AbuDhabi ICIP 2020
IEEE International Conference on Image Processing
25-28 October 2020, United Arab Emirates
IEEE Signal Processing Society

Special Session : Explainable Machine Learning for Image Processing

Contrastive Explanations in Neural Networks




Mohit Prabhushankar (Speaker) | Gukyeong Kwon | Dogancan Temel | Ghassan AlRegib

Georgia Tech
CREATING THE NEXT

Codes

OLIVES
@GeorgiaTech
www.ghassanalregib.com

Website

Explanations

Explanations are a set of rationales used to understand the reasons behind a decision [1]

Question

Name of the bird?

Answer

Spoonbill

Why Spoonbill?

Shallow-water bird with flattened beak and football shaped body. They are pale pink birds with pink shoulders and rump. They have a white neck and a partially feathered, yellow green head.

Language-based explanation

[1] Philip Kitcher and Wesley C Salmon, *Scientific explanation*, vol 13, U of Minnesota Press, 1962

Visual Explanations

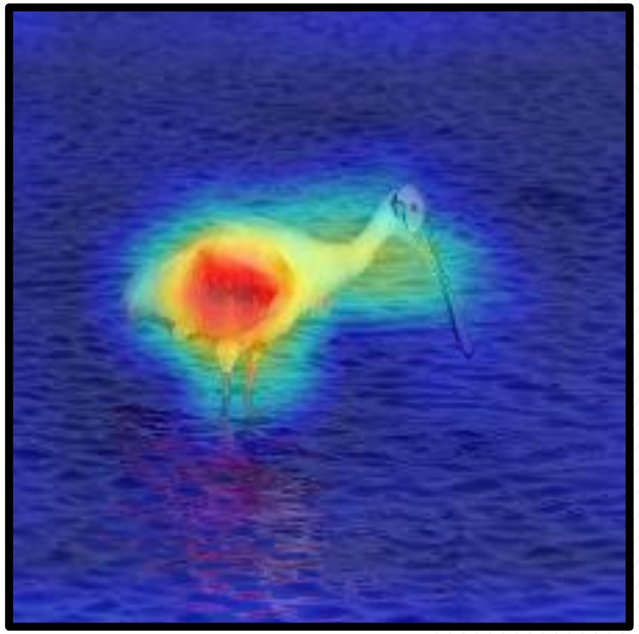
Visual characteristics that are used to justify decisions are termed as visual explanations

Question	Answer
Name of the bird?	Spoonbill

Why Spoonbill?

Shallow-water bird with flattened beak and football shaped body. They are pale pink birds with pink shoulders and rump. They have a white neck and a partially feathered, yellow green head.

Language-based explanation



Visual Explanation

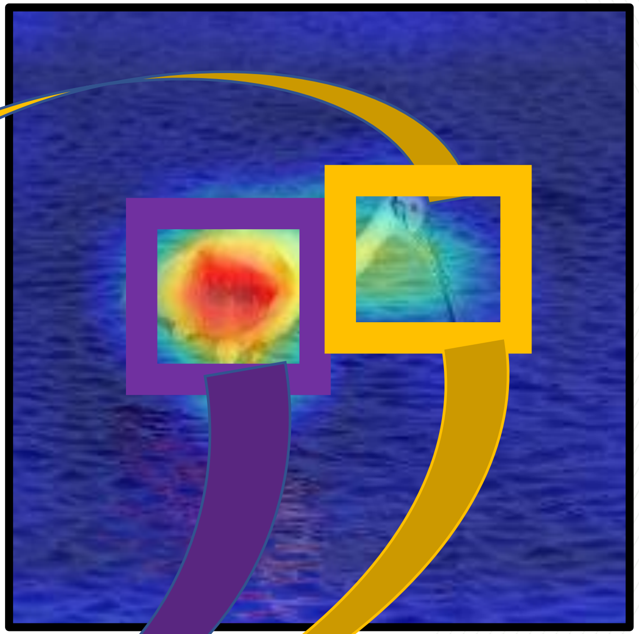
Visual Explanations

Visual characteristics that are used to justify decisions are termed as visual explanations

Question	Answer
Name of the bird?	Spoonbill

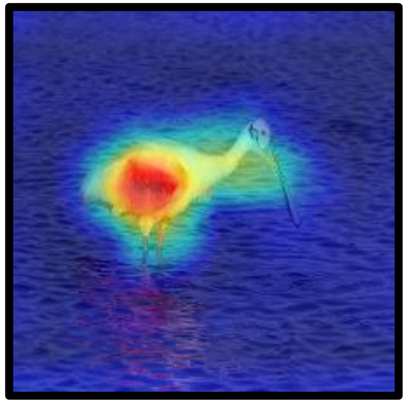
Why Spoonbill?

Shallow-water bird with flattened beak and football shaped body. They are pale pink birds with pink shoulders and rump. They have a white neck and a partially feathered, yellow-green head.



Causal factors based visual explanations – answers to *`Why?'* Questions

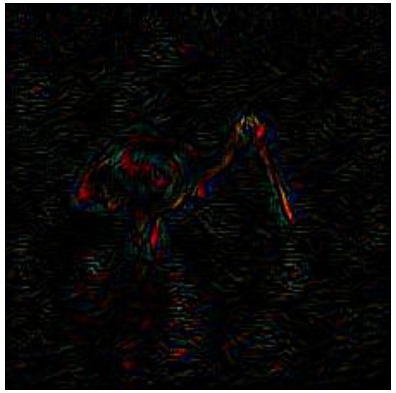
Visual Explanations



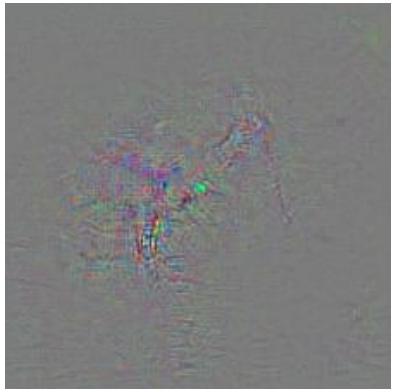
Grad-CAM

`Why P?'

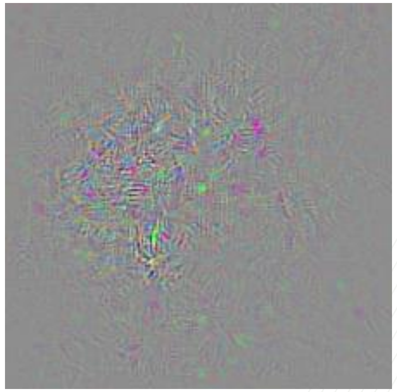
Guided Backpropagation



Positive saliency



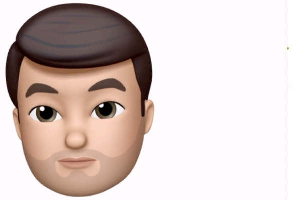
Smooth Gradients



Vanilla Backpropagation

Contrastive Explanations

ICIP 2020
IEEE Internatic
on Image Proc
25-28 October 2020



CONTRASTIVE EXPLANATIONS IN NEURAL NETWORKS

Mohit Prabhushankar, Gukyeong Kwon, Dogancan Temel, and Ghassan AlRegib

OLIVES at the Center for Signal and Information Processing,
School of Electrical and Computer Engineering,
Georgia Institute of Technology, Atlanta, GA, 30332-0250
{mohit.p, gukyeong.kwon, cantemel, alregib}@gatech.edu

ABSTRACT

Visual explanations are logical arguments based on visual features that justify the predictions made by neural networks. Current modes of visual explanations answer questions of the form *'Why P?'*. These *Why* questions operate under broad contexts thereby providing answers that are irrelevant in some cases. We propose to constrain these *Why* questions based on some context *Q* so that our explanations answer contrastive questions of the form *'Why P, rather than Q?'*. In this paper, we formalize the structure of contrastive visual explanations for neural networks. We define contrast based on neural networks and propose a methodology to extract defined contrasts. We then use the extracted contrasts as a plug-in on top of existing *'Why P?'* techniques, specifically Grad-CAM. We demonstrate their value in analyzing both networks and data in applications of large-scale recognition, fine-grained recognition, subsurface seismic analysis, and image quality assessment.

Index Terms— Interpretability, Gradients, Deep Learning, Fine-Grained Recognition, Image Quality Assessment

1. INTRODUCTION

Explanations are a set of rationales used to understand the reasons behind a decision [1]. When these rationales are based on visual characteristics in a scene, the justifications used to understand the decision are termed as visual explanations [2]. Visual explanations can be used as a means to interpret deep neural networks. While deep networks have surpassed human level performance in traditional computer vision tasks like recognition [3], their lack of transparency in decision making has presented obstacles to their widespread adoption. We first formalize the structure of visual explanations to motivate the need for the proposed contrastive explanations. Hempel and Oppenheim [4] were the first to provide formal structure to explanations [5]. They argued that explanations are like proofs in a logical system [6] and that explanations elucidate decisions of hitherto un-interpretable systems. Typically, explanations involve an answer to structured questions of the form *'Why P?'*, where *P* refers to any decision. For instance, in recognition algorithms, *P* refers to the predicted class. In image quality assessment, *P* refers to the estimated quality. *Why*-questions are generally thought of to be *causal-like* in their explanations [7]. In this paper, we refer to them as visual causal explanations for simplicity. Note that these visual causal explanations do not allow causal inference as described by [8].

Consider an example shown in Fig. 1 where we classify between two birds - a spoonbill, and a flamingo. Given a spoonbill, a trained neural network classifies the input correctly as a spoonbill. A visual explanation of its decision generally assumes the form of a heat

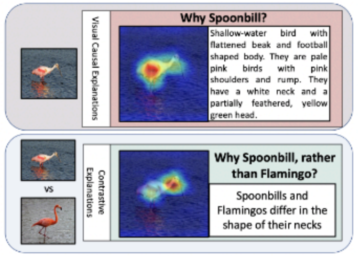


Fig. 1. The visual explanation to *Why Spoonbill?* is answered through Grad-CAM. The proposed contrastive explanatory method explains *Why Spoonbill, rather than Flamingo?* by highlighting the neck region in the same input image. Figure best viewed in color.

map that is overlaid on the image. In the visual explanations shown in Fig. 1, the red regions answer the posed question. If the posed question takes the form of *'Why Spoonbill?'*, then the regions corresponding to the body shape and color of the spoonbill are highlighted. Such an explanation is based on features that describe a Spoonbill irrespective of the context. Instead of *'Why Spoonbill?'*, if the posed question were *'Why Spoonbill, rather than Flamingo?'*, then the visual explanation points to the most contrastive features between the two birds, which in this case is the neck of the Spoonbill. Flamingos have a longer S-shaped neck not prevalent in Spoonbills. The answers to such *'Why P, rather than Q?'* questions are *contrastive explanations* where *Q* is the contrast.

The question of *'Why P, rather than Q?* provides context to the answer and hence relevance [9]. In some cases, such context can be more descriptive for interpretability. For instance, in autonomous driving applications that recognize traffic signs, knowing why a particular traffic sign was chosen over another is informative in contexts of analyzing decisions in case of accidents. Similarly, in the application of image quality assessment where an algorithm predicts the score of an image as 0.25, knowing *'Why 0.25, rather than 0.5?'* or *'Why 0.25, rather than 1?'* can be beneficial to analyze both the image and the method itself. In applications like seismic analysis where geophysicists interpret subsurface images, visualizing *'Why fault, rather than salt dome?'* can help evaluating the model, thereby increasing the trust in such systems. In this paper, we set the frame-

NOVELTY DETECTION THROUGH MODEL-BASED CHARACTERIZATION OF NEURAL NETWORKS

Gukyeong Kwon, Mohit Prabhushankar, Dogancan Temel, and Ghassan AlRegib

OLIVES at the Center for Signal and Information Processing,
School of Electrical and Computer Engineering,
Georgia Institute of Technology, Atlanta, GA, 30332-0250
{gukyeong.kwon, mohit.p, cantemel, alregib}@gatech.edu

ABSTRACT

In this paper, we propose a model-based characterization of neural networks to detect novel input types and conditions. Novelty detection is crucial to identify abnormal inputs that can significantly degrade the performance of machine learning algorithms. Majority of existing studies have focused on activation-based representations to detect abnormal inputs, which limits the characterization of abnormality from a data perspective. However, a model perspective can also be informative in terms of the novelties and abnormalities. To articulate the significance of the model perspective in novelty detection, we utilize backpropagated gradients. We conduct a comprehensive analysis to compare the representation capability of gradients with that of activation and show that the gradients outperform the activation in novel class and condition detection. We validate our approach using four image recognition datasets including MNIST, Fashion-MNIST, CIFAR-10, and CURE-TSR. We achieve a significant improvement on all four datasets with an average AUROC of 0.953, 0.918, 0.582, and 0.746, respectively.

Index Terms— Gradients, Novelty detection, Anomaly detection, Representation learning.

