Visual Explainability in Machine Learning

Lecture 9: Causality and 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 7, 2023

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

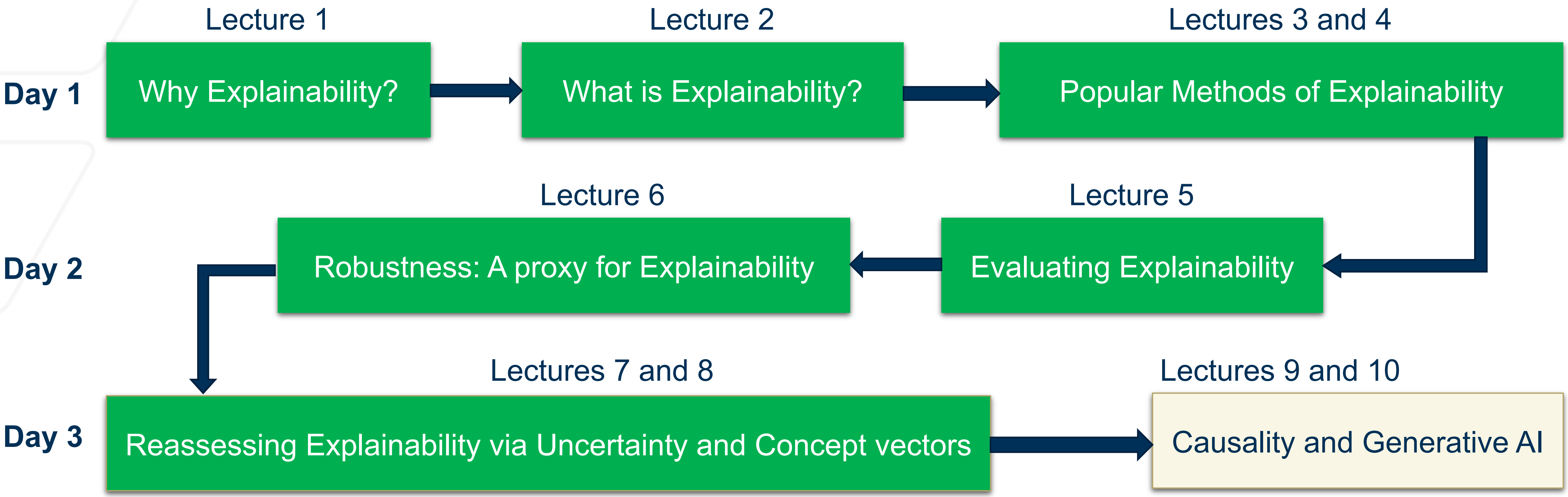
Georgia Institute of Technology, Atlanta, USA

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

Short Course

Course Outline

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



Outline

Lecture 9: Causality and Visual Explainability

- Causality: Forbidden Word
  - Causality in constructing explanations
  - Causality in evaluating explanations
- Causal assessment via Interventions
  - Three rules of Causality
  - Challenges in Deep Learning
- Case Studies
  - Visual Causal Feature Learning
  - Causal Interventional Training
  - CausalCAM: Causal Visual Features
- Takeaways

Outline

Lecture 9: Causality and Visual Explainability

- Causality: Forbidden Word
  - Causality in constructing explanations
  - Causality in evaluating explanations
- Causal assessment via Interventions
  - Three rules of Causality
  - Challenges in Deep Learning
- Case Studies
  - Visual Causal Feature Learning
  - Causal Interventional Training
  - CausalCAM: Causal Visual Features
- Takeaways

Causality

The Forbidden Word

In Lecture 2, we skirted around causal definition of 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:

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

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.

Causality

The Forbidden Word

In Lecture 4, we called *`Why Bullmastiff?'* as observed correlation question rather than causal

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

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

Causality

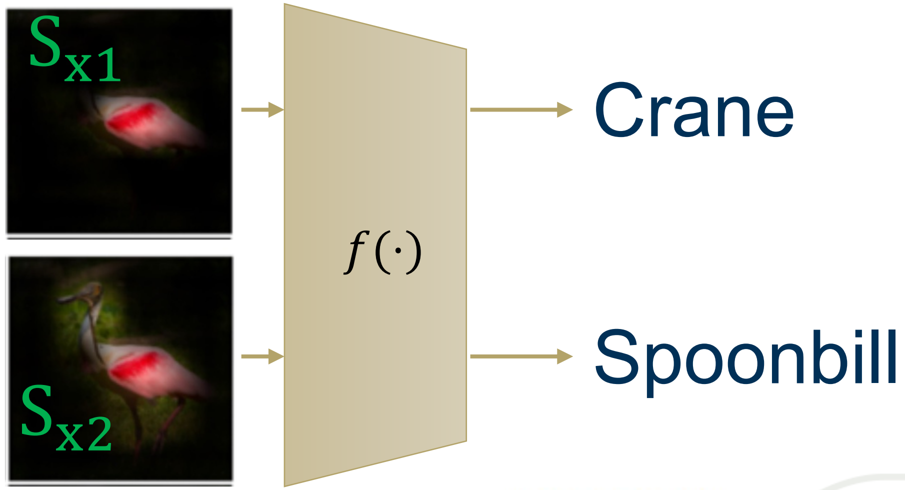
The Forbidden Word

In Lecture 5, we assumed interventions on data and validated the outputs without commenting on causes

