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"*

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, *109*(5), 612-634.

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**"*

Hansen, M.: Top 10 mistakes in system design from a privacy perspective and privacy protection goals. In: Camenisch, J., Crispo, B., Fischer-Hübner, S., Leenes, R., Russello, G. (eds.) Privacy and Identity Management for Life. IFIP AICT, vol. 375, pp. 14–31. Springer, Heidelberg (2012)

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**"*

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

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..."*

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, *109*(5), 612-634.

Case Study: Intervenability in Interpretability

Challenges in Intervenability

The amenability of neural network decisions to human interventions

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

Challenges:

- **Choosing the type of Intervention: Explanation Evaluation**
- Residuals of Interventions: Uncertainty

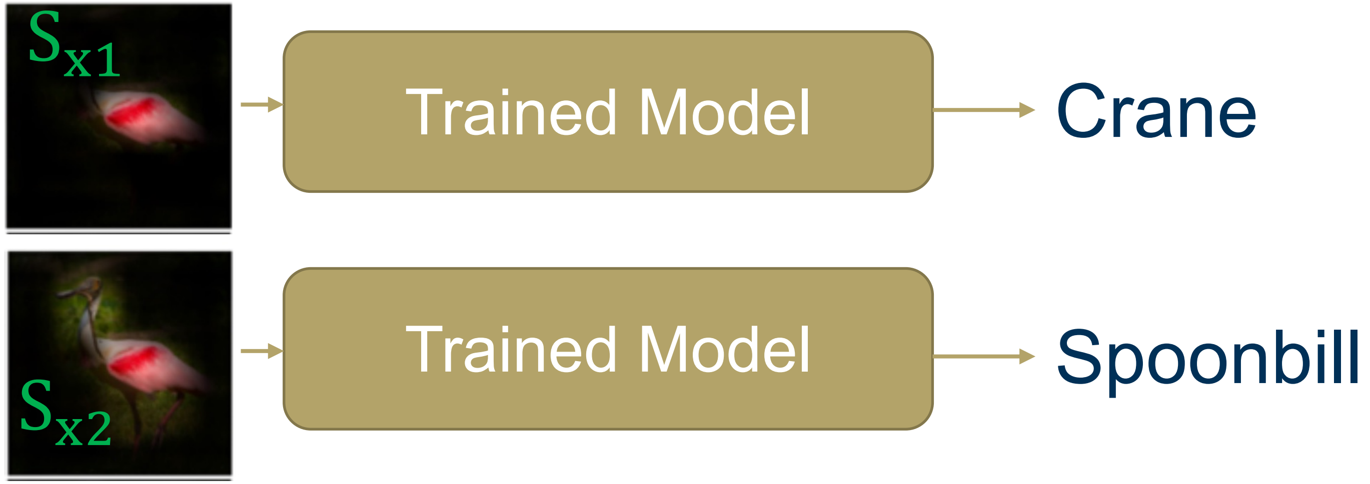
Case Study: Intervenability in Interpretability

Explanation Evaluation

Visual explanations are evaluated via masking the important regions in the image and passing it through the network

Three types of Masking:

1. **Masking using explanation heatmap**
2. Pixel-wise masking using explanation as importance
3. Structure-wise masking using information encoded in explanation



Masking = Intelligent Intervention

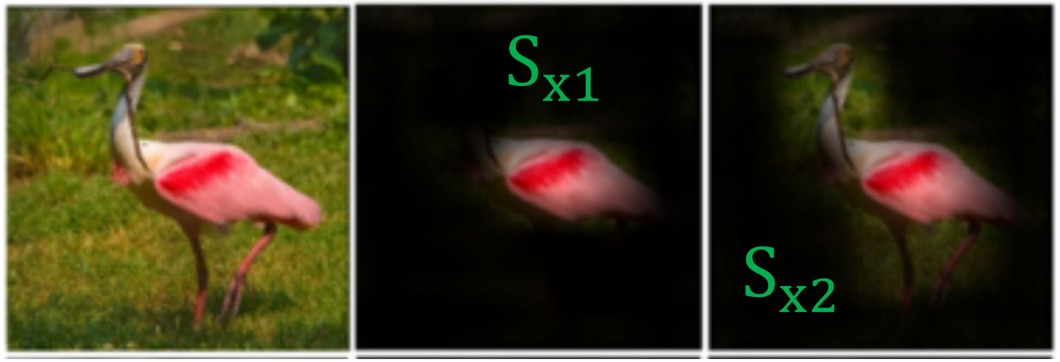
Case Study: Intervenability in Interpretability

Evaluation 1: Explanation Evaluation via Masking

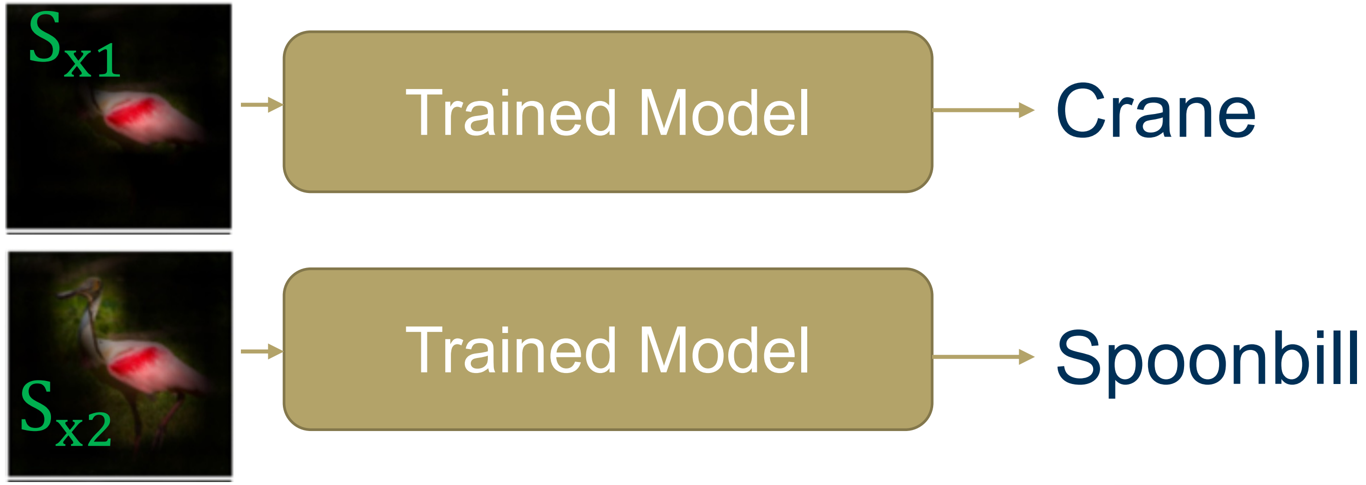
Common evaluation technique is masking the image and checking for prediction correctness

y = Prediction
$\mathrm{S_x}$ = Explanation masked data

$\mathrm{E(Y|S_x)}$ = Expectation of class given $\mathrm{S_x}$



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



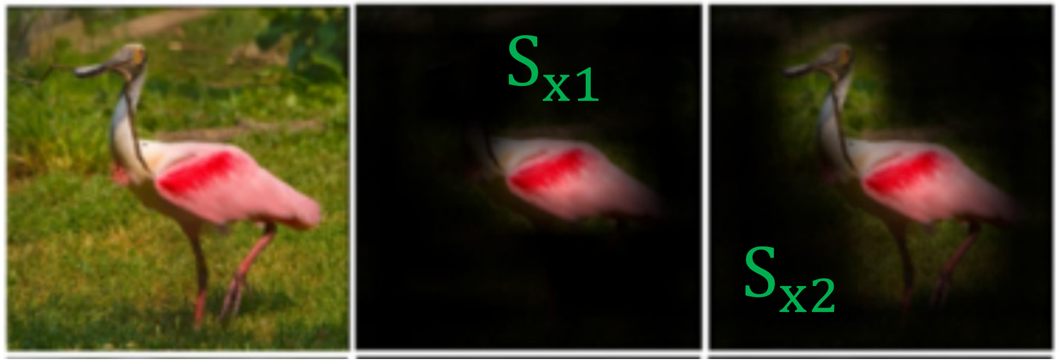
Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018.

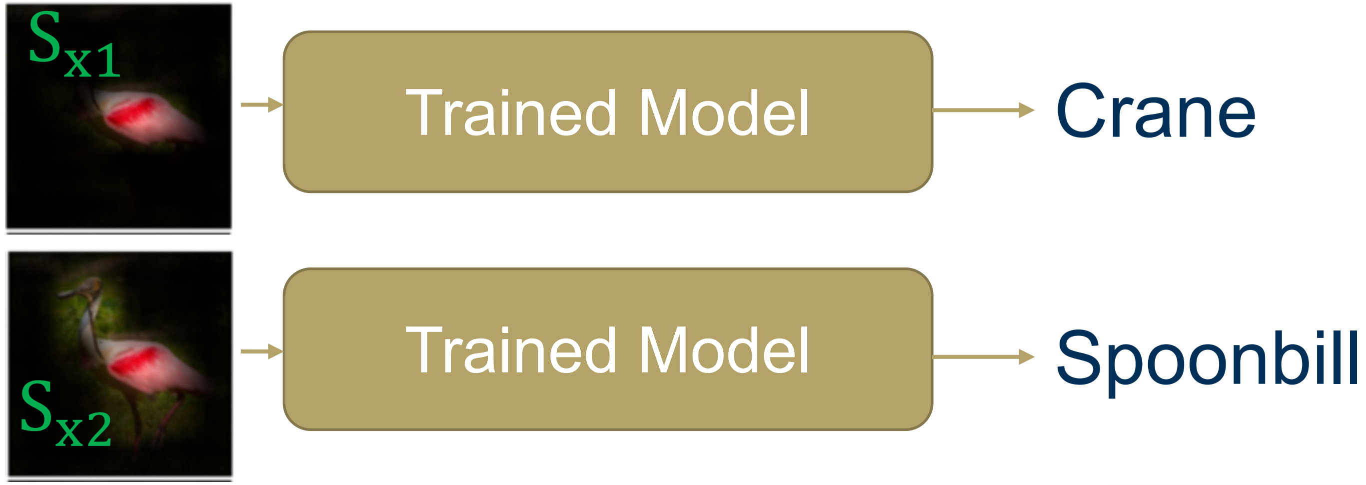
Case Study: Intervenability in Interpretability

Evaluation 1: Explanation Evaluation via Masking

However, explanation masking encourages ‘larger’ explanations

- Larger explanations imply more features in masked images are intact (unmasked)
- This increases likelihood of a correct prediction
- ‘Fine-grained’ explanations are not promoted





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.

