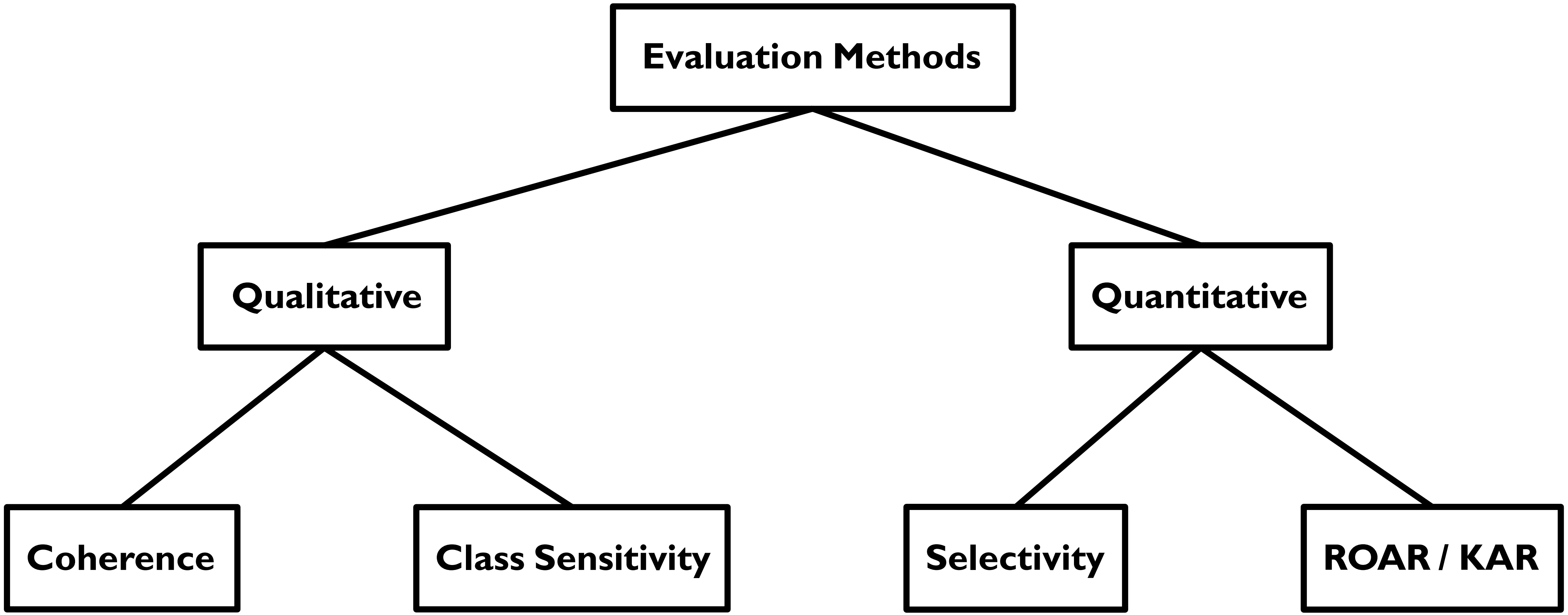
Attribution Method Review

Given an image $\boldsymbol{x} \in \mathbb{R}^n$ and a decision $f(\boldsymbol{x})$,
assign to each pixel $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$ attribution values $R_1(\boldsymbol{x}), R_2(\boldsymbol{x}), \dots, R_n(\boldsymbol{x})$.



