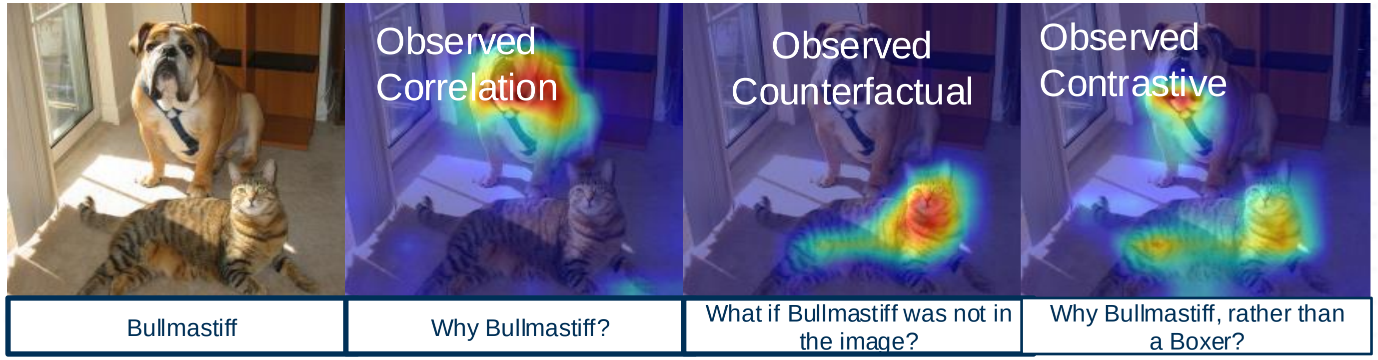
ECE 4252/8803: Fundamentals of Machine Learning (FunML)
Fall 2024

Lecture 23: Explainability Paradigms and Evaluation



OLIVES @GeorgiaTech

Georgia Tech

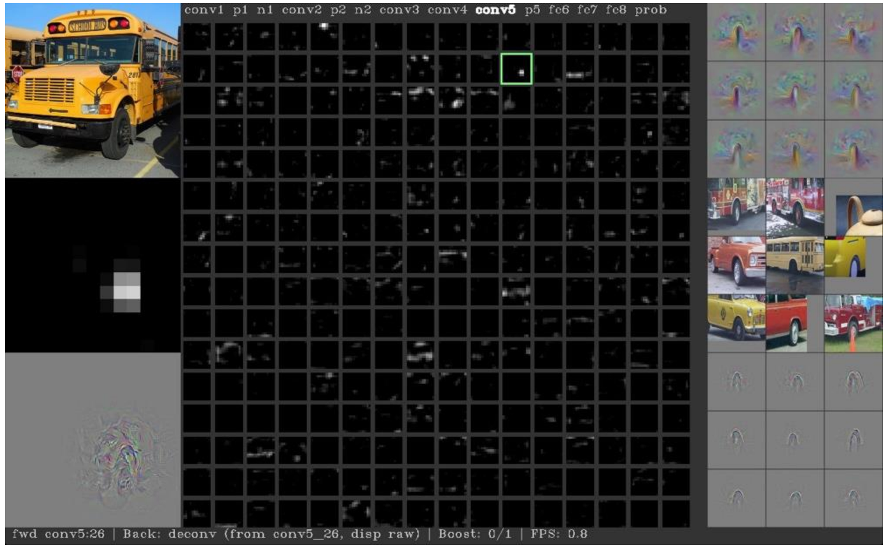
- What constitutes an explanation?
- What makes some explanations better than the others?

Explainability

Understanding Explanations

- What constitutes an explanation?
  - Visualizing weights across different layers
- What makes some explanations better than the others?

- What constitutes an explanation?
  - Visualizing weights across different layers
  - Visualizing activations in intermediate layers
- What makes some explanations better than the others?

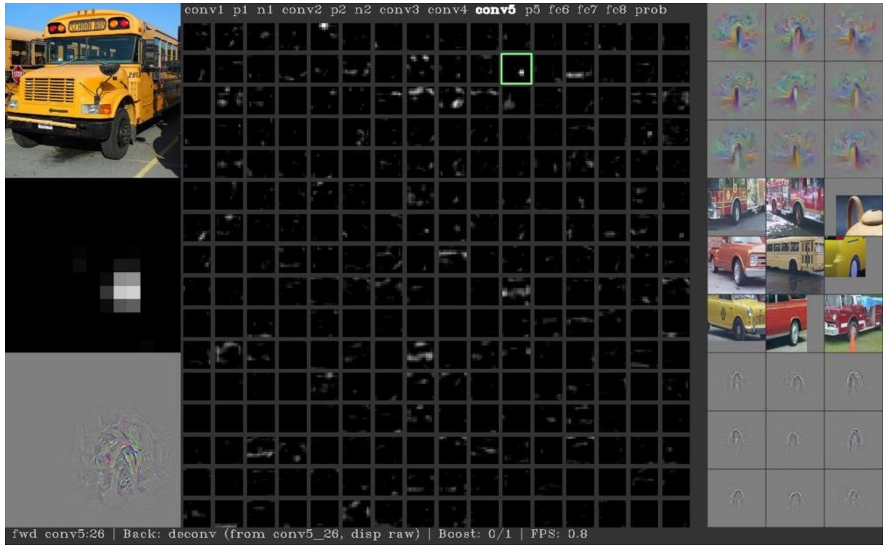


Conv 5

Explainability

Understanding Explanations

- What constitutes an explanation?
  - Visualizing weights across different layers
  - Visualizing activations in intermediate layers
- What makes some explanations better than the others?
  - Subjectively, visualizing activations provide better explanations than weights in intermediate layers



Conv 5

Explainability

Understanding Explanations

- What constitutes an explanation?
  - Visualizing weights across different layers
  - Visualizing activations in intermediate layers
  - Visualizing maximally activation patches
  - Visualizing last layer embedding
  - Nearest neighbor samples
  - Dimensionality reduction
  - Pixel saliency through intervention
  - Pixel saliency through via feature importance
- What makes some explanations better than the others?
  - Subjectively, visualizing activations provide better explanations than weights in intermediate layers
  - Pixel saliency via feature importance generalizes to different tasks and are computationally less expensive

Explainability

Understanding Explanations

- What constitutes an explanation?
  - Visualizing weights across different layers
  - Visualizing activations in intermediate layers
  - Visualizing maximally activation patches
  - Visualizing last layer embedding
  - Nearest neighbor samples
  - Dimensionality reduction
  - Pixel saliency through intervention
  - Pixel saliency through via feature importance
- What makes some explanations better than the others?
  - Subjectively, visualizing activations provide better explanations than weights in intermediate layers
  - Pixel saliency via feature importance generalizes to different tasks and are computationally less expensive

Types of explanations

Evaluating explanations

Overview

In this Lecture..

Explainability

Visualization of Convolutional Neural Networks

Types of Explanations

- Categorization of existing explanations
- Types of explanations
- Indirect and Direct Explanations
- Targeted Explanations
- Explanatory Paradigms

Explanatory Evaluation

Types of Explanations

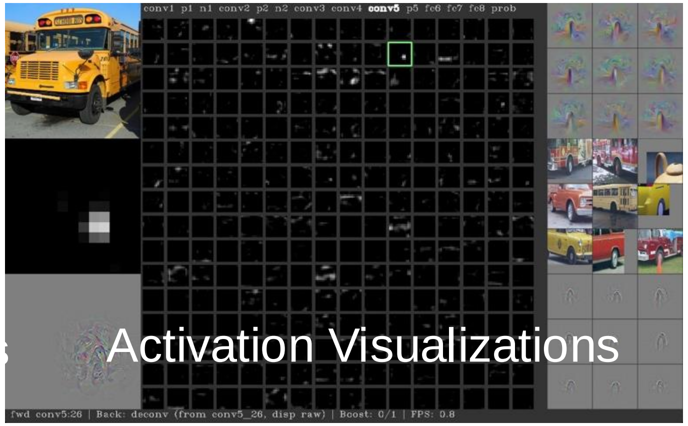
- What constitutes an explanation?
  - Visualizing weights across different layers
  - Visualizing activations in intermediate layers
  - Visualizing maximally activation patches
  - Visualizing last layer embeddings
  - Nearest neighbor samples
  - Dimensionality reduction
  - Pixel saliency through intervention
  - Pixel saliency through feature importance

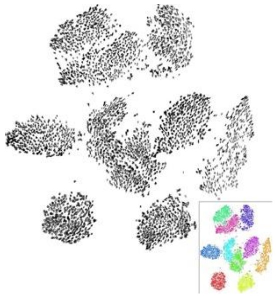
Two types of explanations

Types of Explanations

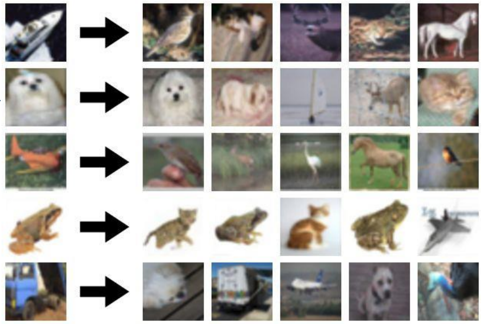
Indirect Explanations

Indirect Explanations	Direct Explanations





Dimensionality Reduction



Nearest Neighbor

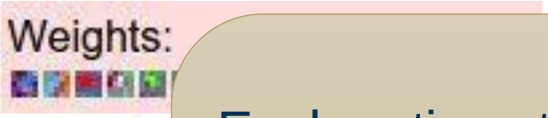
OLIVES @GeorgiaTech Georgia Tech

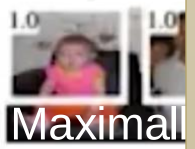
AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks. *arXiv preprint arXiv:2202.11838*.

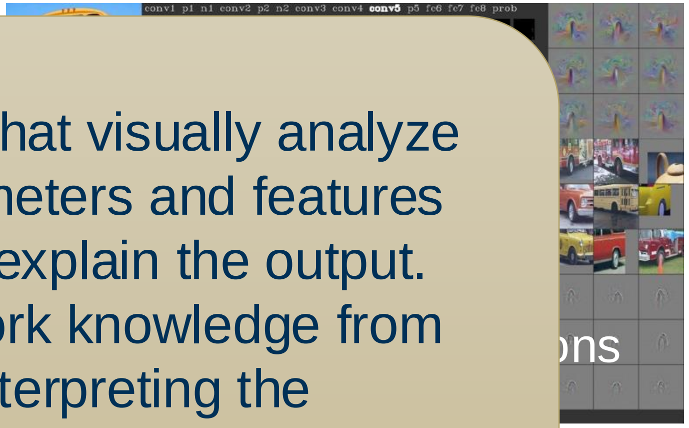
Types of Explanations

Indirect Explanations

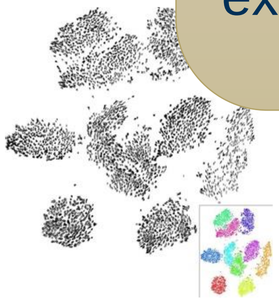
Indirect Explanations	Direct Explanations



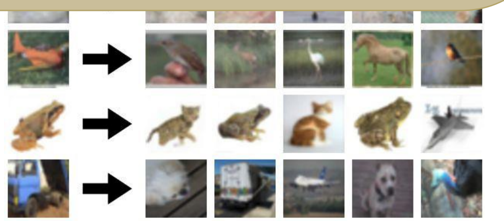




Explanations that visually analyze network parameters and features and indirectly explain the output. Require network knowledge from the humans interpreting the explanations



Dimensionality Reduction



Nearest Neighbor

OLIVES @GeorgiaTech

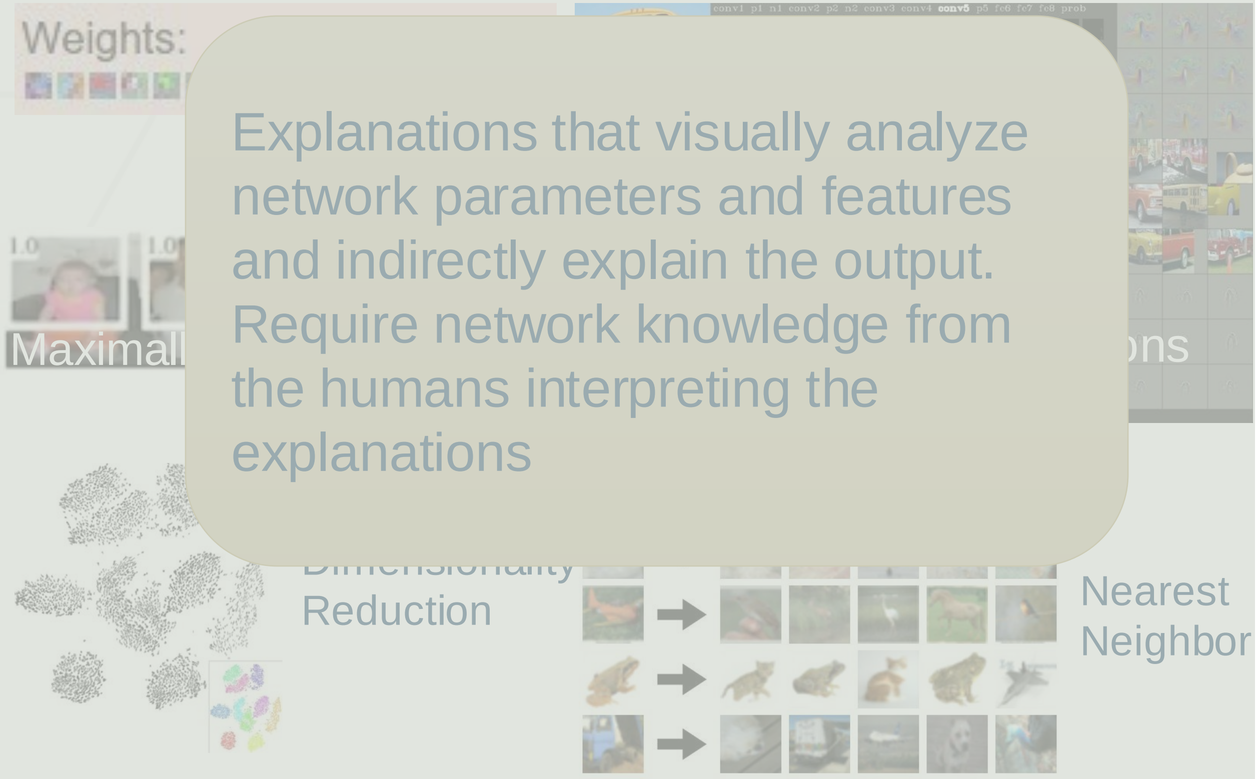
Georgia Tech

AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks. *arXiv preprint arXiv:2202.11838*.