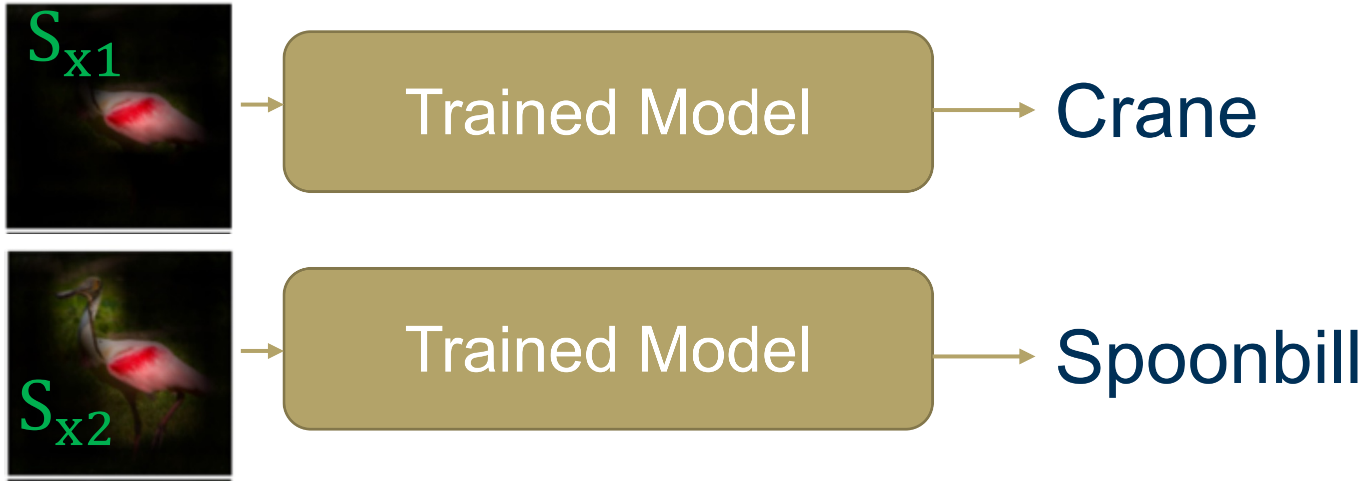
Case Study: Intervenability in Interpretability

Explanation Evaluation

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

Three types of Masking:

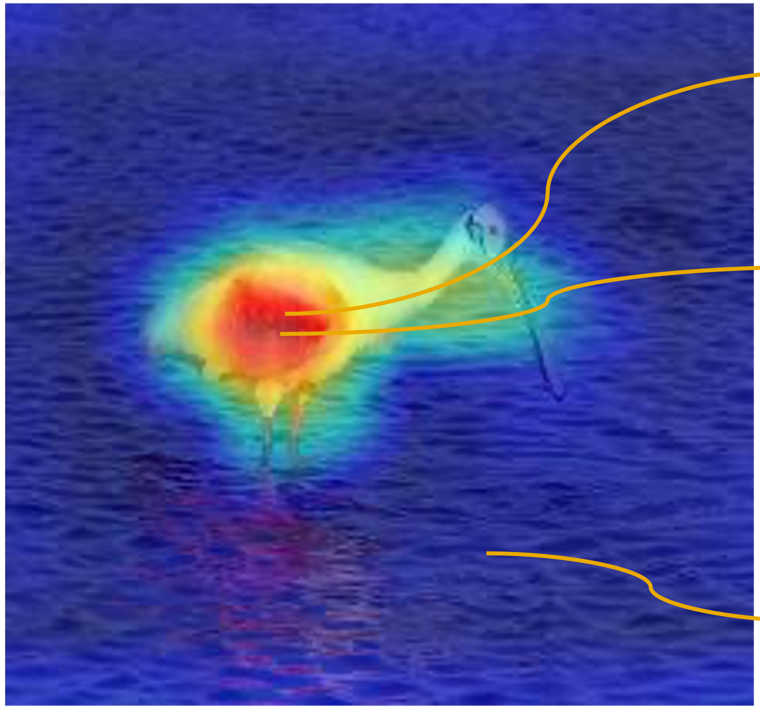
1. Masking using explanation heatmap
2. **Pixel-wise masking using explanation as importance**
3. Structure-wise masking using information encoded in explanation



Case Study: Intervenability in Interpretability

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



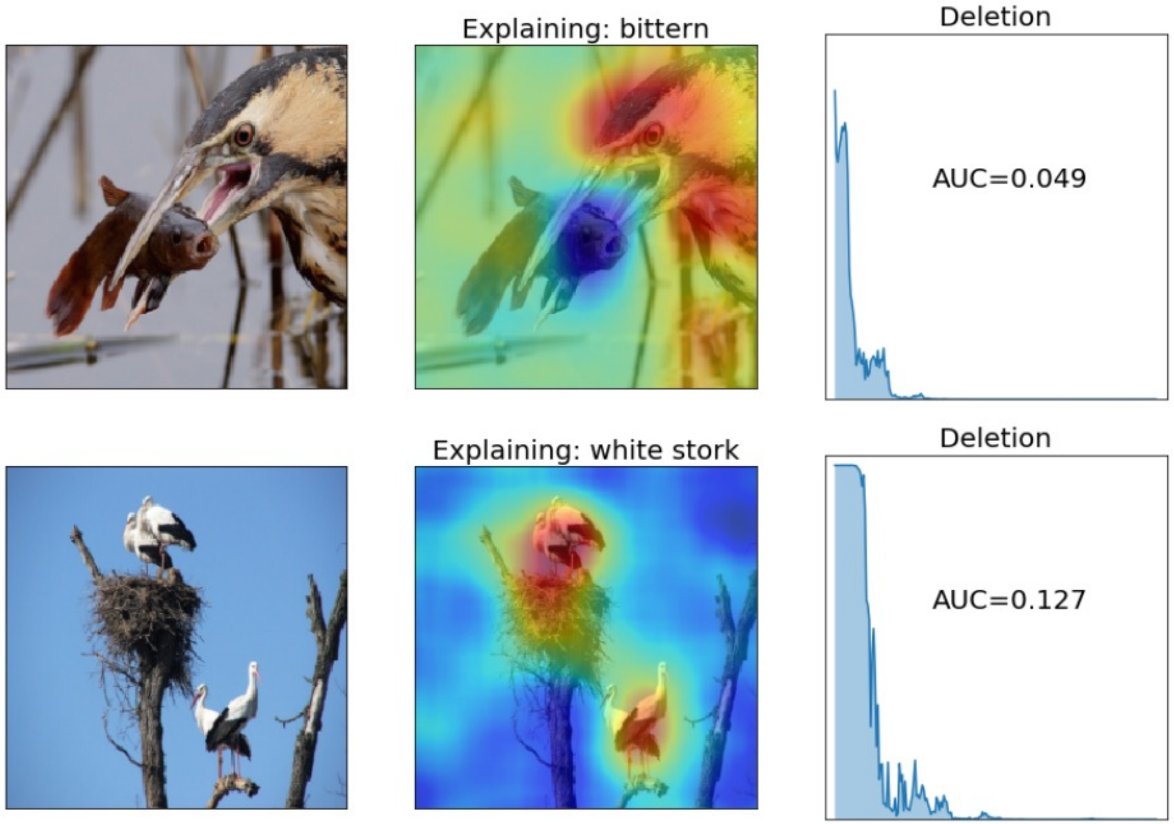
Step 1: Mask highest importance pixel and pass the image through the network. Note the probability of spoonbill.
Step 2: Mask the second highest importance pixel from the image in Step 1 and pass the image through the network. Note the probability of spoonbill.
Step 3: Repeat until all pixels are deleted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

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

Case Study: Intervenability in Interpretability

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.