Slides by Beomsu Kim, KAIST

Evaluating Attribution Methods

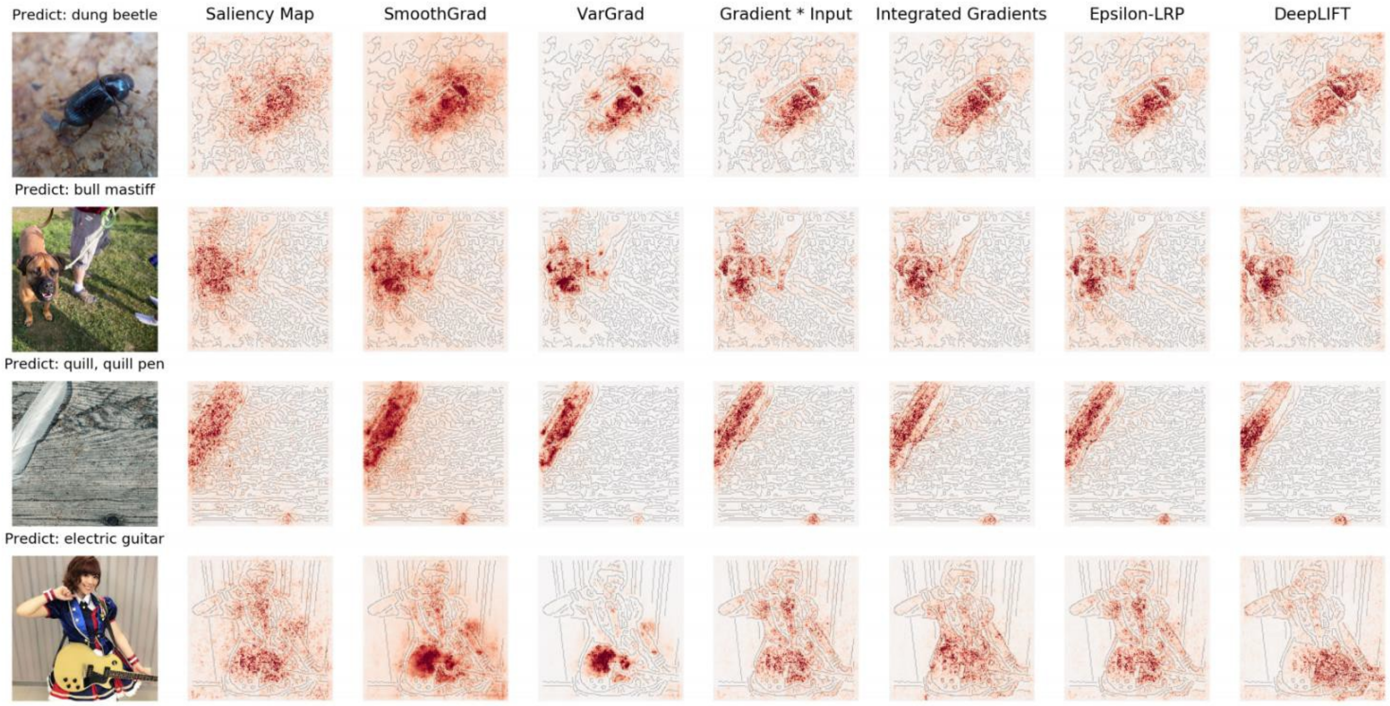


3a – Qualitative: coherence

Evaluating Attribution Methods

Coherence | **Class Sensitivity** | **Selectivity** | **ROAR / KAR**

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

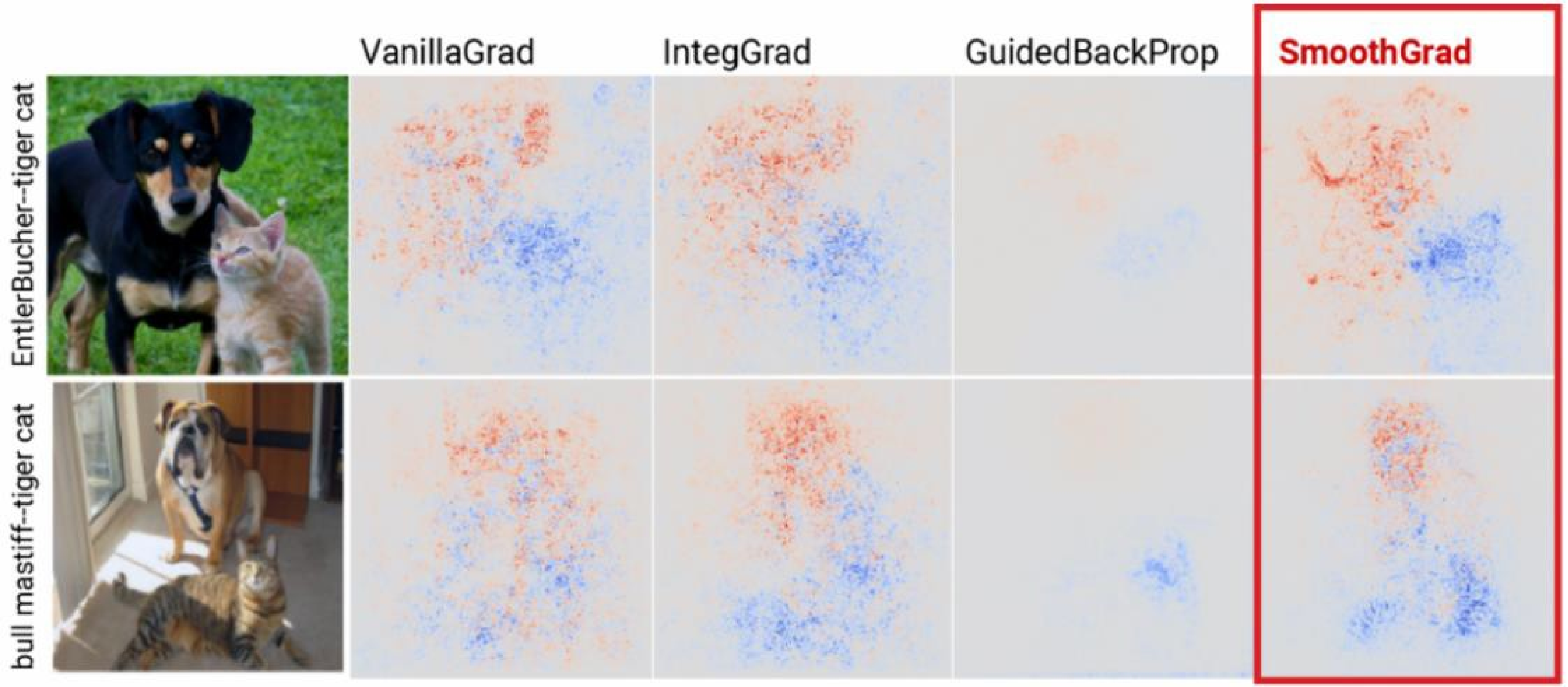


"Noise-adding Methods of Saliency Maps as Series of Higher Order Partial Derivative", https://arxiv.org/pdf/1806.03000.pdf

Evaluating Attribution Methods

Coherence | **Class Sensitivity** | **Selectivity** | **ROAR / KAR**

- Attributions should be sensitive to class labels



"SmoothGrad: removing noisy by adding noise", https://arxiv.org/pdf/1706.03825.pdf

Evaluating 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 $R_i(\boldsymbol{x})$

Iterate:

