# **Robust Neural Networks Part 2: Explainability at Inference**







# **Objective** Objective of the Tutorial

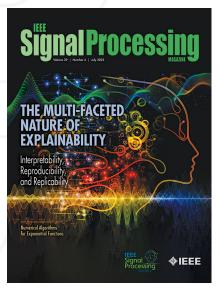
### To discuss methodologies that promote robustness in neural networks at inference

- 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







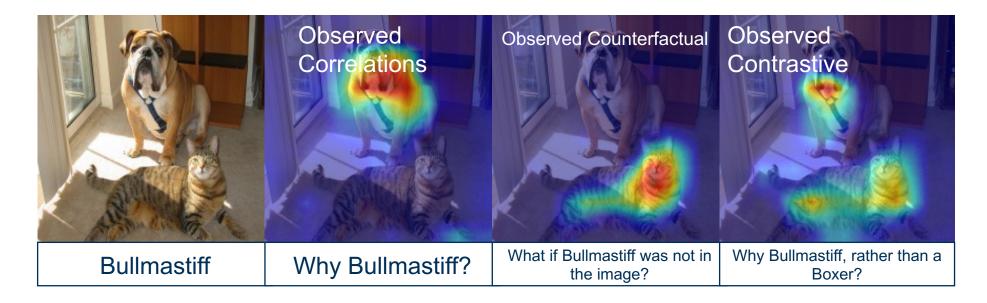


### **Explanations** Visual Explanations



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





[Tutorial@AAAI'24] | [Ghassan AlRegib and Mohit Prabhushankar] | [Feb 21, 2024]

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



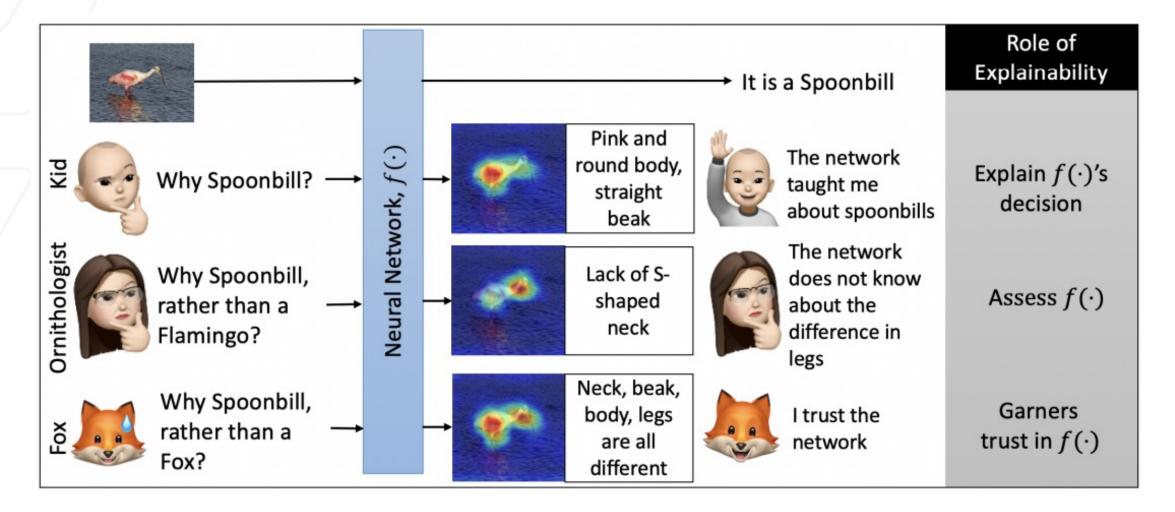


### **Explanations**

Role of Explanations – context and relevance



Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations



37 of 172

[Tutorial@AAAI'24] | [Ghassan AlRegib and Mohit Prabhushankar] | [Feb 21, 2024]

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





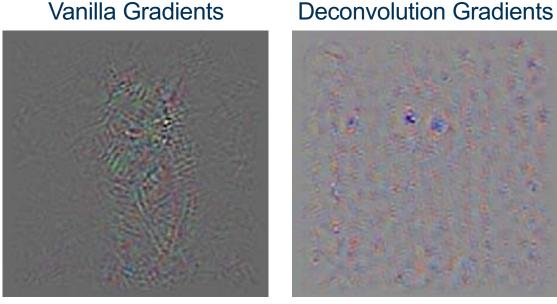
### **Explanations** Gradient-based Explanations

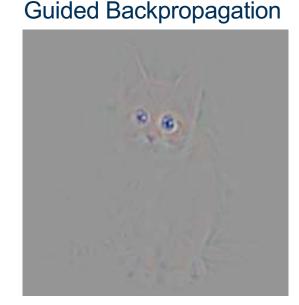


Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

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







### However, localization remains an issue





[Tutorial@AAAI'24] | [Ghassan AlRegib and Mohit Prabhushankar] | [Feb 21, 2024]

