

Robust Neural Networks

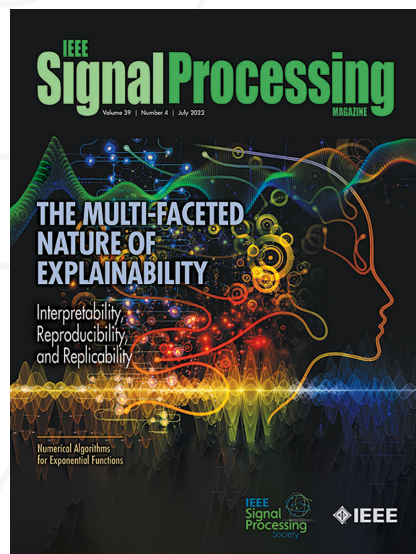
Part 2: Explainability at Inference

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**
 - Visual Explanations
 - Gradient-based Explanations
 - GradCAM
 - CounterfactualCAM
 - ContrastCAM
- Part 3: Uncertainty at Inference
- Part 4: Intervenability at Inference
- Part 5: Conclusions and Future Directions



Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations



Mohit Prabhushankar, PhD
Postdoc



Ghassan AlRegib, PhD
Professor



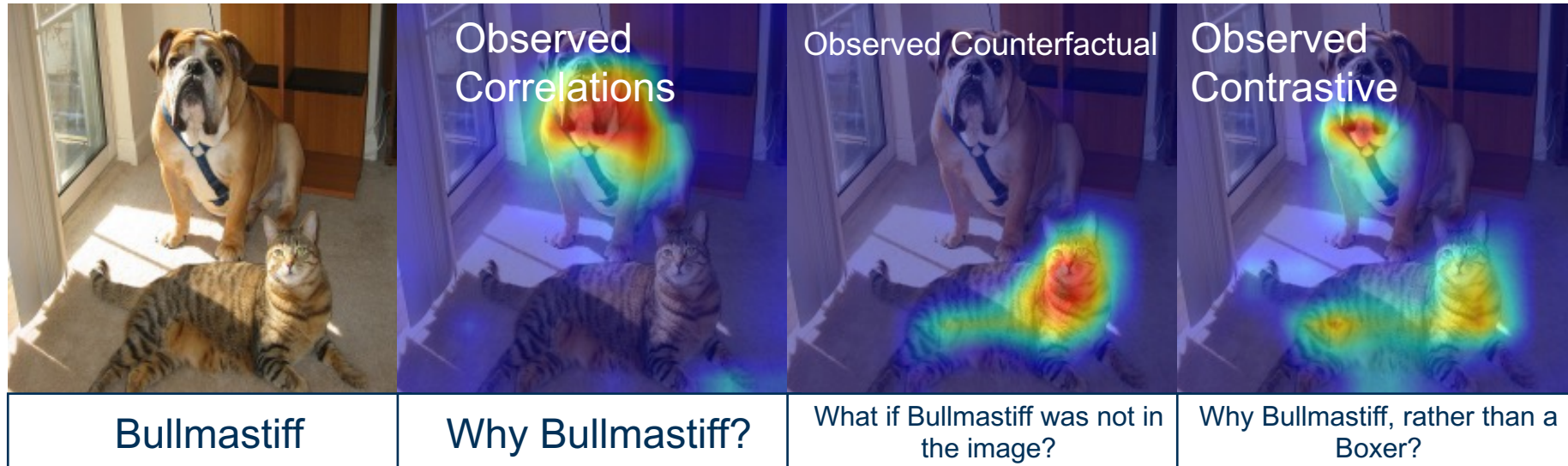