y = Prediction
S_x = Explanation masked data

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

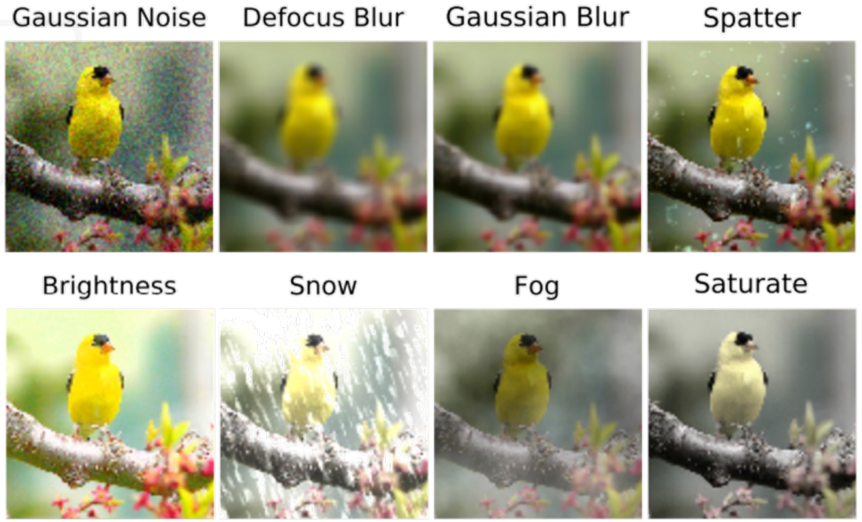
If across N images,
$\mathbf{E(Y|}\boldsymbol{S_{x2}}\mathbf{)} > \mathbf{E(Y|}\boldsymbol{S_{x1}}\mathbf{)}$,
explanation technique 2 is better than explanation technique 1



M. Prabhushankar, and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," *Journal of Selected Topics in Signal Processing*, submitted on Aug. 27, 2023.

Causality
The Forbidden Word

In Lecture 6, we assumed *good* features provide *correct* predictions. Ideally, *causal* features provide correct predictions under corruptions



Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Causality

The Forbidden Word

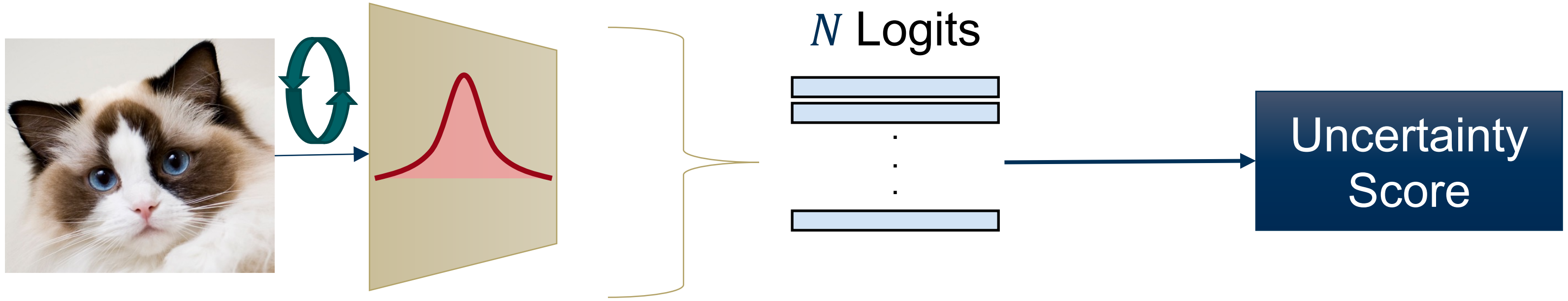
In Lecture 7, we made multiple interventions within data but do not claim causal analysis

Different forward passes with dropout simulate $f_1(\cdot), f_2(\cdot), f_3(\cdot)$.

Challenge: intractable denominator

$$p(\boldsymbol{W}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{W})p(\boldsymbol{W})}{\int p(\boldsymbol{x}|\boldsymbol{W})p(\boldsymbol{W})d\boldsymbol{W}}$$