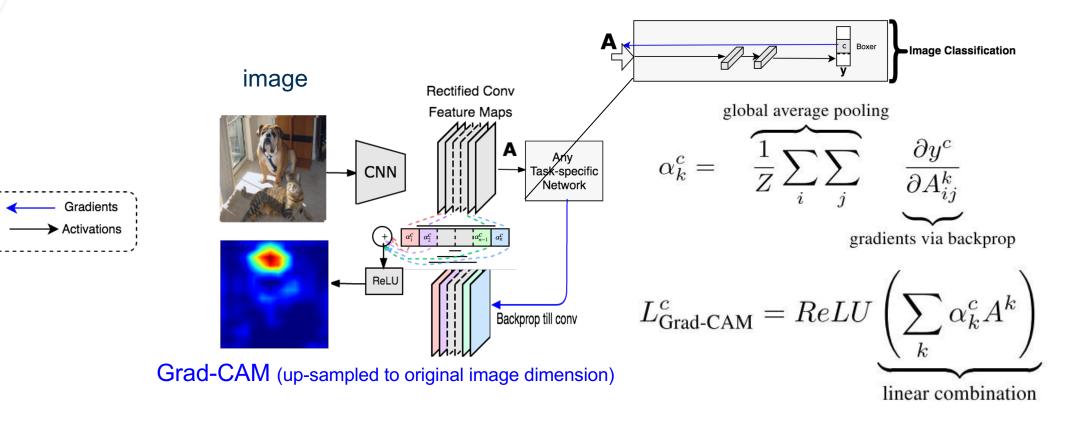


Springenberg, Dosovitskiy, et al., Striving for Simplicity: The all convolutional net, 2015



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.





[Tutorial@AAAI'24] | [Ghassan AlRegib and Mohit Prabhushankar] | [Feb 21, 2024]

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



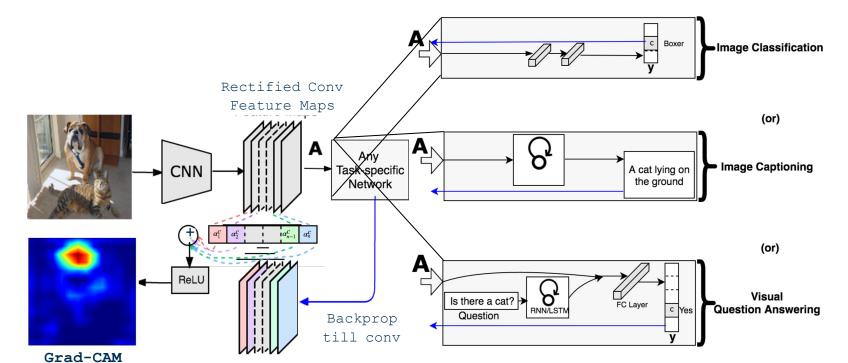
### Grad-CAM generalizes to any task:

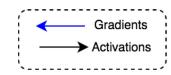
- Image classification
- Image captioning
- Visual question answering



**Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations** 

SCAN ME







• etc.



[Tutorial@AAAI'24] | [Ghassan AlRegib and Mohit Prabhushankar] | [Feb 21, 2024]

