Explainable AI

Motivation

Understanding how an AI or a machine learning algorithms makes its decision can be beneficial from a variety of aspects:

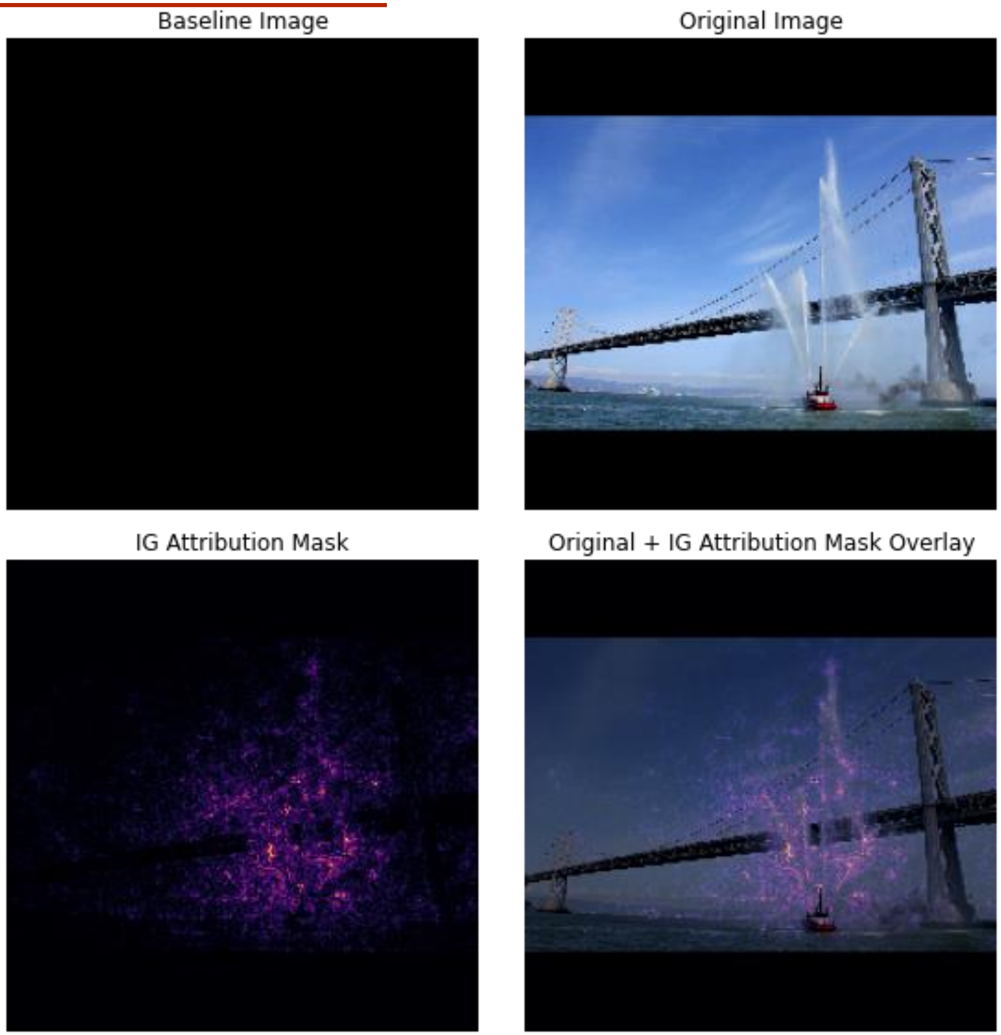
- Algorithm development: find bugs in the code
- Model development: find shortcomings in the training data set
- Determining how reasonable to determination is

Even though machine learning algorithms tend to be a black box, different methods exist to shed some light on their decision making.

Integrated Gradients

This approach is designed for machine learning methods for image recognition. It tries to highlight areas of importance that the algorithm used to base its decision on.

https://www.tensorflow.org/tutorials/interpretability/integrated_gradients



Integrated Gradients

Gradients: tell you which pixels have the steepest local relative to your model's prediction at a given point along your model's prediction function

Gradients only describe *local* changes in your model's prediction function with respect to pixel values and do not fully describe your entire model prediction function. As your model fully "learns" the relationship between the range of an individual pixel and the correct ImageNet class, the gradient for this pixel will *saturate*, meaning become increasingly small and even go to zero.

