

Robust Neural Networks

Part 4: Intervenability at Inference

Objective

Objective of the Tutorial

To discuss methodologies that promote robustness in neural networks at inference

- Part 1: Inference in Neural Networks
- Part 2: Explainability at Inference
- Part 3: Uncertainty at Inference
- **Part 4: Intervenability at Inference**
 - Definitions of Intervenability
 - Causality
 - Privacy
 - Interpretability
 - Prompting
 - Benchmarking
 - Case Study: Intervenability in Interpretability
- Part 5: Conclusions and Future Directions

Intervenability

Through the Causal Glass

Assess: The amenability of neural network decisions to human interventions



“Interventions in data are manipulations that are designed to test for causal factors”

Intervenability

Through the Privacy Glass

Assure: The amenability of neural network decisions to human interventions



*“Intervenability aims at the possibility for parties involved in any **privacy-relevant** data processing to **interfere** with the ongoing or planned data processing”*

Intervenability

Through the Interpretability Glass

Interpret: The amenability of neural network decisions to human interventions



*“The post-hoc field of explainability, that previously only justified decisions, becomes **active** by being involved in the decision making process and providing limited, but relevant and contextual interventions”*

Intervenability

Through the Benchmarking Glass

Verify: The amenability of neural network decisions to human interventions



*“... new **benchmarks** were proposed to specifically test generalization of classification and detection methods with respect to **simple** algorithmically generated interventions like spatial shifts, blur, changes in brightness or contrast...”*

Intervenability

Through the Human Glass

The amenability of neural network decisions to human interventions



- **Assess: Causality**
- **Assure: Privacy**
- **Interpret: Interpretability**
- **Verify: Benchmarking**

Case Study: Intervenability in Interpretability

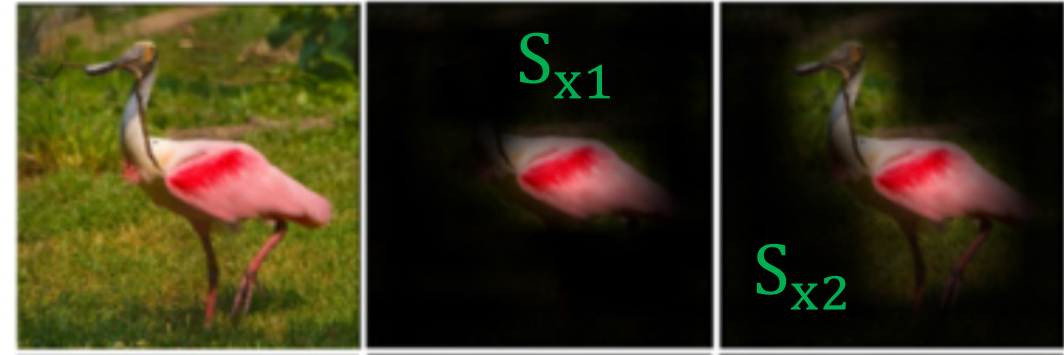
Explanation Evaluation via Masking

Common evaluation technique is masking the image and checking for prediction correctness

y = Prediction

S_x = Explanation masked data

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



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



VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability



Mohit Prabhushankar, PhD
Postdoc



Ghassan AlRegib, PhD
Professor

Case Study: Intervenability in Interpretability

Predictive Uncertainty in Explanations

Explanatory techniques have predictive uncertainty

Explanation of Prediction

Uncertainty of Explanation



Why Bullmastiff?

Uncertainty in answering
Why Bullmastiff?

Case Study: Intervenability in Interpretability

Predictive Uncertainty

Uncertainty due to variance in prediction when model is kept constant



$$V[y|S_x] = V[E(y|S_x)] + E(V[y|S_x])$$

y = Prediction

$V[y]$ = Variance of prediction (Predictive Uncertainty)

S_x = Subset of data (Some intervention)

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

$V(Y|S_x)$ = Variance of class given all other residuals

Case Study: Intervenability in Interpretability

Visual Explanations (partially) reduce Predictive Uncertainty

A 'good' explanatory technique is evaluated to have zero $V[E(y|S_x)]$



zero ←

$$V[y|S_x] = V[E(y|S_x)] + E(V[y|S_x])$$

y = Prediction

$V[y]$ = Variance of prediction (Predictive Uncertainty)

S_x = Subset of data (Some intervention)