1. INTRODUCTION

Characterization of novel data for machine learning algorithms has become an increasingly important topic for diverse applications including but not limited to visual recognition [1], speech processing [2], and medical diagnosis [3]. In particular, when trained models are deployed in diverse environments [4, 5], new classes of input (e.g. unknown objects) or conditions (e.g. inclement conditions such as rain and snow) [6, 7] that the models have not been exposed to during training can cause a significant performance degradation. To ensure the safety of machine learning algorithms in real-world scenarios, it is essential to characterize and detect novel data.

Novelty detection, often also referred to as one-class classification or anomaly detection, is a research topic which aims

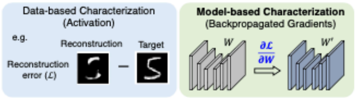


Fig. 1. Data-based and model-based characterization for novelty detection.

to classify input data that is different in some aspects from training data [8]. A key element for the success of novelty detection is to learn a representation that can clearly separate normal and abnormal data. Most of existing works have focused on learning representations obtained in a form of activation. Novelty detection based on activation-based representations characterizes *how much of the input corresponds to the learned information of the model*. For instance, assume that we input digit '5' (abnormal data) to an autoencoder trained to accurately reconstruct digit '0' (normal data). Based on the reconstructed image, which is the activation-based representation of the autoencoder, we calculate the reconstruction error as shown in the left side of Fig. 1. Since the autoencoder has learned round shape information from '0', the curved edges at the top and the bottom of '5' are reconstructed but straight edges in the middle cannot be accurately reconstructed. The reconstruction error captures what the autoencoder has not learned and quantifies the abnormality. We can interpret this novelty detection based on activation-based representation as the characterization of abnormality from a data perspective.

In this paper, we propose to characterize novelty from a model perspective. In particular, we use backpropagated gradients from neural networks to obtain the model-based characterization of abnormality. A gradient is generated through backpropagation to train neural networks by minimizing designed loss functions [9]. The gradient with respect to the weights provides directional information to update the neural network. Also, the abnormal data requires more drastic updates on neural networks compared to the normal data. There-

IMPLICIT SALIENCY IN DEEP NEURAL NETWORKS

Yutong Sun, Mohit Prabhushankar, and Ghassan AlRegib

OLIVES at the Center for Signal and Information Processing
School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA, 30332-0250
{ysun465, mohit.p, alregib}@gatech.edu

ABSTRACT

In this paper, we show that existing recognition and localization deep architectures, that have not been exposed to eye tracking data or any saliency datasets, are capable of predicting the human visual saliency. We term this as *implicit saliency* in deep neural networks. We calculate this *implicit saliency* using expectancy-mismatch hypothesis in an *unsupervised fashion*. Our experiments show that extracting saliency in this fashion provides comparable performance when measured against the state-of-art *supervised* algorithms. Additionally, the robustness outperforms those algorithms when we add large noise to the input images. Also, we show that semantic features contribute more than low-level features for human visual saliency detection. Based on these properties and performances, our proposed method greatly lowers the threshold for saliency detection in terms of required data and bridges the gap between human visual saliency and model saliency.

Index Terms— Saliency, Implicit Saliency, Expectation Mismatch, Recognition, Deep Learning

1. INTRODUCTION

Saliency is defined as those regions in a visual scene that are 'most noticeable' and attract significant attention [1]. Human visual saliency detection has been deployed in an extensive set of image processing applications including but not limited to data compression, image segmentation, recognition, image quality assessment (IQA) and object recognition [2]. Broadly, saliency detection algorithms can be classified into two categories. The first is bottom-up approaches where saliency detection techniques extract features from data and compute saliency based on extracted features [3, 4, 5]. The second is top-down approaches where the algorithms have a prior target for which features are to be calculated [6]. Both these approaches derive from the expectancy mismatch hypothesis [7].

The expectancy-mismatch hypothesis for a sensory system is based on receiving information which is in conflict with the system's prior expectation. The authors in [7] show that a message which is unexpected, captures human attention and is hence salient. Extensive work in the field of cognitive sciences has been conducted to study the impact of expectancy-mismatch in human attention and visual saliency [8, 9, 10, 11, 7]. Based on these works, human attention mechanism suppresses expected messages and focuses on the unexpected ones. During this process, human visual system checks whether the input scenario matches the observers' expectation and past experience. When they are conflicting, error neurons in human brain encode the prediction error and pass the error message back to the representational neurons. Existing work applies this concept of expectancy-mismatch to saliency detection. The authors in [10, 11] show how unexpected colors impact human eye fixations. [7] indicates that a motion singleton captures attention.

Previous works that define expectations and calculate mismatches are based on low-level representations like colors and edges. However, the advent of deep learning has shown the importance of semantic information that combines low-level features for complicated tasks like recognition. Neural networks have shown an aptitude for learning higher-order semantic representations. In [12], the authors claim that it is crucial to consider semantic representations in saliency detection. In this paper, we propose to create expectancy based on high level semantic features and calculate mismatch from input information to obtain saliency. To set expectancy, we use neural networks. To calculate mismatch, we provide conflicting information to the network along with the input image to search for those regions in the input image that are affected by the conflict. In this work, conflicting information refers to labels that conflict with predicted classes. For instance, consider Fig. 1. The network has learned the low-level features like edges and colors and their combinatorial high-level semantics to recognize a car. However, by providing a conflicting label such as 'airplane', we force the network to reexamine its decision process. The network reconciles its expectation of finding a car and the conflicting label that it is an airplane by encoding the error within the gradients. These gradients are backpropagated throughout the network to resolve the conflict. The change brought about by the gradients is indicative of regions within the image that are used for expecting the output. We postulate that these regions are thereby salient.

In this work, we use commonly used recognition and localization pre-trained networks to set expectancy. These networks have not been exposed to either saliency datasets or eye-tracking data. Hence, the proposed method is completely unsupervised. We extract saliency that is implicitly embedded within any given network. Hence, the proposed approach is termed implicit saliency in neural networks. The contributions in this paper are three-fold:

1) we extract implicit saliency from pre-trained networks that have not seen eye-tracking data in an *unsupervised* fashion.
2) we show that the proposed implicit saliency is robust to noise.
3) we show that semantic features combined with unexpected stimuli have a higher correlation with human visual saliency than low-level features or semantic features without unexpected stimuli.

We introduce the background for the pre-trained deep networks in Section 2. In Section 3, we detail the proposed method to extract implicit saliency. In Section 4, we compare the performance of proposed method against state-of-art supervised methods and model saliency methods. We conclude in Sec. 5.

Providing novelty/contributions by contrasting against existing related works!

OLIVES
@GeorgiaTech
www.ghassanalregib.com

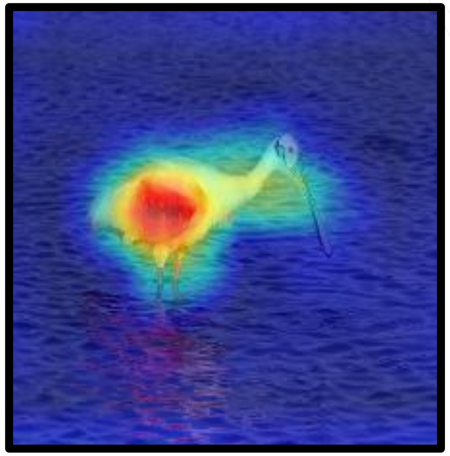
Georgia Tech
CREATING THE NEXT

Visual Explanations

Causal factors based visual explanations – answers to *`Why?'* Questions

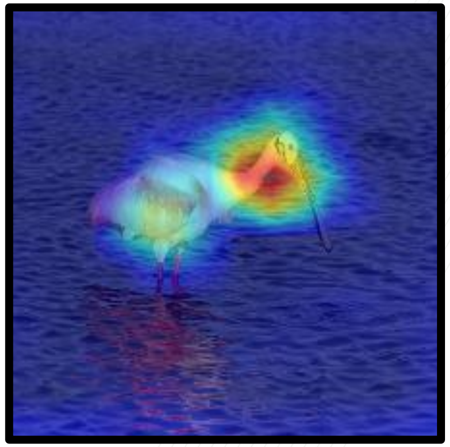
Why Spoonbill?

Shallow-water bird with flattened beak and football shaped body. They are pale pink birds with pink shoulders and rump. They have a white neck and a partially feathered, yellow green head.



Why Spoonbill, rather than Flamingo?

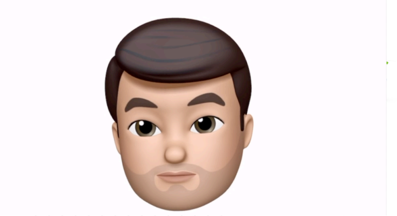
Spoonbills have shorter legs and necks compared to Flamingos



OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

IEEE Internatic
on Image Proc
25-28 October 2020

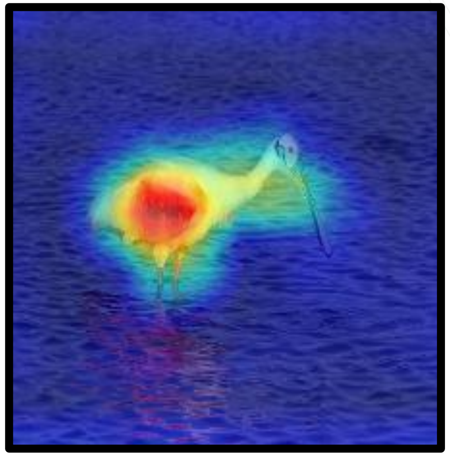


Contrastive Visual Explanations

Causal factors based visual explanations – answers to *`Why?'* Questions

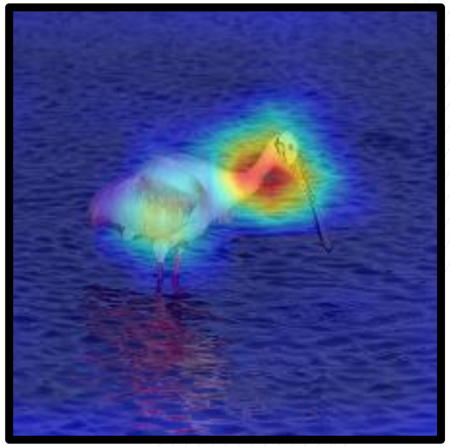
Why Spoonbill?

Shallow-water bird with flattened beak and football shaped body. They are pale pink birds with pink shoulders and rump. They have a white neck and a partially feathered, yellow green head.



Why Spoonbill, rather than Flamingo?

Spoonbills have shorter legs and necks compared to Flamingos



OLIVES
@GeorgiaTech

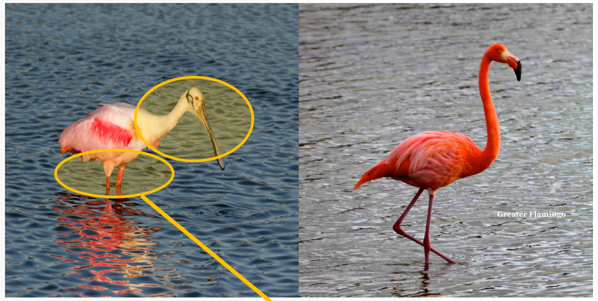
Contrastive visual explanations – answers to *`Why P, rather than Q?'* Questions

Georgia Tech
CREATING THE NEXT

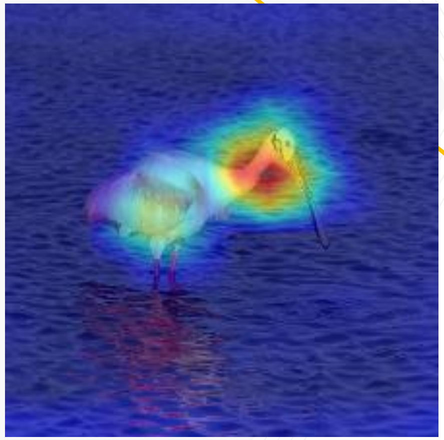
IEEE Internatic
on Image Proc
25-28 October 2020

Paper Objective

Contrast B/w Spoonbill and Flamingo



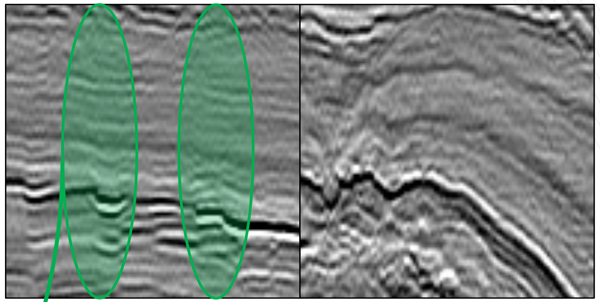
Contrastive Explanation



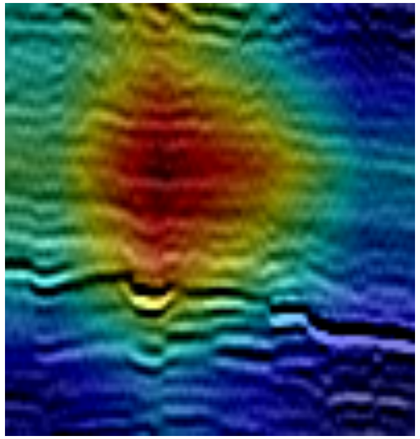
Contrast B/w Bugatti Convertible and Coupe

