Visual Explainability in Machine Learning

Lecture 5: Evaluating Visual Explanations

Ghassan AlRegib, PhD
Professor

Mohit Prabhushankar, PhD
Postdoctoral Fellow

Omni Lab for Intelligent Visual Engineering and Science (OLIVES)
School of Electrical and Computer Engineering
Georgia Institute of Technology
{alregib, mohit.p}@gatech.edu

Dec 5, 2023

Short Course Materials

Accessible Online

SCAN ME

https://alregib.ece.gatech.edu/sps-education-short-course/

{alregib, mohit.p}@gatech.edu

IEEE Signal Processing Society

Title: Visual Explainability in Machine Learning

Presented by: *Ghassan AlRegib, and Mohit Prabhushankar*

Omni Lab for Intelligent Visual Engineering and Science (OLIVES)

School of Electrical and Computer Engineering

Georgia Institute of Technology, Atlanta, USA

https://alregib.ece.gatech.edu/

Short Course

Course Outline

Day 1: Define and Detail; Day 2: Evaluate; Day 3: Reassess

Day 1
- Lecture 1: Why Explainability? →
- Lecture 2: What is Explainability? →
- Lectures 3 and 4: Popular Methods of Explainability

Day 2
- Lecture 5: Evaluating Explainability →
- Lecture 6: Robustness: A proxy for Explainability

Day 3
- Lectures 7 and 8: Reassessing Explainability via Uncertainty and Concept vectors →
- Lectures 9 and 10: Causality and Generative AI

Outline

Lecture 5: Evaluating Visual Explanations

- Explanatory Evaluation Taxonomy
- Human Evaluation
  - Challenges
  - Methodology
- Application Evaluation
  - Methodology
  - Gaze Prediction
  - Pointing Game
  - Localization
- Network Evaluation
  - Intervention-based Evaluation
  - Masking
  - Progressive Pixel-wise masking
  - Progressive Structure-wise masking
- Challenges in Explanatory Evaluation
  - Human and Application Evaluation
  - Network Evaluation

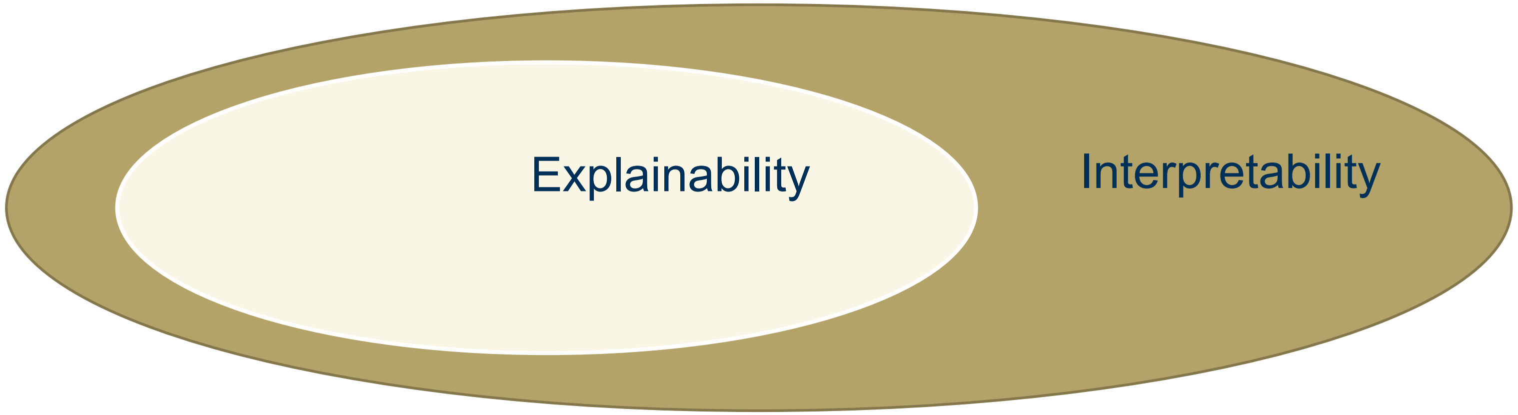
Explanations

What is Explainability?

The ability of an entity to explain or justify its decisions or predictions in human-understandable terms

Interpretability: Goal of Interpretability research is to understand the inner workings of the model

Explainability: Goal of Explainability research is to explain the network decisions **to humans**

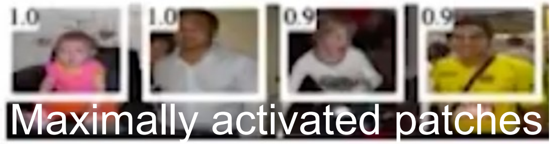


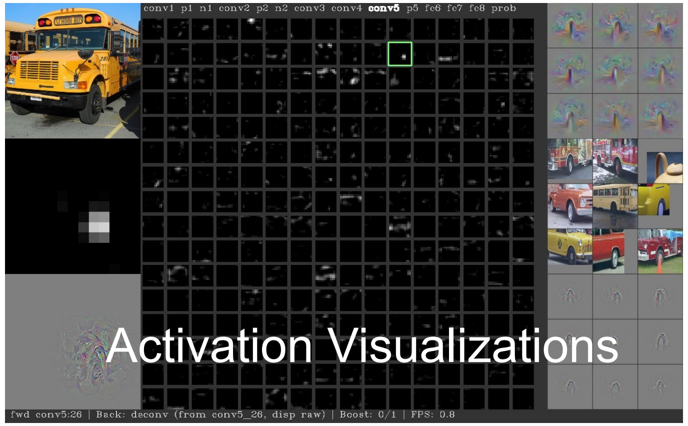
Explanations

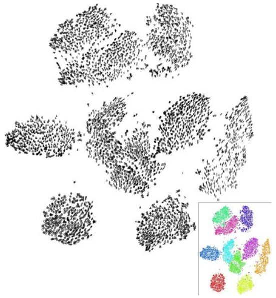
Human-centric Explanations

Explanations can be characterized based on the knowledge of the audience they cater to

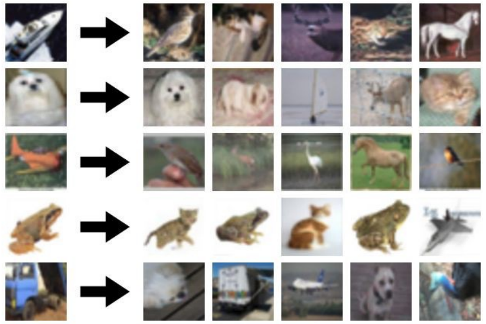
Lecture 3: Indirect and Direct Explanations





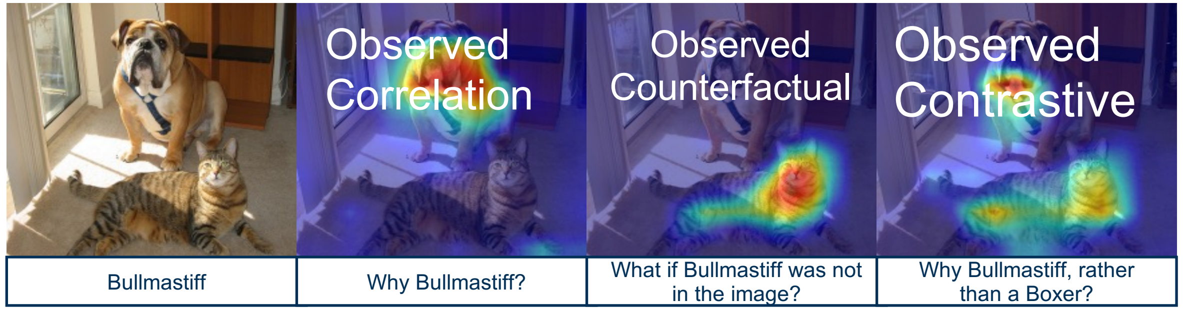


Dimensionality Reduction



Nearest Neighbor

Lecture 4: Targeted Explanations



IEEE Signal Processing Society CELEBRATING 75 YEARS

AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.

OLIVES @GeorgiaTech

Georgia Tech

Outline

Lecture 5: Evaluating Visual Explanations

- Explanatory Evaluation Taxonomy
- Human Evaluation
  - Challenges
  - Methodology
- Application Evaluation
  - Methodology
  - Gaze Prediction
  - Pointing Game
  - Localization
- Network Evaluation
  - Intervention-based Evaluation
  - Masking
  - Progressive Pixel-wise masking
  - Progressive Structure-wise masking
- Challenges in Explanatory Evaluation
  - Human and Application Evaluation
  - Network Evaluation

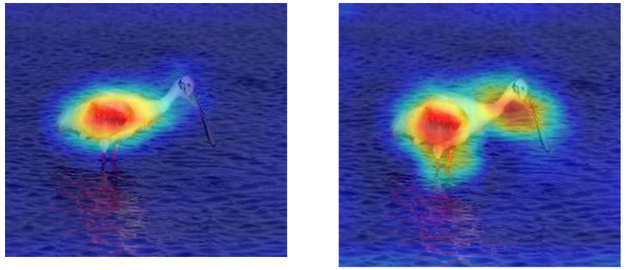
Evaluating Explanations

Explanatory Evaluation Taxonomy

The ability of an entity to explain or justify its decisions or predictions in human-understandable terms

Human Evaluation

Tasks : Humans directly evaluate explanations.



Which explanation is better for answering Why Spoonbill?

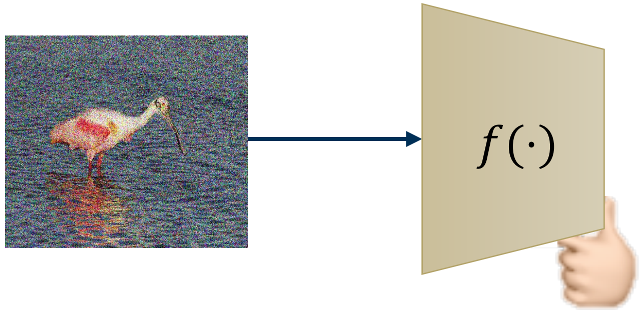
Application Evaluation

Tasks : Any task that requires humans-in-the-evaluation-loop without directly having humans measure explainability.

Which regions in the image are salient to the human visual system?

Network Evaluation

Tasks : Any intervention based on explanation techniques that does not require humans for evaluation.



Is this intervened image still a spoonbill?

AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.

Outline

Lecture 5: Evaluating Visual Explanations

- Explanatory Evaluation Taxonomy
- Human Evaluation
  - Challenges
  - Methodology
- Application Evaluation
  - Methodology
  - Gaze Prediction
  - Pointing Game
  - Localization
- Network Evaluation
  - Intervention-based Evaluation
  - Masking
  - Progressive Pixel-wise masking
  - Progressive Structure-wise masking
- Challenges in Explanatory Evaluation
  - Human and Application Evaluation
  - Network Evaluation

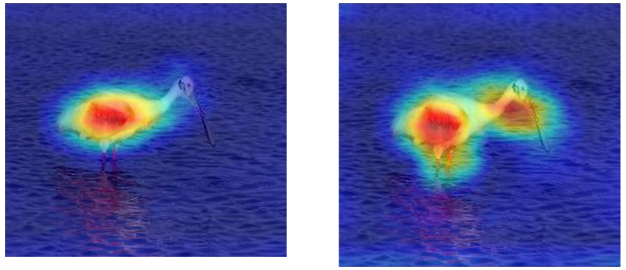
Human Evaluation

Methodology

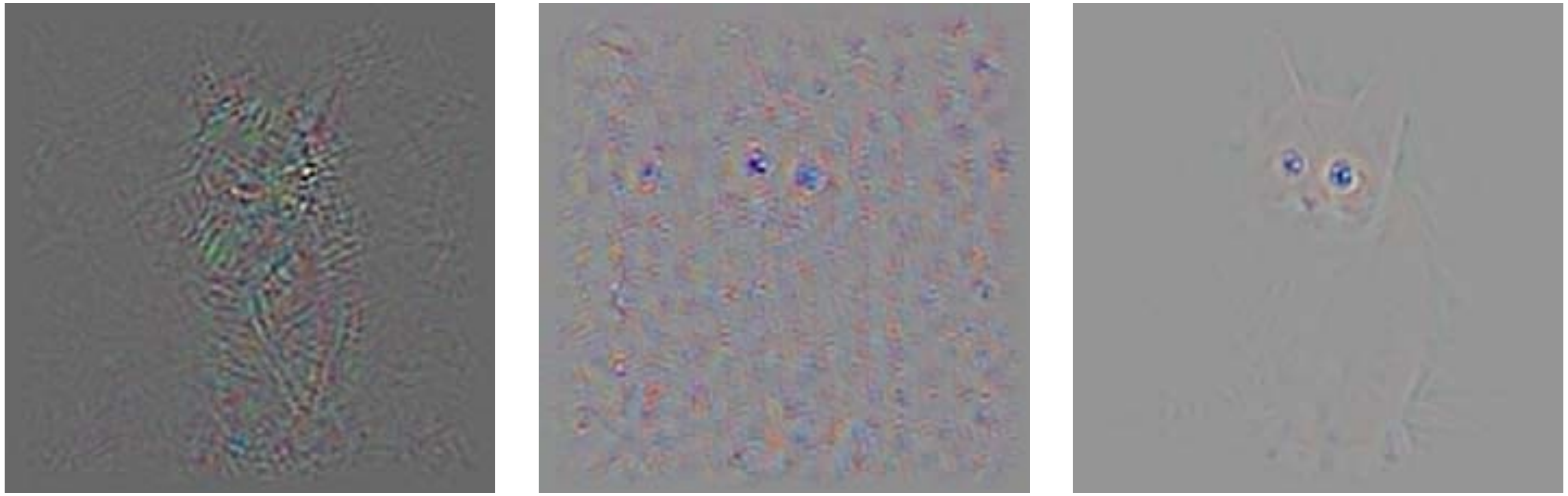
Humans are directly asked to evaluate explanatory techniques

Human Evaluation

Tasks : Humans directly evaluate explanations.



Which explanation is better for answering Why Spoonbill?



Backprop

Deconv

Guided Backprop

Which of the three techniques is better?