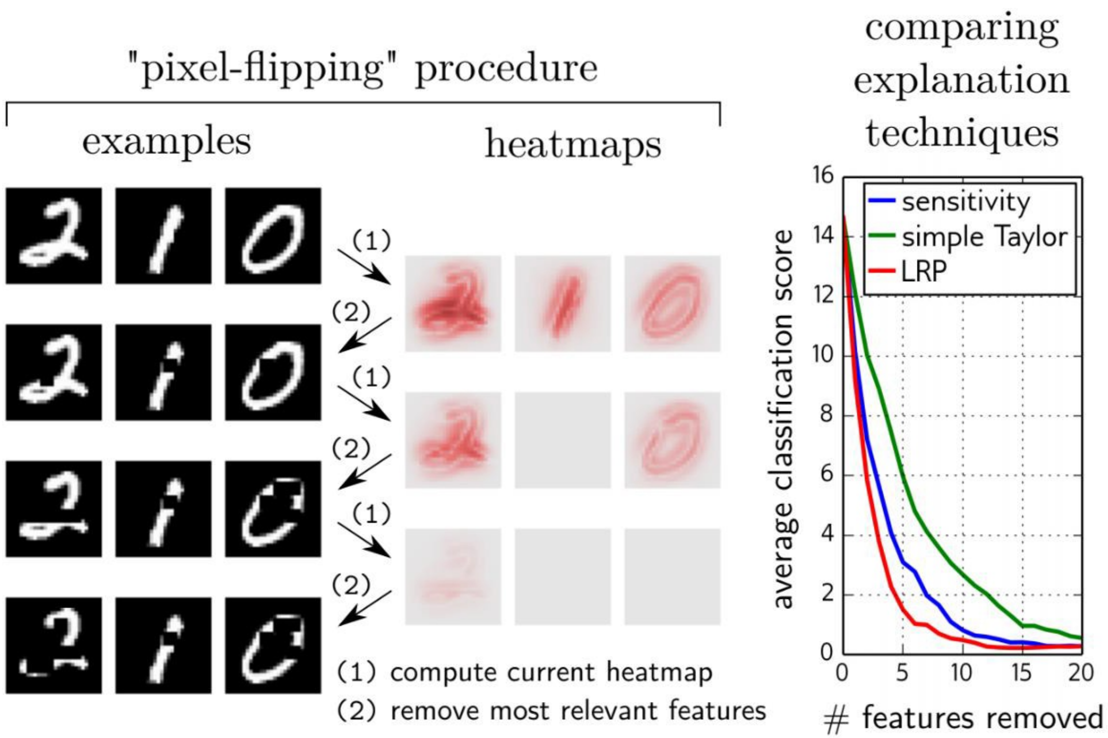
  Remove pixels

  Evaluate $f(\boldsymbol{x})$

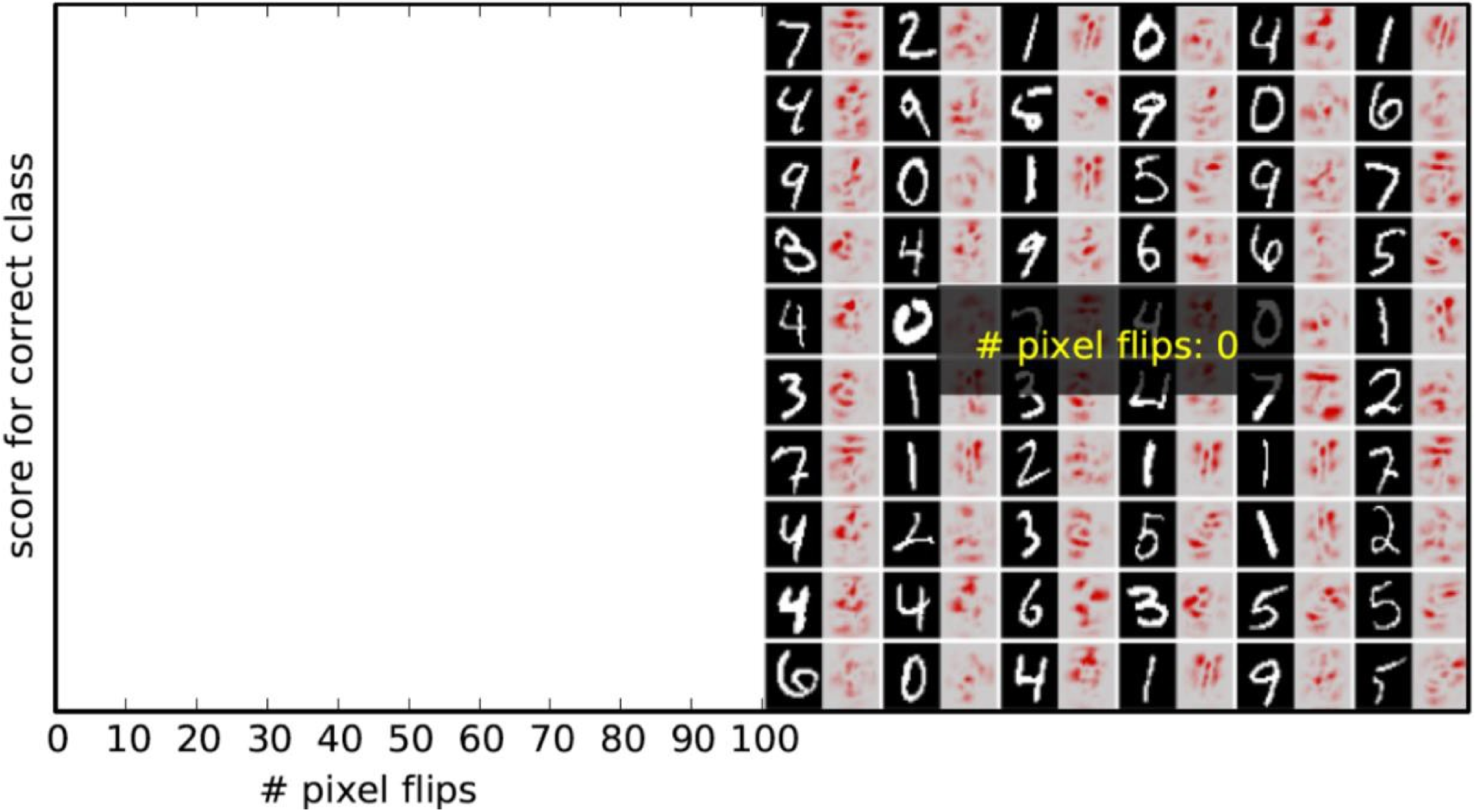
Measure decrease of $f(\boldsymbol{x})$