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

$V(Y|S_x)$ = Variance of class given all other residuals

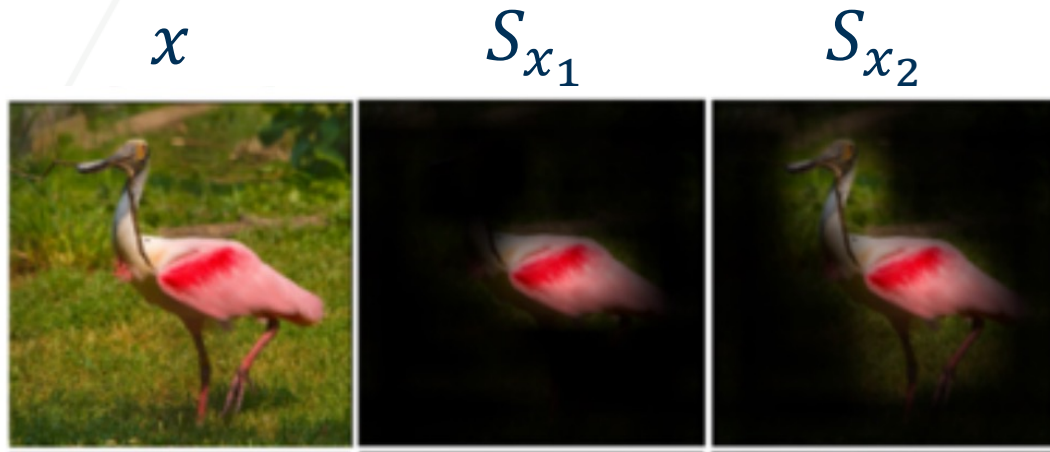
Key Observation 1: Visual Explanations are evaluated to partially reduce the predictive uncertainty in a neural network

Network evaluations have nothing to do with human Explainability!

Case Study: Intervenability in Interpretability

Predictive Uncertainty in Explanations is the Residual

All other subsets 'not' chosen by the explanatory technique contributes to uncertainty



$$V[y|S_x] = V[E(y|S_x)] + E(V[y|S_x])$$

y = Prediction

$V[y]$ = Variance of prediction (Predictive Uncertainty)

S_x = Subset of data (Some intervention)

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

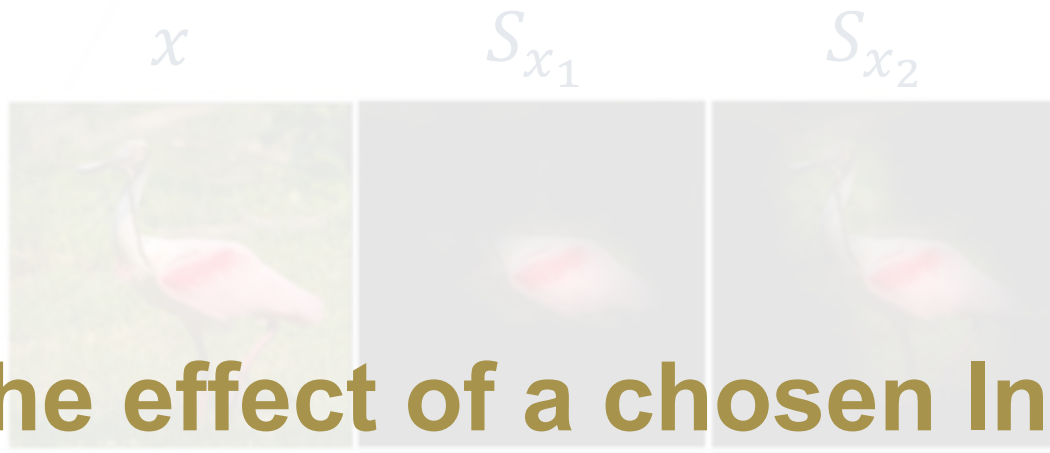
$V(Y|S_x)$ = Variance of class given all other residuals

Key Observation 2: Uncertainty in Explainability occurs due to all combinations of features that the explanation did not attribute to the network's decision

Case Study: Intervenability in Interpretability

Predictive Uncertainty in Explanations is the Residual

All other subsets 'not' chosen by the explanatory technique contributes to uncertainty



$$V[y|S_x] = V[E(y|S_x)] + E(V[y|S_x])$$

The effect of a chosen Interventions can be measured based on *all the Interventions that were not chosen*

Key Observation 2: Uncertainty in Explainability occurs due to all combinations of features that the explanation did not attribute to the network's decision

Case Study: Intervenability in Interpretability

Predictive Uncertainty in Explanations is the Residual

All other subsets 'not' chosen by the explanatory technique contributes to uncertainty

Explanation of Prediction Uncertainty of Explanation



Snout is not as highlighted as the jowls in explanation (not as important for decision)

However, snout is an important characteristic that is used to differentiate against other dogs. Hence, there is uncertainty on why this feature is not included in the attribution

Key Observation 2: Uncertainty in Explainability occurs due to all combinations of features that the explanation did not attribute to the network's decision

Case Study: Intervenability in Interpretability

Predictive Uncertainty in Explanations is the Residual

All other subsets **'not' chosen** by the explanatory technique contributes to uncertainty

Explanation of Prediction Uncertainty of Explanation



Snout is not as highlighted as the jowls in explanation (not as important for decision)

However, snout is an important characteristic that is used to differentiate against other dogs. Hence, there is uncertainty on why this feature is not included in the attribution

Not chosen features are intractable!

Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability

Contrastive explanations are an intelligent way of obtaining other subsets



$$V[y|S_x] = V[E(y|S_x)] + E(V[y|S_x])$$

Make it finite by only considering the subsets that change y

$$\left. \begin{array}{l} Y_1|S_{x1} \\ Y_2|S_{x2} \\ Y_3|S_{x3} \\ Y_4|S_{x4} \\ Y_5|S_{x5} \\ \cdot \\ \cdot \\ Y_N|S_{xN} \end{array} \right\} \text{Variance}$$

Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability

Uncertainty in Explainability can be used to analyze Explanatory methods and Networks

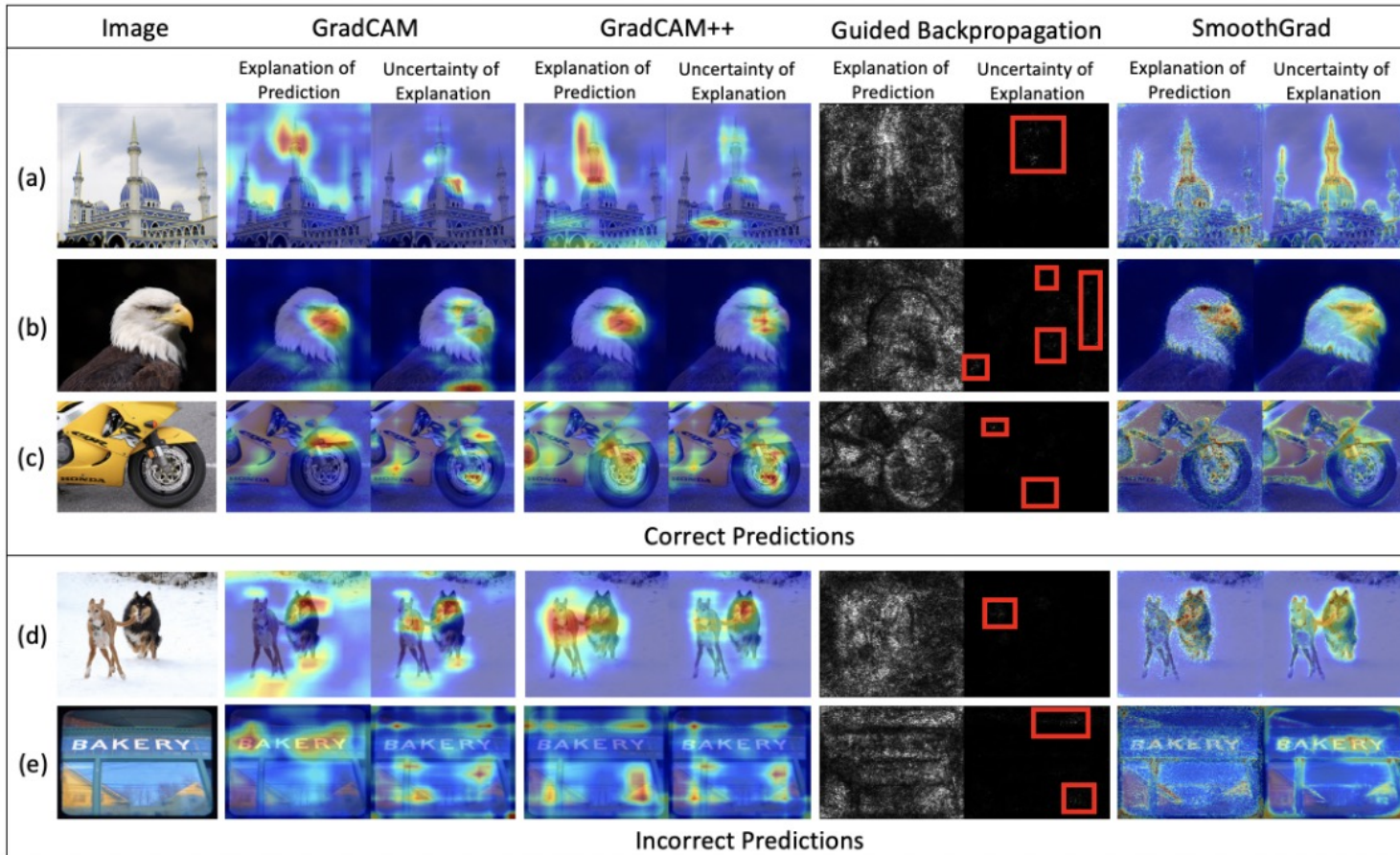
- Is GradCAM better than GradCAM++?
- Is a SWIN transformer more reliable than VGG-16?