Contrastive Explanation

Contrast B/w Fault and Salt Dome



Contrastive Explanation



No Contrastive Ground Truths

Objective :
- Define Contrast from a visual and representational sense
- Extract contrast in an unsupervised fashion

OLIVES @GeorgiaTech

Georgia Tech
CREATING THE NEXT

Paper Structure and Novelty

IEEE Internatic
on Image Proc
25-28 October 2020

Structure :

Formalize structure of explanations → Define Contrast → Extract Contrastive explanations from Neural Nets → Applicability in Recognition and Image Quality Assessment

Novelty
Formalize explanations through First-Order logic Provide an unsupervised and intuitive methodology to extract contrastive explanation

OLIVES @GeorgiaTech

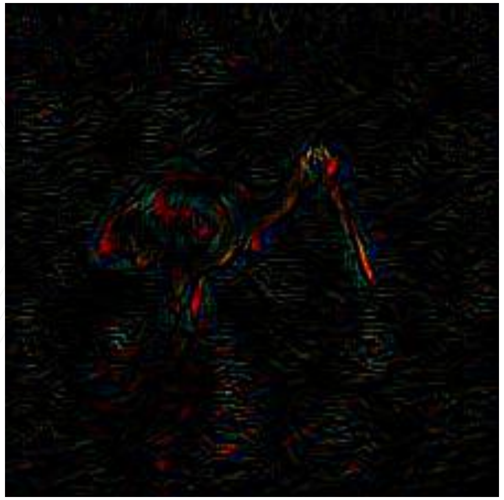
Georgia Tech CREATING THE NEXT

Nature of Explanations

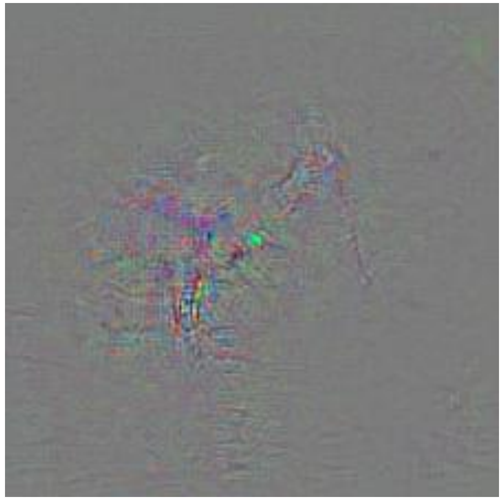
IEEE International Conference on Image Processing
25-28 October 2020

Formalize structure of explanations	Define Contrast	Extract Contrastive explanations from Neural Nets	Applicability in Recognition and Image Quality Assessment

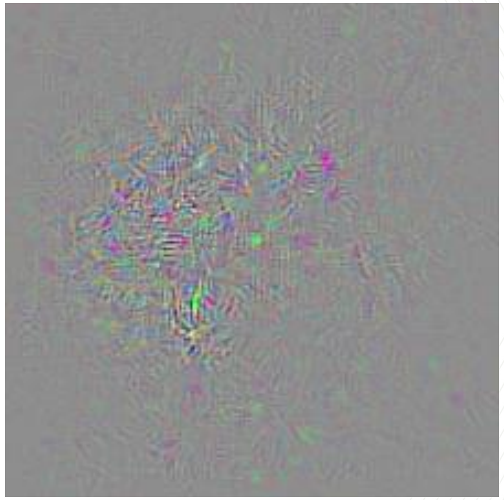
Guided Backpropagation



Positive saliency



Smooth Gradients



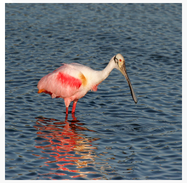
Vanilla Backpropagation

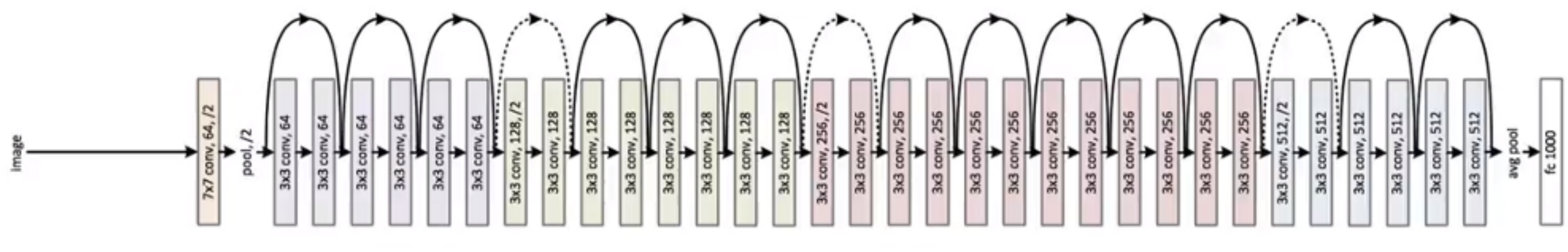
All existing approaches answer *`Why Spoonbill?'*

Contrast in Neural Networks

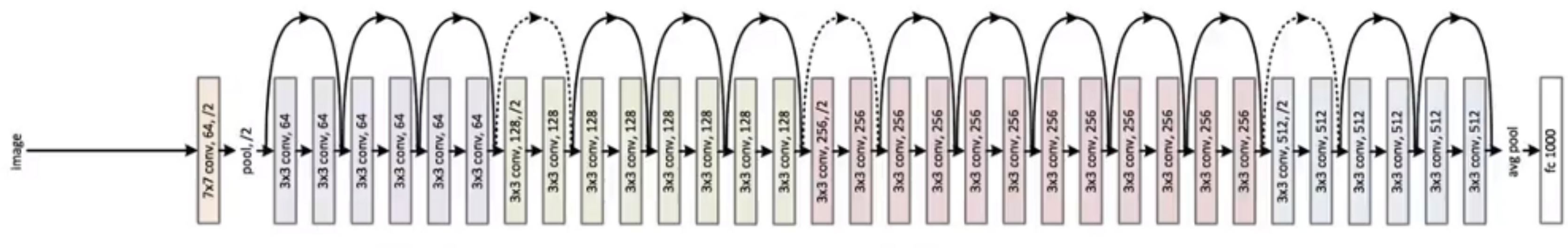
Formalize structure of explanations	Define Contrast	Extract Contrastive explanations from Neural Nets	Applicability in Recognition and Image Quality Assessment

In Visual space, contrast is the perceived difference between two known quantities





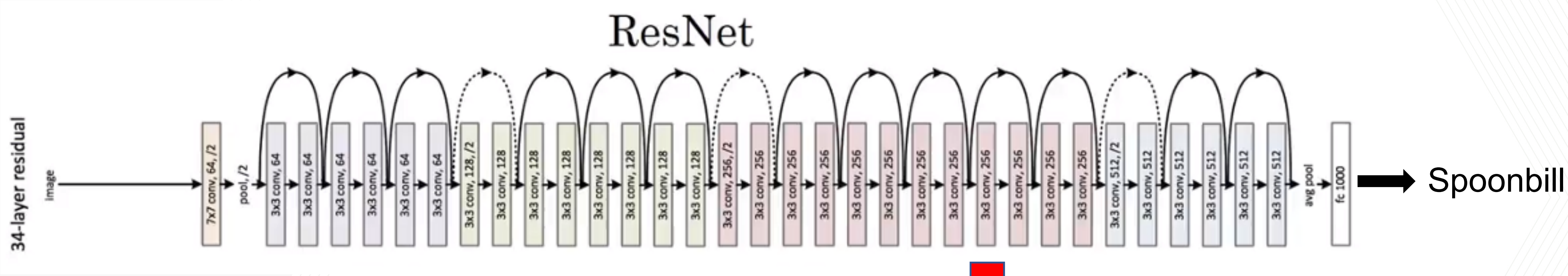
→ Spoonbill



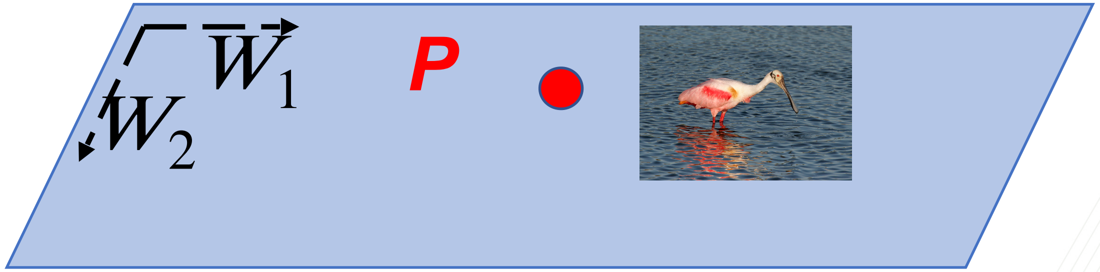
→ Flamingo

Contrast in Neural Networks

Formalize structure of explanations	**Define Contrast**	Extract Contrastive explanations from Neural Nets	Applicability in Recognition and Image Quality Assessment



Projection from an arbitrary layer spans a manifold

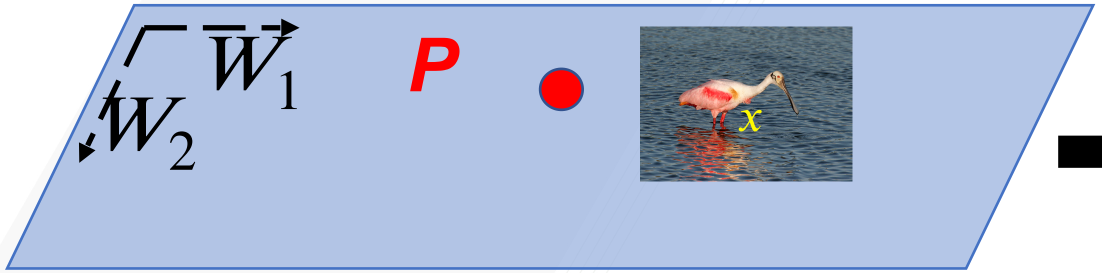


Learned Manifold : ***spoonbill*** labeled a ***spoonbill***

Contrast in Neural Networks

Formalize structure of explanations	**Define Contrast**	Extract Contrastive explanations from Neural Nets	Applicability in Recognition and Image Quality Assessment

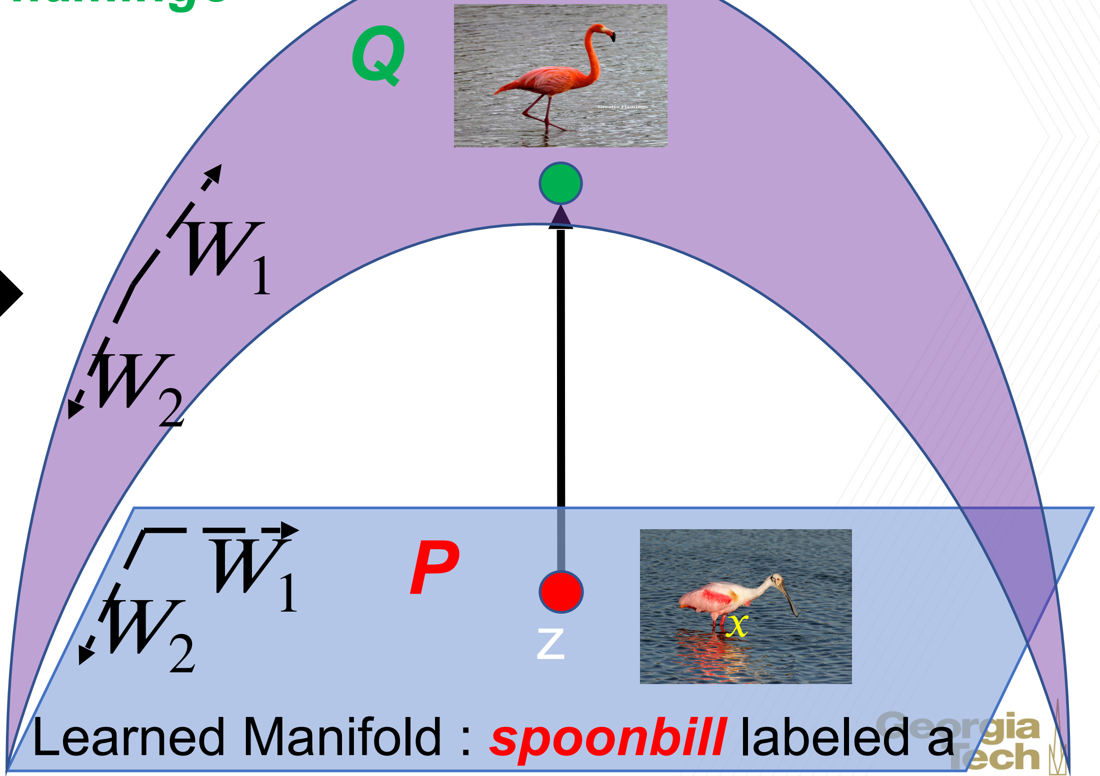
In representation space, contrast is the distance between the manifolds that predicts the input image x as P vs x as Q