Types of Explanations

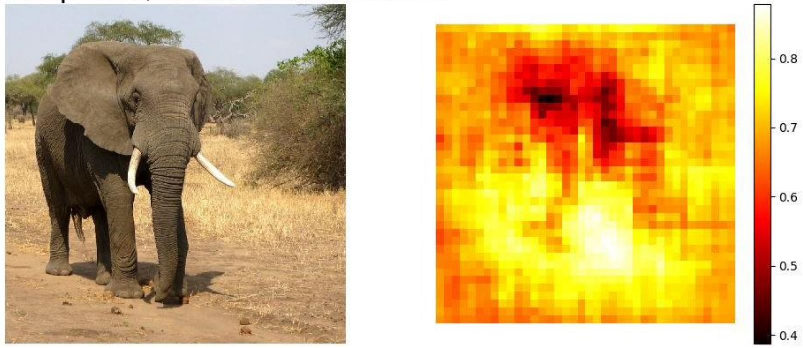
Direct Explanations

Indirect Explanations	Direct Explanations



Explanations that visually analyze network parameters and features and indirectly explain the output. Require network knowledge from the humans interpreting the explanations

African elephant, Loxodonta africana



Saliency via occlusion

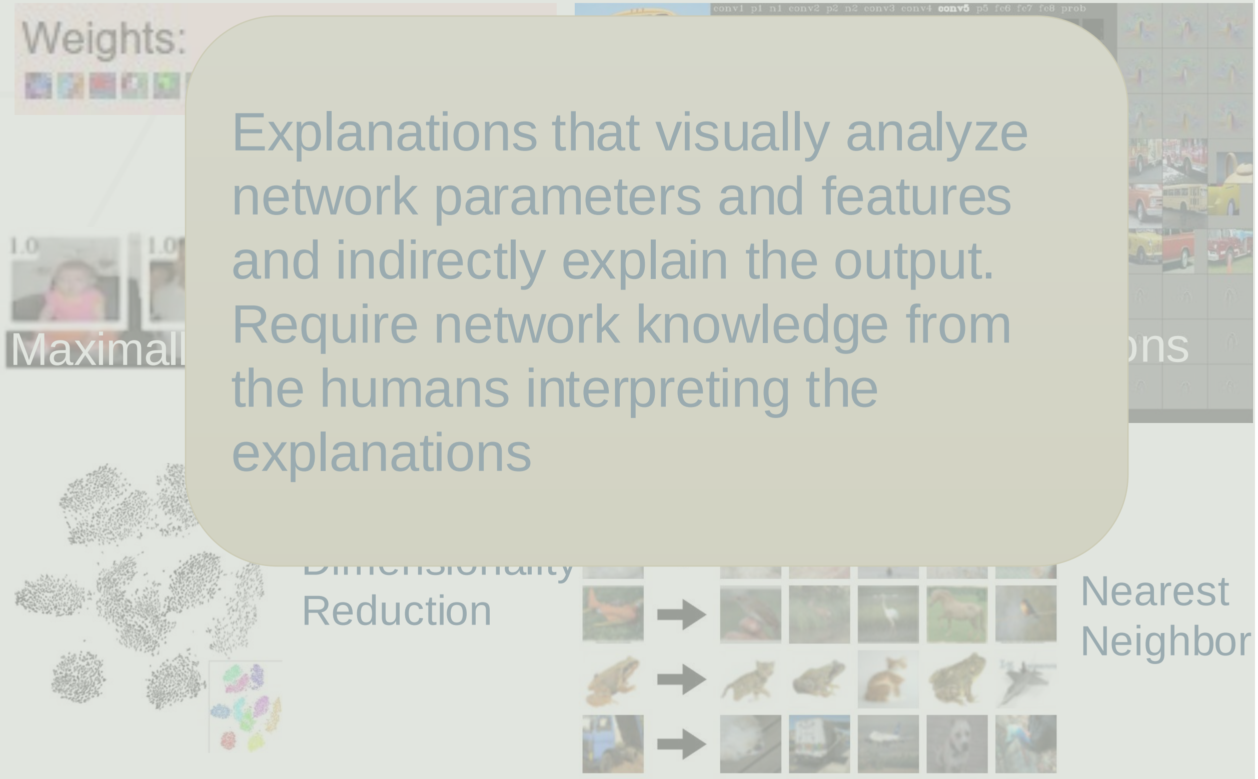
Saliency via Feature Importance

AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks. *arXiv preprint arXiv:2202.11838*.

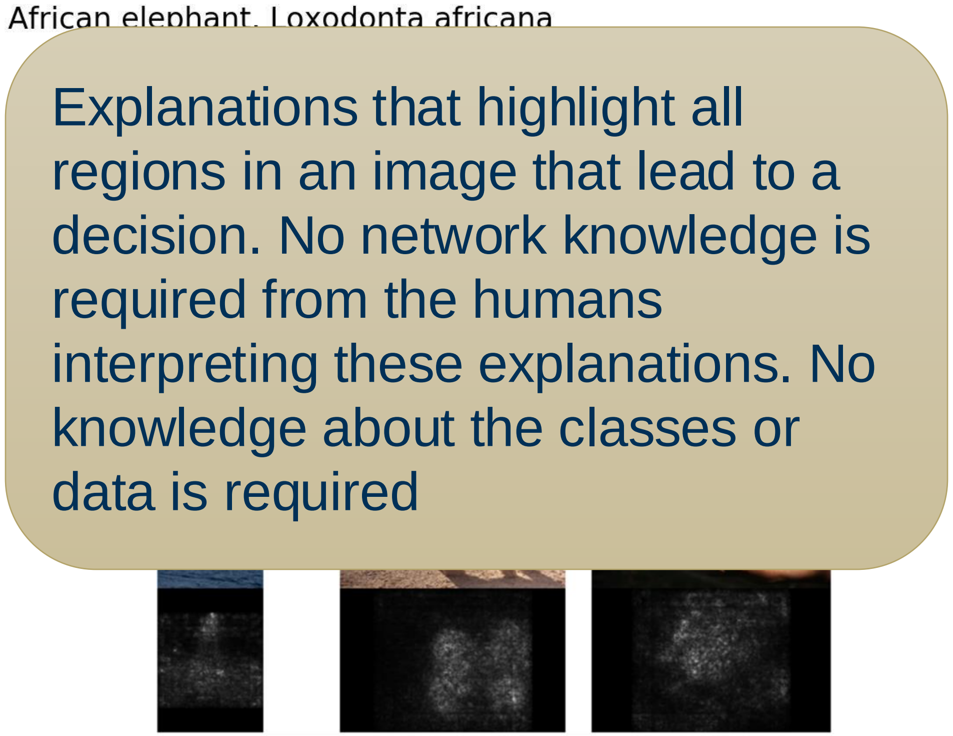
Types of Explanations

Direct Explanations

Indirect Explanations	Direct Explanations



Explanations that visually analyze network parameters and features and indirectly explain the output. Require network knowledge from the humans interpreting the explanations



Explanations that highlight all regions in an image that lead to a decision. No network knowledge is required from the humans interpreting these explanations. No knowledge about the classes or data is required

Saliency via Feature Importance

AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks. *arXiv preprint arXiv:2202.11838*.

Types of Explanations

Direct Explanations

No knowledge of elephants or any other class is required to interpret the heat map!

Indirect Explanations

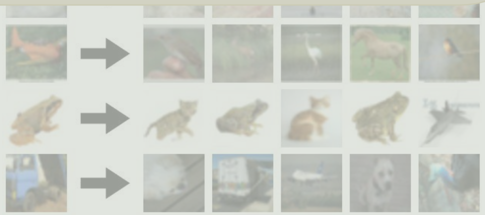
Direct Explanations

Explanations that visually analyze network parameters and features and indirectly explain the output. Require network knowledge from the humans interpreting the explanations

Weights:

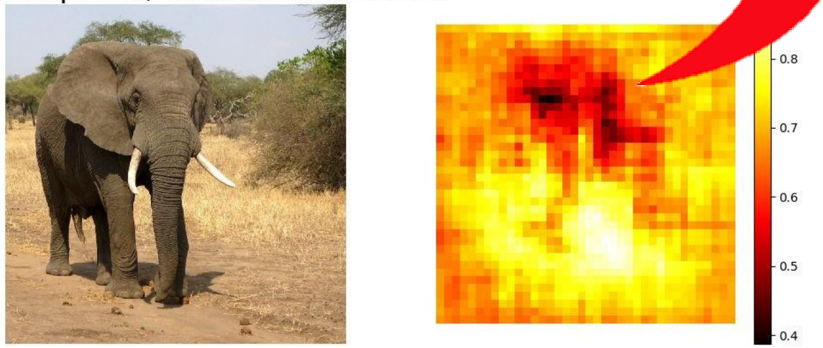
Maximall

Dimensionality Reduction



Nearest Neighbor

African elephant, Loxodonta africana



Saliency via occlusion

Explanations that highlight all regions in an image that lead to a decision. No network knowledge is required from the humans interpreting these explanations. No knowledge about the classes or data is required

AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks. *arXiv preprint arXiv:2202.11838*.

Types of Explanations

Targeted Explanations

Neural Network, $f(\cdot)$

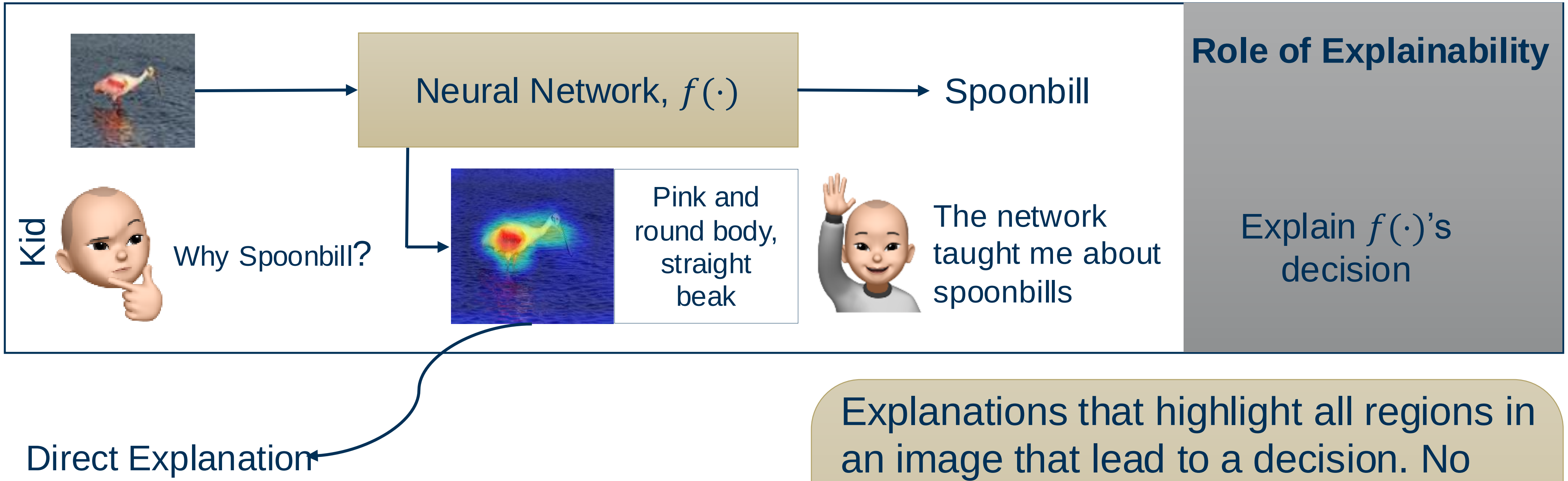
Spoonbill

Consider a trained Neural Network

AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks. *arXiv preprint arXiv:2202.11838*.

Types of Explanations

Targeted Explanations



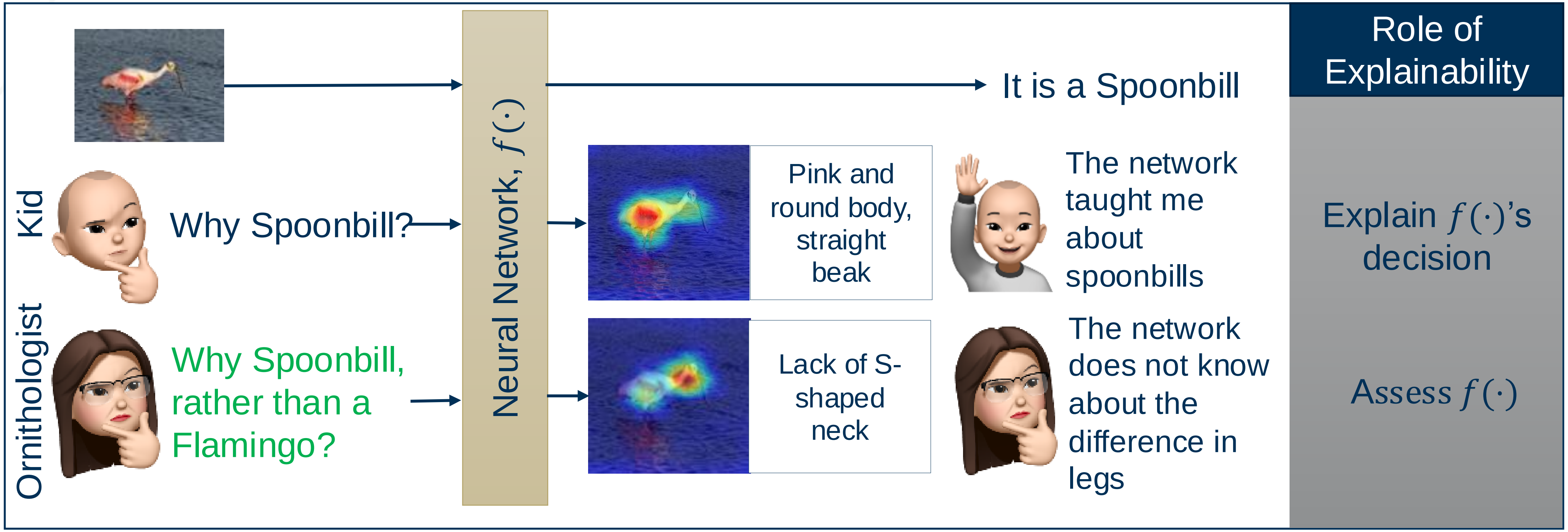
Kid does not have knowledge about Spoonbill!

Explanations that highlight all regions in an image that lead to a decision. No network knowledge is required from the humans interpreting these explanations. No knowledge about the classes or data is required

AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks. *arXiv preprint arXiv:2202.11838*.

