Visual Explainability in Machine Learning

Lecture 2: Basics of Visual Explainability

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

IEEE Signal Processing Society
CELEBRATING 75 YEARS

OLIVES @GeorgiaTech

Georgia Tech

Short Course Materials

Accessible Online

SCAN ME

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

{alregib, mohit.p}@gatech.edu

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

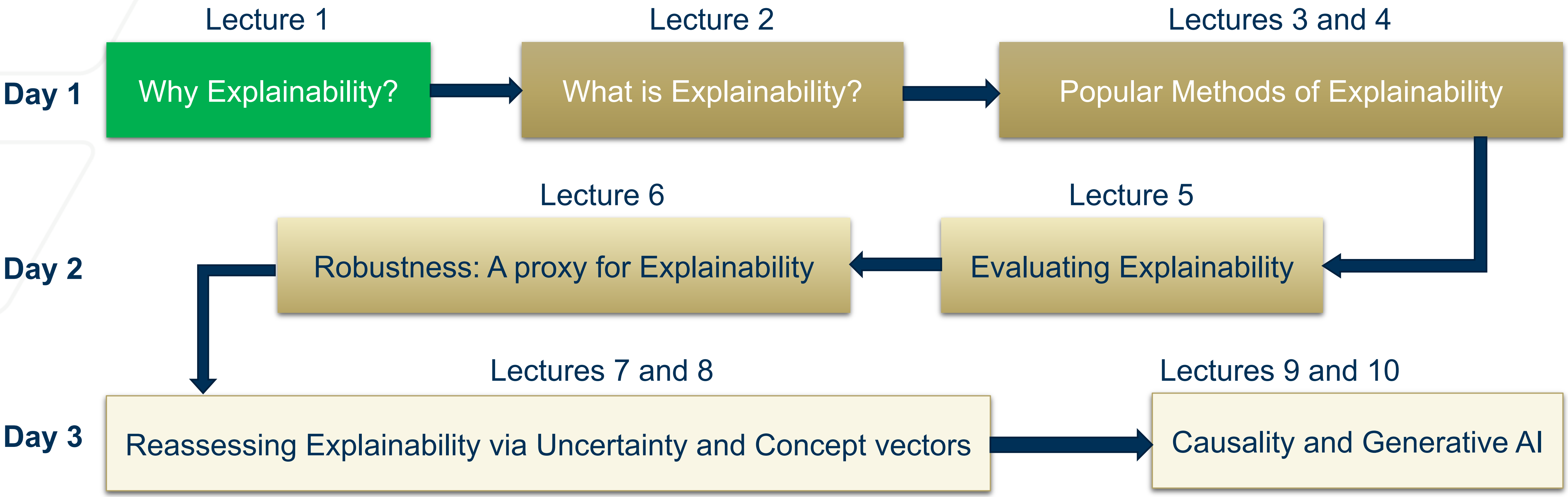
Georgia Institute of Technology, Atlanta, USA

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

Short Course

Course Outline

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



Outline

Lecture 2: Basics of Visual Explainability

- Explanations
    - Interpretability vs Explainability
- Categorization of Explanations
- Method-based Categorization
    - Implicit vs Explicit
    - Interventionist vs Non-interventionist
    - White-box vs Black-box
    - Gradient-based vs Non gradient-based
- Human-centric Categorization
    - Indirect
    - Direct
    - Targeted
- Properties-based Categorization
    - Necessity
    - Sufficiency
    - Importance
- Reasoning-based Categorization
    - Deductive
    - Inductive
    - Abductive
- Mathematical Formulations
    - Probabilistic
        - Complete Explanations

Outline

Lecture 2: Basics of Visual Explainability

- Explanations
  - Interpretability vs Explainability
- Categorization of Explanations
- Method-based Categorization
  - Implicit vs Explicit
  - Interventionist vs Non-interventionist
  - White-box vs Black-box
  - Gradient-based vs Non gradient-based
- Human-centric Categorization
  - Indirect
  - Direct
  - Targeted
- Properties-based Categorization
  - Necessity
  - Sufficiency
  - Importance
- Reasoning-based Categorization
  - Deductive
  - Inductive
  - Abductive
- Mathematical Formulations
  - Probabilistic
    - Complete Explanations

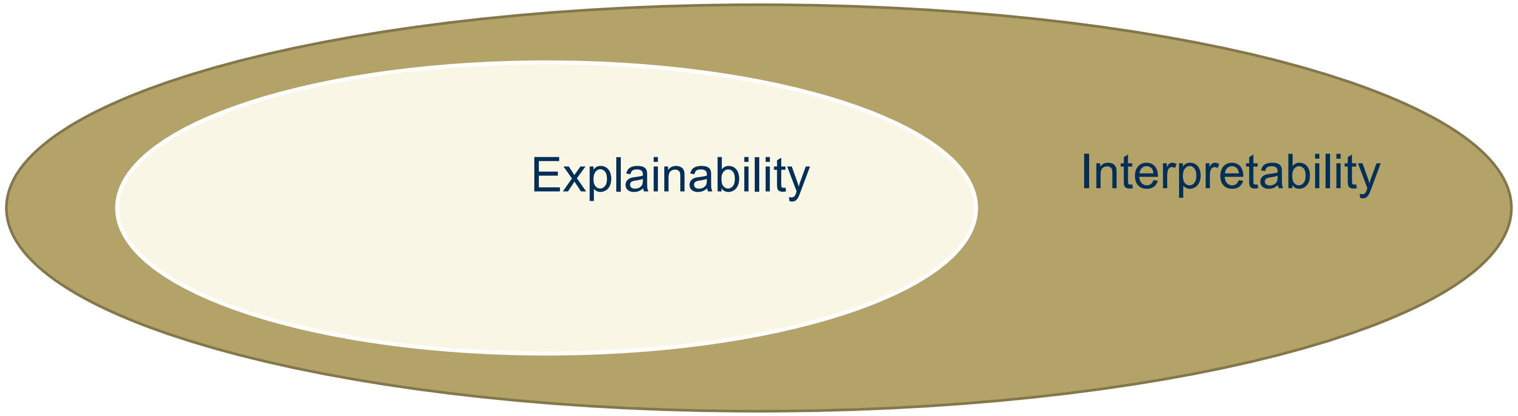
Explanations

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



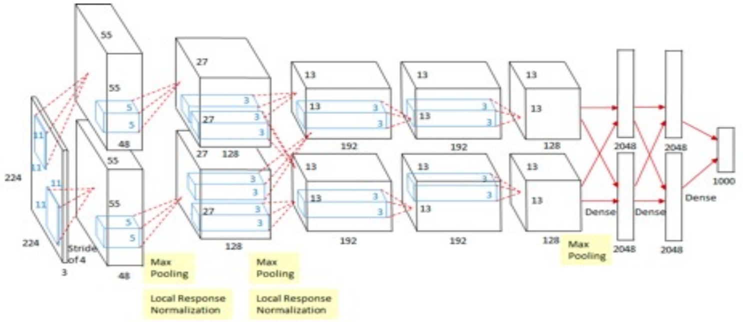
Explanations

Why does Explainability matter?