N forward passes



Final prediction is the mean of the outputs

Variation or entropy of logits is the uncertainty

$$q\,(\boldsymbol{W}_N) \approx p(\boldsymbol{W}_N|\boldsymbol{x})$$

Y Gal, Z Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning", ICML 2016

Causality

The Forbidden Word

In Lecture 8, we perform hypothesis testing but not for causal factors

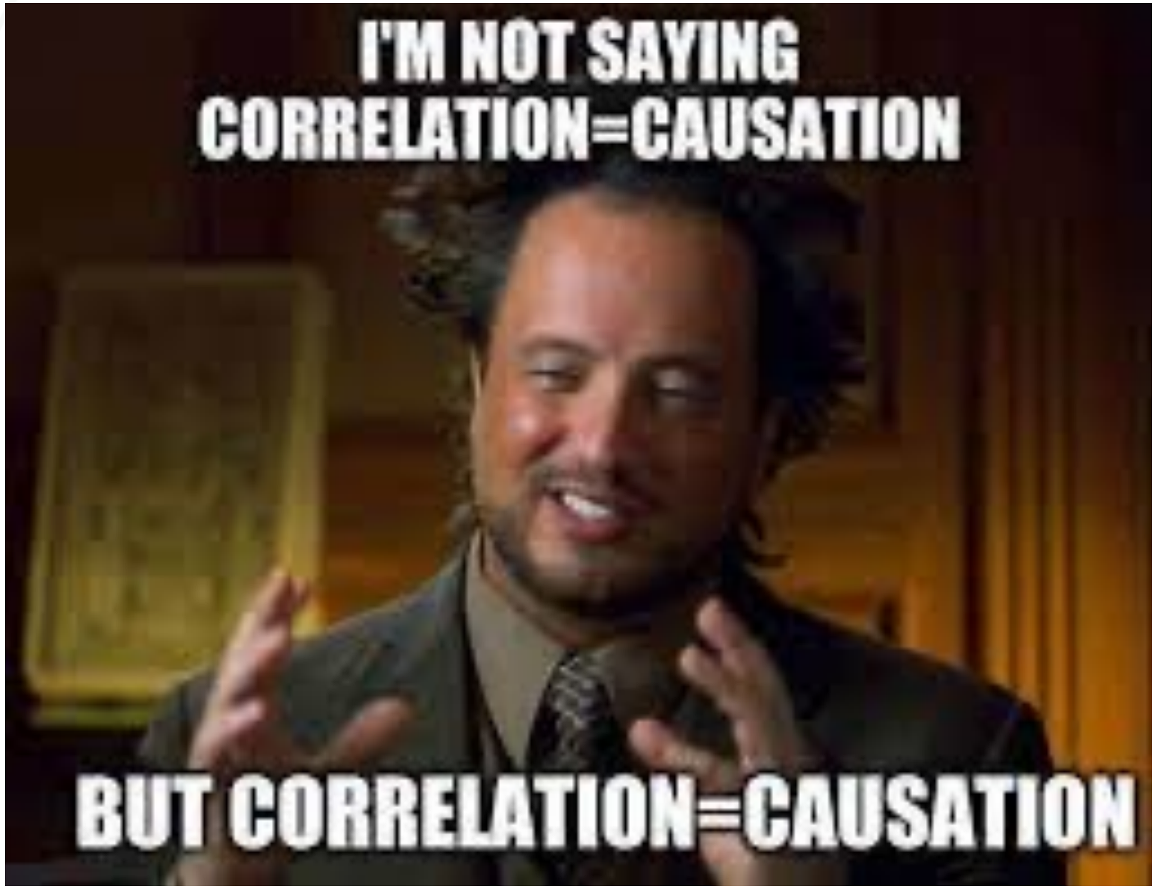
(a) Texture image
81.4% **Indian elephant**
10.3% indri
8.2% black swan

(b) Content image
71.1% **tabby cat**
17.3% grey fox
3.3% Siamese cat

(c) Texture-shape cue conflict
63.9% **Indian elephant**
26.4% indri
9.6% black swan

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Definition, methodology, and evaluation of explanations stem from causal literature



All Explainability techniques, probably

Outline

Lecture 9: Causality and Visual Explainability

- Causality: Forbidden Word
  - Causality in constructing explanations
  - Causality in evaluating explanations
- Causal assessment via Interventions
  - Three rules of Causality
  - Challenges in Deep Learning
- Case Studies
  - Visual Causal Feature Learning
  - Causal Interventional Training
  - CausalCAM: Causal Visual Features
- Takeaways

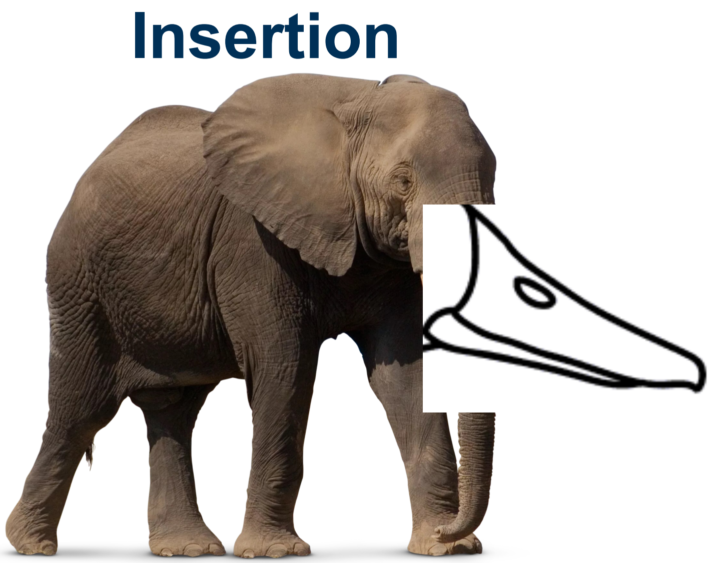
Causal Assessment via Interventions

3 Rules of Causal Inference

Rule 1: Insertion and Deletion of Causal Features

Rule 1 (Insertion/deletion of observations):

$$P(y|do(x), z, w) = P(y|do(x), w)$$



- Fix a causal feature (or a feature that is being tested for causality) in the data

Key Differences:

- There are **no causal features;** approximate using pixels/structures
- The underlying network is **not a structured causal model**

Pearl, Judea. "The do-calculus revisited." *arXiv preprint arXiv:1210.4852* (2012).