Types of Explanations

Targeted Explanations

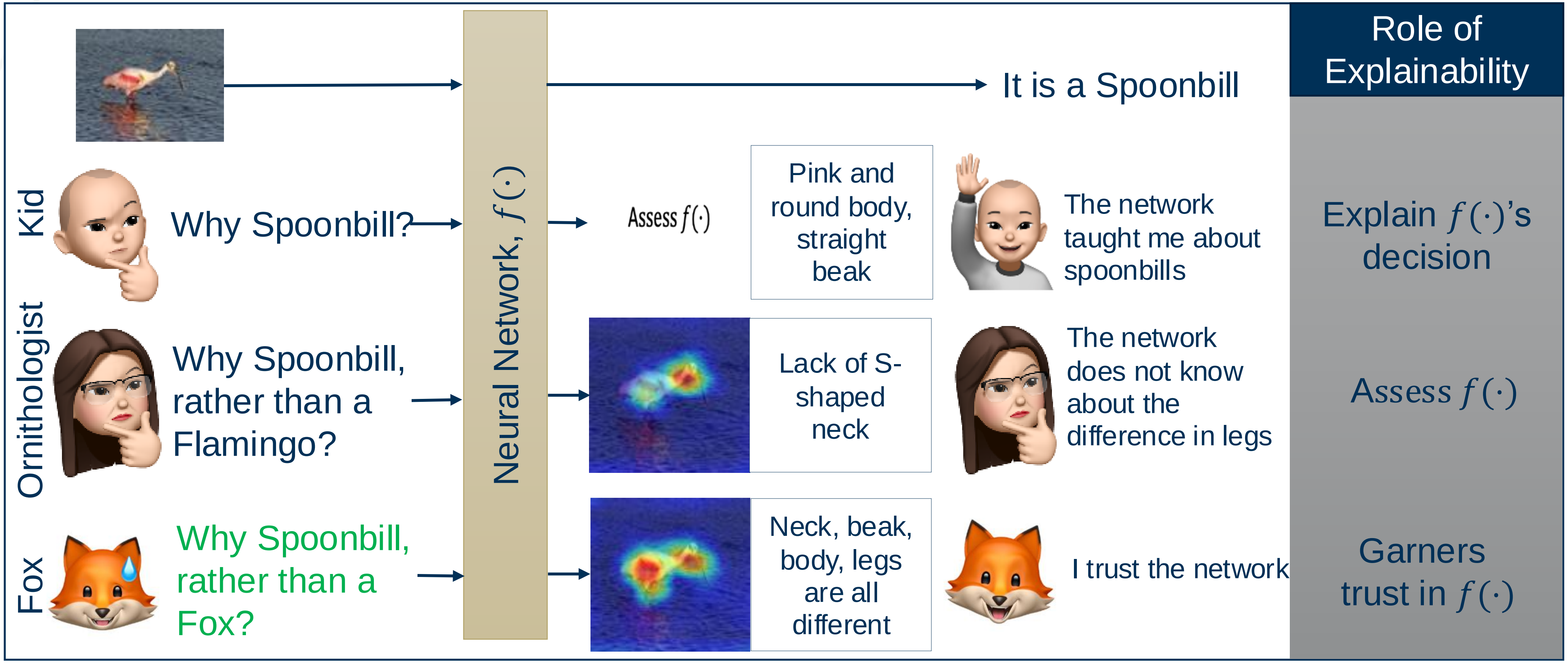


The ornithologist has knowledge of Spoonbills and Flamingos and uses this knowledge to ask **targeted** questions!

AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks. *arXiv preprint arXiv:2202.11838*.

Types of Explanations

Targeted Explanations

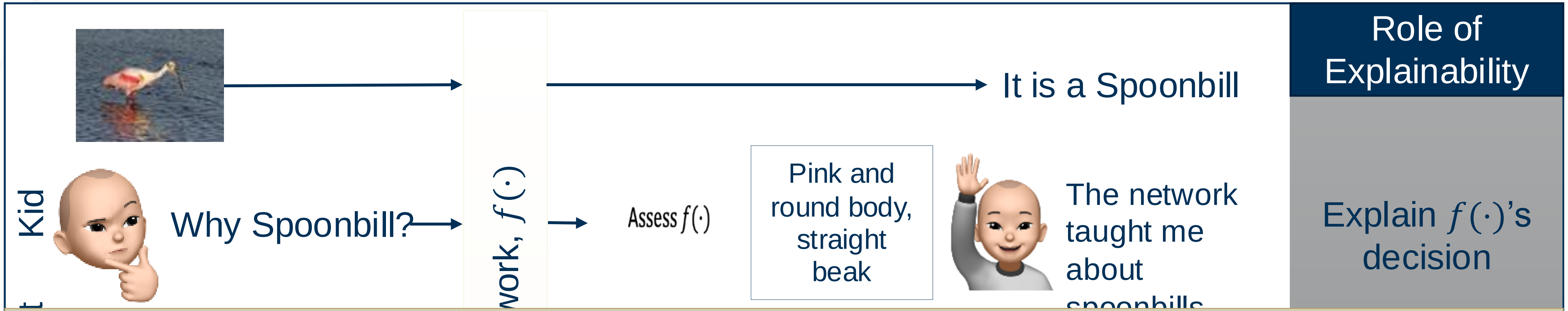


The fox has knowledge of Spoonbills and foxes and uses this knowledge to ask **targeted** questions!

AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks. *arXiv preprint arXiv:2202.11838*.

Types of Explanations

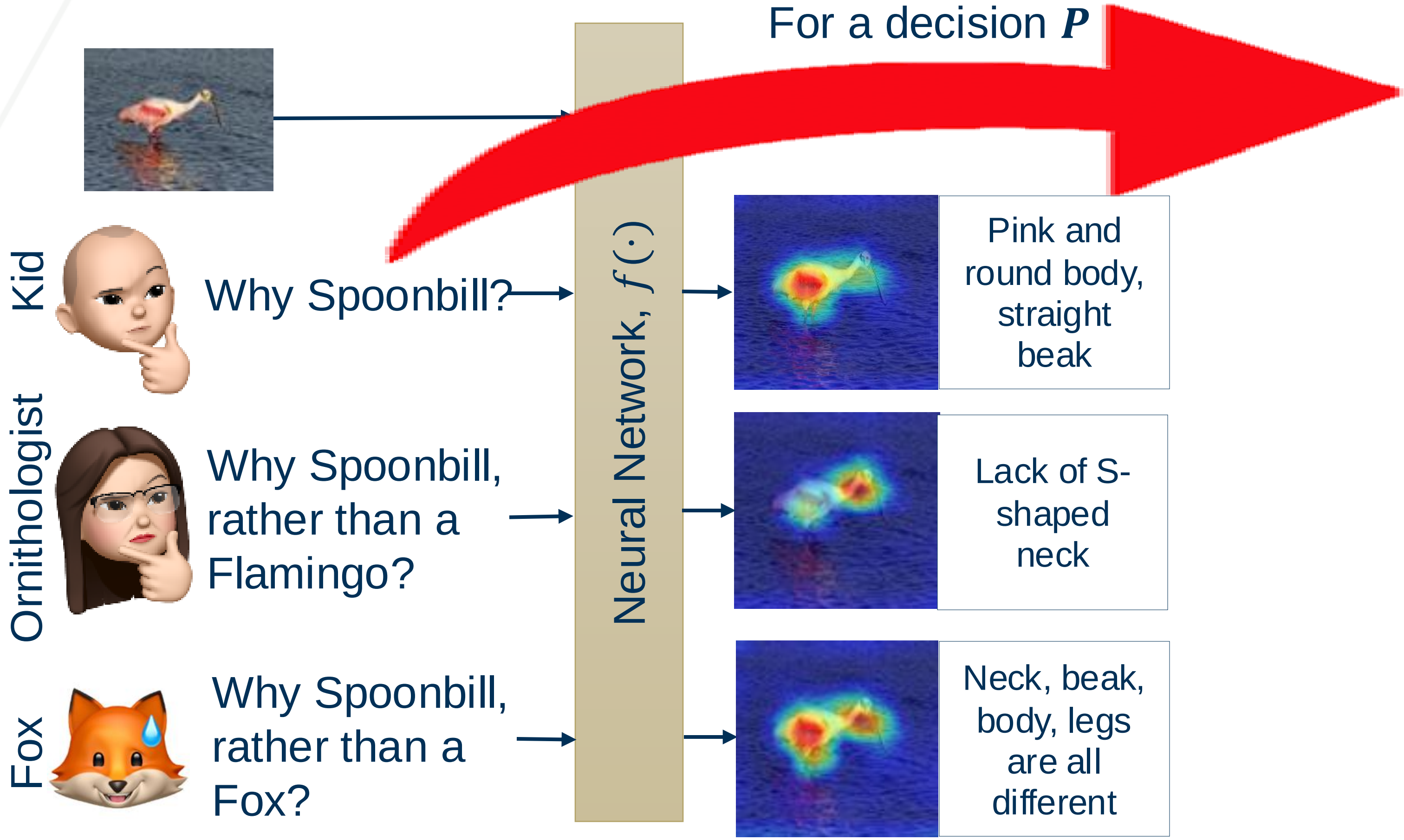
Targeted Explanations



Targeted Explanations : Explanations that highlight contextually relevant regions in an image. No network knowledge is required from the humans interpreting these explanations. Knowledge about the classes or data is required by the humans seeking explanations.

AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks. *arXiv preprint arXiv:2202.11838*.

Targeted Explanations
Questions



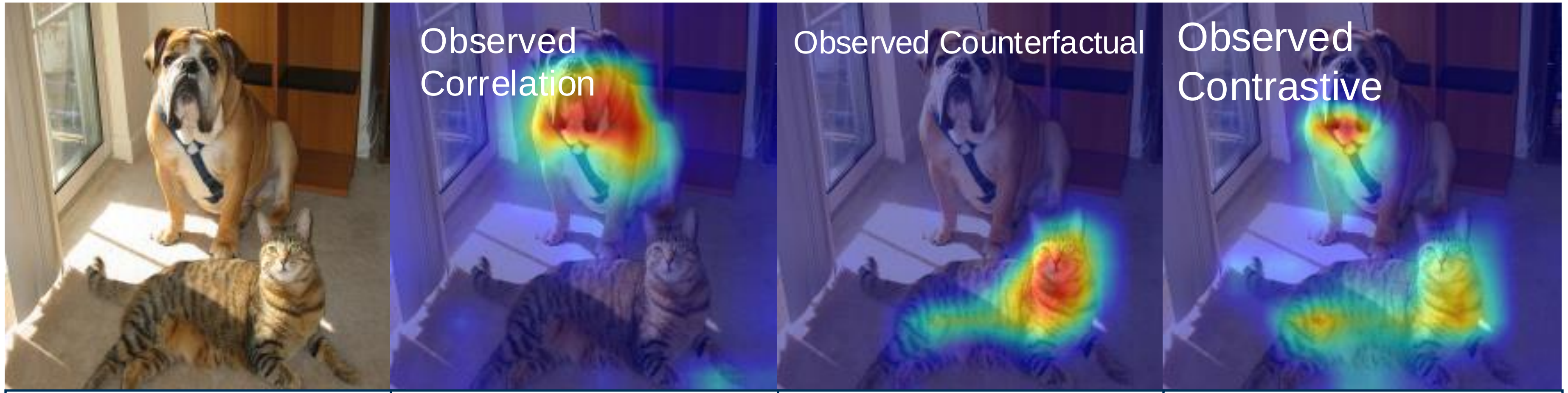
3 forms of questions

- *`Why P?'* - *Correlations*
- *`What if?'*- *Counterfactual*
- *`Why P, rather than Q?'*- *Contrastive*

AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks. *arXiv preprint arXiv:2202.11838*.

Targeted Explanations

Explanatory Paradigms

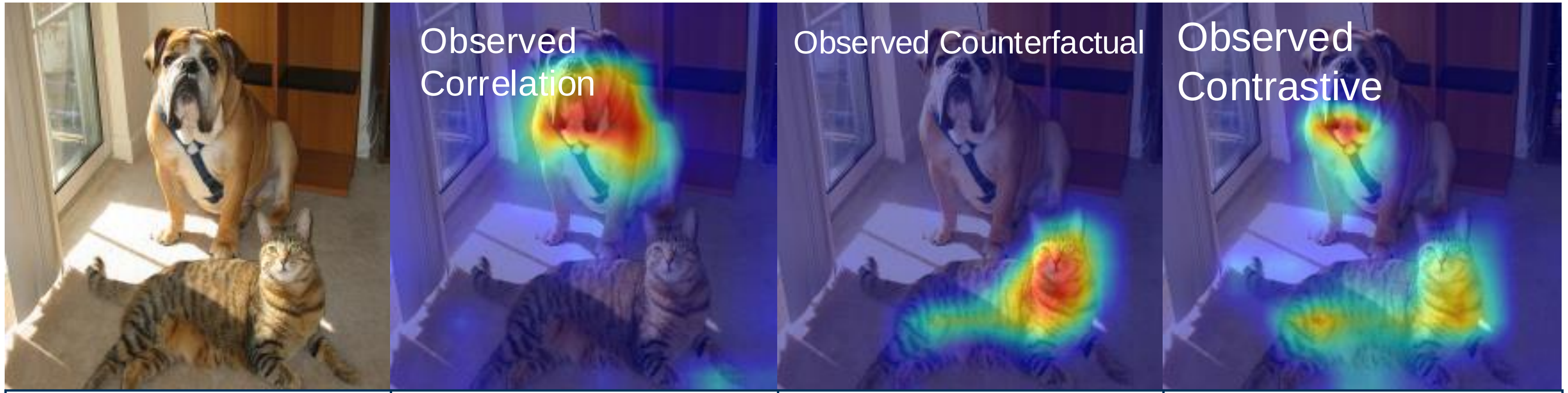


Bullmastiff	Why Bullmastiff?	What if Bullmastiff was not in the image?	Why Bullmastiff, rather than a Boxer?

AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks. *arXiv preprint arXiv:2202.11838*.

Targeted Explanations

Questions



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

As an aside..

The three paradigms are instances of abductive reasoning technique

`When you have eliminated all which is impossible, that whatever remains, however improbable, must be the truth' - Sherlock Holmes' reasoning technique is abductive

OLIVES @GeorgiaTech

Georgia Tech

AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks. *arXiv preprint arXiv:2202.11838*.

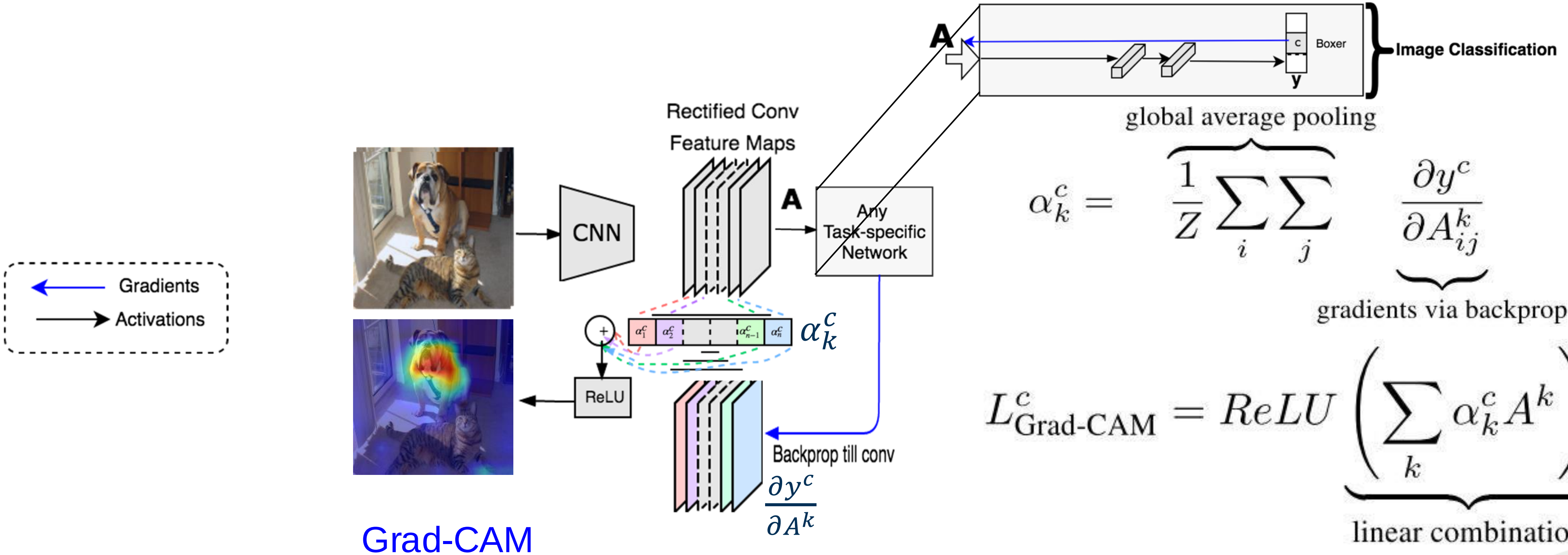
Targeted Explanations

Correlations through Grad-CAM

Why P?