Explainability establishes trust in deep learning systems by developing *transparent* models that can explain *why they predict what they predict* to humans

Algorithm

Data



Output

class scores

Explainability is useful in*:*

- Medical: help doctors diagnose
- Seismic: help interpreters label seismic data
- Autonomous Systems: build appropriate trust and confidence

Deep models act as algorithms that take data and output something **without** being able to **explain** their methodology

Outline

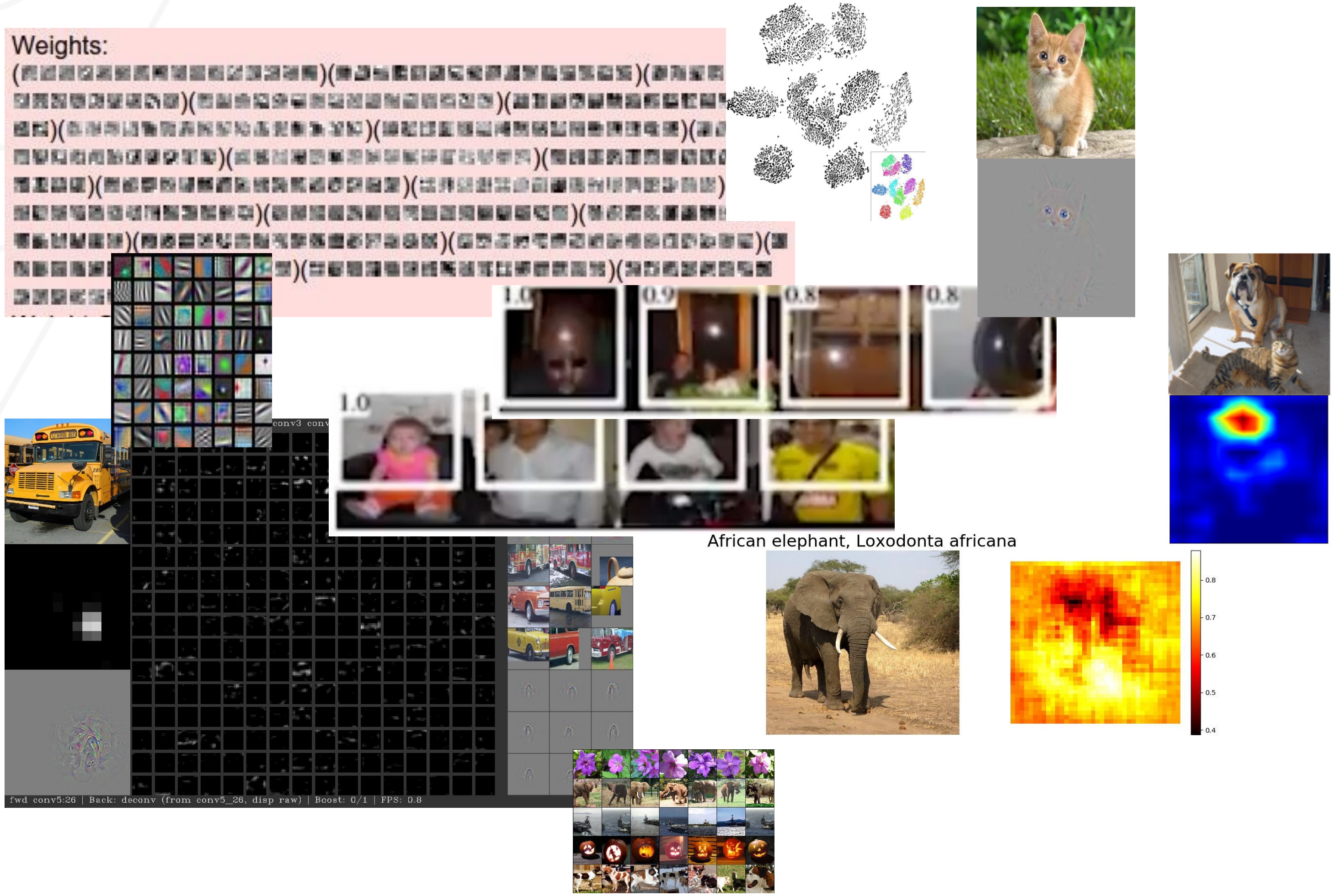
Lecture 2: Basics of Visual Explainability

- Explanations
  - Interpretability vs Explainability
- Categorization of Explanations
- Method-based Categorization
  - Implicit vs Explicit
  - Interventionist vs Non-interventionist
  - White-box vs Black-box
  - Gradient-based vs Non gradient-based
- Human-centric Categorization
  - Indirect
  - Direct
  - Targeted
- Properties-based Categorization
  - Necessity
  - Sufficiency
  - Importance
- Reasoning-based Categorization
  - Deductive
  - Inductive
  - Abductive
- Mathematical Formulations
  - Probabilistic
    - Complete Explanations

Explanations

Method-centric categorization of Explanations

Explanations can be characterized based on the techniques they employ



Four categorizations[1] of explanations based on methods:

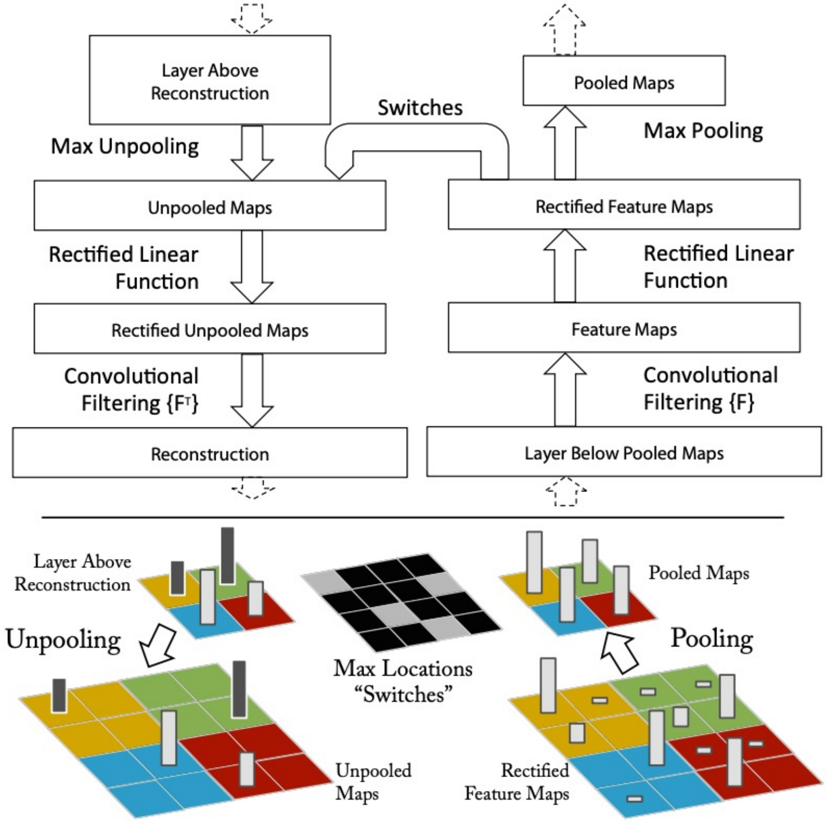
1. Implicit vs Explicit
2. Interventionist vs non-interventionist
3. Black-box vs White-box
4. Gradient-based vs activation-based

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

Explanation Categorizations

Implicit vs Explicit Explanations

Explanations that require architectural change in the models are explicit



Left: Deconvolution network, Right: Convolutional encoder

- **DeconvNet**: An additional deconvolution network is added to map features back into input space[1]
- Addition/change of network architecture creates explicit explanations
- **Implicit explanations justify decisions without additional network components**

Other implicit and explicit explanations are detailed in [2]

