

Visual Explainability in Machine Learning

Lecture 4: Visual Explanations II



Ghassan AlRegib, PhD
Professor



Mohit Prabhushankar, PhD
Postdoctoral Fellow

Omni Lab for Intelligent Visual Engineering and Science (OLIVES)
School of Electrical and Computer Engineering
Georgia Institute of Technology
{alregib, mohit.p}@gatech.edu

Dec 5, 2023

Short Course Materials

Accessible Online



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

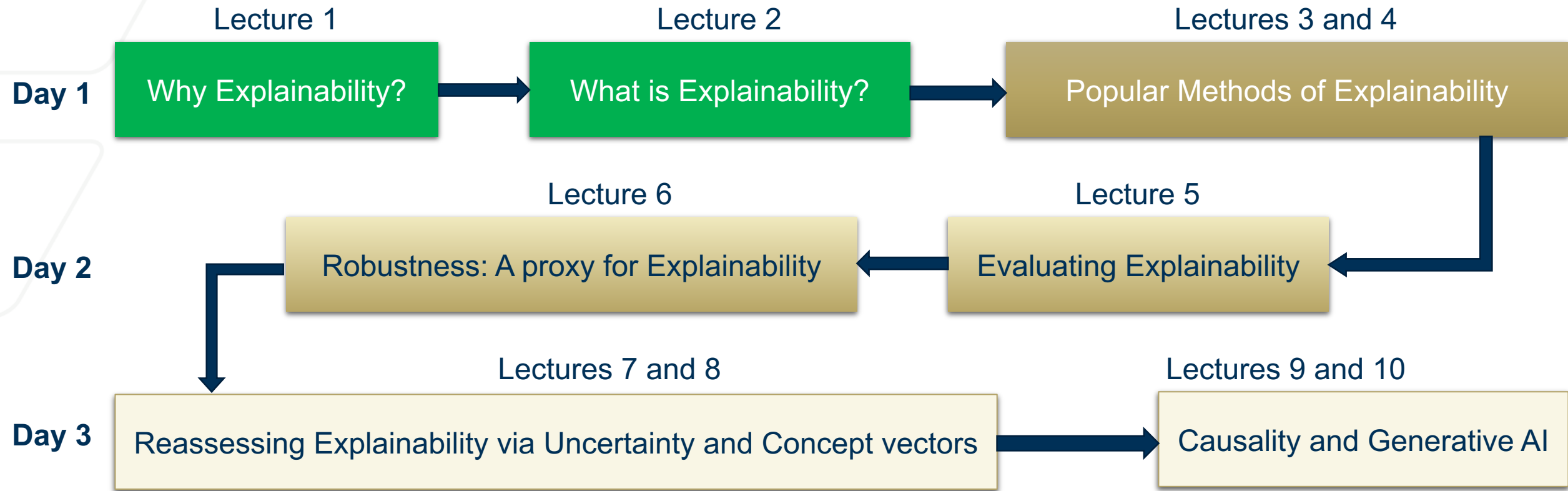
<https://alregib.ece.gatech.edu/>

<https://alregib.ece.gatech.edu/sps-education-short-course/>
{alregib, mohit.p}@gatech.edu

Short Course

Course Outline

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



Outline

Lecture 4: Visual Explanations II

- Targeted Explanations
 - Role of Explainability
 - Definition
 - Contextual Questions
- GradCAM: Gradient weighted Class Activation Maps
 - Methodology
 - Results
- Explanatory Paradigms
 - CounterfactualCAM
 - ContrastCAM
 - Results
- Case Study: Image Quality Assessment
- Takeaways

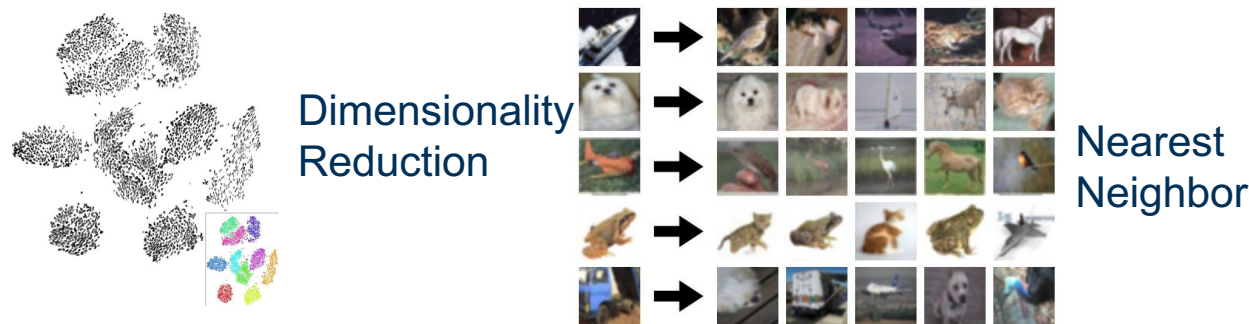
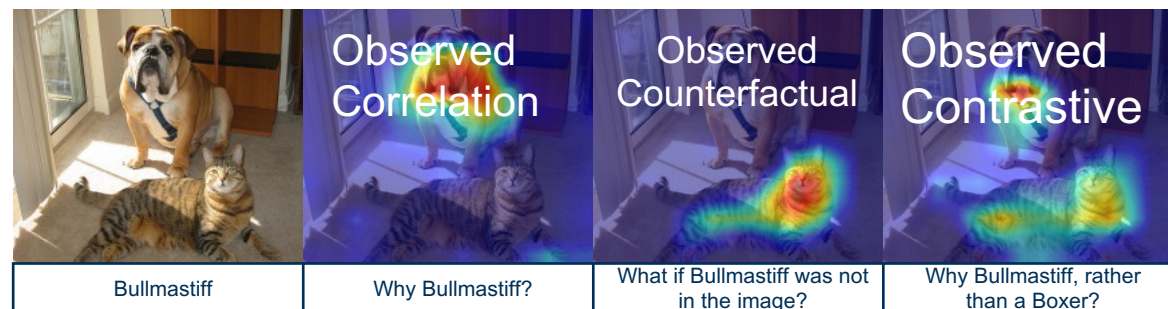
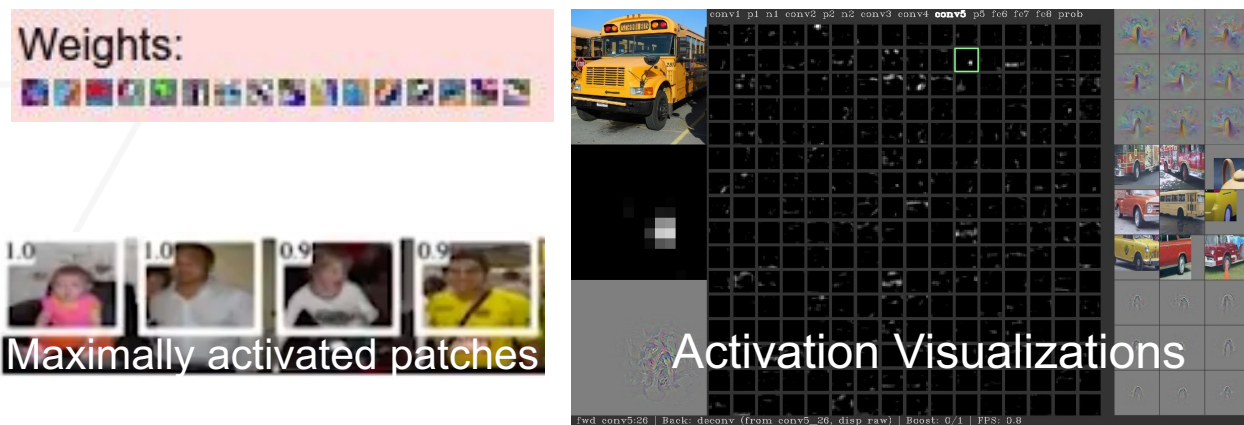
Explanations

Human-centric Explanations

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

Lecture 3: Indirect and Direct Explanations

Lecture 4: Targeted Explanations



Outline

Lecture 4: Visual Explanations II

- Targeted Explanations
 - Role of Explainability
 - Definition
 - Contextual Questions
- GradCAM: Gradient weighted Class Activation Maps
 - Methodology
 - Results
- Explanatory Paradigms
 - CounterfactualCAM
 - ContrastCAM
 - Results
- Case Study: Image Quality Assessment
- Takeaways

Explanations

Targeted Explanations

Targeted explanations highlight contextually relevant regions in an image

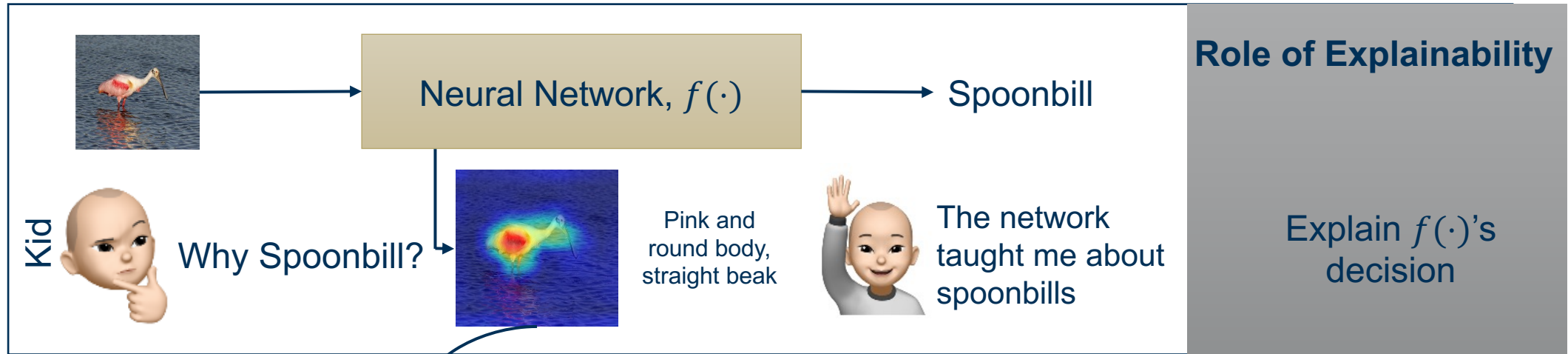
- **Required knowledge to understand explanations:** Knowledge about data and classes
- **Required knowledge to obtain explanations:** Models, model parameters, and training data
- **Explanations audience:** Researchers and Engineers, users, policymakers, and general public
- **Explanatory chronology:** Newer explanatory techniques are targeted

Methods	Definition		
	Indirect	Direct	Targeted
Deconvolution [21]	✓	—	—
Inverted Representations [22]	✓	—	—
Guided-Backpropagation [18]	—	✓	—
SmoothGrad [17]	—	✓	—
LIME [39]	—	✓	—
CAM [24]	—	✓	—
Graph-CNN [23]	✓	—	—
GradCAM [12]	—	—	✓
TCAV [40]	—	✓	—
GradCAM++ [16]	—	—	✓
RISE [35]	—	✓	—
Causal-CAM [15]	—	—	✓
Counterfactual-CAM [12]	—	—	✓
Goyal et al. [26]	—	—	✓
CEM [29]	—	—	✓
Contrast-CAM [13]	—	—	✓
Contrastive reasoning [14]	—	—	✓