Integrated Gradients

Computing gradients:

- interpolate small steps along a straight line in the feature space between 0 (a baseline or starting point) and 1 (input pixel's value)
- compute gradients at each step between your model's predictions with respect to each step
- approximate the integral between your baseline and input by accumulating (cumulative average) these local gradients.

Integrated Gradients

Establish a baseline

- A baseline is an input image used as a starting point for calculating feature importance.
- The role of the baseline is to represent the impact of the absence of each pixel on the feature prediction to contrast with its impact of each pixel on the feature prediction when present in the input image
- The choice of the baseline plays a central role in interpreting and visualizing pixel feature importances.
- The baseline image could be an all black or all white image, or a random image.

Integrated Gradients

Computing integrated gradients:

$$IntegratedGradients_i(x)$$
$$= (x_i - x_{i'}) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \frac{dy}{dx} \times (x - x')}{\partial xi}$$

where:

i = feature

x = input

x'= baseline

α= interpolation constant to perturb features by

Integrated Gradients

Computing integrated gradients:

$$IntegratedGradients_i(x)$$

$$= (x_i - x_{i'}) \times \sum_{k=1}^{m} \frac{\partial F(x' + \alpha \frac{dy}{dx} \times (x - x'))}{\partial xi} \times \frac{1}{m}$$

where:

i = feature

x = input

x'= baseline

α= interpolation constant to perturb features by

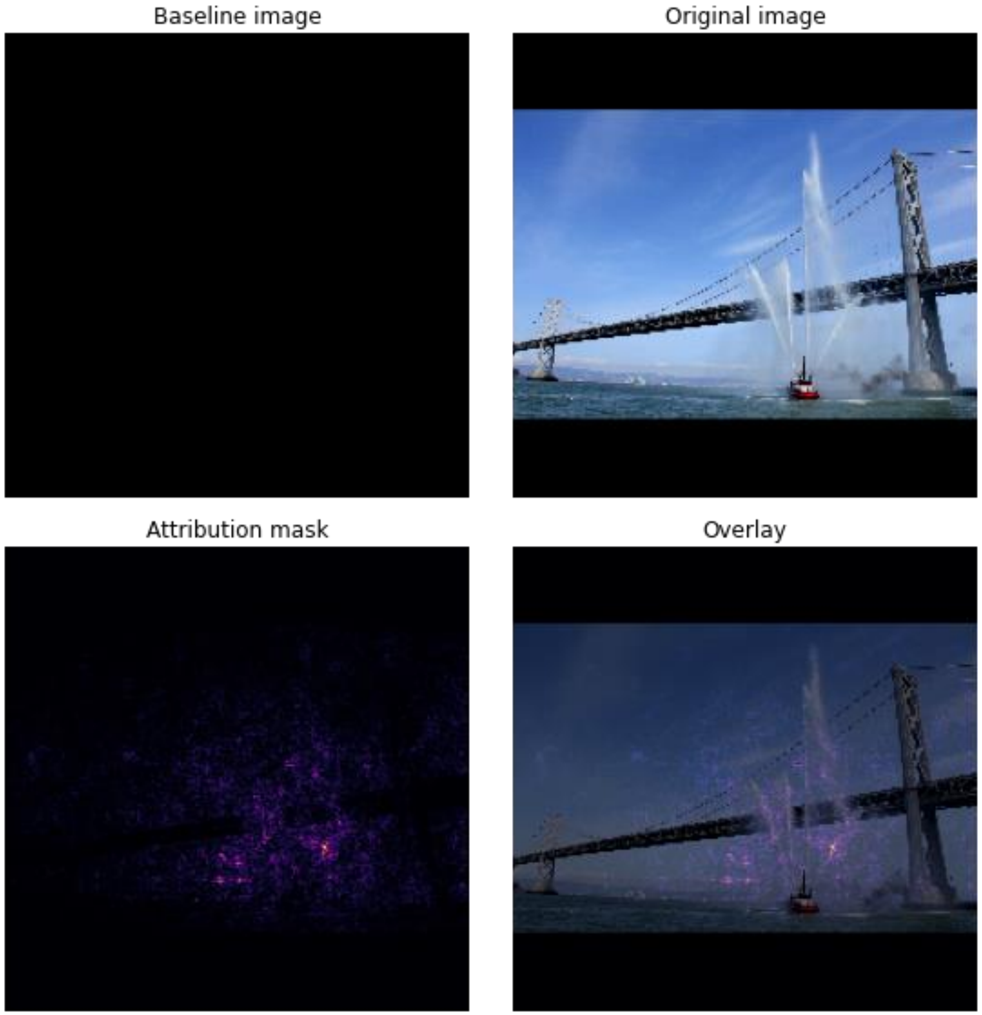
k = scaled feature perturbation constant

