Visual Explainable AI for Convolutional Neural Networks

ECE 1512: DIGITAL IMAGE PROCESSING AND APPLICATIONS

SEMESTER: WINTER 2022 PROJECT "A" TUTORIAL

PRESENTER: AHMAD SAJEDI

Based on NeurIPS'2020/AAAI'21 tutorial on Explainable AI: https://explainml-tutorial.github.io/

Outline

A. Tutorial on visual explainable AI (XAI)
   - Motivation
   - Primer on explainability in Artificial Intelligence (AI)
   - Approaches for visual explanation generation

B. Project “A” Description
   - Project Goal
   - Datasets and Models
   - Evaluation Metrics

C. Your Questions!

Outline

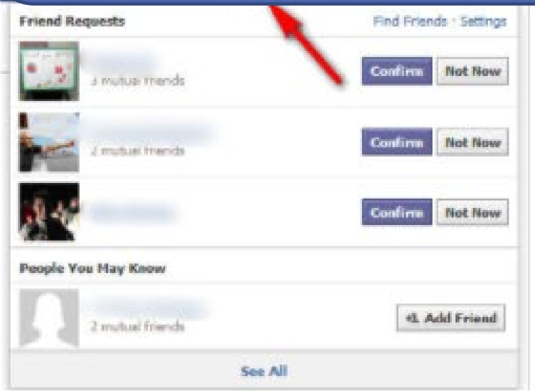
A. Tutorial on visual explainable AI (XAI)
- Motivation
- Primer on explainability in Artificial Intelligence (AI)
- Approaches for visual explanation generation

B. Project “A” Description
- Project Goal
- Datasets and Models
- Evaluation Metrics

C. Your Questions!

Motivation

Machine Learning is everywhere!



Friend Requests Find Friends · Settings

3 mutual friends Confirm Not Now

2 mutual friends Confirm Not Now

Confirm Not Now

People You May Know

2 mutual friends +1 Add Friend

See All

amazon.com **More to Explore**

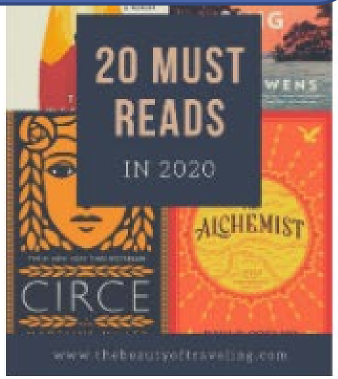
Customers who have shown an interest in point-and-shoot cameras might like to see this week's bestselling models.

Canon PowerShot A495 10.0 MP Digital Camera with 3.3x Optical Zoom and 2.5-Inch LCD (Blue)

Canon PowerShot A3000IS 10 MP Digital Camera with 4x Optical Image Stabilized Zoom and 2.7-Inch LCD

Canon PowerShot ELPH 300 HS 12 MP CMOS Digital Camera with Full 1080p HD Video (Black)

Canon PowerShot S95 10 MP Digital Camera with 3.8x Wide Angle Optical Image Stabilized Zoom and 3.0-Inch inch LCD



https://explainml-tutorial.github.io/

Motivation

https://explainml-tutorial.github.io/

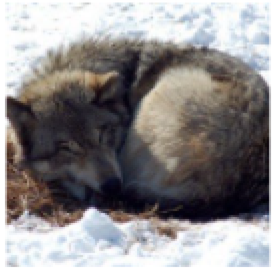
Model understanding is critical in several domains, such as:

- Autonomous driving
- Healthcare
- Criminal justice

Why model understanding?

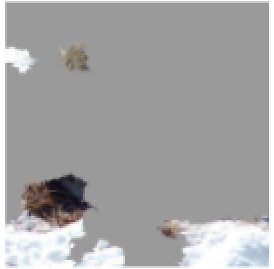
A model may make correct predictions based on irrelevant/incorrect evidence.

Input



Predictive Model

Model Understanding

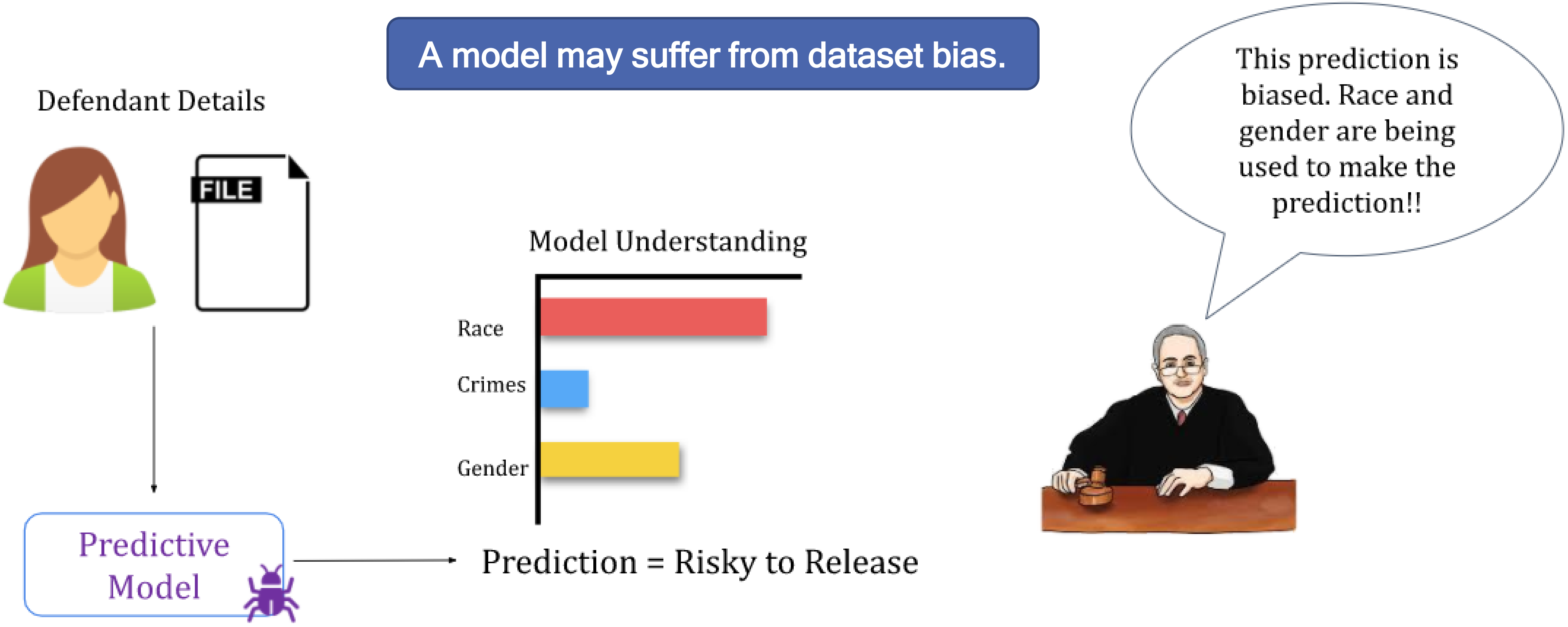


Prediction = Siberian Husky

This model is relying on incorrect features to make this prediction!! Let me fix the model

https://explainml-tutorial.github.io/

Why model understanding?



A model may suffer from dataset bias.

Defendant Details

Predictive Model

Model Understanding

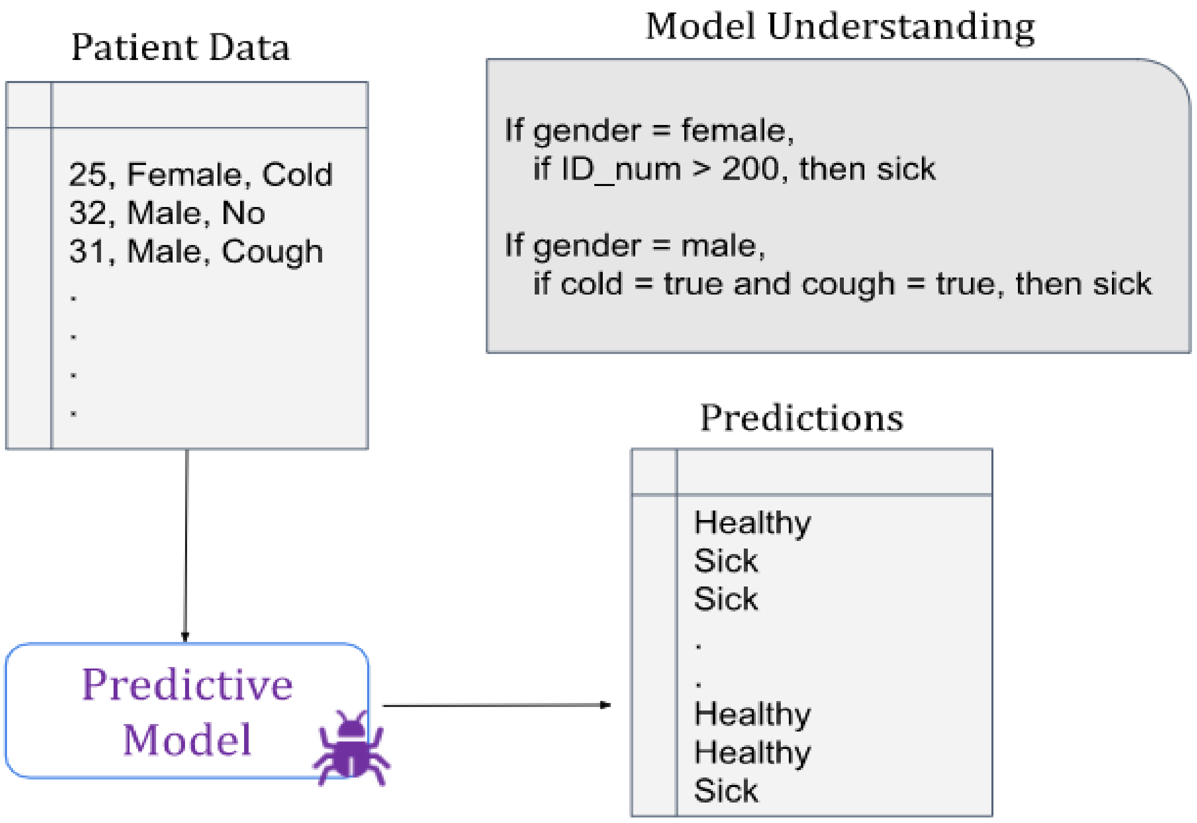
Prediction = Risky to Release

This prediction is biased. Race and gender are being used to make the prediction!!

https://explainml-tutorial.github.io/

Why model understanding?

A model may be reliable in some cases, while being unreliable in some other cases



Patient Data

25, Female, Cold
32, Male, No
31, Male, Cough
.
.
.
.

Predictive Model

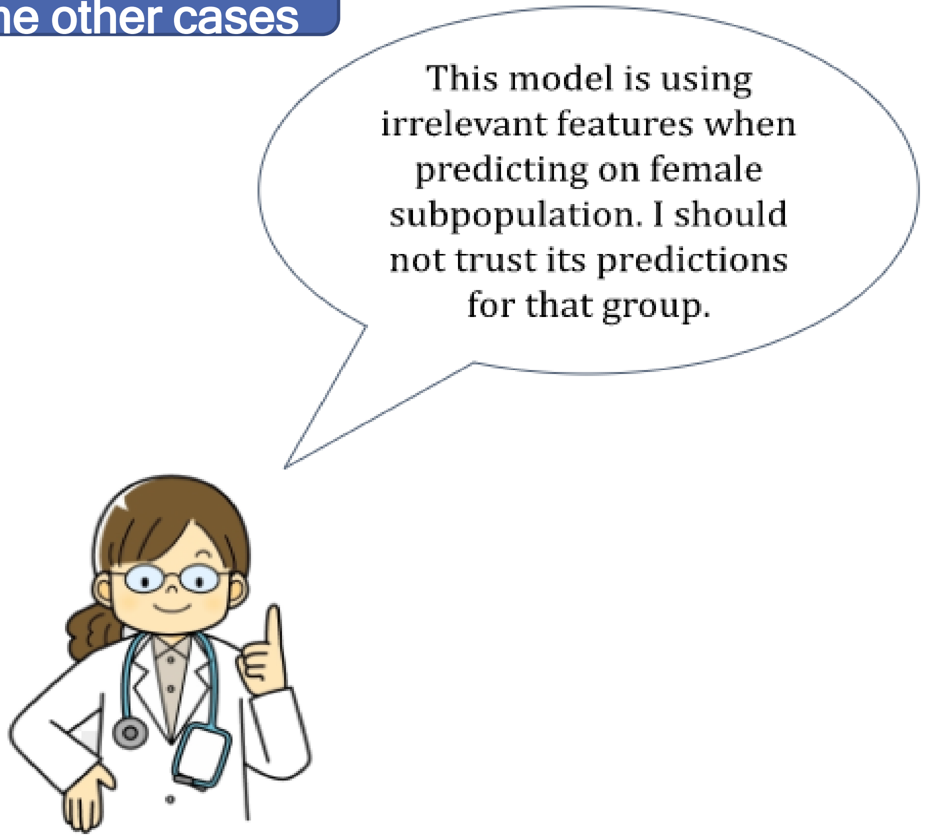
Model Understanding

```
If gender = female,
   if ID_num > 200, then sick

If gender = male,
   if cold = true and cough = true, then sick
```

Predictions

Healthy
Sick
Sick
.
.
Healthy
Healthy
Sick



This model is using irrelevant features when predicting on female subpopulation. I should not trust its predictions for that group.

https://explainml-tutorial.github.io/

Why model understanding?

Utility:

- Debugging
- Detecting dataset biases
- Assessing the mode’s applicability in real world
- Realizing if and when to trust model predictions

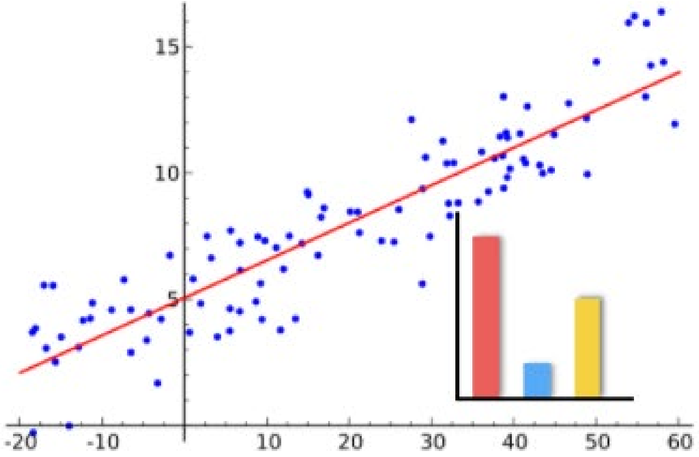
Primer on Explainability in Artificial Intelligence (AI)

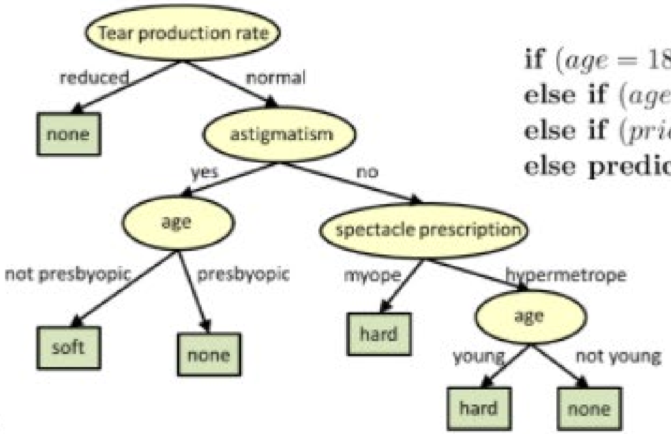
How to achieve model understanding?