Targeted Explanations

Role of Explainability

Targeted explanations enhance the role of Explainability



Direct Explanation

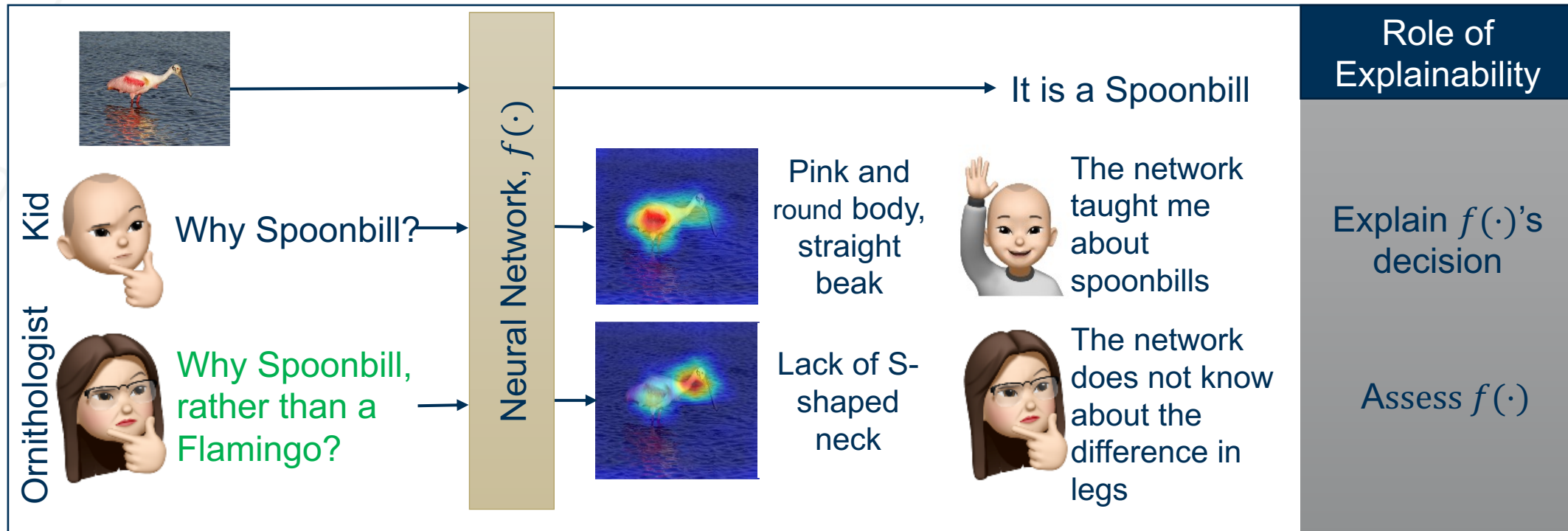
Kid does not have knowledge about Spoonbill!

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

Targeted Explanations

Role of Explainability

Targeted explanations enhance the role of Explainability

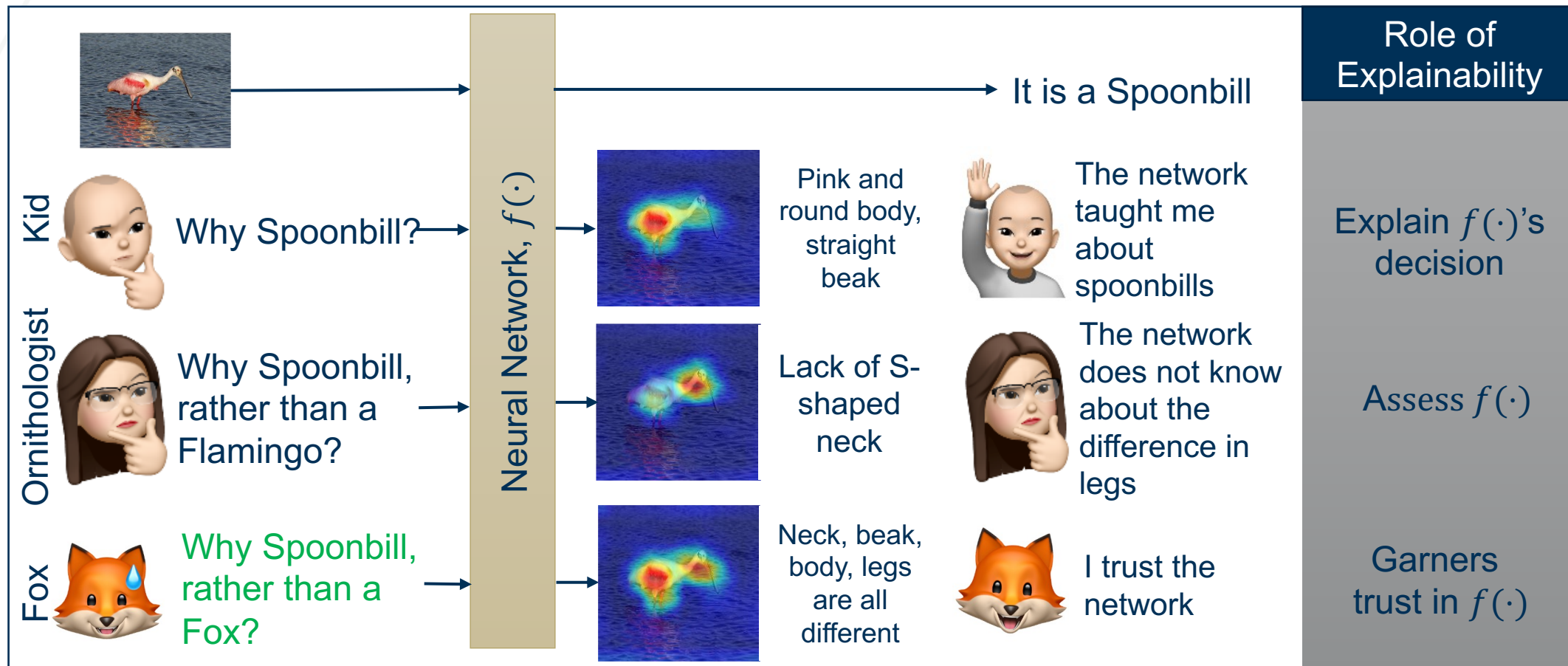


The ornithologist has knowledge of Spoonbills and Flamingos and uses this knowledge to ask **targeted** questions!

Targeted Explanations

Role of Explainability

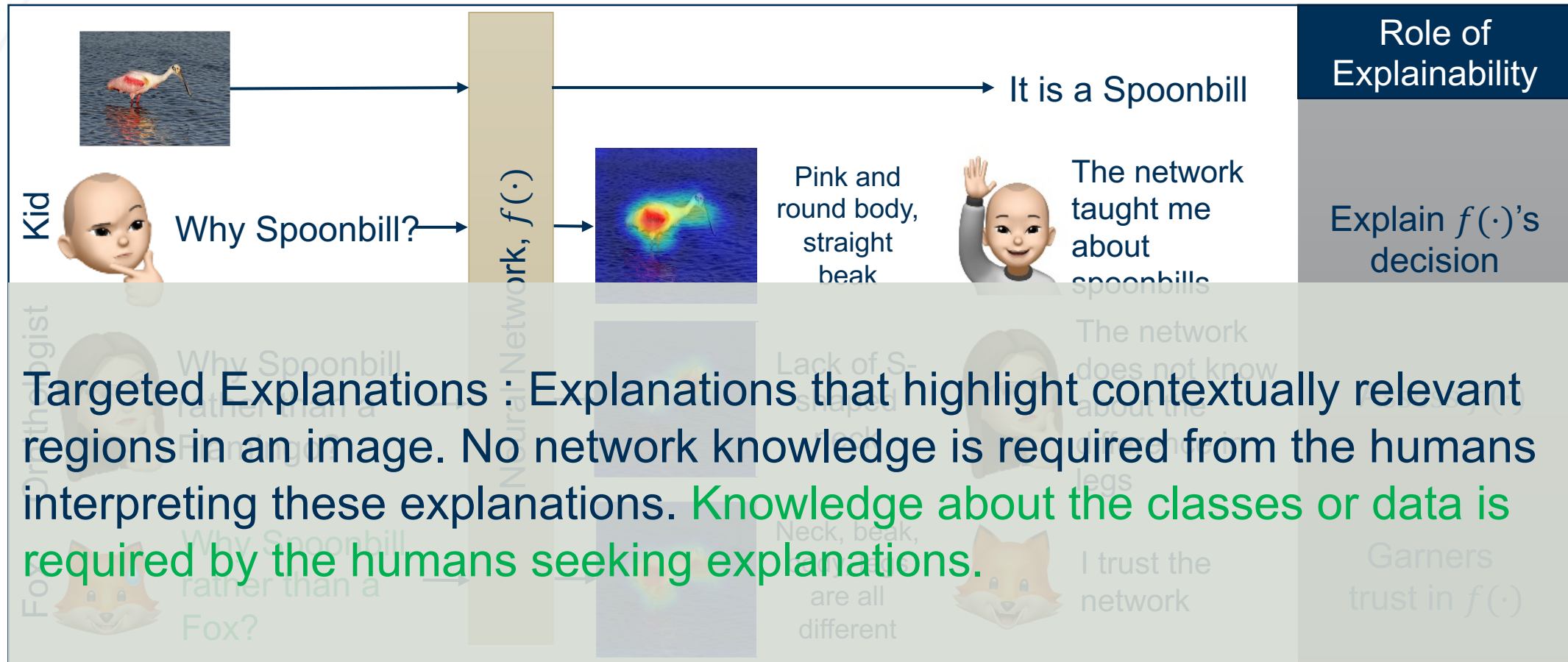
More the interaction, more is the trust in the base network