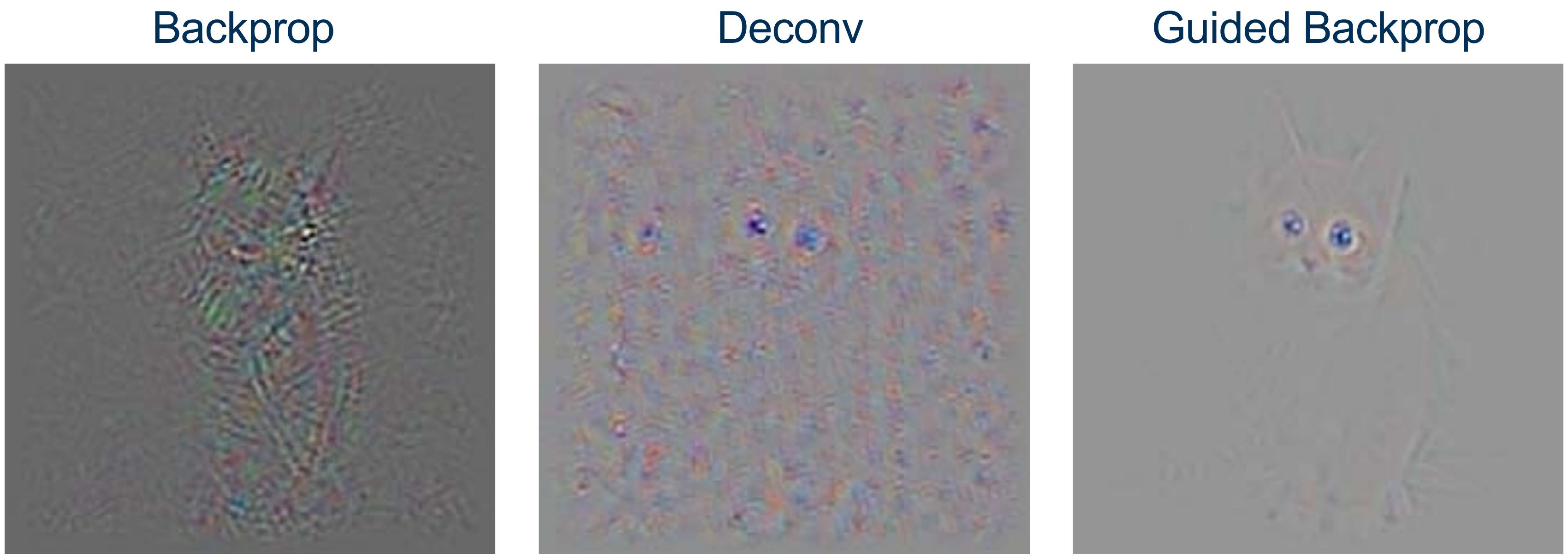
Human Evaluation

Challenges

Humans are directly asked to evaluate explanatory techniques

This evaluation is subjective

- **Cleaner explanation** or **class-discriminative** explanation?
- Should it highlight the **whole cat or only the face**?



Which of the three techniques is better?

Human Evaluation

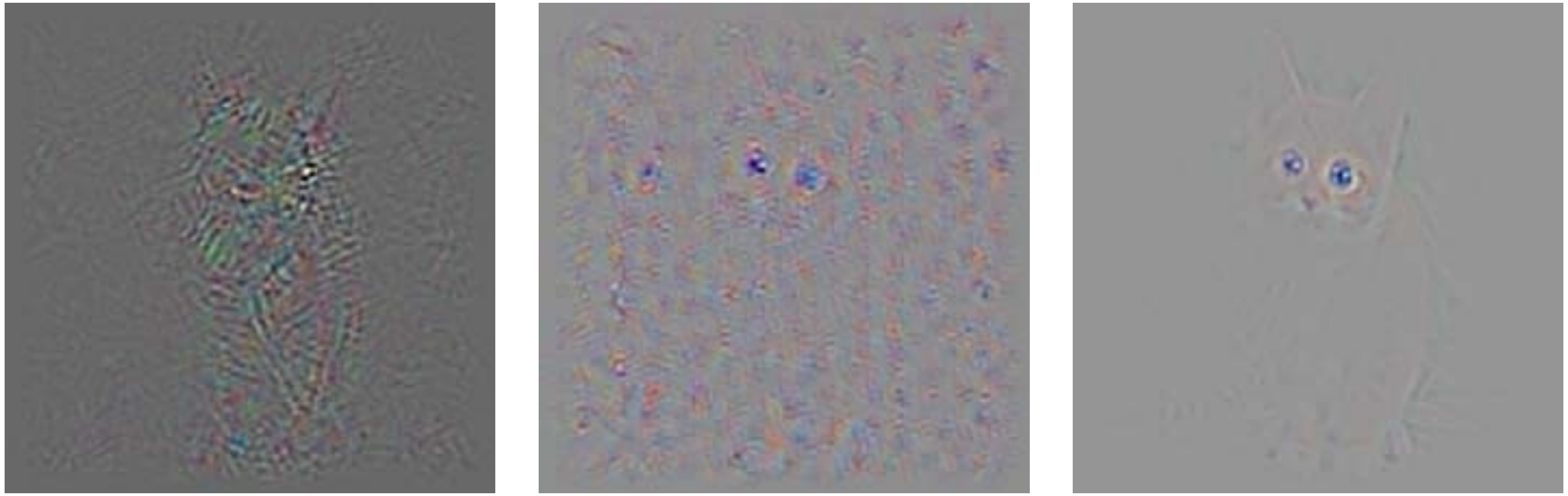
Challenges

Humans are directly asked to evaluate explanatory techniques

This evaluation is subjective
- Cleaner explanation or class-discriminative explanation?
- Should it highlight the whole cat or only the face?

- Ask a `**large number**' of humans, the same question
- Make sure that the **humans are unaware of the goal of the researchers**
- **Guide** them through with **questions**

Backprop Deconv Guided Backprop



Which of the three techniques is better?

Human Evaluation

Methodology

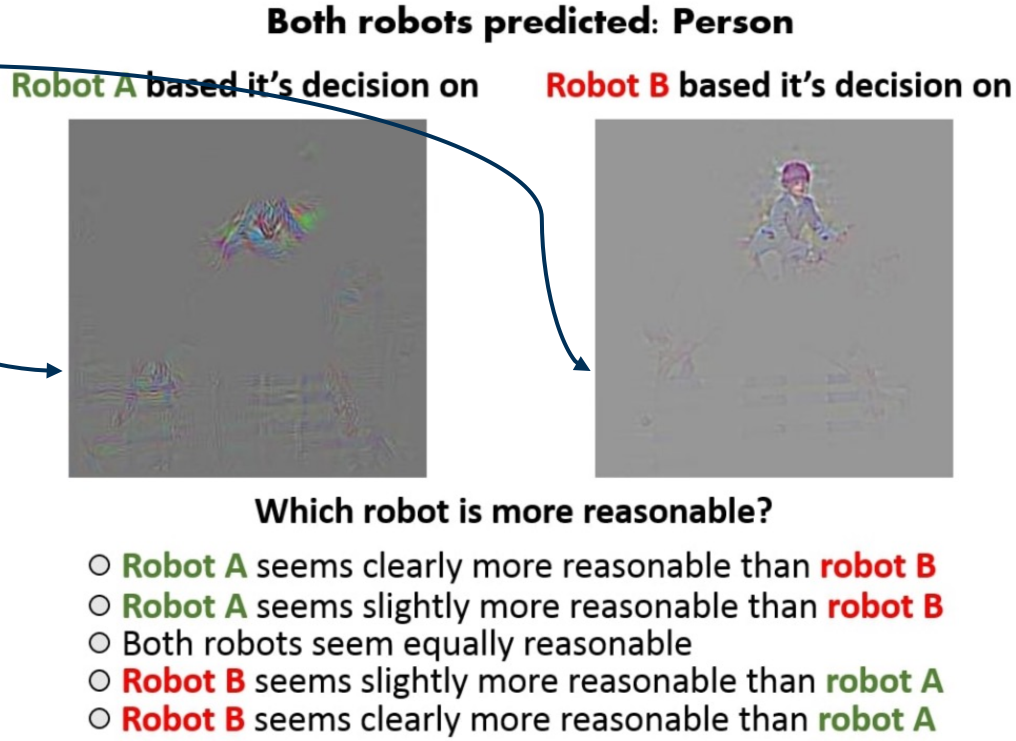
Humans are directly asked to evaluate explanatory techniques

- Ask a `**large number**' of humans, the same question
- Make sure that the **humans are unaware of the goal of the researchers**
- **Guide** them through with **questions**

2 explanations from separate techniques

- 43 AMT workers were asked the adjoining question for each image-question pair
- This experiment was repeated over 90 image-question pairs

Input



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.

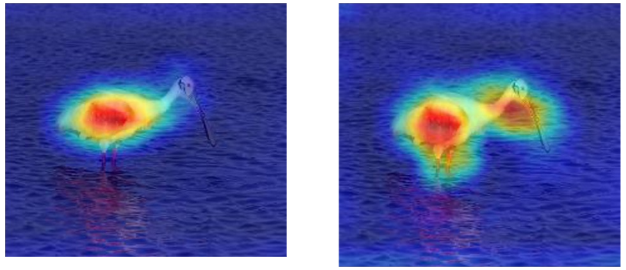
Human Evaluation

Human Evaluation Summary

Humans are directly asked to evaluate explanatory techniques

Human Evaluation

Tasks : Humans directly evaluate explanations.



Which explanation is better for answering Why Spoonbill?

- Evaluates based on the definition of Explainability – as answers from humans

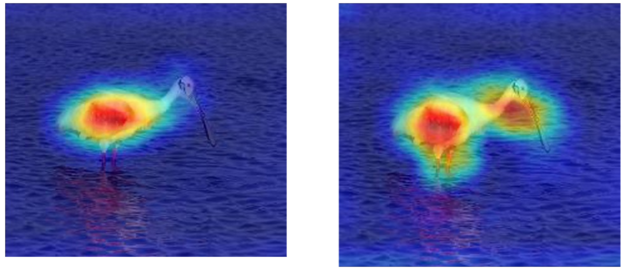
- **Expensive**: Requires a lot of manual effort for large scale datasets
- **Systematic Bias**: Systematic bias cannot be removed
- **Experimental Design**: Evaluation procedure itself is an experimental design problem
- **Domain Knowledge**: Human evaluation works for natural images. However, domain specific knowledge-dependent tasks require specialized explanations and annotations.

AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.

Humans are directly asked to evaluate explanatory techniques

Human Evaluation

Tasks : Humans directly evaluate explanations.



Which explanation is better for answering Why Spoonbill?

Methods	Human	Application	Network
Deconvolution [21]	✓	–	–
Inverted Representations [22]	✓	–	–
Guided-Backpropagation [18]	–	✓	–
SmoothGrad [17]	–	✓	–
LIME [39]	✓	✓	–
CAM [24]	–	✓	–
Graph-CNN [23]	✓	✓	–
GradCAM [12]	✓	✓	–
TCAV [40]	✓	✓	–
GradCAM++ [16]	✓	✓	–
RISE [35]	–	✓	✓
Causal-CAM [15]	✓	–	✓
Counterfactual-CAM [12]	✓	–	–
Goyal et al. [26]	✓	✓	–
CEM [29]	✓	✓	–
Contrast-CAM [13]	✓	–	–
Contrastive reasoning [14]	✓	–	✓

- Human evaluation is the **most popular** Explainability evaluation
- However, in most cases, **authors forego** Mturk style **large-scale evaluation** and only **show randomly** selected exemplar **images**
- Doing so introduces bias and requires additional evaluation techniques

AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.

Outline

Lecture 5: Evaluating Visual Explanations

- Explanatory Evaluation Taxonomy
- Human Evaluation
  - Challenges
  - Methodology
- Application Evaluation
  - Methodology
  - Gaze Prediction
  - Pointing Game
  - Localization
- Network Evaluation
  - Intervention-based Evaluation
  - Masking
  - Progressive Pixel-wise masking
  - Progressive Structure-wise masking
- Challenges in Explanatory Evaluation
  - Human and Application Evaluation
  - Network Evaluation

Application Evaluation

Methodology

Applications that require human annotations are designed or selected. Explanations are then sought retrospectively

Application Evaluation

Tasks : Any task that requires humans-in-the-evaluation-loop without directly having humans measure explainability.

Which regions in the image are salient to the human visual system?

To reduce bias due to the questions themselves, humans are not directly asked to evaluate between explanations.

Instead applications that have already used human annotations are used to evaluate explanations

- Detection and Localization
- Gaze tracking
- Pointing game

AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.

Application Evaluation

Gaze Tracking

Assumption: Human focus on regions that allows some inference. Hence, salient regions are explanations

Application Evaluation
Tasks : Any task that requires humans-in-the-evaluation-loop without directly having humans measure explainability.

Which regions in the image are salient to the human visual system?

Gaze Tracking

Given an image, humans tend to focus on the the salient regions.

Tracking human visual gaze without a specific objective results in salient objects being focused on. From a neuroscience perspective, the inference engine, that is the brain, uses these salient regions to make any inference. Hence, the salient regions are an explanation for any inference.

Y. Sun, M. Prabhushankar, and G. AlRegib, ”Implicit Saliency in Deep Neural Networks,” in IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, Oct. 2020.

Application Evaluation

Gaze Tracking

Assumption: Human focus on regions that allows some inference. Hence, salient regions are explanations

Application Evaluation

Tasks : Any task that requires humans-in-the-evaluation-loop without directly having humans measure explainability.

Which regions in the image are salient to the human visual system?

Gaze Tracking

Given an image, humans tend to focus on the the salient regions.

The explanations from various methods are evaluated against this ground truth using Correlation Coefficient (CC) and Normalized Scanpath Saliency (NSS)

Y. Sun, M. Prabhushankar, and G. AlRegib, "Implicit Saliency in Deep Neural Networks," in IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, Oct. 2020.