Evaluating Attribution Methods

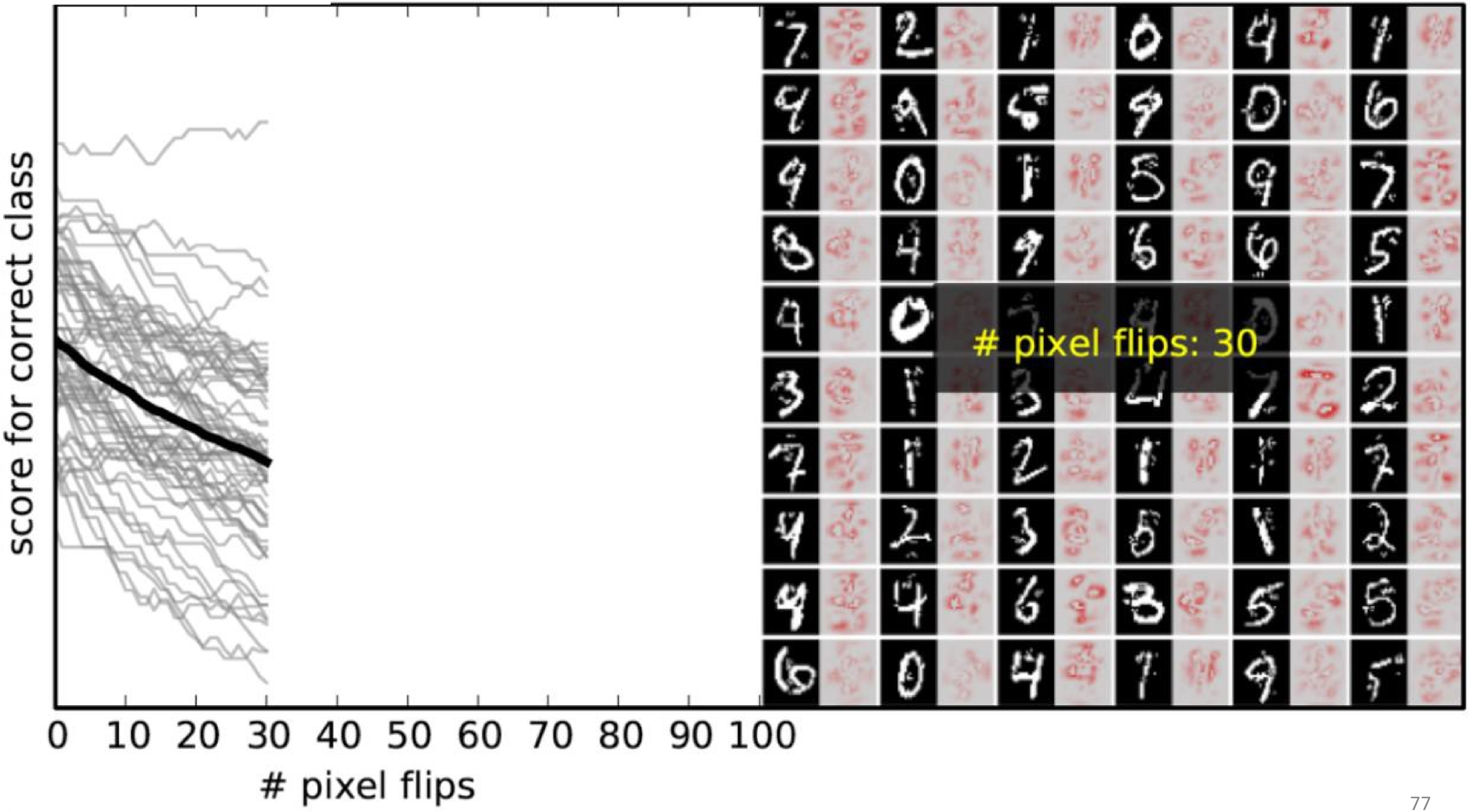
Coherence | **Class Sensitivity** | **Selectivity** | **ROAR / KAR**