Recap of Grad-CAM

- Given an image, feed forward through CNN
- Final convolutional layer output feature maps for later task-specific layers
- Backward pass the predicted logit y^c to last conv layer
- Compute gradients w.r.t. last conv activations
- Global average pool the gradients to obtain α^c for each kernel k
- Multiply each α^c with activations at that kernel A^k and add the resultant before passing through RELU
- Up-sample to original size and normalize



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

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

Grad-CAM

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.

Targeted Explanations

Correlations through Grad-CAM

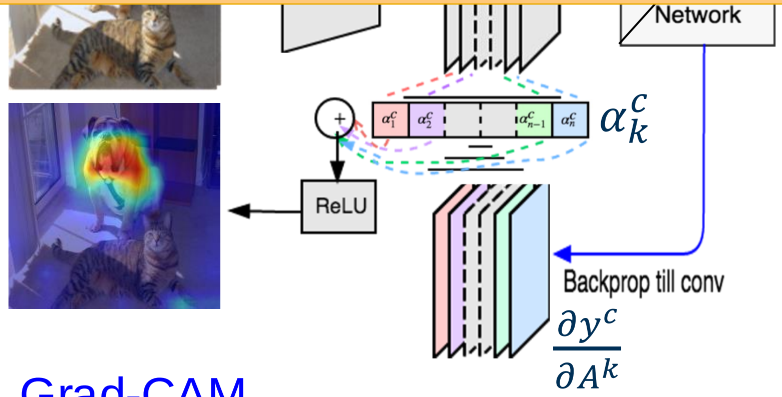
Why P?

Recap of Grad-CAM

- Given an image, feed forward through CNN
- Final convolutional layer output feature maps for later task-specific layers
- Backward pass the predicted logit y^c to last conv layer
- Compute gradients w.r.t. last conv activations
- Global average pool the gradients to obtain α^c for each kernel k
- Multiply each α^c with activations at that kernel A^k and add the resultant before passing through RELU
- Up-sample to original size and normalize

All direct explanations answer *'Why P?'*

α_k^c



Grad-CAM

$i \quad j$

gradients via backprop

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

OLIVES @GeorgiaTech

Georgia Tech

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.

Targeted Explanations

Correlations through Grad-CAM

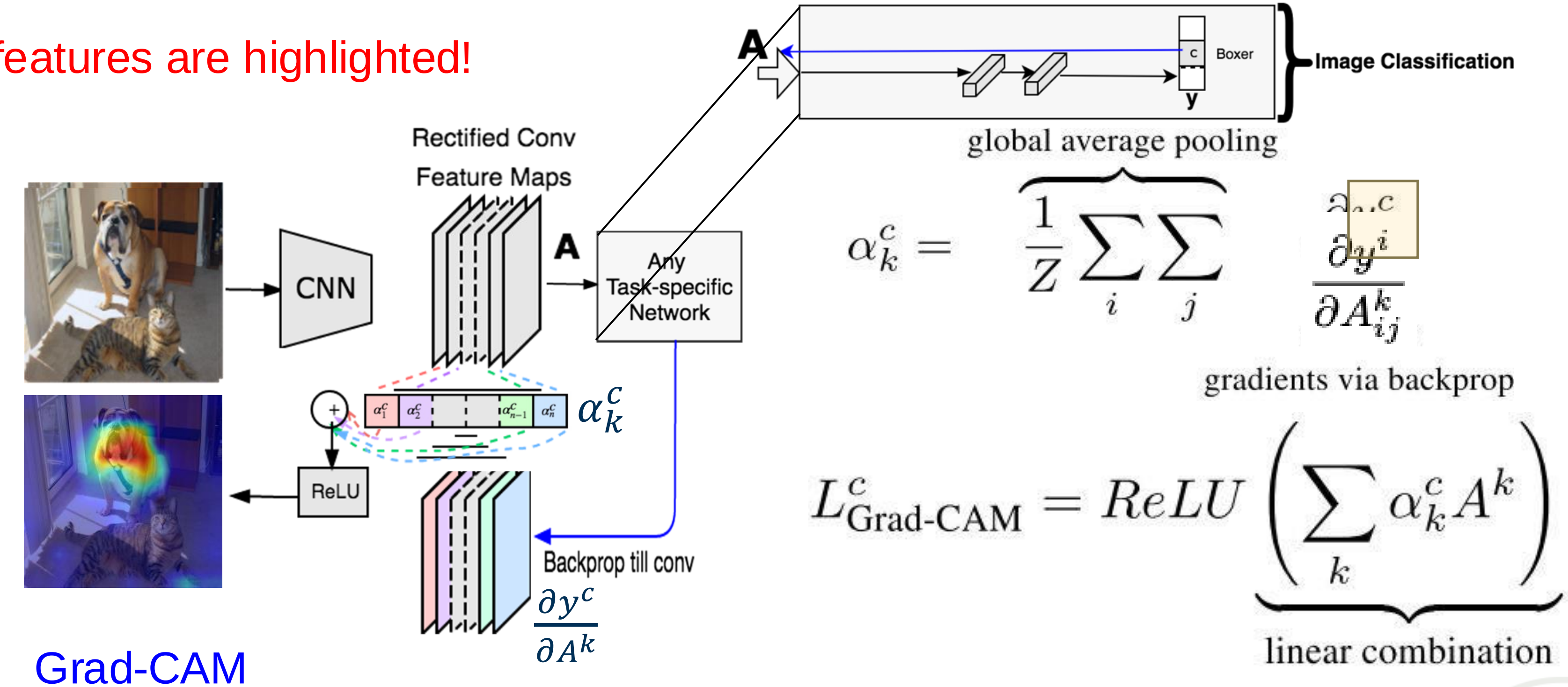
Recap of Grad-CAM

- Given an image, feed forward through CNN
- Final convolutional layer output feature maps for later task-specific layers
- Backward pass the any logit y^i to last conv layer
- Compute gradients w.r.t. last conv activations
- Global average pool the gradients to obtain α^c for each kernel k
- Multiply each α^c with activations at that kernel A^k and add the resultant before passing through RELU
- Up-sample to original size and normalize

Can ask `*Why Blue Jay?*'

Generally the same features are highlighted!

α_k^c



Grad-CAM

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.

Targeted Explanations

Counterfactual-CAM

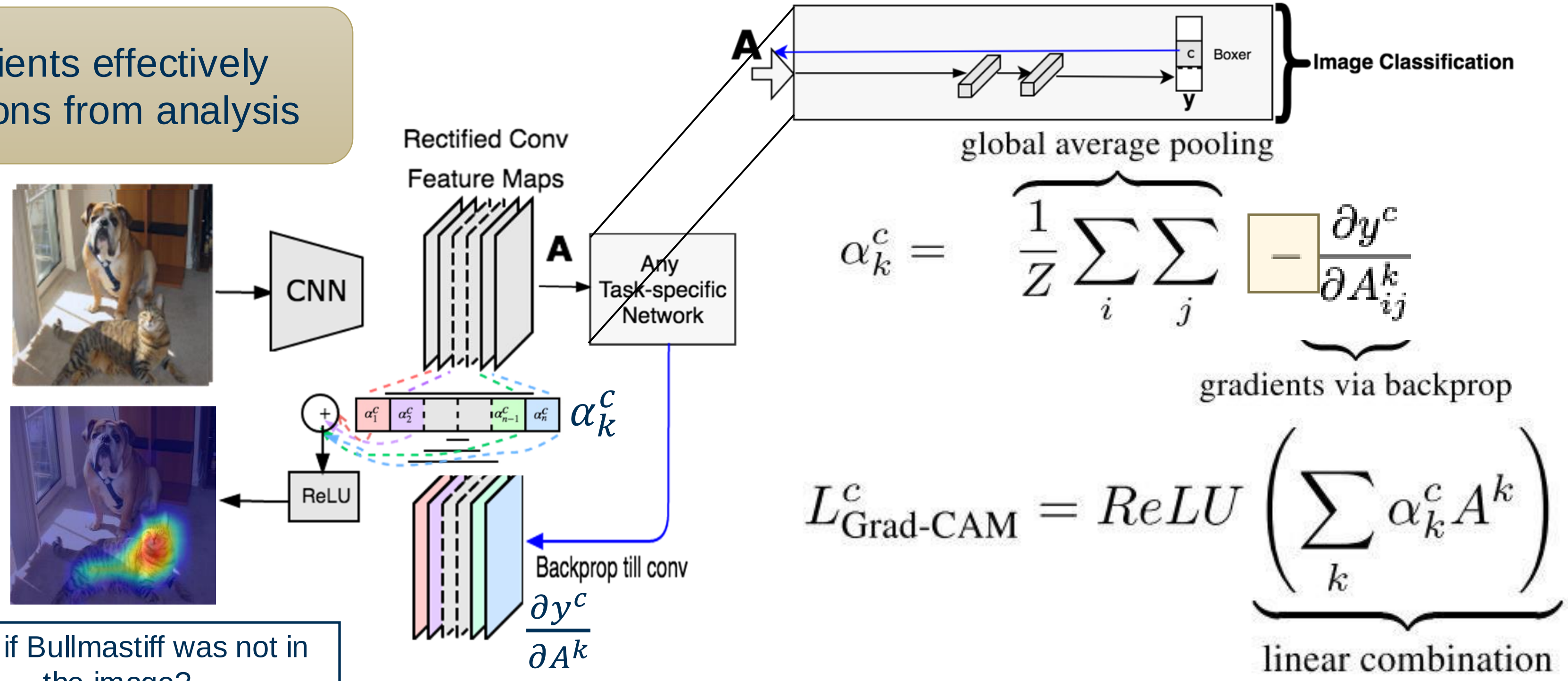
Recap of Grad-CAM

- Given an image, feed forward through CNN
- Final convolutional layer output feature maps for later task-specific layers
- Backward pass the predicted logit y^c to last conv layer
- Compute gradients w.r.t. last conv activations
- Global average pool the negative of gradients to obtain α^c for each kernel k
- Multiply each α^c with activations at that kernel A^k and add the resultant before passing through RELU
- Up-sample to original size and normalize

What if P was not in the image?

Negating the gradients effectively removes these regions from analysis

α_k^c



What if Bullmastiff was not in the image?

Counterfactual-CAM

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.

Targeted Explanations

Contrast-CAM

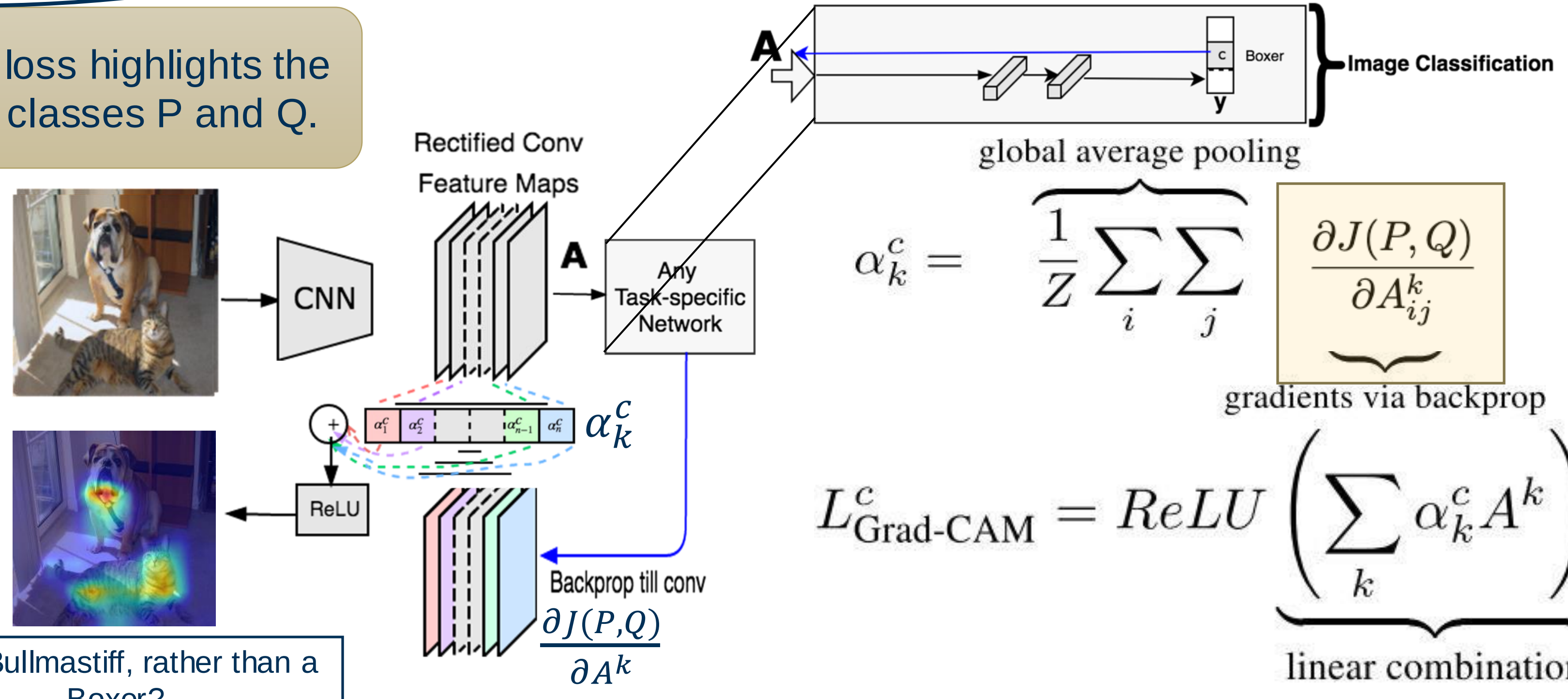
Recap of Grad-CAM

- Given an image, feed forward through CNN
- Final convolutional layer output feature maps for later task-specific layers
- Backward pass the loss between predicted class P and some contrast class Q to last conv layer
- Compute gradients w.r.t. last conv activations
- Global average pool the negative of gradients to obtain α^c for each kernel k
- Multiply each α^c with activations at that kernel A^k and add the resultant before passing through RELU
- Up-sample to original size and normalize

Why P, rather than Q?

Backpropagating the loss highlights the differences between classes P and Q.

α_k^c



$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{\boxed{\frac{\partial J(P,Q)}{\partial A_{ij}^k}}}_{\text{gradients via backprop}}$$

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

Contrast-CAM

Why Bullmastiff, rather than a Boxer?

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.