Primer on Explainability in Artificial Intelligence (AI)

How to achieve model understanding?

Option 1: Build inherently interpretable models (e.g., decision trees, linear classifiers).



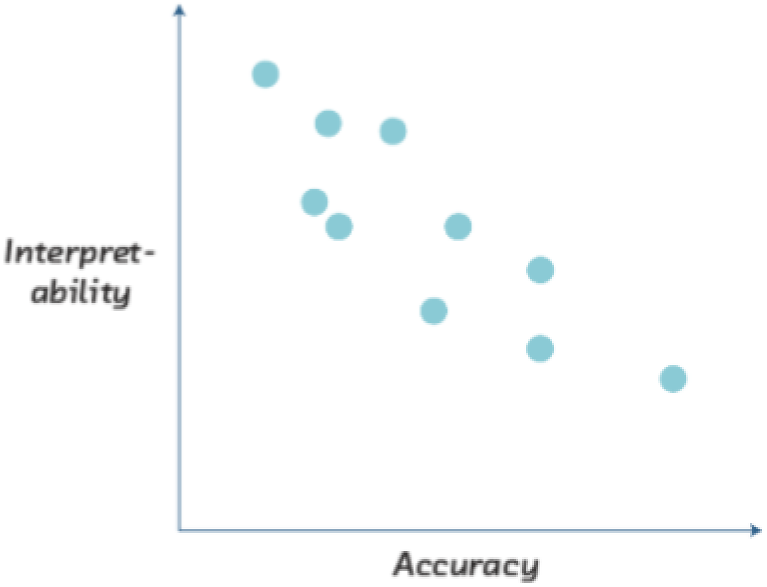


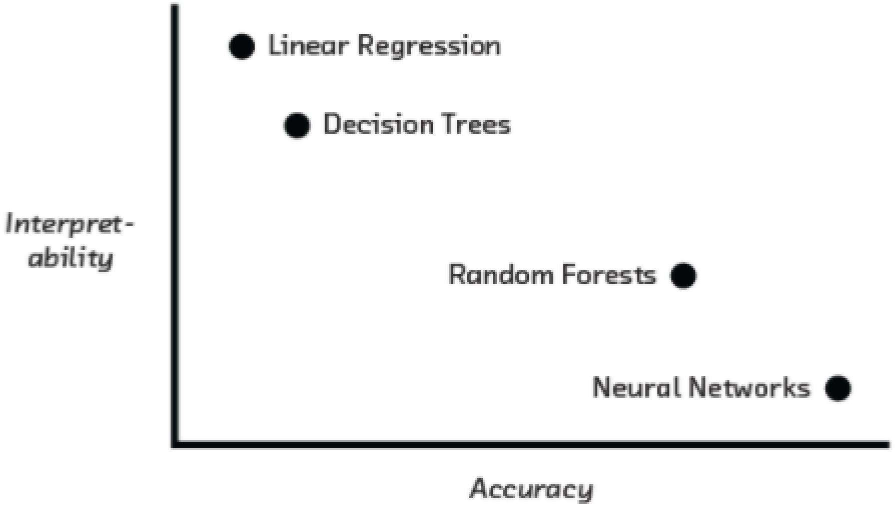
if $(age = 18 - 20)$ **and** $(sex = male)$ **then predict** *yes*
else if $(age = 21 - 23)$ **and** $(priors = 2 - 3)$ **then predict** *yes*
else if $(priors > 3)$ **then predict** *yes*
else predict *no*

https://explainml-tutorial.github.io/

Primer on Explainability in Artificial Intelligence (AI)

Drawbacks of the option 1:
A trade-off may exist between interpretability and prediction accuracy.





https://explainml-tutorial.github.io/

Primer on Explainability in Artificial Intelligence (AI)

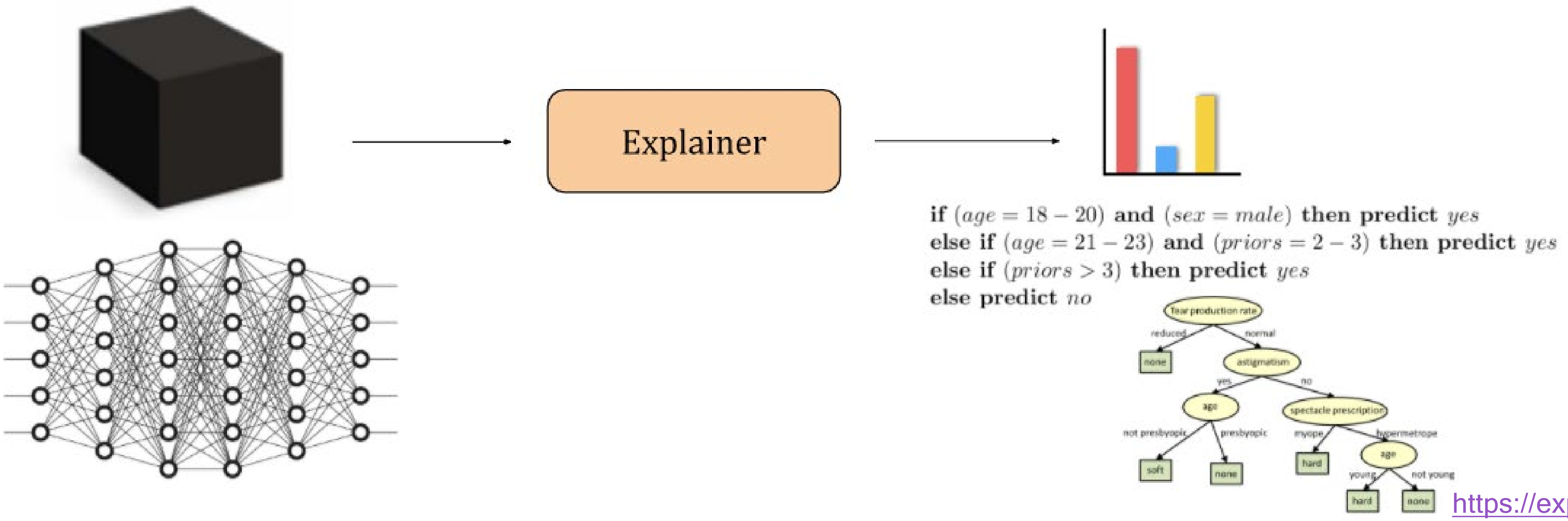
If you an build interpretable models that are adequately accurate, DO IT!

Otherwise, ***post-hoc explanations*** come to rescue!

Primer on Explainability in Artificial Intelligence (AI)

How to achieve model understanding?

Option 2: Explain cumbersome built models in a *"post-hoc"* manner.



if $(age = 18-20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21-23)$ **and** $(priors = 2-3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

https://explainml-tutorial.github.io/

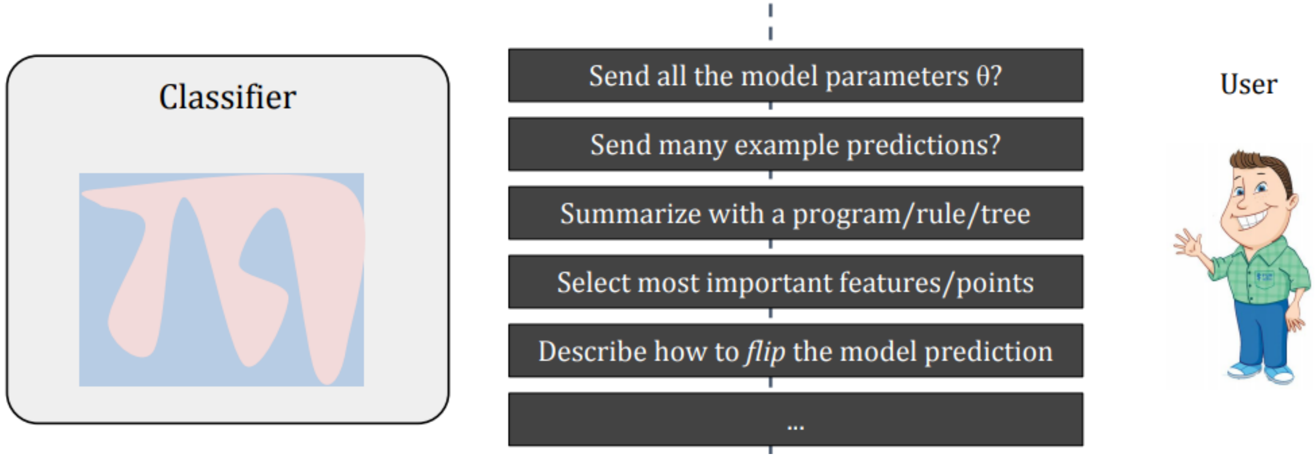
What is an explanation?

Definition:

- Interpretable description of the model's behavior

Properties:

- *Faithful*: from the model's end
- *Understandable*: from the user's end



Send all the model parameters θ?

Send many example predictions?

Summarize with a program/rule/tree

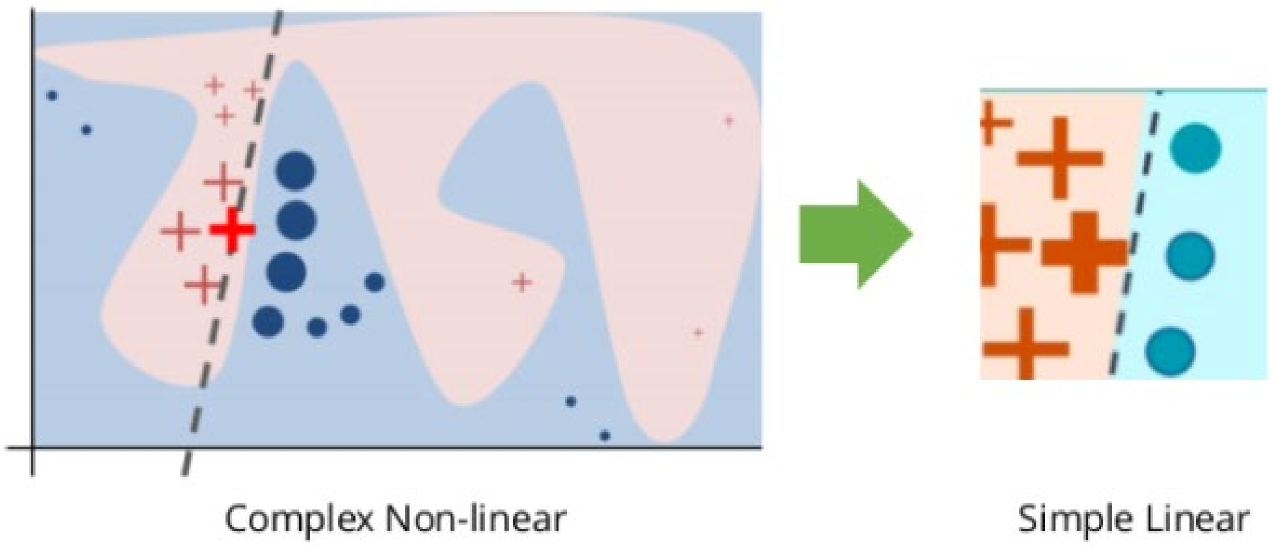
Select most important features/points

Describe how to *flip* the model prediction

...

https://explainml-tutorial.github.io/

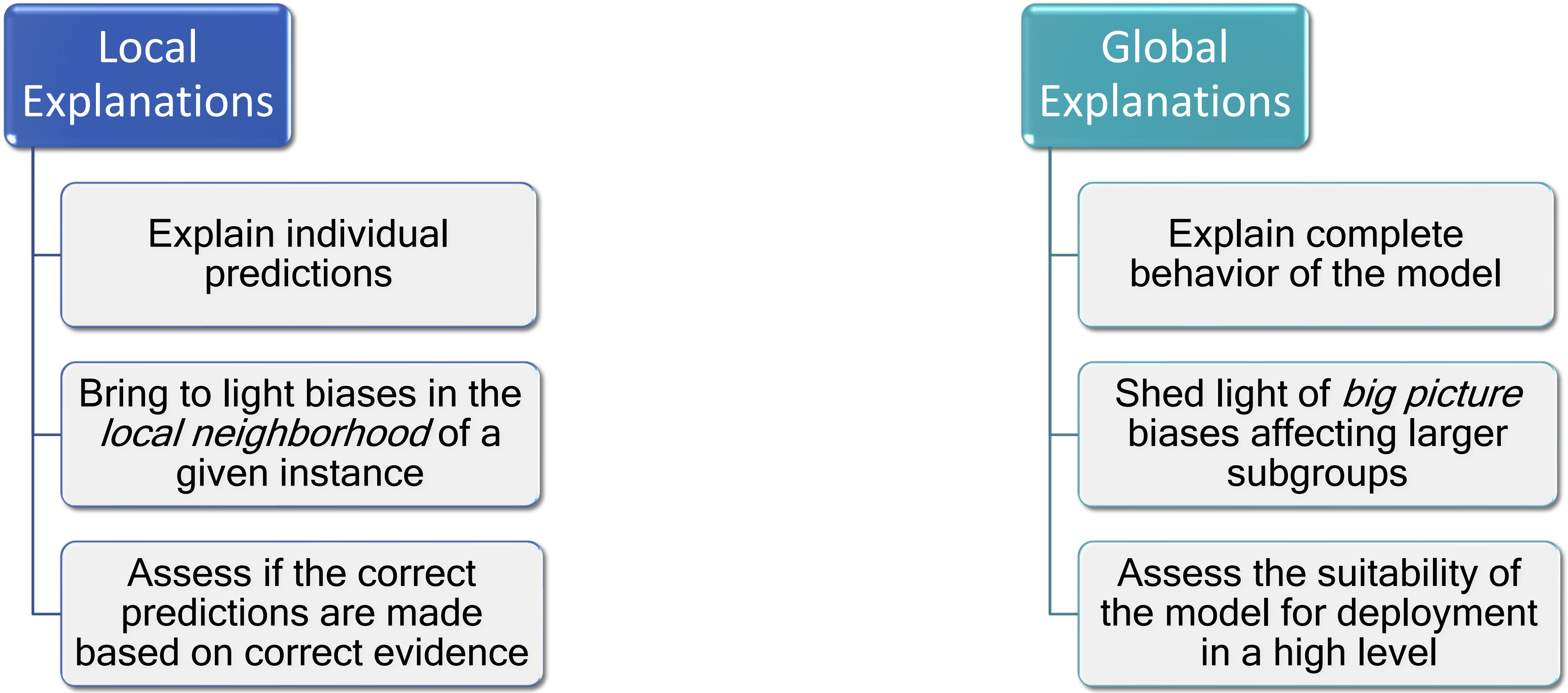
Local Explanation vs. Global Explanations



https://www.kdnuggets.com/2019/12/interpretability-part-3-lime-shap.html

Local Explanations: Interpretable description of the model's behavior *in a target neighborhood.*

Local Explanation vs. Global Explanations



Local Explanations

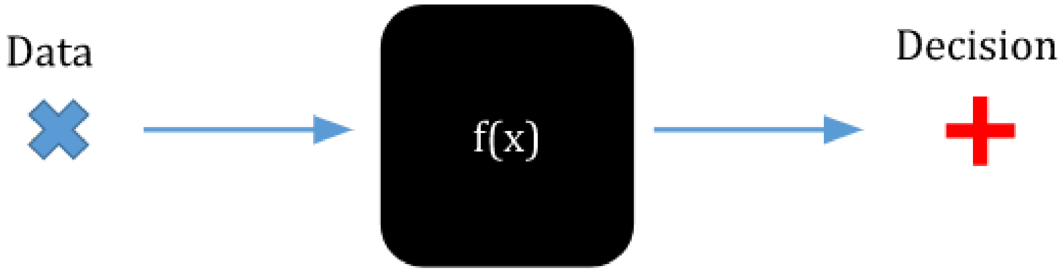
- Explain individual predictions
- Bring to light biases in the *local neighborhood* of a given instance
- Assess if the correct predictions are made based on correct evidence

Global Explanations

- Explain complete behavior of the model
- Shed light of *big picture* biases affecting larger subgroups
- Assess the suitability of the model for deployment in a high level

Model-Agnostic approaches

No access to the internal structure…



Not restricted to specific models

Practically easy: not tied to PyTorch, Tflow, etc.