Learned Manifold : ***spoonbill*** labeled a ***spoonbill***

Introduce Contrast

Contrastive Manifold : ***spoonbill*** labeled a ***flamingo***

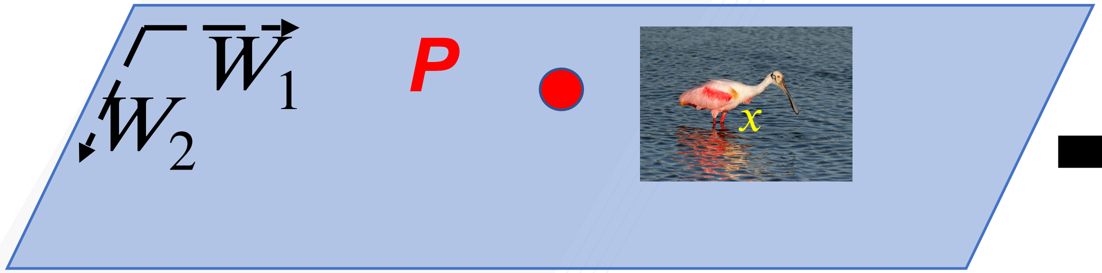


Learned Manifold : ***spoonbill*** labeled a ***spoonbill***

Contrast in Neural Networks

Formalize structure of explanations | **Define Contrast** | Extract Contrastive explanations from Neural Nets | Applicability in Recognition and Image Quality Assessment

In representation space, contrast is the distance between the manifolds that predicts the input image x as P vs x as Q



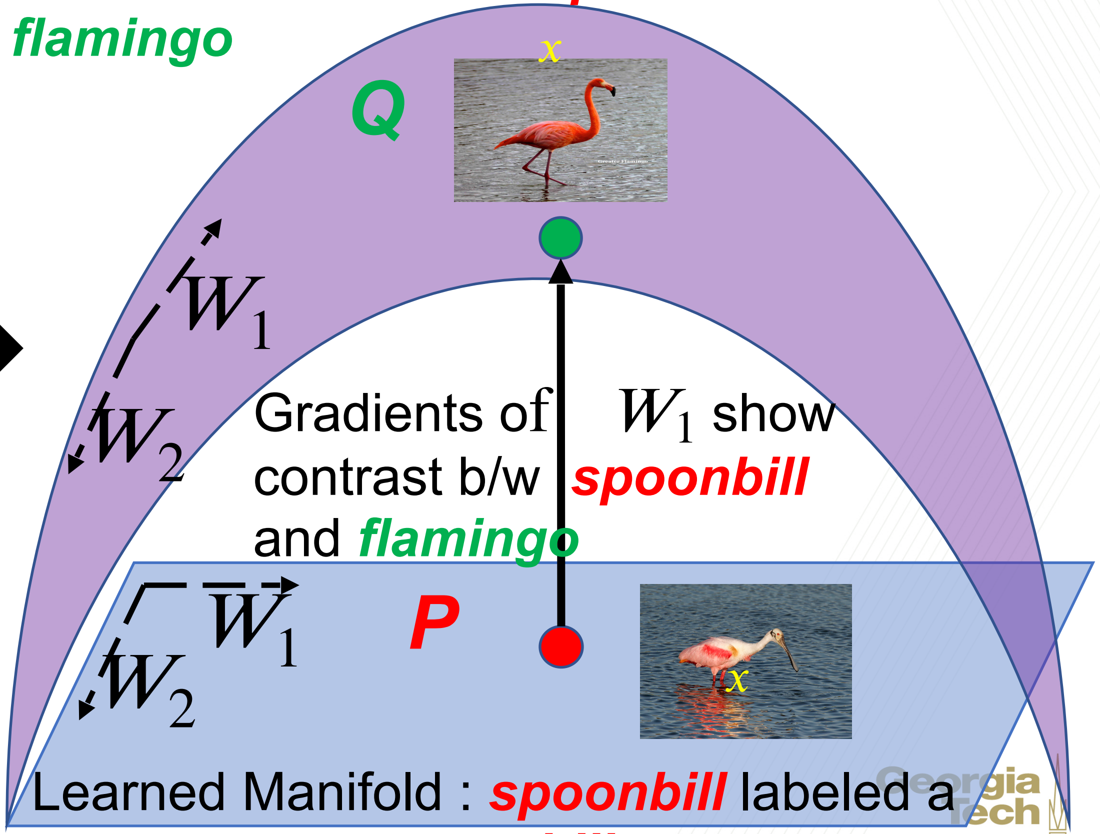
Learned Manifold : ***spoonbill*** labeled a ***spoonbill***

Introduce

Contrast

Gradients as Contrastive Features

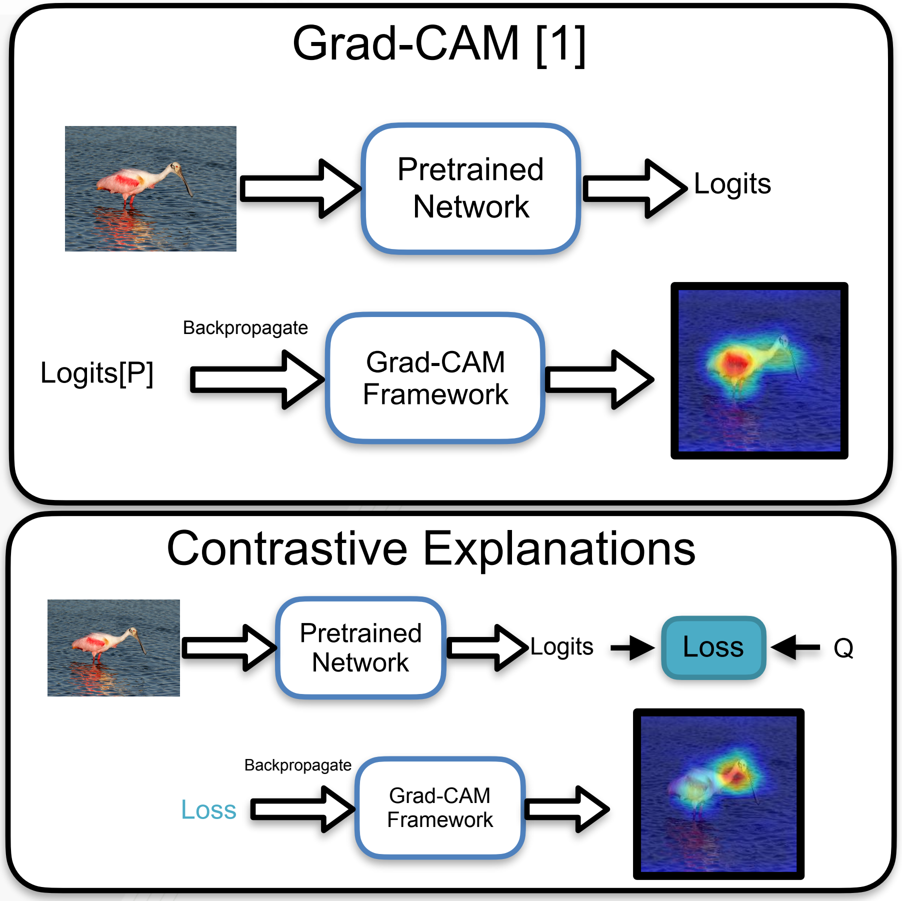
Contrastive Manifold : ***spoonbill*** labeled a ***flamingo***



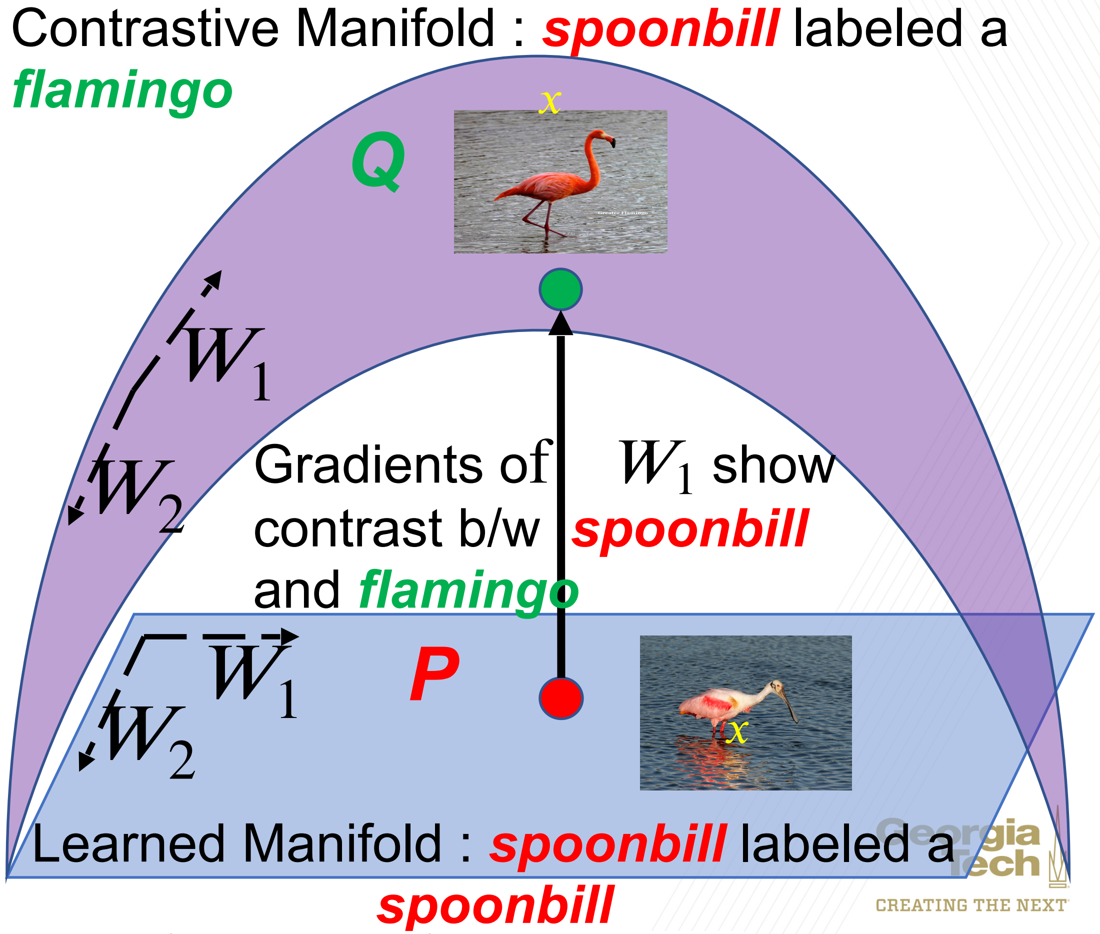
spoonbill

Contrast in Neural Networks

Formalize structure of explanations | Define Contrast | **Extract Contrastive explanations from Neural Nets** | Applicability in Recognition and Image Quality Assessment



Gradients as Contrastive Features



[1] 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.

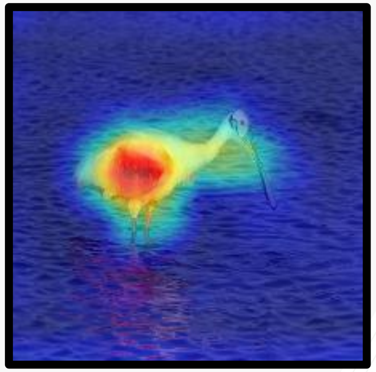
Contrast in Neural Networks

Formalize structure of explanations	Define Contrast	Extract Contrastive explanations from Neural Nets	Applicability in Recognition and Image Quality Assessment

Implementation : Within Grad-CAM framework

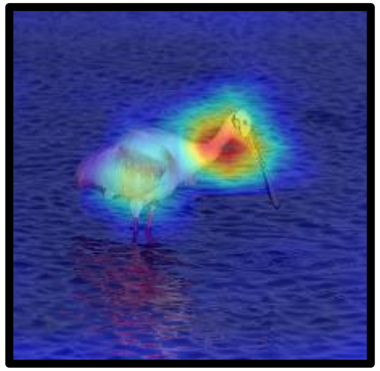
Grad-CAM

```
logit = self.model_arch(input)
#Grad-CAM gradient initialization
if class_idx is None:
    score = logit[:, logit.max(1)[-1]].squeeze()
else:
    score = logit[:, class_idx].squeeze()

self.model_arch.zero_grad()
score.backward(retain_graph=retain_graph)
```