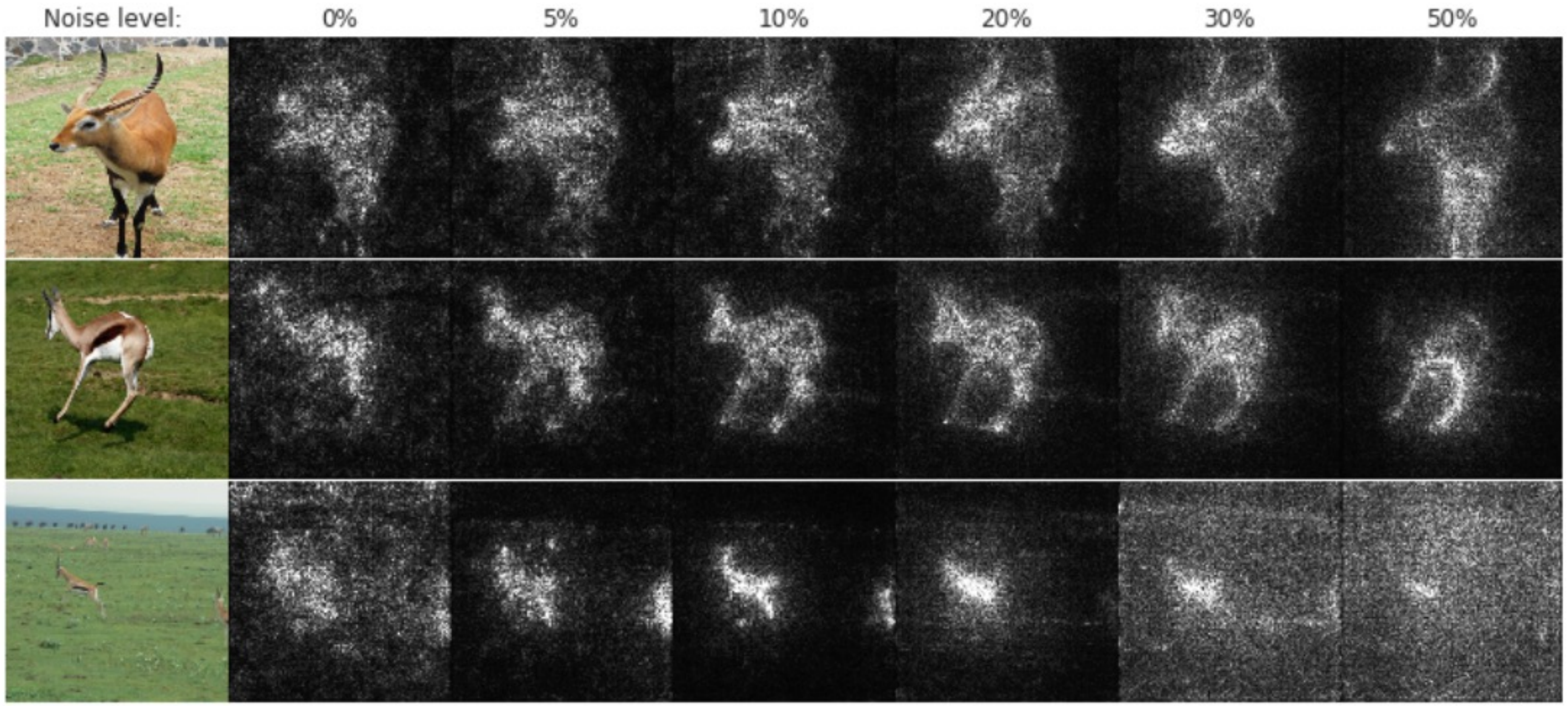
[1] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in European conference on computer vision. Springer, 2014, pp. 818–833

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

Explanation Categorizations

Interventionist vs Non-interventionist Explanations

Explanations that require change in the inputs are interventionist



Gradients from different noise levels from [1]

- **SmoothGrad**: Noise is added to the same input multiple times and the gradients of the outputs are averaged across the pixel space[1]. This is an interventionist explanation
- **Non-interventionist explanations justify decisions without changes to inputs**

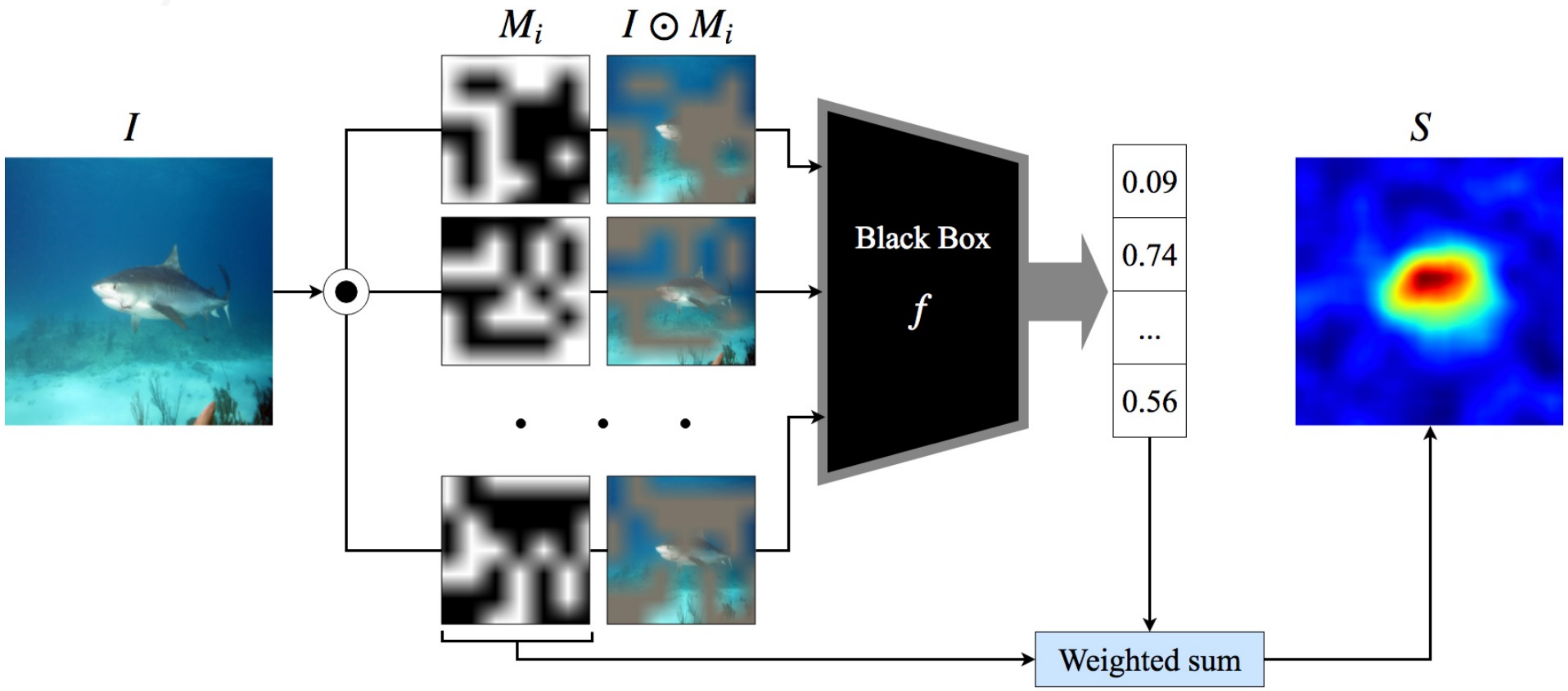
Other explicit explanations are detailed in [2]

[1] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg, “Smoothgrad: removing noise by adding noise,” arXiv preprint arXiv:1706.03825, 2017.
[2] AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.

Explanation Categorizations

White-box vs Black-box Explanations

Black-box explanations do not assume access to activations, gradients or any network parameters



Random input masking and output logit-weighted sum from [1].

- **RISE**: Inputs are masked randomly and a weighted sum of their output logits are calculated[1]. This is a black-box explanation
- **White-box explanations assume access to network parameters**