Study models that you don’t have access to!

LIME

Definition:

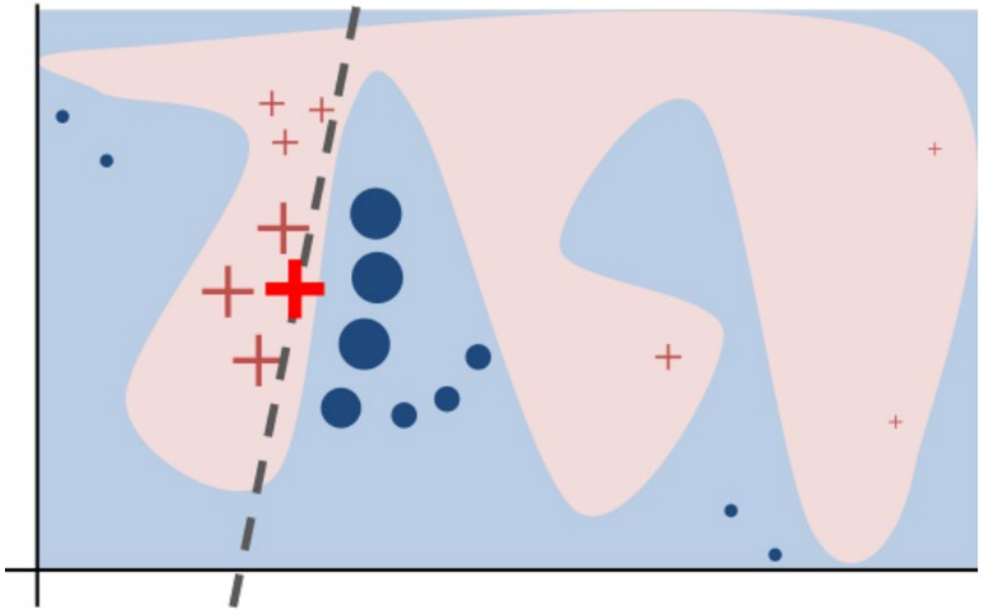
- Local Interpretable Model-Agnostic Explanations

Idea:

- Identifying the important dimensions and quantifying their relative importance.

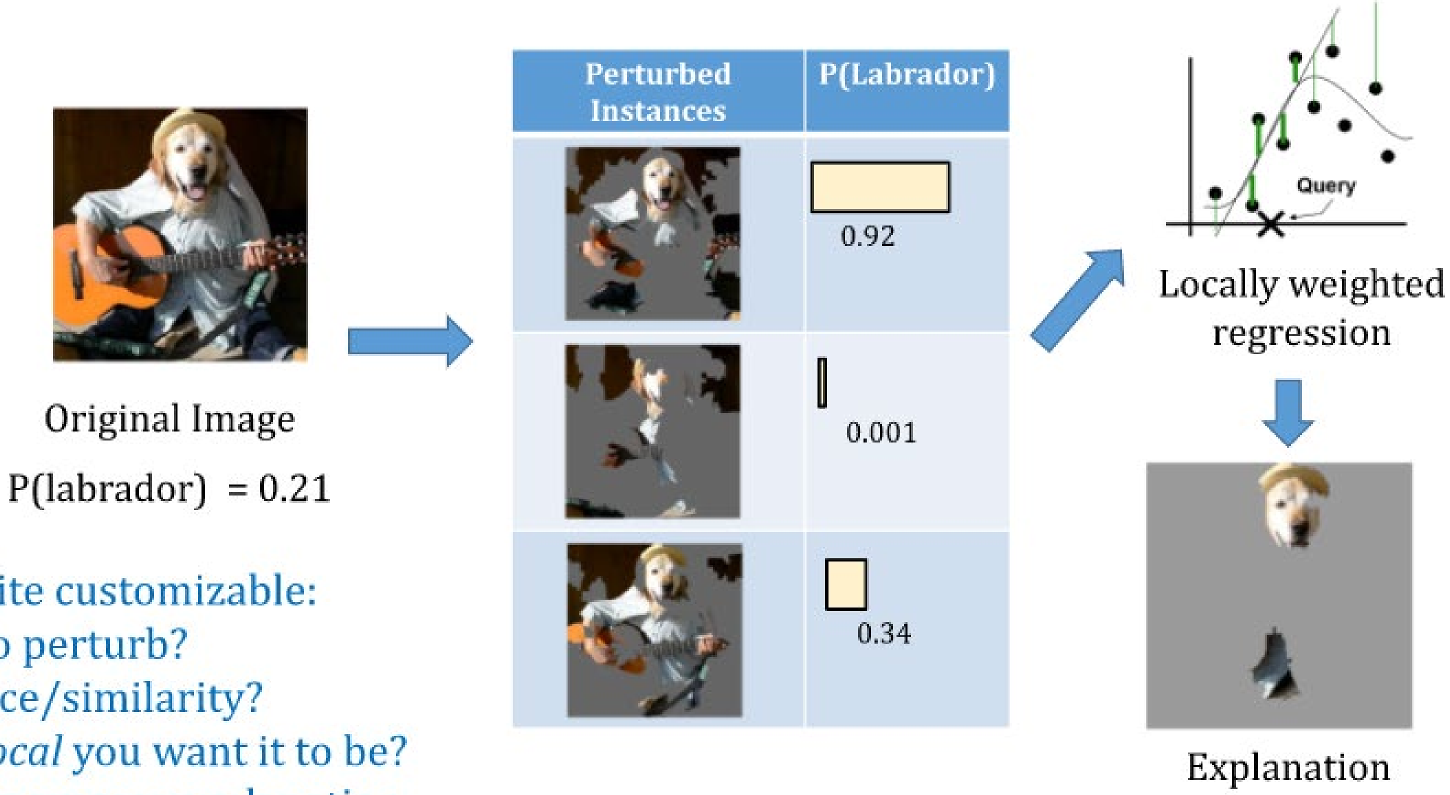
Utility:

- Initially proposed for tabular data, but also applicable to image/text data.



Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.

LIME



Perturbed Instances	P(Labrador)
	0.92
	0.001
	0.34

https://explainml-tutorial.github.io/

LIME is quite customizable:

- How to perturb?
- Distance/similarity?
- How *local* you want it to be?
- How to express explanation

SHAP

Definition:

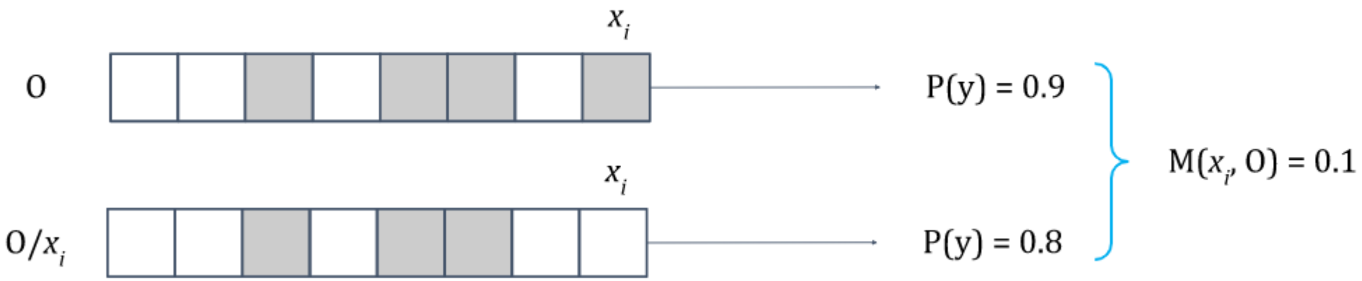
- Shapely Additive Explanations

Idea:

- Computing *marginal contribution* of each input feature towards the prediction, averaged over all possible permutations.

Utility:

- Initially proposed for tabular data, but also applicable to image/text data.



Lundberg, Scott M., and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* 30 (2017): 4765-4774.

SHAP

Definition:

- Shapely Additive Explanations

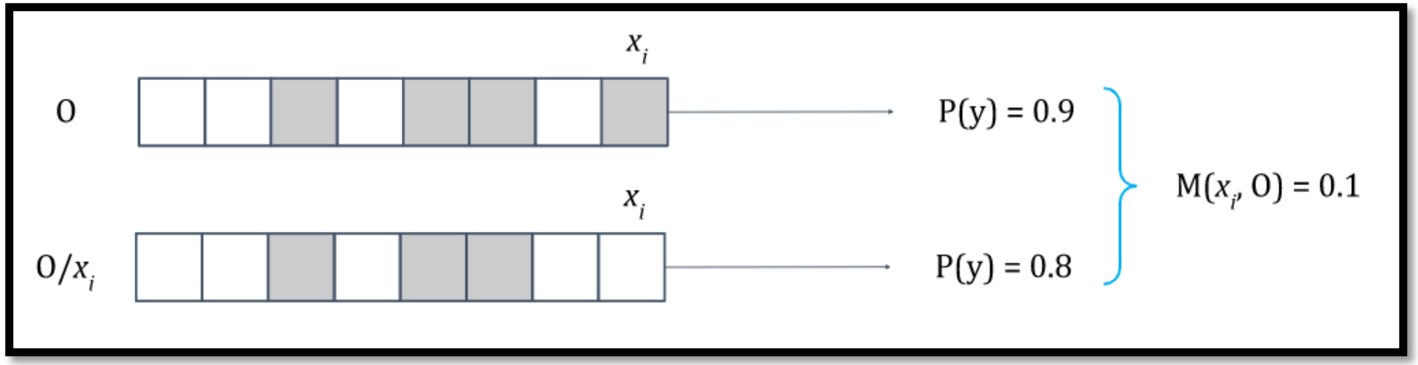
Idea:

- Computing *marginal contribution* of each input feature towards the prediction, averaged over all possible permutations.

Utility:

- Initially proposed for tabular data, but also applicable to image/text data.

$$g(x_i) = \mathbb{E}_O\,\{M(x_i, O)\}$$

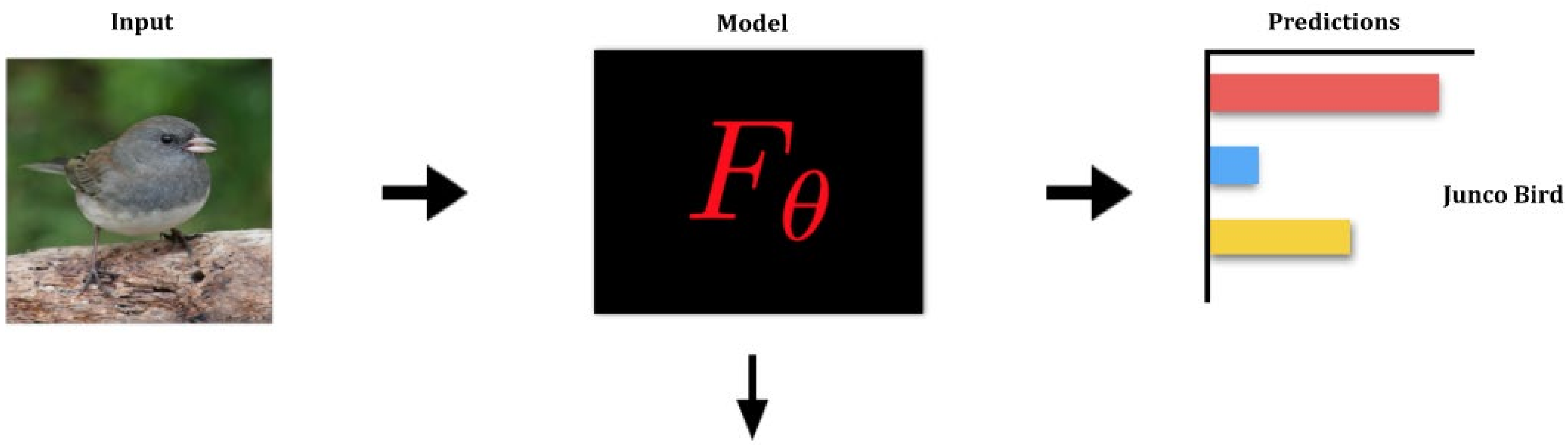


$g(x_i)$ :: Importance of the feature x_i

$M(x_i, O)$ :: Marginal contribution of x_i over the permutation set O

Backpropagation-based approaches

Assumption: we are explaining a *differentiable* model.