Targeted Explanations

Contrast-CAM

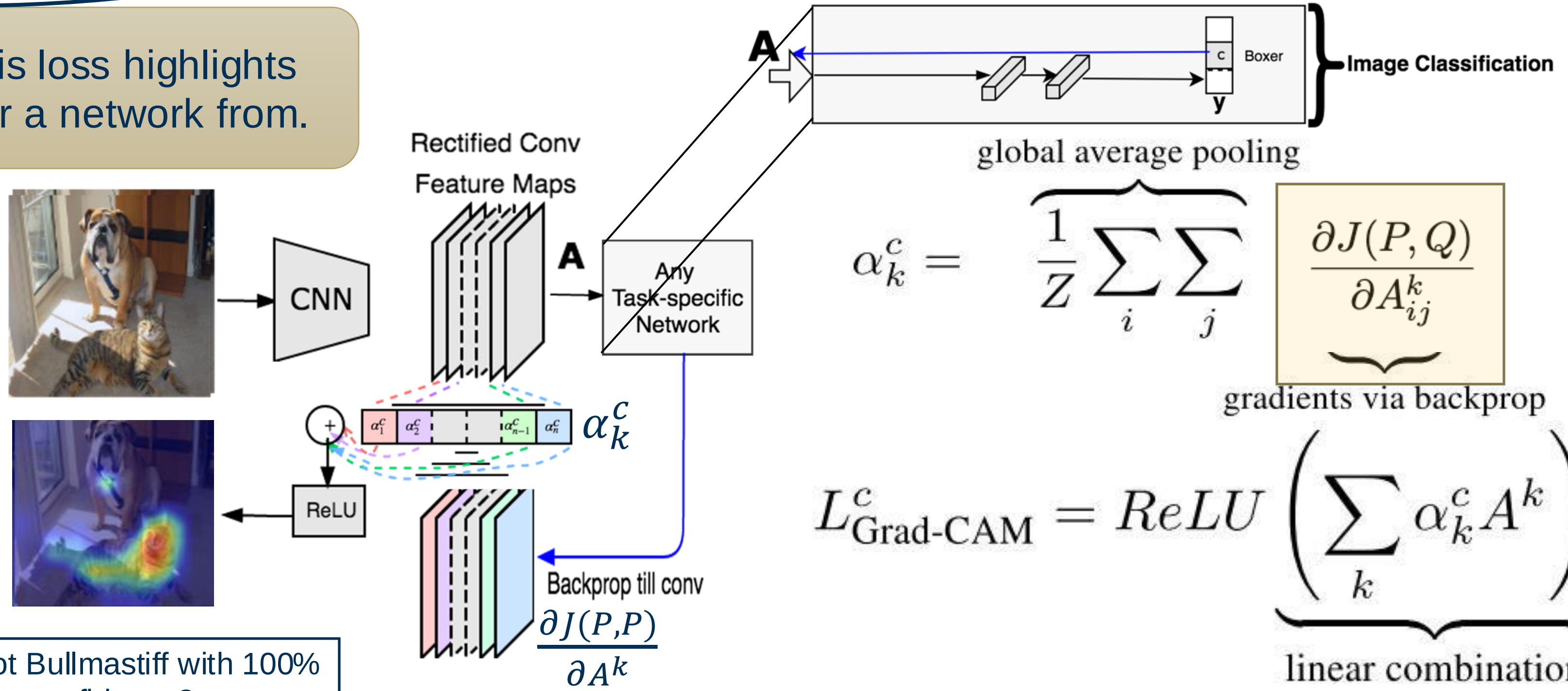
Recap of Grad-CAM

- Given an image, feed forward through CNN
- Final convolutional layer output feature maps for later task-specific layers
- Backward pass the loss between predicted class P and same class P to last conv layer
- Compute gradients w.r.t. last conv activations
- Global average pool the negative of gradients to obtain α^c for each kernel k
- Multiply each α^c with activations at that kernel A^k and add the resultant before passing through RELU
- Up-sample to original size and normalize

Why P, rather than P?

Backpropagating this loss highlights confusing regions for a network from.

α_k^c



$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{\boxed{\frac{\partial J(P,Q)}{\partial A_{ij}^k}}}_{\text{gradients via backprop}}$$

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

Contrast-CAM

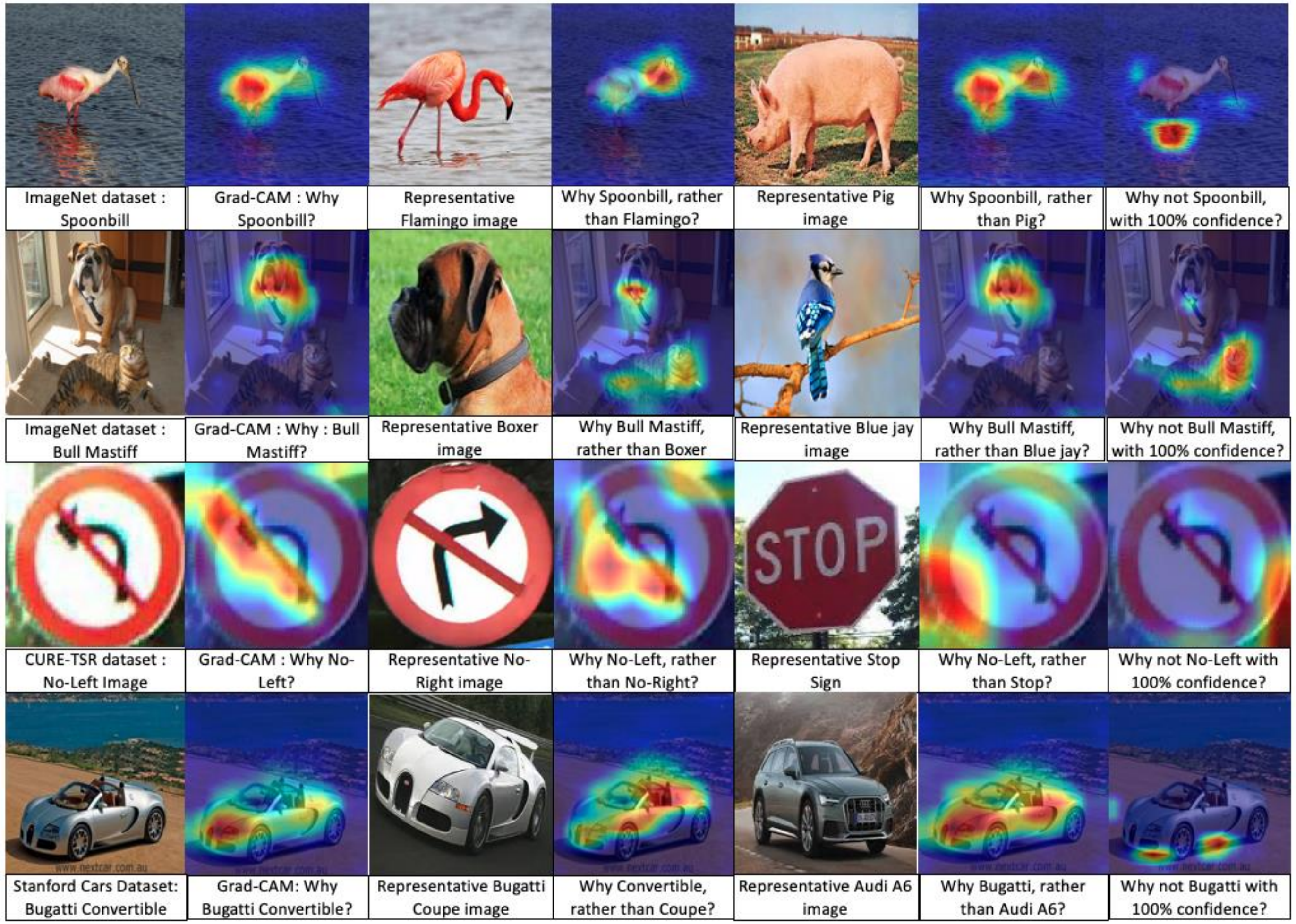
Why not Bullmastiff with 100% confidence?

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.

Targeted Explanations

Contrast-CAM

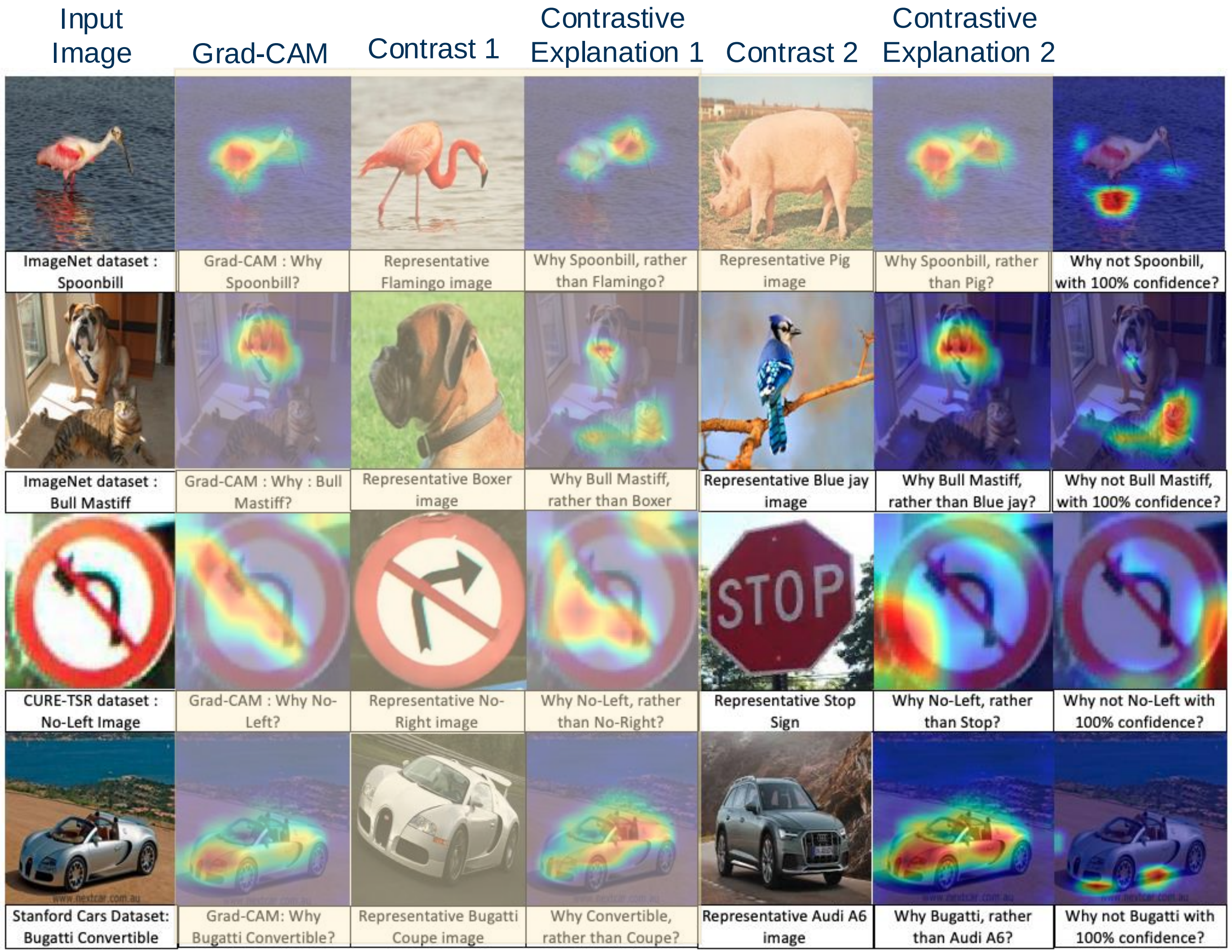
Input Image | Grad-CAM | Contrast 1 | Contrastive Explanation 1 | Contrast 2 | Contrastive Explanation 2



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.

Targeted Explanations

Contrast-CAM



Human Interpretable

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.

Targeted Explanations

Contrast-CAM

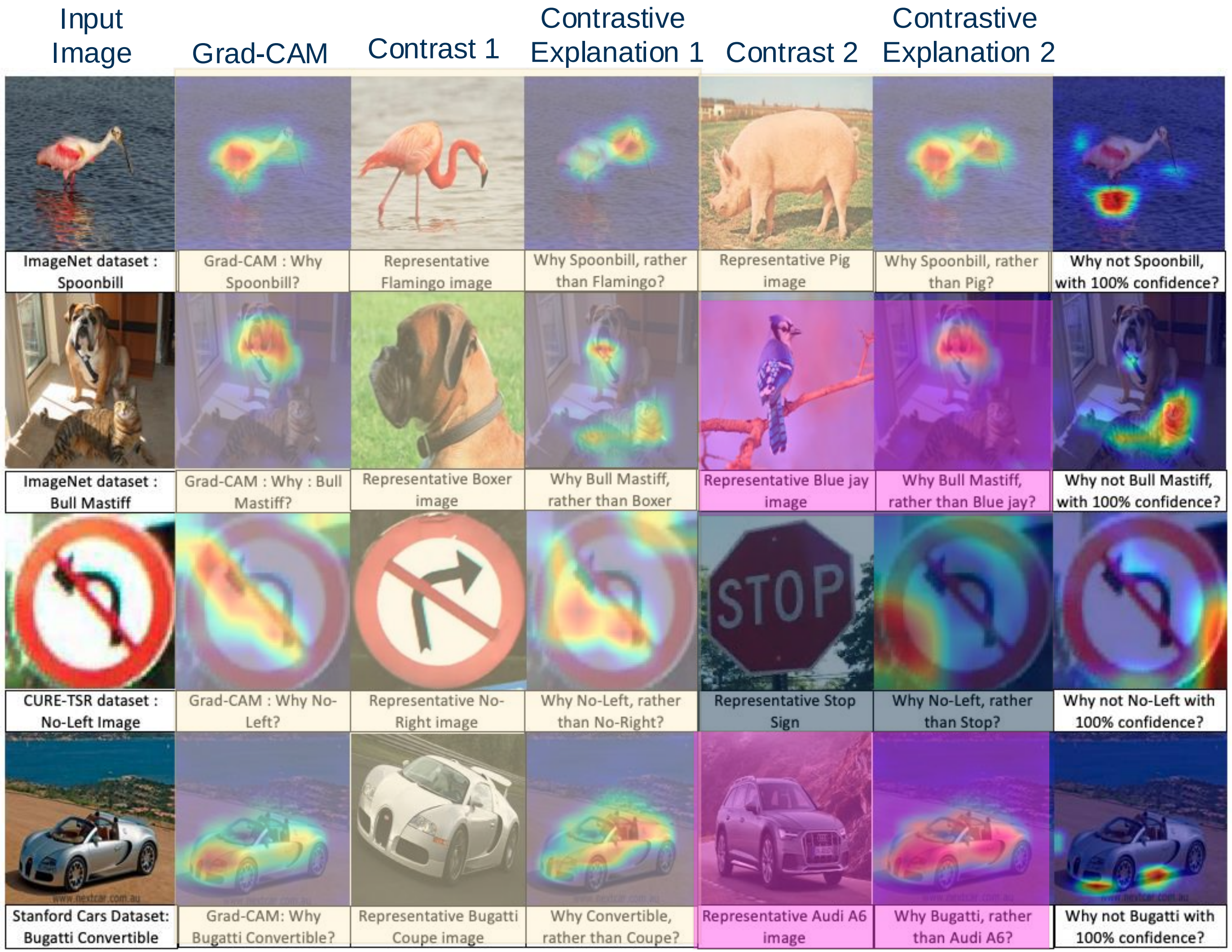
Input Image | Grad-CAM | Contrast 1 | Contrastive Explanation 1 | Contrast 2 | Contrastive Explanation 2

Human Interpretable

Same as Grad-CAM

Targeted Explanations

Contrast-CAM



Human Interpretable

Same as Grad-CAM

Not Human Interpretable

Targeted Explanations

Contrast-CAM

Human Interpretable

Same as Grad-

OLIVES

Georgia Tech

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.

Targeted Explanations

Contrast-CAM

Overview

In this Lecture..