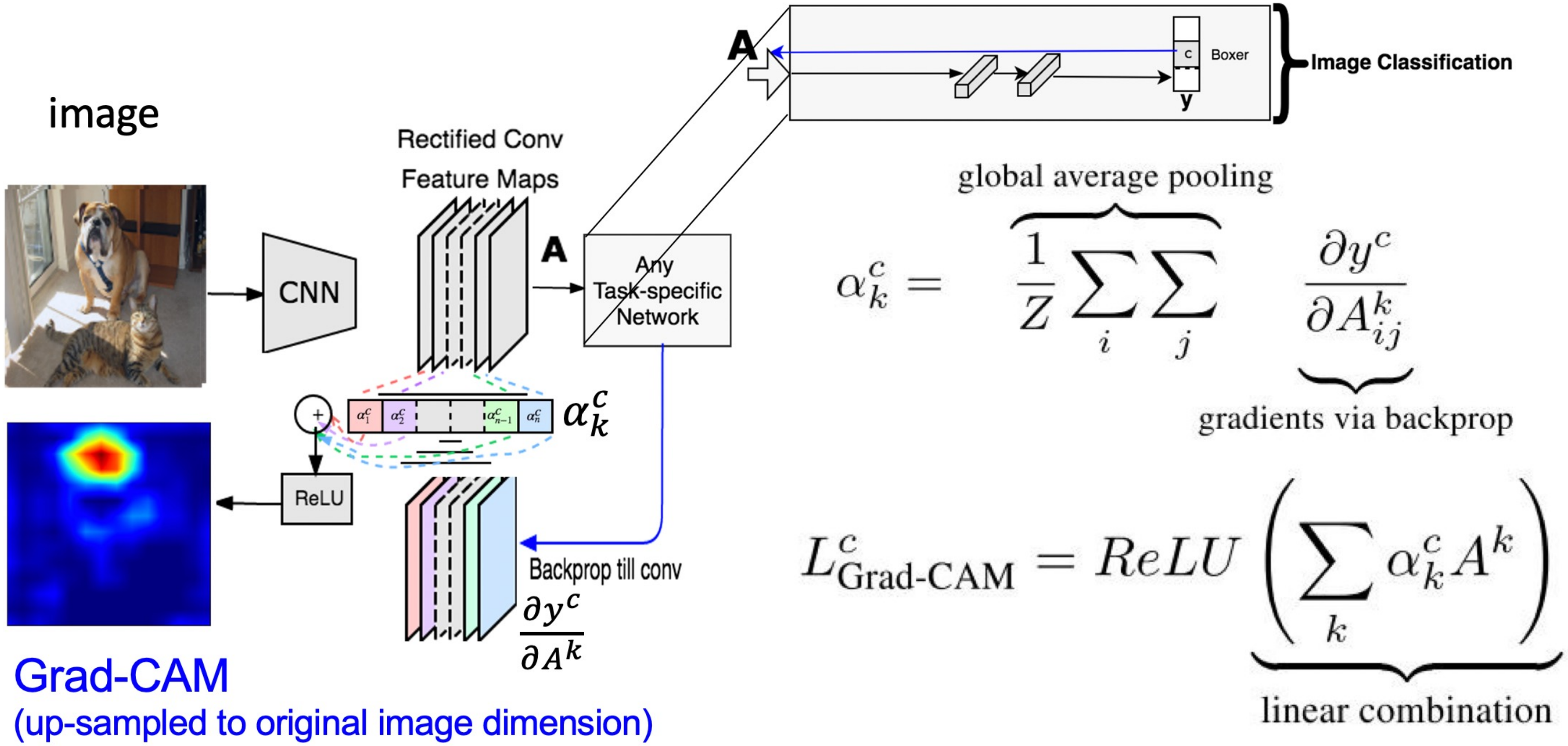
Other black and white-box explanations are detailed in [2]

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

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

Gradient-based vs non gradient-based Explanations

Gradient-based explanations use gradients as features to obtain explanations



$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

$$L^c_{\text{Grad-CAM}} = ReLU \underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$

Utility of gradients as weights from [1]

- **GradCAM**: Backpropagated logit gradients weigh the activations from convolution layers[1]
- **Non-gradient explanations only utilize forward propagation parameters and do not backpropagate**

Other gradient and non gradient-based explanations are detailed in [2]

[1] 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.
[2] AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.

Explanation Categorizations

All Method-based categorizations

	Technique Categorization							
Methods	Implicit	Explicit	Black Box	White Box	Intervention	Nonintervention	Gradient Based	Nongradient Based
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]	✓	—	—	✓	—	✓	✓	—

All explanatory techniques can be described based on their method choices

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

Explanation Categorizations

All Method-based categorizations

Each categorization has its pros and cons

- Architectural changes in the explicit explanations may change the original decisions or its confidence and uncertainty
- Implicit explanations are only post-hoc and cannot be used to bootstrap the network
- Interventions may not be possible in certain scenarios like biomedical images
- Network parameters may not be available in white-box explanations
- Black-box explanations are generally computationally expensive
- Gradient-based explanations are sensitive to noise and input challenges
- Activation-based explanations and deconvolution nets generally reconstruct the image and are not true explanations

No explanations are one size fits all!

Outline

Lecture 2: Basics of Visual Explainability

- Explanations
  - Interpretability vs Explainability
- Categorization of Explanations
- Method-based Categorization
  - Implicit vs Explicit
  - Interventionist vs Non-interventionist
  - White-box vs Black-box
  - Gradient-based vs Non gradient-based
- Human-centric Categorization
  - Indirect
  - Direct
  - Targeted
- Properties-based Categorization
  - Necessity
  - Sufficiency
  - Importance
- Reasoning-based Categorization
  - Deductive
  - Inductive
  - Abductive
- Mathematical Formulations
  - Probabilistic
    - Complete Explanations

Explanations

Human-centric categorization of Explanations

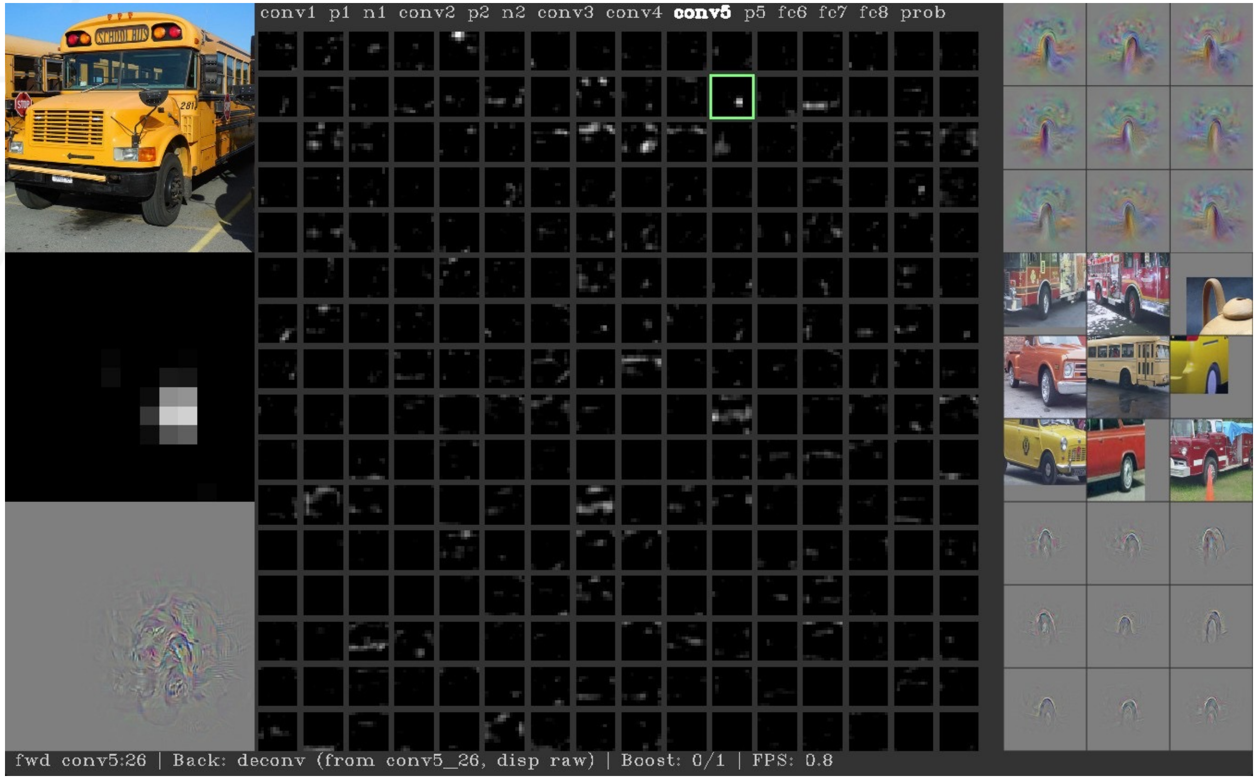
Explanations can be characterized based on the knowledge of the audience they cater to

Three categorizations[1] of explanations based on audience:

1. Direct Explanations
2. Indirect Explanations
3. Targeted 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.

Indirect explanations visually analyze network parameters and features and indirectly explain the output



The filter in conv 5 layer is activated when it sees a wheel[1]

- **Require network knowledge from the humans interpreting the explanations**
- Example of Indirect Explanations: Visualizing hidden layer representations and finding the concepts that maximally activate patches (in this image, wheels activate the filter in green box)

More details regrading indirect explanations are detailed in [2]

[1] Yosinski et al, “Understanding Neural Networks Through Deep Visualization”, ICML DL Workshop 2014.
[2] AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.