Assumption: Human focus on regions that allows some inference. Hence, salient regions are explanations

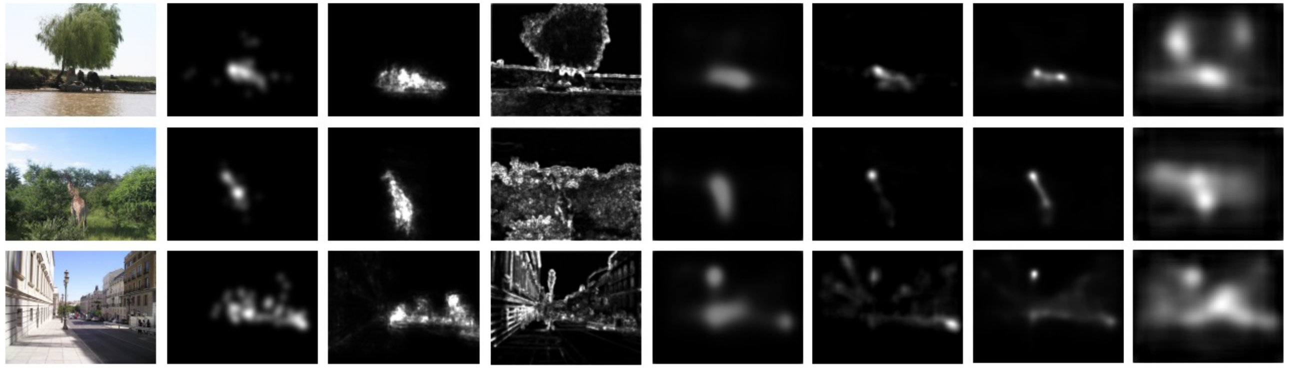


Fig. 3. Saliency map visualization. (a) Input image (b) Groudtruth (c) Proposed Method (d) Feed-forward feature (e) SalGan [21] (f) ML-Net [5] (g) DeepGazeII [22] (h) ShallowDeep [23]

Table 1. Human visual saliency vs Model Saliency

	NSS				CC			
Networks	ResNet-18	ResNet-34	ResNet-50	ResNet-101	ResNet-18	ResNet-34	ResNet-50	ResNet-101
GradCam	0.7657	0.7545	0.7203	0.7335	0.3496	0.3396	0.3190	0.3210
GBP	0.3862	0.4191	0.3898	0.3415	0.2474	0.2453	0.2443	0.2233
ImplicitSaliency	**0.8274**	**0.8018**	**0.7659**	**0.7981**	**0.4132**	**0.4112**	**0.3868**	**0.4051**

Y. Sun, M. Prabhushankar, and G. AlRegib, ”Implicit Saliency in Deep Neural Networks,” in IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, Oct. 2020.

Application Evaluation

Gaze Tracking

Assumption: Human focus on regions that allows some inference. Hence, salient regions are explanations

Gaze Tracking

- Tracks true salient regions in an image **without bias of questions or questionnaire**

- Requires **expensive eye-tracking equipment**
- May lead to **spurious noise and center-bias**
- **No targeted answers or explanations** can be obtained

Y. Sun, M. Prabhushankar, and G. AlRegib, "Implicit Saliency in Deep Neural Networks," in IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, Oct. 2020.

Application Evaluation

Pointing Game

Given a blurry image and a question, humans are asked to sharpen the regions in the image that lead to their decision

Application Evaluation

Tasks : Any task that requires humans-in-the-evaluation-loop without directly having humans measure explainability.

Which regions in the image are salient to the human visual system?

- Saliency through gaze tracking is completely unsupervised
- Pointing game adds questions to it

Question: How many players are visible in the image?

Das, Abhishek, et al. "Human attention in visual question answering: Do humans and deep networks look at the same regions?." Computer Vision and Image Understanding 163 (2017): 90-100.

Application Evaluation

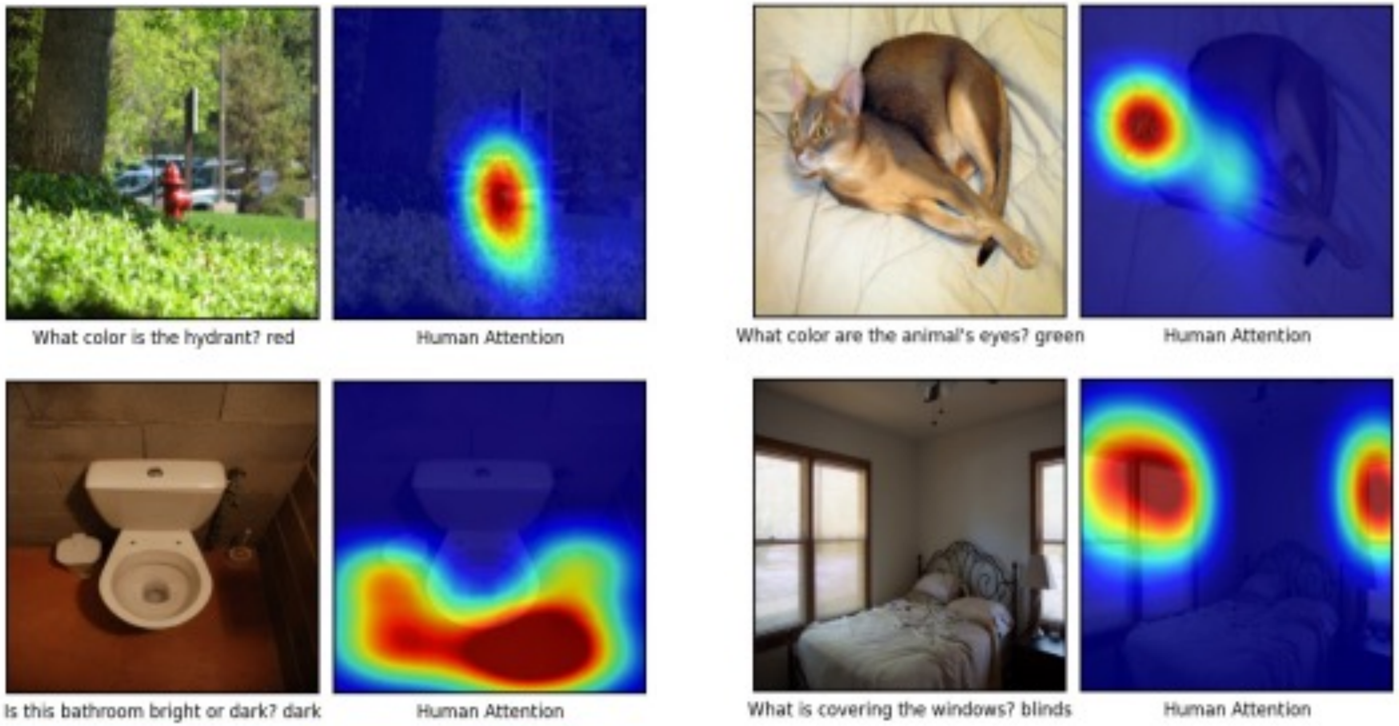
Pointing Game

Applications that require human annotations are designed or selected. Explanations are then sought retrospectively

Application Evaluation

Tasks : Any task that requires humans-in-the-evaluation-loop without directly having humans measure explainability.

Which regions in the image are salient to the human visual system?



Explanations from various methods are evaluated against this ground truth using Mean Intersection over Union, CC or other segmentation metrics

Das, Abhishek, et al. "Human attention in visual question answering: Do humans and deep networks look at the same regions?." Computer Vision and Image Understanding 163 (2017): 90-100.

Application Evaluation

Pointing Game

Given a blurry image and a question, humans are asked to sharpen the regions in the image that lead to their decision

Question: How many players are visible in the image?

- Does **not require specialized equipment** and can be performed on Mechanical Turk
- Targeted answers and explanations can be obtained based on **targeted questions**

- May **introduce bias** in viewer
- Viewers may **miss details** based on blur levels

Das, Abhishek, et al. "Human attention in visual question answering: Do humans and deep networks look at the same regions?." Computer Vision and Image Understanding 163 (2017): 90-100.

Application Evaluation

Detection and Localization

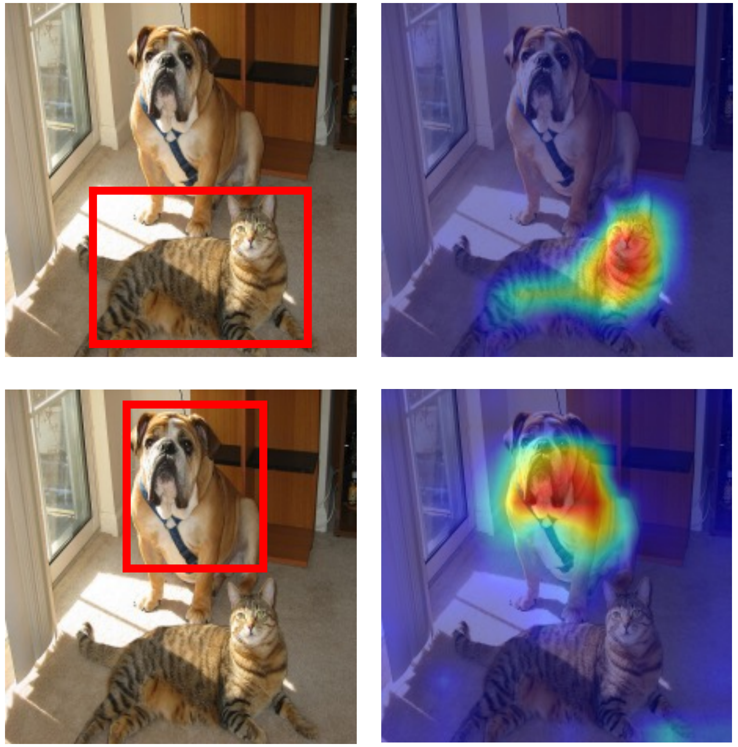
Assumption: Humans localize objects in an image. Explanations (for those object predictions) must lie within the localized region

Application Evaluation

Tasks : Any task that requires humans-in-the-evaluation-loop without directly having humans measure explainability.

Which regions in the image are salient to the human visual system?

Object Labels | Explanations



- Object localizations are provided by annotators beforehand
- **Assumption:** Explanation for *"Why Dog?"* and *"Why Cat?"* must **highlight the cat and dog features** within the bounding box
- Explanations are **evaluated** against this ground truth using **Detection/Segmentation** metrics

Application Evaluation

Application Evaluation Summary

Assumption: Humans localize objects in an image. Explanations (for those object predictions) must lie within the localized region

Application Evaluation

Tasks : Any task that requires humans-in-the-evaluation-loop without directly having humans measure explainability.

Which regions in the image are salient to the human visual system?

Methods	Human	Application	Network
Deconvolution [21]	✓	–	–
Inverted Representations [22]	✓	–	–
Guided-Backpropagation [18]	–	✓	–
SmoothGrad [17]	–	✓	–
LIME [39]	✓	✓	–
CAM [24]	–	✓	–
Graph-CNN [23]	✓	✓	–
GradCAM [12]	✓	✓	–
TCAV [40]	✓	✓	–
GradCAM++ [16]	✓	✓	–
RISE [35]	–	✓	✓
Causal-CAM [15]	✓	–	✓
Counterfactual-CAM [12]	✓	–	–
Goyal et al. [26]	✓	✓	–
CEM [29]	✓	✓	–
Contrast-CAM [13]	✓	–	–
Contrastive reasoning [14]	✓	–	✓

- **Application evaluation in conjunction with human evaluation is the most common validation of Explainability**
- Application evaluation provides an **indirect objective** for explanation
- Application evaluation ties decision making with Explainability

AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.

Outline

Lecture 5: Evaluating Visual Explanations

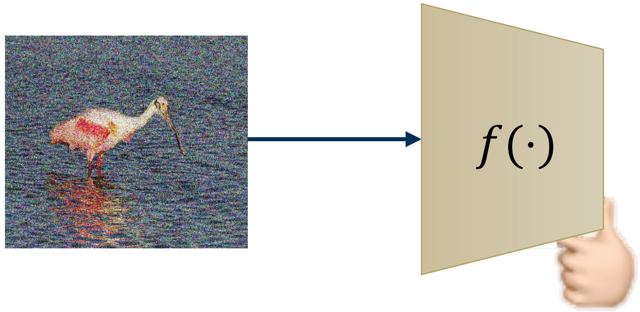
- Explanatory Evaluation Taxonomy
- Human Evaluation
  - Challenges
  - Methodology
- Application Evaluation
  - Methodology
  - Gaze Prediction
  - Pointing Game
  - Localization
- Network Evaluation
  - Intervention-based Evaluation
  - Masking
  - Progressive Pixel-wise masking
  - Progressive Structure-wise masking
- Challenges in Explanatory Evaluation
  - Human and Application Evaluation
  - Network Evaluation

Intervening within data and objectively evaluating the *effect of Explainability* on networks

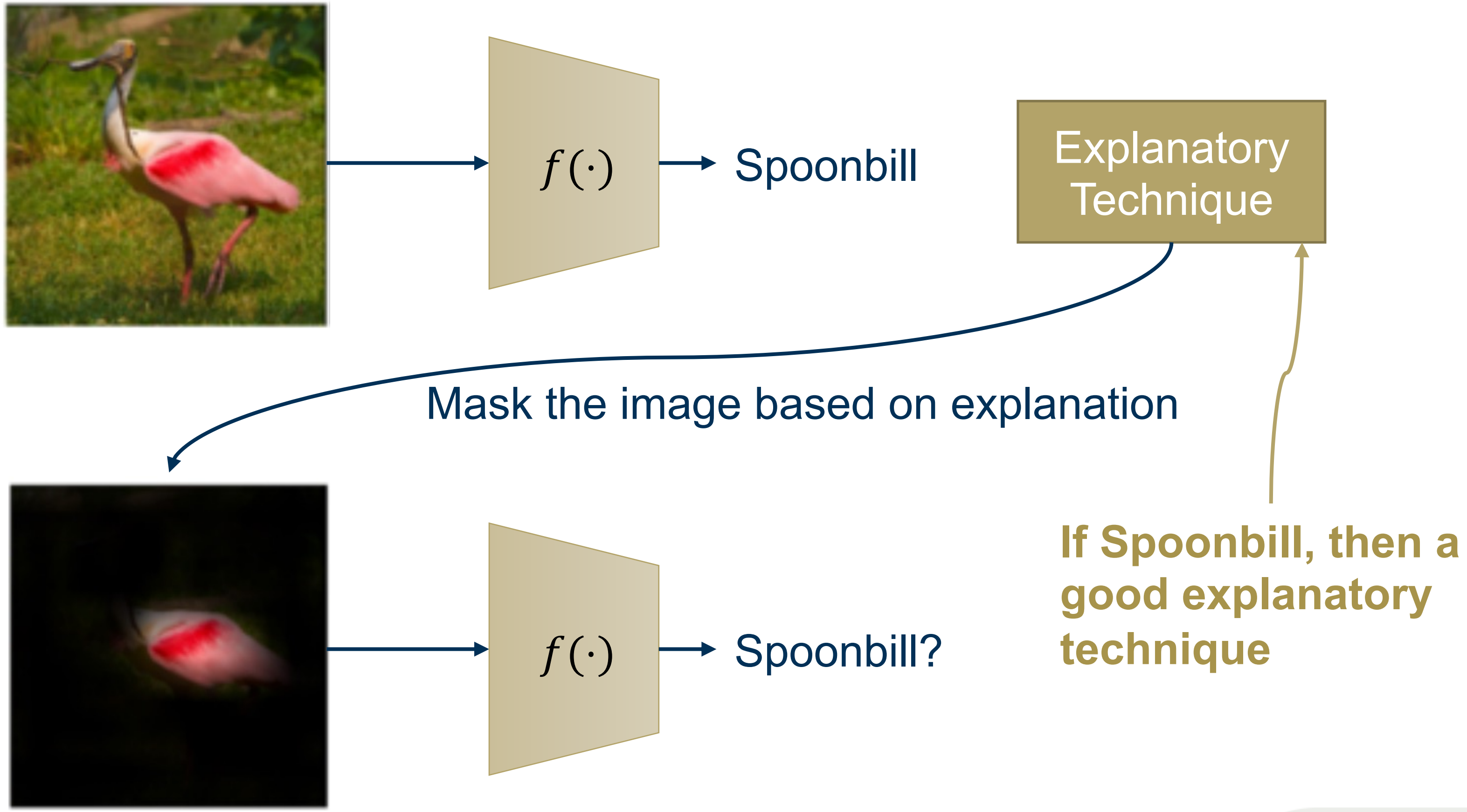
Ex: Masking

Network Evaluation

Tasks : Any intervention based on explanation techniques that does not require humans for evaluation.