Targeted Explanations

Definition

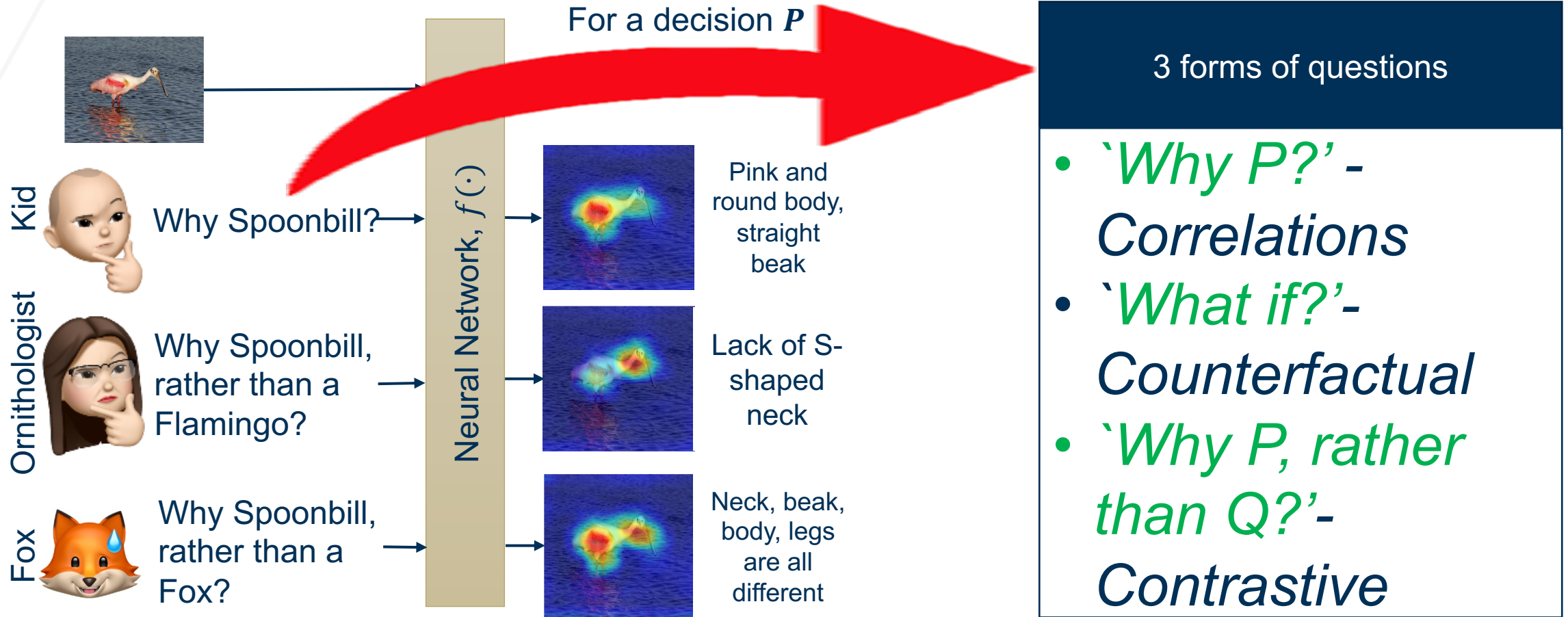
Targeted Explanations highlight contextually relevant regions in an image



Targeted Explanations

Contextually-relevant Questions

Targeted Explanations highlight contextually relevant regions in an image



Outline

Lecture 4: Visual Explanations II

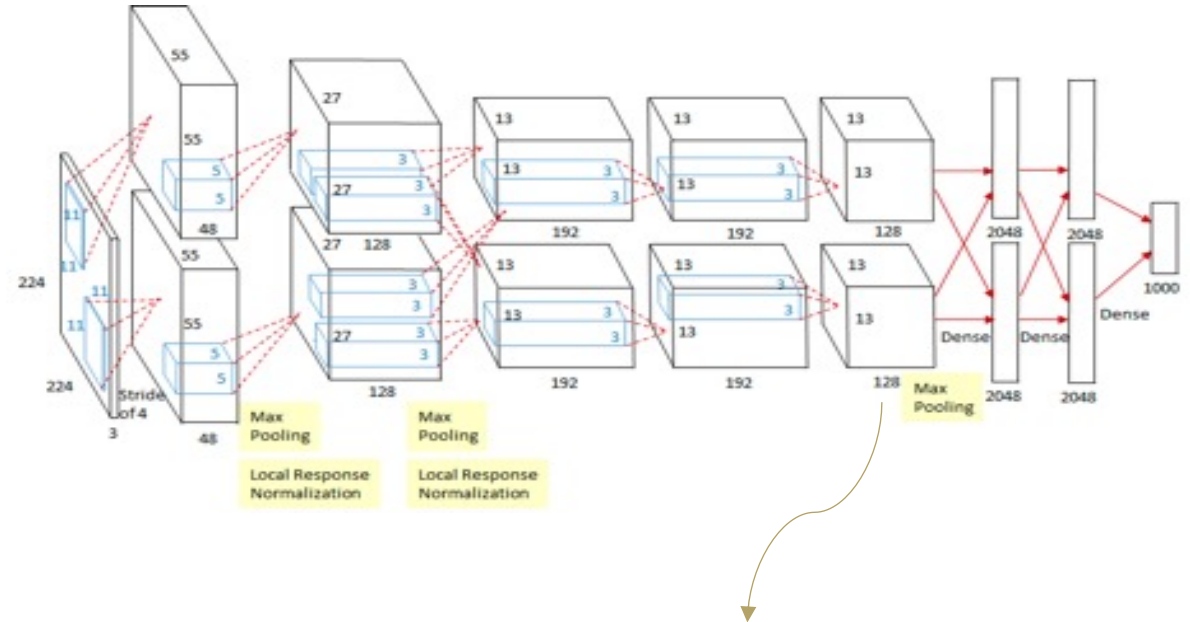
- Targeted Explanations
 - Role of Explainability
 - Definition
 - Contextual Questions
- **GradCAM: Gradient weighted Class Activation Maps**
 - **Methodology**
 - **Results**
- Explanatory Paradigms
 - CounterfactualCAM
 - ContrastCAM
 - Results
- Case Study: Image Quality Assessment
- Takeaways

Targeted Explanations

Gradient-weighted Class Activation Mapping: GradCAM

Gradients provide feature importance *for all* available features; Activations provide class-discriminative features

- GradCAM combines activations and gradients
- Which layer to extract activations:
 - **Higher layers** capture **class specific information**
 - **Spatial information is lost in fully-connected layers**
 - **Last convolutional layer** forms a best compromise between high-level semantics and detailed spatial resolution



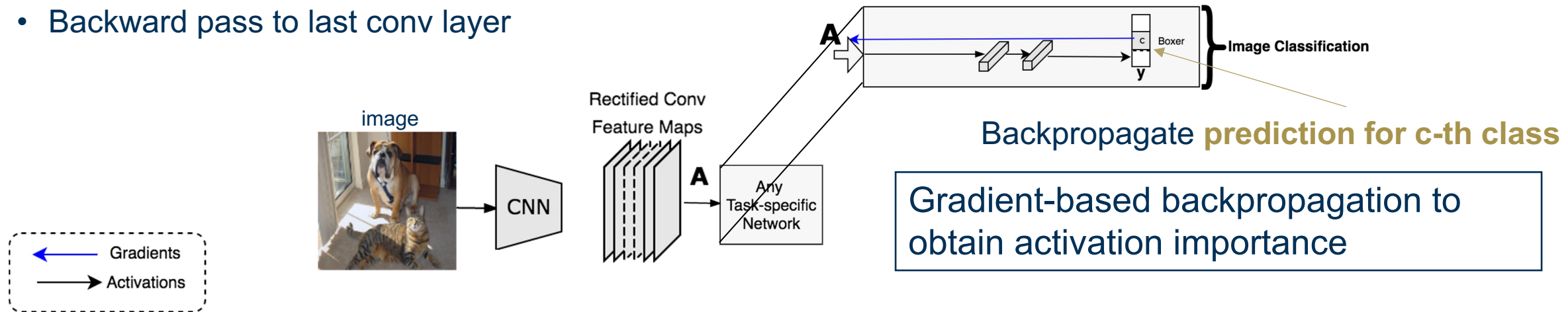
Ideal layer for activation extraction

Targeted Explanations

Gradient-weighted Class Activation Mapping: GradCAM Methodology

Gradients provide feature importance *for all* available features; Activations provide class-discriminative features

- Given an image, feed forward through CNN
- Final convolutional layer output feature maps for later task-specific layers, i.e., fc layer for classification
- Backward pass to last conv layer

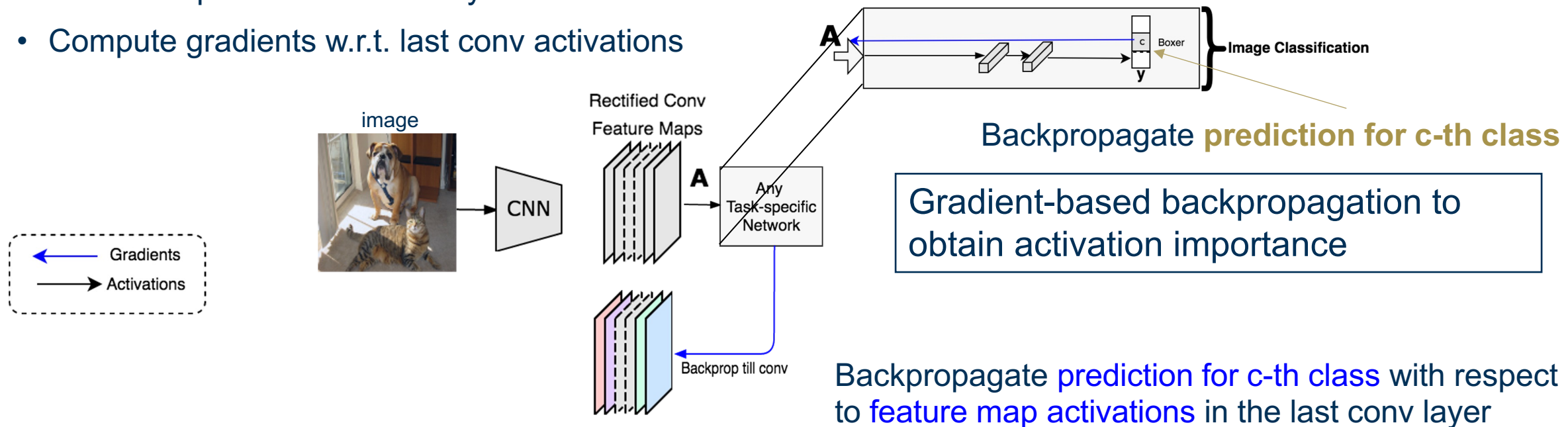


Targeted Explanations

Gradient-weighted Class Activation Mapping: GradCAM Methodology

Gradients provide feature importance *for all* available features; Activations provide class-discriminative features

- Given an image, feed forward through CNN
- Final convolutional layer output feature maps for later task-specific layers
- Backward pass to last conv layer
- Compute gradients w.r.t. last conv activations

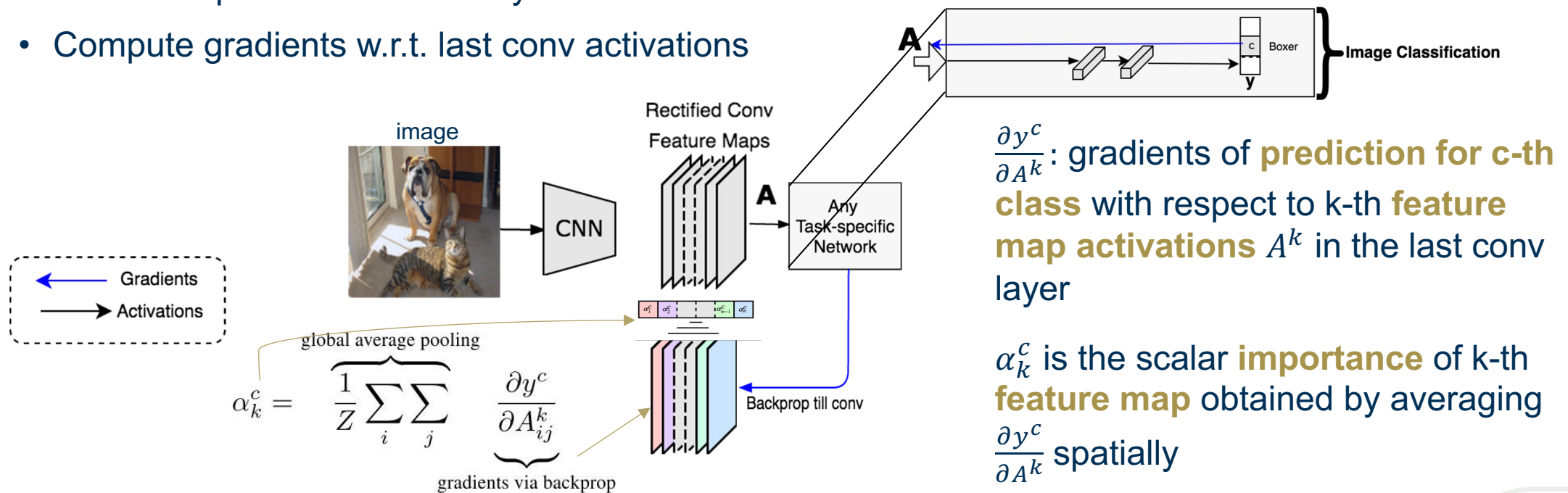


Targeted Explanations

Gradient-weighted Class Activation Mapping: GradCAM Methodology

Gradients provide feature importance for all available features; Activations provide class-discriminative features

