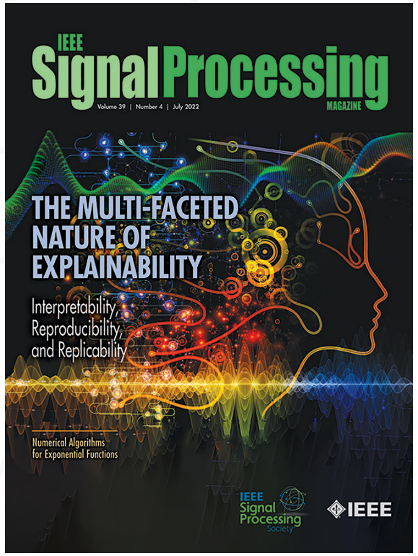
Robust Neural Networks
Part 2: Explainability at Inference

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

SCAN ME

Explanations

Visual Explanations

SCAN ME

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**

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

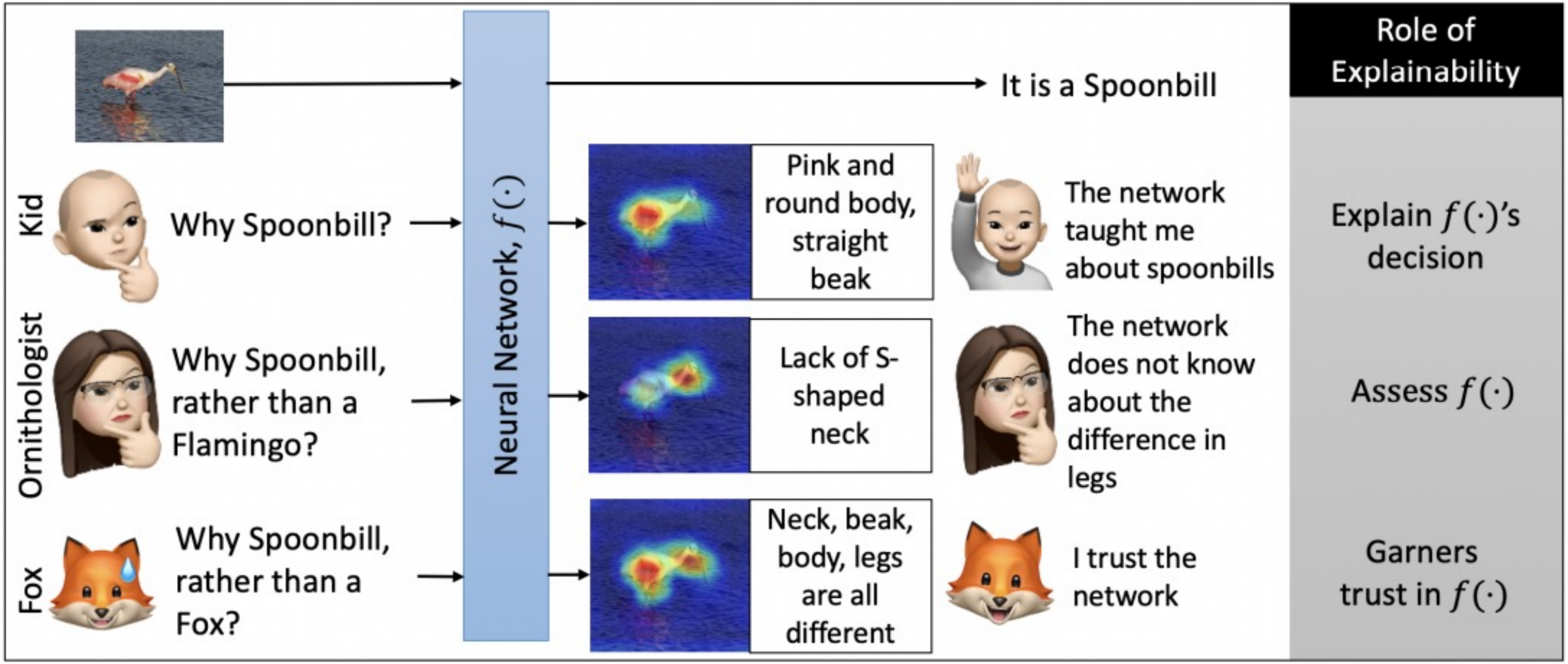
AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks: Towards relevant and contextual explanations. *IEEE Signal Processing Magazine*, *39*(4), 59-72.

Explanations

Role of Explanations – context and relevance

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME



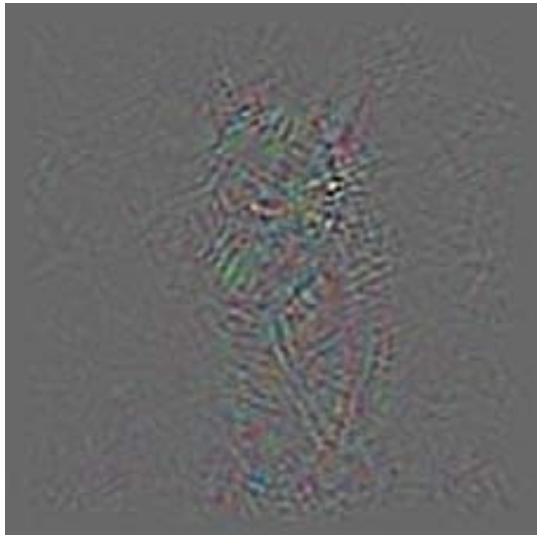
AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks: Towards relevant and contextual explanations. *IEEE Signal Processing Magazine*, *39*(4), 59-72.