Explanations

Visual Explanations



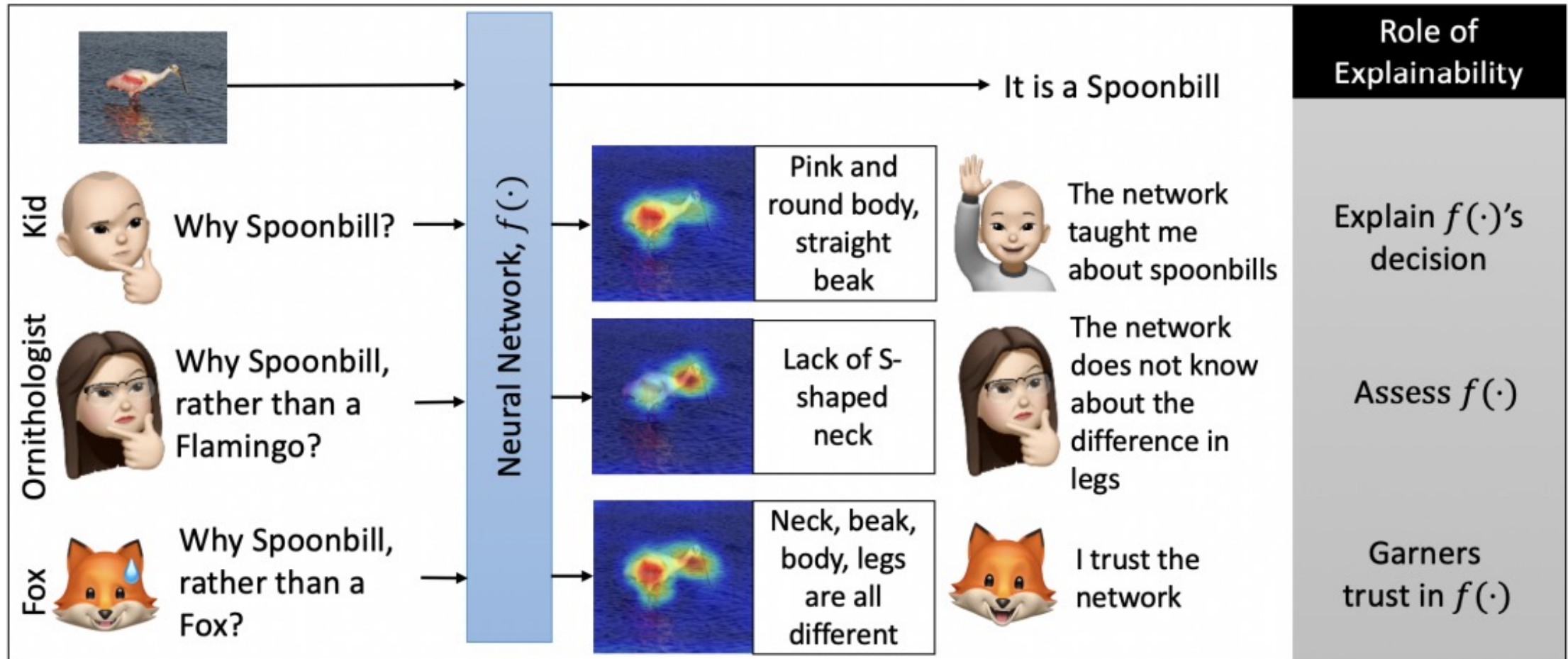
Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

- Explanations are defined as a set of rationales used to understand the reasons behind a decision
- If the decision is based on visual characteristics within the data, the decision-making reasons are visual explanations



Explanations

Role of Explanations – context and relevance



Explanations

Gradient-based Explanations



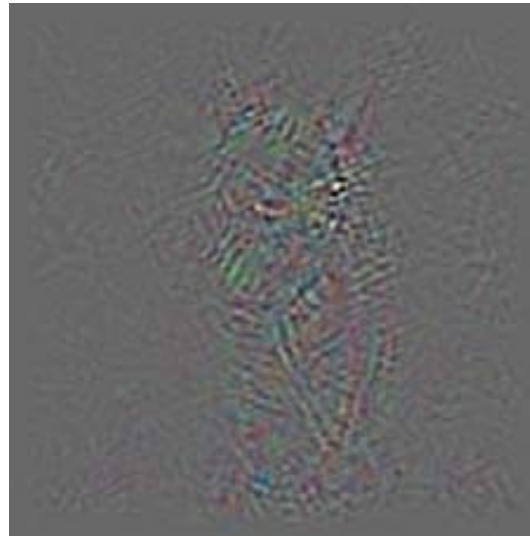
Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