- Given an image, feed forward through CNN
- Final convolutional layer output feature maps for later task-specific layers
- Backward pass to last conv layer
- Compute gradients w.r.t. last conv activations

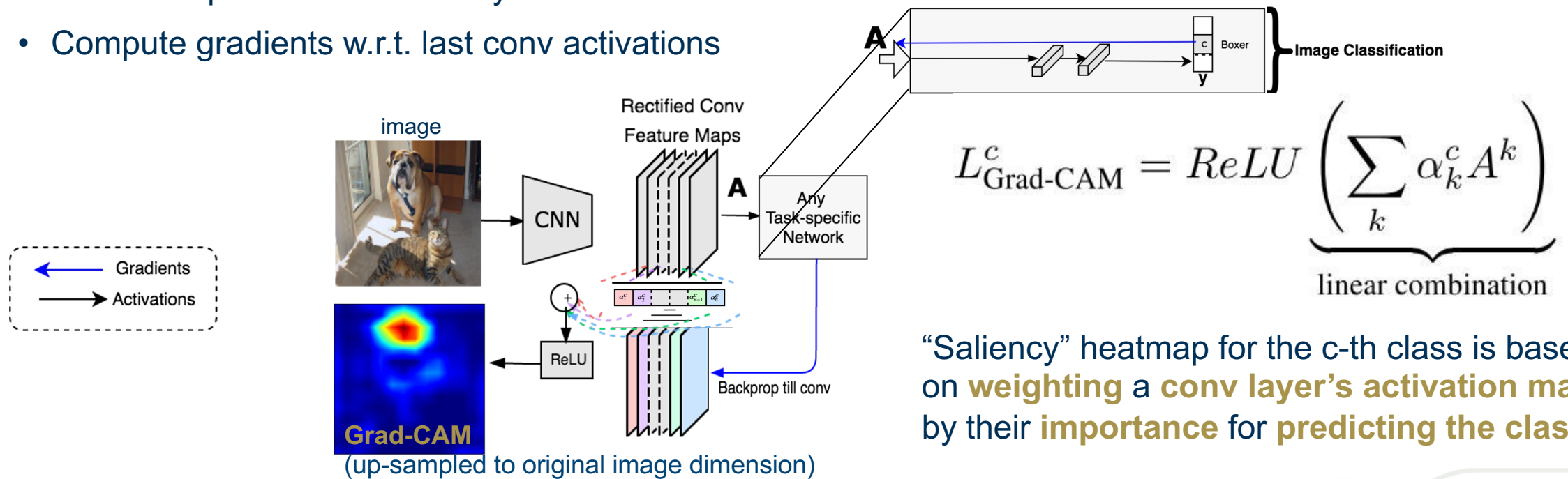


Targeted Explanations

Gradient-weighted Class Activation Mapping: GradCAM Methodology

Gradients provide feature importance for all available features; Activations provide class-discriminative features

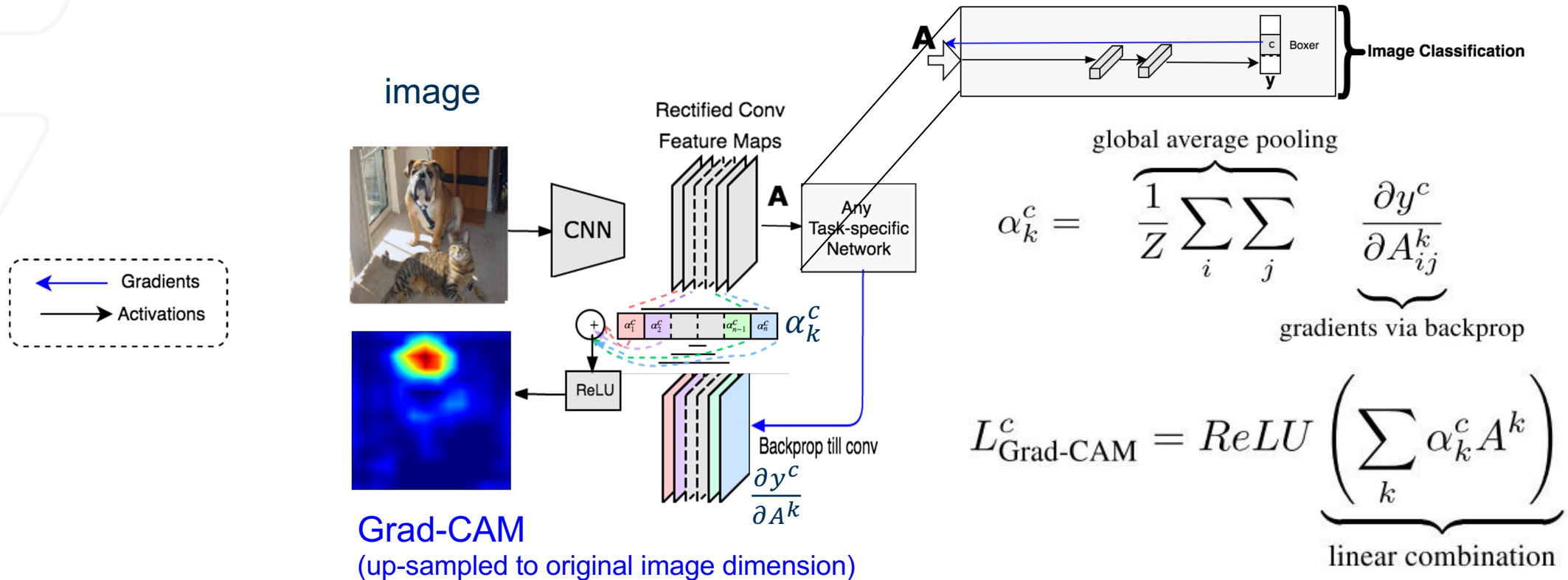
- Given an image, feed forward through CNN
- Final convolutional layer output feature maps for later task-specific layers
- Backward pass to last conv layer
- Compute gradients w.r.t. last conv activations



Targeted Explanations

Gradient-weighted Class Activation Mapping: GradCAM Methodology

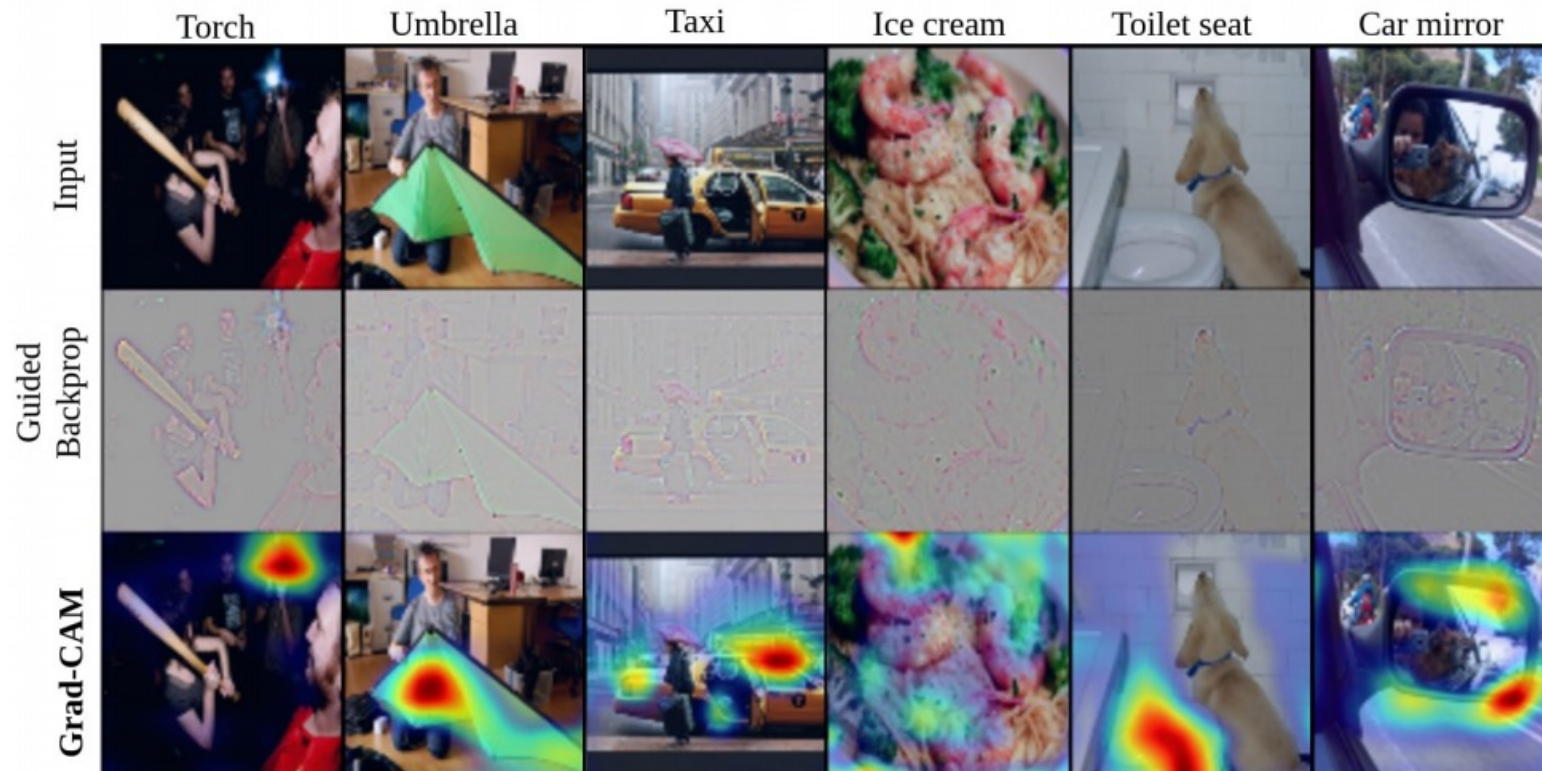
Grad-CAM uses the gradient information in the last convolutional layer of the CNN to assign importance values to each activation for any class c