m = number of steps in the Riemann sum approximation of the integral

Integrated Gradients

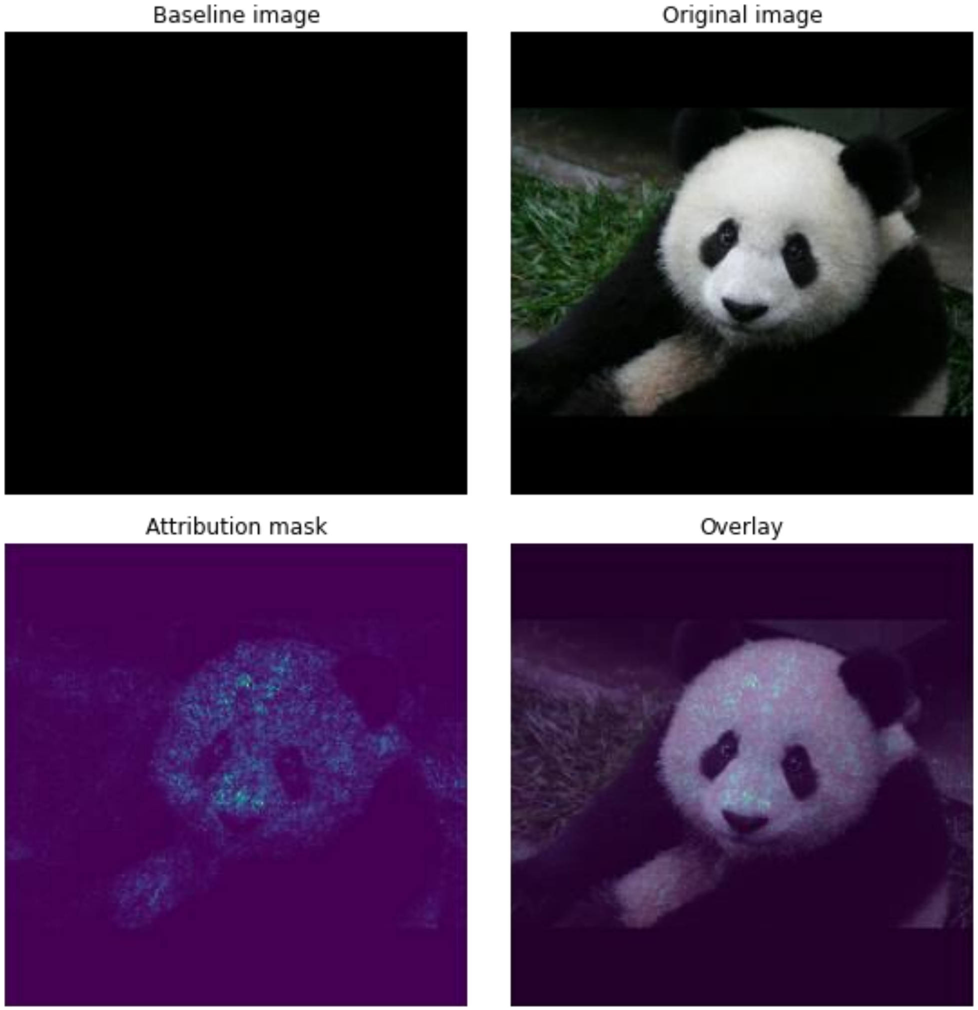
Visualize attributions

To visualize attributions, overlay them on the original image. Compute the sums of the absolute values of the integrated gradients across the color channels to produce an attribution mask.



Integrated Gradients

On the "Giant Panda" image, the attributions highlight the texture, nose, and the fur of the Panda's face.