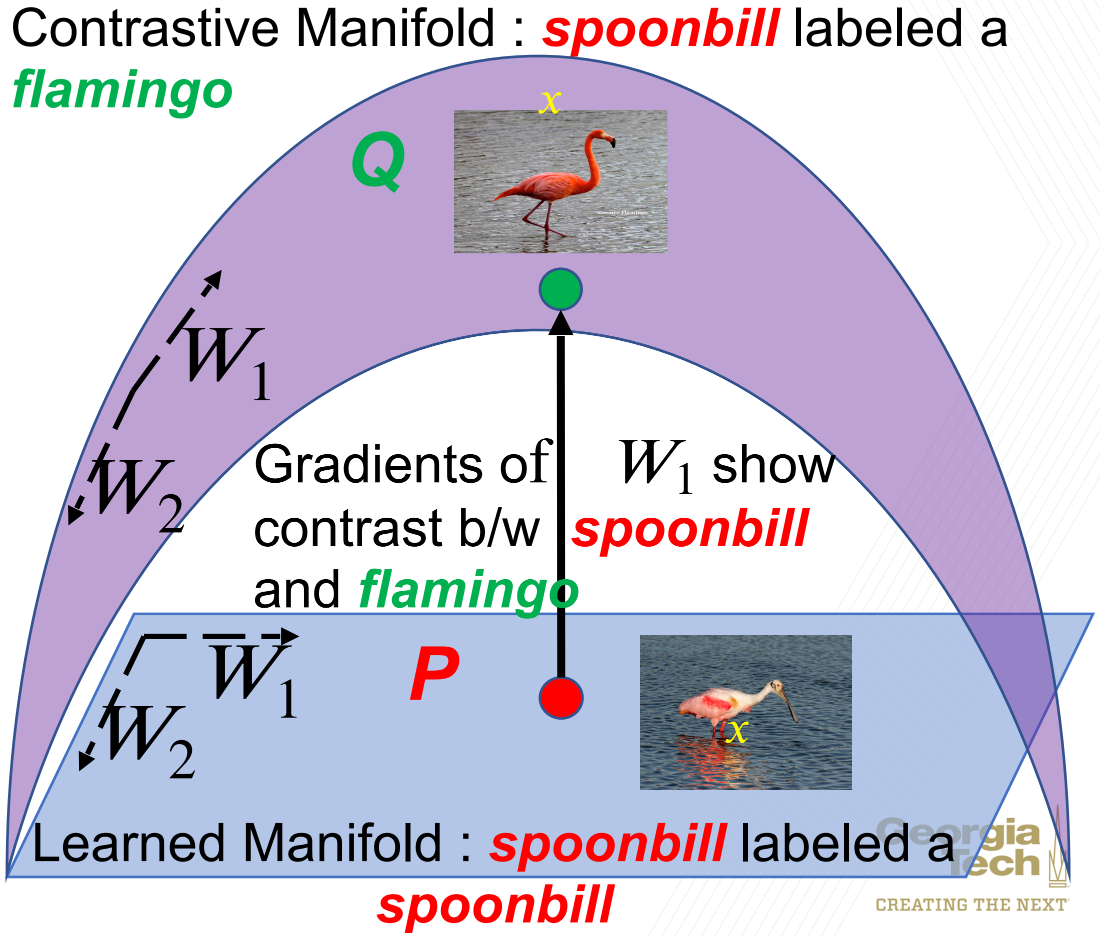


Contrastive Explanation

```
logit = self.model_arch(input)
# The only change to Grad-CAM code
ce_loss = nn.CrossEntropyLoss()
im_label_as_var = Variable(torch.from_numpy(np.asarray([Q])))
pred_loss = ce_loss(logit.cuda(), im_label_as_var.cuda())

self.model_arch.zero_grad()
pred_loss.backward()
```



Gradients as Contrastive Features



Recognition

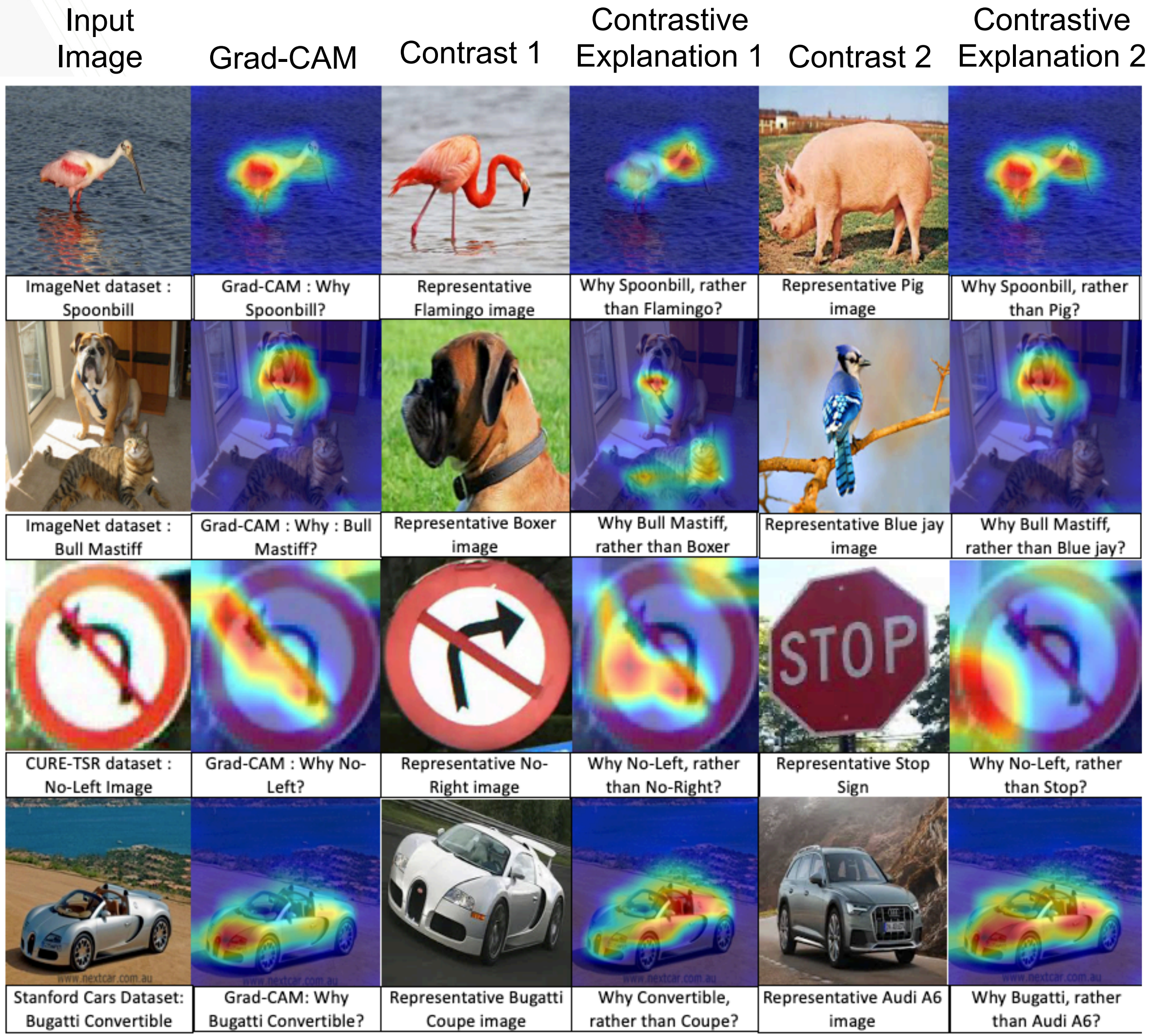
IEEE Internatio
on Image Proc
25-28 October 2020

Formalize structure of explanations

Define Contrast

Extract Contrastive explanations from Neural Nets

Applicability in Recognition and Image Quality Assessment



OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

Recognition

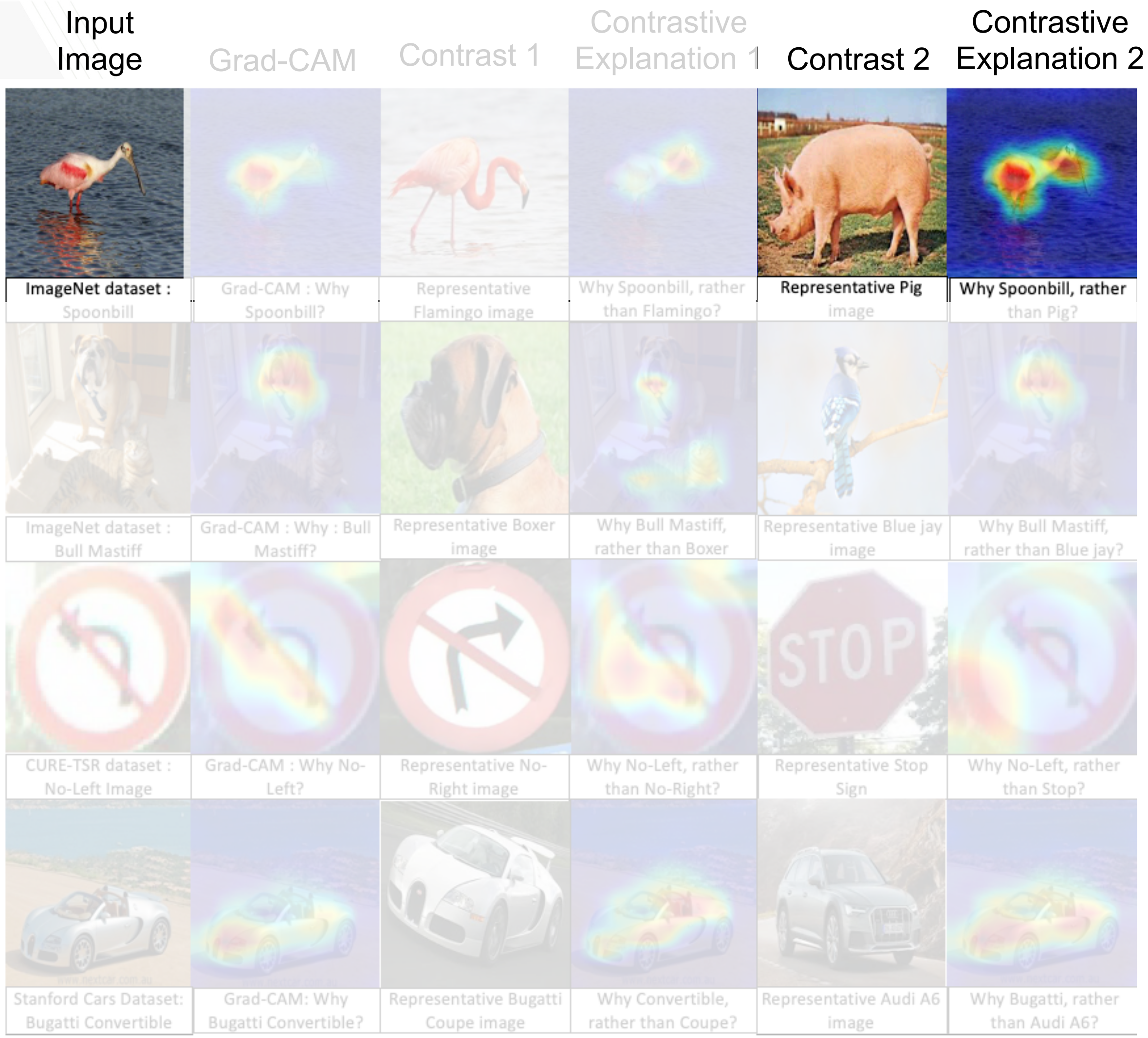
IEEE International Conference on Image Processing
25-28 October 2020, United Arab Emirates

Formalize structure of explanations

Define Contrast

Extract Contrastive explanations from Neural Nets

Applicability in Recognition and Image Quality Assessment

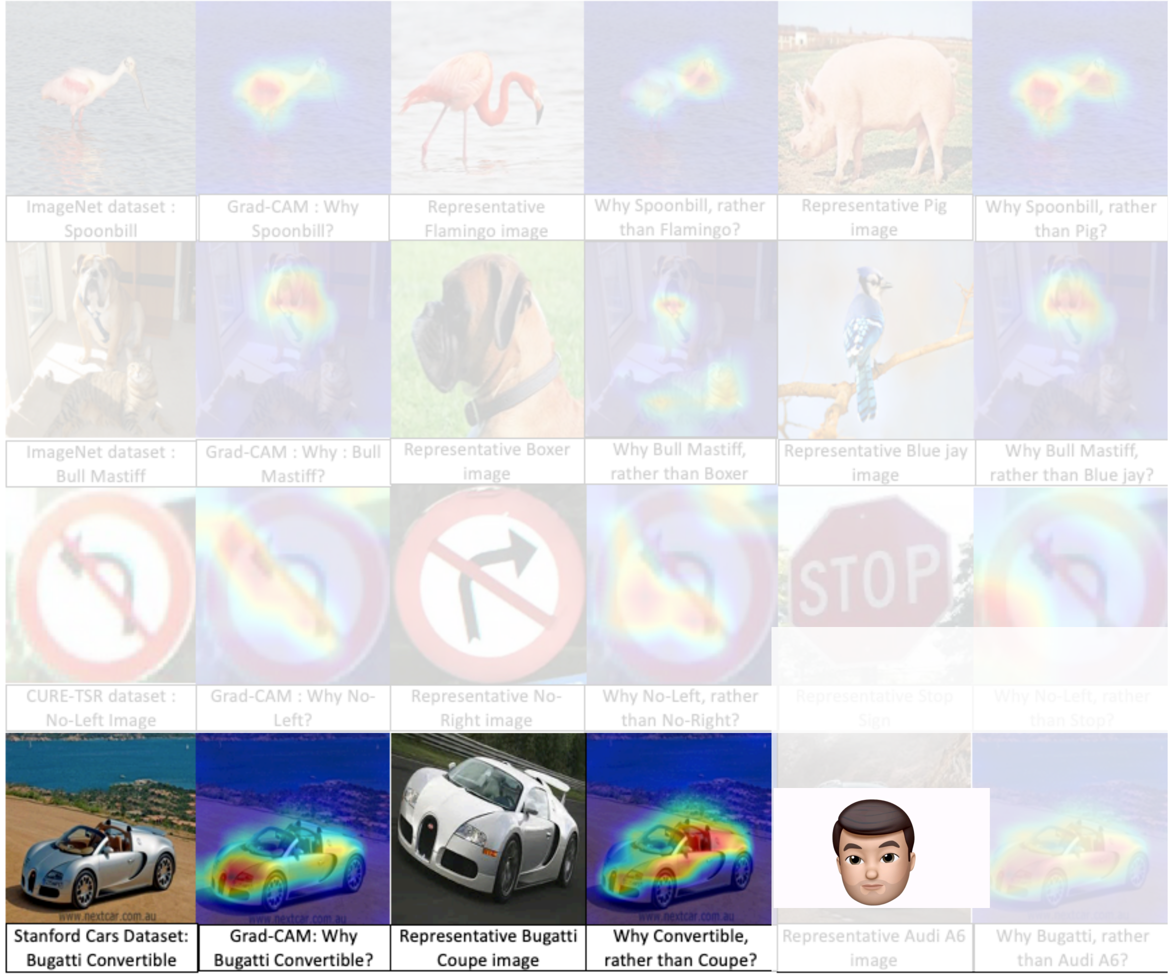


Recognition

IEEE International Conference on Image Processing
25-28 October 2020, United Arab Emirates