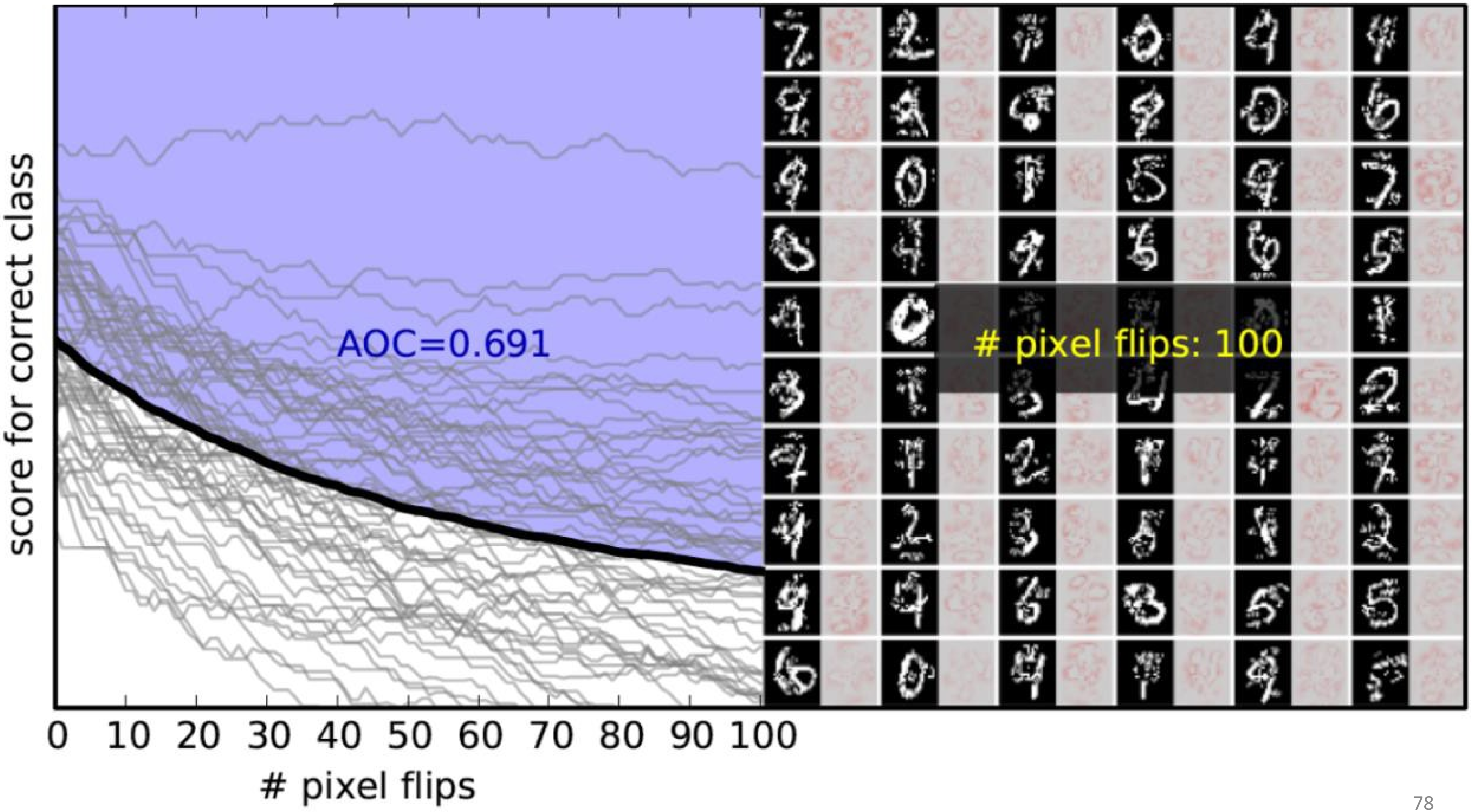
Selectivity on Saliency Map



Selectivity on Saliency Map



Selectivity on Saliency Map



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)