Explainability

Visualization of Convolutional Neural Networks

Types of Explanations

- Categorization of existing explanations
- Types of explanations
- Indirect and Direct Explanations
- Targeted Explanations
- Explanatory Paradigms

Explanatory Evaluation

- Evaluation Taxonomy
- Human Evaluation
- Application Evaluation
- Network Evaluation

Explanatory Evaluation

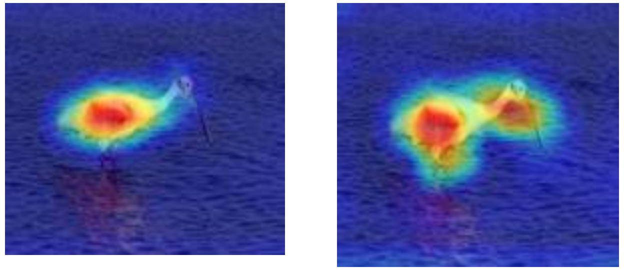
- What constitutes an explanation?
  - Visualizing weights across different layers
  - Visualizing activations in intermediate layers
  - Visualizing maximally activation patches
  - Visualizing last layer embedding
  - Nearest neighbor samples
  - Dimensionality reduction
  - Pixel saliency through intervention
  - Pixel saliency through via feature importance
- What makes some explanations better than the others?
  - Subjectively, visualizing activations provide better explanations than weights in intermediate layers
  - Pixel saliency via feature importance generalizes to different tasks and are computationally less expensive

Types of explanations

Evaluating explanations

Explanatory Evaluation Taxonomy

Human Evaluation	Application Evaluation	Network Evaluation
Tasks : Humans directly evaluate explanations.	**Tasks** : Any task that requires humans-in-the-evaluation-loop without directly having humans measure explainability.	**Tasks** : Any task intersecting with explainability that does not require humans for evaluation. Ex : Robustness of neural nets.



Which explanation is better for answering Why Spoonbill?

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

Is this noisy image still a spoonbill?

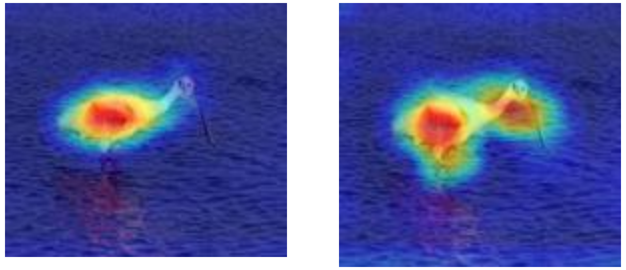
AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks. *arXiv preprint arXiv:2202.11838*.

Explanatory Evaluation

Human Evaluation

Human Evaluation

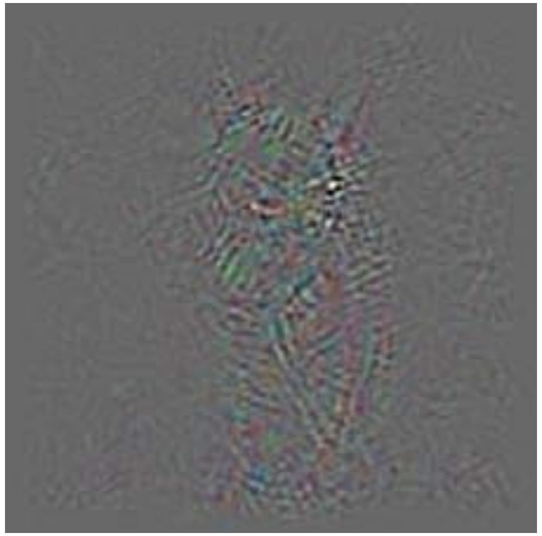
Tasks : Humans directly evaluate explanations.



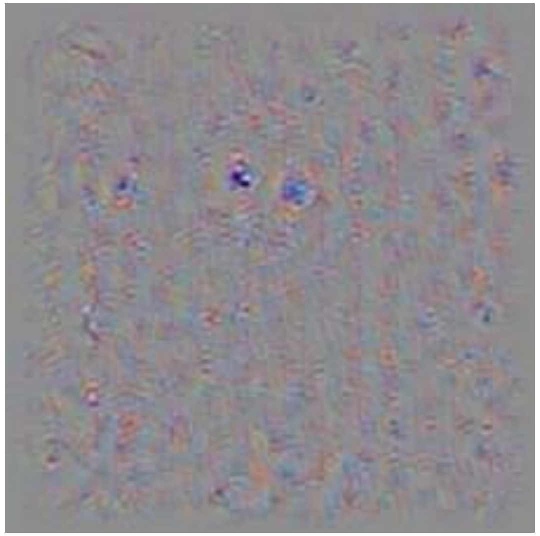
Which explanation is better for answering Why Spoonbill?

Humans are directly asked to evaluate explanatory techniques

Backprop



Deconv



Guided Backprop

Which of the three techniques is better?

AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks. *arXiv preprint arXiv:2202.11838*.

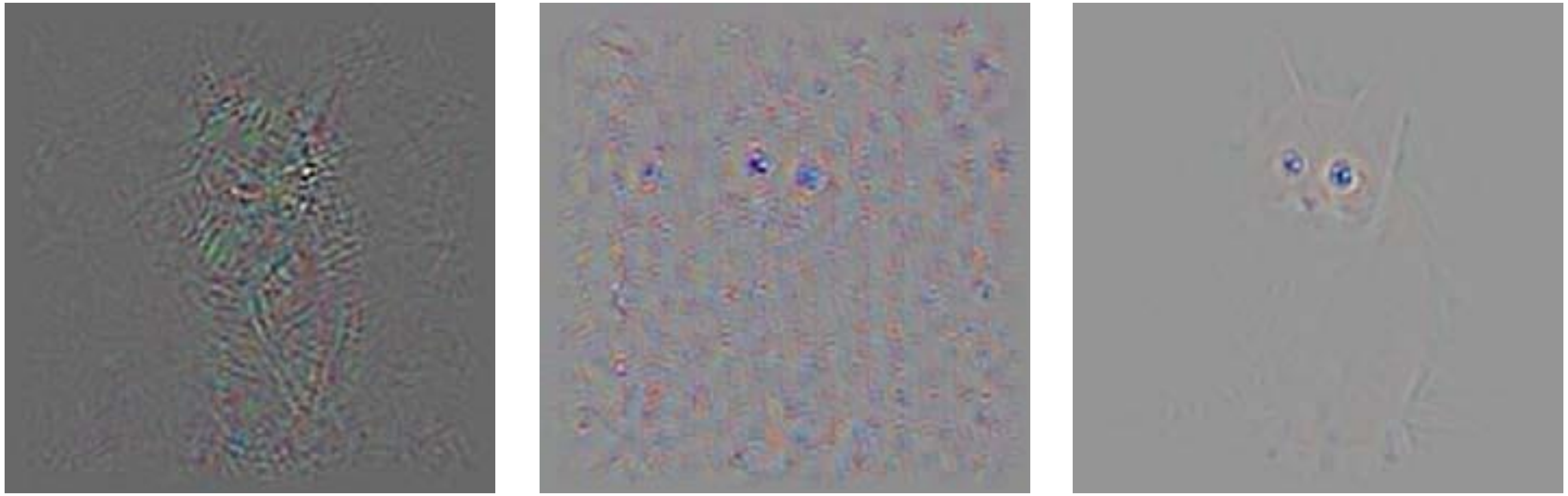
Explanatory Evaluation

Human Evaluation

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

Humans are directly asked to evaluate explanatory techniques

Backprop | Deconv | Guided Backprop



Which of the three techniques is better?

OLIVES @GeorgiaTech

Georgia Tech

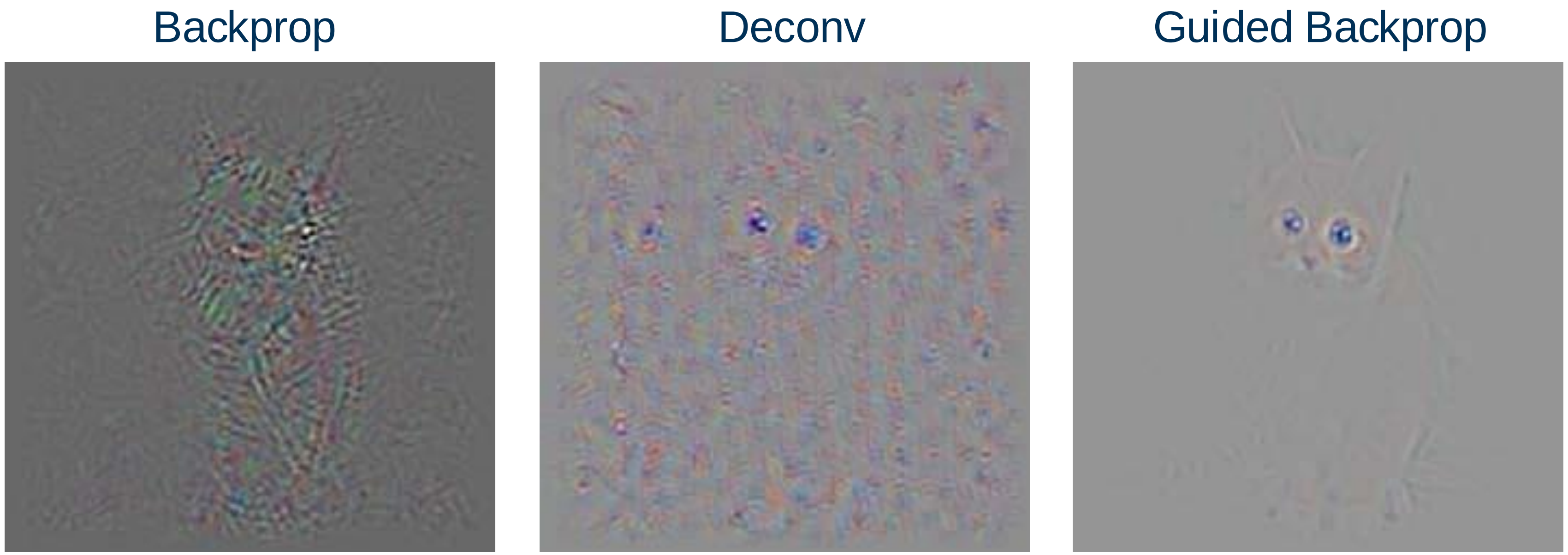
AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks. *arXiv preprint arXiv:2202.11838*.

Explanatory Evaluation
Human Evaluation

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

How to overcome?
- Ask a `large number’ of humans the same question
- Make sure that the humans are unaware of the goal of the researchers
- Guide them through with questions

Humans are directly asked to evaluate explanatory techniques



Which of the three techniques is better?

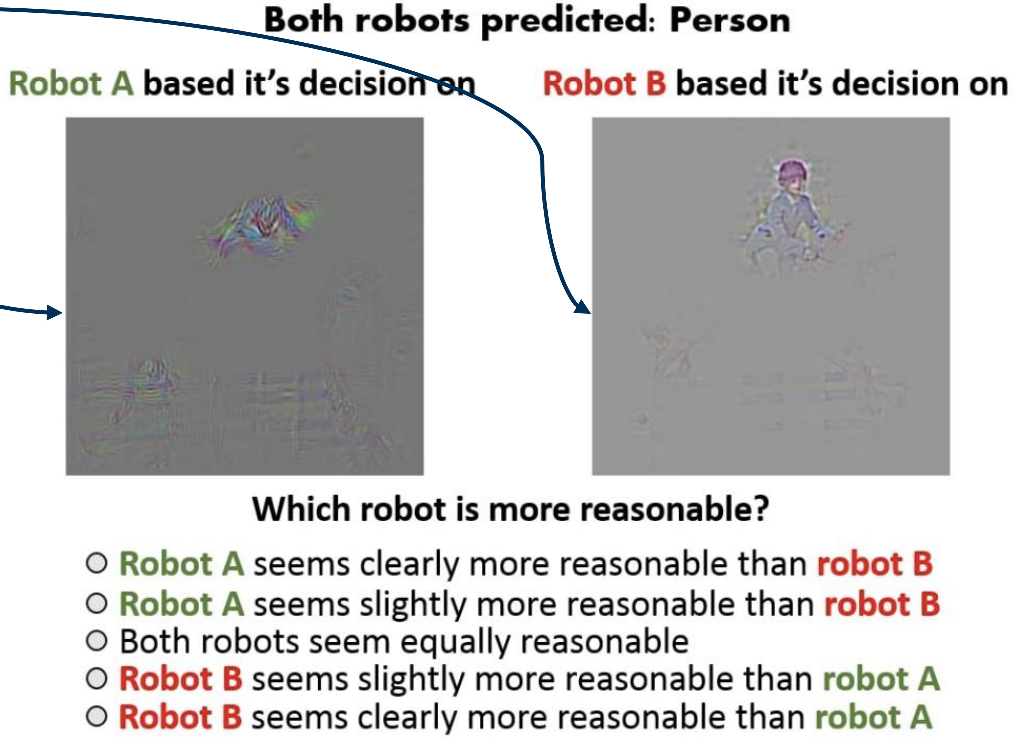
AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks. *arXiv preprint arXiv:2202.11838*.

Explanatory Evaluation

Human Evaluation

Input

2 explanations from separate techniques



How to overcome?

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

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

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.

Explanatory Evaluation

Human Evaluation

How to overcome?

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

Input

2 explana
from sep
techniques

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

Robot A based it's decision on **Robot B** based it's decision on

Aesthetically pleasing vs class discriminative?
Bias based on the questions themselves?
How many answers are sufficient?

- 43 AMT work
asked the ad
question for e
image-question pair
- This experiment was repeated over 90 image-question pairs

Which robot is more reasonable?

- ○ **Robot A** seems clearly more reasonable than **robot B**
- ○ **Robot A** seems slightly more reasonable than **robot B**
- ○ Both robots seem equally reasonable
- ○ **Robot B** seems slightly more reasonable than **robot A**
- ○ **Robot B** seems clearly more reasonable than **robot A**

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.