Explanation Categorization

Direct Explanations

Direct explanations highlight all regions in an image that lead to a decision

Input image predicted as bullmastiff

GradCAM explanation[1]

- **No network knowledge is required from the humans interpreting these explanations.**
- **No knowledge about the classes or data is required**

- Example of Direct Explanations: Visualizing the face of the dog to explain the prediction of bullmastiff

More details regrading direct explanations are detailed in [2]

[1] 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.
[2] AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.

Explanation Categorization

Targeted Explanations

Targeted explanations highlight contextually relevant regions in an image

Input image predicted as spoonbill

ContrastCAM explanation for '*Why Spoonbill, rather than Flamingo?*'[1]

- **No network knowledge is required from the humans interpreting these explanations.**
- **Knowledge about the classes or data is required by the humans seeking explanations.**

- Example of Targeted Explanations: Visualizing the lack of S-shaped neck in the Spoonbill to answer why it is not a Flamingo[1]

More details regrading targeted explanations are detailed in [2]

[1] Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.

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

Explanation Categorizations

All Human-centric categorizations

Methods	Definition: Indirect	Definition: Direct	Definition: Targeted
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]	–	–	✓

- The rows are ordered chronologically
- Human-centric explanation categorization provides an **evolution of Explainability research**

- **Initial goal of Explainability**: To indirectly understand decisions to **facilitate understanding the network**

- **Subsequent goal of Explainability:** To facilitate direct and targeted **understanding** of decisions among **all stakeholders**

Note: Many of the listed targeted explanations can also act as direct explanations with slight modifications

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

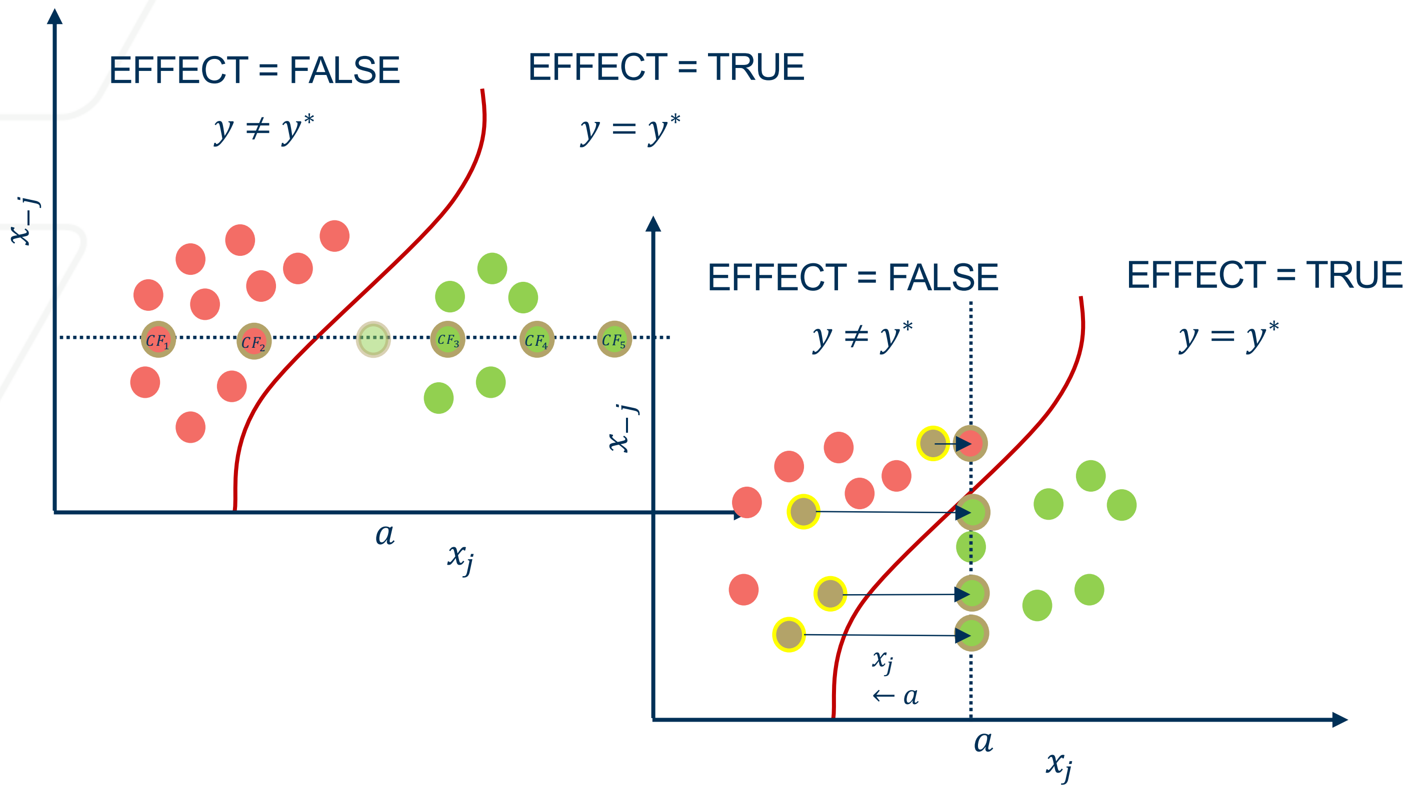
Lecture 2: Basics of Visual Explainability

- Explanations
  - Interpretability vs Explainability
- Categorization of Explanations
- Method-based Categorization
  - Implicit vs Explicit
  - Interventionist vs Non-interventionist
  - White-box vs Black-box
  - Gradient-based vs Non gradient-based
- Human-centric Categorization
  - Indirect
  - Direct
  - Targeted
- Properties-based Categorization
  - Necessity
  - Sufficiency
  - Importance
- Reasoning-based Categorization
  - Deductive
  - Inductive
  - Abductive
- Mathematical Formulations
  - Probabilistic
    - Complete Explanations

Explanations

Property-based categorization of Explanations

Explanations can be characterized based on the property that they suffice



Two categorizations[1] of explanations based on properties:

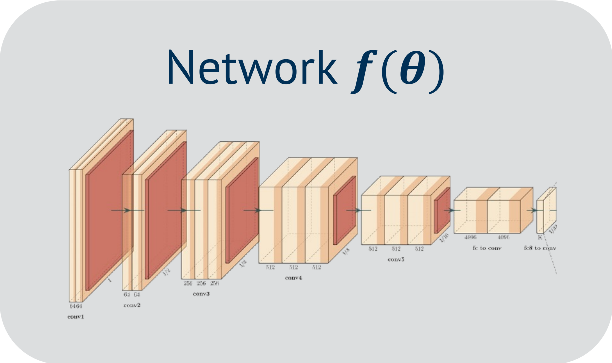
1. Necessity
2. Sufficiency

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

Explanation Categorizations

'Necessary' Property of Explanations

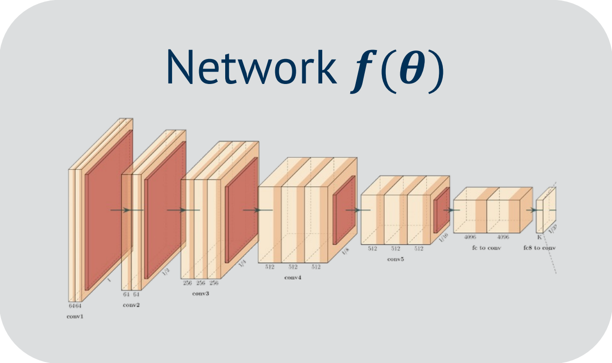
Features are said to be necessary if their deletion causes a misclassification[1]



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

Note: This is a approximation of a more formal definition in [1]

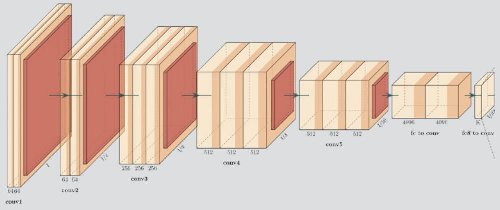
[1] Chowdhury, Prithwijit, Mohit Prabhushankar, and Ghassan AlRegib. "Explaining Explainers: Necessity and Sufficiency in Tabular Data." *NeurIPS 2023 Second Table Representation Learning Workshop*. 2023.