SCAN ME

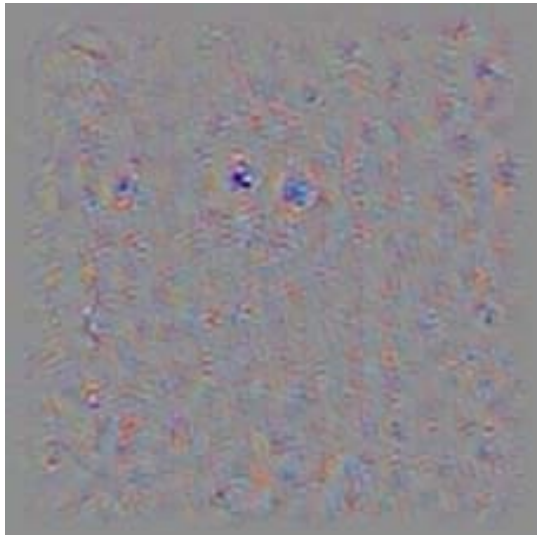
Gradients provide a one-shot means of perturbing the input that changes the output; They provide pixel-level importance scores

Input

Vanilla Gradients



Deconvolution Gradients



Guided Backpropagation

However, localization remains an issue

Springenberg, Dosovitskiy, et al., Striving for Simplicity: The all convolutional net, 2015

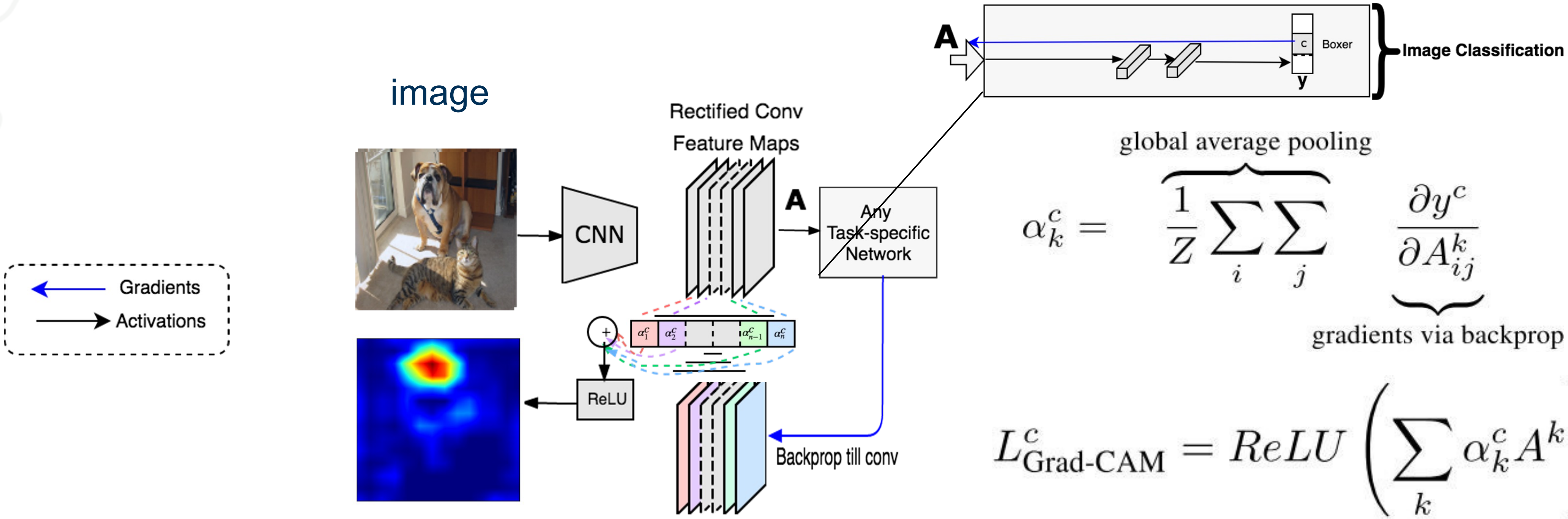
Gradient and Activation-based Explanations

GradCAM

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

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.