Explanatory Evaluation

Application Evaluation

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

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

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

Instead, evaluation is indirectly done by designing other applications

- Gaze tracking
- Pointing game

Explanatory Evaluation

Application Evaluation

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

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

Gaze

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

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

OLIVES @GeorgiaTech

Georgia Tech

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

Explanatory Evaluation
Application Evaluation

Gaze

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

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

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

The explanations from various methods are evaluated against this ground truth using Mean Intersection over Union or other segmentation metrics

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

Explanatory Evaluation

Application Evaluation – Performance metrics

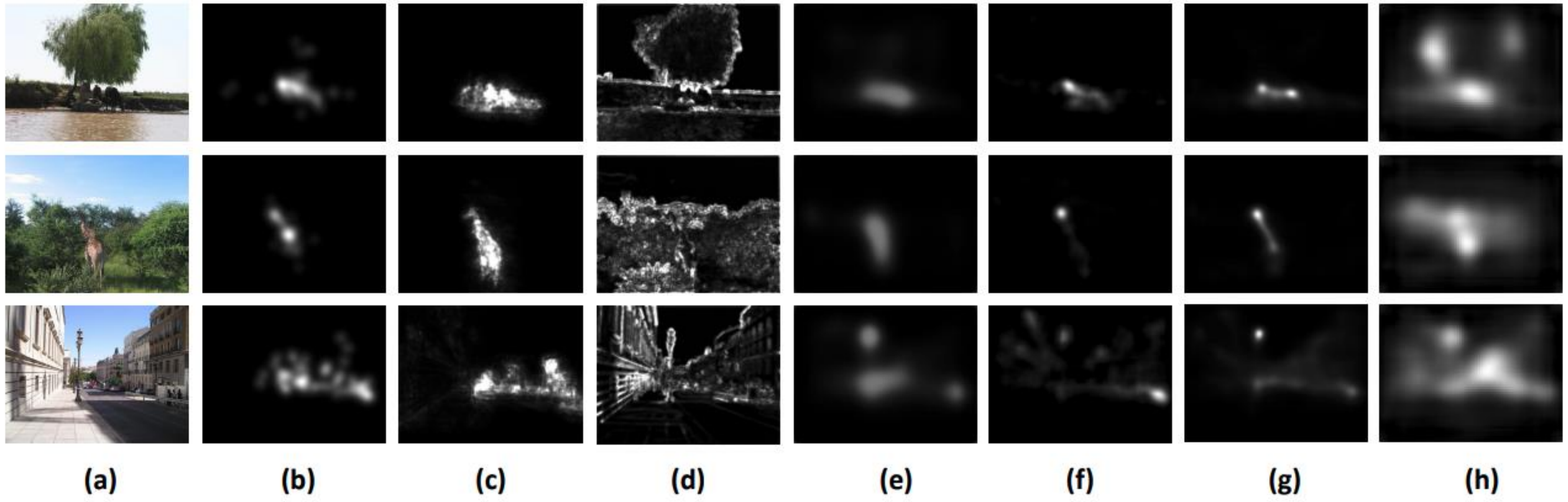


Fig. 3. Saliency map visualization. (a) Input image (b) Groudtruth (c) Proposed Method (d) Feed-forward feature (e) SalGan [21] (f) ML-Net [5] (g) DeepGazeII [22] (h) ShallowDeep [23]

- **Mean Intersection over Union**: Pixel-area of overlap between predicted and ground truth segmentation masks divided by the area of union between predicted and ground truth masks
- **Correlation coefficient**: Measures the linear relationship between each variable
- **Pixel accuracy**: Number of correctly predicted pixels divided by the total number of pixels in ground truth mask

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

Explanatory Evaluation

Application Evaluation

Pointing Game

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

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

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

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

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

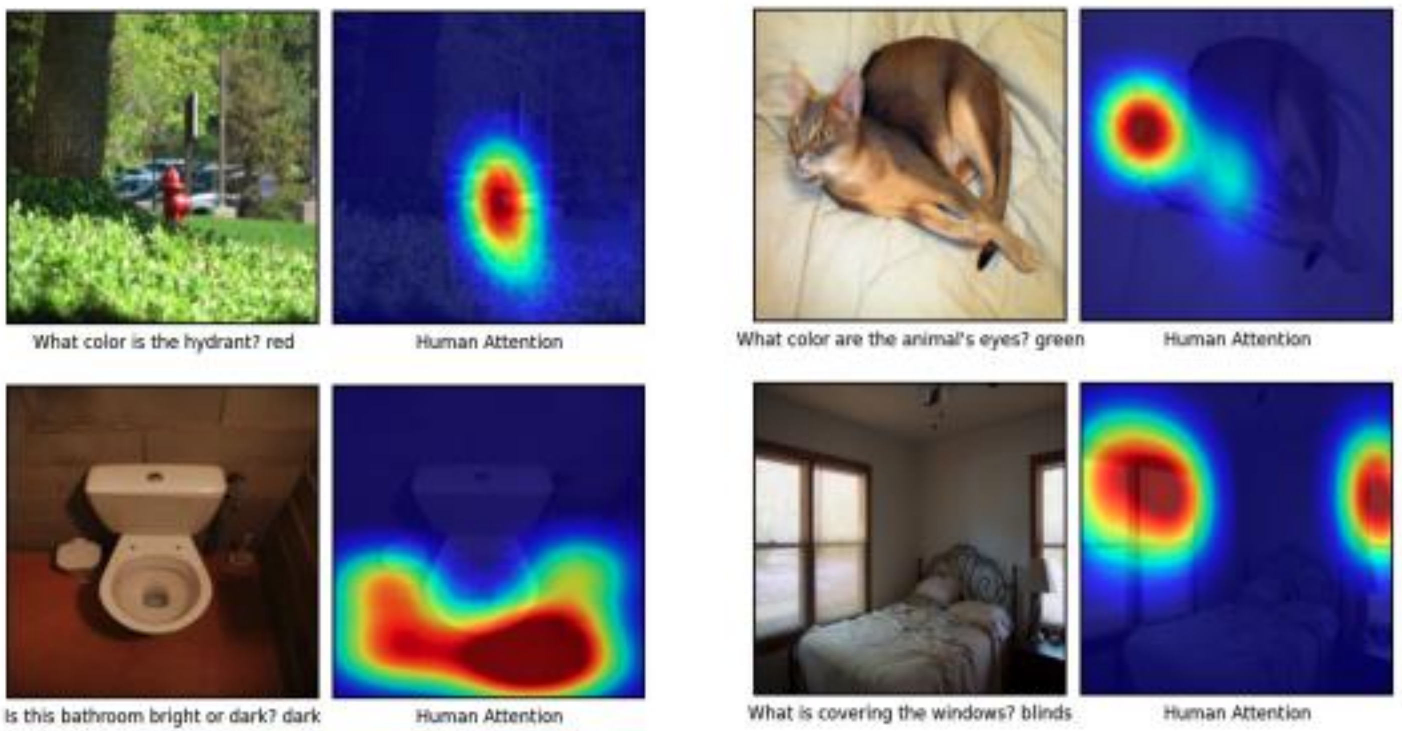
Explanatory Evaluation

Application Evaluation

Pointing Game

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

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



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

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

Explanatory Evaluation

Application Evaluation

Application Evaluation

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

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

Gaze tracking:

- Tracks true salient regions in an image without bias of questions or questionnaire
- Requires expensive eye-tracking equipment
- May lead to spurious noise and center-bias
- No targeted answers or explanations can be obtained

Pointing game:

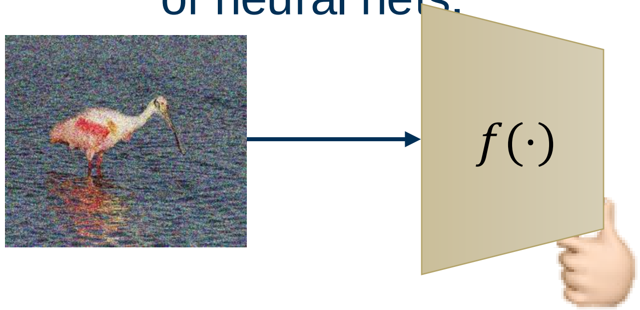
- May introduce bias in viewer
- Does not require specialized equipment and can be performed on Mechanical Turk
- Clean results since people are asked to sharpen from an already blurred image
- Targeted answers and explanations can be obtained based on targeted questions

Explanatory Evaluation

Network Evaluation

Network Evaluation

Tasks : Any task intersecting with explainability that does not require humans for evaluation. Ex : Robustness of neural nets.



Is this noisy image still a spoonbill?

Does not require visualizations or human annotations at any stage

Based on the idea that networks that perform certain tasks are implicitly explanatory

Tasks:

- Robust classification [1] : Classification in the presence of noise, adversarial examples, and unseen data during testing after training on clean data
- Anomaly Detection [2] : Detecting data drawn from out-of-distribution during testing
- Machine Teaching [3] : Networks are used to teach humans
- …

[1] Prabhushankar, Mohit, and Ghassan AlRegib. "Contrastive Reasoning in Neural Networks." *arXiv preprint arXiv:2103.12329* (2021).
[2] Kwon, G., Prabhushankar, M., Temel, D., & AlRegib, G. (2020, August). Backpropagated gradient representations for anomaly detection. In *European Conference on Computer Vision* (pp. 206-226). Springer, Cham.
[3] Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., & Lee, S. (2019, May). Counterfactual visual explanations. In *International Conference on Machine Learning* (pp. 2376-2384). PMLR.

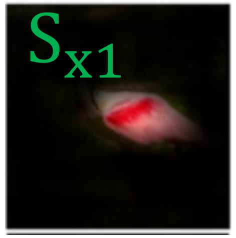
Explanatory Evaluation

Network evaluation via masking

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

Two types of Masking:

1. **Masking using explanation heatmap**
2. Pixel-wise masking using explanation as importance



$\rightarrow$ Trained Model $\rightarrow$ Crane

$\rightarrow$ Trained Model $\rightarrow$ Spoonbill

S_{x1} = Guided Backpropagation masked data

S_{x2} = GradCAM masked data

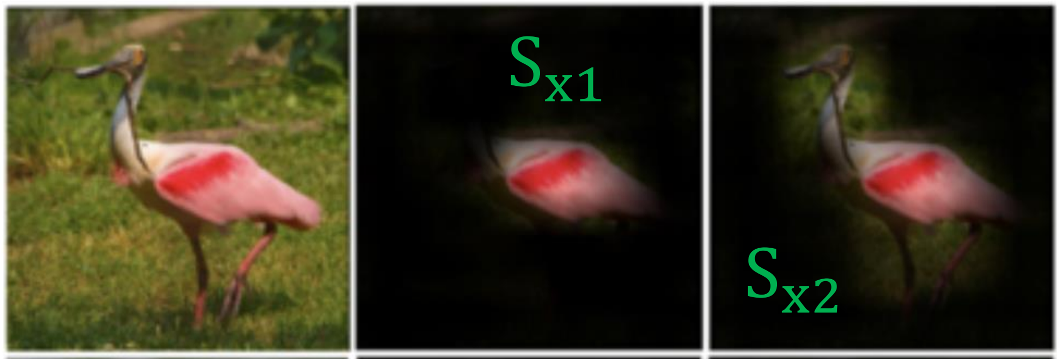
Network Evaluation

Evaluation 1: Explanation Evaluation via Masking

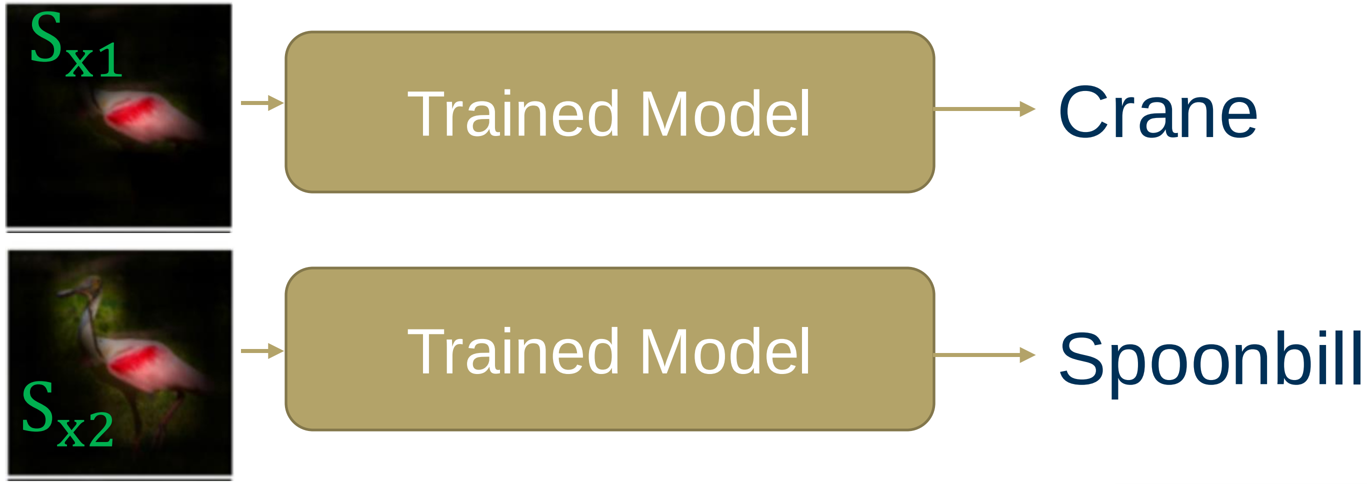
Common evaluation technique is masking the image and checking for prediction correctness

y = Prediction
S_x = Explanation masked data

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



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

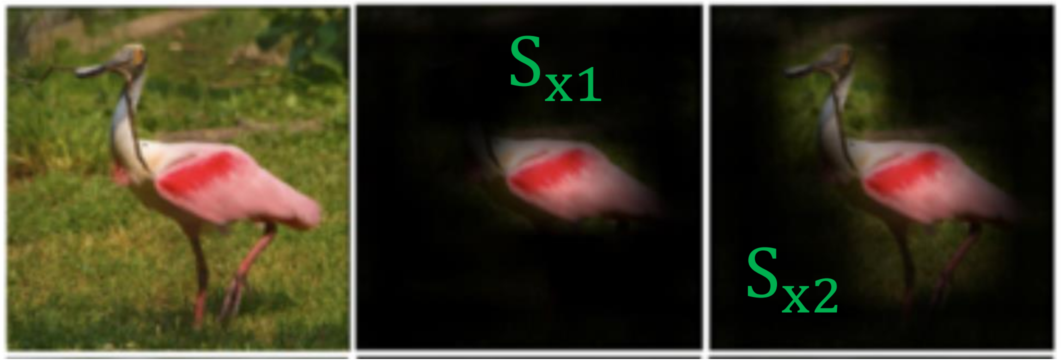


Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018.

Network Evaluation

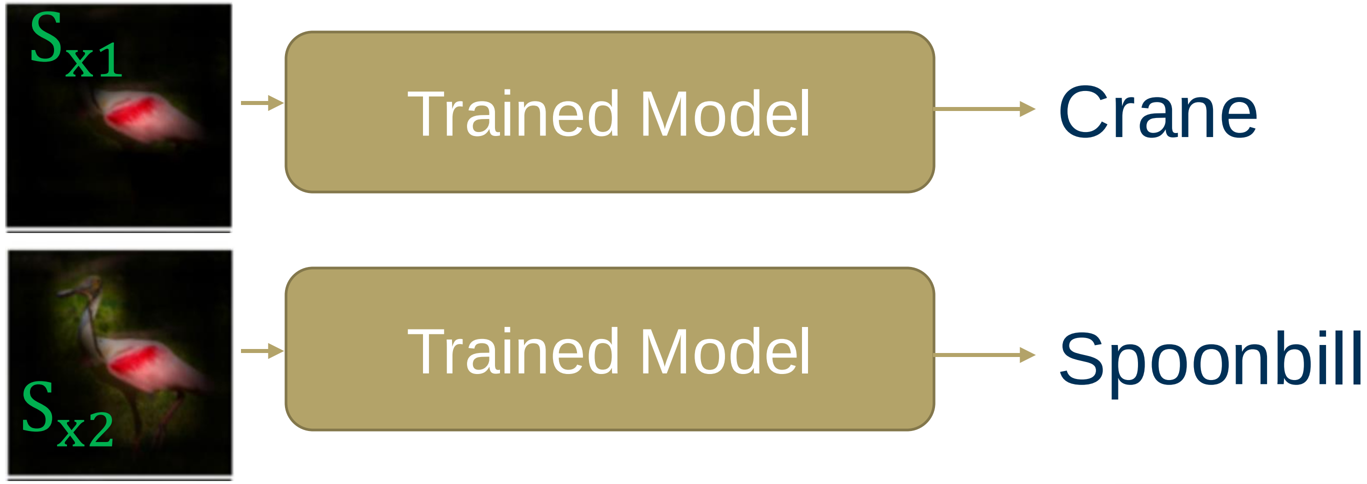
Evaluation 1: Explanation Evaluation via Masking

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.