Targeted Explanations

Gradient-weighted Class Activation Mapping: GradCAM Results

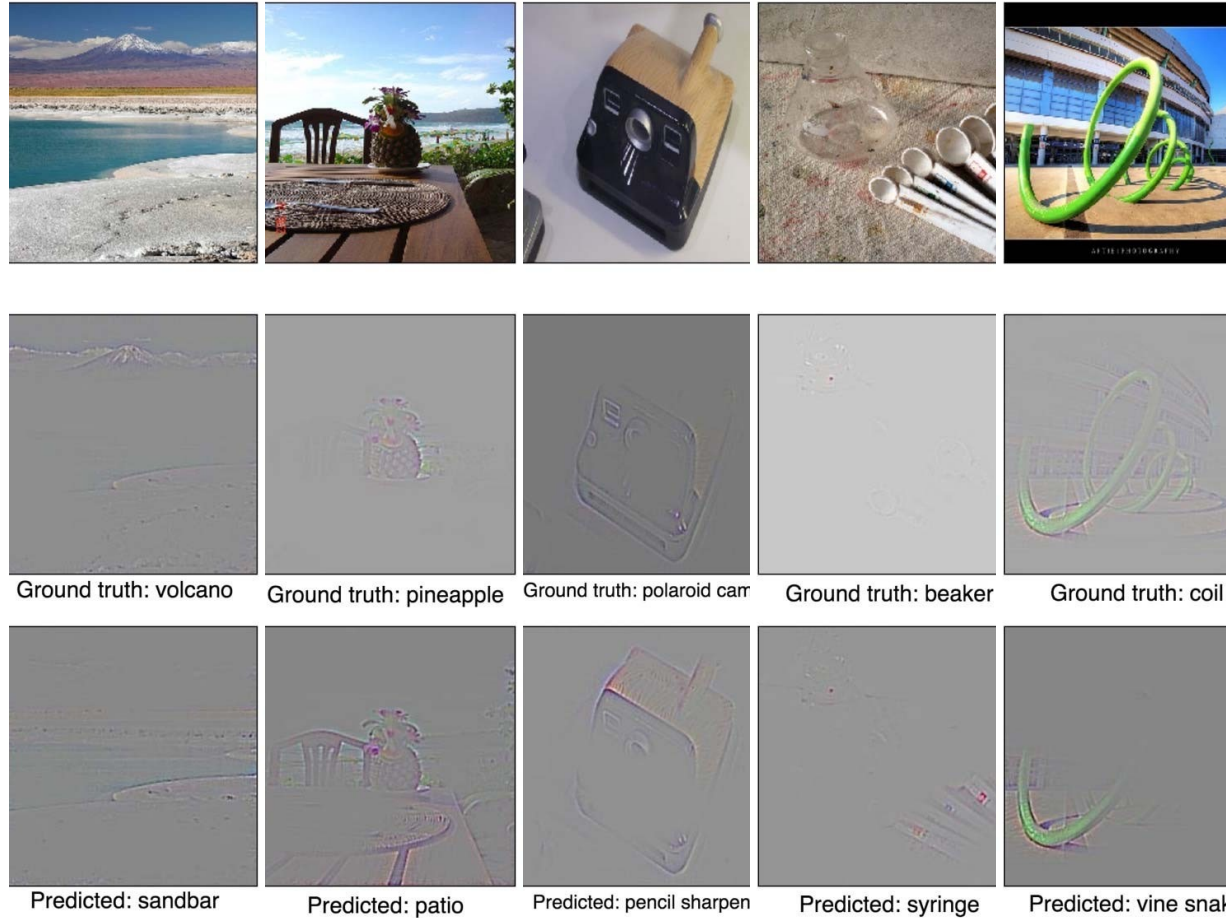
Results in image classification



Targeted Explanations

Gradient-weighted Class Activation Mapping: GradCAM Results

Results when the model has misclassified



Outline

Lecture 4: Visual Explanations II

- Targeted Explanations
 - Role of Explainability
 - Definition
 - Contextual Questions
- GradCAM: Gradient weighted Class Activation Maps
 - Methodology
 - Results
- **Explanatory Paradigms**
 - **CounterfactualCAM**
 - **ContrastCAM**
 - **Results**
- Case Study: Image Quality Assessment
- Takeaways

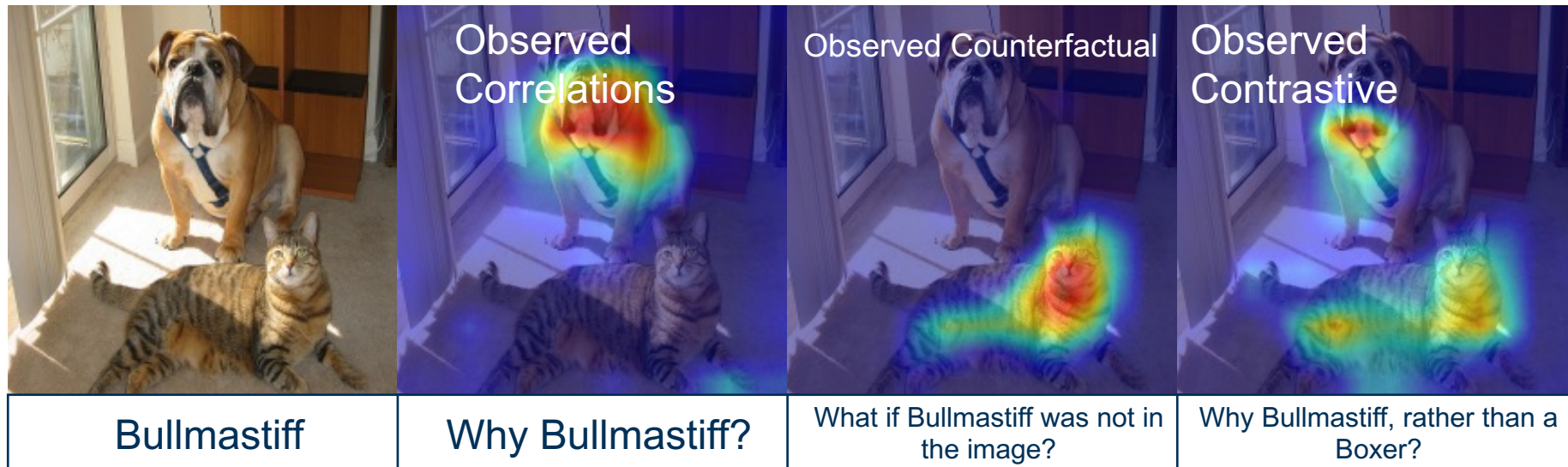
Gradient and Activation-based Explanations

Explanatory Paradigms



Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

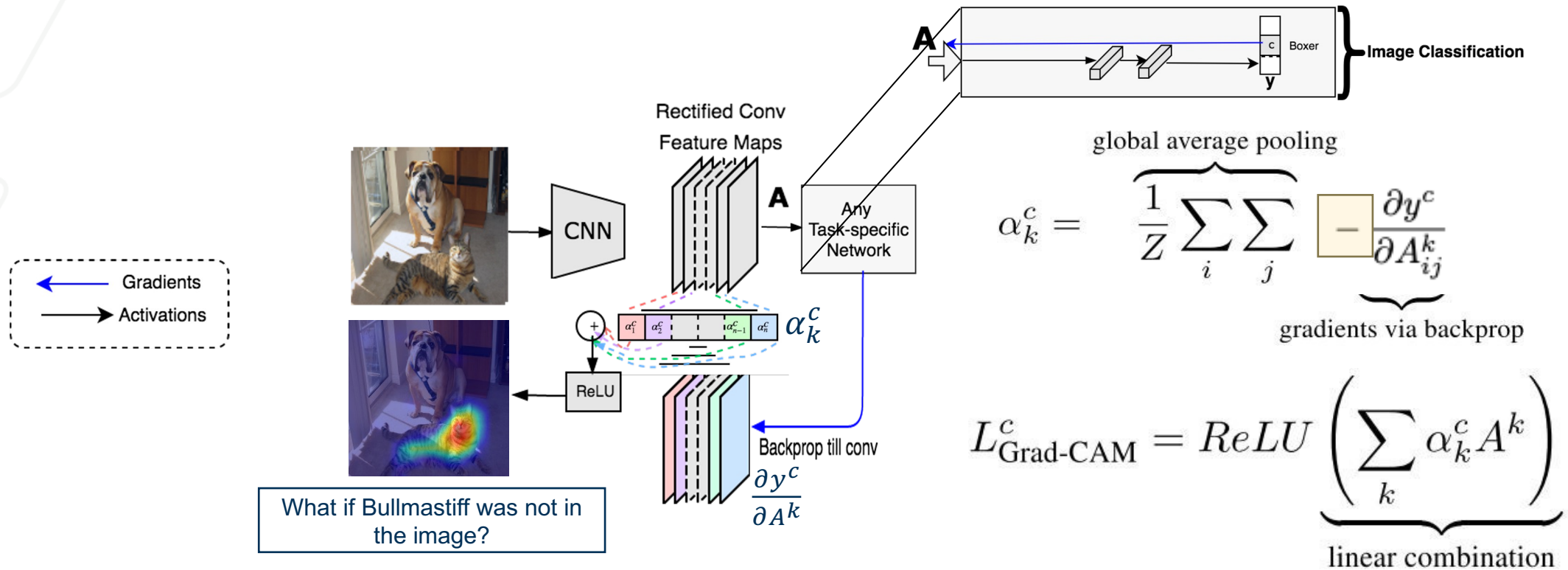
GradCAM provides answers to ‘Why P?’ questions. But different stakeholders require relevant and contextual explanations



Gradient and Activation-based Explanations

CounterfactualCAM: What if this region were absent in the image?