Grad-CAM (up-sampled to original image dimension)

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

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

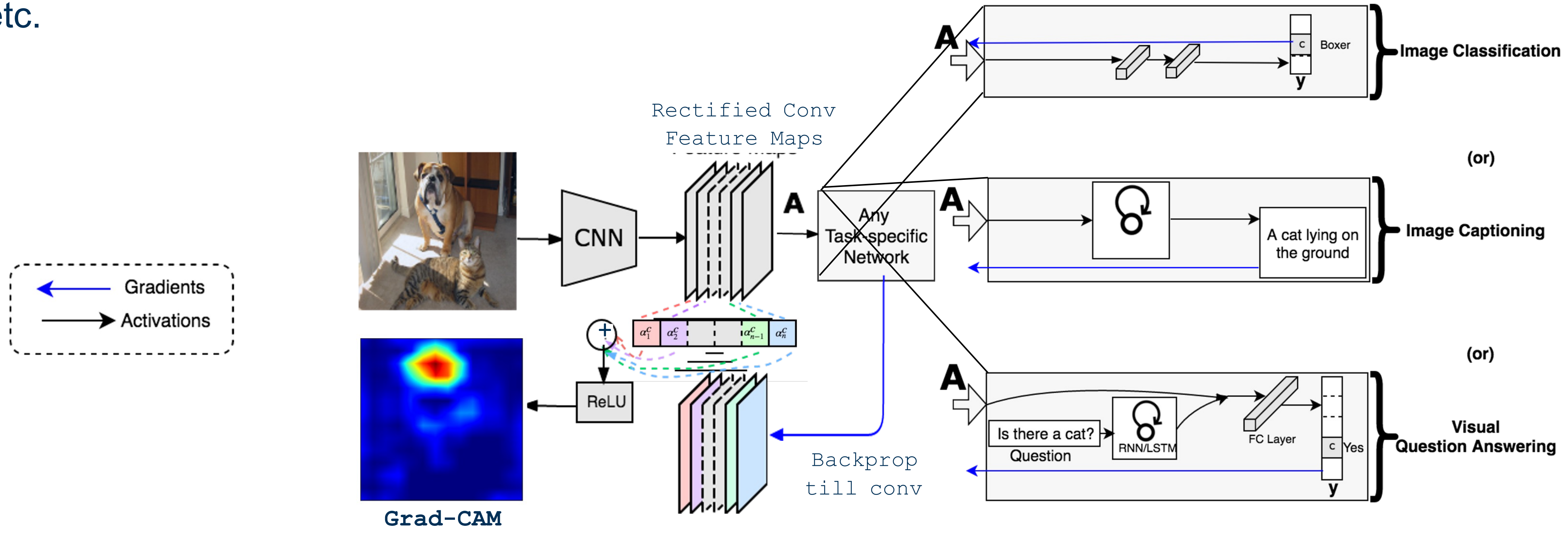
Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

Gradient and Activation-based Explanations

GradCAM

Grad-CAM generalizes to any task:

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



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.

Gradient and Activation-based Explanations

Explanatory Paradigms

SCAN ME

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

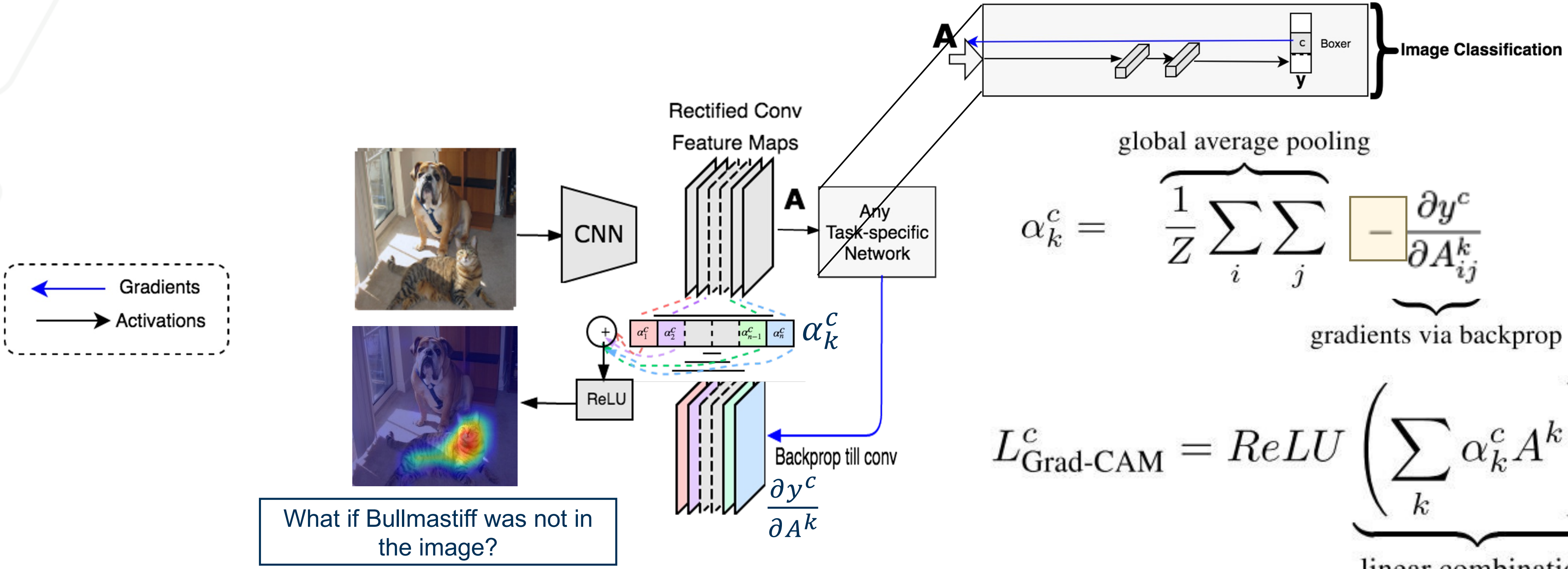
Bullmastiff	Why Bullmastiff?	What if Bullmastiff was not in the image?	Why Bullmastiff, rather than a Boxer?

AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks: Towards relevant and contextual explanations. *IEEE Signal Processing Magazine*, *39*(4), 59-72.

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



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

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

Negating the gradients effectively removes these regions from analysis

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.

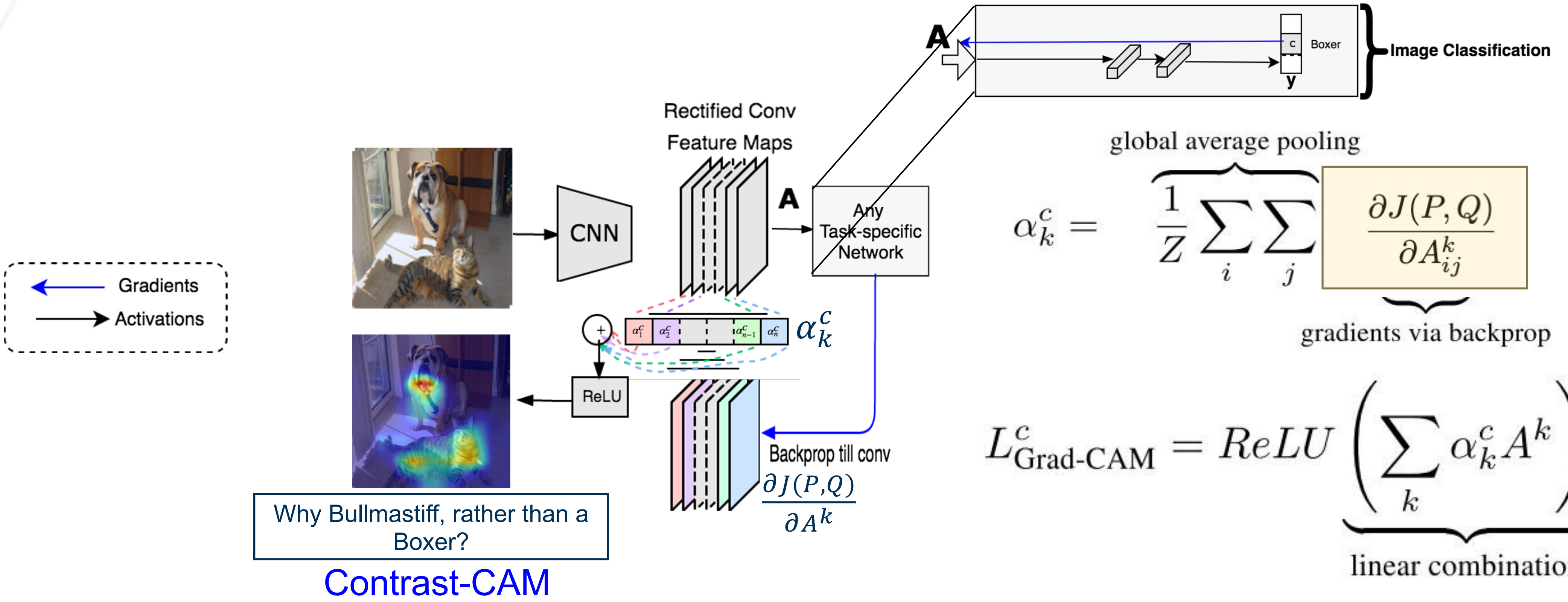
Gradient and Activation-based Explanations

ContrastCAM: Why P, rather than Q?