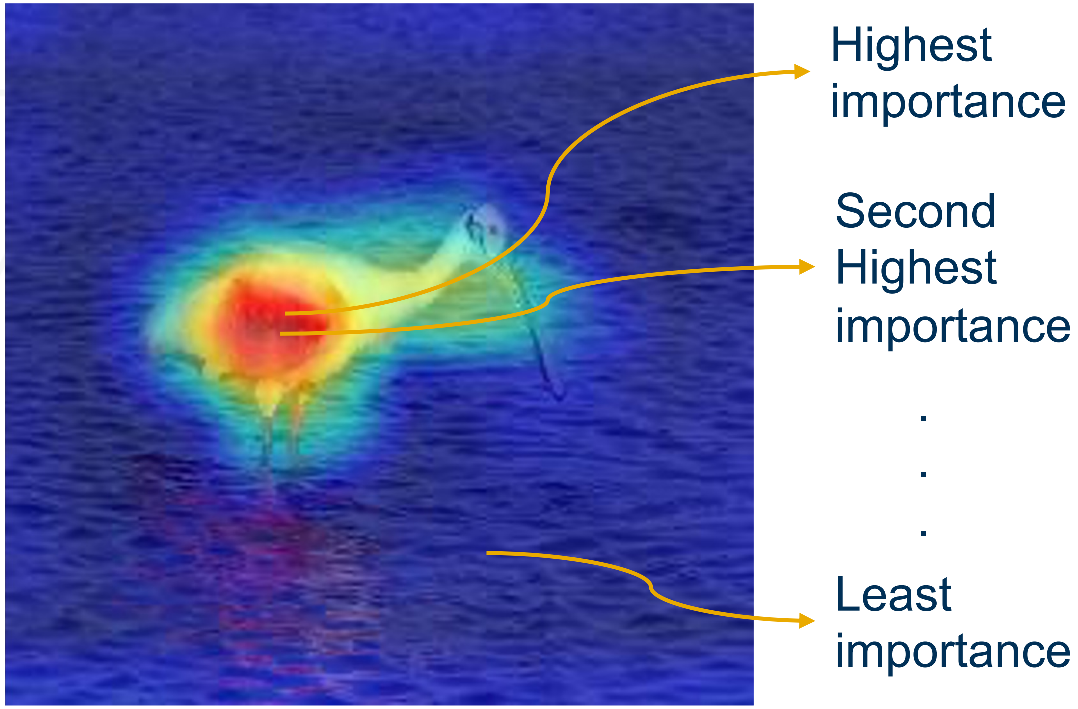
OLIVES @GeorgiaTech

Georgia Tech

Case Study: Intervenability in Interpretability

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

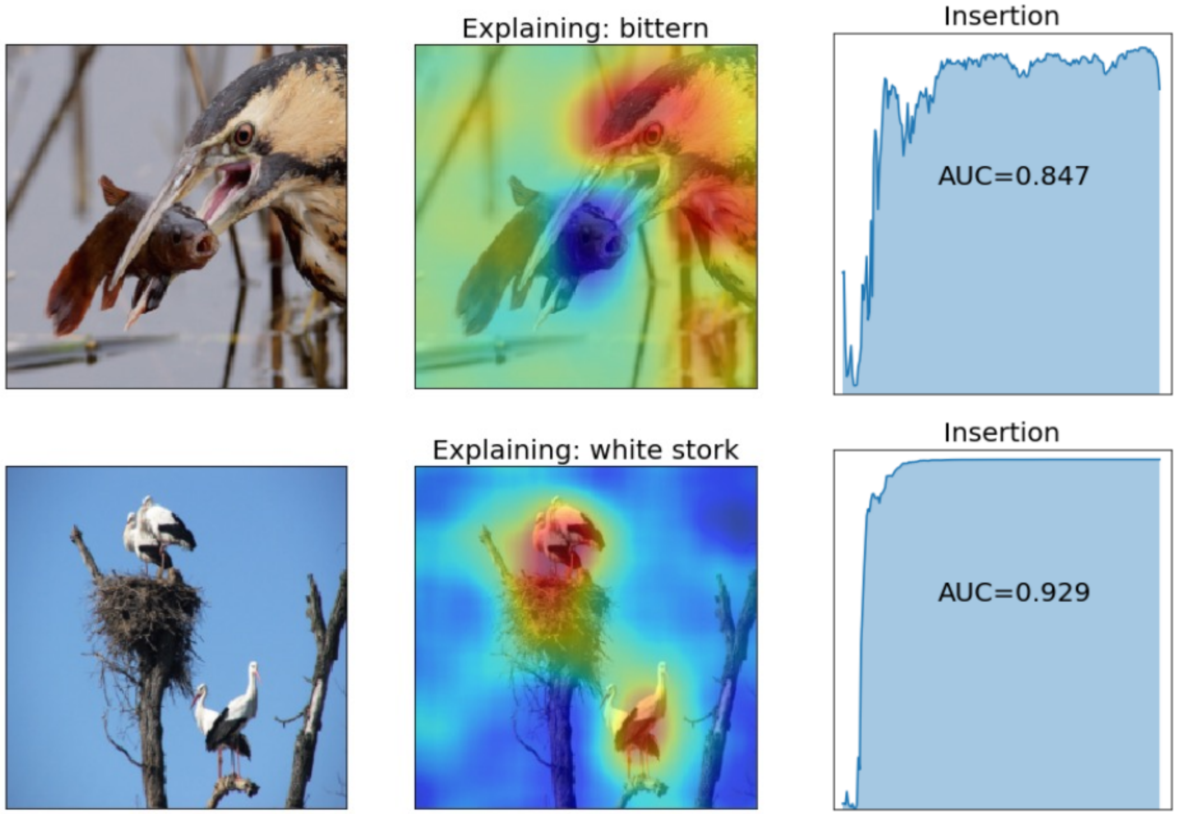
Step 1: Add the highest importance pixel to the mean image and pass it through the network. Note the probability of spoonbill.
Step 2: Add the second highest importance pixel to the image in Step 1 and pass the image through the network. Note the probability of spoonbill.
Step 3: Repeat until all pixels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

Case Study: Intervenability in Interpretability

Evaluation 2: Progressive Pixel-wise Insertion and Deletion

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



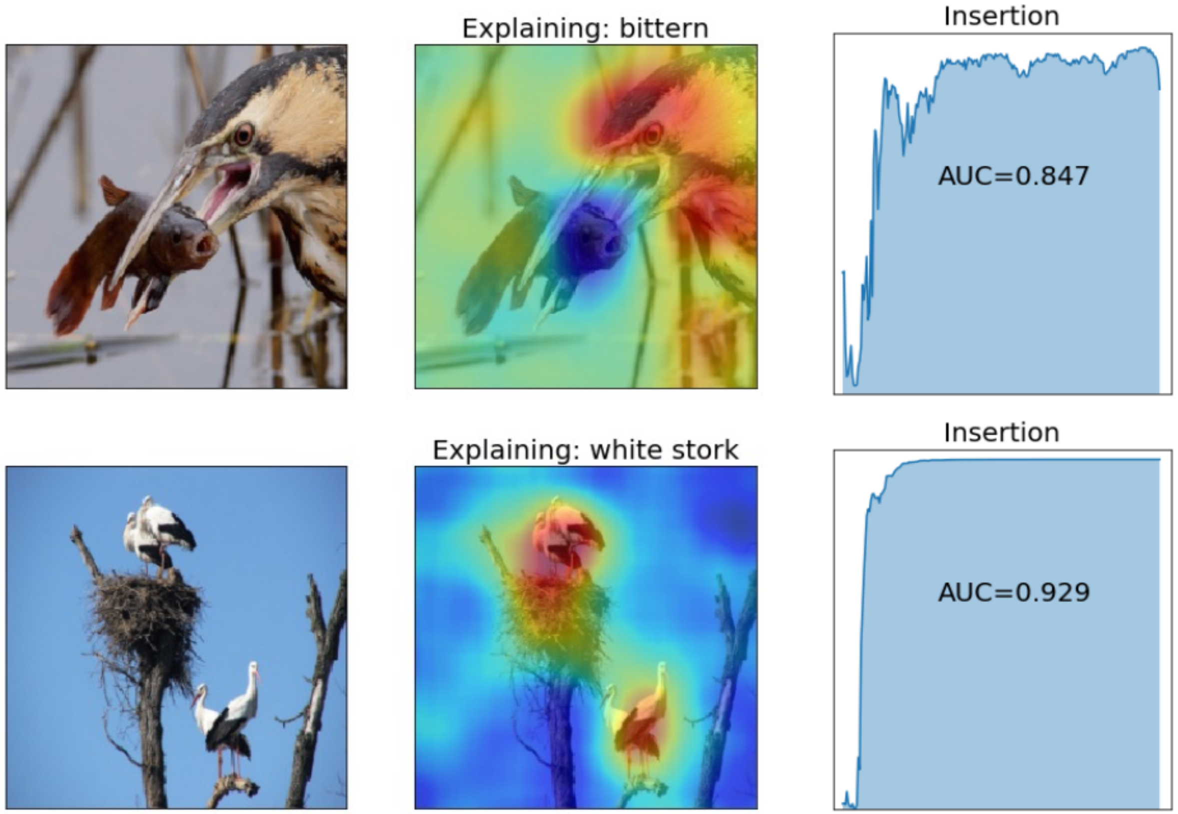
- **Insertion approximates Sufficiency** criterion of a "good" explanation
- **AUC** for a good explanation will be **high**
- **Insertion** encourages **fine-grained explanations** by choosing those heatmaps that select the most relevant pixels

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

Case Study: Intervenability in Interpretability

Evaluation 2: Progressive Pixel-wise Insertion and Deletion

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



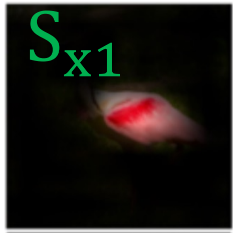
- **However, humans do not "see" in pixels**
- Rather they view scenes in a "**structure-wise**" fashion
- While **heatmap masking** encourages **large explanations**, **pixel-wise masking** encourages **unrealistic and non-human** like explanations

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

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

Three types of Masking:

1. Masking using explanation heatmap
2. Pixel-wise masking using explanation as importance
3. **Structure-wise masking using information encoded in explanation**



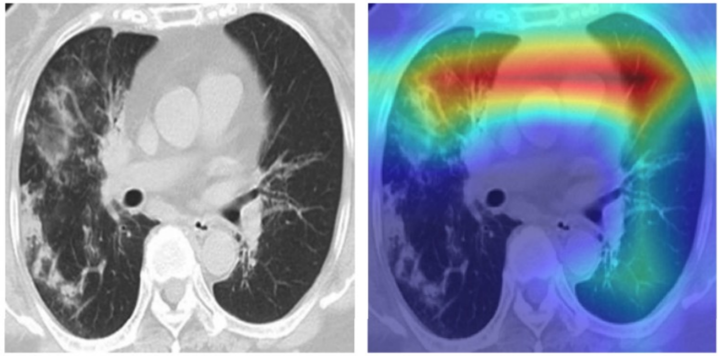
→ Trained Model → Crane

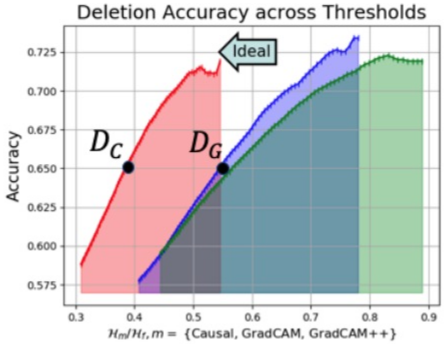
→ Trained Model → Spoonbill

Case Study: Intervenability in Interpretability

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

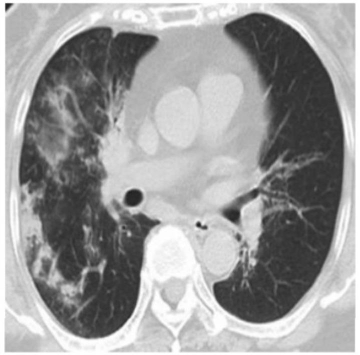
- **CausalCAM encodes dense structure-rich features in lesser bits, that aid accuracy**

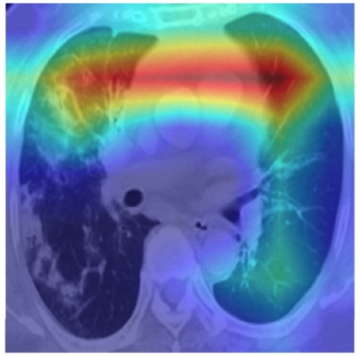
Prabhushankar, Mohit, and Ghassan AlRegib. "Extracting causal visual features for limited label classification." *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021.

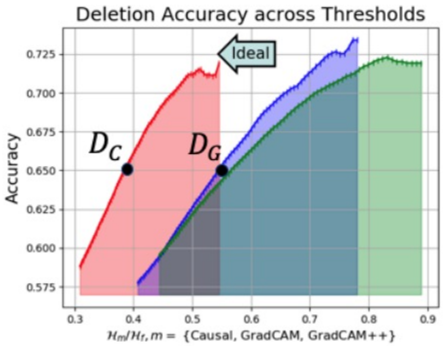
Case Study: Intervenability in Interpretability

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

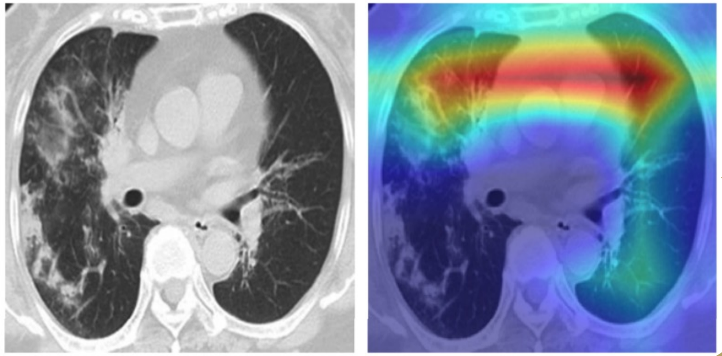
Step 1: Choose a threshold in the explanation (say 0.1) and delete (mask) all the pixels in the original image below the threshold. Pass the masked image through the network and note the change in prediction (if any)

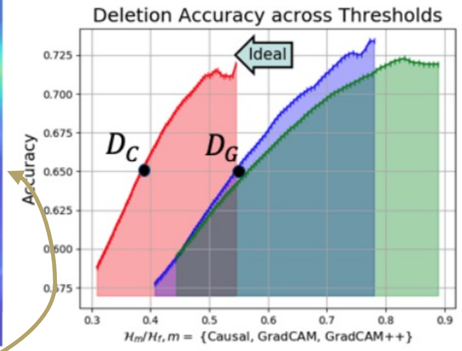
Prabhushankar, Mohit, and Ghassan AlRegib. "Extracting causal visual features for limited label classification." *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021.