In GradCAM, global average pool the **negative of** gradients to obtain α^c for each kernel k

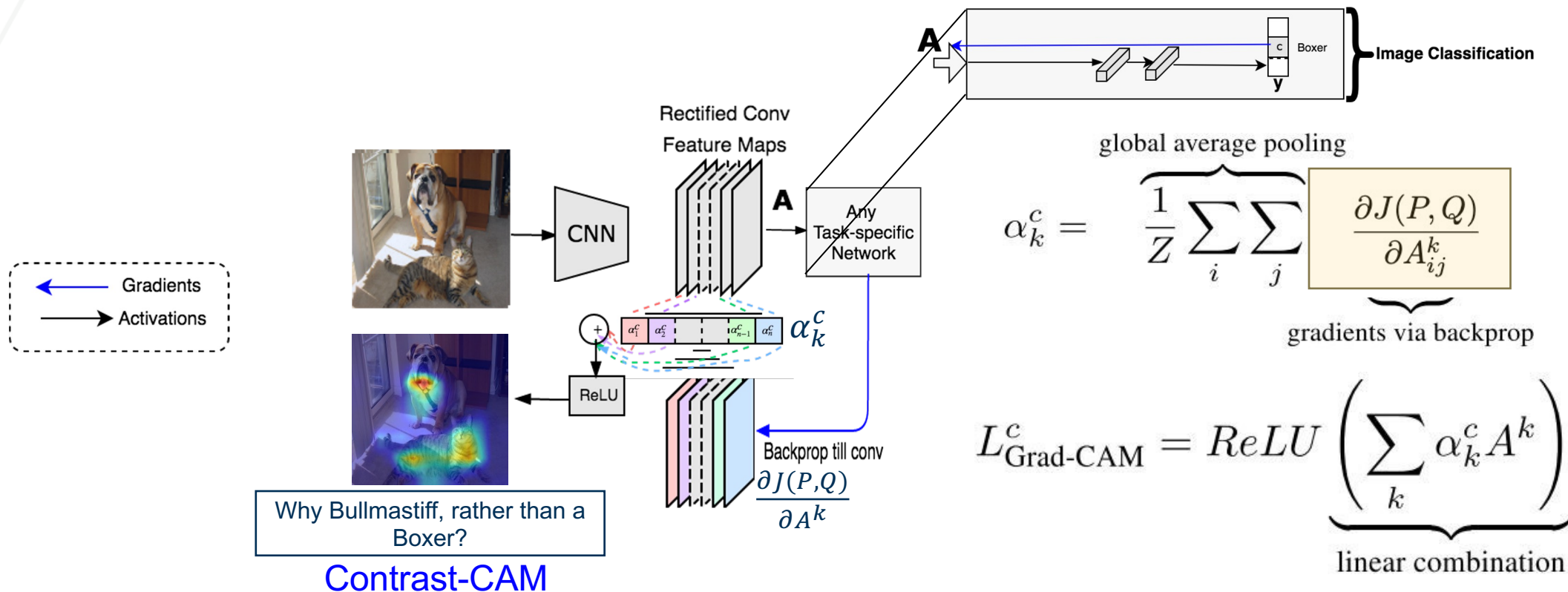


Negating the gradients effectively removes these regions from analysis

Gradient and Activation-based Explanations

ContrastCAM: Why P, rather than Q?


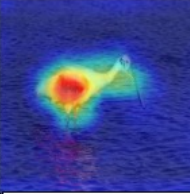

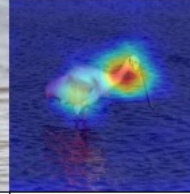

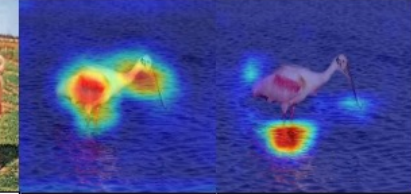





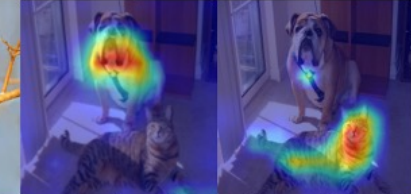

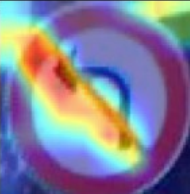

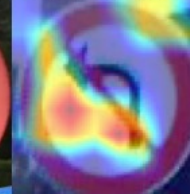







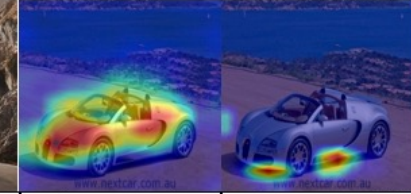
In GradCAM, backward pass the **loss between predicted class P and some contrast class Q** to last conv layer



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


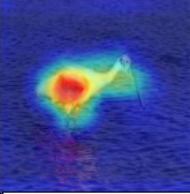

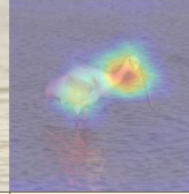
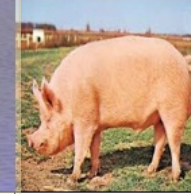
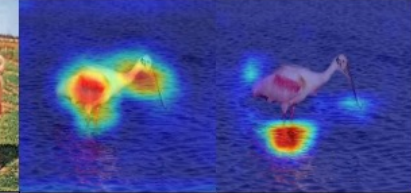





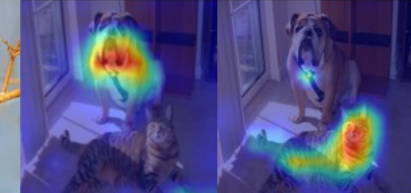

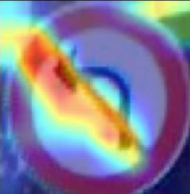

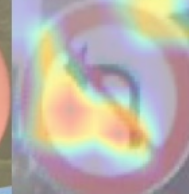







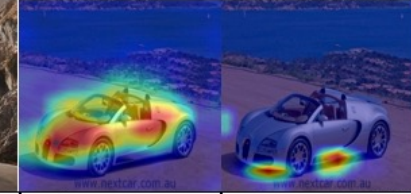
Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

Input Image	Grad-CAM	Contrast 1	Contrastive Explanation 1	Contrast 2	Contrastive Explanation 2
					
ImageNet dataset : Spoonbill	Grad-CAM : Why Spoonbill?	Representative Flamingo image	Why Spoonbill, rather than Flamingo?	Representative Pig image	Why Spoonbill, rather than Pig? Why not Spoonbill, with 100% confidence?
					
ImageNet dataset : Bull Mastiff	Grad-CAM : Why : Bull Mastiff?	Representative Boxer image	Why Bull Mastiff, rather than Boxer?	Representative Blue jay image	Why Bull Mastiff, rather than Blue jay? Why not Bull Mastiff, with 100% confidence?
					
CURE-TSR dataset : No-Left Image	Grad-CAM : Why No-Left?	Representative No-Right image	Why No-Left, rather than No-Right?	Representative Stop Sign	Why No-Left, rather than Stop? Why not No-Left with 100% confidence?
					
Stanford Cars Dataset: Bugatti Convertible	Grad-CAM: Why Bugatti Convertible?	Representative Bugatti Coupe image	Why Convertible, rather than Coupe?	Representative Audi A6 image	Why Bugatti, rather than Audi A6? Why not Bugatti with 100% confidence?

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

Input Image	Grad-CAM	Contrast 1	Contrastive Explanation 1	Contrast 2	Contrastive Explanation 2
					
ImageNet dataset : Spoonbill	Grad-CAM : Why Spoonbill?	Representative Flamingo image	Why Spoonbill, rather than Flamingo?	Representative Pig image	Why Spoonbill, rather than Pig? Why not Spoonbill, with 100% confidence?
					
ImageNet dataset : Bull Mastiff	Grad-CAM : Why : Bull Mastiff?	Representative Boxer image	Why Bull Mastiff, rather than Boxer?	Representative Blue jay image	Why Bull Mastiff, rather than Blue jay? Why not Bull Mastiff, with 100% confidence?
					
CURE-TSR dataset : No-Left Image	Grad-CAM : Why No-Left?	Representative No-Right image	Why No-Left, rather than No-Right?	Representative Stop Sign	Why No-Left, rather than Stop? Why not No-Left with 100% confidence?
					
Stanford Cars Dataset: Bugatti Convertible	Grad-CAM: Why Bugatti Convertible?	Representative Bugatti Coupe image	Why Convertible, rather than Coupe?	Representative Audi A6 image	Why Bugatti, rather than Audi A6? Why not Bugatti with 100% confidence?

Human Interpretable

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM


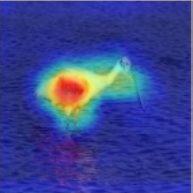
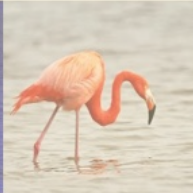
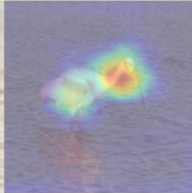

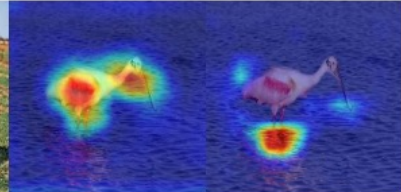

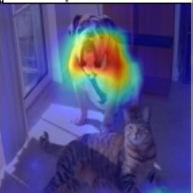

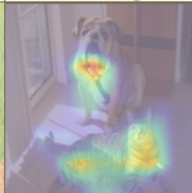
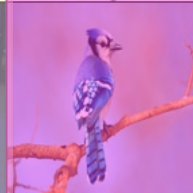


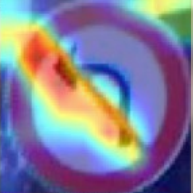

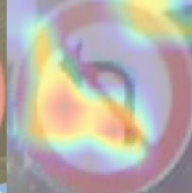



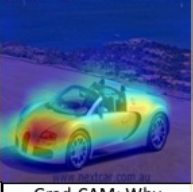
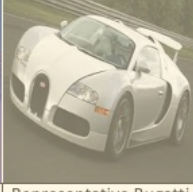

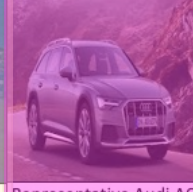

Input Image	Grad-CAM	Contrast 1	Contrastive Explanation 1	Contrast 2	Contrastive Explanation 2
ImageNet dataset : Spoonbill	Grad-CAM : Why Spoonbill?	Representative Flamingo image	Why Spoonbill, rather than Flamingo?	Representative Pig image	Why Spoonbill, rather than Pig? Why not Spoonbill, with 100% confidence?
ImageNet dataset : Bull Mastiff	Grad-CAM : Why : Bull Mastiff?	Representative Boxer image	Why Bull Mastiff, rather than Boxer	Representative Blue jay image	Why Bull Mastiff, rather than Blue jay? Why not Bull Mastiff, with 100% confidence?
CURE-TSR dataset : No-Left Image	Grad-CAM : Why No-Left?	Representative No-Right image	Why No-Left, rather than No-Right?	Representative Stop Sign	Why No-Left, rather than Stop? Why not No-Left with 100% confidence?
Stanford Cars Dataset: Bugatti Convertible	Grad-CAM: Why Bugatti Convertible?	Representative Bugatti Coupe image	Why Convertible, rather than Coupe?	Representative Audi A6 image	Why Bugatti, rather than Audi A6? Why not Bugatti with 100% confidence?

Human Interpretable

Same as Grad-CAM

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

Input Image	Grad-CAM	Contrast 1	Contrastive Explanation 1	Contrast 2	Contrastive Explanation 2
					
ImageNet dataset : Spoonbill	Grad-CAM : Why Spoonbill?	Representative Flamingo image	Why Spoonbill, rather than Flamingo?	Representative Pig image	Why Spoonbill, rather than Pig? Why not Spoonbill, with 100% confidence?
					
ImageNet dataset : Bull Mastiff	Grad-CAM : Why : Bull Mastiff?	Representative Boxer image	Why Bull Mastiff, rather than Boxer	Representative Blue jay image	Why Bull Mastiff, rather than Blue jay? Why not Bull Mastiff, with 100% confidence?
					
CURE-TSR dataset : No-Left Image	Grad-CAM : Why No-Left?	Representative No-Right image	Why No-Left, rather than No-Right?	Representative Stop Sign	Why No-Left, rather than Stop? Why not No-Left with 100% confidence?
					
Stanford Cars Dataset: Bugatti Convertible	Grad-CAM: Why Bugatti Convertible?	Representative Bugatti Coupe image	Why Convertible, rather than Coupe?	Representative Audi A6 image	Why Bugatti, rather than Audi A6? Why not Bugatti with 100% confidence?