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

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



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

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

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

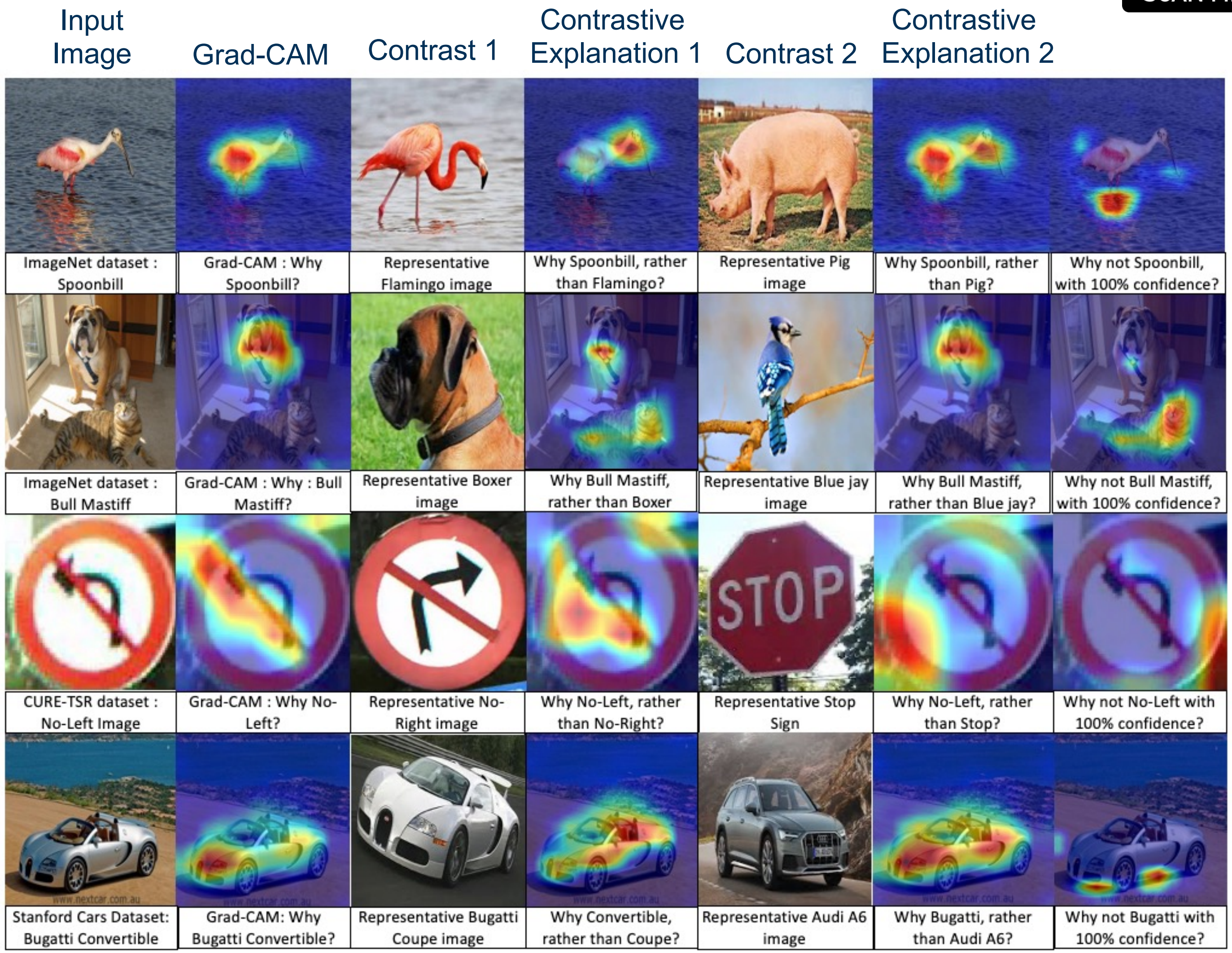
Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