Is this intervened image still a spoonbill?



AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.

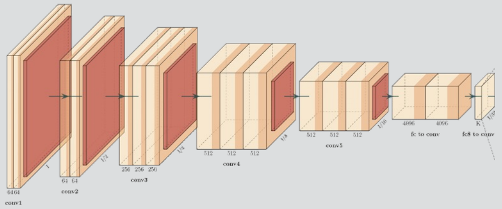
Network Evaluation

Justification for Intervention: Necessity

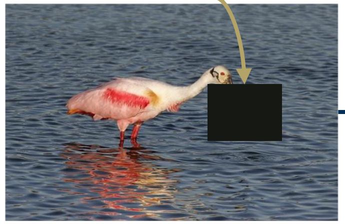
Lecture 2: Properties of Explanations are Necessity and Sufficiency

Conditions for Necessity

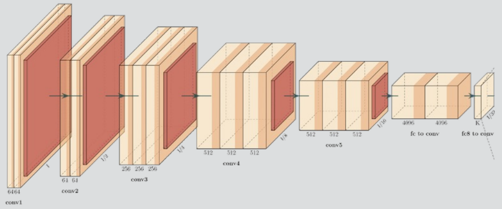
Network $\boldsymbol{f}(\boldsymbol{\theta})$



Intervention



Network $\boldsymbol{f}(\boldsymbol{\theta})$



Beak
Neck
Legs
Feathers
Water
Grass
Teeth
.
.

Features $\mathcal{T}_P$

P is Spoonbill

If still Spoonbill, then beak is not a necessary condition for bird

Chowdhury, Prithwijit, Mohit Prabhushankar, and Ghassan AlRegib. "Explaining Explainers: Necessity and Sufficiency in Tabular Data." *NeurIPS 2023 Second Table Representation Learning Workshop*. 2023.

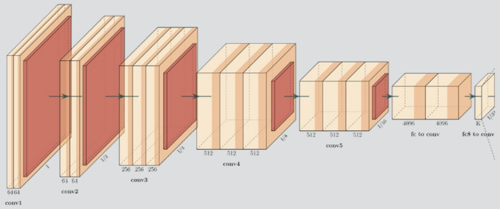
Network Evaluation

Justification for Intervention: Sufficiency

Lecture 2: Properties of Explanations are Necessity and Sufficiency

Conditions for Sufficiency

Network $\boldsymbol{f}(\boldsymbol{\theta})$

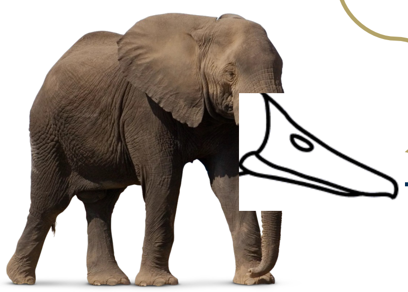


Beak
Neck
Legs
Feathers
Water
Grass
Teeth
.
.

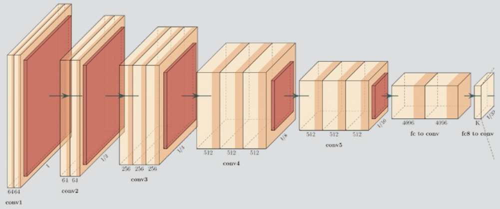
Features $\mathcal{T}_P$

P is Spoonbill

Adding a beak



Network $\boldsymbol{f}(\boldsymbol{\theta})$



If this is Spoonbill, then beak is a sufficient condition of Spoonbill

Chowdhury, Prithwijit, Mohit Prabhushankar, and Ghassan AlRegib. "Explaining Explainers: Necessity and Sufficiency in Tabular Data." *NeurIPS 2023 Second Table Representation Learning Workshop*. 2023.

Network Evaluation

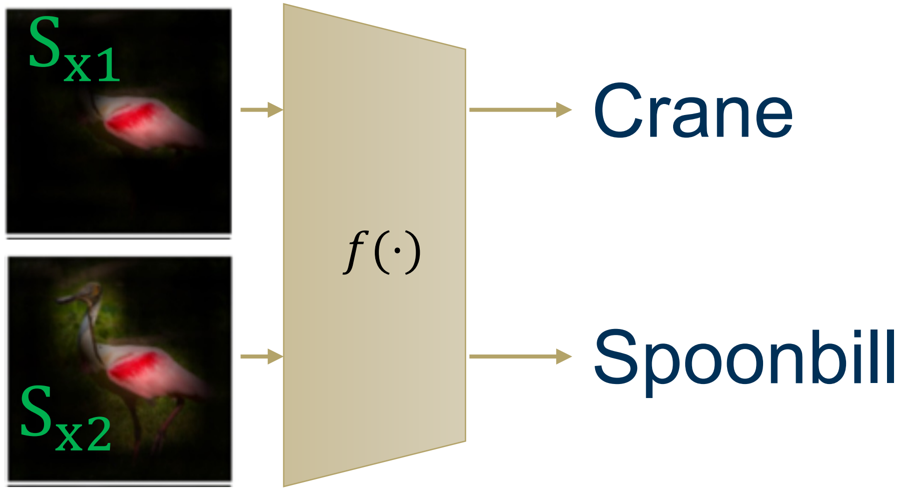
Evaluation 1: Masking

Intervening within data and objectively evaluating the *effect of Explainability* on networks

y = Prediction
S_x = Explanation masked data

$E(Y|S_x)$ = Expectation of class given S_x

If across N images, $\mathbf{E(Y|}\boldsymbol{S}_{\mathbf{x2}}\mathbf{)} > \mathbf{E(Y|}\boldsymbol{S}_{\mathbf{x1}}\mathbf{)}$, explanation technique 2 is better than explanation technique 1



AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.

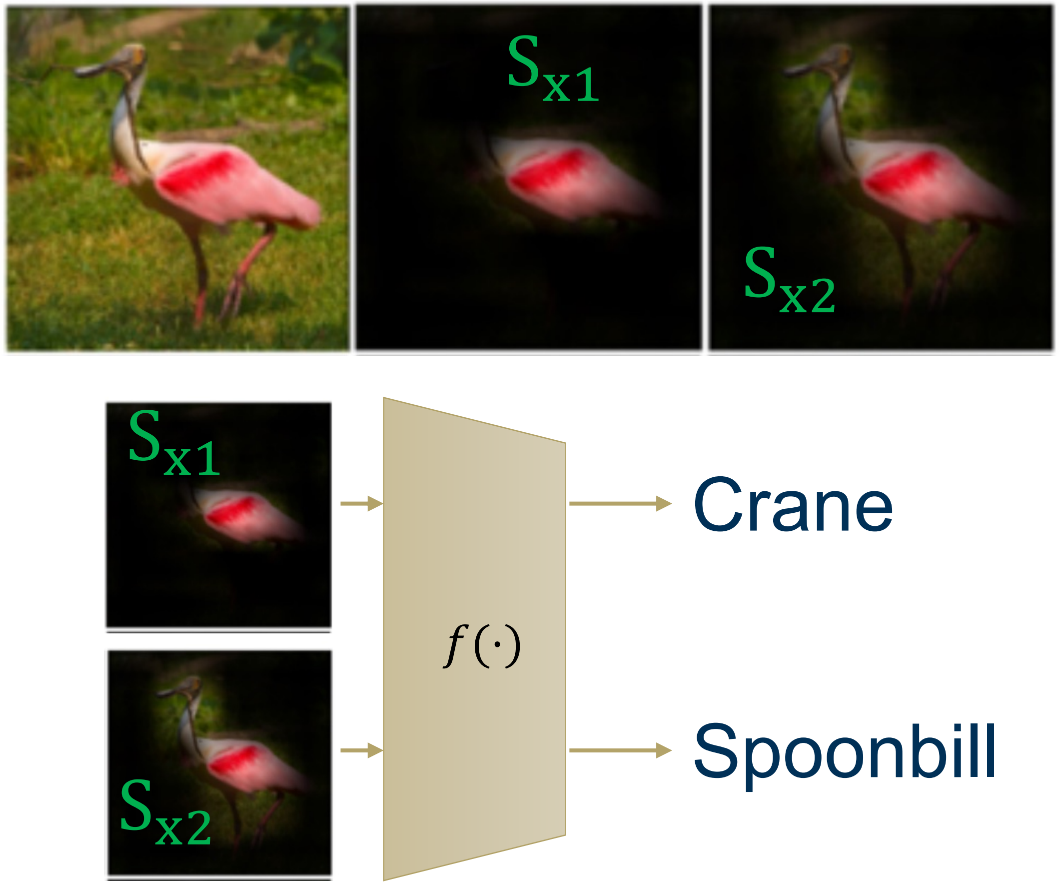
Network Evaluation

Evaluation 1: Masking

Intervening within data and objectively evaluating the *effect of Explainability* on networks

- Masking is an intuitive methodology for objective evaluation
- **Mean masks** are used instead of black masks to overcome the **network preprocessing**

- However, **larger explanation** leads to **better classification**
- Masking evaluation **encourages larger and less fine-grained explanations**

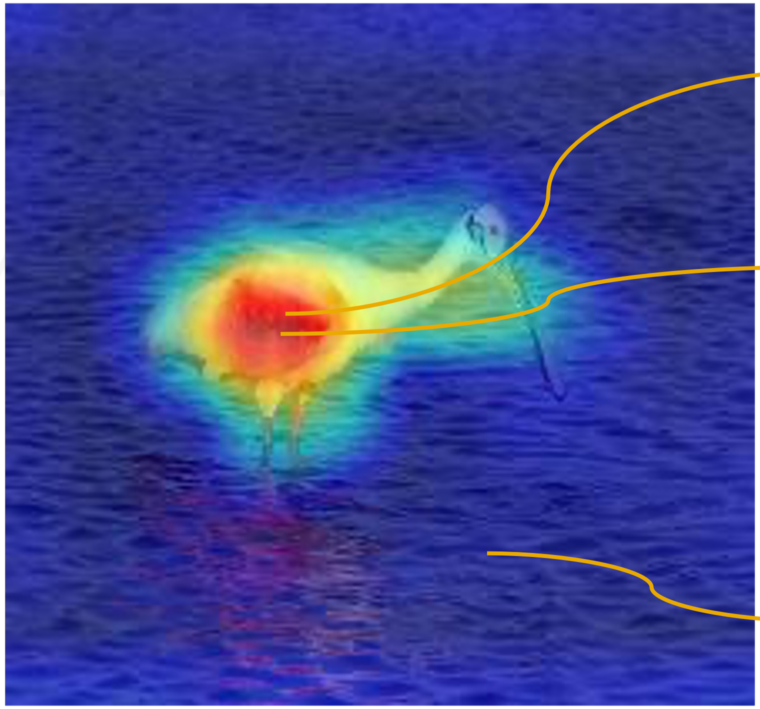


Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018.

Network Evaluation

Evaluation 2: Progressive Pixel-wise Insertion and Deletion

Pixel-wise Deletion: Sequentially delete (mask) pixels in an image based on their explanation assigned importance scores



Highest importance

Second Highest importance

.

.

.

Least importance

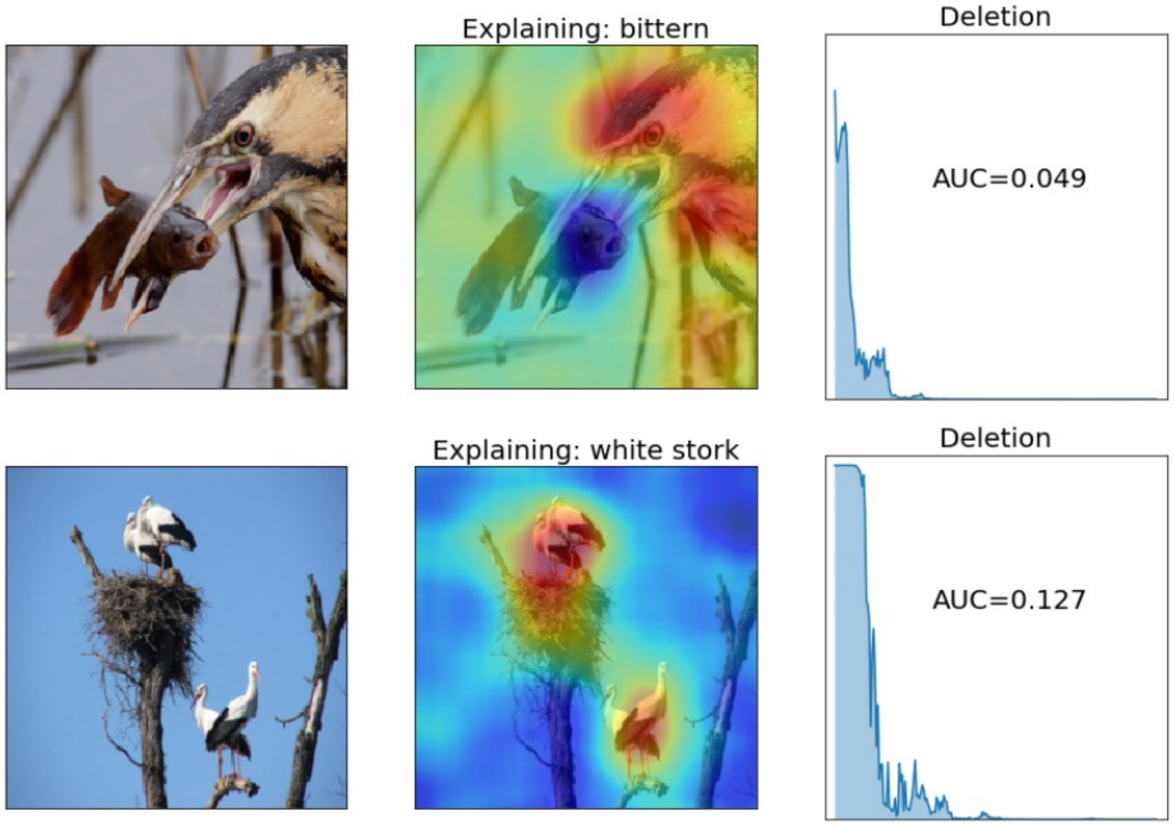
Step 1: Mask highest importance pixel and pass the image through the network. Note the probability of spoonbill.
Step 2: Mask the second highest importance pixel from the image in Step 1 and pass the image through the network. Note the probability of spoonbill.
Step 3: Repeat until all pixels are deleted (masked)

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

Network Evaluation

Evaluation 2: Progressive Pixel-wise Insertion and Deletion

The removal of the "cause" (important pixels) will force the base model to change its decision.