Human Interpretable

Same as Grad-CAM

Not Human Interpretable

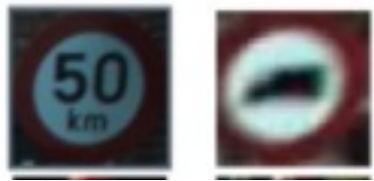
Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

Input Image	Grad-CAM	Contrast 1	Contrastive Explanation 1	Contrast 2	Contrastive Explanation 2
ImageNet dataset : Spoonbill	Grad-CAM : Why Spoonbill?	Representative Flamingo image	Why Spoonbill, rather than Flamingo?	Representative Pig image	Why Spoonbill, rather than Pig? Why not Spoonbill, with 100% confidence?

Human Interpretable

Same as Grad-CAM



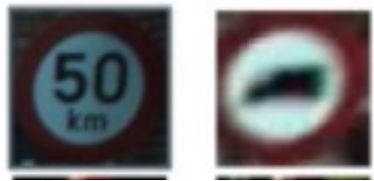
CURE-TSR dataset : No-Left Image	Grad-CAM : Why No-Left?	Representative No-Right image	Why No-Left, rather than No-Right?	Representative Stop Sign	Why No-Left, rather than Stop?	Why not No-Left with 100% confidence?
Stanford Cars Dataset: Bugatti Convertible	Grad-CAM: Why Bugatti Convertible?	Representative Bugatti Coupe image	Why Convertible, rather than Coupe?	Representative Audi A6 image	Why Bugatti, rather than Audi A6?	Why not Bugatti with 100% confidence?

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

Input Image	Grad-CAM	Contrast 1	Contrastive Explanation 1	Contrast 2	Contrastive Explanation 2
ImageNet dataset : Spoonbill	Grad-CAM : Why Spoonbill?	Representative Flamingo image	Why Spoonbill, rather than Flamingo?	Representative Pig image	Why Spoonbill, rather than Pig? with 100% confidence?

Only traffic sign with a straight bottom-left edge – enough to say 'Not STOP Sign'



CURE-TSR dataset : No-Left Image	Grad-CAM : Why No-Left?	Representative No-Right image	Why No-Left, rather than No-Right?	Representative Stop Sign	Why No-Left, rather than Stop?	Why not No-Left with 100% confidence?
Stanford Cars Dataset: Bugatti Convertible	Grad-CAM: Why Bugatti Convertible?	Representative Bugatti Coupe image	Why Convertible, rather than Coupe?	Representative Audi A6 image	Why Bugatti, rather than Audi A6?	Why not Bugatti with 100% confidence?

Outline

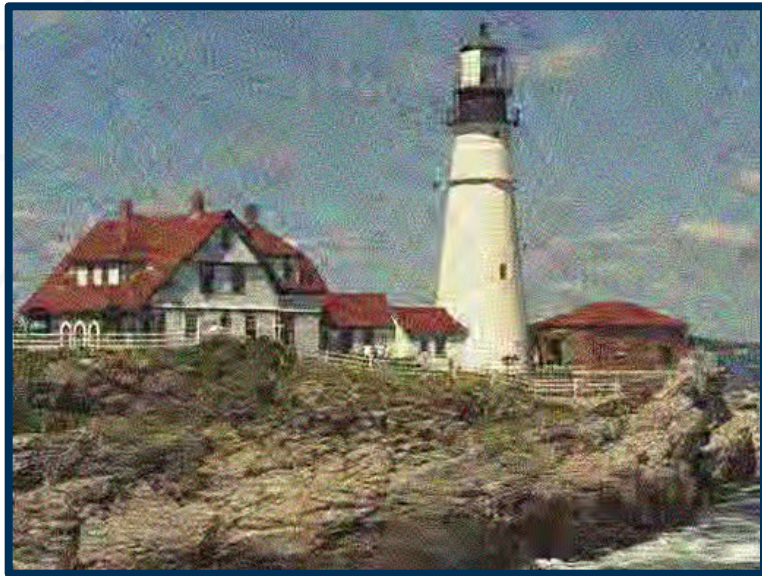
Lecture 4: Visual Explanations II

- Targeted Explanations
 - Role of Explainability
 - Definition
 - Contextual Questions
- GradCAM: Gradient weighted Class Activation Maps
 - Methodology
 - Results
- Explanatory Paradigms
 - CounterfactualCAM
 - ContrastCAM
 - Results
- Case Study: Image Quality Assessment
- Takeaways

Case Study: Image Quality Assessment

What is IQA?

IQA is the objective Assessment of Subjective Quality



Lighthouse image with level 5 lossy compression from TID 2013 dataset



Image Quality Assessment
Algorithm :
DIQaM [1]

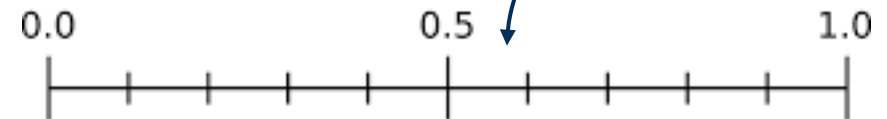


Score : 0.58

The given image is
somewhat OK quality

Bad
Quality

Good
Quality



Case Study: Image Quality Assessment

Dataset Construction



Subjects make contrastive choices during Dataset Construction

- Subjects are shown a reference image in a controlled setting
- Based on the reference image, they are asked pick one of the images on the top that differs least from the reference image
- Reference image sets the expectancy
- The task of subjectively picking the least mismatched image is IQA

This requires **Fine-grained** Analysis!

Case Study: Image Quality Assessment

Expectancy-Mismatch in Dataset Construction



Subjects make contrastive choices during Dataset Construction

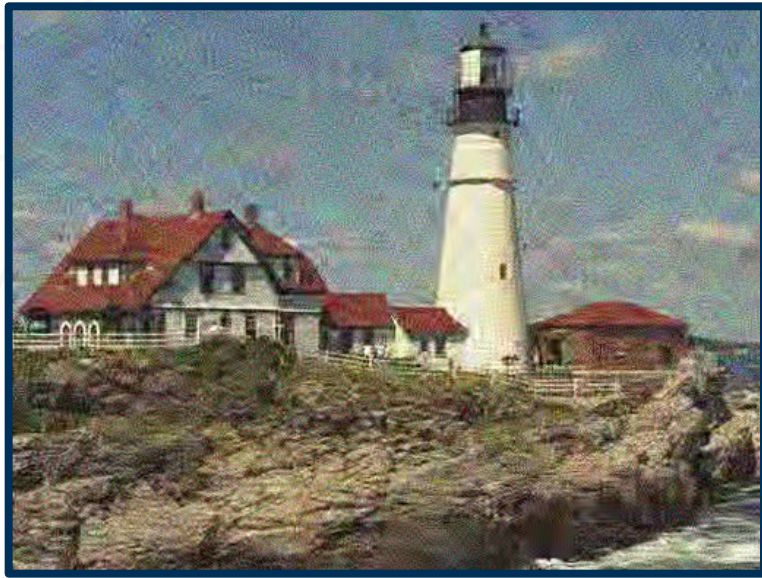
This requires **Fine-grained** Analysis on the part of the subjects!

Our Goal: To determine if a trained IQA detector understands the fine-grained nature of contrast in quality

Case Study: Image Quality Assessment

GradCAM in IQA

GradCAM explanation for Why 0.58?



Lighthouse image with level 5 lossy compression from TID 2013 dataset

The given image is somewhat OK quality

DIQaM :
0.58

Grad-CAM

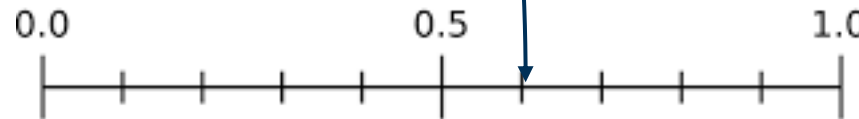
Why 0.58?



Bad
Quality

Good
Quality

Add heatmap
Explain blue
Yellow, red, green



Case Study: Image Quality Assessment

GradCAM in IQA

GradCAM explanation may not be useful for fine-grained analysis

Grad-CAM explanation tells us that the quality score was decided based on all parts of the image and specifically based on the base of the lighthouse

Lighthouse image with level 5 lossy compression from TID 2013 dataset

Bad Quality

Good Quality

0.0

0.5

1.0

Grad-CAM

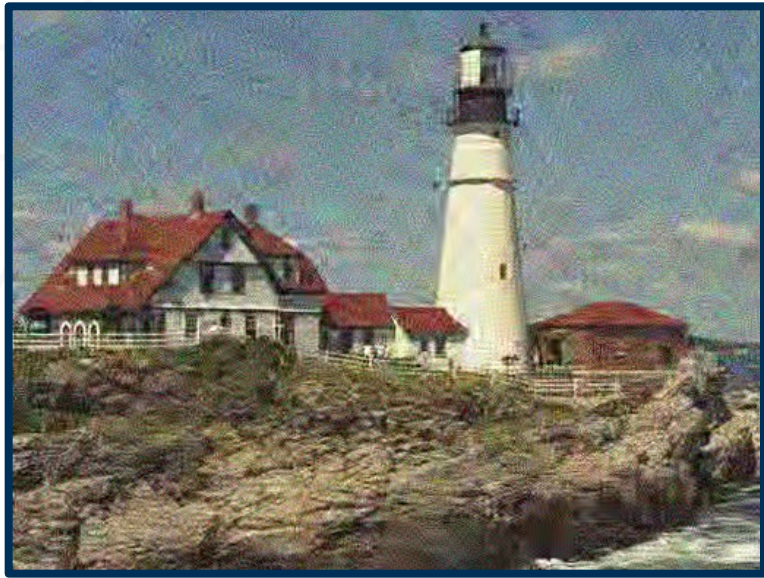
Why 0.58?



Case Study: Image Quality Assessment

ContrastCAM in IQA

All the distortions in the foreground prevent a quality score of 1



Lighthouse image with level 5 lossy compression from TID 2013 dataset



DIQaM :
0.58

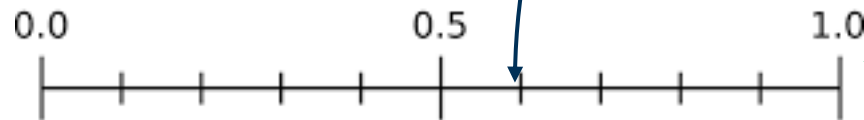
Contrastive explanation



Why 0.58,
rather than 1?