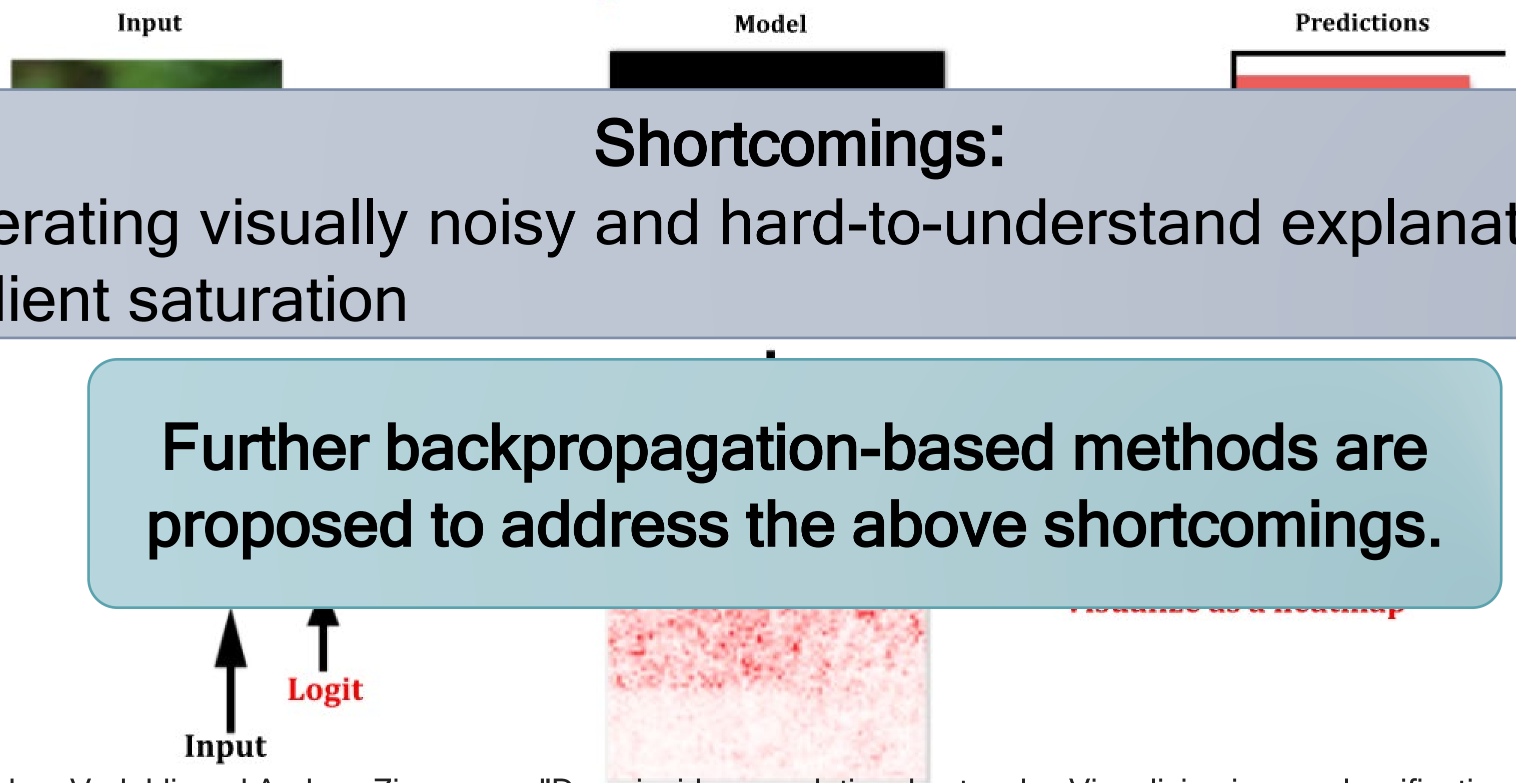


$$F : \mathbb{R}^d \rightarrow \mathbb{R}^c$$ **Model**

$$F_i : \mathbb{R}^d \rightarrow \mathbb{R}$$ **class specific logit**

Backpropagation-based approaches

Simplest approach: Vanilla Gradient



Shortcomings:

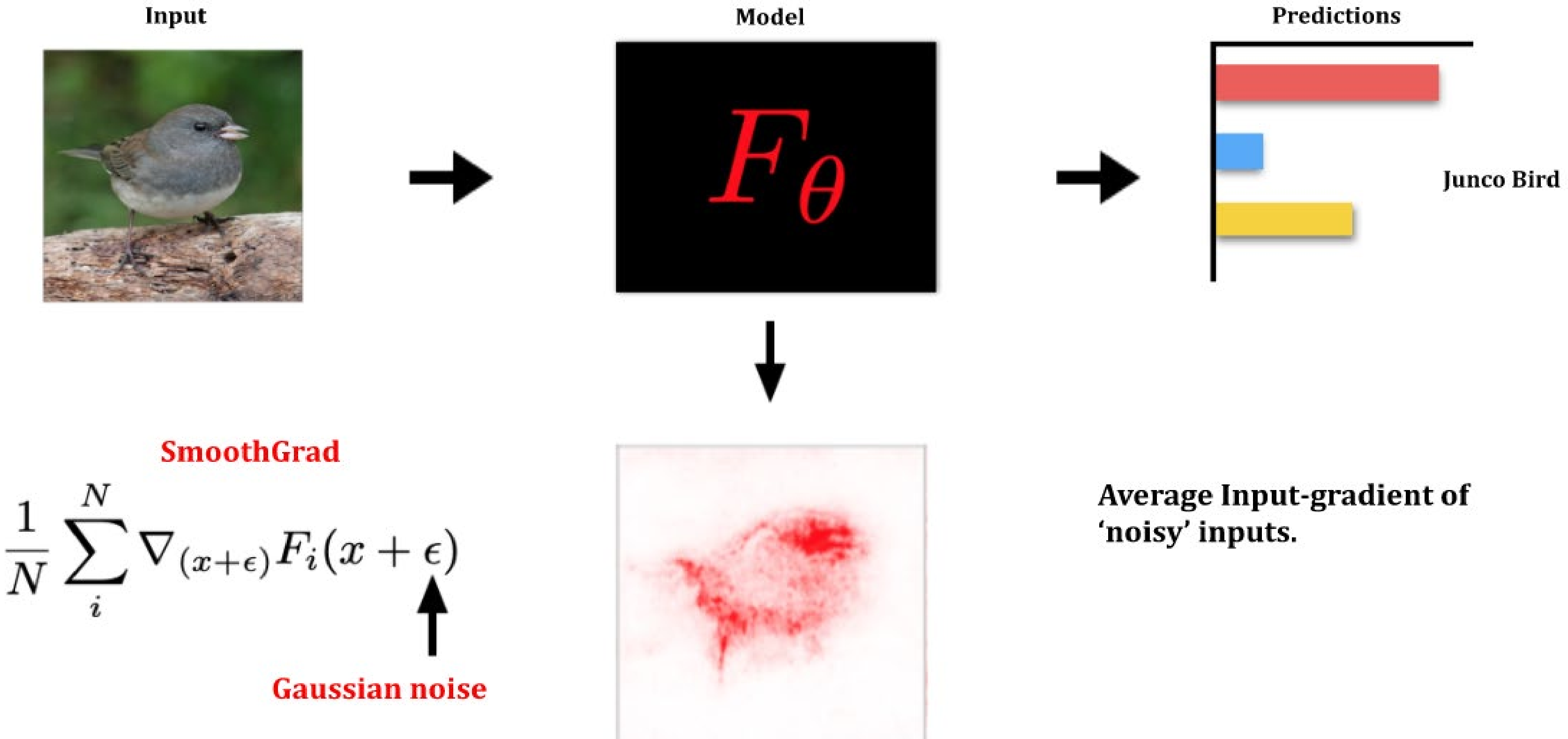
- ○ Generating visually noisy and hard-to-understand explanation maps
- ○ Gradient saturation

Further backpropagation-based methods are proposed to address the above shortcomings.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps

Backpropagation-based approaches

SmoothGrad



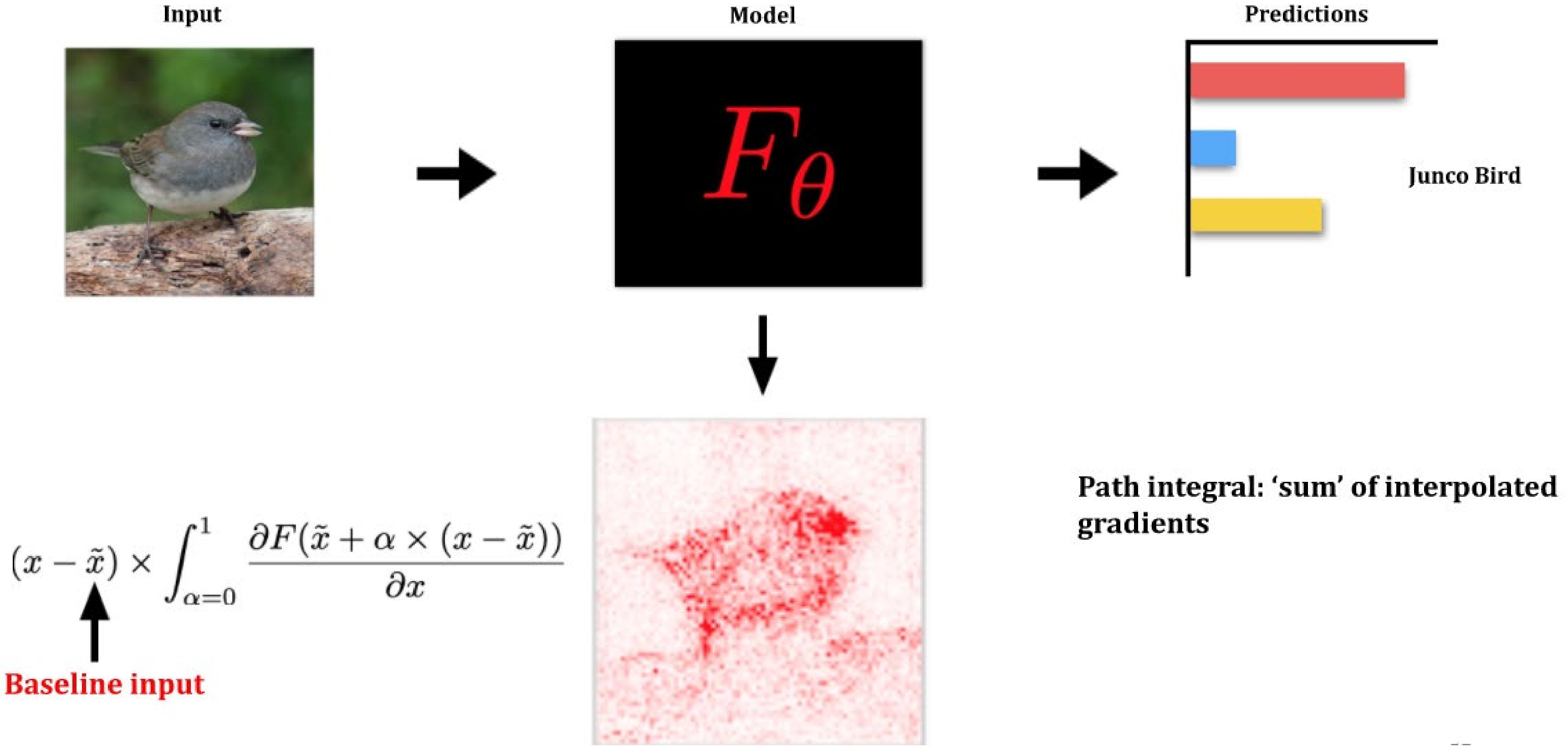
$$\frac{1}{N}\sum_{i}^{N} \nabla_{(x+\epsilon)} F_i(x+\epsilon)$$

Average Input-gradient of 'noisy' inputs.

Smilkov, Daniel, et al. "Smoothgrad: removing noise by adding noise." *arXiv preprint arXiv:1706.03825* (2017).

Backpropagation-based approaches

Integrated Gradients



$$(x - \tilde{x}) \times \int_{\alpha=0}^{1} \frac{\partial F(\tilde{x} + \alpha \times (x - \tilde{x}))}{\partial x}$$

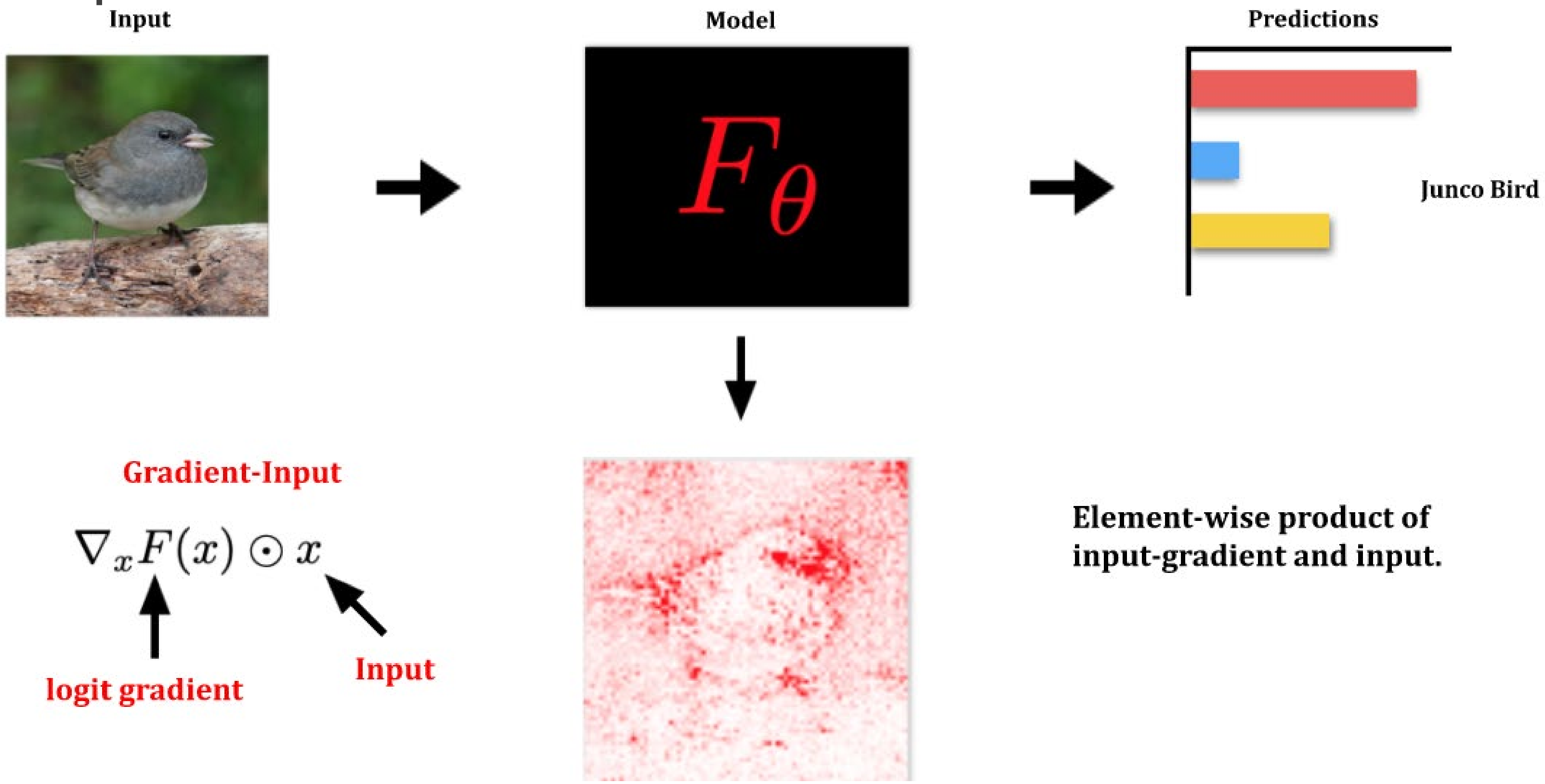
Baseline input

Path integral: 'sum' of interpolated gradients

Sundararajan, M., Taly, A., & Yan, Q. (2017, July). Axiomatic attribution for deep networks. In *International Conference on Machine Learning* (pp. 3319-3328). PMLR.