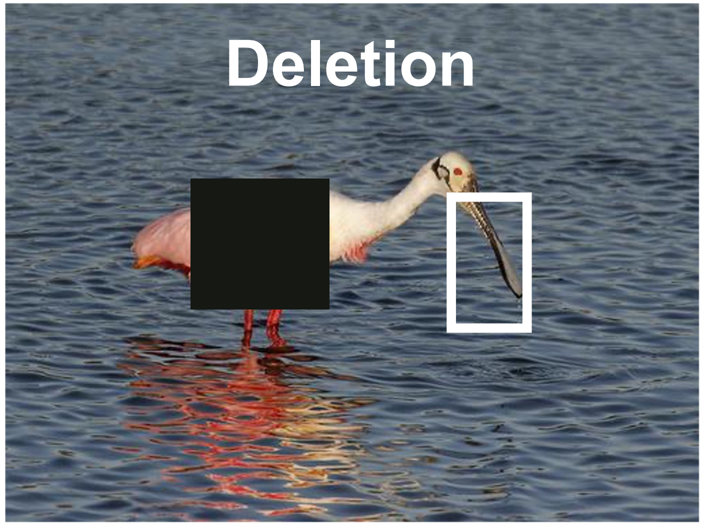
Causal Assessment via Interventions

3 Rules of Causal Inference

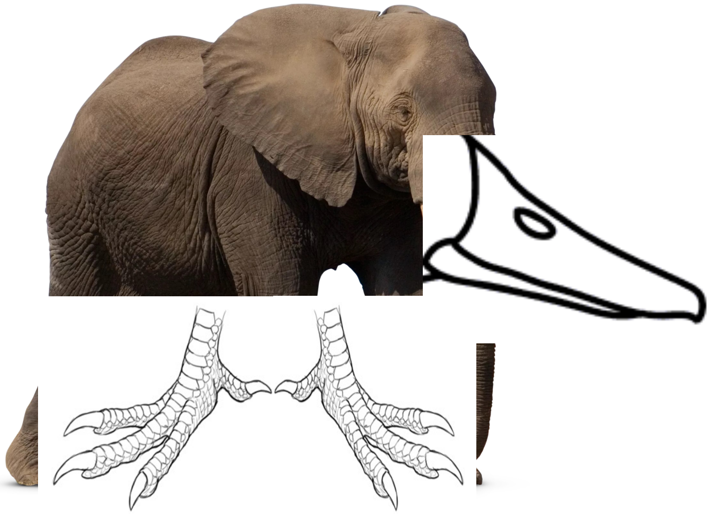
Rule 2: Intervene on all other factors keeping the causal factor constant

Rule 2 (Action/observation exchange):

$$P(y|do(x), do(z), w) = P(y|do(x), z, w)$$



Insertion



- Keeping the causal factor constant from rule 1, change all available factors

Key Differences:

- There are **no causal features;** approximate using pixels/structures
- The underlying network is **not a structured causal model**
- **Impossible** to intervene on all pixels

Pearl, Judea. "The do-calculus revisited." *arXiv preprint arXiv:1210.4852* (2012).

Causal Assessment via Interventions

3 Rules of Causal Inference

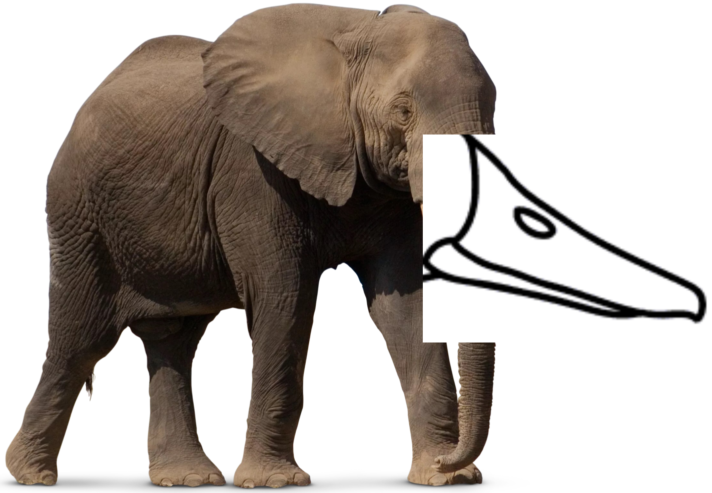
Rule 3: Insertion/Deletion of interventional actions

Rule 3 (Insertion/deletion of actions):

$$P(y|do(x), do(z), w) = P(y|do(x), w)$$

Once causal factors are determined, the interventions from rule 2 are reverted and the causal attribution is noted

Insertion



Key Differences:

- There are **no causal features;** approximate using pixels/structures
- The underlying network is **not a structured causal model**
- **Impossible** to intervene on all pixels

Pearl, Judea. "The do-calculus revisited." *arXiv preprint arXiv:1210.4852* (2012).

Causal Assessment via Interventions

Challenges in Deep Learning

Rules 1 and 2 are not feasible in deep learning

Goal of causal assessment: To determine if a feature is causal

Rule 1: Fix the feature
Challenge: There is no defined set of pixels that are considered features across images

Rule 2: Intervene on other features
Challenge: There is no defined set of pixels that are considered features across images. Since neural networks are not structured causal models, there is no simple mechanism to intervene.

Pearl, Judea. "The do-calculus revisited." *arXiv preprint arXiv:1210.4852* (2012).

Outline

Lecture 9: Causality and Visual Explainability

- Causality: Forbidden Word
  - Causality in constructing explanations
  - Causality in evaluating explanations
- Causal assessment via Interventions
  - Three rules of Causality
  - Challenges in Deep Learning
- Case Studies
  - Visual Causal Feature Learning
  - Causal Interventional Training
  - CausalCAM: Causal Visual Features
- Takeaways

Case Study: Causality via Interventions

Visual Causal Feature Learning

Construct a manipulator function that changes an image sufficiently to be classified incorrectly by a human

- **Images** from MNIST are **manipulated** according to causal intervention principles
  - The methodology is provided in [1]
- The manipulated images are shown to **humans on Amazon Mechanical Turk**
- **Constraint** on the manipulator function is to learn **the least possible modifications** on the original image
- **Modifications** are done on a set of pixels termed **macro-variables**

The manipulator function is an intelligent way of reducing the number of possible interventions

[1] Chalupka, Krzysztof, Pietro Perona, and Frederick Eberhardt. "Visual causal feature learning." *arXiv preprint arXiv:1412.2309* (2014).