Integrated Gradients

Uses and limitations

Use cases

- Employing techniques like Integrated Gradients before deploying your model can help you develop intuition for how and why it works. Do the features highlighted by this technique match your intuition? If not, that may be indicative of a bug in your model or dataset, or overfitting.

Limitations

- Integrated Gradients provides feature importances on individual examples, however, it does not provide global feature importances across an entire dataset.
- Integrated Gradients provides individual feature importances, but it does not explain feature interactions and combinations.

The explainable AI toolkit

The explainable AI toolkit (XAITK) was developed by kitware in response to the DARPA’s Explainable Artificial Intelligence (XAI) program in 2015 with three major technical areas:

1. the development of new explainable models and explanation interfaces for generating effective explanations
2. understanding the psychology of explanation by summarizing, extending and applying psychological models of explanation
3. evaluation of the new XAI techniques in two challenge problem areas: data analytics and autonomy.

The explainable AI toolkit

XAITK is a collaboration between kitware and researchers from several universities and other companies.

It provides access to a collection of algorithms available most of which are based on current research. Supported languages vary and include python and JavaScript (Node.js). These algorithms are available publicly through separate repositories.

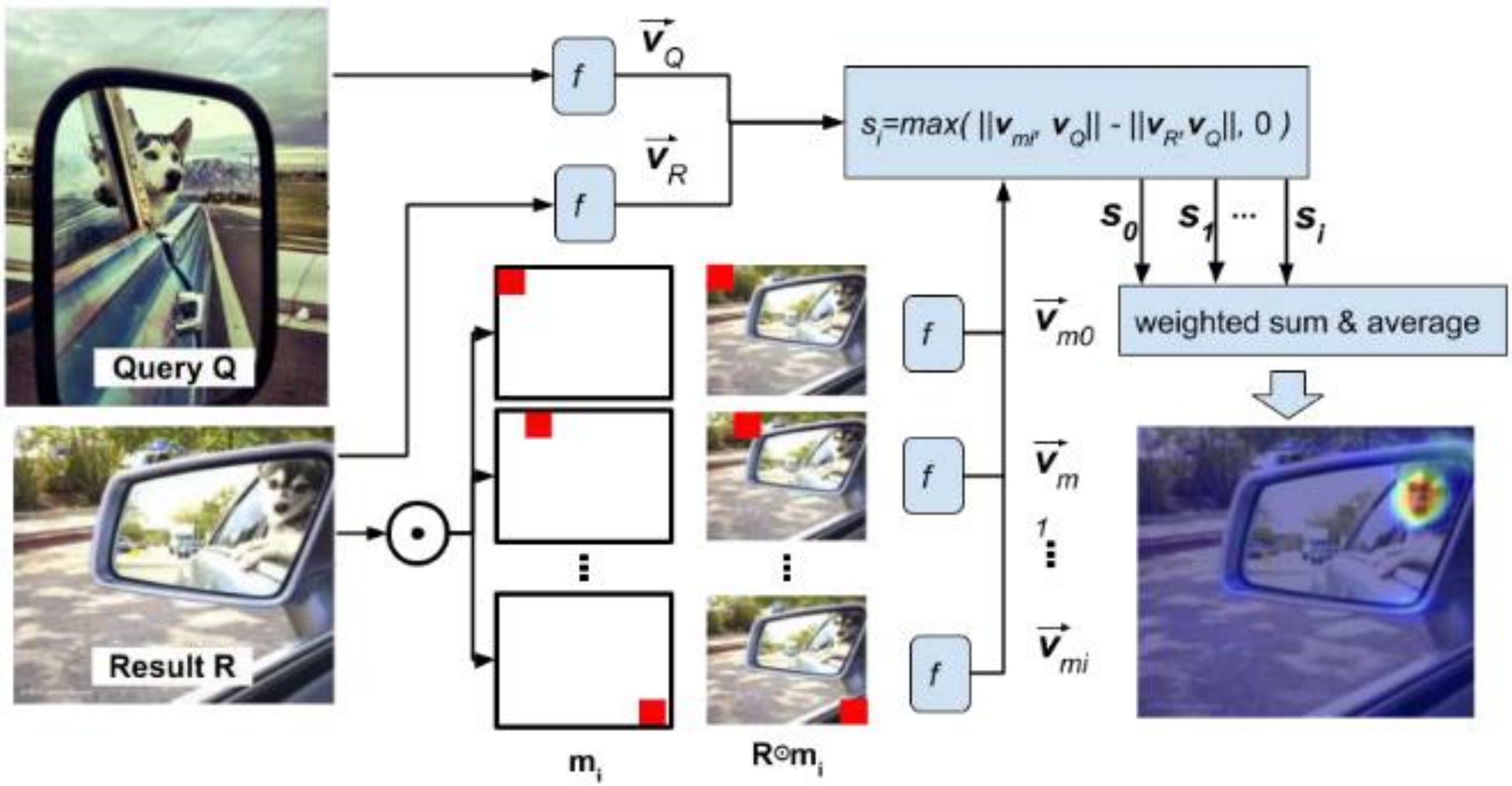
The explainable AI toolkit

Similarity Based Saliency Maps

- a saliency algorithm that compares image descriptors in the embedding space in an attempt to reason about retrieval performance between two reference images.
- Similarity Based Saliency Maps (SBSM) illustrate which areas are used when comparing and ranking match
- Informally, the SBSM is a heatmap: “hotter” regions contribute more to the match score with the query whereas “cooler” areas have less impact