Backpropagation-based approaches

Gradient ⊙ Input



Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje. "Learning important features through propagating activation differences." *International Conference on Machine Learning*. PMLR, 2017.

: Class Activation Mapping

What CAM stands for?

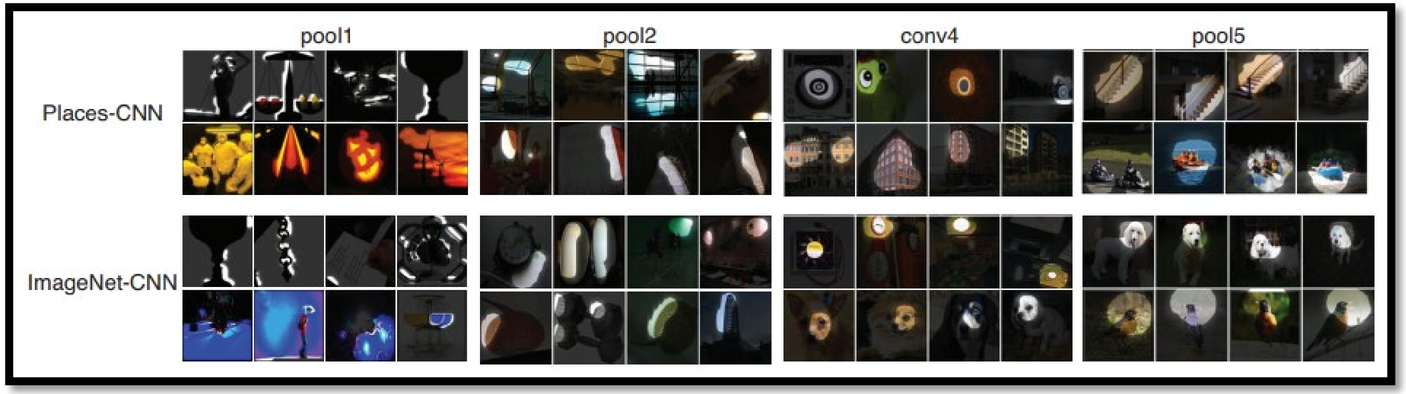
- Class Activation Mapping

Intuition:

- The convolutional units of various layers of convolutional neural networks (CNNs) actually behave as object detectors despite no supervision on the location of the object was provided (Zhou et al. 2014)

Utility:

- Model-specific explanation method
- Specialized for convolutional neural networks (CNNs).



Zhou, Bolei, et al. "Object Detectors Emerge in Deep Scene CNNs." *ICLR*. 2015.

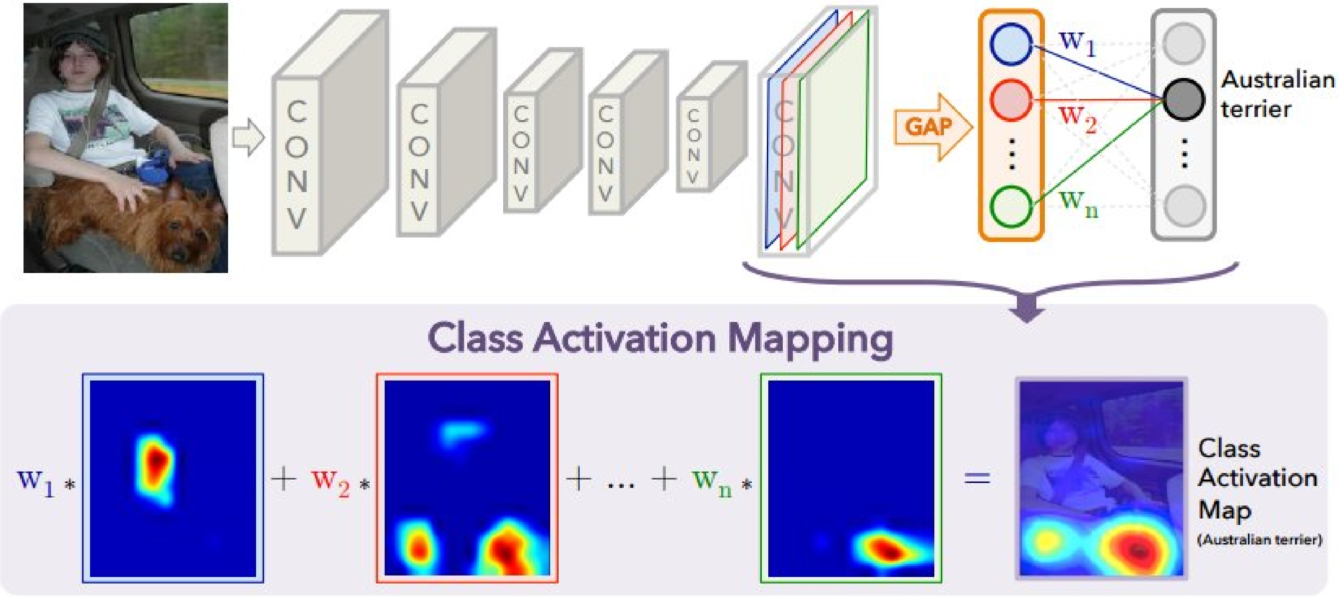
CAM-based methods

Aim

- Visualizing the features extracted by the CNN.
- NOT explanation.

Methodology

- Replacing the classifier units of the target model with a Global Average Pooling (GAP) layer, followed by an output layer.
- Training the model.
- Linear combination of the feature maps in the last convolutional layer, using the weights in the output layer.

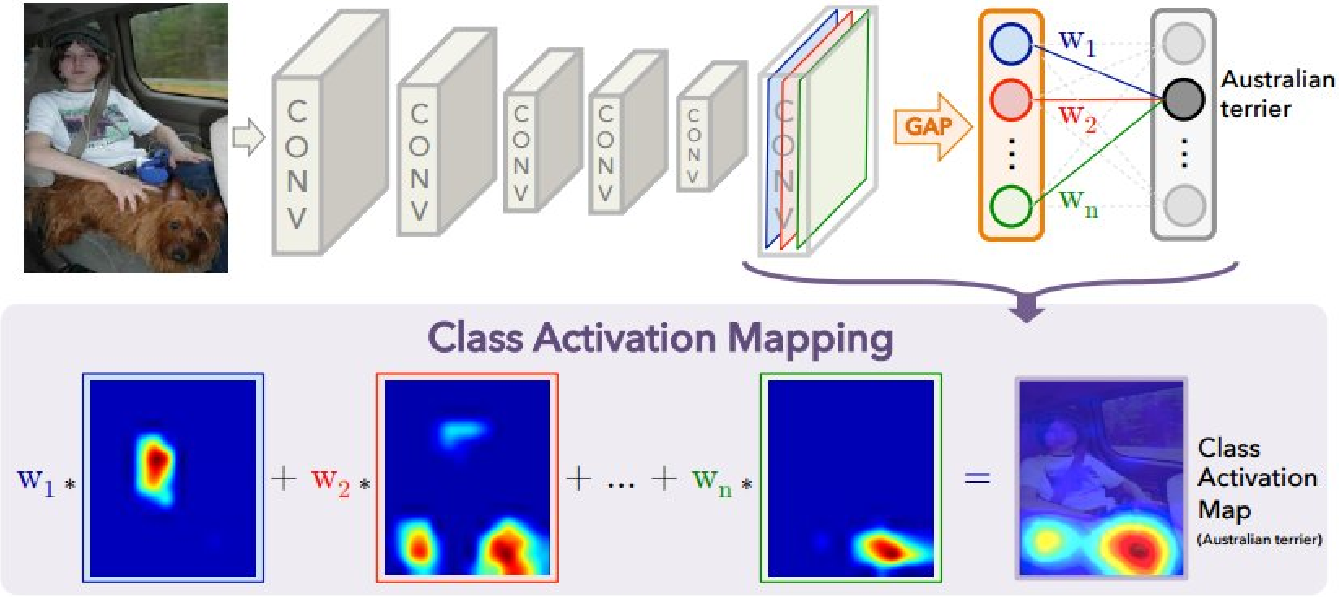


Zhou, Bolei, et al. "Learning deep features for discriminative localization." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

CAM-based methods

Limitations

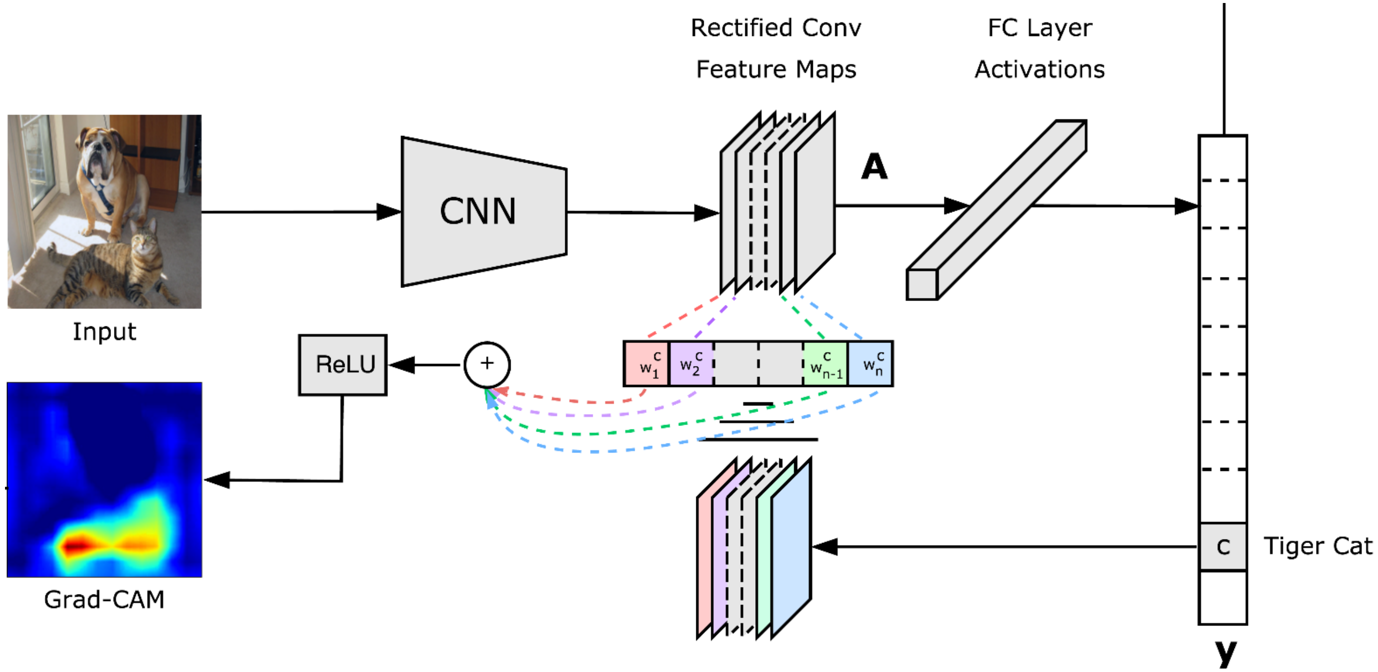
- Fails to visualize CNNs with complicated classifiers.
- Inapplicable for CNN explanation.
- Performance degradation because of the alteration in the CNN's classifier units.
- Inapplicable to the models trained for more complex image recognition tasks.



CAM-based methods

Grad-CAM

- Addresses the mentioned limitations in CAM.
- Uses "average gradient" values to score the feature maps in the last convolutional layer.
- Backpropagates the signal from the output, to calculate the average gradient values
- Runs in a forward pass and a backward pass.

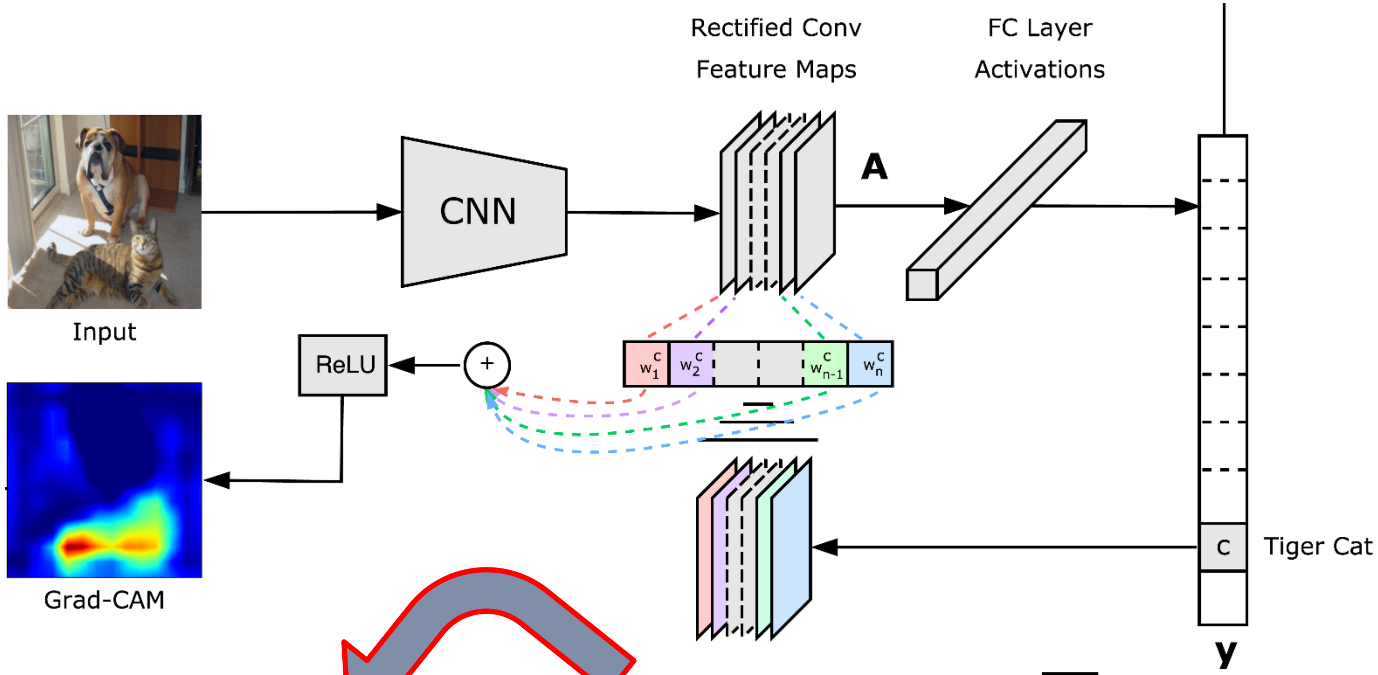


Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

Grad-CAM

Grad-CAM

- ○ Addresses the mentioned limitations in CAM.
- ○ Uses “average gradient” values to score the feature maps in the last convolutional layer.
- ○ Backpropagates the signal from the output, to calculate the average gradient values
- ○ Runs in a forward pass and a backward pass.