Case Study: Causality via Interventions

Explanations via Interventions

Assumption: Predictions and explanations are based on Causal and Context Features

Label: Dog

Causal Features for Dog

Context Features for Dog

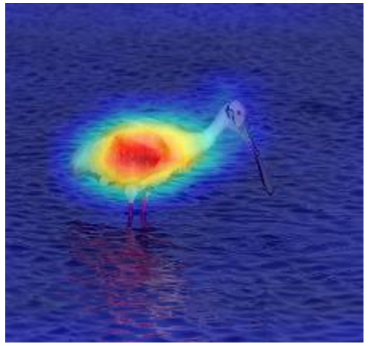
Goal: To **model context features** and remove them out of existing explanation techniques

Case Study: Causality via Interventions

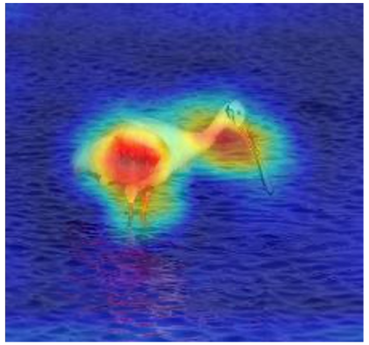
Explanations via Interventions

Causal and Context Features separation via Contrastive Interventions

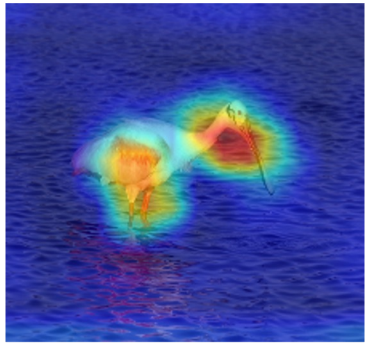
Region to determine
`Why not Flamingo?'



Region to determine
`Why not Crane?'



Region to determine
`Why not Fox?'



Region to determine
`Why not Band-Aid?'

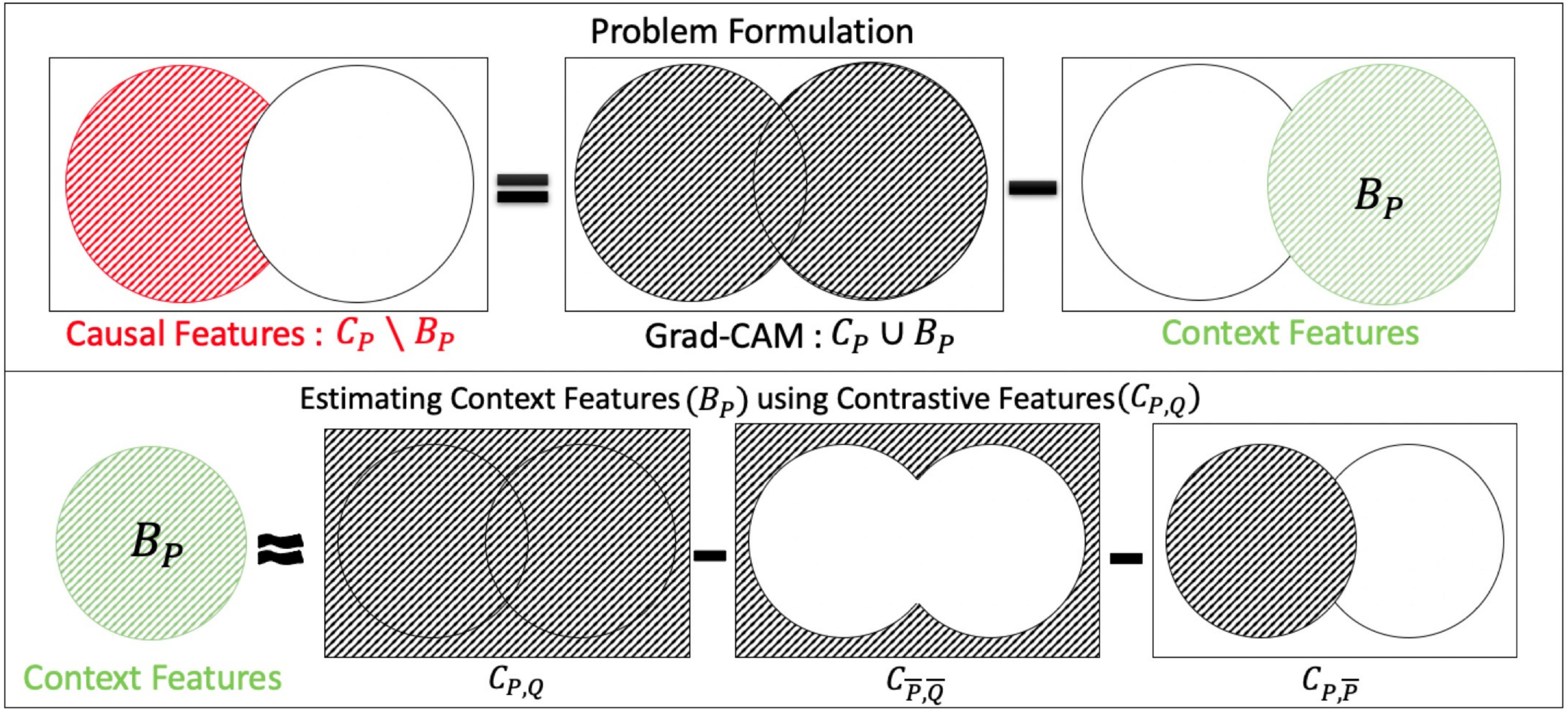
Overlap between relevant contrastive explanations!