SCAN ME

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

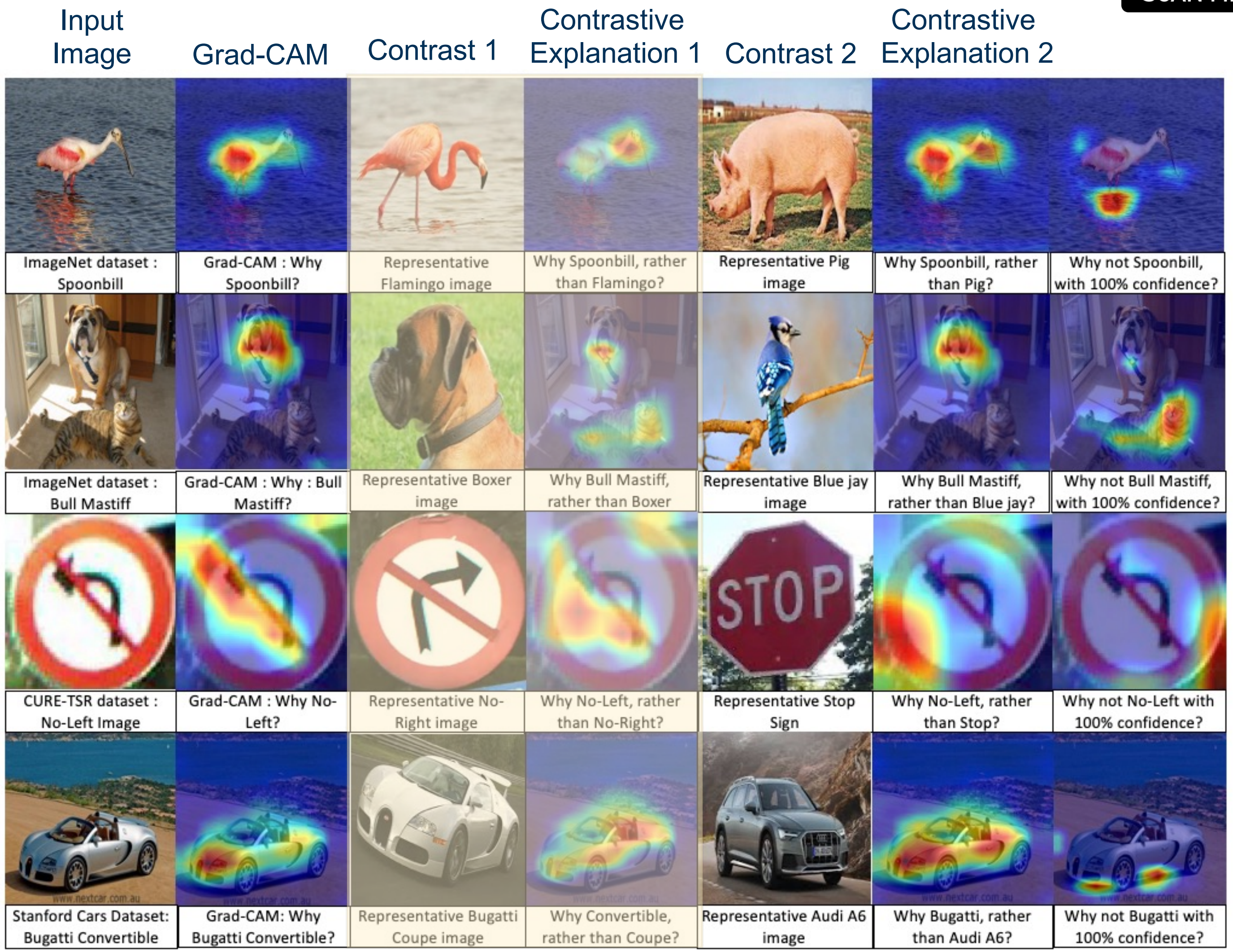


Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations



Human Interpretable

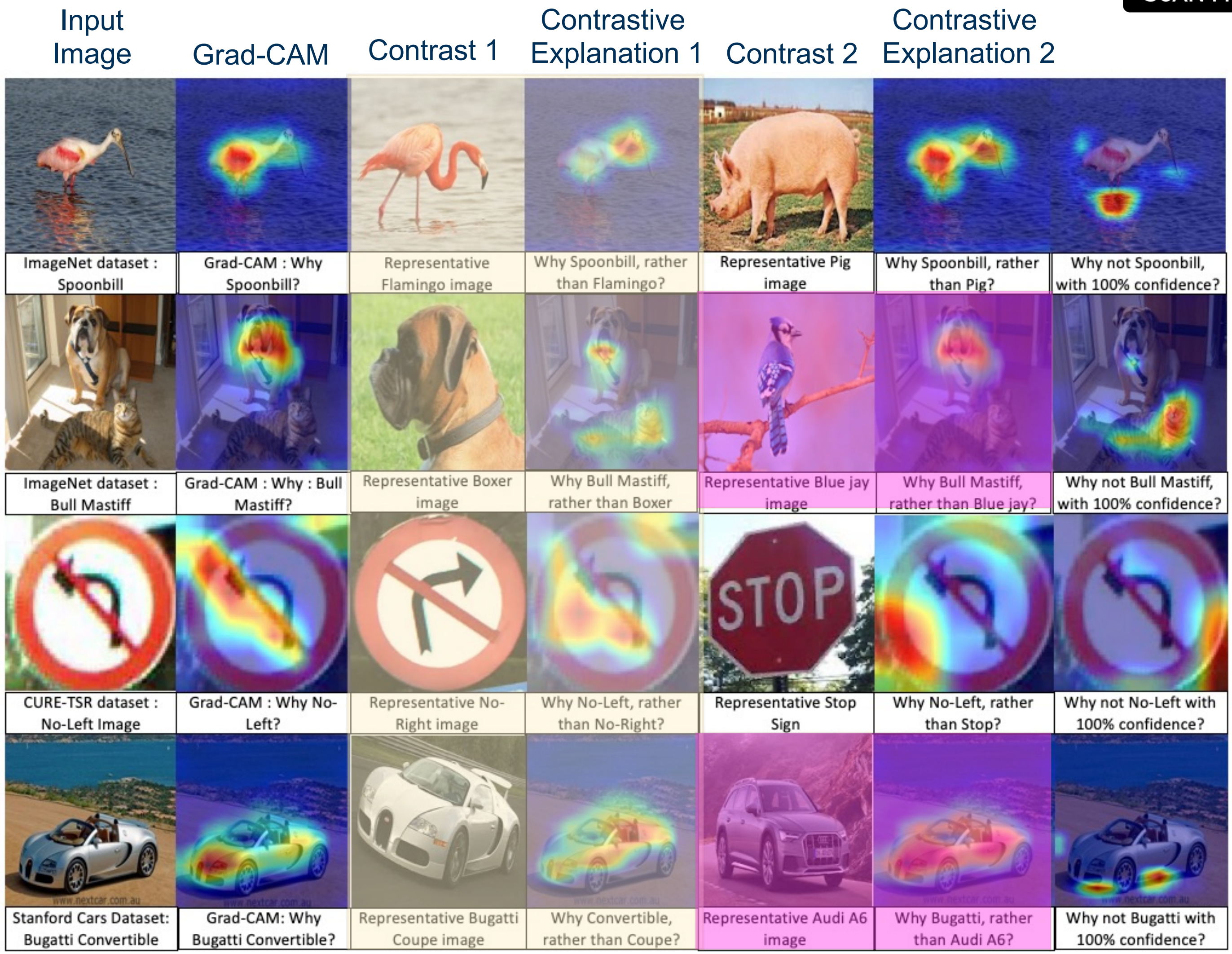
Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