Explanation Categorizations

'Sufficiency' Property of Explanations

Features are said to be sufficient if their addition causes the required classification[1]

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

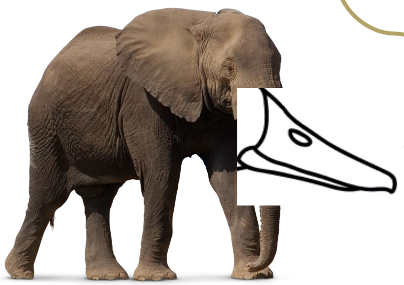


Beak
Neck
Legs
Feathers
Water
Grass
Teeth
.
.

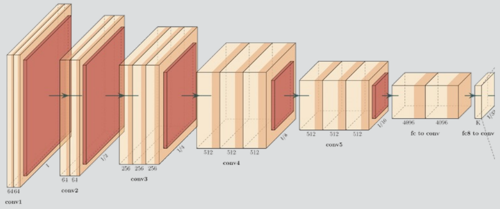
Features $\mathcal{T}_P$

P is Spoonbill

Adding a beak



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



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

Note: This is a approximation of a more formal definition in [1]

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

Explanations

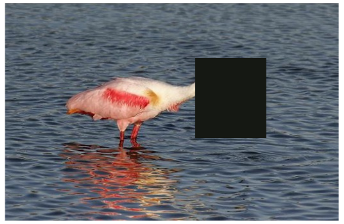
Property-based categorization of Explanations

Explanations can be evaluated based on either necessity or sufficiency properties

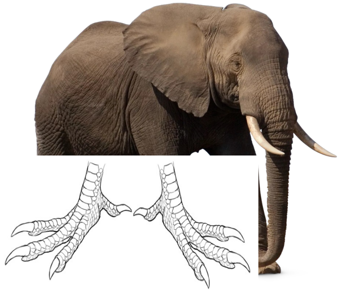
Necessity according to Explanation 1

Which explanation is better?

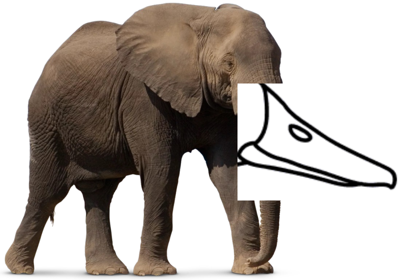
Necessity according to Explanation 2



Sufficiency according to Explanation 1



Sufficiency according to Explanation 2



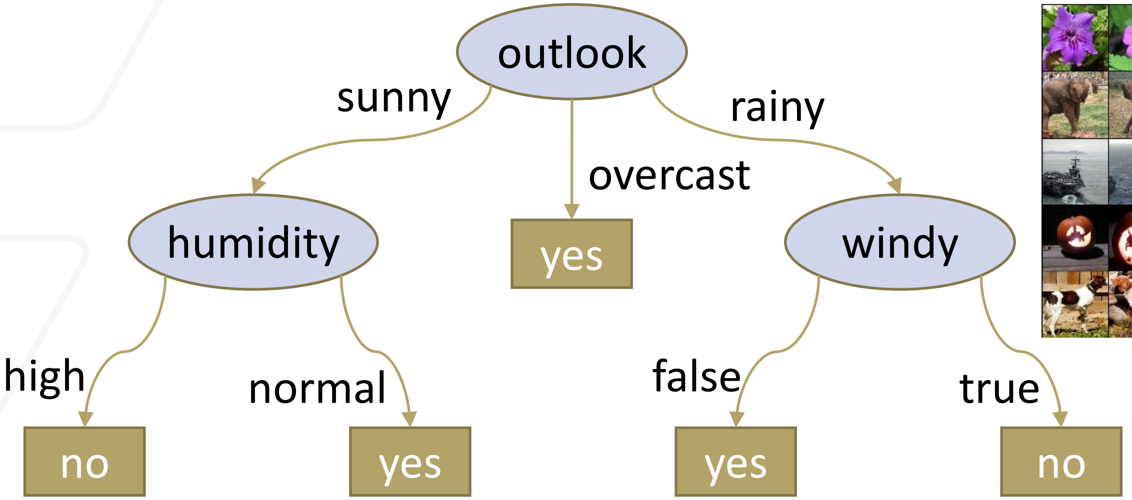
In Lecture 5, we will detail objective approximations of necessary and sufficient conditions for evaluation

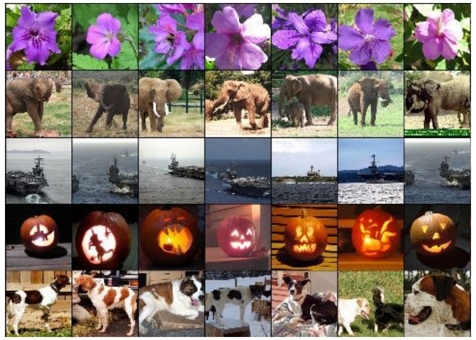
AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.

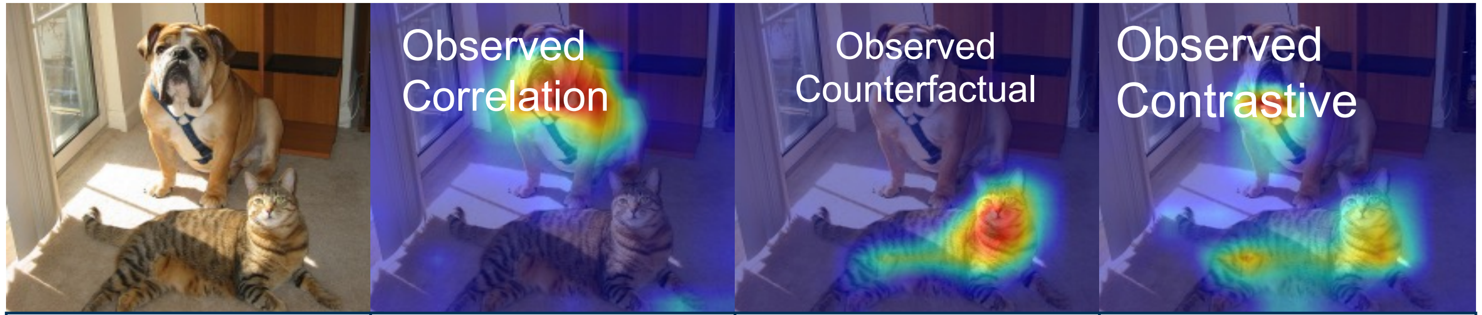
Explanations

Reasoning-based categorization of Explanations

Explanations can be characterized based on the reasoning paradigm they are derived from







Bullmastiff	Why Bullmastiff?	What if Bullmastiff was not in the image?	Why Bullmastiff, rather than a Boxer?