Model’s prediction for class *c*

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y_c}{\partial A_{ij}^k}$$

Feature map values

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right)$$

Grad-CAM explanation map

Perturbation-based methods: RISE

Definition:

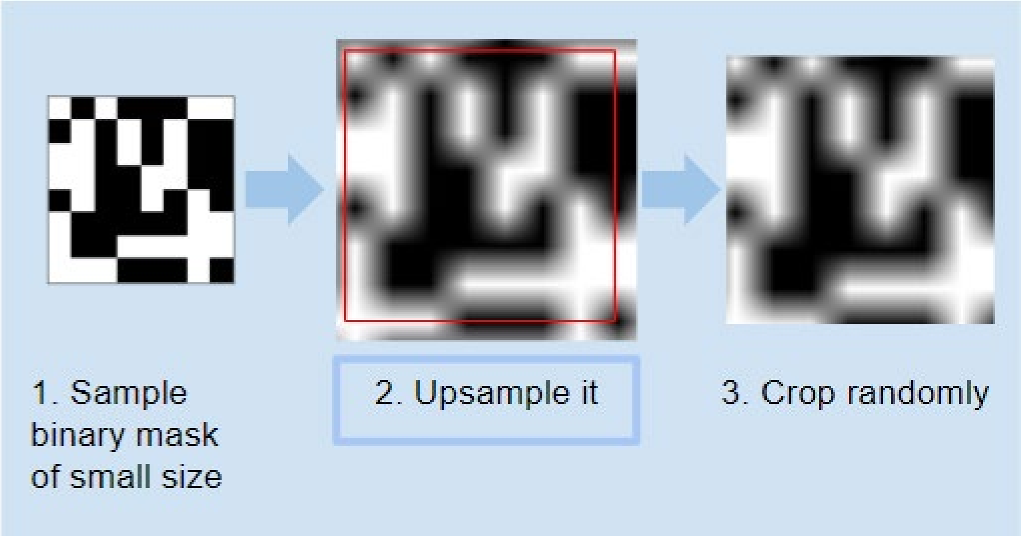
- Randomized Input Sampling for Explanation

Idea:

- Feeding the target model with the perturbed copies of the input image.
- Perturbation is performed using a set of smooth *random masks*. (why smooth?)

Utility:

- Initially proposed for image data, but also applicable to tabular /text data.
- Model-agnostic explanation method.

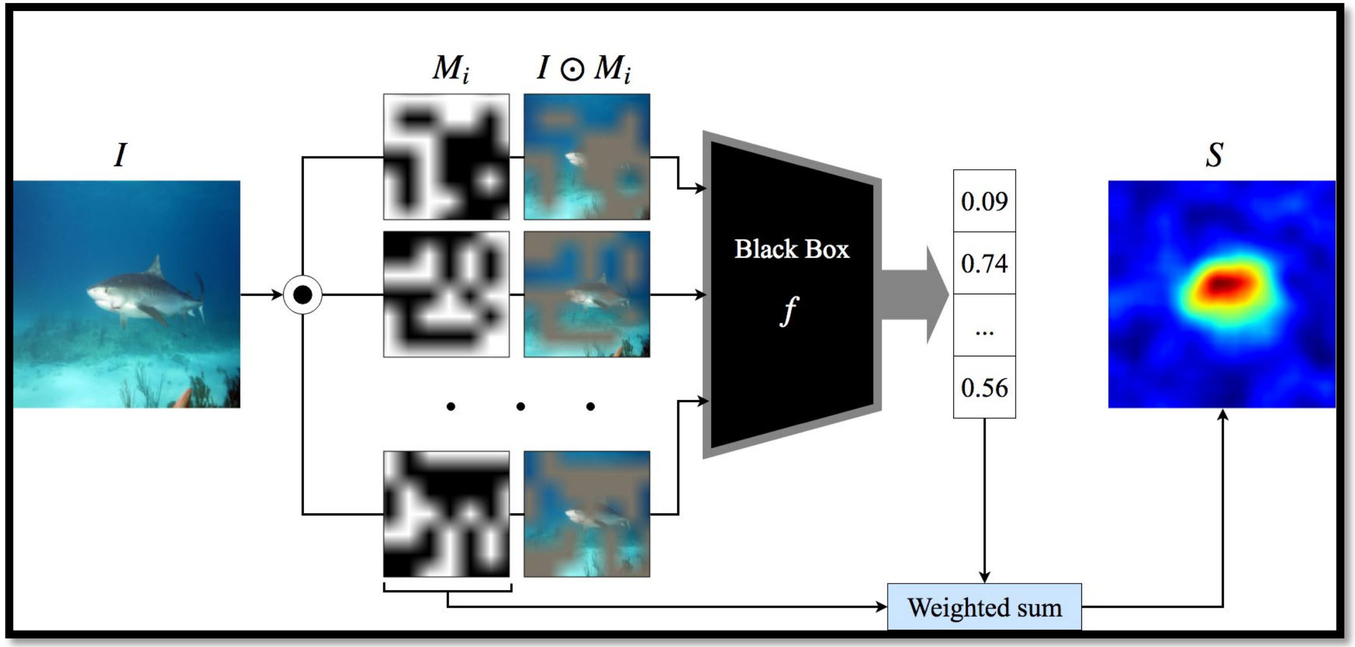


https://cs-people.bu.edu/vpetsiuk/rise/#

Perturbation-based methods: RISE

Methodology:

- ○ Generating a set of random masks
- ○ Masking the input with the random masks
- ○ Feeding the model with the masked images
- ○ Weighted linear combination of the masks.



Petsiuk, Vitali, Abir Das, and Kate Saenko. "Rise: Randomized input sampling for explanation of black-box models." *arXiv preprint arXiv:1806.07421* (2018).

Perturbation-based methods: RISE

Pros:

○ Applicability of the method to the AI models beyond the family of CNNs.

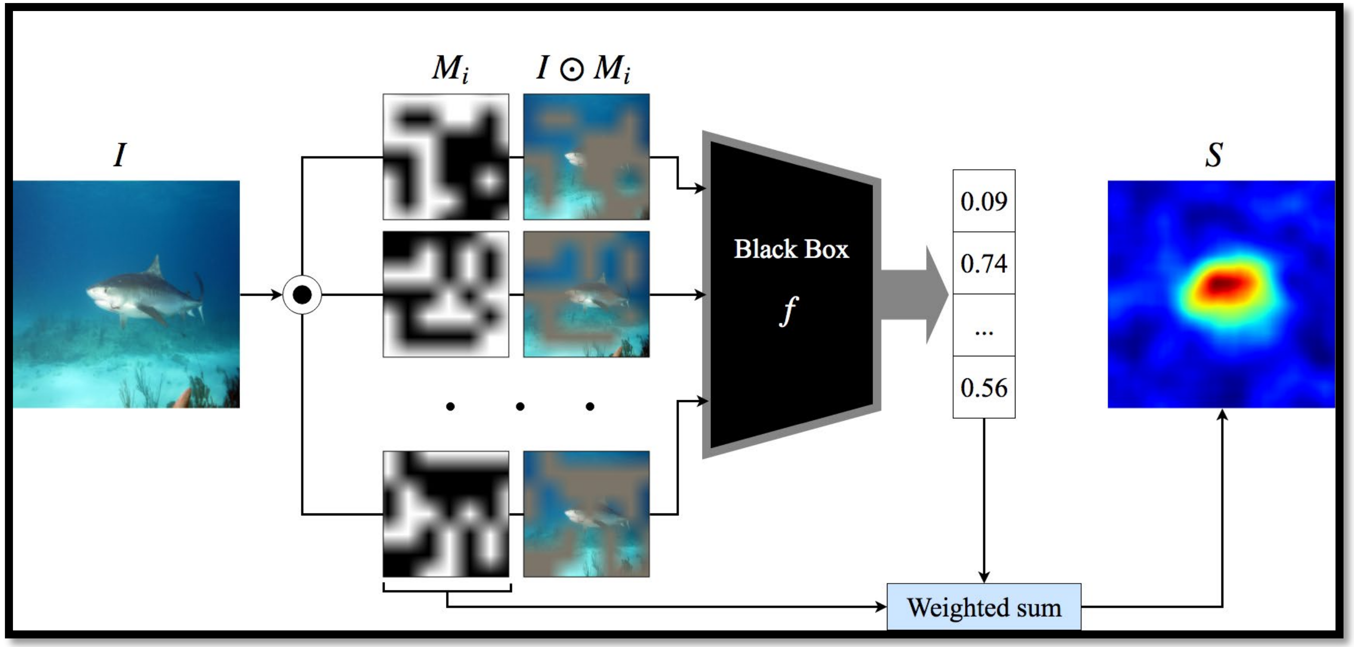
○ Shows the superior preciseness of perturbation rather than backpropagation, in forming explanation map.

Cons:

○ Low visual quality of RISE explanation maps.

○ Increase of failure chance, while dealing with small object instances

○ Slow runtime, as it passes numerous (4000-8000) masked images through a model.



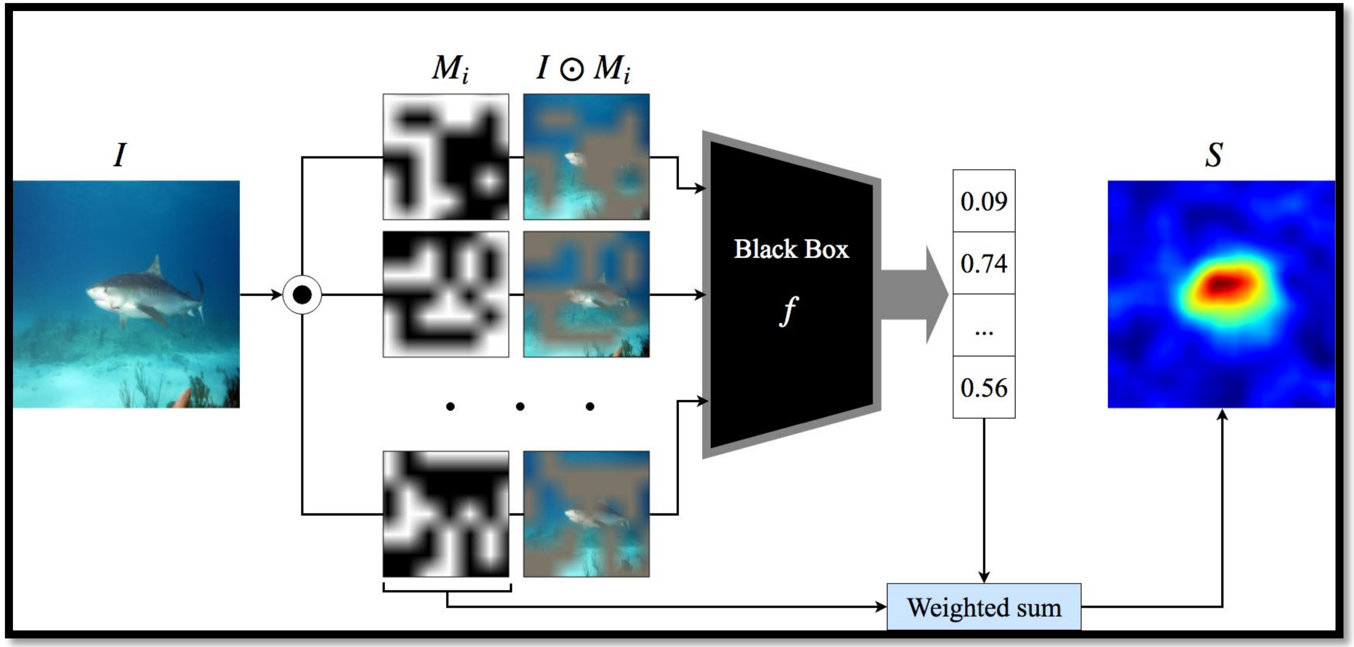
Perturbation-based methods: RISE

Pros:

- Applicability of the method to the AI models beyond the family of CNNs.
- Shows the superior preciseness of perturbation rather than backpropagation, in forming explanation map.

Cons:

- Low visual quality of RISE explanation maps.
- Increase of failure chance, while dealing with small object instances
- Slow runtime, as it passes numerous (4000-8000) masked images through a model.



Perturbation-based methods: SISE

Definition:

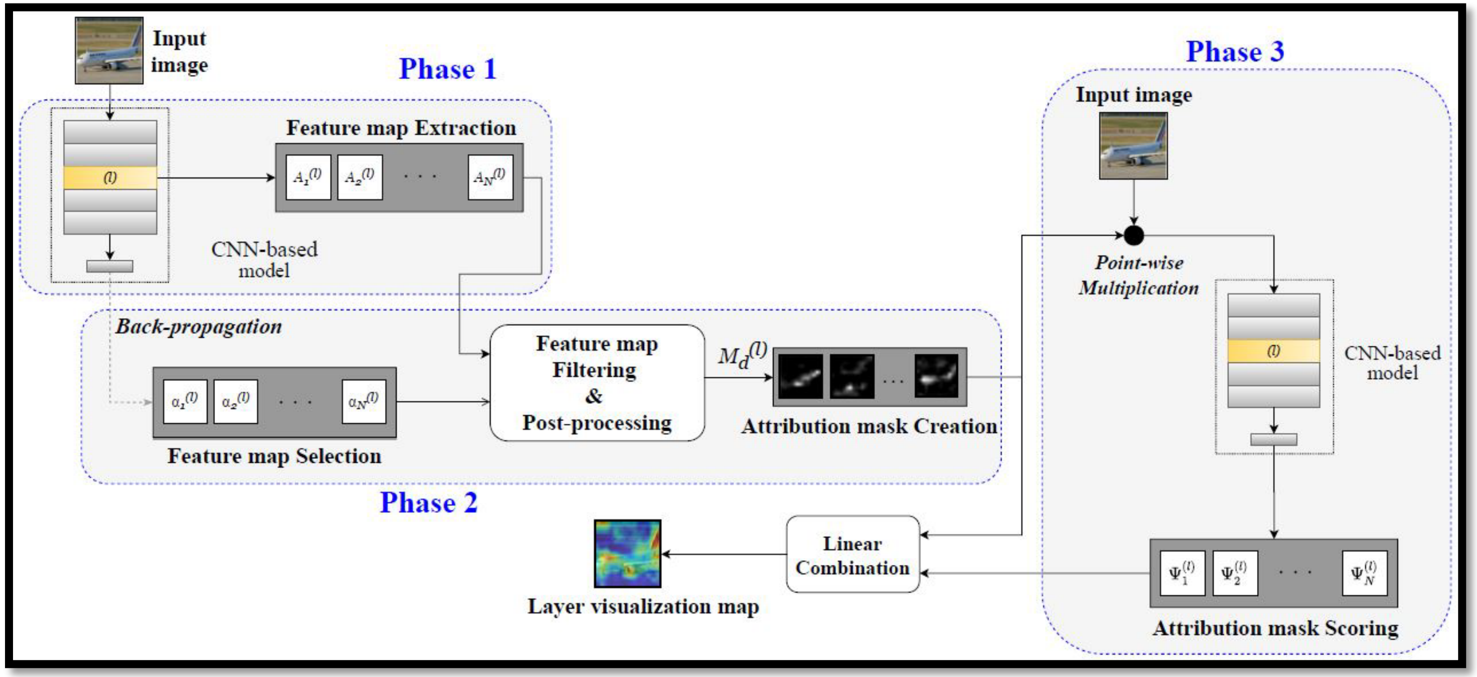
- Semantic Input Sampling for Explanation

Idea:

- Feeding the target model with the perturbed copies of the input image.
- Random masks in RISE are replaced with *attribution masks.*

Utility:

- Model-specific explanation method
- Specialized for convolutional neural networks (CNNs).