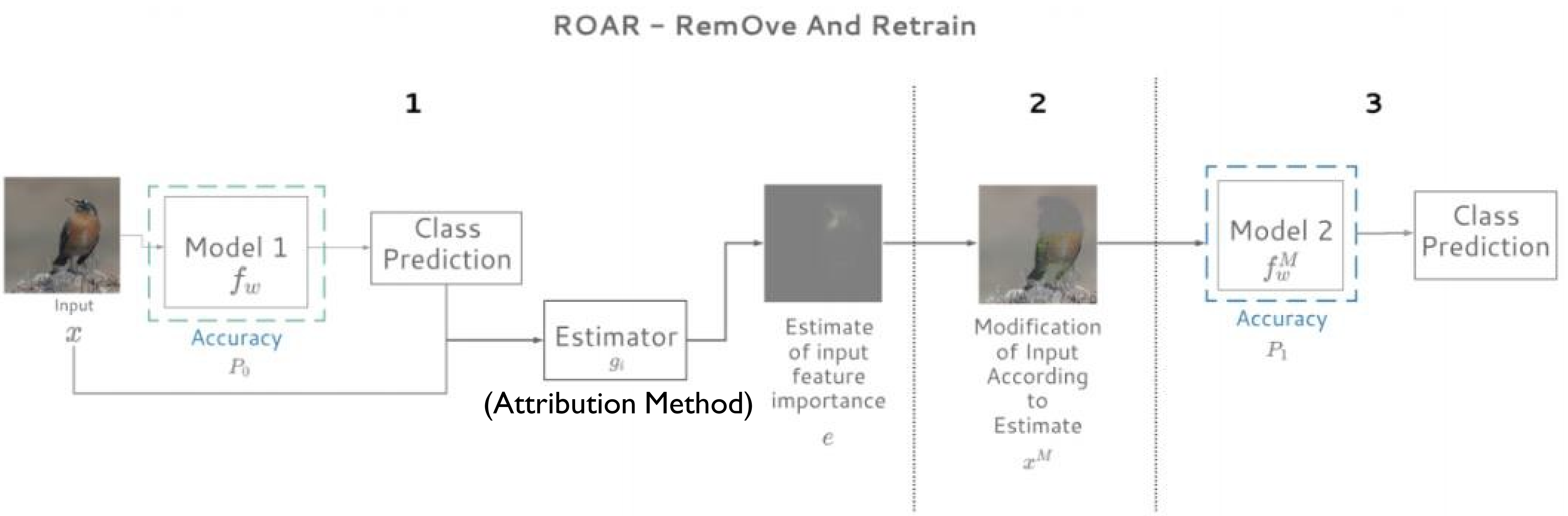
Measure how the performance of the classifier changes as features are removed based on the attribution method

- 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 Feature Importance Estimates”, https://arxiv.org/pdf/1806.10758.pdf