Formalize structure of explanations	Define Contrast	Extract Contrastive explanations from Neural Nets	**Applicability in Recognition and Image Quality Assessment**

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



OLIVES @GeorgiaTech

Georgia Tech
CREATING THE NEXT

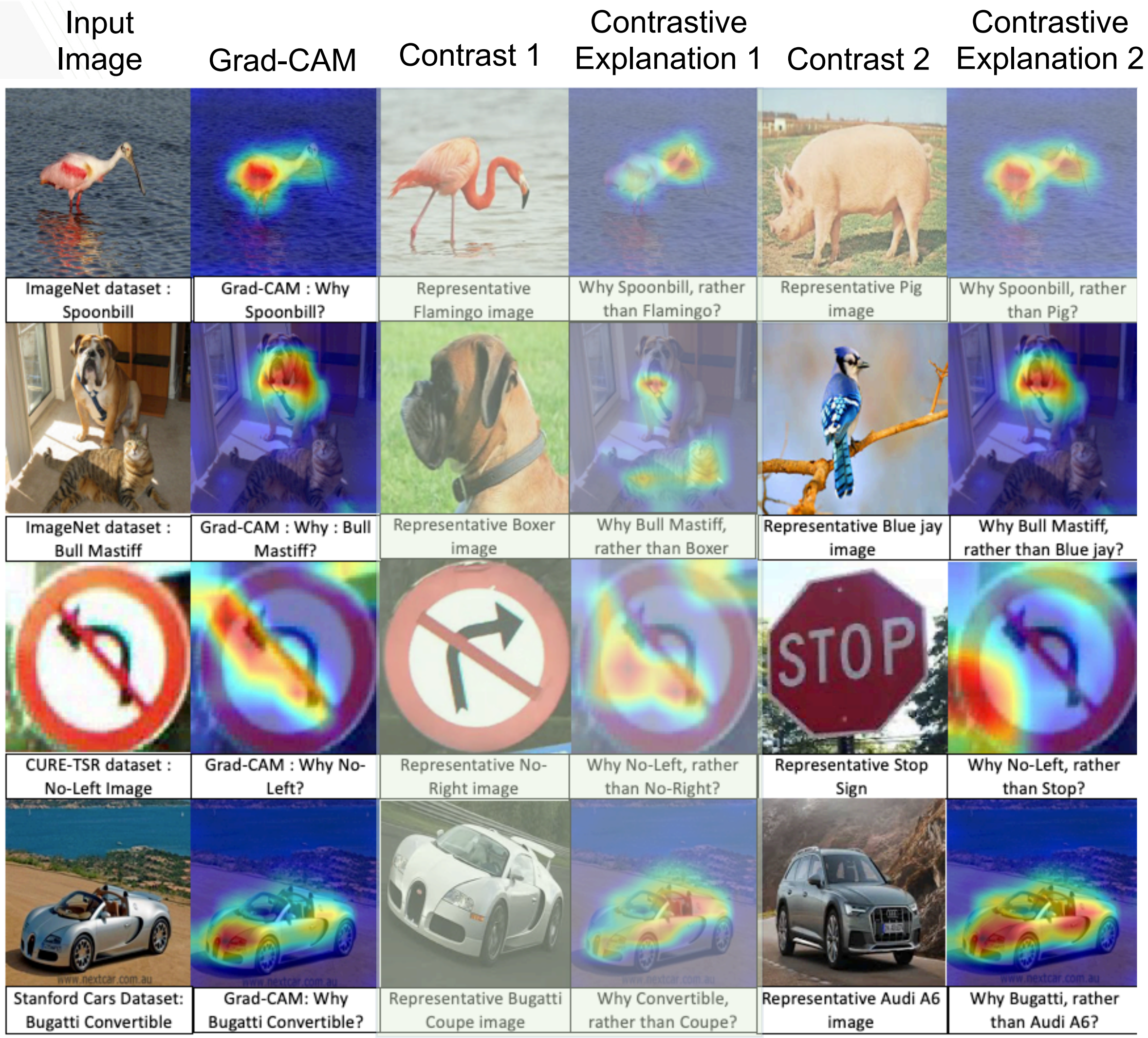
Recognition

Formalize structure of explanations

Define Contrast

Extract Contrastive explanations from Neural Nets

Applicability in Recognition and Image Quality Assessment



Human Interpretable

Recognition

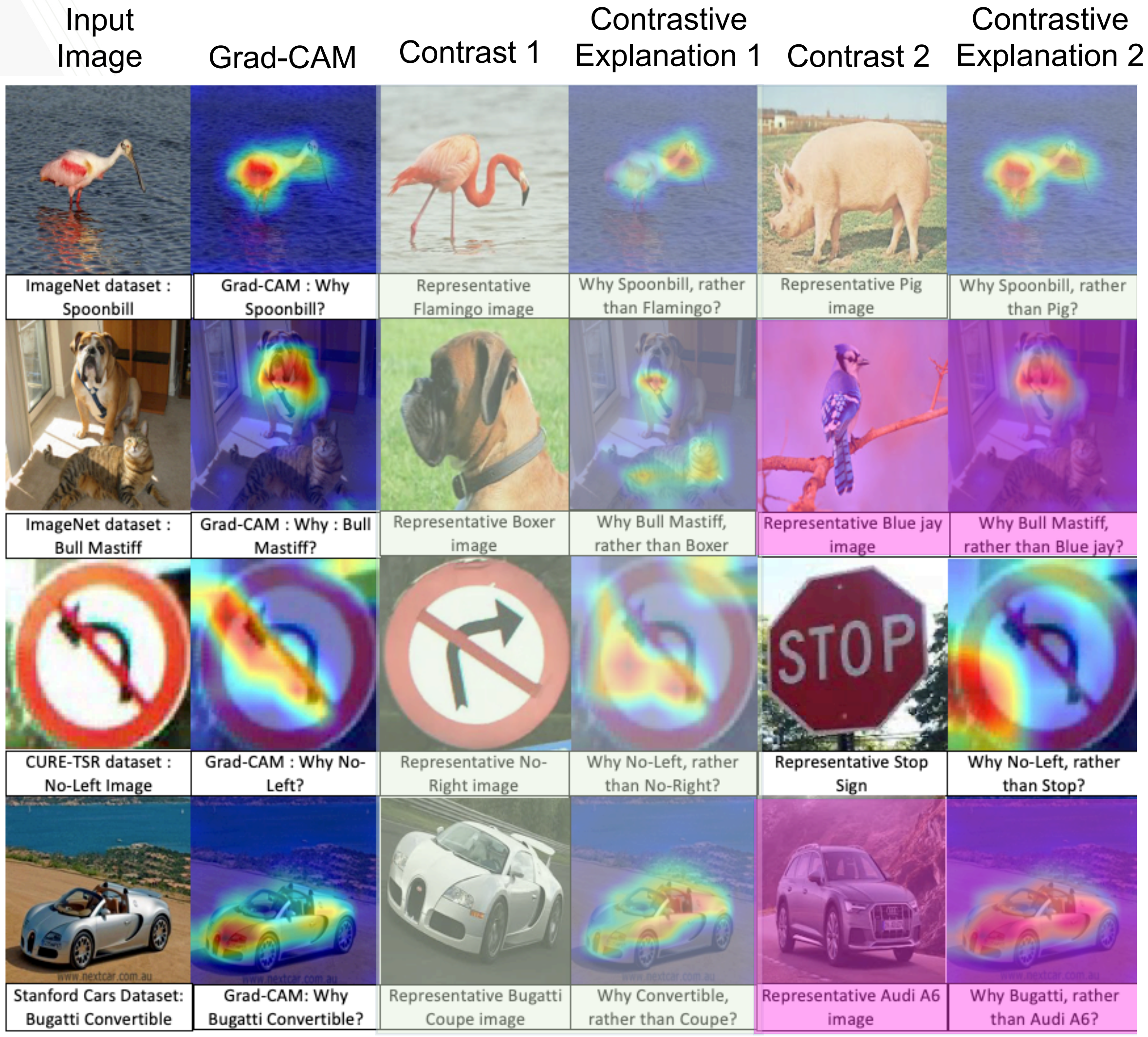
IEEE International Conference on Image Processing
25-28 October 2020

Formalize structure of explanations

Define Contrast

Extract Contrastive explanations from Neural Nets

Applicability in Recognition and Image Quality Assessment



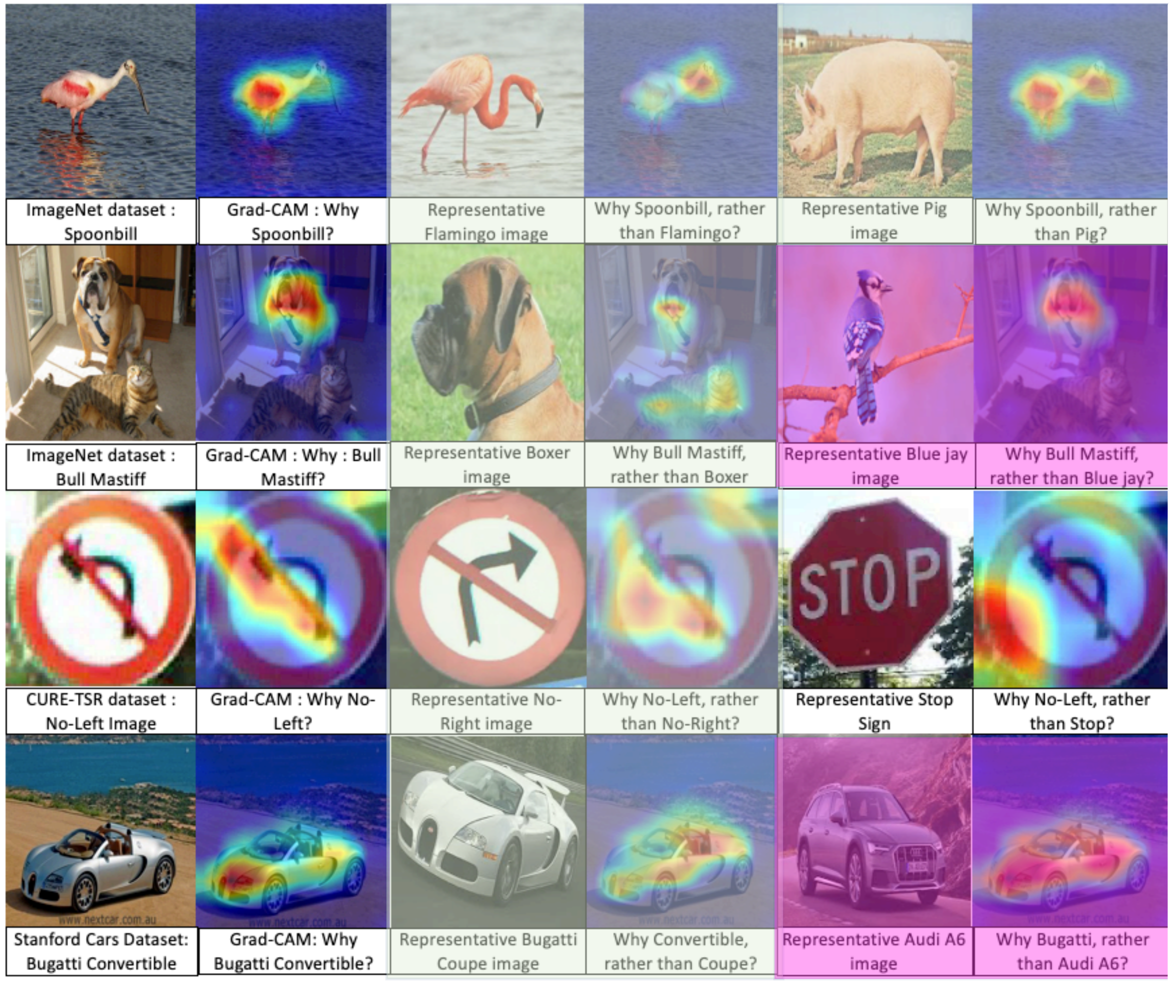
Human Interpretable

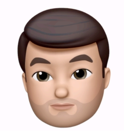
Same as Grad-CAM

Recognition

IEEE International Conference on Image Processing
25-28 October 2020, United Arab Emirates