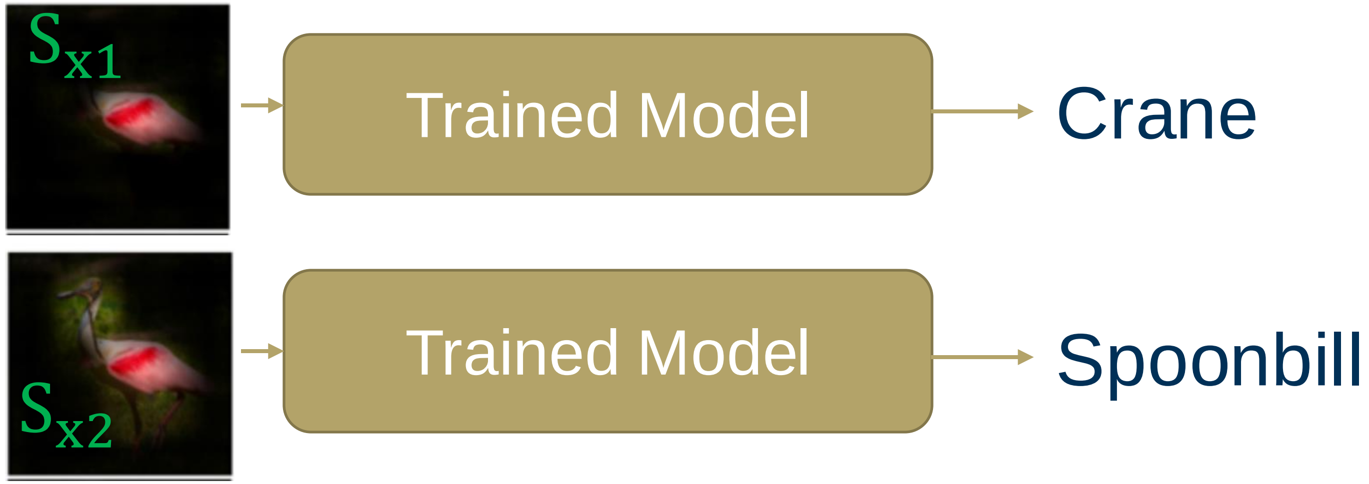
Network Evaluation

Network evaluation via masking

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

Two types of Masking:

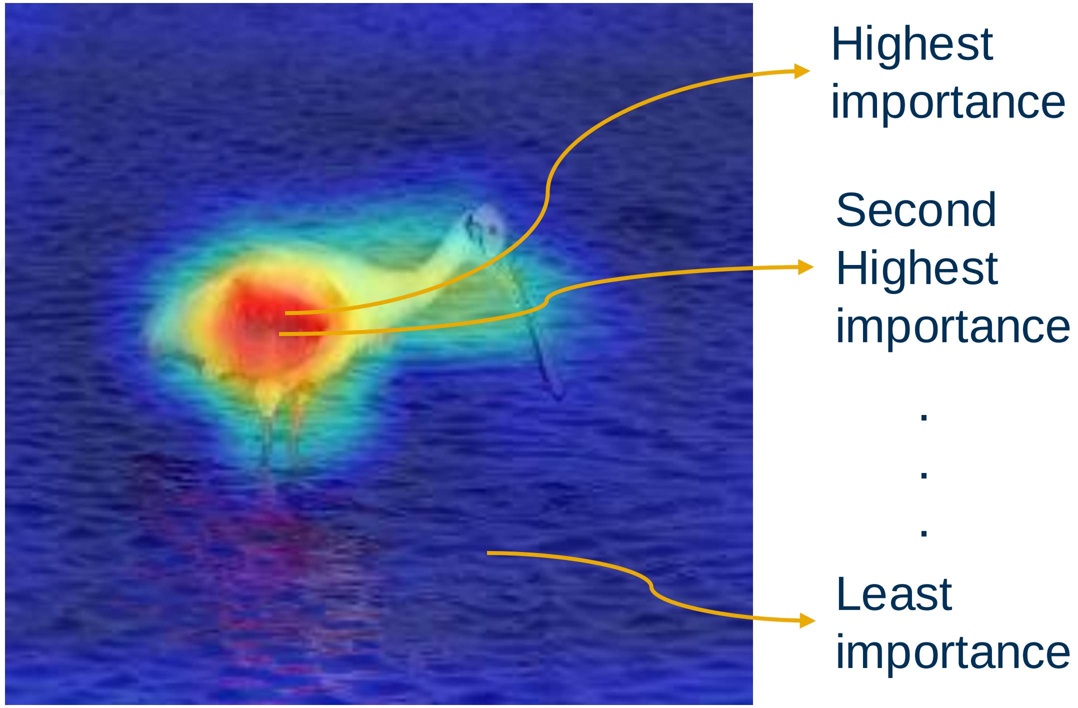
1. Masking using explanation heatmap
2. **Pixel-wise masking using explanation as importance**



Explanatory Evaluation

Evaluation 2: Progressive Pixel-wise Insertion and Deletion

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



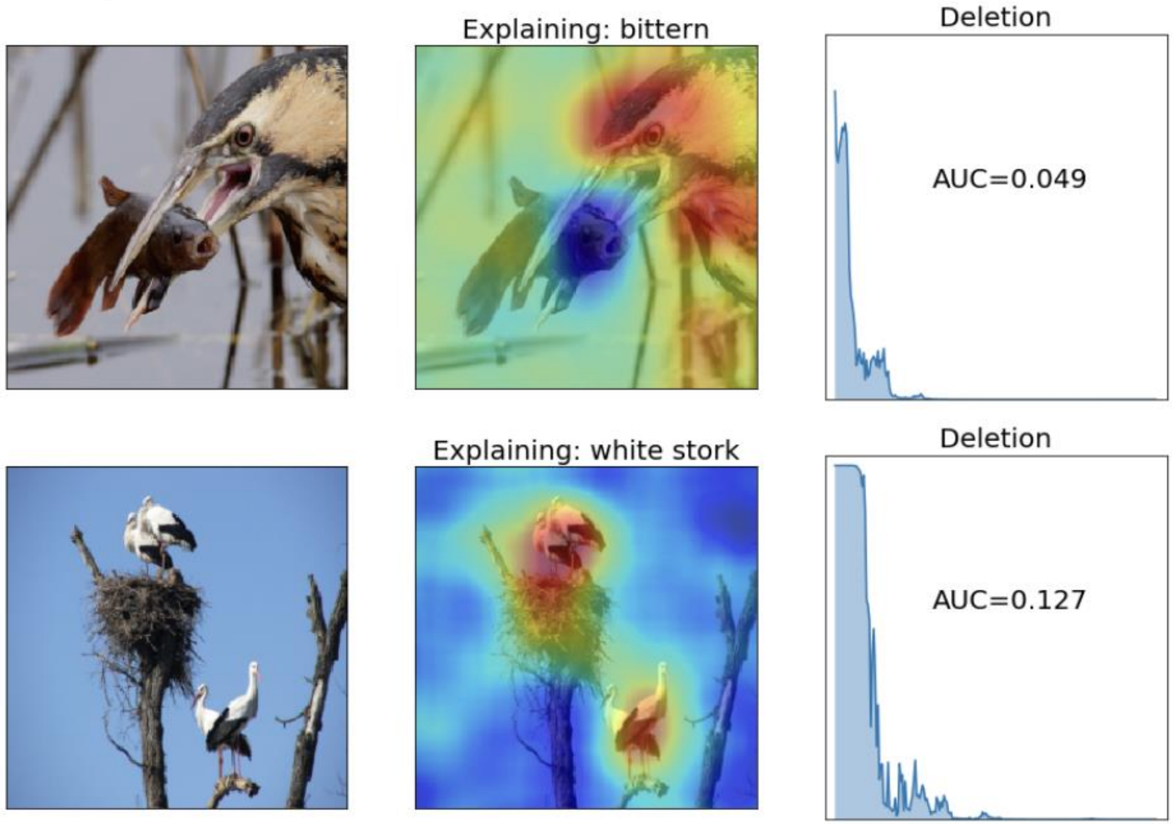
Step 1: Mask highest importance pixel and pass the image through the network. Note the probability of spoonbill.
Step 2: Mask the second highest importance pixel from the image in Step 1 and pass the image through the network. Note the probability of spoonbill.
Step 3: Repeat until all pixels are deleted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

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

Network Evaluation

Evaluation 2: Progressive Pixel-wise Insertion and Deletion

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



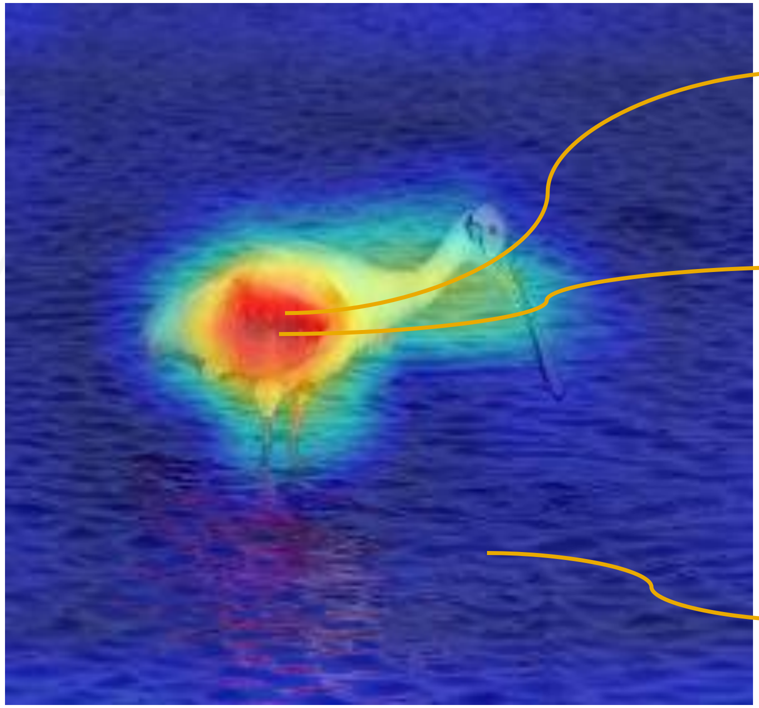
- **Deletion approximates Necessity** criterion of a "good" explanation
- **AUC** for a good explanation will be **low**
- **Deletion** encourages **fine-grained explanations** by choosing those heatmaps that select the most relevant pixels

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

Network Evaluation

Evaluation 2: Progressive Pixel-wise Insertion and Deletion

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



Highest importance

Second Highest importance

.
.
.

Least importance

Take a mean (grayscale) image

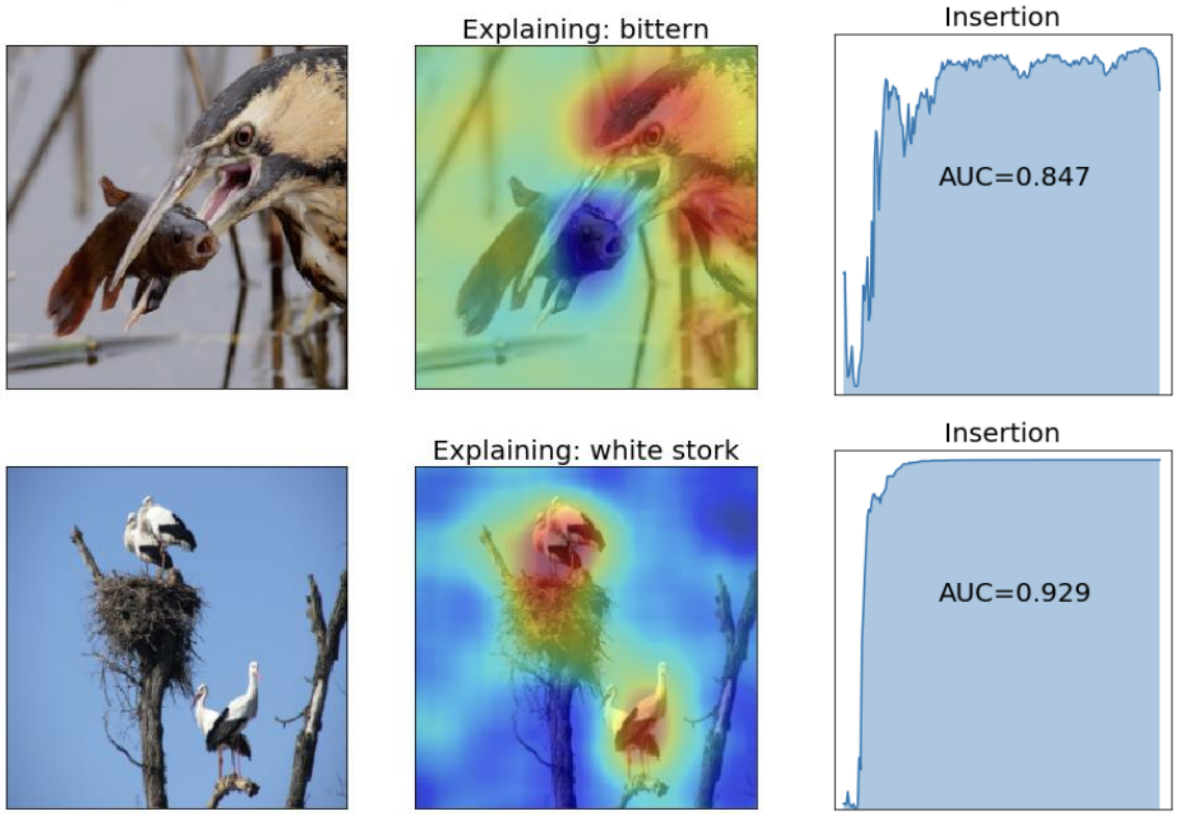
Step 1: Add the highest importance pixel to the mean image and pass it through the network. Note the probability of spoonbill.
Step 2: Add the second highest importance pixel to the image in Step 1 and pass the image through the network. Note the probability of spoonbill.
Step 3: Repeat until all pixels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

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

Network Evaluation

Evaluation 2: Progressive Pixel-wise Insertion and Deletion

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



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

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

Appendix

Resources

- Explanatory Paradigms: AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory Paradigms in Neural Networks." *arXiv preprint arXiv:2202.11838* (2022).
- Abductive Reasoning and Neural Networks: Prabhushankar, Mohit, and Ghassan AlRegib. "Contrastive Reasoning in Neural Networks." *arXiv preprint arXiv:2103.12329* (2021).
- Grad-CAM and Counterfactual-CAM: 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.
- Contrast-CAM: 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.
- Explanatory Evaluation Taxonomy: AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory Paradigms in Neural Networks." *arXiv preprint arXiv:2202.11838* (2022).
- CURE-TSR dataset: Temel, Dogancan, et al. "CURE-TSR: Challenging unreal and real environments for traffic sign recognition." *arXiv preprint arXiv:1712.02463* (2017).
- Saliency and Gaze Tracking: Sun, Yutong, Mohit Prabhushankar, and Ghassan AlRegib. "Implicit saliency in deep neural networks." *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020.

Appendix
Notations

- x_i: a single feature
- $\boldsymbol{x}_i$: feature vector (a data sample)
- $\boldsymbol{x}_{:,i}$: feature vector of all data samples
- $\boldsymbol{X}$: matrix of feature vectors (dataset)
- N: number of data samples
- $\boldsymbol{W}$: weight matrix
- $\boldsymbol{b}$: bias vector
- $\boldsymbol{v}(t)$: first moment at time t
- $\boldsymbol{G}(t)$: second moment at time t
- $\boldsymbol{H}(\boldsymbol{\theta})$: Hessian matrix
- P: number of features in a feature vector
- α: learning rate
- Bold letter/symbol: vector
- Bold capital letters/symbol: matrix