Three categorizations[1] of explanations based on reasoning:

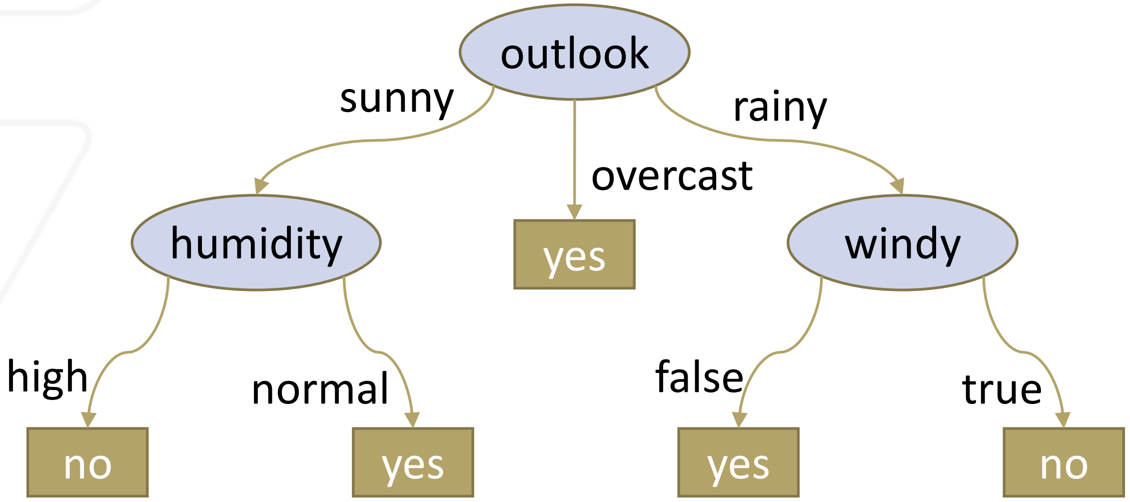
1. Deductive Reasoning
2. Inductive Reasoning
3. Abductive Reasoning

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

Explanations

Deductive Reasoning-based categorization of Explanations

Deductive Explanations are logic-based reasoning paradigms



Final decision tree computed based on data in table

- **Comprises logic-based formalization of Explainability**
- **Provides rigorous explanations that are cardinal and *np-complete***

- However, applicability to large scale neural networks is an ongoing area of research

Extensions to visual data in neural networks are presented in [2]

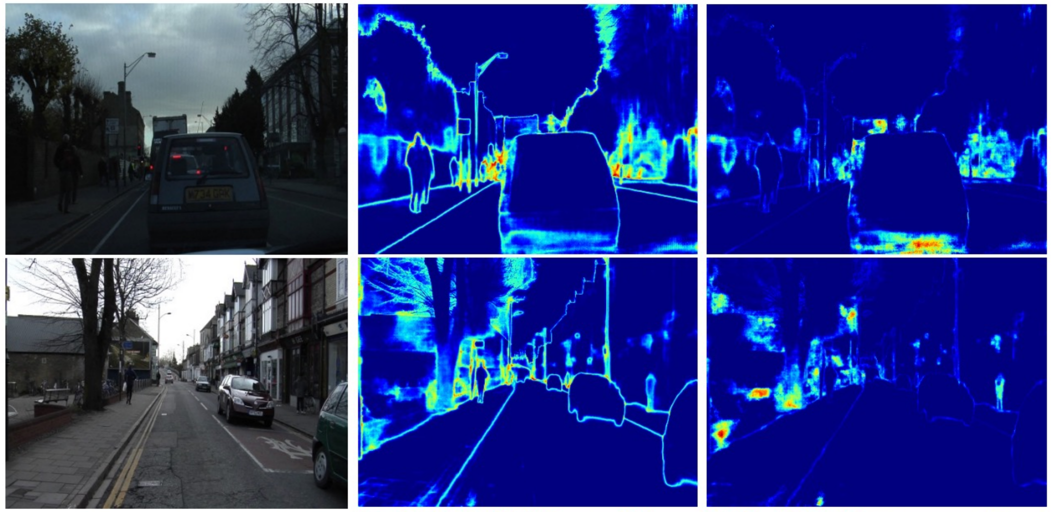
[1] Marques-Silva, Joao. "Logic-based explainability in machine learning." *Reasoning Web. Causality, Explanations and Declarative Knowledge: 18th International Summer School 2022, Berlin, Germany, September 27–30, 2022, Tutorial Lectures*. Cham: Springer Nature Switzerland, 2023. 24-104.
[2] Abduction-Based Explanations for Machine Learning Models

Inductive Reasoning-based categorization of Explanations

Inductive Explanations draw *seen* examples to explain *unseen* data

Example 1

- **Example 1:** Nearest neighbors (of the current test data) from the training data are shown to explain the classification of the test data[1]
- **Example 2:** Visualizing uncertainty in data and models to explain what the network does not know[2]



Example 2: Data uncertainty Model uncertainty

[1] Material adapted from : Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

[2] Kendall, Alex, and Yarin Gal. "What uncertainties do we need in bayesian deep learning for computer vision?." *Advances in neural information processing systems* 30 (2017).

Explanations

Abductive Reasoning-based categorization of Explanations

Abductive Explanations are *post-hoc:* They justify a hypothesis or a prediction

Bullmastiff	Why Bullmastiff?	What if Bullmastiff was not in the image?	Why Bullmastiff, rather than a Boxer?

Abductive questions and their visual explanations[1]

- **Deductive and Inductive** explanations are tied to the **decision-making process**
- **Abductive** explanations **justify** an already made decision or any other hypotheses
- [1] poses **abductive** explanations as **answers** to contextual and relevant **questions**

[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.

Outline

Lecture 2: Basics of Visual Explainability

- Explanations
  - Interpretability vs Explainability
- Categorization of Explanations
- Method-based Categorization
  - Implicit vs Explicit
  - Interventionist vs Non-interventionist
  - White-box vs Black-box
  - Gradient-based vs Non gradient-based
- Human-centric Categorization
  - Indirect
  - Direct
  - Targeted
- Properties-based Categorization
  - Necessity
  - Sufficiency
  - Importance
- Reasoning-based Categorization
  - Deductive
  - Inductive
  - Abductive
- Mathematical Formulations
  - Probabilistic
    - Complete Explanations

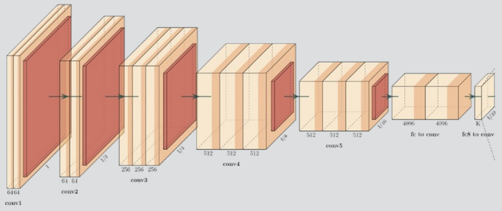
Mathematical Formulation

Probabilistic Interpretation of Explanations

Explanations are probabilities conditioned on features

Let $\mathcal{T}$ be the set of all features learned by a trained network

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



Beak
Neck
Legs
Feathers
Water
Grass
Teeth
.
.

Features $\mathcal{T}_P$

P is Spoonbill

Goal of any explanation $\mathcal{M}(\cdot)$: Find the set of features $\mathcal{T}_P$ that lead to a decision P

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

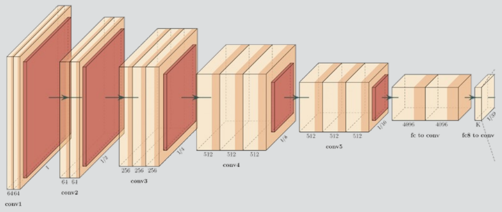
Mathematical Formulation

Probabilistic Interpretation of Explanations

Explanations are probabilities conditioned on features

Let $\mathcal{T}$ be the set of all features learned by a trained network

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



Beak
Neck
Legs
Feathers
Water
Grass
Teeth
.
.

Features $\mathcal{T}_P$

P is Spoonbill

Goal of any explanation $\mathcal{M}(\cdot)$: Find the set of features $\mathcal{T}_P$ that lead to a decision P

~~Causal Explanation,~~ $\cancel{\mathcal{M}(\cdot) = \mathbb{P}(P|\mathcal{T}_P)}$

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

Mathematical Formulation

Probabilistic Interpretation of Explanations

Explanations are probabilities conditioned on features

Let $\mathcal{T}$ be the set of all features learned by a trained network

Explanations maximize the probability of selecting a combination of features $\cup_{i=1}^{P} \mathcal{T}_i$ given that there is already a decision P:

$$\mathcal{M}(\cdot) = \mathbb{P}(\cup_{i=1}^{P} \mathcal{T}_i | P)$$

$$\left.\begin{array}{l} \text{Beak} \\ \text{Neck} \\ \text{Legs} \\ \text{Feathers} \\ \text{Water} \\ \text{Grass} \\ \text{Teeth} \\ \vdots \end{array}\right\} \text{Features } \mathcal{T}_P$$

P is Spoonbill

Goal of any explanation $\mathcal{M}(\cdot)$: Find the set of features $\mathcal{T}_P$ that lead to a decision P

~~Causal Explanation,~~ $\cancel{\mathcal{M}(\cdot) = \mathbb{P}(P|\mathcal{T}_P)}$

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

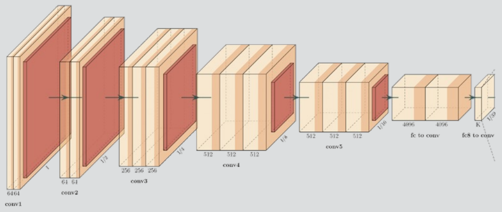
Mathematical Formulation

Probabilistic Interpretation of Explanations

Explanations are probabilities conditioned on features

Let $\mathcal{T}$ be the set of all features learned by a trained network

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



Beak
Neck
Legs
Feathers
Water
Grass
Teeth
.
.