Need objective quantification of Intervention Residuals

Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability: mIOU

On incorrect predictions, the overlap of explanations and uncertainty is higher



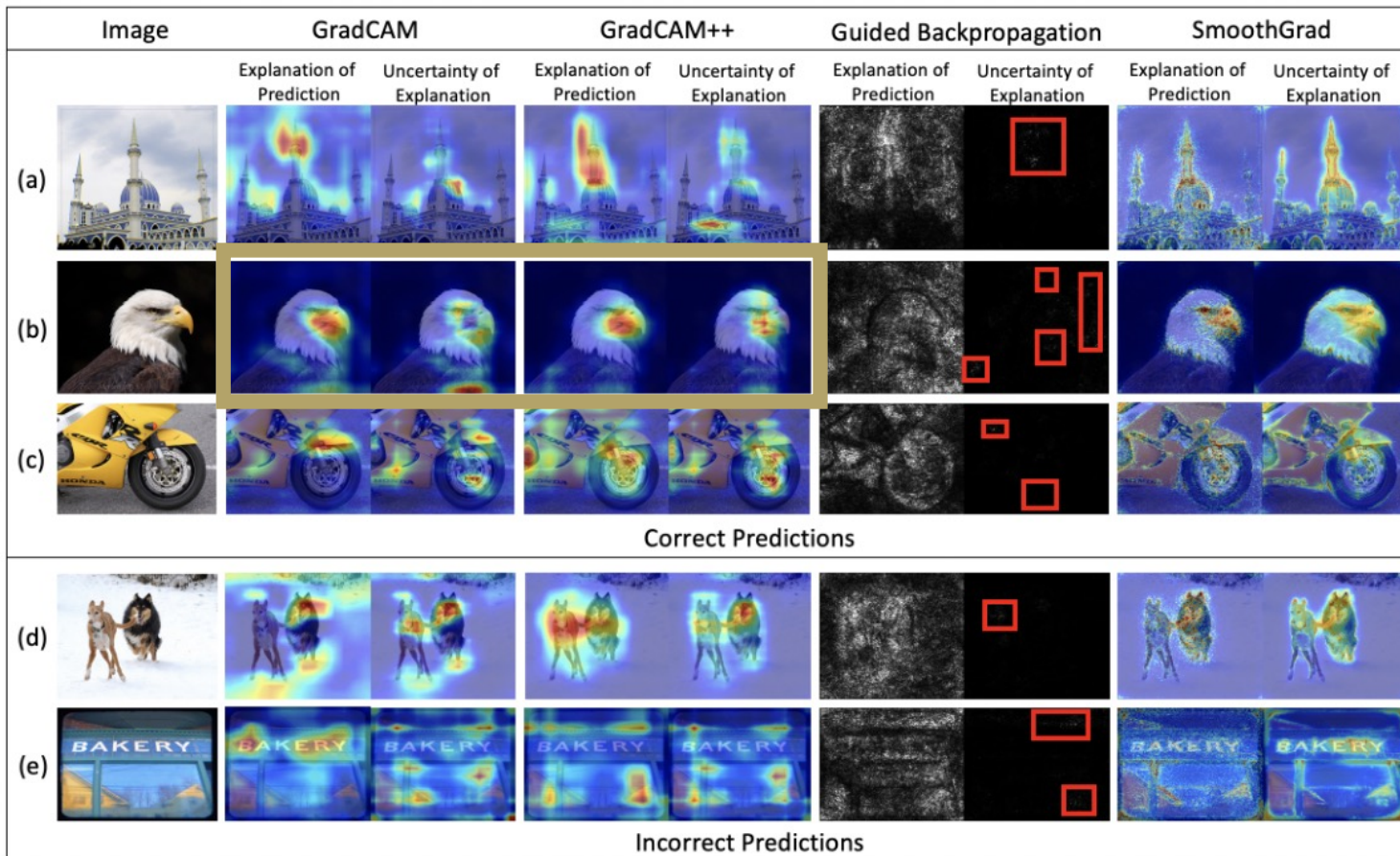
Objective Metric:
Intersection over Union (IoU)
between
explanation and
Uncertainty