Gradients provide a one-shot means of perturbing the input that changes the output; They provide pixel-level importance scores

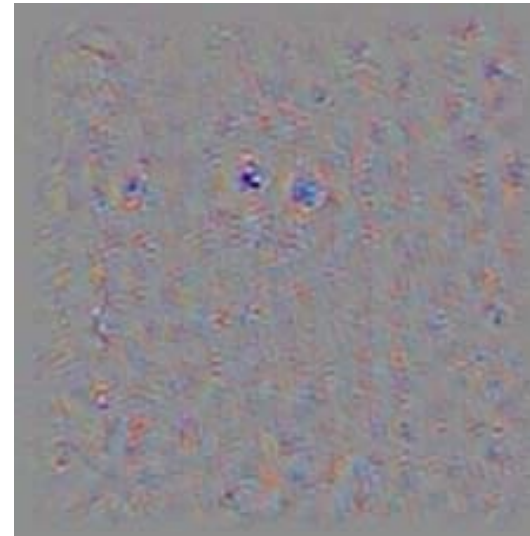
Input



Vanilla Gradients



Deconvolution Gradients



Guided Backpropagation



However, localization remains an issue

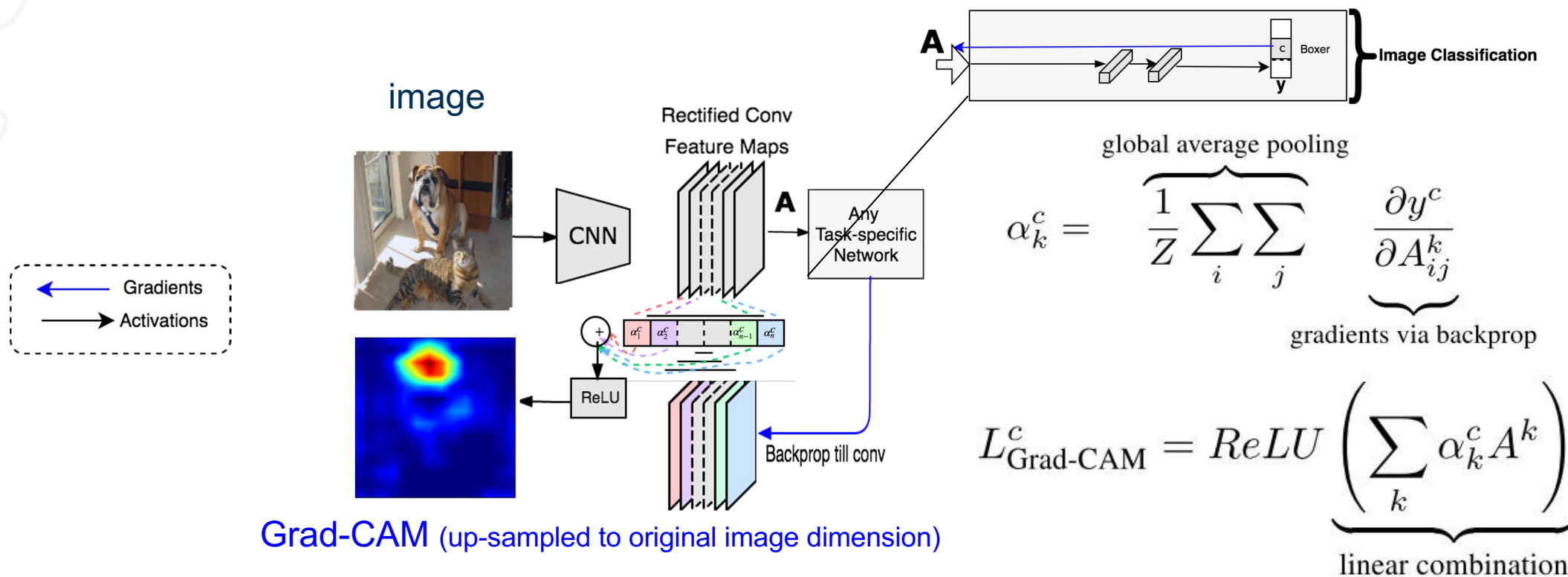
Gradient and Activation-based Explanations

GradCAM



Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

Grad-CAM uses the gradient information flowing into the last convolutional layer of the CNN to assign importance values to each activation for a particular decision of interest.