Evaluating Attribution Methods

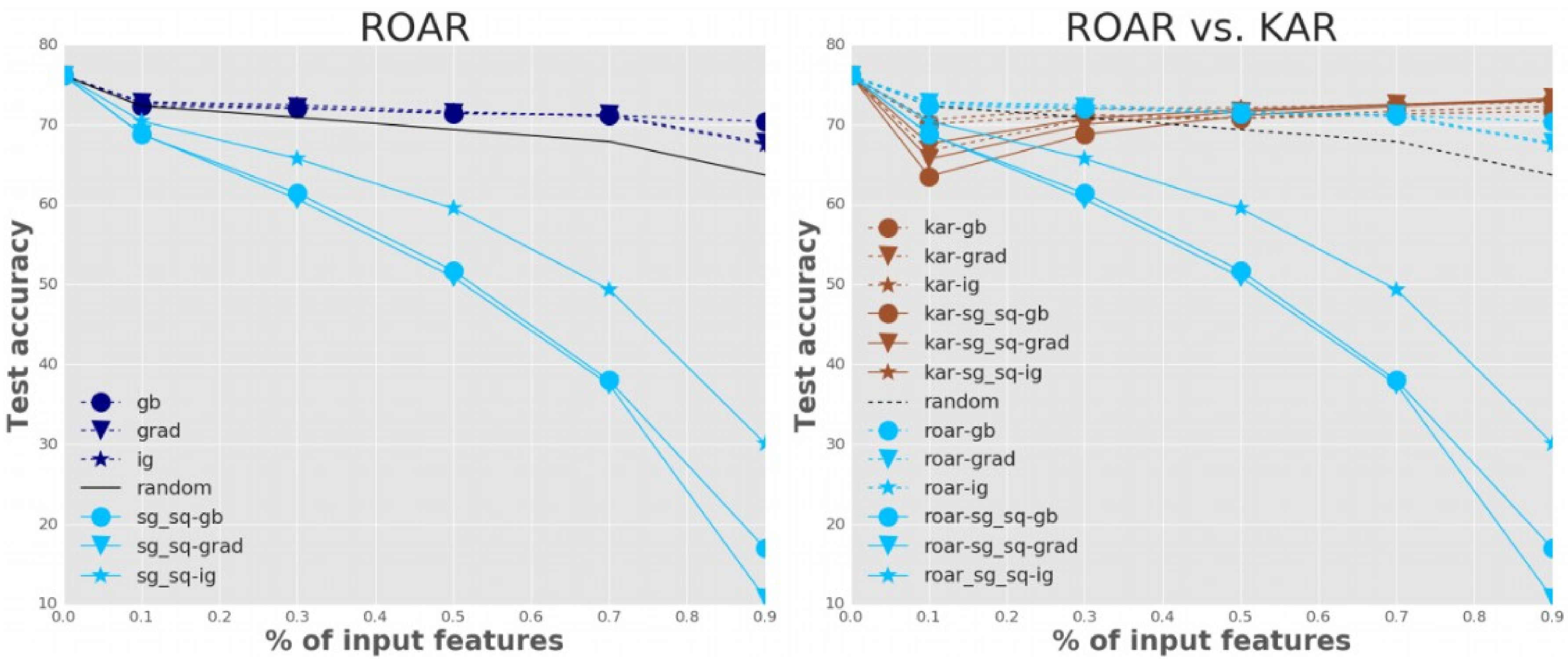
Coherence | **Class Sensitivity** | **Selectivity** | **ROAR / KAR**



"Evaluating Feature Importance Estimates", https://arxiv.org/pdf/1806.10758.pdf

Evaluating Attribution Methods

Coherence | **Class Sensitivity** | **Selectivity** | **ROAR / KAR**



"Evaluating Feature Importance Estimates", https://arxiv.org/pdf/1806.10758.pdf

Interpretable Deep Learning

1. Intro to Interpretability

1a. Interpretability definition: Convert implicit NN information to human-interpretable information

1b. Motivation: Verify model works as intended; debug classifier; make discoveries; Right to explanation

1c. Ante-hoc (train interpretable model) vs. **Post-hoc** (interpret complex model; degree of “locality”)

2. Interpreting Deep Neural Networks

2a. Interpreting Models (macroscopic, understand internals) vs. **decisions** (microscopic, practical applications)

2b. Interpreting Models: Weight visualization, Surrogate model, Activation maximization, Example-based

2c. Interpreting Decisions:

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

3. Evaluating Attribution Methods

3a. Qualitative: Coherence: Attributions should highlight discriminative features / objects of interest

3b. Qualitative: Class Sensitivity: Attributions should be sensitive to class labels

3c. Quantitative: Sensitivity: Removing feature with high attribution ➔ large decrease in class probability

3d. Quantitative: ROAR & KAR. Low class prob cuz image unseen ➔ remove pixels, retrain, measure acc. drop

Summary

1. 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. Interpreting 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. Evaluating Attribution Methods

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

Slides by Beomsu Kim, KAIST