Higher the IoU, higher the
uncertainty in explanation (or
less trustworthy is the
prediction)

Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability: mIOU

On incorrect predictions, the overlap of explanations and uncertainty is higher



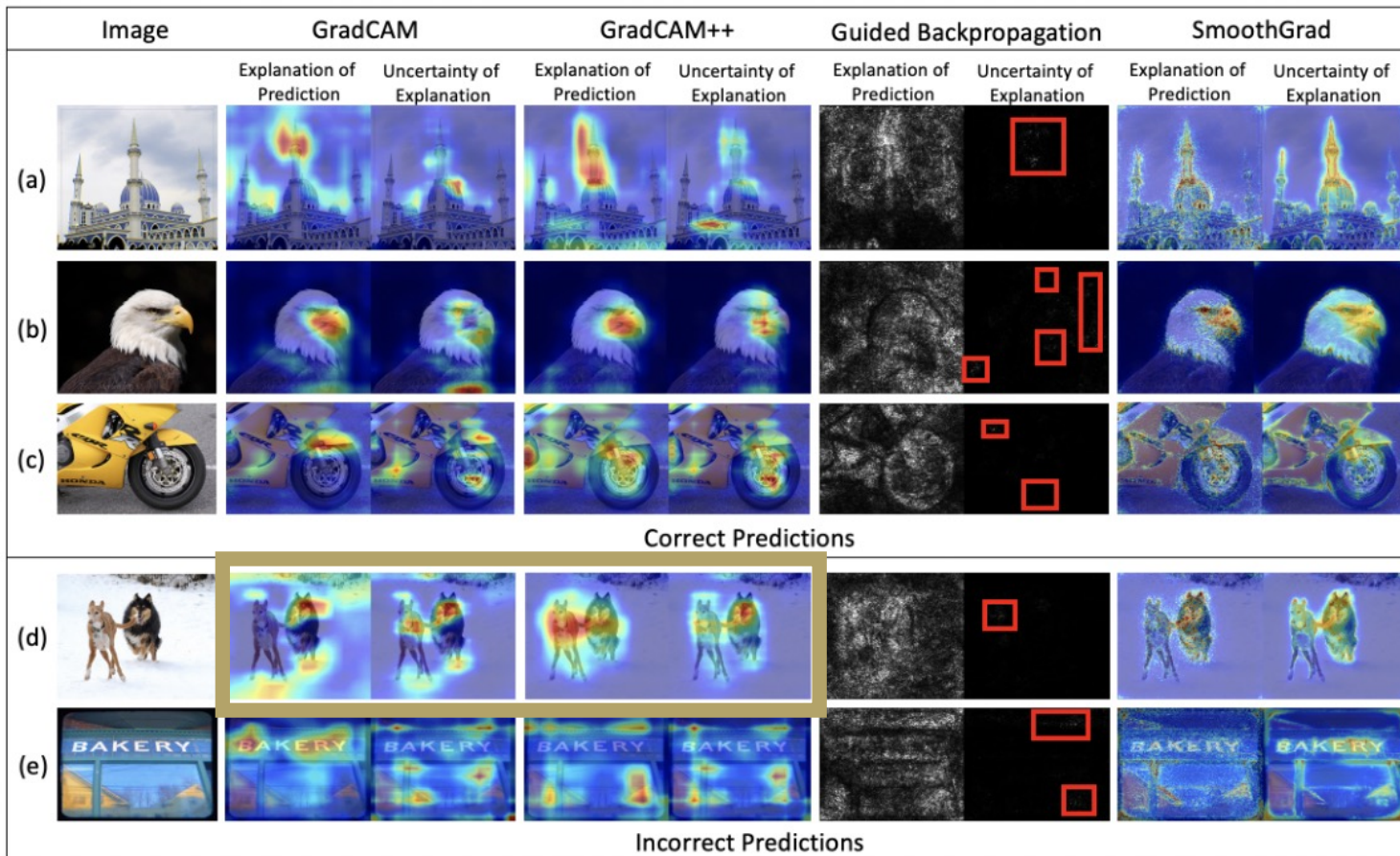
Objective Metric 1:
Intersection over Union (IoU)
between
explanation and
Uncertainty

Higher the IoU, higher the
uncertainty in explanation (or
less trustworthy is the
prediction)

Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability: mIOU

On incorrect predictions, the overlap of explanations and uncertainty is higher



Objective Metric 1:
Intersection over Union (IoU)
between
explanation and
Uncertainty

Higher the IoU, higher the
uncertainty in explanation (or
less trustworthy is the
prediction)