Case Study: Intervenability in Interpretability

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)

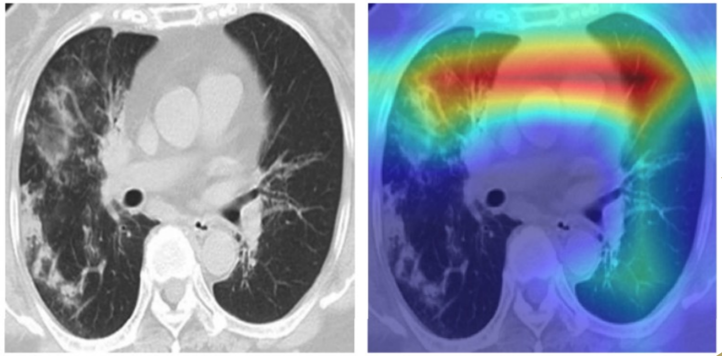
Step 2: Calculate the Huffman code for the original and the masked image. The ratio between the codes of masked and original image is taken on the x-axis and the corresponding accuracy across all images is shown on the y-axis

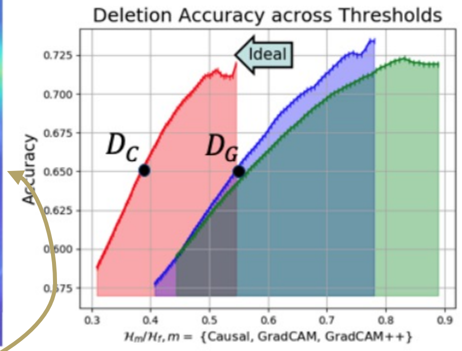
Prabhushankar, Mohit, and Ghassan AlRegib. "Extracting causal visual features for limited label classification." *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021.

Case Study: Intervenability in Interpretability

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

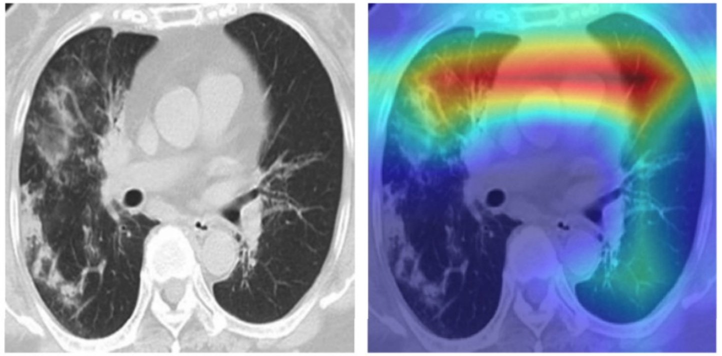
Step 1: Choose a threshold in the explanation (say 0.1) and delete (mask) all the pixels in the original image below the threshold. Pass the masked image through the network and note the change in prediction (if any)
Step 2: Calculate the Huffman code for the original and the masked image. The ratio 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

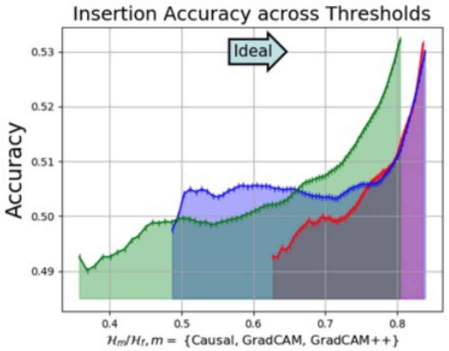
Prabhushankar, Mohit, and Ghassan AlRegib. "Extracting causal visual features for limited label classification." *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021.

Case Study: Intervenability in Interpretability

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.

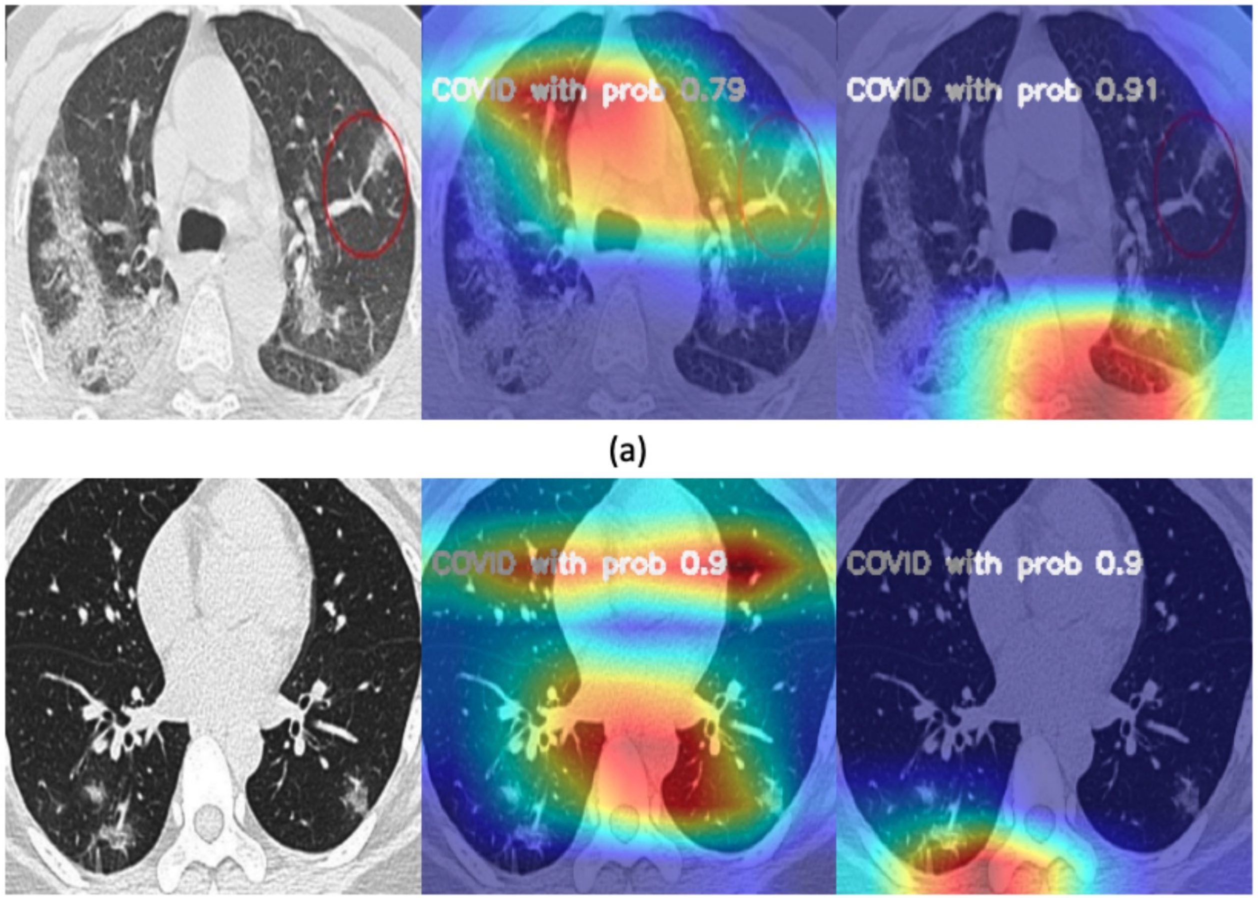
OLIVES @GeorgiaTech

Georgia Tech

Case Study: Intervenability in Interpretability

Evaluation 3: Progressive Structure-wise Insertion and Deletion

Structure-wise insertion and deletion can sometimes promote adversarial explanations



- Best explanations according to structure-wise insertion and deletion.
- Corroborated by high probabilities

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

Case Study: Intervenability in Interpretability

Pros and Cons

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 and dispersed 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 (and sometimes adversarial) explanations without considering context information

Case Study: Intervenability in Interpretability

Challenges in Intervenability

The amenability of neural network decisions to human interventions

- Hence, there is **no single-best interventional** strategy
- Choosing the **right** intervention is still an **art**

Challenges:

- **Choosing the type of Intervention: Explanation Evaluation**
- Residuals of Interventions: Uncertainty

Case Study: Intervenability in Interpretability

Challenges in Intervenability

The amenability of neural network decisions to human interventions

- Hence, there is **no single-best interventional** strategy
- Choosing the **right** intervention is still an **art**

Challenges:

- Choosing the type of Intervention: Explanation Evaluation
- **Residuals of Interventions: Uncertainty**

VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability

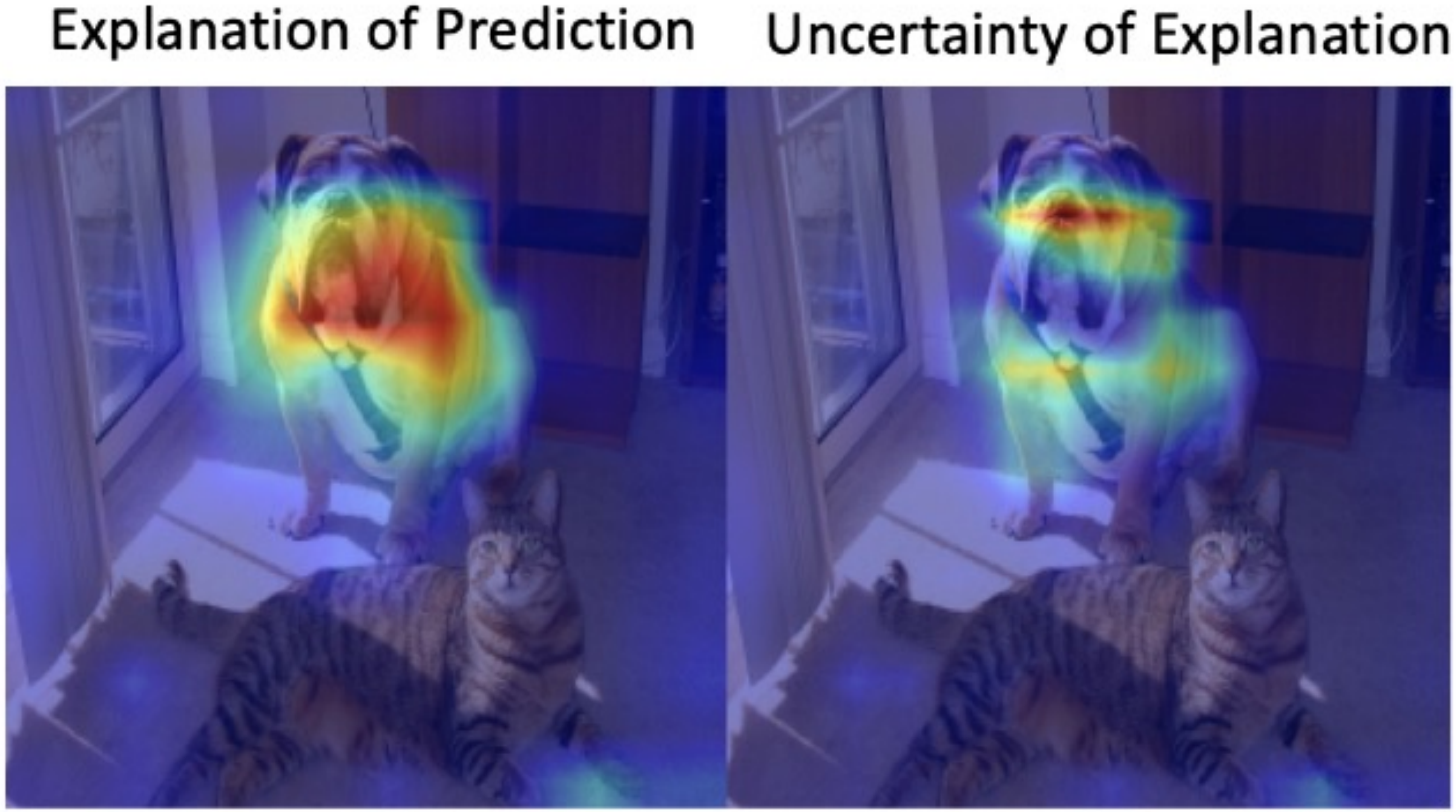

Mohit Prabhushankar, PhD
Postdoc

Ghassan AlRegib, PhD
Professor

Case Study: Intervenability in Interpretability
Predictive Uncertainty in Explanations

Explanatory techniques have predictive uncertainty

Why Bullmastiff?



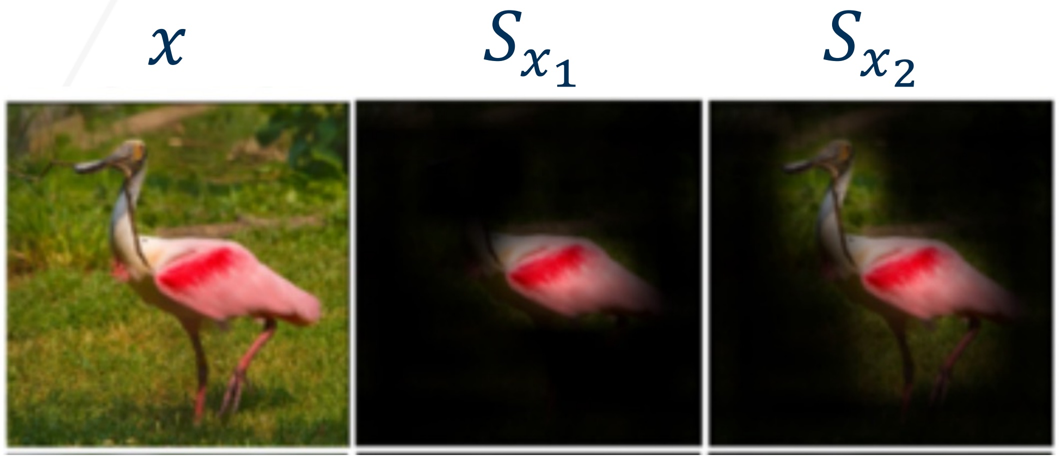
Uncertainty in answering Why Bullmastiff?

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.

Case Study: Intervenability in Interpretability

Predictive Uncertainty

Uncertainty due to variance in prediction when model is kept constant



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

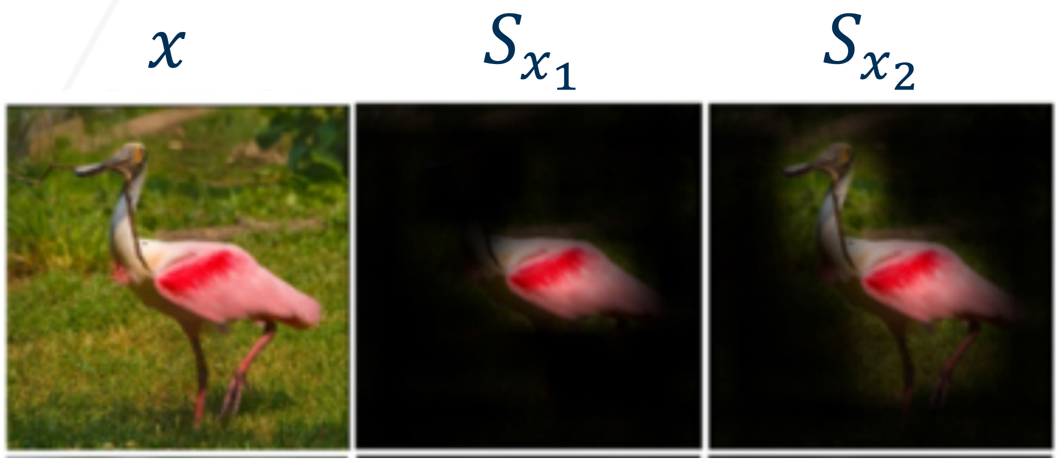
y = Prediction
$V[y]$ = Variance of prediction (Predictive Uncertainty)
S_x = Subset of data (Some intervention)
$E(Y|S_x)$ = Expectation of class given a subset
$V(Y|S_x)$ = Variance of class given all other residuals

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.

Case Study: Intervenability in Interpretability

Visual Explanations (partially) reduce Predictive Uncertainty

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



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

zero

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

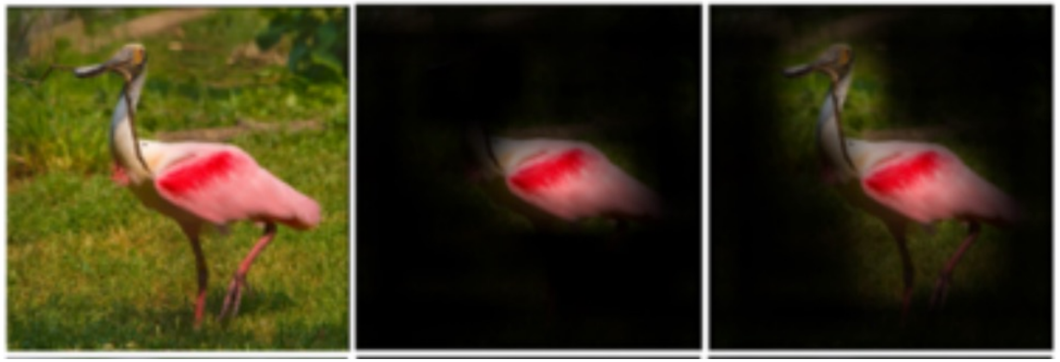
Network evaluations have nothing to do with human Explainability!

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.

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

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.

Case Study: Intervenability in Interpretability

Predictive Uncertainty in Explanations is the Residual

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

x $\quad S_{x_1}$ $\quad S_{x_2}$

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

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

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

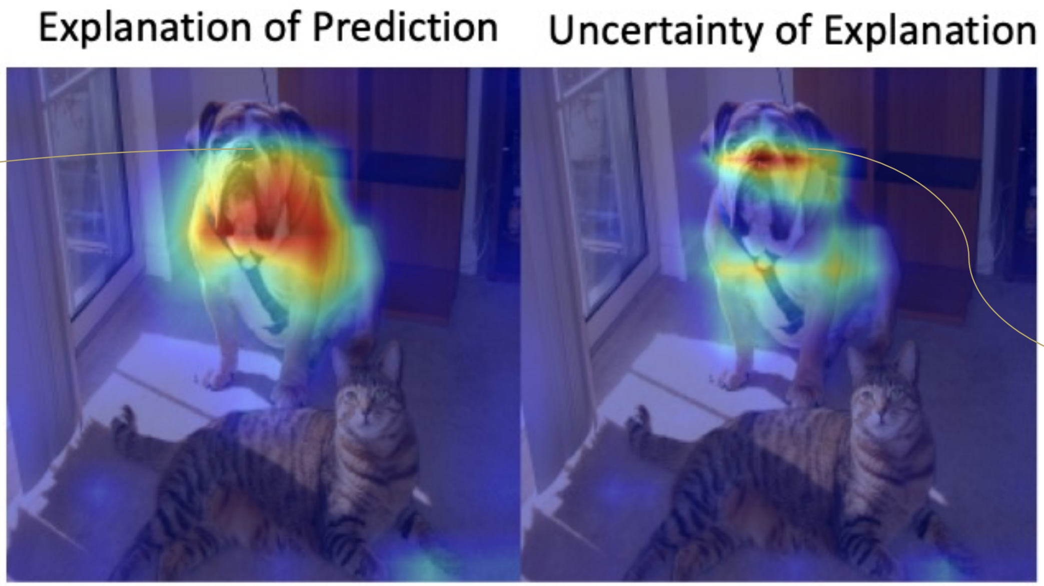
M. Prabhushankar, and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," *Journal of Selected Topics in Signal Processing*, submitted on Aug. 27, 2023.

Case Study: Intervenability in Interpretability

Predictive Uncertainty in Explanations is the Residual

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

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

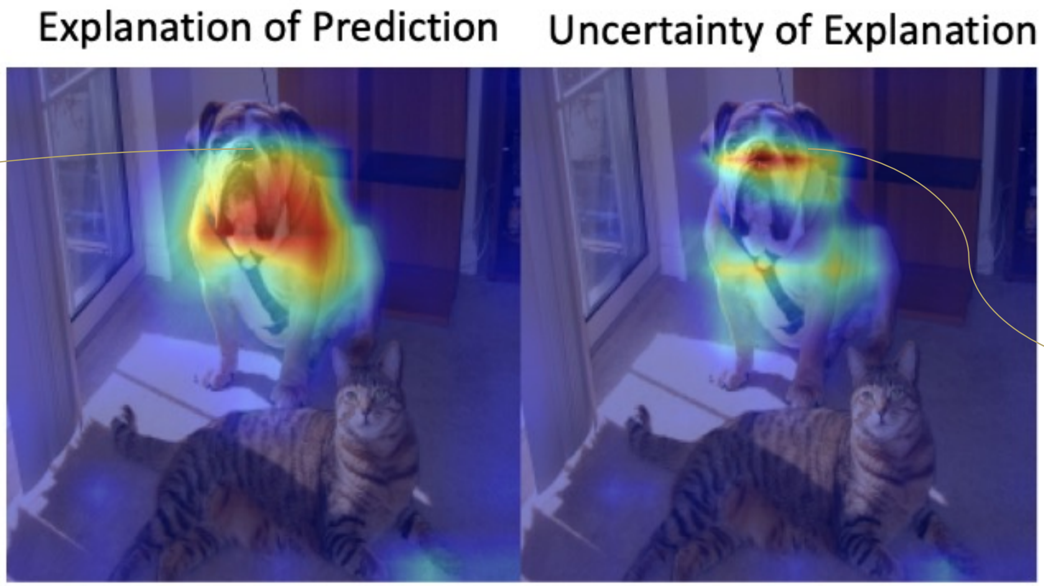
M. Prabhushankar, and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," *Journal of Selected Topics in Signal Processing*, submitted on Aug. 27, 2023.