Gradient and Activation-based Explanations

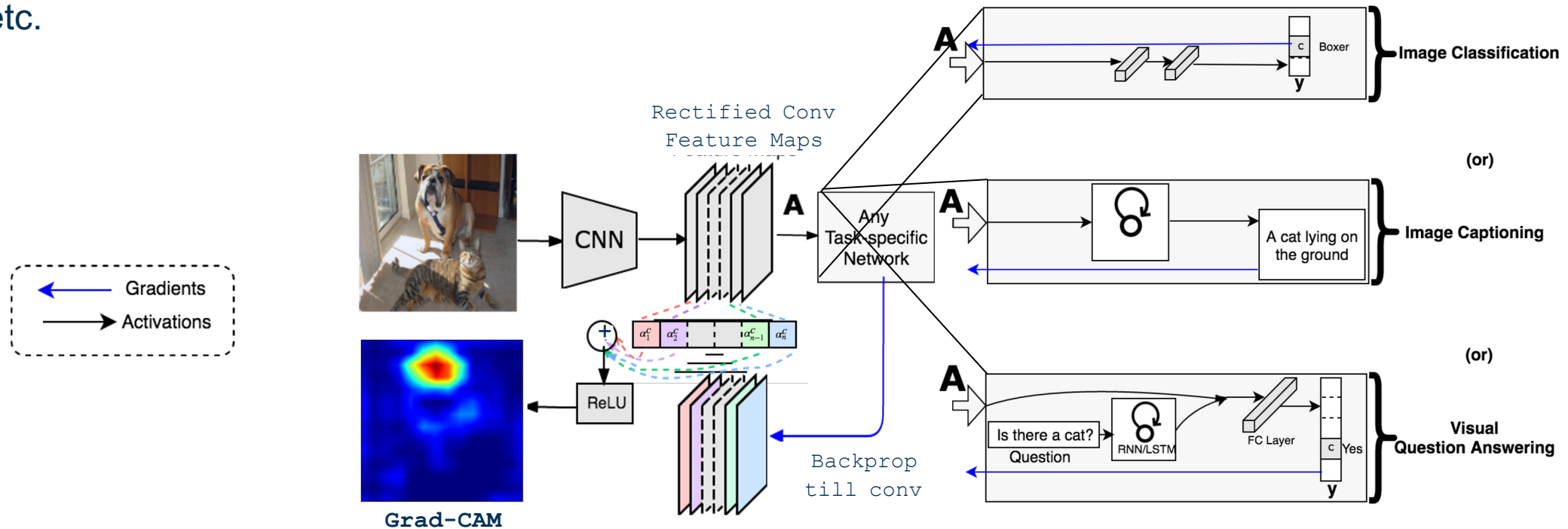
GradCAM

Grad-CAM generalizes to any task:

- Image classification
- Image captioning
- Visual question answering
- etc.



Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations



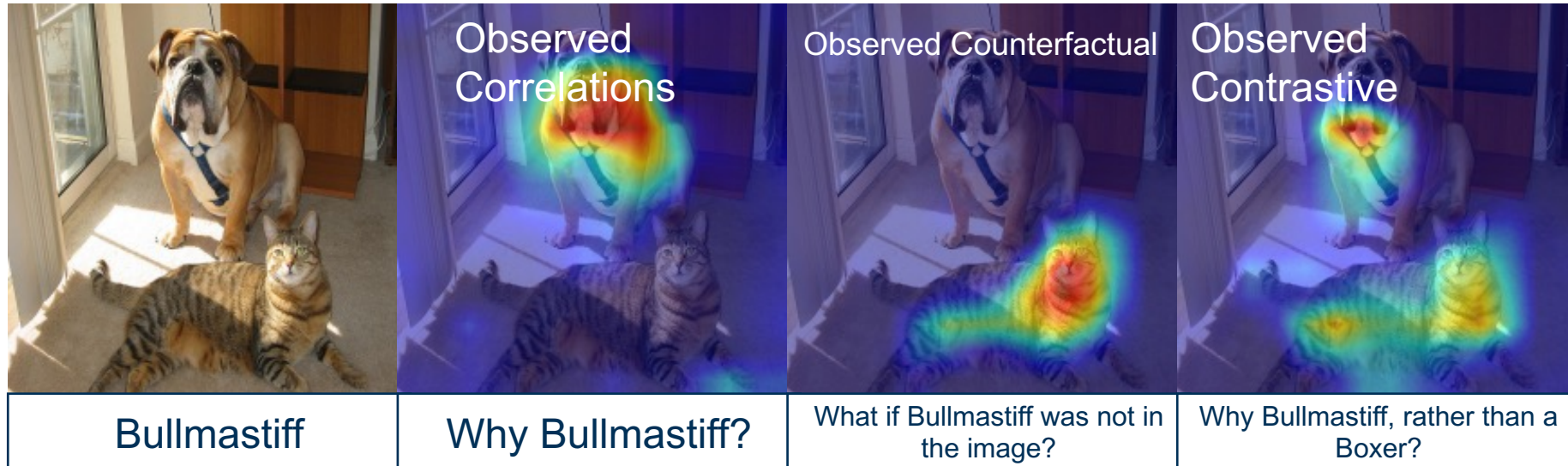
Gradient and Activation-based Explanations

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



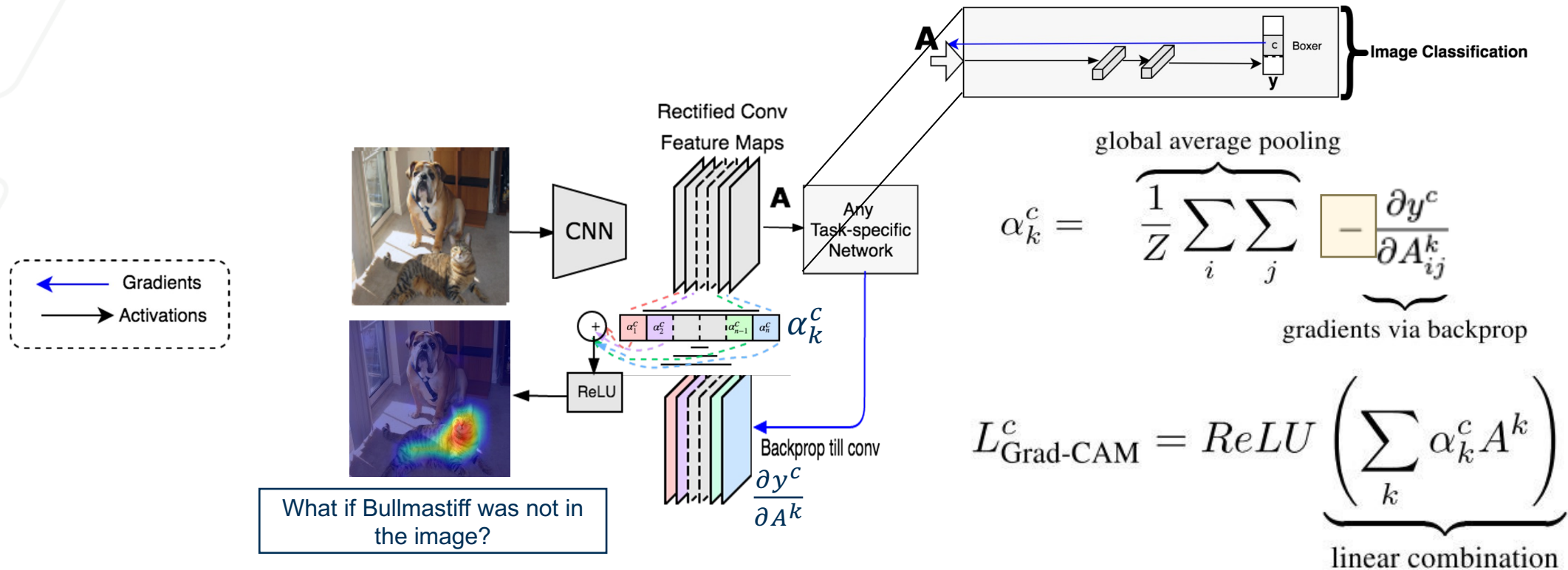
Gradient and Activation-based Explanations