M. Prabhushankar, and G. AlRegib, "Extracting Causal Visual Features for Limited Label Classification," *IEEE International Conference on Image Processing (ICIP)*, Sept. 2021.

Case Study: Causality via Interventions

Explanations via Interventions

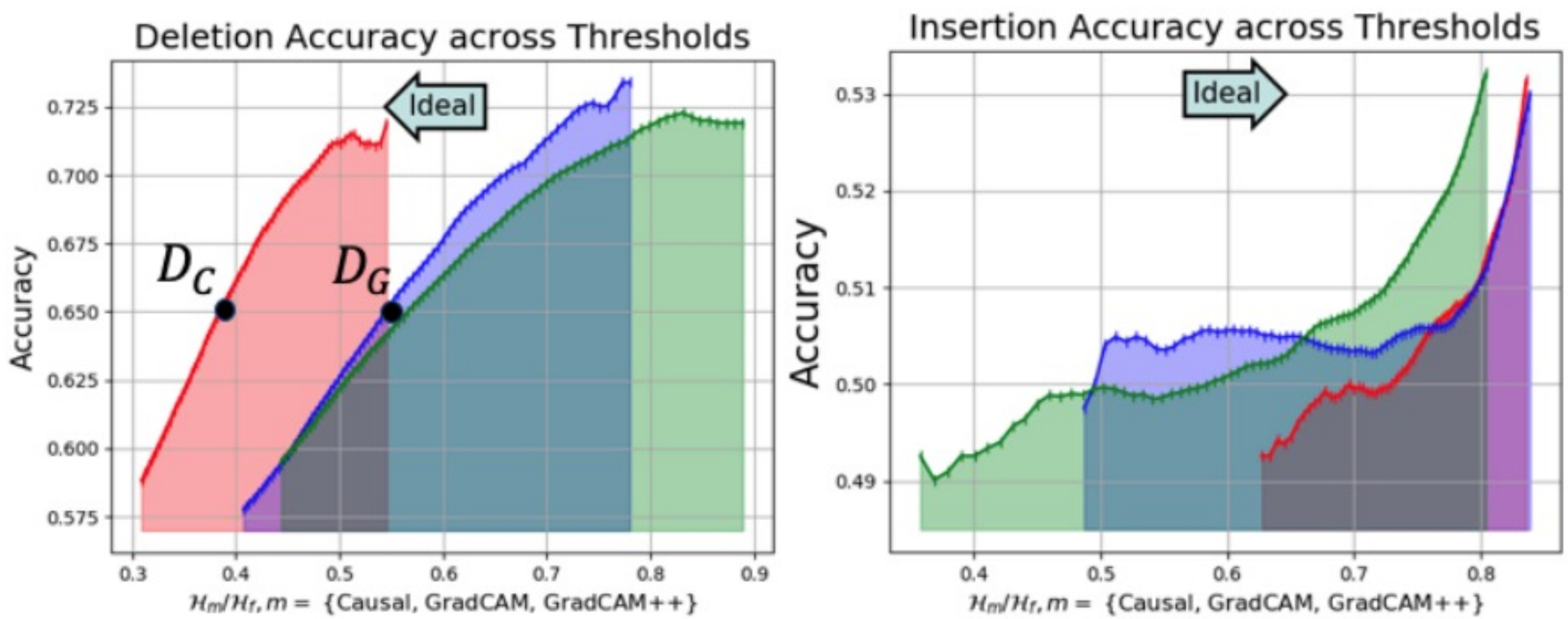
Modeling context features via contrastive features[1]



M. Prabhushankar, and G. AlRegib, ”Extracting Causal Visual Features for Limited Label Classification,” *IEEE International Conference on Image Processing (ICIP)*, Sept. 2021.