Case Study: Intervenability in Interpretability

Predictive Uncertainty in Explanations is the Residual

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

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!

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.

Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability

Contrastive explanations are an intelligent way of obtaining other subsets

S_{x_1} S_{x_2} .... S_{x_N}

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

Make it finite by only considering the subsets that change y

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

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.

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

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.

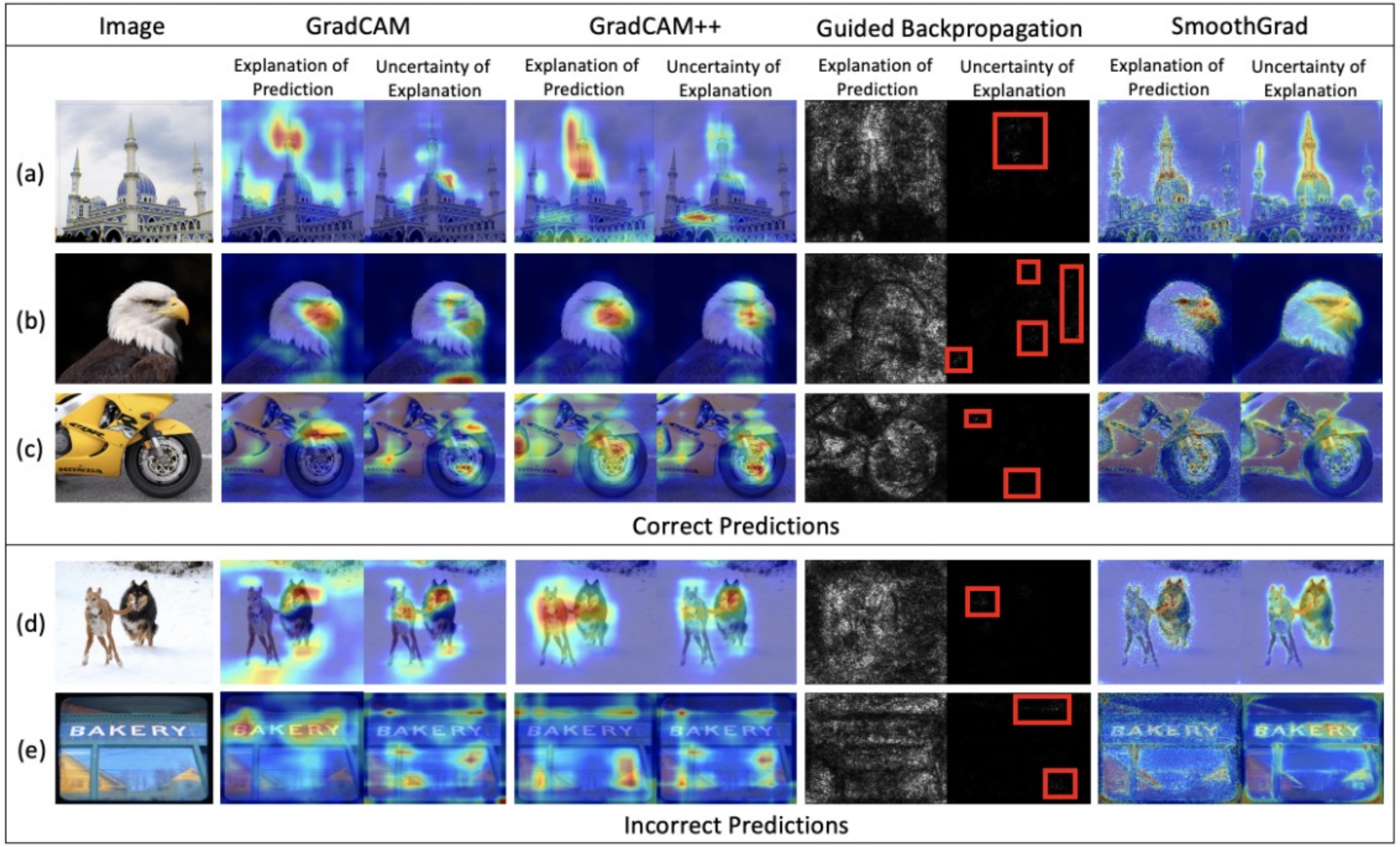
OLIVES @GeorgiaTech

Georgia Tech

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

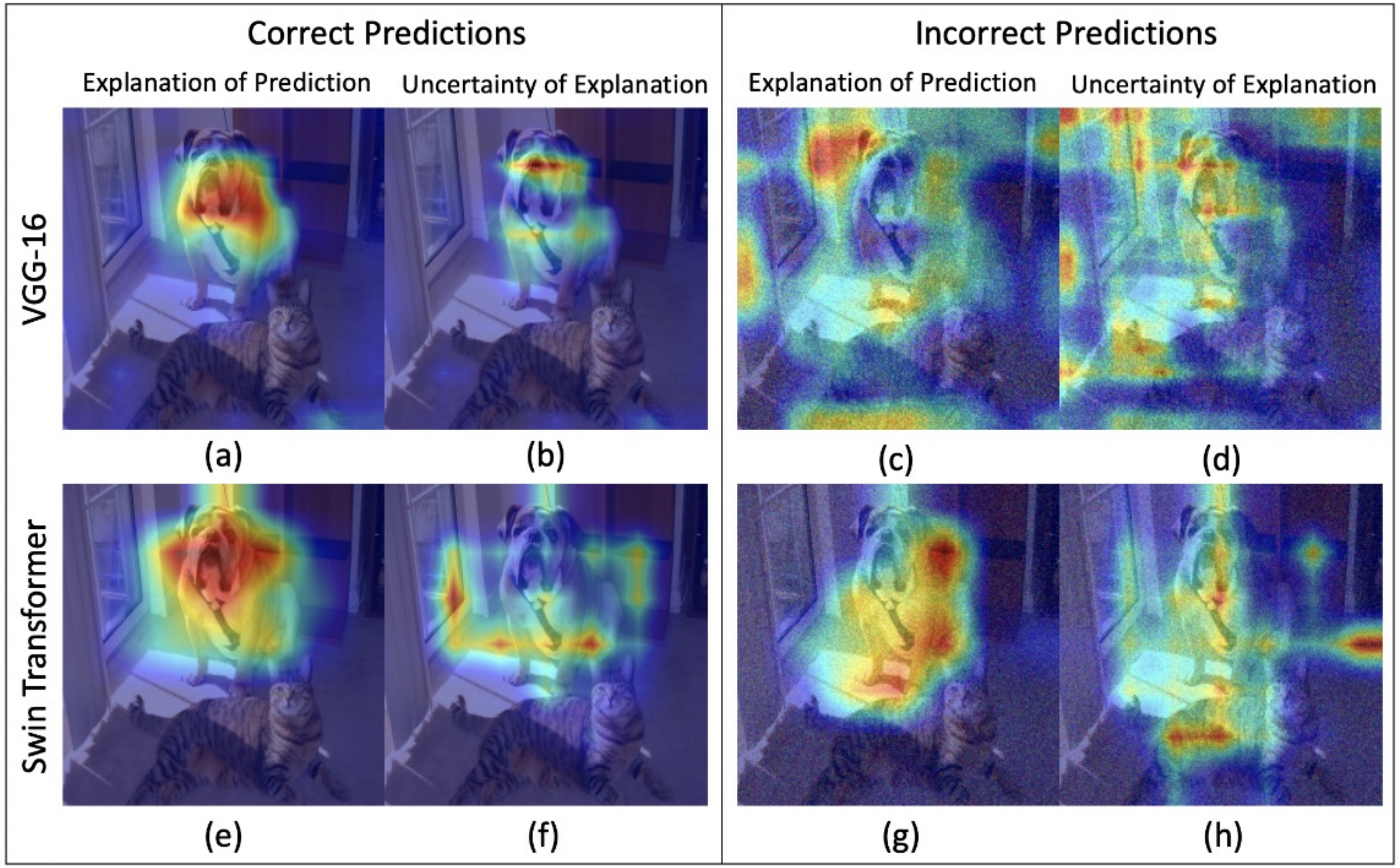
Higher the IoU, higher the uncertainty in explanation (or less trustworthy is the prediction)

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.

Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability: SNR

Explanation and uncertainty are dispersed under noise (under low prediction confidence)



Objective Metric: Signal to Noise Ratio of the Uncertainty map

Higher the SNR of uncertainty, more is the dispersal (or less trustworthy is the prediction)

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.