The explainable AI toolkit

Similarity Based Saliency Maps



The explainable AI toolkit

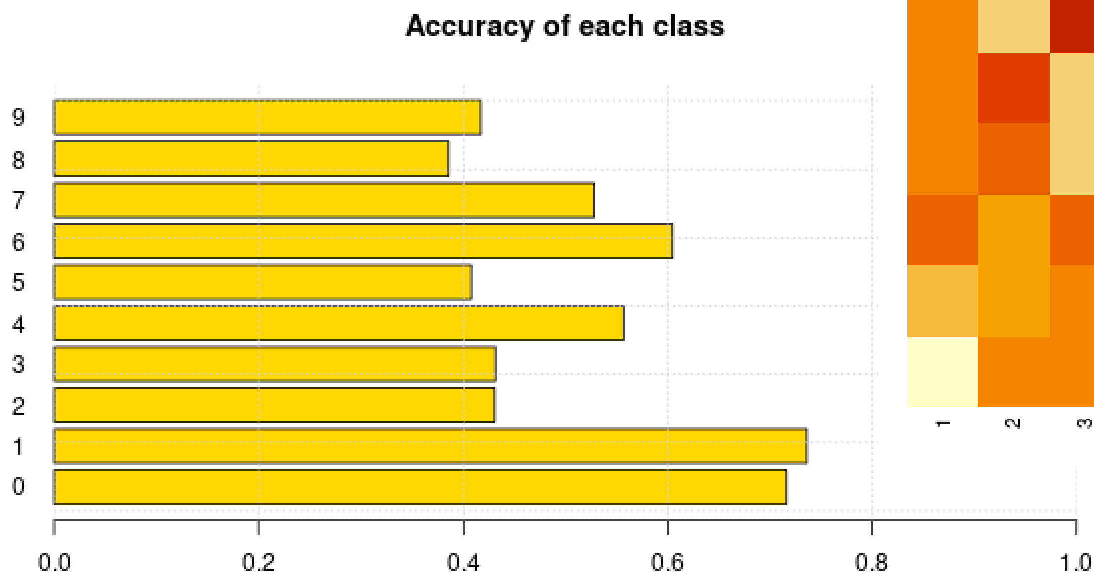
XAI Discovery Platform

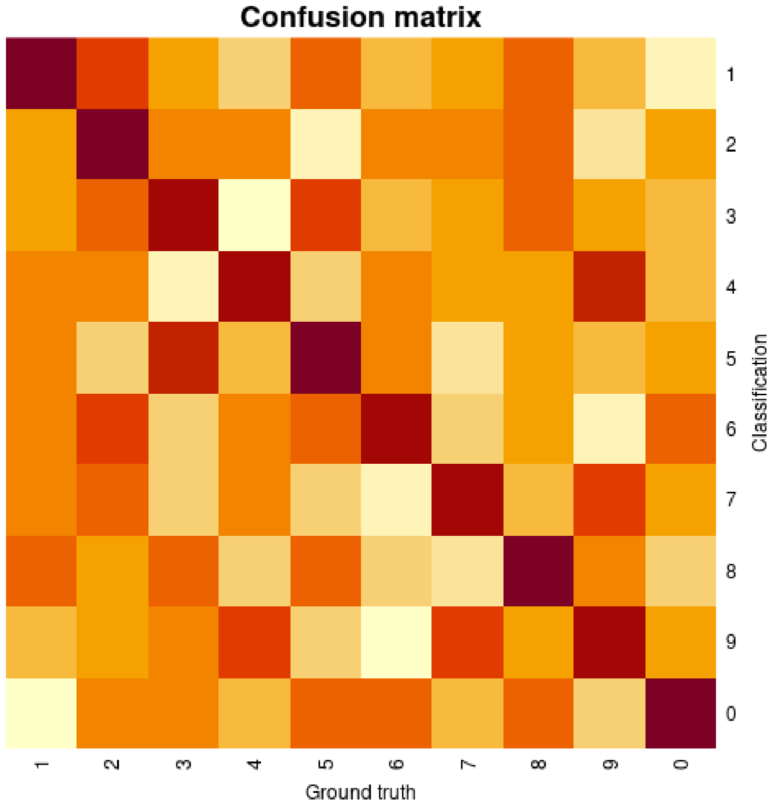
The XAI Discovery Platform provides a customizable interface for exploring image classification data sets. Its goal is to help explore strengths and weaknesses of an image classifier, focusing on consistent errors, and patterns that help predict performance. It does not attempt to provide explanations on its own, but rather helps users understand the things that need to be explained, and to test and compare ideas.