SCAN ME

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations



Human Interpretable

Same as Grad-CAM

JAN 4-8 WACV 2024 WAIKOLOA HAWAII

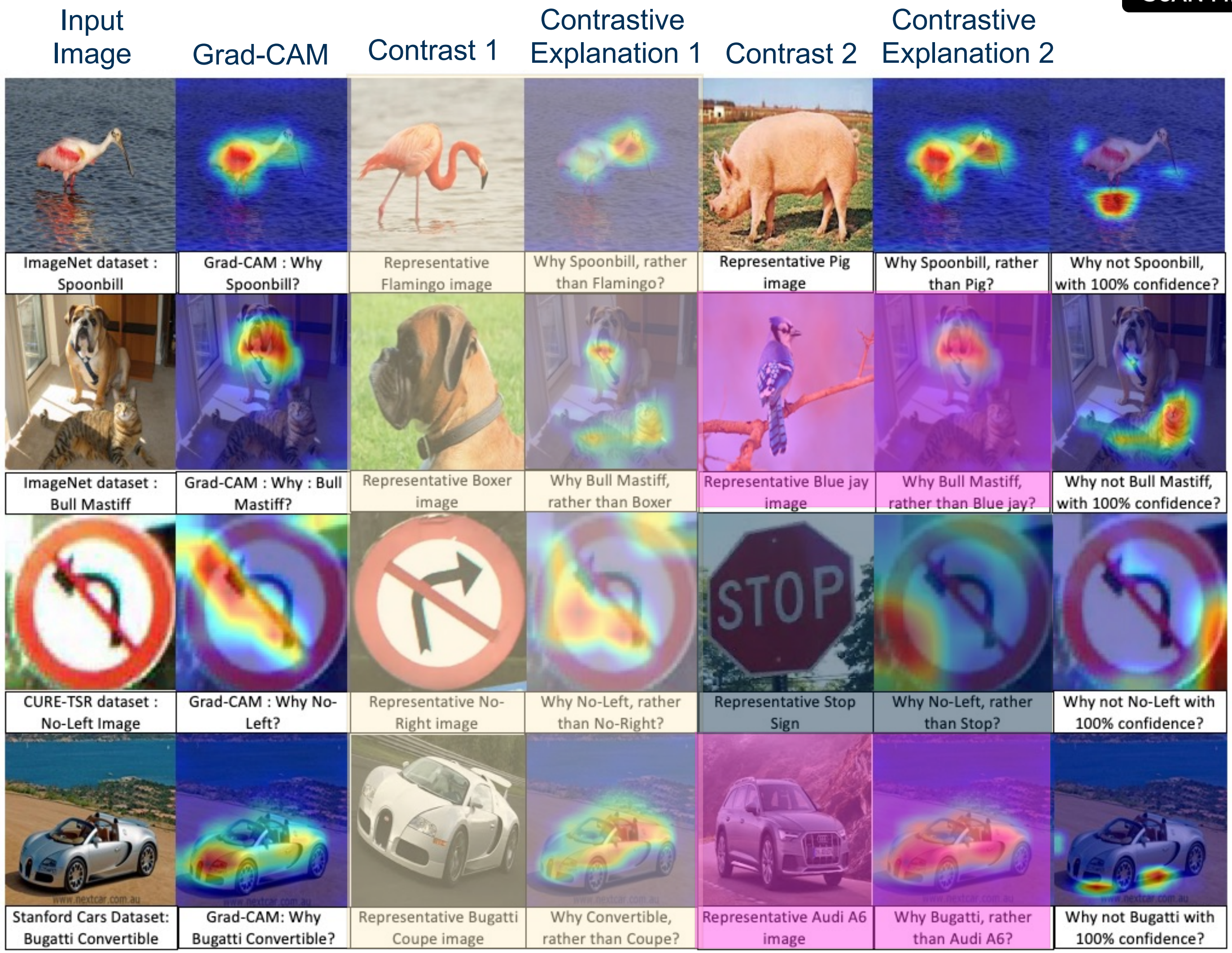
Prabhushankar, M., Kwon, G., Temel, D., & AlRegib, G. (2020, October). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)* (pp. 3289-3293). IEEE.

OLIVES @GeorgiaTech

Georgia Tech

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM



Human Interpretable

Same as Grad-CAM

Not Human Interpretable

SCAN ME

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

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.

Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

SCAN ME

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.