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

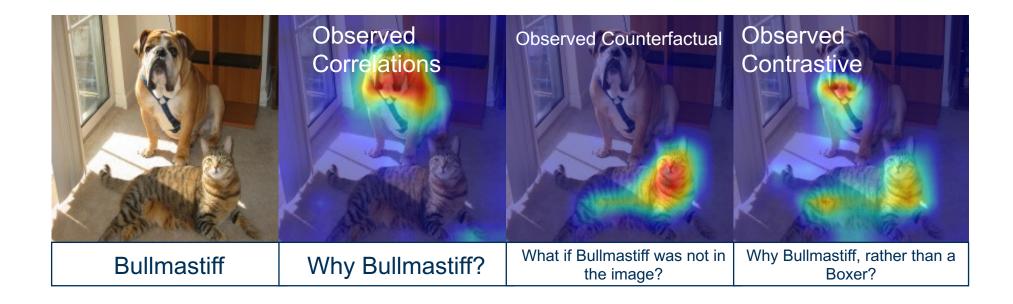


**Explanatory Paradigms** 



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





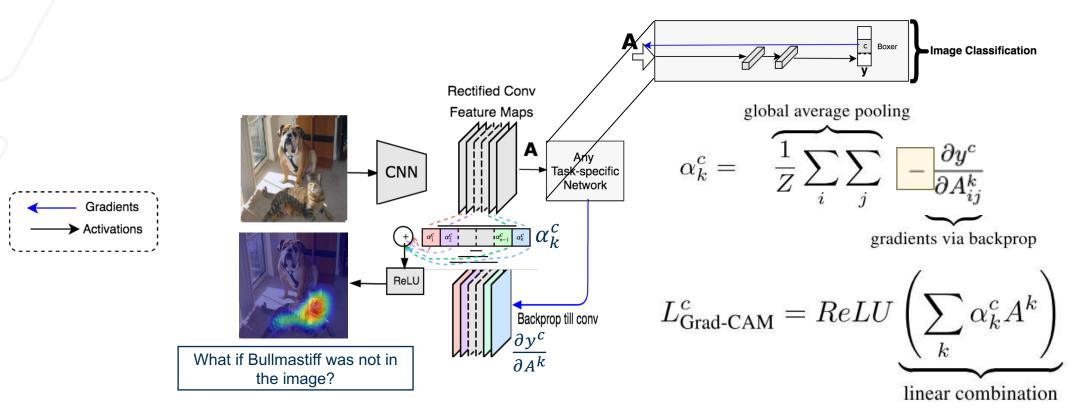
[Tutorial@AAAI'24] | [Ghassan AlRegib and Mohit Prabhushankar] | [Feb 21, 2024]

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



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

### In GradCAM, global average pool the negative of gradients to obtain $\alpha^c$ for each kernel k



### Negating the gradients effectively removes these regions from analysis



[Tutorial@AAAI'24] | [Ghassan AlRegib and Mohit Prabhushankar] | [Feb 21, 2024]



Georgia Tech

Explanatory Paradigms in Neural Networks: Towards Relevant and

**Contextual Explanations** 

SCAN ME

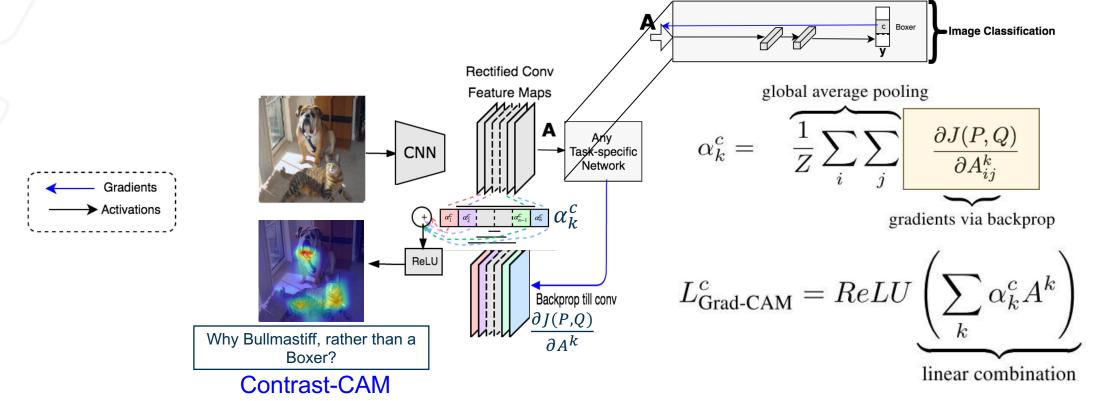
Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradientbased localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

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.



[Tutorial@AAAI'24] | [Ghassan AlRegib and Mohit Prabhushankar] | [Feb 21, 2024]

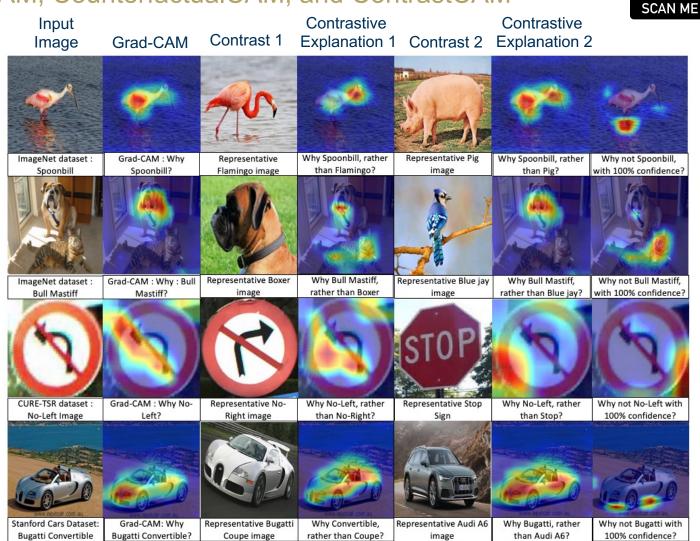




### Results from GradCAM, CounterfactualCAM, and ContrastCAM



Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations



47 of 172

[Tutorial@AAAI'24] | [Ghassan AlRegib and Mohit Prabhushankar] | [Feb 21, 2024]