Formalize structure of explanations | Define Contrast | Extract Contrastive explanations from Neural Nets | **Applicability in Recognition and Image Quality Assessment**





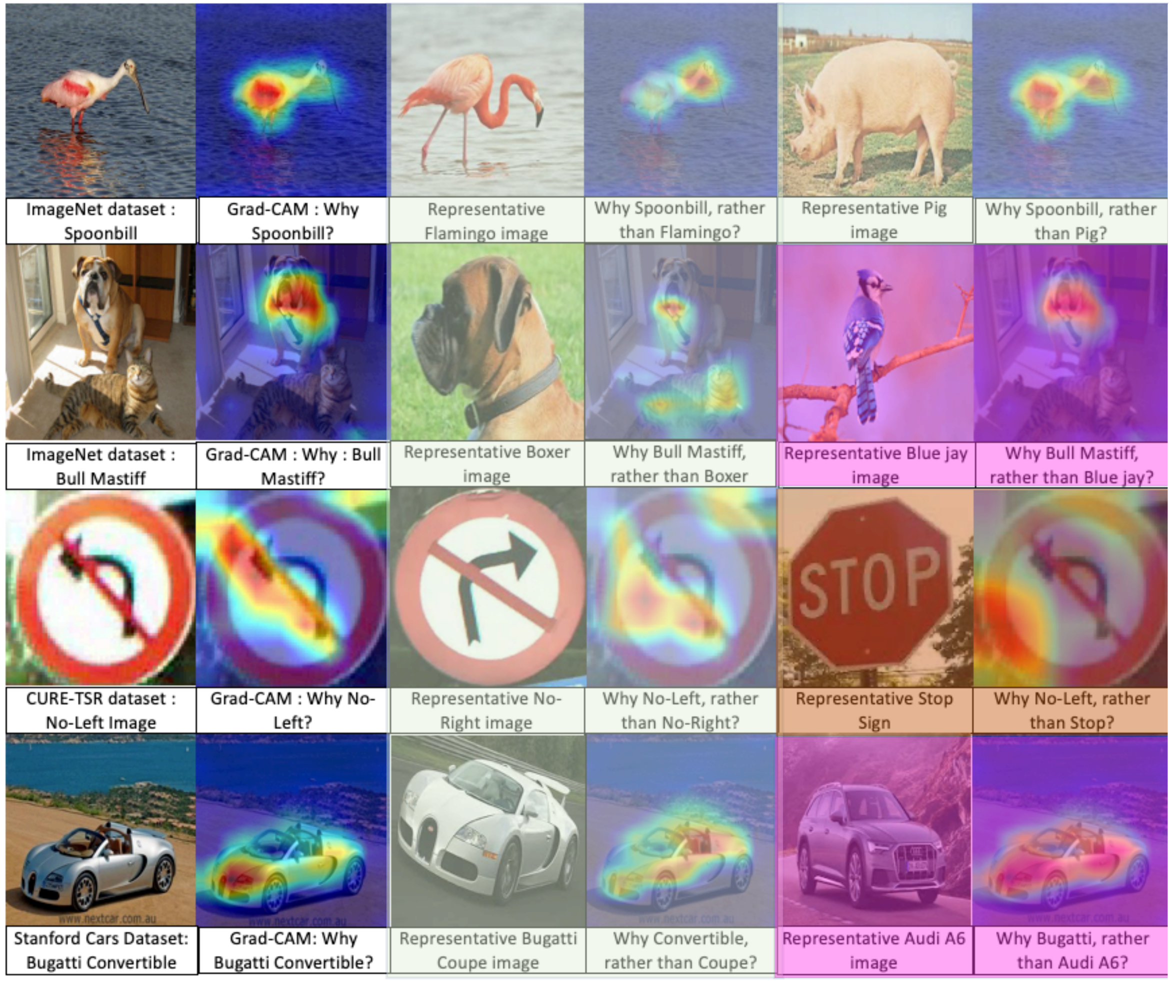
Human Interpretable

Same as Grad-CAM

Recognition

Formalize structure of explanations | Define Contrast | Extract Contrastive explanations from Neural Nets | **Applicability in Recognition and Image Quality Assessment**

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



Human Interpretable

Same as Grad-CAM

Not Human Interpretable

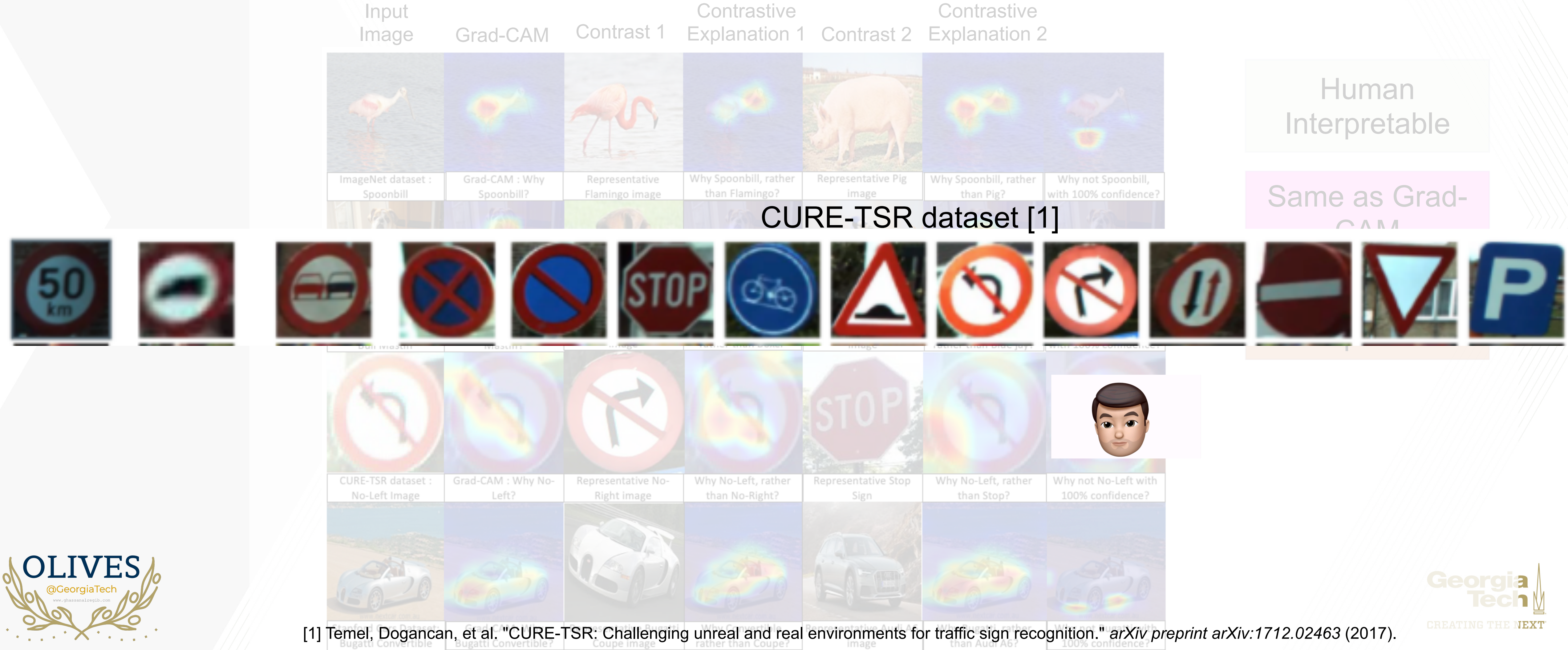
Recognition

Formalize structure of explanations | Define Contrast | Extract Contrastive explanations from Neural Nets | **Applicability in Recognition and Image Quality Assessment**

Human Interpretable

Same as Grad-CAM

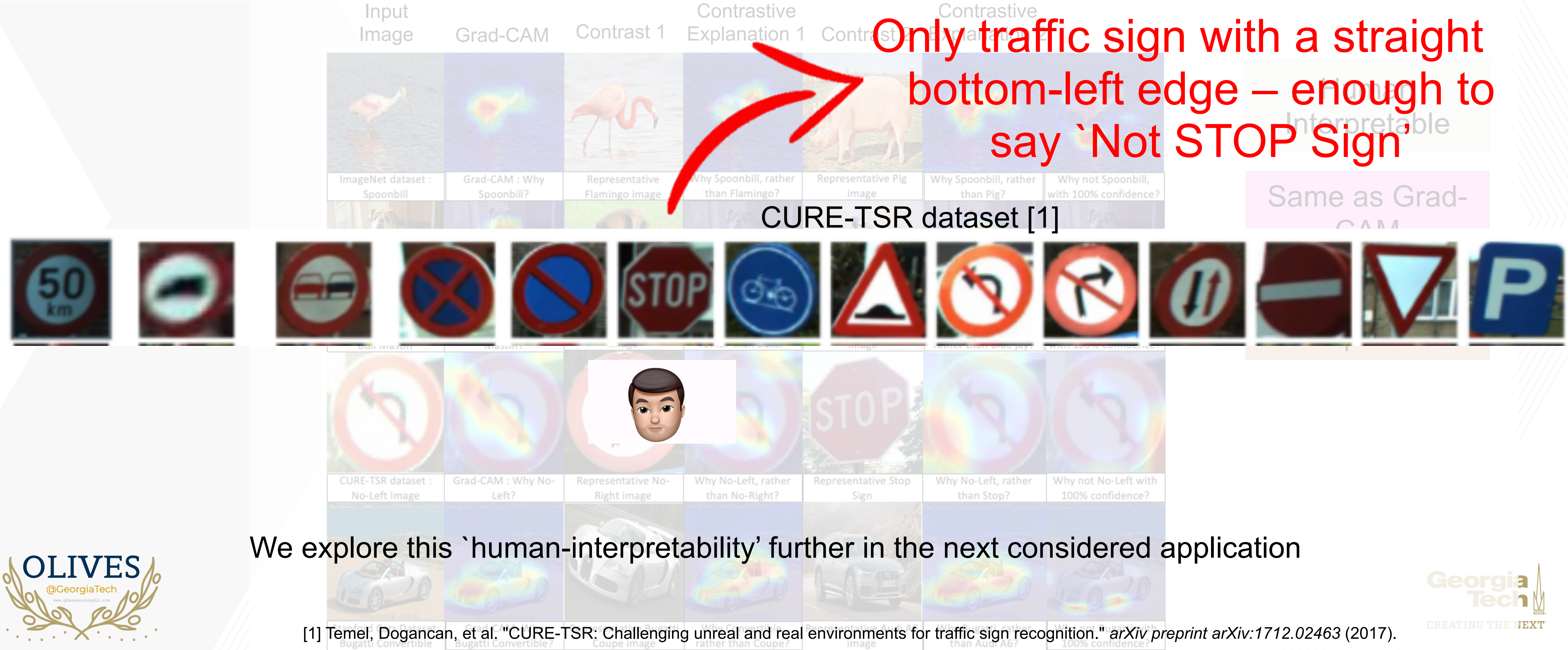
CURE-TSR dataset [1]



[1] Temel, Dogancan, et al. "CURE-TSR: Challenging unreal and real environments for traffic sign recognition." *arXiv preprint arXiv:1712.02463* (2017).

Recognition

Formalize structure of explanations | Define Contrast | Extract Contrastive explanations from Neural Nets | **Applicability in Recognition and Image Quality Assessment**



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

CURE-TSR dataset [1]

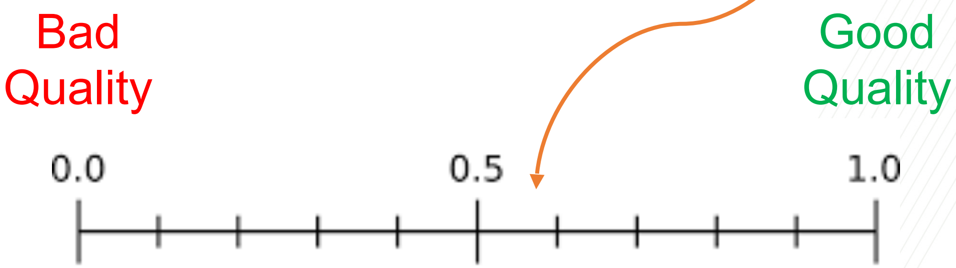
We explore this `human-interpretability' further in the next considered application

[1] Temel, Dogancan, et al. "CURE-TSR: Challenging unreal and real environments for traffic sign recognition." *arXiv preprint arXiv:1712.02463* (2017).