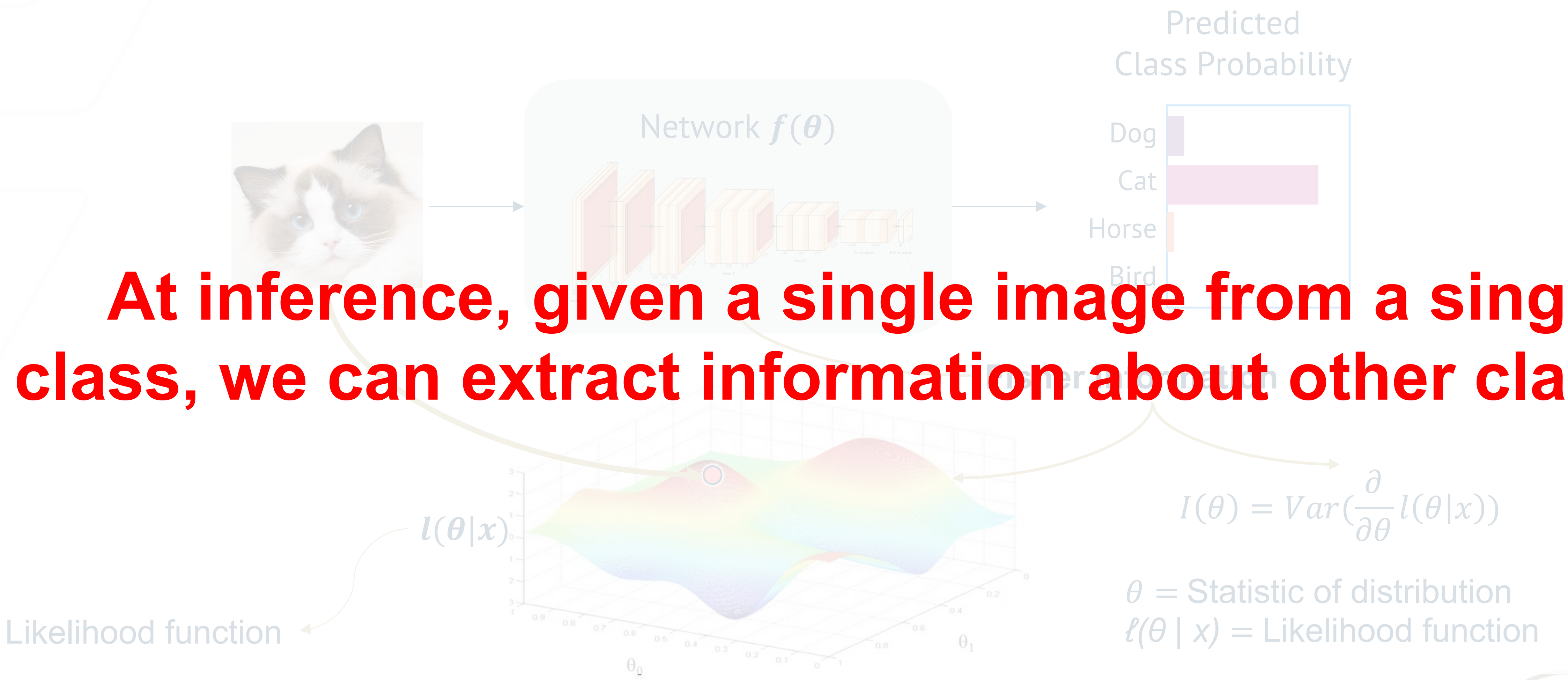
Gradient and Activation-based Explanations

Results from GradCAM, CounterfactualCAM, and ContrastCAM

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.

A Callback…

Information at Inference

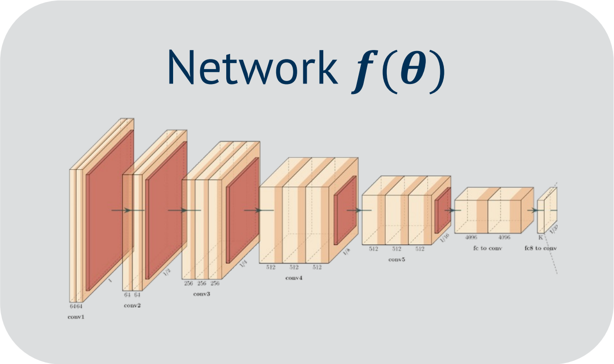


At inference, given a single image from a single class, we can extract information about other classes

Information at Inference

Case Study: Explainability

$\mathcal{T}$ is the set of all features learned by a trained network



Beak
Neck
Legs
Feathers
Water
Grass
Teeth
.
.

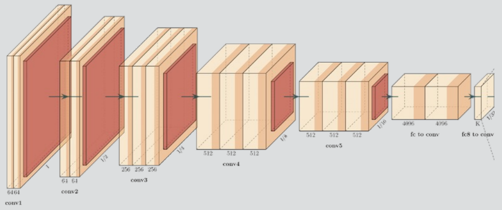
Features $\mathcal{T}$

Information at Inference

Case Study: Explainability

Given only an image of a spoonbill, we can extract information about a Flamingo

Network $\boldsymbol{f(\theta)}$



Beak
Neck
Legs
Feathers
Water
Grass
Teeth
.
.

Features $\mathcal{T}$

All the requisite Information is stored within $\boldsymbol{f(\theta)}$

Goal: To extract and quantify this information at inference