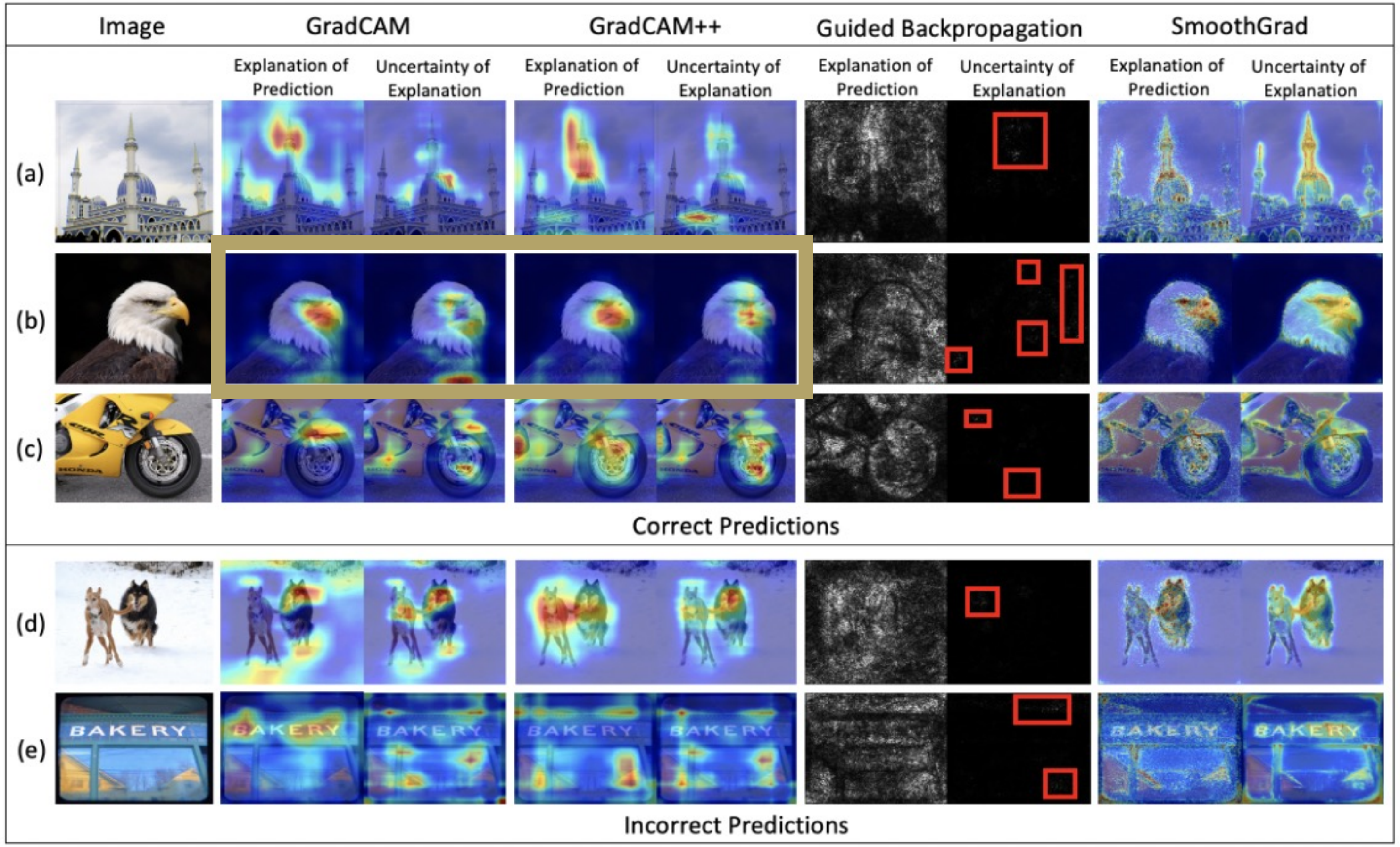
OLIVES @GeorgiaTech

Georgia Tech

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

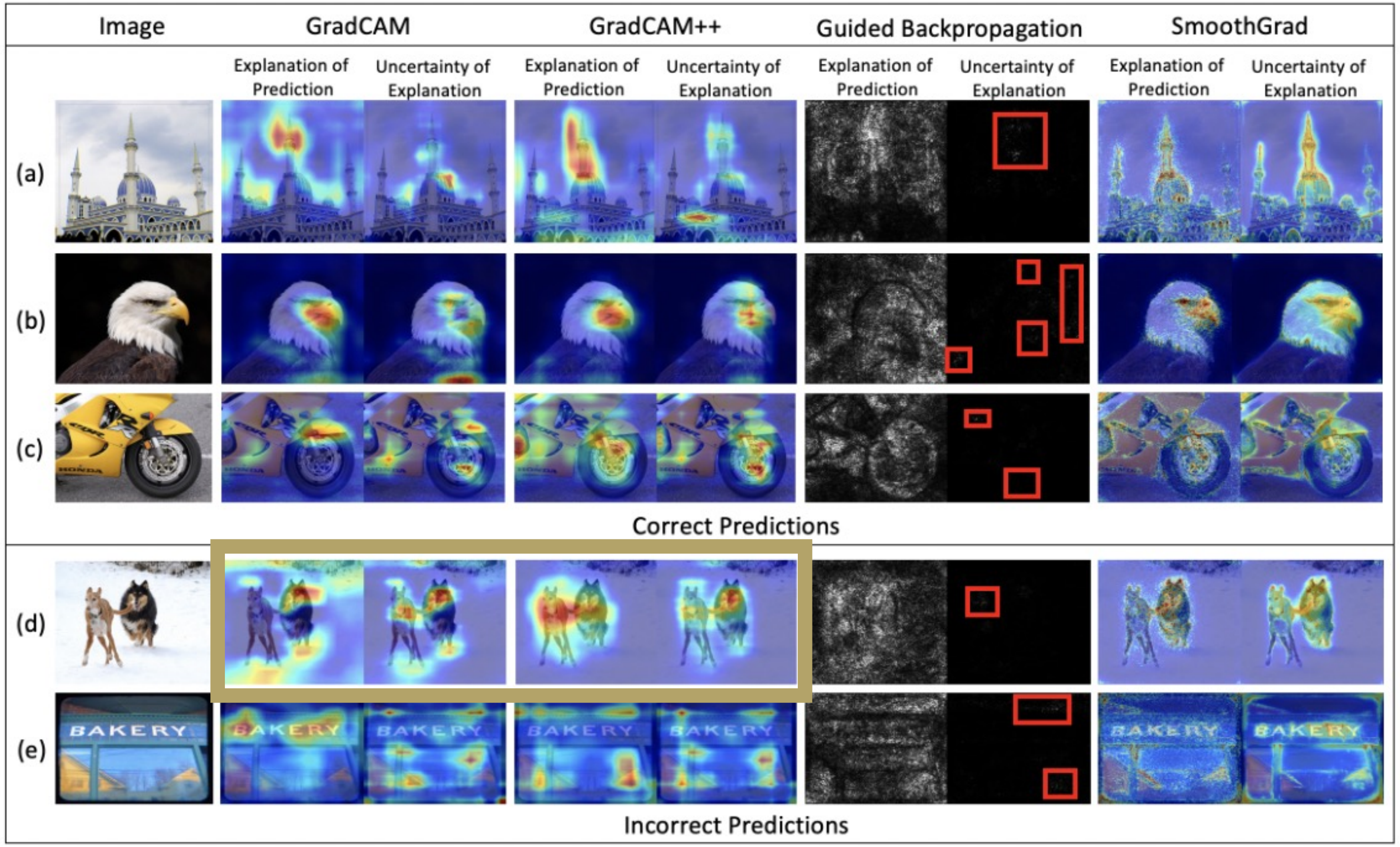
Higher the IoU, higher the uncertainty in explanation (or less trustworthy is the prediction)

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.

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)

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.

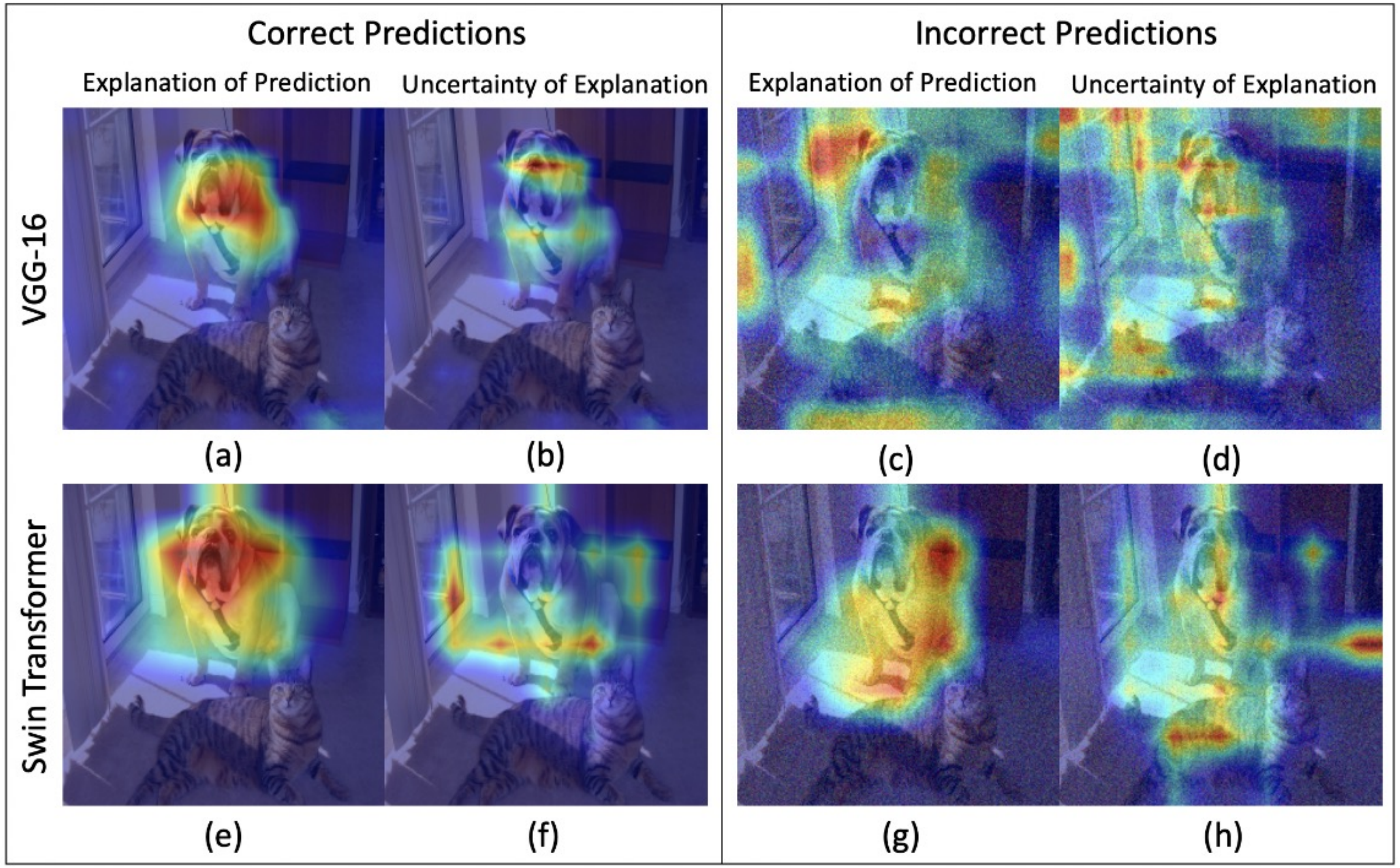
OLIVES @GeorgiaTech

Georgia Tech

Case Study: Intervenability in Interpretability

Quantifying Interventions in Explainability: SNR

Explanation and uncertainty are dispersed under noise (under low prediction confidence)



Objective Metric 2: Signal to Noise Ratio of the Uncertainty map

Higher the SNR of uncertainty, more is the dispersal (or less trustworthy is the prediction)

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.

Case Study: Intervenability in Interpretability

Challenges in Intervenability

The amenability of neural network decisions to human interventions

- **Not choosing interventions** causes **uncertainty** in the chosen interventions
- **Residuals** must be **analyzed** intelligently to **'trust or not to trust'** predictions at inference
- Gradients quantify residual uncertainty

Challenges:

- Choosing the type of Intervention: Explanation Evaluation
- **Residuals of Interventions: Uncertainty**

Intervenability

Through the Human Glass

The amenability of neural network decisions to human interventions

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

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, *109*(5), 612-634.

CURE-TSD: Challenging Unreal and Real Environments for Traffic Sign Detection

Data Characteristics:

- 49 real and virtual sequences
- 300 frames in each sequence
- 12 different challenges including decolorization, codec error, lens blur etc.
- 5 progressively increasingly levels in each challenge
- **Goal**: Detect and localize traffic signs

Temel, Dogancan, et al. "Cure-tsd: Challenging unreal and real environments for traffic sign detection." *IEEE Transactions on Intelligent Transportation Systems* (2017).