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



Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

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

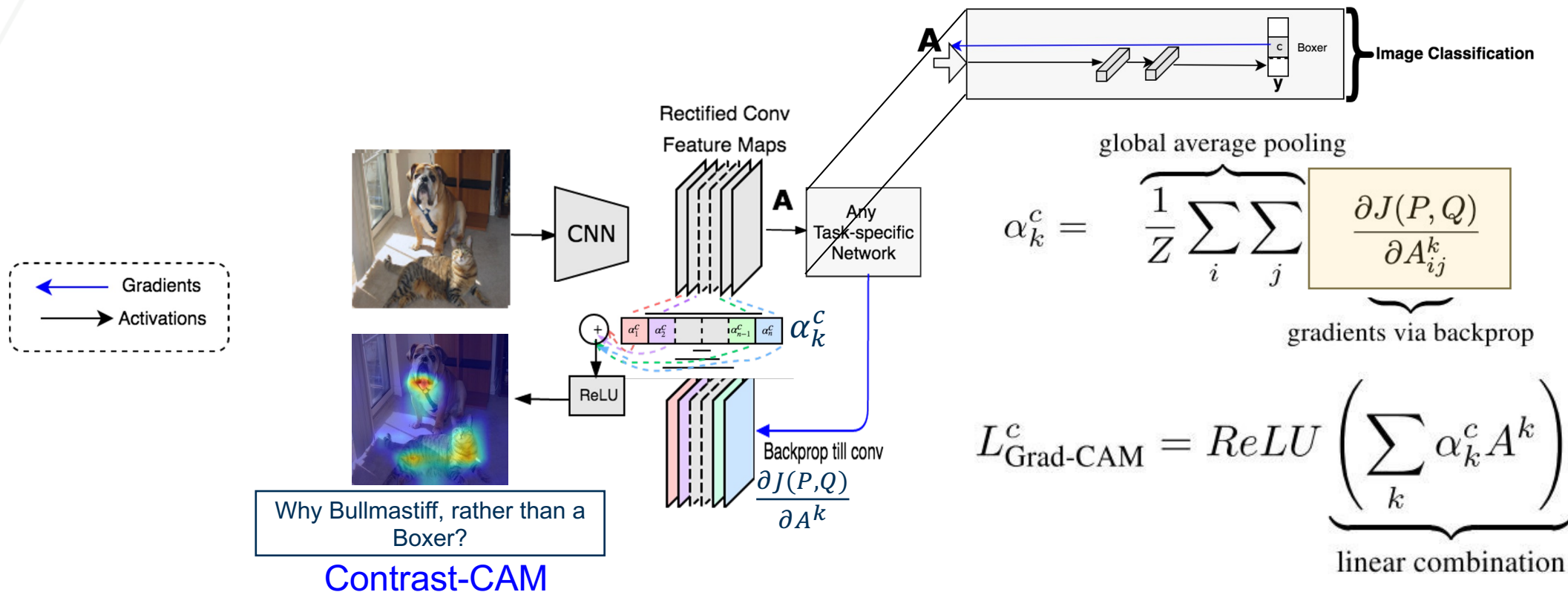
Gradient and Activation-based Explanations

ContrastCAM: Why P, rather than Q?



Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

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



Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

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



Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

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



Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

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



Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

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

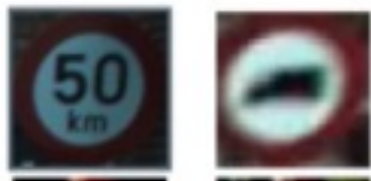


Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

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

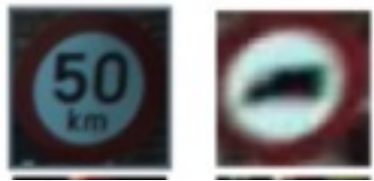


Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

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'

Human
e...all
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?	