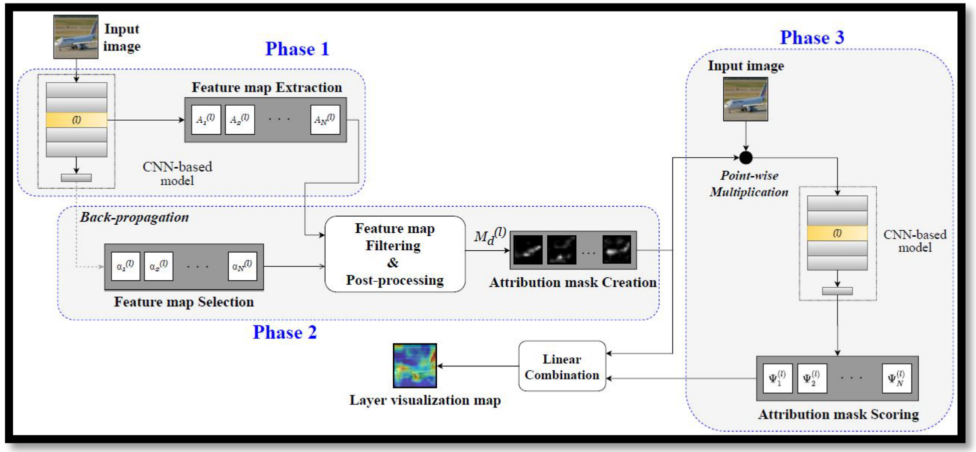


Sattarzadeh, Sam, et al. "Explaining Convolutional Neural Networks through Attribution-Based Input Sampling and Block-Wise Feature Aggregation." *arXiv preprint arXiv:2010.00672* (2020).

Perturbation-based methods: SISE

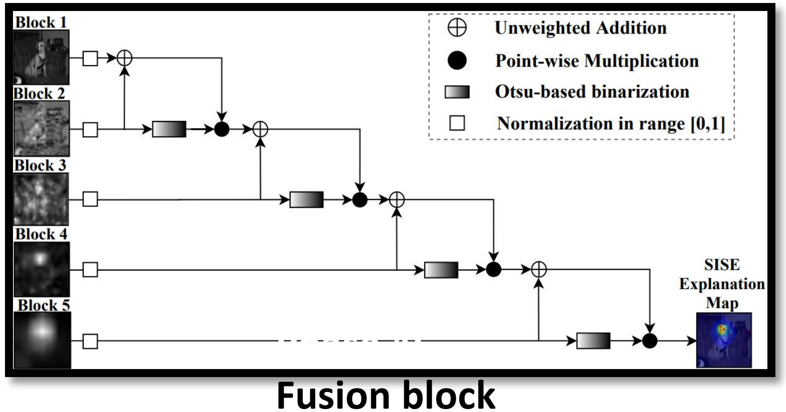
Methodology: Consists four consecutive phases

- ○ Feature map extraction
- ○ Feature map selection
- ○ Attribution mask scoring
- ○ Feature aggregation



Input image
Phase 1
Feature map Extraction
$A_1^{(l)}$ $A_2^{(l)}$ · · · $A_N^{(l)}$
(l)
CNN-based model
Back-propagation
$\alpha_1^{(l)}$ $\alpha_2^{(l)}$ · · · $\alpha_N^{(l)}$
Feature map Selection
Feature map Filtering & Post-processing
$M_d^{(l)}$
Attribution mask Creation
Phase 2
Phase 3
Input image
Point-wise Multiplication
CNN-based model
Linear Combination
Layer visualization map
$\Psi_1^{(l)}$ $\Psi_2^{(l)}$ · · · $\Psi_N^{(l)}$
Attribution mask Scoring

- The first phases are applied on the last layers in all convolutional blocks (why?). Corresponding to each layer, the third phase outputs a 2-dimensional map called ***visualization map***.
- The visualization maps are **aggregated** in the last phase to form the desires explanation map.



Block 1
Block 2
Block 3
Block 4
Block 5
⊕ Unweighted Addition
● Point-wise Multiplication
▬ Otsu-based binarization
□ Normalization in range [0,1]
SISE Explanation Map

Fusion block

Outline

A. Tutorial on visual explainable AI (XAI)
   - Motivation
   - Primer on explainability in Artificial Intelligence (AI)
   - Approaches for visual explanation generation

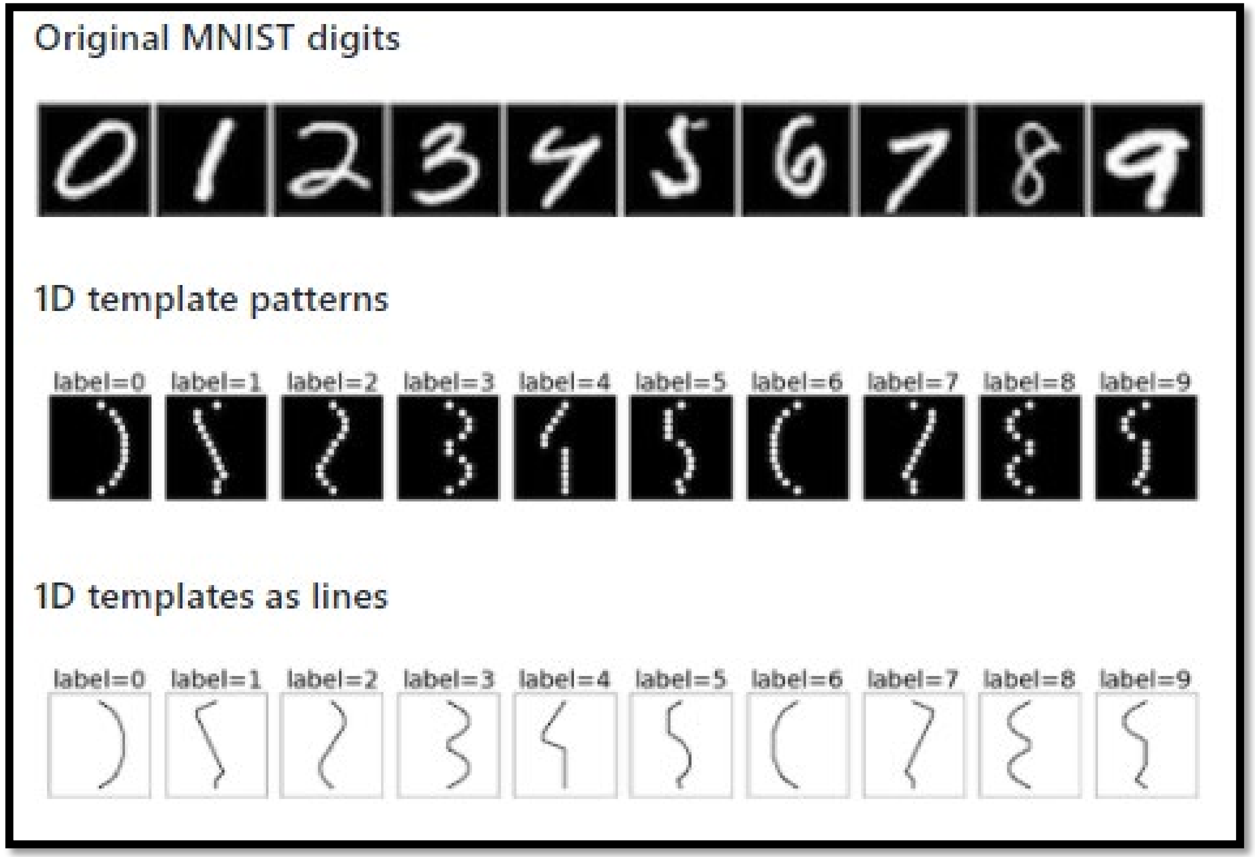
B. Project “A” Description
   - Project Goal
   - Datasets and Models
   - Evaluation Metrics

C. Your Questions!

Datasets: MNIST-1D

- A 1-dimensional analogue for the famous **MNIST** digit classification dataset.
- Small-scale, low-memory, synthetic
- 4000 train data + 1000 test data (partitioned by the dataset promoters).
- Each data is generating by applying random transformations to one of the 1-D templates.

- Transformations include:
  - Random padding
  - Random translation
  - Random 1-D shearing
  - Gaussian noise addition.



https://greydanus.github.io/2020/12/01/scaling-down/

Greydanus, Sam. "Scaling* down* Deep Learning." *arXiv preprint arXiv:2011.14439* (2020).

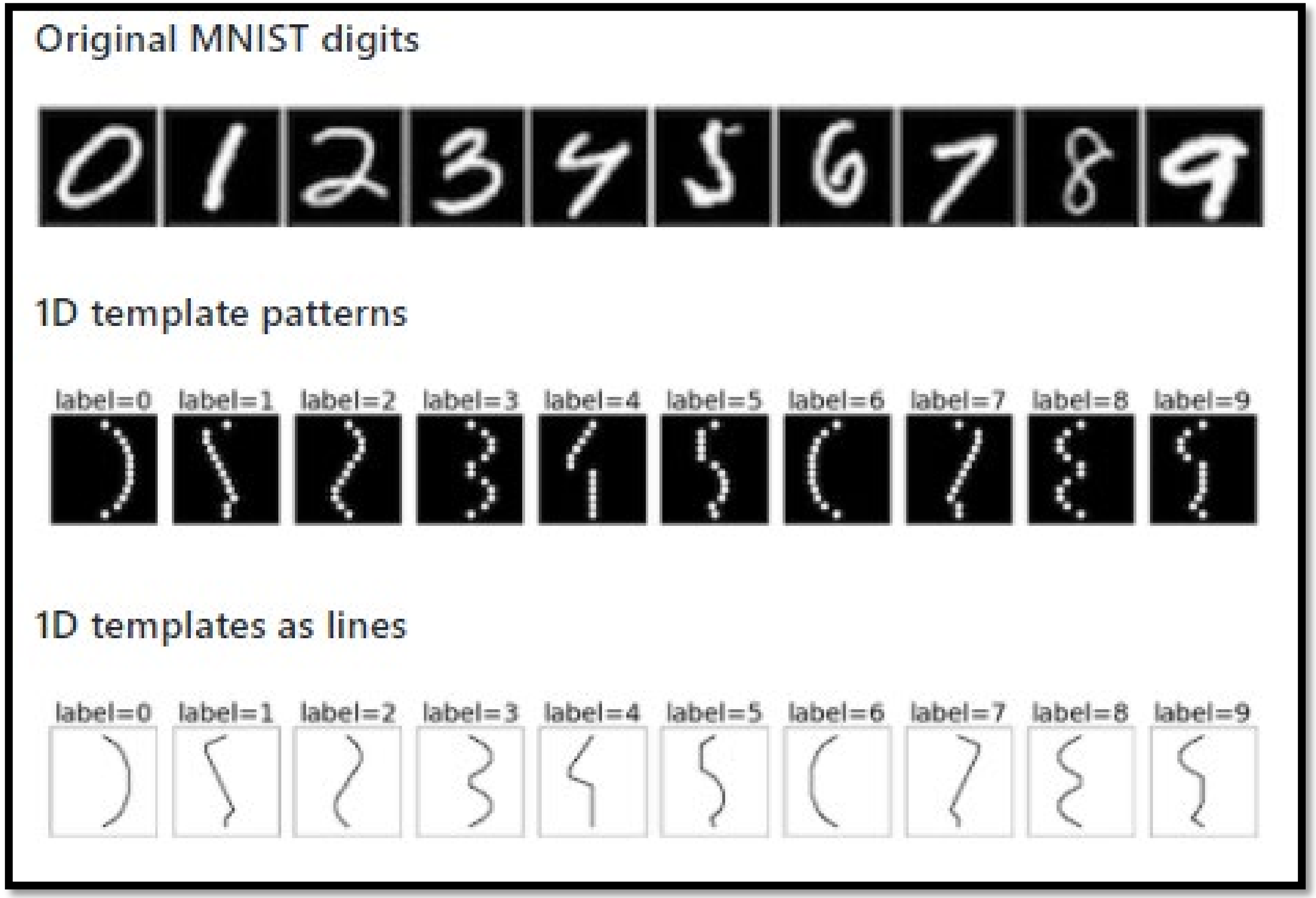
Project Goal

- Inspect the ability of one/two explanation methods(s) in providing correct explanations for CNNs in two different scenarios.
- **Specifications:**
  - Datasets:
    - MNIST-1D: Generic digit pattern classification
    - HMT: Histopathological tissue classification
  - Models:
    - A shallow CNN trained on the MNIST-1D dataset.
    - A VGG-7 network trained on the HMT dataset.
  - Evaluation metrics:
    - Drop%
    - Increase%