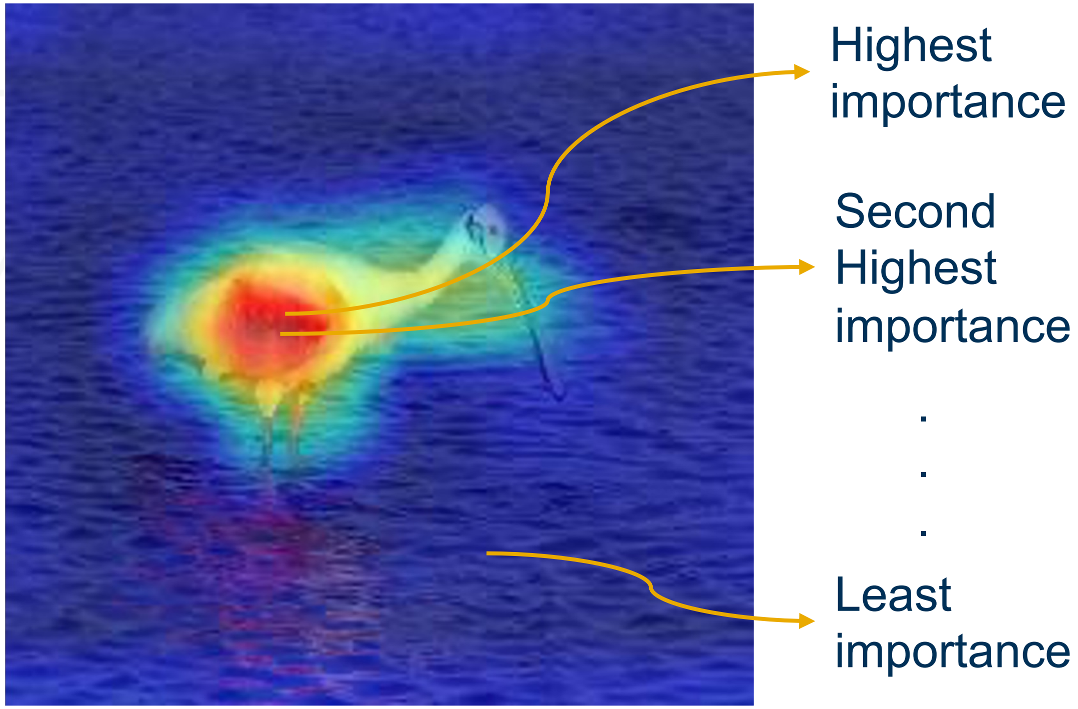
- **Deletion approximates Necessity** criterion of a "good" explanation
- **AUC** for a good explanation will be **low**
- **Deletion** encourages **fine-grained explanations** by choosing those heatmaps that select the most relevant pixels

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

Network Evaluation

Evaluation 2: Progressive Pixel-wise Insertion and Deletion

Pixel-wise Insertion: Sequentially add pixels to a mean image based on their explanation assigned importance scores



Take a mean (grayscale) image

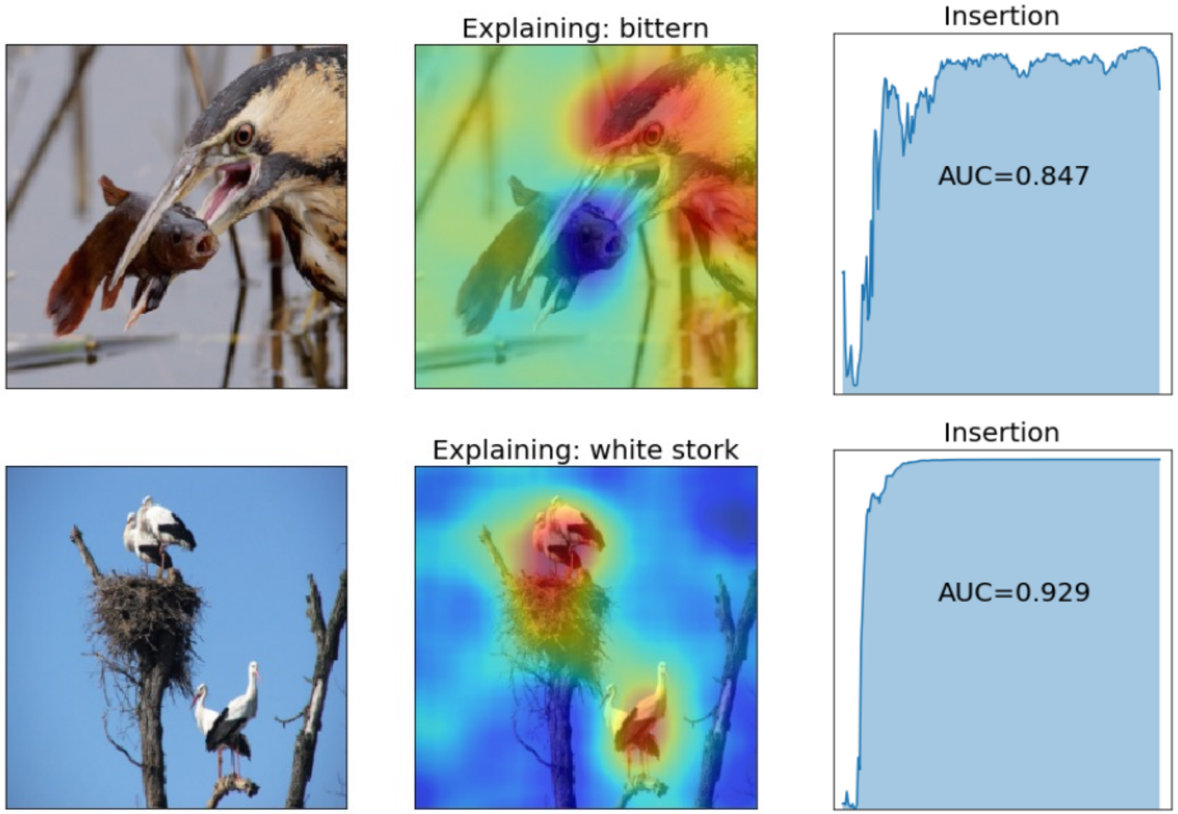
Step 1: Add the highest importance pixel to the mean image and pass it through the network. Note the probability of spoonbill.
Step 2: Add the second highest importance pixel to the image in Step 1 and pass the image through the network. Note the probability of spoonbill.
Step 3: Repeat until all pixels are inserted

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

Network Evaluation

Evaluation 2: Progressive Pixel-wise Insertion and Deletion

The addition of the "cause" (important pixels) will force the base model to change its decision.



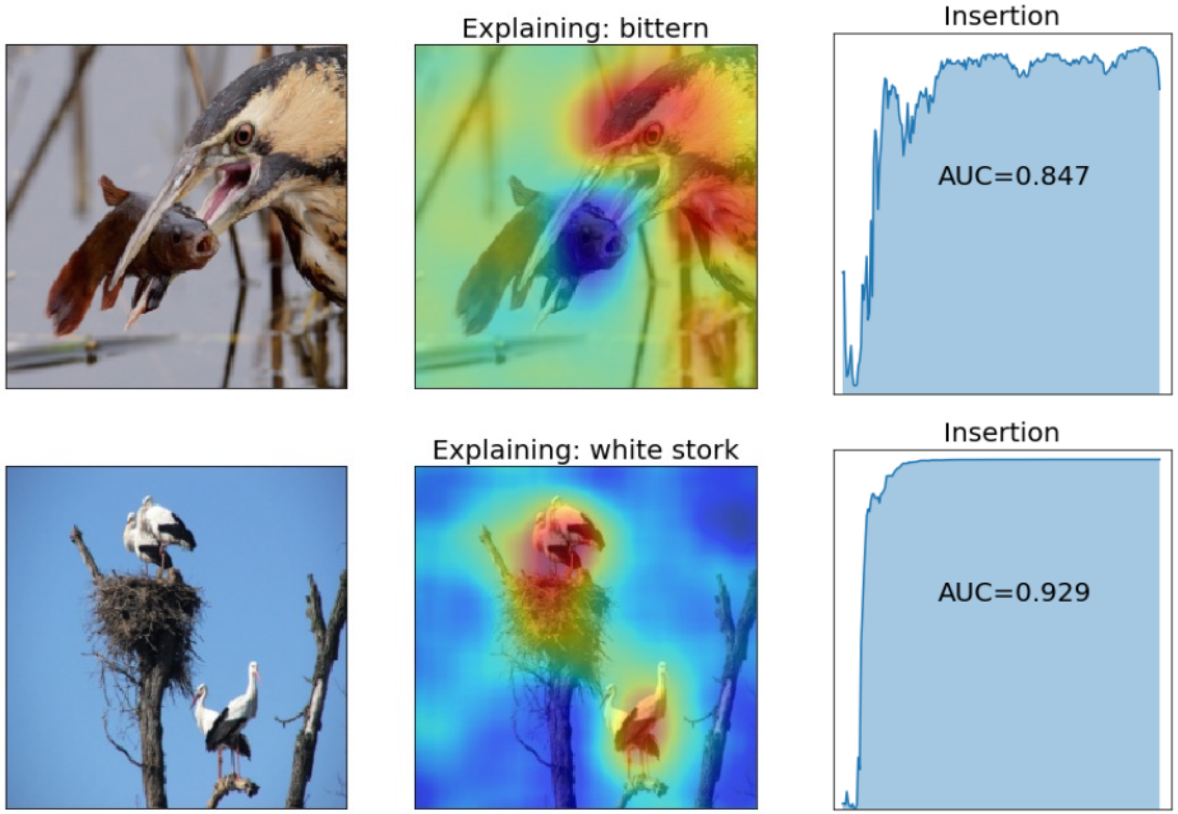
- **Insertion approximates Sufficiency** criterion of a "good" explanation
- **AUC** for a good explanation will be **high**
- **Insertion** encourages **fine-grained explanations** by choosing those heatmaps that select the most relevant pixels

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

Network Evaluation

Evaluation 2: Progressive Pixel-wise Insertion and Deletion

Insertion and Deletion evaluation metrics encourage pixel-wise analysis of explanations



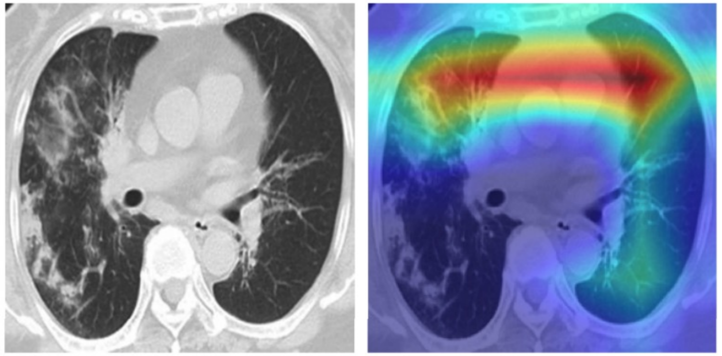
- **However, humans do not "see" in pixels**
- Rather they view scenes in a "**structure-wise**" fashion
- While **heatmap masking** encourages **large explanations**, **pixel-wise masking** encourages **unrealistic and non-human** like explanations

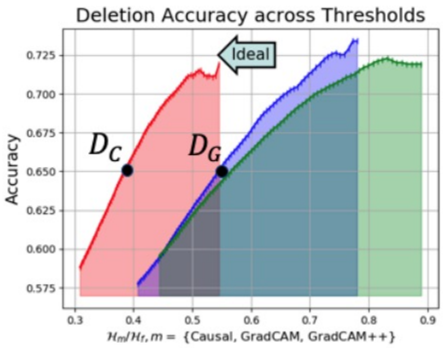
Vitali Petsiuk, Abir Das, and Kate Saenko, "Rise: Randomized input sampling for explanation of black-box models," arXiv preprint arXiv:1806.07421, 2018.

Network Evaluation

Evaluation 3: Progressive Structure-wise Insertion and Deletion

Structure-wise Deletion: Sequentially delete (mask) pixels in an image based on the number of bits used to represent the region





Ideal scenario: The explanation encodes the most important information in the least possible bits

CausalCAM in Red[1]
GradCAM in Purple
GradCAM++ in Green

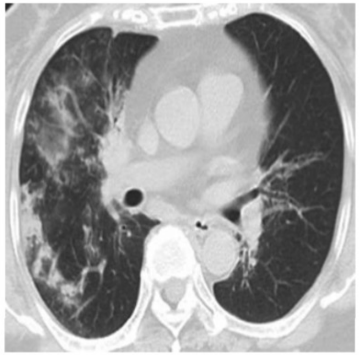
- D_C and D_G represent 65% accuracy for CausalCAM and GradCAM respectively
- **CausalCAM encodes dense structure-rich features in lesser bits, that aid accuracy**

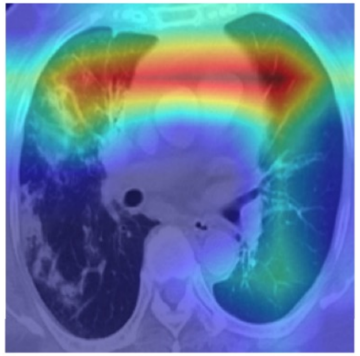
Prabhushankar, Mohit, and Ghassan AlRegib. "Extracting causal visual features for limited label classification." *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021.