OLIVES @GeorgiaTech

Georgia Tech

CURE-TSR: Challenging Unreal and Real Environments for Traffic Sign Recognition

Data Characteristics:

- 2 million real and virtual traffic sign images
- 14 Traffic signs including common signs like stop, no-right, no-left etc. and uncommon signs like goods-vehicles, priority lanes etc.
- 12 different challenges including decolorization, codec error, lens blur etc.
- 5 progressively increasingly levels in each challenge

D. Temel, G. Kwon*, M. Prabhushankar*, and G. AlRegib, “CURE-TSR: Challenging unreal and real environments for traffic sign recognition,” in *Neural Information Processing Systems (NIPS) Workshop on Machine Learning for Intelligent Transportation Systems,* Long Beach, U.S., December 2017, (*: equal contribution)

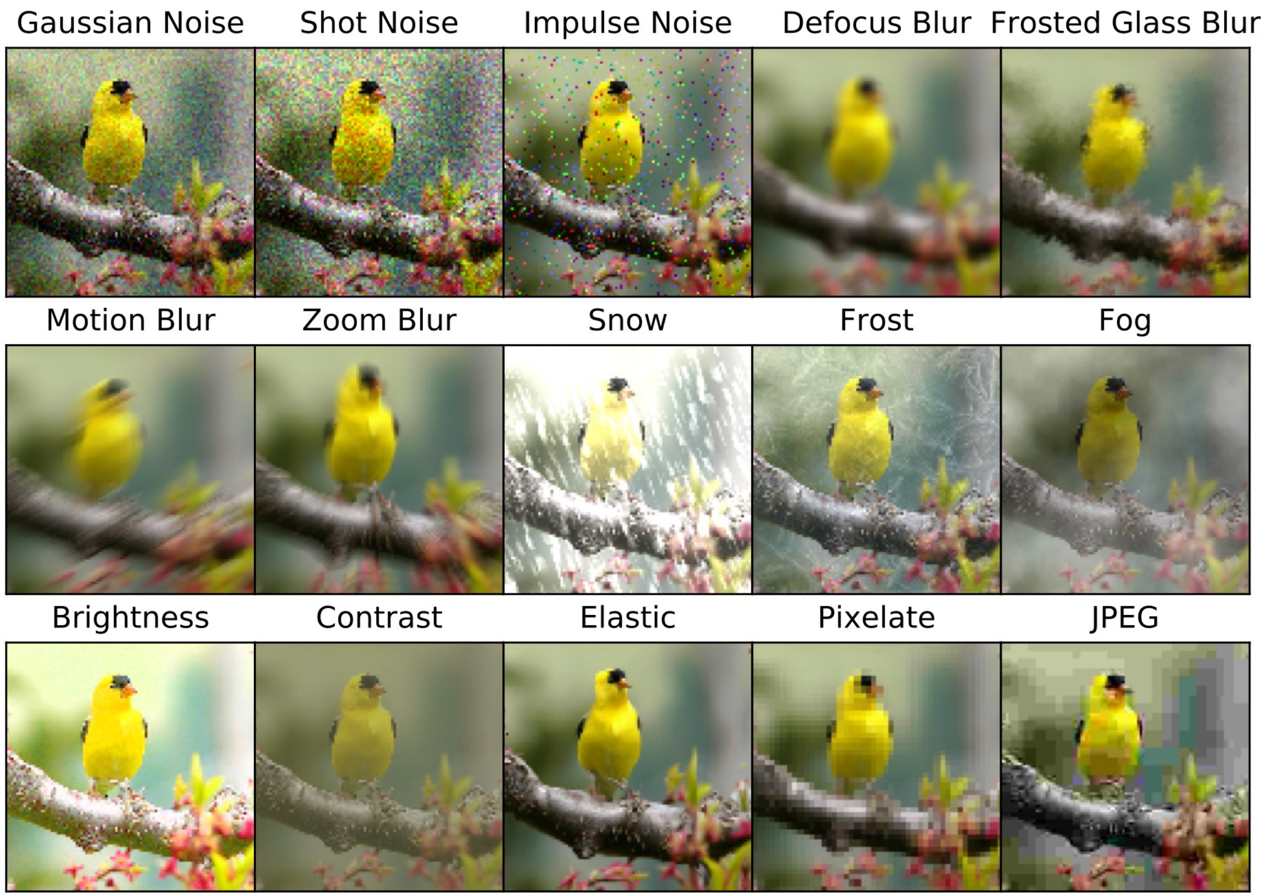
Intervenability in Benchmarking

Recognition

ImageNet-C: ImageNet-Corruptions

Data Characteristics:

- 3.75 million images
- 15 different challenges including decolorization, codec error, lens blur etc. for testing
- 4 different challenges for validation and training
- 5 progressively increasingly levels in each challenge
- **Goal**: Recognize 1000 classes from ImageNet using pretrained networks



Hendrycks, Dan, and Thomas Dietterich. "Benchmarking neural network robustness to common corruptions and perturbations." *arXiv preprint arXiv:1903.12261* (2019).

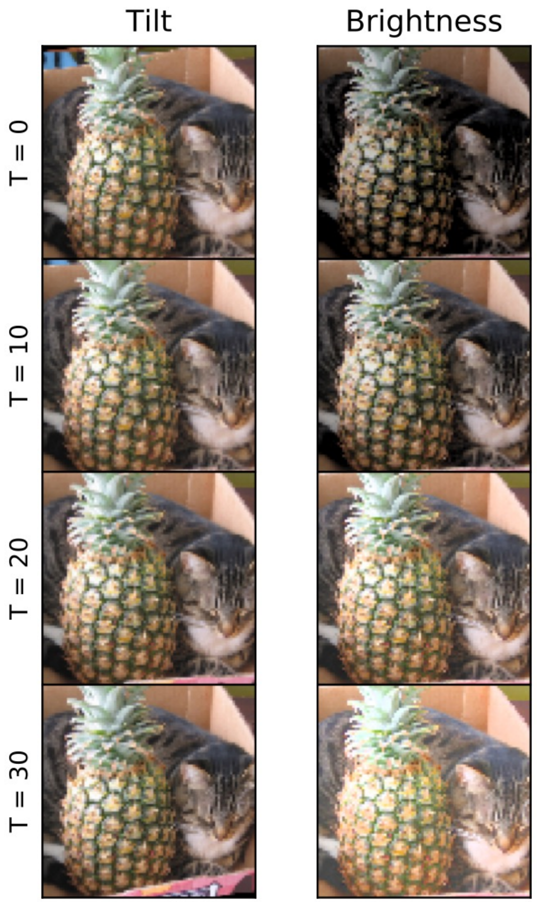
OLIVES @GeorgiaTech

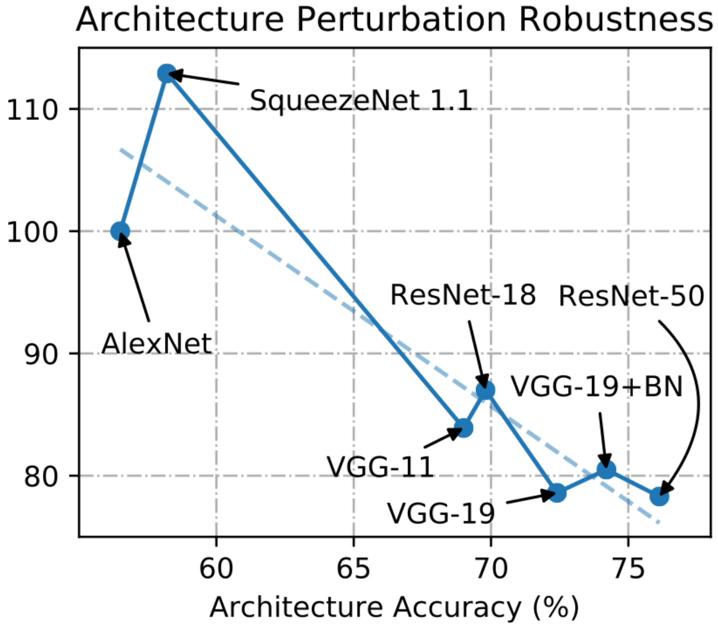
Georgia Tech

ImageNet-P: ImageNet-Perturbations

Data Characteristics:

- 5 million images
- 100 perturbations of 50000 images
- 10 frames of algorithmically generated perturbations for each image in ImageNet validation testset
- 10 common perturbations including brightness, tilt, motion etc.





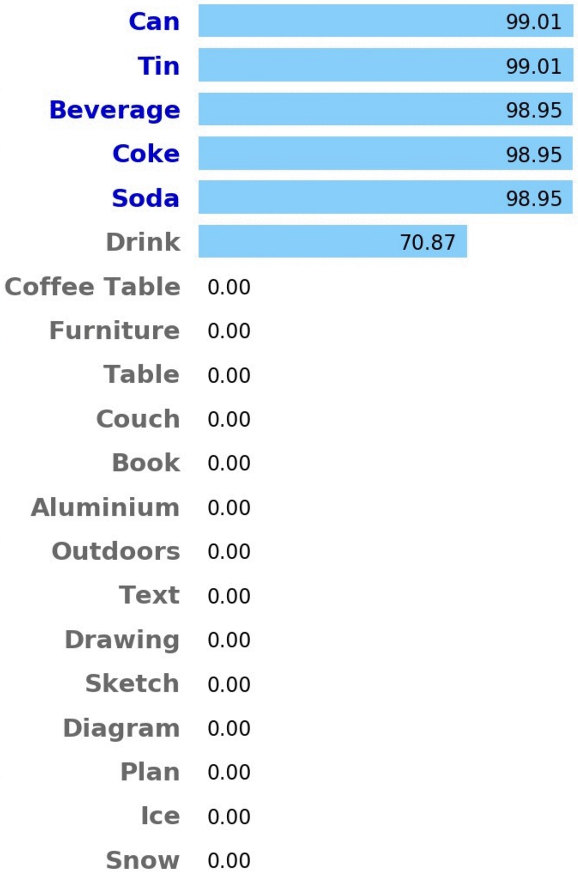
Hendrycks, Dan, and Thomas Dietterich. "Benchmarking neural network robustness to common corruptions and perturbations." *arXiv preprint arXiv:1903.12261* (2019).

CURE-OR: Challenging Unreal and Real Environments for Object Recognition

Data Characteristics:

- 1 million images
- 100 common household objects and 10000 images per object
- 5 backgrounds, 5 object orientations, 5 devices, and 78 challenging conditions
- **Goal**: To recognize and retrieve the same object across backgrounds, orientations, devices, and challenging conditions

Challenge Type: **None**



D. Temel*, J. Lee*, and G. AlRegib, “CURE-OR: Challenging unreal and real environments for object recognition,” IEEE International Conference on Machine Learning and Applications, Orlando, Florida, USA, December 2018, (*: equal contribution)

OLIVES @GeorgiaTech

Georgia Tech