Image Quality Assessment

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

OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

[1] 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.

Image Quality Assessment

IEEE Internatic
on Image Proc
25-28 October 2020

Formalize structure of explanations | Define Contrast | Extract Contrastive explanations from Neural Nets | **Applicability in Recognition and Image Quality Assessment**

Lighthouse image with level 5 lossy compression from TID 2013 dataset

DIQaM : 0.58

Grad-CAM

Why 0.58?

Bad Quality

Good Quality

The given image is somewhat OK quality

OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

Image Quality Assessment

IEEE International Conference on Image Processing 25-28 October 2020

Formalize structure of explanations | Define Contrast | Extract Contrastive explanations from Neural Nets | **Applicability in Recognition and Image Quality Assessment**

DIQaM : 0.58

Grad-CAM

Why 0.58?

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

Lighthouse image with level 5 lossy compression from TID 2013 dataset

Bad Quality

Good Quality

0.0 0.5 1.0

The given image is somewhat OK quality

OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

Image Quality Assessment

Formalize structure of explanations | Define Contrast | Extract Contrastive explanations from Neural Nets | **Applicability in Recognition and Image Quality Assessment**

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

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

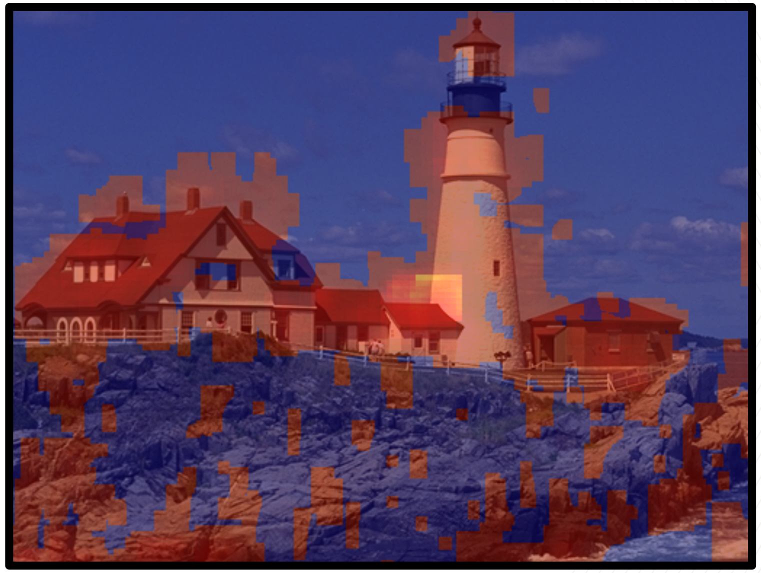
Image Quality Assessment

Lighthouse image with level 5 lossy compression from TID 2013 dataset

DIQaM : 0.58

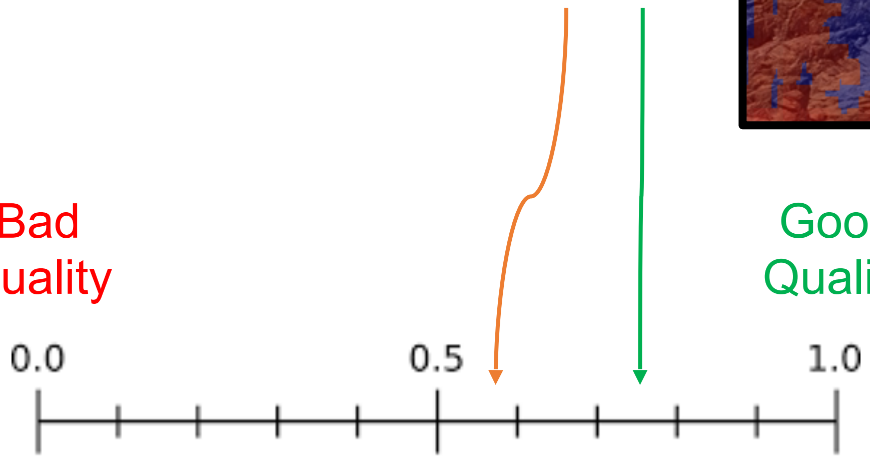
Contrastive explanation

Why 0.58, rather than 0.75?



Bad Quality

Good Quality



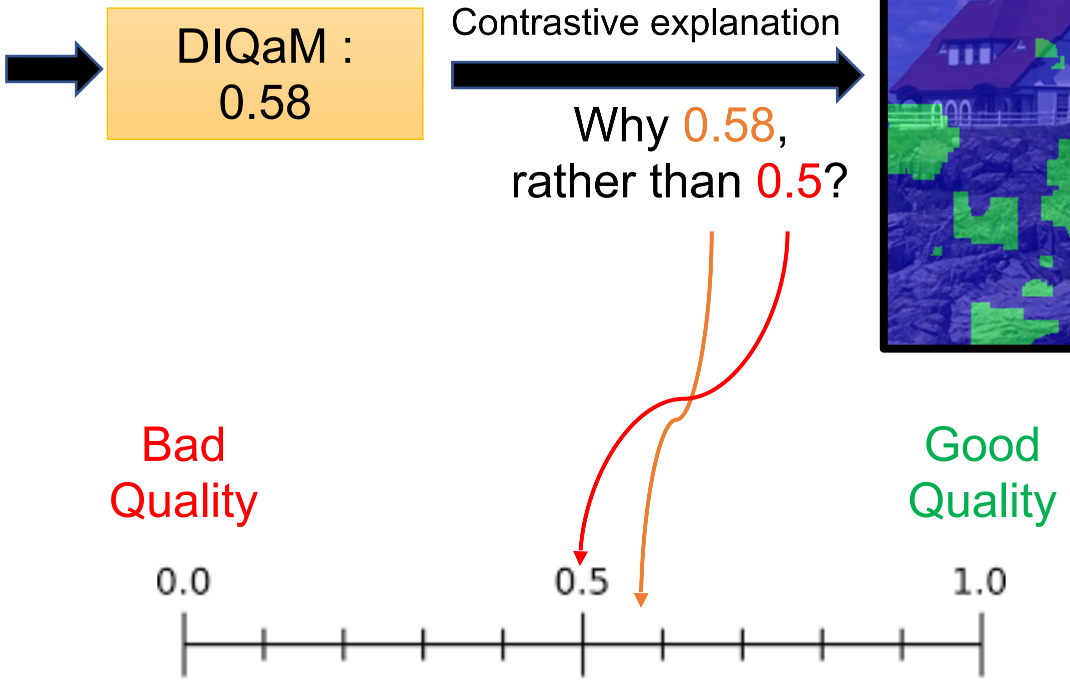
The distortions on the lighthouse and houses prevent a higher score of 0.75

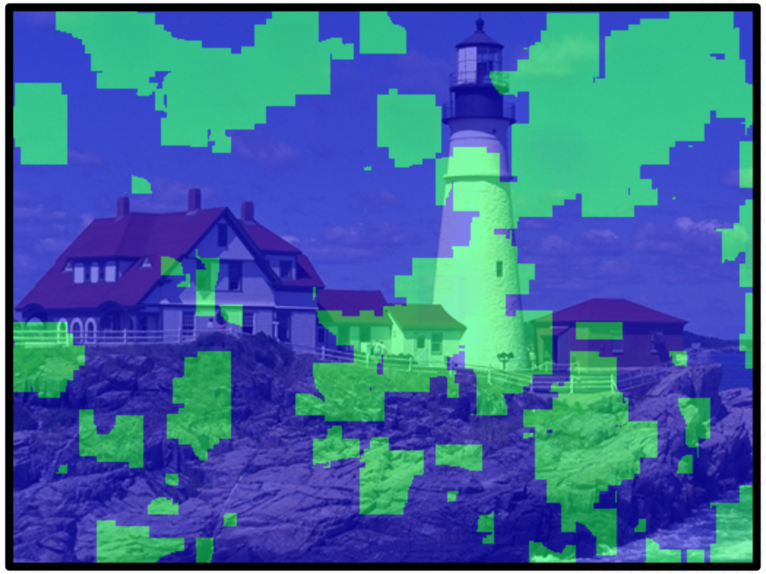
Image Quality Assessment

IEEE International Conference on Image Processing
25-28 October 2020

Formalize structure of explanations | Define Contrast | Extract Contrastive explanations from Neural Nets | **Applicability in Recognition and Image Quality Assessment**

Lighthouse image with level 5 lossy compression from TID 2013 dataset





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

OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

Image Quality Assessment

IEEE International Conference on Image Processing
25-28 October 2020

Formalize structure of explanations | Define Contrast | Extract Contrastive explanations from Neural Nets | **Applicability in Recognition and Image Quality Assessment**

Lighthouse image with level 5 lossy compression from TID 2013 dataset

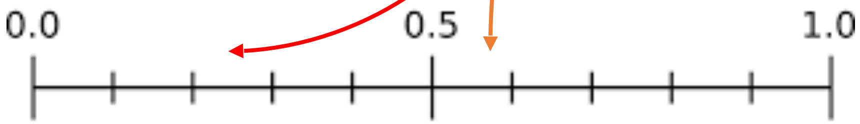
DIQaM : 0.58

Contrastive explanation

Why 0.58, rather than 0.25?

Bad Quality

Good Quality

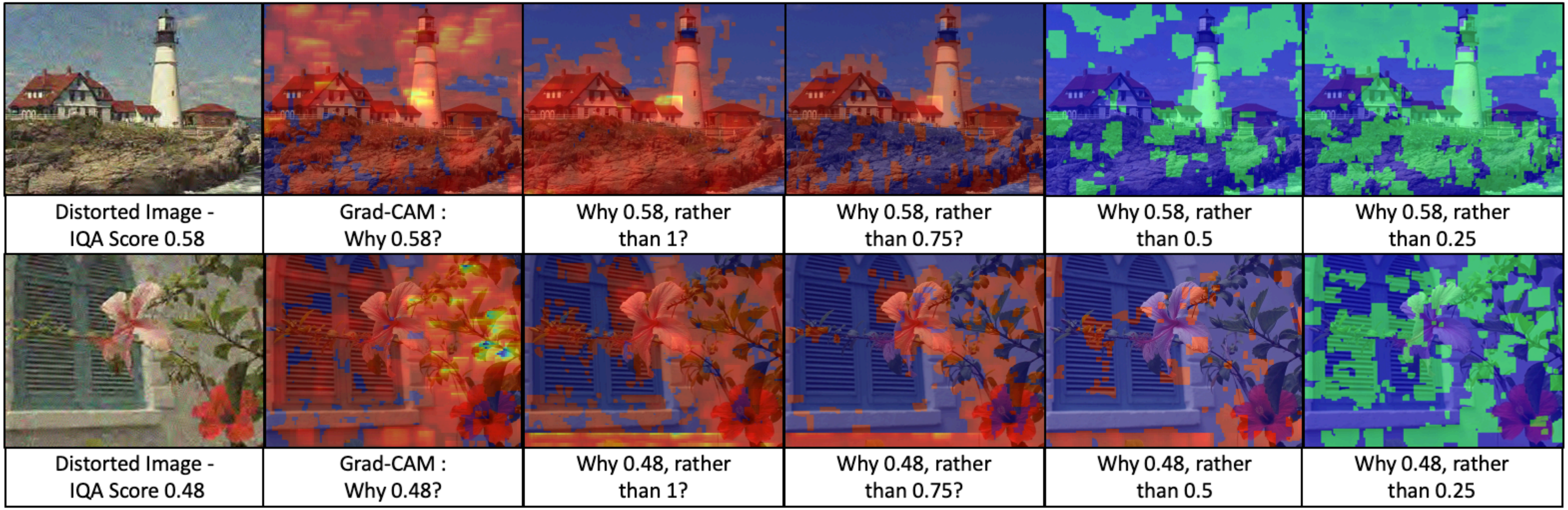


The sky, lighthouse, and cliff merit a quality higher than 0.25

Image Quality Assessment

IEEE Internatic
on Image Proc
25-28 October 2020

Formalize structure of explanations	Define Contrast	Extract Contrastive explanations from Neural Nets	**Applicability in Recognition and Image Quality Assessment**



Contrastive explanations elicit the fine-grained decisions made by the network

OLIVES @GeorgiaTech

Georgia Tech CREATING THE NEXT

Contributions

- Provide structure to existing explanations
- Questioned the nature of existing explanations based on structure
- Defined contrast from a visual and representational perspective
- Extracted contrast in an unsupervised fashion from pre-trained neural network

IEEE International Conference on Image Processing
25-28 October 2020, United Arab Emirates

Thank You

Paper

Codes

Website

Codes : https://github.com/olivesgatech/Contrastive-Explanations

Paper : https://arxiv.org/abs/2008.00178

Lab Website : https://ghassanalregib.info