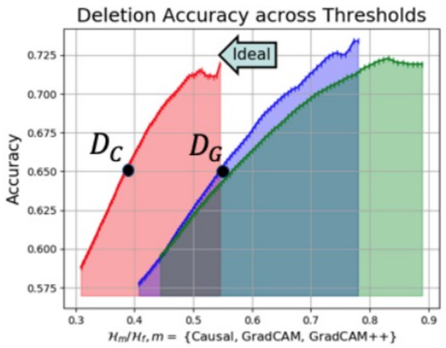
Network Evaluation

Evaluation 3: Progressive Structure-wise Insertion and Deletion

Structure-wise Deletion: Sequentially delete (mask) pixels in an image based on the number of bits used to represent the region







Ideal scenario: The explanation encodes the most important information in the least possible bits

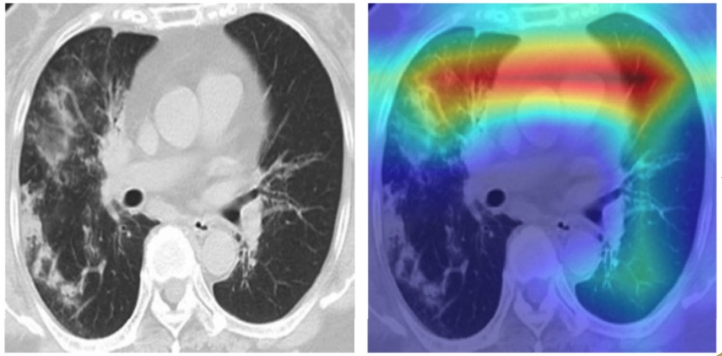
Step 1: Choose a threshold in the explanation (say 0.1) and delete (mask) all the pixels in the original image below the threshold. Pass the masked image through the network and note the change in prediction (if any)

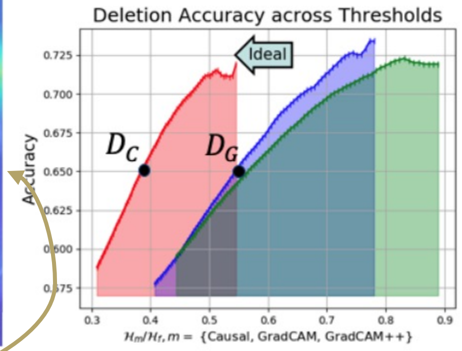
Prabhushankar, Mohit, and Ghassan AlRegib. "Extracting causal visual features for limited label classification." *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021.

Network Evaluation

Evaluation 3: Progressive Structure-wise Insertion and Deletion

Structure-wise Deletion: Sequentially delete (mask) pixels in an image based on the number of bits used to represent the region





Y-axis: Performance accuracy across all ratios

X-axis: Ratio of Huffman encoded masked and original images for all explanations. Smaller the ratio, less is the number of bits encoding the masked image

Ideal scenario: The explanation encodes the most important information in the least possible bits

Step 1: Choose a threshold in the explanation (say 0.1) and delete (mask) all the pixels in the original image below the threshold. Pass the masked image through the network and note the change in prediction (if any)

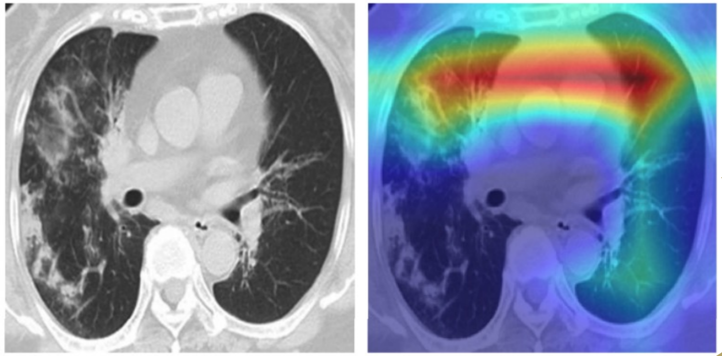
Step 2: Calculate the Huffman code for the original and the masked image. The ratio between the codes of masked and original image is taken on the x-axis and the corresponding accuracy across all images is shown on the y-axis

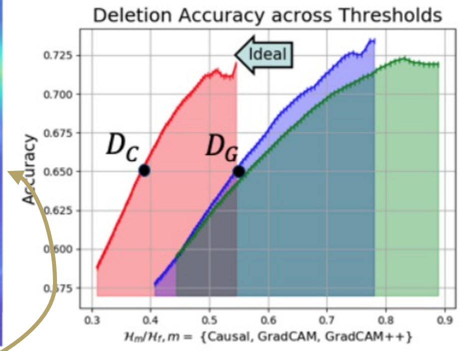
Prabhushankar, Mohit, and Ghassan AlRegib. "Extracting causal visual features for limited label classification." *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021.

Network Evaluation

Evaluation 3: Progressive Structure-wise Insertion and Deletion

Structure-wise Deletion: Sequentially delete (mask) pixels in an image based on the number of bits used to represent the region





Y-axis: Performance accuracy across all ratios

X-axis: Ratio of Huffman encoded masked and original images for all explanations. Smaller the ratio, less is the number of bits encoding the masked image

Ideal scenario: The explanation encodes the most important information in the least possible bits

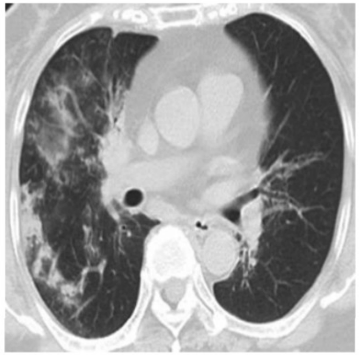
Step 1: Choose a threshold in the explanation (say 0.1) and delete (mask) all the pixels in the original image below the threshold. Pass the masked image through the network and note the change in prediction (if any)
Step 2: Calculate the Huffman code for the original and the masked image. The ratio between the codes of masked and original image is taken on the x-axis and the corresponding accuracy across all images is shown on the y-axis
Step 3: Repeat across thresholds

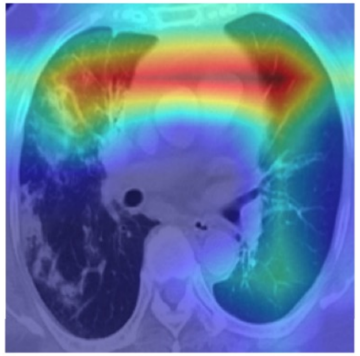
Prabhushankar, Mohit, and Ghassan AlRegib. "Extracting causal visual features for limited label classification." *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021.

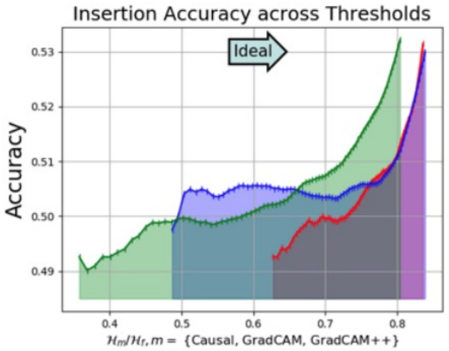
Network Evaluation

Evaluation 3: Progressive Structure-wise Insertion and Deletion

Structure-wise Insertion: Sequentially add (insert) pixels in an image based on the number of bits used to represent the region







Ideal scenario: The explanation encodes the most important information in the least possible bits

CausalCAM in Red[1]
GradCAM in Purple
GradCAM++ in Green

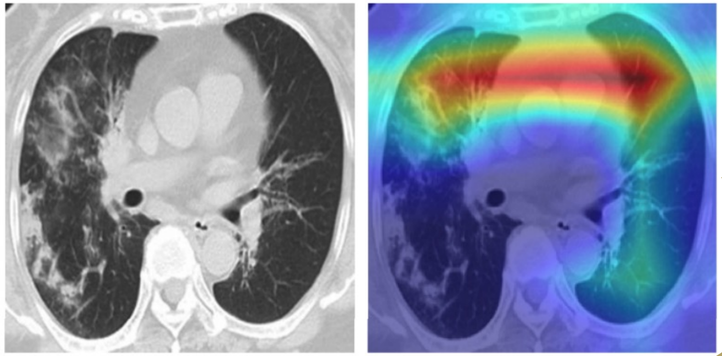
- **CausalCAM encodes dense structure-rich features in at the lowest threshold, that aid accuracy**

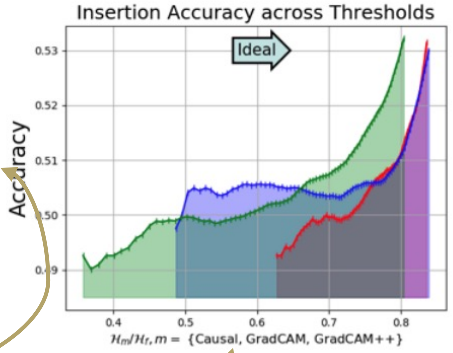
Prabhushankar, Mohit, and Ghassan AlRegib. "Extracting causal visual features for limited label classification." *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021.

Network Evaluation

Evaluation 3: Progressive Structure-wise Insertion and Deletion

Structure-wise Insertion: Sequentially add (insert) pixels in an image based on the number of bits used to represent the region





Y-axis: Performance accuracy across all ratios

X-axis: Ratio of Huffman encoded inserted and original images for all explanations. Larger the ratio, more is the number of bits encoding the inserted image

Ideal scenario: The explanation encodes the most important information in the least possible bits

Step 1: Choose a threshold in the explanation (say 0.1) and insert (add) all the pixels in the original image above the threshold. Pass the inserted image through the network and note the change in prediction (if any)

Step 2: Calculate the Huffman code for the original and the inserted image. The ratio between the codes of inserted and original image is taken on the x-axis and the corresponding accuracy across all images is shown on the y-axis

Step 3: Repeat across thresholds

Prabhushankar, Mohit, and Ghassan AlRegib. "Extracting causal visual features for limited label classification." *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021.

Network Evaluation

Evaluation 3: Progressive Structure-wise Insertion and Deletion

Evaluation 1: Explanation heatmap masking

- **Pro**: Structures are visible in the explanations
- **Con**: Encourages large non-fine grained explanations

Evaluation 2: Pixel-wise insertion and deletion

- **Pro**: Progressively assigns importance to pixels
- **Con**: Encourages unrealistic fine-grained explanations

Evaluation 3: Structure-wise insertion and deletion

- **Pro**: Encourages structures while progressively assigning importance to structures based on information bits
- **Pro**: Other human-centric measures including SSIM, saliency etc. can be used on x-axis

- **Con**: Encourages causal explanations without considering context information (More in Lecture 9)

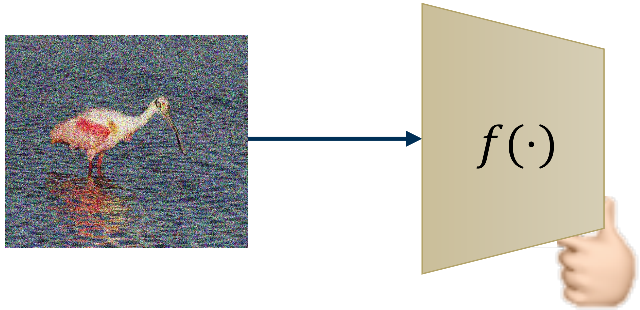
Network Evaluation

Network Evaluation Summary

Intervening within data and objectively evaluating the *effect of Explainability* on networks

Network Evaluation

Tasks : Any intervention based on explanation techniques that does not require humans for evaluation.



Is this intervened image still a spoonbill?

Methods	Human	Application	Network
Deconvolution [21]	✓	–	–
Inverted Representations [22]	✓	–	–
Guided-Backpropagation [18]	–	✓	–
SmoothGrad [17]	–	✓	–
LIME [39]	✓	✓	–
CAM [24]	–	✓	–
Graph-CNN [23]	✓	✓	–
GradCAM [12]	✓	✓	–
TCAV [40]	✓	✓	–
GradCAM++ [16]	✓	✓	–
RISE [35]	–	✓	✓
Causal-CAM [15]	✓	–	✓
Counterfactual-CAM [12]	✓	–	–
Goyal et al. [26]	✓	✓	–
CEM [29]	✓	✓	–
Contrast-CAM [13]	✓	–	–
Contrastive reasoning [14]	✓	–	✓

- Network evaluation is a relatively **new evaluation measure** of Explainability
- While **masking** was introduced in Deconvolution, it was **used to create explanations – not evaluate them**
- **More robust the network, better is its explanatory power, as measured by Network Evaluation**

AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.

Outline

Lecture 5: Evaluating Visual Explanations

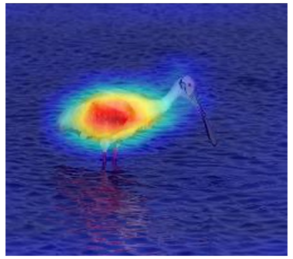
- Explanatory Evaluation Taxonomy
- Human Evaluation
    - Challenges
    - Methodology
- Application Evaluation
    - Methodology
    - Gaze Prediction
    - Pointing Game
    - Localization
- Network Evaluation
    - Intervention-based Evaluation
    - Masking
    - Progressive Pixel-wise masking
    - Progressive Structure-wise masking
- Challenges in Explanatory Evaluation
    - Human and Application Evaluation
    - Network Evaluation

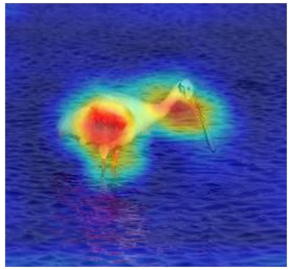
Challenges in Explanatory Evaluation

Challenges

Human Evaluation

Tasks : Humans directly evaluate explanations.





Which explanation is better for answering Why Spoonbill?

Application Evaluation

Tasks : Any task that requires humans-in-the-evaluation-loop without directly having humans measure explainability.

Which regions in the image are salient to the human visual system?

Network Evaluation

Tasks : Any intervention based on explanation techniques that does not require humans for evaluation.

Is this intervened image still a spoonbill?

Data Domain Knowledge

Network Domain Adaptation

AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.