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

### **Gradient and Activation-based Explanations**

### Results from GradCAM, CounterfactualCAM, and ContrastCAM



Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

### Human Interpretable

48 of 172



[Tutorial@AAAI'24] | [Ghassan AlRegib and Mohit Prabhushankar] | [Feb 21, 2024]



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

### **Gradient and Activation-based Explanations**

### Results from GradCAM, CounterfactualCAM, and ContrastCAM



Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations

> Human Interpretable

Same as Grad-CAM



[Tutorial@AAAI'24] | [Ghassan AlRegib and Mohit Prabhushankar] | [Feb 21, 2024]





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

### **Gradient and Activation-based Explanations**

### Results from GradCAM, CounterfactualCAM, and ContrastCAM



**Explanatory Paradigms in Neural Networks: Towards Relevant and Contextual Explanations** 

> Human Interpretable

Same as Grad-CAM

Not Human Interpretable



[Tutorial@AAAI'24] | [Ghassan AIRegib and Mohit Prabhushankar] | [Feb 21, 2024]







51 of 172

[Tutorial@AAAI'24] | [Ghassan AlRegib and Mohit Prabhushankar] | [Feb 21, 2024]







52 of 172

[Tutorial@AAAI'24] | [Ghassan AlRegib and Mohit Prabhushankar] | [Feb 21, 2024]



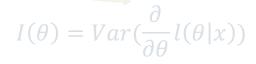


### A Callback... Information at Inference

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

Network  $f(\theta)$ 

Likelihood function



Predicted

 $\theta$  = Statistic of distribution  $\ell(\theta \mid x)$  = Likelihood function





 $l(\theta|x)$ 

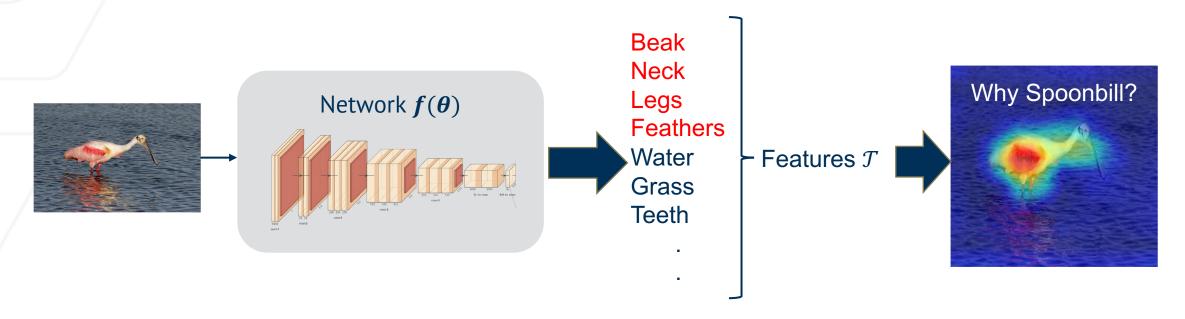
[Tutorial@AAAI'24] | [Ghassan AlRegib and Mohit Prabhushankar] | [Feb 21, 2024]



### **Information at Inference**

Case Study: Explainability

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





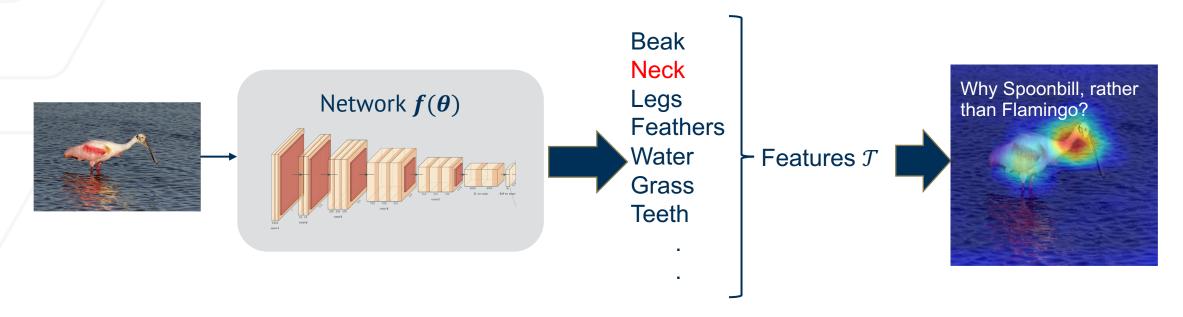




### **Information at Inference**

Case Study: Explainability

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



# All the requisite Information is stored within $f(\theta)$

**Goal: To extract and quantify this information at inference** 



[Tutorial@AAAI'24] | [Ghassan AlRegib and Mohit Prabhushankar] | [Feb 21, 2024]