Explanations via Interventions

Network Evaluation via structure-wise insertion and deletions



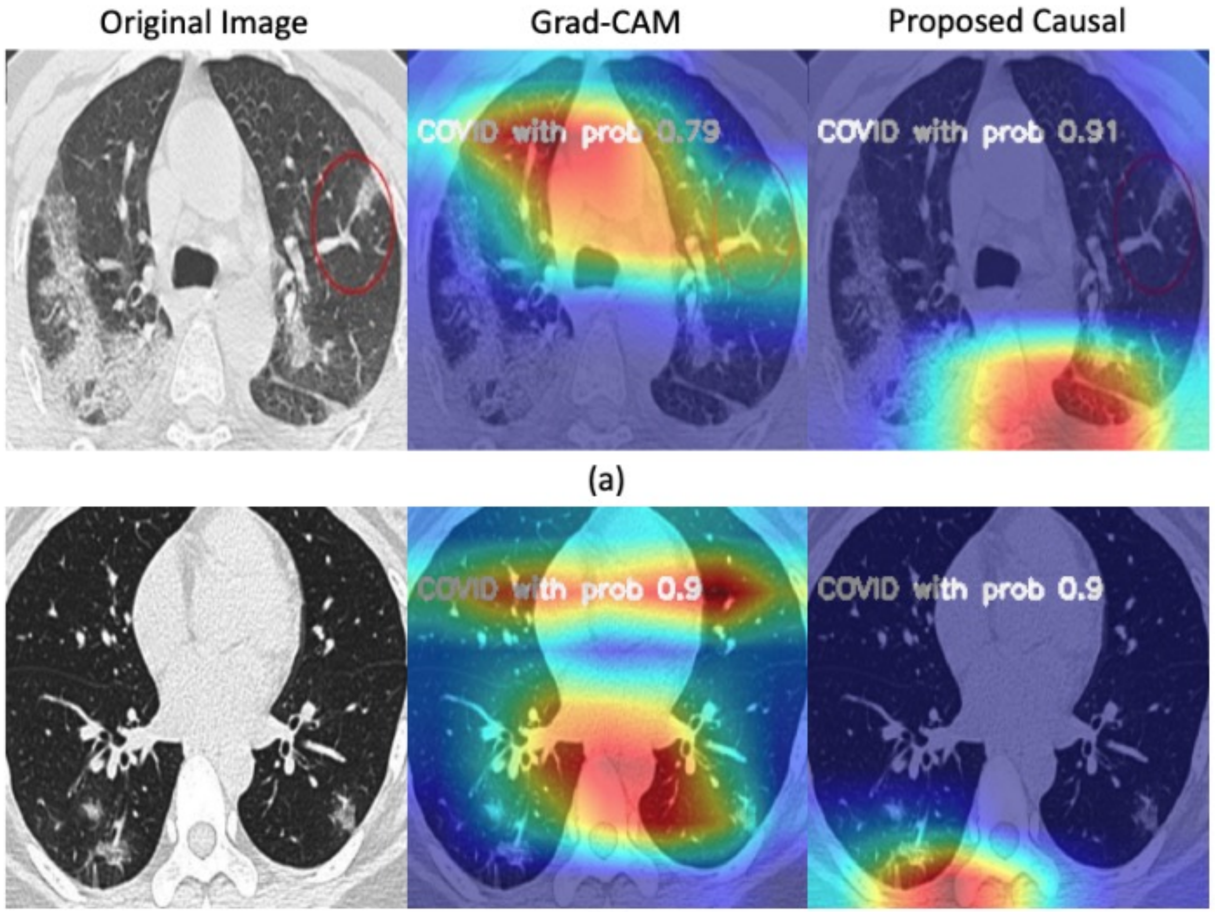
CausalCAM in Red
GradCAM in Purple
GradCAM++ in Green

M. Prabhushankar, and G. AlRegib, "Extracting Causal Visual Features for Limited Label Classification," *IEEE International Conference on Image Processing (ICIP)*, Sept. 2021.

Case Study: Causality via Interventions

Explanations via Interventions

Visual Causal Features increase the confidence in the wrong regions!



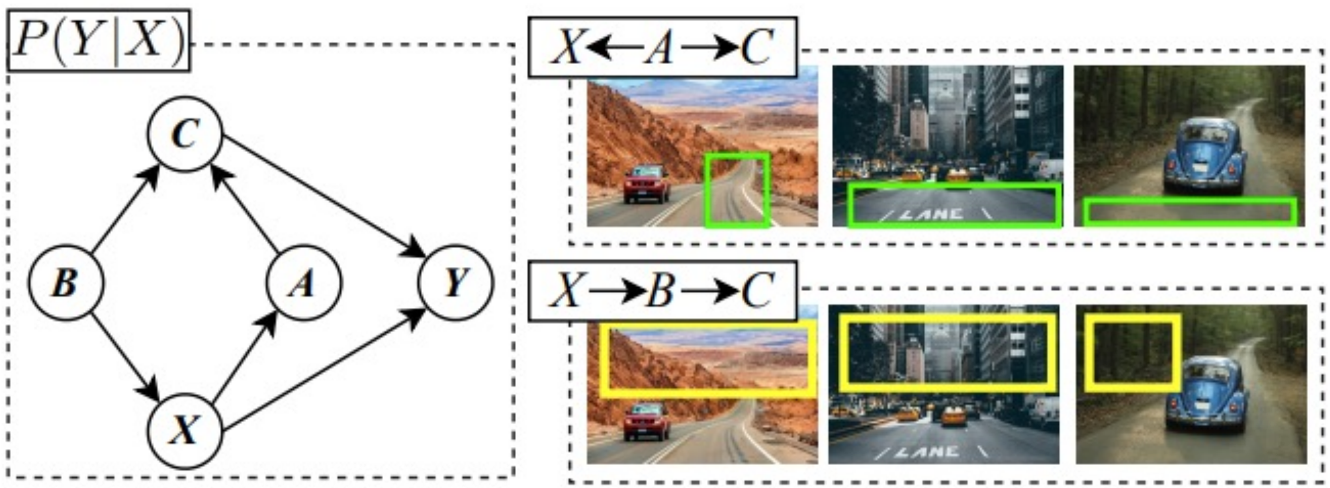
The networks are looking at incorrect regions to make the best predictions!

M. Prabhushankar, and G. AlRegib, "Extracting Causal Visual Features for Limited Label Classification," *IEEE International Conference on Image Processing (ICIP)*, Sept. 2021.