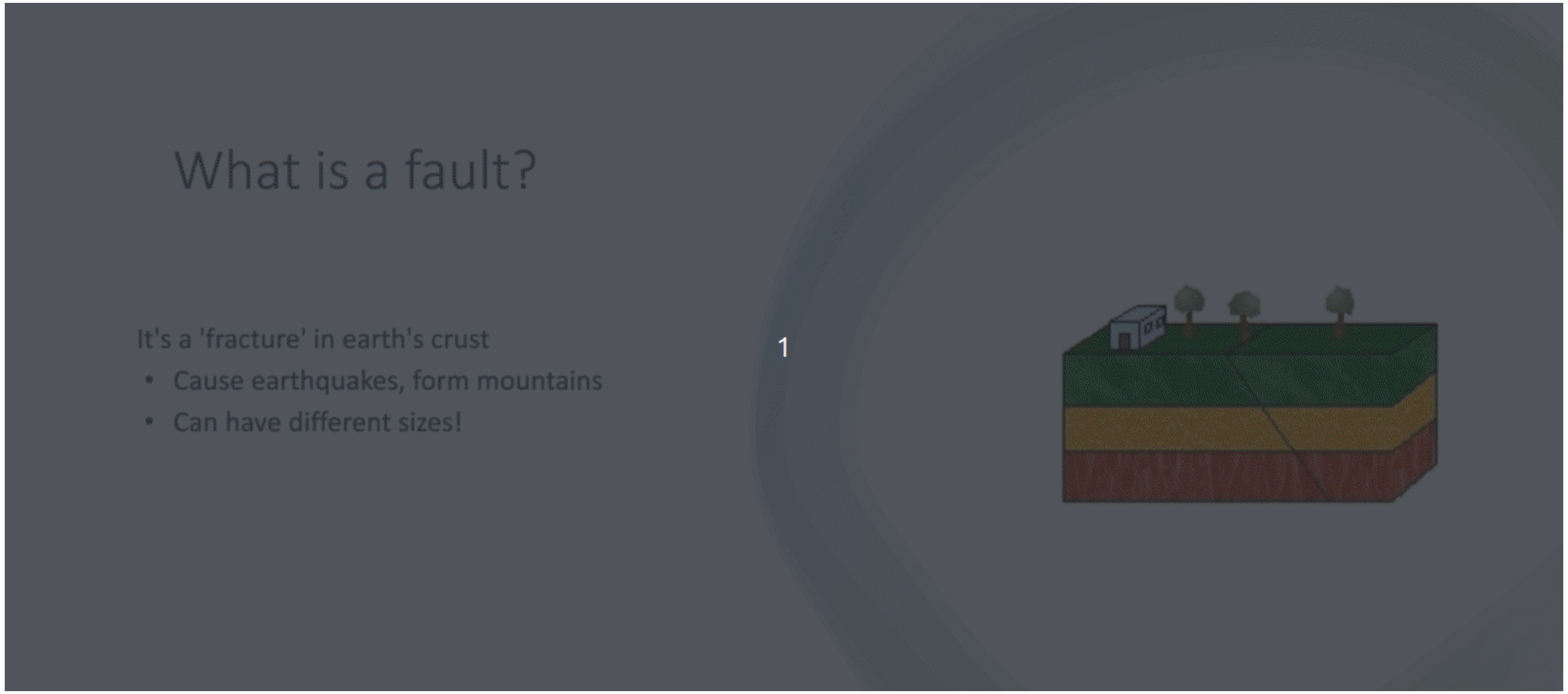
Challenges in Explanatory Evaluation

Case Study: Seismic Data

Amazon Mturk is used to obtain Fault annotations from Subsurface data

We provide an instructional video in the task website and inside the layout, with the following details:

- Fault definition
- Sample image and label meaning
- Platform usage
- Payment scheme



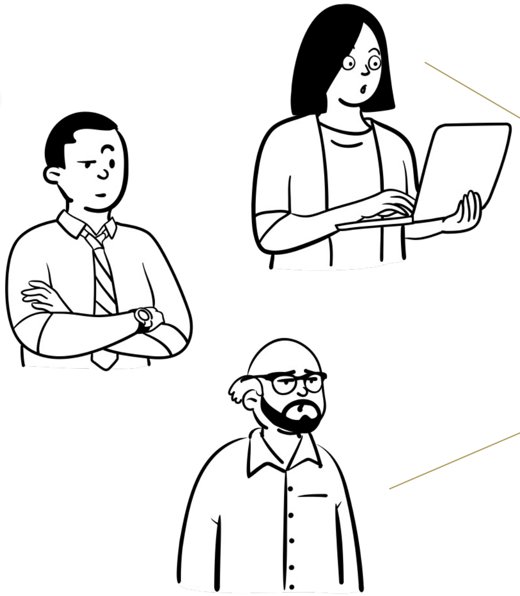
https://alregib.ece.gatech.edu/fun-ml-fault-uncertainty-for-machine-learning/

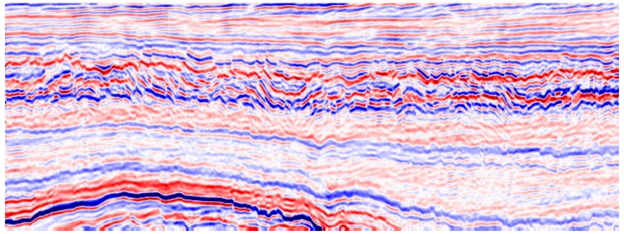
Challenges in Explanatory Evaluation

Case Study: Seismic Data

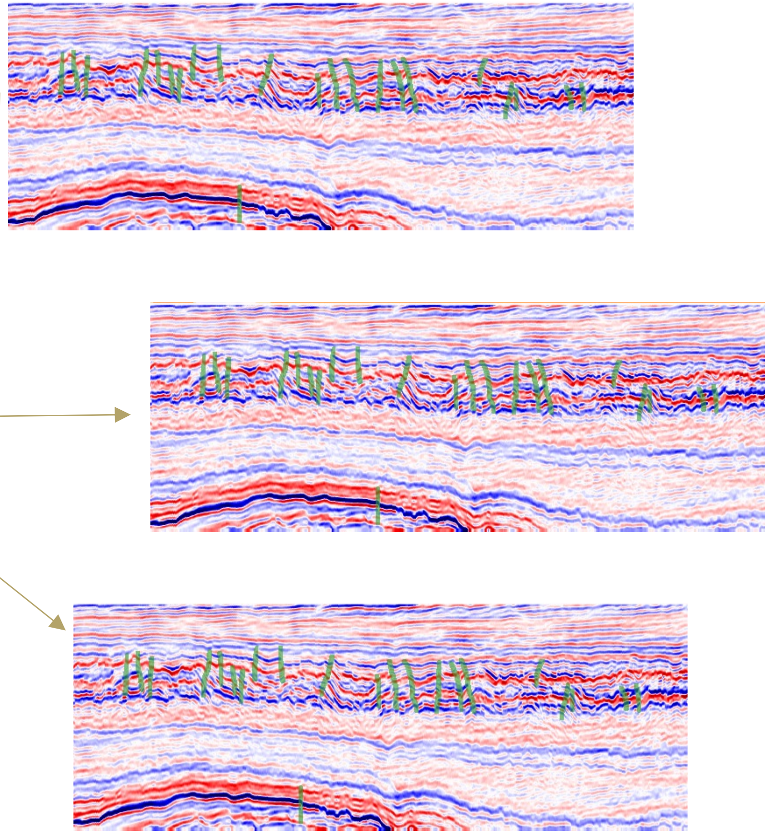
Amazon Mturk is used to obtain Fault annotations from Subsurface data

- Every annotator provides the same labels





Processed Subsurface Image



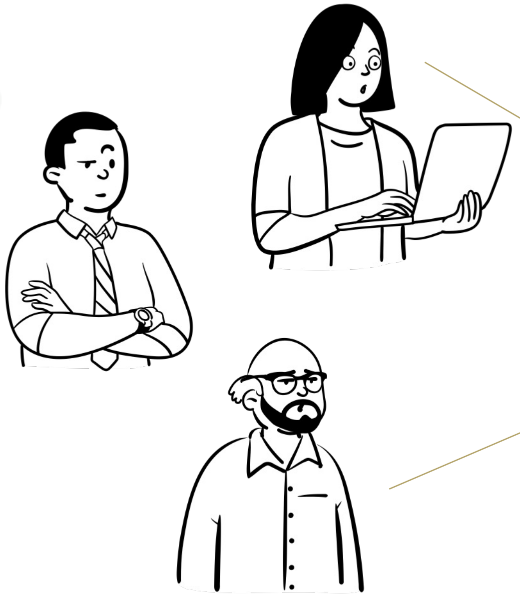
https://alregib.ece.gatech.edu/fun-ml-fault-uncertainty-for-machine-learning/

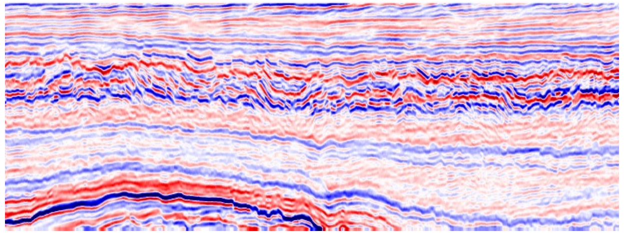
Challenges in Explanatory Evaluation

Case Study: Seismic Data

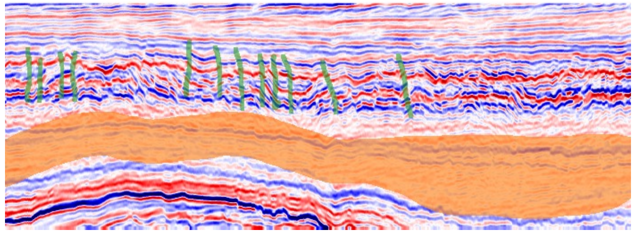
Amazon Mturk is used to obtain Fault annotations from Subsurface data

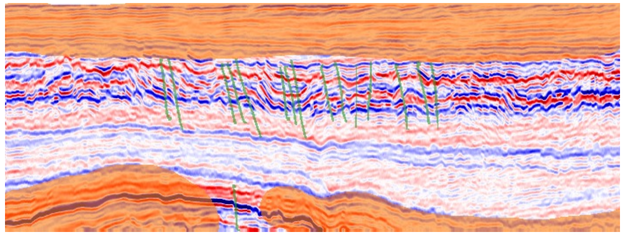
- Every annotator provides the same labels
- Large differences between annotations

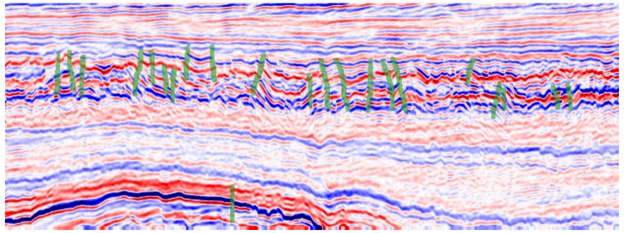




Processed Subsurface Image







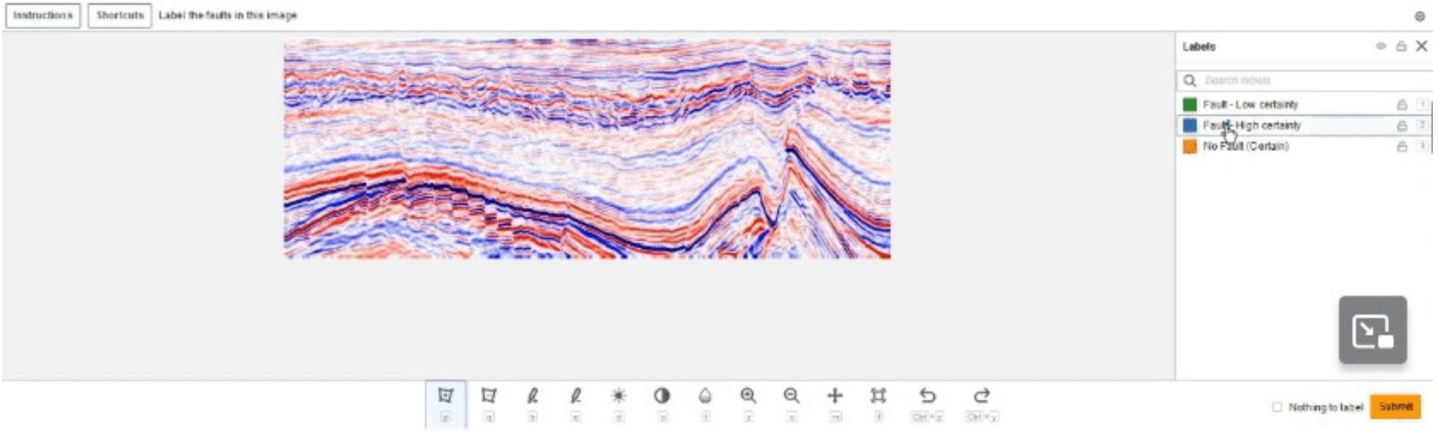
https://alregib.ece.gatech.edu/fun-ml-fault-uncertainty-for-machine-learning/

Challenges in Explanatory Evaluation

Case Study: Seismic Data

Annotation setup on Amazon MTurk

- 400 images, divided into 20 batches
- For each batch, 2 images are repeated 3 times for quality assessment
- 2 monetary bonuses:
    - Number of images, promotes full dataset completion
    - Consistency, promotes thorough labeling