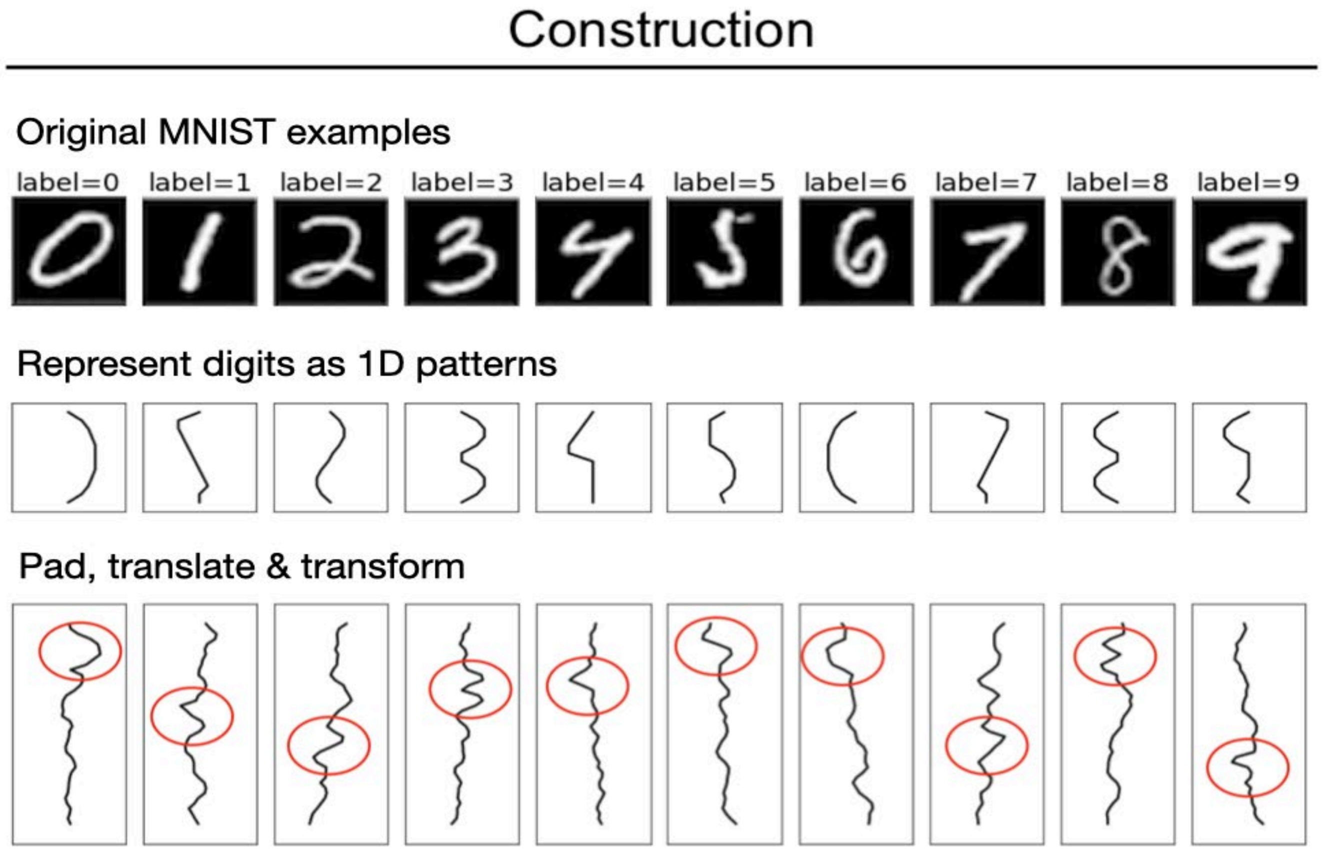
Datasets: MNIST-1D

- A 1-dimensional analogue for the famous **MNIST** digit classification dataset.
- Small-scale, low-memory, synthetic
- 4000 train data + 1000 test data (partitioned by the dataset promoters).
- Each data is generating by applying random transformations to one of the 1-D templates.

- Transformations include:
  - Random padding
  - Random translation
  - Random 1-D shearing
  - Gaussian noise addition.

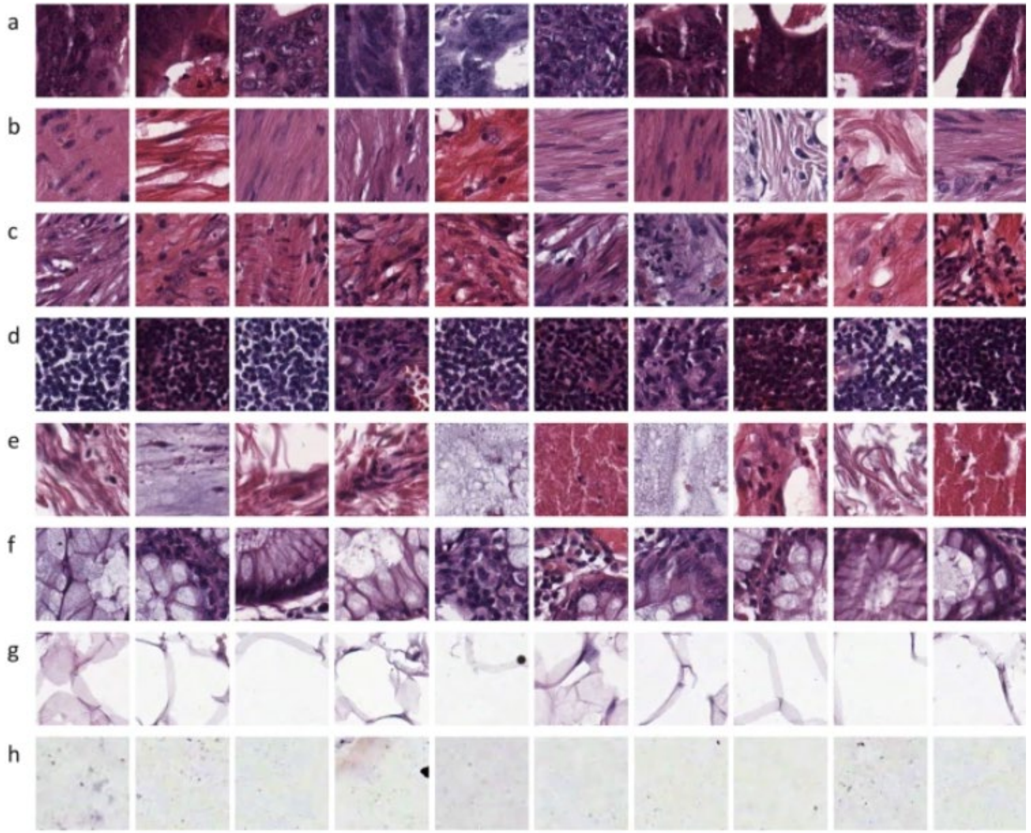


Datasets: MNIST-1D



Datasets: HMT

- Multi-class texture analysis in colorectal cancer histology.
- An equally-balanced dataset.
- classes:
  - (a) tumor epithelium,
  - (b) simple stroma,
  - (c) complex stroma,
  - (d) immune cell conglomerates,
  - (e) debris and mucus,
  - (f) mucosal glands,
  - (g) adipose tissue,
  - (h) background.
- 4504 train images + 496 test images.



https://www.nature.com/articles/srep27988

Greydanus, Sam. "Scaling* down* Deep Learning." *arXiv preprint arXiv:2011.14439* (2020).

Evaluation Metrics

Properties of a “good” explanation method:

- Faithfulness: correlataion of the generated explanation maps with the behavior of the model (the model’s end)
- Understandability: Clarity of the generated explanation maps (the user’s end)

Model truth-based metrics:

- Aim to evaluate the faithfulness of the explanation methods.

The focus of
Project A

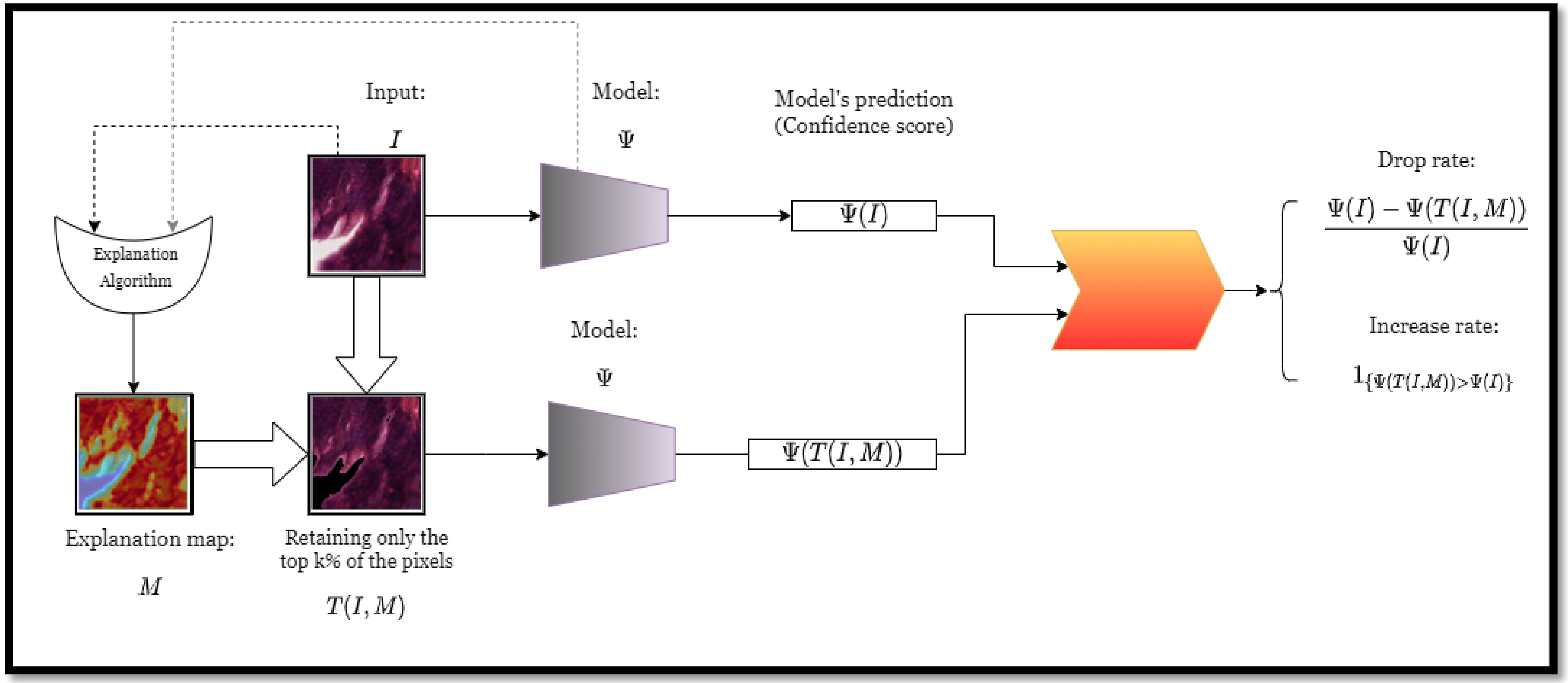
Evaluation Metrics

Drop%

- When the most important features are retained and the other features are removed, the model’s output should not “drop” significantly.

Increase

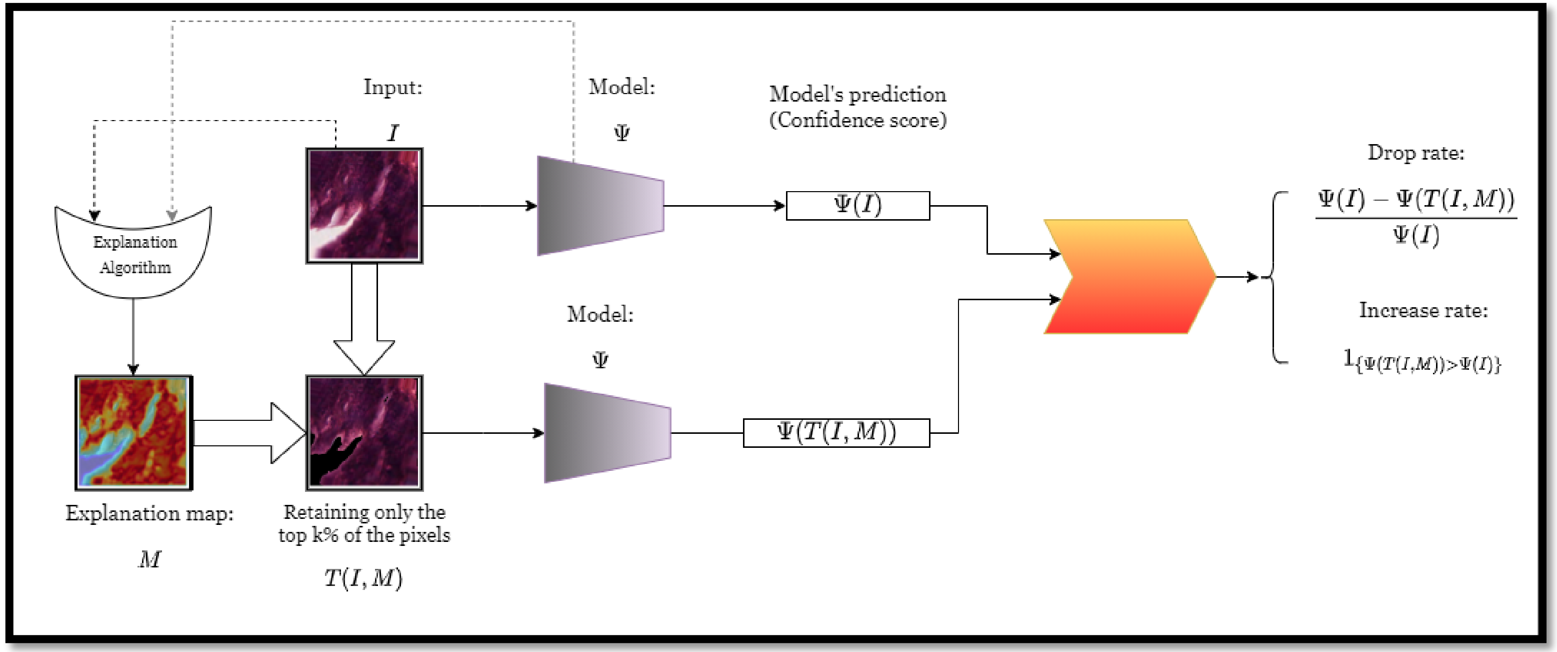
- When the most important features are retained and the other features are removed, the model’s output may increase in some cases



Evaluation Metrics

How Drop/Increase% evaluate your explanation algorithm?

- Retain the top-most *k%* pixels of the image, and set the other pixels to black.
- Pass the main image and the masked image to the model.
- Compare the model's output in the two cases.



Input:
I
Model:
Ψ
Model's prediction (Confidence score)
$\Psi(I)$
Explanation Algorithm
Model:
Ψ
$\Psi(T(I,M))$
Drop rate:
$\frac{\Psi(I)-\Psi(T(I,M))}{\Psi(I)}$
Increase rate:
$1_{\{\Psi(T(I,M))>\Psi(I)\}}$
Explanation map:
M
Retaining only the top k% of the pixels
$T(I,M)$

k= { The MNIST-1D dataset: 30%

The HMT dataset: 90%

Outline

A. Tutorial on visual explainable AI (XAI)
  - Motivation
  - Primer on explainability in Artificial Intelligence (AI)
  - Approaches for visual explanation generation

B. Project “A” Description
  - Project Goal
  - Datasets and Models
  - Evaluation Metrics

C. Your Questions!

14. ‘Th [illegible] (1 slide)

THANK YOU

Questions?