Bad
Quality

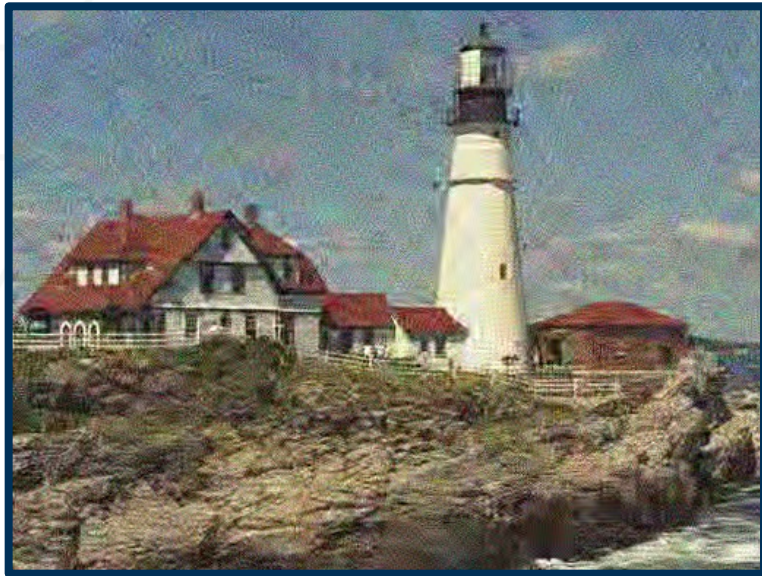


Good
Quality

Case Study: Image Quality Assessment

ContrastCAM in IQA

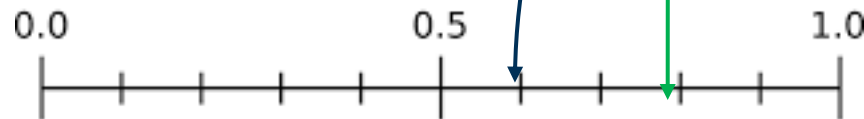
The distortions on the lighthouse and houses prevent a higher score of 0.75



Lighthouse image with level 5 lossy compression from TID 2013 dataset



Bad Quality

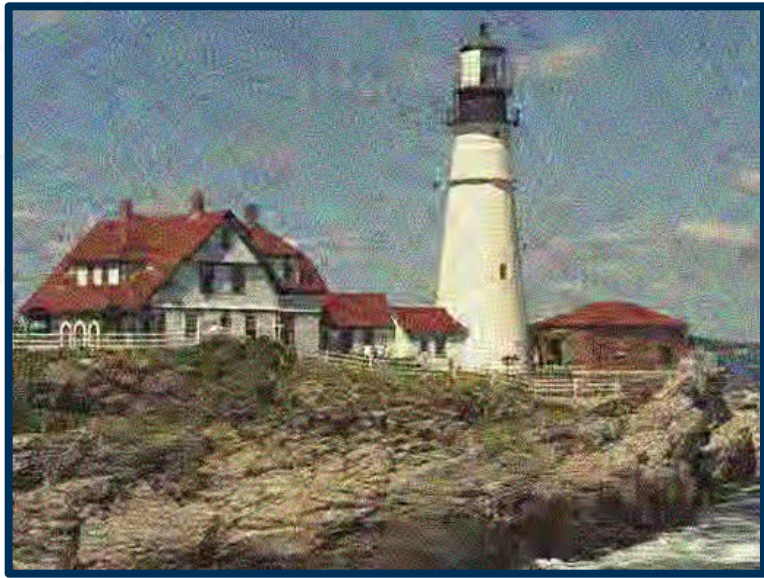


Good Quality

Case Study: Image Quality Assessment

ContrastCAM in IQA

The quality of the lighthouse and sky is better than a score of 0.5



Lighthouse image with level 5 lossy compression from TID 2013 dataset



DIQaM :
0.58

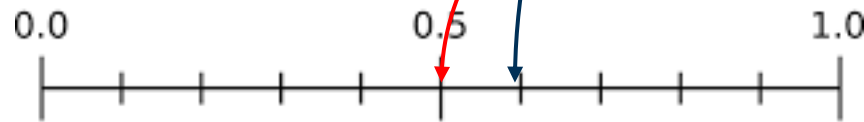
Contrastive explanation



Why 0.58,
rather than 0.5?



Bad
Quality

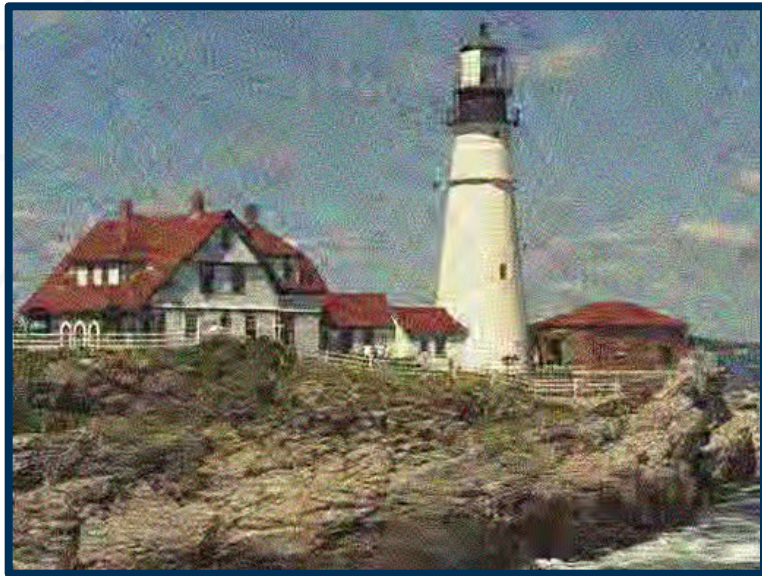


Good
Quality

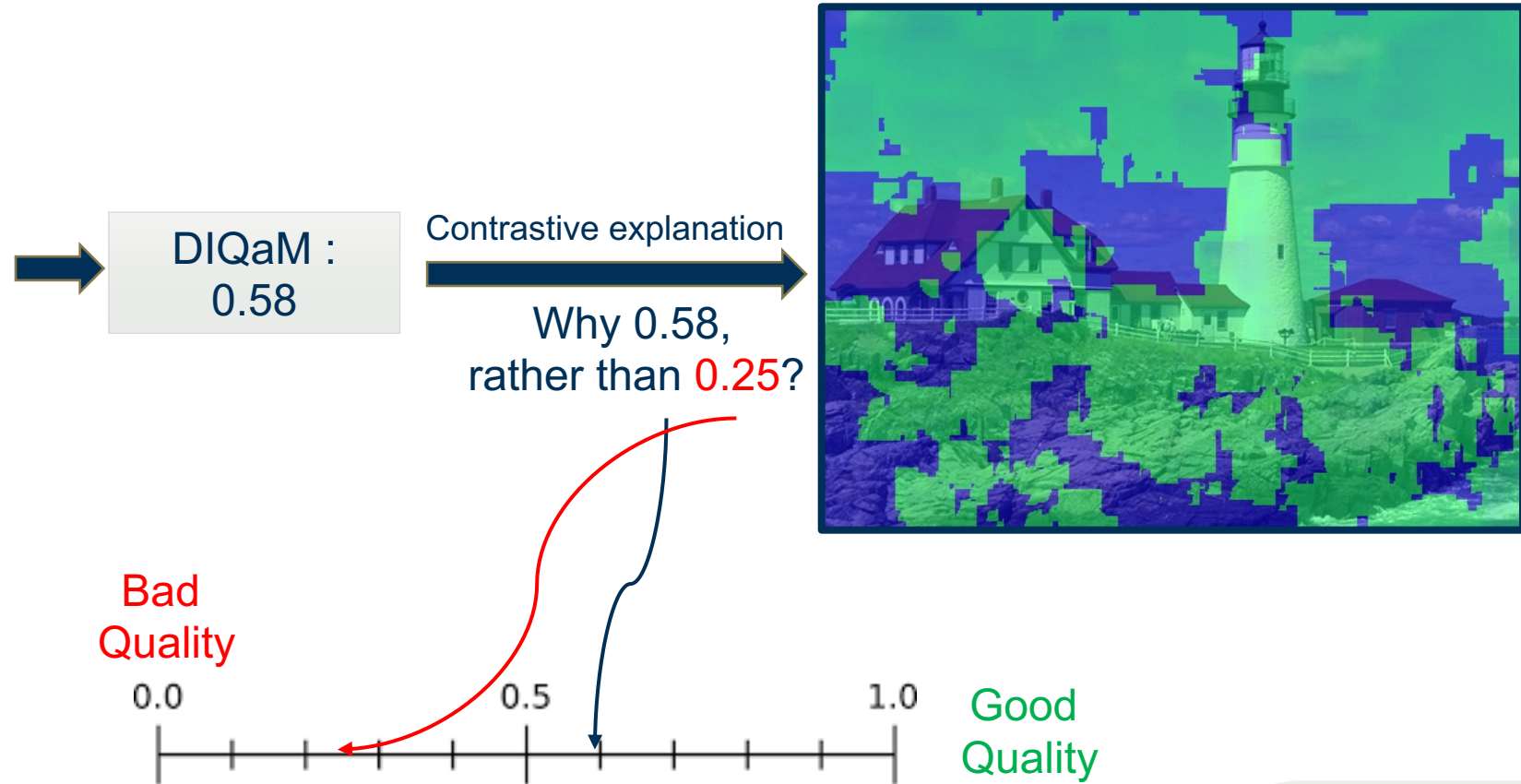
Case Study: Image Quality Assessment

ContrastCAM in IQA

The sky, lighthouse, and cliff merit a quality higher than 0.25









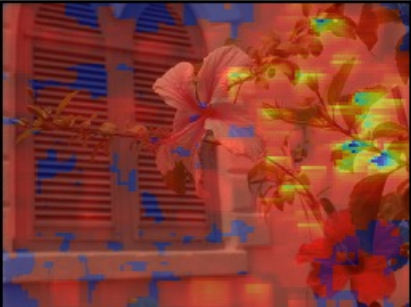




Lighthouse image with level 5 lossy compression from TID 2013 dataset



Case Study: Image Quality Assessment

ContrastCAM in IQA

Contrastive IQA elicits the fine-grained decisions made by the network

					
Distorted Image - IQA Score 0.58	Grad-CAM : Why 0.58?	Why 0.58, rather than 1?	Why 0.58, rather than 0.75?	Why 0.58, rather than 0.5	Why 0.58, rather than 0.25
					
Distorted Image - IQA Score 0.48	Grad-CAM : Why 0.48?	Why 0.48, rather than 1?	Why 0.48, rather than 0.75?	Why 0.48, rather than 0.5	Why 0.48, rather than 0.25

Takeaways

Takeaways from Lecture 4

- There are **no “one size fits all” explanations** and techniques
- **Targeted explanations requires knowledge of data**
 - They are only accessible to most
- **GradCAM uses gradients and activations to highlight class-specific features**
 - **Logits are backpropagated** and the resulting gradients are used as feature weights
 - It's a single forward and backward pass explanatory method
- **CounterfactualCAM backpropagates the negative of the logit** and the resulting gradients are used as feature weights
- **ContrastCAM backpropagates a loss between predicted class and some contrast class** and the resulting gradients are used as feature weights
- Image Quality Assessment benefits from fine-grained analysis of explanations

References

Lecture 4: Visual Explanations II

- AIRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.
- 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.
- Prabhushankar, M., Kwon, G., Temel, D., & AIRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.
- Bosse S, Maniry D, Müller K R, et al. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 2018, 27(1): 206-219.
- Ponomarenko, Nikolay, et al. "Image database TID2013: Peculiarities, results and perspectives." *Signal processing: Image communication* 30 (2015): 57-77