The explainable AI toolkit

XAI Discovery Platform

MNIST Digit Classification

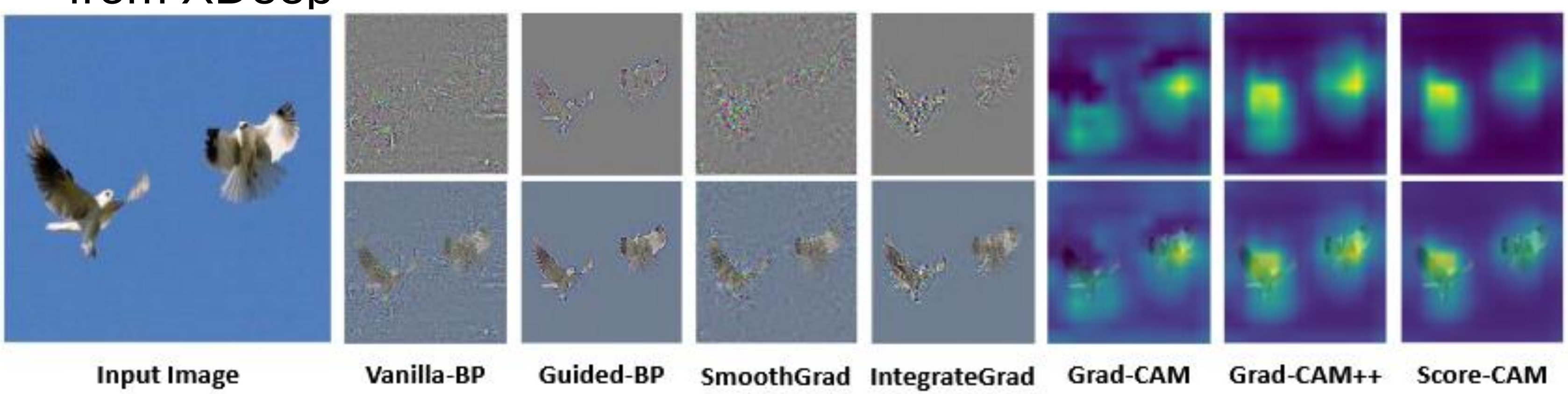




The explainable AI toolkit

XDeep: An Interpretation Tool for Deep Neural Networks

- Interpretation of deep neural networks (DNN)
- integrates a wide range of interpretation algorithms from the state-of-the-arts, covering different types of methodologies, and is capable of providing both local explanation and global explanation for DNN
- Visualizations of local gradient-based interpretation for VGG16 from XDeep



The explainable AI toolkit