Features $\mathcal{T}_{P,Q}$

P is Spoonbill,
Q is Flamingo

Goal of contrastive technique $\mathcal{M}(\cdot)$: Find the set of features $\mathcal{T}_{P,Q}$ that lead to a decision P but not to Q

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

Mathematical Formulation

Probabilistic Interpretation of Explanations

Explanations are probabilities conditioned on features

Let $\mathcal{T}$ be the set of all features learned by a trained network

Explanations maximize the probability of selecting a combination of features $\cup_{i=1}^{P} \mathcal{T}_i$ conditioned on some decision Y:

$$\mathcal{M}(\cdot) = \mathbb{P}\left(\cup_{i=1}^{P} \mathcal{T}_i \middle| \boldsymbol{Y}\right), \boldsymbol{Y} \in [\mathbf{1}, \boldsymbol{N}]$$

Beak
Neck
Legs
Feathers
Water
Grass
Teeth
.
.

Features $\mathcal{T}_{P,Q}$

P is Spoonbill,
Q is Flamingo

Goal of any explanation $\mathcal{M}(\cdot)$: Find the set of features $\mathcal{T}_P$ that lead to a decision P

~~Causal Explanation,~~ $\cancel{\mathcal{M}(\cdot) = \mathbb{P}(\boldsymbol{P}|\mathcal{T}_P)}$

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

Mathematical Formulation

Probabilistic Interpretation of Explanations

Explanations are probabilities conditioned on features

Let $\mathcal{T}$ be the set of all features learned by a trained network

Explanations maximize the probability of selecting a combination of features $\cup_{i=1}^{P} \mathcal{T}_i$ conditioned on some decision Y:

$$\mathcal{M}(\cdot) = \mathbb{P}\left(\cup_{i=1}^{P} \mathcal{T}_i \middle| Y\right), Y \in [1, N]$$

We obtain information about a class even when that class is absent

$$\left.\begin{array}{l} \text{Beak} \\ \color{red}{\text{Neck}} \\ \text{Legs} \\ \text{Feathers} \\ \text{Water} \\ \text{Grass} \\ \text{Teeth} \\ . \\ . \end{array}\right\} \text{Features } \mathcal{T}_{P,Q}$$

P is Spoonbill,
Q is Flamingo

Goal of any explanation $\mathcal{M}(\cdot)$: Find the set of features $\mathcal{T}_P$ that lead to a decision P

~~Causal Explanation,~~ $\cancel{\mathcal{M}(\cdot) = \mathbb{P}(P|\mathcal{T}_P)}$

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

Mathematical Formulation

Complete Explanations

Complete explanations describe all learned features, irrespective of their presence or absence in a given image

For a binary classifier, with P and Q as the possible classes, probabilistic completeness is given by,

$$1 = \mathbb{P}(P) + \mathbb{P}(Q)$$

Using Law of total probability,

$$1 = \mathbb{P}(P|\mathcal{T}_P)\mathbb{P}(\mathcal{T}_P) + \mathbb{P}(P|\mathcal{T}_P^c)\mathbb{P}(\mathcal{T}_P^c) + \mathbb{P}\big(Q|\mathcal{T}_Q\big)\mathbb{P}(\mathcal{T}_Q) + \mathbb{P}\big(Q|\mathcal{T}_Q^c\big)\mathbb{P}(\mathcal{T}_Q^c)$$

Using Bayes theorem and eliminating the probabilities of the features,

$$1 = \mathbb{P}(\mathcal{T}_P|P)\mathbb{P}(P) + \mathbb{P}(\mathcal{T}_P^c|P)\mathbb{P}(P) + \mathbb{P}\big(\mathcal{T}_Q|Q\big)\mathbb{P}(Q) + \mathbb{P}\big(\mathcal{T}_Q^c|Q\big)\mathbb{P}(Q)$$

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

Mathematical Formulation

Complete Explanations

Complete explanations describe all learned features, irrespective of their presence or absence in a given image

For a binary classifier, with P and Q as the possible classes, probabilistic completeness is given by,

$$1 = \mathbb{P}(P) + \mathbb{P}(Q)$$

Using Law of total probability,

$$1 = \mathbb{P}(P|\mathcal{T}_P)\mathbb{P}(\mathcal{T}_P) + \mathbb{P}(P|\mathcal{T}_P^c)\mathbb{P}(\mathcal{T}_P^c) + \mathbb{P}\left(Q|\mathcal{T}_Q\right)\mathbb{P}(\mathcal{T}_Q) + \mathbb{P}\left(Q|\mathcal{T}_Q^c\right)\mathbb{P}(\mathcal{T}_Q^c)$$

Using Bayes theorem and eliminating the probabilities of the features,

Contrastive explanation

$$1 = \mathbb{P}(\mathcal{T}_P|P)\mathbb{P}(P) + \mathbb{P}(\mathcal{T}_P^c|P)\mathbb{P}(P) + \mathbb{P}\left(\mathcal{T}_Q|Q\right)\mathbb{P}(Q) + \mathbb{P}\left(\mathcal{T}_Q^c|Q\right)\mathbb{P}(Q)$$

Correlation explanation

Counterfactual 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.

Takeaways

Takeaways from Lecture 2

- There are **no "one size fits all" explanations** and techniques
- Explanatory techniques can be categorized based on:
  - Methods they employ
  - Human knowledge requirements
  - Explanation properties
  - Reasoning about decisions
- **These are not disjoint categorizations. The goal of categorization is to simplify the operational requirements of Explainability**
- **Human-centric explanations provide an intuitive probabilistic interpretation**
- **Complete explanations** describe all features in an image, even if said **features are not involved in decision making**

References

Lecture 2: Basics of 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.
- Matthew D Zeiler and Rob Fergus, “Visualizing and understanding convolutional networks,” in European conference on computer vision. Springer, 2014, pp. 818–833 Lee, D. D., and H. S. Seung, 1999, Learning the parts of objects by non-negative matrix factorization.: Nature, 401, 788–91
- Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg, “Smoothgrad: removing noise by adding noise,” arXiv preprint arXiv:1706.03825, 2017.
- Vitali Petsiuk, Abir Das, and Kate Saenko, “Rise: Randomized input sampling for explanation of black-box models,” arXiv preprint arXiv:1806.07421, 2018.
- 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.
- Yosinski et al, “Understanding Neural Networks Through Deep Visualization”, ICML DL Workshop 2014.
- Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE
- Chowdhury, Prithwijit, Mohit Prabhushankar, and Ghassan AlRegib. "Explaining Explainers: Necessity and Sufficiency in Tabular Data." *NeurIPS 2023 Second Table Representation Learning Workshop*. 2023.
- Marques-Silva, Joao. "Logic-based explainability in machine learning." *Reasoning Web. Causality, Explanations and Declarative Knowledge: 18th International Summer School 2022, Berlin, Germany, September 27–30, 2022, Tutorial Lectures*. Cham: Springer Nature Switzerland, 2023. 24-104.
- Kendall, Alex, and Yarin Gal. "What uncertainties do we need in bayesian deep learning for computer vision?." *Advances in neural information processing systems* 30 (2017).