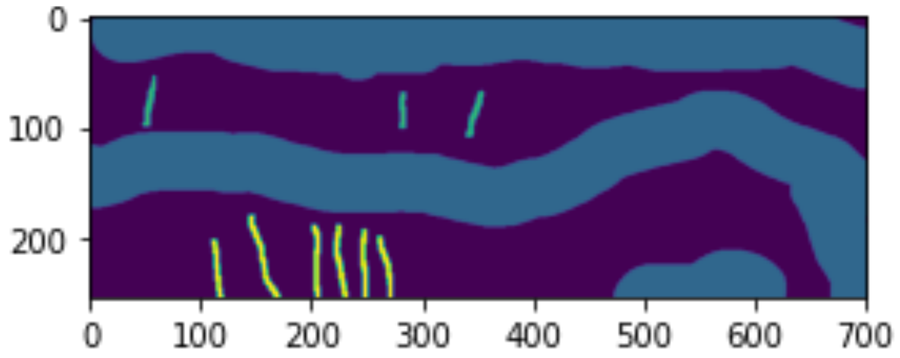


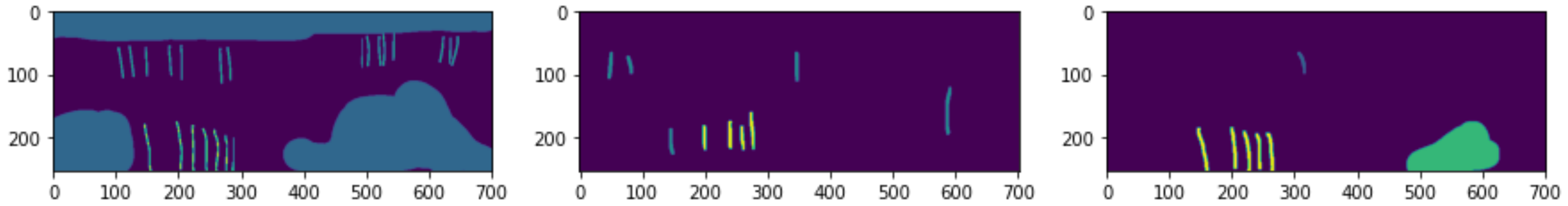
Challenges in Explanatory Evaluation

Case Study: Seismic Data

Disagreement between annotators for the same seismic section

Ground Truth Annotation



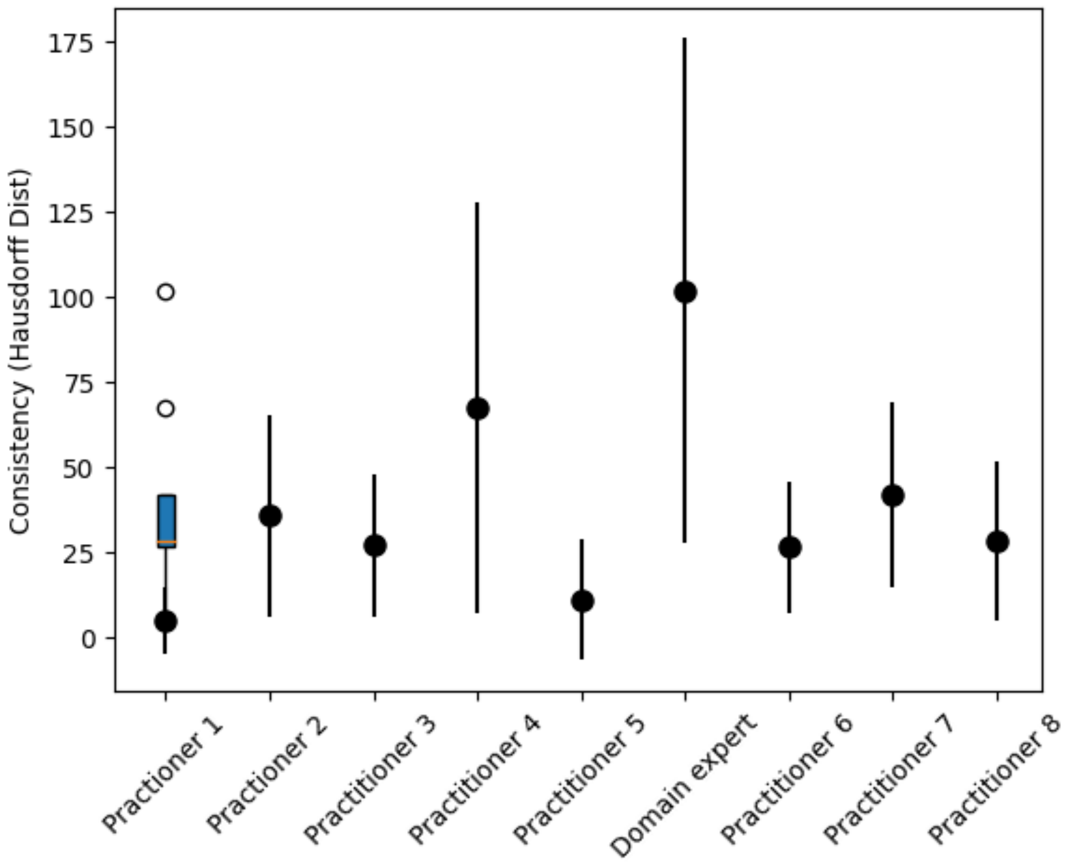


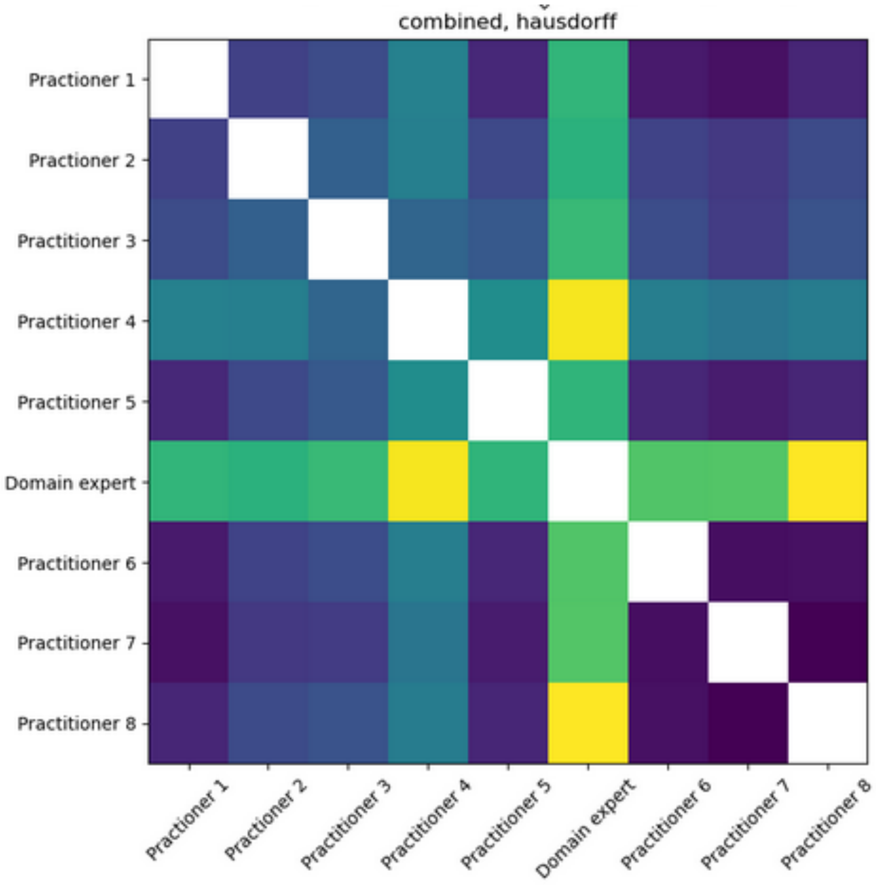
https://alregib.ece.gatech.edu/fun-ml-fault-uncertainty-for-machine-learning/

Challenges in Explanatory Evaluation

Case Study: Seismic Data

Consistency for the repeated images in the same batch



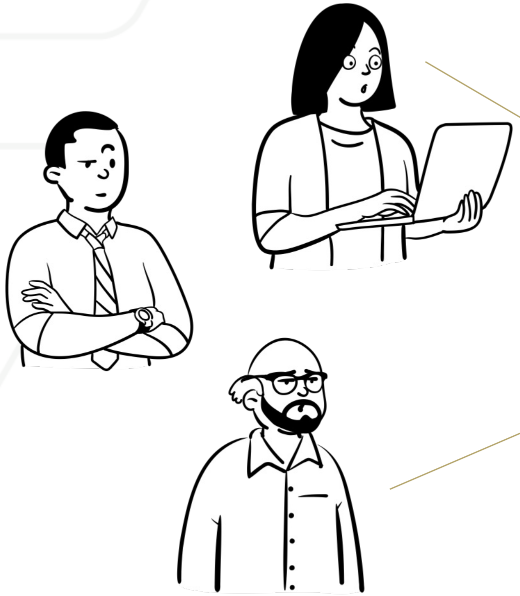


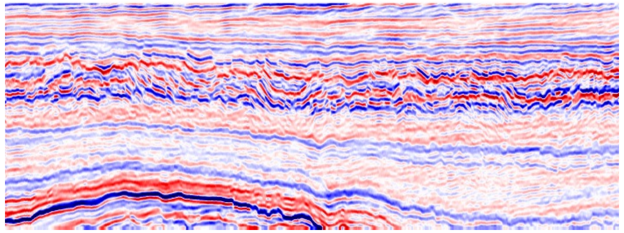
https://alregib.ece.gatech.edu/fun-ml-fault-uncertainty-for-machine-learning/

Challenges in Explanatory Evaluation

Case Study: Human Evaluation Summary

Humans are directly asked to evaluate explanatory techniques





Processed Subsurface Image

- **Human knowledge** plays a key role in Explainability
- **Experimental design logistics** include:
  - Setting up the platform with **no label bias**
  - Setting up **educational materials** for the non-domain experts (example: instructional video)
  - Setting up **intra-annotator consistency** and **inter-annotator disagreement** metrics
  - Setting up a **pay scale** to discourage quick and incorrect annotations

https://alregib.ece.gatech.edu/fun-ml-fault-uncertainty-for-machine-learning/

Challenges in Explanatory Evaluation

Challenges

Human Evaluation

Tasks : Humans directly evaluate explanations.

Which explanation is better for answering Why Spoonbill?

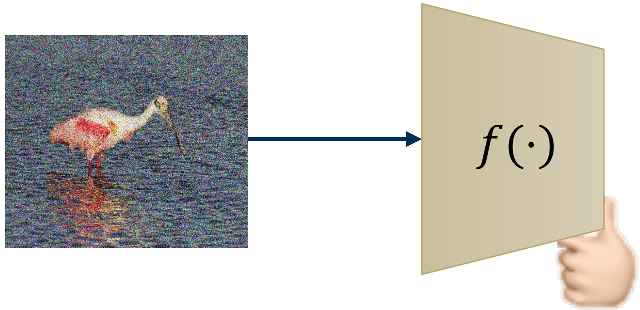
Application Evaluation

Tasks : Any task that requires humans-in-the-evaluation-loop without directly having humans measure explainability.

Which regions in the image are salient to the human visual system?

Network Evaluation

Tasks : Any intervention based on explanation techniques that does not require humans for evaluation.



Is this intervened image still a spoonbill?

Data Domain Knowledge

Network Domain Adaptation

AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.

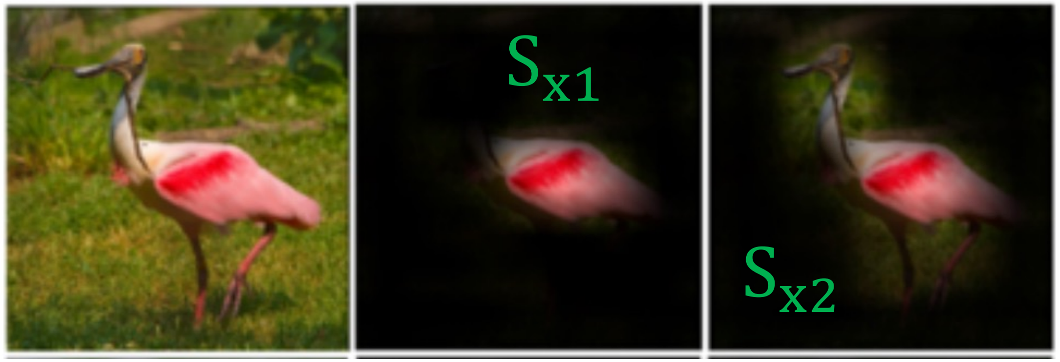
Challenges in Explanatory Evaluation

Challenges

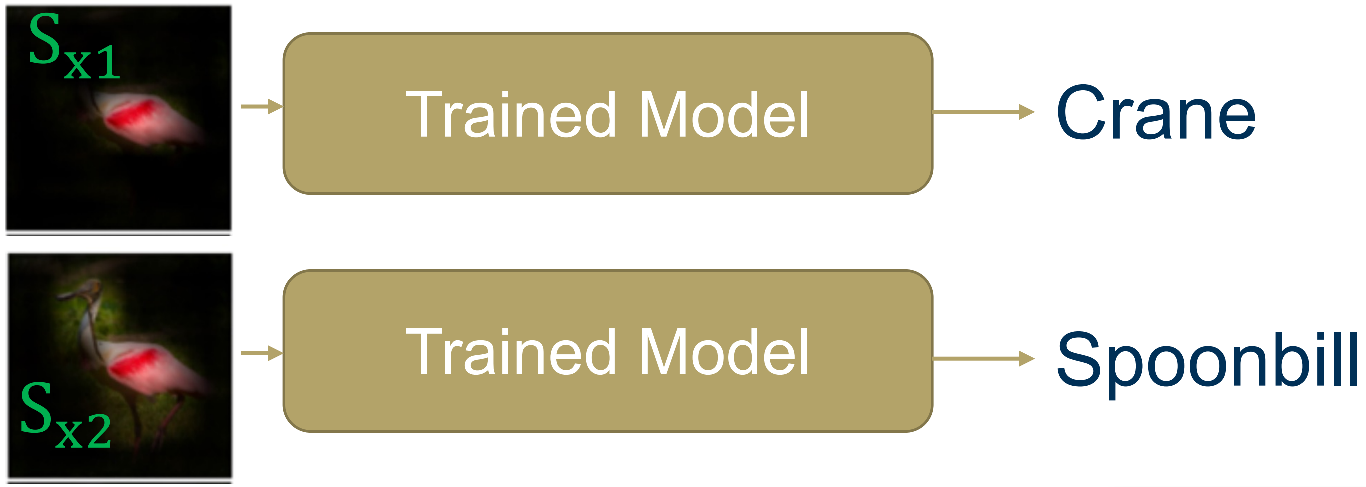
Interventions on data may push the data out of the domain of the trained network

In such cases, the prediction Y cannot be trusted

Y = Prediction
S_x = Explanation masked data



Analyzing Y under domain shift: Lecture 7



AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.

Takeaways

Takeaways from Lecture 5

- There are **no "one size fits all" explanations** and techniques
- **There are no "one evaluation fits all" for explanations and techniques**
- **Human evaluation** can be very **subjective and knowledge dependent**
    - **Large-scale** evaluation **removes subjective bias**
    - However, large-scale evaluation **cannot remove systemic bias** in the evaluation design
- **Application evaluation** can turn into **experimental design research** areas
- Network evaluation **provides objective assessment of subjective explanations**
    - Explanation masking mimics deletion but encourages large explanations
    - Pixel-wise insertion and deletion encourages unrealistic explanations
    - Structure-wise masking and insertion-deletion provides a compromise

References

Lecture 5: Evaluating Explanations

- AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.
- 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.
- Y. Sun, M. Prabhushankar, and G. AlRegib, ”Implicit Saliency in Deep Neural Networks,” in IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, Oct. 2020.
- Das, Abhishek, et al. "Human attention in visual question answering: Do humans and deep networks look at the same regions?." Computer Vision and Image Understanding 163 (2017): 90-100.
- Chowdhury, Prithwijit, Mohit Prabhushankar, and Ghassan AlRegib. "Explaining Explainers: Necessity and Sufficiency in Tabular Data." *NeurIPS 2023 Second Table Representation Learning Workshop*. 2023.
- Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018.
- Vitali Petsiuk, Abir Das, and Kate Saenko, “Rise: Randomized input sampling for explanation of black-box models,” arXiv preprint arXiv:1806.07421, 2018.
- Prabhushankar, Mohit, and Ghassan AlRegib. "Extracting causal visual features for limited label classification." *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021.
- https://alregib.ece.gatech.edu/fun-ml-fault-uncertainty-for-machine-learning/