Case Study: Causality via Interventions

Causality via Interventional Training

Causal and Context features can be modeled via a structural causal model[1]



(a) our proposed SCM (b) *car* samples

X: object content ***A***: common sense ***B***: noisy context factor
C: object context ***Y***: prediction (label)

Assumptions:

X = Causal features
C = Context Features
Y = Label

X connects to C via 2 branches:

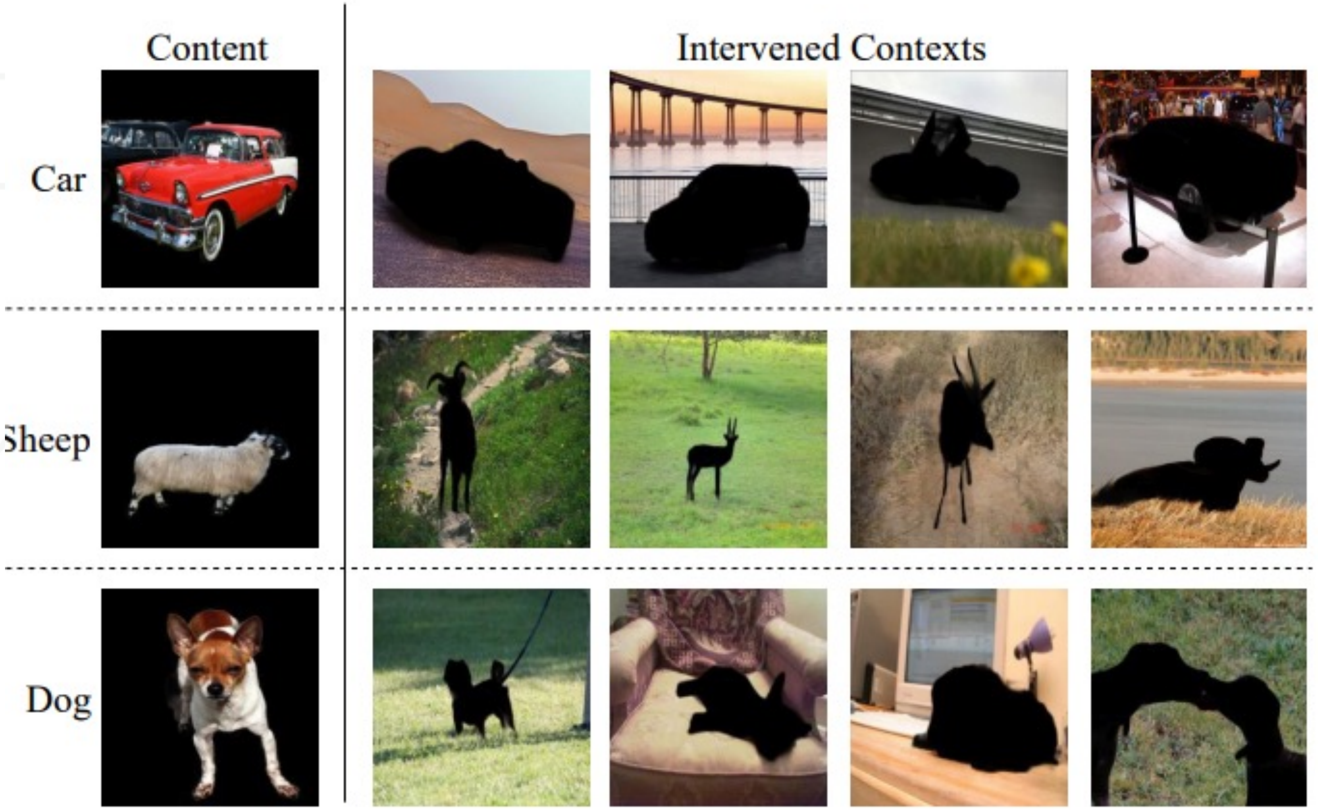
- Via mode A which are the common sense factors (green boxes in (b))
- Via node B which are noisy contexts (yellow boxes in (b))

[1] Qin, W., Zhang, H., Hong, R., Lim, E. P., & Sun, Q. (2021). Causal interventional training for image recognition. *IEEE Transactions on Multimedia*.

Case Study: Causality via Interventions

Causality via Interventional Training

Causal and Context features can be modeled via a structural causal model[1]



- Saliency maps are used to separate X, A, and B features
  - More details in [1]
  - **Saliency algorithms have an object bias**
  - **Results require a well-centered object**
  - Methodology is vulnerable to choice of saliency algorithms and networks

[1] Qin, W., Zhang, H., Hong, R., Lim, E. P., & Sun, Q. (2021). Causal interventional training for image recognition. *IEEE Transactions on Multimedia*.

Takeaways

Lecture 9: Causality and Visual Explainability

- **Causal literature has played a pivotal role is constructing, defining, and evaluating explanations**
- **However, explanations do not highlight causal features**
  - Rather they act as **feature attribution methods,** after knowing the predicted class
- Visual factor causal assessment is challenging in deep learning networks
  - **Disjoint features are not available**
  - **The neural network is not a structured causal model**
- Existing techniques that perform causal modeling
  - Construct manipulator functions that find causal factors
  - Model context features out of post hoc explanations
  - Construct a simplified structured causal model by extracting saliency map based objects

References

Lecture 9: Causality and Visual 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.
- M. Prabhushankar, and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," *Journal of Selected Topics in Signal Processing*, submitted on Aug. 27, 2023.
- Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.
- Y Gal, Z Ghahramani, “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning”, ICML 2016
- Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.
- Pearl, Judea. "The do-calculus revisited." *arXiv preprint arXiv:1210.4852* (2012).
- Chalupka, Krzysztof, Pietro Perona, and Frederick Eberhardt. "Visual causal feature learning." *arXiv preprint arXiv:1412.2309* (2014).
- M. Prabhushankar, and G. AlRegib, ”Extracting Causal Visual Features for Limited Label Classification,” *IEEE International Conference on Image Processing (ICIP)*, Sept. 2021.
- Qin, W., Zhang, H., Hong, R., Lim, E. P., & Sun, Q. (2021). Causal interventional training for image recognition. *IEEE Transactions on Multimedia*.